Statistics for High-Dimensional Data

Methods, Theory and Applications

Springer

Springer Series in Statistics

Advisors:
P. Bickel, P. Diggle, S. Fienberg, U. Gather,
I. Olkin, S. Zeger

Peter Bühlmann • Sara van de Geer

Statistics for High-Dimensional Data

Methods, Theory and Applications

Springer

Peter Bühlmann
Seminar for Statistics
ETH Zürich
CH-8092 Zürich
Switzerland
buhlmann@stat.math.ethz.ch

Sara van de Geer
Seminar for Statistics
ETH Zürich
CH-8092 Zürich
Switzerland
geer@stat.math.ethz.ch

Cover design: deblik, Berlin

Printed on acid-free paper

To

Anthony and Sigi

and

Tabea, Anna, Sophia, Simon and Lukas

Preface

High-dimensional data are nowadays rule rather than exception in areas like information technology, bioinformatics or astronomy, to name just a few. The word "high-dimensional" refers to the situation where the number of unknown parameters which are to be estimated is one or several orders of magnitude larger than the number of samples in the data. Classical statistical inference cannot be used for high-dimensional problems. For example, least-squares fitting of a linear model having many more unknown parameters than observations and assigning corresponding standard errors and measures of significance is ill-posed. It is rather obvious that without additional assumptions, or say restricting to a certain class of models, high-dimensional statistical inference is impossible. A well-established framework for fitting many parameters is based on assuming structural smoothness, enabling estimation of smooth functions. The last years have witnessed a revolution of methodological, computational and mathematical advances which allow for high-dimensional statistical inference based on assuming certain notions of sparsity. Shifting the focus from smoothness to sparsity constraints, or combining the two, opens the path for many more applications involving complex data. For example, the sparsity assumption that the health status of a person is depending only on a few among several thousands of biomarkers appears much more realistic than considering a model where all the thousands of variables would contribute in a smooth way to the state of health.

This book brings together methodological concepts, computational algorithms, a few applications and mathematical theory for high-dimensional statistics. The mathematical underpinning of methodology and computing has implications on exploring exciting possibilities and understanding fundamental limitations. In this sense, the combination of methodology and theory builds the foundation of the book. We present the methods and their potential for data analysis with a view on the underlying mathematical assumptions and properties and vice-versa, the theoretical derivations are motivated by applicability and implications to real data problems. The mathematical results yield additional insights and allow to categorize different methods and algorithms in terms of what they can achieve and what not. The book

is not meant as an overview of the state-of-the-art, but rather as a selective treatment with emphasis on our own work.

It is possible to read the book with more emphasis on methods and applications or on theory; but of course, one can also focus on all aspects with equal intensity. As such, we hope that the book will be useful and appealing to statisticians, data analysts and other researchers who appreciate the possibilities to learn about methods and algorithms, mathematical theory and the combination of both of them.

This book emerged from a very nice collaboration between the authors. We acknowledge many people who have contributed in various ways to its completion. Wolfgang Härdle proposed to write a book on high-dimensional statistics while hiking in the black forest at Oberwolfach, and we are thankful for it. Alain Hauser, Mohamed Hebiri, Markus Kalisch, Johannes Lederer, Lukas Meier, Nicolai Meinshausen, Patric Müller, Jürg Schelldorfer and Nicolas Städler have contributed with many original ideas and concepts as collaborators of joint research projects or making some thoughtful suggestions for the book. Finally, we would like to express our gratitude to our families for providing a different, interesting, supportive and beautiful environment.

Zürich, December 2010   *Peter Bühlmann and Sara van de Geer*

Contents

Chapter 1
Introduction

Abstract High-dimensional statistics refers to statistical inference when the number of unknown parameters is of much larger order than sample size. We present some introductory motivation and a rough picture about high-dimensional statistics.

1.1 The framework

High-dimensional statistics refers to statistical inference when the number of unknown parameters p is of much larger order than sample size n, that is: $p \gg n$. This encompasses supervised regression and classification models where the number of covariates is of much larger order than n, unsupervised settings such as clustering or graphical modeling with more variables than observations or multiple testing where the number of considered testing hypotheses is larger than sample size. Among the mentioned examples, we discuss in this book regression and classification, graphical modeling and a few aspects of multiple testing.

High-dimensional statistics has relations to other areas. The methodological concepts share some common aspects with nonparametric statistics and machine learning, all of them involving a high degree of complexity making regularization necessary. An early and important book about statistics for complex data is Breiman et al. (1984) with a strong emphasis placed on the CART algorithm. The influential book by Hastie et al. (2001) covers a very broad range of methods and techniques at the interface between statistics and machine learning, also called "statistical learning" and "data mining". From an algorithmic point of view, convex optimization is a key ingredient for regularized likelihood problems which are a central focus of our book, and such optimization arises also in the area of kernel methods from machine learning, cf. Schölkopf and Smola (2002). We include also some deviations where non-convex optimization or iterative algorithms are used. Regarding many aspects of optimization, the book by Bertsekas (1995) has been an important

source for our use and understanding. Furthermore, the mathematical analysis of high-dimensional statistical inference has important connections to approximation theory, cf. Temlyakov (2008), in particular in the context of sparse approximations.

1.2 The possibilities and challenges

A simple yet very useful model for high-dimensional data is a linear model

$$Y_i = \mu + \sum_{j=1}^{p} \beta_j X_i^{(j)} + \varepsilon_i \ (i = 1, \ldots, n), \tag{1.1}$$

with $p \gg n$. It is intuitively clear that the unknown intercept μ and parameter vector $\beta = (\beta_1, \ldots, \beta_p)^T$ can only be estimated reasonably well, based on n observations, if β is sparse in some sense. Sparsity can be quantified in terms of the ℓ_q-norm for $1 \le q \le \infty$, the analogue (which is not a norm) with $0 < q < 1$, or the ℓ_0-analogue (which is not a norm) $\|\beta\|_0^0 = |\{j; \ \beta_j \ne 0\}|$ which counts the number of non-zero entries of the parameter. Note that the notation $\|\beta\|_0^0 = \sum_{j=1}^{p} |\beta_j|^0$ (where $0^0 = 0$) is in analogy to $\|\beta\|_q^q = \sum_{j=1}^{p} |\beta_j|^q$ for $0 < q < \infty$. In contrast to ℓ_0, the ℓ_1-norm $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$ measures sparsity in a different way and has a computational advantage of being a convex function in β.

Roughly speaking, high-dimensional statistical inference is possible, in the sense of leading to reasonable accuracy or asymptotic consistency, if

$$\log(p) \cdot (\text{sparsity}(\beta)) \ll n,$$

depending on how we define sparsity and the setting under consideration.

Early progress of high-dimensional statistical inference has been achieved a while ago: Donoho and Johnstone (1994) present beautiful and clean results for the case of orthogonal design in a linear model where $p = n$. A lot of work has been done to analyze much more general designs in linear or generalized linear models where $p \gg n$, as occurring in many applications nowadays, cf. Donoho and Huo (2001), Donoho and Elad (2003), Fuchs (2004) and many other references given later. We present in this book a detailed treatment for high-dimensional linear and generalized linear models. Much of the methodology and techniques relies on the idea of ℓ_1-penalization for the negative log-likelihood, including versions of such regularization methods. Such ℓ_1-penalization has become tremendously popular due to its computational attractiveness and its statistical properties which reach optimality under certain conditions. Other problems involve more complicated models with e.g. some nonparametric components or some more demanding likelihood functions as occurring in e.g. mixture models. We also describe results and aspects when going beyond generalized linear models.

For sound statistical inference, we would like to quantify uncertainty of estimates or predictions. In particular, if statistical results cannot be validated with a scientific experiment, as for example in bio-medicine where say biomarkers of patients cannot be manipulated, the scientific conclusions hinge on statistical results only. In such cases, high-dimensional statistical inference must be equipped with measures of uncertainty, stability or significance. Our book presents some early ideas in this direction but more refined answers need to be developed in the future.

1.3 About the book

The book is intended for graduate students and researchers in statistics or related fields who are interested in methodological themes and/or detailed mathematical theory for high-dimensional statistics. It is possible to read the methodology and theory parts of the book separately.

Besides methodology and theory, the book touches on applications, as suggested by its title. Regarding the latter, we present illustrations largely without detailed scientific interpretation. Thus, the main emphasis is clearly on methodology and theory. We believe that the theory has its implications on using methods in practice and the book interweaves these aspects. For example, when using the so-called Lasso (ℓ_1-penalization) method for high-dimensional regression, the theory gives some important insights about variable selection and more particularly about false positive and false negative selections.

The book presents important advances in high-dimensional statistical inference. Some of them, like the Lasso and some of its versions, are treated comprehensively with details on practical methodology, computation and mathematical theory. Other themes, like boosting algorithms and graphical modeling with covariance estimation, are discussed from a more practical view point and with less detailed mathematical theory. However, all chapters include a supporting mathematical argumentation.

1.3.1 Organization of the book

The book combines practical methodology and mathematical theory. For the so-called Lasso and group Lasso and versions thereof in linear, generalized linear and additive models, there are separate theory and methods chapters with cross-references to each other.

Other chapters on non-convex negative likelihood problems, stable solutions, p-values for high-dimensional inference, boosting algorithms or graphical modeling

with covariance estimation are presenting in each chapter the methods and some mathematical theory. The last chapter on probability inequalities presents mathematical results and theory which are used at various places in the book. Figure 1.1 gives an overview which parts belong closely to each other.
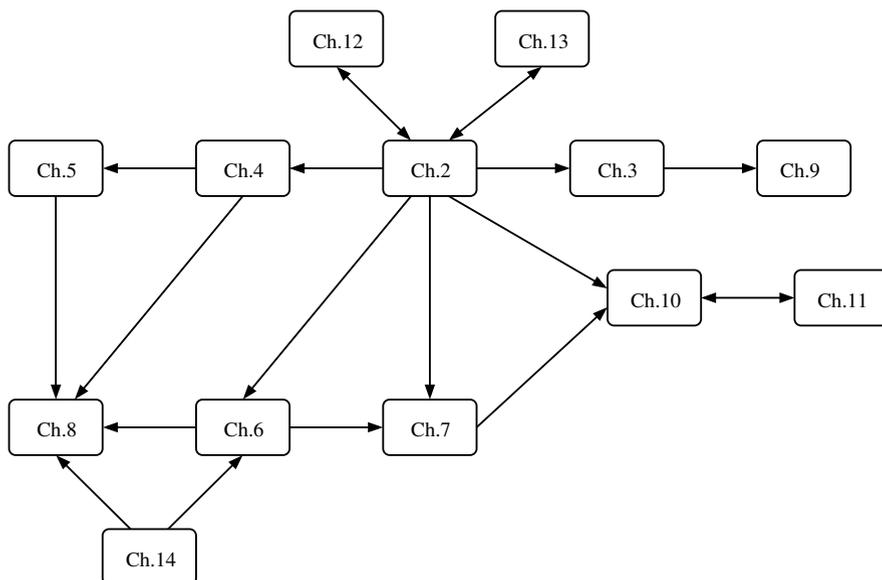
Fig. 1.1 Organization of the book. The arrowheads indicate the directions in which the chapters relate to each other. Chapters 2, 3, 4 and 5 describe statistical methodology and computation, Chapters 6, 7, 8 and 14 present detailed mathematical theory, and the remaining Chapters 9, 10, 11, 12 and 13 each contain methodological, theoretical and computational aspects.

1.4 Some examples

High-dimensional data arises nowadays in a wide variety of applications. The book contains illustrations and applications to problems from biology, a field of our own interest. However, the presented material includes models, methods, algorithms and theory whose relevance is very generic. In particular, we consider high-dimensional linear and generalized linear models as well as the more flexible generalized additive models, and both of them cover a very broad range of applications. Other areas of high-dimensional data applications include text mining, pattern recognition in imaging, astronomy and climate research.

1.4.1 Prediction and biomarker discovery in genomics

In genomics with high-throughput measurements, thousands of variables such as expressions of genes and abundances of proteins can be measured for each person in a (pre-)clinical study. A typical goal is to classify the health status of a person, e.g. healthy or diseased, based on its bio-molecular profile, i.e., the thousands of bio-molecular variables measured for the person.

1.4.1.1 Further biology applications treated in the book

We briefly describe now examples from genomics which will be considered in the book.

We consider motif regression in Chapters 2, 5, 10 and 11. The goal is to infer short DNA-words of approximate length $8 - 16$ base pairs, e.g., "ACCGTTAC", where a certain protein or transcription factor binds to the DNA. We have supervised data available with a continuous response variable Y_i and p-dimensional covariates X_i with continuous values. Thereby Y_i measures e.g. binding intensity of the protein of interest in the ith region of the whole DNA sequence and X_i contains abundance scores of p candidate motifs (or DNA words) in the ith region of the DNA. We relate the response Y_i and the covariates X_i with a linear model as in (1.1) (or an additive model as in Chapter 5), where $X_i^{(j)}$ denotes the abundance score of candidate word j in DNA region i. The task is to infer which candidate words are relevant for explaining the response Y. Statistically, we want to find the variables $X^{(j)}$ whose corresponding regression coefficients β_j are substantial in absolute value or significantly different from zero. That is, motif regression is concerned about variable or feature selection. The typical sizes for motif regression are $n \approx 50 - 1'000$ and $p \approx 100 - 2'000$ and hence, the number of variables or the dimensionality p is about of the same order as sample size n. In this sense, motif regression is a fairly but not truly high-dimensional problem.

Another example is the prediction of DNA splice sites which are the regions between coding and non-coding DNA segments. The problem is discussed in Chapter 4. We have binary response variables $Y_i \in \{0, 1\}$, encoding whether there is a splice site or not at a certain position i of the DNA sequence, and categorical p-dimensional covariates $X_i \in \{A, C, G, T\}^p$ with four categories corresponding to the letters of the DNA alphabet. The p categorical variables correspond to p neighboring values of a certain position i of the DNA sequence: for example, 3 positions to the left and 4 positions to the right from i, corresponding to $p = 7$ and e.g. $X_i = (A, A, T, G, G, C, G)$. We model the data as a binary logistic regression whose covariates consist of 7 factors each having 4 levels. The primary goal here is prediction or classification of a new, unknown splice site. The typical sizes for DNA splice site prediction is $n \approx 10'000 - 50'000$ and $p \approx 5 - 20$. When allowing for all interactions, the num-

ber of parameters in the logistic model is 4^p which can be huge in comparison to n, e.g., $4^{10} \approx 1.05 \cdot 10^6$. Depending on how many interactions we allow, the problem may involve a million unknown parameters which is of larger order than the typical sample size.

In Chapters 9 and 10 we illustrate some methods for a problem about riboflavin production with bacillus subtilis. The data consists of continuous response variables Y_i, measuring the log-concentration of riboflavin, and p-dimensional covariates X_i containing the log-expressions from essentially all genes from bacillus subtilis, for the ith individual. The goal is primarily variable selection to increase understanding which genes are relevant for the riboflavin production rate. A linear model as in (1.1) is often a reasonable approximation but we will also discuss in Chapter 9 a mixture model which is an attempt to model inhomogeneity of the data. The size of the data is about $n \approx 70 - 150$ and $p = 4088$, and hence it is a real high-dimensional problem.

Finally, we consider in Chapter 13 an unsupervised problem about genes in two biosynthesis pathways in arabidopsis thaliana. The data consists of continuous gene expressions from 39 genes for $n = 118$ samples of different arabidopsis plants. We illustrate covariance estimation and aspects of graphical modeling which involve $39 \cdot 40/2 = 780$ covariance parameters, i.e., more parameters than sample size.

Chapter 2
Lasso for linear models

Abstract The Lasso, proposed by Tibshirani (1996), is an acronym for Least Absolute Shrinkage and Selection Operator. Among the main reasons why it has become very popular for high-dimensional estimation problems are its statistical accuracy for prediction and variable selection coupled with its computational feasibility. Furthermore, since the Lasso is a penalized likelihood approach, the method is rather general and can be used in a broad variety of models. In the simple case of a linear model with orthonormal design, the Lasso equals the soft thresholding estimator introduced and analyzed by Donoho and Johnstone (1994). The Lasso for linear models is the core example to develop the methodology for ℓ_1-penalization in high-dimensional settings. We discuss in this chapter some fundamental methodological and computational aspects of the Lasso. We also present the adaptive Lasso, an important two-stage procedure which addresses some bias problems of the Lasso. The methodological steps are supported by describing various theoretical results which will be fully developed in Chapters 6 and 7.

2.1 Organization of the chapter

We present in this chapter the Lasso for linear models from a methodological point of view. Theoretical results are loosely described to support methodology and practical steps for analyzing high-dimensional data. After an introduction in Section 2.2 with the definition of the Lasso for linear models, we focus in Section 2.4 on prediction of a new response when given a new covariate. Afterwards, we discuss in Section 2.5 the Lasso for estimating the regression coefficients which is rather different from prediction. An important implication will be that under certain conditions, the Lasso will have the screening property for variable selection saying that it will include all relevant variables whose regression coefficients are sufficiently large (besides potentially false positive selections). In Section 2.6 we discuss the more ambitious goal of variable selection in terms of exact recovery of all the rele-

vant variables. Some of the drawbacks of the Lasso can be addressed by two-stage or multi-stage procedures. Among them are the adaptive Lasso (Zou, 2006) and the relaxed Lasso (Meinshausen, 2007), discussed in Sections 2.8 and 2.10, respectively. Finally, we present concepts and ideas for computational algorithms in Section 2.12.

2.2 Introduction and preliminaries

We consider here the setting where the observed data are realizations of

$$(X_1, Y_1), \ldots, (X_n, Y_n)$$

with p-dimensional covariates $X_i \in \mathscr{X} \subset \mathbb{R}^p$ and univariate response variables $Y_i \in \mathscr{Y} \subset \mathbb{R}$. The covariates are either deterministic fixed values or random variables: regarding the methodology, there is no difference between these two cases. Typically, we assume that the samples are independent but the generalization to stationary processes poses no essential methodological or theoretical problems.

Modeling high-dimensional data is challenging. For a continuous response variable $Y \in \mathbb{R}$, a simple yet very useful approach is given by a linear model

$$Y_i = \sum_{j=1}^{p} \beta_j X_i^{(j)} + \varepsilon_i \ (i = 1, \ldots, n), \tag{2.1}$$

where $\varepsilon_1, \ldots, \varepsilon_n$ i.i.d., independent of $\{X_i; \ i = 1, \ldots, n\}$ and with $\mathbb{E}[\varepsilon_i] = 0$.

For simplicity and without loss of generality, we usually assume that the intercept is zero and that all covariates are centered and measured on the same scale. Both of these assumptions can be approximately achieved by empirical mean centering and scaling with the standard deviation, and the standardized data then satisfies $\overline{Y} = n^{-1} \sum_{i=1}^{n} Y_i = 0$ and $\hat{\sigma}_j^2 := n^{-1} \sum_{i=1}^{n} (X_i^{(j)} - \overline{X}^{(j)})^2 = 1$ for all j. The only unusual aspect of the linear model in (2.1) is the fact that $p \gg n$.

We often use for (2.1) the matrix- and vector-notation

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

with response vector $\mathbf{Y}_{n \times 1}$, design matrix $\mathbf{X}_{n \times p}$, parameter vector $\beta_{p \times 1}$ and error vector $\varepsilon_{n \times 1}$. If the model is correct, we denote the true underlying parameter by β^0. We denote the best approximating parameter, in a sense to be specified, by β^*: this case will be discussed from a theory point of view in Chapter 6 in Section 6.2.3.

2.2.1 The Lasso estimator

If $p > n$, the ordinary least squares estimator is not unique and will heavily overfit the data. Thus, a form of complexity regularization will be necessary. We focus here on regularization with the ℓ_1-penalty. The parameters in model (2.1) are estimated with the Lasso (Tibshirani, 1996):

$$\hat{\beta}(\lambda) = \arg\min_{\beta} \left(\|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + \lambda \|\beta\|_1 \right), \tag{2.2}$$

where $\|\mathbf{Y} - \mathbf{X}\beta\|_2^2 = \sum_{i=1}^n (Y_i - (\mathbf{X}\beta)_i)^2$, $\|\beta\|_1 = \sum_{j=1}^p |\beta_j|$ and where $\lambda \geq 0$ is a penalty parameter. The estimator has the property that it does variable selection in the sense that $\hat{\beta}_j(\lambda) = 0$ for some j's (depending on the choice of λ) and $\hat{\beta}_j(\lambda)$ can be thought as a shrunken least squares estimator; hence, the name Least Absolute Shrinkage and Selection Operator (LASSO). An intuitive explanation for the variable selection property is given below.

The optimization in (2.2) is convex, enabling efficient computation of the estimator, see Section 2.12. In addition, the optimization problem in (2.2) is equivalent to

$$\hat{\beta}_{\text{primal}}(R) = \arg\min_{\beta; \|\beta\|_1 \leq R} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n, \tag{2.3}$$

with a one-to-one correspondence between λ in (2.2) and R in (2.3), depending on the data $(X_1, Y_1), \ldots, (X_n, Y_n)$. Such an equivalence holds since $\|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n$ is convex in β with convex constraint $\|\beta\|_1 \leq R$, see for example Bertsekas (1995, Ch. 5.3).

Because of the ℓ_1-geometry, the Lasso is performing variable selection in the sense that an estimated component can be exactly zero. To see this, we consider the representation in (2.3) and Figure 2.1: the residual sum of squares reaches a minimal value (for certain constellations of the data) if its contour lines hit the ℓ_1-ball in its corner which corresponds to the first component $\hat{\beta}_{\text{primal},1}$ being equal to zero. Figure 2.1 indicates that such a phenomenon does not occur with say Ridge regression,

$$\hat{\beta}_{\text{Ridge}}(\lambda) = \arg\min_{\beta} \left(\|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + \lambda \|\beta\|_2^2 \right),$$

with its equivalent primal solution

$$\hat{\beta}_{\text{Ridge};\text{primal}}(R) = \arg\min_{\beta; \|\beta\|_2 \leq R} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n, \tag{2.4}$$

with again a data-dependent one-to-one correspondence between λ and R.
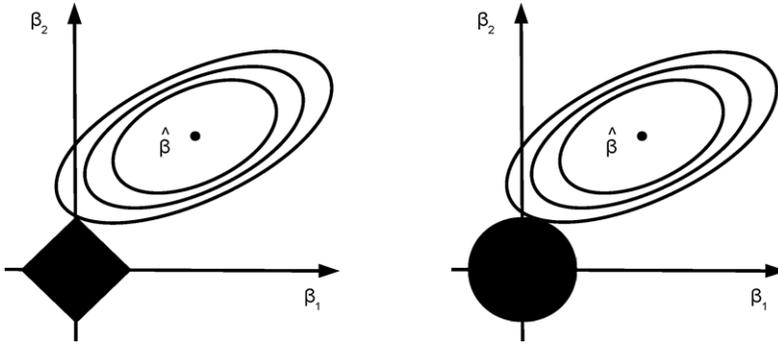
Fig. 2.1 Left: Contour lines of residual sum of squares, with $\hat{\beta}$ being the least squares estimator, and ℓ_1-ball corresponding to the Lasso problem in (2.3). Right: Analogous to left panel but with ℓ_2-ball corresponding to Ridge regression in (2.4). The figure is as in Tibshirani (1996).

2.2.1.1 Estimation of the error variance

The estimator in (2.2) does not directly provide an estimate for the error variance σ^2. One can construct an estimator using the residual sum of squares and the degrees of freedom of the Lasso (Section 2.11). Alternatively, and rigorously developed, we can estimate β and σ^2 simultaneously using a reparametrization: this is discussed in detail in Section 9.2.2.1 from Chapter 9.

2.3 Orthonormal design

It is instructive to consider the orthonormal design where $p = n$ and the design matrix satisfies $n^{-1}\mathbf{X}^T\mathbf{X} = I_{p \times p}$. In this case, the Lasso estimator is the soft-threshold estimator

$$\hat{\beta}_j(\lambda) = \text{sign}(Z_j)(|Z_j| - \lambda/2)_+, \; Z_j = (\mathbf{X}^T\mathbf{Y})_j/n \;\; (j = 1, \ldots, p = n), \quad (2.5)$$

where $(x)_+ = \max(x, 0)$ denotes the positive part and Z_j equals the ordinary least squares estimator for β_j. This follows from the general characterization in Lemma 2.1 below and we leave a direct derivation (without using Lemma 2.1) as Problem 2.1. Thus, the estimator can be written as

$$\hat{\beta}_j(\lambda) = g_{\text{soft}, \lambda/2}(Z_j),$$

where $g_{\text{soft}, \lambda}(z) = \text{sign}(z)(|z| - \lambda)_+$, is the soft-threshold function depicted in Figure 2.2. There, we also show for comparison the hard-threshold and the adaptive Lasso
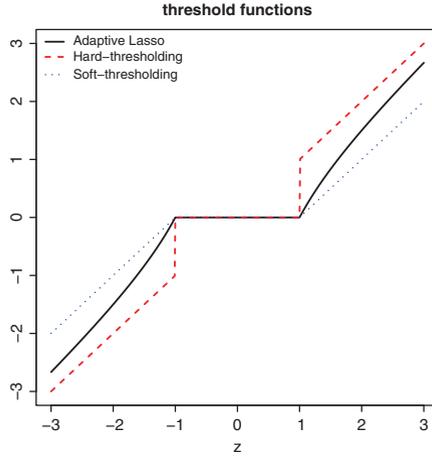
Fig. 2.2 Various threshold functions $g(\cdot)$ for orthonormal design: soft-threshold (dashed line), hard-threshold (dotted line), Adaptive Lasso (solid line). The estimators are of the form $\hat{\beta}_j = g(Z_j)$ with Z_j as in (2.5).

estimator (see Section 2.8) for β_j defined by

$$\hat{\beta}_{\text{hard}, j}(\lambda) = g_{\text{hard}, \lambda/2}(Z_j), \quad g_{\text{hard}, \lambda}(z) = z\mathbb{1}(|z| \le \lambda),$$
$$\hat{\beta}_{\text{adapt}, j}(\lambda) = g_{\text{adapt}, \lambda/2}(Z_j), \quad g_{\text{adapt}, \lambda}(z) = z(1 - \lambda/|z|^2)_+ = \text{sign}(z)(|z| - \lambda/|z|)_+.$$

2.4 Prediction

We refer to prediction whenever the goal is estimation of the regression function $\mathbb{E}[Y|X = x] = \sum_{j=1}^{p} \beta_j x^{(j)}$ in model (2.1). This is also the relevant quantity for predicting a new response.

2.4.1 Practical aspects about the Lasso for prediction

From a practical perspective, prediction with the Lasso is straightforward and easy. One often uses a cross-validation (CV) scheme, e.g., 10-fold CV, to select a reasonable tuning parameter λ minimizing the cross-validated squared error risk. In addition, we can validate the accuracy of the performance by using again some cross-validation scheme. Regarding the latter, we should cross-validate the whole procedure including the selection of the tuning parameter λ. In particular, by comparing the cross-validated risk, we can roughly see whether the Lasso yields a performance which is better, equal or worse than another prediction algorithm. However, when aiming for more formal conclusions, it is not straightforward to test statistically whether the performances of two prediction algorithms are significantly different, see for example van de Wiel et al. (2009).

2.4.1.1 Binary classification of lymph node status using gene expressions

We consider a classification problem involving a binary response variable $Y \in \{0,1\}$, describing the lymph node status of a cancer patient, and we have a covariate with $p = 7129$ gene expression measurements. There are $n = 49$ breast cancer tumor samples. The data is taken from West et al. (2001). It is known that this is a difficult, high noise classification problem. The best methods achieve a cross-validated misclassification error of about 20%.

Despite the binary nature of the classification problem, we can use the Lasso as in (2.2) which yields an estimate of the conditional class probability $f(x) = \mathbf{P}[Y = 1|X = x] = \mathbb{E}[Y|X = x]$:

$$\hat{f}_\lambda(x) = x\hat{\beta}(\lambda) = \sum_{j=1}^{p} \hat{\beta}_j(\lambda)x^{(j)}.$$

Of course, we could use the Lasso also for logistic regression as described later in Chapter 3. In either case, having an estimate of the conditional class probability, denoted here by $\hat{f}_\lambda(\cdot)$, we classify as follows:

$$\mathscr{C}_\lambda(x) = \begin{cases} 1 & \hat{f}_\lambda(x) > 1/2, \\ 0 & \hat{f}_\lambda(x) \le 1/2. \end{cases}$$

For comparison, we consider a forward variable selection method in penalized linear logistic regression with ℓ_2-norm (Ridge-type) penalty. The optimal regularization parameter, for Lasso and forward penalized logistic regression, is chosen by 10-fold cross-validation. For evaluating the performance of the tuned algorithms, we use a cross-validation scheme for estimating the test-set misclassification error. We randomly divide the sample into 2/3 training- and 1/3 test-data and we repeat this

100 times: the average test-set misclassification error is reported in Table 2.1. Note that we run a double cross-validation: one inner level for choosing the regularization parameter and one outer level for assessing the performance of the algorithm.

Table 2.1 illustrates that the forward selection approach yields, in this example, much poorer performance than the Lasso. Forward selection methods tend to be

Lasso	forw. penalized logist. regr.
21.1%	35.25%

Table 2.1 Misclassification test set error using cross-validation

unstable (Breiman, 1996): they are of a very greedy nature striving for maximal improvement of the objective function (e.g. residual sum of squares) in every step.

Finally, we report that the Lasso selected on cross-validation average 13.12 out of $p = 7129$ variables (genes). Thus, the fitted linear model is very sparse with respect to the number of selected variables.

2.4.2 Some results from asymptotic theory

We now describe results which are developed and described in detail in Chapter 6. For simplicity, we take here an asymptotic point of view instead of finite sample results in Chapter 6. To capture high-dimensional scenarios, the asymptotics is with respect to a triangular array of observations:

$$Y_{n;i} = \sum_{j=1}^{p_n} \beta^0_{n;j} X^{(j)}_{n;i} + \varepsilon_{n;i}, \ i = 1, \ldots, n; \ n = 1, 2, \ldots \tag{2.6}$$

Thereby, we allow that $p = p_n \gg n$. The assumptions about $\varepsilon_{n;i}$ are as in the linear model in (2.1). A consistency result requires a sparsity assumption of the form

$$\|\beta^0\|_1 = \|\beta^0_n\|_1 = o\left(\sqrt{\frac{n}{\log(p)}}\right),$$

see Corollary 6.1 in Chapter 6. Assuming further mild regularity conditions on the error distribution, the following holds: for a suitable range of $\lambda = \lambda_n \asymp \sqrt{\log(p)/n}$, the Lasso is consistent for estimating the underlying regression function:

$$(\hat{\beta}(\lambda) - \beta^0)^T \Sigma_X (\hat{\beta}(\lambda) - \beta^0) = o_P(1) \ (n \to \infty), \tag{2.7}$$

where Σ_X equals $n^{-1} \mathbf{X}^T \mathbf{X}$ in case of a fixed design. In the case of random design, Σ_X is the covariance of the covariate X, and then (2.7) holds assuming $\|\beta^0\|_1 =$

$o((n/\log(n))^{1/4})$, see Greenshtein and Ritov (2004). This condition on the growth of $\|\beta^0\|_1$ is relaxed in Bartlett et al. (2009). Note that the left hand side in (2.7) can be written as the average squared error loss:

$$\|\mathbf{X}(\hat{\beta} - \beta^0)\|_2^2/n \text{ for fixed design,}$$
$$\mathbb{E}[(X_{\text{new}}(\hat{\beta}(\lambda) - \beta^0))^2] \text{ for random design,}$$

where $\mathbb{E}$ is with respect to the new test observation X_{new} (a $1 \times p$ vector) and $X(\hat{\beta}(\lambda) - \beta^0)$ is the difference between the estimated and true regression function $\hat{f}(X) - f^0(X)$.

Under certain compatibility (or restricted eigenvalue) conditions on the design $\mathbf{X}$, for λ in a suitable range of the order $\sqrt{\log(p)/n}$, one can show a so-called oracle inequality for fixed design,

$$\mathbb{E}[\|\mathbf{X}(\hat{\beta}(\lambda) - \beta^0)\|_2^2/n] = O\left(\frac{s_0 \log(p)}{n\phi^2}\right), \tag{2.8}$$

where $s_0 = \text{card}(S_0) = |S_0|$ and $S_0 = \{j; \beta_j^0 \neq 0\}$ is the active set of variables, and ϕ^2 is the so-called compatibility constant or restricted eigenvalue which is a number depending on the compatibility between the design and the ℓ_1-norm of the regression coefficient. At best, it is bounded below by a positive constant. See Corollary 6.2. An analogous result holds for random design as well. This means that, up to the $\log(p)$-term (and the compatibility constant ϕ^2), the mean-squared prediction error is of the same order as if one knew a-priori which of the covariates are relevant and using ordinary least squares estimation based on the true, relevant s_0 variables only. The rate in (2.8) is optimal, up to the factor $\log(p)$ and the inverse compatibility constant $1/\phi^2$.

2.5 Variable screening and $\|\hat{\beta} - \beta^0\|_q$-norms

We consider now the estimation accuracy for the parameter β, a different task than prediction. Under compatibility assumptions on the design $\mathbf{X}$ and on the sparsity $s_0 = |S_0|$ in a linear model, it can be shown that for λ in a suitable range of order $\lambda \asymp \sqrt{\log(p)/n}$,

$$\|\hat{\beta}(\lambda) - \beta^0\|_q \to 0 \text{ in probability } (n \to \infty), \tag{2.9}$$

where $q \in \{1, 2\}$ and $\|\beta\|_q = (\sum_j |\beta_j|^q)^{1/q}$. The asymptotic framework is again with respect to the triangular array described in (2.6). The derivation of such results is given in Chapter 6, Section 6.8.

The result in (2.9) has fairly direct and interesting implications in terms of variable screening. Consider the active set of variables

$$S_0 = \{j;\ \beta_j^0 \neq 0.\ j = 1, \ldots, p\}$$

which contains all covariates with non-zero corresponding regression coefficients. Note that in a setting as in (2.6), the active set $S_0 = S_{0;n}$ is allowed to depend on n. Since the Lasso estimator in (2.2) is selecting some variables, in the sense that some of the coefficients are exactly zero, we use it as screening estimator:

$$\hat{S}(\lambda) = \{j;\ \hat{\beta}_j(\lambda) \neq 0,\ j = 1, \ldots, p\}. \tag{2.10}$$

It is worth pointing out that no significance testing is involved. Furthermore, the variables with corresponding non-zero coefficients remain the same across different solutions $\hat{\beta}(\lambda)$ of the optimization in (2.2), see Lemma 2.1. Note that different solutions occur if the optimization is not strictly convex as in the case where $p > n$.

An important characterization of the solution $\hat{\beta}(\lambda)$ in (2.2) can be derived from the Karush-Kuhn-Tucker conditions (and some additional reasoning regarding uniqueness of zeroes).

Lemma 2.1. *Denote the gradient of $n^{-1}\|\mathbf{Y} - \mathbf{X}\beta\|_2^2$ by $G(\beta) = -2\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\beta)/n$. Then a necessary and sufficient condition for $\hat{\beta}$ to be a solution of (2.2) is:*

$$G_j(\hat{\beta}) = -\text{sign}(\hat{\beta}_j)\lambda \text{ if } \hat{\beta}_j \neq 0,$$
$$|G_j(\hat{\beta})| \leq \lambda \text{ if } \hat{\beta}_j = 0.$$

Moreover, if the solution of (2.2) is not unique (e.g. if $p > n$) and $G_j(\hat{\beta}) < \lambda$ for some solution $\hat{\beta}$, then $\hat{\beta}_j = 0$ for all solutions of (2.2).

Proof. For the first statements regarding a necessary and sufficient characterization of the solution, we invoke subdifferential calculus (Bertsekas, 1995), see also Problem 4.2 in Chapter 4. Denote the criterion function by

$$Q_\lambda(\beta) = \|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + \lambda\|\beta\|_1.$$

For a minimizer $\hat{\beta}(\lambda)$ of $Q_\lambda(\cdot)$ it is necessary and sufficient that the subdifferential at $\hat{\beta}(\lambda)$ is zero. If the jth component $\hat{\beta}_j(\lambda) \neq 0$, this means that the ordinary first derivative at $\hat{\beta}(\lambda)$ has to be zero:

$$\frac{\partial Q_\lambda(\beta)}{\partial \beta_j}\Big|_{\beta = \hat{\beta}(\lambda)} = -2\mathbf{X}_j^T(\mathbf{Y} - \mathbf{X}\beta)/n + \lambda\,\text{sign}(\beta_j)\big|_{\beta = \hat{\beta}(\lambda)} = 0,$$

where $\mathbf{X}_j$ is the $n \times 1$ vector $(X_1^{(j)}, \ldots, X_n^{(j)})^T$. Of course, this is equivalent to

$$G_j(\hat{\beta}(\lambda)) = -2\mathbf{X}_j^T(\mathbf{Y} - \mathbf{X}\hat{\beta}(\lambda))/n = -\lambda\,\text{sign}(\hat{\beta}_j(\lambda)) \text{ if } \hat{\beta}_j(\lambda) \neq 0.$$

On the other hand, if $\hat{\beta}_j(\lambda) = 0$, the subdifferential at $\hat{\beta}(\lambda)$ has to include the zero element (Bertsekas, 1995). That is:

$$\text{if } \hat{\beta}_j(\lambda) = 0: \ G_j(\hat{\beta}(\lambda)) + \lambda e = 0 \text{ for some } e \in [-1, 1].$$

But this is equivalent to

$$|G_j(\hat{\beta}(\lambda))| \leq \lambda \text{ if } \hat{\beta}_j(\lambda) = 0.$$

And this is the second statement in the characterization of the solution of $\hat{\beta}(\lambda)$.

Regarding uniqueness of the zeroes among different solutions we argue as follows. Assume that there exist two solutions $\hat{\beta}^{(1)}$ and $\hat{\beta}^{(2)}$ such that for a component j we have $\hat{\beta}_j^{(1)} = 0$ with $|G_j(\hat{\beta}^{(1)})| < \lambda$ but $\hat{\beta}_j^{(2)} \neq 0$. Because the set of all solutions is convex,

$$\hat{\beta}_\rho = (1 - \rho)\hat{\beta}^{(1)} + \rho\hat{\beta}^{(2)}$$

is also a minimizer for all $\rho \in [0, 1]$. By assumption and for $0 < \rho < 1$, $\hat{\beta}_{\rho,j} \neq 0$ and hence, by the first statement from the KKT conditions, $|G_j(\hat{\beta}_\rho)| = \lambda$ for all $\rho \in (0, 1)$. Hence, it holds for $g(\rho) = |G_j(\hat{\beta}_\rho)|$ that $g(0) < \lambda$ and $g(\rho) = \lambda$ for all $\rho \in (0, 1)$. But this is a contradiction to the fact that $g(\cdot)$ is continuous. Hence, a non-active (i.e. zero) component j with $|G_j(\hat{\beta})| < \lambda$ can not be active (i.e. non-zero) in any other solution.   □

Ideally, we would like to infer the active set S_0 from data. We will explain in Section 2.6 that the Lasso as used in (2.10) requires fairly strong conditions on the design matrix **X** (Theorem 7.1 in Chapter 7 gives the precise statement.)

A less ambitious but still relevant goal in practice is to find at least the covariates whose corresponding absolute values of the regression coefficients $|\beta_j|$ are substantial. More formally, for some $C > 0$, define the substantial (relevant) covariates as

$$S_0^{\text{relevant}(C)} = \{j; \ |\beta_j^0| \geq C, \ j = 1, \ldots, p\}.$$

Using the result in (2.9), one can show (Problem 2.2) that

$$\text{for any fixed } 0 < C < \infty: \ \mathbf{P}[\hat{S}(\lambda) \supset S_0^{\text{relevant}(C)}] \to 1 \ (n \to \infty). \quad (2.11)$$

This result can be generalized as follows. Assume that

$$\|\hat{\beta}_n(\lambda_n) - \beta^0\|_1 \leq a_n \text{ with high probability.} \quad (2.12)$$

We note that under compatibility conditions on the design matrix **X**, with λ_n in the range of order $\sqrt{\log(p_n)/n}$, it holds that $a_n = O(s_0\sqrt{\log(p_n)/n})$ with $s_0 = |S_0|$. The details are described in Theorem 6.1 in Chapter 6. Then,

$$\text{for } C_n > a_n: \ \text{with high probability } \hat{S}_n(\lambda_n) \supset S_0^{\text{relevant}(C_n)}. \quad (2.13)$$

The proof is elementary and we leave it as Problem 2.3. (Actually, under stronger restricted eigenvalue conditions, we have $\|\hat{\beta}_n(\lambda_n) - \beta^0\|_2 \leq b_n = O(\sqrt{s_0 \log(p_n)/n})$ leading to $C_n > b_n = O(\sqrt{s_0 \log(p_n)/n})$; see Section 6.8.) It may happen that $S_0^{\text{relevant}(C_n)} = S_0$, that is, all non-zero coefficients are at least as large as C_n in absolute value. (We call this a beta-min condition, see later.) Then, $\hat{S}(\lambda_n) \supset S_0$ with high probability.

We refer to the property in (2.11) or in (2.13) as *variable screening*: with high probability, the Lasso estimated model includes the substantial (relevant) covariates. Variable screening with the Lasso has a great potential because of the following fact. Every Lasso estimated model has cardinality smaller or equal to $\min(n, p)$: this follows from the analysis of the LARS algorithm (Efron et al., 2004). If $p \gg n$, $\min(n, p) = n$ which is a small number as compared to p and hence, we achieve a typically large dimensionality reduction in terms of the original covariates. For example, in the lymph node status classification problem in Section 2.4.1, we reduce from $p = 7129$ to at most $n = 49$ covariates.

2.5.1 Tuning parameter selection for variable screening

In practice, the tuning parameter λ is usually chosen via some cross-validation scheme aiming for prediction optimality. Such prediction optimality is often in conflict with variable selection where the goal is to recover the underlying set of active variables S_0: for the latter (if at all possible), we often need a larger penalty parameter than for good (or optimal) prediction. It is generally rather difficult to choose a proper amount of regularization for identifying the true active set S_0. In Section 2.11, the BIC criterion is described which, however, has no theoretical justification for variable selection with the Lasso; Chapters 10 and 11 describe alternatives to choosing the amount of regularization.

When sticking to cross-validation yielding a value $\hat{\lambda}_{CV}$, the Lasso often selects too many variables. This ties in nicely with the screening property in (2.11) or (2.13). We summarize that the Lasso screening procedure is very useful and easy to implement. The Lasso screening procedure yields an estimated set of selected variables $\hat{S}(\hat{\lambda}_{CV})$ containing with high probability S_0, or at least its relevant variables from $S_0^{\text{relevant}(C)}$, and whose cardinality is bounded by $|\hat{S}(\hat{\lambda}_{CV})| \leq \min(n, p)$.

As an alternative, we may pursue a Lasso screening procedure by including all $\min(n, p)$ variables using a value λ sufficiently close to zero (e.g. using the LARS algorithm until the end of the regularization path (Efron et al., 2004)) and hence, no tuning parameter needs to be chosen. If $p \gg n$, this tuning-free dimensionality reduction can be very worthwhile for a first stage.

The empirical fact that often $\hat{S}(\hat{\lambda}_{CV}) \supseteq S_0$ (or replacing S_0 by $S^{\text{relevant}(C_n)}$) is supported by theory. Consider the prediction optimal tuning parameter supplied by an

oracle, for random design,

$$\lambda^* = \lambda_n^* = \operatorname{argmin}_\lambda \mathbb{E}\left[(Y_{\text{new}} - \sum_{j=1}^{p} \hat{\beta}_j(\lambda)X_{\text{new}}^{(j)})^2\right], \qquad (2.14)$$

where $(X_{\text{new}}, Y_{\text{new}})$ is an independent copy of (X_i, Y_i) $(i = 1, \ldots, n)$. Meinshausen and Bühlmann (2006, Prop.1) present a simple example, with random design generated from uncorrelated variables, where

$$\mathbf{P}[\hat{S}(\lambda^*) \supseteq S_0] \to 1 \ (p \geq n \to \infty),$$
$$\limsup_{n \to \infty} \mathbf{P}[\hat{S}(\lambda^*) = S_0] < 1 \ (p \geq n \to \infty). \qquad (2.15)$$

2.5.2 Motif regression for DNA binding sites

We illustrate the Lasso on a motif regression problem (Conlon et al., 2003) for finding the binding sites in DNA sequences of the so-called HIF1α transcription factor. Such transcription factor binding sites, also called motifs, are short "words" of DNA letters denoted by $\{A, C, G, T\}$, typically 6-15 base pairs long.

The data consists of a univariate response variable Y measuring the binding intensity of the HIF1α transcription factor on coarse DNA segments which are a few thousands base pairs long. This data is collected using CHIP-chip experiments. In order to get to the exact short DNA "words" or motifs, short candidate DNA "words" of length 6–15 base pairs are generated and their abundance scores are measured within coarse DNA regions. This can be done using computational biology algorithms based on DNA sequence data only, and we use a variant of the MDScan algorithm (Liu et al., 2002). In our particular application, we have the following data:

Y_i measuring the binding intensity of HIF1α in coarse DNA segment i,

$X_i^{(j)}$ measuring the abundance score of candidate motif j in DNA segment i,

$i = 1, \ldots, n = 287;\ j = 1, \ldots, p = 195.$

A linear model fits reasonably well (see below) for relating the response to the covariates:

$$Y_i = \mu + \sum_{j=1}^{195} \beta_j X_i^{(j)} \ (i = 1, \ldots, n = 287).$$

The main goal in this application is variable selection to infer the relevant covariates and hence the relevant motifs (short DNA "words"). Having scaled the covariates to the same empirical variance, we use the Lasso with regularization parameter $\hat{\lambda}_{CV}$

from 10-fold cross-validation for optimal prediction. The fitted model has an $R^2 \approx$ 50% which is rather high for this kind of application. There are $|\hat{S}(\hat{\lambda}_{CV})| = 26$ non-zero coefficient estimates $\hat{\beta}_j(\hat{\lambda}_{CV})$ which are plotted in Figure 2.3. Based on the methodology and theory described informally in Section 2.5 (rigorous mathematical arguments are given in Chapters 6 and 7), there is evidence that the truly relevant variables are a subset of the 26 selected variables shown in Figure 2.3.
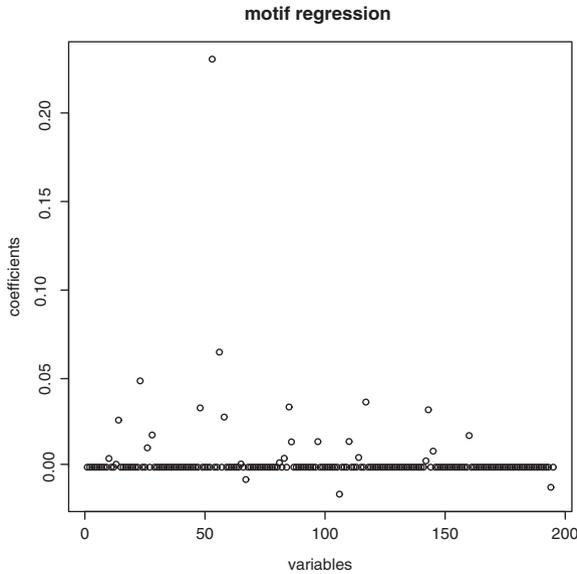
Fig. 2.3 Coefficient estimates $\hat{\beta}(\hat{\lambda}_{CV})$ for the motif regression data, aiming to find the HIF1α binding sites. Sample size and dimensionality are $n = 287$ and $p = 195$, respectively, and the cross-validation tuned Lasso selects 26 variables.

2.6 Variable selection

The problem of variable selection for a high-dimensional linear model in (2.1) is important since in many areas of applications, the primary interest is about the relevance of covariates. As there are 2^p possible sub-models, computational feasibility is crucial. Commonly used variable selection procedures are based on least squares and a penalty which involves the number of parameters in the candidate sub-model:

$$\hat{\beta}(\lambda) = \text{argmin}_\beta \left(\|\mathbf{Y} - \mathbf{X}\beta\|_2^2 / n + \lambda \|\beta\|_0^0 \right), \tag{2.16}$$

where the ℓ_0-penalty is $\|\beta\|_0^0 = \sum_{j=1}^p 1(\beta_j \neq 0)$. Many well known model selection criteria such as the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC) or the Minimum Description Length (MDL) fall into this framework. For example, when the error variance is known, AIC and BIC correspond to $\lambda = 2\sigma^2/n$ and $\lambda = \log(n)\sigma^2/n$, respectively. The estimator in (2.16) is infeasible to compute when p is of medium or large size since the ℓ_0-penalty is a non-convex function in β. Computational infeasibility remains even when using branch-and-bound techniques, cf. Hofmann et al. (2007) or Gatu et al. (2007). Forward selection strategies are computationally fast but they can be very instable (Breiman, 1996), as illustrated in Table 2.1 where forward selection produced a poor result. Other ad-hoc methods may be used to get approximations for the ℓ_0-penalized least squares estimator in (2.16). On the other hand, the requirement of computational feasibility and statistical accuracy can be met by the Lasso defined in (2.2): it can also be viewed as a convex relaxation of the optimization problem with the ℓ_0 analogue of a norm in (2.16).

We will first build up the methodology and theory by using the Lasso in a single stage. We describe later in Section 2.8 how to use the Lasso not just once but in two (or more) stages. Consider the set of estimated variables using the Lasso as in (2.10):

$$\hat{S}(\lambda) = \{j; \ \hat{\beta}_j(\lambda) \neq 0, \ j = 1,\ldots,p\}.$$

In particular, we can compute all possible Lasso sub-models

$$\widehat{\mathscr{S}} = \{\hat{S}(\lambda); \text{ all } \lambda\} \tag{2.17}$$

with $O(np\min(n,p))$ operation counts, see Section 2.12. As pointed out above in Section 2.5, every sub-model in $\widehat{\mathscr{S}}$ has cardinality smaller or equal to $\min(n,p)$. Furthermore, the number of sub-models in $\widehat{\mathscr{S}}$ is typically of the order $O(\min(n,p))$ (Rosset and Zhu, 2007). Thus, in summary, each Lasso estimated sub-model contains at most $\min(n,p)$ variables,

$$|\hat{S}(\lambda)| \leq \min(n,p) \text{ for every } \lambda,$$

which is a small number if $p \gg n$, and the number of different Lasso estimated sub-models is typically

$$|\widehat{\mathscr{S}}| = O(\min(n,p)),$$

which represents a huge reduction compared to all 2^p possible sub-models if $p \gg n$. The question of interest is whether the true set of effective variables $S_0 = \{j; \ \beta_j^0 \neq 0, \ j = 1,\ldots,p\}$ is contained in $\widehat{\mathscr{S}}$ and if yes, which particular choice of λ will identify the true underlying set of active variables S_0.

An asymptotic result described below shows, assuming rather restrictive conditions, that with probability tending to 1, $S_0 \in \widehat{\mathscr{S}}$ and that the Lasso is appropriate for addressing the problem of variable selection. As in Section 2.4, to capture high-dimensionality of the model (2.1) in an asymptotic sense, we consider the triangular array scheme in (2.6). The main and restrictive assumption for consistent variable selection concerns the (fixed or random) design matrix $\mathbf{X}$. The condition, called neighborhood stability or irrepresentable condition, is described with more rigor in Section 2.6.1. Under such a neighborhood stability condition, and assuming that the non-zero regression coefficients satisfy

$$\inf_{j \in S_0^c} |\beta_j^0| \gg \sqrt{s_0 \log(p)/n}, \tag{2.18}$$

Meinshausen and Bühlmann (2006, Theorems 1 and 2) show the following: for a suitable $\lambda = \lambda_n \gg \sqrt{\log(p_n)/n}$,

$$\mathbf{P}[\hat{S}(\lambda) = S_0] \to 1 \ (n \to \infty). \tag{2.19}$$

We note that in general, the regularization parameter $\lambda = \lambda_n$ needs to be chosen of a larger order than $\sqrt{\log(p)/n}$ to achieve consistency for variable selection and hence, the regularization parameter λ should be chosen larger for variable selection than for prediction, see Section 2.5.1. See also Problem 7.5.

It is worth mentioning here, that the neighborhood stability condition on the design is sufficient and necessary (see also the next subsection) and hence, we have a sharp result saying when the Lasso is consistent for variable selection and when not (for sufficiency of the condition, we implicitly assume that (2.18) holds). It should represent a warning that the restrictive assumptions on the design have some relevant implications on the statistical practice for high-dimensional model selection: with strongly correlated design, the Lasso can perform very poorly for variable selection. In addition, the requirement in (2.18), which we call a beta-min condition, that all non-zero coefficients are sufficiently large may be unrealistic in practice. Small non-zero coefficients cannot be detected (in a consistent way) and their presence is related to the phenomenon of super-efficiency: Leeb and Pötscher (2005) discuss many aspects of model selection, covering in particular the issue of small regression coefficients and its implications and challenges. Without a condition as in (2.18), we can still have a variable screening result as in (2.13). More details about the beta-min condition are given in Section 7.4. Finally, another difficulty comes with the choice of the regularization parameter as indicated in (2.15).

More detailed mathematical formulations and statements are provided in Chapter 7. Furthermore, we will describe in Chapter 13 the relation between Gaussian graphical modeling and variable selection in a linear model.

2.6.1 Neighborhood stability and irrepresentable condition

There is certainly an interesting potential to use the Lasso for variable selection in high-dimensional models, as described in (2.19). However, the so-called neighborhood stability condition is crucial for consistent variable selection with the Lasso: in fact, it is sufficient and essentially necessary for (2.19), see Theorems 1, 2 and Proposition 3 in Meinshausen and Bühlmann (2006). The word "essentially" refers to the fact that the necessary condition requires a quantity to be "≤ 1" while the sufficient condition requires strict "< 1", analogously to the explanation after formula (2.20).

The neighborhood stability condition is equivalent to the so-called irrepresentable condition (at least for the case where $n > p$ is fixed) which has been introduced by Zou (2006) and Zhao and Yu (2006) and which is easier to describe. We denote by $\hat{\Sigma} = n^{-1}\mathbf{X}^T\mathbf{X}$. Without loss of generality, we assume that the active set $S_0 = \{j; \beta_j^0 \neq 0\} = \{1, \ldots, s_0\}$ consists of the first s_0 variables. Let

$$\hat{\Sigma} = \begin{pmatrix} \hat{\Sigma}_{1,1} & \hat{\Sigma}_{1,2} \\ \hat{\Sigma}_{2,1} & \hat{\Sigma}_{2,2} \end{pmatrix},$$

where $\hat{\Sigma}_{1,1}$ is a $s_0 \times s_0$ matrix corresponding to the active variables, $\hat{\Sigma}_{1,2} = \hat{\Sigma}_{2,1}^T$ is a $s_0 \times (p - s_0)$ matrix and $\hat{\Sigma}_{2,2}$ a $(p - s_0) \times (p - s_0)$ matrix. The irrepresentable condition then reads:

$$\|\hat{\Sigma}_{2,1}\hat{\Sigma}_{1,1}^{-1}\text{sign}(\beta_1^0, \ldots, \beta_{s_0}^0)\|_\infty \leq \theta \text{ for some } 0 < \theta < 1, \tag{2.20}$$

where $\|x\|_\infty = \max_j |x^{(j)}|$ and $\text{sign}(\beta_1^0, \ldots, \beta_p^0) = (\text{sign}(\beta_1^0), \ldots, \text{sign}(\beta_p^0))^T$. As with the neighborhood stability condition, the irrepresentable condition in (2.20) is sufficient and "essentially" necessary for consistent model selection with the Lasso: the word "essentially" refers to the fact that the necessary condition requires the relation "≤ 1", while the sufficient condition requires "$\leq \theta$" for some $0 < \theta < 1$, see Theorem 7.1 in Chapter 7. At first sight, the difference between "≤ 1" and "$\leq \theta$ for some $0 < \theta < 1$" seems rather small: however, examples like the case with equal positive correlation below show that this difference may be substantial. For the high-dimensional setting and in terms of the triangular array as in (2.6), it is understood that the right-hand side of (2.20) is bounded by θ for all $n \in \mathbb{N}$. Furthermore, the bound $\theta < 1$ in general requires that the regularization parameter $\lambda = \lambda_n$ needs to be chosen of a larger order than $\sqrt{\log(p)/n}$ to achieve consistency for variable selection.

Roughly speaking, the neighborhood stability or irrepresentable condition fails to hold if the design matrix $\mathbf{X}$ is too much "ill-posed" and exhibits a too strong degree of linear dependence within "smaller" sub-matrices of $\mathbf{X}$. We now give some examples where the irrepresentable condition holds: we formulate them in terms of a

general covariance matrix Σ. For the consequences for a sample covariance matrix, we refer to Problem 7.6.

Equal positive correlation. $\Sigma_{j,j} = 1$ for all $j = 1, \ldots, p$ and $\Sigma_{j,k} = \rho$ for all $j \neq k$ with $0 \leq \rho \leq \frac{\theta}{s_0(1-\theta)+\theta} < 1$ $(0 < \theta < 1)$. Then the irrepresentable condition holds with the constant θ. (Note the difference to $\theta = 1$ corresponding to the necessity of the irrepresentable condition: then, $0 \leq \rho \leq 1$ would be allowed). We leave the derivation as Problem 2.4 (see also Problem 6.14).

Toeplitz structure. $\Sigma_{j,k} = \rho^{|j-k|}$ for all j,k with $|\rho| \leq \theta < 1$. Then the irrepresentable condition holds with the constant θ (Problem 2.4).

Bounded pairwise correlation. If

$$\frac{\sqrt{s_0} \max_{j \notin S_0} \sqrt{\sum_{k \in S_0} \Sigma_{j,k}^2}}{\Lambda_{\min}^2(\Sigma_{1,1})} \leq \theta < 1,$$

where $\Lambda_{\min}^2(\Sigma_{1,1})$ is the minimal eigenvalue of $\Sigma_{1,1}$, then the irrepresentable condition holds with the constant θ.

It is shown in Chapter 7 that the condition on bounded pairwise correlations implies the irrepresentable condition (Corollary 7.2) and that the irrepresentable condition implies the compatibility condition (Theorem 7.2). The latter allows to establish oracle results for prediction and estimation as in (2.8) and (2.12), respectively and hence, this indicates that variable selection is a harder problem than prediction or parameter estimation.

2.7 Key properties and corresponding assumptions: a summary

We summarize here in a rough way Sections 2.4, 2.5 and 2.6 about the key properties of the Lasso in a linear model

$$\mathbf{Y} = \mathbf{X}\beta^0 + \varepsilon$$

with fixed design, as in (2.1). Thereby, we do not give a precise specification of the regularization parameter λ.

For prediction with a slow rate of convergence,

$$\|\mathbf{X}(\hat{\beta} - \beta^0)\|_2^2 / n = O_P(\|\beta^0\|_1 \sqrt{\log(p)/n}) \tag{2.21}$$

where $O_P(\cdot)$ is with respect to $p \geq n \to \infty$. That is, we achieve consistency for prediction if $\|\beta^0\|_1 \ll \sqrt{n/\log(p)}$. Oracle optimality improves the statement (2.21) to a considerably faster convergence rate and estimation error bounds with respect to the ℓ_1- or ℓ_2-norm:

$$\|\mathbf{X}(\hat{\beta} - \beta^0)\|_2^2/n = O_P(s_0 \phi^{-2} \log(p)/n),$$
$$\|\hat{\beta} - \beta^0\|_q = O_P(s_0^{1/q} \phi^{-2} \sqrt{\log(p)/n}), \ q \in \{1,2\}, \tag{2.22}$$

where s_0 equals the number of non-zero regression coefficients and ϕ^2 denotes a restricted eigenvalue of the design matrix $\mathbf{X}$. The rate in (2.22) is optimal up to the $\log(p)$ factor and the restricted eigenvalue ϕ^2: oracle least squares estimation where the relevant variables would be known would have rate $O_P(s_0/n)$. (We note that the result for $q = 2$ requires stronger conditions on the design than for $q = 1$). From (2.22), when assuming the beta-min condition (see also (2.18) and Section 7.4)

$$\min_{j \in S_0^c} |\beta_j^0| \gg \phi^{-2} \sqrt{s_0 \log(p)/n}, \tag{2.23}$$

we obtain the variable screening property:

$$\mathbf{P}[\hat{S} \supseteq S_0] \to 1 \ (p \geq n \to \infty) \tag{2.24}$$

where $\hat{S} = \{j; \ \hat{\beta}_j \neq 0, \ j = 1, \ldots, p\}$ and $S_0 = \{j; \ \beta_j^0 \neq 0, \ j = 1, \ldots, p\}$. A quite different problem is variable selection for inferring the true underlying active set S_0. Consistent variable selection then means that

$$\mathbb{P}[\hat{S} = S_0] \to 1 \ (p \geq n \to \infty), \tag{2.25}$$

These basic facts are summarized in Table 2.2, and we note that the results can be refined as shown in Chapters 6 and 7.

property	design condition	size of non-zero coeff.
slow convergence rate as in (2.21)	no requirement	no requirement
fast convergence rate as in (2.22) with $q = 1$	compatibility	no requirement
variable screening as in (2.24)	restricted eigenvalue	beta-min condition (2.23)
variable selection as in (2.25)	neighborhood stability $\Leftrightarrow$ irrepresentable cond.	beta-min condition (2.23)

Table 2.2 Properties of the Lasso and sufficient conditions to achieve them. The neighborhood stability condition and the equivalent irrepresentable condition are discussed in Section 2.6.1. The restricted eigenvalue assumption or the slightly weaker compatibility condition, see Section 6.2.2 in Chapter 6, are weaker than the neighborhood stability or irrepresentable condition, see Section 7.5.4 in Chapter 7.

2.8 The adaptive Lasso: a two-stage procedure

An interesting approach to correct Lasso's overestimation behavior, see formulae (2.11), (2.13) and (2.15), is given by the adaptive Lasso (Zou, 2006) which replaces the ℓ_1-penalty by a re-weighted version. For a linear model as in (2.1), it is defined as a two-stage procedure:

$$\hat{\beta}_{\text{adapt}}(\lambda) = \text{argmin}_\beta \left(\|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + \lambda \sum_{j=1}^{p} \frac{|\beta_j|}{|\hat{\beta}_{\text{init},j}|} \right), \tag{2.26}$$

where $\hat{\beta}_{\text{init}}$ is an initial estimator.

In the high-dimensional context, we propose to use the Lasso from a first stage as the initial estimator, tuned in a prediction optimal way. Typically, we use cross-validation to select the tuning parameter, denoted here by $\hat{\lambda}_{\text{init},CV}$. Thus, the initial estimator is $\hat{\beta}_{\text{init}} = \hat{\beta}(\hat{\lambda}_{\text{init},CV})$ from (2.2). For the second stage, we use again cross-validation to select the parameter λ in the adaptive Lasso (2.26). Proceeding this way, we select the regularization parameters in a sequential way: this is computationally much cheaper since we optimize twice over a single parameter instead of simultaneous optimization over two tuning parameters. The procedure is also described in Section 2.8.5 (when using $k = 2$).

The adaptive Lasso has the following obvious property:

$$\hat{\beta}_{\text{init},j} = 0 \implies \hat{\beta}_{\text{adapt},j} = 0. \tag{2.27}$$

Furthermore, if $|\hat{\beta}_{\text{init},j}|$ is large, the adaptive Lasso employs a small penalty (i.e. little shrinkage) for the jth coefficient β_j which implies less bias. Thus, the adaptive Lasso yields a sparse solution and it can be used to reduce the number of false positives (selected variables which are not relevant) from the first stage. This is a desirable property since the Lasso from the first stage has the screening property that $\hat{S} \supseteq S_0$ with high probability. Further details about variable selection with the adaptive Lasso are described below in Section 2.8.3, Section 6.10 and Chapter 7.

2.8.1 An illustration: simulated data and motif regression

We illustrate the Lasso and adaptive Lasso on some simulated data from a linear model as in (2.1) with $p = 1000$ and $n = 50$. We choose $\beta_1 = 2$, $\beta_2 = 1$, $\beta_3 = 0.5$ and $\beta_4 = \ldots \beta_{1000} = 0$, $\varepsilon \sim \mathcal{N}(0,1)$ and $X^{(1)}, \ldots, X^{(1000)}$ i.i.d. $\sim \mathcal{N}(0,1)$. This amounts to a "medium-size" (squared) signal to noise ratio

$$\frac{\text{Var}(f(X))}{\sigma^2} = 5.5,$$

where $f(x) = x\beta$.

Figure 2.4 shows the coefficient estimates for the Lasso and the adaptive Lasso, with initial estimator from the Lasso, respectively. The tuning parameters are selected as follows. For the Lasso, we use the optimal λ from 10-fold cross-validation. This Lasso fit is used as initial estimator and we then choose λ for the second stage in adaptive Lasso by optimizing 10-fold cross-validation again. We empirically exploit
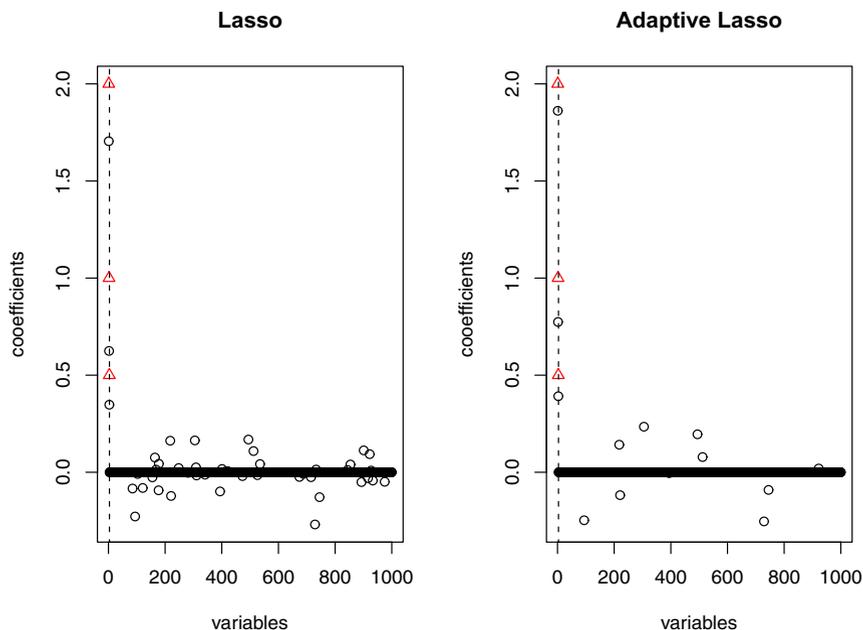
Fig. 2.4 Estimated regression coefficients in the linear model with $p = 1000$ and $n = 50$. Left: Lasso. Right: Adaptive Lasso with Lasso as initial estimator. The 3 true regression coefficients are indicated with triangles. Both methods used with tuning parameters selected from 10-fold cross-validation.

here the fact that Lasso is a powerful screening method: all three relevant variables are selected, i.e., $\hat{S} \supseteq S_0$, but it also selects 41 noise covariates. The adaptive Lasso yields a substantially sparser fit: it selects all of the 3 relevant variables and 10 noise covariates in addition.

We briefly described in Section 2.5.2 a problem from biology about finding binding sites of the HIF1α transcription factor. We recall that a linear model is a reasonable approximation for relating a univariate response about binding intensities on long DNA segments and a p-dimensional covariate measuring abundance scores of short candidate motifs within long DNA segments. Figure 2.5 displays the estimated regression coefficients using the Lasso with 10-fold cross-validation and using the

adaptive Lasso with the first-stage Lasso fit as initial estimator and choosing λ in the second adaptive stage with another 10-fold cross-validation. The adaptive Lasso fit is substantially sparser than using the Lasso only.
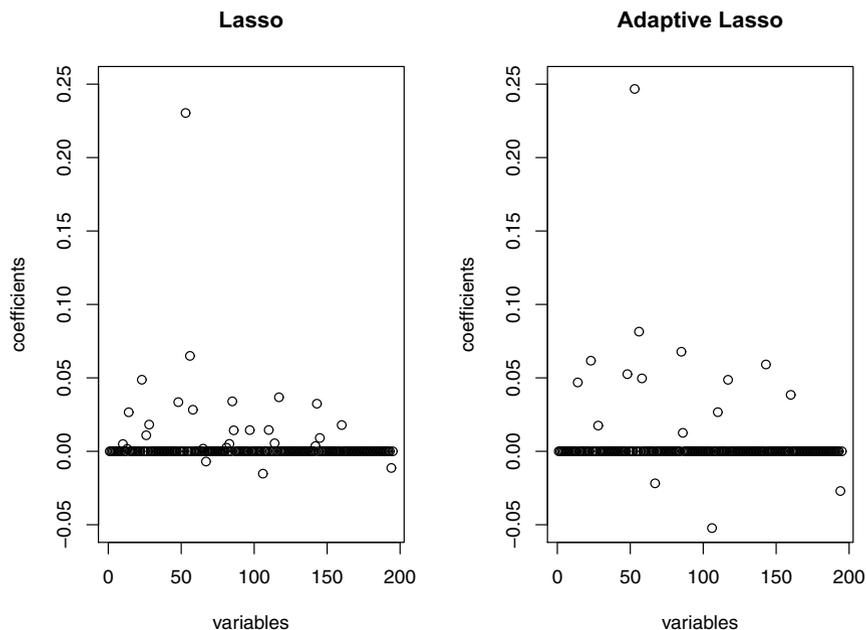
Fig. 2.5 Motif regression for HIF1α transcription factor ($n = 287$, $p = 195$), see also Section 2.5.2. Coefficient estimates using Lasso (left) or Adaptive Lasso (right), selecting 26 or 16 variables, respectively. Both methods used with tuning parameters selected from 10-fold cross-validation.

2.8.2 Orthonormal design

In the special case of an orthonormal design with $p = n$ and $\hat{\Sigma} = n^{-1}\mathbf{X}^T\mathbf{X} = I$, the adaptive Lasso has an explicit solution. We consider the case with the ordinary least squares initial estimator $\hat{\beta}_{\text{init},j} = n^{-1}(\mathbf{X}^T\mathbf{Y})_j = Z_j$ ($j = 1, \ldots, p = n$). Then the adaptive Lasso equals (Problem 2.5)

$$\hat{\beta}_{\text{adapt},j} = \text{sign}(Z_j)(|Z_j| - \frac{\lambda}{2|Z_j|})_+, \; Z_j = \mathbf{X}_j^T\mathbf{Y}/n \;\; (j = 1, \ldots, p = n), \quad (2.28)$$

where $(x)_+ = \max(x, 0)$ denotes the positive part of x. This is again a thresholding-type estimator $\hat{\beta}_{\text{adapt},j} = g(Z_j)$, where the thresholding function $g(\cdot)$ is depicted in Figure 2.2.

Figure 2.2 has the following interpretation. Hard-thresholding $g_{\text{hard},\lambda/2}(Z_j)$ where $g_{\text{hard},\lambda}(z) = z\mathbb{1}(|z| \le \lambda)$ yields a truncated least-squares estimator and hence, its bias is only due to the truncation (thresholding). Soft-thresholding, corresponding to Lasso, involves shrinkage, either to zero or to a value which is in absolute value smaller than the least squares estimate by λ. Hence, even if the least squares estimate is large in absolute value, soft-thresholding shrinks by the additive amount λ. Finally, the adaptive Lasso "adapts" to the least squares estimate whenever the latter is large in absolute value and thus in this sense, the adaptive Lasso is less biased than the Lasso.

There is an interesting connection to the nonnegative garrote estimator (Breiman, 1995) which is defined as

$$\hat{\beta}_{\text{nn-gar}} = \hat{c}_j \hat{\beta}_{\text{init},j},$$
$$\hat{c} = \operatorname{argmin}_c (n^{-1} \sum_{i=1}^{n} (Y_i - \sum_{j=1}^{p} c_j \hat{\beta}_{\text{init},j} X_i^{(j)})^2$$
$$\text{subject to } c_j \ge 0 \ (j = 1, \ldots, p) \text{ and } \sum_{j=1}^{p} c_j \le \lambda.$$

In the special case of an orthonormal design and using ordinary least squares as initial estimator, the nonnegative garrote estimator is equal to the adaptive Lasso in (2.28).

2.8.3 The adaptive Lasso: variable selection under weak conditions

For (consistent) variable selection in a linear model, the Lasso needs, as a sufficient and essentially necessary condition, that the design matrix satisfies the neighborhood stability or irrepresentable condition described in Section 2.6.1. On the other hand, we have argued in Section 2.5 and formula (2.9) that under weaker design conditions, the Lasso is reasonable for estimating the true underlying β^0 in terms of the $\|.\|_q$-norm with $q \in \{1, 2\}$. As an implication, the Lasso has the screening property where $\hat{S} \supseteq S_0$ with high probability. Thereby, we assume, depending on the Gram matrix $\hat{\Sigma} = n^{-1}\mathbf{X}^T\mathbf{X}$, that the non-zero regression coefficients are not too small, i.e.,

$$\min\{|\beta_j|; \ \beta_j \ne 0, \ j = 1, \ldots, p\} \ge Cs_0 \sqrt{\log(p)/n}$$

for some constant $C > 0$, see also formula (2.13) and the discussion afterwards how to relax the lower bound to the order $O(\sqrt{s_0 \log(p)/n})$.

With the adaptive Lasso, the hope is that the two-stage process would be sufficient to correct Lasso's overestimation behavior. This can be mathematically proved (see Corollary 7.8 and Corollary 7.9 in Chapter 7), assuming compatibility conditions on the design $\mathbf{X}$ which are weaker than the neighborhood stability or irrepresentable condition. When assuming sufficiently large non-zero regression coefficients as above (and in general, the lower bound cannot be relaxed to the order $\sqrt{s_0 \log(p)/n}$, see Section 7.5.9), these compatibility conditions are sufficient to achieve variable selection consistency in the $p \gg n$ setting: denoting by $\hat{S}_{\text{adapt},n}(\lambda) = \hat{S}_{\text{adapt}}(\lambda) = \{j;\ \hat{\beta}_{\text{adapt},j}(\lambda) \neq 0\}$,

$$\mathbf{P}[\hat{S}_{\text{adapt},n}(\lambda) = S_0] \to 1\ (n \to \infty),$$

for λ in the range of order $\sqrt{\log(p)/n}$. The fact that we can achieve consistent variable selection with the adaptive Lasso for cases where the Lasso is inconsistent for estimating the set S_0 is related to the issue that the adaptive Lasso exhibits less bias than the Lasso, as mentioned in Section 2.8.2. A detailed mathematical treatment for the adaptive Lasso is given in Section 6.10 in Chapter 6 and Sections 7.8 and 7.9 in Chapter 7.

2.8.4 Computation

The optimization for the adaptive Lasso in (2.26) can be re-formulated as a Lasso problem. We reparametrize by re-scaling the covariates as follows:

$$\tilde{X}^{(j)} = |\hat{\beta}_{\text{init},j}| X^{(j)}, \quad \tilde{\beta}_j = \frac{\beta_j}{|\hat{\beta}_{\text{init},j}|}.$$

Then the objective function in (2.26) becomes

$$\|\mathbf{Y} - \tilde{\mathbf{X}}\tilde{\beta}\|_2^2/n + \lambda \|\tilde{\beta}\|_1.$$

This is a Lasso-problem (where we omit all variables j with $\hat{\beta}_{\text{init},j} = 0$). Denote a solution by $\hat{\tilde{\beta}}$ and by back-transformation, we obtain a solution for the adaptive Lasso in (2.26):

$$\hat{\beta}_{\text{adapt}} = |\hat{\beta}_{\text{init},j}|\hat{\tilde{\beta}}_j.$$

In particular, any algorithm for solving the Lasso can be used for computation of the adaptive Lasso. We refer to Section 2.12 for Lasso algorithms.

2.8.5 Multi-step adaptive Lasso

For regularization in high-dimensional problems, we may want to use more than
one or two tuning parameters. This can be achieved by pursuing more adaptive (or
weighted) Lasso iterations where every iteration involves a separate tuning param-
eter (and as described below, these parameters are "algorithmically" constrained).
The multi-step adaptive Lasso (Bühlmann and Meier, 2008) works as follows.

<div align="center">Multi-Step Adaptive Lasso (MSA-Lasso)</div>

1. Initialize the weights $w_j^{[0]} \equiv 1$ $(j = 1\ldots,p)$.
2. For $k = 1, 2\ldots, M$:
   Use the adaptive Lasso with penalty term

$$\lambda^{*[k]} \sum_{j=1}^{p} w_j^{[k-1]} |\beta_j|. \tag{2.29}$$

where $\lambda^{*[k]}$ is the regularization parameter leading to prediction optimality. De-
note the corresponding estimator by $\hat{\beta}^{[k]} = \hat{\beta}^{[k]}(\lambda^{*[k]})$. In practice, the value $\lambda^{*[k]}$
can be chosen with a cross-validation scheme.
Up-date the weights

$$w_j^{[k]} = \frac{1}{|\hat{\beta}^{[k-1]}(\lambda^{*[k-1]})_j|}, \ j = 1, \ldots, p.$$

For $k = 1$ (one-stage), we do an ordinary Lasso fit and $k = 2$ (two-stage) corresponds
to the adaptive Lasso.

We will illustrate below the MSA-Lasso on a small simulated model and a real data
set from molecular biology. Before doing so, we describe some properties of the
method which are straightforward to derive.

First, MSA-Lasso increases the sparsity in every step in terms of the number of se-
lected variables although there is not necessarily a strict decrease of this number.
This follows immediately from (2.27). Second, MSA-Lasso can be computed using
an algorithm for the Lasso problem in every step, see also Section 2.8.4. The com-
putational complexity of computing all Lasso solutions over the whole range of the
tuning parameter λ is of the order $O(np\min(n, p))$, see formula (2.37) below. Thus,
MSA-Lasso has computational complexity $O(Mnp\min(n, p))$ since we select the
regularization parameters $\lambda^{*[k]}$ $(k = 1, 2, \ldots, M)$ sequentially instead of simultane-
ously. Due to the increase of sparsity, a later step is faster to compute than an earlier
one. The computational load is in sharp contrast to computing all solutions over
the whole range of all M tuning parameters: this would require $O(np(\min(n, p))^M)$
essential operations.

MSA-Lasso is related to approximating a non-convex optimization with a non-convex penalty function. This will be discussed in Section 2.8.6, Section 6.11 and Section 7.13.

2.8.5.1 Motif regression in computational biology

Reducing the number of false positives is often very desirable in biological or biomarker discovery applications since follow-up experiments can be costly and laborious. In fact, it can be appropriate to do conservative estimation with a low number of selected variables since we still see more selections than what may be validated in a laboratory.

We illustrate the MSA-Lasso method on a problem of motif regression for finding transcription factor binding sites in DNA sequences (Conlon et al., 2003), see also Section 2.5.2. Such transcription factor binding sites, also called motifs, are short "words" of DNA base pairs denoted by $\{A,C,G,T\}$, typically 6-15 base pairs long. Beer and Tavazoie (2004) contains a collection of microarray data and a collection of motif candidates for yeast. The latter is typically extracted from computational algorithms based on DNA sequence data only: for every of the n genes, we have a score for each of the p candidate motifs which describes the abundance of occurrences of a candidate motif up-stream of every gene. This yields a $n \times p$ design matrix $\mathbf{X}$ with motif scores for every gene (i.e. rows of $\mathbf{X}$) and every candidate motif (i.e. columns of $\mathbf{X}$). The idea is to predict the gene expression value of a gene based on motif scores.

The dataset which we consider consists of $n = 2587$ gene expression values of a heat-shock experiment with yeast and $p = 666$ motif scores. We use a training set of size 1300 and a validation set of size 650 for selecting the regularization parameters. The remaining data is used as a test-set. We use a linear model and the MSA-Lasso for fitting the model which is fairly high-dimensional exhibiting $n_{train} \approx 2p$.

The squared prediction error on the test-set, approximating $\mathbb{E}[(\hat{Y}_{new} - Y_{new})] = (\hat{\beta} - \beta)^T \Sigma (\hat{\beta} - \beta) + \text{Var}(\varepsilon)$ with $\Sigma = \text{Cov}(X)$, remains essentially constant for all estimators. This is probably due to high noise, i.e., large value of $\text{Var}(\varepsilon)$. But the number of selected variables decreases substantially as k increases:

	Lasso ($k = 1$)	1-Step ($k = 2$)	2-Step ($k = 3$)
test set squared prediction error	0.6193	0.6230	0.6226
number of selected variables	91	42	28

The list of top-ranked candidate motifs, i.e., the selected covariates ranked according to $|\hat{\beta}_j|$, gets slightly rearranged between the different estimators. The hope (and in part a verified fact) is that estimators with $k = 2$ or 3 stages yield more stable lists with fewer false positives than using the Lasso corresponding to $k = 1$.

2.8.6 Non-convex penalty functions

MSA-Lasso from Section 2.8.5 is related to approximating a non-convex optimization with a non-convex penalty function:

$$\hat{\beta} = \arg\min_{\beta} \left(\|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + \sum_{j=1}^{p} \mathrm{pen}(\beta_j) \right),$$

where $\mathrm{pen}(\cdot)$ is a non-convex penalty function which typically involves one or several tuning parameters.

One example is the ℓ_r-penalty for r close to 0 with the corresponding estimator

$$\hat{\beta}_{\ell_r}(\lambda) = \mathrm{argmin}_{\beta}(\|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + \lambda\|\beta\|_r^r), \qquad (2.30)$$

where $\|\beta\|_r^r = \sum_{j=1}^{p}|\beta_j|^r$. We note that the typical value of λ is now of the order $\sqrt{\log(p)/n}^{2-r}$, see Section 6.11 and Section 7.13. The penalty function is non-convex and not differentiable at zero: we define

$$\mathrm{pen}'_{\ell_r}(u) = \lambda \mathrm{sign}(u)|u|^{r-1}1(u \neq 0) + \infty 1(u = 0).$$

We discuss in Chapter Sections 6.11 and 7.13 theoretical properties of the ℓ_r-penalized least squares method with $0 < r < 1$.

Another prominent example is the SCAD (Smoothly Clipped Absolute Deviation) with the following penalty function: for $a > 2$,

$$\mathrm{pen}_{\lambda,a}(u) = \begin{cases} \lambda|u| & |u| \leq \lambda, \\ -(u^2 - 2a\lambda|u| + \lambda^2)/[2(a-1)] & \lambda < |u| \leq a\lambda, \\ (a+1)\lambda^2/2 & |u| > a\lambda, \end{cases}$$

where $u \in \mathbb{R}$, $\lambda \geq 0$ and the usual choice for the a-parameter is $a = 3.7$ (Fan and Li, 2001). The SCAD-penalized regression estimator is then

$$\hat{\beta}_{\mathrm{SCAD}}(\lambda) = \arg\min_{\beta} \left(\|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + \sum_{j=1}^{p} \mathrm{pen}_{\lambda,a}(\beta_j) \right). \qquad (2.31)$$

The SCAD penalty function is non-differentiable at zero and non-convex. The derivative (without the point zero) is: for $a > 2$,

$$\mathrm{pen}'_{\lambda,a}(u) = \mathrm{sign}(u) \left(\lambda 1(|u| \leq \lambda) + \frac{(a\lambda - |u|)_+}{a-1}1(|u| > \lambda) \right).$$

For both examples, an iterative local linear approximation for computing the estimator in (2.30) or (2.31) is related to a multi-step weighted Lasso procedure as follows: in the kth iteration,

$$\hat{\beta}^{[k]} = \arg\min_{\beta} \left(\|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + \sum_{j=1}^{p} w_j|\beta_j| \right), \quad w_j^{[k-1]} = |\text{pen}'_{\lambda,a}(\hat{\beta}_j^{[k-1]})|.$$

This is similar to (2.29), except that the tuning parameter λ is not depending on the iteration. The difference between the ℓ_r- and the SCAD-penalty is whether the point zero is an "absorbing state". If $\hat{\beta}_j^{[k-1]} = 0$, the weight $w_j^{[k-1]} = \infty$ for the ℓ_r-penalty, as appearing also in (2.29); for SCAD, however, $w_j^{[k-1]} = \lambda$ in contrast to an infinite weight in (2.29) (note that the subdifferential of $\text{pen}_{\lambda,a}(\cdot)$ at zero is in $[-\lambda, \lambda]$, see also Problem 4.2 in Chapter 4). For further details we refer the interested reader to Zou and Li (2008).

2.9 Thresholding the Lasso

Instead of using the adaptive Lasso to obtain a sparser model fit than the initial Lasso, see formula (2.27), we can simply threshold the coefficient estimates from the initial Lasso estimator $\hat{\beta}_{\text{init}} = \hat{\beta}_{\text{init}}(\lambda_{\text{init}})$:

$$\hat{\beta}_{\text{thres},j}(\lambda_{\text{init}}, \delta) = \hat{\beta}_{\text{init},j} \mathbb{1}(|\hat{\beta}_{\text{init},j}| > \delta).$$

The selected variables are then given by $\hat{S}_{\text{thres}} = \{j; \ \hat{\beta}_{\text{thres},j} \neq 0\}$. Furthermore, for estimation, we should refit the selected variables by ordinary least squares:

$$\hat{\beta}_{\text{thres}-\text{refit}} = (\mathbf{X}_{\hat{S}_{\text{thres}}}^T \mathbf{X}_{\hat{S}_{\text{thres}}})^{-1} \mathbf{X}_{\hat{S}_{\text{thres}}}^T \mathbf{Y},$$

where for $S \subset \{1, \ldots, p\}$, $\mathbf{X}_S$ is the restriction of $\mathbf{X}$ to columns in S.

Despite that this thresholding and refitting method is rather simple, its theoretical properties are as good or even slightly better than for the adaptive Lasso, see Section 7.8.4 in Chapter 7. In contrast to the Lasso-OLS hybrid estimator (see Section 2.10), the method above involves an additional thresholding stage. The theory perspective (Section 7.8) indicates that the additional thresholding step leads to better performance.

Regarding the selection of tuning parameters, we can proceed sequentially as for the adaptive Lasso. Using cross-validation for optimizing prediction, first select a regularization parameter λ_{init} and then, for fixed λ_{init} select the threshold parameter δ for the refitted estimator $\hat{\beta}_{\text{thres}-\text{refit}}$.

2.10 The relaxed Lasso

The relaxed Lasso (Meinshausen, 2007) is similar to the adaptive or thresholded Lasso in the sense that it addresses the bias problems of the Lasso. The method works as follows. In a first stage, all possible Lasso sub-models in $\widehat{\mathscr{S}}$ defined in (2.17) are computed. Then, in a second stage, every sub-model $\hat{S} \in \widehat{\mathscr{S}}$ is considered and the Lasso with smaller penalty parameter is used on such sub-models. That is, we consider the estimator

$$\hat{\beta}_{\hat{S}}(\lambda, \phi) = \arg\min_{\beta_{\hat{S}}} \left\{ \|\mathbf{Y} - \mathbf{X}_{\hat{S}} \beta_{\hat{S}}\|_2^2 / n + \phi \cdot \lambda \|\beta_{\hat{S}}\|_1 \right\} \ (0 \le \phi \le 1),$$
$$\hat{\beta}_{\hat{S}^c}(\lambda, \phi) = 0, \tag{2.32}$$

where $\hat{S}(\lambda)$ is the estimated sub-model from the first stage (see (2.17)), and where we denote by $\beta_S = \{\beta_j; \ j \in S\}$ and $\mathbf{X}_S$ the $n \times |S|$ matrix whose columns correspond to S, for some subset $S \subseteq \{1, \ldots, p\}$. It is worth pointing out that once we have computed the Lasso with parameter λ in the first stage, it is often very fast to compute the relaxed estimator in (2.32). A special case occurs with $\phi = 0$ which is known as the Lasso-OLS hybrid (Efron et al., 2004), using an OLS estimator in the second stage. The tuning parameters λ and ϕ can be selected by a cross-validation scheme. However, unlike as for the adaptive Lasso, we should select them simultaneously.

The relaxed and the adaptive Lasso appear to perform similarly in practice. Both procedures can be generalized to other penalties and models.

2.11 Degrees of freedom of the Lasso

Degrees of freedom are often used to quantify the complexity of a model fit and we can use them for choosing the amount of regularization. So far, we have always mentioned cross-validation for selecting tuning parameters of the Lasso and its multi-stage extensions. Another possibility is to use information criteria, such as the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC), which penalize the likelihood by the degrees of freedom of the fitted model. For example, for a Gaussian linear model as in (2.1), the estimated model with fitted values $\hat{Y}_i$ ($i = 1, \ldots, n$) has BIC-score:

$$\mathrm{BIC} = n \log(\hat{\sigma}^2) + \log(n) \cdot \mathrm{df}(\hat{\mathbf{Y}}),$$
$$\hat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2,$$

where df($\hat{\mathbf{Y}}$) denotes here the degrees of freedom of the fitted model, see below. We note that $\hat{\mathbf{Y}}$ is not necessarily an ordinary least squares fit, as discussed next.

Degrees of freedom can be defined in various ways, particularly when using different estimators than maximum likelihood. Stein's theory about unbiased risk estimation leads to a rigorous definition of degrees of freedom in a Gaussian linear model as in (2.1) with fixed design and errors $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. We denote by $\mathscr{H}\mathbf{Y} = \hat{\mathbf{Y}}$ the hat-operator which maps the response vector $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ to its fitted values $\hat{\mathbf{Y}} = (\hat{Y}_1, \ldots, \hat{Y}_n)^T$. The degrees of freedom for a possibly non-linear hat-operator $\mathscr{H}$ are then defined as

$$\mathrm{df}(\mathscr{H}) = \sum_{i=1}^{n} \mathrm{Cov}(\hat{Y}_i, Y_i)/\sigma^2, \qquad (2.33)$$

where the values $\hat{Y}_i$ arise from any model fitting method, see Efron (2004).

When using maximum likelihood estimation in parametric models, the degrees of freedom equal the number of estimated parameters. Alternatively, for linear hat-operators where $\hat{\mathbf{Y}} = \mathscr{H}\mathbf{Y}$ with a hat-matrix $\mathscr{H}$, the degrees of freedom in (2.33) equal

$$\mathrm{df}(\mathscr{H}) = \mathrm{trace}(\mathscr{H}) \qquad (2.34)$$

which is a standard formula for degrees of freedom of linear hat-operators, see Hastie and Tibshirani (1990). The derivation of (2.34) is left as Problem 2.6.

It is unknown how to assign degrees of freedom for the Lasso, except for the low-dimensional case where $p \leq n$. First, it is a nonlinear fitting method, e.g., soft-thresholding in the special case of an orthonormal design, and hence, formula (2.34) cannot be used. Secondly, counting the number of parameters seems wrong. A bit surprisingly though, it is this second view which leads to a simple formula, although only for the low-dimensional case where $p \leq n$.

We can easily count the number of non-zero estimated parameters, i.e., $|\hat{S}|$. It is plausible that shrinkage estimators involve less degrees of freedom than non-shrunken maximum likelihood estimates. On the other hand, the Lasso is estimating the sub-model with the active set $\hat{S}$, i.e., $\hat{S}$ is random, which adds variability and degrees of freedom in comparison to the situation where the model would be fixed. Surprisingly, the cost of search for selecting the model and the fact that shrinkage instead of maximum likelihood estimators are used compensate each other. The following result holds: for the Lasso with penalty parameter λ and associated hat-operator $\mathscr{H} = \mathscr{H}(\lambda)$, and if $\mathrm{rank}(\mathbf{X}) = p$ (i.e. not covering the high-dimensional case), the degrees of freedom are

$$\mathrm{df}(\mathscr{H}) = \mathbb{E}[|\hat{S}|],$$

see Zou et al. (2007). It is not known whether this simple formula would also hold for the case where $\mathrm{rank}(\mathbf{X}) < p$. In words, the expected number of selected vari-

ables from the Lasso(λ) estimator equals the degree of freedom. A simple unbiased estimator for the degrees of freedom of the Lasso is then:

$$\widehat{\mathrm{df}}(\mathscr{H}) = |\hat{S}|.$$

Needless to say that this formula is extremely easy to use. We can now choose the regularization parameter λ according to e.g. the BIC criterion

$$\hat{\lambda}_{\mathrm{BIC}} = \ \operatorname{argmin}_{\lambda} \left(n \log(n^{-1} \| \mathbf{Y} - \mathscr{H}(\lambda)\mathbf{Y} \|^2) + \log(n) \cdot |\hat{S}(\lambda)| \right). \qquad (2.35)$$

As we will see in Section 2.12, the regularization path of $\hat{\beta}(\lambda)$ is piecewise linear as a function of λ. Hence, the minimizer of (2.35) can be evaluated exactly.

2.12 Path-following algorithms

Usually, we want to compute the estimator $\hat{\beta}(\lambda)$ in (2.2) for many values of λ. For example, selecting a good λ, e.g., by using cross-validation, requires the computation over many different candidate values.

For the estimator in (2.2), it is possible to compute the whole regularized solution path over all values of λ in the following sense. The regularized solution path $\{\hat{\beta}(\lambda);\ \lambda \in \mathbb{R}^+\}$ is piecewise linear with respect to λ. That is:

there exist $\lambda_0 = 0 < \lambda_1 < \lambda_{m-1} < \lambda_m = \infty$, $\gamma_0, \gamma_1, \ldots, \gamma_{m-1} \in \mathbb{R}^p$ such that
$$\hat{\beta}(\lambda) = \hat{\beta}(\lambda_k) + (\lambda - \lambda_k)\gamma_k \text{ for } \lambda_k \le \lambda < \lambda_{k+1} \ (0 \le k \le m-1). \qquad (2.36)$$

There is a maximal value $\lambda_{\max} = \lambda_{m-1}$ where $\hat{\beta}(\lambda) = 0$ for all $\lambda \ge \lambda_{\max}$ and $\hat{\beta}_j(\lambda) \ne 0$ for $\lambda < \lambda_{\max}$ and some j. The value $\lambda_{\max}$ is characterized by

$$\lambda_{\max} = \max_{1 \le j \le p} |2\mathbf{X}_j^T \mathbf{Y}|/n.$$

This follows from the characterization of the Lasso solutions in Lemma 2.1. Typically, every λ_k is a "kink point" (marking piecewise linear segments) for only a single component of the coefficient paths $\hat{\beta}(\cdot)$. The number of different λ_k-values is typically of the order $m = O(n)$, see Rosset and Zhu (2007).

The fact that the estimator in (2.2) has a piecewise linear solution path as in (2.36) has computational consequences. All what we need to compute are the values (λ_k, γ_k) $(k = 0, \ldots, m-1)$. Having these, we can easily reconstruct the whole regularized solution path by linear interpolation. The (modified) LARS algorithm from Efron et al. (2004) can be used for this task which bears some similarities to the approach from Osborne et al. (2000). Its computational complexity, for computing the whole regularization path is:

$$O(np\min(n,p)) \text{ essential operation counts.} \tag{2.37}$$

Hence, if $p \gg n$, $O(np\min(n,p)) = O(p)$ and we have a computational complexity which is linear in the dimensionality p.

Figure 2.6 shows the whole regularization path for Lasso in a linear model, based on a real data example with $n = 71$ samples and $p = 4088$ covariates which is described in more detail in Section 9.2.6 (but here with a smaller more homogeneous sub-sample). The coefficients $\hat{\beta}_j(\lambda)$ are plotted as a function of a re-scaled λ parameter.
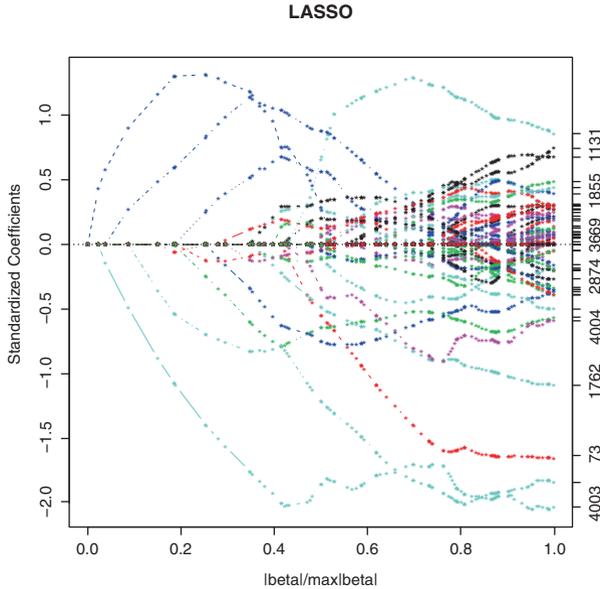
Fig. 2.6 Regularization path for Lasso in a linear model with $n = 71$ and $p = 4088$. x-axis: $\|\hat{\beta}(\lambda)\|_1 / \max\{\|\hat{\beta}(\lambda)\|_1; \lambda\}$, and y-axis: $\hat{\beta}_j\sqrt{\hat{\sigma}_j^2(n-1)}$ where $\hat{\sigma}_j^2$ denotes the empirical variance of $X^{(j)}$.

Although the LARS algorithm is exact for the whole piecewise linear regularization path, other algorithms described in Section 2.12.1 can be considerably faster for computing the Lasso over a large grid of λ-values (Friedman et al., 2007a, 2010). In addition, for other models and penalties, there is no piecewise linear regularization path any more and other algorithms are needed.

2.12.1 Coordinatewise optimization and shooting algorithms

For very high-dimensional but sparse problems, coordinate descent algorithms are often much faster than exact path-following methods such as the LARS-algorithm (Efron et al., 2004). This happens because coordinatewise up-dates are often very fast, e.g., explicit as in (2.38) in case of squared error loss, and they also exploit sparsity when using an active set modification as outlined in Section 2.12.1.1. In addition, when using other loss functions than squared error or when having a group-structure in the penalty function, exact path-following algorithms are not available and other optimization algorithms are needed. These two facts are the main motivation to focus on coordinatewise methods. We refer to Efron et al. (2004) for a description of the LARS algorithm for solving the Lasso optimization in (2.2).

Despite the fact that the regularized solution path for $\hat{\beta}(\lambda)$ in (2.2) is piecewise linear, see (2.36), it is often sufficient (or even better) for practical purposes to compute $\hat{\beta}(\lambda)$ on a grid of values $\Lambda = \{0 \leq \lambda_{\mathrm{grid},1} < \lambda_{\mathrm{grid},2} < \lambda_{\mathrm{grid},g}\}$. In particular, the value λ_k in (2.36) are data-dependent and hence, they change for say every cross-validation run. Therefore, when determining the best regularization parameter λ with cross-validation, we have to use fixed (data-independent) candidate values for λ anyway (or work with a fixed parameter on another scale).

We recommend to choose the grid to be equi-distant on the log-scale as follows. Choose

$$\lambda_{\mathrm{grid},g} = \lambda_{\max} = \max_{1 \leq j \leq p} |2\mathbf{X}_j^T \mathbf{Y}|/n,$$

$$\lambda_{\mathrm{grid},k-1} = \lambda_{\mathrm{grid},k} \exp(-C),$$

where $C > 0$ is a constant. Typically, we would choose C as a function of $\lambda_{\mathrm{grid},1}$: for the latter, we recommend

$$\lambda_{\mathrm{grid},1} \approx n^{-1},$$

and hence

$$C = \frac{\log(\lambda_{\max}) - \log(\lambda_{\mathrm{grid},1})}{g-1}.$$

The general idea is to compute a solution $\hat{\beta}(\lambda_{\mathrm{grid},g})$ and use it as a starting value for the computation of $\hat{\beta}(\lambda_{\mathrm{grid},g-1})$ and so on: the value $\hat{\beta}(\lambda_{\mathrm{grid},k})$ is used as a warm-start for the computation of $\hat{\beta}(\lambda_{\mathrm{grid},k-1})$. Hence, we will focus in the sequel on the computation for a single regularization parameter λ.

The simplest algorithm which exploits the characterization from Lemma 2.1 pursues coordinate descent minimization. Denote by

$$Q_\lambda(\beta) = \|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + \lambda \|\beta\|_1$$

the criterion function in (2.2). Furthermore, let

$$G_j(\beta) = -2\mathbf{X}_j^T(\mathbf{Y} - \mathbf{X}\beta)/n \quad (j = 1, \ldots, p)$$

be the gradient of $\|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n$. Consider the following algorithm.

Algorithm 1 Coordinate descent minimization

1: Let $\beta^{[0]} \in \mathbb{R}^p$ be an initial parameter vector. Set $m = 0$.
2: **repeat**
3:    Increase m by one: $m \leftarrow m + 1$.
      Denote by $\mathscr{S}^{[m]}$ the index cycling through the coordinates $\{1, \ldots, p\}$:
      $\mathscr{S}^{[m]} = \mathscr{S}^{[m-1]} + 1 \mod p$. Abbreviate by $j = \mathscr{S}^{[m]}$ the value of $\mathscr{S}^{[m]}$.
4:    if $|G_j(\beta_{-j}^{[m-1]})| \leq \lambda$: set $\beta_j^{[m]} = 0$,
      otherwise: $\beta_j^{[m]} = \underset{\beta_j}{\arg\min} \, Q_\lambda(\beta_{+j}^{[m-1]})$,
      where $\beta_{-j}^{[m-1]}$ is the parameter vector where the jth component is set to zero and $\beta_{+j}^{[m-1]}$ is
      the parameter vector which equals $\beta^{[m-1]}$ except for the jth component where it is equal to
      β_j (i.e. the argument we minimize over).
5: **until** numerical convergence

Due to the nature of the squared error loss, the up-date in Step 4 in Algorithm 1 is explicit (Problem 2.7): for $j = \mathscr{S}^{[m]}$,

$$\beta_j^{[m]} = \frac{\text{sign}(Z_j)(|Z_j| - \lambda/2)_+}{\hat{\Sigma}_{j,j}},$$

$$Z_j = \mathbf{X}_j^T(\mathbf{Y} - \mathbf{X}\beta_{-j}^{[m-1]})/n, \quad \hat{\Sigma} = n^{-1}\mathbf{X}^T\mathbf{X}, \tag{2.38}$$

where $\beta_{-j}^{[m-1]}$ denotes the parameter vector whose jth component is set to zero. Thus, we are doing componentwise soft-thresholding. For more details about such an algorithm and variations for other Lasso-related problems, we refer to Friedman et al. (2007a). Fu's shooting algorithm for the Lasso (Fu, 1998) is a special case of a coordinate descent approach.

Numerical convergence of the coordinate descent minimization algorithm can be established as follows. First, coordinatewise minima are attained since $Q_\lambda(\beta) = \|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + \lambda\|\beta\|_1 \to \infty$ if $\|\beta\|_1 \to \infty$. Second, we now argue that Step 4 minimizes the convex function $h(\beta_j) = Q_\lambda(\beta_{+j}^{[m-1]})$ with respect to β_j, where $\beta_{+j}^{[m-1]}$ denotes the parameter vector which equals $\beta^{[m-1]}$ except for the jth component where it is equal to β_j; note that $h(\beta_j)$ serves only as a notational abbreviation where all other parameters $\beta_k^{[m-1]}$ ($k \neq j$) are fixed. Since $h(\cdot)$ is not differentiable everywhere, we invoke subdifferential calculus (Bertsekas, 1995). The subdifferential of $h(\cdot)$ is the set $\partial h(\beta_j) = \{G_j(\beta_{+j}^{[m-1]}) + \lambda e, e \in E(\beta_j)\}$, $E(\beta_j) = \{e \in \mathbb{R} :$

sign(β_j) if $\beta_j \neq 0$ and $\|e\| \leq 1$ if $\beta_j = 0$}. The parameter β_j minimizes $h(\beta_j)$ if and only if $0 \in \partial h(\beta_j)$ which is equivalent to the formulation in Step 4.

Thirdly, cycling through the coordinates $\mathscr{S}^{[m]} = 1, \ldots, p, 1, \ldots$ ($m = 1, 2, \ldots$), i.e., a Gauss-Seidel algorithm, can be shown to converge to a stationary point. Numerical convergence of such a Gauss-Seidel algorithm seems plausible, but exact mathematical arguments are more involved crucially exploiting that the penalty function $\lambda \|\beta\|_1$ is a separable function of β.[1] We refer for details to the theory in Tseng (2001). In particular, conditions (A1), (B1) - (B3) and (C2) from Tseng (2001) hold and furthermore, by Lemma 3.1 and Proposition 5.1 in Tseng (2001), every cluster point of the sequence $\{\hat{\beta}^{[m]}\}_{m \geq 0}$ is a stationary point of the convex function $Q_\lambda(\cdot)$ and hence a minimum point.

Taking the three steps together, we summarize the result as follows.

Proposition 2.1. *Denote by $\hat{\beta}^{[m]}$ the parameter vector from Algorithm 1 after m iterations. Then every cluster point of the sequence $\{\hat{\beta}^{[m]}; m = 0, 1, 2, \ldots\}$ is a minimum point of $Q_\lambda(\cdot)$.*

We note that the iterates $\beta^{[m]}$ can be shown to stay in a compact set (because of the penalty term) and thus, the existence of a cluster point is guaranteed. Proposition 2.1 also follows from the more general result in Proposition 4.1 in Chapter 4.

The coordinatewise optimization above can easily incorporate the more general case where some parameters are unpenalized, i.e.,

$$\hat{\beta} = \arg\min_{\beta} Q_\lambda(\beta),$$

$$Q_\lambda(\beta) = \|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + \lambda \sum_{j=r+1}^{p} |\beta_j|,$$

and thus, $\beta_1, \ldots, \beta_r$ are unpenalized. The up-dating step in the optimization algorithm then looks as follows:

$$\text{if } j \in \{1, \ldots, r\} : \ \beta_j^{[m]} = \arg\min_{\beta_j} Q_\lambda(\beta_{+j}^{[m-1]}),$$

$$\text{if } j \in \{r+1, \ldots, p\} :$$

$$\text{if } |G_j(\beta_{-j}^{[m-1]})| \leq \lambda : \ \text{set } \beta_j^{[m]} = 0,$$

$$\text{otherwise: } \beta_j^{[m]} = \arg\min_{\beta_j} Q_\lambda(\beta_{+j}^{[m-1]}).$$

---

[1] A function $f(\beta)$ is called separable (into convex functions) if $f(\beta) = \sum_{j=1}^{p} f_j(\beta_j)$ for some convex functions $f_j(\cdot)$.

2.12.1.1 Active set strategy

An active set strategy can greatly improve computational efficiency for sparse, high-dimensional problems where only few among very many variables are active.

When cycling through the coordinates, we focus on the current active set and visit only "rarely" the remaining variables (e.g. every 10th iteration) to update the active set.

2.13 Elastic net: an extension

A double penalization using a combination of the ℓ_1- and ℓ_2-penalties has been proposed by Zou and Hastie (2005):

$$\hat{\beta}_{\text{naiveEN}}(\lambda_1,\lambda_2) = \arg\min_{\beta}\left(\|\mathbf{Y}-\mathbf{X}\beta\|_2^2/n+\lambda_1\|\beta\|_1+\lambda_2\|\beta\|_2^2\right), \qquad (2.39)$$

where $\lambda_1,\lambda_2 \geq 0$ are two regularization parameters and $\|\beta\|_2^2 = \sum_{j=1}^{p}\beta_j^2$. Zou and Hastie (2005) called the estimator in (2.39) the "naive elastic net". A correction leading to the elastic net estimator is then:

$$\hat{\beta}_{\text{EN}}(\lambda_1,\lambda_2) = (1+\lambda_2)\hat{\beta}_{\text{naiveEN}}(\lambda_1,\lambda_2). \qquad (2.40)$$

The correction factor $(1+\lambda_2)$ in (2.40) is best motivated from the orthonormal design where $n^{-1}\mathbf{X}^T\mathbf{X} = I$. Then, see Problem 2.8,

$$\hat{\beta}_{\text{naiveEN},j}(\lambda_1,\lambda_2) = \frac{\text{sign}(Z_j)(|Z_j|-\lambda/2)_+}{1+\lambda_2},$$
$$Z_j = \mathbf{X}_j^T\mathbf{Y}/n \ \ (j=1,\ldots,p=n), \qquad (2.41)$$

where $(x)_+ = \max(x,0)$. This should be compared to the Lasso in formula (2.5). We see that the naive elastic net estimator has very bad bias behavior if $Z_j = \hat{\beta}_{\text{OLS},j}$ is large. The correction factor then leads to the fact that the elastic net estimator in (2.40) equals the Lasso for the case of orthonormal design.

The reason for adding an additional squared ℓ_2-norm penalty is motivated by Zou and Hastie (2005) as follows. For strongly correlated covariates, the Lasso may select one but typically not both of them (and the non-selected variable can then be approximated as a linear function of the selected one). From the point of view of sparsity, this is what we would like to do. However, in terms of interpretation, we may want to have two even strongly correlated variables among the selected variables: this is motivated by the idea that we do not want to miss a "true" variable due to selection of a "non-true" which is highly correlated with the true one. For more details, we refer the reader to Zou and Hastie (2005). From the prediction

point of view, there can be gains as well by using the elastic net in comparison to the Lasso (Bunea, 2008; Hebiri and van de Geer, 2010).

Computation of the elastic net estimator can be done by using an algorithm for the Lasso, see Problem 2.9.

Problems

2.1. Threshold estimator
(a) Show that in the orthonormal case, the Lasso equals the soft-threshold estimator which is depicted in Figure 2.2.
(b) Show that the ℓ_0-penalized estimator in (2.16) equals the hard-threshold estimator which is depicted in Figure 2.2.

2.2. Variable screening
Assume that (2.9) holds. For fixed $0 < C < \infty$, prove formula (2.11).

2.3. Variable screening (Similar to Problem 2.2).
Assume that (2.12) holds. Prove formula (2.13).

2.4. Irrepresentable condition
(a) Consider the case of equicorrelation for Σ: $\Sigma_{j,j} = 1$ for all $j = 1,\ldots,p$ and $\Sigma_{j,k} = \rho$ for all $j \neq k$ with $0 \leq \rho \leq \frac{\theta}{s_0(1-\theta)+\theta} < 1$ ($0 < \theta < 1$). Show that the irrepresentable condition (2.20) holds with constant θ.
Hint: the inverse is given by

$$\Sigma^{-1} = \frac{1}{1-\rho}(I_{p \times p} - \frac{\rho}{1+(p-1)\rho}\tau\tau^T), \ \tau = \tau_{p \times 1} = (1,1,\ldots,1).$$

See also Problem 6.14 and Problem 10.4.
(b) Consider equicorrelation for Σ as in (a) but now with potentially negative values: the range $-1/(p-1) < C_1 \leq \rho \leq C_2 < 1$ is the set where Σ is strictly positive definite. Consider now the restricted range $-\theta/(2s_0-1) \leq \rho \leq \frac{\theta}{s_0(1-\theta)+\theta}$ ($0 < \theta < 1$). Show that the irrepresentable condition (2.20) holds with constant θ.
Hint: use again the formula for the inverse in (a).
(c) Consider the case of Toeplitz structure for Σ: $\Sigma_{j,j} = 1$ for all $j = 1,\ldots,p$ and $\Sigma_{j,k} = \rho^{|j-k|}$ for all $j \neq k$ with $0 \leq |\rho| \leq \theta < 1$. Show that the irrepresentable condition (2.20) holds with constant θ. Use the fact that Σ^{-1} is a banded matrix with a diagonal with equal entries and two side-diagonals with equal entries, i.e., $\Sigma_{j,j}^{-1} = a$ ($1 \leq j \leq p$), $\Sigma_{j,k}^{-1} = b$ for $k = j+1$ ($1 \leq j \leq p-1$) and $k = j-1$ ($2 \leq j \leq p$), and $\Sigma_{j,k}^{-1} = 0$ for $k \geq j+2$ ($1 \leq j \leq p-2$) and $k \leq j-2$ ($3 \leq j \leq p$), and exploiting the identity $\Sigma\Sigma^{-1} = I$.

2.5. Adaptive Lasso
(a) For the orthonormal case, derive the threshold function for the adaptive Lasso with ordinary least squares initial estimator. This threshold function is depicted in Figure 2.2.
Hint: Consider every component and the parameter $\lambda_j = \lambda / |Z_j|$.
(b) For the orthonormal case, show that the nonnegative garrote estimator with ordinary least squares initial estimate equals the adaptive Lasso.

2.6. Degrees of freedom for linear hat-operators
Prove that formula (2.34) holds for linear hat-operators $\hat{\mathbf{Y}} = \mathcal{H}\mathbf{Y}$ where $\mathcal{H}$ is linear (i.e. $\mathcal{H}$ is a $n \times n$ matrix).

2.7. Coordinate descent algorithm
Prove formula (2.38).

2.8. Prove the threshold formula (2.41) for the elastic net in the orthonormal case.

2.9. Show that the elastic net estimator for fixed λ_2 can be computed by using a Lasso algorithm.
Hint: Consider the definition in (2.39) and make an appropriate enlargement of the design matrix using the (additional) matrix $\sqrt{n\lambda_2}I_{p \times p}$.

Chapter 3
Generalized linear models and the Lasso

Abstract Generalized linear models build a unified framework containing many extensions of a linear model. Important examples include logistic regression for binary responses, Poisson regression for count data or log-linear models for contingency tables. Penalizing the negative log-likelihood with the ℓ_1-norm, still called the Lasso, is in many examples conceptually similar to the case with squared error loss in linear regression due to the convexity of the negative log-likelihood. This implies that the statistical properties as well as the computational complexity of algorithms are attractive. A noticeable difference, however, occurs with log-linear models for large contingency tables where the computation is in general much more demanding. We present in this chapter the models and estimators while computational algorithms and theory are described in more details in Chapters 4 and 6, respectively.

3.1 Organization of the chapter

This is a short chapter. After an introduction with the description of the Lasso and the adaptive Lasso for generalized linear models, we describe the loss functions for binary responses and logistic regression in Section 3.3.1, for Poisson regression in Section 3.3.2 and for multi-category responses with multinomial distributions in Section 3.3.3. We develop in Section 3.3.3.1 a few more details for the rather different case of contingency tables.

3.2 Introduction and preliminaries

Generalized linear models (GLMs) (McCullagh and Nelder, 1989) are very useful to treat many extensions of a linear model in a unified way. We consider a model

with univariate response Y and p-dimensional covariates $X \in \mathcal{X} \subseteq \mathbb{R}^p$:

$$Y_1, \ldots, Y_n \text{ independent}$$

$$g(\mathbb{E}[Y_i|X_i = x]) = \mu + \sum_{j=1}^{p} \beta_j x^{(j)}, \qquad (3.1)$$

where $g(\cdot)$ is a real-valued, known link function, μ denotes the intercept term and the covariates X_i are either fixed or random. We use the notation

$$f(x) = f_{\mu,\beta}(x) = \mu + \sum_{j=1}^{p} \beta_j x^{(j)}$$

to denote the linear predictor. An implicit assumption of the model in (3.1) is that the (conditional) distribution of Y_i (given X_i) is depending on X_i only through the function $g(\mathbb{E}[Y_i|X_i]) = f_{\mu,\beta}(X_i) = \mu + \sum_{j=1}^{p} \beta_j X_i^{(j)}$. That is, the (conditional) probability or density of $Y|X = x$ is of the form

$$p(y|x) = p_{\mu,\beta}(y|x). \qquad (3.2)$$

Obviously, a linear model is a special case of a generalized linear model with the identity link function $g(x) = x$. Other well-known examples are described below.

3.2.1 The Lasso estimator: penalizing the negative log-likelihood

For generalized linear models, the Lasso estimator is defined by penalizing the negative log-likelihood with the ℓ_1-norm.

The negative log-likelihood equals

$$-\sum_{i=1}^{n} \log(p_{\mu,\beta}(Y_i|X_i)),$$

where $p_{\mu,\beta}(y|x)$ is as in (3.2). This expression can be re-written (and scaled by the factor n^{-1}) as an empirical risk with a loss function $\rho(.,.)$:

$$n^{-1} \sum_{i=1}^{n} \rho_{\mu,\beta}(X_i, Y_i),$$

$$\rho_{\mu,\beta}(x, y) = -\log(p_{\mu,\beta}(y|x)).$$

For many examples and models, the loss function $\rho_{\mu,\beta}(x, y)$ is convex in μ, β for all values x, y. In particular, if the (conditional) distribution of $Y|X = x$ is from a subclass of the exponential family model (see McCullagh and Nelder (1989, Section

2.2)), we obtain convexity of $\rho_{\mu,\beta}(x,y) = \rho_{h(\mu,\beta)}(x,y)$ which depends on μ, β only through some linear function $h(\mu, \beta)$. Rather than striving for the most general set-up, we will present important examples below.

The ℓ_1-norm penalized Lasso estimator is then defined as:

$$\hat{\mu}(\lambda), \hat{\beta}(\lambda) = \underset{\mu,\beta}{\arg\min} \left(n^{-1} \sum_{i=1}^{n} \rho_{\mu,\beta}(X_i, Y_i) + \lambda \|\beta\|_1 \right).$$

Usually, we do not penalize the intercept term. Sometimes, we absorb μ (and $\hat{\mu}$) in the notation with β (and $\hat{\beta}$) where the intercept is then denoted by β_0 (and $\hat{\beta}_0$) and left unpenalized. Computation algorithms for solving the above optimization problem are described in Sections 4.7.1 and 4.7.2 in Chapter 4, noting that ℓ_1-norm penalization is a special case of the Group ℓ_1-penalty (described there).

The two-stage adaptive Lasso, introduced in Section 2.8 can also be used for this more general framework:

$$(\hat{\mu}(\lambda), \hat{\beta}(\lambda))_{\text{adapt}} = \underset{\mu,\beta}{\arg\min} \left(n^{-1} \sum_{i=1}^{n} \rho_{\mu,\beta}(X_i, Y_i) + \lambda \sum_{j=1}^{p} \frac{|\beta_j|}{|\hat{\beta}_{\text{init},j}|} \right),$$

where $\hat{\beta}_{\text{init}}$ is an initial estimator, for example from the Lasso above.

The properties of the Lasso in generalized linear models are very similar to the linear model case. We have again high-dimensional consistency, oracle inequalities (and hence optimality) and variable screening (and selection) properties. The theory can be derived in a similar fashion as for the Lasso in linear models, see Sections 6.3 - 6.8 in Chapter 6.

3.3 Important examples of generalized linear models

We discuss here a few prominent examples of generalized linear models which are often used in practice.

3.3.1 Binary response variable and logistic regression

Consider the case of logistic regression where $Y_i|X_i = x \sim$ Bernoulli($\pi(x)$), i.e., Binomial($1, \pi(x)$), with

$$\log \left(\frac{\pi(x)}{1 - \pi(x)} \right) = \mu + \sum_{j=1}^{p} \beta_j x^{(j)} = f_{\mu,\beta}(x).$$

This is a GLM with link function $g(\pi) = \log(\frac{\pi}{1-\pi})$, where $\pi \in (0,1)$.

The negative log-likelihood equals

$$-\sum_{i=1}^{n} \log(p_{\mu,\beta}(Y_i|X_i)) = \sum_{i=1}^{n} \left\{-Y_i f_{\mu,\beta}(X_i) + \log(1 + \exp(f_{\mu,\beta}(X_i)))\right\}, \quad (3.3)$$

see Problem 3.1, and the corresponding loss function is

$$\rho_{\mu,\beta}(x,y) = -y(\mu + \sum_{j=1}^{p} \beta_j x^{(j)}) + \log(1 + \exp(\mu + \sum_{j=1}^{p} \beta_j x^{(j)})).$$

In terms of the linear predictor $f(x)$, this loss function is of the form

$$\rho(x,y) = h_y(f(x)) = -yf + \log(1 + \exp(f)),$$

where we abbreviate $f(x) = f$ on the right hand side. This is a convex function in f since the first term is linear, the second term has positive second derivative and the sum of convex functions is convex. Furthermore, $f = f_{\mu,\beta}(x) = \mu + \sum_{j=1}^{p} \beta_j x^{(j)}$ is linear and hence

$$\rho_{\mu,\beta}(x,y) = h_y(f_{\mu,\beta}(x))$$

is convex in μ, β as a composition of a convex function $h_y(\cdot)$ (convex for all y) and a linear function.

For describing the margin-type interpretation, cf. Hastie et al. (2001) or Schölkopf and Smola (2002), we can rewrite the loss function as

$$\rho(f,y) = \log(1 + \exp(-(2y-1)f)) = \log(1 + \exp(-\tilde{y}f)),$$
$$\tilde{y} = 2y - 1 \in \{-1, 1\}, \quad (3.4)$$

see Problem 3.2. This formulation shows that the loss function is a function of the single argument $\tilde{y}f$, the so-called margin in binary classification. By scaling, the equivalent loss function is often used:

$$\rho(f,y) = \log_2(1 + \exp(-\tilde{y}f)), \quad (3.5)$$

which equals one at the value zero and hence, it becomes an upper bound of the misclassification error, see Figure 3.1. Regarding the latter, note that the natural classifier is

$$\mathscr{C}(x) = \begin{cases} 1 & \text{if } f(x) > 0, \\ 0 & \text{if } f(x) \leq 0 \end{cases}$$

since $f(x) > 0$ is equivalent to $\pi(x) > 0.5$. Hence the misclassification loss (where both misclassification errors are assigned a loss being equal to one) is given by

$$\rho_{\text{misclass}}(f, y) = 1(\tilde{y}f < 0), \ \tilde{y} = 2y - 1,$$

where $1(\cdot)$ denotes the indicator function.

Loss functions for binary classification
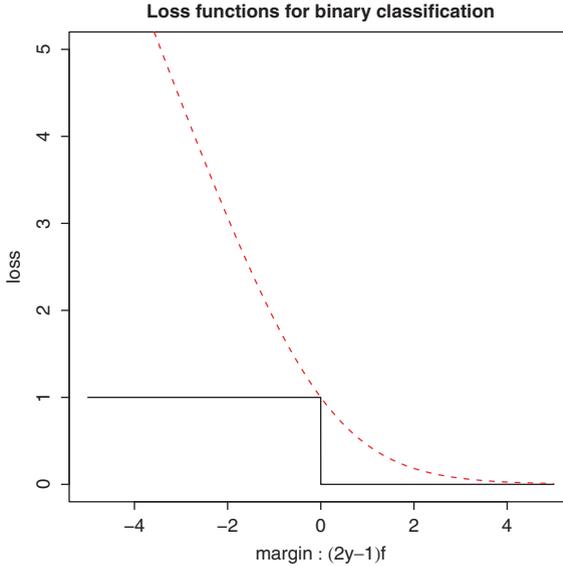
Fig. 3.1 Misclassification loss (solid line) and logistic loss in (3.5) (dashed line) as a function of the margin $\tilde{y}f = (2y - 1)f$.

3.3.2 Poisson regression

For a response variable Y taking values in $0, 1, 2, \ldots$, i.e., count data, we consider Poisson regression where the (conditional) distribution $Y_i | X_i = x \sim \text{Poisson}(\lambda(x))$. Using the link function

$$\log(\lambda(x)) = \mu + \sum_{j=1}^{p} \beta_j x^{(j)} = f_{\mu, \beta}(x)$$

we have a GLM as in (3.1).

The negative log-likelihood equals

$$-\sum_{i=1}^{n} \log(p_{\mu, \beta}(Y_i | X_i)) = \sum_{i=1}^{n} \{-Y_i f_{\mu, \beta}(X_i) + \exp(f_{\mu, \beta}(X_i))\},$$

and the corresponding loss function is

$$\rho_{\mu,\beta}(x,y) = -y(\mu + \sum_{j=1}^{p} \beta_j x^{(j)}) + \exp(\mu + \sum_{j=1}^{p} \beta_j x^{(j)}).$$

The first term is linear and hence convex in μ, β, the second term is a composition of a convex and a linear function and hence convex in μ, β, and since the sum of convex functions is convex, the loss function is convex in μ, β.

3.3.3 Multi-category response variable and multinomial distribution

The multinomial distribution is an example with a vector-valued link function. Consider a categorical response $Y \in \{0, 1, \ldots, k-1\}$ with labels $0, 1, \ldots, k-1$, as appearing in multi-category classification problems. We assume that the (conditional) distribution of $Y|X = x \sim \text{Multinom}(\pi(x))$, where $\pi(x) = \pi_0(x), \ldots, \pi_{k-1}(x)$ with $\sum_{r=0}^{k-1} \pi_r(x) = 1$ for all x. A link function

$$g : (0,1)^k \to \mathbb{R}^k, \ \pi = (\pi_0, \ldots, \pi_{k-1}) \mapsto f = (f_0, \ldots, f_{k-1})$$

is easier to describe by its inverse

$$g_r^{-1}(f) = \pi_r = \frac{\exp(f_r)}{\sum_{s=0}^{k-1} \exp(f_s)}, \ r = 0, \ldots, k-1.$$

This automatically ensures that $\sum_{r=0}^{k-1} \pi_r = 1$. Thus,

$$\log(\pi_r) = f_r - \log(\sum_{s=0}^{k-1} \exp(f_s)).$$

The linear predictors are parametrized as

$$f_r(x) = \mu_r + \sum_{j=1}^{p} \beta_{r;j} x^{(j)}, \ r = 0, \ldots, k-1.$$

Note that this is over-parametrized: it would suffice to determine say $f_1, \ldots, f_{k-1}$ (without f_0), but the constraint $\sum_{r=0}^{k-1} \pi_r(x) = 1$ for all x is automatically enforced with such an over-parametrized formulation.

The negative log-likelihood is

$$-\sum_{i=1}^{n} \sum_{r=0}^{k-1} \log(\pi_r(X_i)) \mathbb{1}(Y_i = r) = \sum_{i=1}^{n} \log(\sum_{s=0}^{k-1} \exp(f_s(X_i))) - \sum_{r=0}^{k-1} \mathbb{1}(Y_i = r) f_r(X_i),$$

$$f_r(X_i) = \mu_r + \sum_{j=1}^{p} \beta_{r;j} X_i^{(j)}.$$

The corresponding loss function is

$$\rho_{\mu,\beta}(x,y) = \log\left(\sum_{s=0}^{k-1} \exp(\mu_s + \sum_{j=1}^{p} \beta_{s;j} x^{(j)})\right) - \sum_{r=0}^{k-1} 1(y = r)(\mu_r + \sum_{j=1}^{p} \beta_{r;j} x^{(j)}).$$

This is again a convex function in $\{\mu_r, \beta_{r;j}; \ r = 0, \ldots, k-1, \ j = 1, \ldots, p\}$. The reasoning is as follows. The second term includes linear functions only and hence convexity follows since the sum of convex functions is convex. The first term is of the form

$$\log\left(\sum_{s=0}^{k-1} \exp(\mu_s + \sum_{j=1}^{p} \beta_{s;j} x^{(j)})\right) = \log\left(\sum_s \exp(f_s(\mu_s, \beta_s))\right), \quad f_s = \mu_s + \sum_{j=1}^{p} \beta_{s;j} x^{(j)}.$$

The so-called "log-sum-exp" function, see Section 3.1.5 in Boyd and Vandenberghe (2004),

$$\log\left(\sum_s \exp(f_s)\right) \tag{3.6}$$

is convex in $f_0, \ldots, f_{k-1}$, see Problem 3.3. Hence, the composition of linear functions $f_s(\mu_s, \beta_s)$ ($s = 0, \ldots, k-1$) with the convexity of the "log-sum-exp" function (see Problem 3.3) implies that the first term is convex in the parameters $\{\mu_r, \beta_{r;j}; \ r = 0, \ldots, k-1, \ j = 1, \ldots, p\}$ as well, and hence we have convexity of the loss function (since sums of convex functions are convex).

3.3.3.1 Contingency tables

The multinomial distribution also arises when modeling contingency tables. Consider q categorical factor variables $Z^{(1)}, \ldots, Z^{(q)}$ where each factor $Z^{(j)} \in \mathscr{I}^{(j)}$, $\mathscr{I}^{(j)}$ denoting a categorical space of $d^{(j)}$ levels (labels). Thus, the q factors take values in the categorical space

$$\mathscr{I} = \mathscr{I}^{(1)} \times \ldots \times \mathscr{I}^{(q)},$$

and we can enumerate $\mathscr{I} = \{r; \ r = 0, 1, \ldots, k-1\}$ where $k = \sum_{j=1}^{q} |\mathscr{I}^{(j)}|$. We then denote by

$$Y = (Z^{(1)}, \ldots, Z^{(q)}) \in \mathscr{I}.$$

The observations in a contingency table are $Y_1, \ldots, Y_n$ i.i.d. with $Y_i \in \mathscr{I}$ and $Y_i \sim$ Multinom(π) with $k = |\mathscr{I}|$-dimensional π satisfying $\sum_{r=0}^{k-1} \pi_r = 1$. Very often, a

log-linear model is used:

$$\log(\pi) = \mu + \mathbf{X}\beta,$$

with $k \times p$ ($k = |\mathscr{I}|$) design matrix $\mathbf{X}$ which encodes the full saturated model (with $p = k$) or some sub-model including only interaction terms up to a certain order (with $p < k$). Typically, an intercept term μ is used to ensure that $\sum_{r=0}^{k-1} \pi_r = 1$. This can be enforced in the same way as for multinomial regression described above. We use

$$\pi_r = \frac{\exp(\mu + (\mathbf{X}\beta)_r)}{\sum_{s \in \mathscr{I}} \exp(\mu + (\mathbf{X}\beta)_s)}, \ r \in \mathscr{I} \tag{3.7}$$

which implies

$$\log(\pi_r) = \mu + (\mathbf{X}\beta)_r - \log(\sum_{s \in \mathscr{I}} \exp(\mu + (\mathbf{X}\beta)_s)), \ r \in \mathscr{I}.$$

With the parametrization in (3.7), the negative log-likelihood equals

$$-\sum_{i=1}^{n} \log(\mathrm{p}_{\mu,\beta}(Y_i)) = -\sum_{i=1}^{n} \sum_{r \in \mathscr{I}} 1(Y_i = r)\{\mu + (\mathbf{X}\beta)_r - \log(\sum_{s \in \mathscr{I}} \exp(\mu + (\mathbf{X}\beta)_s))\},$$
$$\tag{3.8}$$

see Problem 3.4, and the corresponding loss function, involving y only, is

$$\rho_{\mu,\beta}(y) = \log(\sum_{s \in \mathscr{I}} \exp(\mu + (\mathbf{X}\beta)_s)) - \sum_{r \in \mathscr{I}} 1(y = r)(\mu + (\mathbf{X}\beta)_r).$$

The loss function is convex in μ, β by the same argument as for the corresponding loss for multinomial regression.

The Lasso estimator is then

$$\hat{\mu}, \hat{\beta} = \arg\min_{\mu,\beta} \left(n^{-1} \sum_{i=1}^{n} \rho_{\mu,\beta}(Y_i) + \lambda \|\beta\|_1 \right).$$

This Lasso estimator has the interesting property that it can be used for problems where many cells have zero counts, i.e., $\sum_{i=1}^{n} 1(Y_i = r) = 0$ for many $r \in \mathscr{I}$, which arises when having a moderate number q of factors implying that $k = |\mathscr{I}|$ is very large. This is in sharp contrast to the unpenalized maximum likelihood estimator, see Problem 3.5. From a conceptual point of view, one would often aim for an estimator where whole main or interactions terms (with respect to the structure of the factors $Z^{(1)}, \ldots, Z^{(q)}$) are zero or not: this can be naturally achieved with the group Lasso described in Chapter 4, see Dahinden et al. (2007).

A major drawback of the Lasso estimator as defined above (also without penalty; and also of the Group Lasso) is its computational cost. Even when restricting the

model to lower-order interactions (with $p < k$), the row-dimension of $\mathbf{X}$ remains to be $k = |\mathscr{I}|$ and the computation of the estimator is at least linear in k. Thus, this naive Lasso strategy can only work for say k up to say 10^6. For example, if every factor has 2 levels only, this would require approximately $2^q \leq 10^6$ and hence $q \leq \log_2(10^6) \approx 20$: that is, we cannot handle more than 20 factors with such an approach. For special cases with binary factor variables, fast componentwise ℓ_1-penalization is possible (Ravikumar et al., 2009b). Alternatively, decomposition approaches based on graphical models can be used (Dahinden et al., 2010).

Problems

3.1. Derive the negative log-likelihood in (3.3) for the binary response case with the logistic link function.

3.2. Derive formula (3.4), i.e., the margin point of view of logistic regression.

3.3. Prove that the log-sum-exp function in (3.6) is a convex function in its k arguments $f_0, \ldots, f_{k-1}$. Hint: Prove this by directly verifying the definition of a convex function

$$f(ax + (1-a)y) \leq af(x) + (1-a)f(y)$$

for all x, y, $0 \leq a \leq 1$.

3.4. Derive the negative log-likelihood in (3.8) for contingency tables.

3.5. Consider a contingency table as in Section 3.3.3.1.
(a) Assume that all cell counts are non-zero, i.e., $\sum_{i=1}^n 1(Y_i = r) \neq 0$ for all $r \in \mathscr{I}$. Derive the maximum likelihood estimator for π_r $(r = 0, \ldots, k-1)$.
(b) Construct an example of a contingency table where the maximum likelihood estimator does not exist.
Hint: use an example with zero cell counts.

Chapter 4
The group Lasso

Abstract In many applications, the high-dimensional parameter vector carries a structure. Among the simplest is a group structure where the parameter is partitioned into disjoint pieces. This occurs when dealing with factor variables or in connection with basis expansions in high-dimensional additive models as discussed in Chapters 5 and 8. The goal is high-dimensional estimation in linear or generalized linear models being sparse with respect to whole groups. The group Lasso, proposed by Yuan and Lin (2006) achieves such group sparsity. We discuss in this chapter methodological aspects, and we develop the details for efficient computational algorithms which are more subtle than for non-group problems.

4.1 Organization of the chapter

We present in this chapter the group Lasso penalty and its use for linear and generalized linear models. The exposition is primarily from a methodological point of view but some theoretical results are loosely described to support methodology and practice. After an introduction in Section 4.2 with the definition of the group Lasso penalty, we present in Section 4.3 the important case with factor variables including a specific example. In Section 4.4 we sketch the statistical properties of the group Lasso estimator while a mathematically rigorous treatment is presented later in Chapter 8. In Section 4.5 we discuss a slight generalization of the group Lasso penalty which is more flexible and we explain more about parametrizations and their invariances. In Section 4.7 we give a detailed treatment of computational algorithms for the Group Lasso. Thereby, the case with squared error loss is substantially simpler than for non-squared error losses as arising in generalized linear models.

4.2 Introduction and preliminaries

In some applications, a high-dimensional parameter vector β in a regression model is structured into groups $\mathcal{G}_1, \ldots, \mathcal{G}_q$ which build a partition of the index set $\{1, \ldots, p\}$. That is, $\cup_{j=1}^{q} \mathcal{G}_j = \{1, \ldots, p\}$ and $\mathcal{G}_j \cap \mathcal{G}_k = \emptyset$ $(j \neq k)$. The parameter vector β then carries the structure

$$\beta = (\beta_{\mathcal{G}_1}, \ldots, \beta_{\mathcal{G}_q}), \quad \beta_{\mathcal{G}_j} = \{\beta_r; \ r \in \mathcal{G}_j\}. \tag{4.1}$$

An important class of examples where some group structure occurs is in connection with factor variables. For example, consider a real-valued response variable Y and p categorical covariates $X^{(1)}, \ldots, X^{(p)}$ where each $X^{(j)} \in \{0, 1, 2, 3\}$ has 4 levels encoded with the labels $0, 1, 2, 3$. Then, for encoding a main effect describing the deviation from the overall mean, we need 3 parameters, encoding a first-order interaction requires 9 parameters and so on. Having chosen such a parametrization, e.g., with sum contrasts, the group structure is as follows. The main effect of $X^{(1)}$ corresponds to $\beta_{\mathcal{G}_1}$ with $|\mathcal{G}_1| = 3$; and likewise, the main effect of all other factors $X^{(j)}$ corresponds to $\beta_{\mathcal{G}_j}$ with $|\mathcal{G}_j| = 3$ for all $j = 1, \ldots, p$. Furthermore, a first-order interaction of $X^{(1)}$ and $X^{(2)}$ corresponds to $\beta_{\mathcal{G}_{p+1}}$ with $|\mathcal{G}_{p+1}| = 9$, and so on. More details are described in Section 4.3.

Another example is a nonparametric additive regression model where the groups $\mathcal{G}_j$ correspond to basis expansions for the jth additive function of the jth covariate $X^{(j)}$. A detailed treatment is given in Chapter 5.

4.2.1 The group Lasso penalty

When estimating models with a group structure for the parameter vector, we often want to encourage sparsity on the group-level. Either all entries of $\hat{\beta}_{\mathcal{G}_j}$ should be zero or all of them non-zero. This can be achieved with the group Lasso penalty

$$\lambda \sum_{j=1}^{q} m_j \|\beta_{\mathcal{G}_j}\|_2, \tag{4.2}$$

where $\|\beta_{\mathcal{G}_j}\|_2$ denotes the standard Euclidean norm. The multiplier m_j serves for balancing cases where the groups are of very different sizes. Typically we would choose

$$m_j = \sqrt{T_j},$$

where T_j denotes the cardinality $|\mathcal{G}_j|$.

The group Lasso estimator in a linear or generalized linear model as in (2.1) or (3.1) respectively is then defined as

$$\hat{\beta}(\lambda) = \arg\min_{\beta} Q_\lambda(\beta),$$

$$Q_\lambda(\beta) = n^{-1}\sum_{i=1}^{n}\rho_\beta(X_i, Y_i) + \lambda\sum_{j=1}^{q}m_j\|\beta_{\mathscr{G}_j}\|_2, \tag{4.3}$$

where $\rho_\beta(x, y)$ is a loss function which is convex in β. Examples are $\rho_\beta(x, y) = |y - x\beta|^2$ or one of the loss functions described in Chapter 3 where $\rho_\beta(x, y) = -\log_\beta(\mathrm{p}(y|x))$ with $\mathrm{p}(\cdot|x)$ denoting the density of Y given $X = x$. We discuss in Section 4.5.1 a different penalty which is invariant under reparametrization whereas in (4.3), we only have invariance with respect to orthonormal reparametrizations within each group $\mathscr{G}_j$. As in Chapter 3, we often include an unpenalized intercept term: the estimator is then

$$\hat{\mu}(\lambda), \hat{\beta}(\lambda) = \arg\min_{\mu,\beta} Q_\lambda(\mu,\beta),$$

$$Q_\lambda(\mu,\beta) = n^{-1}\sum_{i=1}^{n}\rho_{\mu,\beta}(X_i, Y_i) + \lambda\sum_{j=1}^{q}m_j\|\beta_{\mathscr{G}_j}\|_2. \tag{4.4}$$

In the sequel, we often focus on the notationally simpler case without intercept; in practice the intercept term is often important but there is no conceptual difficulty in including it (in an unpenalized way) as described in (4.4).

Lemma 4.1. *Assume that $\rho_\beta(X_i, Y_i) \geq C > -\infty$ for all β, X_i, Y_i $(i = 1,\ldots,n)$ and $\rho_\beta(x, y)$ is a convex function in β for all X_i, Y_i $(i = 1,\ldots,n)$. Then, for $\lambda > 0$ and $m_j > 0$ for all j, the minimum in the optimization problem (4.3) is attained.*

Proof. Because $Q_\lambda(\beta)$ is continuous and $Q_\lambda(\beta) \to \infty$ as $\|(\beta_{\mathscr{G}_1},\ldots,\beta_{\mathscr{G}_q})\|_2 \to \infty$, the minimum is attained. □

The boundedness assumption in the lemma is very mild and holds for the commonly used loss functions for (generalized) regression or classification. Furthermore, we could replace the convexity assumption by requiring instead a continuous loss function (in β).

The group Lasso estimator has the following properties. Depending on the value of the regularization parameter λ, the estimated coefficients within a group $\mathscr{G}_j$ satisfy: either $(\hat{\beta}_{\mathscr{G}_j})_r \equiv 0$ for all components $r = 1,\ldots,T_j$ or $(\hat{\beta}_{\mathscr{G}_j})_r \neq 0$ for all components $r = 1,\ldots,T_j$. This is a consequence of the non-differentiability of the $\sqrt{\cdot}$ function at zero: an exact characterization of the solutions of the optimization problem in (4.3) is given in Lemma 4.2 Furthermore, with trivial groups consisting of singletons $\mathscr{G}_j = \{j\}$ for all $j = 1,\ldots,q = p$, and using $m_j = T_j \equiv 1$, the penalty function in (4.2) equals the standard Lasso penalty. Finally, the group Lasso penalty is invariant under orthonormal transformations within the groups. Often we would choose any

orthonormal basis for parametrization leading to orthonormal sub-matrices $\mathbf{X}_{\mathcal{G}_j}^T \mathbf{X}_{\mathcal{G}_j}$ for each group $\mathcal{G}_j$ ($\mathbf{X}_{\mathcal{G}_j}$ denotes the $n \times T_j$ submatrix of $\mathbf{X}$ whose columns correspond to $\mathcal{G}_j$). This has computational advantages (see Section 4.7.1.1) but one should keep in mind that in principle, the estimator in general depends on the possibly non-orthonormal parametrizations.

The group Lasso estimator has similar qualitative properties as the Lasso. It exhibits good accuracy for prediction and parameter estimation, and it has the groupwise variable screening property saying that all relevant groups with corresponding parameter vector $\beta_{\mathcal{G}} \neq 0$ are also estimated as active groups with corresponding parameter vector $\hat{\beta}_{\mathcal{G}} \neq 0$. We give more details in Section 4.4 and present rigorous mathematical theory in Chapter 8.

4.3 Factor variables as covariates

As mentioned earlier at the beginning of Chapter 4, grouping of the parameter vector occurs naturally with factor variables. We consider here the simple case with just two covariates $X^{(1)}, X^{(2)} \in \{0, 1, 2, 3\}$, where $\{0, 1, 2, 3\}$ denotes a set of four categorical labels, i.e., we consider two factors each having 4 levels. Consider a linear model with real-valued response Y and dummy variables encoding the contribution of the two factors:

$$Y_i = \mu + \sum_{k=0}^{3} \gamma_k \mathbb{1}(X_i^{(1)} = k) + \sum_{k=0}^{3} \delta_k \mathbb{1}(X_i^{(2)} = k)$$

$$+ \sum_{k,\ell=0}^{3} \kappa_{k,\ell} \mathbb{1}(X_i^{(1)} = k, X_i^{(2)} = \ell) + \varepsilon_i \quad (i = 1, \ldots, n), \tag{4.5}$$

where we assume sum-constraints $\sum_k \gamma_k = \sum_k \delta_k = 0, \sum_k \kappa_{k,\ell} = \sum_\ell \kappa_{k,\ell} = 0$ for all k, ℓ, $\mathbb{1}(\cdot)$ denotes the indicator function and $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. variables with $\mathbb{E}[\varepsilon_i] = 0$. This model can be parametrized as

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon, \tag{4.6}$$

with $\mathbf{Y} = (Y_1, \ldots, Y_n)$, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ and $n \times 16$ design matrix $\mathbf{X}$ which ensures the sum-constraints from above.

The parametrization in (4.6) can be achieved as follows. A first model matrix $\tilde{\mathbf{X}}$ can be constructed which ensures the sum-constraints (by dropping the redundant parameters and keeping the non-redundant parameters in the model). For example, in the R software environment for statistical computing, the function `model.matrix` provides such a first design matrix $\tilde{\mathbf{X}}$ (see also Problem 4.1). Next, we center all columns of $\tilde{\mathbf{X}}$ to mean zero. This is appropriate whenever we do not want to penalize the intercept term (thus, we project onto the space of variables which are not

penalized). Afterward, we parametrize using orthonormal bases for the sub-spaces corresponding to the two main effects (parametrized in (4.5) with γ, δ) and to the interaction effect (parametrized in (4.5) with κ). As a result, we end up with a design matrix $\mathbf{X}$ as in (4.6) and we can apply the group Lasso for estimation of β. It is worth pointing out that the sum-constraint plays no special role here: other constraints such as Helmert contrasts can be parametrized with orthonormal bases for the sub-spaces of the main effects and interactions. Since the group Lasso penalty is invariant under orthonormal transformations of the parameter vector, the estimation results (for $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}$) are not affected by the choice of the contrast.

4.3.1 Prediction of splice sites in DNA sequences

Prediction of short DNA motifs plays an important role in many areas of computational biology. Gene finding algorithms such as GENIE (Burge and Karlin, 1997) often rely on the prediction of splice sites. Splice sites are the regions between coding (exons) and non-coding (introns) DNA segments. The 5' end of an intron is called a donor splice site and the 3' end an acceptor splice site. A donor site whose first two intron positions are the letters "GT" is called canonical, whereas an acceptor site is called canonical if the corresponding intron ends with "AG". An overview of the splicing process and of some models that are used for detecting splice sites can be found in Burge (1998).

We analyze here the so-called MEMset Donor dataset. It consists of a training set of 8'415 true (encoded as $Y = 1$) and 179'438 false (encoded as $Y = 0$) human donor sites. An additional test set contains 4'208 true and 89'717 false donor sites. The covariates or predictor variables are 7 factors with values in $\{A, C, G, T\}^7$, namely 3 bases from the exon and 4 bases from the intron part. The data are available at `http://genes.mit.edu/burgelab/maxent/ssdata/`. A more detailed description can be found in Yeo and Burge (2004).

We fit a logistic regression model using the group Lasso penalty for the main effects and higher-order interactions among the 7 factors $X^{(1)}, \ldots, X^{(7)}$. For $\pi(x) = \mathbf{P}[Y = 1 | X = x]$, we model $\text{logit}(\pi(x))$ analogously as in (4.5), but now in the logistic setting with 7 factors. We use the sum-constraint as encoding scheme for the dummy variables, i.e., the coefficients have to add up to zero. The entire predictor space has dimension $4^7 = 16'384$ but we restrict ourselves to interactions of at most order 2 which are sometimes also called 3-way interactions. After reparametrization with orthonormal bases for all groups $\mathscr{G}_j$ corresponding to the sub-spaces from main effects or interaction terms, we end up with a model

$$\text{logit}(\pi) = \mu + \mathbf{X}\beta$$

with $n \times 1155$ design matrix $\mathbf{X}$. We then use the group Lasso estimator

$$\hat{\beta}(\lambda) = \underset{\beta}{\arg\min} \left(-\ell(\beta; Y_1, \ldots, Y_n) + \lambda \sum_{j=1}^{q} \sqrt{T_j} \|\beta_{\mathscr{G}_j}\|_2 \right), \qquad (4.7)$$

where the intercept μ is unpenalized and $T_j = |\mathscr{G}_j|$.

The original training dataset is used to build a smaller balanced training dataset (5'610 true, 5'610 false donor sites) and an unbalanced validation set (2'805 true, 59'804 false donor sites). All sites are chosen randomly without replacement such that the two sets are disjoint. The additional test set (4'208 true and 89'717 false donor sites) remains unchanged. Note that the ratios of true to false sites (i.e. $Y = 1$ or $Y = 0$, respectively) are equal for the validation and the test set.

All models are fitted on the balanced training dataset. As the ratio of true to false splice sites strongly differs from the training to the validation and the test set, the intercept is corrected as follows (King and Zeng, 2001):

$$\hat{\mu}_0^{corr} = \hat{\mu} - \log\left(\frac{\bar{y}}{1 - \bar{y}}\right) + \log\left(\frac{\pi_{\text{val}}}{1 - \pi_{\text{val}}}\right),$$

where π_{val} is the proportion of true sites in the validation set. The penalty parameter λ is chosen according to the (unpenalized) log-likelihood score on the validation set using the corrected intercept estimate.

For a threshold $\tau \in (0, 1)$ we assign observation i to class 1 if $\pi_{\hat{\mu}^{corr}, \hat{\beta}}(x_i) > \tau$ and to class 0 otherwise. Note that the class assignment can also be constructed without intercept correction by using a different threshold.

The correlation coefficient ρ_τ corresponding to a threshold τ is defined as the Pearson correlation between the binary random variable of the true class membership and the binary random variable of the predicted class membership. In Yeo and Burge (2004) the maximal correlation coefficient

$$\rho_{\max} = \max\{\rho_\tau \mid \tau \in (0, 1)\}$$

is used as a goodness of fit statistics on the test set.

The candidate model that was used for the Logistic group Lasso consists of all 3-way and lower order interactions involving 64 terms resulting in $p = 1156$ parameters. Such a group Lasso fitted model achieves $\rho_{max} = 0.6593$ on the test set which is very competitive with published results from Yeo and Burge (2004) whose best ρ_{max} equals 0.6589 based on a maximum entropy approach.

In the spirit of the adaptive Lasso in Section 2.8 or the relaxed Lasso in 2.10, we consider here also some two-stage procedures. Instead of an adaptive group ℓ_1-penalization (see Section 4.6), we consider the following. The first stage is group Lasso yielding a parameter vector $\hat{\beta}(\lambda_{\text{init}})$. Denote by $\hat{S}(\lambda_{\text{init}}) = \{j; \hat{\beta}_j(\lambda_{\text{init}}) \neq 0\}$ which is the set of variables from the selected groups. In the second stage, we either use maximum likelihood estimation (group Lasso/MLE hybrid) or ℓ_2-penalization (group Lasso/Ridge hybrid) on the reduced space given by the selected variables

(from the selected groups) in $\hat{S}(\lambda_{\mathrm{init}})$. The latter amounts to the following: when splitting the parameter vector into the components $(\beta_{\hat{S}(\lambda_{\mathrm{init}})}, \beta_{\hat{S}(\lambda_{\mathrm{init}})^c})$ where the estimator $\hat{\beta}(\lambda_{\mathrm{init}})$ is non-zero and zero, respectively, we define:

$$\hat{\beta}_{\hat{S}(\lambda_{\mathrm{init}})}(\lambda_{\mathrm{init}}, \lambda_{\mathrm{hybrid}})$$
$$= \underset{\beta_{\hat{S}(\lambda_{\mathrm{init}})}}{\arg\min} \left(-\ell((\beta_{\hat{S}(\lambda_{\mathrm{init}})}, 0_{\hat{S}(\lambda_{\mathrm{init}})^c}); Y_1, \ldots, Y_n) + \lambda_{\mathrm{hybrid}} \|\beta_{\hat{S}(\lambda)}\|_2^2 \right),$$

and for $\lambda_{\mathrm{hybrid}} = 0$, we have the group Lasso/MLE hybrid. The penalty parameters λ_{init} and $\lambda_{\mathrm{hybrid}}$ are again chosen according to the (unpenalized) log-likelihood score on the validation set using the corrected intercept estimate.

In terms of predictive accuracy, there is no benefit when using such two-stage procedures. On the other hand, while the group Lasso solution has some active 3-way interactions, the group Lasso/Ridge hybrid and the Group Lasso/MLE hybrid only contain 2-way interactions. Figure 4.1 shows the ℓ_2-norms of each parameter group for the three estimators. The 3-way interactions of the group Lasso solution seem to be very weak, and the two-stage procedures select slightly fewer terms. Decreasing the candidate model size at the beginning to only contain 2-way interactions gives similar results which are not shown here.

In summary, the prediction performance of the group Lasso estimate in a simple logistic regression factor model is competitive with a maximum entropy approach that was used in Yeo and Burge (2004). Advantages of the group Lasso include selection of terms corresponding to main effects and interactions, and the logistic model is a natural framework to include other predictor variables.

4.4 Properties of the group Lasso for generalized linear models

Recall that we denote by $f^0(x) = x\beta^0$ and $\hat{f}_\lambda(x) = x\hat{\beta}(\lambda)$ the linear predictor and its estimate in a generalized linear model as in (3.1). Furthermore, we denote the set of true underlying active groups by

$$S_{\mathrm{group}}^0 = \{j; \beta_{\mathscr{G}_j}^0 \neq 0\},$$

with the understanding that $\beta_{\mathscr{G}_j}^0 \neq 0$ if there exists at least one component say k where $(\beta_{\mathscr{G}_j}^0)_k \neq 0$. The cardinality of the set of active groups is $s_0 = |S_{\mathrm{group}}^0|$.

For prediction, when choosing an appropriate regularization parameter λ, the group Lasso estimator is consistent in high-dimensional settings where $p = p_n$ is of much larger order than sample size n:

$$(\hat{\beta}(\lambda) - \beta^0)^T \Sigma_X (\hat{\beta}(\lambda) - \beta^0) = o_P(1) \ (n \to \infty),$$
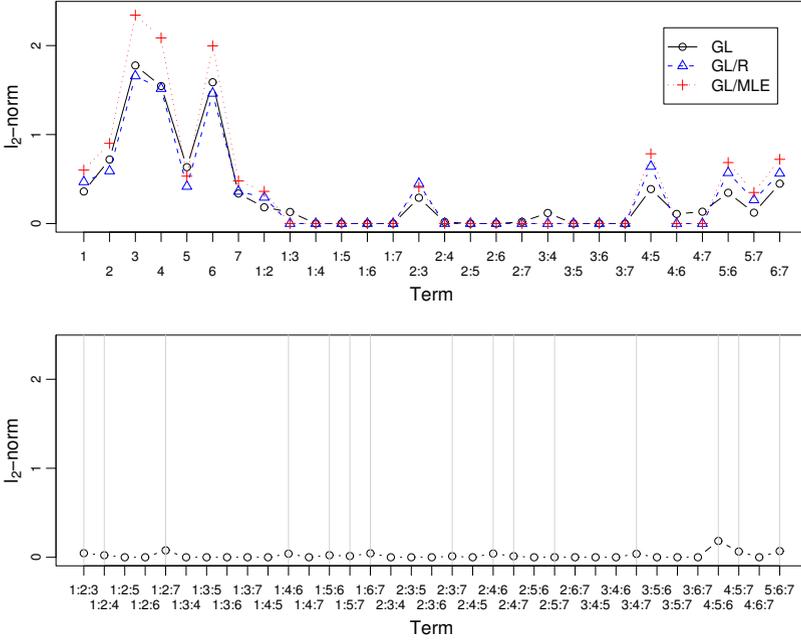
Fig. 4.1 DNA splice site prediction. ℓ_2-norms $\|\hat{\beta}_{\mathcal{G}_j}\|_2$, $j \in \{1, \ldots, q\}$ of the parameter groups with respect to the groupwise orthonormalized design matrix when using a candidate model with all 3-way interactions. $i : j : k$ denotes the 3-way interaction between the ith, jth and kth sequence position. The same scheme applies to the 2-way interactions and the main effects. Active 3-way interactions are additionally marked with vertical lines. The figure is taken from Meier et al. (2008).

where Σ_X is $n^{-1}\mathbf{X}^T\mathbf{X}$ in case of a fixed design or equals the covariance of the covariate X in case of a random design. Thereby, the asymptotic framework is set-up in an analogous way as in (2.6), and we implicitly assume here a mild condition on the distribution of $Y|X$ (e.g. sub-Gaussian distribution). Recall that the quantity on the left-hand side can be interpreted as

$$\text{for fixed design}: \quad n^{-1}\sum_{i=1}^{n}(\hat{f}(X_i) - f^0(X_i))^2 = \|\mathbf{X}(\hat{\beta}(\lambda) - \beta^0)\|_2^2/n,$$
$$\text{for random design}: \quad \mathbb{E}[(\hat{f}(X_{new}) - f^0(X_{new}))^2] = \mathbb{E}[\{X_{new}(\hat{\beta}(\lambda) - \beta^0)\}^2],$$

where $\mathbb{E}$ is with respect to the new test observation X_{new}. Under an additional compatibility assumption on the design matrix $\mathbf{X}$, and using $\lambda \asymp n^{-1/2}(1 \vee \sqrt{\log(q)/T})$, we obtain the convergence rate

$$(\hat{\beta}(\lambda) - \beta_0)^T \Sigma_X (\hat{\beta}(\lambda) - \beta_0) = O_P\left(\frac{s_0}{n\phi^2}(\log(q) \vee T)\right). \tag{4.8}$$

Here, for simplicity, we assumed equal group-size $T \equiv |\mathcal{G}_j|$ for all $j = 1, \ldots, q$, where q and s_0 denote the number of groups or active groups, respectively. Furthermore, ϕ^2 is a so-called compatibility constant, depending on the compatibility of the design, which at best is bounded below by a positive constant. In addition to the prediction error above, assuming a compatibility condition on the design matrix $\mathbf{X}$, the estimation error is

$$\sum_{j=1}^{q} \|\hat{\beta}_{\mathcal{G}_j}(\lambda) - \beta^0_{\mathcal{G}_j}\|_2 = O_P\left(\frac{s_0}{\sqrt{n}\phi^2}\sqrt{\log(q) \vee T}\right). \tag{4.9}$$

Mathematical details are given in Theorem 8.1 in Chapter 8.

When comparing the convergence rate in (4.8) with (2.8) for the Lasso (where for the latter, the number ϕ^2 is typically smaller than here), we see that we gain a $\log(p)$-factor if T is larger than $\log(q)$ (noting that $s_0 T$ corresponds to the number of non-zero regression coefficients). We also see from (4.8) that if the group-sizes are large, say in the order of sample size n, the group Lasso is not consistent for prediction. For such cases, we need additional assumptions such as smoothness to achieve consistency of predictions. This is treated in greater detail in Chapters 5 and 8.

The variable screening property on the groupwise level, analogous to the description in Section 2.5, also holds for the Group Lasso. As before, we denote by $S^0_{\mathrm{group}} = \{j; \beta^0_{\mathcal{G}_j} \neq 0\}$ the set of groups whose corresponding coefficient vector is not equal to the 0-vector (i.e. at least one component is different from zero) and analogously, $\hat{S}_{\mathrm{group}}(\lambda)$ is the estimated version using the group Lasso estimator. Then, for suitable $\lambda = \lambda_n$, typically $\lambda_n \asymp n^{-1/2}(1 \vee \sqrt{\log(q)/T})$ (see Theorem 8.1):

$$\mathbf{P}[\hat{S}_{\mathrm{group}}(\lambda) \supseteq S^0_{\mathrm{group}}] \to 1 \ (n \to \infty). \tag{4.10}$$

Such a result follows from the convergence rate (4.9) for $\sum_{j=1}^{q} \|\hat{\beta}_{\mathcal{G}_j}(\lambda) - \beta^0_{\mathcal{G}_j}\|_2$ (Theorem 8.1) and assuming that the smallest non-zero group norm $\inf_j\{\|\beta^0_{\mathcal{G}_j}\|_2; \ j \in S^0_{\mathrm{group}}\}$ is larger than a certain detection limit. More details are given in Theorem 8.1 in Chapter 8. Usually, when choosing $\hat{\lambda}_{CV}$ from cross-validation, the screening property in (4.10) still holds, in analogy to the results from Section 2.5.1 for the Lasso. The variable screening property on the groupwise level from (4.10) is very useful to do effective dimensionality reduction while keeping the relevant groups in the model. Typically, the number of groups $|\hat{S}_{\mathrm{group}}|$ is much smaller than the total number q of groups. Furthermore, if the group-sizes are relatively small, the total number of parameters in $\hat{S}_{\mathrm{group}}$ is often smaller than sample size n. As pointed out above, if the group-sizes are large, additional smoothness assumptions still yield statistically meaningful (or even optimal) results. This topic is treated in greater detail in Chapters 5 and 8 (Sections 8.4 and 8.5). We emphasize that in addition to a prediction gain when $T > \log(q)$ with the group Lasso in comparison to the Lasso, it is worthwhile to use it since it encourages sparsity for whole groups, and cor-

responding group selection may be desirable in practical applications, for example when dealing with factor variables.

4.5 The generalized group Lasso penalty

The group Lasso penalty in (4.2) is

$$\lambda \sum_{j=1}^{q} m_j \|\beta_{\mathcal{G}_j}\|_2 = \lambda \sum_{j=1}^{q} m_j \sqrt{\beta_{\mathcal{G}_j}^T \beta_{\mathcal{G}_j}}.$$

In some applications, we want a penalty of the form

$$\lambda \sum_{j=1}^{q} m_j \sqrt{\beta_{\mathcal{G}_j}^T A_j \beta_{\mathcal{G}_j}}, \tag{4.11}$$

where A_j are positive definite $T_j \times T_j$ matrices. A concrete example is an additive model treated in more detail in Chapter 5.

Due to the fact that A_j is positive definite, we can reparametrize:

$$\tilde{\beta}_{\mathcal{G}_j} = A_j^{1/2} \beta_{\mathcal{G}_j},$$

and hence, an ordinary group Lasso penalty arises of the form

$$\lambda \sum_{j=1}^{q} m_j \|\tilde{\beta}_{\mathcal{G}_j}\|_2.$$

The matrix $A_j^{1/2}$ can be derived using e.g. the Cholesky decomposition $A_j = R_j^T R_j$ for some quadratic matrix R_j which we denote by $A_j^{1/2} = R_j$. Of course, we also need to reparametrize the (generalized) linear model part:

$$\mathbf{X}\beta = \sum_{j=1}^{q} \mathbf{X}_{\mathcal{G}_j} \beta_{\mathcal{G}_j}.$$

The reparametrization is then for every sub-design matrix $\mathbf{X}_{\mathcal{G}_j}$:

$$\tilde{\mathbf{X}}_{\mathcal{G}_j} = \mathbf{X}_{\mathcal{G}_j} R_j^{-1} = \mathbf{X}_{\mathcal{G}_j} A_j^{-1/2}, \ j = 1, \ldots, q$$

such that $\mathbf{X}\beta = \sum_{j=1}^{q} \tilde{\mathbf{X}}_{\mathcal{G}_j} \tilde{\beta}_{\mathcal{G}_j}$.

The generalized group Lasso estimator in a linear model is then defined by:

$$\hat{\beta} = \underset{\beta}{\arg\min} \left(\|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + \lambda \sum_{j=1}^{q} m_j \sqrt{\beta_{\mathscr{G}_j}^T A_j \beta_{\mathscr{G}_j}} \right).$$

Equivalently, we have:

$$\hat{\tilde{\beta}}_{\mathscr{G}_j} = A_j^{-1/2} \hat{\tilde{\beta}}_{\mathscr{G}_j},$$

$$\hat{\tilde{\beta}} = \underset{\tilde{\beta}}{\arg\min} \left(\|\mathbf{Y} - \sum_{j=1}^{q} \tilde{\mathbf{X}}_{\mathscr{G}_j} \tilde{\beta}_{\mathscr{G}_j}\|_2^2/n + \lambda \sum_{j=1}^{q} m_j \|\tilde{\beta}_{\mathscr{G}_j}\|_2 \right).$$

4.5.1 Groupwise prediction penalty and parametrization invariance

The Lasso and group Lasso estimator depend on the parametrization used in a linear or generalized linear model. The estimated active set $\hat{S}$ or $\hat{S}_{\text{group}}$, respectively and the estimated linear predictor $\hat{f}(x) = x\hat{\beta}$ depend on the parametrization (e.g. on the choice of basis functions).

In the setting where the group sizes are not too large, we can use the following penalty:

$$\lambda \sum_{j=1}^{q} m_j \|\mathbf{X}_{\mathscr{G}_j} \beta_{\mathscr{G}_j}\|_2 = \lambda \sum_{j=1}^{q} m_j \sqrt{\beta_{\mathscr{G}_j}^T \mathbf{X}_{\mathscr{G}_j}^T \mathbf{X}_{\mathscr{G}_j} \beta_{\mathscr{G}_j}}$$

which is a generalized group Lasso penalty if $\mathbf{X}_{\mathscr{G}_j}^T \mathbf{X}_{\mathscr{G}_j}$ is positive definite for every j (we throughout require that the group sizes T_j are smaller than sample size n). We call this the groupwise prediction penalty since we penalize the norms of the linear predictors $\mathbf{X}_{\mathscr{G}_j} \beta_{\mathscr{G}_j}$. Trivially, the penalty is invariant under reparametrization within every group $\mathscr{G}_j$, i.e., we can use $\tilde{\beta}_{\mathscr{G}_j} = B_j \beta_{\mathscr{G}_j}$ where B_j is any invertible $T_j \times T_j$ matrix. Therefore, using such a groupwise prediction penalty in a generalized linear model, we have the property that the estimated set of active groups and the predictions are invariant under any one-to-one reparametrization within the groups. That is, for $\tilde{\beta}_{\mathscr{G}_j} = B_j \beta_{\mathscr{G}_j}$ with any invertible B_j's:

$$\mathbf{X}_{\mathscr{G}_j} \hat{\beta}_{\mathscr{G}_j} = \tilde{\mathbf{X}}_{\mathscr{G}_j} \hat{\tilde{\beta}}_{\mathscr{G}_j}, \quad \tilde{\mathbf{X}}_{\mathscr{G}_j} = \mathbf{X}_{\mathscr{G}_j} B_j^{-1} \ (j = 1, \ldots, q),$$

and if $B_j 0 = 0$ for all j, then also

$$\hat{S}_{\text{group}} = \{j; \ \hat{\beta}_{\mathscr{G}_j} \neq 0\} = \{j; \ \hat{\tilde{\beta}}_{\mathscr{G}_j} \neq 0\}.$$

This invariance is a nice property suggesting to use the groupwise prediction penalty more often than what is common practice.

4.6 The adaptive group Lasso

The idea of the adaptive Lasso in Section 2.8 can also be applied to the generalized group Lasso. As a starting point, we assume to have an initial estimator $\hat{\beta}_{\text{init}}$. Ideally, it is tailored for the structure with groups $\mathscr{G}_1, \ldots, \mathscr{G}_q$ as in (4.1) so that we have sparsity in the sense that a whole sub-vector estimate $\hat{\beta}_{\text{init},\mathscr{G}_j}$ is zero or all components thereof are non-zero. A natural candidate for an initial estimator is the Group Lasso estimate in (4.7) or the generalized group Lasso estimate with the penalty in (4.11). From a practical perspective, we would tune the regularization parameter for the initial estimator according to prediction optimality using a cross-validation scheme. Thereby, we would measure prediction accuracy with the squared error loss for linear models or negative log-likelihood loss for generalized linear models.

The adaptive group Lasso is then defined with the following re-weighted penalty. Instead of (4.2), we take

$$\lambda \sum_{j=1}^{q} m_j \frac{\|\beta_{\mathscr{G}_j}\|_2}{\|\hat{\beta}_{\text{init},\mathscr{G}_j}\|_2}.$$

In terms of computation, we can simply re-scale the covariates in a linear or generalized linear model:

$$\tilde{X}^{(j)} = X^{(j)} \|\hat{\beta}_{\text{init},\mathscr{G}_r}\|_2 \text{ if } j \in \mathscr{G}_r.$$

Then, $\sum_{j=1}^{p} \beta_j X^{(j)} = \sum_{j=1}^{p} \tilde{\beta}_j \tilde{X}^{(j)}$ with

$$\tilde{\beta}_j = \frac{\beta_j}{\|\hat{\beta}_{\text{init},\mathscr{G}_r}\|_2} \text{ if } j \in \mathscr{G}_r,$$

to achieve that

$$\lambda \sum_{j=1}^{q} m_j \frac{\|\beta_{\mathscr{G}_j}\|_2}{\|\hat{\beta}_{\text{init},\mathscr{G}_j}\|_2} = \lambda \sum_{j=1}^{q} m_j \|\tilde{\beta}_{\mathscr{G}_j}\|_2.$$

Hence, we can use the same program to compute the adaptive group Lasso as for the plain non-adaptive case (omitting all the groups with $\mathscr{G}_r$ with $\hat{\beta}_{\text{init},\mathscr{G}_j} \equiv 0$).

Obviously, we can also use an adaptive generalized group Lasso. Instead of (4.11) we use

$$\lambda \sum_{j=1}^{q} m_j \frac{\sqrt{\beta_{\mathscr{G}_j}^T A_j \beta_{\mathscr{G}_j}}}{\sqrt{\hat{\beta}_{\text{init},\mathscr{G}_j}^T A_j \hat{\beta}_{\text{init},\mathscr{G}_j}}}.$$

Regarding computation, we can rewrite this with rescaled matrices of the form

$$\tilde{A}_j = c_j A_j, \quad c_j = \frac{1}{\hat{\beta}_{\text{init},\mathcal{G}_j}^T A_j \hat{\beta}_{\text{init},\mathcal{G}_j}}, \quad j = 1,\ldots,q.$$

Thereby, if $c_j = \infty$, we omit the variables from group $\mathcal{G}_j$.

The adaptive group Lasso is primarily recommended to be used for better selection of groups of variables. The heuristics and motivation are the same as for the adaptive Lasso described in Section 2.8. Moreover, when using the group Lasso as initial estimator, the adaptive group Lasso is always at least as sparse in terms of non-zero coefficients (and number of groups with non-zero coefficients). This is desirable if the underlying true structure is indeed very sparse with a few strong signals, and we then would get better prediction results as well.

4.7 Algorithms for the group Lasso

The group Lasso estimator $\hat{\beta}(\lambda)$ in (4.3) is given by minimizing the convex objective function

$$Q_\lambda(\beta) = n^{-1} \sum_{i=1}^n \rho_\beta(X_i, Y_i) + \lambda \sum_{j=1}^q m_j \|\beta_{\mathcal{G}_j}\|_2, \tag{4.12}$$

where $\rho_\beta(X_i, Y_i)$ is a loss function which is convex in β. For the squared error loss, we consider

$$\rho_\beta(x, y) = |y - x\beta|^2, \quad (y \in \mathbb{R}, \ x \in \mathbb{R}^p),$$

and for the logistic loss in a binary classification problem we have (see formula (3.4) in Chapter 3),

$$\rho_\beta(x, y) = -y f_\beta(x) + \log(1 + \exp(f_\beta(x))), \quad (y \in \{0, 1\}, \ x \in \mathbb{R}^p),$$
$$f_\beta(x) = x\beta.$$

In both these examples, the loss functions are of the form $\rho_\beta(x, y) = \rho(f_\beta(x), y)$ as a composition of a linear function in β and a convex function $f \mapsto \rho(f, y)$ in f for all y. This class of loss functions, covering the case of generalized linear models, is considered in Chapter 6 using the notation ρ_{f_β}.

We denote in the sequel the empirical risk by

$$\rho(\beta) = n^{-1} \sum_{i=1}^n \rho_\beta(X_i, Y_i).$$

The penalized version then decomposes as

$$Q_\lambda(\beta) = \rho(\beta) + \lambda \sum_{j=1}^{q} m_j \|\beta_{\mathscr{G}_j}\|_2.$$

As a consequence of the Karush-Kuhn-Tucker (KKT) conditions (Bertsekas, 1995) we have the following result which generalizes the first statements in Lemma 2.1 from Section 2.5.

Lemma 4.2. *Assume that $\rho(\beta)$ is differentiable and convex. Then, a necessary and sufficient condition for $\hat{\beta}$ to be a solution of (4.12) is*

$$\nabla \rho(\hat{\beta})_{\mathscr{G}_j} + \lambda m_j \frac{\hat{\beta}_{\mathscr{G}_j}}{\|\hat{\beta}_{\mathscr{G}_j}\|_2} = 0 \text{ if } \hat{\beta}_{\mathscr{G}_j} \neq 0 \text{ (i.e. not equal to the 0-vector)},$$

$$\|\nabla \rho(\hat{\beta})_{\mathscr{G}_j}\|_2 \leq \lambda m_j \text{ if } \hat{\beta}_{\mathscr{G}_j} \equiv 0,$$

where $\nabla \rho(\hat{\beta})$ denotes the gradient vector of $\rho(\beta)$ evaluated at $\hat{\beta}$.

Proof. If $\hat{\beta}_{\mathscr{G}_j} \neq 0$, the criterion function $Q_\lambda(\cdot)$ is partially differentiable with respect to $\beta_{\mathscr{G}_j}$ and it is necessary and sufficient that these partial derivatives are zero (there are no local minima due to convexity): this is the first equation in the characterization. If $\hat{\beta}_{\mathscr{G}_j} \equiv 0$, the criterion function $Q_\lambda(\cdot)$ is not differentiable but we can invoke subdifferential calculus (Bertsekas, 1995). The subdifferential of $Q_\lambda(\cdot)$ with respect to $\beta_{\mathscr{G}_j}$ is the set

$$\partial Q_\lambda(\beta)_{\mathscr{G}_j} = \{\nabla \rho(\beta)_{\mathscr{G}_j} + \lambda e; \ e \in E(\beta_{\mathscr{G}_j})\},$$

$$E(\beta_{\mathscr{G}_j}) = \{e \in \mathbb{R}^{T_j}; \ e = m_j \frac{\beta_{\mathscr{G}_j}}{\|\beta_{\mathscr{G}_j}\|_2} \text{ if } \beta_{\mathscr{G}_j} \neq 0 \text{ and } \|e\|_2 \leq m_j \text{ if } \beta_{\mathscr{G}_j} \equiv 0\},$$

$$(4.13)$$

see Problem 4.2. Note that the latter case with $\beta_{\mathscr{G}_j} \equiv 0$ is of special interest here: then, e is any vector within the ball having Euclidean radius m_j. Finally, a standard result from subdifferential calculus says that the parameter vector $\hat{\beta}_{\mathscr{G}_j}$ minimizes $Q_\lambda(\beta)_{\mathscr{G}_j}$ if and only if $0 \in \partial Q_\lambda(\hat{\beta})$ (Bertsekas, 1995), see also Problem 4.2, which is equivalent to the first and second statement in the characterization (the subdifferential $\partial Q_\lambda(\hat{\beta})$ consists of all block components $\partial Q_\lambda(\hat{\beta})_{\mathscr{G}_j}$ $(j = 1, \ldots, q)$). $\qquad\square$

4.7.1 Block coordinate descent

For the squared error loss, we can proceed in a simple way using a block coordinate descent algorithm, also known as Gauss-Seidel type method, as proposed by Yuan and Lin (2006). The idea of block coordinate descent is more general, however, and

we can use it also for other loss functions $\rho_\beta(\cdot,\cdot)$, as in formula (4.12), which are convex and differentiable with respect to β.

We cycle through the groups (blocks) $j = 1,\ldots,q,1,\ldots,q,1,\ldots$ and in every of these cycling steps, we optimize the objective function with respect to the corresponding group (block) $\mathcal{G}_j$ while keeping all but the current parameters corresponding to a group fixed. This leads us to the computation presented in Algorithm 2, where we denote by $\beta_{-\mathcal{G}_j}$ the vector β whose components in $\mathcal{G}_j$ are set to zero:

$$(\beta_{-\mathcal{G}_j})_k = \begin{cases} \beta_k, & k \notin \mathcal{G}_j, \\ 0, & k \in \mathcal{G}_j. \end{cases} \tag{4.14}$$

Similarly, $\mathbf{X}_{\mathcal{G}_j}$ denotes the $n \times T_j$ matrix consisting of the columns of the design matrix $\mathbf{X}$ corresponding to the predictors from the group $\mathcal{G}_j$. For notational simplicity, we drop in the following the hat-notation for β.

Algorithm 2 Block Coordinate Descent Algorithm

1: Let $\beta^{[0]} \in \mathbb{R}^p$ be an initial parameter vector. Set $m = 0$.
2: **repeat**
3:    Increase m by one: $m \leftarrow m+1$.
      Denote by $\mathscr{S}^{[m]}$ the index cycling through the block coordinates $\{1,\ldots,q\}$:
      $\mathscr{S}^{[m]} = \mathscr{S}^{[m-1]} + 1 \bmod q$. Abbreviate by $j = \mathscr{S}^{[m]}$ the value of $\mathscr{S}^{[m]}$.
4:    if $\|(-\nabla\rho(\beta_{-\mathcal{G}_j}^{[m-1]}))_{\mathcal{G}_j}\|_2 \le \lambda m_j$: set $\beta_{\mathcal{G}_j}^{[m]} = 0$,
      otherwise: $\beta_{\mathcal{G}_j}^{[m]} = \underset{\beta_{\mathcal{G}_j}}{\arg\min} Q_\lambda(\beta_{+\mathcal{G}_j}^{[m-1]})$,
      where $\beta_{-\mathcal{G}_j}^{[m-1]}$ is defined in (4.14) and $\beta_{+\mathcal{G}_j}^{[m-1]}$ is the parameter vector which equals $\beta^{[m-1]}$
      except for the components corresponding to group $\mathcal{G}_j$ whose entries are equal to $\beta_{\mathcal{G}_j}$ (i.e.
      the argument we minimize over).
5: **until** numerical convergence

In Step 4 of Algorithm 2, the ℓ_2-norm of the negative gradient and the corresponding inequality look as follows for the squared error and logistic loss, respectively (see Problem 4.3): denoting by $j = \mathscr{S}^{[m]}$,

$$\|2n^{-1}\mathbf{X}_{\mathcal{G}_j}^T(\mathbf{Y} - \beta_{-\mathcal{G}_j}^{[m-1]})\|_2 \le \lambda m_j \quad \text{for the squared error loss,} \tag{4.15}$$

$$\|n^{-1}\mathbf{X}_{\mathcal{G}_j}^T(\mathbf{Y} - \pi_{\beta_{-\mathcal{G}_j}^{[m-1]}})\|_2 \le \lambda m_j \quad \text{for the logistic loss,} \tag{4.16}$$

where for the latter, $(\pi_\beta)_i = \mathbf{P}_\beta[Y_i = 1|X_i]$.

Step 4 is an explicit check whether the minimum is at the non-differentiable point with $\beta_{\mathcal{G}_j} \equiv 0$. If not, we can use a standard numerical minimizer, e.g., a gradient-type algorithm, to find the optimal solution with respect to $\beta_{\mathcal{G}_j}$.

4.7.1.1 Squared error loss

In case of squared error loss, the block-update in Step 4 in Algorithm 2 is explicit if $n^{-1}\mathbf{X}_{\mathscr{G}_j}^T\mathbf{X}_{\mathscr{G}_j} = I_{T_j}$. Note that this assumption is quite harmless since the penalty term is invariant under orthonormal transformations, that is, it does not matter how we proceed to orthonormalize the design sub-matrices corresponding to the different groups; see also the short discussion at the end of Section 4.2.1 and Section 4.5.1. It then holds that the minimizer in Step 4 is as follows: denoting by $j = \mathscr{S}^{[m]}$,

$$\text{if } \|(-\nabla\rho(\beta_{-\mathscr{G}_j}^{[m-1]})_{\mathscr{G}_j}\|_2 = \|2\mathbf{X}_{\mathscr{G}_j}^T(\mathbf{Y} - \mathbf{X}\beta_{-\mathscr{G}_j}^{[m-1]})\|_2 > \lambda m_j :$$

$$\beta_{\mathscr{G}_j}^{[m]} = \arg\min_{\beta_{\mathscr{G}_j}} Q_\lambda(\beta_{+\mathscr{G}_j}^{[m-1]}) = (1 - \frac{\lambda m_j}{\|U_{\mathscr{G}_j}\|_2})U_{\mathscr{G}_j},$$

$$U_{\mathscr{G}_j} = 2n^{-1}\mathbf{X}_{\mathscr{G}_j}^T(\mathbf{Y} - \mathbf{X}\beta_{-\mathscr{G}_j}^{[m-1]}). \tag{4.17}$$

In short, the entire up-date is

$$\beta_{\mathscr{G}_j}^{[m]} = (1 - \frac{\lambda m_j}{\|U_{\mathscr{G}_j}\|_2})_+U_{\mathscr{G}_j},$$

where $(x)_+ = \max(x,0)$. Thus, the block coordinate descent algorithm amounts to some form of iterative thresholding. See Problem 4.4.

4.7.1.2 Active set strategy

For sparse problems with a large number of groups q but only few of them being active, an active set strategy can speed up the algorithm considerably. An active set is here defined as the set of groups whose coefficient vector is non-zero. .

When cycling through the coordinate blocks (or groups), we restrict ourselves to the current active set and visit only "rarely" the remaining blocks (or groups), e.g., every 10th iteration, to up-date the active set. This is especially useful for very high-dimensional settings and it easily allows for $p \approx 10^4 - 10^6$. For the high-dimensional example in Section 4.3.1, this modification decreases the computation time by about 40% (the example uses the logistic loss whose block up-dates are not explicit, as discussed below).

4.7.1.3 General convex loss

In case of other than squared error loss, we need to do numerical optimization for a block up-date in Step 4 in Algorithm 2. Then, the value of the last iteration can

be used as starting value to save computing time. If the group was not active in the last iteration (e.g. $\beta_{\mathscr{G}_j}^{[m-1]} = 0$ for group j) we first go a small step in the opposite direction of the gradient of the negative log-likelihood function to ensure that we start at a differentiable point. We will discuss later that full groupwise optimization in Step 4 is not necessary and an approximate minimization will be sufficient, as discussed in detail in Section 4.7.2.

We show now that the block gradient descent algorithm converges to a global optimum, a result which generalizes Proposition 2.1 from Section 2.12.1. The arguments are completely analogous to the derivation of this proposition. We discuss first that the block up-dates are well-defined. Consider the function

$$h = h_{\beta_{\mathscr{G}_j^c}} : \ \beta_{\mathscr{G}_j} \mapsto Q_\lambda(\{\beta_{\mathscr{G}_j}, \beta_{\mathscr{G}_j^c}\}) \ (j = 1, \ldots, q),$$

(where we use a slight abuse of notation with the ordering of coordinates in the argument of the function $Q_\lambda(\cdot)$). That is, $h(\beta_{\mathscr{G}_j})$ describes the function $Q_\lambda(\beta)$ as a function of $\beta_{\mathscr{G}_j}$ for fixed $\beta_{\mathscr{G}_j^c}$, where we denote by $\mathscr{G}_j^c = \{1, \ldots, p\} \setminus \mathscr{G}_j$. Note that $h(\beta_{\mathscr{G}_j})$ is just a different notation for $Q_\lambda(\beta_{+\mathscr{G}_j})$ in Step 4 from Algorithm 2. We observe that if $Q_\lambda(\beta)$ is convex in β, then also $h(\beta_{\mathscr{G}_j}) = h_{\beta_{\mathscr{G}_j^c}}(\beta_{\mathscr{G}_j})$ is convex in $\beta_{\mathscr{G}_j}$, for all $\beta_{\mathscr{G}_j^c}$. See Problem 4.5. Using the same argument as in the short proof of Lemma 4.1, the groupwise minima are attained. In a next step, one needs to show that Step 4 of Algorithm 2 minimizes the convex function $h(\beta_{\mathscr{G}_j}) = h_{\beta_{\mathscr{G}_j^c}}(\beta_{\mathscr{G}_j})$ with respect to $\beta_{\mathscr{G}_j}$ ($j = \mathscr{S}^{[m]}$). Analogously as in Section 2.12.1, since $h(\beta_{\mathscr{G}_j})$ is not differentiable everywhere, we invoke subdifferential calculus(Bertsekas, 1995). Here, the subdifferential of $h(\cdot)$ is the set $\partial h(\beta_{\mathscr{G}_j}) = \{\nabla\rho(\beta)_{\mathscr{G}_j} + \lambda e; \ e \in E(\beta_{\mathscr{G}_j})\}$, $E(\beta_{\mathscr{G}_j}) = \{e \in \mathbb{R}^{T_j}; \ e = m_j \frac{\beta_{\mathscr{G}_j}}{\|\beta_{\mathscr{G}_j}\|_2} \text{ if } \beta_{\mathscr{G}_j} \neq 0 \text{ and } \|e\|_2 \leq m_j \text{ if } \beta_{\mathscr{G}_j} \equiv 0\}$. The parameter vector $\beta_{\mathscr{G}_j}$ minimizes $h(\beta_{\mathscr{G}_j})$ if and only if $0 \in \partial h(\beta_{\mathscr{G}_j})$, and this leads to the formulation in Step 4. Finally, as a Gauss-Seidel algorithm which cycles through the coordinates $\mathscr{S}^{[m]} = 1, \ldots, q, 1, \ldots$ ($m = 1, 2, \ldots$), one can establish numerical convergence to a stationary point. The mathematical details are not entirely trivial. One can exploit the fact that the penalty term is block-separable[1] and then make use of a general theory from Tseng (2001) for numerical convergence of Gauss-Seidel type algorithms (conditions (A1), (B1) - (B3) and (C2) from Tseng (2001) hold; this then implies that every cluster point of the sequence $\beta^{[m]}\}_{m \geq 0}$ is a stationary point of the convex function $Q_\lambda(\cdot)$ and hence a minimum point).

We summarize the derivation by the following result.

Proposition 4.1. *For the quantities in formula (4.12), assume that the loss function satisfies $\rho_\beta(X_i, Y_i) \geq C > -\infty$ for all β, X_i, Y_i ($i = 1, \ldots n$), it is continuously differentiable with respect to β, and that $Q_\lambda(\cdot)$ is convex. Then, Step 4 of the block coor-*

---

[1] A function $f(\beta)$ is called block separable (into convex functions) with blocks $\mathscr{G}_1, \ldots, \mathscr{G}_q$ if $f(\beta) = \sum_{j=1}^{q} f_j(\beta_{\mathscr{G}_j})$ with convex functions $f_j(\cdot)$.

dinate descent Algorithm 2 performs groupwise minimizations of $Q_\lambda(\cdot)$ (i.e. of the functions $h_{\beta_{\mathscr{G}_j^c}}(\cdot)$) and is well defined in the sense that the corresponding minima are attained. Furthermore, if we denote by $\hat{\beta}^{[m]}$ the parameter vector from Algorithm 2 after m iterations, then every cluster point of the sequence $\{\hat{\beta}^{[m]}\}_{m\geq0}$ is a minimum point of $Q_\lambda(\cdot)$.

As pointed out earlier, the boundedness assumption for the loss function holds for the commonly used loss functions in (generalized) regression and classification. Furthermore, the iterates $\hat{\beta}^{[m]}$ can be shown to stay in a compact set (because of the penalty term) and thus, the existence of a cluster point is guaranteed.

The main drawback of such a block gradient descent Algorithm 2 is for cases other than squared error loss where the blockwise minimizations of the active groups in Step 4 have to be performed numerically. However, for small and moderate sized problems in the dimension p and group sizes T_j, this turns out to be sufficiently fast. Improvements in computational efficiency are possible by replacing an exact groupwise minimization in Step 4 with a suitable approximation whose computation is explicit. This will be discussed next.

4.7.2 Block coordinate gradient descent

As described in (4.17), the blockwise up-dates are available in closed form for squared error loss. For other loss functions, the idea is to use a quadratic approximation which then allows for some rather explicit blockwise up-dates. More technically, the key idea is to combine a quadratic approximation of the empirical loss with an additional line search. In fact, this then equals the block coordinate gradient descent method from Tseng and Yun (2009). The description here closely follows Meier et al. (2008).

Using a second order Taylor expansion at $\beta^{[m]}$, the estimate in the mth iteration, and replacing the Hessian of the empirical risk $\rho(\beta)$ by a suitable matrix $H^{[m]}$ we define

$$M_\lambda^{[m]}(d) = \rho(\beta^{[m]}) + d^T\nabla\rho(\beta^{[m]}) + \frac{1}{2}d^T H^{[m]}d + \lambda\sum_{j=1}^{q}m_j\|\beta_{\mathscr{G}_j}^{[m]} + d_{\mathscr{G}_j}\|_2$$

$$\approx Q_\lambda(\beta^{[m]} + d), \tag{4.18}$$

where $d \in \mathbb{R}^p$.

Now we consider minimization of $M_\lambda^{[m]}(\cdot)$ with respect to the jth parameter group. This means that we restrict ourselves to vectors d with $d_k = 0$ for $k \notin \mathscr{G}_j$. Moreover, we assume that the corresponding $T_j \times T_j$ submatrix $H_{\mathscr{G}_j,\mathscr{G}_j}^{[m]}$ is of the form $H_{\mathscr{G}_j,\mathscr{G}_j}^{[m]} = h_j^{[m]} \cdot I_{T_j}$ for some scalar $h_j^{[m]} \in \mathbb{R}$; more discussion about the choice of the matrix $H^{[m]}$

is given below in formula (4.22). If $\|\nabla\rho(\beta^{[m]})_{\mathcal{G}_j} - h_j^{[m]}\beta_{\mathcal{G}_j}^{[m]}\|_2 \leq \lambda m_j$, the minimizer of $M_\lambda^{[m]}(d)$ in (4.18) with respect to the components in group $\mathcal{G}_j$ is

$$d_{\mathcal{G}_j}^{[m]} = -\beta_{\mathcal{G}_j}^{[m]}, \tag{4.19}$$

see Problem 4.6. Note that this is similar to Lemma 4.2, due to the KKT conditions, where we also examine the absolute value of the gradient, and here, the non-differentiable point is at $d_{\mathcal{G}_j}^{[m]} = -\beta_{\mathcal{G}_j}$. Otherwise, the minimizer of (4.18) with respect to the components in $\mathcal{G}_j$ is (Problem 4.6)

$$d_{\mathcal{G}_j}^{[m]} = -\frac{1}{h_{\mathcal{G}_j}^{[m]}}\left\{\nabla\rho(\beta^{[m]})_{\mathcal{G}_j} - \lambda m_j \frac{\nabla\rho(\beta^{[m]})_{\mathcal{G}_j} - h_{\mathcal{G}_j}^{[m]}\beta_{\mathcal{G}_j}^{[m]}}{\|\nabla\rho(\beta^{[m]})_{\mathcal{G}_j} - h_{\mathcal{G}_j}^{[m]}\beta_{\mathcal{G}_j}^{[m]}\|_2}\right\}. \tag{4.20}$$

If $d_{\mathcal{G}_j}^{[m]} \neq 0$, an inexact line search using the Armijo rule has to be performed for up-dating the parameter vector.

Armijo Rule: Let $\alpha_j^{[m]}$ be the largest value among the grid-points $\{\alpha_0 \delta^l\}_{l \geq 0}$ such that

$$Q_\lambda(\beta^{[m]} + \alpha_j^{[m]} d_{\mathcal{G}_j}^{[m]}) - Q_\lambda(\beta^{[m]}) \leq \alpha_j^{[m]}\sigma\Delta^{[m]}, \tag{4.21}$$

where $0 < \delta < 1$, $0 < \sigma < 1$, $\alpha_0 > 0$, and $\Delta^{[m]}$ is the improvement in the objective function $Q_\lambda(\cdot)$ when using a linear approximation for the objective function, i.e.,

$$\Delta^{[m]} = (d_{\mathcal{G}_j}^{[m]})^T \nabla\rho(\beta^{[m]})_{\mathcal{G}_j} + \lambda m_j \|\beta_{\mathcal{G}_j}^{[m]} + d_{\mathcal{G}_j}^{[m]}\|_2 - \lambda m_j\|\beta_{\mathcal{G}_j}^{[m]}\|_2.$$

When writing $\beta^{[m]} + \alpha_j^{[m]} d_{\mathcal{G}_j}^{[m]}$, we implicitly mean that only the block of parameters corresponding to the group $\mathcal{G}_j$ are affected by the summation. We remark that $\Delta^{[m]} < 0$ for $d_{\mathcal{G}_j}^{[m]} \neq 0$, as shown in Tseng and Yun (2009).

Finally, we define

$$\beta^{[m+1]} = \beta^{[m]} + \alpha_j^{[m]} d_{\mathcal{G}_j}^{[m]},$$

where only the block or parameters corresponding to the group $\mathcal{G}_j$ is up-dated (j denotes the index of the block component in iteration $m+1$). A summary is outlined in Algorithm 3. Standard choices for the tuning parameters are for example $\alpha_0 = 1$, $\delta = 0.5$, $\sigma = 0.1$ (Bertsekas, 1995; Tseng and Yun, 2009). Other definitions of $\Delta^{[m]}$ as for example to include the quadratic part of the improvement are also possible. We refer the reader to Tseng and Yun (2009) for more details and the fact that the line search can always be performed. It is worth pointing out that the block up-dates are fairly explicit. However, in comparison to the block coordinate descent Algorithm 2

for the squared error loss, we need to implement an additional line search using the Armijo rule as in (4.21).

For a general matrix $H^{[m]}$ the minimization with respect to the jth parameter group depends on $H^{[m]}$ only through the corresponding submatrix $H^{[m]}_{\mathcal{G}_j,\mathcal{G}_j}$. To ensure a reasonable quadratic approximation in (4.18), $H^{[m]}_{\mathcal{G}_j,\mathcal{G}_j}$ is ideally chosen to be close to the corresponding submatrix of the Hessian of the empirical risk function. However, there is a trade-off between accuracy of the quadratic approximation and computational efficiency. The latter is the reason to restrict ourselves to matrices of the form $H^{[m]}_{\mathcal{G}_j,\mathcal{G}_j} = h^{[m]}_j \cdot I_{T_j}$, and a possible choice is (Tseng and Yun, 2009)

$$
h^{[m]}_j = \min\left(\max\left\{\operatorname{diag}\left\{\nabla^2\rho(\beta^{[m]})_{\mathcal{G}_j,\mathcal{G}_j}\right\}, c_{\min}\right\}, c_{\max}\right), \qquad (4.22)
$$

where $0 < c_{\min} < c_{\max} < \infty$ are bounds (e.g. $c_{\min} = 10^{-6}$ and $c_{\max} = 10^8$) to ensure convergence (see Proposition 4.2). The matrix $H^{[m]}$ does not necessarily have to be recomputed in each iteration. Under some mild conditions on $H^{[m]}$ (which are satisfied for the choice in (4.22)), convergence of the algorithm is assured and we refer for the details to Tseng and Yun (2009).

When minimizing $M^{[m]}_\lambda(\cdot)$ with respect to a group showing up in the penalty term, we first have to check whether the minimum is at a non-differentiable point as outlined above. For an unpenalized intercept β_0, this is not necessary and the solution can be directly computed:

$$
d^{[m]}_0 = -\frac{1}{h^{[m]}_0}\nabla\rho(\beta^{[m]})_0.
$$

Algorithm 3 Block Coordinate Gradient Descent Algorithm

1: Let $\beta^{[0]} \in \mathbb{R}^p$ be an initial parameter vector. Set $m = 0$.
2: **repeat**
3:     Increase m by one: $m \leftarrow m+1$.
       Denote by $\mathscr{S}^{[m]}$ the index cycling through the block coordinates $\{1,\ldots,q\}$:
       $\mathscr{S}^{[m]} = \mathscr{S}^{[m-1]} + 1 \bmod q$. Abbreviate by $j = \mathscr{S}^{[m]}$ the value of $\mathscr{S}^{[m]}$.
4:     $H^{[m-1]}_{\mathcal{G}_j,\mathcal{G}_j} = h^{[m-1]}_j \cdot I_{T_j} = h_j(\beta^{[m-1]}) \cdot I_{T_j}$ as in (4.22),
       $d^{[m-1]}_{\mathcal{G}_j} = (d^{[m-1]})_{\mathcal{G}_j}, \ d^{[m-1]} = \underset{d;\, d_{\mathcal{G}_k}=0 \ (k\neq j)}{\arg\min} \ M^{[m-1]}_\lambda(d)$ with $M^{[m-1]}_\lambda(\cdot)$ as in (4.18),
       if $d^{[m-1]}_{\mathcal{G}_j} \neq 0$:
           $\alpha^{[m-1]}_j \leftarrow$ line search,
           $\beta^{[m]} \leftarrow \beta^{[m-1]} + \alpha^{[m-1]}_j \cdot d^{[m-1]}_{\mathcal{G}_j}$,
5: **until** numerical convergence

Proposition 4.2. *Assume that the loss function $\rho_\beta(X_i, Y_i) \geq C > -\infty$ for all β, X_i, Y_i ($i = 1, \ldots, n$), it is continuously differentiable with respect to β, and that the empirical risk $\rho(\beta)$ is convex. Denote by $\hat{\beta}^{[m]}$ the parameter vector from the block coordinate gradient descent Algorithm 3 after m iterations. If $H^{[m]}_{\mathscr{G}_j, \mathscr{G}_j}$ is chosen according to (4.22), then every cluster point of the sequence $\{\hat{\beta}^{[m]}\}_{m \geq 0}$ is a minimum point of $Q_\lambda(\cdot)$.*

This result is a consequence of a more general theory on the coordinate gradient descent method, see Tseng and Yun (2009, Theorem 1(e), Section 4). Linking this theory to the case of the group Lasso is rigorously described in Meier et al. (2008, Proposition 2). We remark that the block coordinate gradient descent Algorithm 3 can be applied to the group Lasso in any generalized linear model where the response Y has a distribution from the exponential family.

To calculate the solutions $\hat{\beta}(\lambda)$ for various penalty parameters from a grid $\Lambda = \{0 \leq \lambda_{\text{grid},1} < \lambda_{\text{grid},2} < \lambda_{\text{grid},g}\}$ we can for example start at

$$\lambda_{\text{grid},g} = \lambda_{max} = \max_{j \in \{1, \ldots, q\}} \frac{1}{m_j} \|\nabla\rho(\beta)_{\mathscr{G}_j}|_{\beta \equiv 0}\|_2,$$

where all parameters in all the groups are equal to zero. We then use $\hat{\beta}(\lambda_{\text{grid},k})$ as a starting value for $\hat{\beta}(\lambda_{\text{grid},k-1})$ and proceed iteratively until $\hat{\beta}(\lambda_{\text{grid},1})$ with $\lambda_{\text{grid},1}$ close or equal to zero. Instead of up-dating the approximation of the Hessian $H^{[m]}$ in each iteration, we can use a constant matrix based on the previous parameter estimates $\hat{\beta}_{\lambda_{\text{grid},k}}$ to save computing time, i.e.,

$$H^{[m]}_{\mathscr{G}_j, \mathscr{G}_j} = h_j(\hat{\beta}(\lambda_{\text{grid},k}))I_{T_j},$$

for the estimation of $\hat{\beta}_{\lambda_{\text{grid},k-1}}$ (and this matrix does not depend on iterations m). A cross-validation scheme can then be used for choosing the parameter λ among the candidate values from the grid Λ.

Problems

4.1. Factor variables
Consider a linear model with 3 factor variables, each of them having 4 categorical levels. Construct the design matrix, denoted in Section 4.3 as $\tilde{\mathbf{X}}$, for the case when considering main effects and first-order interactions only and requiring that the sum of the effects (for each main and each first-order effect) equals zero (sum to zero constraints). You may want to use the R statistical software using the command `model.matrix`.

4.2. Subdifferential and Subgradient
Consider a convex function

$$f : \mathbb{R}^p \to \mathbb{R}.$$

A vector $d \in \mathbb{R}^p$ is called a *subgradient* of f at point $x \in \mathbb{R}^p$ if

$$f(y) \geq f(x) + (y - x)^T d.$$

The set of all subgradients of the convex function f at $x \in \mathbb{R}^p$ is called the *subdifferential* of f at x, denoted by $\partial f(x)$. A necessary and sufficient condition for $x \in \mathbb{R}^p$ to be a minimum of f is: $0 \in \partial f(x)$, see p. 736 in Bertsekas (1995).

Verify that the expression $E(\beta_{\mathcal{G}_j})$ in (4.13) is the subdifferential of $m_j \|\beta_{\mathcal{G}_j}\|_2$ at $\beta_{\mathcal{G}_j}$. Therefore, $\partial Q_\lambda(\beta)$ in (4.13) is the subdifferential of Q_λ at β.

4.3. For the block coordinate up-dates, derive the expressions in (4.15) and (4.16).

4.4. Derive the up-date formula in (4.17).

4.5. Convexity of block coordinate functions (Step 4 of Algorithm 2)
Assume that

$$g(\cdot) : \mathbb{R}^2 \to \mathbb{R}$$
$$\beta = (\beta_1, \beta_2) \mapsto g(\beta)$$

is a convex function in $\beta = (\beta_1, \beta_2) \in \mathbb{R}^2$. Show that

$$h_{\beta_2}(\cdot) : \mathbb{R} \to \mathbb{R}$$
$$\beta_1 \mapsto h_{\beta_2}(\beta_1) = g(\beta_1, \beta_2)$$

is a convex function in β_1, for all fixed values of β_2.

4.6. For the block coordinate gradient descent Algorithm 3, derive formulae (4.19) and (4.20) by using the KKT conditions as in Lemma 4.2. First, work out the case where the group-size $|\mathcal{G}_j| = 1$.

4.7. Group Lasso for logistic regression
Consider a binary response variable Y and logistic regression as in Section 3.3.1 from Chapter 3. We focus on the group Lasso with loss function given by the negative log-likelihood as in (3.3). Write the block coordinate gradient descent algorithm (Algorithm 3 in this chapter) with explicit formulae for (4.20) and (4.22).

Chapter 5

Additive models and many smooth univariate functions

Abstract Additive models build a nonparametric extension of linear models and as such, they exhibit a substantial degree of flexibility. While the most important effects may still be detected by a linear model, substantial improvements are potentially possible by using the more flexible additive model class. At first sight, it seems very ambitious to fit additive models with high-dimensional covariates but sparsity implies feasible computations and good statistical properties. Besides encouraging sparsity, it is important to control smoothness as well. This can be achieved by a sparsity-smoothness penalty function. The combination of sparsity and smoothness is crucial for mathematical theory as well as for better performance on data. We discuss in this chapter methodology and computational aspects which are related to the group Lasso presented in Chapter 4.

5.1 Organization of the chapter

We consider in Section 5.2 the framework using basis expansions for every additive function. Section 5.3 presents the idea of combining sparsity and smoothness in a penalty term and it also describes an algorithm for computation. The penalty described there is the one which has clearer mathematical properties. An alternative penalty of group Lasso type is presented in Section 5.4: it is a bit easier to use since the corresponding optimization algorithm is very efficient. However, the drawback of such a group Lasso type penalty is its potential sub-optimality in terms of statistical accuracy. Other alternative penalties are discussed in this section as well. We then present some numerical examples in Section 5.5. In Section 5.6 we loosely describe statistical properties which are useful to know for justifying the methodology and some practical steps. A mathematically rigorous treatment is given in Section 8.4 in Chapter 8. Generalized additive models are treated in Section 5.7. We also

consider two related models: varying coefficient models are discussed in Section 5.8 and Section 5.9 considers multivariate and multitask models.

5.2 Introduction and preliminaries

Additive and generalized additive models are perhaps the first extension from linear to nonlinear (generalized) regression functions. The model class became very popular in the low-dimensional $p \ll n$ setting and the concepts and methodology have been well established, see for example Hastie and Tibshirani (1990).

We will show here that fitting additive models for high-dimensional covariates is quite easily possible and hence, flexible additive models should be a standard tool for high-dimensional generalized regression. Besides encouraging sparsity, it is important to control smoothness of the function estimates as well. This can be achieved by a sparsity-smoothness penalty function. The combination of sparsity and smoothness is crucial for mathematical theory as well as for better performance on data. One version of a sparsity-smoothness penalty function (see Section 5.4) amounts to an optimization with a generalized group Lasso penalty introduced in Section 4.5.

5.2.1 Penalized maximum likelihood for additive models

We consider high-dimensional additive regression models with a continuous response $Y \in \mathbb{R}$ and p covariates $X^{(1)}, \ldots, X^{(p)} \in \mathbb{R}$ connected through the model

$$Y_i = \mu + \sum_{j=1}^{p} f_j(X_i^{(j)}) + \varepsilon_i, \; i = 1, \ldots, n, \tag{5.1}$$

where μ is the intercept term, ε_i are i.i.d. random variables, independent of $\{X_i; \; i = 1, \ldots, n\}$, with $\mathbb{E}[\varepsilon_i] = 0$, and f_j are smooth univariate functions. For identification purposes we assume that all f_j are centered, i.e.

$$\sum_{i=1}^{n} f_j(X_i^{(j)}) = 0$$

for $j = 1, \ldots, p$. The design points X_i are allowed to be either fixed or random. If the model is correct, we denote by $f_j^0(\cdot)$ the true underlying additive functions. With some slight abuse of notation we also denote by f_j the n-dimensional vector $(f_j(X_1^{(j)}), \ldots, f_j(X_n^{(j)}))^T$. For a vector $f \in \mathbb{R}^n$ we denote by $\|f\|_n^2 = f^T f / n$.

The basic idea is to expand each function $f_j(\cdot)$

$$f_j(\cdot) = \sum_{k=1}^{K} \beta_{j,k} h_{j,k}(\cdot) \qquad (5.2)$$

using basis functions $h_{j,k}(\cdot)$ $(k = 1, \ldots, K)$ and estimate the unknown parameter vector β by penalized least squares:

$$\hat{\beta} = \arg\min_{\beta} \|\mathbf{Y} - \sum_{j=1}^{p} H_j \beta_j\|_2^2/n + \text{pen}(\beta),$$

where for $j = 1, \ldots, p$, H_j is an $n \times K$ matrix defined by

$$(H_j)_{i,k} = h_{j,k}(X_i^{(j)}), \ i = 1, \ldots, n; \ k = 1, \ldots, K,$$

$\beta_j = (\beta_{j,1}, \ldots, \beta_{j,K})^T$, $\beta = (\beta_1, \ldots, \beta_p)^T$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$. The exact form of suitable penalty functions will be discussed below. If $\mathbf{Y}$ is centered, we can omit an unpenalized intercept term and the nature of the objective function automatically forces the function estimates $\hat{f}_1, \ldots, \hat{f}_p$ to be centered. In other words, the estimate for μ in model (5.1) equals $\hat{\mu} = n^{-1} \sum_{i=1}^{n} Y_i$.

Typically, if p is large, we aim for a solution which is sparse in terms of whole $K \times 1$ parameter vectors $\beta_j = \{\beta_{j,k}; \ k = 1, \ldots, K\}$. This could be achieved with a group Lasso penalty of the form $\lambda \sum_{j=1}^{p} \|\beta_j\|_2$. We rather prefer here (a scaled version of) the prediction group Lasso penalty from Section 4.5.1

$$\lambda \sum_{j=1}^{p} \|H_j \beta_j\|_2/\sqrt{n} = \lambda \sum_{j=1}^{p} \|f_j\|_n. \qquad (5.3)$$

However, the penalty in (5.3) does not include any aspect of smoothness of the functions $f_j(\cdot)$. This could be addressed implicitly by a careful choice of the number K of basis functions. It is often better though to include smoothness into the penalty function. This then allows to use a large number of basis functions for every function $f_j(\cdot)$, which is necessary to capture some functions at high complexity, and an additional penalty term can then be used for appropriate regularization with respect to smoothness.

5.3 The sparsity-smoothness penalty

In order to construct an estimator which encourages sparsity at the function level, the norms $\|f_j\|_n$ $(j = 1, \ldots, p)$ should be penalized, see formula (5.3). As mentioned above, this alone is often not sufficient and we want additional control of the smoothness of the estimated functions.

We translate the smoothness of the functions f_j in terms of some quadratic norm. An important example is the Sobolev space of continuously differentiable functions

with the squared Sobolev semi-norm $I^2(f_j) = \int |f_j''(x)|^2 dx$ assumed to be finite. Then, using the basis expansion in (5.2) we obtain

$$I^2(f_j) = \int |f_j''(x)|^2 dx = \beta_j^T W_j \beta_j,$$

with W_j a given $K \times K$ matrix of weights

$$(W_j)_{k,\ell} = \int h_{j,k}''(x) h_{j,\ell}''(x) dx, \ k, \ell = 1, \ldots, K.$$

In Chapter 8, we write $W_j = B_j^T B_j$. More generally, we use

$$I^2(f_j) = \beta_j^T W_j \beta_j = \|B_j \beta_j\|_2^2$$

as a measure of squared smoothness of the function f_j using essentially any kind of $K \times K$ smoothing matrices $W_j = B_j^T B_j$, see e.g. Section 5.3.1 below.

In order to get sparse and sufficiently smooth function estimates, we consider the sparsity-smoothness penalty

$$\begin{aligned}
\text{pen}_{\lambda_1, \lambda_2}(\beta) &= \lambda_1 \sum_{j=1}^{p} \|f_j\|_n + \lambda_2 \sum_{j=1}^{p} I(f_j) \\
&= \lambda_1 \sum_{j=1}^{p} \|H_j \beta_j\|_2 / \sqrt{n} + \lambda_2 \sum_{j=1}^{p} \sqrt{\beta_j^T W_j \beta_j}.
\end{aligned} \tag{5.4}$$

The two tuning parameters $\lambda_1, \lambda_2 \geq 0$ control the amount of penalization. It is important to have two tuning parameters since the penalization for sparsity and of smoothness live on very different scales. The theory in Section 8.4 assumes that $\lambda_2 \asymp \lambda_1^2$.

Having chosen basis functions as in (5.2), the additive model estimator is defined by the following penalized least squares problem:

$$\hat{\beta}(\lambda_1, \lambda_2) = \arg\min_{\beta} \left\| Y - \sum_{j=1}^{p} H_j \beta_j \right\|_2^2 / n + \lambda_1 \sum_{j=1}^{p} \|H_j \beta_j\|_2 / \sqrt{n} + \lambda_2 \sum_{j=1}^{p} \sqrt{\beta_j^T W_j \beta_j},$$

$$\tag{5.5}$$

where $\beta = (\beta_1, \ldots, \beta_p)^T$ and $\beta_j = (\beta_{j,1}, \ldots, \beta_{j,K})^T$.

5.3.1 Orthogonal basis and diagonal smoothing matrices

We consider now the case where the basis functions $h_{j,k}(\cdot)$ are orthogonal with

$$n^{-1}H_j^T H_j = I \ (j = 1, \dots, p)$$

and with diagonal smoothing matrices (diagonalized smoothness)

$$W_j = \mathrm{diag}(d_1^2, \dots, d_K^2) \ (j = 1, \dots, p),$$
$$d_k = k^m,$$

where $m > 1/2$ is a chosen degree of smoothing. Most often, we assume $m = 2$. Since there is no dependence on j, we write

$$D^2 = \mathrm{diag}(d_1^2, \dots, d_K^2), \ d_k = k^m.$$

Then, the sparsity-smoothness penalty in (5.4) becomes

$$\mathrm{pen}_{\lambda_1, \lambda_2}(\beta) = \lambda_1 \sum_{j=1}^p \|\beta_j\|_2 + \lambda_2 \sum_{j=1}^p \|D\beta_j\|_2,$$

where $\|D\beta_j\|_2 = \sqrt{\sum_{k=1}^K k^{2m}\beta_{j,k}^2}$, and the additive model estimator is defined by

$$\hat{\beta}(\lambda_1, \lambda_2) = \arg\min_{\beta} \left\|\mathbf{Y} - \sum_{j=1}^p H_j\beta_j\right\|_2^2/n + \lambda_1 \sum_{j=1}^p \|\beta_j\|_2 + \lambda_2 \sum_{j=1}^p \|D\beta_j\|_2. \quad (5.6)$$

A concrete example for an orthonormal basis is given by orthogonal polynomials. (Roughly speaking, the functions $h_{j,k}(\cdot)$ and values d_k can be thought as eigenfunctions and eigenvalues of the space of functions with m derivatives). Clearly, the sparsity penalty $\lambda_1 \sum_{j=1}^p \|\beta_j\|_2$ remains invariant under orthonormal transformation of the basis. However, the smoothness penalty $\lambda_2 \sum_{j=1}^p \|D\beta_j\|_2$ depends on the type of orthonormal basis which is used.

5.3.2 Natural cubic splines and Sobolev spaces

We consider here additive functions f_j belonging to the Sobolev space of continuously differentiable functions on $[a,b]$ with squared smoothness semi-norms $I^2(f_j) = \int_a^b |f_j''(x)|^2 dx$. Using the penalty in (5.4) in terms of functions f_j, the additive model estimator can be written as a penalized least squares problem over the Sobolev class of functions:

$$\hat{f}_1, \dots, \hat{f}_p = \arg\min_{f_1, \dots, f_p \in \mathscr{F}} \left\|\mathbf{Y} - \sum_{j=1}^p f_j\right\|_n^2 + \lambda_1 \sum_{j=1}^p \|f_j\|_n + \lambda_2 \sum_{j=1}^p I(f_j), \quad (5.7)$$

where $\mathscr{F}$ is the Sobolev class of functions on $[a,b]$. Note that as before, we assume the same level of regularity for each function f_j.

Proposition 5.1. *Let* $a, b \in \mathbb{R}$ *such that* $a < \min_{i,j}\{X_i^{(j)}\}$ *and* $b > \max_{i,j}\{X_i^{(j)}\}$. *Let* $\mathscr{F}$ *be the Sobolev space of functions that are continuously differentiable on* $[a, b]$ *with square integrable second derivatives, and assume that there exist minimizers* $\hat{f}_j \in \mathscr{F}$ *of (5.7). Then the* $\hat{f}_j$'s *are natural cubic splines with knots at* $X_i^{(j)}, i = 1, \dots, n$.

Proof. Because of the additive structure of f and the penalty, it suffices to analyze each component f_j, $j = 1, \dots, p$ independently. Let $\hat{f}_1, \dots, \hat{f}_p$ be a solution of (5.7) and assume that some or all $\hat{f}_j$ are not natural cubic splines with knots at $X_i^{(j)}, i = 1, \dots, n$. By Theorem 2.2 in Green and Silverman (1994) we can construct natural cubic splines $\hat{g}_j$ with knots at $X_i^{(j)}, i = 1, \dots, n$ such that

$$\hat{g}_j(X_i^{(j)}) = \hat{f}_j(X_i^{(j)})$$

for $i = 1, \dots, n$ and $j = 1, \dots, p$. Hence

$$\left\| \mathbf{Y} - \sum_{j=1}^{p} \hat{g}_j \right\|_n^2 = \left\| Y - \sum_{j=1}^{p} \hat{f}_j \right\|_n^2$$

and

$$\| \hat{g}_j \|_n^2 = \| \hat{f}_j \|_n^2.$$

But by Theorem 2.3 in Green and Silverman (1994), $I^2(\hat{g}_j) \leq I^2(\hat{f}_j)$. Therefore, the value in the objective function (5.7) can be decreased. Hence, the minimizer of (5.7) must lie in the space of natural cubic splines. $\square$

Due to Proposition 5.1, we can restrict ourselves to the finite dimensional space of natural cubic splines instead of considering the infinite dimensional space of continuously differentiable functions with square integrable second derivatives.

5.3.3 Computation

The optimization problem in (5.5) is convex in the $pK \times 1$-dimensional parameter vector $\beta = (\beta_1, \dots, \beta_p)^T$. For large p and K (the latter depending on n such as $K \asymp \sqrt{n}$ or even larger), the optimization is very high-dimensional and efficient algorithms are desirable. In principle, many convex optimization algorithms could be used including interior point methods or second-order cone programming. The problem in (5.5) is of the form

$$\arg\min_{\beta}(g(\beta) + \text{pen}(\beta)),$$

with continuously differentiable function $g(\cdot)$ and a penalty function $\text{pen}(\cdot)$ which is convex and separable with respect to the parameter blocks $\beta_1, \dots, \beta_p$. Hence, a sim-

ple block coordinate descent (Gauss-Seidel) algorithm can be used, similarly as the block coordinate descent Algorithm 2 from Chapter 4, having provable numerical convergence properties in the sense that every cluster point of the iteration sequence is a stationary and hence a minimum point of the convex objective function. This follows from a more general theory in Tseng (2001).

The optimization problem becomes easier when considering orthogonal basis functions and diagonalized smoothing matrices, as used in the estimator defined in (5.6). Then, an adaptation of Algorithm 2 in Section 4.7.1 can be used. However, the block coordinate descent algorithm for the problem in (5.6) is a bit more complex. In particular, despite the squared error loss we are dealing with, the block up-date for non-zero parameters is not explicit.

From the Karush-Kuhn-Tucker (KKT) conditions we have that a necessary and sufficient condition for the solution is:

$$-2H_j^T(\mathbf{Y} - H\hat{\beta}_j) + \lambda_1 e_j + \lambda_2 D t_j = 0 \quad \text{if } \hat{\beta}_j \equiv 0,$$
$$\text{for vectors } e_j, t_j \text{ with } \|e_j\|_2 \leq 1, \|t_j\|_2 \leq 1,$$
$$-2H_j^T(\mathbf{Y} - H\hat{\beta}_j) + \lambda_1 \frac{\hat{\beta}_j}{\|\hat{\beta}_j\|_2} + \lambda_2 \frac{D^2\hat{\beta}_j}{\|D\hat{\beta}_j\|_2} = 0 \quad \text{if } \hat{\beta}_j \neq 0. \tag{5.8}$$

Here, $\hat{\beta}_j \neq 0$ means not equal to the 0-vector. See Problem 5.2.

The following Gauss-Seidel algorithm can be used. Assume that we have a parameter value $\beta^{[m-1]}$ in iteration $m-1$, and we cycle through the groups with indices $j = 1, \ldots, p$. Consider the computation for the jth group in iteration m (i.e. the up-date for obtaining $\beta_j^{[m]}$) keeping all other parameter values fixed. We can then up-date the block of parameters corresponding to the jth group as described in the next subsection.

5.3.3.1 Determining the zero estimates

The idea is now analogous as in Section 4.7.1: we want to determine in an efficient manner whether the parameter up-date $\beta_j^{[m]}$ in the mth iteration of the algorithm should be set to zero. We denote by β_{-j} the parameter vector β whose block components $\beta_j \equiv 0$, i.e. the coefficients of the jth block are set to zero. As in Section 4.7.1, we use the characterization in (5.8). Set $\beta_j^{[m]} = 0$ if for some vectors e_j, t_j with $\|e_j\|_2 \leq 1, \|t_j\|_2 \leq 1$:

$$-2H_j^T(\mathbf{Y} - H\hat{\beta}_{-j}^{[m-1]}) + \lambda_1 e_j + \lambda_2 D t_j = 0. \tag{5.9}$$

Algorithmically, this can be determined as follows. Abbreviate by

$$V_j = 2H_j^T(\mathbf{Y} - H\hat{\beta}_{-j}^{[m-1]}),$$

and hence

$$V_j - \lambda_2 D t_j = \lambda_1 e_j.$$

Now, minimize

$$\hat{t}_j = \underset{t_j; \; \|t_j\|_2 \leq 1}{\arg\min} \|V_j - \lambda_2 D t_j\|_2^2. \qquad (5.10)$$

Clearly, formula (5.9) can only hold if and only if

$$\|V_j - \lambda_2 D \hat{t}_j\|_2 \leq \lambda_1, \qquad (5.11)$$

and hence:

$$\text{if (5.11) holds: set } \beta_j^{[m]} \equiv 0.$$

Note that (5.10) is a standard convex optimization problem.

We elaborate briefly how (5.10) could be computed. The optimization is related to Ridge regression and can be solved as follows:

$$\hat{t}_j(\gamma) = \underset{\gamma_j}{\arg\min} \|V_j - \lambda_2 D t_j\|_2^2 + \gamma \|t_j\|_2^2.$$

The solution is explicit:

$$\hat{t}_j(\gamma)_k = \frac{\lambda_2 d_k}{\lambda_2^2 d_k^2 + \gamma}(V_j)_k \quad (k = 1, \ldots, K).$$

If the unconstrained solution with $\gamma = 0$ satisfies

$$\|\hat{t}_j(0)\|_2 \leq 1, \qquad (5.12)$$

then it must be the solution of (5.10). Otherwise, if $\|\hat{t}_j(0)\|_2 > 1$, the minimum in (5.10) is attained for $\|t_j\|_2 = 1$ and we optimize over γ:

$$\hat{\gamma} = \underset{\gamma > 0}{\arg\min}(\|\hat{t}_j(\gamma)\|_2 - 1)^2, \qquad (5.13)$$

and the solution of (5.10) then equals:

$$(\hat{t}_j)_k = \frac{\lambda_2 d_k}{\lambda_2^2 d_k^2 + \hat{\gamma}}(V_j)_k \quad (k = 1, \ldots, K),$$

where $\hat{\gamma}$ is either 0, if (5.12) holds, or as in (5.13) otherwise.

5.3.3.2 Up-dates for the non-zero estimates

If $\beta_j^{[m]}$ has not been determined to be zero, we make a numerical up-date:

$$\beta_j^{[m]} = \arg\min_{\beta_j} \|\mathbf{U} - H_j\beta_j\|_2^2/n + \lambda_1\|\beta_j\|_2 + \lambda_2\|D\beta_j\|_2,$$

where $\mathbf{U} = \mathbf{Y} - \sum_{k\neq j} H_k\beta_k^{[m-1]}$.

We now summarize the description in Algorithm 4.

Algorithm 4 Block Coordinate Descent Algorithm for estimator in (5.6)

1: Let $\beta^{[0]} \in \mathbb{R}^{pK}$ be an initial parameter vector. Set $m = 0$.
2: **repeat**
3:     Increase m by one: $m \leftarrow m+1$.
       Denote by $\mathscr{S}^{[m]}$ the index cycling through the coordinates $\{1,\dots,p\}$:
       $\mathscr{S}^{[m]} = \mathscr{S}^{[m-1]} + 1 \bmod p$. Abbreviate by $j = \mathscr{S}^{[m]}$ the value of $\mathscr{S}^{[m]}$.
4:     if (5.11) holds: set $\beta_j^{[m]} \equiv 0$,
       else $\beta_j^{[m]} = \arg\min_{\beta_j} \|\mathbf{U} - H_j\beta_j\|_2^2/n + \lambda_1\|\beta_j\|_2 + \lambda_2\|D\beta_j\|_2$,
       where $\mathbf{U} = \mathbf{Y} - \sum_{k\neq j} H_k\beta_k^{[m-1]}$.
5: **until** numerical convergence

An active set strategy as described in Section 4.7.1 should be used here as well to drastically speed up computations.

5.4 A sparsity-smoothness penalty of group Lasso type

Instead of the penalty in (5.4), we can use a modification which allows for more efficient computation relying on a group Lasso structure. Consider the following sparsity-smoothness penalty of Group Lasso type:

$$\text{pen}_{\lambda_1,\lambda_2}(\beta) = \lambda_1 \sum_{j=1}^{p} \sqrt{\|f_j\|_n^2 + \lambda_2^2 I^2(f_j)}$$

$$= \lambda_1 \sum_{j=1}^{p} \sqrt{\|H_j\beta_j\|_2^2/n + \lambda_2^2 \beta_j^T W_j \beta_j}, \qquad (5.14)$$

with smoothing matrices W_j as before. The difference to the penalty in (5.4) is that here, both sparsity and smoothness norms appear in their squared values under the single square root, whereas (5.4) involves both norms in non-squared form and with

no square root common to both of them. From a mathematical point of view, the penalty in (5.4) is more convenient and potentially better than the penalty in (5.14) (compare with Problem 8.3 in Chapter 8).

Analogously as before in formula (5.7), we can also optimize over the Sobolev space of continuously differentiable functions with square integrable second derivatives to obtain the following additive model estimator:

$$\hat{f}_1, \ldots, \hat{f}_p = \operatorname*{arg\,min}_{f_1, \ldots, f_p \in \mathscr{F}} \left\| \mathbf{Y} - \sum_{j=1}^{p} f_j \right\|_n^2 + \lambda_1 \sum_{j=1}^{p} \sqrt{\|f_j\|_n^2 + \lambda_2^2 I^2(f_j)}, \quad (5.15)$$

where $\mathscr{F}$ is the Sobolev space of continuously differentiable functions and $I^2(f_j) = \int |f''(x)|^2 dx$ assumed to be finite. Proposition 5.1 also applies to the problem in (5.15), with exactly the same proof. Thus, the solution of the optimization in (5.15) is given by finite-dimensional basis expansions with natural cubic splines having knots at the observations.

5.4.1 Computational algorithm

In view of the estimator defined in (5.15), leading to solutions based on expansions with natural cubic splines, we consider a cubic B-spline parametrization for every function $f_j(\cdot)$ with a reasonable amount of knots or basis functions. A typical choice would be to use $K - 4 \asymp \sqrt{n}$ interior knots that are placed at the empirical quantiles of $X_1^{(j)}, \ldots, X_n^{(j)}$. Hence, we parametrize

$$f_j(x) = \sum_{k=1}^{K} \beta_{j,k} h_{j,k}(x),$$

where $h_{j,k} : \mathbb{R} \to \mathbb{R}$ are the B-spline basis functions and $\beta_j = (\beta_{j,1}, \ldots, \beta_{j,K})^T$ is the parameter vector corresponding to f_j. With the basis functions we construct an $n \times pK$ design matrix $H = [H_1 | H_2 | \ldots | H_p]$, where H_j is the $n \times K$ design matrix of the B-spline basis of the jth predictor, i.e. $(H_j)_{i,k} = h_{j,k}(X_i^{(j)})$.

The estimator is then defined analogously as in (5.5):

$$\hat{\beta}(\lambda_1, \lambda_2) = \operatorname*{arg\,min}_{\beta} \left\| \mathbf{Y} - H\beta \right\|_n^2 + \lambda_1 \sum_{j=1}^{p} \sqrt{\beta_j^T H_j^T H_j \beta_j / n + \lambda_2^2 \beta_j^T W_j \beta_j}, \quad (5.16)$$

where the $K \times K$ matrix W_j contains the inner products of the second derivatives of the B-spline basis functions, i.e.

$$(W_j)_{k,\ell} = \int h''_{j,k}(x) h''_{j,\ell}(x) \, dx$$

for $k, \ell \in \{1, \ldots, K\}$.

Hence, (5.16) can be re-written as a generalized group Lasso problem, treated in Section 4.5.

$$\hat{\beta} = \underset{\beta=(\beta_1,\ldots,\beta_p)}{\arg\min} \|\mathbf{Y} - H\beta\|_n^2 + \lambda_1 \sum_{j=1}^p \sqrt{\beta_j^T M_j \beta_j}, \qquad (5.17)$$

where $M_j = \frac{1}{n} H_j^T H_j + \lambda_2^2 W_j$. That is, for any fixed λ_2, this is a generalized group Lasso problem. In particular, the existence of a solution is guaranteed (see also the assumption in Proposition 5.1).

Coordinate-wise approaches as described in Section 4.7 are efficient and have rigorous convergence properties. Thus, we are able to compute the estimator exactly, even if p is very large. As described in Section 4.5, we reparametrize to compute the estimator in (5.17). By decomposing (e.g. using the Cholesky decomposition) $M_j = R_j^T R_j$ for some quadratic $K \times K$ matrix R_j and by defining $\tilde{\beta}_j = R_j \beta_j$, $\tilde{H}_j = H_j R_j^{-1}$, (5.17) reduces to

$$\hat{\tilde{\beta}} = \underset{\tilde{\beta}=(\tilde{\beta}_1,\ldots,\tilde{\beta}_p)}{\arg\min} \|\mathbf{Y} - \tilde{H}\tilde{\beta}\|_n^2 + \lambda_1 \sum_{j=1}^p \|\tilde{\beta}_j\|_2. \qquad (5.18)$$

Moreover, there exists a value $\lambda_{1,max} < \infty$ such that $\hat{\tilde{\beta}}_1 = \ldots = \hat{\tilde{\beta}}_p = 0$ for $\lambda_1 \geq \lambda_{1,max}$. This is especially useful to construct a grid of λ_1 candidate values for cross-validation (usually on the log-scale).

Regarding the uniqueness of the identified components, the results are analogous as for the Lasso, see Lemma 2.1. Define $W_{\tilde{H}}(\tilde{\beta}) = \|\mathbf{Y} - \tilde{H}\tilde{\beta}\|_n^2$. We have the following Proposition.

Proposition 5.2. *If $pK \leq n$ and if $\tilde{H}$ has full rank, a unique solution of (5.18) exists. If $pK > n$, there exists a convex set of solutions of (5.18). Moreover, if $\|\nabla W_{\tilde{H}}(\hat{\tilde{\beta}})_j\|_2 < \lambda_1$ then $\hat{\tilde{\beta}}_j = 0$ and all other solutions $\hat{\tilde{\beta}}_{other}$ satisfy $\hat{\tilde{\beta}}_{other,j} = 0$.*

Proof. The first part follows due to the strict convexity of the objective function. Consider now the case $pK > n$. The (necessary and sufficient) conditions for $\hat{\tilde{\beta}}$ to be a solution of the Group-Lasso problem (5.18) are (see Lemma 4.2)

$$\|\nabla W_{\tilde{H}}(\hat{\tilde{\beta}})_j\|_2 = \lambda_1 \text{ for } \hat{\tilde{\beta}}_j \neq 0$$
$$\|\nabla W_{\tilde{H}}(\hat{\tilde{\beta}})_j\|_2 \leq \lambda_1 \text{ for } \hat{\tilde{\beta}}_j = 0.$$

Regarding uniqueness of the pattern of zeroes, we argue as in the proof of Lemma 2.1. Suppose that there are two solutions $\hat{\tilde{\beta}}^{(1)}$ and $\hat{\tilde{\beta}}^{(2)}$ having $\hat{\tilde{\beta}}_j^{(1)} = 0$ and $\|\nabla W_{\tilde{H}}(\hat{\tilde{\beta}}^{(1)})_j\|_2 = c < \lambda_1$ but $\hat{\tilde{\beta}}_j^{(2)} \neq 0$. Use that the set of all solutions is convex, and thus

$$\hat{\hat{\beta}}_\rho = (1-\rho)\hat{\hat{\beta}}^{(1)} + \rho\hat{\hat{\beta}}^{(2)}$$

is also a minimizer for all $\rho \in [0,1]$. Since $\hat{\hat{\beta}}_{\rho,j} \neq 0$ (by assumption) we have that $\|\nabla W_{\tilde{H}}(\hat{\hat{\beta}}_\rho)_j\|_2 = \lambda_1$ for all $\rho \in (0,1)$. Therefore, it holds for $g(\rho) = \|\nabla W_{\tilde{H}}(\hat{\hat{\beta}}_\rho)_j\|_2$ that $g(0) = c < \lambda_1$ and $g(\rho) = \lambda_1$ for all $\rho \in (0,1)$. But this is a contradiction to the fact that $g(\cdot)$ is continuous. Therefore, if the regression coefficients in a component j are all equal to zero (non-active) with $\|\nabla W_{\tilde{H}}(\hat{\hat{\beta}})_j\|_2 < \lambda_1$, such a component j can not be active (non-zero) in any other solution.   □

5.4.2 Alternative approaches

Another, more direct approach to incorporate smoothness could be achieved by applying appropriate regularization in the basis expansions in (5.2) using a suitable choice of K. For example, if each additive function $f_j(\cdot)$ is twice continuously differentiable, we would use a basis expansion (e.g. spline functions) with $K \asymp n^{1/5}$ basis functions $h_{j,k}(\cdot)$ $(k = 1, \ldots K)$:

$$f_j(x) = \sum_{k=1}^{K} \beta_{j,k} h_{j,k}(x).$$

We could then use the generalized group Lasso penalty (i.e. the sparsity penalty)

$$\lambda \sum_{j=1}^{p} \|f_j\|_n = \lambda \sum_{j=1}^{p} \sqrt{\beta_j^T H_j^T H_j \beta_j / n},$$

where $H_j = [h_{j,k}(X_i^{(j)})]_{i=1,\ldots,n;k=1,\ldots,K}$ and $\beta_j = (\beta_{j,1}, \ldots, \beta_{j,K})^T$. Note that this procedure involves two tuning parameters as well, namely λ and K. When having unequally spaced design points $X_i^{(j)}$, the penalty in (5.4) or (5.14) typically performs better as it is more flexible to adapt to unequal spacings and also to varying degree of roughness of the true underlying additive functions.

Alternative possibilities of the penalty in (5.4) or (5.14) include the proposal $\text{pen}_{\lambda_1,\lambda_2} = \lambda_1 \sum_{j=1}^{p} \|f_j\|_n + \lambda_2 \sum_{j=1}^{p} I^2(f_j)$ which basically leads again to a group Lasso problem with an additional Ridge-type quadratic norm regularization (see Problem 5.3). However, it appears to have theoretical drawbacks leading to severely sub-optimal rates of convergence, i.e. the term $\lambda_2 I^2(f_j)$ should appear within the square root or without the power 2, see Chapter 8, Section 8.4.5.

5.5 Numerical examples

In this section, we always use the penalized least squares estimator as in (5.14) with B-spline basis expansions using $K = \lfloor \sqrt{n} \rfloor$. The regularization parameters λ_1 and λ_2 are chosen via cross-validation.

5.5.1 Simulated example

We consider first a simulated example with $n = 150$, $p = 200$, $s_0 = 4$ active functions and signal to noise ratio approximately equal to 15. The model is

$$Y_i = f_1(X_i^{(1)}) + f_2(X_i^{(2)}) + f_3(X_i^{(3)}) + f_4(X_i^{(4)}) + \varepsilon_i, \ \varepsilon_i \text{ i.i.d } \mathcal{N}(0,1),$$

with

$$f_1(x) = -\sin(2x), \ \ f_2(x) = x_2^2 - 25/12,$$
$$f_3(x) = x, \ \ f_4(x) = e^{-x} - 2/5 \cdot \sinh(5/2).$$

The covariates are simulated from independent Uniform(-2.5, 2.5) distributions (which is a relatively "easy" scenario due to independence of the covariates). The
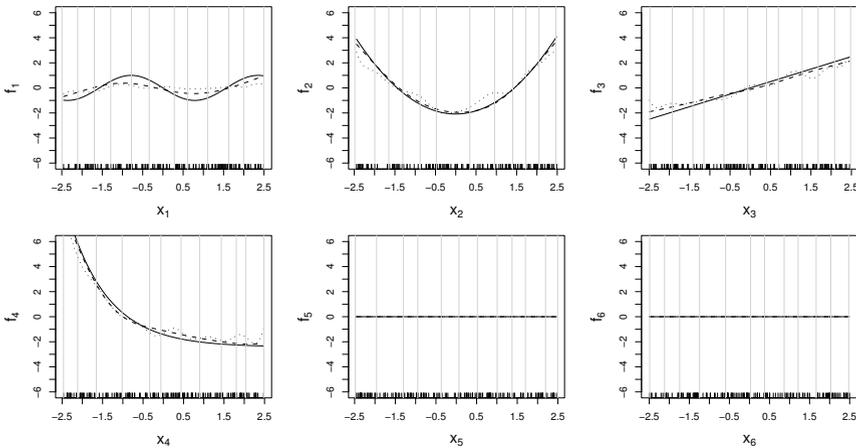
Fig. 5.1 True functions f_j (solid) and estimated functions $\hat{f}_j$ (dashed) for the first 6 components of a single simulation run from the model in Section 5.5.1. Small vertical bars indicate original data and gray vertical lines knot positions. The dotted lines are the function estimates when no smoothness penalty is used, i.e. when setting $\lambda_2 = 0$. The figure is taken from Meier et al. (2009).

true and the estimated functions of a single simulation run are illustrated in Figure 5.1. We see that the additional smoothness penalty term with $\lambda_2 \neq 0$ yields indeed smoother function estimates, although mostly at the fine scale. In addition, the sparsity of the estimator is visible as it (correctly) infers that the fifth and sixth covariates are noise variables (with corresponding smooth function being the trivial zero function). Results of a larger simulation study (with in particular correlated covariates) are reported in Meier et al. (2009) showing that the estimator with the sparsity-smoothness penalty performs very well in comparison to other methods.

5.5.2 Motif regression

We introduced in Section 2.5.2 the problem of motif regression for the HIF1α transcription factor. For our specific dataset, we have binding intensities Y_i of the HIF1α transcription factor at $n = 287$ regions of the DNA sequence. Moreover, for each region i, motif scores $X_i^{(1)}, \ldots, X_i^{(p)}$ of $p = 195$ candidates are available. A motif itself is a candidate for the binding site of the HIF1α transcription factor on the DNA sequence, typically a 5–15bp (base pairs) long DNA sequence. The score $X_i^{(j)}$ measures how well the jth motif is represented in the ith DNA region. The candidate list of motifs and their corresponding scores were created with a variant of the MDScan algorithm (Liu et al., 2002). The main goal is to find the relevant covariates.

We fit an additive model to this data with $n = 287$, $p = 195$ motif scores $X_i = (X_i^{(1)}, \ldots X_i^{(p)})^T$ and with response Y_i describing the real-valued log-transformed binding intensity of the transcription factor in DNA region i:

$$Y_i = \mu + \sum_{j=1}^{195} f_j(X_i^{(j)}) + \varepsilon_i.$$

We used 5 fold cross-validation to determine the prediction optimal tuning parameters yielding 28 active functions: that is, $\hat{S} = \{j; \|\hat{f}_j\|_n \neq 0\} = \{j; \hat{\beta}_j \neq 0\}$ has cardinality 28. To assess the stability of the estimated model, we performed a nonparametric bootstrap analysis. At each of the 100 bootstrap samples, we fit the model with the fixed optimal tuning parameters from above. The two functions which are selected most often in the bootstrapped estimates are depicted in Figure 5.2, i.e. the two most "stable" selections and and their refit using the original data. Chapter 10 will discuss in more detail about bootstrapping and stable variable selection. While the left-hand side plot shows an approximate linear relationship, the effect of the other motif seems to diminish for larger values. In fact, the right panel corresponds to a true (known) binding site. Regarding the motif corresponding to the left panel, a follow-up experiment showed that the transcription factor does not directly bind to this motif, and we may view it as an interesting candidate for a binding site of

a co-factor (another transcription factor) which would need further experimental validation.
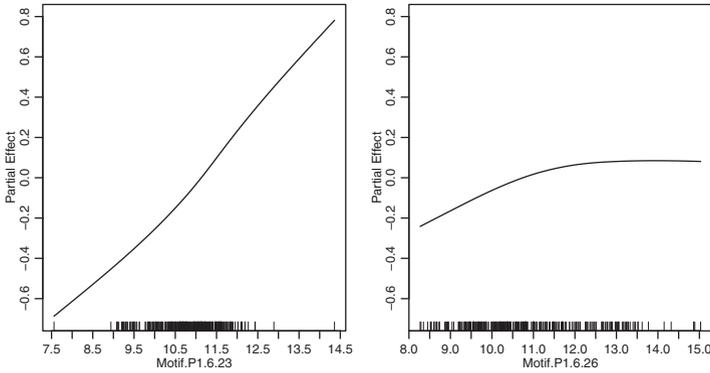
Fig. 5.2 Motif regression. Estimated functions $\hat{f}_j$ (refitted on original data) of the two most stable motifs. Small vertical bar indicate original data. The right panel corresponds to a true known motif while the left panel indicates an interesting candidate for a binding site of a co-factor. The figure is taken from Meier et al. (2009).

We use this real-data example as an illustration that fitting additive models in high dimensions is indeed feasible. It has the potential for better prediction, in this case yielding an improvement of the cross-validated prediction error of about 20% in comparison to a Lasso-estimated linear model, or for better variable selection than with regularized linear modeling.

5.6 Prediction and variable selection

For prediction, we measure the squared discrepancy $\|\hat{f} - f^0\|_n^2$ between the estimated additive function $\hat{f}$ and the true function f^0. This is closely related to the squared test-sample prediction error $\mathbb{E}[(\hat{f}(X_{new}) - Y_{new})^2] = \mathbb{E}[(\hat{f}(X_{new}) - f(X_{new}))^2] + \mathrm{Var}(\varepsilon)$, where $\mathbb{E}$ is with respect to a new test observation (X_{new}, Y_{new}). The theory about prediction and function estimation in the high-dimensional additive modeling framework is presented in Chapter 8, Section 8.4.

A key assumption for deriving consistency $\|\hat{f} - f^0\|_n^2 = o_P(1)$ of the estimator in (5.5), and (5.6) as a special case, is sparsity in the sense that

$$\lambda_1 \sum_{j=1}^p \|f_j^0\|_n + \lambda_2 \sum_{j=1}^p I(f_j^0) = o(1)$$

for $\lambda_1 \asymp n^{-2/5}$ and $\lambda_2 \asymp n^{-4/5}\sqrt{\log(pn)}$ and where $I(.)$ is a smoothness semi-norm such as $I(f) = \int |f''(x)|^2 dx$. The asymptotic relations for λ_1 and λ_2 are implicitly assuming that all f_j^0's are twice continuously differentiable. (See the Basic Inequality in Section 8.4 appearing just before formula (8.5)).

For oracle (optimality) results, an additional compatibility condition on the design is needed ensuring that it is not too strongly ill-posed. Furthermore, we require sparsity of the active set

$$S_0 = \{j;\ \|f_j^0\|_n \neq 0\}$$

whose cardinality is $s_0 = |S_0|$. Then, a result of the following form holds: if all f_j^0's are twice continuously differentiable

$$\|\hat{f} - f^0\|_n^2 = O_P(s_0 \sqrt{\log(p)} n^{-4/5} / \phi^2)$$

assuming that $p \gg n$ and where ϕ^2 depends on the compatibility condition for the design. Thus, we achieve the optimal rate $O(s_0 n^{-4/5})$ up to the factor $\sqrt{\log(p)}$ (and $1/\phi^2$) which is the price of not knowing which additive functions are active. See Theorem 8.2 in Section 8.4.

Analogous in spirit as in Section 2.5 for the Lasso, we can derive a variable screening property, assuming that the non-zero ℓ_2-norms of the coefficient vectors within groups (corresponding to the non-zero functions f_j^0) are sufficiently large (the analogue of the beta-min condition in formula (2.23) in Chapter 2) and requiring a compatibility condition for the design matrix. Then, with high probability, for suitably chosen penalty parameters λ_1 and λ_2, the sparsity-smoothness additive model estimator selects at least all non-zero functions:

$$\hat{S} = \{j;\ \|\hat{f}_j\|_n \neq 0\} \supseteq S_0 = \{j;\ \|f_j^0\|_n \neq 0\}.$$

In view of the restrictive assumptions one needs for the Lasso in a linear model for consistent variable selection, see Section 2.6 and Section 7.5.1, we do not pursue this topic for additive models. From a practical point of view, it is much more realistic that we achieve a reasonably good variable screening result with much reduced dimensionality where $|\hat{S}|$ is much smaller than the number p of all covariates.

5.7 Generalized additive models

The extension to generalized additive models is straightforward. The model relates a univariate response variable Y and a high-dimensional covariate X as follows:

$$Y_1, \ldots, Y_n \text{ independent,}$$

$$g(\mathbb{E}[Y_i|X_i = x]) = f(x) = \mu + \sum_{j=1}^{p} f_j(x^{(j)}), \qquad (5.19)$$

where $g(\cdot)$ is a specified known link function, μ is the intercept term and f_j are smooth univariate functions as in (5.1). For identification purposes we assume that all f_j's are centered, i.e.

$$\sum_{i=1}^{n} f_j(X_i^{(j)}) = 0$$

for $j = 1, \ldots, p$. The design points X_i are allowed to be either fixed or random.

As in generalized linear models in (3.1), examples include Bernoulli- or Poisson-distributed response variables. Typical link functions are described in Chapter 3.

Estimation of a high-dimensional generalized additive model can be done analogously to (5.4) or (5.7):

$$\hat{\mu}, \hat{f}_1, \ldots, \hat{f}_p = \operatorname*{arg\,min}_{\mu; f_1, \ldots, f_p \in \mathscr{F}} -n^{-1} \sum_{i=1}^{n} \log(p_{f(X_i)}(Y_i|X_i)) + \sum_{j=1}^{p} \mathrm{pen}_{\lambda_1, \lambda_2}(f_j), \quad (5.20)$$

where $p_{f(x)}(y|x)$ is the density of $Y|X = x$ which depends only on $f(x) = \mu + \sum_{j=1}^{p} f_j(x^{(j)})$ and

$$\mathrm{pen}_{\lambda_1, \lambda_2}(f_j) = \lambda_1 \sum_{j=1}^{p} \|f_j\|_n + \lambda_2 \sum_{j=1}^{p} I(f_j)$$

is as in (5.4) or (5.7). Note that we have to deal here with an unpenalized intercept term μ. This does not create any further difficulties. The optimization can be done along the lines described in Section 5.3.3 but using the negative gradient of the negative log-likelihood loss and allowing for an unpenalized intercept term.

When using the alternative sparsity-smoothness penalty of group Lasso type from (5.14), the optimization in (5.20) becomes a Group Lasso problem, as discussed in Section 5.4.1, which involves here an additional unpenalized intercept term. The algorithms presented in Section 4.7 can be used and they easily allow to deal with additional unpenalized terms.

5.8 Linear model with varying coefficients

Another useful extension of the linear model in (2.1), but based on an additional "time" component of the data structure, is a regression model observed at different units, such as time, whose coefficients are smoothly changing:

$$Y_i(t_r) = \mu + \sum_{j=1}^{p} \beta_j(t_r) X_i^{(j)}(t_r) + \varepsilon_i(t_r), \ i = 1, \ldots, n, \ r = 1, \ldots, T, \qquad (5.21)$$

where $\{\varepsilon_i(t_r); \ i = 1, \ldots, n, r = 1, \ldots T\}$ are i.i.d., independent of $\{X_i(t_r); \ i = 1, \ldots, n, r = 1, \ldots, T\}$ and with $\mathbb{E}[\varepsilon_i(t_r)] = 0$. The covariates are either random or fixed and the regression coefficients change smoothly with respect to t_r, that is $\beta_j(\cdot)$ are smooth univariate functions. If the first covariate represents an intercept term, the model (assuming p covariates plus an intercept) can be represented as

$$Y_i(t_r) = \mu + \beta_0(t_r) + \sum_{j=1}^{p} \beta_j(t_r) X_i^{(j)}(t_r) + \varepsilon_i(t_r),$$

with identifiability constraint $\sum_{r=1}^{T} \beta_0(t_r) = 0$.

The model (5.21) involves the estimation of p univariate smooth functions $\beta_j(\cdot)$ and thus, we exploit a close relation to the additive model in (5.1). We proceed similarly by using the estimator:

$$\hat{\mu}, \hat{\beta}_1, \ldots, \hat{\beta}_p =$$

$$\underset{\mu; \beta_1, \ldots, \beta_p \in \mathscr{F}}{\arg\min} \left(T^{-1} n^{-1} \sum_{r=1}^{T} \sum_{i=1}^{n} (Y_i(t_r) - \mu - \sum_{j=1}^{p} \beta_j(t_r) X_i^{(j)}(t_r))^2 + \sum_{j=1}^{p} \mathrm{pen}_{\lambda_1, \lambda_2}(\beta_j) \right),$$

$$(5.22)$$

where $\mathscr{F}$ is a suitable class of functions (e.g. the Sobolev space of continuously differentiable functions with square integrable second derivatives) and the sparsity-smoothness penalty is as in (5.4),

$$\mathrm{pen}_{\lambda_1, \lambda_2}(\beta_j) = \lambda_1 \|\beta_j\|_n + \lambda_2 I(\beta_j),$$

or using the alternative sparsity-smoothness penalty of group Lasso type as in (5.14):

$$\mathrm{pen}_{\lambda_1, \lambda_2}(\beta_j) = \lambda_1 \sqrt{\|\beta_j\|_n^2 + \lambda_2^2 I^2(\beta_j)}. \qquad (5.23)$$

Analogously to Proposition 5.1, when optimizing in (5.22) over the space of continuously differentiable functions with square integrable second derivatives, the solutions are natural cubic splines with knots at t_r, $r = 1, \ldots T$. This fact can be derived analogously to the proof of Proposition 5.1. Due to the sparsity-smoothness penalty, the solution will be sparse in the sense that some functions $\hat{\beta}_j(\cdot) \equiv 0$, depending on the data and the magnitude of the tuning parameters λ_1 and λ_2.

When using the penalty in (5.23), the estimator in (5.22) can be re-written in terms of a (generalized) group Lasso problem, in an analogous way as in Section 5.4.1 for the additive modeling estimator. We leave the details as Problem 5.4.

5.8.1 Properties for prediction

The accuracy of the estimator in (5.22) can be measured in terms of estimating the functions $\{\beta_j(\cdot); \ j = 1, \ldots, p\}$ or in terms of predicting a new response Y_{new}. We will discuss in Chapter 8, Section 8.5 the prediction properties in the high-dimensional context. Key assumptions for deriving consistency or oracle results are again of the same nature as for the group Lasso or the Lasso: we need a sparsity assumption and for oracle (optimality) results, we require an additional compatibility condition which excludes ill-posed designs.

5.8.2 Multivariate linear model

It is interesting to make a connection to the multivariate regression model: if $X_i^{(j)}(t_r) \equiv X_i^{(j)}$ for all t_r, then the model in (5.21) becomes a multivariate linear model

$$Y_i(t_r) = \mu + \sum_{j=1}^{p} \beta_j(t_r)X_i^{(j)} + \varepsilon_i(t_r).$$

If we make no smoothness assumption on $\beta_j(\cdot)$, that is the pT parameters $\beta_j(t_r)$ ($j = 1, \ldots, p; t = 1, \ldots, T$) are unrelated of each other, we obtain the standard multivariate linear model. The indices $t_r \in \{1, \ldots T\}$ could then be replaced by indices $t \in \{1, \ldots, T\}$ since without smoothness, there is no need to consider closeness of different t_r's.

We will show in Chapter 8, Sections 8.5 and 8.6, that without smoothness, we gain relatively little over running the Lasso for all T regression problems separately. Simultaneous estimation as in (5.22) is advantageous in presence of smoothness structure for the regression coefficients with respect to t_r. The theory shows for example that when optimizing over the space of continuously differentiable functions, the estimator has much better performance than T single Lasso's for $T \gg (n/\log(p))^{1/4}$.

5.9 Multitask learning

We have briefly touched in the previous section on the multivariate linear model:

$$Y_i(t) = \sum_{j=1}^{p} \beta_j(t)X_i^{(j)} + \varepsilon_i(t), \ i = 1, \ldots, n, \ t = 1, \ldots, T, \tag{5.24}$$

where $\{\varepsilon_i(t);\ i = 1,\ldots,n, t = 1,\ldots,T\}$ are i.i.d., independent of $\{X_i;\ i = 1,\ldots,n\}$, with $\mathbb{E}[\varepsilon_i(t)] = 0$, and the covariates are either fixed or random covariates. Note the relation to the model in (5.21). There, the regression coefficients $\beta_j(t_r)$ are structured as smooth univariate functions, and hence the appearance of the indices t_r which denote potentially non-equispaced points. Furthermore, the model in (5.21) allows for different covariates $X_i^{(j)}(t_r)$ when varying r, a feature which we will include in the following extension of (5.24).

Consider the slight extension of (5.24) where the covariates may also change over the indices t:

$$Y_i(t) = \sum_{j=1}^{p} \beta_j(t)X_i^{(j)}(t) + \varepsilon_i(t),\ i = 1,\ldots,n,\ t = 1,\ldots,T, \tag{5.25}$$

where $\varepsilon_i(t)$ are as in (5.24) and $X_i^{(j)}(t)$ are either fixed or random covariates. This model is the analogue of (5.21) but with no structure on the coefficients $\beta_j(t)$. Estimation of the coefficients from the model in (5.25) can be seen as a special case of a high-dimensional multitask learning problem (Caruana, 1997).

We can estimate the unknown parameters $\beta_j(t)$ either by using the Lasso for each of the $t = 1,\ldots,T$ regression separately, i.e.,

$$\hat{\beta}(t) = \underset{\beta_1,\ldots,\beta_p}{\arg\min}\left(n^{-1}\sum_{i=1}^{n}(Y_i(t) - \sum_{j=1}^{p}\beta_j X_i^{(j)}(t))^2 + \lambda\|\beta\|_1\right),\ t = 1,\ldots,T.$$

As usual, the Lasso will be sparse that some $\hat{\beta}_j(t) = 0$ for some indices j and t but without exploiting any structure among the $t = 1,\ldots T$ different regressions. The Lasso approach is easy and straightforward as it works by considering the T regression problems separately.

Alternatively, we can use the group Lasso by considering the group vectors $\beta_{\mathcal{G}_j} = (\beta_j(1),\ldots,\beta_j(T))^T$ for each $j = 1,\ldots,p$:

$$\{\hat{\beta}_j(t);\ j = 1,\ldots,p, t = 1,\ldots,T\}$$
$$= \underset{\{\beta_j(t);\ j,t\}}{\arg\min}\left(n^{-1}T^{-1}\sum_{i=1}^{n}\sum_{t=1}^{T}(Y_i(t) - \sum_{j=1}^{p}\beta_j(t)X_i^{(j)}(t))^2 + \lambda\sum_{j=1}^{p}\|\beta_{\mathcal{G}_j}\|_2\right).$$

This group Lasso estimator has the sparsity property that for some covariates j, the parameter estimates $\hat{\beta}_j(t) \equiv 0$ for all indices t. In some applications this is a desirable property, for example when doing variable selection in terms of covariates j over all time points t.

We will show in Chapter 8, Section 8.6 that in terms of prediction, there is a gain by a $\log(p)$ factor if T is large, in comparison to using the decoupled Lasso approach over T individual univariate response regressions; see the discussion of Theorem 8.4 in Chapter 8. Furthermore, there are some differences in the underlying assumptions

(about compatibility conditions or restricted eigenvalues for the underlying design matrices): Lemma 8.6 (Chapter 8) shows that the conditions are becoming weaker when considering the multivariate case simultaneously.

Problems

5.1. Consider the B-spline basis: for literature, see e.g. in Hastie et al. (2001). Examine formula (5.2) with $K = 10$. Where would you place the knots for the $10 \cdot p$ different B-spline basis functions?

5.2. Block coordinate descent Algorithm 4
Prove the characterization of a solution given in formula (5.8).

5.3. Consider the additive model estimator minimizing the function

$$\|\mathbf{Y} - H\beta\|_2^2/n + \lambda_1 \sum_{j=1}^{p} \|f_j\|_n + \lambda_2 \sum_{j=1}^{p} I^2(f_j),$$

where f_j is as in (5.2) and H as in Section 5.2.1. Show that this can be re-written as a generalized group Lasso estimator by an appropriate extension of the design matrix H.

5.4. Show that the optimization in (5.22) with the penalty in (5.23) can be re-written as a generalized group Lasso problem.

5.5. Consider a multivariate linear model with correlated errors (and fixed covariates):

$$Y_i(t) = \sum_{j=1}^{p} \beta_j(t)X_i^{(j)} + \varepsilon_i(t), \ i = 1,\dots,n, \ t = 1,\dots,T,$$

where for $\varepsilon_i = (\varepsilon_i(1),\dots,\varepsilon_i(T))^T$, $\varepsilon_1,\dots,\varepsilon_n$ are i.i.d. $\mathcal{N}_T(0,\Sigma_\varepsilon)$.
(a) Assume that Σ_ε is known. Write down the penalized maximum likelihood estimator with group Lasso penalty.
(b) Consider the case where Σ_ε is unknown. Assume that T is small relative to sample size n. What kind of procedure would you use to estimate Σ_ε (and this estimate could then be plugged-in to the expression derived in (a)).

Chapter 6
Theory for the Lasso

Abstract We study the Lasso, i.e., ℓ_1-penalized empirical risk minimization, for general convex loss functions. The aim is to show that the Lasso penalty enjoys good theoretical properties, in the sense that its prediction error is of the same order of magnitude as the prediction error one would have if one knew a priori which variables are relevant. The chapter starts out with squared error loss with fixed design, because there the derivations are the simplest. For more general loss, we defer the probabilistic arguments to Chapter 14. We allow for misspecification of the (generalized) linear model, and will consider an oracle that represents the best approximation within the model of the truth. An important quantity in the results will be the so-called *compatibility constant*, which we require to be non-zero. The latter requirement is called the *compatibility condition*, a condition with eigenvalue-flavor to it. Our bounds (for prediction error, etc.) are given in explicit (non-asymptotic) form.

6.1 Organization of this chapter

We start out in Section 6.2 with the case of squared error loss with fixed design. The theoretical properties then follow from some rather straightforward inequalities and a probability inequality for the error term. A key condition, encountered here and also more generally, is the so-called compatibility condition, which is an identifiability assumption in terms of the ℓ_1-norm of the coefficients in the model.

Once the theory for squared error loss is established, it is relatively clear how to extend the results to general loss. We will consider convex loss throughout this chapter. With convexity, we are able to localize the problem (that is, one only needs to consider a neighborhood of the "truth", or more generally, a neighborhood of a linear approximation of the "truth"). The results can be found in Sections 6.3- 6.7. After an introduction to general convex loss in Section 6.3, Section 6.4 discusses the

so-called margin condition, which is the behavior of the theoretical risk near its minimizer. Section 6.5 provides a benchmark for how good empirical risk minimizers behave if one knows the relevant variables. Section 6.6 gives conditions for consistency of the Lasso, and Section 6.7 presents the main result: an oracle inequality for the Lasso (see Theorem 6.4). Section 6.8 examines the ℓ_q-error of the Lasso, for $1 \leq q \leq 2$. A brief discussion of the weighted Lasso is given in Section 6.9.

This chapter is only about prediction error, and ℓ_q-error of estimated coefficients, $1 \leq q \leq 2$, and not about the selection of variables (the latter topic will be treated in Chapter 7). The results follow from a so-called Basic Inequality, which is a reformulation of the "arg min" property. This property is based on the fact that the Lasso estimator is a penalized empirical risk minimizer and hence its penalized empirical risk is not larger than that of any other parameter choice, in particular, that of a suitable "oracle". For the selection of variables, one takes the KKT conditions as starting point (see Lemma 2.1, and also Section 7.5), instead of the "arg min" property. For selection of variables, under only a compatibility condition, the Lasso appears to be not well-calibrated. More refined two stage procedures (the adaptive Lasso, see Section 7.6) have been proposed. As an intermediate zone, we study the prediction error using the adaptively weighted Lasso and the related concave penalties in Sections 6.10 and 6.11, returning to squared error loss with fixed design for simplicity. These results will be applied in Chapter 7 to the variable selection problem.

The results are presented in a non-asymptotic form. To better understand their implications, we sometimes present a short asymptotic formulation.

Actually, much of the effort goes into showing that the ℓ_1-penalty resembles the ℓ_0-penalty. One could argue that this is a topic belonging to approximation theory, rather than statistics. We refer to Candès and Tao (2005) and Donoho (2006) for important results in a noiseless setting. We consider in this chapter the prediction error in a noisy setting. Due to the noise, this statistical problem is in a sense harder than the corresponding noiseless deterministic problem. Yet, most of the random part can be separated from the approximation theory part. The probabilistic arguments are only briefly addressed in this chapter. They come down to showing that a certain event, which we generically denote as $\mathscr{T}$, has large probability. Most probabilistic arguments are deferred to Chapter 14.

Some simple random matrices are considered in Section 6.12. There, we show that the conditions on the empirical design matrix when the design is random can be replaced by population counterparts on the population design, provided there is enough sparsity (see Corollary 6.8).

We give a more detailed account of the compatibility condition in the last section of this chapter, Section 6.13. There, its relation with conditions in the literature are given, such as restricted eigenvalue conditions (Koltchinskii, 2009b; Bickel et al., 2009) restricted isometry conditions (Candès and Tao, 2005), and coherence conditions (Bunea et al., 2007a,b,c).

6.2 Least squares and the Lasso

The aim of this section is to describe the main theoretical arguments for establishing oracle inequalities for the Lasso. The easiest context to do this is the one of squared error loss with fixed design. However, virtually all results carry over to more general loss functions and to random design: see Sections 6.3- 6.7.

6.2.1 Introduction

The linear model, as described in Chapter 2, is

$$Y_i = \sum_{j=1}^{p} \beta_j X_i^{(j)} + \varepsilon_i, \ i = 1, \dots, n,$$

or, in matrix notation,

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon,$$

with $\mathbf{Y}_{n \times 1}$ the vector of responses, $\mathbf{X}_{n \times p}$ the design matrix, and $\varepsilon_{n \times 1}$ the vector of measurement errors. To simplify in this section, we assume that the design $\mathbf{X}$ is fixed, and that ε is $\mathcal{N}(0, \sigma^2 I)$-distributed. We moreover assume in Subsection 6.2.2 that the linear model holds exactly, with some "true parameter value" β^0. One may argue that this can be done without loss of generality by a projection argument.[1]

We first sketch what we mean by an " oracle inequality". Suppose for the moment that $p \le n$ and that $\mathbf{X}$ has full rank p. Consider the least squares estimator in the linear model

$$\hat{b} := (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}.$$

Then from standard least squares theory, we know that the prediction error

$$\|\mathbf{X}(\hat{b} - \beta^0)\|_2^2 / \sigma^2$$

is χ_p^2-distributed, i.e., it has the chi-square distribution with p degrees of freedom. In particular, this means that

$$\frac{\mathbb{E}\|\mathbf{X}(\hat{b} - \beta^0)\|_2^2}{n} = \frac{\sigma^2}{n} p.$$

---

[1] Let $\mathbb{E}\mathbf{Y} := \mathbf{f}^0$ be the regression of $\mathbf{Y}$ on $\mathbf{X}$. One can study the estimator of the projection $\mathbf{f}^0_{\mathrm{LM}} := \mathbf{X}\beta^0$ of $\mathbf{f}^0$ on the span of the columns of $\mathbf{X}$, and, by orthogonality (Pythagoras' Theorem), separate the estimation of $\mathbf{f}^0_{\mathrm{LM}}$ from the model bias $\|\mathbf{f}^0_{\mathrm{LM}} - \mathbf{f}^0\|_2$. (See also Koltchinskii et al. (2010b) for refined projection arguments.) We note however that projection arguments are less straightforward when minimizing over only a subset of β's, and that moreover, for other loss functions, projection arguments will be more involved, or that in fact a separation between estimation and approximation error is not possible.

In words: after reparametrizing to orthonormal design, each parameter β_j^0 is estimated with squared accuracy σ^2/n, $j = 1, \ldots, p$. The overall squared accuracy is then $(\sigma^2/n) \times p$.

We now turn to the situation where possibly $p > n$. The philosophy that will generally rescue us, is to "believe" that in fact only a few, say s_0, of the β_j^0 are non-zero. We use the notation

$$S_0 := \{j : \beta_j^0 \neq 0\},$$

so that $s_0 = |S_0|$. We call S_0 the *active set*, and s_0 the *sparsity index* of β^0. If we would know S_0, we could simply neglect all variables $\mathbf{X}^{(j)}$ with $j \notin S_0$. Then, by the above argument, the overall squared accuracy would be $(\sigma^2/n) \times s_0$.

Because S_0 is not known, one needs some kind of regularization penalty. A, both mathematically and computationally, sensible choice is the ℓ_1-penalty, i.e., the Lasso

$$\hat{\beta} := \arg\min_{\beta} \left\{ \frac{\|\mathbf{Y} - \mathbf{X}\beta\|_2^2}{n} + \lambda \|\beta\|_1 \right\}.$$

Indeed, the Lasso turns out to have good theoretical properties, as we will show in Corollary 6.2. Loosely speaking, we show that, with a proper choice for λ (of order $\sigma\sqrt{\log p/n}$), one has the "oracle inequality"

$$\frac{\|\mathbf{X}(\hat{\beta} - \beta^0)\|_2^2}{n} \leq \text{const.} \frac{\sigma^2 \log p}{n} s_0,$$

with large probability. For the "const." here, one can find explicit values which only depend on p or n through the (scaled) Gram matrix $\hat{\Sigma} := \mathbf{X}^T\mathbf{X}/n$. Note that we have inserted an additional $(\log p)$-factor, which can be seen as the price to pay for not knowing the active set S_0 (see also Donoho and Johnstone (1994)).

One may argue that our oracle is somewhat too modest. There may for example be very many non-zero $|\beta_j^0|$ which are actually very small (say smaller than the noise level $\sqrt{\sigma^2/n}$). Indeed, in that case, one would rather want to have an oracle bound which is proportional to the number of significantly non-zero β_j^0 times $\sigma^2 \log p/n$. This extension is mathematically of the same nature as the extension where the linear model is not assumed to hold exactly, which will be treated in Subsection 6.2.3.

6.2.2 The result assuming the truth is linear

The basis of all our derivations (as far as prediction error is concerned; for model selection we will use additional arguments) is the following so-called Basic Inequality.

Lemma 6.1. (Basic Inequality) *We have*

$$\|\mathbf{X}(\hat{\beta}-\beta^0)\|_2^2/n + \lambda\|\hat{\beta}\|_1 \leq 2\varepsilon^T\mathbf{X}(\hat{\beta}-\beta^0)/n + \lambda\|\beta^0\|_1.$$

Proof. This is simply rewriting the inequality

$$\|\mathbf{Y}-\mathbf{X}\hat{\beta}\|_2^2/n + \lambda\|\hat{\beta}\|_1 \leq \|\mathbf{Y}-\mathbf{X}\beta^0\|_2^2/n + \lambda\|\beta^0\|_1.$$

$\square$

For the case of quadratic loss, the term

$$2\varepsilon^T\mathbf{X}(\hat{\beta}-\beta^0)/n \tag{6.1}$$

will be called the "empirical process" part of the problem. It is the term where the measurement error plays a role, i.e., the random part. The empirical process part for quadratic loss can be easily bounded in terms of the ℓ_1-norm of the parameters involved:

$$2|\varepsilon^T\mathbf{X}(\hat{\beta}-\beta^0)| \leq \left(\max_{1\leq j\leq p} 2|\varepsilon^T\mathbf{X}^{(j)}|\right)\|\hat{\beta}-\beta^0\|_1. \tag{6.2}$$

The idea of the penalty is that it should typically "overrule" the empirical process part. Let us therefore introduce the set

$$\mathscr{T} := \left\{\max_{1\leq j\leq p} 2|\varepsilon^T\mathbf{X}^{(j)}|/n \leq \lambda_0\right\}.$$

where we assume (quite arbitrarily) that $\lambda \geq 2\lambda_0$ to make sure that on $\mathscr{T}$ we can get rid of the random part of the problem.

It is not difficult to show that for a suitable value of λ_0, the set $\mathscr{T}$ has large probability. Indeed, with Gaussian errors, the argument goes as follows. Let us denote the diagonal elements of the Gram matrix (scaled by $1/n$) $\hat{\Sigma} := \mathbf{X}^T\mathbf{X}/n$, by

$$\hat{\sigma}_j^2 := \hat{\Sigma}_{j,j}, \; j = 1,\ldots,p.$$

Lemma 6.2. *Suppose that $\hat{\sigma}_j^2 = 1$ for all j. Then we have for all $t > 0$, and for*

$$\lambda_0 := 2\sigma\sqrt{\frac{t^2 + 2\log p}{n}},$$

$$\mathbf{P}(\mathscr{T}) \geq 1 - 2\exp[-t^2/2].$$

Proof. As $\hat{\sigma}_j^2 = 1$, we know that $V_j := \varepsilon^T\mathbf{X}^{(j)}/(\sqrt{n\sigma^2})$ is $\mathcal{N}(0,1)$-distributed. So

$$\mathbf{P}\left(\max_{1\leq j\leq p}|V_j| > \sqrt{t^2 + 2\log p}\right) \leq 2p\exp\left[-\frac{t^2 + 2\log p}{2}\right] = 2\exp\left[-\frac{t^2}{2}\right].$$

$\square$

We now first apply Lemmas 6.1 and 6.2 to establish consistency of the Lasso. The result is analogous to Greenshtein (2006) (see also our discussion in Chapter 2).

Corollary 6.1. *(Consistency of the Lasso) Assume that $\hat{\sigma}_j^2 = 1$ for all j. For some $t > 0$, let the regularization parameter be*

$$\lambda = 4\hat{\sigma}\sqrt{\frac{t^2 + 2\log p}{n}},$$

where $\hat{\sigma}$ is some estimator of σ. Then with probability at least $1 - \alpha$, where

$$\alpha := 2\exp[-t^2/2] + \mathbf{P}(\hat{\sigma} \le \sigma),$$

we have

$$2\|\mathbf{X}(\hat{\beta} - \beta^0)\|_2^2/n \le 3\lambda\|\beta^0\|_1.$$

Asymptotics We conclude that taking the regularization parameter λ of order $\sqrt{\log p/n}$, and assuming that the ℓ_1-norm for the true β^0 is of smaller order than $\sqrt{n/\log p}$, results in consistency of the Lasso. The paper Bartlett et al. (2009) presents - modulo log-terms - the corresponding result for the case of random design (see also Theorem 14.6 in Chapter 14, which is from Guédon et al. (2007)).

We need that the estimator $\hat{\sigma}$ of σ is well-chosen, i.e., not too small, but also not much too large. One may consider the estimator

$$\hat{\sigma}^2 := \mathbf{Y}^T\mathbf{Y}/n$$

(after centering $\mathbf{Y}$, see also Section 6.9[2]). The signal-to-noise ratio is

$$SNR := \frac{\|\mathbf{X}\beta^0\|_2}{\sqrt{n}\sigma}.$$

By Problem 6.1, for any reasonable signal-to-noise ratio SNR, the estimator $\hat{\sigma}^2 = \mathbf{Y}^T\mathbf{Y}/n$ satisfies $\sigma \le \hat{\sigma} \le \text{const.}\sigma$, with the "const." well under control.

We now return to the more refined oracle inequality. To exploit the sparsity of β^0, we need to introduce some more notation. Let us write, for an index set $S \subset \{1,\ldots,p\}$,

$$\beta_{j,S} := \beta_j 1\{j \in S\},$$

and (hence)

---

[2] Generally, in practice, one first centers both the $\mathbf{X}^{(j)}$ as well as $\mathbf{Y}$, and then proceeds with the Lasso on the centered data. Equivalently, one may keep an intercept, but leave the intercept unpenalized. As the centered $\mathbf{Y}$ no longer has independent components, and we also as yet do not consider unpenalized terms, our theory does not immediately go through. However, no essentially new arguments are required to deal with the situation. To avoid digressions, we do not treat the issue here, but defer it to Section 6.9.

$$\beta_{j,S^c} := \beta_j 1\{j \notin S\}.$$

Thus β_S has zeroes outside the set S, and the elements of β_{S^c} can only be non-zero in the complement S^c of S. Clearly, $\beta = \beta_S + \beta_{S^c}$. The next lemma is a starting point for the deterministic part of the problem.

Lemma 6.3. *We have on $\mathscr{T}$, with $\lambda \geq 2\lambda_0$,*

$$2\|\mathbf{X}(\hat{\beta} - \beta^0)\|_2^2/n + \lambda \|\hat{\beta}_{S_0^c}\|_1 \leq 3\lambda \|\hat{\beta}_{S_0} - \beta_{S_0}^0\|_1.$$

Proof. On $\mathscr{T}$, by the Basic Inequality, and using $2\lambda_0 \leq \lambda$,

$$2\|\mathbf{X}(\hat{\beta} - \beta^0)\|_2^2/n + 2\lambda \|\hat{\beta}\|_1 \leq \lambda \|\hat{\beta} - \beta^0\|_1 + 2\lambda \|\beta^0\|_1.$$

But on the left-hand side, using the triangle inequality,

$$\|\hat{\beta}\|_1 = \|\hat{\beta}_{S_0}\|_1 + \|\hat{\beta}_{S_0^c}\|_1 \geq \|\beta_{S_0}^0\|_1 - \|\hat{\beta}_{S_0} - \beta_{S_0}^0\|_1 + \|\hat{\beta}_{S_0^c}\|_1,$$

whereas on the right-hand side, we can invoke

$$\|\hat{\beta} - \beta^0\|_1 = \|\hat{\beta}_{S_0} - \beta_{S_0}^0\|_1 + \|\hat{\beta}_{S_0^c}\|_1.$$

$\square$

We will need certain conditions on the design matrix to make the theory work. These conditions will be referred to as "compatibility conditions", as they require a certain compatibility of ℓ_1-norms with ℓ_2-norms. In fact, a compatibility condition is nothing else than simply an assumption that makes our proof go through.

In Lemma 6.3 above, a term involving the ℓ_1-norm $\|\hat{\beta}_{S_0} - \beta_{S_0}^0\|_1$ occurs on the right-hand side. To get rid of it, we want to somehow incorporate it into the term $2\|\mathbf{X}(\hat{\beta} - \beta^0)\|_2^2/n$ at the left-hand side. Now clearly, by the Cauchy-Schwarz inequality, we can replace the ℓ_1-norm by the ℓ_2-norm, paying a price $\sqrt{s_0}$:

$$\|\hat{\beta}_{S_0} - \beta_{S_0}^0\|_1 \leq \sqrt{s_0}\|\hat{\beta}_{S_0} - \beta_{S_0}^0\|_2.$$

We are now in the ℓ_2-world, where we have to relate $\|\hat{\beta}_{S_0} - \beta_{S_0}^0\|_2$ to $\|\mathbf{X}(\hat{\beta} - \beta^0)\|_2$. Recall the (scaled) Gram matrix

$$\hat{\Sigma} := \mathbf{X}^T \mathbf{X}/n,$$

so that

$$\|\mathbf{X}(\hat{\beta} - \beta^0)\|_2^2/n = (\hat{\beta} - \beta^0)^T \hat{\Sigma} (\hat{\beta} - \beta^0).$$

If for some constant $\phi_0 > 0$,

$$\|\hat{\beta}_{S_0}^0 - \beta_{S_0}^0\|_2^2 \leq (\hat{\beta} - \beta^0)^T \hat{\Sigma} (\hat{\beta} - \beta^0)/\phi_0^2, \tag{6.3}$$

we can continue our chain of inequalities in a desirable way. However, as $\hat{\beta}$ is random, we need inequalities for a whole class of β's. It may be too rough to require (6.3) for **all** β, as this needs $\hat{\Sigma}$ to be non-singular (which is of course troublesome: with $p > n$, $\hat{\Sigma}$ is always singular). However, we know from Lemma 6.3, that on $\mathscr{T}$,

$$\|\hat{\beta}_{S_0^c}\|_1 \leq 3\|\hat{\beta}_{S_0} - \beta_{S_0}^0\|_1.$$

Thus we may restrict ourselves to such β (provided $\mathscr{T}$ has large probability). The compatibility condition is exactly this.

Compatibility condition *We say that the* compatibility condition *is met for the set* S_0, *if for some* $\phi_0 > 0$, *and for all* β *satisfying* $\|\beta_{S_0^c}\|_1 \leq 3\|\beta_{S_0}\|_1$, *it holds that*

$$\|\beta_{S_0}\|_1^2 \leq \left(\beta^T \hat{\Sigma} \beta\right) s_0/\phi_0^2. \tag{6.4}$$

We remark that the constant 3 appearing in this definition is quite arbitrary. It can be replaced by any constant bigger than 1, when we adjust some other constants (in particular in the lower bound for λ).

The next question is then of course: when does this compatibility condition actually hold? We first recall that if in (6.4), we replace $\|\beta_{S_0}\|_1^2$ by its upper bound $s_0\|\beta_{S_0}\|_2^2$, the condition is similar to a condition on the smallest eigenvalue of $\hat{\Sigma}$. But the restriction $\|\beta_{S_0^c}\|_1 \leq 3\|\beta_{S_0}\|_1$ puts a limitation on the set of β's for which (6.4) is required, so that it is in fact weaker than imposing non-zero eigenvalues. Note further that it can not be checked in practice, as S_0 is unknown. If its cardinality $s_0 = |S_0|$ were known, it is of course sufficient to check the inequalities for **all** sets $S \subset \{1, \ldots, p\}$ with cardinality s_0. This is called the *restricted eigenvalue assumption* in Bickel et al. (2009). The assumption can also be found in Koltchinskii (2009a,b). We call ϕ_0^2 a *compatibility constant* (and following Bickel et al. (2009), sometimes refer to it as a (lower bound for the) *restricted eigenvalue*) of the matrix $\hat{\Sigma}$. Finally, observe that we can merge s_0/ϕ_0^2 into one constant, say $\psi_0^2 := s_0/\phi_0^2$. The reason why we do not use this notation is solely to facilitate the interpretation. In Section 6.13, we will further clarify the relation of compatibility conditions with restricted eigenvalue conditions, coherence conditions (Bunea et al., 2006, 2007a,b,c) and restricted isometry conditions (Candès and Tao, 2007). Furthermore, we show in Section 6.12 that it suffices to have the compatibility condition with $\hat{\Sigma}$ replaced by a suitable approximation (Corollary 6.8) (for example a population variant Σ of the empirical Gram matrix $\hat{\Sigma}$).

An oracle inequality now reads as follows.

Theorem 6.1. *Suppose the compatibility condition holds for* S_0. *Then on* $\mathscr{T}$, *we have for* $\lambda \geq 2\lambda_0$,

$$\|\mathbf{X}(\hat{\beta} - \beta^0)\|_2^2/n + \lambda\|\hat{\beta} - \beta^0\|_1 \leq 4\lambda^2 s_0/\phi_0^2.$$

The theorem combines two results. Firstly, it shows the bound

$$\|\mathbf{X}(\hat{\beta} - \beta^0)\|_2^2/n \leq 4\lambda^2 s_0/\phi_0^2$$

for the prediction error. And secondly, it gives the bound

$$\|\hat{\beta} - \beta^0\|_1 \leq 4\lambda s_0/\phi_0^2$$

for the ℓ_1-error. (Both bounds hold on $\mathscr{T}$.) In Section 6.8, we present the implied bounds on the ℓ_q-error, $1 < q \leq 2$, assuming there a stronger compatibility condition.

Proof of Theorem 6.1. We continue with Lemma 6.3. This gives

$$2\|\mathbf{X}(\hat{\beta} - \beta^0)\|_2^2/n + \lambda\|\hat{\beta} - \beta^0\|_1$$

$$= 2\|\mathbf{X}(\hat{\beta} - \beta^0)\|_2^2/n + \lambda\|\hat{\beta}_{S_0} - \beta^0_{S_0}\|_1 + \lambda\|\hat{\beta}_{S_0^c}\|_1$$

$$\leq 4\lambda\sqrt{s_0}\|\mathbf{X}(\hat{\beta} - \beta^0)\|_2/(\sqrt{n}\phi_0)$$

$$\leq \|\mathbf{X}(\hat{\beta} - \beta^0)\|_2^2/n + 4\lambda^2 s_0/\phi_0^2,$$

where we inserted the compatibility condition in the first inequality, and used

$$4uv \leq u^2 + 4v^2$$

in the second inequality. □

Combining Theorem 6.1 with the probabilistic statement of Lemma 6.2 to handle the set $\mathscr{T}$, gives the following corollary.

Corollary 6.2. *Assume that $\hat{\sigma}_j^2 = 1$ for all j and that the compatibility condition holds for S_0, with $\hat{\Sigma}$ normalized in this way. For some $t > 0$, let the regularization parameter be*

$$\lambda := 4\hat{\sigma}\sqrt{\frac{t^2 + 2\log p}{n}},$$

where $\hat{\sigma}^2$ is an estimator of the noise variance σ^2. Then with probability at least $1 - \alpha$, where

$$\alpha := 2\exp[-t^2/2] + \mathbf{P}(\hat{\sigma} \leq \sigma),$$

we have

$$\|\mathbf{X}(\hat{\beta} - \beta^0)\|_2^2/n + \lambda\|\hat{\beta} - \beta^0\|_1 \leq 4\lambda^2 s_0/\phi_0^2.$$

We remark that it is straightforward to extend the situation to the case of independent, centered, non-Gaussian errors. One may use for example the moment inequality

$$\mathbb{E}\left(\max_{1 \leq j \leq p}|\varepsilon^T\mathbf{X}^{(j)}|\right)^2 \leq 8\log(2p)\sum_{i=1}^n\left(\max_{1 \leq j \leq p}|X_i^{(j)}|\right)^2\mathbb{E}\varepsilon_i^2 \qquad (6.5)$$

(this is up to constants Nemirovski's inequality, see Dümbgen et al. (2010), and see also Example 14.3), i.e., it is possible to prove similar results, assuming only second moments of the errors, and (say) bounded co-variables $X^{(j)}$ (Problem 6.2).

The normalization with $\hat{\sigma}_j^2 = 1$ for all j is quite natural, and often used in practice. Tacitly assuming the covariables are centered, $\hat{\sigma}_j^2$ is simply the (sample) variance of $\mathbf{X}^{(j)}$, and the normalization puts all covariables on the same scale. It is equivalent to defining $\hat{\beta}$ as

$$\hat{\beta} := \arg\min_{\beta} \left\{ \frac{\|\mathbf{Y} - \mathbf{X}\beta\|_2^2}{n} + \lambda \sum_{j=1}^{p} \hat{\sigma}_j |\beta_j| \right\},$$

i.e., to taking the weighted ℓ_1-penalty $\text{pen}(\beta) := \sum_{j=1}^{p} \hat{\sigma}_j |\beta_j|$. Of course, with all ones on the diagonal, the matrix $\hat{\Sigma}$ is rather a correlation matrix. In some situations (in particular with random $\mathbf{X}^{(j)}$), the distinction will be stressed by writing $\hat{R} := \text{diag}(\hat{\Sigma})^{-1/2} \hat{\Sigma} \, \text{diag}(\hat{\Sigma})^{-1/2}$ for the normalized $\hat{\Sigma}$. Recall that the normalization is only used in Lemma 6.2. It leads to good bounds for the probability of the set $\mathcal{T}$. Such bounds of course also hold under more general conditions.

Note finally that

$$\hat{\sigma}_j^2 |\beta_j|^2 = \|\mathbf{X}^{(j)} \beta_j\|_2^2 / n,$$

in other words, the weighted ℓ_1-norm is the ℓ_1-norm of the vector of ℓ_2-norms of the individual linear functions $\mathbf{X}^{(j)} \beta_j$. Such a viewpoint will help us to generalize to the group Lasso penalty (see Subsection 4.5.1 and Section 8.3). Section 6.9 provides a discussion of more general weights.

6.2.3 Linear approximation of the truth

We extend the theory to the case where

$$\mathbb{E}\mathbf{Y} := \mathbf{f}^0$$

is possibly not a sparse linear combination of the variables $\mathbf{X}^{(j)}$. One then still has the Basic Inequality of Lemma 6.1, albeit with an additional term representing the approximation error: for any vector β^*,

$$\|\mathbf{X}\hat{\beta} - \mathbf{f}^0\|_2^2 / n + \lambda \|\hat{\beta}\|_1 \le 2\varepsilon^T \mathbf{X}(\hat{\beta} - \beta^*)/n + \lambda \|\beta^*\|_1 + \|\mathbf{X}\beta^* - \mathbf{f}^0\|_2^2 / n. \quad (6.6)$$

We can carry out our chain of inequalities as before, taking along the additional term $\|\mathbf{X}\beta^* - \mathbf{f}^0\|_2^2 / n$ at every step. One then finds on $\mathcal{T}$, and assuming $\lambda \ge 4\lambda_0$ (instead of $\lambda \ge 2\lambda_0$ as we did in the previous section),

$$4\|\mathbf{X}\hat{\beta} - \mathbf{f}^0\|_2^2 / n + 3\lambda \|\hat{\beta}_{S_*^c}\|_1 \le 5\lambda \|\hat{\beta}_{S_*} - \beta_{S_*}^*\|_1 + 4\|\mathbf{X}\beta^* - \mathbf{f}^0\|_2^2 / n,$$

where $S_* := \{j : \beta_j^* \neq 0\}$. (Note that S_* depends on β^*, and that β^* is as yet arbitrary.) Due to the approximation error term $\|\mathbf{X}\beta^* - \mathbf{f}^0\|_2$, we cannot directly conclude anymore that $\|\hat{\beta}_{S_*^c}\|_1 \leq 3\|\hat{\beta}_{S_*} - \beta_{S_*}^*\|_1$. To handle this, we take the following approach. One of the two expressions on the right-hand side is the larger one. Either (**Case i**))

$$\lambda\|\hat{\beta}_{S_*} - \beta_{S_*}^*\|_1 \geq \|\mathbf{X}\beta^* - \mathbf{f}^0\|_2^2/n,$$

or (**Case ii**))

$$\lambda\|\hat{\beta}_{S_*} - \beta_{S_*}^*\|_1 < \|\mathbf{X}\beta^* - \mathbf{f}^0\|_2^2/n.$$

So it must hold that either (**Case i**))

$$4\|\mathbf{X}\hat{\beta} - \mathbf{f}^0\|_2^2/n + 3\lambda\|\hat{\beta}_{S_*^c}\|_1 \leq 9\lambda\|\hat{\beta}_{S_*} - \beta_{S_*}^*\|_1,$$

or (**Case ii**))

$$4\|\mathbf{X}\hat{\beta} - \mathbf{f}^0\|_2^2/n + 3\lambda\|\hat{\beta}_{S_*^c}\|_1 \leq 9\|\mathbf{X}\beta^* - \mathbf{f}^0\|_2^2/n,$$

(or both). In the first case, we again find $\|\hat{\beta}_{S_*^c}\|_1 \leq 3\|\hat{\beta}_{S_*} - \beta_{S_*}^*\|_1$, and we can go on as before, invoking the compatibility condition for general sets (see below). In the second case, we consider the argument done, as the obtained inequality is already quite good if β^* well-chosen (see below).

The compatibility condition considers sets $\{j : \beta_j^* \neq 0\}$ for general β^*, i.e., it considers general sets S.

Compatibility condition (for general sets) *We say that the compatibility condition holds for the set S, if for some constant $\phi(S) > 0$, and for all β, with $\|\beta_{S^c}\|_1 \leq 3\|\beta_S\|_1$, one has*

$$\|\beta_S\|_1^2 \leq \left(\beta^T \hat{\Sigma} \beta\right) |S|/\phi^2(S).$$

Let $\mathscr{S}$ be some given collection of index sets S for which the compatibility condition holds.

Definition of the oracle *We define the oracle β^* as*

$$\beta^* = \arg\min_{\beta:\, S_\beta \in \mathscr{S}} \left\{ \|\mathbf{X}\beta - \mathbf{f}^0\|_2^2/n + \frac{4\lambda^2 s_\beta}{\phi^2(S_\beta)} \right\}, \tag{6.7}$$

where $S_\beta := \{j : \beta_j \neq 0\}$, and where $s_\beta := |S_\beta|$ denotes the cardinality of S_β.

The factor 4 in the right hand side of (6.7) comes primarily from our choice $\lambda \geq 4\lambda_0$.

Note that we may define the oracle as minimizing over **all** β, with the convention $\phi(S) = 0$ if the compatibility condition does not hold for S. On the other hand, if $f^0 = f_{\beta^0}$ is linear, we may want to take $\mathscr{S} = \{S_0\}$.

We use the shorthand notation $S_* := S_{\beta^*}$, $s_* = |S_*|$, and $\phi_* := \phi(S_*)$. In other words, given a set S, we first look for the best approximation of $\mathbf{f}^0$ using only non-zero

coefficients inside the set S:

$$b^S := \arg\min_{\beta=\beta_S} \|\mathbf{X}\beta - \mathbf{f}^0\|_2.$$

We write $f_S := \mathbf{X}b^S$. Thus f_S is the projection of $\mathbf{f}^0$ on the space spanned by the variables $\{X^{(j)}\}_{j\in S}$. We then minimize over all $S \in \mathscr{S}$, with an ℓ_0-penalty, i.e., a penalty on the size of S, penalizing as well a small compatibility constant:

$$S_* := \arg\min_{S\in\mathscr{S}} \left\{ \|f_S - \mathbf{f}^0\|_2^2/n + \frac{4\lambda^2|S|}{\phi^2(S)} \right\}.$$

The oracle is $\beta^* = b^{S_*}$.

Theorem 6.2. *Assume that $\hat{\sigma}_j^2 = 1$ for all j and that the compatibility condition holds for all $S \in \mathscr{S}$, with $\hat{\Sigma}$ normalized in this way. For some $t > 0$, let the regularization parameter be*

$$\lambda = 8\hat{\sigma}\sqrt{\frac{t^2 + 2\log p}{n}},$$

where $\hat{\sigma}^2$ is some estimator of the noise variance σ^2. Then with probability at least $1 - \alpha$, where

$$\alpha := 2\exp[-t^2/2] + \mathbf{P}(\hat{\sigma} \le \sigma),$$

we have

$$2\|\mathbf{X}\hat{\beta} - \mathbf{f}^0\|_2^2/n + \lambda\|\hat{\beta} - \beta^*\|_1 \le 6\|\mathbf{X}\beta^* - \mathbf{f}^0\|_2^2/n + \frac{24\lambda^2 s_*}{\phi_*^2}. \qquad (6.8)$$

We emphasize that in our definition of the oracle β^*, one is free to choose the collection $\mathscr{S}$ over which is minimized (assuming the compatibility condition for all $S \in \mathscr{S}$, or even some $S \in \mathscr{S}$). In particular, when $\mathbf{f}^0 = \mathbf{X}\beta^0$ and $\mathscr{S} = \{S_0\}$, with $S_0 := S_{\beta_0}$, gives $\beta^* = \beta_0$ and hence, we obtain

$$2\|\mathbf{X}\hat{\beta} - \mathbf{f}^0\|_2^2/n + \lambda\|\hat{\beta} - \beta^0\|_1 \le \frac{24\lambda^2 s_0}{\phi_0^2},$$

where $\phi_0 = \phi(S_0)$ and $s_0 = |S_0|$. For a larger collection of sets $\mathscr{S}$, the right-hand side of (6.8) can be much smaller, i.e., then one has a better bound for the prediction error $\|\mathbf{X}\hat{\beta} - \mathbf{f}^0\|_2^2/n$. In addition, one now has the bound for the ℓ_1-error $\|\hat{\beta} - \beta^*\|_1$. One may object that the oracle β^* is difficult to interpret, so that $\|\hat{\beta} - \beta^*\|_1$ is a less interesting quantity to look at. However, one can use of course the triangle inequality

$$\|\hat{\beta} - \beta^0\|_1 \le \|\hat{\beta} - \beta^*\|_1 + \|\beta^* - \beta^0\|_1.$$

The term $\|\beta^* - \beta^0\|_1$ can easily be bounded by the right-hand side of (6.8) plus $\lambda\|\beta_{S_0\setminus S_*}^0\|_1$ (see Problem 6.4). Finally, by the triangle inequality, (6.8) implies

$$2\|\mathbf{X}\hat{\beta} - \mathbf{f}^0\|_2^2/n + \lambda\|\hat{\beta} - \beta^0\|_1 \leq 6\|\mathbf{X}\beta^* - \mathbf{f}^0\|_2^2/n + \frac{24\lambda^2 s_*}{\phi_*^2} + \lambda\|\beta^* - \beta^0\|_1.$$

In view of the latter it would also make sense to reformulated the oracle to

$$\beta^* = \arg\min_{\beta:\, S_\beta \in \mathscr{S}}\left\{\|\mathbf{X}\beta - \mathbf{f}^0\|_2^2/n + \frac{4\lambda^2 s_\beta}{\phi^2(S_\beta)} + \lambda\|\beta - \beta^0\|_1/6\right\}.$$

In fact, the particular definition of the oracle plays only a limited role. The vector β^* is chosen to optimize the result in a certain way. Throughout, our statements may also be seen as holding for any fixed β^*. We note however that when considering general loss in the next sections, we will need that the chosen β^* has a small enough approximation error, as this will be used to localize the problem.

Proof of Theorem 6.2.

Case i) On $\mathscr{T}$, whenever

$$\lambda\|\hat{\beta}_{S_*} - \beta^*_{S_*}\|_1 \geq \|\mathbf{X}\beta^* - \mathbf{f}^0\|_2^2/n,$$

we have

$$4\|\mathbf{X}\hat{\beta} - \mathbf{f}^0\|_2^2/n + 3\lambda\|\hat{\beta}_{S_*^c}\|_1 \leq 9\lambda\|\hat{\beta}_{S_*} - \beta^*_{S_*}\|_1,$$

so then

$$4\|\mathbf{X}\hat{\beta} - \mathbf{f}^0\|_2^2/n + 3\lambda\|\hat{\beta} - \beta^*\|_1 \leq 12\lambda\|\hat{\beta}_{S_*} - \beta^*_{S_*}\|_1$$

$$\leq \frac{12\lambda\sqrt{s_*}\|\mathbf{X}(\hat{\beta} - \beta^*)\|_2}{\sqrt{n}\phi_*} \leq \frac{24\lambda^2 s_*}{\phi_*^2} + 2\|\mathbf{X}\hat{\beta} - \mathbf{f}^0\|_2^2/n + 6\|\mathbf{X}\beta^* - \mathbf{f}^0\|_2^2/n.$$

Here, we use $12uv \leq 18u^2 + 2v^2$, and $12uv \leq 6u^2 + 6v^2$. Hence

$$2\|\mathbf{X}\hat{\beta} - \mathbf{f}^0\|_2^2/n + 3\lambda\|\hat{\beta} - \beta^*\|_1 \leq 6\|\mathbf{X}\beta^* - \mathbf{f}^0\|_2^2/n + \frac{24\lambda^2 s_*}{\phi_*^2}.$$

Case ii) If on the other hand, on $\mathscr{T}$,

$$\lambda\|\hat{\beta}_{S_*} - \beta^*_{S_*}\|_1 < \|\mathbf{X}\beta^* - \mathbf{f}^0\|_2^2/n,$$

we get

$$4\|\mathbf{X}\hat{\beta} - \mathbf{f}^0\|_2^2/n + 3\lambda\|\hat{\beta}_{S_*^c}\|_1 \leq 9\|\mathbf{X}\beta^* - \mathbf{f}^0\|_2^2/n,$$

and hence

$$4\|\mathbf{X}\hat{\beta} - \mathbf{f}^0\|_2^2/n + 3\lambda\|\hat{\beta} - \beta^*\|_1 \leq 12\|\mathbf{X}\beta^* - \mathbf{f}^0\|_2^2/n.$$

$\square$

6.2.4 A further refinement: handling smallish coefficients

The oracle, as defined in the previous section, trades off the approximation error with an ℓ_0-penalty including a restricted eigenvalue. We now refine this to a trade-off including both ℓ_0- and ℓ_1-penalties. Namely, for each S, we define

$$S^{\text{sub}} := \arg\min_{S^\circ \subset S}\left\{ \frac{3\lambda^2|S^\circ|}{\phi^2(S^\circ)} + \lambda\,\|(b^S)_{S\setminus S^\circ}\|_1 \right\}.$$

This means that the smaller coefficients b_j^S go into the ℓ_1-penalty, and the larger ones in the ℓ_0-penalty. Putting fewer coefficients into the ℓ_0-penalty will generally increase (and hence improve) the value for the restricted eigenvalue $\phi^2(S^\circ)$. Indeed, for $S^\circ \subset S$, one has

$$\phi^2(S^\circ)/|S^\circ| \ge \phi^2(S)/|S|$$

(see Lemma 6.19).

Definition of the oracle *Let*

$$S_* := \arg\min_{S\in\mathscr{S}}\left\{ 3\|\mathbf{f}_S - \mathbf{f}^0\|_2^2/n \right.$$

$$\left. + \frac{12\lambda^2|S^{\text{sub}}|}{\phi^2(S^{\text{sub}})} + 4\lambda\,\|(b^S)_{S\setminus S^{\text{sub}}}\|_1 \right\}.$$

The oracle is $\beta^* := b^{S_*}$.

We use the short-hand notation $s_*^{\text{sub}} := |S_*^{\text{sub}}|$ and $\phi_*^{\text{sub}} := \phi(S_*^{\text{sub}})$.

Theorem 6.3. *On the set*

$$\mathscr{T} := \left\{ \max_{1\le j\le p} 2|\varepsilon^T\mathbf{X}^{(j)}|/n \le \lambda_0 \right\},$$

we have, for $\lambda \ge 4\lambda_0$,

$$2\|\mathbf{X}\hat{\beta} - \mathbf{f}^0\|_2^2/n + \lambda\|\hat{\beta} - \beta^*\|_1$$

$$\le 6\|\mathbf{X}\beta^* - \mathbf{f}^0\|_2^2/n + \frac{24\lambda^2 s_*^{\text{sub}}}{(\phi_*^{\text{sub}})^2} + 8\lambda\|\beta^*_{S_*\setminus S_*^{\text{sub}}}\|_1.$$

By replacing (for some S) S^{sub} by the (sub-optimal) choice $S^\circ := \emptyset$, Theorem 6.3 implies the bound

$$2\|\mathbf{X}\hat{\beta} - \mathbf{f}^0\|_2^2/n + \lambda\|\hat{\beta} - b^S\|_1 \le 6\|\mathbf{X}b^S - \mathbf{f}^0\|_2^2/n + 8\lambda\|b^S\|_1 \ \forall\, S,$$

on $\mathscr{T}$. On the other hand, replacing S^{sub} by the (sup-optimal) choice $S^\circ := S$ gives

$$2\|\mathbf{X}\hat{\beta} - \mathbf{f}^0\|_2^2/n + \lambda\|\hat{\beta} - b^S\|_1 \le 6\|\mathbf{X}b^S - \mathbf{f}^0\|_2^2/n + \frac{24\lambda^2|S|}{\phi^2(S)} \ \forall \ S.$$

In other words, Theorem 6.3 combines a consistency result (compare with Corollary 6.1) with an oracle result (Theorem 6.2).

Proof of Theorem 6.3. Throughout, we assume we are on $\mathcal{T}$. By the Basic Inequality

$$\|\mathbf{X}\hat{\beta} - \mathbf{f}^0\|_2^2/n + \lambda\|\hat{\beta}\|_1 \le \lambda_0\|\hat{\beta} - \beta^*\|_1 + \lambda\|\beta^*\|_1 + \|\mathbf{X}\beta^* - \mathbf{f}^0\|_2^2/n.$$

So

$$\|\mathbf{X}\hat{\beta} - \mathbf{f}^0\|_2^2/n + \lambda\|\hat{\beta}_{S_*^c}\|_1 + \lambda\|\hat{\beta}_{S_*^{\text{sub}}}\|_1 + \lambda\|\hat{\beta}_{S_* \setminus S_*^{\text{sub}}}\|_1$$

$$\le \lambda_0\|\hat{\beta}_{S_*^c}\|_1 + \lambda_0\|\hat{\beta}_{S_*^{\text{sub}}} - \beta^*_{S_*^{\text{sub}}}\|_1 + \lambda_0\|\hat{\beta}_{S_* \setminus S_*^{\text{sub}}}\|_1$$

$$+\lambda\|\beta^*_{S_*^{\text{sub}}}\|_1 + (\lambda + \lambda_0)\|\beta^*_{S_* \setminus S_*^{\text{sub}}}\|_1 + \|\mathbf{X}\beta^* - \mathbf{f}^0\|_2^2/n.$$

Take the term $\lambda\|\hat{\beta}_{S_*^{\text{sub}}}\|_1$ to the right hand side, and use the triangle inequality

$$\lambda\|\beta^*_{S_*^{\text{sub}}}\|_1 - \lambda\|\hat{\beta}_{S_*^{\text{sub}}}\|_1 \le \lambda\|\hat{\beta}_{S_*^{\text{sub}}} - \beta_{S_*^{\text{sub}}}\|_1.$$

Moreover, take both terms $\lambda_0\|\hat{\beta}_{S_*^c}\|_1$ and $\lambda_0\|\hat{\beta}_{S_* \setminus S_*^{\text{sub}}}\|_1$ to the left hand side. This gives

$$\|\mathbf{X}\hat{\beta} - \mathbf{f}^0\|_2^2/n + (\lambda - \lambda_0)\|\hat{\beta}_{S_*^c}\|_1 + (\lambda - \lambda_0)\|\hat{\beta}_{S_* \setminus S_*^{\text{sub}}}\|_1$$

$$\le (\lambda + \lambda_0)\|\hat{\beta}_{S_*^{\text{sub}}} - \beta^*_{S_*^{\text{sub}}}\|_1 + (\lambda + \lambda_0)\|\beta^*_{S_* \setminus S_*^{\text{sub}}}\|_1 + \|\mathbf{X}\beta^* - \mathbf{f}^0\|_2^2/n.$$

Now, on the left hand side, use the inequality

$$\|\hat{\beta}_{S_* \setminus S_*^{\text{sub}}} - \beta^*_{S_* \setminus S_*^{\text{sub}}}\|_1 \le \|\hat{\beta}_{S_* \setminus S_*^{\text{sub}}}\|_1 + \|\beta^*_{S_* \setminus S_*^{\text{sub}}}\|_1.$$

We then obtain

$$\|\mathbf{X}\hat{\beta} - \mathbf{f}^0\|_2^2/n + (\lambda - \lambda_0)\|\hat{\beta}_{(S_*^{\text{sub}})^c} - \beta^*_{(S_*^{\text{sub}})^c}\|_1$$

$$\le (\lambda + \lambda_0)\|\hat{\beta}_{S_*^{\text{sub}}} - \beta^*_{S_*^{\text{sub}}}\|_1 + 2\lambda\|\beta^*_{S_* \setminus S_*^{\text{sub}}}\|_1 + \|\mathbf{X}\beta^* - \mathbf{f}^0\|_2^2/n.$$

Next, we use our assumption $\lambda \ge 4\lambda_0$. We arrive at

$$4\|\mathbf{X}\hat{\beta} - \mathbf{f}^0\|_2^2/n + 3\lambda\|\hat{\beta}_{(S_*^{\text{sub}})^c} - \beta^*_{(S_*^{\text{sub}})^c}\|_1$$

$$\le 5\lambda\|\hat{\beta}_{S_*^{\text{sub}}} - \beta^*_{S_*^{\text{sub}}}\|_1 + 8\lambda\|\beta^*_{S_* \setminus S_*^{\text{sub}}}\|_1 + 4\|\mathbf{X}\beta^* - \mathbf{f}^0\|_2^2/n.$$

We now invoke the "Case i)/Case ii)" argument. One of the two expressions in the right-hand side is the larger one. Either (**Case i)**)

$$\lambda \|\hat{\beta}_{S_*^{\mathrm{sub}}} - \beta_{S_*^{\mathrm{sub}}}^*\|_1 \geq 2\lambda \|\beta_{S_* \setminus S_*^{\mathrm{sub}}}^*\|_1 + \|\mathbf{X}\beta^* - \mathbf{f}^0\|_2^2/n,$$

or (**Case ii**)

$$\lambda \|\hat{\beta}_{S_*^{\mathrm{sub}}} - \beta_{S_*^{\mathrm{sub}}}^*\|_1 < 2\lambda \|\beta_{S_* \setminus S_*^{\mathrm{sub}}}^*\|_1 + \|\mathbf{X}\beta^* - \mathbf{f}^0\|_2^2/n.$$

So it must hold that either (**Case i**))

$$4\|\mathbf{X}\hat{\beta} - \mathbf{f}^0\|_2^2/n + 3\lambda \|\hat{\beta}_{(S_*^{\mathrm{sub}})^c} - \beta_{(S_*^{\mathrm{sub}})^c}^*\|_1 \leq 9\lambda \|\hat{\beta}_{S_*^{\mathrm{sub}}} - \beta_{S_*^{\mathrm{sub}}}^*\|_1,$$

or (**Case ii**))

$$4\|\mathbf{X}\hat{\beta} - \mathbf{f}^0\|_2^2/n + 3\lambda \|\hat{\beta}_{(S_*^{\mathrm{sub}})^c} - \beta_{(S_*^{\mathrm{sub}})^c}^*\|_1 \leq 10\lambda \|\beta_{S_* \setminus S_*^{\mathrm{sub}}}^*\|_1 + 9\|\mathbf{X}\beta^* - \mathbf{f}^0\|_2^2/n,$$

(or both). In **Case i**), we can use the same argument as its version in the proof of Theorem 6.2. In **Case ii**), we have

$$4\|\mathbf{X}\hat{\beta} - \mathbf{f}^0\|_2^2/n + 3\lambda \|\hat{\beta} - \beta^*\|_1 \leq 16\lambda \|\beta_{S_* \setminus S_*^{\mathrm{sub}}}^*\|_1 + 12\|\mathbf{X}\beta^* - \mathbf{f}^0\|_2^2/n,$$

$$\square$$

6.3 The setup for general convex loss

The results in this section are based on Loubes and van de Geer (2002), van de Geer (2007) and van de Geer (2008). We extend the framework for squared error loss with fixed design to the following scenario. Consider data $\{Z_i\}_{i=1}^n$, with (for $i = 1, \ldots, n$) Z_i in some space $\mathcal{Z}$. Let $\mathbf{F}$ be a (rich) parameter space, and, for each $f \in \mathbf{F}$, $\rho_f : \mathcal{Z} \to \mathbb{R}$ be a loss function. We assume that $\mathbf{F} := (\mathbf{F}, \|\cdot\|)$ is a normed real vector space, and that the map $f \mapsto \rho_f(z)$ is convex for all $z \in \mathcal{Z}$. For example, in a regression setup, the data are (for $i = 1, \ldots, n$) $Z_i = (X_i, Y_i)$, with $X_i \in \mathcal{X}$ and $Y_i \in \mathcal{Y} \subset \mathbb{R}$, and f is a regression function. Examples are then quadratic loss, where

$$\rho_f(\cdot, y) = (y - f(\cdot))^2,$$

or logistic loss, where

$$\rho_f(\cdot, y) = -yf(\cdot) + \log(1 + \exp[f(\cdot)]),$$

etc. See Chapter 3 for more examples. The regression f may also be vector-valued (for example in multi-category classification problems, or in the case of normally distributed responses with both the mean and the log-variance depending on a large number of co-variables). The vector-valued situation however will typically have

separate penalties for each component. This will be examined explicitly in Chapter 9, for possibly non-convex loss.

We denote, for a function $\rho : \mathscr{Z} \to \mathbb{R}$, the empirical average by

$$P_n \rho := \sum_{i=1}^{n} \rho(Z_i)/n,$$

and the theoretical mean by

$$P\rho := \sum_{i=1}^{n} \mathbb{E}\rho(Z_i)/n.$$

Thus, P_n is the empirical measure that puts mass $1/n$ at each observation Z_i ($i = 1, \ldots, n$), and P is the "theoretical" measure. The empirical risk, and theoretical risk, at f, is defined as $P_n \rho_f$, and $P\rho_f$, respectively. We furthermore define the *target* as the minimizer of the theoretical risk

$$f^0 := \arg\min_{f \in \mathbf{F}} P\rho_f.$$

We assume for simplicity that the minimum is attained (and that it is unique for the $\|\cdot\|$-norm). The target f^0 plays the rule of the "truth", as in the previous section with squared error loss.

For $f \in \mathbf{F}$, the excess risk is

$$\mathscr{E}(f) := P(\rho_f - \rho_{f^0}).$$

Note that by definition, $\mathscr{E}(f) \geq 0$ for all $f \in \mathbf{F}$.

Consider a linear subspace $\mathscr{F} := \{f_\beta : \beta \in \mathbf{R}^p\} \subset \mathbf{F}$, where the map $\beta \mapsto f_\beta$ is linear. The collection $\mathscr{F}$ will be the model class, over which we perform empirical risk minimization. When the loss ρ_f is a minus-log-likelihood, the situation corresponds to a *generalized linear model* (*GLM*) as described in Chapter 3. The class of linear functions $\mathscr{F}$ is generally *strictly* smaller than $\mathbf{F}$. In other words, the model may be misspecified. Formally, one is allowed to take $\mathbf{F} = \mathscr{F}$, but this can affect our *margin conditions* (see Section 6.4 for a definition of the margin condition).

Let us denote the the best linear approximation of the target f^0 by

$$f^0_{\text{GLM}} := f_{\beta^0_{\text{GLM}}}, \quad \beta^0_{\text{GLM}} := \arg\min_{\beta} P\rho_{f_\beta}.$$

Throughout this chapter, it is tacitly assumed that the approximation error $\mathscr{E}(f^0_{\text{GLM}})$ is "small". Again formally, this can be achieved by taking $\mathscr{F} = \mathbf{F}$. Thus f^0 and f^0_{GLM} may be different, but very close in terms of the excess risk, and typically, f^0 is the target of interest with good *margin behavior*. In the case of squared error loss with fixed design, one may without loss of generality assume that $f^0 = f^0_{\text{GLM}}$ and thus

$\mathscr{E}(f^0_{\text{GLM}}) = 0$ (as, by Pythagoras' Theorem, the margin remains quadratic near the projection of f^0 on $\mathscr{F}$).

The Lasso is[3]

$$\hat{\beta} = \arg\min_{\beta} \left\{ P_n \rho_{f_\beta} + \lambda \|\beta\|_1 \right\}. \tag{6.9}$$

We write $\hat{f} = f_{\hat{\beta}}$. We will mainly examine the excess risk $\mathscr{E}(\hat{f})$ of the Lasso.

The program for the rest of this chapter was already sketched in the introduction. Let us briefly recall it here. In the next section (Section 6.4), we introduce the so-called margin condition. This condition describes the sensitivity of $P\rho_f$ to changes in $f \in (\mathbf{F}, \|\cdot\|)$, locally near f^0. In Section 6.5, we consider the case where p is relatively small, as compared to n, so that no regularization is needed. The result serves as a benchmark for the case where p is large, and where the above ℓ_1-regularization penalty is invoked. After presenting a general consistency result in Section 6.6, we investigate more refined oracle inequalities, imposing further conditions. In Section 6.7, we show that the Lasso behaves as if it knew which variables are relevant for a linear approximation of the target f^0, assuming a *compatibility condition*. Section 6.8 examines the ℓ_q-error $\|\hat{\beta} - \beta^*\|_q$, for $1 \leq q \leq 2$.

In (6.9), the ℓ_1-penalty weights all coefficients equally. In practice, certain terms (e.g., the constant term) will not be penalized, and a normalized version of the ℓ_1-norm will be used, with e.g. more weight assigned to variables with large sample variance. In essence, (possibly random) weights that stay within a reasonable range do not alter the main points in the theory. We therefore initially consider the non-weighted ℓ_1-penalty to clarify these main points. The weighted version is studied in Section 6.9. Here, the weights are assumed to be within certain bounds, or to be zero. In the latter case, ideas of Section 6.5 and Section 6.7 are combined.

In Sections 6.10 and 6.11, we return to squared error loss with fixed design. Section 6.10 studies an adaptively weighted penalty, where the weights may be based on an initial estimator (the adaptive Lasso). Section 6.11, replaces the ℓ_1-penalty by an ℓ_r-penalty, with $0 \leq r \leq 1$. These two sections still only consider prediction error. The results are to be understood as complementing the results on selection of variables in Chapter 7.

The compatibility condition is also of interest when comparing the properties of the sample covariance matrix with those of the theoretical one. In fact, one may generally replace the norm used in the compatibility condition by an approximation (see Section 6.12, Corollary 6.8).

Finally, Section 6.13 examines under what circumstances the compatibility condition is met.

---

[3] More generally, the minimization may be over $\beta \in \mathscr{B}$ where $\mathscr{B}$ is a **convex** subset of $\mathbb{R}^p$. All β's considered (in particular the "oracle") are then restricted to lie in $\mathscr{B}$. (For the linear model with squared error loss, the subset $\mathscr{B}$ need not be convex in order that the theory goes through.)

Throughout, the stochastic part of the problem (involving the empirical process: see below) is set aside. It can be handled using the theory in Chapter 14. In the current chapter, we simply restrict ourselves to a set $\mathscr{T}$ (which may be different in different sections) where the empirical process (defined in the particular context of that section) behaves "well". This allows us to proceed with purely deterministic, and relatively straightforward, arguments.

We now give some details concerning the random part. The *empirical process* is defined as

$$\left\{ v_n(\beta) := (P_n - P)\rho_{f_\beta} : \ \beta \in \mathbb{R}^p \right\}. \tag{6.10}$$

By the definition of $\hat{\beta}$, we have for any β^*,

$$P_n \rho_{f_{\hat{\beta}}} + \lambda \|\hat{\beta}\|_1 \leq P_n \rho_{f_{\beta^*}} + \lambda \|\beta^*\|_1.$$

We can rewrite this, in the same spirit as the Basic Inequality for the case of squared error loss (see (6.6) in Subsection 6.2.3). This gives

Lemma 6.4. (Basic inequality) *For any* β^*, *it holds that*

$$\mathscr{E}(f_{\hat{\beta}}) + \lambda \|\hat{\beta}\|_1 \leq -\left[v_n(\hat{\beta}) - v_n(\beta^*) \right] + \lambda \|\beta^*\|_1 + \mathscr{E}(f_{\beta^*}).$$

The proof is left as exercise (Problem 6.5).

The Basic Inequality implies that again, we need to control the behavior of the *increments* of the empirical process $[v_n(\beta) - v_n(\beta^*)]$ in terms of $\|\beta - \beta^*\|_1$.

Let us briefly discuss the least squares example, to clarify the various concepts.

Example 6.1. **Least squares with fixed design** Let $Z_i := (X_i, Y_i)$, $X_i \in \mathscr{X}$ a fixed co-variable, $Y_i \in \mathbb{R}$ a response variable, and

$$Y_i = f^0(X_i) + \varepsilon_i, \ i = 1, \ldots, n,$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. $\mathscr{N}(0, \sigma^2)$-distributed (say). The target is now the regression function $f^0 : \mathscr{X} \to \mathbb{R}$, which we considered as an n-dimensional vector $\mathbf{f}^0 := (f^0(X_1), \ldots, f^0(X_n))^T$ in the previous section. The linear space is

$$\mathscr{F} = \{ f_\beta(\cdot) = \sum_{j=1}^p \beta_j \psi_j(\cdot) : \ \beta \in \mathbb{R}^p \}.$$

In the previous section, we assumed that $X_i \in \mathbb{R}^p$ and took $\psi_j(X_i) := X_i^{(j)}$, the j-th component of $X_i := (X_i^{(1)}, \ldots, X_i^{(p)})$. More generally, the space $\mathscr{X}$ can be arbitrary, and the functions ψ_j are given base functions (often called *feature mappings*, and the collection $\{\psi_j\}_{j=1}^p$ is often called a *dictionary*).

The squared error loss is
$$\rho_f(x,y) := (y - f(x))^2,$$

and the empirical risk is

$$P_n \rho_f = \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2 = \varepsilon^T \varepsilon / n - 2(\varepsilon, f - f^0)_n + \|f - f^0\|_n^2,$$

where we use the notation

$$(\varepsilon, f)_n := \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(X_i) = \varepsilon^T \mathbf{f} / n,$$

and

$$\|f\|_n^2 := \frac{1}{n} \sum_{i=1}^{n} f^2(X_i) = \|\mathbf{f}\|_2^2 / n,$$

with ε and $\mathbf{f}$ being, respectively, the vectors

$$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}, \ \mathbf{f} = \begin{pmatrix} f(X_1) \\ \vdots \\ f(X_n) \end{pmatrix}.$$

Thus, for f a function on $\mathscr{X}$, $\|f\|_n$ is its $L_2(Q_n)$-norm, with Q_n the empirical measure of the co-variables $\{X_i\}_{i=1}^{n}$. The theoretical risk is

$$P \rho_f = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(Y_i - f(X_i))^2 = \sigma^2 + \|f - f^0\|_n^2.$$

The excess risk is
$$\mathscr{E}(f) = P(\rho_f - \rho_{f^0}) = \|f - f^0\|_n^2. \tag{6.11}$$

Finally, the empirical process is

$$\nu_n(\beta) := (P_n - P)\rho_{f_\beta} = \varepsilon^T \varepsilon / n - \sigma^2 - 2 \sum_{j=1}^{p} \beta_j(\varepsilon, \psi_j)_n + 2(\varepsilon, f^0)_n.$$

Note thus that its increments are

$$\nu_n(\beta^*) - \nu_n(\beta) = 2 \sum_{j=1}^{p} (\beta_j - \beta_j^*)(\varepsilon, \psi_j)_n,$$

which corresponds for $\beta^* = \beta^0$ to the empirical process part (6.1).

Returning to general loss functions, we recall that we assume that $f \mapsto \rho_f$ is convex. This will be used in the Basic Inequality as follows. Let for some $0 < t < 1$, $\tilde{\beta} := t\hat{\beta} + (1-t)\beta^*$. Invoking that the ℓ_1-penalty is also convex, and that the map $\beta \mapsto f_\beta$

is linear, one easily checks that the Basic Inequality remains to hold, with $\hat{\beta}$ replaced by $\tilde{\beta}$. Choosing

$$t := \frac{M^*}{M^* + \|\hat{\beta} - \beta^*\|_1},$$

for some $M^* > 0$ gives

$$\|\tilde{\beta} - \beta^*\|_1 \leq M^*.$$

We apply this with a fixed value for M^*, which will depend on the behavior of the "oracle" β^*. It will be assumed that this value is sufficiently small. Then we have the Basic Inequality, locally, i.e., we are already within ℓ_1-distance M^* of β^*. As we will show, a consequence is that we need not control the empirical process $[\nu_n(\beta) - \nu_n(\beta^*)]$ globally, but only the local supremum

$$\mathbf{Z}_{M^*} := \sup_{\|\beta - \beta^*\|_1 \leq M^*} |\nu_n(\beta) - \nu_n(\beta^*)|.$$

6.4 The margin condition

We will make frequent use of the so-called *margin condition* (see below for its definition), which is assumed for a "neighborhood", denoted by $\mathbf{F}_{\text{local}}$, of the target f^0. This neighborhood is typically with respect to some distance which is stronger than the one induced by the norm $\|\cdot\|$ on $\mathbf{F}$. (In fact, we usually take an L_∞-neighborhood.) Some of the effort goes in proving that the estimator is indeed in this neighborhood. Here, we use arguments that rely on the assumed convexity of $f \mapsto \rho_f$. Thus, the convexity is actually used twice: it is exploited to show that the stochastic part of the problem needs only to be studied locally, near the target, and to show that the (deterministic) margin condition is only needed locally.

The margin condition requires that in the neighborhood $\mathbf{F}_{\text{local}} \subset \mathbf{F}$ of f^0 the excess risk $\mathscr{E}$ is bounded from below by a strictly convex function. This is true in many particular cases for an L_∞-neighborhood $\mathbf{F}_{\text{local}} = \{\|f - f^0\|_\infty \leq \eta\}$ (for $\mathbf{F}$ being a class of functions).

Definition *We say that the* margin condition *holds with strictly convex function G, if for all $f \in \mathbf{F}_{\text{local}}$, we have*

$$\mathscr{E}(f) \geq G(\|f - f^0\|).$$

Indeed, in typical cases, the margin condition holds with quadratic function G, that is, $G(u) = cu^2$, $u \geq 0$, where c is a positive constant. For example, for the case of quadratic loss with fixed design, we may take $\|\cdot\| = \|\cdot\|_n$, and $G(u) = u^2$ (see

(6.11)). More generally, G is of the form $G(u) = G_1(u^2)$ with G_1 convex, which means that identification can be worse than the quadratic case.

Let us now consider further regression examples.

Example 6.2. Suppose that $\{Z_i\}_{i=1}^n := \{(X_i, Y_i)\}_{i=1}^n$ are i.i.d. copies of a random variable $Z := (X, Y)$. Let Q be the distribution of X, and let $(\mathbf{F}, \|\cdot\|) \subset L_2(Q)$ be a class of real-valued functions on $\mathcal{X}$. For $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, let the loss be of the form

$$\rho_f(x, y) := \rho(f(x), y), \ f \in \mathbf{F} .$$

Set

$$l(a, \cdot) = \mathbb{E}(\rho(a, Y)|X = \cdot), \ a \in \mathbb{R} .$$

We moreover write $l_f(\cdot) := l(f(\cdot), \cdot)$. As target we take the overall minimizer

$$f^0(\cdot) := \arg\min_{a \in \mathbb{R}} l(a, \cdot) .$$

Now, fix some $x \in \mathcal{X}$. If $l(a, x)$ has two derivatives (with respect to its first argument) near $a_0 := f^0(x)$, and if the second derivatives near a_0 are positive and bounded away from zero, then $l(a, x)$ behaves quadratically near its minimum, i.e., for some $\tau(x) > 0$, some constant $\eta > 0$, and all $|f(x) - f^0(x)| \leq \eta$,

$$l_f(x) - l_{f^0}(x) \geq \tau(x)|f(x) - f^0(x)|^2 .$$

We now assume this to hold for all x, i.e., that for some strictly positive function τ on $\mathcal{X}$

$$l_f(\cdot) - l_{f^0}(\cdot) \geq \tau(\cdot)|f(\cdot) - f^0(\cdot)|^2, \ \forall \ \|f - f^0\|_\infty \leq \eta. \tag{6.12}$$

Consider two cases.

Quadratic margin Assume that the function $\tau(\cdot)$ defined in (6.12) has $\tau(\cdot) \geq 1/K$ for some constant K. Then it follows that for all $\|f - f^0\|_\infty \leq \eta$,

$$\mathscr{E}(f) \geq c\|f - f^0\|^2 ,$$

with $c = 1/K$.

General margin Consider functions

$$H_1(v) \leq vQ\{x : \tau(x) < v\}, \ v > 0 ,$$

and

$$G_1(u) = \sup_{v > 0}\{uv - H_1(v)\}, \ u > 0.$$

Lemma 6.5. *Assume (6.12). For $\|f - f^0\|_\infty \leq \eta$, the inequality*

$$\mathscr{E}(f) \geq \eta^2 G_1\left(\|\rho_f - \rho_{f^0}\|^2/\eta^2\right)$$

holds.

Proof. By (6.12), we have, for any $v > 0$,

$$\mathscr{E}(f) = P(\rho_f - \rho_{f^0}) \geq \int \tau |f - f^0|^2 dQ$$

$$\geq \int_{\tau \geq v} \tau |f - f^0|^2 dQ \geq v \int_{\tau \geq v} |f - f^0|^2 dQ$$

$$= v \|f - f^0\|^2 - \eta^2 v Q\{x : \tau(x) < v\}$$

$$= \eta^2 \left(v \|f - f^0\|^2 / \eta^2 - H_1(v) \right).$$

We now maximize over v to obtain the required result. $\qquad\square$.

If $H_1(v) = 0$ for v sufficiently small, we are back in the case of locally quadratic margin behavior. More generally, the Tsybakov margin condition (see Tsybakov (2004)) assumes that one may take, for some $C_1 \geq 1$ and $\gamma \geq 0$,

$$H_1(v) = v(C_1 v)^{1/\gamma},$$

Then one has

$$G_1(u) = u^{1+\gamma}/C^{1+\gamma}$$

where

$$C = C_1^{\frac{1}{1+\gamma}} \gamma^{-\frac{\gamma}{1+\gamma}} (1 + \gamma).$$

Thus, the margin condition then holds with

$$G(u) = \eta^{-2\gamma} u^{2(1+\gamma)}/C^{1+\gamma}.$$

In the above example, we actually encountered the notion of a *convex conjugate* of a function. As this notion will play a crucial role in the oracle bounds as well, we present here its formal definition.

Definition *Let G be a strictly convex function on $[0, \infty)$, with $G(0) = 0$. The* convex conjugate *H of G is defined as*

$$H(v) = \sup_u \{uv - G(u)\}, \; v \geq 0.$$

When G is quadratic, its convex conjugate H is quadratic as well: for $G(u) = cu^2$ one has $H(v) = v^2/(4c)$. So we have

$$uv \leq cu^2 + v^2/(4c).$$

This inequality was used in the proof of Theorem 6.1 (with $c = 1/4$), as well as in the proof of Theorem 6.2 (twice: with $c = 3/2$ and $c = 1/2$). More generally, for $G(u) = cu^{2(1+\gamma)}$ ($\gamma \geq 0$), we have $H(v) = dv^{\frac{2(1+\gamma)}{1+2\gamma}}$, where $d = (1+2\gamma)(2(1+\gamma))^{-\frac{2(1+\gamma)}{1+2\gamma}} c^{-\frac{1}{1+2\gamma}}$. The function H will appear in the estimation error term. We will see that the larger γ, the larger the estimation error can be.

6.5 Generalized linear model without penalty

This section is relevant when $\mathscr{F}$ is low-dimensional, i.e., when $p < n$. The idea is to provide a benchmark for the case where $p \geq n$, and where one aims at mimicking an oracle that knows which variables are to be included in the model.

The estimator in the generalized linear model, that we will study in this section, is

$$\hat{b} := \arg\min_{\beta} P_n \rho_{f_\beta}. \tag{6.13}$$

We recall that the "best approximation" in the class of linear functions $\mathscr{F}$ is

$$f_{\text{GLM}}^0 := f_{\beta_{\text{GLM}}^0}, \quad \beta_{\text{GLM}}^0 := \arg\min_{\beta} P \rho_{f_\beta},$$

and that the excess risk $\mathscr{E}(f_{\text{GLM}}^0)$, can be regarded as the *approximation error* due to considering a (generalized) linear model. To assess the *estimation error* of $\hat{b}$, we assume the margin condition with strictly convex and increasing function G (with $G(0) = 0$). We moreover assume that G satisfies, for a positive constant η_0,

$$\alpha_0 := \inf_{0 < \delta \leq \eta_0} \frac{G(\delta/4)}{G(\delta)} > 0. \tag{6.14}$$

Condition 6.14 holds for power-functions, so in particular it is met when $G(u) = cu^2$ is quadratic. In the latter case

$$\alpha_0 = 1/16.$$

For all $\delta > 0$, we define the random variable

$$\bar{\mathbf{Z}}_\delta := \sup_{\|f_\beta - f_{\text{GLM}}^0\| \leq \delta} |v_n(\beta) - v_n(\beta_{\text{GLM}}^0)|/\sqrt{p},$$

where $v_n(\cdot)$ is the empirical process (see (6.10)). Consider for a fixed δ^0 the set

$$\mathscr{T} := \{\bar{\mathbf{Z}}_{\delta^0} \leq \bar{\lambda}\delta^0\}, \tag{6.15}$$

for a suitable $\bar{\lambda}$. Under general (moment) conditions, and for each (sufficiently small) δ, the random variable $\bar{\mathbf{Z}}_\delta$ is concentrated near a value of order $\delta/\sqrt{n}$ (see Section 14.8, Lemma 14.19), so then the choice $\bar{\lambda} \asymp 1/\sqrt{n}$ equips the set $\mathscr{T}$ in (6.15) with large probability. We take the particular choice $\delta^0 := G^{-1}(\varepsilon^0)$, with ε^0 given in (6.16) below.

Example 6.3. **(Least squares with fixed design)** In this reference example, one may without loss of generality assume $\beta^0 = \beta^0_{\text{GLM}}$, and $f^0 = f^0_{\text{GLM}}$. Then, when taking $\|\cdot\| := \|\cdot\|_n$,

$$\bar{\mathbf{Z}}_\delta = \sup_{\|f_\beta - f_{\beta^0}\|_n \leq \delta} 2|(\varepsilon, f_\beta - f_{\beta^0})_n|/\sqrt{p}$$

$$= \sup_{(\beta-\beta^0)^T \hat{\Sigma}(\beta-\beta^0) \leq \delta^2} 2|(\varepsilon, \psi)_n(\beta - \beta^0)|/\sqrt{p}.$$

Here, $\psi := (\psi_1, \ldots, \psi_p)$, and

$$\hat{\Sigma} := \int \psi^T \psi dQ_n$$

is the Gram matrix. Assuming $\hat{\Sigma}$ has rank p (actually without loss of generality, as when $\text{rank}(\hat{\Sigma}) = r < p$ one may replace p by r), we may reparametrize, showing that without loss of generality we may assume that the ψ_j are orthonormal in $L_2(Q_n)$, i.e., that $\hat{\Sigma} = I$. Now, define $V_j := \sqrt{n}(\varepsilon, \psi_j)_n$, $j = 1, \ldots, p$. Then $V_1, \ldots, V_p$ are i.i.d. $\mathscr{N}(0, \sigma^2)$, and

$$|(\varepsilon, \psi)_n(\beta - \beta^0)\sqrt{p}| = |V^T(\beta - \beta^0)|/\sqrt{np}$$

$$\leq \|V\|_2 \|\beta - \beta^0\|_2/\sqrt{np} = \|V\|_2 \|f_\beta - f_{\beta^0}\|_n/\sqrt{np}.$$

So

$$\bar{\mathbf{Z}}_\delta \leq 2\|V\|_2 \delta/\sqrt{np}.$$

Moreover, $\|V\|_2^2/\sigma^2$ is χ_p^2-distributed. So, for instance, we can bound the second moment of $\bar{\mathbf{Z}}_\delta$ by

$$\mathbb{E}\bar{\mathbf{Z}}_\delta^2 \leq 4\sigma^2 \delta^2/n.$$

Hence, to take care that $\mathscr{T}$ has large probability, we can take $\bar{\lambda} = 2t\sigma/\sqrt{n}$ with some large value for t, as

$$\mathbf{P}\left(\bar{\mathbf{Z}}_\delta > (2t\sigma/\sqrt{n})\delta\right) \leq \frac{1}{t^2}.$$

(See Lemma 8.1, for exponential probability bounds for χ^2 random variables.)

Let H be the convex conjugate of the function G of the margin condition. Let $\bar{\lambda} > 0$ be the constant employed in (6.15) to handle the empirical process. Define

$$\varepsilon^0 := \frac{2\mathscr{E}(f_{\mathrm{GLM}}^0)}{\alpha_0} + H\left(\frac{2\bar{\lambda}\sqrt{p}}{\alpha_0}\right).\tag{6.16}$$

The first term in (6.16) is - up to constants - the approximation error, and the second term may be seen - up to constants - as the estimation error.

Lemma 6.6. *Let $\hat{b}$ be given in (6.13), and let $\hat{f} := f_{\hat{b}}$. Assume the margin condition with G satisfying (6.14) with constants α_0 and η_0, where $\eta_0 \geq \delta^0$, with $\delta^0 := G^{-1}(\varepsilon^0)$. Assume that f_{GLM}^0 is in the neighborhood $\mathbf{F}_{\mathrm{local}}$ of the target f^0, and also that $f_\beta \in \mathbf{F}_{\mathrm{local}}$ for all β satisfying $\|f_\beta - f_{\mathrm{GLM}}^0\| \leq \delta^0$. Then, on the set $\mathscr{T}$ defined in (6.15), it holds that*

$$\mathscr{E}(\hat{f}) \leq \alpha_0\varepsilon^0 = 2\mathscr{E}(f_{\mathrm{GLM}}^0) + \alpha_0 H\left(\frac{2\bar{\lambda}\sqrt{p}}{\alpha_0}\right).$$

The assumption $f_\beta \in \mathbf{F}_{\mathrm{local}}$ for all $\|f_\beta - f_{\mathrm{GLM}}^0\| \leq \delta^0$ actually requires that the functions that are in a local $\|\cdot\|$-environment of f^0 for are also in $\mathbf{F}_{\mathrm{local}}$ which is typically an $\|\cdot\|_\infty$-environment. Again typically, the assumption depends on the dictionary and on how large δ^0 is. Such requirements on the connection of norms are more detailed in e.g. van de Geer (2002).

If the linear model is assumed to hold exactly, there is no approximation error, i.e., then $\mathscr{E}(f_{\mathrm{GLM}}^0) = 0$. The excess risk is then bounded on the set $\mathscr{T} := \{\bar{\mathbf{Z}}_{\delta^0} \leq \bar{\lambda}\delta^0\}$, where $\delta^0 = G^{-1}(H(2\bar{\lambda}\sqrt{p}/\alpha_0))$, by

$$\mathscr{E}(\hat{f}) \leq \alpha_0 H\left(\frac{2\bar{\lambda}\sqrt{p}}{\alpha_0}\right).$$

Quadratic margin In the case $G(u) = cu^2$, we have $H(v) = v^2/(4c)$, and $\alpha_0 = 1/16$ so that we get

$$\varepsilon^0 = 16\left(2\|f_{\mathrm{GLM}}^0 - f^0\|^2 + \frac{16\bar{\lambda}^2 p}{c}\right),$$

and

$$\delta^0 = \sqrt{\varepsilon^0/c},$$

and on $\mathscr{T}$,

$$\mathscr{E}(\hat{f}) \leq 2\mathscr{E}(f_{\mathrm{GLM}}^0) + \frac{16\bar{\lambda}^2 p}{c}.$$

Asymptotics As noted above, generally one can take $\bar{\lambda} \asymp 1/\sqrt{n}$ (see Section 14.8, Lemma 14.19, and see also Example 6.3). This means that under quadratic margin behavior, the estimation error $H(2\bar{\lambda}\sqrt{p}/\alpha_0)$ behaves like p/n.

Proof of Lemma 6.6 Let

$$t := \frac{\delta^0}{\delta^0 + \|\hat{f} - f^0_{\text{GLM}}\|},$$

and $\tilde{\beta} := t\hat{b} + (1-t)\beta^0_{\text{GLM}}$. Write $\tilde{f} := f_{\tilde{\beta}}$, $\tilde{\mathscr{E}} := \mathscr{E}(\tilde{f})$, and $\mathscr{E}^0 := \mathscr{E}(f^0_{\text{GLM}})$. Then by the convexity of $f \mapsto \rho_f$,

$$P_n\rho_{\tilde{f}} \le tP_n\rho_{\hat{f}} + (1-t)P_n\rho_{f^0_{\text{GLM}}} \le P_n\rho_{f^0_{\text{GLM}}}.$$

Thus, we arrive at the Basic Inequality

$$\tilde{\mathscr{E}} = -\left[v_n(\tilde{\beta}) - v_n(\beta^0_{\text{GLM}})\right] + P_n(\rho_{\tilde{f}} - \rho_{f^0_{\text{GLM}}}) + \mathscr{E}^0 \le -\left[v_n(\tilde{\beta}) - v_n(\beta^0_{\text{GLM}})\right] + \mathscr{E}^0.$$
$$(6.17)$$

So on $\mathscr{T}$

$$\tilde{\mathscr{E}} \le \bar{\lambda}\sqrt{p}\delta^0 + \mathscr{E}^0.$$

Now, by the definition of the convex conjugate, it holds that $uv \le G(u) + H(v)$, for all positive u and v. Apply this with $u := \delta^0 = G^{-1}(\varepsilon^0)$ and $v := 2\bar{\lambda}\sqrt{p}/\alpha_0$, to find

$$\tilde{\mathscr{E}} \le \alpha_0\varepsilon^0/2 + \alpha_0 H(2\bar{\lambda}\sqrt{p}/\alpha_0)/2 + \mathscr{E}^0 = \alpha_0\varepsilon^0, \qquad (6.18)$$

by the definition (6.16) of ε^0. So, by the margin condition, and using that $\tilde{f} \in \mathbf{F}_{\text{local}}$, we get by (6.14),

$$\|\tilde{f} - f^0\| \le G^{-1}(\tilde{\mathscr{E}}) \le G^{-1}(\alpha_0\varepsilon^0) \le \delta^0/4.$$

Also, as by the definition of ε^0, $\mathscr{E}^0 \le \alpha_0\varepsilon^0/2 \le \alpha_0\varepsilon^0$, so that

$$\|f^0_{\text{GLM}} - f^0\| \le G^{-1}(\mathscr{E}^0) \le G^{-1}(\alpha_0\varepsilon^0) \le \delta^0/4.$$

Thus we have shown that
$$\|\tilde{f} - f^0_{\text{GLM}}\| \le \delta^0/2.$$

But
$$\|\tilde{f} - f^0_{\text{GLM}}\| = t\|\hat{f} - f^0_{\text{GLM}}\| = \frac{\delta^0\|\hat{f} - f^0_{\text{GLM}}\|}{\delta^0 + \|\hat{f} - f^0_{\text{GLM}}\|}.$$

So $\|\tilde{f} - f^0_{\text{GLM}}\| \le \delta^0/2$ implies $\|\hat{f} - f^0_{\text{GLM}}\| \le \delta^0$. Repeat the argument, with $\tilde{f}$ replaced by $\hat{f}$, to arrive in (6.18) with $\tilde{\mathscr{E}}$ replaced by $\hat{\mathscr{E}} := \mathscr{E}(\hat{f})$. $\qquad\square$

6.6 Consistency of the Lasso for general loss

We return to the Lasso estimator $\hat{\beta}$ defined in (6.9), and let $\hat{f} := f_{\hat{\beta}}$. We first show consistency (in terms of the excess risk $\mathscr{E}(\hat{f})$). Consistency is a rough result, in particular, it requires neither the margin condition, nor any compatibility condition.

Here, and throughout the rest of this chapter, we use the notation (for $M > 0$),

$$\mathbf{Z}_M := \sup_{\|\beta - \beta^*\|_1 \leq M} |v_n(\beta) - v_n(\beta^*)|, \tag{6.19}$$

where β^* is fixed, and may be different in different sections. Sections 14.8 and 14.9 show that (under some conditions), with large probability, the empirical process increment $\mathbf{Z}_M$ is proportional to M. That is, for a value λ_0 depending the confidence level $1 - \alpha$, as well as on the problem considered, we prove that for all M sufficiently small[4]

$$\mathbf{P}(\mathbf{Z}_M \leq \lambda_0 M) \geq 1 - \alpha,$$

see Section 14.8, Lemma 14.20 and Theorem 14.5.

Lemma 6.7. *(Consistency of the Lasso) Let*

$$\beta^* := \arg\min_{\beta}\{\mathscr{E}(f_\beta) + \lambda\|\beta\|_1\}, \tag{6.20}$$

and $f^* = f_{\beta^*}$. Take $\mathbf{Z}_M$ as defined in (6.19), with the value (6.20) for β^*. Define

$$M^* := \frac{1}{\lambda_0}\{\mathscr{E}(f^*) + 2\lambda\|\beta^*\|_1\}.$$

and let

$$\mathscr{T} := \{\mathbf{Z}_{M^*} \leq \lambda_0 M^*\}, \tag{6.21}$$

where

$$4\lambda_0 \leq \lambda.$$

Then on the set $\mathscr{T}$, we have

$$\mathscr{E}(\hat{f}) + \lambda\|\hat{\beta}\|_1 \leq 2\{\mathscr{E}(f^*) + 2\lambda\|\beta^*\|_1\}.$$

Hence, if the target $f^0 = f_{\beta^0}$ is linear, this gives

$$\mathscr{E}(\hat{f}) + \lambda\|\hat{\beta}\|_1 \leq 4\lambda\|\beta^0\|_1,$$

---

[4] The probability of the set $\{\mathbf{Z}_M \leq \lambda_0 M\}$ generally does not depend on β^*. However, the set $\{\mathbf{Z}_M \leq \lambda_0 M\}$ itself generally does depend on β^* (with an exception for the case of squared error loss with fixed design), so considering several β^*'s simultaneously typically reduces the confidence level of our statements.

and more generally, for the best linear approximation

$$f_{\mathrm{GLM}}^0 := f_{\beta_{\mathrm{GLM}}^0}, \ \beta_{\mathrm{GLM}}^0 := \arg\min_{\beta} P\rho_{f_\beta},$$

we have

$$\mathscr{E}(\hat{f}) + \lambda \|\hat{\beta}\|_1 \le 2\{\mathscr{E}(f_{\mathrm{GLM}}^0) + 2\lambda \|\beta_{\mathrm{GLM}}^0\|_1\}.$$

We stress that it typically pays off to use the smallest possible value M^* in the definition (6.21) of $\mathscr{T}$ (because typically, the smaller M^*, the easier it is to handle the probability of $\mathscr{T}$).

Asymptotics As follows from Section 14.8 and Section 14.9, under general conditions one may take $\lambda_0 \asymp \sqrt{\log p/n}$. Choosing λ of the same order as λ_0, one sees that the Lasso is consistent (in terms of prediction error), if $\mathscr{E}(f^*) = o(1)$ and if in addition $\|\beta^*\|_1 = o(\sqrt{n/\log p})$. In other words, the consistency result for squared error loss with fixed design (see Lemma 6.1) lets itself be extended to general convex loss.

Proof of Lemma 6.7. This again follows quite easily from the Basic Inequality. Define

$$t := \frac{M^*}{M^* + \|\hat{\beta} - \beta^*\|_1}.$$

We use the short-hand notation $\tilde{\beta} := t\hat{\beta} + (1-t)\beta^*$. Then $\|\tilde{\beta} - \beta^*\|_1 \le M^*$. Define $\tilde{\mathscr{E}} := \mathscr{E}(f_{\tilde{\beta}})$, and $\mathscr{E}^* := \mathscr{E}(f^*)$. Then, as in (6.17), we may apply a convexity argument to obtain

$$\tilde{\mathscr{E}} + \lambda \|\tilde{\beta}\|_1 \le \mathbf{Z}_{M^*} + \mathscr{E}^* + \lambda \|\beta^*\|_1. \tag{6.22}$$

So if $\mathbf{Z}_{M^*} \le \lambda_0 M^*$, we have

$$\tilde{\mathscr{E}} + \lambda \|\tilde{\beta}\|_1 \le \lambda_0 M^* + \mathscr{E}^* + \lambda \|\beta^*\|_1.$$

But then

$$\tilde{\mathscr{E}} + \lambda \|\tilde{\beta} - \beta^*\|_1 \le \lambda_0 M^* + \mathscr{E}^* + 2\lambda \|\beta^*\|_1 = 2\lambda_0 M^* \le \lambda \frac{M^*}{2}.$$

This implies

$$\|\tilde{\beta} - \beta^*\|_1 \le \frac{M^*}{2},$$

which in turn implies

$$\|\hat{\beta} - \beta^*\|_1 \le M^*.$$

Repeat the argument with $\tilde{\beta}$ replaced by $\hat{\beta}$, to find

$$\mathscr{E}(\hat{f}) + \lambda \|\hat{\beta}\|_1 \le \lambda_0 M^* + \mathscr{E}^* + \lambda \|\beta^*\|_1 \le 2\lambda_0 M^*.$$

□

6.7 An oracle inequality

Our goal in this section is to show that the Lasso estimator (6.9) has oracle properties. We hope that the target f^0 can be well approximated by an f_β with only a few non-zero coefficients β_j ($j = 1, \ldots, p$), that is, by a *sparse* f_β. Recall that a natural choice would be to base a penalty on the number of non-zero coefficients

$$s_\beta := |S_\beta|, \ S_\beta := \{j : \beta_j \neq 0\}.$$

Remember also that s_β can be regarded as $\|\beta\|_q^q := \sum_{j=1}^q |\beta_j|^q$, with $q = 0$, which is why penalties based on s_β are often called ℓ_0-penalties. In our context, the ℓ_0-penalty takes the form $H(\tilde\lambda \sqrt{s_\beta}/\phi(S_\beta))$, where $\tilde\lambda$ is a regularization parameter, $\phi^2(S)$ is a *compatibility constant*, and H is the convex conjugate of G. This is inspired by the benchmark result of Section 6.5. Indeed, for $S \subset \{1, \ldots, p\}$ a given index set, write, as before,

$$\beta_{j,S} := \beta_j 1\{j \in S\}, \ j = 1, \ldots, p,$$

and consider the estimator, restricted to β's with only non-zero coefficients in S:

$$\hat{b}^S := \arg\min_{\beta=\beta_S} P_n \rho_{f_\beta}.$$

Write $\hat{f}_S := f_{\hat{b}^S}$. Restricted to β_S's, the best approximation of the target f^0 is $f_S := f_{b^S}$, where

$$b^S := \arg\min_{\beta=\beta_S} P \rho_{f_\beta}.$$

In view of Lemma 6.6 one has, under general conditions, on a set with large probability, the inequality

$$\mathscr{E}(\hat{f}_S) \leq 2\mathscr{E}(f_S) + \alpha_0 H\left(\frac{2\tilde\lambda \sqrt{|S|}}{\alpha_0}\right).$$

Choosing the best index set S amounts to minimizing the right-hand side over $S \in \mathscr{S}$, for a suitable, hopefully large collection of index sets $\mathscr{S}$. For the interpretation of the oracle we will use, it may be helpful to recall that

$$\min_{S \in \mathscr{S}} \left\{ 2\mathscr{E}(f_S) + \alpha_0 H\left(\frac{2\tilde\lambda \sqrt{|S|}}{\alpha_0}\right) \right\}$$

$$= \min_{\beta:\, S_\beta \in \mathscr{S}} \left\{ 2\mathscr{E}(f_\beta) + \alpha_0 H\left(\frac{2\bar{\lambda}\sqrt{s_\beta}}{\alpha_0} \right) \right\}.$$

In summary, our goal is to show that for $\hat{f} = f_{\hat{\beta}}$, with $\hat{\beta}$ the Lasso estimator, one mimics the above minimizer, with large probability, accepting some additional $\log p$-factors and/or constants from a compatibility condition.

Indeed, we need a compatibility condition to deal with the ℓ_1-norm $\|\cdot\|_1$ on the one hand, and on the other hand the norm $\|\cdot\|$ on the vector space $\mathbf{F}$.

Definition *We say that the* compatibility condition *is met for the set S, with constant* $\phi(S) > 0$, *if for all* $\beta \in \mathbf{R}^p$, *that satisfy* $\|\beta_{S^c}\|_1 \leq 3\|\beta_S\|_1$, *it holds that*

$$\|\beta_S\|_1^2 \leq \|f_\beta\|^2 |S| / \phi^2(S).$$

More details on this condition are given in Section 6.13.

Definition of the oracle *We define the oracle* β^* *as*

$$\beta^* := \arg\min_{\beta:\, S_\beta \in \mathscr{S}} \left\{ 3\mathscr{E}(f_\beta) + 2H\left(\frac{4\lambda\sqrt{s_\beta}}{\phi(S_\beta)} \right) \right\}. \tag{6.23}$$

where $S_\beta := \{j : \beta_j \neq 0\}$, *and where* $s_\beta := |S_\beta|$ *denotes the cardinality of* S_β.

This is the generalization of the oracle defined in (6.7) (for squared error loss with fixed design) to general loss and design.

We again use the short-hand notation $S_* := S_{\beta^*}$, and $\phi_* = \phi(S_*)$, and we set $f^* := f_{\beta^*}$. Thus, $\beta^* = b^{S_*}$, where

$$S_* = \arg\min_{S \in \mathscr{S}} \left\{ 3\mathscr{E}(f_S) + 2H\left(\frac{4\lambda\sqrt{|S|}}{\phi(S)} \right) \right\}.$$

The minimum is denoted as

$$2\varepsilon^* := 3\mathscr{E}(f_{\beta^*}) + 2H\left(\frac{4\lambda\sqrt{s_*}}{\phi_*} \right).$$

We let $\mathbf{Z}_M$ be given as in (6.19), but now with the newly defined β^*, i.e.,

$$\mathbf{Z}_M := \sup_{\|\beta - \beta^*\|_1 \leq M} |v_n(\beta) - v_n(\beta^*)|. \tag{6.24}$$

with the value (6.23) for β^*. Set

$$M^* := \varepsilon^* / \lambda_0. \tag{6.25}$$

and
$$\mathscr{T} := \{\mathbf{Z}_{M^*} \leq \lambda_0 M^*\} = \{\mathbf{Z}_{M^*} \leq \varepsilon^*\}. \tag{6.26}$$

Here, the idea is again to choose λ_0 in such a way that this set $\mathscr{T}$ has large probability: see Section 14.8 and Section 14.9.

Theorem 6.4. *(Oracle inequality for the Lasso) Assume the compatibility condition for all $S \in \mathscr{S}$. Assume the margin condition with strictly convex function G, and that $f_\beta \in \mathbf{F}_{local}$ for all $\|\beta - \beta^*\|_1 \leq M^*$, as well as $f^* \in \mathbf{F}_{local}$. Suppose that λ satisfies the inequality*
$$\lambda \geq 8\lambda_0. \tag{6.27}$$

Then on the set $\mathscr{T}$ given in (6.26), we have

$$\mathscr{E}(\hat{f}) + \lambda \|\hat{\beta} - \beta^*\|_1 \leq 4\varepsilon^* = 6\mathscr{E}(f^*) + 4H\left(\frac{4\lambda\sqrt{s_*}}{\phi_*}\right).$$

The condition $f_\beta \in \mathbf{F}_{local}$ for all $\|\beta - \beta^*\|_1 \leq M^*$ is again, typically, a condition on the connection between norms, as in Lemma 6.6. We note that $\|\beta - \beta^*\|_1 \leq M^*$ implies $\|f_\beta - f_{\beta^*}\|_\infty \leq M^* K$, where

$$K := \max_{1 \leq j \leq p} \|\psi_j\|_\infty.$$

This means that typically (with $\mathbf{F}_{local}$ being an $\|\cdot\|_\infty$-neighborhood) we need $K < \infty$ (but possibly growing with n and/or p, depending on how large M^* is). See Example 6.4 for the details in the case of logistic regression.

Corollary 6.3. *Assume the conditions of Theorem 6.4, with quadratic margin behavior, i.e., with $G(u) = cu^2$. Then $H(v) = v^2/(4c)$, and we obtain on $\mathscr{T}$,*

$$\mathscr{E}(\hat{f}) + \lambda \|\hat{\beta} - \beta^*\|_1 \leq 6\mathscr{E}(f^*) + \frac{16\lambda^2 s_*}{c\phi_*^2}.$$

The definition of the oracle β^* allows much flexibility, in the sense that the choice of the collection $\mathscr{S}$ is left unspecified. We may apply the result to the best linear approximation

$$f^0_{\mathrm{GLM}} := f_{\beta^0_{\mathrm{GLM}}}, \ \beta^0_{GLM} := \arg\min_\beta P\rho_{f_\beta}.$$

Suppose that $S_{0,\mathrm{GLM}} := S_{\beta^0_{\mathrm{GLM}}}$ satisfies the compatibility condition. Then we obtain under the conditions of Theorem 6.4, with $\mathscr{S} := \{S_{0,\mathrm{GLM}}\}$,

$$\mathscr{E}(\hat{f}) + \lambda \|\hat{\beta} - \beta^0_{\mathrm{GLM}}\|_1 \leq 6\mathscr{E}(f^0_{\mathrm{GLM}}) + 4H\left(\frac{4\lambda\sqrt{|S_{0,\mathrm{GLM}}|}}{\phi(S_{0,\mathrm{GLM}})}\right).$$

This thus gives a bound for both the excess risk $\mathcal{E}(\hat{f})$, as well as the ℓ_1-error $\|\hat{\beta} - \beta_{\mathrm{GLM}}^0\|_1$. If the target $f^0 = f_{\beta^0}$ is linear, one has $f^0 = f_{\mathrm{GLM}}^0$, i.e., then the approximation error $\mathcal{E}(f_{\mathrm{GLM}}^0)$ vanishes.

In order to improve the bound for the excess risk $\mathcal{E}(\hat{f})$, one may choose to minimize over a larger collection $\mathcal{S}$. One then ends up with the ℓ_1-error $\|\hat{\beta} - \beta^*\|_1$, between the estimator $\hat{\beta}$ and the less easy to interpret β^*. But again, a triangle inequality

$$\|\hat{\beta} - \beta_{\mathrm{GLM}}^0\|_1 \leq \|\hat{\beta} - \beta^*\|_1 + \|\beta^* - \beta_{\mathrm{GLM}}^0\|_1,$$

can overcome the interpretation problem (see Problem 6.4).[5]

Asymptotics We generally can take $\lambda_0 \asymp \sqrt{\log(p)/n}$ (see Section 14.8 and Section 14.9). Taking λ also of this order, and assuming a quadratic margin, and in addition that ϕ_* stays away from zero, the estimation error $H(4\lambda\sqrt{s_*}/\phi_*)$ behaves like $s_* \log p/n$.

Proof of Theorem 6.4. Throughout the proof, we assume we are on the set $\mathcal{T}$.

Let, as in the proof of Lemma 6.7,

$$t := \frac{M^*}{M^* + \|\hat{\beta} - \beta^*\|_1}.$$

We again use the short-hand notation $\tilde{\beta} := t\hat{\beta} + (1-t)\beta^*$, and $\tilde{\mathcal{E}} := \mathcal{E}(f_{\tilde{\beta}})$, $\mathcal{E}^* := \mathcal{E}(f_{\beta^*})$. Then, as in (6.17), a convexity argument gives the Basic Inequality

$$\tilde{\mathcal{E}} + \lambda\|\tilde{\beta}\|_1 \leq \mathbf{Z}_{M^*} + \mathcal{E}^* + \lambda\|\beta^*\|_1 \leq \lambda_0 M^* + \mathcal{E}^* + \lambda\|\beta^*\|_1. \tag{6.28}$$

Now, for any β,

$$\beta = \beta_{S_*} + \beta_{S_*^c},$$

Note thus that $\beta_{S_*}^* = \beta^*$ and $\beta_{S_*^c}^* \equiv 0$.

So we have

$$\tilde{\mathcal{E}} + \lambda\|\tilde{\beta}_{S_*^c}\|_1 \leq \mathcal{E}^* + \mathcal{E}^* + \lambda\|\tilde{\beta}_{S_*} - \beta^*\|_1 \leq 2\mathcal{E}^* + \lambda\|\tilde{\beta}_{S_*} - \beta^*\|_1. \tag{6.29}$$

Case i) If

$$\lambda\|\tilde{\beta}_{S_*} - \beta^*\|_1 \geq \mathcal{E}^*,$$

we get from (6.29), that

$$\lambda\|\tilde{\beta}_{S_*^c}\|_1 \leq 2\mathcal{E}^* + \lambda\|\tilde{\beta}_{S_*} - \beta^*\|_1 \leq 3\lambda\|\tilde{\beta}_{S_*} - \beta^*\|_1. \tag{6.30}$$

---

[5] In connection with this, we remark that one may for instance want to choose $\mathcal{S}$ to contain only subsets of $S_{0,\mathrm{GLM}}$. The oracle is then not allowed to choose non-zeroes where the best approximation has zeroes .

This means that we can apply the compatibility condition.

We find
$$\|\tilde{\beta}_{S_*} - \beta^*\|_1 \le \sqrt{s_*}\|\tilde{f} - f^*\|/\phi_*.$$

Thus
$$\tilde{\mathscr{E}} + \lambda\|\tilde{\beta}_{S_*^c}\|_1 + \lambda\|\tilde{\beta}_{S_*} - \beta^*_{S_*}\|_1$$
$$\le \lambda_0 M^* + \mathscr{E}^* + 2\lambda\sqrt{s_*}\|\tilde{f} - f^*\|/\phi_*.$$

Now, because we assumed $f^* \in \mathbf{F}_{\text{local}}$, and since also $\tilde{f} \in \mathbf{F}_{\text{local}}$, we can invoke the margin condition to arrive at

$$2\lambda\sqrt{s_*}\|\tilde{f} - f^*\|/\phi^* \le H\left(\frac{4\lambda\sqrt{s_*}}{\phi_*}\right) + \tilde{\mathscr{E}}/2 + \mathscr{E}^*/2.$$

It follows that

$$\tilde{\mathscr{E}} + \lambda\|\tilde{\beta} - \beta^*\|_1 \le \lambda_0 M^* + 3\mathscr{E}^*/2 + H\left(\frac{4\lambda\sqrt{s_*}}{\phi_*}\right) + \tilde{\mathscr{E}}/2$$

$$\le \lambda_0 M^* + \varepsilon^* + \tilde{\mathscr{E}}/2 = 2\lambda_0 M^* + \tilde{\mathscr{E}}/2 = 2\varepsilon^* + \tilde{\mathscr{E}}/2,$$

or
$$\tilde{\mathscr{E}}/2 + \lambda\|\tilde{\beta} - \beta^*\|_1 \le 2\varepsilon^*. \tag{6.31}$$

This yields
$$\|\tilde{\beta} - \beta^*\|_1 \le \frac{2\lambda_0}{\lambda}M^* \le \frac{M^*}{2},$$

where the last inequality is ensured by the assumption $\lambda \ge 8\lambda_0 \ge 4\lambda_0$. But $\|\tilde{\beta} - \beta^*\|_1 \le M^*/2$ implies
$$\|\hat{\beta} - \beta^*\|_1 \le M^*.$$

Case ii) If
$$\lambda\|\tilde{\beta}_{S_*} - \beta^*\|_1 < \varepsilon^*,$$

this implies
$$\tilde{\mathscr{E}} + \lambda\|\tilde{\beta}_{S_*^c}\|_1 \le 3\varepsilon^*,$$

and hence
$$\tilde{\mathscr{E}} + \lambda\|\tilde{\beta} - \beta^*\|_1 \le 4\varepsilon^*, \tag{6.32}$$

so that
$$\|\tilde{\beta} - \beta^*\|_1 \le 4\frac{\varepsilon^*}{\lambda} = 4\frac{\lambda_0}{\lambda}M^* \le \frac{M^*}{2},$$

since $\lambda \ge 8\lambda_0$. This implies that $\|\hat{\beta} - \beta^*\|_1 \le M^*$.

So in both cases, we arrive at $\|\hat{\beta} - \beta^*\|_1 \le M^*$.

Now, repeat the above arguments with $\tilde{\beta}$ replaced by $\hat{\beta}$. Then, with this replacement, either from (6.30) we can apply the compatibly assumption to arrive at (6.31) (**Case i)**), or we reach (6.32) (**Case ii)**).

$\square$

Example 6.4. **Logistic regression**

As an illustration, with characteristics shared by various other examples, we examine the logistic regression model. Suppose $\{(X_i, Y_i)\}_{i=1}^n$ are independent copies of (X, Y), with $X \in \mathcal{X}$ and $Y \in \{0, 1\}$. The distribution of X is denoted by Q, and we let $\mathbf{F} = L_2(Q)$, so that $\|\cdot\|$ is the $L_2(Q)$-norm. The logistic loss is

$$\rho_f(\cdot, y) = \rho(f(\cdot), y) = -yf(\cdot) + \log(1 + \exp[f(\cdot)]),$$

see also Section 3.3.1. Define

$$\pi(\cdot) := \mathbf{P}(Y = 1 | X = \cdot).$$

Then for $x \in \mathcal{X}$,

$$l(a, x) := \mathbb{E}(\rho(a, Y) | X = x) = -\pi(x)a + \log(1 + \exp[a]),$$

which is minimized at

$$a = \log\left(\frac{\pi(x)}{1 - \pi(x)}\right).$$

Therefore, we take $f^0 := \log(\pi/(1-\pi))$ as target. To check the margin condition, we consider the second derivative

$$\ddot{l}(a, \cdot) := \frac{d^2}{da^2} l(a, \cdot) = \frac{e^a}{1 + e^a}\left(1 - \frac{e^a}{1 + e^a}\right).$$

Let $\mathbf{F}_{\text{local}}$ be the neighborhood $\mathbf{F}_{\text{local}} := \{\|f - f^0\|_\infty \leq \eta\}$.

Lemma 6.8. *Assume that for some constant $0 < \varepsilon_0 < 1$,*

$$\varepsilon_0 < \pi(x) < 1 - \varepsilon_0, \ \forall \, x, \tag{6.33}$$

and furthermore, that for some constant K,

$$\max_{1 \leq j \leq p} \|\psi_j\|_\infty \leq K.$$

Take, for some constant L, $8\lambda_0 \leq \lambda \leq L\lambda_0$. Suppose that

$$\frac{8KL^2(e^\eta/\varepsilon_0 + 1)^2}{\eta} \frac{\lambda_0 s_*}{\phi_*^2} \leq 1. \tag{6.34}$$

and that $\|f^* - f^0\|_\infty \le \eta/2$ and $\mathscr{E}(f^*)/\lambda_0 \le \eta/4$. Then the conditions of Theorem 6.4 are met, and on $\mathscr{T}$, we have

$$\mathscr{E}(\hat{f}) + \lambda\|\hat{\beta} - \beta^*\|_1 \le 6\mathscr{E}(f_{\beta^*}) + \frac{16\lambda^2 s_*(e^\eta/\varepsilon_0 + 1)^2}{\phi_*^2}.$$

We can relax condition (6.33), but then the margin is possibly no longer quadratic (see Example 6.2). One furthermore sees that in logistic regression, the loss function $\rho(f,y)$ is Lipschitz in f for all y. This will greatly facilitate the handling of the set $\mathscr{T}$, namely, it allows one to apply a so-called *contraction inequality* (see Theorem 14.4). By Theorem 14.5 of Section 14.8, under the conditions of Lemma 6.8, and with the additional normalizing condition that $\|\psi_j\| \le 1$ for all j, we have the probability inequality

$$\mathbf{P}(\mathscr{T}) \ge 1 - \exp[-t],$$

for

$$\lambda_0 := \left[4\lambda\left(\frac{K}{3}, n, p\right) + \frac{tK}{3n} + \sqrt{\frac{2t}{n}}\sqrt{1 + 8\lambda\left(\frac{K}{3}, n, p\right)}\right],$$

where

$$\lambda\left(\frac{K}{3}, n, p\right) := \sqrt{\frac{2\log(2p)}{n}} + \frac{K\log(2p)}{3n}.$$

Proof of Lemma 6.8. For $f \in \mathbf{F}_\eta$,

$$\ddot{l}(f, \cdot) \ge (e^{|f^0| + \eta} + 1)^{-2} \ge (e^\eta/\varepsilon_0 + 1)^{-2}.$$

Hence, the quadratic margin condition holds, with $G(u) = cu^2$, where

$$c = (e^\eta/\varepsilon_0 + 1)^{-2}.$$

For $\|\beta - \beta^*\|_1 \le M^*$, we have

$$\|f_\beta - f^*\|_\infty \le KM^*,$$

and hence

$$\|f_\beta - f^0\|_\infty \le \eta/2 + KM^*.$$

Now,

$$M^* := \varepsilon^*/\lambda_0 = 2\mathscr{E}(f_{\beta^*})/\lambda_0 + \frac{4\lambda^2 s_*}{c\phi_*^2}/\lambda_0.$$

Condition (6.34) ensures that

$$KM^* \le \eta/2.$$

$\square$

6.8 The ℓ_q-error for $1 \leq q \leq 2$

Theorem 6.4 provides us with rates for the ℓ_1-error $\|\hat{\beta} - \beta^*\|_1$, which in turn can be used to derive rates for (for instance) $\|\hat{\beta} - \beta^0_{\text{GLM}}\|_1$, where β^0_{GLM} are the coefficients of the best linear approximation of f^0 (see also Problem 6.4).

To derive rates for $\|\hat{\beta} - \beta^*\|_q$ (and $\|\hat{\beta} - \beta^0_{\text{GLM}}\|_q$), with $1 < q \leq 2$, we need a stronger compatibility condition.

For $\mathcal{N} \supset S$, and $L \geq 0$, define the restricted set of β's

$$\mathcal{R}(L, S, \mathcal{N}) := \left\{ \|\beta_{S^c}\|_1 \leq L\|\beta_S\|_1, \ \|\beta_{\mathcal{N}^c}\|_\infty \leq \min_{j \in \mathcal{N} \setminus S} |\beta_j| \right\}.$$

If $\mathcal{N} = S$, we necessarily have $\mathcal{N} \setminus S = \emptyset$. In that case, we let $\min_{j \in \mathcal{N} \setminus S} |\beta_j| = \infty$.

The restricted eigenvalue condition (Bickel et al. (2009); Koltchinskii (2009b)) is essentially the following condition.

Definition *Let S be an index set with cardinality s, L be a non-negative constant, and N $\geq$ s be an integer. We say that the (L, S, N)-restricted eigenvalue condition is satisfied, with constant $\phi(L, S, N) > 0$, if for all $\mathcal{N} \supset S$, with $|\mathcal{N}| = N$, and all $\beta \in \mathcal{R}(L, S, \mathcal{N})$, it holds that*

$$\|\beta_{\mathcal{N}}\|_2 \leq \|f_\beta\|/\phi(L, S, N).$$

For the case where we only have a linear approximation of the truth, we need another version, which we call the (minimal) adaptive restricted eigenvalue condition.

Define the restricted set of β's

$$\mathcal{R}_{\text{adap}}(L, S, \mathcal{N}) := \left\{ \|\beta_{S^c}\|_1 \leq L\sqrt{s}\|\beta_S\|_2, \ \|\beta_{\mathcal{N}^c}\|_\infty \leq \min_{j \in \mathcal{N} \setminus S} |\beta_j| \right\}.$$

Definition *Let S be an index set with cardinality s and N $\geq$ s be an integer. We say that the adaptive (L, S, N)-restricted eigenvalue condition is satisfied, with constant $\phi_{\text{adap}}(L, S, N) > 0$, if for all $\mathcal{N} \supset S$, with $|\mathcal{N}| = N$, and all $\beta \in \mathcal{R}_{\text{adap}}(L, S, \mathcal{N})$, it holds that*

$$\|\beta_{\mathcal{N}}\|_2 \leq \|f_\beta\|/\phi_{\text{adap}}(L, S, N).$$

Definition *For S a set with cardinality s, L a non-negative constant, and N $\geq$ s an integer, the minimal adaptive restricted eigenvalue is*

$$\phi^2_{\text{min}}(L, S, N) = \min_{\mathcal{N} \supset S, |\mathcal{N}| = N} \phi^2_{\text{adap}}(L, \mathcal{N}, N).$$

The next lemma is our tool to go from the ℓ_1-norm to an ℓ_q-norm (see Candès and Tao (2007)).

Lemma 6.9. *Let $b_1 \geq b_2 \geq \ldots \geq 0$. For $1 < q < \infty$, and $s \in \{1,\ldots\}$, we have*

$$\left(\sum_{j \geq s+1} b_j^q \right)^{1/q} \leq \sum_{k=1}^{\infty} \left(\sum_{j=ks+1}^{(k+1)s} b_j^q \right)^{\frac{1}{q}} \leq s^{-(q-1)/q} \|b\|_1.$$

Proof. Clearly,

$$\left(\sum_{j \geq s+1} b_j^q \right)^{\frac{1}{q}} = \left(\sum_{k=1}^{\infty} \sum_{j=ks+1}^{(k+1)s} b_j^q \right)^{\frac{1}{q}} \leq \sum_{k=1}^{\infty} \left(\sum_{j=ks+1}^{(k+1)s} b_j^q \right)^{\frac{1}{q}}.$$

Moreover, for $ks+1 \leq j \leq (k+1)s$,

$$b_j \leq \sum_{l=(k-1)s+1}^{ks} b_l / s.$$

So

$$\sum_{j=ks+1}^{(k+1)s} b_j^q \leq s^{-(q-1)} \left(\sum_{l=(k-1)s}^{ks} b_l \right)^q.$$

Therefore

$$\sum_{k=1}^{\infty} \left(\sum_{j=ks+1}^{(k+1)s} b_j^q \right)^{\frac{1}{q}} \leq s^{-(q-1)/q} \sum_{k=1}^{\infty} \sum_{l=(k-1)s}^{ks} b_l = s^{-(q-1)/q} \|b\|_1.$$

$\square$

6.8.1 Application to least squares assuming the truth is linear

The results in this subsection are from Bickel et al. (2009). Consider the situation of Subsection 6.2.2. The following results can be invoked in the same way as in Problem 2.3, to improve variable screening, as discussed in Subsection 2.5.

Lemma 6.10. *Assume $\lambda \geq 2\lambda_0$. Then on*

$$\mathscr{T} := \left\{ \max_{1 \leq j \leq p} 2|\varepsilon^T \mathbf{X}^{(j)}|/n \leq \lambda_0 \right\},$$

and for $1 \leq q \leq 2$, we have

$$\|\hat{\beta} - \beta^0\|_q^q \le \frac{(2^q + 4^q)\lambda^q s_0}{\phi^{2q}(3, S_0, 2s_0)}.$$

Proof. We showed in Lemma 6.3, that for $\lambda \ge 2\lambda_0$, on the set $\mathscr{T}$,

$$\|\hat{\beta}_{S_0^c}\|_1 \le 3\|\hat{\beta}_{S_0} - \beta^0\|_1.$$

By Theorem 6.1 moreover, on $\mathscr{T}$,

$$\|\mathbf{X}(\hat{\beta} - \beta^0)\|_2^2/n + \lambda\|\hat{\beta} - \beta^0\|_1 \le 4\lambda^2 s_0/\phi_0^2,$$

where $\phi_0 = \phi(S_0) \ge \phi(3, S_0, 2s_0)$. The $(3, S_0, 2s_0)$-restricted eigenvalue condition implies that if we take $\mathcal{N} \backslash S_0$ as the indices of the s largest in absolute value coefficients $|\hat{\beta}_j|$, $j \notin S_0$,

$$\|\hat{\beta}_{\mathcal{N}} - \beta^0\|_2^2 \le \frac{1}{\phi^2(3, S_0, 2s_0)}\left(4\lambda^2 s_0/\phi_0^2\right) \le \frac{4\lambda^2 s_0}{\phi^4(3, S_0, 2s_0)}.$$

So

$$\|\hat{\beta}_{\mathcal{N}} - \beta^0\|_q \le s_0^{-(2q-1)/2q}\|\hat{\beta}_{\mathcal{N}} - \beta^0\|_2 \le s_0^{1/q}2\lambda/\phi^2(3, S_0, 2s_0).$$

Moreover, invoking Lemma 6.9,

$$\|\hat{\beta}_{\mathcal{N}^c}\|_q \le s_0^{-(q-1)/q}\|\beta_{S_0^c}\|_1 \le 4s_0^{1/q}\lambda/\phi_0^2 \le 4s_0^{1/q}\lambda/\phi^2(3, S_0, 2s_0).$$

Thus,

$$\|\hat{\beta} - \beta^0\|_q^q = \|\hat{\beta}_{\mathcal{N}^c}\|_q^q + \|\hat{\beta}_{\mathcal{N}} - \beta^0\|_q^q \le \frac{2^q\lambda^q s_0}{\phi^{2q}(3, S_0, 2s_0)} + \frac{4^q\lambda^q s_0}{\phi^{2q}(3, S_0, 2s_0)}$$

$$= \frac{(2^q + 4^q)\lambda^q s_0}{\phi^{2q}(3, S_0, 2s_0)}.$$

$\square$

6.8.2 Application to general loss and a sparse approximation of the truth

Recall from the previous section, that S_* is the active set of the oracle $f^* := f_{\beta^*}$ and $\phi_* = \phi(S_*)$, and that

$$2\varepsilon^* := 3\mathscr{E}(f_{\beta^*}) + 2H\left(\frac{4\lambda\sqrt{s_*}}{\phi_*}\right).$$

We now define

$$2e^* := 3\mathscr{E}(f_{\beta^*}) + 2H\left(\frac{4\lambda\sqrt{2s_*}}{\phi_{\min}(3, S_*, 2s_*)}\right).$$

Lemma 6.11. *Suppose the conditions of Theorem 6.4 are met. Then on the set* $\mathscr{T} = \{\mathbf{Z}_{M^*} \le \lambda_0 M^*\}$ *defined in (6.26), we have for* $1 \le q \le 2$,

$$\|\hat{\beta} - \beta^*\|_q^q \le (4^q + 2^{q+1})s_*^{-(q-1)}(e^*/\lambda)^q. \tag{6.35}$$

We remark that when the truth f^0 is itself linear, i.e., when $f^0 = f_{\mathrm{GLM}}^0$, and one takes $f^* = f_{\mathrm{GLM}}^0$, then one can replace the minimal adaptive restricted eigenvalue $\phi_{\min}(3, S_0, 2s_0)$ by the (not smaller and perhaps larger) restricted eigenvalue $\phi(3, S_0, 2s_0)$, as in the case of least squares error loss considered in Lemma 6.10. One then needs to adjust the arguments in the proof of Theorem 6.4, considering the sets

$$\left\{\sup_{\|\beta - \beta^*\|_1 \le M} |v_n(\beta) - v_n(\beta^*)|/\|\beta - \beta^*\|_1 \le \lambda_0\right\}.$$

With the help of the so-called *peeling device* (see e.g. van de Geer (2000)), one can show that such sets have large probability.

Proof of Lemma 6.11. By Theorem 6.4,

$$\lambda\|\hat{\beta} - \beta^*\|_1 \le 4\varepsilon^*. \tag{6.36}$$

In the proof of Theorem 6.4, we keep β^* as it is, and we replace S_* by $\mathscr{N}$, with $\mathscr{N}\backslash S_*$ being the set indices of of the largest $|\hat{\beta}_j|$, $j \notin S_*$. We moreover replace $\|\hat{\beta}_\mathscr{N} - \beta^*\|_1$ by $\sqrt{2s_*}\|\hat{\beta}_\mathscr{N} - \beta^*\|_2$. Instead of 6.36, we then obtain

$$\lambda\|\hat{\beta}_{\mathscr{N}^c}\|_1 + \lambda\sqrt{2s_*}\|\hat{\beta}_\mathscr{N} - \beta^*\|_2 \le 4e^*.$$

(see also Lemma 6.12 for this line of reasoning). It follows that

$$\|\hat{\beta}_{\mathscr{N}^c}\|_q \le s^{-(q-1)/q}\|\hat{\beta}_{S_*^c}\|_1 \le 4s^{-(q-1)/q}e^*/\lambda.$$

We also have

$$\|\hat{\beta}_\mathscr{N} - \beta^*\|_2 \le 4e^*/(\lambda\sqrt{2s_*}),$$

and thus

$$\|\hat{\beta}_\mathscr{N} - \beta^*\|_q \le (2s_*)^{(q-2)/(2q)}\|\hat{\beta}_\mathscr{N} - \beta^*\|_2 \le (2s_*)^{-(q-1)/q}4e^*/\lambda.$$

Hence

$$\|\hat{\beta} - \beta^*\|_q^q \le 4^q s_*^{-(q-1)}(e^*/\lambda)^q + 4^q(2s_*)^{-(q-1)}(e^*/\lambda)^q$$

$$= (1 + 2^{-(q-1)})4^q s_*^{-(q-1)}(e^*/\lambda)^q.$$

$\square$

The following corollary takes $f^* = f^0$.

Corollary 6.4. *If $f^0 = f_{\beta^0}$ is linear, we have for $\mathscr{S} = \{S_0\}$, under the conditions of Lemma 6.11, and on $\mathscr{T}$ defined in (6.26),*

$$\|\hat{\beta} - \beta^0\|_q^q \leq (4^q + 2^{q+1}) s_0^{-(q-1)} \lambda^{-q} H^q \left(\frac{4\lambda \sqrt{2s_0}}{\phi_{\min}(3, S_0, 2s_0)} \right).$$

In the case of quadratic margin behavior, with $G(u) = cu^2$, we then get on $\mathscr{T}$,

$$\|\hat{\beta} - \beta^0\|_q^q \leq (4^q + 2^{q+1}) \lambda^q s_0 \left(\frac{8}{c^2 \phi_{\min}^2(3, S_0, 2s_0)} \right)^q$$

(compare with Lemma 6.10).

Of special interest is the case $q = 2$. We present the situation with quadratic margin behavior.

Corollary 6.5. *Assume the conditions of Lemma 6.11, and that $G(u) = cu^2$. Then on $\mathscr{T}$ defined in (6.26),*

$$\|\hat{\beta} - \beta^*\|_2^2 \leq 6\lambda^2 s_* \left(\frac{3\mathscr{E}(f_{\beta^*})}{\lambda^2 s_*} + \frac{16}{c^2 \phi_{\min}^2(3, S_*, 2s_*)} \right)^2.$$

6.9 The weighted Lasso

The weighted Lasso is

$$\hat{\beta} = \arg\min_{\beta} \left\{ P_n \rho_{f_\beta} + \lambda \sum_{j=1}^p w_j |\beta_j| \right\}, \tag{6.37}$$

where $\{w_j\}_{j=1}^p$ are (possibly random) non-negative weights. Consider $f_\beta(\cdot) = \sum_{j=1}^p \beta_j \psi_j(\cdot)$, with ψ_j given base functions on $\mathscr{X}$. Inserting weights in the penalty is equivalent to a normalization of the base functions ψ_j. The normalization is implicit in the previous sections, namely in the sets $\mathscr{T}$ considered. Only with a proper normalization will these sets have large probability.

Suppose that $\psi_1 \equiv 1$, and that the other ψ_j's are centered. A reasonable choice for the weights is then to take $w_1 = 0$ and, for $j = 2, \ldots, p$,

$$w_j^2 = P_n \psi_j^2 := \hat{\sigma}_j^2. \tag{6.38}$$

Corollary 6.6 below will deal with zero weights, that is, unpenalized terms. Further-more, weights that uniformly stay away from zero and infinity, do not require a new argument. That is, with the weights as in (6.38) (say) on the set

$$\Omega := \{\delta \le w_j \le 1/\delta, \ \forall \ j \in \{2,\ldots,p\}\},$$

with $0 < \delta < 1$ fixed but otherwise arbitrary, one can easily adjust Theorem 6.4, combining it with Corollary 6.6, with the constant δ now appearing in the bounds. We will furthermore show in Example 14.1 (Section 14.5.1, see also Problem 14.4) that, under some moment conditions, with random design, for the choice (6.38), the set Ω has large probability.

The situation is more involved if weights can be arbitrarily large or arbitrarily small. Assuming the truth $f^0 = f_{\beta^0}$ is linear, with active set S_0, the adaptive Lasso attempts to make the weights w_j close to zero for indices j in S_0, and very large for j not in S_0. This means the weights are chosen "adaptively", a situation which is deferred to Section 6.10.

The remainder of this section studies the case where certain weights are a priori set to zero. As pointed out, the constant term is often left unpenalized, and there may also be other unpenalized terms; say $w_1 = \ldots = w_r = 0$ and all other weights are equal to one. The Lasso then becomes

$$\hat{\beta} = \arg\min_{\beta} \left\{ P_n \rho_{f_\beta} + \lambda \sum_{j=r+1}^{p} |\beta_j| \right\}, \tag{6.39}$$

where $r \le p$. One can rather easily see from the proof of Theorem 6.4 that if one forces the oracle to keep the unpenalized coefficients, there is no additional argument needed to handle this situation. This gives the following corollary.

Corollary 6.6. *Assume the conditions of Theorem 6.4. Suppose the oracle is re-stricted to sets S containing the indices $1,\ldots,r$ of the unpenalized coefficients. Then on the set $\mathcal{T}$ given in (6.26), we have for the estimator $\hat{\beta}$ given in (6.39),*

$$\mathcal{E}(f_{\hat{\beta}}) + \lambda \|\hat{\beta} - \beta^*\|_1 \le 4\varepsilon^* = 6\mathcal{E}(f_{\beta^*}) + 4H\left(\frac{4\lambda\sqrt{s_*}}{\phi_*}\right).$$

Again, when the penalty is a weighted version $\sum_{j=r+1}^{p} w_j |\beta_j|$, with the non-zero weights w_j bounded away from zero and infinity (with large probability), the theory goes through with no substantial changes.

If none of the coefficients are penalized, we are back in the situation of Section 6.5 (assuming $p \le n$). Comparing the result with those of Section 6.5, we see that we now rely on the perhaps restrictive compatibility condition, and that the result is possibly less good, in the sense that $\bar{\lambda} < \lambda$ (i.e., we generally now have a superfluous log-term).

Also with some penalized coefficients, the above corollary does not always give the best result. With squared error loss, one may use projection arguments. For instance, centering all ψ_j makes them orthogonal to the constant term. More generally, one may want to choose the $\{\psi_j\}_{j=r+1}^{p}$ orthogonal to the $\{\psi_j\}_{j=1}^{r}$. In the least squares case, one can then simply separate the estimation of the $\{\beta_j\}_{j=1}^{r}$ from the estimation of the penalized coefficients $\{\beta_j\}_{j=r+1}^{p}$ (see also Problem 6.9). In other situations, this is however not always possible.

An extension occurs when there are additional "nuisance" parameters, which do not occur as coefficients in the linear model, but rather nonlinearly (and which are not penalized, say). With the above approach, assuming some smoothness in the nuisance parameters, such a situation can be incorporated as well. More precise derivations for possibly non-convex models are given in Section 9.4.

6.10 The adaptively weighted Lasso

We again look at squared error loss (other loss can be treated similarly), and consider, as in Section 6.9, the weighted Lasso

$$\hat{\beta}_{\text{weight}} = \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - f_\beta(X_i))^2 + \lambda_{\text{init}} \lambda_{\text{weight}} \sum_{j=1}^{p} w_j |\beta_j| \right\},$$

where $\{w_j\}_{j=1}^{p}$ are (possibly random) non-negative weights. This section is tailored for the situation where the weights may be chosen depending on some initial estimator, as is for example the case for the adaptive Lasso (see Section 2.8 and, for further theoretical results, Sections 7.8 and 7.9). This generally means that the regularization parameter λ_{init} is of the same order of magnitude as the regularization parameter λ for the standard Lasso. An appropriate choice for the regularization parameter λ_{weight} depends on the (random) weights (see (6.40)).
Fix $\lambda_0 > 0$ and define

$$\mathscr{T} := \left\{ \max_{1 \le j \le p} 2|(\varepsilon, \psi_j)_n| \le \lambda_0 \right\}.$$

For sets $S \subset \{1, \ldots, n\}$ we define

$$\|w_S\|_2 := \sqrt{\sum_{j \in S} w_j^2}, \ w_{S^c}^{\min} := \min_{j \notin S} w_j.$$

Let us recall the definition of the adaptive restricted eigenvalue (see Section 6.8). For $L \ge 0$, $\mathscr{N} \supset S$, we define the set of restrictions

$$\mathscr{R}_{\text{adap}}(L,S,\mathscr{N}) := \left\{ \|\beta_{S^c}\|_1 \le L\sqrt{s}\|\beta_S\|_2, \ \|\beta_{\mathscr{N}^c}\|_\infty \le \min_{j \in \mathscr{N}\backslash S}|\beta_j| \right\}.$$

Definition *We say that the adaptive (L,S,N)-restricted eigenvalue condition is satisfied, with constant $\phi_{\text{adap}}(L,S,N) > 0$, if for all $\mathscr{N} \supset S$, with $|\mathscr{N}| = N$, and all $\beta \in \mathscr{R}_{\text{adap}}(L,S,\mathscr{N})$, it holds that*

$$\|\beta_{\mathscr{N}}\|_2 \le \|f_\beta\|/\phi_{\text{adap}}(L,S,N).$$

The following lemma is proved in the same way as Theorem 6.2. It essentially reduces to Theorem 6.2 when one takes $\lambda_w = 1$ and $w_j = 1$ for all j. The adaptive restricted eigenvalue condition is however somewhat stronger than our "usual" compatibility condition. See Section 6.13 for the relations between various restricted eigenvalues and the compatibility condition.

Lemma 6.12. *Suppose we are on $\mathscr{T}$. Let $\lambda_{\text{init}} \ge 2\lambda_0$. Let S be a set with cardinality $|S| := s$, that satisfies*

$$Lw_{S^c}^{\min} \ge \|w_S\|_2/\sqrt{s},$$

and

$$\lambda_{\text{weight}}(\|w_S\|_2/\sqrt{s} \wedge w_{S^c}^{\min}) \ge 1. \tag{6.40}$$

Write $\hat{f}_{\text{weight}} := f_{\hat{\beta}_{\text{weight}}}$. For all β we have

$$\|\hat{f}_{\text{weight}} - f^0\|_n^2 + \lambda_{\text{init}}\lambda_{\text{weight}} \sum_{j \notin S} w_j|\hat{\beta}_{\text{weight},j}|/2$$

$$\le 2\|f_{\beta_S} - f^0\|_n^2 + \frac{14\lambda_{\text{init}}^2\lambda_{\text{weight}}^2}{\phi_{\text{adap}}^2(6L,S,s)}\|w_S\|_2^2,$$

and

$$\lambda_{\text{init}}\lambda_{\text{weight}}\|w_S\|_2\|(\hat{\beta}_{\text{weight}})_S - \beta_S\|_2 + \lambda_{\text{init}}\lambda_{\text{weight}} \sum_{j \notin S} w_j|\hat{\beta}_{\text{weight},j}|$$

$$\le 5\|f_{\beta_S} - f^0\|_n^2 + \frac{7\lambda_{\text{init}}^2\lambda_{\text{weight}}^2}{\phi_{\text{adap}}^2(6L,S,s)}\|w_S\|_2^2.$$

Note that by (6.40), the bound for the prediction error of the weighted Lasso is always at least $14\lambda_{\text{init}}^2|S|^2/\phi_{\text{adap}}^2(6L,S,s)$ (compare with the prediction error for the initial Lasso, and compare also with Condition C in Section 7.8.6).

Proof of Lemma 6.12. We have on $\mathscr{T}$

$$\|\hat{f}_{\text{weight}} - f^0\|_n^2 + \lambda_{\text{init}}\lambda_{\text{weight}} \sum_{j=1}^p w_j|\hat{\beta}_{\text{weight},j}|$$

$$\leq \|f_{\beta_S} - f^0\|_n^2 + \lambda_0\|\hat{\beta}_{\text{weight}} - \beta\|_1 + \lambda_{\text{init}}\lambda_{\text{weight}}\sum_{j\in S} w_j|\beta_j|,$$

and hence

$$\|\hat{f}_{\text{weight}} - f^0\|_n^2 + \lambda_{\text{init}}\lambda_{\text{weight}}\sum_{j\notin S} w_j|\hat{\beta}_{\text{weight},j}|/2$$

$$\leq \|f_{\beta_S} - f^0\|_n^2 + 3\lambda_{\text{init}}\lambda_{\text{weight}}\|w_S\|_2\|(\hat{\beta}_{\text{weight}})_S - \beta_S\|_2/2.$$

Case i) If

$$\|f_{\beta_S} - f^0\|_n^2 \leq 3\lambda_{\text{init}}\lambda_{\text{weight}}\|w_S\|_2\|(\hat{\beta}_{\text{weight}})_S - \beta_S\|_2/2,$$

we get

$$\|\hat{f}_{\text{weight}} - f^0\|_n^2 + \lambda_{\text{init}}\lambda_{\text{weight}}\sum_{j\notin S} w_j|\hat{\beta}_{\text{weight},j}|/2 \leq 3\lambda_{\text{init}}\lambda_{\text{weight}}\|w_S\|_2\|(\hat{\beta}_{\text{weight}})_S - \beta_S\|_2.$$

$$\tag{6.41}$$

It follows that

$$\|(\hat{\beta}_{\text{weight}})_{S^c}\|_1 \leq 6L\sqrt{s}\|(\hat{\beta}_{\text{weight}})_S - (\beta)_S\|_2.$$

But then

$$\|w_S\|_2\|(\hat{\beta}_{\text{weight}})_S - \beta_S\|_2$$

$$\leq \|w_S\|_2\|\hat{f}_{\text{weight}} - f_{\beta_S}\|_n/\phi_{\text{adap}}(6L,S,s)$$

$$\leq \|w_S\|_2\|\hat{f}_{\text{weight}} - f^0\|_n/\phi_{\text{adap}}(6L,S,s) + \|w_S\|_2\|f_{\beta_S} - f^0\|_n/\phi_{\text{adap}}(6L,S,s).$$

This gives

$$\|\hat{f}_{\text{weight}} - f^0\|_n^2 + \lambda_{\text{init}}\lambda_{\text{weight}}\sum_{j\notin S} w_j|\hat{\beta}_{\text{weight},j}|/2$$

$$\leq 3\lambda_{\text{init}}\lambda_{\text{weight}}\|w_S\|_2\|\hat{f}_{\text{weight}} - f^0\|_n/\phi_{\text{adap}}(6L,S,s)$$

$$+ 3\lambda_{\text{init}}\lambda_{\text{weight}}\|w_S\|_2\|f_{\beta_S} - f^0\|_n/\phi_{\text{adap}}(6L,S,s)$$

$$\leq \frac{1}{2}\|\hat{f}_{\text{weight}} - f^0\|_n^2 + \|f_{\beta_S} - f^0\|_n^2 + \frac{27\lambda_{\text{init}}^2\lambda_{\text{weight}}^2\|w_S\|_2^2}{4\phi_{\text{adap}}^2(6L,S,s)}.$$

Hence,

$$\|\hat{f}_{\text{weight}} - f^0\|_n^2 + \lambda_{\text{init}}\lambda_{\text{weight}}\sum_{j\notin S} w_j|\hat{\beta}_{\text{weight},j}| \leq 2\|f_{\beta_S} - f^0\|_n^2 + \frac{27\lambda_{\text{init}}^2\lambda_{\text{weight}}^2\|w_S\|_2^2}{2\phi_{\text{adap}}^2(6L,S,s)}.$$

We now apply $27/2 \leq 14$.

Case ii) If

$$\|\hat{f}_{\beta_S} - f^0\|_n^2 > \lambda_{\text{init}}\lambda_{\text{weight}}\|w_S\|_2\|(\hat{\beta}_{\text{weight}})_S - \beta_S\|_2,$$

we get

$$\|\hat{f}_{\text{weight}} - f^0\|_n^2 + \lambda_{\text{init}}\lambda_{\text{weight}} \sum_{j \notin S} w_j |\hat{\beta}_{\text{weight},j}| \leq 2\|f_{\beta_S} - f^0\|_n^2.$$

For the second result, we add in Case i), $\lambda_{\text{init}}\lambda_{\text{weight}}\|w_S\|_2\|(\hat{\beta}_{\text{weight}})_S - \beta_S\|_2/2$ to the left and right hand side of (6.10):

$$\|\hat{f}_{\text{weight}} - f^0\|_n^2 + \lambda_{\text{init}}\lambda_{\text{weight}}\|w_S\|_2\|(\hat{\beta}_{\text{weight}})_S - \beta_S\|_2/2 + \lambda_{\text{init}}\lambda_{\text{weight}} \sum_{j \notin S} w_j |\hat{\beta}_{\text{weight},j}|/2$$

$$\leq 4\lambda_{\text{init}}\lambda_{\text{weight}}\|w_S\|_2\|(\hat{\beta}_{\text{weight}})_S - \beta_S\|_2.$$

The same arguments now give

$$4\lambda_{\text{init}}\lambda_{\text{weight}}\|w_S\|_2\|(\hat{\beta}_{\text{weight}})_S - \beta_S\|_2 \leq$$

$$\|\hat{f}_{\text{weight}} - f^0\|_n^2 + 14\|f_{\beta_S} - f^0\|_n^2/3 + \frac{44\lambda_{\text{init}}^2\lambda_{\text{weight}}^2\|w_S\|^2}{7\phi_{\text{adap}}^2(6L,S,s)}.$$

In Case ii), we have

$$\lambda_{\text{init}}\lambda_{\text{weight}} \sum_{j \notin S} w_j |\hat{\beta}_{\text{weight},j}| \leq 4\|f_{\beta_S} - f^0\|_n^2,$$

and also

$$\lambda_{\text{init}}\lambda_{\text{weight}}\|w_S\|_2\|(\hat{\beta}_{\text{weight}})_S - \beta_S\|_2 < 2\|f_{\beta_S} - f^0\|_n^2/3.$$

So then

$$\lambda_{\text{init}}\lambda_{\text{weight}}\|w_S\|_2\|(\hat{\beta}_{\text{weight}})_S - \beta_S\|_2 + \lambda_{\text{init}}\lambda_{\text{weight}} \sum_{j \notin S} w_j |\hat{\beta}_{\text{weight},j}|$$

$$< 14\|f_{\beta_S} - f^0\|_n^2/3.$$

We now use $14/3 \leq 5$ and $44/7 \leq 7$. $\square$

6.11 Concave penalties

We study squared error loss and fixed design (for simplicity), but now with ℓ_r-penalty

$$\|\beta\|_r^r = \sum_{j=1}^p |\beta_j|^r,$$

where $0 \leq r \leq 1$ (in particular, $\|\beta\|_0^0 := \#\{\beta_j \neq 0\}$). We consider data $\{(X_i, Y_i)\}_{i=1}^n$, with response variable $Y_i \in \mathbf{R}$ and covariable $X_i \in \mathcal{X}$ $(i = 1, \ldots, n)$.

Let, again as in Section 6.3,

$$f_\beta := \sum_{j=1}^{p} \beta_j \psi_j(\cdot), \ \beta \in \mathbf{R}^p,$$

be real-valued linear functions on $\mathscr{X}$. The estimator with ℓ_r-penalty is

$$\hat{\beta} := \arg\min_\beta \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - f_\beta(X_i))^2 + \lambda^{2-r} \|\beta\|_r^r \right\},$$

where $0 \le r \le 1$ is fixed. We let $\hat{f} := f_{\hat{\beta}}$. Again, λ is a (properly chosen) regularization parameter. We will typically choose it of order $\lambda \asymp \sqrt{\log p/n}$. We show that the least squares estimator with concave penalty has oracle properties similar to the Lasso. The proofs for this section can be found in Subsection 6.11.2.

The idea is that the ℓ_r-penalized estimator ($r < 1$) is in some sense less biased than the Lasso. In particular, the ℓ_0-penalty is often considered as theoretically ideal, but unfortunately computationally infeasible. Also the ℓ_r-penalty with $0 < r < 1$ is computationally difficult. Moreover, its theoretical merits seem only to become apparent when looking at variable selection (see Section 7.13). Nevertheless, in this section we only treat the prediction error and ℓ_r-error. Our results are to be understood as showing that one does not loose in prediction error (but that there is possibly a gain in model selection). The results of this section only serve as a theoretical benchmark of what can be accomplished. In practice (for $0 < r < 1$), one may apply a one step approximation of the concave penalty by using an adaptive Lasso procedure (see e.g. Subsection 2.8)

$$\min_\beta \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - f_\beta(X_i))^2 + r\lambda^{2-r} \sum_{j=1}^{p} \frac{|\beta_j|}{|\hat{\beta}_{j,\text{init}}|^{1-r}} \right\},$$

where $\hat{\beta}_{\text{init}}$ is an initial estimate, obtained e.g. by the (standard ℓ_1-) Lasso. In other words, the concave penalty is connected to the adaptive Lasso, indeed indicating that it will be less biased than the (one stage, non-adaptive) Lasso. However, there may be a great discrepancy between the ℓ_r-penalized estimator, and adaptive (two stage) Lasso. We refer to Zhang (2010) for some important contributions for concave penalties, in particular rigorous theory for a related algorithm, and for the role of the initial estimates. Theory for the standardly weighted adaptive Lasso is provided in Chapter 7.

The true regression is

$$f^0(X_i) := \mathbb{E}Y_i,$$

and $\varepsilon_i := Y_i - f^0(X_i)$, $(i = 1, \ldots, n)$ denotes the measurement error. We let for $f : \mathscr{X} \to \mathbf{R}$,

$$\|f\|_n^2 := \frac{1}{n} \sum_{i=1}^{n} f^2(X_i),$$

and define the empirical inner product

$$(\varepsilon, f)_n := \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(x_i).$$

6.11.1 Sparsity oracle inequalities for least squares with ℓ_r-penalty

With some abuse of notation, we write for a positive, semi-definite $(p \times p)$-matrix,

$$\|f_\beta\|_\Sigma^2 := \beta^T \Sigma \beta, \ \beta \in \mathbb{R}^p.$$

Thus, $\|f_\beta\|_n = \|f_\beta\|_{\hat{\Sigma}}$.

Definition We say that the (Σ_0, ℓ_r)-compatibility condition *is satisfied for the set S, with constant* $\phi_{\Sigma_0, r}(S) > 0$, *if for all* $\beta \in \mathbb{R}^p$, *that satisfy* $\|\beta_{S^c}\|_r^r \le 3\|\beta_S\|_r^r$, *it holds that*

$$\|\beta_S\|_r^r \le \|f_\beta\|_{\Sigma_0}^r |S|^{\frac{2-r}{2}} / \phi_{\Sigma_0, r}^r(S). \tag{6.42}$$

Obviously, for $r = 0$, the (Σ_0, ℓ_r)-compatibility condition is trivially fulfilled for all S, with $\phi_{\Sigma_0, r}(S) = 1$. In this sense, the ℓ_0-penalty does not need any compatibility condition.

To go from the ℓ_r-world to the ℓ_2-world, we may use

Lemma 6.13. *We have for all index sets S,*

$$\|\beta_S\|_r^r \le |S|^{\frac{2-r}{2}} \|\beta\|_2^r.$$

For $0 < r \le 1$, the (Σ_0, ℓ_r)-compatibility condition on an index set S can be used to compare $\|\cdot\|_{\Sigma_0}$ and $\|\cdot\|_{\hat{\Sigma}}$. Define

$$\|\hat{\Sigma} - \Sigma_0\|_\infty := \max_{j,k} |(\hat{\Sigma})_{j,k} - (\Sigma_0)_{j,k}|,$$

The norms $\|\cdot\|_{\Sigma_0}$ and $\|\cdot\|_{\hat{\Sigma}}$ are close if, with $\tilde{\lambda} \ge \|\hat{\Sigma} - \Sigma_0\|_\infty$, the expression $\tilde{\lambda}|S|^{\frac{2-r}{r}} / \phi_{\Sigma_0, r}^2(S)$ is small. This is shown in the next lemma. This lemma is actually a generalization of Lemma 6.17 which is given in the next section, and which considers only the case $r = 1$.

Lemma 6.14. *Let* $0 < r \le 1$. *Suppose that the* (Σ_0, ℓ_r)-*compatibility condition holds for S, with compatibility constant* $\phi_{\Sigma_0, r}(S)$. *For all* $\beta \in \mathbb{R}^p$ *satisfying* $\|\beta_{S^c}\|_r^r \le 3\|\beta_S\|_r^r$,

$$\left|\frac{\|f_\beta\|_n^2}{\|f_\beta\|_{\Sigma_0}^2} - 1\right| \le 4^{\frac{2}{r}}\|\hat{\Sigma} - \Sigma_0\|_\infty |S|^{\frac{2-r}{r}}/\phi_{\Sigma_0,r}^2(S).$$

We now state the oracle inequality for the case of concave penalties. The result is a straightforward extension of Lemma 6.3. We assume to be on the set $\mathscr{T}$ given in (6.43), where the empirical process behaves well. In Corollary 14.7, it is shown that, under general conditions, $\mathscr{T}$ has large probability, for λ_0 of order $\sqrt{\log p/n}$.

Our definition of the oracle is a generalization of Subsection 6.2.3. Let S be an index set. Recall the projection $f_S = f_{b_S}$, in $L_2(Q_n)$, of f^0 on the linear space spanned by $\{\psi_j\}_{j \in S}$:

$$f_S := \arg\min_{f = f_{\beta_S}} \|f - f^0\|_n.$$

Definition of the oracle *We define the oracle as* $\beta^* := b^{S_*}$, *with*

$$S_* := \arg\min_{S \in \mathscr{S}}\left\{4\|f_S - f^0\|_n^2 + 12(9\lambda)^2 |S|^2/\phi_{\Sigma_0,r}^{\frac{2r}{2-r}}(S)\right\},$$

and set $f^* := f_{\beta^*}$, *and* $\phi_{*,r} := \phi_{\Sigma_0,r}(S_*)$.

Lemma 6.15. *(ℓ_r-penalty) Let $\mathscr{T}$ be the set*

$$\mathscr{T} := \left\{\sup_\beta \frac{|2(\varepsilon, f_\beta)_n|}{\|f_\beta\|_n^{\frac{2(1-r)}{2-r}}\|\beta\|_r^{\frac{r}{2-r}}} \le \lambda_0\right\}, \tag{6.43}$$

where for $r = 0$,

$$\|\beta\|_r^{\frac{r}{2-r}} := \sqrt{s_\beta}, \ s_\beta = \|\beta\|_0^0.$$

Suppose $\lambda^{2-r} \ge 5\lambda_0^{2-r}4^{1-r}$. Let the oracle minimize over the set

$$\mathscr{S} \subset \{S: 4^{\frac{2}{r}}\tilde{\lambda}|S|^{\frac{2-r}{2}}/\phi_{\Sigma_0,r}^2(S) \le 1/2\}.$$

We then have on $\mathscr{T} \cap \{\|\hat{\Sigma} - \Sigma_0\|_\infty \le \tilde{\lambda}\}$,

$$\|\hat{f} - f^0\|_n^2 + 4\lambda^{2-r}\|\hat{\beta} - \beta^*\|_r^r/5$$

$$\le 4\|f^* - f^0\|_n^2 + 12(9\lambda)^2|S_*|/\phi_{*,r}^{\frac{2r}{2-r}}.$$

6.11.2 Proofs for this section (Section 6.11)

We start with some simple inequalities (compare with our calculations of convex conjugates, see Section 6.4).

Lemma 6.16. *Let u and v be positive numbers. Then for $0 \le r \le 1$,*

$$u^r v \le u^2 + \left(\frac{r}{2}\right)^{\frac{r}{2-r}} \left(1 - \frac{r}{2}\right) v^{\frac{2}{2-r}} \le u^2 + v^{\frac{2}{2-r}}.$$

Moreover,

$$u^{\frac{2(1-r)}{2-r}} v^{\frac{r}{2-r}} \le u^2 + \frac{(1-r)^{1-r}}{(2-r)^{2-r}} v^r \le u^2 + v^r.$$

Proof. For $r = 0$, the results are trivial. For $0 < r \le 1$, it holds that

$$\max_u \left(uv - u^{2/r} \right) = \left(\frac{r}{2}\right)^{\frac{r}{2-r}} \left(1 - \frac{r}{2}\right) v^{\frac{2}{2-r}},$$

which implies the first result.

The second result follows from replacing r by $\frac{2(1-r)}{2-r}$ and v by $v^{\frac{r}{2-r}}$. $\qquad \square$

Corollary 6.7. *For all $0 \le r \le 1$ and $c > 0$, it holds that*

$$u^r v \le cu^2 + \frac{v^{\frac{2}{2-r}}}{c^{\frac{r}{2-r}}},$$

and

$$u^{\frac{2(1-r)}{2-r}} v^{\frac{r}{2-r}} \le cu^2 + \frac{v^r}{c^{1-r}}.$$

Proof of Lemma 6.13. For $r = 0$, the result is trivial. For $0 < r \le 1$, let $p := 2/r$, and

$$\frac{1}{q} := 1 - \frac{1}{p} = \frac{2-r}{2}.$$

Then, by Hölder's inequality, for any sequence $\{a_j\}$ of numbers

$$\sum_{j \in S} |a_j| \le |S|^{\frac{1}{q}} \|a\|_p.$$

Apply this with $a_j = |\beta_j|^r$. $\qquad \square$.

Proof of Lemma 6.14. For all β,

$$\left| \|f_\beta\|_{\hat{\Sigma}}^2 - \|f_\beta\|_{\Sigma_0}^2 \right| = |\beta^T \hat{\Sigma} \beta - \beta^T \Sigma_0 \beta|$$

$$= |\beta^T (\hat{\Sigma} - \Sigma_0)\beta| \le \tilde{\lambda} \|\beta\|_1^2 \le \tilde{\lambda} \|\beta\|_r^2,$$

since $0 < r \le 1$. If $\|\beta_{S^c}\|_r^r \le 3\|\beta_S\|_r^r$, we also have

$$\|\beta\|_r^2 \le 4^{\frac{2}{r}} \|\beta_S\|_r^2 \le 4^{\frac{2}{r}} |S|^{\frac{2-r}{r}} \|f_\beta\|_{\Sigma_0}^2 / \phi_{\Sigma_0, r}^2(S).$$

$$\square.$$

Proof of Lemma 6.15. Clearly, on $\mathscr{T}$, by the Basic Inequality,

$$\|\hat{f} - f^0\|_n^2 + \lambda^{2-r}\|\hat{\beta}\|_r^r \le \lambda_0 \|\hat{f} - f^*\|_n^{\frac{2(1-r)}{2-r}} \|\hat{\beta} - \beta^*\|_r^{\frac{r}{2-r}} + \lambda^{2-r}\|\beta^*\|_r^r + \|f^* - f^0\|_n^2$$

$$\le \frac{1}{4}\|\hat{f} - f_{\beta^*}\|_n^2 + \lambda_0^{2-r}\|\hat{\beta} - \beta^*\|_r^r 4^{1-r} + \lambda^{2-r}\|\beta^*\|_r^r + \|f^* - f^0\|_n^2.$$

We now invoke that for $0 \le r \le 1$, and all β and $\tilde{\beta}$, and for all S,

$$\|\beta_S\|_r^r - \|\tilde{\beta}_S\|_r^r \le \|\beta_S - \tilde{\beta}_S\|_r^r.$$

This gives

$$\frac{3}{4}\|\hat{f} - f^0\|_n^2 + (\lambda^{2-r} - 4^{1-r}\lambda_0^{2-r})\|\hat{\beta}_{S_*^c}\|_r^r$$

$$\le (\lambda^{2-r} + \lambda_0^{2-r}4^{1-r})\|\hat{\beta}_{S_*} - \beta_{S_*}^*\|_r^r + \|f^* - f^0\|_n^2.$$

Case i) Whenever

$$(\lambda^{2-r} + \lambda_0^{2-r}4^{1-r})\|\hat{\beta}_{S_*} - \beta_{S_*}^*\|_r^r \ge \|f^* - f^0\|_n^2,$$

we have

$$\frac{3}{4}\|\hat{f} - f^0\|_n^2 + (\lambda^{2-r} - 4^{1-r}\lambda_0^{2-r})\|\hat{\beta}_{S_*^c}\|_r^r \le 2(\lambda^{2-r} + \lambda_0^{2-r}4^{1-r})\|\hat{\beta}_{S_*} - \beta_{S_*}^*\|_r^r.$$

But, as $\lambda^{2-r} \ge 5\lambda_0^{2-r}4^{1-r}$,

$$\frac{2(\lambda^{2-r} + \lambda_0^{2-r}4^{1-r})}{\lambda^{2-r} - 4^{1-r}\lambda_0^{2-r}} \le 3,$$

so then we may apply the (Σ_0, ℓ_r)-compatibility condition, or actually, its implied $(\hat{\Sigma}, \ell_r)$-compatibility condition. This gives on $\mathscr{T} \cap \{\|\hat{\Sigma} - \Sigma_0\|_\infty \le \tilde{\lambda}\}$,

$$\frac{3}{4}\|\hat{f} - f^0\|_n^2 + (\lambda^{2-r} - 4^{1-r}\lambda_0^{2-r})\|\hat{\beta} - \beta^*\|_r^r \le 3(\lambda^{2-r} + \lambda_0^{2-r}4^{1-r})\|\hat{\beta}_{S_*} - \beta_{S_*}^*\|_r^r$$

$$\le 2^{\frac{r}{2}}(\lambda^{2-r} + \lambda_0^{2-r}4^{1-r})|S_*|^{\frac{2-r}{2}}\|\hat{f}^* - f^*\|_n^r / \phi_{*,r}^r$$

because $S_* \in \mathscr{S}$ implies $\phi_{\hat{\Sigma},r}^2(S_*) \ge 2\phi_{*,r}^2$. Application of Corollary 6.7 gives

$$\frac{3}{4}\|\hat{f} - f^0\|_n^2 + (\lambda^{2-r} - 4^{1-r}\lambda_0^{2-r})\|\hat{\beta} - \beta^*\|_r^r$$

$$\le \frac{1}{2}\|f^* - f_0\|_n^2 + 2 \times 2^{\frac{r}{2-r}}(3(\lambda^{2-r} + \lambda_0^{2-r}4^{1-r}))^{\frac{2}{2-r}}|S_*|/\phi_{*,r}^{\frac{2r}{2-r}}$$

$$+ \|f^* - f^0\|_n^2 + 2^{\frac{r}{2-r}}(3(\lambda^{2-r} + \lambda_0^{2-r}4^{1-r}))^2|S_*|/\phi_{*,r}^{\frac{2r}{2-r}}.$$

So then

$$\|\hat{f} - f^0\|_n^2 + 4(\lambda^{2-r} - 4^{1-r}\lambda_0^{2-r})\|\hat{\beta} - \beta^*\|_r^r$$

$$\leq 4\|f^* - f^0\|_n^2 + 12 \times 2^{\frac{r}{2-r}}(3(\lambda^{2-r} + \lambda_0^{2-r}4^{1-r}))^{\frac{2}{2-r}}|S_*|/\phi_{*,r}^{\frac{2r}{2-r}}$$

$$\leq 4\|f^* - f^0\|_n^2 + 12(9\lambda)^2|S_*|/\phi_{*,r}^{\frac{2r}{2-r}},$$

where we use

$$3(\lambda^{2-r} + \lambda_0^{2-r}4^{1-r}) \leq 3(\lambda^{2-r} + \lambda^{2-r}/5) = 18\lambda^{2-r}/5,$$

and

$$2^{\frac{r}{2-r}}(18/5)^{\frac{2}{2-r}} \leq 9^2.$$

Case ii) If on the other hand,

$$(\lambda^{2-r} + \lambda_0^{2-r}4^{1-r})\|\hat{\beta}_{S_*} - \beta_{S_*}^*\|_r^r < \|f^* - f^0\|_n^2,$$

we get

$$\frac{3}{4}\|\hat{f} - f^0\|_n^2 + (\lambda^{2-r} - 4^{1-r}\lambda_0^{2-r})\|\hat{\beta}_{S_*^c}\|_r^r \leq 2\|f^* - f^0\|_n^2,$$

so also

$$\frac{3}{4}\|\hat{f} - f^0\|_n^2 + (\lambda^{2-r} - 4^{1-r}\lambda_0^{2-r})\|\hat{\beta}^* - \beta^*\|_r^r \leq 3\|f^* - f^0\|_n^2,$$

and hence

$$\|\hat{f} - f^0\|_n^2 + (\lambda^{2-r} - 4^{1-r}\lambda_0^{2-r})\|\hat{\beta}^* - \beta^*\|_r^r \leq 4\|f^* - f^0\|_n^2.$$

$\square$

6.12 Compatibility and (random) matrices

As in Section 6.3, let $(\mathbf{F}, \|\cdot\|)$ be a normed space of real-valued functions on $\mathcal{Z}$, and $\mathscr{F} := \{f_\beta = \sum_{j=1}^p \beta_j \psi_j : \beta \in \mathbb{R}^p\} \subset \mathbf{F}$ be a linear subset. Observe that both the margin condition (see Section 6.4), as well as the compatibility condition, depend on the norm $\|\cdot\|$. It is generally the $L_2(P)$-norm (or, in regression problems, the $L_2(Q)$-norm where Q is the marginal distribution of the co-variables), but also other norms may be of interest. In this section, we consider the situation where $\|f_\beta\|^2 = \beta^T \Sigma_0 \beta := \|f_\beta\|_{\Sigma_0}^2$ is a quadratic form in β. The matrix Σ_0 is some $p \times p$ symmetric, positive semi-definite matrix. To stress the dependence on Σ_0, we call the compatibility condition for this norm the Σ_0-*compatibility condition*.

We will show that when two matrices Σ_0 and Σ_1 are "close" to each other, the Σ_0-compatibility condition implies the Σ_1-compatibility condition. This is particularly useful when Σ_0 is a population covariance matrix Σ and Σ_1 is its sample variant $\hat{\Sigma}$. This allows then an easy switch from the Lasso with least squares loss and fixed design, to random design. For convex Lipschitz loss, whether or not the design is fixed or random plays a less prominent role.

Of particular interest are the choices $\Sigma_0 = \Sigma$, where Σ is the $p \times p$ Gram matrix

$$\Sigma = \int \psi^T \psi dP, \; \psi := (\psi_1, \ldots, \psi_p).$$

Also of interest is the choosing the empirical version $\Sigma_0 = \hat{\Sigma}$, where

$$\hat{\Sigma} := \int \psi^T \psi dP_n,$$

where P_n is the empirical distribution of n observations $Z_1, \ldots, Z_n$. One may also consider a normalized version $\Sigma_0 := \hat{R}$, with $\hat{R} := \mathrm{diag}(\hat{\Sigma})^{-1/2} \hat{\Sigma} \, \mathrm{diag}(\hat{\Sigma})^{-1/2}$.

Definition *We say that the Σ_0-compatibility condition is met for the set S, with constant $\phi_{\Sigma_0}(S) > 0$, if for all $\beta \in \mathbf{R}^p$, that satisfy $\|\beta_{S^c}\|_1 \leq 3\|\beta_S\|_1$, it holds that*

$$\|\beta_S\|_1^2 \leq \|f_\beta\|_{\Sigma_0}^2 |S| / \phi_{\Sigma_0}^2(S).$$

For two (positive semi-definite) matrices Σ_0 and Σ_1, we define the supremum distance

$$\|\Sigma_1 - \Sigma_0\|_\infty := \max_{j,k} |(\Sigma_1)_{j,k} - (\Sigma_0)_{j,k}|,$$

Lemma 6.17. *Suppose that the Σ_0-compatibility condition holds for the set S with cardinality s, with compatibility constant $\phi_{\Sigma_0}(S)$. Assume*

$$\|\Sigma_1 - \Sigma_0\|_\infty \leq \tilde{\lambda}.$$

Then for all β satisfying $\|\beta_{S^c}\|_1 \leq 3\|\beta_S\|_1$,

$$\left| \frac{\|f_\beta\|_{\Sigma_1}^2}{\|f_\beta\|_{\Sigma_0}^2} - 1 \right| \leq 16\tilde{\lambda}s / \phi_{\Sigma_0}^2(S).$$

Proof of Lemma 6.17. For all β,

$$\left| \|f_\beta\|_{\Sigma_1}^2 - \|f_\beta\|_{\Sigma_0}^2 \right| = |\beta^T \Sigma_1 \beta - \beta^T \Sigma_0 \beta|$$

$$= |\beta^T (\Sigma_1 - \Sigma_0)\beta| \leq \tilde{\lambda} \|\beta\|_1^2.$$

But if $\|\beta_{S^c}\|_1 \leq 3\|\beta_S\|_1$, we also have by the Σ_0-compatibility condition,

$$\|\beta\|_1 \leq 4\|\beta_S\|_1 \leq 4\|f_\beta\|_{\Sigma_0}\sqrt{s}/\phi_{\Sigma_0}(S).$$

This gives

$$\left| \|f_\beta\|_{\Sigma_1}^2 - \|f_\beta\|_{\Sigma_0}^2 \right| \leq 16\tilde{\lambda}\|f_\beta\|_{\Sigma_0}^2 s/\phi_{\Sigma_0}^2(S).$$

$\square$.

Corollary 6.8. *Suppose that the Σ_0-compatibility condition holds for the set S with cardinality s, with compatibility constant $\phi_{\Sigma_0}(S)$, and that $\|\Sigma_1 - \Sigma_0\|_\infty \leq \tilde{\lambda}$, where*

$$32\tilde{\lambda}s/\phi_{\Sigma_0}^2(S) \leq 1. \tag{6.44}$$

Then, for the set S, the Σ_1-compatibility condition holds as well, with $\phi_{\Sigma_1}^2(S) \geq \phi_{\Sigma_0}^2(S)/2$. Moreover, for all β satisfying $\|\beta_{S^c}\|_1 \leq 3\|\beta_S\|_1$, we have

$$\|f_\beta\|_{\Sigma_0}^2 \leq 2\|f_\beta\|_{\Sigma_1}^2 \leq 3\|f_\beta\|_{\Sigma_0}^2.$$

A similar statement can be made for the (adaptive) (L,S,N)-restricted eigenvalue condition, as considered in Section 6.8 and Section 6.10. (Section 6.13 gathers the various conditions.) See Problem 6.10 for a general statement.

We remark that for the case where Σ_1 is the sample covariance matrix $\hat{\Sigma}$ of a sample from a population with covariance matrix Σ, the results can be refined (e.g. in the sub-Gaussian case), leading to a major relaxation of (6.44). The details can be found in Zhou (2009a).

Asymptotics In the case $\Sigma_0 = \Sigma$, and $\Sigma_1 = \hat{\Sigma}$, and under (moment) conditions, one has $\|\Sigma_1 - \Sigma_0\|_\infty \leq \tilde{\lambda}$ with large probability, where $\tilde{\lambda}$ is of order $\sqrt{\log p/n}$ (see Problem 14.3). We conclude that if

$$s/\phi_\Sigma^2(S) = O\left(\sqrt{\frac{n}{\log p}}\right), \tag{6.45}$$

then the metrics $\|\cdot\|_\Sigma$ and $\|\cdot\|_{\hat{\Sigma}}$ show similar behavior on the set of functions f_β with $\|\beta_{S^c}\|_1 \leq 3\|\beta_S\|_1$.

Let us now look at some further implications of Corollary 6.8 for the Lasso. Let, for some measure Q_1,

$$\Sigma_1 := \int \psi^T \psi dQ_1.$$

By definition, $\|f_\beta\|_{\Sigma_1}^2 = \beta^T \Sigma_1 \beta$. We now also use the same notation $\|\cdot\|_{\Sigma_1}$ for the $L_2(Q_1)$-norm, so that $\|f\|_{\Sigma_1}$ is defined for all $f \in \mathbf{F}$. This is not completely correct, as Σ_1 generally does not characterize the $L_2(Q_1)$-norm, but we believe confusion is not likely.

Recall that the excess risk is $\mathscr{E}(f) := P(f - f^0)$ (see Section 6.3). We introduce a set $\mathbf{F}_{\text{local}} \subset \mathbf{F}$ as in Section 6.4.

Definition *We say that the Σ_1-margin condition holds with strictly convex function G, if for all $f \in \mathbf{F}_{\text{local}}$, we have*

$$\mathscr{E}(f) \geq G(\|f - f^0\|_{\Sigma_1}).$$

Suppose the Σ_1-margin condition holds, and the Σ_0-compatibility condition for a certain collection of sets S. Let

$$\mathscr{S} \subseteq \{S : 32\tilde{\lambda}|S|/\phi^2_{\Sigma_0}(S) \leq 1\}.$$

When $\|\Sigma_1 - \Sigma_0\|_\infty \leq \tilde{\lambda}$, we know from Corollary 6.8 that the Σ_1-compatibility condition also holds for all $S \in \mathscr{S}$. Let us recall some definitions, now with the implied bound for the Σ_1-compatibility constant.

Set, as in Section 6.7,

$$b^S := \arg\min_{\beta=\beta_S} \mathscr{E}(f_\beta), \; f_S := f_{b^S}.$$

Let

$$S_* := \arg\min_{S \in \mathscr{S}} \left\{ 3\mathscr{E}(f_S) + 2H\left(\frac{4\sqrt{2}\lambda\sqrt{|S|}}{\phi_{\Sigma_0}(S)}\right) \right\}.$$

$$\beta^* := b^{S_*}, \; f^* := f_{\beta^*},$$

$$\varepsilon^* := 3\mathscr{E}(f^*)/2 + H\left(\frac{4\sqrt{2}\lambda\sqrt{|S_*|}}{\phi_{\Sigma_0}(S_*)}\right), \; M^* := \varepsilon^*/\lambda_0.$$

An immediate corollary of Theorem 6.4 is now:

Corollary 6.9. *Assume the Σ_1-margin condition with strictly convex function G, with convex conjugate H. Take*

$$\mathscr{S} \subseteq \{S \in \mathscr{S}_0 : 32\tilde{\lambda}|S|/\phi^2_{\Sigma_0}(S) \leq 1\}.$$

Assume $f_\beta \in \mathbf{F}_{\text{local}}$ for all $\|\beta - \beta^*\|_1 \leq M^*$, as well as $f^* \in \mathbf{F}_{\text{local}}$, and that $\lambda \geq 8\lambda_0$. Let $\mathscr{T} := \{\mathbf{Z}_{M^*} \leq \lambda_0 M^*\}$. Then on the set $\mathscr{T} \cap \{\|\Sigma_1 - \Sigma_0\|_\infty \leq \tilde{\lambda}\}$, we have

$$\mathscr{E}(\hat{f}) + \lambda\|\hat{\beta} - \beta^*\|_1 \leq 4\varepsilon^* = 6\mathscr{E}(f^*) + 4H\left(\frac{4\sqrt{2}\lambda\sqrt{|S_*|}}{\phi_{\Sigma_0}(S_*)}\right).$$

If $f^0 = f_{\beta^0}$, and if, for $S_0 := S_{\beta^0}$, it holds that $32\tilde{\lambda}|S_0|\phi^2_{\Sigma_0}(S_0) \leq 1$, then we have under the above assumptions with $\mathscr{S} = \{S_0\}$, that on $\mathscr{T} \cap \{\|\Sigma_1 - \Sigma_0\|_\infty \leq \tilde{\lambda}\}$,

$$\mathcal{E}(\hat{f}) \le 4\varepsilon^0 := 4H\left(\frac{4\sqrt{2}\lambda\sqrt{|S_0|}}{\phi_{\Sigma_0}(S_0)}\right).$$

Also, then on $\mathcal{T} \cap \{\|\Sigma_1 - \Sigma_0\|_\infty \le \tilde{\lambda}\}$,

$$\left| \|\hat{f} - f^0\|_{\Sigma_0}^2 - \|\hat{f} - f^0\|_{\Sigma_1}^2 \right| \le \tilde{\lambda}\left(\frac{4\varepsilon^0}{\lambda}\right)^2.$$

The assumption $32\tilde{\lambda}|S_0|/\phi_{\Sigma_0}^2(S_0) \le 1$ is related to the possibility to consistently estimate β^0 in ℓ_1-norm. Corollary 6.9 implies that

$$\|\hat{\beta} - \beta^0\|_1 \le 4\varepsilon^0/\lambda.$$

Assume now for simplicity that the margin behavior is quadratic, say $G(u) = cu^2$. Then the convex conjugate is $H(v) = v^2/(4c)$, and we find

$$\varepsilon^0/\lambda := \frac{8\lambda|S_0|}{c\phi_{\Sigma_0}^2(S_0)}.$$

With λ being of order $\sqrt{\log p/n}$, the condition $|S_0|/\phi_{\Sigma_0}^2(S_0) = O(\sqrt{n/\log p})$ is needed for a value ε^0/λ that remains bounded.

In fact, a bounded (actually "small") value for ε^0/λ, or more generally for $M^* = \varepsilon^*/\lambda_0$, will also be helpful to verify the assumptions. We assumed in Corollary 6.9 (and similarly in Theorem 6.4), that

$$f_\beta \in \mathbf{F}_{\text{local}} \ \forall \ \|\beta - \beta^*\|_1 \le M^*. \tag{6.46}$$

If the base functions are bounded, say $\|\psi_j\|_\infty \le K$ for all j, one easily verifies that $\|f_\beta - f_{\beta^*}\|_\infty \le K\|\beta - \beta^*\|_1$. Thus, if $\mathbf{F}_{\text{local}}$ is an L_∞-neighborhood of f^0, and the oracle f_{β^*} is L_∞-close enough to f^0, then condition (6.46) will be met when M^* is small enough. A detailed illustration of this line of reasoning is given in Example 6.4 (where, in particular, we require (6.34)).

Example 6.5. **Application to squared error loss**

Consider quadratic loss

$$\rho_f(\cdot, y) = (y - f(\cdot))^2.$$

Let $Q^{(i)}$ be the distribution of X_i, $Q := \sum_{i=1}^n Q^{(i)}/n$, and Q_n the empirical distribution based on $X_1, \ldots, X_n$. Define the Gram matrices

$$\Sigma := \int \psi^T \psi dQ, \ \hat{\Sigma} := \int \psi^T \psi dQ_n.$$

For $i = 1, \ldots, n$, let the target $f^0(X_i) := E(Y_i|X_i)$ be the conditional expectation of Y_i given X_i.

We write (for $i = 1, \ldots, n$), $\varepsilon_i = Y_i - f^0(X_i)$, and as before (Example 6.1), for $f : \mathcal{X} \to \mathbb{R}$,

$$(\varepsilon, f)_n := \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(X_i),$$

and we let $\| \cdot \|_n$ ($\| \cdot \|$) be the $L_2(Q_n)$- ($L_2(Q)$-)norm. Thus,

$$\|f_\beta\|_n = \|f_\beta\|_{\hat{\Sigma}}, \ \|f_\beta\| = \|f_\beta\|_{\Sigma}.$$

We assume $f^0 = f_{\beta^0}$ is linear (or consider the estimation of the linear projection). The (within sample) prediction error of a regression function f is $\|f - f^0\|_{\hat{\Sigma}}^2$ (for the case of fixed design) or $\|f - f^0\|_{\Sigma}^2$ (for the case of random design). More generally, we may be interested in out-of-sample prediction error, or *transductive* situations (Vapnik (2000)). This means one is interested in

$$\|f - f^0\|_{\Sigma_0}^2,$$

where Σ_0 is some positive definite matrix possibly other than $\hat{\Sigma}$ or Σ.

As for any design, the $\hat{\Sigma}$-margin condition holds, we obtain as consequence of Corollary 6.9:

Corollary 6.10. *Assume the Σ_0-compatibility condition for S_0 for some Σ_0. Let $\tilde{\lambda}$ satisfy $32\tilde{\lambda}|S_0|/\phi_{\Sigma_0}^2(S_0) \leq 1$. Let*

$$\mathcal{T} := \left\{ \max_j 2|(\varepsilon, \psi_j)_n| \leq \lambda_0 \right\}.$$

Then on $\mathcal{T} \cap \{\|\hat{\Sigma} - \Sigma_0\|_\infty \leq \tilde{\lambda}\}$, and for $\lambda \geq 8\lambda_0$, we have

$$\|\hat{f} - f^0\|_n^2 + \lambda \|\hat{\beta} - \beta^0\|_1 \leq \frac{32\lambda^2|S_0|}{\phi_{\Sigma_0}^2(S_0)},$$

and

$$\left| \|\hat{f} - f^0\|_{\Sigma_0}^2 - \|\hat{f} - f^0\|_n^2 \right| \leq \frac{32\lambda\tilde{\lambda}|S_0|}{\phi_{\Sigma_0}^2(S_0)}.$$

Note that the straightforward application of the rather general Corollary 6.9 resulted in somewhat bigger constants than in Theorem 6.1, whose proof is tailored for the situation considered.

One may alternatively directly apply Theorem 6.4 to general design, as the Σ-margin condition holds as well. Note however that the set $\mathcal{T}$ then becomes

$$\mathcal{T} = \left\{ \sup_{\|\beta - \beta^0\|_1 \leq M^*} \left| 2(\varepsilon, f_\beta - f^0)_n \right. \right.$$

$$+\|f_\beta - f^0\|_n^2 - \|f_\beta - f^0\|^2 \bigg| \leq \lambda_0 M^* \bigg\}.$$

We will treat this set in Section 14.9.

6.13 On the compatibility condition

The compatibility condition we discuss here partly follows van de Geer (2007), but we also present some extensions. For a further discussion, we refer to van de Geer and Bühlmann (2009).

Let $\mathcal{X}$ be some measurable space, P be a probability measure on $\mathcal{X}$, $\|\cdot\|$ be the $L_2(P)$ norm, and

$$\mathcal{F} := \left\{ f_\beta(\cdot) = \sum_{j=1}^p \beta_j \psi_j(\cdot) : \ \beta \in \mathbb{R}^p \right\}$$

be a linear subspace of $L_2(P)$. The Gram matrix is

$$\Sigma := \int \psi^T \psi dP,$$

so that

$$\|f_\beta\|^2 = \beta^T \Sigma \beta := \|f_\beta\|_\Sigma^2.$$

The entries of Σ are denoted by $\sigma_{j,k} := (\psi_j, \psi_k)$, with $(\cdot, \cdot)$ being the inner product in $L_2(P)$.

To clarify the notions we shall use below, consider for a moment a partition of the form

$$\Sigma := \begin{pmatrix} \Sigma_{1,1} & \Sigma_{1,2} \\ \Sigma_{2,1} & \Sigma_{2,2} \end{pmatrix},$$

where $\Sigma_{1,1}$ is an $s \times s$ matrix, $\Sigma_{2,1}$ is a $(p-s) \times s$ matrix and $\Sigma_{1,2} := \Sigma_{2,1}^T$ is its transpose, and where $\Sigma_{2,2}$ is a $(p-s) \times (p-s)$ matrix.

Such partitions will play an important role in this section. We need them for general index sets S with cardinality s, sets which are not necessarily the first s indices $\{1, \ldots, s\}$. We introduce the $s \times s$ matrix

$$\Sigma_{1,1}(S) := (\sigma_{j,k})_{j,k \in S},$$

the $(p-s) \times s$ matrix

$$\Sigma_{2,1}(S) = (\sigma_{j,k})_{j \notin S, k \in S},$$

and the $(p-s) \times (p-s)$ matrix

$$\Sigma_{2,2}(S) := (\sigma_{j,k})_{j,k \notin S}.$$

We let $\Lambda_{\max}^2(\Sigma_{1,1}(S))$ and $\Lambda_{\min}^2(\Sigma_{1,1}(S))$ be the largest and smallest eigenvalue of $\Sigma_{1,1}(S)$ respectively. Throughout, we assume that $\Lambda_{\min}^2(\Sigma_{1,1}(S)) > 0$, i.e., that $\Sigma_{1,1}(S)$ is non-singular.

We consider an index set $S \subset \{1,\ldots,p\}$, with cardinality $s := |S|$. With β being a vector in $\mathbb{R}^p$, we denote by

$$\beta_{j,S} := \beta_j \mathrm{l}\{j \in S\}, \ j = 1,\ldots,p,$$

the vector with only non-zero entries in the set S. We will sometimes identify β_S with $\{\beta_j\}_{j \in S} \in \mathbb{R}^s$.

For a vector v, we invoke the usual notation

$$\|v\|_q = \begin{cases} (\sum_j |v_j|^q)^{1/q}, & 1 \le q < \infty \\ \max_j |v_j|, & q = \infty \end{cases}.$$

We introduce *minimal ℓ_1-eigenvalues*, and re-introduce the compatibility constant.

Definition *The minimal ℓ_1-eigenvalue of $\Sigma_{1,1}(S)$ is*

$$\Lambda_{\min,1}^2(\Sigma_{1,1}(S)) := \min\left\{ \frac{s\beta_S^T \Sigma_{1,1}(S)\beta_S}{\|\beta_S\|_1^2} \ : \ \|\beta_S\|_1 \ne 0 \right\}.$$

Let $L > 0$ be some constant. The (L,S)-compatibility constant is

$$\phi_{\mathrm{comp}}^2(L,S) := \min\left\{ \frac{s\beta^T \Sigma\beta}{\|\beta_S\|_1^2} \ : \ \|\beta_{S^c}\|_1 \le L\|\beta_S\|_1 \ne 0 \right\}.$$

By this definition, for all $\|\beta_{S^c}\|_1 \le L\|\beta_S\|_1 \ne 0$,

$$\|\beta_S\|_1^2 \le s\|f_\beta\|^2 / \phi_{\mathrm{comp}}^2(L,S).$$

In the previous sections, the constant L was chosen as $L = 3$ for ease of exposition.[6] For applications to the Lasso, this may be replaced by some other value bigger than one, but then one has to adjust the choice of the regularization parameter.

Definition
The (L,S)-compatibility condition is satisfied for the set S, if $\phi_{\mathrm{comp}}(L,S) > 0$.

This definition is as in the previous sections. However, we now define $\phi_{\mathrm{comp}}(L,S)$ as the largest possible constant for which the compatibility condition holds.

---

[6] If λ is the regularization parameter, and $\lambda_0 < \lambda$ is the noise level we used in the set $\mathscr{T}$ of Subsection 6.2.3, then one may replace $L = 3$ by $L = (2\lambda + \lambda_0)/(\lambda - \lambda_0)$. The latter expression is bounded by 3 when one assumes $\lambda \ge 4\lambda_0$. (In Subsection 6.2.2, we took $\beta^* = \beta^0$, $\lambda \ge 2\lambda_0$ and $L = (\lambda + \lambda_0)/(\lambda - \lambda_0)$.)

In this section, we will present conditions for the compatibility condition to hold. The organization is as follows. We first present some direct bounds for the compatibility constant in Subsection 6.13.1. We show that the compatibility constant is a minimizer of a Lasso problem.

In Subsection 6.13.2, we replace, in the definition of the compatibility constant, the ℓ_1-norm $\|\beta_S\|_1$ of β_S by its bound $\sqrt{s}\|\beta_S\|_2$, and call the result the (adaptive) restricted eigenvalue. The conditions are then similar to the restricted eigenvalue condition of Bickel et al. (2009). We impose conditions on the length of the projection of $-f_{\beta_{S^c}}$ on f_{β_S}. Subsection 6.13.3 finally uses conditions from Candès and Tao (2007), and also some extensions. It considers supsets $\mathcal{N}$ of S, and places restrictions on the minimal eigenvalues of $\Sigma_{1,1}(\mathcal{N})$, and on the elements of $\Sigma_{1,2}(\mathcal{N})$ (uniformly in all $\mathcal{N} \supset S$ with cardinality at most some given value N).

6.13.1 Direct bounds for the compatibility constant

A first, rather trivial observation is that if $\Sigma = I$, the compatibility condition holds for all L and S, with $\phi_{\text{comp}}(L,S) = 1$. The case $\Sigma = I$ corresponds to uncorrelated variables. If there is correlation, this may be due to some common underlying latent variables. Because this will become important later on, we present this simple situation in a lemma (see also Problem 6.14).

Lemma 6.18. *Suppose that*

$$\Sigma = \Sigma_0 + \tilde{\Sigma},$$

where Σ_0 and $\tilde{\Sigma}$ are both positive semi-definite. If the $\|\cdot\|_{\Sigma_0}$-compatibility condition holds for S with constant $\phi_{\text{comp},\Sigma_0}(L,S)$, then also the $\|\cdot\|_{\Sigma}$-compatibility holds for S with constant $\phi_{\text{comp},\Sigma}(L,S) \geq \phi_{\text{comp},\Sigma_0}(L,S)$.

Proof of Lemma 6.18. This is clear, as

$$\|f_\beta\|_{\Sigma}^2 = \beta^T \Sigma \beta \geq \beta^T \Sigma_0 \beta^T = \|f_\beta\|_{\Sigma_0}$$

$\square$

An important special case is

$$\Sigma_0 = (I - \Theta),$$

where $\Theta = \text{diag}(\theta_1, \ldots, \theta_p)$, with $0 \leq \theta_j \leq 1$ for all j. Then $\phi_{\text{comp},\Sigma_0}^2(L,S) \geq \Lambda_{\min}^2(\Sigma_0) = 1 - \max_{k \in S} \theta_k$.

After normalizing by $|S|$, the larger S, the smaller the compatibility constant (and hence the harder the compatibility condition).

Lemma 6.19. *For $S \supset S°$,*

$$\phi_{\text{comp}}^2(L,S)/|S| \leq \phi_{\text{comp}}^2(L,S^\circ)/|S^\circ|.$$

Proof of Lemma 6.19. Suppose that $\|\beta_{(S^\circ)^c}\|_1 \leq L\|\beta_{S^\circ}\|_1$. Then $\|\beta_{S^c}\|_1 \leq \|\beta_{(S^\circ)^c}\|_1 \leq L\|\beta_{S^\circ}\|_1 \leq L\|\beta_S\|_1$. Hence, for all $\|\beta_{(S^\circ)^c}\|_1 \leq L\|\beta_{S^\circ}\|_1$,

$$\phi_{\text{comp}}(L,S)/\sqrt{|S|} = \min\left\{ \frac{\|f_{\tilde{\beta}}\|}{\|\tilde{\beta}_S\|_1} : \|\tilde{\beta}_{S^c}\|_1 \leq L\|\tilde{\beta}_S\|_1 \neq 0 \right\} \leq \frac{\|f_\beta\|}{\|\beta_S\|_1} \leq \frac{\|f_\beta\|}{\|\beta_{S^\circ}\|_1}.$$

But then also

$$\phi_{\text{comp}}(L,S)/\sqrt{|S|} \leq \min\left\{ \frac{\|f_\beta\|}{\|\beta_{S^\circ}\|_1} : \quad \|\beta_{(S^\circ)^c}\|_1 \leq L\|\beta_{S^\circ}\|_1 \neq 0 \right\}$$

$$= \phi_{\text{comp}}(L,S^\circ)/\sqrt{|S^\circ|}.$$

$\square$

Compatibility constants may be smaller than the ℓ_1-eigenvalues of the matrix $\Sigma_{1,1}(S)$, as the following lemma shows.

Lemma 6.20. *We have*

$$\phi_{\text{comp}}^2(L,S) \leq \Lambda_{\min,1}^2(\Sigma_{1,1}(S)).$$

Proof of Lemma 6.20. This follows from the obvious fact that for all β, $(\beta_S)_{S^c} \equiv 0$, so certainly $\|(\beta_S)_{S^c}\|_1 \leq L\|\beta_S\|_1$. Hence

$$\frac{s\|f_{\beta_S}\|^2}{\|\beta_S\|_1^2} \geq \phi_{\text{comp}}^2(L,S).$$

$\square$

The question is of course: how can we calculate $\phi_{\text{comp}}^2(L,S)$? It turns out that this question is related to how the Lasso behaves as deterministic approximation scheme.

Let us define, for $f \in L_2(P)$ and $L > 0$,

$$\text{LASSO}(f,L,S) := \min_{\|\beta_S\|_1 \leq L} \|f_{\beta_S} - f\|^2.$$

The restricted minimum can be derived using Lagrange calculus. Let $\lambda \geq 0$ be the Lagrange parameter, and set

$$\beta_S(\lambda) := \arg\min_\beta \left\{ \|f_{\beta_S} - f\|^2 + \lambda\|\beta_S\|_1 \right\}.$$

By duality, there exists a value λ_L such that

$$\text{LASSO}(f,L,S) = \|f_{\beta_{S,\text{primal}}} - f\|^2, \ \beta_{S,\text{primal}} := \beta_S(\lambda_L).$$

Lemma 6.21. *The (L,S)-compatibility constant is the solution of a Lasso problem, namely*

$$\phi_{\text{comp}}^2(L,S)/s = \min_{\|\beta_S\|_1=1} \text{LASSO}(-f_{\beta_S},L,S^c).$$

Lemma 6.21 does not provide us with explicit lower bounds for the compatibility constant. This is the reason why we present further bounds in the subsections to come.

The lemma illustrates that the problem of finding the compatibility constant is at least as hard as finding the Lasso approximation. We actually hope that the Lasso approximation $\text{LASSO}(-f_{\beta_S},L,S^c)$ is not very good, i.e., that we cannot approximate $-f_{\beta_S}$ very well by a function $f_{\beta_{S^c}}$ with $\|\beta_{S^c}\|_1 \leq L$.

Proof of Lemma 6.21. We have

$$\min_{\|\beta_{S_c}\|_1 \leq L \|\beta_S\|_1 \neq 0} \frac{\|f_\beta\|^2}{\|\beta_S\|_1^2} = \min_{\|\beta_{S^c}\|_1 \leq L \|\beta_S\|_1 \neq 0} \frac{\|f_{\beta_{S^c}} + f_{\beta_S}\|^2}{\|\beta_S\|_1^2}$$

$$= \min_{\|\beta_{S^c}\|_1 \leq L \|\beta_S\|_1 \neq 0} \|f_{\beta_{S^c}/\|\beta_S\|_1} + f_{\beta_S/\|\beta_S\|_1}\|^2 = \min_{\|\beta_S\|_1=1} \min_{\|\beta_{S^c}\|_1 \leq L} \|f_{\beta_{S^c}} + f_{\beta_S}\|^2$$

$$= \min_{\|\beta_S\|_1=1} \text{LASSO}(-f_{\beta_S},L,S^c).$$

$\square$

Let us now have a closer look at the minimal ℓ_1-eigenvalue. Consider an $s \times s$ symmetric, positive definite matrix $\Sigma_{1,1}$. One easily checks that

$$\Lambda_{\min,1}^2(\Sigma_{1,1}) \geq \Lambda_{\min}^2(\Sigma_{1,1}).$$

The lower bound $\Lambda_{\min}^2(\Sigma_{1,1})$ can generally be improved, as the following lemma shows.

Lemma 6.22. *Let $\Sigma_{1,1}$ be some $s \times s$ symmetric, positive definite matrix. Then for some vector τ satisfying $\|\tau\|_\infty \leq 1$,*

$$\Lambda_{\min,1}^2(\Sigma_{1,1}) = \frac{\tau^T \Sigma_{1,1}^{-1} \tau}{s \|\Sigma_{1,1}^{-1} \tau\|_1^2}.$$

More precisely, a solution of

$$b := \arg \min_{\|b\|_1=1} b^T \Sigma_{1,1} b.$$

is $b = \Sigma_{1,1}^{-1}\tau/\|\Sigma_{1,1}^{-1}\tau\|_1$, $\text{sign}(b_j) = \tau_j \mathbb{1}\{|b_j| \neq 0\}$, $j = 1,\ldots,s$.

Proof of Lemma 6.22. Introduce a Lagrange parameter $\lambda \in \mathbb{R}$. Then the problem

$$\min_{\|b\|_1=1} b^T \Sigma_{1,1} b$$

is equivalent to minimizing, for a suitable λ,

$$\min\{b^T \Sigma_{1,1} b + 2\lambda \|b\|_1\}.$$

By the KKT conditions (see Lemma 2.1), for $b_j \neq 0$,

$$(\Sigma_{1,1}b)_j + \lambda \operatorname{sign}(b_j) = 0,$$

and for $b_j = 0$,

$$|(\Sigma_{1,1}b)_j| \leq |\lambda|.$$

Therefore, there exists a vector τ with $\|\tau\|_\infty \leq 1$ such that

$$\Sigma_{1,1} b = -\lambda \tau,$$

and $\operatorname{sign}(b_j) = \tau_j$ if $b_j \neq 0$, and $b_j = 0$ if $|\tau_j| < 1$. It follows that

$$b^T \Sigma_{1,1} b = -\lambda \|b\|_1.$$

So, because $\|b\|_1 = 1$, we have $\lambda = -b^T \Sigma_{1,1} b < 0$. Furthermore,

$$b = -\lambda \Sigma_{1,1}^{-1} \tau.$$

It follows that

$$\lambda = -1/\|\Sigma_{1,1}^{-1}\tau\|_1.$$

So we may take the solution

$$b = \Sigma_{1,1}^{-1}\tau / \|\Sigma_{1,1}^{-1}\tau\|_1.$$

We now have

$$b^T \Sigma_{1,1} b = \tau^T \Sigma_{1,1}^{-1}\tau / \|\Sigma_{1,1}^{-1}\tau\|_1^2.$$

$\square$

6.13.2 Bounds using $\|\beta_S\|_1^2 \leq s\|\beta_S\|_2^2$

Let S be a set with cardinality s. In the compatibility condition, one may replace $\|\beta_S\|_1$ by its ℓ_2-bound $\sqrt{s}\|\beta_S\|_2$. This leads to the following definitions.

Definition *The (L,S,s)-restricted eigenvalue of Σ is*

$$\phi^2(L,S,s) := \min\left\{\frac{\|f_\beta\|^2}{\|\beta_S\|_2^2} \; : \; \|\beta_{S^c}\|_1 \le L\|\beta_S\|_1 \ne 0\right\}.$$

The adaptive $(L,S,|S|)$-restricted eigenvalue *of Σ is*

$$\phi_{\mathrm{adap}}^2(L,S,s) := \min\left\{\frac{\|f_\beta\|^2}{\|\beta_S\|_2^2} \; : \; \|\beta_{S^c}\|_1 \le L\sqrt{s}\|\beta_S\|_2 \ne 0\right\}.$$

Definition
We say that the (L,S,s)-restricted eigenvalue condition is met if $\phi(L,S,s) > 0$. The adaptive (L,S,s)-restricted eigenvalue condition holds if $\phi_{\mathrm{adap}}(L,S,s) > 0$.

The restricted eigenvalue condition from Bickel et al. (2009) assumes the (L,S,s)-compatibility condition for **all** S with size s.

It is clear that $\phi_{\mathrm{adap}}^2(L,S,s) \le \phi^2(L,S,s) \le \phi_{\mathrm{comp}}^2(L,S)$, i.e.,

   adaptive $(L,S,|S|)$-restricted eigenvalue condition $\Rightarrow$

   (L,S,s)-restricted eigenvalue condition $\Rightarrow$

   (L,S)-compatibility condition.

We first present a simple lower bound, for the case where Σ is non-singular.

Lemma 6.23. *Suppose that Σ has smallest eigenvalue $\Lambda_{\min}^2(\Sigma) > 0$. Then the adaptive $(L,S,|S|)$-restricted eigenvalue condition holds for all index sets $S \subset \{1,\dots,p\}$, with $\phi_{\mathrm{adap}}(L,S,|S|) \ge \Lambda_{\min}(\Sigma)$.*

Proof of Lemma 6.23. Let $\beta \in \mathbf{R}^p$ be arbitrary. It is clear that

$$\|\beta\|_2 \le \|f_\beta\|/\Lambda_{\min}(\Sigma).$$

The result now follows from the trivial inequality $\|\beta_S\|_2 \le \|\beta\|_2$.   □

Lemma 6.24. *It holds that*

$$\phi_{\mathrm{adap}}^2(L,S,s) = \min_{\|\beta_S\|_2=1} \mathrm{LASSO}(-f_{\beta_S}, L\sqrt{s}, S^c).$$

Proof of Lemma 6.24. We have

$$\min_{\|\beta_{S_c}\|_1 \le L\sqrt{s}\|\beta_S\|_2 \ne 0} \frac{\|f_\beta\|^2}{\|\beta_S\|_2^2} = \min_{\|\beta_{S^c}\|_1 \le L\sqrt{s}\|\beta_S\|_2 \ne 0} \|f_{\beta_{S^c}/\|\beta_S\|_2} + f_{\beta_S/\|\beta_S\|_2}\|^2$$

$$= \min_{\|\beta_S\|_2=1} \min_{\|\beta_{S^c}\|_1 \le L\sqrt{s}} \|f_{\beta_{S^c}} + f_{\beta_S}\|^2$$

$$= \min_{\|\beta_S\|_2=1} \mathrm{LASSO}(-f_{\beta_S}, L\sqrt{s}, S^c).$$

$\square$

One may verify that, as with the compatibility constants,

$$\phi(L,S,|S|) \leq \phi(L,S^\circ,|S^\circ|)$$

for $S \supset S^\circ$. The same is true for adaptive restricted eigenvalues.

Next, we show that our assumption $\Lambda^2_{\min}(\Sigma_{1,1}(S)) > 0$ is not restrictive, in the context of the present subsection.

Lemma 6.25. *We have*

$$\phi^2(L,S,s) \leq \Lambda^2_{\min}(\Sigma_{1,1}(S)).$$

The proof is Problem 6.11.

Let f_1 and f_2 by two functions in $L_2(P)$. The next lemma assumes that the regression of f_2 on f_1 is strictly larger than -1. This rules out the possibility to cancel out f_1 by f_2.

Lemma 6.26. *Suppose for some $0 < \vartheta < 1$.*

$$-(f_1,f_2) \leq \vartheta \|f_1\|^2.$$

Then

$$(1 - \vartheta)\|f_1\| \leq \|f_1 + f_2\|.$$

Proof. Write the projection of f_2 on f_1 as

$$f^{\mathrm{P}}_{2,1} := (f_2, f_1)/\|f_1\|^2 f_1.$$

Similarly, let

$$(f_1 + f_2)^{\mathrm{P}}_1 := (f_1 + f_2, f_1)/\|f_1\|^2 f_1$$

be the projection of $f_1 + f_2$ on f_1. Then

$$(f_1 + f_2)^{\mathrm{P}}_1 := f_1 + f^{\mathrm{P}}_{2,1} = \left(1 + (f_2,f_1)/\|f_1\|^2\right)f_1,$$

so that

$$\|(f_1 + f_2)^{\mathrm{P}}_1\| = \left|1 + (f_2,f_1)/\|f_1\|^2\right|\|f_1\|$$

$$= \left(1 + (f_2,f_1)/\|f_1\|^2\right)\|f_1\| \geq (1 - \vartheta)\|f_1\|$$

Moreover, by Pythagoras' Theorem

$$\|f_1 + f_2\|^2 \geq \|(f_1 + f_2)^{\mathrm{P}}_1\|^2.$$

□

This result leads to the definition of the (adaptive) S-restricted regression, given below. At this stage however, it is important to recall Lemma 6.18. If

$$\Sigma = \Sigma_0 + \tilde{\Sigma}, \tag{6.47}$$

with both Σ_0 and $\tilde{\Sigma}$ positive semi-definite, then for verifying the compatibility condition, or (adaptive) restricted eigenvalue condition, one may replace throughout Σ by any of the two matrices Σ_0 or $\tilde{\Sigma}$, say Σ_0. A special case is $\Sigma_0 = (I - \Theta)$, where $\Theta = \text{diag}(\theta_1, \ldots, \theta_p)$, $0 \leq \theta_j \leq 1$. The compatibility conditions and it variants are very easy to verify for the diagonal matrix $(I - \Theta)$. However, if we then look at the inner products $\Sigma_{1,2}(S)$ (as we essentially do when studying the (adaptive) restricted regression defined below), this only involves the matrix $\tilde{\Sigma}$. This point should be kept in mind throughout the rest of this section: lower bounds for compatibility constants or (adaptive) restricted eigenvalues that are based on the inner products in $\Sigma_{1,2}(S)$ can be unnecessarily pessimistic.

Definition *The S-restricted regression is*

$$\vartheta(S) := \max_{\|\beta_{S^c}\|_1 \leq \|\beta_S\|_1} \frac{|(f_{\beta_S}, f_{\beta_{S^c}})|}{\|f_{\beta_S}\|^2}.$$

The adaptive S-restricted regression *is*

$$\vartheta_{\text{adap}}(S) := \max_{\|\beta_{S^c}\|_1 \leq \sqrt{s}\|\beta_S\|_2} \frac{|(f_{\beta_S}, f_{\beta_{S^c}})|}{\|f_{\beta_S}\|^2}.$$

Corollary 6.11. *Suppose that* $\vartheta(S) < 1/L$ *($\vartheta_{\text{adap}}(S) < 1/L$). Then the (L, S, s)-restricted eigenvalue (adaptive (L, S, s)-restricted eigenvalue) condition holds, with* $\phi(L, S, s) \geq (1 - L\vartheta(S))\Lambda_{\min}(\Sigma_{1,1}(S))$ *($\phi_{\text{adap}}(L, S, s) \geq (1 - L\vartheta_{\text{adap}}(S)\Lambda_{\min}(\Sigma_{1,1}(S)))$).*

It will be shown in Theorem 7.3 of Section 7.5.5, that for all $\|\tau_S\|_\infty \leq 1$,

$$\|\Sigma_{2,1}(S)\Sigma_{1,1}^{-1}(S)\tau_S\|_\infty \leq \vartheta_{\text{adap}}(S).$$

Hence, bounds for $\phi_{\text{adap}}(L, S, s)$, (with $L < 1$), based on the above corollary, actually imply the irrepresentable condition.

There is always an upper bound for $\vartheta_{\text{adap}}(S)$. This bound will play its role for proving variable selection of the adaptive Lasso under general conditions.

Lemma 6.27. *If $\|\psi_j\| \leq 1$ for all j, we have*

$$\vartheta_{\text{adap}}(S) \leq \frac{\sqrt{s}}{\Lambda_{\min}(\Sigma_{1,1}(S))}.$$

Corollary 6.12. *Suppose* $\|\psi_j\| \leq 1$ *for all* j. *Then for* $2L \leq \sqrt{s}/\Lambda_{\min}(\Sigma_{1,1}(S))$,

$$\phi_{\mathrm{adap}}(L,S,s) \geq \Lambda_{\min}(\Sigma_{1,1}(S))/2.$$

Proof of Lemma 6.27. We have for $\|\beta_{S^c}\|_1 \leq \sqrt{s}\|\beta_S\|_2$,

$$\|f_{\beta_{S^c}}\| = \|\sum_{j \notin S} \beta_j \psi_j\| \leq \sum_{j \notin S} |\beta_j|\|\psi_j\|$$

$$= \|\beta_{S^c}\|_1 \leq \sqrt{s}\|\beta_S\|_2$$

$$\leq \sqrt{s}\|f_{\beta_S}\|/\Lambda_{\min}(\Sigma_{1,1}(S)).$$

This implies

$$\frac{|(f_{\beta_S}, f_{\beta_{S^c}})|}{\|f_{\beta_S}\|^2} \leq \frac{\|f_{\beta_{S^c}}\|}{\|f_{\beta_S}\|} \leq \frac{\sqrt{s}}{\Lambda_{\min}(\Sigma_{1,1}(S))}.$$

$\square$

We now present some further bounds on the S-restricted regression $\vartheta(S)$ and its adaptive version. For this purpose, we introduce some matrix norms. Let $1 \leq q \leq \infty$, and r be its conjugate, i.e.,

$$\frac{1}{q} + \frac{1}{r} = 1.$$

Define

$$\|\Sigma_{1,2}(S)\|_{\infty,q} := \max_{\|\beta_{S^c}\|_r \leq 1} \|\Sigma_{1,2}(S)\beta_{S^c}\|_\infty,$$

and

$$\|\Sigma_{1,2}(S)\|_{2,q} := \max_{\|\beta_{S^c}\|_r \leq 1} \|\Sigma_{1,2}(S)\beta_{S^c}\|_2.$$

In this subsection, we actually only consider the case $q = \infty$, i.e., the quantities

$$\|\Sigma_{1,2}(S)\|_{\infty,\infty} := \max_{\|\beta_{S^c}\|_1 \leq 1} \|\Sigma_{1,2}(S)\beta_{S^c}\|_\infty,$$

and

$$\|\Sigma_{1,2}(S)\|_{2,\infty} := \max_{\|\beta_{S^c}\|_1 \leq 1} \|\Sigma_{1,2}(S)\beta_{S^c}\|_2.$$

The case where q is taken to be finite is studied in the next subsection.

Lemma 6.28. *We have the upper bound*

$$\vartheta(S) \leq \frac{\|\Sigma_{1,2}(S)\|_{\infty,\infty}}{\Lambda_{\min,1}^2(\Sigma_{1,1}(S))}.$$

Similarly,

$$\vartheta_{\mathrm{adap}}(S) \leq \frac{\sqrt{s}\|\Sigma_{1,2}(S)\|_{2,\infty}}{\Lambda_{\min}^2(\Sigma_{1,1}(S))}. \tag{6.48}$$

Remember that $\Lambda^2_{\min,1}(\Sigma_{1,1}(S)) \geq \Lambda^2_{\min}(\Sigma_{1,1}(S))/s$. It is moreover clear that

$$\|\Sigma_{1,2}(S)\|_{2,\infty} \leq \sqrt{s}\|\Sigma_{1,2}(S)\|_{\infty,\infty} \leq \sqrt{s}\max_{j\notin S}\max_{k\in S}|\sigma_{j,k}|.$$

In addition (see Problem 6.12),

$$\|\Sigma_{1,2}(S)\|_{2,\infty} \leq \max_{j\notin S}\sqrt{\sum_{k\in S}\sigma^2_{j,k}} \leq \sqrt{s}\max_{j\notin S}\max_{k\in S}|\sigma_{j,k}|.$$

The consequences are in the spirit of the maximal local coherence condition in Bunea et al. (2007c).

Corollary 6.13. *(Coherence with $q = \infty$) Assume that*

$$\frac{\sqrt{s}\max_{j\notin S}\sqrt{\sum_{k\in S}\sigma^2_{j,k}}}{\Lambda^2_{\min}(\Sigma_{1,1}(S))} \leq \theta < 1/L.$$

Then $\phi_{\mathrm{adap}}(L,S,|S|) \geq (1-L\theta)\Lambda_{\min}(\Sigma_{1,1}(S))$.

Proof of Lemma 6.28.

$$\max_{\|\beta_{S^c}\|_1 \leq \|\beta_S\|_1} \frac{|(f_{\beta_S}, f_{\beta_{S^c}})|}{\|f_{\beta_S}\|^2} = \max_{\|\beta_{S^c}\|_1 \leq \|\beta_S\|_1} \frac{|(f_{\beta_S}, f_{\beta_{S^c}})|}{\|\beta_S\|^2_1} \frac{\|\beta_S\|^2_1}{\|f_{\beta_S}\|^2}$$

$$\leq \max_{\|\beta_{S^c}\|_1 \leq \|\beta_S\|_1} \frac{|(f_{\beta_S}, f_{\beta_{S^c}})|}{\|\beta_S\|^2_1}/\Lambda^2_{\min,1}(\Sigma_{1,1}(S))$$

$$= \max_{\|\beta_{S^c}\|_1 \leq 1} \|\Sigma_{1,2}(S)\beta_{S^c}\|_\infty/\Lambda^2_{\min,1}(\Sigma_{1,1}(S)).$$

$$= \|\Sigma_{1,2}(S)\|_{\infty,\infty}/\Lambda^2_{\min,1}(\Sigma_{1,1}(S)).$$

The second result we derive similarly:

$$\max_{\|\beta_{S^c}\|_1 \leq \sqrt{s}\|\beta_S\|_2} \frac{|(f_{\beta_S}, f_{\beta_{S^c}})|}{\|f_{\beta_S}\|^2} = \max_{\|\beta_{S^c}\|_1 \leq \sqrt{s}\|\beta_S\|_2} \frac{|(f_{\beta_S}, f_{\beta_{S^c}})|}{\|\beta_S\|^2_2} \frac{\|\beta_S\|^2_2}{\|f_{\beta_S}\|^2}$$

$$\leq \max_{\|\beta_{S^c}\|_1 \leq \sqrt{s}\|\beta_S\|_2} \frac{|(f_{\beta_S}, f_{\beta_{S^c}})|}{\|\beta_S\|^2_2}/\Lambda^2_{\min}(\Sigma_{1,1}(S))$$

$$= \max_{\|\beta_{S^c}\|_1 \leq \sqrt{s}} \|\Sigma_{1,2}(S)\beta_{S^c}\|_2/\Lambda^2_{\min}(\Sigma_{1,1}(S)).$$

$$= \sqrt{s}\|\Sigma_{1,2}(S)\|_{2,\infty}/(\Lambda^2_{\min}(\Sigma_{1,1}(S))).$$

$\square$

Example 6.6. Let $S = \{1, \ldots, s\}$ be the active set, and suppose that

$$\Sigma := \begin{pmatrix} \Sigma_{1,1} & \Sigma_{1,2} \\ \Sigma_{2,1} & \Sigma_{2,2} \end{pmatrix},$$

and in fact, that for some $0 < \rho < 1$,

$$\Sigma := \begin{pmatrix}
1 & 0 & \cdots & 0 & \rho/\sqrt{s} & 0 & \cdots & 0 \\
0 & 1 & \cdots & 0 & \rho/\sqrt{s} & 0 & \cdots & 0 \\
0 & 0 & \ddots & 0 & \vdots & \vdots & & \vdots \\
0 & 0 & \cdots & 1 & \rho/\sqrt{s} & 0 & \cdots & 0 \\
\rho/\sqrt{s} & \rho/\sqrt{s} & \cdots & \rho/\sqrt{s} & 1 & 0 & \cdots & 0 \\
0 & 0 & \cdots & 0 & 0 & 1 & \cdots & 0 \\
\vdots & \vdots & & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 1
\end{pmatrix}.$$

It can easily be shown that the compatibility condition holds, with $\phi_{\mathrm{comp}}(L, S) \geq 1 - \rho$. Moreover, for $\rho > 1/\sqrt{s}$, the irrepresentable condition (defined in Subsection 7.5.1) does not hold (see also Problem 7.2). Because $\|f_{\beta_S}\| = \|\beta_S\|_2$, and $\Lambda_{\min}(\Sigma_{1,1}) = 1$, we moreover have

$$\vartheta_{\mathrm{adap}}(S) = \sqrt{s}\|\Sigma_{1,2}\|_{2,\infty},$$

i.e., the bounds of Lemma 6.27 and (6.48) are achieved in this example. If fact, $\|\Sigma_{1,2}\|_{2,\infty} = \rho$, so

$$\vartheta_{\mathrm{adap}}(S) = \sqrt{s}\rho.$$

6.13.3 Sets $\mathcal{N}$ containing S

Recall (6.47), i.e., one may first want to replace Σ by some simpler Σ_0 if possible.

To derive rates for the ℓ_q-norm with $1 < q \leq 2$ (see Section 6.8), we needed stronger versions of the (L, S)-compatibility condition. Let us cite these versions here.

For $\mathcal{N} \supset S$, we have introduced the restricted set

$$\mathcal{R}(L, S, \mathcal{N}) := \left\{ \|\beta_{S^c}\|_1 \leq L\|\beta_S\|_1, \ \|\beta_{\mathcal{N}^c}\|_\infty \leq \min_{j \in \mathcal{N} \setminus S} |\beta_j| \right\}.$$

We complemented this with an adaptive version

$$\mathcal{R}_{\mathrm{adap}}(L, S, \mathcal{N}) := \left\{ \|\beta_{S^c}\|_1 \leq L\sqrt{s}\|\beta_S\|_2, \ \|\beta_{\mathcal{N}^c}\|_\infty \leq \min_{j \in \mathcal{N} \setminus S} |\beta_j| \right\},$$

and a minimal adaptive version

$$\mathcal{R}_{\min}(L,S,\mathcal{N}) := \left\{ \|\beta_{\mathcal{N}^c}\|_1 \le L\sqrt{N}\|\beta_{\mathcal{N}}\|_2, \ \|\beta_{\mathcal{N}^c}\|_\infty \le \min_{j \in \mathcal{N}\setminus S}|\beta_j| \right\},$$

Definition *Let S be an index set with cardinality s and $N \ge s$ be an integer. We say that the (L,S,N)-restricted eigenvalue condition holds, with constant $\phi(L,S,N)$, if*

$$\phi(L,S,N) := \min\left\{ \frac{\|f_\beta\|}{\|\beta_{\mathcal{N}}\|_2} : \ \beta \in \mathcal{R}(L,S,\mathcal{N}), \ \mathcal{N} \supset S, \ |\mathcal{N}| = N \right\} > 0. \quad (6.49)$$

We say that the adaptive (L,S,N)-restricted eigenvalue condition *holds, with constant $\phi_{\mathrm{adap}}(L,S,N)$, if*

$$\phi_{\mathrm{adap}}(L,S,N) := \min\left\{ \frac{\|f_\beta\|}{\|\beta_{\mathcal{N}}\|_2} : \ \beta \in \mathcal{R}_{\mathrm{adap}}(L,S,\mathcal{N}), \ \mathcal{N} \supset S, \ |\mathcal{N}| = N \right\} > 0.$$
$$(6.50)$$

We say that the minimal adaptive (L,S,N)-restricted eigenvalue condition *holds, with constant $\phi_{\min}(L,S,N)$, if*

$$\phi_{\min}(L,S,N) := \min\left\{ \frac{\|f_\beta\|}{\|\beta_{\mathcal{N}}\|_2} : \ \beta \in \mathcal{R}_{\min}(L,S,\mathcal{N}), \ \mathcal{N} \supset S, \ |\mathcal{N}| = N \right\} > 0.$$
$$(6.51)$$

The constants $\phi(L,S,N)$, $\phi_{\mathrm{adap}}(L,S,N)$ and $\phi_{\min}(L,S,N)$ will again be be referred to as ((minimal) adaptive) restricted eigenvalues. Clearly (for $N \ge s$),

$$\phi_{\mathrm{comp}}(L,S) \ge \phi(L,S,s) \ge \phi(L,S,N)$$

and similarly

$$\phi_{\mathrm{comp}}(L,S) \ge \phi_{\mathrm{adap}}(L,S,s) \ge \phi_{\mathrm{adap}}(L,S,N),$$

and

$$\phi_{\mathrm{comp}}(L,S) \ge \phi_{\min}(L,S,s) \ge \phi_{\min}(L,S,N).$$

Furthermore,

$$\phi(L,S,N) \ge \phi_{\mathrm{adap}}(L,S,N) \ge \phi_{\min}(L,S,N).$$

We recall that, apart for the case where we assume the truth itself is sparse, our Lasso results for the ℓ_2-error mainly needed lower bounds for $\phi_{\min}(L,S,N)$ (see Section 6.8).

Definition *For $N \ge s$, the (L,S,N)-restricted regression is*

$$\vartheta(L,S,N) := \max_{\mathcal{N} \supset S, \ |\mathcal{N}| = N} \max_{\beta \in \mathcal{R}(L,S,\mathcal{N})} \frac{|(f_{\beta_{\mathcal{N}}}, f_{\beta_{\mathcal{N}^c}})|}{\|f_{\beta_{\mathcal{N}}}\|^2}.$$

The adaptive (L,S,N)-restricted regression *is*

$$\vartheta_{\text{adap}}(L,S,N) := \max_{\mathcal{N} \supset S,\ |\mathcal{N}|=N} \max_{\beta \in \mathscr{R}_{\text{adap}}(L,S,\mathcal{N})} \frac{|(f_{\beta_{\mathcal{N}}}, f_{\beta_{\mathcal{N}^c}})|}{\|f_{\beta_{\mathcal{N}}}\|^2}.$$

The minimal adaptive (L,S,N)-regression *is*

$$\vartheta_{\min}(L,S,N) := \max_{\mathcal{N} \supset S,\ |\mathcal{N}|=N} \max_{\beta \in \mathscr{R}_{\min}(L,S,\mathcal{N})} \frac{|(f_{\beta_{\mathcal{N}}}, f_{\beta_{\mathcal{N}^c}})|}{\|f_{\beta_{\mathcal{N}}}\|^2}.$$

Note that

$$\vartheta(L,S,s) = L\vartheta(S),$$

and similarly for the (minimal) adaptive variant. For $N > s$, this scaling no longer holds.

Definition *For $N \geq s$, the (S,N)-uniform eigenvalue is*

$$\Lambda_{\min}(S,N) := \min_{\mathcal{N} \supset S,\ |\mathcal{N}|=N} \Lambda_{\min}(\Sigma_{1,1}(\mathcal{N})).$$

One easily verifies that, similar to Corollary 6.11, for $\vartheta(L,S,N) < 1$,

$$\phi(L,S,N) \geq (1 - \vartheta_{\text{adap}}(L,S,N))\Lambda_{\min}(S,N).$$

Similarly,

$$\phi_{\text{adap}}(L,S,N) \geq (1 - \vartheta_{\text{adap}}(L,S,N))\Lambda_{\min}(S,N). \tag{6.52}$$

and

$$\phi_{\min}(L,S,N) \geq (1 - \vartheta_{\min}(L,S,N))\Lambda_{\min}(S,N).$$

6.13.4 Restricted isometry

The *restricted isometry property* (Candès and Tao (2005) or Candès and Tao (2007)) is the condition

$$\delta_s + \theta_{s,s} + \theta_{s,2s} < 1,$$

where δ_s (assumed to be in $[0,1)$) is the smallest value such that for all S with $|S| = s$, and all β

$$(1 - \delta_s)\|\beta_S\|_2^2 \leq \|f_{\beta_S}\|^2 \leq (1 + \delta_s)\|\beta_S\|_2^2,$$

and for $N \geq s$, $\theta_{s,N}$ is the smallest value such that for all $\mathcal{N}$ with $|\mathcal{N}| \leq N$, all $\mathcal{M} \subset \mathcal{N}^c$ with $|\mathcal{M}| \leq s$ and all β

$$\frac{|(f_{\beta_{\mathcal{M}}}, f_{\beta_{\mathcal{N}}})|}{\|\beta_{\mathcal{M}}\|_2 \|\beta_{\mathcal{N}}\|_2} \leq \theta_{s,N}.$$

The constants δ_s are called *restricted isometry constants*, and the $\theta_{s,N}$ are called *orthogonality constants*. It is clear that

$$1 - \delta_N \leq \Lambda_{\min}^2(S,N).$$

Candès and Tao (2005) show that

$$\delta_{2s} \leq \delta_s + \theta_{s,s}.$$

In Koltchinskii (2009b); Bickel et al. (2009), it is shown that

$$\vartheta_{\mathrm{adap}}(1,S,2s) \leq \frac{\theta_{s,2s}}{1 - \delta_s - \theta_{s,s}}.$$

Hence, the restricted isometry property implies that $\vartheta_{\mathrm{adap}}(1,S,2s) < 1$.

6.13.5 Sparse eigenvalues

Sparse eigenvalues can play an important role for the variable selection problem, see Subsection 7.8 and 10.5.1. They are similar to the restricted isometry constants. We present a relation between the adaptive restricted regression and sparse eigenvalue conditions. This leads to a bound for the adaptive restricted eigenvalue (see Corollary 6.14), and hence also for the restricted eigenvalue (the latter result is as in Bickel et al. (2009)), and for the compatibility constant.

Definition *For a given $N \in \{1, \ldots, p\}$, the* maximal sparse eigenvalue *is*

$$\Lambda_{\max}(N) := \max_{|\mathcal{N}|=N} \Lambda_{\max}(\Sigma_{1,1}(\mathcal{N})).$$

We also recall the (S,N) uniform eigenvalue

$$\Lambda_{\min}(S,N) := \min_{\mathcal{N} \supset S, \, |\mathcal{N}|=N} \Lambda_{\min}(\Sigma_{1,1}(\mathcal{N})).$$

A quantity that one also encounters in literature is the *the minimal sparse eigenvalue*

$$\Lambda_{\min}(N) := \min_{|\mathcal{N}|=N} \Lambda_{\min}(\Sigma_{1,1}(\mathcal{N})).$$

We will however base our results on the uniform eigenvalue, and not on the sparse minimal eigenvalue. Note that

$$\Lambda_{\min}(S,N) \geq \Lambda_{\min}(N) \; \forall \; |S| \leq N.$$

Lemma 6.29. *Let S have cardinality s and let $N \geq s$. We have*

$$\vartheta_{\text{adap}}(L,S,N) \leq \frac{L\Lambda_{\max}^2(2N-s)}{\Lambda_{\min}^2(S,N)}\sqrt{\frac{s}{N-s}},$$

and

$$\vartheta_{\min}(L,S,N) \leq \frac{(L+1)\Lambda_{\max}^2(2N-s)}{\Lambda_{\min}^2(S,N)}\sqrt{\frac{N}{N-s}},$$

Proof. Fix some $\beta \in \mathbb{R}^p$ that satisfies $\|\beta_{S^c}\|_1 \leq L\sqrt{s}\|\beta_S\|_2$. Let $\mathcal{N} := \mathcal{N}_0 \supset S$ be the set which has $\mathcal{N}\setminus S$ as the $N-s$ largest coefficients $|\beta_j|$, $j \notin S$. Let for $k = 1,\dots,$ $\mathcal{N}_k$ be the set of $N-s$ largest coefficients $|\beta_j|$, $j \notin \mathcal{N}_{k-1}$. Then by Lemma 6.9

$$\sum_{k\geq 1} \|\beta_{\mathcal{N}_k}\|_2 \leq \|\beta_{S^c}\|_1/\sqrt{N-s}$$

$$\leq L\|\beta_S\|_2\sqrt{\frac{s}{N-s}} \leq L\|\beta_{\mathcal{N}}\|_2\sqrt{\frac{s}{N-s}}.$$

It is moreover not difficult to see that for all $k \geq 1$,

$$\frac{|(f_{\beta_{\mathcal{N}_k}},f_{\beta_{\mathcal{N}}})|}{\|\beta_{\mathcal{N}_k}\|_2\|\beta_{\mathcal{N}}\|_2} \leq \Lambda_{\max}^2(2N-s)$$

(we use: $|(f_1,f_2)| \leq \|f_1+f_2\|^2 \vee \|f_1-f_2\|^2$ for any two functions f_1 and f_2). It follows that

$$|(f_{\beta_{\mathcal{N}^c}},f_{\beta_{\mathcal{N}}})| \leq \sum_{k\geq 1}|(f_{\beta_{\mathcal{N}_k}},f_{\beta_{\mathcal{N}}})| \leq L\Lambda_{\max}^2(2N-s)\|\beta_{\mathcal{N}}\|_2^2\sqrt{\frac{s}{N-s}}$$

$$\leq L\Lambda_{\max}^2(2N-s)\|f_{\beta_{\mathcal{N}}}\|^2\sqrt{\frac{s}{N-s}}\Lambda_{\min}^{-2}(S,N).$$

For the second result, we refer to Problem 6.15. $\qquad\square$

Corollary 6.14. *Suppose that*

$$\frac{L\Lambda_{\max}^2(2N-s)}{\Lambda_{\min}^2(S,N)}\sqrt{\frac{s}{N-s}} < 1. \tag{6.53}$$

Then by (6.52)

$$\phi_{\text{adap}}(L,S,N) \geq \left(1 - \frac{L\Lambda_{\max}^2(2N-s)}{\Lambda_{\min}^2(S,N)}\sqrt{\frac{s}{N-s}}\right)\Lambda_{\min}(S,N).$$

Condition (6.53) is used in Zhang and Huang (2008) and Meinshausen and Bühlmann (2010) with a suitable choice of N (and with $\Lambda_{\min}(S,N)$ replaced by $\Lambda_{\min}(N)$), for deriving variable selection properties of the Lasso or randomized Lasso. (see also Subsection 7.8 and 10.6).

6.13.6 Further coherence notions

We first consider the matrix norms in some detail.

Lemma 6.30. *The quantity* $\|\Sigma_{1,2}(\mathcal{N})\|_{2,2}^2$ *is the largest eigenvalue of the matrix* $\Sigma_{1,2}(\mathcal{N})\Sigma_{2,1}(\mathcal{N})$. *We further have*

$$\|\Sigma_{1,2}(\mathcal{N})\|_{2,q} \leq \sqrt{N}\|\Sigma_{1,2}(\mathcal{N})\|_{\infty,q},$$

and

$$\|\Sigma_{1,2}(\mathcal{N})\|_{2,q} \leq \left(\sum_{j\notin\mathcal{N}}\left(\sqrt{\sum_{k\in\mathcal{N}}\sigma_{j,k}^2}\right)^q\right)^{1/q}.$$

Moreover,

$$\|\Sigma_{1,2}(\mathcal{N})\|_{\infty,q} \leq \left(\sum_{j\notin\mathcal{N}}\max_{k\in\mathcal{N}}|\sigma_{j,k}|^q\right)^{\frac{1}{q}}.$$

Finally,

$$\|\Sigma_{1,2}(\mathcal{N})\|_{\infty,q} \geq \|\Sigma_{1,2}(\mathcal{N})\|_{\infty,\infty},$$

and

$$\|\Sigma_{1,2}(\mathcal{N})\|_{2,q} \geq \|\Sigma_{1,2}(\mathcal{N})\|_{2,\infty}.$$

The proof is Problem 6.13. Hence, for replacing $\|\Sigma_{1,2}(\mathcal{N})\|_{\infty,\infty}$ ($\|\Sigma_{1,2}(\mathcal{N}\|_{2,\infty})$ by $\|\Sigma_{1,2}(\mathcal{N})\|_{\infty,q}$ ($\|\Sigma_{1,2}(\mathcal{N}\|_{2,q})$, $q < \infty$, one might have to pay a price.

In the next lemma, we assume $q < \infty$, as the case $q = \infty$ was already treated in Subsection 6.13.2. Recall that $\Lambda_{\min,1}^2(\Sigma_{1,1}(\mathcal{N}))$ is the minimal ℓ_1-eigenvalue of the matrix $\Sigma_{1,1}(\mathcal{N})$. It was defined in the beginning of this section.

Lemma 6.31. *(Coherence lemma) Let* $N > s$ *and* $1 \leq q < \infty$ *and* $1/q + 1/r = 1$. *Then*

$$\vartheta(L,S,N) \leq \max_{\mathcal{N}\supset S,\,|\mathcal{N}|=N}\frac{L\|\Sigma_{1,2}(\mathcal{N})\|_{\infty,q}}{(N-s)^{1/q}\Lambda_{\min,1}^2(\Sigma_{1,1}(\mathcal{N}))}.$$

Moreover,

$$\vartheta_{\mathrm{adap}}(L,S,N) \leq \max_{\mathcal{N}\supset S,\,|\mathcal{N}|=N}\frac{L\sqrt{s}\|\Sigma_{1,2}(\mathcal{N})\|_{2,q}}{(N-s)^{1/q}\Lambda_{\min}^2(\Sigma_{1,1}(\mathcal{N}))}.$$

Finally,

$$\vartheta_{\min}(L,S,N) \leq \max_{\mathcal{N} \supset S,\, |\mathcal{N}|=N} \frac{(L+1)\sqrt{N}\|\Sigma_{1,2}(\mathcal{N})\|_{2,q}}{(N-s)^{1/q}\Lambda_{\min}^2(\Sigma_{1,1}(\mathcal{N}))}.$$

Proof of Lemma 6.31. Let $\mathcal{N} \supset S$ the set with $\mathcal{N}\backslash S$ being the set of indices of the $N-s$ largest $|\beta_j|$ with $j \notin S$. We let $f_{\mathcal{N}} := f_{\beta_{\mathcal{N}}}$, $f_{\mathcal{N}^c} := f_{\beta_{\mathcal{N}^c}}$, and $f := f_\beta$.

We have

$$|(f_{\mathcal{N}}, f_{\mathcal{N}^c})| = |\beta_{\mathcal{N}}^T \Sigma_{1,2}(\mathcal{N})\beta_{\mathcal{N}^c}|$$

$$\leq \|\Sigma_{1,2}(\mathcal{N})\|_{\infty,q}\|\beta_{\mathcal{N}^c}\|_r\|\beta_{\mathcal{N}}\|_1.$$

By Lemma 6.9,

$$\|\beta_{\mathcal{N}^c}\|_r \leq \|\beta_{S^c}\|_1/(N-s)^{1/q}. \tag{6.54}$$

We now use: if $\|\beta_{S^c}\|_1 \leq L\|\beta_S\|_1$, $\|\beta_{S^c}\|_1 \leq L\|\beta_{\mathcal{N}}\|_1$. This yields

$$|(f_{\mathcal{N}}, f_{\mathcal{N}^c})| \leq \frac{L}{(N-s)^{1/q}}\|\beta_{\mathcal{N}}\|_1^2.$$

Alternatively,

$$|(f_{\mathcal{N}}, f_{\mathcal{N}^c})| \leq \|\Sigma_{1,2}(\mathcal{N})\|_{2,q}\|\beta_{\mathcal{N}^c}\|_r\|\beta_{\mathcal{N}}\|_2,$$

and if $\|\beta_{S^c}\|_1 \leq L\sqrt{s}\|\beta_S\|_2$, $\|\beta_{S^c}\|_1 \leq L\sqrt{s}\|\beta_{\mathcal{N}}\|_2$. So then

$$|(f_{\mathcal{N}}, f_{\mathcal{N}^c})| \leq \frac{L\sqrt{s}}{(N-s)^{1/q}}\|\beta_{\mathcal{N}}\|_2^2.$$

The third result follows from

$$\|\beta_{S^c}\|_1 \leq \|\beta_{\mathcal{N}^c}\|_1 + \|\beta_{\mathcal{N}}\|_1.$$

Hence, for $\beta \in \mathscr{R}_{\min}(L,S,\mathcal{N})$,

$$\|\beta_{S^c}\|_1 \leq (L+1)\sqrt{N}\|\beta_{\mathcal{N}}\|_2.$$

$\square$

With $q = 1$ and $N = s$, the coherence lemma is similar to the cumulative local coherence condition in Bunea et al. (2007b). We also consider the case $N = 2s$. We confine ourselves to the adaptive restricted eigenvalues. (For the minimal adaptive restricted eigenvalues it is straightforward to adjust the constants appropriately.)

Corollary 6.15. *(Coherence with $q = 1$) Suppose that*

$$\frac{\sqrt{s}\sqrt{\sum_{k\in S}\left(\sum_{j\notin S}|\sigma_{j,k}|\right)^2}}{\Lambda_{\min}^2(\Sigma_{1,1}(S))} \le \vartheta < 1/L,$$

or

$$\max_{\mathcal{N}\supset S,\,|\mathcal{N}|=2s}\frac{\sqrt{\sum_{k\in\mathcal{N}}\left(\sum_{j\notin\mathcal{N}}|\sigma_{j,k}|\right)^2}}{\sqrt{s}\Lambda_{\min}^2(\Sigma_{1,1}(\mathcal{N}))} \le \vartheta < 1/L,$$

then $\phi_{\mathrm{adap}}(L,S,N) \ge (1-L\vartheta)\Lambda_{\min}(\Sigma_{1,1}(S))$.

The coherence lemma with $q=2$ is a condition about eigenvalues. It is stronger than the restricted isometry property in Candès and Tao (2005) or Candès and Tao (2007). Taking moreover $N=2s$ in Lemma 6.31 gives

Corollary 6.16. *(Coherence with $q=2$) Suppose that*

$$\max_{\mathcal{N}\supset S,\,|\mathcal{N}|=2s}\frac{\|\Sigma_{1,2}(\mathcal{N})\|_{2,2}}{\Lambda_{\min}^2(\Sigma_{1,1}(\mathcal{N}))} \le \vartheta < 1/L$$

Then $\phi_{\mathrm{adap}}(L,S,2s) \ge (1-L\vartheta)\Lambda_{\min}(\Sigma_{1,1}(S))$.

6.13.7 An overview of the various eigenvalue flavored constants

We put the various constants used for proving oracle results together.

Let S be an index set with cardinality s. For $\mathcal{N}\supset S$, we define the restricted sets

$$\mathcal{R}(L,S,\mathcal{N}) := \left\{\beta:\ \|\beta_{S^c}\|_1 \le L\|\beta_S\|_1,\ \max_{j\notin\mathcal{N}}|\beta_j| \le \min_{j\in\mathcal{N}\setminus S}|\beta_j|\right\},$$

$$\mathcal{R}_{\mathrm{adap}}(L,S,\mathcal{N}) := \left\{\beta:\ \|\beta_{S^c}\|_1 \le L\sqrt{s}\|\beta_S\|_2,\ \max_{j\notin\mathcal{N}}|\beta_j| \le \min_{j\in\mathcal{N}\setminus S}|\beta_j|\right\},$$

and

$$\mathcal{R}_{\min}(L,S,\mathcal{N}) := \left\{\beta:\ \|\beta_{\mathcal{N}^c}\|_1 \le L\sqrt{|\mathcal{N}|}\|\beta_{\mathcal{N}}\|_2,\ \max_{j\notin\mathcal{N}}|\beta_j| \le \min_{j\in\mathcal{N}\setminus S}|\beta_j|\right\}.$$

In Chapter 7, we will also use

$$\mathcal{R}_{\mathrm{varmin}}(L,S,\mathcal{N}) := \left\{\beta:\ \|\beta_{\mathcal{N}^c}\|_1 \le L\sqrt{|\mathcal{N}|}\|\beta_{\mathcal{N}}\|_2\right\} = \mathcal{R}_{\mathrm{adap}}(L,\mathcal{N},|\mathcal{N}|),$$

and the variant of the minimal adaptive restricted eigenvalue given below.

For $N \geq s$, we have defined the following constants:

the **maximal eigenvalue**

$$\Lambda_{\max}^2(\Sigma_{1,1}(S)) := \max\left\{\frac{\|f_{\beta_S}\|^2}{\|\beta_S\|_2^2} : \beta \neq 0\right\}.$$

the **maximal sparse eigenvalue**

$$\Lambda_{\max}^2(N) := \max\left\{\Lambda_{\max}^2(\Sigma_{1,1}(\mathcal{N})) : |\mathcal{N}| = N\right\}.$$

the **minimal eigenvalue**

$$\Lambda_{\min}^2(\Sigma_{1,1}(S)) := \min\left\{\frac{\|f_{\beta_S}\|^2}{\|\beta_S\|_2^2} : \beta \neq 0\right\},$$

the **uniform eigenvalue** (which is generally larger than the minimal sparse eigenvalue $\Lambda_{\min}^2(N)$ used in literature)

$$\Lambda_{\min}^2(S,N) := \min\left\{\Lambda_{\min}^2(\Sigma_{1,1}(\mathcal{N})) : \mathcal{N} \supset S, |\mathcal{N}| = N\right\}.$$

the **compatibility constant**

$$\phi_{\text{comp}}^2(L,S) := \min\left\{\frac{s\|f_{\beta}\|^2}{\|\beta_S\|_1^2} : \beta \in \mathcal{R}(L,S,S)\right\},$$

the **restricted eigenvalue**

$$\phi^2(L,S,N) := \min\left\{\frac{\|f_{\beta}\|^2}{\|\beta_{\mathcal{N}}\|_2^2} : \beta \in \mathcal{R}(L,S,\mathcal{N}), \mathcal{N} \supset S, |\mathcal{N}| = N\right\},$$

the **adaptive restricted eigenvalue**

$$\phi_{\text{adap}}^2(L,S,N) := \min\left\{\frac{\|f_{\beta}\|^2}{\|\beta_{\mathcal{N}}\|_2^2} : \beta \in \mathcal{R}_{\text{adap}}(L,S,\mathcal{N}), \mathcal{N} \supset S, |\mathcal{N}| = N\right\},$$

and the **minimal adaptive restricted eigenvalue**

$$\phi_{\min}^2(L,S,N) := \min\left\{\frac{\|f_{\beta}\|^2}{\|\beta_{\mathcal{N}}\|_2^2} : \beta \in \mathcal{R}_{\min}(L,S,\mathcal{N}), \mathcal{N} \supset S, |\mathcal{N}| = N\right\}.$$

We also introduced the minimal ℓ_1-eigenvalue, which we skip here as it was not further exploited.

Moreover, in Chapter 7 we employ

the **variant of the minimal adaptive restricted eigenvalue**

$$\phi^2_{\text{varmin}}(L,S,N) := \min\left\{ \frac{\|f_\beta\|^2}{\|\beta_{\mathscr{N}}\|_2^2} : \beta \in \mathscr{R}_{\text{varmin}}(L,S,\mathscr{N}), \ \mathscr{N} \supset S, \ |\mathscr{N}| = N \right\}.$$

$$= \min_{\mathscr{N} \supset S, \ |\mathscr{N}|=N} \phi^2_{\text{adap}}(L,\mathscr{N},N).$$

We note that

$$\phi_{\text{comp}}(L,S) \geq \phi(L,S,N), \ \forall N \geq s,$$

and that

$$\phi(L,S,N) \geq \phi_{\text{adap}}(L,S,N) \geq \phi_{\text{min}}(L,S,N) \geq \phi_{\text{varmin}}(L,S,N).$$

This follows from

$$\mathscr{R}(L,S,\mathscr{N}) \subset \mathscr{R}_{\text{adap}}(L,S,\mathscr{N}) \subset \mathscr{R}_{\text{min}}(L,S,\mathscr{N}) \subset \mathscr{R}_{\text{varmin}}(L,S,\mathscr{N}).$$

We have also introduced

the **restricted regression**

$$\vartheta(L,S,N) := \max\left\{ \frac{|(f_{\beta_{\mathscr{N}}},f_{\beta_{\mathscr{N}^c}})|}{\|f_{\beta_{\mathscr{N}}}\|^2} : \beta \in \mathscr{R}(L,S,\mathscr{N}), \ \mathscr{N} \supset S, \ |\mathscr{N}| = N \right\},$$

and the **adaptive restricted regression**

$$\vartheta_{\text{adap}}(L,S,N) := \max\left\{ \frac{|(f_{\beta_{\mathscr{N}}},f_{\beta_{\mathscr{N}^c}})|}{\|f_{\beta_{\mathscr{N}}}\|^2} : \beta \in \mathscr{R}_{\text{adap}}(L,S,\mathscr{N}), \ \mathscr{N} \supset S, \ |\mathscr{N}| = N \right\}.$$

The minimal adaptive restricted regression $\vartheta_{\text{min}}(L,S,N)$ and variant of the minimal adaptive restricted regression $\vartheta_{\text{varmin}}(L,S,N)$ can be defined analogously:

$$\vartheta_{\text{min}}(L,S,N) := \max\left\{ \frac{|(f_{\beta_{\mathscr{N}}},f_{\beta_{\mathscr{N}^c}})|}{\|f_{\beta_{\mathscr{N}}}\|^2} : \beta \in \mathscr{R}_{\text{min}}(L,S,\mathscr{N}), \ \mathscr{N} \supset S, \ |\mathscr{N}| = N \right\}.$$

$$\vartheta_{\text{varmin}}(L,S,N) := \max\left\{ \frac{|(f_{\beta_{\mathscr{N}}},f_{\beta_{\mathscr{N}^c}})|}{\|f_{\beta_{\mathscr{N}}}\|^2} : \beta \in \mathscr{R}_{\text{varmin}}(L,S,\mathscr{N}), \ \mathscr{N} \supset S, \ |\mathscr{N}| = N \right\}.$$

Then clearly

$$\vartheta(L,S,N) \leq \vartheta_{\text{adap}}(L,S,N) \leq \vartheta_{\text{min}}(L,S,N) \leq \vartheta_{\text{varmin}}(L,S,N).$$

Some further relations are summarized in Figure 6.1.

As an implication of Lemma 6.26, we have, whenever the restricted regression is less than 1,

$$\phi(L,S,N) \geq (1 - \vartheta(L,S,N))\Lambda_{\min}(S,N),$$

and similarly

$$\phi_{\mathrm{adap}}(L,S,N) \geq (1 - \vartheta_{\mathrm{adap}}(L,S,N))\Lambda_{\min}(S,N),$$

$$\phi_{\min}(L,S,N) \geq (1 - \vartheta_{\min}(L,S,N))\Lambda_{\min}(S,N),$$

and

$$\phi_{\mathrm{varmin}}(L,S,N) \geq (1 - \vartheta_{\mathrm{varmin}}(L,S,N))\Lambda_{\min}(S,N).$$

We have seen in Section 6.4 that bounds for the excess risk and the ℓ_1-error involve the compatibility constant $\phi_{\mathrm{comp}}(L,S)$, where we throughout have chosen (quite arbitrarily) $L = 3$, which related to the required lower bounds on the tuning parameter λ.
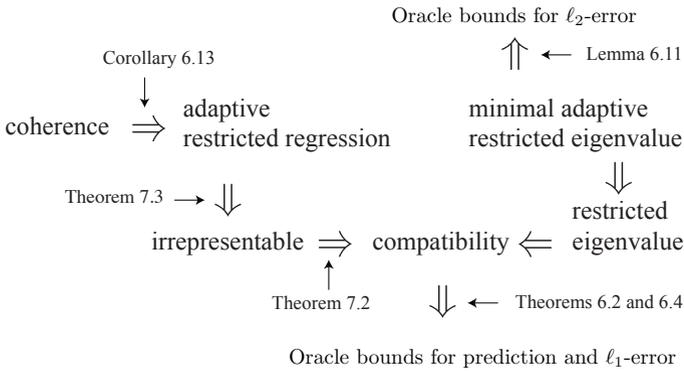
Fig. 6.1 Some important relations between the various concepts. Theorem 6.4 and Lemma 6.11 consider general convex loss functions. Theorem 6.2 considers squared error loss. When applied to the true active set S_0, the restricted eigenvalue conditions also imply oracle bounds for the ℓ_2-error (see Lemma 6.10). The irrepresentable condition is defined in Subsection 7.5.1.

For bounds for the ℓ_2-error, we need the restricted eigenvalue $\phi(L,S_0,2s_0)$ in case we assume the truth is sparse, and the minimal adaptive restricted eigenvalue $\phi_{\min}(L,S_*,2s_*)$ in case we use a sparse oracle approximation. Problem 6.15 shows that $\vartheta_{\mathrm{varmin}}(L,S,N)$ can be bounded in terms of the orthogonality constants $\theta_{s,N}$ and uniform eigenvalues.

Problems

6.1. In the context of Section 6.2, let

$$SNR := \frac{\|\mathbf{X}\beta^0\|_2}{\sqrt{n}\sigma}.$$

be the signal-to-noise ratio, and

$$\hat{\sigma}^2 := \mathbf{Y}^T\mathbf{Y}/n,$$

where $\mathbf{Y}$ is the response variable (assuming for simplicity that $\mathbf{Y}$ has mean zero).
Verify that, for any $t > 0$, one has with probability at least $1 - 2\exp[-t^2/2]$,

$$1 + SNR(SNR - 2t/\sqrt{n}) - b_n \leq \frac{\hat{\sigma}^2}{\sigma^2} \leq 1 + SNR(SNR + 2t/\sqrt{n}) + b_n,$$

where

$$b_n := \left|\frac{\varepsilon^T\varepsilon}{n\sigma^2} - 1\right|.$$

6.2. Suppose that $|X_i^{(j)}| \leq 1$ for all $1 \leq i \leq n$ and $1 \leq j \leq p$, and that $\varepsilon_1,\ldots,\varepsilon_n$ are independent centered random variables with second moment uniformly bounded by 1:

$$\max_{1 \leq i \leq n} \mathbb{E}\varepsilon_i^2 \leq 1.$$

Let

$$\mathcal{T} := \left\{\max_{1 \leq j \leq p} 2|\varepsilon^T\mathbf{X}^{(j)}|/n \leq \lambda_0\right\},$$

with, for some $t > 0$,

$$\lambda_0 := 4t\sqrt{\log(2p)/n}.$$

Show that

$$\mathbf{P}(\mathcal{T}) \geq 1 - 2/t^2,$$

using the result of Dümbgen et al. (2010) (see (6.5), and Section 14.10).

6.3. Theorem 6.2 is a corollary of

Theorem 6.5. *Let*

$$\mathcal{T} := \left\{\max_{1 \leq j \leq p} 2|\varepsilon^T\mathbf{X}^{(j)}|/n \leq \lambda_0\right\}.$$

Take $\lambda \geq 4\lambda_0$. *Then on* $\mathcal{T}$,

$$2\|\mathbf{X}\hat{\beta} - \mathbf{f}^0\|_2^2/n + \lambda\|\hat{\beta} - \beta^*\|_1 \leq 6\|\mathbf{X}\beta^* - \mathbf{f}^0\|_2^2/n + \frac{24\lambda^2 s_*}{\phi_*^2}.$$

Use this to prove an oracle inequality assuming the conditions of Problem 6.2.

6.4. Theorem 6.4 gives a bound for $\|\hat{\beta} - \beta^*\|_1$ where β^* is the "oracle". Because β^* may not be the parameter of interest, we consider the following. Let Σ be some symmetric, positive definite matrix. For an index set $S \subset \{1, \ldots, p\}$, with cardinality s let $\Sigma_{1,1}(S) := (\Sigma_{j,k})_{j,k \in S}$ be the $s \times s$ matrix consisting of the rows and columns corresponding to the indices in S. Let furthermore $\Lambda^2_{\min}(\Sigma_{1,1}(S))$ be the smallest eigenvalue of $\Sigma_{1,1}(S)$. Define, for some β^0 and β^*, the sets $S_0 := S_{\beta^0}$, and $S_* := S_{\beta^*}$. Let $s_* := |S_*|$. Show that (for any $\lambda > 0$)

$$\lambda \|\beta^* - \beta^0\|_1 \leq (\beta^* - \beta^0)^T \Sigma (\beta^* - \beta^0) + \frac{\lambda^2 s_*}{4\Lambda^2_{\min}(\Sigma_{1,1}(S_*))} + \lambda \|\beta^0_{S_0 \setminus S_*}\|_1.$$

Let $\phi^2_\Sigma(S_*)$ be the Σ-compatibility constant (see Section 6.12). Since $\phi_\Sigma(S_*) \leq \Lambda_{\min}(\Sigma_{1,1}(S_*))$, the same result holds if we replace $\Lambda^2_{\min}(\Sigma_{1,1}(S_*))$ by the compatibility constant $\phi^2_\Sigma(S_*)$. Extend the result to

$$\lambda \|\beta^* - \beta^0\|_1 \leq G(\|f_{\beta^*} - f_{\beta^0}\|) + H\left(\frac{\lambda \sqrt{s_*}}{\phi_\Sigma(S_*)}\right) + \lambda \|\beta^0_{S_0 \setminus S_*}\|_1.$$

Here, $f_\beta = \sum_{j=1}^p \beta_j \psi_j$, $\|\cdot\| = \|\cdot\|_\Sigma$, and G is an increasing and strictly convex function with convex conjugate H.

6.5. Prove the Basic Inequality of Lemma 6.4.

6.6. Consider the density estimation problem. Let $X_1, \ldots, X_n$ be i.i.d. copies of X. Suppose X has density p^0 with respect to some given σ-finite dominating measure μ. Write $f^0 := \log p^0$, and let

$$\mathscr{F} := \left\{ f_\beta = \sum_{j=1}^p \beta_j \psi_j \right\},$$

where β ranges over the convex set

$$\left\{ \beta : \int \exp[\sum_{j=1}^p \beta_j \psi_j] d\mu < \infty \right\}.$$

Define

$$\mathbf{F} := \left\{ f : \int \exp[f] d\mu < \infty \right\},$$

and for $f \in \mathbf{F}$,

$$b(f) := \log\left(\int \exp[f] d\mu\right).$$

Let the loss function be

$$\rho_f = -f + b(f).$$

First, show that the target is indeed f^0. Show that the excess risk is the Kullback-Leibler information:

$$\mathcal{E}(f) = \int \log\left(\frac{p^0}{p_f}\right) p^0 d\mu,$$

where $p_f = \exp[f - b(f)]$. What norm $\|\cdot\|$ would you use on $\mathbf{F}$? Suppose for some constant $\varepsilon_0 > 0$, that $\varepsilon_0 \le p^0 \le 1/\varepsilon_0$. Let $\mathbf{F}_{\text{local}} := \{\|f - f^0\|_\infty \le \eta\}$. Check the quadratic margin condition.

Consider now the Lasso-density estimator

$$\hat{\beta} := \{-P_n f_\beta + b(f_\beta) + \lambda \|\beta\|_1\}.$$

Using Theorem 6.4, derive an oracle inequality for $\hat{f} = f_{\hat{\beta}}$.

6.7. As in Problem 6.4, we consider the density estimation problem, but now with a different loss function. Let $X_1, \ldots, X_n$ be i.i.d. copies of X. Suppose X has density f^0 with respect to some given σ-finite dominating measure μ. Let $\|\cdot\|$ be the $L_2(\mu)$-norm, and $\mathbf{F}$ be a convex subset of $L_2(\mu)$. We assume $f^0 \in \mathbf{F}$. Let

$$\rho_f := \|f\|^2 - 2f, \ f \in \mathbf{F}.$$

Show that f^0 is indeed the target. Let $\mathscr{F} := \{f_\beta = \sum_{j=1}^{p} \beta_j \psi_j\}$ be some convex subset of $\mathbf{F}$. Consider the Lasso-density estimator (or *SPADES*: see Bunea et al. (2007b))

$$\hat{\beta} := \arg\min\{\|f\|^2 - 2P_n f + \lambda \|\beta\|_1\}.$$

Using Theorem 6.4, derive an oracle inequality for $\hat{f} = f_{\hat{\beta}}$.

6.8. In classification, a popular loss is the so-called hinge loss (used for example in support vector machines). Let $\{X_i, Y_i\}_{i=1}^n$ be i.i.d. copies of (X, Y), with $Y \in \{-1, 1\}$. Hinge loss is

$$\rho_f(x, y) = (1 - yf(x))_+,$$

where $z_+ = z\mathbb{1}\{z \ge 0\}$ is the positive part of z. Define

$$\pi(\cdot) := P(Y = 1|X = \cdot).$$

Show that the target can be taken as Bayes' rule

$$f^0 := \text{sign}(2\pi - 1).$$

Remark The Lasso (in particular the margin condition with Bayes rule as target) is studied in Tarigan and van de Geer (2006). However, typically, Bayes' rule will not be well-approximated by linear functions. Therefore it makes sense to replace Bayes' rule by the alternative target

$$f^0_{\text{GLM}} := \arg\min_{f \in \mathscr{F}} P\rho_f,$$

where $\mathscr{F} := \{f_\beta = \sum_{j=1}^p \beta_j \psi_j\}$ is the class of linear functions under consideration.

6.9. Consider the linear "link function"

$$f_\beta(x) = \sum_{j=1}^p \beta_j \psi_j(x).$$

We fix $r < p$, do not penalize $\beta_1, \dots, \beta_r$, and use the ℓ_1-penalty

$$\lambda \sum_{j=r+1}^p |\beta_j|$$

for the remaining $p - r$ variables. We now show that one can reparametrize to having the penalized base functions orthogonal to the unpenalized ones. Let $\Psi_1 := (\psi_1, \dots, \psi_r)$, and $\Psi_2 := (\psi_{r+1}, \dots, \psi_p)$. Take

$$\tilde{\Psi}_1 := \Psi_1 + \Psi_{2,1}^{\mathrm{P}},$$

where $\Psi_{2,1}^{\mathrm{P}}$ is the projection of Ψ_2 on Ψ_1. Moreover, take

$$\tilde{\Psi}_2 := \Psi_2 - \Psi_{2,1}^{\mathrm{P}}.$$

Then $\tilde{\Psi}_1$ and $\tilde{\Psi}_2$ are clearly orthogonal. Check that for certain coefficients $\{\tilde{\beta}_j\}_{j=1}^r$

$$\sum_{j=1}^p \beta_j \psi_j = \sum_{j=1}^r \tilde{\beta}_j \psi_j + \sum_{j=r+1}^p \beta_j \tilde{\psi}_j.$$

6.10. Let Σ be some positive semi-definite matrix, and write

$$\|f_\beta\|_\Sigma^2 := \beta^T \Sigma \beta.$$

The compatibility and restricted eigenvalue conditions depend on Σ. Let us write the (L, S, N)-restricted eigenvalue as $\phi_\Sigma^2(L, S, N)$. Let $\hat{\Sigma}$ be another positive semi-definite matrix, and write

$$\|\hat{\Sigma} - \Sigma\|_\infty := \max_{j,k} |\hat{\Sigma}_{j,k} - \Sigma_{j,k}|.$$

Show that

$$\phi_{\hat{\Sigma}}^2(L, S, N) \geq \phi_\Sigma^2(L, S, N) - (L+1)^2 \|\hat{\Sigma} - \Sigma\|_\infty s.$$

6.11. Provide a proof for Lemma 6.25, by applying the same arguments as in Lemma 6.20.

6.12. Show that for all β,

$$\sqrt{\sum_{k \in S} \left(\sum_{j \notin S} \sigma_{j,k} \beta_j \right)^2} \leq \sum_{j \notin S} |\beta_j| \sqrt{\sum_{k \in S} \sigma_{j,k}^2}.$$

6.13. By using the definitions of $\|\Sigma_{1,2}(\mathcal{N})\|_{\infty,q}$ and $\|\Sigma_{1,2}(\mathcal{N})\|_{2,q}$, prove Lemma 6.30.

6.14.
(a) Show that if Σ has ones on the diagonal, and all off-diagonal elements of Σ are equal to θ, where $0 < \theta < 1$, then its minimal eigenvalue is at least $1 - \theta$, so the compatibility condition holds with $\phi_{\text{comp}}^2(S) \geq 1 - \theta$. Hint: note that one can write

$$\Sigma = (1 - \theta)I + \theta\tau\tau^T,$$

where τ is a p-vector of ones. Also show that with this Σ, the irrepresentable condition (see Subsection 7.5.1) holds for all S.
(b) More generally, let

$$\Sigma = (I - \Theta) + \Theta^{1/2}\tau\tau^T\Theta^{1/2},$$

where $\tau \in \{-1, 1\}$, and $\Theta := \text{diag}(\theta_1, \ldots, \theta_p)$, with $0 < \theta_j \leq 1$ for all j. Show that

$$\Lambda_{\min}^2(\Sigma_{1,1}(S)) \geq 1 - \max_{k \in S} \theta_k,$$

and in fact, that

$$\phi_{\text{adap}}^2(L, S, s) \geq 1 - \max_{k \in S} \theta_k.$$

Check that

$$\Sigma_{1,1}^{-1}(S) = (I - \Theta(S))^{-1} - \frac{(I - \Theta(S))^{-1}\Theta(S)^{1/2}\tau_S\tau_S^T\Theta(S)^{1/2}(I - \Theta(S))^{-1}}{1 + \tau_S^T\Theta(S)^{1/2}(I - \Theta(S))^{-1}\Theta(S)^{1/2}\tau},$$

and that

$$\Sigma_{2,1}(S)\Sigma_{1,1}^{-1}(S) = \frac{\Theta^{1/2}(S^c)\tau_{S^c}\tau_S^T\Theta^{1/2}(S)(I - \Theta(S))^{-1}}{1 + \tau_S^T\Theta^{1/2}(S)(I - \Theta(S))^{-1}\Theta^{1/2}\tau_S}.$$

Does the irrepresentable condition hold?
(c) Suppose now that

$$\Sigma = (1 - \Theta)I + \Theta^{1/2}R\Theta^{1/2},$$

where R is some correlation matrix. Show that

$$\phi_{\text{adap}}^2(L, S, s) \geq 1 - \max_{k \in S} \theta_k.$$

6.15. Check that in Lemma 6.29, one may replace the sparse eigenvalue $\Lambda_{\max}^2(2N - s)$ by the orthogonality constant $\theta_{N,N-s}$ defined in Subsection 6.13.4. Verify moreover that

$$\vartheta_{\text{varmin}}(L, S, N) \leq \frac{(L+1)\theta_{N,N-s}}{\Lambda_{\min}^2(S, N)}\sqrt{\frac{N}{N-s}}.$$

Chapter 7
Variable selection with the Lasso

Abstract We use the Lasso, its adaptive or its thresholded variant, as procedure for variable selection. This essentially means that for $S_0 := \{j : \beta_j^0 \neq 0\}$ being the true active set, we look for a Lasso procedure delivering an estimator $\hat{S}$ of S_0 such that $\hat{S} = S_0$ with large probability. However, it is clear that very small coefficients $|\beta_j^0|$ cannot be detected by any method. Moreover, irrepresentable conditions show that the Lasso, or any weighted variant, typically selects too many variables. In other words, unless one imposes very strong conditions, false positives cannot be avoided either. We shall therefore aim at estimators with oracle prediction error, yet having not too many false positives. The latter is considered as achieved when $|\hat{S} \backslash S_*| = O(|S_*|)$, where $S_* \subset S_0$ is the set of coefficients the oracle would select. We will show that the adaptive Lasso procedure, and also thresholding the initial Lasso, reaches this aim, assuming sparse eigenvalues, or alternatively, so-called "beta-min" conditions.

7.1 Introduction

In this chapter, we confine ourselves to the linear model:

$$Y_i = \sum_{j=1}^{p} \psi_j(X_i)\beta_j^0 + \varepsilon_i, \ i = 1,\ldots,n,$$

where $\{\psi_j\}_{j=1}^{p}$ is a given dictionary, $X_i \in \mathcal{X}$, $i = 1,\ldots,n$, is fixed design, and where $\varepsilon_1,\ldots,\varepsilon_n$ are noise variables. We use the Lasso, its adaptive or its thresholded variant, as procedure for variable selection.

A key motivation for the exploration of thresholding and the adaptive Lasso for variable selection is to relax the stringent irrepresentable conditions on the design matrix. Thus, we have to tolerate some false positive selections. Furthermore, some

false negative selections can not be avoided either, as one preferably refrains from assumptions saying that the minimal non-zero coefficients β^0 of the "true" regression are "sufficiently large" (we call this the "beta-min" condition, see Section 7.4), since allowing for small non-zero regression coefficients appears to be much more realistic. Consequently, it is impossible to infer the true underlying active set

$$S_0 = \{j : \beta_j^0 \neq 0\},$$

since covariables j whose corresponding absolute coefficient $|\beta_j^0|$ is below a detection limit cannot be inferred from data (say with probability tending to 1 as $n \to \infty$).

7.2 Some results from literature

For consistent variable selection with the Lasso, it is known that the so-called "neighborhood stability condition" (Meinshausen and Bühlmann, 2006) for the design matrix, which has been re-formulated in a nicer form as the "irrepresentable condition" (Zhao and Yu, 2006), is sufficient and essentially necessary, see Section 2.6.1 and also Subsection 7.5.1 in the present chapter. A further refined analysis in given in Wainwright (2007, 2009), which presents under certain incoherence conditions the smallest sample size needed to recover a sparse signal. Because irrepresentable conditions or incoherence conditions are restrictive - they are much stronger than restricted eigenvalue conditions (see Subsection 6.13.7 or van de Geer and Bühlmann (2009) for an overview) - we conclude that the Lasso for variable selection only works in a rather narrow range of problems, excluding many cases where the design exhibits strong (empirical) correlations.

There is moreover a bias problem with ℓ_1-penalization, due to the shrinking of the estimates which correspond to true signal variables. A discussion can be found in Zou (2006), and Meinshausen (2007) (see also Subsection 2.8.2). Regularization with the ℓ_r-"norm" with $r < 1$ (see Section 7.13) mitigates some of the bias problems but are computationally infeasible as the penalty is non-convex. As an interesting alternative, one can consider multi-step procedures where each of the steps involves a convex optimization only. A prime example is the adaptive Lasso which is a two-step algorithm and whose repeated application corresponds in some "loose" sense to a non-convex penalization scheme (see Zou and Li (2008) and Subsection 2.8.6). The adaptive Lasso was originally proposed by Zou (2006). He analyzed the case where p is fixed. Further progress in the high-dimensional scenario has been achieved by Huang et al. (2008). Under a rather strong mutual incoherence condition between every pair of relevant and irrelevant covariables, they prove that the adaptive Lasso recovers the correct model and has an oracle property.

Meinshausen and Yu (2009) examined the variable selection property of the Lasso followed by a thresholding procedure, when all non-zero components are large enough. Under a relaxed incoherence assumption, they show that the estimator is

still consistent in the ℓ_2-norm sense. In addition, they prove that it is possible to achieve variable selection consistency. Thresholding and multistage procedures are also considered in Candès et al. (2006). In Zhou (2009b, 2010), it is shown that a multi-step thresholding procedure can accurately estimate a sparse vector $\beta^0 \in \mathbb{R}^p$ under the restricted eigenvalue condition of Bickel et al. (2009). The two-stage procedure in Zhang (2009b) applies "selective penalization" in the second stage. This procedure is studied assuming incoherence conditions. A more general framework for multi-stage variable selection was studied by Wasserman and Roeder (2009). Their approach controls the probability of false positives (type I error) but pays a price in terms of false negatives (type II error). Chapter 11 describes the details.

7.3 Organization of this chapter

Section 7.4 discusses the so-called "beta-min" condition, which requires that the non-zero (true or oracle) coefficients are sufficiently large.

Section 7.5 considers the irrepresentable condition in the noiseless case (i.e., the case where $\varepsilon = 0$) and its relation with other conditions. For necessity of the conditions it suffices to consider the noiseless case.

Section 7.5 consists of 9 subsections.

A sufficient and essentially necessary condition to perform variable selection with the Lasso is the irrepresentable condition. (A related result is that under additional assumptions, the standard Lasso estimator $\hat{\beta}_{\text{init}}$ is close to β^0, not only in ℓ_q-norm ($q = 1$ or $1 \leq q \leq 2$), but also in ℓ_∞-norm, see Zhang (2009b).) The irrepresentable condition is defined in Subsection 7.5.1. After recalling the KKT conditions (Subsection 7.5.2), we give in Subsection 7.5.3 the necessary and sufficient conditions for (exact) variable selection.

We show in Subsection 7.5.4 that the irrepresentable condition implies the compatibility condition (which in turn is sufficient for oracle inequalities for the prediction error of the Lasso), and in Subsection 7.5.5 that the adaptive restricted regression condition (the latter being introduced in Subsection 6.13.2 as sufficient condition for the compatibility condition) implies the irrepresentable condition. A simple generalization of the irrepresentable condition, allowing for a given number of false positives, is given in Subsection 7.5.6.

The adaptive Lasso is a special case of the weighted Lasso. After a reparametrization, one easily sees that a sufficient condition for variable selection by the weighted Lasso is the weighted irrepresentable condition. The weighted irrepresentable condition is moreover essentially necessary if the non-zero true coefficients are large enough, that is, if certain beta-min conditions hold. This is elaborated upon in Subsection 7.5.7.

We moreover show in Subsection 7.5.8 that a bound for the adaptive restricted regression is sufficient for proving the weighted irrepresentable condition, and hence variable selection of the weighted Lasso. This bound requires a great amount of

separation between the weights inside and those outside the active set. We present an example (Example 7.3) which implies that one cannot remove the requirement on the amount of separation between weights, without imposing further conditions on the Gram matrix. We show in Subsection 7.5.9, that with "ideal" weights inside the active set, the weights outside the active set should be quite large. With an initial estimator used for the weights, this means that this initial estimator either needs a rather fast convergence rate in sup-norm, or alternatively the sparsity s_0 should be small namely of order $(n/\log p)^{1/3}$. Our conditions cannot be improved, in the sense that there exist Gram matrices where the weighted irrepresentable condition does not hold if our required amount of separation of the weights is not fulfilled (see Example 7.3). We then have reached the conclusion that exact variable selection, even with the adaptive Lasso, is in a sense ill-posed, it can only be accomplished under very strong conditions. We therefore set a different (and perhaps more moderate) aim, targeting at no more than $O(s_*)$ false positives, $s_* \leq s_0$ being the number of variables an oracle would select.

In most of the remainder of the chapter, our results concern the adaptive Lasso and the thresholded Lasso. The two approaches share the problem of the choice of tuning parameter. In Section 7.6 we give the definitions of the adaptive and thresholded Lasso. Section 7.7 recalls our definition of the oracle β^*, and collects the results we obtained in Chapter 6 for the prediction error, ℓ_1-error, and ℓ_2-error of the (weighted) Lasso.

Section 7.8 treats the adaptive and thresholded Lasso, invoking sparse eigenvalue conditions. It consists of six subsections. Subsection 7.8.1 gives the conditions on the tuning parameters, and Subsection 7.8.2 gives the results under sparse eigenvalue conditions. We then compare these with the properties of the standard Lasso in Subsection 7.8.3. The comparison of thresholding and adaptive Lasso is discussed further in Subsection 7.8.4. We look at the implications for the number of false negatives in Subsection 7.8.5. Finally, Subsection 7.8.6 checks how the theory simplifies under so-called beta-min conditions as discussed in Section 7.4. As we will see, beta-min conditions can (partly) replace sparse eigenvalue conditions.

For the proofs of the results in Section 7.8, we refer to van de Geer et al. (2010). They are very much in the spirit of the proofs for the section following it, Section 7.9 (which are given in Sections 7.11 and 7.12).

Section 7.9 gives the results for the adaptive Lasso when sparse eigenvalue conditions are avoided altogether. We again choose the tuning parameters to optimize bounds for the prediction error (Subsection 7.9.1). We show that, depending on the trimmed harmonic mean of the $|\beta_j^*|$, the adaptive Lasso still improves the one-stage Lasso as regards variable selection, and can sometimes maintain a good prediction error. These results are summarized in Theorem 7.10 of Subsection 7.9.2. Section 7.10 contains some concluding remarks.

The technical complements for the (adaptive) Lasso in the noiseless case are derived in Section 7.11. The reason we again omit noise here is that many theoretical issues involved concern the approximation properties of the two stage procedure,

and not so much the fact that there is noise. By studying the noiseless case first, we separate the approximation problem from the stochastic problem. For the noiseless case, Subsection 7.11.1 summarizes the prediction error of the weighted Lasso, and Subsection 7.11.2 gives a bound for the number of false positives in terms of the prediction error. We obtain in Subsection 7.11.3 some simple bounds for the initial Lasso and its thresholded version. In Subsection 7.11.4 we derive results for the adaptive Lasso by comparing it with a "oracle-thresholded" initial Lasso. When the trimmed harmonic mean of the squared coefficients of the target β^* is large enough, the adaptive Lasso combines good variable selection properties with good prediction properties.

The technical complements for the noisy situation are in Section 7.12.

The prediction error of least squares loss with a concave penalty was studied in Section 6.11. Section 7.13 shows under sparse eigenvalue conditions, this method also has $O(s_*)$ false positives, and thus is in that sense comparable to thresholding and to the adaptive Lasso.

7.4 The beta-min condition

It is clear that the larger the smallest non-zero coefficient

$$|\beta^0|_{\min} := \min_{j \in S_0} |\beta_j^0|,$$

the easier is variable selection. We call a condition requiring some lower non-zero bound on $|\beta^0|_{\min}$ a "beta-min" condition. Such a condition is generally not very natural, nor very much in the spirit of uniformity in local neighborhoods (see Leeb and Pötscher (2003)). Nevertheless, beta-min conditions occur in many theoretical works. In our viewpoint, theoretical results that rely on beta-min conditions are primarily useful as a benchmark but should always be considered with some reservation.

A further aspect having to do with beta-min conditions follows from signal-to-noise considerations. Let us explain this here. We study the situation where the truth

$$f^0(X_i) := \mathbb{E}Y_i, \ i = 1, \ldots, n,$$

is linear: $f^0 = \sum_{j=1}^p \psi_j \beta_j^0$. (If this is not the case, our results are to be understood as selection results of the projection of f^0 on the space spanned by $\{\psi_j\}_{j=1}^p$.) The active set of the truth is $S_0 := \{j : \beta_j^0 \neq 0\}$, and $s_0 = |S_0|$ is its cardinality, i.e., the sparsity index of f^0. The typical situation is the one where the ℓ_2-norm $\|\beta^0\|_2$ of the coefficients β^0 is bounded from above (and below). This has certain consequences for the order of magnitude of most $|\beta_j^0|$. But let us first explain why bounds on $\|\beta^0\|_2$ are "typical". Let $Q_n := \sum_{i=1}^n \delta_{X_i}/n$ be the empirical measure of the covariables, and

$\| \cdot \|_n$ be the $L_2(Q_n)$-norm. The Gram matrix is

$$\hat{\Sigma} = \int \psi^T \psi dQ_n, \ \psi := (\psi_1, \ldots, \psi_p).$$

For the case of independent centered noise variables $\{\varepsilon_i\}_{i=1}^n$ with variance σ^2, the signal-to-noise ratio is

$$SNR := \frac{\|f^0\|_n}{\sigma}.$$

A "reasonable" signal-to-noise ratio is an SNR satisfying

$$\eta \le SNR \le 1/\eta,$$

where $\eta > 0$, and in fact, where η is close to one. Clearly,

$$\|\beta^0\|_2 \le \|f^0\|_n / \Lambda_{\min}(\hat{\Sigma}_{1,1}(S_0)),$$

where $\Lambda_{\min}^2(\hat{\Sigma}_{1,1}(S_0))$ is the smallest eigenvalue of the Gram matrix $\hat{\Sigma}_{1,1}(S_0)$ corresponding to the variables in S_0. Thus

$$\|\beta^0\|_2 \le (SNR)\sigma / \Lambda_{\min}(\hat{\Sigma}_{1,1}(S_0)).$$

For the normalized case (i.e., the case where $\text{diag}(\hat{\Sigma}) = I$), it holds that $\Lambda_{\min}(\hat{\Sigma}_{1,1}(S_0)) \le 1$. On the other hand, we generally hope we have a situation where the eigenvalue $\Lambda_{\min}(\hat{\Sigma}_{1,1}(S_0))$ is not very small, actually, that the compatibility constant - or the restricted eigenvalue - is not very small, because this gives good result for the prediction error of the Lasso (see Theorem 6.1). In other words, with a reasonable signal-to-noise ratio and nicely behaved eigenvalues, the ℓ_2-norm $\|\beta^0\|_2$ cannot be too large.[1]

The upper bound for $\|\beta^0\|_2$ has important consequences for variable selection results based on beta-min conditions, as we always have

$$|\beta^0|_{\min} \le \|\beta^0\|_2 / \sqrt{s_0},$$

i.e., the smallest coefficients are not allowed to be larger than $\sigma / \sqrt{s_0}$ in order of magnitude. Put differently, if there are a few large coefficients, this leaves little space for the other coefficients. The latter may need to drop below the noise level. When s_0 is large, a few large coefficients means that the majority of the non-zero coefficients are too small to detect.

---

[1] It can generally not be really small either, e.g., in the normalized case, $\|f^0\|_n = \|\sum_{j=1}^p \psi_j \beta_j^0\|_n \le \|\beta^0\|_1$, so $\|\beta^0\|_2 \ge (SNR)\sigma / \sqrt{s_0}$.

7.5 The irrepresentable condition in the noiseless case

We let Q be some probability measure on $\mathcal{X}$, and $\|\cdot\|$ the $L_2(Q)$ norm. An example is Q being the empirical measure $Q_n := \sum_{i=1}^{n} \delta_{X_i}/n$.

The Gram matrix is

$$\Sigma := \int \psi^T \psi \, dQ.$$

The entries of Σ are denoted by $\sigma_{j,k} := (\psi_j, \psi_k)$, with $(\cdot, \cdot)$ being the inner product in $L_2(Q)$. We furthermore use the notation of Section 6.13. That is, for a given index set S, we consider the submatrices

$$\Sigma_{1,1}(S) := (\sigma_{j,k})_{j,k \in S}, \ \Sigma_{2,2}(S) := (\sigma_{j,k})_{j,k \notin S},$$

and

$$\Sigma_{2,1}(S) := (\sigma_{j,k})_{j \notin S, k \in S}, \ \Sigma_{1,2}(S) := \Sigma_{2,1}^T(S).$$

We let $\Lambda_{\min}^2(\Sigma_{1,1}(S))$ be the smallest eigenvalue of $\Sigma_{1,1}(S)$.

Moreover, as usual, for β being a vector in $\mathbb{R}^p$, we denote by

$$\beta_{j,S} := \beta_j 1\{j \in S\}, \ j = 1, \ldots, p.$$

Thus, β_S is the vector with only non-zero entries in the set S.

The largest eigenvalue of Σ is denoted by $\Lambda_{\max}^2$, i.e.,

$$\Lambda_{\max}^2 := \max_{\|\beta\|_2 = 1} \beta^T \Sigma \beta.$$

We will also need the largest eigenvalue of submatrices containing the inner products of variables in S:

$$\Lambda_{\max}^2(\Sigma_{1,1}(S)) := \max_{\|\beta_S\|_2 = 1} \beta_S^T \Sigma \beta_S.$$

(Minimal adaptive) restricted eigenvalues (defined in Section 6.13) are denoted by $\phi(L, S, N) := \phi_\Sigma(L, S, N)$, $\phi_{\min}(L, S, N) := \phi_{\min, \Sigma}(L, S, N)$, etc.

Let, for some given $\lambda = \lambda_{\text{init}} \geq 0$, β_{init} be the noiseless Lasso

$$\beta_{\text{init}} := \arg\min_\beta \{\|f_\beta - f^0\|^2 + \lambda_{\text{init}} \|\beta\|_1\}.$$

Define its active set as $S_{\text{init}} := \{j : \beta_{\text{init}, j} \neq 0\}$.

7.5.1 Definition of the irrepresentable condition

The irrepresentable condition, as given in (2.20), depends on the Gram matrix Σ, but also on the signs of the unknown "true" parameter β^0, whereas the compatibility condition only depends on Σ and the set S. To compare the two (see Subsection 7.5.4), we assume the irrepresentable condition for **all** sign-vectors, and in fact for all vectors τ with $\|\cdot\|_\infty$-norm at most one.

Definition *We say that the* irrepresentable condition *is met for the set S with cardinality s, if for all vectors $\tau_S \in \mathbb{R}^s$ satisfying $\|\tau_S\|_\infty \leq 1$, we have*

$$\|\Sigma_{2,1}(S)\Sigma_{1,1}^{-1}(S)\tau_S\|_\infty < 1. \tag{7.1}$$

For a fixed $\tau_S \in \mathbb{R}^s$ with $\|\tau_S\|_\infty \leq 1$, the weak irrepresentable condition *holds for τ_S, if*

$$\|\Sigma_{2,1}(S)\Sigma_{1,1}^{-1}(S)\tau_S\|_\infty \leq 1.$$

Moreover, for some $0 < \theta < 1$, the θ-uniform irrepresentable condition *is met for the set S, if*

$$\max_{\|\tau_S\|_\infty \leq 1} \|\Sigma_{2,1}(S)\Sigma_{1,1}^{-1}(S)\tau_S\|_\infty \leq \theta.$$

7.5.2 The KKT conditions

We will frequently make use of the KKT conditions given in Lemma 2.1. In our context, they read as follows:

KKT conditions *We have*

$$2\Sigma(\beta_{\text{init}} - \beta^0) = -\lambda_{\text{init}}\tau_{\text{init}}.$$

Here $\|\tau_{\text{init}}\|_\infty \leq 1$, and moreover

$$\tau_{\text{init},j}1\{\beta_{\text{init},j} \neq 0\} = \text{sign}(\beta_{\text{init},j}), \ j = 1,\ldots,p.$$

We now turn to the noiseless version of the weighted Lasso. Let $W := \text{diag}(w)$, with $w := (w_1,\ldots,w_p)^T$ a diagonal matrix of positive weights. The weighted noiseless Lasso is

$$\beta_{\text{weight}} := \arg\min_\beta \left\{ \|f_\beta - f^0\|^2 + \lambda_{\text{weight}}\lambda_{\text{init}}\|W\beta\|_1 \right\},$$

with active set $S_{\text{weight}} := \{j : \beta_{\text{weight},j} \neq 0\}$.

By the reparametrization $\beta \mapsto \gamma := W\beta$, one sees that the weighted Lasso is a standard Lasso with Gram matrix

$$\Sigma_w := W^{-1}\Sigma W^{-1}.$$

Hence, it inherits all the properties of the standard Lasso. We emphasize however that Σ_w is generally not normalized, i.e., generally $\text{diag}(\Sigma_w) \neq I$. With appropriate weights, this is exactly the strength of the weighted Lasso.

The weighted KKT conditions are:

Weighted KKT conditions *We have*

$$2\Sigma(\beta_{\text{weight}} - \beta^0) = -\lambda_{\text{weight}}\lambda_{\text{init}}W\tau_{\text{weight}}.$$

Here $\|\tau_{\text{weight}}\|_\infty \leq 1$, and moreover

$$\tau_{\text{weight},j}1\{\beta_{\text{weight},j} \neq 0\} = \text{sign}(\beta_{\text{weight},j}), \ j = 1,\ldots,p.$$

Set
$$w_{j,S} := w_j 1\{j \in S\}, \ j = 1,\ldots,p.$$

Note that
$$\|w_S\|_2^2 = \sum_{j\in S} w_j^2,$$

a quantity that will be of importance in our further considerations. We will need conditions on the ratio $\|w_S\|_2/w_{S^c}^{\min}$, where

$$w_{S^c}^{\min} := \min_{j \notin S} w_j.$$

The ratio $\|w_{S_0}\|_2/w_{S_0^c}^{\min}$ should preferably be small, i.e., the weights inside the active set should be small as compared to those outside the active set.

7.5.3 Necessity and sufficiency for variable selection

As we show in the next theorem, the irrepresentable condition for the true active set S_0 is in the noiseless case a sufficient condition for the Lasso to select only variables in the active set S_0 (for the noisy case, see Problem 7.5). We moreover establish that it is essentially a necessary condition. This means that also in the situation with noise, there is no way to get around such a condition (see also Zhao and Yu (2006)).

Let [2]

$$S_0^{\text{relevant}} := \left\{ j : \ |\beta_j^0| > \lambda_{\text{init}} \sup_{\|\tau_{S_0}\|_\infty \le 1} \|\Sigma_{1,1}^{-1}(S_0)\tau_{S_0}\|_\infty/2 \right\}.$$

Theorem 7.1.
Part 1 *Suppose the irrepresentable condition is met for S_0. Then $S_0^{\text{relevant}} \subset S_{\text{init}} \subset S_0$,
and*

$$\|(\beta_{\text{init}})_{S_0} - \beta_{S_0}^0\|_\infty \le \lambda_{\text{init}} \sup_{\|\tau_{S_0}\|_\infty \le 1} \|\Sigma_{1,1}^{-1}(S_0)\tau_{S_0}\|_\infty/2,$$

(so for all $j \in S_0^{\text{relevant}}$, the $\beta_{\text{init},j}$ have the same sign as the β_j^0).
Part 2 *Conversely, suppose that $S_0^{\text{relevant}} = S_0$ and $S_{\text{init}} \subset S_0$. Then the weak irrepresentable condition holds for the sign-vector $\tau_{S_0}^0 := \text{sign}(\beta_{S_0}^0)$.*

Proof of Theorem 7.1.

Part 1 By the KKT conditions, we must have

$$2\Sigma(\beta_{\text{init}} - \beta^0) = -\lambda_{\text{init}}\tau_{\text{init}},$$

where $\|\tau_{\text{init}}\|_\infty \le 1$, and $\tau_{\text{init},j}l\{|\beta_{\text{init},j}| \ne 0\} = \text{sign}(\beta_{\text{init},j})$. This gives

$$2\Sigma_{1,1}(S_0)\left((\beta_{\text{init}})_{S_0} - \beta_{S_0}^0\right) + 2\Sigma_{1,2}(S_0)(\beta_{\text{init}})_{S_0^c} = -\lambda_{\text{init}}(\tau_{\text{init}})_{S_0},$$

$$2\Sigma_{2,1}(S_0)\left((\beta_{\text{init}})_{S_0} - \beta_{S_0}^0\right) + 2\Sigma_{2,2}(S_0)(\beta_{\text{init}})_{S_0^c} = -\lambda_{\text{init}}(\tau_{\text{init}})_{S_0^c}.$$

It follows that

$$2\left((\beta_{\text{init}})_{S_0} - \beta_{S_0}^0\right) + 2\Sigma_{1,1}^{-1}(S_0)\Sigma_{1,2}(S_0)(\beta_{\text{init}})_{S_0^c} = -\lambda_{\text{init}}\Sigma_{1,1}^{-1}(S_0)(\tau_{\text{init}})_{S_0},$$

$$2\Sigma_{2,1}(S_0)\left((\beta_{\text{init}})_{S_0} - \beta_{S_0}^0\right) + 2\Sigma_{2,2}(S_0)(\beta_{\text{init}})_{S_0^c} = -\lambda_{\text{init}}(\tau_{\text{init}})_{S_0^c}$$

(leaving the second equality untouched). Hence, multiplying the first equality by $-(\beta_{\text{init}})_{S_0^c}^T\Sigma_{2,1}(S_0)$, and the second by $-(\beta_{\text{init}})_{S_0^c}^T$,

$$-2(\beta_{\text{init}})_{S_0^c}^T\Sigma_{2,1}(S_0)\left((\beta_{\text{init}})_{S_0} - \beta_{S_0}^0\right) - 2(\beta_{\text{init}})_{S_0^c}^T\Sigma_{2,1}(S_0)\Sigma_{1,1}^{-1}(S_0)\Sigma_{1,2}(S_0)(\beta_{\text{init}})_{S_0^c}$$

$$= \lambda_{\text{init}}(\beta_{\text{init}})_{S_0^c}^T\Sigma_{2,1}(S_0)\Sigma_{1,1}^{-1}(S_0)(\tau_{\text{init}})_{S_0},$$

---

[2] One may invoke the bound $\sup_{\|\tau_{S_0}\|_\infty \le 1} \|\Sigma_{1,1}^{-1}(S_0)\tau_{S_0}\|_\infty \le \sqrt{s_0}\Lambda_{\min}^{-2}(\Sigma_{1,1}(S_0))$. However, there are important examples where the latter bound is too rough.

$$-2(\beta_{\text{init}})_{S_0^c}^T \Sigma_{2,1}(S_0)\left((\beta_{\text{init}})_{S_0} - \beta_{S_0}^0\right) - 2(\beta_{\text{init}})_{S_0^c}^T \Sigma_{2,2}(S_0)(\beta_{\text{init}})_{S_0} = \lambda_{\text{init}}\|(\beta_{\text{init}})_{S_0^c}\|_1,$$

where we invoked that $\beta_{\text{init},j}\tau_{\text{init},j} = |\beta_{\text{init},j}|$. Subtracting the second from the first gives

$$2(\beta_{\text{init}})_{S_0^c}^T \Sigma_{2,2}(S_0)(\beta_{\text{init}})_{S_0^c} - 2(\beta_{\text{init}})_{S_0^c}^T \Sigma_{2,1}(S_0)\Sigma_{1,1}^{-1}(S_0)\Sigma_{1,2}(S_0)(\beta_{\text{init}})_{S_0^c}$$

$$= \lambda_{\text{init}}(\beta_{\text{init}})_{S_0^c}^T \Sigma_{2,1}(S_0)\Sigma_{1,1}^{-1}(S_0)(\tau_{\text{init}})_{S_0} - \lambda_{\text{init}}\|(\beta_{\text{init}})_{S_0^c}\|_1.$$

But by the irrepresentable condition, if $\|(\beta_{\text{init}})_{S_0^c}\|_1 \neq 0$,

$$\left|(\beta_{\text{init}})_{S_0^c}^T \Sigma_{2,1}(S_0)\Sigma_{1,1}^{-1}(S_0)(\tau_{\text{init}})_{S_0}\right| \leq \|(\beta_{\text{init}})_{S_0^c}\|_1 \|\Sigma_{2,1}(S_0)\Sigma_{1,1}^{-1}(S_0)(\tau_{\text{init}})_{S_0}\|_\infty$$

$$< \|(\beta_{\text{init}})_{S_0^c}\|_1.$$

We conclude that if $\|(\beta_{\text{init}})_{S_0^c}\|_1 \neq 0$, then

$$(\beta_{\text{init}})_{S_0^c}^T \Sigma_{2,2}(S_0)(\beta_{\text{init}})_{S_0^c} - (\beta_{\text{init}})_{S_0^c}^T \Sigma_{2,1}(S_0)\Sigma_{1,1}^{-1}(S_0)\Sigma_{1,2}(S_0)(\beta_{\text{init}})_{S_0^c} < 0.$$

The matrix

$$\Sigma_{2,2}(S_0) - \Sigma_{2,1}(S_0)\Sigma_{1,1}^{-1}(S_0)\Sigma_{1,2}(S_0)$$

is positive semi-definite. Hence we arrived at a contradiction. So it must hold that $\|(\beta_{\text{init}})_{S_0^c}\|_1 = 0$, i.e., that $S_{\text{init}} \subset S_0$.

We thus conclude that under the irrepresentable condition, the KKT conditions take the form: for a vector $\tau_{\text{init}} \in \mathbb{R}^p$ with $\|\tau_{\text{init}}\|_\infty \leq 1$, and with $\tau_{\text{init},j}\mathbb{1}\{\beta_{\text{init},j} \neq 0\} = \text{sign}(\beta_{\text{init},j})$,

$$2\Sigma_{1,1}(S_0)\left((\beta_{\text{init}})_{S_0} - \beta_{S_0}^0\right) = -\lambda_{\text{init}}(\tau_{\text{init}})_{S_0},$$

and

$$2\Sigma_{2,1}(S_0)\left((\beta_{\text{init}})_{S_0} - \beta_{S_0}^0\right) = -\lambda_{\text{init}}(\tau_{\text{init}})_{S_0^c}.$$

It follows that

$$\|(\beta_{\text{init}})_{S_0} - \beta_{S_0}^0\|_\infty \leq \lambda_{\text{init}}\|\Sigma_{1,1}^{-1}(S_0)(\tau_{\text{init}})_{S_0}\|_\infty/2$$

$$\leq \lambda_{\text{init}} \sup_{\|\tau_{S_0}\|_\infty \leq 1} \|\Sigma_{1,1}^{-1}(S_0)\tau_{S_0}\|_\infty/2.$$

If $j \in S_0^{\text{relevant}}$ and $\beta_{\text{init},j} = 0$, we would have

$$|\beta_{\text{init},j} - \beta_j^0| = |\beta_j^0| > \lambda_{\text{init}} \sup_{\|\tau_{S_0}\|_\infty \leq 1} \|\Sigma_{1,1}^{-1}(S_0)\tau_{S_0}\|_\infty/2.$$

This is a contradiction. Therefore $j \in S_0^{\text{relevant}}$ implies that $\beta_{\text{init},j} \neq 0$, i.e., that $j \in S_{\text{init}}$. In other words, $S_0^{\text{relevant}} \subset S_{\text{init}}$. In fact, it implies $\text{sign}(\beta_{\text{init},j}) = \text{sign}(\beta_j^0)$ for all $j \in S_0^{\text{relevant}}$.

Part 2 We now show that the weak irrepresentable condition for $\tau_{S_0}^0$ is a necessary condition for variable selection. Suppose that we indeed only select variables in the active set, i.e., that $S_{\text{init}} \subset S_0$. Then $(\beta_{\text{init}})_{S_0^c} \equiv 0$. The KKT conditions then take again the form given above:

$$2\Sigma_{1,1}(S_0)\left((\beta_{\text{init}})_{S_0} - \beta_{S_0}^0 \right) = -\lambda_{\text{init}}(\tau_{\text{init}})_{S_0},$$

and

$$2\Sigma_{2,1}(S_0)\left((\beta_{\text{init}})_{S_0} - \beta_{S_0}^0 \right) = -\lambda_{\text{init}}(\tau_{\text{init}})_{S_0^c}.$$

This implies as before that $\text{sign}(\beta_{\text{init},j}) = \text{sign}(\beta_j^0)$ for all $j \in S_0^{\text{relevant}}$. Hence, as we assumed $S_0 = S_0^{\text{relevant}}$, we have $(\tau_{\text{init}})_{S_0} = \tau_{S_0}^0$. The KKT conditions are thus

$$2\Sigma_{1,1}(S_0)\left((\beta_{\text{init}})_{S_0} - \beta_{S_0}^0 \right) = -\lambda_{\text{init}}\tau_{S_0}^0,$$

and

$$2\Sigma_{2,1}(S_0)\left((\beta_{\text{init}})_{S_0} - \beta_{S_0}^0 \right) = -\lambda_{\text{init}}(\tau_{\text{init}})_{S_0^c}.$$

Hence

$$(\beta_{\text{init}})_{S_0} - \beta_{S_0}^0 = \lambda_{\text{init}}\Sigma_{1,1}^{-1}(S_0)\tau_{S_0}^0/2,$$

and, inserting this in the second KKT-equality,

$$\Sigma_{2,1}(S_0)\Sigma_{1,1}^{-1}(S_0)\tau_{S_0}^0 = (\tau_{\text{init}})_{S_0^c}.$$

But then

$$\|\Sigma_{2,1}(S_0)\Sigma_{1,1}^{-1}(S_0)\tau_{S_0}^0\|_\infty = \|(\tau_{\text{init}})_{S_0^c}\|_\infty \leq 1.$$

$\square$

Corollary 7.1. *Suppose the irrepresentable condition is met for some set S_0. Fix some arbitrary $\beta^0 = \beta_{S_0}^0$ with zeroes outside the set S_0, and let $f^0 := f_{\beta^0} = \sum_{j \in S_0} \psi_j \beta_j^0$. Define, for a fixed $L \geq 0$,*

$$\beta_{\text{primal}} := \arg\min\{\|f_\beta - f^0\| : \|\beta\|_1 \leq L\},$$

i.e., $\|f_{\beta_{\text{primal}}} - f^0\|^2 = \text{LASSO}(f^0, L, \{1, \ldots, p\})$, invoking the notation of Section 6.13. Then for $S_{\text{primal}} := \{\beta_{\text{primal},j} \neq 0\}$, it holds that $S_{\text{primal}} \subset S_0$.

7.5.4 The irrepresentable condition implies the compatibility condition

Theorem 7.1 proves that the irrepresentable condition implies variable selection. One therefore expects it will be more restrictive than the compatibility condition, which only implies a bound for the prediction and estimation error. This turns out to be indeed the case, albeit under the uniform version of the irrepresentable condition.

Recall the compatibility constant

$$\phi_{\text{comp}}^2(L,S) := \min_{\beta}\{s\|f_\beta\|^2 : \|\beta_S\|_1 = 1, \|\beta_{S^c}\|_1 \leq L\}$$

(see Section 6.13).

Theorem 7.2. *Suppose the θ-uniform irrepresentable condition is met for S. Then for $L\theta < 1$,*
$$\phi_{\text{comp}}^2(L,S) \geq (1-L\theta)^2 \Lambda_{\min}^2(\Sigma_{1,1}(S)).$$

Proof of Theorem 7.2. Define

$$\beta^* := \arg\min_{\beta}\{s\|f_\beta\|^2 : \|\beta_S\|_1 = 1, \|\beta_{S^c}\|_1 \leq L\}.$$

Let us write $f^* := f_{\beta^*}$, $f_S^* := f_{\beta_S^*}$ and $f_{S^c}^* := f_{\beta_{S^c}^*}$. Introduce a Lagrange multiplier $\lambda \in \mathbb{R}$. As in Lemma 7.1, there exists a vector τ_S, with $\|\tau_S\|_\infty \leq 1$, such that $\tau_S^T \beta_S^* = \|\beta_S^*\|_1$, and such that

$$\Sigma_{1,1}(S)\beta_S^* + \Sigma_{1,2}(S)\beta_{S^c}^* = -\lambda\tau_S.$$

By multiplying by $(\beta_S^*)^T$, we obtain

$$\|f_S^*\|^2 + (f_S^*, f_{S^c}^*) = -\lambda\|\beta_S^*\|_1.$$

The restriction $\|\beta_S^*\|_1 = 1$ gives

$$\|f_S^*\|^2 + (f_S^*, f_{S^c}^*) = -\lambda.$$

We also have
$$\beta_S^* + \Sigma_{1,1}^{-1}(S)\Sigma_{1,2}(S)\beta_{S^c}^* = -\lambda\Sigma_{1,1}^{-1}\tau_S. \tag{7.2}$$

Hence, by multiplying with τ_S^T,

$$\|\beta_S^*\|_1 + \tau_S^T\Sigma_{1,1}^{-1}(S)\Sigma_{1,2}(S)\beta_{S^c}^* = -\lambda\tau_S^T\Sigma_{1,1}^{-1}\tau_S,$$

or

$$1 = -\tau_S^T\Sigma_{1,1}^{-1}(S)\Sigma_{1,2}(S)\beta_{S^c}^* - \lambda\tau_S^T\Sigma_{1,1}^{-1}(S)\tau_S$$

$$\leq \theta \|\beta_{S^c}^*\|_1 - \lambda \tau_S^T \Sigma_{1,1}^{-1}(S) \tau_S$$

$$\leq \theta L - \lambda \tau_S^T \Sigma_{1,1}^{-1}(S) \tau_S.$$

Here, we applied the θ-uniform irrepresentable condition, and the condition $\|\beta_{S^c}^*\|_1 \leq L$. Thus

$$1 - \theta L \leq -\lambda \tau_S^T \Sigma_{1,1}^{-1}(S) \tau_S.$$

Because $1 - \theta L > 0$ and $\tau_S^T \Sigma_{1,1}^{-1}(S) \tau_S > 0$, this implies that $\lambda < 0$, and in fact that

$$(1 - \theta L) \leq -\lambda s / \Lambda_{\min}^2(\Sigma_{1,1}(S)),$$

where we invoked

$$\tau_S^T \Sigma_{1,1}^{-1}(S) \tau_S \leq \|\tau_S\|_2^2 / \Lambda_{\min}^2(\Sigma_{1,1}(S)) \leq s / \Lambda_{\min}^2(\Sigma_{1,1}(S)).$$

So

$$-\lambda \geq (1 - \theta L) \Lambda_{\min}^2(\Sigma_{1,1}(S)) / s.$$

Continuing with (7.2), we moreover have

$$(\beta_{S^c}^*)^T \Sigma_{2,1}(S) \beta_S^* + (\beta_{S^c}^*)^T \Sigma_{2,1}(S) \Sigma_{1,1}^{-1}(S) \Sigma_{1,2}(S) \beta_{S^c}^*$$

$$= -\lambda (\beta_{S^c}^*)^T \Sigma_{2,1}(S) \Sigma_{1,1}^{-1}(S) \tau_S.$$

In other words,

$$(f_S^*, f_{S^c}^*) + \|(f_{S^c}^*)_S^P\|^2 = -\lambda (\beta_{S^c}^*)^T \Sigma_{2,1}(S) \Sigma_{1,1}^{-1}(S) \tau_S,$$

where $(f_{S^c}^*)_S^P$ is the projection of $f_{S^c}^*$ on the space spanned by $\{\psi_k\}_{k \in S}$. Again, by the θ-uniform irrepresentable condition and by $\|\beta_{S^c}^*\|_1 \leq L$,

$$\left| (\beta_{S^c}^*)^T \Sigma_{2,1}(S) \Sigma_{1,1}^{-1}(S) \tau_S \right| \leq \theta \|\beta_{S^c}^*\|_1 \leq \theta L,$$

so

$$-\lambda (\beta_{S^c}^*)^T \Sigma_{2,1}(S) \Sigma_{1,1}^{-1}(S) \tau_S = |\lambda| (\beta_{S^c}^*)^T \Sigma_{2,1}(S) \Sigma_{1,1}^{-1}(S) \tau_S$$

$$\geq -|\lambda| \left| (\beta_{S^c}^*)^T \Sigma_{2,1}(S) \Sigma_{1,1}^{-1}(S) \tau_S \right| \geq -|\lambda| \theta L = \lambda \theta L.$$

It follows that

$$\|f^*\|^2 = \|f_S^*\|^2 + 2(f_S^*, f_{S^c}^*) + \|f_{S^c}^*\|^2$$

$$= -\lambda + (f_S^*, f_{S^c}^*) + \|f_{S^c}^*\|^2$$

$$\geq -\lambda + (f_S^*, f_{S^c}^*) + \|(f_{S^c}^*)_S^P\|^2 \geq -\lambda + \lambda \theta c = -\lambda(1 - \theta L)$$

$$\geq (1 - \theta L)^2 \Lambda_{\min}^2(\Sigma_{1,1}(S)) / s.$$

$$\square$$

7.5.5 The irrepresentable condition and restricted regression

We recall the definition of the adaptive S-restricted regression (introduced in Subsection 6.13.2, see also the overview in Subsection 6.13.7)

$$\vartheta_{\text{adap}}(S) := \sup_{\|\beta_{S^c}\|_1 \le \sqrt{s}\|\beta_S\|_2} \frac{|(f_{\beta_S}, f_{\beta_{S^c}})|}{\|f_{\beta_S}\|^2}.$$

The adaptive restricted regression was introduced to prove that when $\vartheta_{\text{adap}}(S) < 1/L$, the adaptive (L, S, s)-restricted eigenvalue condition is satisfied for S, with $\phi_{\text{adap}}(L, S, s) \ge (1 - L\vartheta_{\text{adap}}(S))\Lambda_{\min}(\Sigma_{1,1}(S))$ (see Corollary 6.11). We now show that for $L\theta < 1$, the condition $\vartheta_{\text{adap}}(S) < 1/L$ actually implies the θ-uniform irrepresentable condition.

Theorem 7.3. *We have for all $\|\tau_S\|_\infty \le 1$,*

$$\|\Sigma_{2,1}(S)\Sigma_{1,1}^{-1}(S)\tau_S\|_\infty \le \vartheta_{\text{adap}}(S).$$

Proof of Theorem 7.3. First observe that

$$\|\Sigma_{2,1}(S)\Sigma_{1,1}^{-1}(S)\tau_S\|_\infty = \sup_{\|\beta_{S^c}\|_1 \le 1} |\beta_{S^c}^T \Sigma_{2,1}(S)\Sigma_{1,1}^{-1}(S)\tau_S|$$

$$= \sup_{\|\beta_{S^c}\|_1 \le 1} |(f_{\beta_{S^c}}, f_{\beta_S})|,$$

where

$$\beta_S := \Sigma_{1,1}^{-1}(S)\tau_S.$$

We note that

$$\frac{\|f_{\beta_S}\|^2}{\sqrt{s}\|\beta_S\|_2} = \frac{\|\Sigma_{1,1}^{1/2}(S)\beta_S\|_2^2}{\|\Sigma_{1,1}(S)\beta_S\|_2\|\beta_S\|_2} \frac{\|\Sigma_{1,1}(S)\beta_S\|_2}{\sqrt{s}} \le 1.$$

Now, for any constant L,

$$\|\Sigma_{2,1}(S)\Sigma_{1,1}^{-1}(S)\tau_S\|_\infty = \sup_{\|\beta_{S^c}\|_1 \le 1} |(f_{\beta_{S^c}}, f_{\beta_S})|$$

$$= \sup_{\|\beta_{S^c}\|_1 \le L} |(f_{\beta_{S^c}}, f_{\beta_S})|/L.$$

Take $L = \sqrt{s}\|\beta_S\|_2$ to find

$$\|\Sigma_{2,1}(S)\Sigma_{1,1}^{-1}(S)\tau_S\|_\infty = \sup_{\|\beta_{S^c}\|_1 \le \sqrt{s}\|\beta_S\|_2} \frac{|(f_{\beta_{S^c}}, f_{\beta_S})|}{\sqrt{s}\|\beta_s\|_2}$$

$$\leq \sup_{\|\beta_{S^c}\|_1 \leq \sqrt{s}\|\beta_S\|_2} \frac{|(f_{\beta_{S^c}}, f_{\beta_S})|}{\|f_{\beta_s}\|^2}.$$

$\square$

Theorem 7.3 proves the implication

$$\vartheta_{\mathrm{adap}}(S) < 1 \ \Rightarrow \ \text{irrepresentable condition.}$$

The irrepresentable condition is quite restrictive, and illustrates that attempting to prove the compatibility condition by checking whether $\vartheta_{\mathrm{adap}}(S) < 1$, i.e., via the irrepresentable condition, is not the best way to go. Because $\vartheta_{\mathrm{adap}}(S) < 1$ is implied by the coherence condition (with $q = \infty$, see Lemma 6.28), we obtain as a by-product that the irrepresentable conditions follows when correlations are small enough:

Corollary 7.2. *Suppose that for some $\theta \geq 0$,*

$$\frac{\sqrt{s} \max_{j \notin S} \sqrt{\sum_{k \in S} \sigma_{j,k}^2}}{\Lambda_{\min}^2(\Sigma_{1,1}(S))} \leq \theta.$$

Then in view of Corollary 6.13, $\vartheta_{\mathrm{adap}}(S) \leq \theta$ and hence by Theorem 7.3, also

$$\|\Sigma_{2,1}(S)\Sigma_{1,1}^{-1}(S)\tau_S\|_\infty \leq \theta,$$

i.e., for $\theta < 1$ the θ-uniform irrepresentable condition holds.

The next example shows that there are Σ for which the bound given in Theorem 7.3 cannot be improved .

Example 7.1. Let $S_0 = \{1, \ldots, s_0\}$ be the active set, and suppose that

$$\Sigma := \begin{pmatrix} \Sigma_{1,1} & \Sigma_{1,2} \\ \Sigma_{2,1} & \Sigma_{2,2} \end{pmatrix},$$

where $\Sigma_{1,1} := I$ is the $(s_0 \times s_0)$-identity matrix, and

$$\Sigma_{2,1} := \rho(b_2 b_1^T),$$

with $0 \leq \rho < 1$, and with b_1 an s_0-vector and b_2 a $(p - s_0)$-vector, satisfying $\|b_1\|_2 = \|b_2\|_2 = 1$. Moreover, $\Sigma_{2,2}$ is some $((p - s_0) \times (p - s_0))$-matrix, with $\mathrm{diag}(\Sigma_{2,2}) = I$, and with largest eigenvalue $\Lambda_{\max}^2(\Sigma_{2,2})$ and smallest eigenvalue $\Lambda_{\min}^2(\Sigma_{2,2})$.

In Problem 7.2, it is shown that the compatibility condition holds for any S, with $\phi_{\mathrm{comp}}^2(L, S) = \Lambda_{\min}^2(\Sigma_{2,2}) - \rho$. Moreover, for $b_1 := (1, 1, \ldots, 1)^T / \sqrt{s_0}$ and $b_2 := (1, 0, \ldots, 0)^T$, and $\rho > 1/\sqrt{s_0}$, the irrepresentable condition does not hold for S_0. Hence, for example when

$$\Sigma := \begin{pmatrix} 1 & 0 & \cdots & 0 & \rho/\sqrt{s_0} & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & \rho/\sqrt{s_0} & 0 & \cdots & 0 \\ 0 & 0 & \ddots & 0 & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 & \rho/\sqrt{s_0} & 0 & \cdots & 0 \\ \rho/\sqrt{s_0} & \rho/\sqrt{s_0} & \cdots & \rho/\sqrt{s_0} & 1 & \theta & \cdots & \theta \\ 0 & 0 & \cdots & 0 & \theta & 1 & \cdots & \theta \\ \vdots & \vdots & & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & \theta & \theta & \cdots & 1 \end{pmatrix},$$

where $\rho = \theta = 1/4$, then the compatibility condition holds for all S with $\phi_{\mathrm{comp}}(L,S) = 1/2$, the irrepresentable condition does not hold for $s_0 > 16$. We note that the bound for the number of false positives of the initial Lasso, as presented in Lemma 7.2 (Section 7.8.3 ahead), depends on $\Lambda_{\max}^2$. In this example the maximal eigenvalue is $\Lambda_{\max}^2(\Sigma_{2,2})$ of $\Sigma_{2,2}$ is at least as large as $\frac{1}{4}(p - s_0)$.

7.5.6 Selecting a superset of the true active set

Remember that S_0 is defined as the active set of the truth β^0, i.e.,

$$S_0 = \{j : \beta_j^0 \neq 0\}.$$

In Theorem 7.1 Part 1, we have not required a beta-min condition, i.e., for $j \in S_0$, the $|\beta_j^0|$ can be arbitrary small. This means that in fact we may replace S_0 by any set $\mathcal{N}$ containing S_0. If the irrepresentable condition holds for the larger set $\mathcal{N}$, one can conclude by Theorem 7.1 Part 1, that $S_{\mathrm{init}} \subset \mathcal{N}$.

Definition *We say that the (S,N)-irrepresentable condition holds for the set S if for some $\mathcal{N} \supset S$ with size N,*

$$\sup_{\|\tau_{\mathcal{N}}\|_\infty \leq 1} \|\Sigma_{2,1}(\mathcal{N})\Sigma_{1,1}^{-1}(\mathcal{N})\tau_{\mathcal{N}}\|_\infty < 1.$$

Corollary 7.3. *Suppose that the (S_0,N)-irrepresentable condition holds. Then $|S_{\mathrm{init}} \backslash S_0| \leq N - s_0$.*

This approach can be rather useful, as illustrated in the next example.

Example 7.2. We continue with Example 7.1: $S_0 = \{1,\ldots,s_0\}$ and

$$\Sigma := \begin{pmatrix} 1 & 0 & \cdots & 0 & \rho/\sqrt{s_0} & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & \rho/\sqrt{s_0} & 0 & \cdots & 0 \\ 0 & 0 & \ddots & 0 & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 & \rho/\sqrt{s_0} & 0 & \cdots & 0 \\ \rho/\sqrt{s_0} & \rho/\sqrt{s_0} & \cdots & \rho/\sqrt{s_0} & 1 & \theta & \cdots & \theta \\ 0 & 0 & \cdots & 0 & \theta & 1 & \cdots & \theta \\ \vdots & \vdots & & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & \theta & \theta & \cdots & 1 \end{pmatrix}$$

with $\rho < 1 - \theta$ (possibly $\theta = 0$). Variable $s_0 + 1$ is correlated with all the variables in the active set S_0. It will be selected by the Lasso, because it can take care of part of all the coefficients. Now, selecting just one false positive may be thought of as being acceptable. We add variable $s_0 + 1$ to the set S_0:

$$\mathcal{N} := S_0 \cup \{s_0 + 1\}.$$

The set $\mathcal{N}$ obviously satisfies the irrepresentable condition. Hence $S_{\text{init}} \subset \mathcal{N}$.

One may verify that the above example can be extended to having ρs_0 additional variables outside the active set S_0, which is still perhaps an acceptable amount. If the aim is to have no more than $N - s_0$ false positives, one may accomplish this by requiring the (S_0, N)-irrepresentable condition. A choice N proportional to s_0 seems reasonable. (In Lemma 7.2 and Lemma 7.3 respectively, we show by different means that the Lasso has no more than an order of magnitude $O(s_0)$ of false positives, provided $\Lambda_{\max}$ remains bounded, respectively certain rather severe sparse eigenvalue conditions hold.)

7.5.7 The weighted irrepresentable condition

Recall that $W = \text{diag}(w_1, \ldots, w_p)$ is a matrix of positive weights. Let

$$W_S := W_{1,1}(S), \ W_{S^c} := W_{2,2}(S).$$

Definition *We say that the* weighted irrepresentable condition *holds for S if for all vectors $\tau_S \in \mathbb{R}^s$ with $\|\tau_S\|_\infty \le 1$, one has*

$$\|W_{S^c}^{-1} \Sigma_{2,1}(S) \Sigma_{1,1}^{-1}(S) W_S \tau_S\|_\infty < 1.$$

The weak weighted irrepresentable condition *holds for a fixed τ_S with $\|\tau_S\|_\infty \le 1$, if*

$$\|W_{S^c}^{-1} \Sigma_{2,1}(S) \Sigma_{1,1}^{-1}(S) W_S \tau_S\|_\infty \le 1.$$

Define

$$S_{0,W}^{\text{relevant}} := \left\{ j : |\beta_j^0| > \lambda_{\text{weight}} \lambda_{\text{init}} \sup_{\|\tau_{S_0}\|_\infty \leq 1} \|\Sigma_{1,1}^{-1}(S_0) W_{S_0} \tau_{S_0}\|_\infty / 2 \right\}.$$

The reparametrization $\beta \mapsto \gamma := W^{-1}\beta$ leads to the following corollary, which is the weighted variant of Theorem 7.1.

Corollary 7.4.
Part 1 *Suppose the weighted irrepresentable condition is met for S_0. Then $S_{0,W}^{\text{relevant}} \subset S_{\text{weight}} \subset S$, and*

$$\|(\beta_{\text{weight}})_{S_0} - \beta_{S_0}^0\|_\infty \leq \lambda_{\text{weight}} \lambda_{\text{init}} \sup_{\|\tau_{S_0}\|_\infty \leq 1} \|\Sigma_{1,1}^{-1}(S_0) W_{S_0} \tau_{S_0}\|_\infty / 2.$$

Part 2 *Conversely, if $S_0 = S_{0,W}^{\text{relevant}}$ and $S_{\text{weight}} \subset S_0$, then the weak weighted irrepresentable condition holds for $\tau_{S_0}^0$, where $\tau_{S_0}^0 := \text{sign}(\beta_{S_0}^0)$.*

7.5.8 The weighted irrepresentable condition and restricted regression

The weighted irrepresentable condition can be linked to the (unweighted) adaptive restricted regression (a weighted variant of Theorem 7.3), as follows.

Theorem 7.4.

$$\sup_{\|\tau_S\|_\infty \leq 1} \|W_{S^c}^{-1} \Sigma_{2,1}(S) \Sigma_{1,1}^{-1}(S) W_S \tau_S\|_\infty \leq \frac{\|w_S\|_2}{\sqrt{s w_{S^c}^{\min}}} \vartheta_{\text{adap}}(S).$$

Proof of Theorem 7.4. Clearly,

$$\|W_{S^c}^{-1} \Sigma_{2,1}(S) \Sigma_{1,1}^{-1}(S) W_S \tau_S\|_\infty \leq \|\Sigma_{2,1}(S) \Sigma_{1,1}^{-1}(S) W_S \tau_S\|_\infty / w_{S^c}^{\min}.$$

Define

$$\beta_S := \Sigma_{1,1}^{-1}(S) W_S \tau_S.$$

Then

$$\|W_{S^c}^{-1} \Sigma_{2,1}(S) \Sigma_{1,1}^{-1}(S) W_S \tau_S\|_\infty = \sup_{\|\gamma_{S^c}\|_1 \leq 1} |\gamma_{S^c}^T W_{S^c}^{-1} \Sigma_{2,1}(S) \Sigma_{1,1}^{-1}(S) W_S \tau_S|$$

$$= \sup_{\|W_{S^c} \beta_{S^c}\|_1 \leq 1} |\beta_{S^c}^T \Sigma_{2,1}(S) \beta_S| = \sup_{\|W_{S^c} \beta_{S^c}\|_1 \leq 1} |(f_{\beta_{S^c}}, f_{\beta_S})|$$

$$\leq \sup_{\|\beta_{S^c}\|_1 \leq 1/w_{S^c}^{\min}} |(f_{\beta_{S^c}}, f_{\beta_S})|$$

$$= \sup_{\|\beta_{S^c}\|_1 \leq \|w_S\|_2 \|\beta_S\|_2 / w_{S^c}^{\min}} \frac{|(f_{\beta_{S^c}}, f_{\beta_S})|}{\|w_S\|_2 \|\beta_S\|_2}$$

$$= \sup_{\|\beta_{S^c}\|_1 \leq \|w_S\|_2 \|\beta_S\|_2 / w_{S^c}^{\min}} \frac{|(f_{\beta_{S^c}}, f_{\beta_S})|}{\|f_{\beta_S}\|^2} \frac{\|f_{\beta_S}\|^2}{\|w_S\|_2 \|\beta_S\|_2}.$$

But

$$\frac{\|f_{\beta_S}\|^2}{\|w_S\|_2 \|\beta_S\|_2} = \frac{\tau_S^T W_S \Sigma_{1,1}^{-1}(S) W_S \tau_S}{\sqrt{\tau_S^T W_S^2 \tau_S} \sqrt{\tau_S W_S \Sigma_{1,1}^{-2}(S) W_S \tau_S}} \frac{\|W_S \tau_S\|_2}{\|w_S\|_2} \leq 1.$$

We conclude that

$$\|W_{S^c}^{-1} \Sigma_{2,1}(S) \Sigma_{1,1}^{-1}(S) W_S \tau_S\|_\infty \leq \sup_{\|\beta_{S^c}\|_1 \leq \|w_S\|_2 \|\beta_S\|_2 / w_{S^c}^{\min}} \frac{|(f_{\beta_{S^c}}, f_{\beta_S})|}{\|f_{\beta_S}\|^2}$$

$$= \frac{\|w_S\|_2}{\sqrt{s} w_{S^c}^{\min}} \vartheta_{\text{adap}}(S).$$

$\square$

Corollary 7.5. *Suppose that*

$$\frac{\|w_{S_0}\|_2}{\sqrt{s_0} w_{S_0^c}^{\min}} \vartheta_{\text{adap}}(S_0) < 1.$$

Then, by Corollary 7.4, $S_{\text{weight}} \subset S_0$. In particular, if $\Sigma_{j,j} \leq 1$ for all j, then, inserting the general bound $\sqrt{s_0}/\Lambda_{\min}(\Sigma_{1,1}(S_0))$ for $\vartheta_{\text{adap}}(S_0)$ (see Lemma 6.27) gives that the inequality

$$\|w_{S_0}\|_2 / w_{S_0^c}^{\min} < \Lambda_{\min}(\Sigma_{1,1}(S_0)), \tag{7.3}$$

implies $S_{\text{weight}} \subset S_0$.

Example 7.3. Take Σ as in Example 7.1, with b_1 changed to $b_1 = w_{S_0}/\|w_{S_0}\|_2$, and b_2 to $b_2 = (0, \ldots, 1, 0, \ldots)^T$, where the 1 is placed at $\arg\min_{j \in S_0^c} w_j$. Then

$$\sup_{\|\tau_{S_0}\|_\infty \leq 1} \|\Sigma_{2,1} \Sigma_{1,1}^{-1} \tau_{S_0}\|_\infty = \rho \|w_{S_0}\|_1 / \|w_{S_0}\|_2,$$

and furthermore,

$$\sup_{\|\tau_{S_0}\|_\infty \leq 1} \|W_{S_0^c}^{-1} \Sigma_{2,1} \Sigma_{1,1}^{-1} W_{S_0} \tau_{S_0}\|_\infty = \rho \|w_{S_0}\|_2 / w_{S_0^c}^{\min}.$$

The above example shows that there exist Gram matrices Σ which satisfy the compatibility condition with, for all $L > 0$, $\phi_{\text{comp}}^2(L, S_0) = 1 - \rho$, where $\rho \in (0, 1)$, and where the adaptive Lasso needs the separation

$$\|w_{S_0}\|_2/w_{S_0^c}^{\min} \le 1/\rho. \tag{7.4}$$

to perform variable selection. Roughly speaking, this means that the weights in S_0^c should be an order of magnitude $\sqrt{s_0}$ larger than the weights in S_0. See also Corollary 7.8, where the same amount of separation is required for the adaptive Lasso to perform exact variable selection.

We now know from Example 7.3 that a compatibility condition alone does not suffice for proving variable selection with the weighted Lasso. If one aims at substantially relaxing the bound (7.3), one needs more restrictions on Σ, for example as in Theorem 7.4, with the adaptive restricted regression $\vartheta_{\text{adap}}(S_0)$ much less than the generic bound $\sqrt{s_0}/\Lambda_{\min}(\Sigma_{1,1}(S_0))$ of Lemma 6.27.

7.5.9 The weighted Lasso with "ideal" weights

In the case of the adaptive Lasso, the "ideal" weights that we target at have $w_j^0 := 1/|\beta_j^0|$, $j \in S_0$. With "ideal" weights w^0, we obviously have that

$$\|w_{S_0}^0\|_2^2 := s_0/|\beta^0|_{\text{harm}}^2,$$

where

$$|\beta^0|_{\text{harm}}^2 := \left(\frac{1}{s_0} \sum_{j \in S_0} \frac{1}{|\beta_j^0|^2}\right)^{-1}$$

is the harmonic mean of the squared non-zero coefficients.

Example 7.3 in Subsection 7.5.8 proves that the bound (7.3) is also a lower bound. We show in Lemma 7.1 below that with the "ideal" weights in the active set S_0, and with outside the active set $w_j = 1/|\beta_j^{\text{init}}|$, $j \notin S_0$, for some initial β^{init}, the inequality (7.3) implies that $\|\beta_{S_0^c}^{\text{int}}\|_\infty$ has to be of order $\|f^0\|/s_0$.

Lemma 7.1. *Let us take* $w_j = w_j^0$, $j \in S_0$ *and* $w_j = 1/|\beta_j^{\text{init}}|$, $j \notin S_0$, *where* β^{init} *is some initial estimator of* β. *Suppose moreover that the condition (7.3) (which is sufficient for having no false positives) holds. Then*

$$\|\beta_{S_0^c}^{\text{int}}\|_\infty < \|f^0\|/s_0.$$

Proof. It is clear from the Cauchy-Schwarz inequality that

$$s_0 \le \|w_{S_0}^0\|_2 \|\beta^0\|_2.$$

Moreover,

$$\|\beta^0\|_2 \leq \|f^0\|/\Lambda_{\min}(\Sigma_{1,1}(S_0)).$$

Hence we get

$$\|w^0_{S_0}\|_2 \geq s_0 \Lambda_{\min}(\Sigma_{1,1}(S_0))/\|f^0\|.$$

Condition (7.3) now gives

$$\frac{1}{w^{\min}_{S^c_0}} < \Lambda_{\min}(\Sigma_{1,1}(S_0))/\|w_{S_0}\| \leq \|f^0\|/s_0.$$

$\square$

Remark 7.1. With the Lasso, we get an initial estimator β_{init} satisfying

$$\|(\beta_{\text{init}})_{S^c_0}\|_\infty = O(\lambda_{\text{init}}\sqrt{s_0})$$

(see Lemma 7.6). Unless this bound can be improved, and for $\lambda_{\text{init}} \asymp \sigma\sqrt{\log p/n}$, where $\sigma \asymp \|f^0\|$, we thus need that $s_0 = O((n/\log p)^{1/3})$ for variable selection with the adaptive Lasso in the worst case scenario.

7.6 Definition of the adaptive and thresholded Lasso

We now return to the noisy case. We will use the standard Lasso as initial estimator for the second stage adaptive Lasso, and write this initial estimator as

$$\hat{\beta}_{\text{init}} := \arg\min_\beta \left\{ \frac{1}{n}\sum_{i=1}^n (Y_i - f_\beta(X_i))^2 + \lambda_{\text{init}}\|\beta\|_1 \right\},$$

where $f_\beta := \sum_{j=1}^p \beta_j \psi_j$. We let $\hat{f}_{\text{init}} := f_{\hat{\beta}_{\text{init}}}$. The active set of the estimator $\hat{\beta}_{\text{init}}$ is $\hat{S}_{\text{init}} := \{j : \hat{\beta}_{\text{init},j} \neq 0\}$, which has cardinality $\hat{s}_{\text{init}} = |\hat{S}_{\text{init}}|$.

7.6.1 Definition of adaptive Lasso

The adaptive Lasso (Zou (2006)) is defined as

$$\hat{\beta}_{\text{adap}} := \arg\min_\beta \left\{ \frac{1}{n}\sum_{i=1}^n (Y_i - f_\beta(X_i))^2 + \lambda_{\text{adap}}\lambda_{\text{init}}\sum_{j=1}^p w_j|\beta_j| \right\},$$

where the weights $\{w_j\}$ are functions of the initial estimator $\hat{\beta}_{\text{init}}$. The most commonly applied adaptive Lasso takes

$$w_j := \frac{1}{|\hat{\beta}_{\text{init},j}|}, \; j = 1, \ldots, p,$$

with the convention that when $\hat{\beta}_{\text{init},j} = 0$ (i.e., when $w_j = \infty$), the j-th variable is excluded in this second stage. In practice, there may be a threshold, or precision level, $\lambda_{\text{precision}} > 0$, that is, all variables j with $|\hat{\beta}_{\text{init},j}| \leq \lambda_{\text{init}}$ are excluded in the second stage:

$$w_j := \frac{1}{|\hat{\beta}_{\text{init},j}|\mathbb{1}\{|\hat{\beta}_{\text{init},j}| > \lambda_{\text{precision}}\}}, \; j = 1, \ldots, p.$$

This can be thought of as the *ruthless* Lasso.

Another procedure is to give the variables j that are not selected in the first stage, a second chance in the second stage. We then take

$$w_j = \frac{1}{|\hat{\beta}_{\text{init},j}| \vee \lambda_{\text{precision}}}, \; j = 1, \ldots, p.$$

We call this the *conservative* Lasso.

In the sequel, we will only consider the standard adaptive Lasso with weights $w_j = 1/|\hat{\beta}_{\text{init},j}|$ for all j. We write $\hat{f}_{\text{adap}} := f_{\hat{\beta}_{\text{adap}}}$, with active set $\hat{S}_{\text{adap}} := \{j : \hat{\beta}_{\text{adap},j} \neq 0\}$, respectively.

7.6.2 Definition of the thresholded Lasso

Another possibility is the thresholded Lasso with refitting. Define

$$\hat{S}_{\text{thres}} := \{j : |\hat{\beta}_{\text{init},j}| > \lambda_{\text{thres}}\}, \tag{7.5}$$

which is the set of variables having estimated coefficients larger than some given threshold λ_{thres}. The refitting is then done by ordinary least squares:

$$\hat{b}_{\text{thres}} = \arg\min_{\beta = \beta_{\hat{S}_{\text{thres}}}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - f_\beta(X_i))^2.$$

We write

$$\hat{f}_{\text{thres}} := f_{\hat{b}_{\text{thres}}}.$$

7.6.3 Order symbols

The behavior of the Lasso, the thresholded Lasso and adaptive Lasso depends on the tuning parameters, on the design, as well as on the true f^0, and actually on the interplay between these quantities. To keep the exposition clear, we will use order symbols. Our expressions are functions of n, p, $\{\psi_j(X_i), \; j = 1, \dots, p, \; i = 1, \dots, n\}$, and f^0, and also of the tuning parameters λ_{init}, λ_{thres}, and λ_{adap}. For positive functions g and h, we say that $g = O(h)$ if $\|g/h\|_\infty$ is bounded, and $g \asymp h$ if in addition $\|h/g\|_\infty$ is bounded. Moreover, we say that $g = O_{\text{suff}}(h)$ if $\|g/h\|_\infty$ is not larger than a suitably chosen sufficiently small constant, and $g \asymp_{\text{suff}} h$ if in addition $\|h/g\|_\infty$ is bounded.

7.7 A recollection of the results obtained in Chapter 6

The initial Lasso estimator is studied in Section 6.2. and we recall the results here. We combine it with Lemma 6.10, which gives conditions for convergence in ℓ_2. The definition of the compatibility constant $\phi_{\text{comp}}(L,S)$ and of the (minimal adaptive) restricted eigenvalue $\phi(L,S,N)$ ($\phi_{\text{min}}(L,S,N)$) is given in Section 6.13 (see Subsection 6.13.7 for an overview). We write $\phi_{\text{comp}}(L,S) := \phi_{\text{comp},\hat{\Sigma}}(L,S)$, and $\phi(L,S,N) = \phi_{\hat{\Sigma}}(L,S,N)$, and so on. We recall the approximation results of Section 6.12, that is, for a Σ "close enough" to $\hat{\Sigma}$, the $\hat{\Sigma}$-compatibility constants and (minimal adaptive) $\hat{\Sigma}$-restricted eigenvalues inherit their properties from the Σ-counterparts. as long as there is enough sparseness.

Let

$$\mathscr{T} := \left\{ \max_{1 \le j \le p} 2|(\varepsilon, \psi_j)_n| \le \lambda_0 \right\},$$

be the set where the correlation between noise and covariables does not exceed a suitable value λ_0 (the "noise level").

Typically, λ_0 can be taken of order $\sqrt{\log p / n}$. Lemma 6.2 which has normally distributed errors, serves as an example, but the results can clearly be extended to other distributions.

Theorem 7.5. *Suppose the $(3, S_0)$-compatibility condition holds, with constant $\phi_{\text{comp}}(3, S_0)$. Assume moreover we are on $\mathscr{T}$, and that $\lambda_{\text{init}} \ge 2\lambda_0$. Then*

$$\|\hat{f}_{\text{init}} - f^0\|_n^2 + \lambda_{\text{init}}\|\hat{\beta}_{\text{init}} - \beta^0\|_1 \le 4\lambda_{\text{init}}^2 s_0 / \phi_{\text{comp}}^2(3, S_0).$$

If moreover the $(3, S_0, 2s_0)$-restricted eigenvalue condition holds, with restricted eigenvalue $\phi(3, S_0, 2s_0)$, then

$$\|\hat{\beta}_{\text{init}} - \beta^0\|_2^2 \leq \frac{20\lambda_{\text{init}}^2 s_0}{\phi^4(3, S_0, 2s_0)}.$$

See Theorem 6.1 and Lemma 6.10.

One immediately obtains the following screening corollary.

Corollary 7.6. *(**Screening Corollary**) Suppose the beta-min condition*

$$|\beta^0|_{\min} > 4\lambda_{\text{init}} s_0 / \phi_{\text{comp}}^2(3, S_0), \tag{7.6}$$

or alternatively, the beta-min condition

$$|\beta^0|_{\min} > 2\lambda_{\text{init}} \sqrt{5s_0} / \phi^2(3, S_0, 2s_0). \tag{7.7}$$

Then on $\mathcal{T}$, $\hat{S}_{\text{init}} \supset S_0$.

The screening property is discussed in e.g. Section 2.5, and also in Chapter 11 where it is used for obtaining asymptotically correct p-values by (multi) sample splitting. It holds for the one stage Lasso if the truth has large enough coefficients. If $1/\phi(3, S_0, 2s_0) = O(1)$, Corollary 7.6 assumes an order of magnitude $\lambda_{\text{init}} \sqrt{s_0}$ for the smallest coefficient, whereas signal-to-noise arguments, with noise variance σ^2, say that the smallest coefficient cannot be larger than $\sigma/\sqrt{s_0}$ in order of magnitude. We conclude that the beta-min condition (7.7) implicitly assumes that $s_0 = O(\sigma/\lambda_{\text{init}})$. With $\lambda_{\text{init}} \asymp \sigma\sqrt{\log p/n}$, this means $s_0 = O(\sqrt{n/\log p})$. By the same arguments, beta-min condition (7.6) means $s_0 = O((n/\log p)^{1/3})$ (compare with Remark 7.1 in Subsection 7.5.9).

We now leave the beta-min conditions aside, and consider a sparse approximation of β^0. Indeed, the sparse object to recover may not be the "true" unknown parameter β^0 of the linear regression. It may well be that many of the $|\beta_j^0|$ are non-zero, but very small. Thus, its active set

$$S_0 = \{j : \beta_j^0 \neq 0\}$$

can be quite large, and not the set we want to recover. More generally, we aim at recovering a sparse approximation f^* of the regression f^0, when f^0 itself is not necessarily sparse. Our proposal will be to target at the sparse approximation that trades off the number of non-zero coefficients against fit. This target was introduced in Chapter 6, Section 6.2.3. We recall its definition here.

Given a set of indices $S \subset \{1, \dots, p\}$, the best approximation of f^0 using only variables in S is

$$f_S = f_{b^S} := \arg\min_{f = f_{\beta_S}} \|f - f^0\|_n,$$

that is, f_S is the projection of f^0 on the span of the variables in S. Our target is now the projection $f^* := f_{S_*}$, where

$$S_* := \arg\min_{S \subset S_0} \left\{ \|\mathbf{f}_S - f^0\|_n^2 + 7\lambda_{\text{init}}^2 |S|/\phi_{\text{comp}}^2(6,S) \right\}.$$

This is in relation with the oracle result of Theorem 6.2, although we have changed the constants (this is only due to some inconsistencies over the chapters in the choice of the constants). Alternatively, one could insert Lemma 6.12 in Section 6.10, but the latter (as price for considering general weights) has $\phi_{\text{comp}}(3,S)$ replaced by $\phi_{\text{adap}}(6,S,|S|) = \phi_{\text{min}}(6,S,|S|)$.

We minimize here over all $S \subset S_0$, so that the oracle is not allowed to trade non-zero coefficients against compatibility constants. This facilitates the interpretation.

To simplify the expressions, we assume moreover throughout that

$$\|f^* - f^0\|_n^2 = O(\lambda_{\text{init}}^2 s_* / \phi_{\text{comp}}^2(3,S_*)) \tag{7.8}$$

which roughly says that the oracle "squared bias" term is not substantially larger than the oracle "variance" term. For example, in the case of orthonormal design, this condition holds if the small non-zero coefficients are small enough, or if there are not too many of them, i.e., if

$$\sum_{|\beta_j^0|^2 \le 7\lambda_{\text{init}}^2} |\beta_j^0|^2 = O(\lambda_{\text{init}}^2 s_*).$$

We stress that (7.8) is merely to write order bounds for the oracle, which we compare to the ones for the various Lasso versions. If actually the "squared bias" term is the strictly dominating term, this does not alter the theory but makes the presentation less transparent.

As before, we refer to f^* as the "oracle". The set S_* is called the active set of f^*, and $\beta^* := b^{S_*}$ are its coefficients, i.e., $f^* = f_{\beta^*}$.

We assume that S_* has a relatively small number $s_* := |S_*|$ of nonzero coefficients. Inferring the sparsity pattern, i.e. variable selection, refers to the task of correctly estimating the support set S_*, or more modestly, to have a limited number of false positives (type I errors) and false negatives (type II errors)[3]. It can be verified that under reasonable conditions (e.g. i.i.d. standard Gaussian noise and properly chosen tuning parameter λ) the "ideal" estimator (with ℓ_0-penalty)

$$\hat{\beta}_{\text{ideal}} := \arg\min_{\beta} \left\{ \|\mathbf{Y} - f_\beta\|_2^2/n + \lambda^2 |\{j: \beta_j \ne 0\}| \right\},$$

has $O(s_*)$ false positives (see for instance Barron et al. (1999) and van de Geer (2001)). With this in mind, we generally aim at $O(s_*)$ false positives (see also Zhou (2010)), yet keeping the prediction error as small as possible.

---

[3] Recall that, for a generic estimator $\hat{\beta}$ and active set S_*, a type I error is a non-zero estimate $\hat{\beta}_j$ when $j \notin S_*$. A type II error occurs when $\hat{\beta}_j = 0$ while $j \in S_*$.

Theorem 7.6. *Suppose* $\lambda_{\text{init}} \geq 2\lambda_0$. *Let*

$$\delta^2_{\text{comp}} := \|f^* - f^0\|^2_n + \frac{7\lambda^2_{\text{init}}s_*}{\phi^2_{\text{comp}}(3, S_*)},$$

and

$$\delta^2_{\text{min}} := \|f^* - f^0\|^2_n + \frac{7\lambda^2_{\text{init}}s_*}{\phi^2_{\text{min}}(6, S_*, 2s_*)}.$$

On $\mathscr{T}$,

$$\|\hat{f}_{\text{init}} - f^0\|^2_n \leq 2\delta^2_{\text{comp}},$$

and

$$\|\hat{\beta}_{\text{init}} - \beta^*\|_1 \leq 5\delta^2_{\text{comp}}/\lambda_{\text{init}},$$

and

$$\|\hat{\beta}_{\text{init}} - \beta^*\|_2 \leq 10\delta^2_{\text{min}}/(\lambda_{\text{init}}\sqrt{s_*}).$$

Theorem 6.2 contains (modulo the alternative constants) the first part of Theorem 7.6 above. We moreover combined it with Lemma 6.9, much in the spirit of Lemma 6.11. The ℓ_2-result is the same as in Corollary 6.5, with slightly improved constants exploiting the specific situation.

Note that $\phi_{\text{min}}(6, S_*, 2s_*)$ is always at least as small as $\phi_{\text{comp}}(6, S_*)$, so that $\delta_{\text{comp}} \leq \delta_{\text{min}}$.

With order symbols, the important quantities become more visible:

Theorem 7.7. *Let* $\lambda_{\text{init}} \geq 2\lambda_0$. *We have on* $\mathscr{T}$,

$$\|\hat{f}_{\text{init}} - f^0\|^2_n = \left[\frac{1}{\phi^2_{\text{comp}}(6, S_*)}\right] O(\lambda^2_{\text{init}}s_*),$$

and

$$\|\hat{\beta}_{\text{init}} - \beta^*\|_1 = \left[\frac{1}{\phi^2_{\text{comp}}(6, S_*)}\right] O(\lambda_{\text{init}}s_*),$$

and

$$\|\hat{\beta}_{\text{init}} - \beta^*\|_2 = \left[\frac{1}{\phi^2_{\text{min}}(6, S_*, 2s_*)}\right] O(\lambda_{\text{init}}\sqrt{s_*}).$$

We did not yet present a bound for the number of false positives of the initial Lasso: it can be quite large (see Problem 7.7) depending on further conditions as given in Lemma 7.3. A general bound is presented in Lemma 7.2.

7.8 The adaptive Lasso and thresholding: invoking sparse eigenvalues

In this section, we use maximal sparse eigenvalues $\Lambda_{\max}(N)$ and uniform eigenvalues $\Lambda_{\min}(S,N)$, as well as minimal adaptive restricted eigenvalues $\phi_{\min}(S,N)$ and $\phi_{\text{varmin}}(S,N)$. See Subsection 6.13.7 for their definition.

7.8.1 The conditions on the tuning parameters

The following conditions play an important role. Conditions A and AA for thresholding are similar to those in Zhou (2010) (Theorems 1.2, 1.3 and 1.4).

Condition A *For the thresholded Lasso, the threshold level λ_{thres} is chosen sufficiently large, in such a way that*

$$\left[\frac{1}{\phi_{\min}^2(6,S_*,2s_*)}\right]\lambda_{\text{init}} = O_{\text{suff}}(\lambda_{\text{thres}}).$$

Condition AA *For the thresholded Lasso, the threshold level λ_{thres} is chosen sufficiently large, but such that*

$$\left[\frac{1}{\phi_{\min}^2(6,S_*,2s_*)}\right]\lambda_{\text{init}} \asymp_{\text{suff}} \lambda_{\text{thres}}.$$

Condition B *For the adaptive Lasso, the tuning parameter λ_{adap} is chosen sufficiently large, in such a way that*

$$\left[\frac{\Lambda_{\max}(s_*)}{\phi_{\text{varmin}}^3(6,S_*,2s_*)}\right]\lambda_{\text{init}} = O_{\text{suff}}(\lambda_{\text{adap}}).$$

Condition BB *For the adaptive Lasso, the tuning parameter λ_{adap} is chosen sufficiently large, but such that*

$$\left[\frac{\Lambda_{\max}(s_*)}{\phi_{\text{varmin}}^3(6,S_*,2s_*)}\right]\lambda_{\text{init}} \asymp_{\text{suff}} \lambda_{\text{adap}}.$$

Remark 7.2. Note that our conditions on λ_{thres} and λ_{adap} depend on the ϕ's and Λ's, which are unknown. Indeed, our study is of theoretical nature, revealing common

features of thresholding and the adaptive Lasso. Furthermore, it is possible to re-move the dependence of the ϕ's and Λ's, when one imposes stronger sparse eigen-value conditions, along the lines of Zhang and Huang (2008). In practice, the tuning parameters are generally chosen by cross validation.

The above conditions can be considered with a zoomed-out look, neglecting the expressions in the square brackets ($[\cdots]$), and a zoomed-in look, taking into account what is inside the square brackets. One may think of λ_{init} as the noise level. Zooming out, Conditions A and B say that the threshold level λ_{thres} and the tuning parameter λ_{adap} are required to be at least of the same order as λ_{init}, i.e., they should not drop below the noise level. Assumption AA and BB put these parameters exactly at the noise level, i.e., at the smallest value we allow. The reason to do this is that one then can have good prediction and estimation bounds. If we zoom in, we see in the square brackets the role played by the various eigenvalues. It is at first reading perhaps easiest to remember that the ϕ's can be small and the $\Lambda_{\text{max}}(\cdot)$'s can be large, but one hopes they behave well, in the sense that the values in the square brackets are not too large.

7.8.2 The results

The next two theorems contain the main ingredients of the present section. We first discuss thresholding. The results correspond to those in Zhou (2010), and will be invoked to prove similar bounds for the adaptive Lasso, as presented in Theorem 7.9.

The set $\mathcal{T}$ is throughout the set defined in Section 7.7, that is

$$\mathcal{T} := \left\{ \max_{1 \le j \le p} 2|(\varepsilon, \psi_j)_n| \le \lambda_0 \right\}.$$

Theorem 7.8. *Let* $\lambda_{\text{init}} \ge 2\lambda_0$. *Suppose Condition A holds. Then on* $\mathcal{T}$,

$$\|\hat{f}_{\text{thres}} - f^0\|_n^2 = \left[\Lambda_{\text{max}}^2(s_*) \right] \frac{\lambda_{\text{thres}}^2}{\lambda_{\text{init}}^2} O(\lambda_{\text{init}}^2 s_*),$$

and

$$\|\hat{b}_{\text{thres}} - \beta^*\|_1 = \left[\frac{\Lambda_{\text{max}}(s_*)}{\Lambda_{\text{min}}(S_*, 2s_*)} \right] \frac{\lambda_{\text{thres}}}{\lambda_{\text{init}}} O(\lambda_{\text{init}} s_*),$$

and

$$\|\hat{b}_{\text{thres}} - \beta^*\|_2 = \left[\frac{\Lambda_{\text{max}}(s_*)}{\Lambda_{\text{min}}(S_*, 2s_*)} \right] \frac{\lambda_{\text{thres}}}{\lambda_{\text{init}}} O(\lambda_{\text{init}} \sqrt{s_*}),$$

and

$$|\hat{S}_{\text{thres}} \backslash S_*| = \left[\frac{1}{\phi_{\min}^4(6, S_*, 2s_*)}\right] \frac{\lambda_{\text{init}}^2}{\lambda_{\text{thres}}^2} O(s_*).$$

Theorem 7.9. *Suppose Condition B holds. Then on $\mathscr{T}$,*

$$\|\hat{f}_{\text{adap}} - f^0\|_n^2 = \left[\frac{\Lambda_{\max}(s_*)}{\phi_{\text{varmin}}(6, S_*, 2s_*)}\right] \frac{\lambda_{\text{adap}}}{\lambda_{\text{init}}} O(\lambda_{\text{init}}^2 s_*),$$

and

$$\|\hat{\beta}_{\text{adap}} - \beta^*\|_1 = \left[\frac{\Lambda_{\max}^{1/2}(s_*)}{\phi_{\text{varmin}}^{3/2}(6, S_*, 2s_*)}\right] \sqrt{\frac{\lambda_{\text{adap}}}{\lambda_{\text{init}}}} O(\lambda_{\text{init}} s_*),$$

and

$$\|\hat{\beta}_{\text{adap}} - \beta^*\|_2 = \left[\frac{\Lambda_{\max}^{1/2}(s_*)\phi_{\text{varmin}}^{1/2}(6, S_*, 2s_*)}{\phi_{\text{varmin}}^2(6, S_*, 3s_*)}\right] \sqrt{\frac{\lambda_{\text{adap}}}{\lambda_{\text{init}}}} O(\lambda_{\text{init}} \sqrt{s_*}),$$

and

$$|\hat{S}_{\text{adap}} \backslash S_*| = \left[\frac{\Lambda_{\max}^2(s_*)}{\phi_{\min}^4(6, S_*, 2s_*)} \frac{\Lambda_{\max}(s_*)}{\phi_{\text{varmin}}(6, S_*, 2s_*)}\right] \frac{\lambda_{\text{init}}}{\lambda_{\text{adap}}} O(s_*).$$

Theorems 7.8 and 7.9 show how the results depend on the choice of the tuning parameters λ_{thres} and λ_{adap}. The following corollary takes the choices of Conditions AA and BB, as these choices give the smallest prediction and estimation error.

Corollary 7.7. *Suppose we are on $\mathscr{T}$. Then, under Condition AA,*

$$\|\hat{f}_{\text{thres}} - f^0\|_n^2 = \left[\frac{\Lambda_{\max}^2(s_*)}{\phi_{\min}^4(6, S_*, 2s_*)}\right] O(\lambda_{\text{init}}^2 s_*), \tag{7.9}$$

and

$$\|\hat{b}_{\text{thres}} - \beta^*\|_1 = \left[\frac{\Lambda_{\max}(s_*)}{\Lambda_{\min}(S_*, 2s_*)\phi_{\min}^2(6, S_*, 2s_*)}\right] O(\lambda_{\text{init}} s_*),$$

and

$$\|\hat{b}_{\text{thres}} - \beta^*\|_2 = \left[\frac{\Lambda_{\max}(s_*)}{\Lambda_{\min}(S_*, 2s_*)\phi_{\min}^2(6, S_*, 2s_*)}\right] O(\lambda_{\text{init}} \sqrt{s_*}),$$

and

$$|\hat{S}_{\text{thres}} \backslash S_*| = O(s_*). \tag{7.10}$$

Similarly, under Condition BB,

$$\|\hat{f}_{\text{adap}} - f^0\|_n^2 = \left[\frac{\Lambda_{\max}^2(s_*)}{\phi_{\text{varmin}}^4(6, S_*, 2s_*)}\right] O(\lambda_{\text{init}}^2 s_*), \tag{7.11}$$

and

$$\|\hat{\beta}_{\text{adap}} - \beta^*\|_1 = \left[\frac{\Lambda_{\max}(s_*)}{\phi_{\text{varmin}}^3(6, S_*, 2s_*)}\right] O(\lambda_{\text{init}} s_*),$$

and

$$\|\hat{\beta}_{\text{adap}} - \beta^*\|_2 = \left[\frac{\Lambda_{\max}(s_*)}{\phi_{\text{varmin}}^2(6, S_*, 3s_*)\phi_{\text{varmin}}(6, S_*, 2s_*)} \right] O(\lambda_{\text{init}}\sqrt{s_*}),$$

and

$$|\hat{S}_{\text{adap}}\backslash S_*| = \left[\frac{\Lambda_{\max}^2(s_*)\phi_{\text{varmin}}^2(6, S_*, 2s_*)}{\phi_{\min}^4(6, S_*, 2s_*)} \right] O(s_*). \qquad (7.12)$$

7.8.3 Comparison with the Lasso

At the zoomed-out level, where all ϕ's and Λ's are neglected, we see that the thresh-olded Lasso (under Condition AA) and the adaptive Lasso (under Condition BB) achieve the same order of magnitude for the prediction error as the initial, one-stage Lasso discussed in Theorem 7.7. The same is true for their estimation errors. Zoom-ing in on the ϕ's and the Λ's, their error bounds are generally larger than for the initial Lasso.

For comparison in terms of false positives, we need a corresponding bound for the initial Lasso. In the paper of Zhang and Huang (2008), one can find results that ensure that also for the initial Lasso, modulo ϕ's and Λ's, the number of false pos-itives is of order s_*. However, this result requires rather involved conditions which also improve the bounds for the adaptive and thresholded Lasso. We briefly address this refinement in Lemma 7.3, imposing a condition of similar nature as the one used in Zhang and Huang (2008). Also under these stronger conditions, the general mes-sage remains that thresholding and the adaptive Lasso can have similar prediction and estimation error as the initial Lasso, and are often far better as regards variable selection

Recall that $\Lambda_{\max}^2$ is the largest eigenvalue of $\hat{\Sigma}$. It can generally be quite large.

Lemma 7.2. *On $\mathscr{T}$,*

$$|\hat{S}_{\text{init}}\backslash S_*| \leq \left[\frac{\Lambda_{\max}^2}{\phi_{\text{comp}}^2(6, S_*)} \right] O(s_*).$$

See Problem 7.7 for an example where the bound of Lemma 7.2 is sharp.

We now discuss the above announced refinement, assuming a condition correspond-ing to the one used in Zhang and Huang (2008) (compare with Subsection 6.13.5).

Condition D *It holds for some $t \geq s_*$, that*

$$D(t, s_*) := \left\{ \frac{\Lambda_{\max}^2(t)s_*}{\phi_{\text{comp}}^2(6, S_*)t} \right\} = O_{\text{suff}}(1).$$

Lemma 7.3. *Suppose we are on* $\mathscr{T}$. *Then under Condition D,*

$$|\hat{S}_{\mathrm{init}} \backslash S_*| = \left[\frac{\Lambda_{\max}^2(t)}{\phi_{\mathrm{comp}}^2(6, S_*)} \right] \left(1 - \frac{D(t, s_*)}{O_{\mathrm{suff}}(1)} \right)^{-1} O(s_*).$$

Moreover, under Condition B,

$$|\hat{S}_{\mathrm{adap}} \backslash S_*| = \Lambda_{\max}(t) \left[\frac{\Lambda_{\max}(s_*)}{\phi_{\mathrm{varmin}}(6, S_*, 2s_*) \phi_{\min}^4(6, S_*, 2s_*)} \right]^{1/2} \sqrt{\frac{\lambda_{\mathrm{init}}}{\lambda_{\mathrm{adap}}}} O(s_*)$$

$$+ \left[\frac{\Lambda_{\max}(t) \phi_{\mathrm{comp}}^2(6, S_*)}{\phi_{\mathrm{varmin}}(6, S_*, 2s_*) \phi_{\min}^4(6, S_*, 2s_*)} \right] D(t, s_*) \frac{\lambda_{\mathrm{init}}}{\lambda_{\mathrm{adap}}} O(s_*).$$

Under Condition BB, this becomes

$$|\hat{S}_{\mathrm{adap}} \backslash S_*| = \left[\frac{\Lambda_{\max}(t)}{\phi_{\mathrm{comp}}(6, S_*)} \right] \left[\frac{\phi_{\mathrm{varmin}}^2(6, S_*, 2s_*) \phi_{\mathrm{comp}}^2(6, S_*)}{\phi_{\min}^2(6, S_*, 2s_*)} \right]^{1/2} O(s_*) \qquad (7.13)$$

$$+ \left[\frac{\phi_{\mathrm{varmin}}^2(6, S_*, 2s_*) \phi_{\mathrm{comp}}^2(6, S_*)}{\phi_{\min}^4(6, S_*, 2s_*)} \right] D(t, s_*) O(s_*).$$

Under Condition D, the first term in the right hand side of (7.13) is generally the leading term. We thus see that the adaptive Lasso replaces the potentially very large constant

$$\left(1 - \frac{D(t, s_*)}{O_{\mathrm{suff}}(1)} \right)^{-1}$$

in the bound for the number of false positives of the initial Lasso by

$$\left[\frac{\phi_{\mathrm{varmin}}^2(6, S_*, 2s_*) \phi_{\mathrm{comp}}^2(6, S_*)}{\phi^4(6, S_*, 2s_*)} \right]^{1/2},$$

a constant which is close to 1 if the ϕ's do not differ too much.

Admittedly, Condition D is difficult to interpret. On the one hand, it wants t to be large, but on the other hand, a large t also can render $\Lambda_{\max}(t)$ large. We refer to Zhang and Huang (2008) for examples where Condition D is met.

7.8.4 Comparison between adaptive and thresholded Lasso

When zooming-out, we see that the adaptive and thresholded Lasso have bounds of the same order of magnitude, for prediction, estimation and variable selection.

At the zoomed-in level, the adaptive and thresholded Lasso also have very similar bounds for the prediction error (compare (7.9) with (7.11)) in terms of the ϕ's and Λ's. A similar conclusion holds for their estimation error. We remark that our choice of Conditions AA and BB for the tuning parameters is motivated by the fact that according to our theory, these give the smallest prediction and estimation errors. It then turns out that the "optimal" errors of the two methods match at a quite detailed level. However, if we zoom-in even further and look at the definition of $\Lambda_{\max}(\cdot)$, $\phi_{\min}$, and ϕ_{varmin} in Section 6.13.7, it will show up that the bounds for the adaptive Lasso prediction and estimation error are (slightly) larger.

Regarding variable selection, at zoomed-out level the results are also comparable (see (7.9) and (7.12)). Zooming-in on the the ϕ's and Λ's, the adaptive Lasso may have more false positives than the thresholded version.

A conclusion is that at the zoomed-in level, the adaptive Lasso has less favorable bounds as the refitted thresholded Lasso. However, these are still only bounds, which are based on focusing on a direct comparison between the two methods, and we may have lost the finer properties of the adaptive Lasso. Indeed, the non-explicitness of the adaptive Lasso makes its analysis a non-trivial task. The adaptive Lasso is a quite popular practical method, and we certainly do not advocate that it should be replaced by thresholding and refitting.

7.8.5 Bounds for the number of false negatives

The ℓ_q-error ($1 \le q \le \infty$) has immediate consequences for the number of false negatives: if for some estimator $\hat{\beta}$, some target β^*, and some constant δ_q^{upper} one has

$$\|\hat{\beta} - \beta^*\|_q \le \delta_q^{\text{upper}}$$

then the number of undetected yet large coefficients cannot be very large, in the sense that

$$|\{j: \ \hat{\beta}_j = 0, |\beta_j^*| > \delta\}|^{1/q} \le \frac{\delta_q^{\text{upper}}}{\delta}.$$

In other words, ℓ_q-bounds imply the screening property for the large - in absolute value - coefficients.

Therefore, on $\mathscr{T}$, for example

$$\left|\left\{j: \ \hat{\beta}_{\text{init},j} = 0, \left[\frac{1}{\phi_{\min}^2(6, S_*, 2s_*)}\right] \sqrt{s_*} \lambda_{\text{init}} = O_{\text{suff}}(|\beta_j^*|)\right\}\right| = 0.$$

Similar bounds hold for the thresholded and the adaptive Lasso (considering now, in terms of the ϕ's and Λ's, somewhat larger $|\beta_j^*|$).

Typically, one thinks here of β^0 as target, although one may argue that one should not aim at detecting variables that the oracle considers as irrelevant. In any case, given an estimator $\hat{\beta}$ and alternative target β^*, it is straightforward to bound $\|\hat{\beta} - \beta^0\|_q$ in terms of $\|\hat{\beta} - \beta^*\|_q$: apply the triangle inequality

$$\|\hat{\beta} - \beta^0\|_q \leq \|\hat{\beta} - \beta^*\|_q + \|\beta^* - \beta^0\|_q.$$

Moreover, for $q = 2$, one has the inequality

$$\|\beta^* - \beta^0\|_2^2 \leq \frac{\|f^* - f^0\|_n^2}{\Lambda_{\min}^2(\Sigma_{1,1}(S_0))}$$

(recall that $\Lambda_{\min}^2(\Sigma_{1,1}(S))$ is the smallest eigenvalue of the Gram matrix corresponding to the variables in S). In other words, choosing β^0 as target instead of β^* does in our approach not lead to an improvement in the bounds for $\|\hat{\beta} - \beta^0\|_2$. See also Problem 6.4 for related derivations.

7.8.6 Imposing beta-min conditions

Let us have a closer look at what conditions on the size of the coefficients can bring us. We only discuss the adaptive Lasso (thresholding again giving similar results, see also Zhou (2010)).

We define

$$|\beta^*|_{\min} := \min_{j \in S_*} |\beta_j^*|.$$

Moreover, we let

$$|\beta^*|_{\text{harm}}^2 := \left(\frac{1}{s_*} \sum_{j \in S_*} \frac{1}{|\beta_j^*|^2} \right)^{-1}$$

be the harmonic mean of the squared coefficients.

Note that $|\beta^*|_{\text{harm}} \geq |\beta^*|_{\min}$. In words: assuming that the non-zero coefficients are all sufficiently large is more severe than assuming they are sufficiently large "on average".

Condition C *For the adaptive Lasso, take λ_{adap} sufficiently large, such that*

$$|\beta^*|_{\text{harm}} = O_{\text{suff}}(\lambda_{\text{adap}}).$$

Condition CC *For the adaptive Lasso, take λ_{adap} sufficiently large, but such that*

$$|\beta^*|_{\text{harm}} \asymp_{\text{suff}} \lambda_{\text{adap}}.$$

Lemma 7.4. *Suppose that for some constant $\delta_\infty^{\text{upper}}$, on $\mathcal{T}$,*

$$\|\hat{\beta}_{\text{init}} - \beta^*\|_\infty \leq \delta_\infty^{\text{upper}}.$$

Assume in addition that

$$|\beta^*|_{\min} > 2\delta_\infty^{\text{upper}}. \tag{7.14}$$

Then under Condition C, on $\mathcal{T}$,

$$\|\hat{f}_{\text{adap}} - f^0\|_n^2 = \left[\frac{1}{\phi_{\text{comp}}^2(6, S_*)}\right]\frac{\lambda_{\text{adap}}^2}{|\beta^*|_{\text{harm}}^2}O(\lambda_{\text{init}}^2 s_*),$$

and

$$\|\hat{\beta}_{\text{adap}} - \beta^*\|_1 = \left[\frac{1}{\phi_{\text{comp}}^2(6, S_*)}\right]\frac{\lambda_{\text{adap}}}{|\beta^*|_{\text{harm}}}O(\lambda_{\text{init}} s_*),$$

and

$$\|\hat{\beta}_{\text{adap}} - \beta^*\|_2 = \left[\frac{1}{\phi_{\min}^2(6, S_*, 2s_*)}\right]\frac{\lambda_{\text{adap}}}{|\beta^*|_{\text{harm}}}O(\lambda_{\text{init}}\sqrt{s_*}),$$

and

$$|\hat{S}_{\text{adap}}\setminus S_*| = \left(s_* \vee \left[\frac{\Lambda_{\max}^2(s_*)}{\phi_{\text{comp}}^2(6, S_*)\phi_{\min}^4(6, S_*, 2s_*)}\right]O\left(\frac{\lambda_{\text{init}}^2 s_*}{|\beta^*|_{\text{harm}}^2}\right)\right)$$

$$\wedge\left[\frac{1}{\phi_{\text{comp}}^2(6, S_*)\phi_{\min}^4(6, S_*, 2s_*)}\right]O\left(\frac{\lambda_{\text{init}}^2 s_*^2}{|\beta^*|_{\text{harm}}^2}\right).$$

We show in Theorem 7.10 that the condition (7.14) on $|\beta^*|_{\min}$ can be relaxed (essentially replacing $|\beta^*|_{\min}$ by $|\beta^*|_{\text{harm}}$).

It is clear that by Theorem 7.7, one may insert

$$\delta_\infty = \left[\frac{\sqrt{s_*}}{\phi_{\text{comp}}^2(6, S_*)}\wedge\frac{1}{\phi_{\min}^2(6, S_*, 2s_*)}\right]O(\lambda_{\text{init}}\sqrt{s_*}). \tag{7.15}$$

This can be improved under coherence conditions on the Gram matrix. To simplify the exposition, we will not discuss such improvements in detail (see Lounici (2008)).

Under Condition CC, the bound for the prediction error and estimation error is again the smallest. We moreover have the following corollary for the number of false positives.

Corollary 7.8. *Assume the conditions of Lemma 7.4 and (say)*

$$\lambda_{\text{init}} = O(|\beta^*|_{\text{harm}}),$$

then we have on $\mathcal{T}$,

$$|\hat{S}_{\text{adap}} \backslash S_*| = \left[\frac{\Lambda_{\max}^2(s_*)}{\phi_{\text{comp}}^2(6, S_*) \phi_{\min}^4(6, S_*, 2s_*)} \right] O(s_*).$$

Moreover, when (say)

$$\frac{\lambda_{\text{init}} \sqrt{s_*}}{\phi_{\text{comp}}^2(6, S_*)} = O(|\beta^*|_{\text{harm}}),$$

one can remove the maximal sparse eigenvalues in the bound for $|\hat{S} \backslash S_*|$: on $\mathcal{T}$,

$$|\hat{S}_{\text{adap}} \backslash S_*| = \left[\frac{\phi_{\text{comp}}^2(6, S_*)}{\phi_{\min}^4(6, S_*, 2s_*)} \right] O(s_*).$$

By assuming that $|\beta^*|_{\text{harm}}$ is sufficiently large, that is,

$$\left[\frac{1}{\phi_{\text{comp}}(6, S_*) \phi_{\min}^2(6, S_*, 2s_*)} \right] \lambda_{\text{init}} s_* = O_{\text{suff}}(|\beta^*|_{\text{harm}}), \qquad (7.16)$$

one can bring $|\hat{S}_{\text{adap}} \backslash S_0|$ down to zero, i.e., no false positives (compare with Corollary 7.9).

Thus, even if the design is strongly ill-posed with the ϕ's very small, the adaptive Lasso behaves well in terms of false positives and prediction error if the regression coefficients are sufficiently large (measured in terms of $|\beta^*|_{\text{harm}}$). The assumption (7.16) corresponds to the bound (7.3) coming from the weighted irrepresentable condition: we need a separation of order $\sqrt{s_*}$ between the weights inside the active set and outside the active set. Indeed, modulo the ϕ's, outside the active set the weights may be of order $1/(\lambda_{\text{init}} \sqrt{s_*})$ (see (7.15)), whereas assumption (7.16) roughly says that the weights inside the active set are of order $1/(\lambda_{\text{init}} s_*)$. In this sense, the results are sharp. See also Section 7.5.9 for a further discussion of "ideal" weights.

7.9 The adaptive Lasso without invoking sparse eigenvalues

In this section, we refrain from using sparse eigenvalues. As a result, we have less control of the prediction error of the thresholded or adaptive Lasso. In practice, the tuning parameters are generally chosen by cross validation. With this in mind, we again discuss choices of λ_{adap} which optimize bounds for the prediction error (without having very explicit bounds for this prediction error): see Condition EE. Thus, our choice of the tuning parameter λ_{adap} is in the spirit of Condition BB and CC, but now without assuming sparse eigenvalues.

Our results depend on the variant of the minimal adaptive restricted eigenvalue

$$\phi_* := \phi_{\text{varmin}}(6, S_*, 2s_*)$$

which we generally think of as being not too small, i.e., $1/\phi_* = O(1)$. Moreover, to simplify the expressions, we do not distinguish between $\phi_{comp}(6, S_*)$ and $\phi_{varmin}(6, S_*, 2s_*)$ (i.e., we take the smaller value $\phi_* := \phi_{varmin}(6, S_*, 2s_*)$). We also keep throughout the assumption $\|f^* - f^0\|_n^2 = O(\lambda_{init}^2 s_*/\phi_*^2)$.

7.9.1 The condition on the tuning parameter

We define for $\delta > 0$, the set of thresholded coefficients

$$S_*^{\delta} := \{j : |\beta_j^*| > \delta\}.$$

We let $|\beta^*|_{trim}^2$ be the trimmed harmonic mean

$$|\beta^*|_{trim}^2 := \left(\frac{1}{s_*} \sum_{|\beta_j^*| > 2\delta_\infty^{upper}} \frac{1}{|\beta_j^*|^2} \right)^{-1},$$

where $\delta_\infty^{upper} > 0$ is to be specified (see Theorem 7.10 below). Recall that f_S is defined as the projection (in $L_2(Q_n)$) of f^0 on the linear space spanned by $\{\psi_j\}_{j \in S}$.

Condition EE *Assume the following condition on the tuning parameter:*

$$\lambda_{adap}^2 \asymp \left(\left\| f_{S_*^{4\delta_\infty^{upper}}} - f^0 \right\|_n^2 + \frac{\lambda_{init}^2 s_*}{\phi_*^2} \right) \frac{\phi_*^2 |\beta^*|_{trim}^2}{s_* \lambda_{init}^2}. \tag{7.17}$$

7.9.2 The results

In the next theorem, result 3) contains the main ingredients of the present section. Results 1) and 2) are recaptures, they were presented in Section 6.2.3 and Section 6.8, and also summarized in Section 7.7.

Theorem 7.10. *Suppose that $\|\psi_j\|_n \leq 1$ for all $j = 1, \ldots, p$. Let*

$$\mathscr{T} := \{2|(\varepsilon, \psi_j)_n| \leq \lambda_0\}.$$

Take $\lambda_{init} \geq 2\lambda_0$. Then on $\mathscr{T}$, the following statements hold.
1) There exists a bound $\delta_{init}^{upper} = O(\lambda_{init}\sqrt{s_*}/\phi_*)$ such that

$$\|\hat{f}_{init} - f^0\|_n^2 \leq \delta_{init}^{upper}.$$

2) For $q \in \{1, 2, \infty\}$, there exist bounds $\delta_q^{\mathrm{upper}}$ satisfying

$$\delta_1^{\mathrm{upper}} = O(\lambda_{\mathrm{init}} s_* / \phi_*^2), \ \delta_2^{\mathrm{upper}} = O(\lambda_{\mathrm{init}} \sqrt{s_*} / \phi_*^2), \ \delta_\infty^{\mathrm{upper}} = O(\lambda_{\mathrm{init}} \sqrt{s_*} / \phi_*^2),$$

such that

$$\|\hat{\beta}_{\mathrm{init}} - \beta^*\|_q \le \delta_q^{\mathrm{upper}}, \ q \in \{1, 2, \infty\}.$$

3) Let $\delta_2^{\mathrm{upper}}$ and $\delta_\infty^{\mathrm{upper}}$ be such bounds, satisfying $\delta_\infty^{\mathrm{upper}} \ge \delta_2^{\mathrm{upper}} / \sqrt{s_*}$, and $\delta_2^{\mathrm{upper}} = O(\lambda_{\mathrm{init}} \sqrt{s_*} / \phi_*^2)$.

Suppose that λ_{adap} is chosen according to (7.17) in Condition EE. Then

$$\|\hat{f}_{\mathrm{adap}} - f^0\|_n^2 = O\left(\left\|f_{S_*^{4\delta_\infty^{\mathrm{upper}}}} - f^0\right\|_n^2 + \frac{\lambda_{\mathrm{init}}^2 s_*}{\phi_*^2}\right),$$

and

$$|\hat{S}_{\mathrm{adap}} \backslash S_*| = O\left(\frac{\lambda_{\mathrm{init}}^2 s_*^2}{\phi_*^6 |\beta^*|_{\mathrm{trim}}^2}\right).$$

Theorem 7.10 is a reformulation of part of Theorem 7.11 in Section 7.12, which contains the proof for the noisy case. According to Theorem 7.10, the larger the trimmed harmonic mean $|\beta^*|_{\mathrm{trim}}^2$, the better the variable selection properties of the adaptive Lasso are. A large value for $\delta_\infty^{\mathrm{upper}}$ will make $|\beta^*|_{\mathrm{trim}}$ large, but on the other hand can increase the bound for the prediction error $\|\hat{f}_{\mathrm{adap}} - f^0\|_n^2$.

Corollary 7.9. *Assume the conditions of Theorem 7.10. Note that*

$$|\beta^*|_{\mathrm{trim}} \ge 2\delta_\infty^{\mathrm{upper}}.$$

This implies that when we take $\delta_\infty^{\mathrm{upper}} \asymp \lambda_{\mathrm{init}} \sqrt{s_*} / \phi_*^2$, then

$$(\lambda_{\mathrm{init}} \sqrt{s_*}) / \phi_*^2 = O(|\beta^*|_{\mathrm{trim}}),$$

and hence, with large probability,

$$|\hat{S}_{\mathrm{adap}} \backslash S_*| = O(s_* / \phi_*^2).$$

If in fact

$$(\lambda_{\mathrm{init}} s_*) / \phi_*^3 = O(|\beta^*|_{\mathrm{trim}}), \tag{7.18}$$

we get that with large probability

$$|\hat{S}_{\mathrm{adap}} \backslash S_*| = O(1),$$

as in Corollary 7.8 of Subsection 7.8.6.

The bound we provide above for $\|\hat{f}_{\mathrm{adap}} - f^0\|_n^2$ may be subject to improvement. In fact, we shall show that the threshold $\delta_\infty^{\mathrm{upper}}$ can be replaced by an "oracle" threshold

which minimizes (for a given λ_{adap}) bounds for the prediction error (see (7.22)). The choice of λ_{adap} we then advocate is the one which minimizes the prediction error obtained with the oracle threshold. This refinement is more involved and therefore postponed to Subsection 7.11.4 (for the noiseless case) and Section 7.12 (for the noisy case).

Note that Theorem 7.10 allows for a large choice of $\delta_\infty^{\text{upper}}$, larger than a tight bound for $\|\hat{\beta}_{\text{init}} - \beta^*\|_\infty$. However, with such a large choice, the choice (7.17) for the tuning parameter is also much too large. Thus, a too large threshold will not reflect in any way a choice for λ_{adap} yielding - given the procedure - an optimal prediction error, or mimic a cross validation choice for λ_{adap}. We always may take $\delta_\infty^{\text{upper}} = O(\lambda_{\text{init}}\sqrt{s_*}/\phi_*^2)$. Under incoherence conditions, one may prove that one can take $\delta_\infty^{\text{upper}}$ as small as $\delta_\infty^{\text{upper}} = \text{constant} \times \lambda_{\text{init}}$, where the constant depends on the incoherence conditions (see Lounici (2008)).

7.10 Some concluding remarks

Estimating the support of the non-zero coefficients is a hard statistical problem. The irrepresentable condition, which is essentially a necessary condition for exact recovery of the non-zero coefficients by the one step Lasso, is much too restrictive in many cases. Our main focus is therefore on having $O(s_*)$ false positives while achieving good prediction. This is inspired by the behavior of the "ideal" ℓ_0-penalized estimator. As noted in Section 7.7, such a viewpoint describes the performance of variable selection in settings where some of the regression coefficients may be smaller than the detection limit.

When using cross validation, the best one can expect is a choice of the tuning parameters that reflects the optimal prediction error of the procedure. We have examined thresholding with least squares refitting and the adaptive Lasso, optimizing the bounds on the prediction error for choosing the tuning parameters. According to our theory (and for simplicity not exploiting the fact that the adaptive Lasso mimics thresholding and refitting using an "oracle" threshold), the two methods are comparable.

The adaptive Lasso with cross validation does fitting and variable selection in one single standard algorithm. It follows from Section 2.12 that the solution path for all λ_{adap} can be derived with $O(n|S_{\text{init}}|\min(n,|S_{\text{init}}|))$ essential operation counts. The tuning parameter λ_{adap} is then chosen based on the performance in the validation sets. Cross validation for thresholding and refitting amounts to removing, for each k, the k smallest estimated initial coefficients $|\hat{\beta}_{\text{init},j}|$, and evaluating the least squares solution based on the remaining variables on the validation sets. Therefore, the two methods are also computationally comparable.

7.11 Technical complements for the noiseless case without sparse eigenvalues

We return to the noiseless setting as considered in Section 7.5. Our purpose here is to provide the arguments for the results, presented in Subsection 7.9.2, for the adaptive Lasso, without assuming sparse eigenvalues. The section can be considered as a proofs section and can be safely skipped.

Recall that $f_S := \arg\min_{f=f_{\beta_S}} \|f_{\beta_S} - f^0\|$ is the projection of f^0 on the $|S|$-dimensional linear space spanned by the variables $\{\psi_j\}_{j \in S}$. The coefficients of f_S are denoted by b^S, i.e.,

$$f_S = \sum_{j \in S} \psi_j b_j^S = f_{b^S}.$$

We fix the set S_* as

$$S_* := \arg\min_{S \subset S_0} \left\{ \|f_S - f^0\|^2 + \frac{3\lambda_{\text{init}}^2 |S|}{\phi_{\min}^2(2, S, |S|)} \right\}, \tag{7.19}$$

where the constants are now from Lemma 7.6. We call $f^* := f_{S_*}$ the oracle, and S_* the oracle active set with cardinality $s_* := |S_*|$, and we let $\beta^* := b^{S_*}$.

For simplicity, we assume throughout that

$$\|f^* - f^0\|^2 = O(\lambda_{\text{init}}^2 s_* / \phi_*^2),$$

where

$$\phi_* := \phi_{\text{varmin}}(2, S_*, 2s_*).$$

Define

$$\delta_{\text{init}} := \|f_{\text{init}} - f^*\|.$$

For $q \geq 1$, we define

$$\delta_q := \|\beta_{\text{init}} - \beta^*\|_q.$$

To avoid too many details, we use here the restricted eigenvalue $\phi_{\text{varmin}}(2, S_*, 2s_*)$ instead of $\phi_{\text{comp}}^2(2, S_*)$.

7.11.1 Prediction error for the noiseless (weighted) Lasso

Let $L > 0$ be some constant.

Lemma 7.5. *For all S satisfying $\|w_S\|_2 / w_{S^c}^{\min} \leq L\sqrt{|S|}$, and all β, we have*

$$\|f_{\text{weight}} - f^0\|^2 + \lambda_{\text{init}}\lambda_{\text{weight}}\sum_{j\notin S}w_j|\beta_{\text{weight},j}|$$

$$\leq 2\|f_{\beta_S} - f^0\|^2 + \frac{6\lambda_{\text{init}}^2\lambda_{\text{weight}}^2}{\phi_{\min}^2(2L,S,|S|)}\|w_S\|_2^2,$$

and

$$\lambda_{\text{init}}\lambda_{\text{weight}}\|w_S\|_2\|(\beta_{\text{weight}})_S - \beta_S\|_2 + \lambda_{\text{init}}\lambda_{\text{weight}}\sum_{j\notin S}w_j|\beta_{\text{weight},j}|$$

$$\leq 3\|f_{\beta_S} - f^0\|^2 + \frac{3\lambda_{\text{init}}^2\lambda_{\text{weight}}^2}{\phi_{\min}^2(2L,S,|S|)}\|w_S\|_2^2.$$

A guided proof is given in Problem 7.10. (Recall that for $N = |S|$, it holds that $\phi_{\min}(2L,S,N) = \phi_{\text{adap}}(2L,S,|N|)$.)

Lemma 7.6. *Let*

$$\delta_*^2 := \|f^* - f^0\|^2 + \frac{3\lambda_{\text{init}}^2 s_*}{\phi_*^2}.$$

We have

$$\delta_{\text{init}}^2 + \lambda_{\text{init}}\|(\beta_{\text{init}})_{S_*^c}\|_1 \leq 2\|f^* - f^0\|^2 + \frac{6\lambda_{\text{init}}^2 s_*}{\phi_{\min}(2,S_*,s_*)} \leq 2\delta_*^2.$$

Moreover

$$\delta_1 \leq 3\|f^* - f^0\|^2/\lambda_{\text{init}} + \frac{3\lambda_{\text{init}} s_*}{\phi_{\min}^2(2,S_*,s_*)} \leq 3\delta_*^2/\lambda_{\text{init}},$$

and

$$\delta_2 \leq 6\delta_*^2/(\lambda_{\text{init}}\sqrt{s_*}).$$

This result follows from Lemma 7.5, and Lemma 6.9. We present some hints in Problem 7.11. (For the prediction and ℓ_1-error, the minimal restricted eigenvalue may in fact be replaced by the compatibility constant.)

We thus conclude that

$$\delta_{\text{init}}^2 = O(\lambda_{\text{init}}^2 s_*/\phi_*^2),$$

and

$$\delta_1 = O(\lambda_{\text{init}} s_*/\phi_*^2), \quad \delta_2 = O(\lambda_{\text{init}}\sqrt{s_*}/\phi_*^2).$$

But then also

$$\delta_\infty = O(\delta_2) = O(\lambda_{\text{init}}\sqrt{s_*}/\phi_*^2).$$

7.11.2 The number of false positives of the noiseless (weighted) Lasso

The KKT conditions can be invoked to derive the next lemma. Set

$$\|(1/w)_S\|_2^2 := \sum_{j \in S} \frac{1}{w_j^2}.$$

Lemma 7.7. *Suppose* $\|\psi_j\| \le 1$ *for all* $j = 1, \ldots, p$. *We have*

$$|S_{\text{weight}} \backslash S_*| \le 4 \frac{\|f_{\text{weight}} - f^0\|^2}{\lambda_{\text{weight}}^2} \frac{\|(1/w)_{S_{\text{weight}} \backslash S_*}\|_2^2}{\lambda_{\text{init}}^2}$$

$$\wedge 2\Lambda_{\max} \frac{\|f_{\text{weight}} - f^0\|}{\lambda_{\text{weight}}} \frac{\|(1/w)_{S_{\text{weight}} \backslash S_*}\|_2}{\lambda_{\text{init}}}.$$

Proof of Lemma 7.7. By the weighted KKT conditions (see Subsection 7.5.2), for all j

$$2(\psi_j, f_{\text{weight}} - f^0) = -\lambda_{\text{init}} \lambda_{\text{weight}} w_j \tau_{\text{weight}, j}.$$

Hence,

$$\sum_{j \in S_{\text{weight}} \backslash S_*} 4|(\psi_j, f_{\text{weight}} - f^0)|^2 \ge \lambda_{\text{init}}^2 \lambda_{\text{weight}}^2 \|w_{S_{\text{weight}} \backslash S_*}\|_2^2$$

$$\ge \lambda_{\text{init}}^2 \lambda_{\text{weight}}^2 |S_{\text{weight}} \backslash S_*|^2 / \|(1/w)_{S_{\text{weight}} \backslash S_*}\|_2^2.$$

On the other hand

$$\sum_{j \in S_{\text{weight}} \backslash S_*} |(\psi_j, f_{\text{weight}} - f^0)|^2 \le \Lambda_{\max}^2(\Sigma_{1,1}(S_{\text{weight}} \backslash S_*)) \|f_{\text{weight}} - f^0\|^2.$$

Clearly,

$$\Lambda_{\max}^2(\Sigma_{1,1}(S_{\text{weight}} \backslash S_*)) \le \Lambda_{\max}^2 \wedge |S_{\text{weight}} \backslash S_*|.$$

$\square$

Application of this result to the initial Lasso gives:

Corollary 7.10. *Suppose that* $\|\psi_j\| \le 1$ *for all* $j = 1, \ldots, p$. *It holds that*

$$|S_{\text{init}} \backslash S_*| \le 4\Lambda_{\max}^2 \frac{\delta_{\text{init}}^2}{\lambda_{\text{init}}^2}.$$

Hence, the initial estimator has number of false positives

$$|S_{\text{init}} \backslash S_*| = \Lambda_{\max}^2 O(s_* / \phi_*^2).$$

This result is the noiseless version of Lemma 7.2. Recall that the eigenvalue $\Lambda^2_{\max}$ can be quite large; it can even be almost as large as p (see Example 7.4 below). Therefore, from the result of Corollary 7.10 one generally cannot deduce good variable selection properties of the initial Lasso.

In the next example, Example 7.4, Σ satisfies the compatibility condition, but $\Sigma_{2,2}(S_*)$ has largest eigenvalue of order $p - s_*$.

Example 7.4. Suppose that

$$\Sigma = (1 - \rho)I + \rho(bb^T),$$

where $0 < \rho < 1$ and where $b = (1, 1, \ldots, 1)^T$. Then Σ has smallest eigenvalue $1 - \rho$, so that we can take $\phi^2_* \geq (1 - \rho)$ (see also Problem 6.14). Problem 7.1 verifies that

$$\Lambda^2_{\max}(\Sigma_{2,2}(S_*)) \geq (1 - \rho) + \rho(p - s_*).$$

Problem 7.7 examines the situation further.

7.11.3 Thresholding the noiseless initial estimator

The adaptive Lasso inherits some of its properties from the initial Lasso. In addition, we derive theory for the adaptive Lasso via the thresholded and refitted initial Lasso.

Let for $\delta > 0$,
$$S^\delta_{\text{init}} := \{j : |\beta_{\text{init},j}| > \delta\},$$

and
$$f^\delta_{\text{init}} := f_{(\beta_{\text{init}})_{S^\delta_{\text{init}}}} = \sum_{j \in S^\delta_{\text{init}}} \psi_j \beta_{\text{init},j}.$$

Recall that f_S is defined as the projection of f^0 on the linear space spanned by $\{\psi_j\}_{j \in S}$. Thus $f_{S^\delta_{\text{init}}}$ is the refitted estimator after thresholding at δ, whereas f^δ_{init} is using the thresholded coefficients without refitting.

The following lemma presents a bound for the prediction error of the thresholded and refitted initial estimator. We stress that under sparse eigenvalue conditions, the result of this lemma can be improved.

Lemma 7.8. *We have*
$$\|f_{S^\delta_{\text{init}}} - f^0\| \leq \|f^\delta_{\text{init}} - f^0\|,$$

and moreover,
$$\|f_{S^\delta_{\text{init}}} - f^0\| \leq \|f_{S^{\delta+\delta_\infty}_*} - f^0\|$$

$$\leq \|f^* - f^0\| + \Lambda_{\max}(\Sigma_{1,1}(S_*\backslash S_*^{\delta_\infty + \delta}))\sqrt{|S_*\backslash S_*^{\delta + \delta_\infty}|}(\delta + \delta_\infty).$$

Proof of Lemma 7.8. The first inequality is trivial, as the refitted version is the projection of f^0 on the space spanned by the variables in S_{init}^δ, and f_{init}^δ is in this space.

For the second result, we note that if $|\beta_{\text{init},j}| \leq \delta$, then $|\beta_j^*| \leq \delta + \delta_\infty$. In other words

$$S_*^{\delta + \delta_\infty} \subset S_{\text{init}}^\delta.$$

Hence

$$\|\mathbf{f}_{S_{\text{init}}^\delta} - f^0\| \leq \|\mathbf{f}_{S_*^{\delta + \delta_\infty}} - f^0\|.$$

Moreover,

$$\|f^* - f_{(\beta^*)_{S_*^{\delta+\delta_\infty}}}\|^2 = (\beta_{S_*\backslash S_*^{\delta+\delta_\infty}}^*)^T \Sigma \beta_{S_*\backslash S_*^{\delta+\delta_\infty}}^*$$

$$\leq \Lambda_{\max}^2(S_*\backslash S_*^{\delta+\delta_\infty})\|\beta_{S_*\backslash S_*^{\delta+\delta_\infty}}^*\|_2^2$$

$$\leq \Lambda_{\max}^2(\Sigma_{1,1}(S_*\backslash S_*^{\delta+\delta_\infty}))|S_*\backslash S_*^{\delta+\delta_\infty}|(\delta+\delta_\infty)^2.$$

But then

$$\|\mathbf{f}_{S_*^{\delta+\delta_\infty}} - f^0\| = \min_{\beta = \beta_{S_*^{\delta+\delta_\infty}}} \|f_\beta - f^0\|$$

$$\leq \|f_{(\beta^*)_{S_*^{\delta+\delta_\infty}}} - f^0\| \leq \|f^* - f^0\| + \|f^* - f_{(\beta^*)_{S_*^{\delta+\delta_\infty}}}\|$$

$$\leq \|f^* - f^0\| + \Lambda_{\max}(\Sigma_{1,1}(S_*\backslash S_*^{\delta+\delta_\infty}))\sqrt{|S_*\backslash S_*^{\delta+\delta_\infty}|}(\delta+\delta_\infty).$$

$\square$

We know that $\|f^* - f^0\| = O(\lambda_{\text{init}}\sqrt{s_*}/\phi_*)$ and $\delta_\infty = O(\lambda_{\text{init}}\sqrt{s_*}/\phi_*^2)$. Therefore, Lemma 7.8 with $\delta = 3\delta_\infty$ (which is the value that will be used in Corollary 7.11 ahead), gives

$$\|\mathbf{f}_{S_{\text{init}}^{3\delta_\infty}} - f^0\| \leq \|\mathbf{f}_{S_*^{4\delta_\infty}} - f^0\|$$

$$= O(\lambda_{\text{init}}\sqrt{s_*}/\phi_*) + 4\Lambda_{\max}(\Sigma_{1,1}(S_*\backslash S_*^{4\delta_\infty}))\sqrt{|S_*\backslash S_*^{4\delta_\infty}|}\delta_\infty$$

$$= O(\lambda_{\text{init}}\sqrt{s_*}/\phi_*)\left(1 + \Lambda_{\max}(\Sigma_{1,1}(S_*\backslash S_*^{4\delta_\infty}))\sqrt{|S_*\backslash S_*^{4\delta_\infty}|}/\phi_*\right). \qquad (7.20)$$

When $|\beta^*|_{\min}$ is larger than $4\delta_\infty$, we obviously have $S_*\backslash S_*^{4\delta_\infty} = \emptyset$. In that case, the prediction error after thresholding at $3\delta_\infty$ is still of the same order as the oracle bound. Without this beta-min condition the situation is less clear. The prediction error can then be worse than the oracle bound, and Lemma 7.8 does not tell us whether it improves by taking the threshold δ small, say $\delta_2/\sqrt{s_*} \leq \delta < \delta_\infty$ (with the lower bound for δ being inspired by the comment following Lemma 7.9).

The number of false positives of the thresholded initial Lasso is examined in the next lemma.

Lemma 7.9. *The thresholded initial estimator has number of false positives*

$$|S_{\text{init}}^{\delta} \setminus S_*| \leq \frac{\delta_2^2}{\delta^2}.$$

Proof of Lemma 7.9. We clearly have

$$\delta^2 |S_{\text{init}}^{\delta} \setminus S_*| \leq \sum_{j \in S_{\text{init}}^{\delta} \setminus S_*} |\beta_{\text{init},j}|^2 \leq \|\beta_{\text{init}} - \beta^*\|_2^2 \leq \delta_2^2.$$

Whence the result. □

If we take $\delta \geq \delta_2/\sqrt{s_*}$, we get from Lemma 7.9 that

$$|S_{\text{init}}^{\delta} \setminus S_*| \leq s_*, \tag{7.21}$$

i.e., then we have at most s_* false positives after thresholding.

Clearly, the larger the threshold δ the smaller the number of false positives. On the other hand, a large δ may result in a bad prediction error. According to Lemma 7.8, the prediction error $\|f_{S_{\text{init}}^{\delta}} - f^0\|^2$ can be quite large for δ much larger than δ_∞. With δ in the range $\delta_2/\sqrt{s_*} \leq \delta \leq 3\delta_\infty$, the prediction error is perhaps not very sensitive to the exact value of δ. Looking ahead to the case with noise, cross validation should moreover prefer a larger threshold due to the additional estimation error that occurs if one keeps too many coefficients.

7.11.4 The noiseless adaptive Lasso

The adaptive Lasso has weights

$$w_j = 1/|\beta_{\text{init},j}|, \ j = 1, \ldots, p.$$

Write

$$f_{\text{adap}} := f_{\beta_{\text{adap}}}, \ S_{\text{adap}} := \{j : \beta_{\text{adap},j} \neq 0\}, \ \delta_{\text{adap}} := \|f_{\text{adap}} - f^0\|.$$

Observe that the adaptive Lasso is somewhat more reluctant than thresholding and refitting: the latter ruthlessly disregards all coefficients with $|\beta_{\text{init},j}| \leq \delta$ (i.e., these coefficients get penalty ∞), and puts zero penalty on coefficients with $|\beta_{\text{init},j}| > \delta$. The adaptive Lasso gives the coefficients with $|\beta_{\text{init},j}| \leq \delta$ a penalty of at least $\lambda_{\text{init}}(\lambda_{\text{adap}}/\delta)$ and those with $|\beta_{\text{init},j}| > \delta$ a penalty of at most $\lambda_{\text{init}}(\lambda_{\text{adap}}/\delta)$.

Lemma 7.10. *We have, for all $\delta \geq \delta_2/\sqrt{s_*}$,*

$$\delta_{\text{adap}}^2 + \lambda_{\text{init}}\lambda_{\text{adap}} \sum_{j\in(S_{\text{init}}^\delta)^c} \frac{|\beta_{\text{adap},j}|}{|\beta_{\text{init},j}|}$$

$$\leq 2\|f_{S_{\text{init}}^\delta} - f^0\|^2 + \frac{6\lambda_{\text{init}}^2}{\phi_*^2}\lambda_{\text{adap}}^2 \sum_{j\in S_{\text{init}}^\delta} \frac{1}{\beta_{\text{init},j}^2}.$$

Proof of Lemma 7.10. This follows from applying Lemma 7.5, with $L = 1$. We only have to show that

$$\phi_{\min}(2, S_{\text{init}}^\delta, |S_{\text{init}}^\delta|) \geq \phi_*.$$

Because (see (7.21)),

$$|S_{\text{init}}^\delta \setminus S_*| \leq s_*,$$

indeed

$$\phi_{\min}(2, S_{\text{init}}^\delta, |S_{\text{init}}^\delta|) \geq \phi_{\min}(2, S_{\text{init}}^\delta \cup S_*, |S_{\text{init}}^\delta \cup S_*|)$$

$$\geq \phi_{\text{varmin}}(2, S_*, 2s_*) = \phi_*.$$

$\square$

The above lemma is an obstructed oracle inequality, where the oracle is restricted to choose the index set as the set of variables that are left over after removing the smallest $|\beta_{\text{init},j}|$. If λ_{adap} is chosen small enough, one sees that the prediction error $\|f_{S_{\text{init}}^\delta} - f^0\|^2$ of the refitted thresholded initial estimator is not overruled by the penalty term on the right hand side. This means that the prediction error of the adaptive Lasso is not of larger order than the prediction error of the refitted thresholded initial Lasso. Note that we may take $\lambda_{\text{adap}} \geq \delta$, because for $\lambda_{\text{adap}} < \delta$, the penalty term in the bound of Lemma 7.10 is not larger than $6\lambda_{\text{init}}^2 s_*/\phi_*^2$ (see also Lemma 7.11), which - in order of magnitude - is the oracle bound (which cannot be improved).

Lemma 7.10 leads to defining the "oracle" threshold as

$$\delta_{\text{oracle}}^2 := \arg\min_{\delta \geq \delta_2/\sqrt{s_*}} \left\{ \|f_{S_{\text{init}}^\delta} - f^0\|^2 + \frac{3\lambda_{\text{init}}^2}{\phi_*^2}\lambda_{\text{adap}}^2 \sum_{j\in S_{\text{init}}^\delta} \frac{1}{\beta_{\text{init},j}^2} \right\}. \qquad (7.22)$$

This oracle has active set $S_{\text{init}}^{\delta_{\text{oracle}}}$, with size $|S_{\text{init}}^{\delta_{\text{oracle}}}| = O(s_*)$. In what follows however, we will mainly choose $\delta = 3\delta_\infty$. Thus, our bounds are good when the oracle threshold δ_{oracle} is not too different from $3\delta_\infty$. If in the range $\delta_2/\sqrt{s_*} \leq \delta \leq 3\delta_\infty$ the prediction error $\|f_{S_{\text{init}}^\delta} - f^0\|$ is roughly constant in δ, the oracle threshold will at least be not much smaller than $3\delta_\infty$. When the oracle threshold is larger than this, it is straightforward to reformulate the situation. We have omitted this to avoid too many cases.

Some further results for the prediction error δ_{adap} follow by inserting bounds for the initial Lasso.

We now define the (squared) trimmed harmonic mean

$$|\beta^*|^2_{\text{trim}} := \left(\frac{1}{s_*} \sum_{|\beta_j^*|>2\delta_\infty} \frac{1}{|\beta_j^*|^2} \right)^{-1}. \tag{7.23}$$

Lemma 7.11. *It holds that*

$$\sum_{j \in S_{\text{init}}^\delta} \frac{1}{\beta_{\text{init},j}^2} \leq \frac{1}{\delta^2} \left\{ \left| \{j: \ \delta - \delta_\infty < |\beta_j^*| \leq 2\delta_\infty\} \right| + 4\delta^2 s_* |\beta^*|^{-2}_{\text{trim}} \right\}.$$

Moreover, for $\delta \geq \delta_2/\sqrt{s_*}$,

$$\sum_{j \in S_{\text{init}}^\delta} \frac{1}{\beta_{\text{init},j}^2} \leq \frac{2s_*}{\delta^2}.$$

Proof of Lemma 7.11. We use that if $|\beta_{\text{init},j}| > \delta$, then $|\beta_j^*| > \delta - \delta_\infty$. Moreover, if $|\beta_j^*| > 2\delta_\infty$, then $|\beta_{\text{init},j}| \geq |\beta_j^*|/2$. Hence,

$$\sum_{j \in S_{\text{init}}^\delta} \frac{1}{\beta_{\text{init},j}^2} = \sum_{|\beta_{j,\text{init}}|>\delta, \ |\beta_j^*| \leq 2\delta_\infty} \frac{1}{\beta_{\text{init},j}^2} + \sum_{|\beta_{\text{init},j}|>\delta, \ |\beta_j^*|>2\delta_\infty} \frac{1}{\beta_{\text{init},j}^2}$$

$$\leq \frac{1}{\delta^2} \left\{ \left| \{j: \ \delta - \delta_\infty < |\beta_j^*| \leq 2\delta_\infty\} \right| + 4\delta^2 \sum_{|\beta_j^*|>2\delta_\infty} \frac{1}{|\beta_j^*|^2} \right\}.$$

The second result follows from

$$\sum_{j \in S_{\text{init}}^\delta} \frac{1}{\beta_{\text{init},j}^2} \leq \frac{1}{\delta^2} |S_{\text{init}}^\delta|,$$

and, invoking (7.21),

$$|S_{\text{init}}^\delta| \leq |S_{\text{init}}^\delta \backslash S_*| + |S_*| \leq 2s_*.$$

$\square$

Corollary 7.11. *With the special choice* $\delta = 3\delta_\infty$, *we get*

$$\frac{1}{s_*} \sum_{j \in S_{\text{init}}^{3\delta_\infty}} \frac{1}{\beta_{\text{init},j}^2} \leq 4|\beta^*|^{-2}_{\text{trim}}.$$

Corollary 7.12. *Using the bound of Lemma 7.11 in Lemma 7.10 gives that for all* $\delta \geq \delta_2/\sqrt{s_*}$,

$$\delta_{\text{adap}}^2 \le 2\|f_{S_{\text{init}}^\delta} - f^0\|^2$$

$$+ \frac{6\lambda_{\text{init}}^2}{\phi_*^2} \frac{\lambda_{\text{adap}}^2}{\delta^2} \left\{ \left| \{j: \ \delta - \delta_\infty < |\beta_j^*| \le 2\delta_\infty\} \right| + 4\delta^2 s_* |\beta^*|_{\text{trim}}^{-2} \right\}.$$

If $\delta_2/\sqrt{s_*} \le 3\delta_\infty$, we may choose $\delta = 3\delta_\infty$ to find

$$\delta_{\text{adap}}^2 \le 2\|f_{S_{\text{init}}^{3\delta_\infty}} - f^0\|^2 + \frac{24\lambda_{\text{init}}^2}{\phi_*^2} \lambda_{\text{adap}}^2 s_* |\beta^*|_{\text{trim}}^{-2}. \qquad (7.24)$$

According to Lemma 7.8,

$$\|f_{S_{\text{init}}^{3\delta_\infty}} - f^0\| \le \|f_{S_*^{4\delta_\infty}} - f^0\|$$

$$= O(\lambda_{\text{init}}\sqrt{s_*}/\phi_*) \left(1 + \Lambda_{\max}(\Sigma_{1,1}(S_* \backslash S_*^{4\delta_\infty})) \sqrt{|S_* \backslash S_*^{4\delta_\infty}|} \delta_\infty \right).$$

Inserting these yields

$$\delta_{\text{adap}}^2 \le 2\|f_{S_*^{4\delta_\infty}} - f^0\|^2 + \frac{24\lambda_{\text{init}}^2}{\phi_*^2} \lambda_{\text{adap}}^2 s_* |\beta^*|_{\text{trim}}^{-2}$$

$$= O\left(\frac{\lambda_{\text{init}}^2 s_*}{\phi_*^2} \right) \left(1 + \Lambda_{\max}^2(\Sigma_{1,1}(S_* \backslash S_*^{4\delta_\infty})) |S_* \backslash S_*^{4\delta_\infty}| \delta_\infty^2 \right)$$

$$+ \frac{24\lambda_{\text{init}}^2}{\phi_*^2} \lambda_{\text{adap}}^2 s_* |\beta^*|_{\text{trim}}^{-2}. \qquad (7.25)$$

We proceed by considering the number of false positives of the adaptive Lasso.

Lemma 7.12. *Suppose that $\|\psi_j\| \le 1$ for all $j = 1, \ldots, p$. We have*

$$|S_{\text{adap}} \backslash S_*| \le 4 \frac{\delta_{\text{adap}}^2}{\lambda_{\text{adap}}^2} \frac{\delta_2^2}{\lambda_{\text{init}}^2} \wedge 2\Lambda_{\max} \frac{\delta_{\text{adap}}}{\lambda_{\text{adap}}} \frac{\delta_2}{\lambda_{\text{init}}}$$

$$\wedge 2 \frac{\delta_{\text{adap}}}{\lambda_{\text{adap}}} \frac{\delta_1}{\lambda_{\text{init}}}.$$

Proof of Lemma 7.12. This is a special case of Lemma 7.7, where we use the inequality

$$|S_{\text{adap}} \backslash S_*|^2 = \left(\sum_{j \in S_{\text{adap}} \backslash S_*} |\beta_{\text{adap},j}|^{-1} |\beta_{\text{adap},j}| \right)^2$$

$$\le \sum_{j \in S_{\text{adap}} \backslash S_*} |\beta_{\text{adap},j}|^{-2} \sum_{j \in S_{\text{adap}} \backslash S_*} |\beta_{\text{adap},j}|^2.$$

Alternatively, one may apply

$$|S_{\text{adap}} \backslash S_*| = \sum_{j \in S_{\text{adap}} \backslash S_*} |\beta_{\text{adap},j}|^{-2/3} |\beta_{\text{adap},j}|^{2/3}$$

$$\le \left(\sum_{j \in S_{\text{adap}} \backslash S_*} |\beta_{\text{adap},j}|^{-2} \right)^{1/3} \left(\sum_{j \in S_{\text{adap}} \backslash S_*} |\beta_{\text{adap},j}| \right)^{2/3}.$$

Hence

$$\sum_{j \in S_{\text{adap}} \backslash S_*} |\beta_{\text{adap},j}|^{-2} \ge |S_{\text{adap}} \backslash S_*|^3 \delta_1^2.$$

One then finds, by the same arguments as in Lemma 7.7,

$$4 \delta_{\text{adap}}^2 \Lambda_{\max}(S_{\text{adap}} \backslash S_*) \ge \lambda_{\text{init}}^2 \lambda_{\text{adap}}^2 |S_{\text{adap}} \backslash S_*|^3 \delta_1^2.$$

This gives

$$|S_{\text{adap}} \backslash S_*| \le 2 \frac{\delta_{\text{adap}}}{\lambda_{\text{adap}}} \frac{\delta_1}{\lambda_{\text{init}}}.$$

$\square$

We will choose $\lambda_{\text{adap}} \ge 3\delta_\infty$ in such a way that that the prediction error and the penalty term in (7.24) are balanced, so that

$$\frac{\delta_{\text{adap}}^2}{\lambda_{\text{adap}}^2} = O\left(\frac{s_* \lambda_{\text{init}}^2}{\phi_*^2 |\beta^*|_{\text{trim}}^2} \right). \tag{7.26}$$

Let us put the consequences of this choice in a corollary, summarizing the main results for the noiseless adaptive Lasso.

Corollary 7.13. *We assume the normalization $\|\psi_j\| \le 1$ for all j and take the choice for λ_{adap} given by (7.26).*

a) It then holds that

$$|S_{\text{adap}} \backslash S_*| = O\left(\frac{\lambda_{\text{init}}^2 s_*^2}{\phi_*^4} |\beta^*|_{\text{trim}}^{-2} \wedge \Lambda_{\max} \frac{\lambda_{\text{init}} s_*}{\phi_*^2} |\beta^*|_{\text{trim}}^{-1} \right).$$

- When $\lambda_{\text{init}}^2 s_* |\beta^*|_{\text{trim}}^{-2} / \phi_*^2 = O(1)$, we get $|S_{\text{adap}} \backslash S_*| = O(s_*/\phi_*^2)$.

- When also $\Lambda_{\max} = O(1)$, we get $|S_{\text{adap}} \backslash S_*| = O(\sqrt{s_*}/\phi_*)$.

- With $\lambda_{\text{init}}^2 s_*^2 |\beta^*|_{\text{trim}}^{-2} / \phi_*^4 = O(1)$, we get $|S_{\text{adap}} \backslash S_*| = O(1)$ (or even $|S_{\text{adap}} \backslash S_*| = 0$ if the constants are small enough). This corresponds with the bound of Corollary 7.5, equation (7.3), implying the weighted irrepresentable condition defined in Subsection 7.5.1, a bound which, according to Example 7.3, cannot be improved.

b) We know that $\delta_\infty = O(\delta_2) = O(\lambda_{\text{init}} \sqrt{s_*}/\phi_*^2)$. Suppose now that this cannot be improved, i.e., that

$$\delta_\infty \asymp \lambda_{\text{init}} \sqrt{s_*}/\phi_*^2.$$

Then we get as above

$$|\beta^*|_{\text{trim}}^{-2} = O\left(\frac{\phi_*^4}{\lambda_{\text{init}}^2 s_*}\right).$$

Hence, when the convergence in sup-norm is slow, we can get relatively few false positives, but possibly a not-so-good prediction error. When $\delta_\infty \geq \delta_2/\sqrt{s_*}$ is small, the bound

$$|\beta^*|_{\text{trim}}^{-2} \leq \frac{1}{s_*} \sum_{|\beta_j^*|>\delta_2/\sqrt{s_*}} \frac{1}{|\beta_j^*|^2}$$

may be appropriate. Assuming this to be $O(\phi_*^4/(\lambda_{\text{init}}^2 s_*))$ amounts to assuming that "on average", the coefficients β_j^* are "not too small". For example, it is allowed that $O(1)$ coefficients are as small as $\lambda_{\text{init}}/\phi_*^2$.

c) Suppose now that

$$|\beta^*|_{\min} := \min_{j\in S_*} |\beta_j^*| \geq \lambda_{\text{init}} \sqrt{s_*}/\phi_*^2.$$

Then again

$$|\beta^*|_{\text{trim}}^2 \geq \lambda_{\text{init}}^2 s_*/\phi_*^4.$$

With this (or larger) values for $|\beta^*|_{\min}$, we also see that the refitted thresholded estimator $\mathrm{f}_{S_{\text{init}}}^{3\delta_\infty}$ has prediction error $O(\delta_{\text{init}}^2) = O(\lambda_{\text{init}}^2 s_*/\phi_*^2)$. If

$$|\beta^*|_{\min} \geq \lambda_{\text{init}} s_*/\phi_*^3,$$

we in fact only have $O(1)$ false positives.

d) More generally, in view of Lemma 7.8, the prediction error of the adaptive Lasso can be bounded by

$$\delta_{\text{adap}}^2 = O\left(\frac{\lambda_{\text{init}}^2 s_0}{\phi_*^2}\right)\left(1 + \Lambda_{\max}^2(\Sigma_{1,1}(S_*\backslash S_*^{4\delta_\infty}))|S_*\backslash S_*^{4\delta_\infty}|\delta_\infty^2\right).$$

(This can be improved under sparse eigenvalue conditions.)

7.12 Technical complements for the noisy case without sparse eigenvalues

This section provides the proof of Theorem 7.10.

Consider an n-dimensional vector of observations

$$\mathbf{Y} = f^0 + \varepsilon,$$

and the weighted (noisy) Lasso

$$\hat{\beta}_{\text{weight}} = \arg\min_{\beta}\left\{\|\mathbf{Y} - f_\beta\|_n^2 + \lambda_{\text{init}}\lambda_{\text{weight}}\sum_{j=1}^{p} w_j|\beta_j|\right\}. \qquad (7.27)$$

Here, f^0, the dictionary $\{\psi_j\}$, and $f_\beta := \sum \psi_j\beta_j$ are now considered as vectors in $\mathbb{R}^n$. The norm we use is the normalized Euclidean norm

$$\|f\| := \|f\|_n := \|f\|_2/\sqrt{n} : f \in \mathbb{R}^n,$$

induced by the inner product

$$(f,\tilde{f})_n := \frac{1}{n}\sum_{i=1}^{n} f_i\tilde{f}_i, \ f,\tilde{f} \in \mathbb{R}^n.$$

We define the projections f_S, in the same way as in the previous section. The ℓ_0-sparse projection $f^* = f_{S^*} = \sum_{j\in S_*} \psi_j\beta_j^*$, is now defined with a larger constant (7 instead of 3) in front of the penalty term, and a larger constant ($L = 6$ instead of $L = 2$) of the (minimal) adaptive restricted eigenvalue condition:

$$S_* := \arg\min_{S\subset S_0}\left\{\|f_S - f^0\|_n^2 + \frac{7\lambda_{\text{init}}^2|S|}{\phi_{\min}^2(6,S,|S|)}\right\}.$$

We also change the constant ϕ_* accordingly:

$$\phi_* := \phi_{\text{varmin}}(6,S_*,2s_*).$$

Let

$$\hat{f}_{\text{weight}} := f_{\hat{\beta}_{\text{weight}}}, \ \hat{S}_{\text{weight}} := \{j : \hat{\beta}_{\text{weight},j} \neq 0\}.$$

We define the estimators $\hat{f}_{\text{init}}$ and $\hat{f}_{\text{adap}}$ as in Section 7.6, with active sets $\hat{S}_{\text{init}}$ and $\hat{S}_{\text{adap}}$. The unpenalized least squares estimator using the variables in S is

$$\hat{f}_S = f_{\hat{\beta}S} := \arg\min_{f=f_{\beta_S}} \|\mathbf{Y} - f_{\beta_S}\|_n.$$

We define for $\delta > 0$,

$$\hat{S}_{\text{init}}^\delta := \{j : |\hat{\beta}_{j,\text{init}}| > \delta\}, \ S_*^\delta := \{j : |\beta_j^*| > \delta\}.$$

The refitted version after thresholding, based on the data $\mathbf{Y}$, is $\hat{f}_{\hat{S}_{\text{init}}^\delta}$.

We let

$$\hat{\delta}_{\text{init}} := \|\hat{f}_{\text{init}} - f^0\|_n, \ \hat{\delta}_{\text{adap}} := \|\hat{f}_{\text{adap}} - f^0\|_n,$$

and moreover, for $q \geq 1$,

$$\hat{\delta}_q := \|\hat{\beta}_{\text{init}} - \beta^*\|_q.$$

To handle the (random) noise, we define the set

$$\mathscr{T} := \left\{ \max_{1 \le j \le p} 2|(\varepsilon, \psi_j)_n| \le \lambda_0 \right\},$$

where λ_0 is chosen in such a way that

$$\mathbf{P}(\mathscr{T}) \ge 1 - \alpha$$

where $(1 - \alpha)$ is the confidence we want to achieve (see Lemma 6.2 for a value of λ_0 in the case of normally distributed errors).

The main point is now to take care that the tuning parameters are chosen in such a way that the noisy part due to variables in S_*^c are overruled by the penalty term. In our situation, this can be done by taking $\lambda_{\text{init}} \ge 2\lambda_0$, and λ_{adap} large enough. A lower bound for λ_{adap} depends on the behavior of the initial estimator (see Corollary 7.15). In Corollary 7.11, we let λ_{adap} depend on λ_{init}, s_* and ϕ_*, on a bound $\delta_\infty^{\text{upper}}$ for $\hat{\delta}_\infty$, on the prediction error of $f_{S_*^{4\delta_\infty^{\text{upper}}}}$, and on the trimmed harmonic mean of the $|\beta_j^*|^2$ defined in (7.23).

After presenting a result for the least squares estimator using only the variables j with large enough $|\hat{\beta}_{\text{init}, j}|$, we give the noisy versions of Lemma 7.7 (the proof is a straightforward adjustment of the noiseless case, see Problem 7.10). We then present the corollaries for the noisy initial and noisy adaptive Lasso, as regards prediction error and variable selection. These corollaries have "random" quantities in the bounds. We end with a corollary containing the main result for the noisy case, where the random bounds are replaced by fixed ones, and where we moreover choose a more specific lower bound for λ_{adap}.

The least squares estimator $\hat{f}_{\hat{S}_{\text{init}}^\delta}$ using only variables in $\hat{S}_{\text{init}}^\delta$ (i.e., the projection of $\mathbf{Y} = f^0 + \varepsilon$ on the linear space spanned by $\{\psi_j\}_{j \in \hat{S}_{\text{init}}^\delta}$) has similar prediction properties as $f_{\hat{S}_{\text{init}}^\delta}$ (the projection of f^0 on the same linear space). This is because, as is shown in the next lemma, their difference is small.

Lemma 7.13. *Suppose we are on $\mathscr{T}$. Let $\delta \ge \hat{\delta}_2 / \sqrt{s_*}$. Then*

$$\|\hat{f}_{\hat{S}_{\text{init}}^\delta} - f_{\hat{S}_{\text{init}}^\delta}\|_n^2 \le \frac{2\lambda_0^2 s_*}{\phi_*^2}.$$

Proof of Lemma 7.13. This follows from

$$\|\hat{f}_{\hat{S}_{\text{init}}^\delta} - f_{\hat{S}_{\text{init}}^\delta}\|_n^2 \le 2(\varepsilon, \hat{f}_{\hat{S}_{\text{init}}^\delta} - f_{\hat{S}_{\text{init}}^\delta})_n,$$

and

$$2(\varepsilon, \hat{\mathrm{f}}_{\hat{S}_{\mathrm{init}}^{\delta}} - \mathrm{f}_{\hat{S}_{\mathrm{init}}^{\delta}})_n \leq \lambda_0 \|\hat{b}^{\hat{S}_{\mathrm{init}}^{\delta}} - b^{\hat{S}_{\mathrm{init}}^{\delta}}\|_1$$

$$\leq \sqrt{2s_*}\lambda_0 \|\hat{b}^{\hat{S}_{\mathrm{init}}^{\delta}} - b^{\hat{S}_{\mathrm{init}}^{\delta}}\|_2 \leq \sqrt{2s_*}\lambda_0 \|\hat{\mathrm{f}}_{\hat{S}_{\mathrm{init}}^{\delta}} - \mathrm{f}_{\hat{S}_{\mathrm{init}}^{\delta}}\|_n / \phi_*.$$

$\square$

Lemma 7.14. *Suppose we are on $\mathscr{T}$. Assume $\lambda_{\mathrm{init}} \geq 2\lambda_0$ and $\lambda_{\mathrm{weight}} w_{S_*^c}^{\min} \geq 1$. We have*

$$|\hat{S}_{\mathrm{weight}} \backslash S_*| \leq 16 \frac{\|\hat{f}_{\mathrm{weight}} - f^0\|_n^2}{\lambda_{\mathrm{weight}}^2} \frac{\|(1/w)_{\hat{S}_{\mathrm{weight}} \backslash S_*}\|_2^2}{\lambda_{\mathrm{init}}^2}$$

$$\wedge 4\Lambda_{\max} \frac{\|\hat{f}_{\mathrm{weight}} - f^0\|_n}{\lambda_{\mathrm{weight}}} \frac{\|(1/w)_{\hat{S}_{\mathrm{weight}} \backslash S_*}\|_2}{\lambda_{\mathrm{init}}}.$$

The proof is Problem 7.8.

Finally[4], we present two corollaries, one recalling the prediction error of the initial Lasso and describing a variable selection result, the other one regarding the prediction error and variable selection of the adaptive Lasso. The consequences of these two corollaries, presented in Theorem 7.11, give qualitatively the same conclusion as in the noiseless case.

Corollary 7.14. *Let*

$$\delta_{\min}^2 := \|f^* - f^0\|_n^2 + \frac{7\lambda_{\mathrm{init}}^2 |S_*|}{\phi_{\min}^2(6, S_*, 2s_*)}.$$

Take $\lambda_{\mathrm{init}} \geq 2\lambda_{\mathrm{noise}}$. We have on $\mathscr{T}$,

$$\hat{\delta}_{\mathrm{init}}^2 \leq 2\delta_{\min}^2.$$

Moreover, on $\mathscr{T}$,

$$\hat{\delta}_1 \leq 5\delta_{\min}^2/\lambda_{\mathrm{init}},$$

and

$$\hat{\delta}_2 \leq 10\delta_{\min}^2/(\lambda_{\mathrm{init}}\sqrt{s_*}).$$

Also, on $\mathscr{T}$, and assuming $\|\psi_j\|_n \leq 1$ for all j,

$$|\hat{S}_{\mathrm{init}} \backslash S_*| \leq 16\Lambda_{\max}^2 \frac{\hat{\delta}_{\mathrm{init}}^2}{\lambda_{\mathrm{init}}^2}.$$

---

[4] Of separate interest is a direct comparison of the noisy initial Lasso with the noisy ℓ_0-penalized estimator. Replacing f^0 by $\mathbf{Y}$ in Lemma 7.6 gives

$$\|\mathbf{Y} - \hat{f}_{\mathrm{init}}\|_n^2 \leq 2 \min_S \left\{ \|\mathbf{Y} - \hat{\mathrm{f}}_S\|_n^2 + \frac{3\lambda_{\mathrm{init}}^2 |S|}{\phi_{\min}^2(2, S, |S|)} \right\}.$$

Corollary 7.15. *Suppose we are on* $\mathcal{T}$. *Take* $\lambda_{\text{init}} \geq 2\lambda_0$ *and* $\delta \geq \hat{\delta}_2/\sqrt{s_*}$. *Let*

$$\lambda_{\text{adap}}^2 \sum_{j \in \hat{S}_{\text{init}}^\delta} \frac{1}{\hat{\beta}_{\text{init},j}^2} \geq |\hat{S}_{\text{init}}^\delta|.$$

The prediction error of the adaptive Lasso then follows from applying Lemma 6.12. We obtain

$$\hat{\delta}_{\text{adap}}^2 + \frac{1}{2}\lambda_{\text{init}}\lambda_{\text{adap}} \sum_{j \notin \hat{S}_{\text{init}}^\delta} \frac{|\hat{\beta}_{\text{adap},j}|}{|\hat{\beta}_{\text{init},j}|}$$

$$\leq 2\|\mathrm{f}_{\hat{S}_{\text{init}}^\delta} - f^0\|_n^2 + \frac{14\lambda_{\text{init}}^2}{\phi_*^2}\lambda_{\text{adap}}^2 \sum_{j \in \hat{S}_{\text{init}}^\delta} \frac{1}{\hat{\beta}_{\text{init},j}^2}.$$

If moreover $\|\psi_j\|_n \leq 1$ *for all* j *and*

$$\lambda_{\text{adap}} \geq \|(\hat{\beta}_{\text{init}})_{S_*^c}\|_\infty,$$

then

$$|\hat{S}_{\text{adap}} \setminus S_*| \leq 16 \frac{\hat{\delta}_{\text{adap}}^2}{\lambda_{\text{adap}}^2} \frac{\hat{\delta}_2^2}{\lambda_{\text{init}}^2} \wedge 4\Lambda_{\max} \frac{\hat{\delta}_{\text{adap}}}{\lambda_{\text{adap}}} \frac{\hat{\delta}_2}{\lambda_{\text{init}}}.$$

The randomness in the bounds for the adaptive Lasso can be easily handled invoking fixed bounds $\delta_2^{\text{upper}} \geq \hat{\delta}_2$ and $\delta_\infty^{\text{upper}} \geq \hat{\delta}_\infty$, that are assumed to hold on $\mathcal{T}$. We recall the notation

$$|\beta^*|_{\text{trim}}^2 := \left(\frac{1}{s_*} \sum_{|\beta_j^*| > 2\delta_\infty^{\text{upper}}} \frac{1}{|\beta_j^*|^2} \right)^{-1}.$$

The special case $\delta = 3\delta_\infty^{\text{upper}}$ then gives

Theorem 7.11. *Let* $\delta_*^2 := \|f^* - f^0\|_n^2 + 7\lambda_{\text{init}}^2 s_*/\phi_*^2$. *Let*

$$\delta_2^{\text{upper}} := \frac{10\delta_*^2}{\lambda_{\text{init}}\sqrt{s_*}}.$$

Suppose we are on $\mathcal{T}$, *and that* $\|\psi_j\|_n \leq 1$ *for all* j. *Suppose* $\hat{\delta}_\infty \leq \delta_\infty^{\text{upper}}$, *where* $3\delta_\infty^{\text{upper}} \geq \delta_2^{\text{upper}}/\sqrt{s_*}$. *Let* $\lambda_{\text{init}} \geq 2\lambda_{\text{noise}}$ *and* $\lambda_{\text{adap}} \geq 3\delta_\infty^{\text{upper}}$. *Then*

$$\hat{\delta}_{\text{adap}}^2 \leq 2\left\|\mathrm{f}_{S_*^{4\delta_\infty^{\text{upper}}}} - f^0\right\|_n^2 + \frac{56\lambda_{\text{init}}^2\lambda_{\text{adap}}^2 s_*}{\phi_*^2|\beta^*|_{\text{trim}}^2}.$$

The choice

$$\lambda_{\text{adap}}^2 = 9\left(\left\|\mathrm{f}_{S_*^{4\delta_\infty^{\text{upper}}}} - f^0\right\|_n^2 + \frac{\lambda_{\text{init}}^2 s_*}{\phi_*^2}\right)\frac{\phi_*^2|\beta^*|_{\text{trim}}^2}{4\lambda_{\text{init}}^2 s_*},$$

indeed has

$$\lambda_{\text{adap}}^2 \geq s_*^2 |\beta^*|_{\text{trim}}^2 \geq (3\delta_\infty^{\text{upper}})^2.$$

With this choice, we find

$$\hat{\delta}_{\text{adap}}^2 \leq 128 \left\{ \left\| \mathsf{f}_{S_*^{4\delta_\infty^{\text{upper}}}} - f^0 \right\|_n^2 + \lambda_{\text{init}}^2 s_* \phi_*^2 \right\},$$

and

$$|\hat{S}_{\text{adap}} \backslash S_*| \leq \frac{(32M)^2 \lambda_{\text{init}}^2 s_*}{\phi_*^6 |\beta^*|_{\text{trim}}^2} s_* \wedge \Lambda_{\text{max}} \frac{32M\sqrt{s_*}\lambda_{\text{init}}}{\phi_*^3 |\beta^*|_{\text{trim}}},$$

where

$$M := \frac{10\delta_{\min}^2 \phi_*^2}{\lambda_{\text{init}}^2 s_*}.$$

The situation is simplified if we assume that the minimal coefficient

$$|\beta^*|_{\min} := \min_{j \in S_*} |\beta_j^*|$$

is sufficiently large. For example, under the beta-min condition

$$|\beta^*|_{\min} > 4\delta_\infty^{\text{upper}},$$

thresholding at $4\delta_\infty^{\text{upper}}$ will not increase the prediction error. The bound of Theorem 7.10 then coincides with the bound for $\hat{\delta}_{\text{init}}^2$, namely

$$\hat{\delta}_{\text{adap}}^2 = O(\lambda_{\text{init}}^2 s_* / \phi_*^2).$$

The number of false positives is again $O(s_*/\phi_*^2)$. If $|\beta^*|_{\min}$ is even larger, the prediction error remains of the same order, but the number of false positives decreases, and may even vanish.

7.13 Selection with concave penalties

Consider now the ℓ_r-"norm"

$$\|\beta\|_r = \left(\sum_{j=1}^p |\beta_j|^r \right)^{1/r},$$

where $0 < r < 1$. The estimator with ℓ_r-penalty, is

$$\hat{\beta} := \underset{\beta}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^n (Y_i - f_\beta(X_i))^2 + \lambda^{2-r} \|\beta\|_r^r \right\}.$$

Again, λ is a (properly chosen) regularization parameter. We let $\hat{f} := f_{\hat{\beta}}$.

We first recall the result of Lemma 6.15 for the prediction error.

Definition *We say that the ℓ_r-compatibility condition is satisfied for the set S, with constant $\phi_{\ell_r}(L,S) > 0$, if for all $\beta \in \mathbf{R}^p$, that satisfy $\|\beta_{S^c}\|_r^r \le L\|\beta_S\|_r^r$, it holds that*

$$\|\beta_S\|_r^r \le \|f_\beta\|_n^r |S|^{\frac{2-r}{2}} / \phi_{\ell_r}^r(L,S). \tag{7.28}$$

In Section 6.11.1, we used the notation

$$\phi_{\hat{\Sigma},r}(S) = \phi_{\ell_r}(3,S),$$

i.e., we took $L = 3$ and explicitly expressed the dependence on the Gram matrix $\hat{\Sigma}$. The latter was to incorporate possible approximations of $\hat{\Sigma}$. This can be done in the present context as well. We omit the details.

Let $\mathscr{T}_r$ be the set

$$\mathscr{T}_r := \left\{ \sup_\beta \frac{|2(\varepsilon, f_\beta)_n|}{\|f_\beta\|_n^{\frac{2(1-r)}{2-r}} \|\beta\|_r^{\frac{r}{2-r}}} \le \lambda_0 \right\},$$

In Corollary 14.7, it is shown that, under general conditions, $\mathscr{T}_r$ has large probability, for λ_0 of order $\sqrt{\log p / n}$.

Recall the projection $f_S = f_{b^S}$, in $L_2(Q_n)$, of f^0 on the linear space spanned by $\{\psi_j\}_{j \in S}$:

$$f_S := \arg \min_{f = f_{\beta_S}} \|f - f^0\|_n.$$

Definition of the oracle *We define the oracle as $\beta^* := b^{S_*}$, with*

$$S_* := \arg \min_{S \in \mathscr{S}} \left\{ 4\|f_S - f^0\|_n^2 + 12(9\lambda)^2 |S|^2 / \phi_{\ell_r}^{\frac{2r}{2-r}}(3,S) \right\},$$

and set $f^* := f_{\beta^*}$, $|S_*| := s_*$ and $\phi_* := \phi_{\ell_r}(S_*)$.

To simplify the exposition, we assume the "estimation error" is the dominating term, i.e., that

$$4\|f^* - f^0\|_n^2 + \frac{12(9\lambda)^2 s_*^2}{\phi_*^{\frac{2r}{2-r}}(S_*)} = \left[\frac{1}{\phi_*^{\frac{2r}{2-r}}(S_*)} \right] O(\lambda^2 s_*).$$

Let us restate Lemma 6.15 for the prediction error.

Lemma 7.15. *Suppose $\lambda^{2-r} \ge 5\lambda_0^{2-r} 4^{1-r}$. We then have on $\mathscr{T}_r$,*

$$\|\hat{f} - f^0\|_n^2 + \lambda^{2-r}\|\hat{\beta} - \beta^*\|_r^r = \left[\frac{1}{\phi_*^{\frac{2r}{2-r}}(S_*)}\right] O(\lambda^2 s_*).$$

For variable selection, we again invoke the KKT conditions. (Due the concavity of the penalty, the KKT conditions do not characterize the solution).

KKT conditions *We have for all $\hat{\beta}_j \neq 0$,*

$$2(\hat{\Sigma}(\hat{\beta} - \beta^0))_j - 2(\varepsilon, \psi_j)_n = -r\lambda^{2-r}|\hat{\beta}_j|^{-(1-r)}\hat{\tau}_j.$$

Here

$$\hat{\tau}_j \mathbb{1}\{\hat{\beta}_j \neq 0\} = \text{sign}(\hat{\beta}_j).$$

We will apply the following auxiliary result.

Lemma 7.16. *For any index set S and vector β with nonzero coefficients in S, and for all $q \geq 1$, we have*

$$\sum_{j \in S} |\beta_j|^{-2(1-r)} \geq |S|^{\frac{q}{q-1}} \left(\sum_{j \in S} |\beta_j|^{2(1-r)(q-1)}\right)^{-\frac{1}{q-1}}.$$

Proof. This follows from

$$|S| = \sum_{j \in S} |\beta_j|^{\frac{2(1-r)(q-1)}{q}} |\beta_j|^{-\frac{2(1-r)(q-1)}{q}}$$

$$\leq \left(\sum_{j \in S} |\beta_j|^{-2(1-r)}\right)^{\frac{q-1}{q}} \left(\sum_{j \in S} |\beta_j|^{-2(1-r)(q-1)}\right)^{\frac{1}{q}},$$

where we applied Hölder's inequality.   □

We now invoke similar arguments as in Lemma 7.7. The definition of the maximal sparse eigenvalue $\Lambda_{\max}^2(N)$ can be found in Subsection 6.13.7.

Lemma 7.17. *Suppose $\|\psi_j\| \leq 1$ for all $j = 1, \ldots, p$. Take $\lambda \geq (5^{\frac{1}{2-r}} 4^{\frac{1-r}{2-r}} \vee 1/(2r))\lambda_0$. We have on $\mathcal{T}_r$,*

$$|\hat{S} \backslash S_*| = \left[\frac{\Lambda_{\max}(s_*)}{\phi_*}\right]^{\frac{r}{1-r}} O(s_*) \wedge \left[\frac{1}{\phi_*}\right]^{\frac{r}{1-r}} O\left(s_*^{1+\frac{r}{2(1-r)}}\right).$$

Thus, under sparse eigenvalue conditions, the concave penalty is capable of selecting the right order of magnitude of variables. If the sparse eigenvalues are very large, we see that the ℓ_r-penalty may still select too many variables, but the number of false positives gets close to $O(s_*)$ as $r \to 0$.

Proof of Lemma 7.17. Throughout, we assume we are on $\mathscr{T}_r$. On $\mathscr{T}_r$, it holds for all j that $2|(\varepsilon, \psi_j)| \leq \lambda_0$ (since $\psi_j = f_\beta$, with $\beta_k = 1\{k = j\}, k = 1, \ldots, p$). Consider the set of estimated coefficients $\hat{S}^\delta$ after thresholding at $\delta := \lambda$. In view of Lemma 7.15, it holds that $\|\hat{\beta}_{S_*^c}\|_r^r = O(\lambda^r s_*)/\phi_*^{\frac{2r}{2-r}}$, and hence

$$|\hat{S}^\delta \backslash S_*| = O(s_*)/\phi_*^{\frac{2r}{2-r}}.$$

By the weighted KKT conditions for the ℓ_r-penalty, when $j \in (\hat{S} \backslash \hat{S}^\delta) \backslash S_*$,

$$2|(\psi_j, \hat{f} - f^0)_n| \leq r\lambda^{2-r}|\hat{\beta}_j|^{-(1-r)}/2,$$

where we use that $r\lambda^{2-r}/\delta^{1-r} = r\lambda \geq 2\lambda_0$.

Hence,

$$\sum_{j \in (\hat{S} \backslash \hat{S}^\delta) \backslash S_*} 16 r^{-2} |(\psi_j, \hat{f} - f^0)_n|^2 \geq \lambda^{2(2-r)} \sum_{j \in (\hat{S} \backslash \hat{S}^\delta) \backslash S_*} |\hat{\beta}_j|^{-2(1-r)}$$

$$\geq \lambda^{2(2-r)} |(\hat{S} \backslash \hat{S}^\delta) \backslash S_*|^{\frac{q}{q-1}} \left(\sum_{j \in \hat{S} \backslash S_*} |\hat{\beta}_j|^{2(1-r)(q-1)} \right)^{-\frac{1}{q-1}},$$

where we used Lemma 7.16. On the other hand

$$\sum_{j \in (\hat{S} \backslash \hat{S}^\delta) \backslash S_*} |(\psi_j, \hat{f} - f^0)_n|^2 \leq \Lambda_{\max}^2((\hat{S} \backslash \hat{S}^\delta) \backslash S_*) \|\hat{f} - f^0\|_n^2.$$

It follows that

$$|(\hat{S} \backslash \hat{S}^\delta) \backslash S_*|^{\frac{1}{q-1}} \leq 16 r^{-2} \lambda^{-2(2-r)} C \|\hat{\beta}_{S_*^c}\|_{2(1-r)(q-1)}^{2(1-r)} \|\hat{f} - f^0\|_n^2$$

$$= 16 r^{-2} C s_* \left(\frac{\|\hat{\beta}_{S_*^c}\|_{2(1-r)(q-1)}^{2(1-r)(q-1)}}{\lambda^{2(1-r)(q-1)} s_*} \right)^{\frac{1}{q-1}} \frac{\|\hat{f} - f^0\|_n^2}{\lambda^2 s_*} s_*^{\frac{1}{q-1}},$$

where

$$C := \frac{\Lambda_{\max}^2(\Sigma_{1,1}((\hat{S} \backslash \hat{S}^\delta) \backslash S_*))}{|(\hat{S} \backslash \hat{S}^\delta) \backslash S_*|}.$$

Hence,

$$|(\hat{S} \backslash \hat{S}^\delta) \backslash S_*|$$

$$\leq 16^{q-1} r^{-2(q-1)} \left(C s_* \right)^{q-1} \left(\frac{\|\hat{\beta}_{S_*^c}\|_{2(1-r)(q-1)}^{2(1-r)(q-1)}}{\lambda^{2(1-r)(q-1)} s_*} \right) \left(\frac{\|\hat{f} - f^0\|_n^2}{\lambda^2 s_*} \right)^{q-1} s_*.$$

We now choose $q - 1 = r/(2(1-r))$, which gives

$$|(\hat{S} \backslash \hat{S}^\delta) \backslash S_*| \leq 16^{\frac{r}{2(1-r)}} r^{-\frac{r}{(1-r)}} \left(C s_* \right)^{\frac{r}{2(1-r)}} \left(\frac{\|\hat{\beta}_{S_*^c}\|_r^r}{\lambda^r s_*} \right) \left(\frac{\|\hat{f} - f^0\|_n^2}{\lambda^2 s_*} \right)^{\frac{r}{2(1-r)}} s_*.$$

By Lemma 7.15,

$$\|\hat{\beta}_{S_*^c}\|_r^r = \left[\frac{1}{\phi_*^{\frac{2r}{2-r}}}\right] O(\lambda^r s_*), \ \|\hat{f} - f^0\|_n^2 = \left[\frac{1}{\phi_*^{\frac{2r}{2-r}}}\right] O(\lambda^r s_*).$$

Moreover, if $|(\hat{S}\backslash\hat{S}^\delta)\backslash S_*| > s_*$, we know that

$$C s_* \le 2\Lambda_{\max}^2(s_*).$$

Thus, then

$$|(\hat{S}\backslash\hat{S}^\delta)\backslash S_*| = \left[\frac{\Lambda(s_*)}{\phi_*}\right]^{\frac{r}{1-r}} O(s_*).$$

We also know that $C \le 1$, so

$$|(\hat{S}\cap\hat{S}_\delta^c)\backslash S_*| = \left[\frac{1}{\phi_*^{\frac{r}{1-r}}}\right] O(s_*^{1+\frac{r}{2(1-r)}}).$$

Finally note that

$$\frac{2r}{2-r} \le \frac{r}{1-r}.$$

$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

Problems

7.1. Let

$$\Sigma := (1-\rho)I + \rho(bb^T),$$

where $0 \le \rho < 1$ and $b = (1,\dots,1)^T$. Take

$$\beta = b/\sqrt{p}.$$

Then clearly $\|\beta\|_2 = 1$. Check that

$$\beta^T \Sigma \beta = (1-\rho) + \rho p.$$

7.2. Here is an example where the compatibility condition holds for all L and S with $\phi_{\text{comp}}(L,S) \ge 1/2$, the irrepresentable condition does not hold for some S_0 with $s_0 := |S_0| > 16$, and the maximal eigenvalue $\Lambda_{\max}^2$ of Σ is at least $\frac{1}{4}(p - s_0)$. The exercise consists in proving Lemmas 7.18, 7.19 and 7.20.

Let $S_0 = \{1,\dots,s_0\}$ be the active set, and suppose that

$$\Sigma := \begin{pmatrix} \Sigma_{1,1} & \Sigma_{1,2} \\ \Sigma_{2,1} & \Sigma_{2,2} \end{pmatrix},$$

where $\Sigma_{1,1} := I$ is the $(s_0 \times s_0)$-identity matrix, and

$$\Sigma_{2,1} := \rho(b_2 b_1^T),$$

with $0 \le \rho < 1$, and with b_1 an s_0-vector and b_2 a $(p - s_0)$-vector, satisfying $\|b_1\|_2 = \|b_2\|_2 = 1$. Moreover, $\Sigma_{2,2}$ is some $(p - s_0) \times (p - s_0)$-matrix, with $\mathrm{diag}(\Sigma_{2,2}) = I$, and with largest eigenvalue $\Lambda^2_{\max}(\Sigma_{2,2})$ and smallest eigenvalue $\Lambda^2_{\min}(\Sigma_{2,2})$.

Lemma 7.18. *Suppose that $\rho < \Lambda^2_{\min}(\Sigma_{2,2})$. Then Σ is positive definite, with smallest eigenvalue $\Lambda^2_{\min} \ge \Lambda^2_{\min}(\Sigma_{2,2}) - \rho$.*

It follows that the compatibility condition holds for any L and S, with $\phi^2_{\mathrm{comp}}(L, S) = \Lambda^2_{\min}(\Sigma_{2,2}) - \rho$.

Next, we illustrate that for a particular choice of b_1 and b_2, and for $\rho > 1/\sqrt{s_0}$, the irrepresentable condition does not hold.

Lemma 7.19. *Let $b_1 := (1, 1, \ldots, 1)^T / \sqrt{s_0}$ and $b_2 := (1, 0, \ldots, 0)^T$. Then for $\tau_{S_0} := (1, \ldots, 1)^T$, we have*

$$\|\Sigma_{2,1} \Sigma_{1,1}^{-1} \tau_{S_0}\|_\infty = \rho \sqrt{s_0}.$$

We now give an example where $\Lambda^2_{\max}(\Sigma_{2,2})$ is huge, but $\Lambda^2_{\min}(\Sigma_{2,2})$ is harmless. We take for some $0 < \theta < 1$,

$$\Sigma_{2,2} := (1 - \theta)I + \theta c_2 c_2^T, \tag{7.29}$$

where $c_2 := (1, \ldots, 1)^T$.

Lemma 7.20. *With $\Sigma_{2,2}$ given in (7.29), one has*

$$\Lambda^2_{\min}(\Sigma_{2,2}) = (1 - \theta),$$

and

$$\Lambda^2_{\max}(\Sigma_{2,2}) = (1 - \theta) + \theta(p - s_0).$$

Corollary 7.16. *Let*

$$\Sigma := \begin{pmatrix}
1 & 0 & \cdots & 0 & \rho/\sqrt{s_0} & 0 & \cdots & 0 \\
0 & 1 & \cdots & 0 & \rho/\sqrt{s_0} & 0 & \cdots & 0 \\
0 & 0 & \ddots & 0 & \vdots & \vdots & & \vdots \\
0 & 0 & \cdots & 1 & \rho/\sqrt{s_0} & 0 & \cdots & 0 \\
\rho/\sqrt{s_0} & \rho/\sqrt{s_0} & \cdots & \rho/\sqrt{s_0} & 1 & \theta & \cdots & \theta \\
0 & 0 & \cdots & 0 & \theta & 1 & \cdots & \theta \\
\vdots & \vdots & & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & 0 & \theta & \theta & \cdots & 1
\end{pmatrix},$$

where $\rho = \theta = 1/4$. Then the compatibility condition holds for all L and S with $\phi_{comp}(L,S) \geq 1/2$, the irrepresentable condition does not hold for $s_0 > 16$, and the maximal eigenvalue Λ_{max}^2 of Σ is at least $\frac{1}{4}(p - s_0)$.

7.3. Consider Example 7.3. Check that

$$\Sigma_{1,1}^{-1}(\mathcal{N}) = \frac{1}{\rho^2 s_0 - 1} \begin{pmatrix} (\rho^2 s_0 - 1)I - \rho^2 \tau_{S_0} \tau_{S_0}^T & \rho \tau_{S_0} \\ \rho \tau_{S_0}^T & \rho^2 s_0 \end{pmatrix}.$$

Hence

$$\sup_{\|(\tau_{S_0}^T, \tau_{s_0+1})\|_\infty \leq 1} \left\| \Sigma_{1,1}^{-1}(\mathcal{N}) \begin{pmatrix} \tau_{S_0} \\ \tau_{s_0+1} \end{pmatrix} \right\|_\infty = 1 + 1/(\rho^2 s_0 - 1).$$

This can be large, e.g., when $\rho^2 = 1/s_0 + 1/s_0^2$ (recall that the irrepresentable condition does not hold for $\rho > 1/\sqrt{s_0}$). By Theorem 7.1, we conclude $\|\beta_{init} - \beta^0\|_\infty$ can be large, which is of course completely due to the $(s_0 + 1)$-th variable.

7.4. In this example, we investigate the effect of adding a variable to the active set on the irrepresentable condition. Consider $S_0 = \{1, \ldots, s_0\}$ and $\mathcal{N} := S_0 \cup \{s_0 + 1\}$. Write $\Sigma_{1,1} := \Sigma_{1,1}(S_0)$. Then

$$\Sigma_{1,1}(\mathcal{N}) = \begin{pmatrix} \Sigma_{1,1} & a \\ a^T & 1 \end{pmatrix}.$$

Verify by straightforward calculations

$$\Sigma_{1,1}^{-1}(\mathcal{N}) = \begin{pmatrix} B & -Ba \\ -a^T B & 1 \end{pmatrix} / (1 - a^T \Sigma_{1,1}^{-1} a),$$

where

$$B := (1 - a^T \Sigma_{1,1}^{-1} a)\Sigma_{1,1}^{-1} + \Sigma_{1,1}^{-1} a a^T \Sigma_{1,1}^{-1} = (\Sigma_{1,1} - aa^T)^{-1}(1 - a^T \Sigma_{1,1}^{-1} a).$$

7.5. In this exercise, we investigate the irrepresentable condition in the noisy setup. Let

$$\mathbf{Y} = \mathbf{X}\beta^0 + \varepsilon.$$

The active set is $S_0 = \{j : \beta_j^0 \neq 0\}$ and the Lasso is

$$\hat{\beta} = \arg\min_\beta \{\|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + \lambda\|\beta\|_1\}.$$

Define

$$\mathcal{T} := \{ \max_{1 \leq j \leq p} 2|\varepsilon^T \mathbf{X}^{(j)}|/n \leq \lambda_0 \}.$$

Let $\hat{S} := \{j : \hat{\beta}_j \neq 0\}$. Show that the KKT conditions yield that on $\mathcal{T}$, for $j \in \hat{S} \backslash S_0$, and $\lambda > \lambda_0$,

$$2(\hat{\Sigma}(\hat{\beta} - \beta^0))_j = -(\lambda - \lambda_0)\tilde{\tau}_j,$$

where $|\tilde{\tau}_j| \le 1$. Here $\hat{\Sigma} := \mathbf{X}^T\mathbf{X}/n$. Show that the irrepresentable condition

$$\sup_{\|\tau_{S_0}\|_\infty \le 1} \|\hat{\Sigma}_{2,1}(S_0)\hat{\Sigma}_{1,1}^{-1}(S_0)\tau_{S_0}\|_\infty < \frac{\lambda - \lambda_0}{\lambda + \lambda_0} \tag{7.30}$$

implies that $\hat{S} \subset S_0$ on $\mathscr{T}$, by arguing along the lines of the proof of Theorem 7.1.

7.6. We study the replacement of the Gram matrix $\hat{\Sigma}$ in the irrepresentable condition by some approximation Σ. Consider the situation of Problem 7.5. Suppose that for some set $\mathscr{T}_X$ and some $\tilde{\lambda}_X$, one has on the set $\mathscr{T} \cap \mathscr{T}_X$ that

$$\|\hat{\Sigma} - \Sigma\|_\infty \|\hat{\beta}_{\text{init}} - \beta^0\|_1 \le \tilde{\lambda}_X.$$

Show that the condition

$$\sup_{\|\tau_{S_0}\|_\infty \le 1} \|\Sigma_{2,1}(S_0)\Sigma_{1,1}^{-1}(S_0)\tau_{S_0}\|_\infty < \frac{\lambda_{\text{init}} - (\lambda_0 + \tilde{\lambda}_X)}{\lambda_{\text{init}} + (\lambda_0 + \tilde{\lambda}_X)}$$

suffices for the initial Lasso to have no false positives on $\mathscr{T} \cap \mathscr{T}_X$.
To appreciate this result, we remark the following. As we have seen (Theorem 7.7), on $\mathscr{T}$, and with $\lambda_{\text{init}} \ge 2\lambda_0$, we have

$$\|\hat{\beta}_{\text{init}} - \beta^*\|_1 = O(\lambda_{\text{init}}s_*)/\phi_{\text{comp}}^2(3, S_*)$$

(where $\phi_{\text{comp}}^2(3, S_*) = \phi_{\text{comp},\hat{\Sigma}}^2(3, S_*)$: see Section 6.12 for conditions for the replacement by $\phi_{\text{comp},\Sigma}(3, S_*)$). Suppose now that the approximation error $\|\beta^* - \beta^0\|_1$ is also of this order (see also Problem 6.4). Then the Σ-irrepresentable condition has slightly larger noise term $\lambda_0 + \tilde{\lambda}_X$ (instead of λ_0), but the additional $\tilde{\lambda}_X$ is of order $\lambda_{\text{init}}\lambda_X/\phi_{\text{comp}}^2(3, S_*)$, where $\lambda_X = \|\hat{\Sigma} - \Sigma\|_\infty$. So we can use the Σ-irrepresentable condition when the sparsity s_* is of order $s_* = O_{\text{suff}}(\lambda_X^{-1}\phi_{\text{comp}}^2(3, S_*))$.

7.7. This exercise (van de Geer et al., 2010) illustrates that the Lasso can select too many false positives, showing that the bound of Lemma 7.2 is sharp. We consider, as in Problem 7.1, the case of equal correlation in an idealized setting. Let P be a probability measure on $\mathscr{X} \times \mathbb{R}$ with marginal distribution Q on $\mathscr{X}$. We study a function $\mathbf{Y} \in L_2(P)$ satisfying $\mathbf{Y} = f^0 + \varepsilon$, where $f^0 = \sum_{j=1}^p \beta^0 \psi_j$, and where $\psi_1, \ldots, \psi_p$ are given functions in $L_2(Q)$. The Gram matrix is $\Sigma := \int \psi^T \psi dQ$, where $\psi := (\psi_1, \ldots, \psi_p)$. The $L_2(P)$ inner product is denoted by $(\cdot, \cdot)$, and $\|\cdot\|$ is the $L_2(P)$-norm. We let

$$\hat{\beta}_{\text{init}} := \arg\min_\beta \left\{ \|\mathbf{Y} - \sum_{j=1}^p \psi_j \beta_j\|^2 + \lambda_{\text{init}}\|\beta\|_1 \right\}.$$

To make our analysis more explicit, we throughout take $\lambda_{\text{init}} = 2\lambda_0$, where $\lambda_0 \geq 2 \max_{1 \leq j \leq p} |(\varepsilon, \psi_j)|$.

Let b be a p-vector of all 1's, and let

$$\Sigma := (1 - \rho)I + \rho bb^T,$$

where $0 \leq \rho < 1$.

(a) Show that for any set S with cardinality s

$$\Lambda_{\max}^2(\Sigma_{1,1}(S)) = 1 - \rho + \rho s.$$

Thus, in this example, the maximal eigenvalue $\Lambda_{\max}^2$ is equal to $1 - \rho + \rho p \geq \rho p$, i.e., it can be vary large.

(b) Verify moreover for any L (and for s and p even),

$$\phi_{\text{varmin}}^2(L, S, 2s) = \phi^2(L, S, 2s) = \Lambda_{\min}^2(\Sigma_{1,1}(S)) = 1 - \rho.$$

Assume that

$$\Delta := \frac{\rho s_0}{1 - \rho + \rho s_0} - \frac{\lambda_{\text{init}} - \lambda_0}{\lambda_{\text{init}} + \lambda_0} > 0$$

(compare with (7.30)), and

$$\frac{2}{2s_0 + 1} \leq \rho \leq \frac{1}{2}.$$

Also assume that

$$2(\varepsilon, \psi_j) = \begin{cases} -\lambda_0 & j \in S_0 \\ +\lambda_0 & j \notin S_0 \end{cases}.$$

The latter can be seen as the "worst case" correlation pattern. Some other and perhaps more typical correlation patterns (that increase the penalty on the true positives and decrease the penalty on the true negatives) will lead to similar conclusions but more involved calculations.

We further simplify the situation by assuming that

$$\beta_j^0 = \beta_0, \ \forall \ j \in S_0,$$

where β_0 is some positive constant. It is not difficult to see that the Lasso is then constant on S_0:

$$\hat{\beta}_{j,\text{init}} = \hat{\beta}_0, \ \forall \ j \in S_0,$$

where $\hat{\beta}_0$ is some non-negative constant. Moreover,

$$\hat{\beta}_{j,\text{init}} = \hat{\gamma}, \ \forall \ j \notin S_0,$$

where $\hat{\gamma}$ is some other constant.

(b) Show that when

$$\beta_0 > \left(\frac{\lambda_{\text{init}} + \lambda_0}{2(1 - \rho + \rho s_0)} + \frac{\rho(p - s_0)\Delta(\lambda_{\text{init}} + \lambda_0)}{2(1 - \rho)(1 - \rho + \rho p)} \right),$$

then

$$\hat{\beta}_0 = \beta_0 - \left(\frac{\lambda_{\text{init}} + \lambda_0}{2(1 - \rho + \rho s_0)} + \frac{\rho(p - s_0)\Delta(\lambda_{\text{init}} + \lambda_0)}{2(1 - \rho)(1 - \rho + \rho p)} \right),$$

and

$$\hat{\gamma} = \frac{\Delta(1 - \rho + \rho s_0)(\lambda_{\text{init}} + \lambda_0)}{2(1 - \rho)(1 - \rho + \rho p)}.$$

7.8. Prove Lemma 7.14, by combining the arguments of Lemma 7.7, with the idea of Problem 7.5.

7.9. Apply Lemma 6.12 to the adaptive Lasso with the *conservative* variant (see Section 7.6) for the choice of the weights. Does it improve the prediction error as compared to the commonly used adaptive Lasso that we studied in Section 7.12 (see Corollary 7.15)?

7.10. We give some guidelines for the proof of Lemma 7.5. First, check that

$$\|f_{\text{weight}} - f^0\|^2 + \lambda_{\text{init}}\lambda_{\text{weight}} \sum_{j=1}^{p} w_j |\beta_{\text{weight},j}| \leq \|f_{\beta_S} - f^0\|^2 + \lambda_{\text{init}}\lambda_{\text{weight}} \sum_{j \in S} w_j |\beta_j|.$$

Consider

Case i).
$$\|f_{\beta_S} - f^0\|^2 \leq \lambda_{\text{init}}\lambda_{\text{weight}} \|w_S\|_2 \|(\beta_{\text{weight}})_S - \beta_S\|_2,$$

and

Case ii)
$$\|f_{\beta_S} - f^0\|^2 > \lambda_{\text{init}}\lambda_{\text{weight}} \|w_S\|_2 \|(\beta_{\text{weight}})_S - \beta_S\|_2.$$

Show that in Case i),

$$\|f_{\text{weight}} - f^0\|^2 + \lambda_{\text{init}}\lambda_{\text{weight}} \|w_S\|_2 \|(\beta_{\text{weight}})_S - \beta_S\|_2 + \lambda_{\text{init}}\lambda_{\text{weight}} \sum_{j \notin S} w_j |\beta_{\text{weight},j}|$$

$$\leq 3\lambda_{\text{init}}\lambda_{\text{weight}} \|w_S\|_2 \|(\beta_{\text{weight}})_S - \beta_S\|_2.$$

In Case ii), verify that

$$\lambda_{\text{init}}\lambda_{\text{weight}} \|w_S\|_2 \|(\beta_{\text{weight}})_S - \beta_S\|_2 + \lambda_{\text{init}}\lambda_{\text{weight}} \sum_{j \notin S} w_j |\beta_{\text{weight},j}|$$

$$< 3\|f_{\beta_S} - f^0\|^2.$$

7.11. In this exercise, we sketch the proof of Lemma 7.6. The first result is a special case of Lemma 7.5, taking $\beta = \beta^*$ and $S = S_*$. The second result then follows from this lemma, as

$$\|\beta_{\text{init}} - \beta^*\|_1 \leq \sqrt{s_*} \|(\beta_{\text{init}})_{S_*} - \beta^*\|_2 + \|(\beta_{\text{init}})_{S_*^c}\|_1.$$

The third result follows from taking $\beta = \beta^*$ and $S = \mathcal{N}$ in Lemma 7.5, where $\mathcal{N}$ is the set S_*, complemented with the s_* largest - in absolute value - coefficients of $(\beta_{\text{init}})_{S_*^c}$. Then $\|f_{\beta_{\mathcal{N}}} - f^0\| = \|f^* - f^0\|$. Moreover $\phi_{\min}(2, \mathcal{N}, 2s_*) \leq \phi_*$. Thus, from Lemma 7.5, we get

$$\lambda_{\text{init}}\sqrt{2s_*}\|(\beta_{\text{init}})_{\mathcal{N}} - \beta_{\mathcal{N}}^*\|_2 + \lambda_{\text{init}}\|(\beta_{\text{init}})_{\mathcal{N}^c}\|_1 \leq 3\|f^* - f^0\|^2 + \frac{6\lambda_{\text{init}}^2 s_*}{\phi_*^2}.$$

Moreover, as is shown in Lemma 6.9 (with original reference Candès and Tao (2005), and Candès and Tao (2007)),

$$\|(\beta_{\text{init}})_{\mathcal{N}^c}\|_2 \leq \|(\beta_{\text{init}})_{S_*^c} - \beta_{S_*^c}^*\|_1/\sqrt{s_*} \leq \frac{3\|f^* - f^0\|^2 + 3\lambda_{\text{init}}^2 s_*/\phi_*^2}{\lambda_{\text{init}}\sqrt{s_*}}.$$

Conclude that

$$\|\beta_{\text{init}} - \beta^*\|_2 \leq \frac{6\delta_*^2}{\lambda_{\text{init}}\sqrt{s_*}}.$$

Chapter 8
Theory for ℓ_1/ℓ_2-penalty procedures

Abstract We study four procedures for regression models with group structure in the parameter vector. The first two are for models with univariate response variable. They are the so-called group Lasso (see Chapter 4), and the smoothed group Lasso for the high-dimensional additive model (see Chapter 5). We also discuss multivariate extensions, namely for the linear model with time-varying coefficients, for multivariate regression, and multitask learning.

8.1 Introduction

Let, for $i = 1, \ldots, n$, Y_i be the response variable in the univariate case, and $Y_{i,t} = Y_i(t)$, $t = 1, \ldots, T$, be the response vector in the multivariate case. The covariables are denoted as $X_i = \{X_i^{(j)}\}_{j=1}^p$, $i = 1, \ldots, n$ (where the $X_i^{(j)}$ may be vectors of, for instance, dummy variables). The co-variables are considered as non-random (i.e., fixed design). The empirical measure of the co-variables is $Q_n := \sum_{i=1}^n \delta_{X_i}$, and $\|\cdot\|_n$ is the $L_2(Q_n)$-norm.

The four models we consider are

Regression with group structure

$$Y_i = \sum_{j=1}^p \left(\sum_{t=1}^{T_j} X_{i,t}^{(j)} \beta_{j,t}^0 \right) + \varepsilon_i, \ i = 1, \ldots, n,$$

where the $\beta_j^0 := (\beta_{j,1}^0, \ldots, \beta_{j,T_j}^0)^T$ have the sparsity property $\beta_j^0 \equiv 0$ for "most" j,

High-dimensional additive model

$$Y_i = \sum_{j=1}^p f_j^0(X_i^{(j)}) + \varepsilon_i, \ i = 1, \ldots, n,$$

where the $f_j^0(X_i^{(j)})$ are (non-parametric) "smooth" functions, with sparsity property $f_j^0 \equiv 0$ for "most" j,

Linear model with time-varying coefficients

$$Y_i(t) = \sum_{j=1}^p X_i^{(j)}(t)\beta_j^0(t) + \varepsilon_i(t), \ i = 1,\ldots,n, \ t = 1,\ldots,T,$$

where the coefficients $\beta_j^0(\cdot)$ are "smooth" functions, with the sparsity property $\beta_j^0 \equiv 0$ for "most" j,

Multivariate linear model

$$Y_{i,t} = \sum_{j=1}^p X_{i,t}^{(j)}\beta_{j,t}^0 + \varepsilon_{i,t}, \ i = 1,\ldots,n, \ t = 1,\ldots,T,$$

with for $\beta_j^0 := (\beta_{j,1}^0,\ldots,\beta_{j,T}^0)^T$, the sparsity property $\beta_j^0 \equiv 0$ for "most" j.

To avoid digressions, we assume throughout that the errors $\{\varepsilon_i, \ i = 1,\ldots,n\}$ and $\{\varepsilon_{i,t} : t = 1,\ldots,T, \ i = 1,\ldots,n\}$ are independent and $\mathcal{N}(0,1)$-distributed, although at the end of Section 8.6, we indicate that results can be generalized assuming only bounded fourth moments for the errors.

The group Lasso was introduced by Yuan and Lin (2006). Oracle theory for small groups was given in Meier et al. (2008), Bach (2008), Chesneau and Hebiri (2008) and Nardi and Rinaldo (2008). With large groups, the standard group Lasso will generally not have good prediction properties. Therefore, one needs to impose a certain structure within groups. Such an approach has been considered by Meier et al. (2009), Ravikumar et al. (2009a), Koltchinskii and Yuan (2008), Koltchinskii and Yuan (2010), and van de Geer (2010). We refer to this approach as the smoothed group Lasso. For theoretical results for group Lasso as well as its multivariate extensions, see Lounici et al. (2009) and Lounici et al. (2010). In Meier and Bühlmann (2007), the model with time-varying coefficients is estimated using smoothing kernels.

8.2 Organization and notation of this chapter

For the four models, we use squared error loss, with an appropriate regularization penalty that matches with the idea of sparsity in each particular case. We will prove oracle inequalities in the same spirit as in Chapter 6, more precisely, Section 6.2. Also bounds for the ℓ_1/ℓ_2-estimation error are derived, which can in turn be used to prove screening properties (of the large coefficients or under beta-min conditions). The ingredients of the proofs are the same as in Chapter 6: an argument to handle

the random part of the problem, which we again call the empirical process part, and a compatibility condition.

We note that the models listed above have their counterparts in the framework of generalized linear models. Indeed, the results of this chapter can be extended, to general error distributions, general (convex) loss functions and general design.

The organization of this chapter is as follows. The group Lasso is studied in Section 8.3. Section 8.4 considers the high-dimensional additive model. There, and also in Section 8.5 which looks at the time-varying model, we partly postpone the handling of the random part of the problem to Chapter 14. Section 8.6, which is on the multivariate model, closes the loop. Each section consists of a subsection introducing the loss function and penalty, a subsection on the empirical process, a subsection introducing the compatibility condition, and then reaches its main result. The compatibility condition in the middle two sections need some further explanation, which is done in the last section of this chapter.

Here is some notation, applied throughout this chapter. We will use both notations with sub-scripts and with arguments, i.e., $\beta_{j,t} = \beta_j(t)$, $X_{i,t}^{(j)} = X_i^{(j)}(t)$, $\varepsilon_{i,t} = \varepsilon_i(t)$, whichever is more convenient. For $j = 1, \ldots, p$, $t = 1, \ldots, T_j$ and $i = 1, \ldots, n$, the $\beta_{j,t}$, $X_{i,t}^{(j)}$ and $\varepsilon_{i,t}$ are real-valued. The vector β_j is throughout the vector $\beta_j = (\beta_{j,1}, \ldots, \beta_{j,T_j})^T$, $j = 1, \ldots, p$. When $T_j := T$ is the same for all j, we can also consider the vector $\beta(t) := (\beta_{1,t}, \ldots, \beta_{p,t})^T$, $t = 1, \ldots, T$. The vector β^T will contain all parameters $\beta_{j,t}$. It is either the vector $(\beta_1^T, \ldots, \beta_p^T)$ or its re-ordered version $(\beta(1)^T, \ldots, \beta(T)^T)$.

For $t = 1, \ldots, T_j$, $j = 1, \ldots, p$, we define

$$\mathbf{X}_t^{(j)} := \begin{pmatrix} X_{1,t}^{(j)} \\ \vdots \\ X_{n,t}^{(j)} \end{pmatrix}.$$

For $j = 1, \ldots, p$, we let

$$\mathbf{X}^{(j)} := (\mathbf{X}_1^{(j)}, \ldots, \mathbf{X}_{T_j}^{(j)}) = \begin{pmatrix} X_{1,1}^{(j)} & \cdots & X_{1,T_j}^{(j)} \\ \vdots & & \vdots \\ X_{n,1}^{(j)} & \cdots & X_{n,T_j}^{(j)} \end{pmatrix},$$

and

$$\hat{\Sigma}^{(j)} := (\mathbf{X}^{(j)})^T (\mathbf{X}^{(j)})/n.$$

The diagonal elements of $\hat{\Sigma}^{(j)}$ are denoted by $\hat{\sigma}_{j,t}^2$. When $T_j := T$ for all j, we also define, for $t = 1, \ldots, T$,

$$\mathbf{X}_t := \mathbf{X}(t) := \begin{pmatrix} X_{1,t}^{(1)} & \cdots & X_{1,t}^{(p)} \\ \vdots & & \vdots \\ X_{n,t}^{(1)} & \cdots & X_{n,t}^{(p)} \end{pmatrix},$$

$$\hat{\Sigma}(t) := \mathbf{X}_t^T \mathbf{X}_t / n, \ t = 1, \ldots, T.$$

To exploit sparsity structures, we need the notation, for an index set $S \subset \{1, \ldots, p\}$,

$$\beta_{j,S} := \beta_j \mathbb{1}\{j \in S\}, \ j = 1, \ldots, p.$$

The cardinality of S is denoted by $s := |S|$. The active set of a vector of coefficients β is denoted by

$$S_\beta := \{j : \ \beta_j \equiv 0\},$$

with cardinality $s_\beta := |S_\beta|$. We will consider various oracles β^*, with active set $S_* := S_{\beta^*}$, with cardinality $s_* = |S_*|$.

Moreover, we define for $t = 1, \ldots T_j$, and $j = 1, \ldots, p$, the random quantities

$$V_{j,t} := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i X_{i,t}^{(j)} = \frac{1}{\sqrt{n}} \varepsilon^T \mathbf{X}_t^{(j)}$$

for the case of univariate response. For the case of multivariate response, where $T_j = T$ for all j, we define

$$W_{j,t} := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_{i,t} X_{i,t}^{(j)} = \frac{1}{\sqrt{n}} \varepsilon_t^T \mathbf{X}_t^{(j)}, \ t = 1, \ldots, T, \ j = 1, \ldots, p.$$

Note that, for each $j = 1, \ldots, p$, both $V_{j,t}$ and $W_{j,t}$ are $\mathcal{N}(0, \hat{\sigma}_{j,t}^2)$-distributed, that $V_j := (V_{j,1}, \ldots, V_{j,T_j})^T$ is $\mathcal{N}(0, \hat{\Sigma}^{(j)})$-distributed, whereas the collection $\{W_{j,t}\}_{t=1}^{T}$ consists of T independent random variables.

We consider penalized loss of the form

$$L_n(\beta) + \lambda \text{pen}(\beta),$$

where $L_n(\cdot)$ is squared error loss, λ is a regularization parameter, and $\text{pen}(\cdot)$ is an ℓ_1/ℓ_2-penalty, to be specified.

8.3 Regression with group structure

The results of this section are along the lines of Meier et al. (2008), and Chesneau and Hebiri (2008).

8.3.1 The loss function and penalty

Recall the model

$$Y_i = \sum_{j=1}^{p} X_i^{(j)} \beta_j^0 + \varepsilon_i, \ i = 1, \dots, n,$$

where $X_i^{(j)} \beta_j = \sum_{t=1}^{T_j} X_{i,t}^{(j)} \beta_{j,t}$.

We can write the model in matrix notation

$$\mathbf{Y} = \mathbf{X}\beta^0 + \varepsilon,$$

with over-all design matrix

$$\mathbf{X} := (\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(p)}) = (\mathbf{X}_1^{(1)}, \dots, \mathbf{X}_{T_1}^{(1)}, \dots, \mathbf{X}_1^{(p)}, \dots, \mathbf{X}_{T_p}^{(p)})$$

$$= \begin{pmatrix} X_{1,1}^{(1)} & \cdots & X_{1,T_1}^{(1)} & \cdots & X_{1,1}^{(p)} & \cdots & X_{1,T_p}^{(p)} \\ \vdots & & \vdots & & \vdots & & \vdots \\ X_{n,1}^{(1)} & \cdots & X_{n,T_1}^{(1)} & \cdots & X_{n,1}^{(p)} & \cdots & X_{n,T_p}^{(p)} \end{pmatrix},$$

an $n \times p\bar{T}$-matrix, with $\bar{T} = \sum_{j=1}^{p} T_j / p$ being the average group size.

The "truth" is now denoted as

$$\mathbb{E}\mathbf{Y} := \mathbf{f}^0 = \mathbf{X}\beta^0.$$

The loss function is

$$L_n(\beta) := \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 / n$$

and the group Lasso penalty is

$$\mathrm{pen}(\beta) := \sum_{j=1}^{p} \left(\|\mathbf{X}^{(j)}\beta_j\|_2 \right) \sqrt{T_j / n},$$

see also Section 4.2.1.

Note that the penalty is invariant under within-group reparametrizations. Therefore, without loss of generality, we assume $\hat{\Sigma}^{(j)} = I$, the $(T_j \times T_j)$-identity matrix. Thus, the penalty is

$$\mathrm{pen}(\beta) = \sqrt{\bar{T}} \|\beta\|_{2,1},$$

where $\|\beta\|_{2,1}$ is the ℓ_1 / ℓ_2-norm

$$\|\beta\|_{2,1} := \sum_{j=1}^{p} \|\beta_j\|_2 \sqrt{T_j / \bar{T}}.$$

The group Lasso is

$$\hat{\beta} := \arg\min_{\beta}\left\{ \|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + \lambda\sqrt{\bar{T}}\|\beta\|_{2,1} \right\}.$$

8.3.2 The empirical process

The empirical process is
$$\nu_n(\beta) := 2\varepsilon^T\mathbf{X}\beta/n$$

$$= 2\varepsilon^T\sum_{j=1}^{p}\mathbf{X}^{(j)}\beta_j/n = \frac{2}{\sqrt{n}}\sum_{j=1}^{p}V_j^T\beta_j,$$

with $V_j^T := \varepsilon^T\mathbf{X}^{(j)}/\sqrt{n}$, $j = 1,\ldots,p$. Because $\hat{\Sigma}^{(j)} = I$, the random variable $\|V_j\|_2^2$ has a chi-square distribution with T_j degrees of freedom. Let for some $\lambda_0 > 0$ (see Lemma 8.1 for a suitable value)

$$\mathcal{T} := \left\{ \max_{1 \le j \le p} 4\|V_j\|_2^2/T_j \le n\lambda_0^2 \right\}.$$

Define $T_{\min} := \min\{T_j : j = 1,\ldots,p\}$.

Lemma 8.1. *Let, for $j = 1,\ldots,p$, the random variables χ_j^2 be chi-square distributed with T_j degrees of freedom. Then for all $x > 0$, and for*

$$\lambda_0^2 := \frac{4}{n}\left(1 + \sqrt{\frac{4x + 4\log p}{T_{\min}}} + \frac{4x + 4\log p}{T_{\min}}\right),$$

we have

$$\mathbf{P}\left(\max_{1 \le j \le p} 4\chi_j^2/T_j \ge n\lambda_0^2\right) \le \exp[-x].$$

Proof. By the inequality of Wallace (1959),

$$\mathbf{P}\left(\chi_j^2 \ge T_j(1+a)\right) \le \exp\left[-\frac{T_j}{2}\left(a - \log(1+a)\right)\right].$$

We now use that
$$a - \log(1+a) \ge \frac{a^2}{2(1+a)}.$$

This gives
$$\mathbf{P}(\chi_j^2 \ge T_j(1+a)) \le \exp\left[-\frac{T_j}{4}\left(\frac{a^2}{1+a}\right)\right].$$

Insert

$$a = \sqrt{\frac{4x}{T_j} + \frac{4x}{T_j}}.$$

Then

$$\frac{a^2}{1+a} \geq \frac{4x}{T_j},$$

so

$$\mathbf{P}\left(\chi_j^2 \geq T_j\left(1 + \sqrt{\frac{4x}{T_j}} + \frac{4x}{T_j}\right)\right) \leq \exp[-x].$$

Finally, apply the union bound. □

We have the following bound for the random part:

$$|v_n(\beta)| \leq \frac{2}{\sqrt{n}} \sum_{j=1}^{p} \|V_j\|_2 \|\beta_j\|_2$$

$$\leq \frac{2}{\sqrt{n}} \max_{1 \leq j \leq p} \frac{\|V_j\|_2}{\sqrt{T_j}} \sum_{j=1}^{p} \sqrt{T_j} \|\beta_j\|_2$$

$$= \frac{2}{\sqrt{n}} \max_{1 \leq j \leq p} \frac{\|V_j\|_2}{\sqrt{T_j}} \mathrm{pen}(\beta).$$

Therefore, on the set $\mathscr{T}$, it holds that

$$|v_n(\beta)| \leq \lambda_0 \mathrm{pen}(\beta).$$

This leads to a choice $\lambda > \lambda_0$ for the regularization parameter, as then the penalty overrules the random part.

8.3.3 The group Lasso compatibility condition

We let, for an index set $S \subset \{1,\ldots,p\}$,

$$\bar{T}_S = \sum_{j \in S} T_j/s,$$

be the average group size in the set S.

Definition *The* group Lasso compatibility condition *holds for the index set $S \subset \{1,\ldots,p\}$, with constant $\phi(S) > 0$, if for all $\|\beta_{S^c}\|_{2,1} \leq 3\|\beta_S\|_{2,1}$, one has that*

$$\bar{T} \|\beta_S\|_{2,1}^2 \leq \left(\|\mathbf{X}\beta\|_2^2/n \right) \bar{T}_s s / \phi^2(S).$$

As the next lemma shows, the group Lasso compatibility condition is not more restrictive than the adaptive restrictive eigenvalue condition.

Lemma 8.2. *Let $S \subset \{1, \ldots, p\}$ be an index set, say, $S = \{1, \ldots, s\}$. Consider the full index set corresponding to S:*

$$S_{\text{full}} := \{(1,1), \ldots, (1, T_1), \ldots, (s, 1), \ldots, (s, T_s)\},$$

with cardinality $\bar{T}_S s = \sum_{j=1}^s T_j$. Assume the adaptive restrictive eigenvalue condition holds, with constant $\phi_{\text{adap}}(3, S_{\text{full}}, \bar{T}_S s)$ (see Subsection 6.13.2 and, for an overview, Subsection 6.13.7). Then the group Lasso compatibility condition holds for S, with $\phi(S) \geq \phi_{\text{adap}}(3, S_{\text{full}}, \bar{T}_S s)$.

Proof of Lemma 8.2. First we observe that

$$\bar{T} \|\beta_S\|_{2,1}^2 = \left(\sum_{j \in S} \sqrt{T_j} \|\beta_j\|_2 \right)^2 \leq \sum_{j \in S} T_j \sum_{j \in S} \|\beta_j\|_2^2 = \bar{T}_S s \|\beta_{S_{\text{full}}}\|_2^2.$$

Suppose now that

$$\|\beta_{S^c}\|_{2,1} \leq 3 \|\beta_S\|_{2,1}.$$

Then

$$\sum_{j \notin S} \|\beta_j\|_1 \leq \sum_{j \notin S} \sqrt{T_j} \|\beta_j\|_2 = \sqrt{\bar{T}} \|\beta_{S^c}\|_{2,1} \leq 3\sqrt{\bar{T}} \|\beta_S\|_{2,1} \leq 3\sqrt{\bar{T}_S s} \|\beta_{S_{\text{full}}}\|_2.$$

So by the adaptive restricted eigenvalue condition

$$\|\beta_{S_{\text{full}}}\|_2^2 \leq \left(\|\mathbf{X}\beta\|_2^2 / n \right) / \phi_{\text{adap}}^2(3, S_{\text{full}}, \bar{T}_S s),$$

and hence

$$\bar{T} \|\beta_S\|_{2,1}^2 \leq \bar{T}_S s \|\beta_{S_{\text{full}}}\|_2^2 \leq \left(\|\mathbf{X}\beta\|_2^2 / n \right) \bar{T}_S s / \phi_{\text{adap}}^2(3, S_{\text{full}}, \bar{T}_S s).$$

$\square$

8.3.4 A group Lasso sparsity oracle inequality

Let $\mathscr{S}$ be a collection of index sets.

Definition of the oracle *Assume the group Lasso compatibility condition for the sets S in $\mathscr{S}$. The oracle β^* is*

$$\beta^* = \arg\min_{\beta:\, S_\beta \in \mathscr{S}} \left\{ \|\mathbf{X}\beta - \mathbf{f}^0\|_2^2/n + \frac{4\lambda^2 \bar{T}_{S_\beta} s_\beta}{\phi^2(S_\beta)} \right\}.$$

We define $\phi_* = \phi(S_{\beta^*})$.

Theorem 8.1. *Take the normalization $\hat{\Sigma}^{(j)} = I$ for all $j = 1, \ldots, p$. Consider the group Lasso*

$$\hat{\beta} = \arg\min_{\beta} \left\{ \|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + \lambda\sqrt{\bar{T}}\|\beta\|_{2,1} \right\},$$

where

$$\lambda \geq 4\lambda_0,$$

with

$$\lambda_0 = \frac{2}{\sqrt{n}} \sqrt{1 + \sqrt{\frac{4x + 4\log p}{T_{\min}}} + \frac{4x + 4\log p}{T_{\min}}}.$$

Then with probability at least $1 - \exp[-x]$, we have

$$\|\mathbf{X}\hat{\beta} - \mathbf{f}^0\|_2^2/n + \lambda\sqrt{\bar{T}}\|\hat{\beta} - \beta^*\|_{2,1} \leq 6\|\mathbf{X}\beta^* - \mathbf{f}^0\|_2^2/n + \frac{24\lambda^2 \bar{T}_{S_*} s_*}{\phi_*^2}.$$

Comparing this result with the one of Theorem 6.2, we see that there is perhaps little gain in the rate of convergence, i.e., it is essentially governed by the number of nonzero coefficients $\bar{T}_{S_*} s_* = \sum_{j \in S_*} T_j$ of the oracle, where $\bar{T}_{S_*} s_*$ is its full set of active variables. Nevertheless, there may be a gain in the compatibility condition (see also Lemma 8.2), i.e., the constant ϕ_* in Theorem 8.1 may be smaller than its counterpart in Theorem 6.2. Moreover, if the smallest group size $T_{\min}$ is bigger than (say) $2\log p$ and if we take $x = \log p$ (say), we get

$$\lambda_0 \leq 2\sqrt{7/n},$$

i.e., we win a $(\log p)$-term in the lower bound for the regularization parameter λ (and the probability of the result is at least $1 - 1/p$). In Huang and Zhang (2010), one can find a further discussion of the advantages of the group Lasso over the Lasso.

Proof of Theorem 8.1. The result follows from the Basic Inequality

$$\|\mathbf{X}\hat{\beta} - \mathbf{f}^0\|_2^2/n + \lambda\sqrt{\bar{T}}\|\hat{\beta}\|_{2,1} \leq v_n(\hat{\beta} - \beta^*) + \lambda\sqrt{\bar{T}}\|\beta^*\|_{2,1} + \|\mathbf{X}\beta^* - \mathbf{f}^0\|_2^2/n.$$

So on the set $\mathscr{T}$,

$$\|\mathbf{X}(\hat{\beta} - \beta^*)\|_2^2/n + \lambda\sqrt{\bar{T}}\|\hat{\beta}\|_{2,1}$$
$$\leq \lambda_0\sqrt{\bar{T}}\|\hat{\beta} - \beta^*\|_{2,1} + \lambda\sqrt{\bar{T}}\|\beta^*\|_{2,1} + \|\mathbf{X}\beta^* - \mathbf{f}^0\|_2^2/n.$$

The rest of the proof is now exactly as the one of Theorem 6.2. $\qquad\qquad\square$.

8.3.5 Extensions

The penalty may be extended to be of the form

$$\text{pen}(\beta) = \sum_{j=1}^{p} \|A_j \beta_j\|_2,$$

where A_j is a given symmetric positive definite $(T_j \times T_j)$-matrix, $j = 1, \ldots, p$, see Section 4.5. In the previous subsection, we took $A_j^T A_j = T_j \hat{\Sigma}^{(j)}$ for all j. The extension to other quadratic forms causes no additional theoretical complications, provided one can still handle the empirical process in terms of the new penalty.

We have

$$v_n(\beta) := \frac{2}{\sqrt{n}} \sum_{j=1}^{p} V_j^T \beta_j$$

$$\leq \frac{2}{\sqrt{n}} \max_{1 \leq j \leq p} \|A_j^{-1} V_j\|_2 \text{pen}(\beta).$$

The set $\mathcal{T}$ is then to be chosen as

$$\mathcal{T} := \{4 \max_{1 \leq j \leq p} \|A_j^{-1} V_j\|_2^2 \leq n\lambda_0^2\}.$$

To prove that the set $\mathcal{T}$ has large probability, we can invoke a straightforward extension of Lemma 8.1 to more general quadratic forms.

A convenient normalization is to suppose that for all j, $\text{trace}(A_j^{-1} \hat{\Sigma}^{(j)} A_j^{-1})$ is constant, say

$$\text{trace}(A_j^{-1} \hat{\Sigma}^{(j)} A_j^{-1}) = 1.$$

Then, for all j,

$$\mathbb{E}\|A_j^{-1} V_j\|_2^2 = 1.$$

8.4 High-dimensional additive model

In this section, we discuss some results obtained in Meier et al. (2009), and van de Geer (2010).

8.4.1 The loss function and penalty

The model is

$$Y_i = \sum_{j=1}^{p} f_j^0(X_i^{(j)}) + \varepsilon_i, \ i = 1,\ldots,n.$$

The functions f_j^0 are defined on the space of the covariables $X_i^{(j)}$. We moreover define $f^0 := \sum_{j=1}^{p} f_j^0$. The functions f_j^0 are assumed to be smooth, in the sense (see Chapter 5) that there exists an expansion

$$f_j^0(X_i^{(j)}) = \sum_{t=1}^{T} b_{j,t}(X_i^{(j)})\beta_{j,t}^0,$$

where the $\{b_{j,t}(\cdot)\}_{t=1}^{T}$ are given base functions (feature mappings). We assume that the number of base functions T is at most n, and, to simplify, that it is the same for each group j. (The latter simplification poses little restrictions, as it will turn out that the group size plays a less prominent role, due to the application of a smoothness penalty.)

Writing

$$X_{i,t}^{(j)} := b_{j,t}(X_i^{(j)}), \ j = 1,\ldots,p, \ t = 1,\ldots,T, \ i = 1,\ldots,n,$$

brings us back to the linear model of the previous subsection:

$$Y_i = \sum_{j=1}^{p}\sum_{t=1}^{T} X_{i,t}^{(j)}\beta_{j,t}^0 + \varepsilon_i \ i = 1,\ldots,n,$$

which reads in matrix notation

$$\mathbf{Y} = \mathbf{X}\beta^0 + \varepsilon.$$

The difference with the previous subsection is that on top of the group structure, there is also some further within-group structure. We translate the smoothness of f_j^0 in terms of bounds for some quadratic norm. For example, when considering the Sobolev space of twice continuously differentiable functions, one may use a finite-dimensional space of natural cubic splines, with some basis $\{b_{j,t}\}$, and write the squared Sobolev semi-norm $\int |f_j''|^2$ of a function $f_j = \sum_t b_{j,t}\beta_{j,t}$ as quadratic form $\beta_j^T W_j \beta_j$, with W_j a given matrix of weights, see Section 5.3.

Returning to the general situation, we let B_j be some given matrix and we apply the penalty

$$\text{pen}(\beta) := \sum_{j=1}^{p} \|\mathbf{X}^j\beta_j\|_2/\sqrt{n} + \mu \sum_{j=1}^{p} \|B_j\beta_j\|_2,$$

where μ is another smoothing parameter. The additional term $\mu\sum_{j=1}^{p}\|B_j\beta_j\|_2$ represents the regularization to achieve smoothness within groups. The matrices B_j will for this reason be called *smoothness* matrices. Note also that we do not merge the two norms into one. The latter gives the alternative penalty

$$\sum_{j=1}^{p} \sqrt{\|\mathbf{X}^j\beta_j\|_2^2/n + \mu^2\|B_j\beta_j\|_2^2}$$

(see Section 5.4). This alternative penalty has some theoretical difficulties. We will make a theoretical comparison of various penalties in Subsection 8.4.5.

Again, without loss of generality, we assume that $\hat{\Sigma}^{(j)} = I$. The penalty then becomes

$$\text{pen}(\beta) = \sum_{j=1}^{p} \|\beta_j\|_2 + \mu \sum_{j=1}^{p} \|B_j\beta_j\|_2$$

$$:= \|\beta\|_{2,1} + \mu\|B\beta\|_{2,1},$$

where $B := (B_1, \ldots, B_p)$.

We examine squared error loss

$$L_n(\beta) := \|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n.$$

The smoothed group Lasso is defined as

$$\hat{\beta} := \arg\min_{\beta}\left\{\|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + \lambda\|\beta\|_{2,1} + \lambda\mu\|B\beta\|_{2,1}\right\}.$$

8.4.2 The empirical process

The empirical process is

$$\nu_n(\beta) = 2\varepsilon^T\mathbf{X}\beta/n = \frac{2}{\sqrt{n}}\sum_{j=1}^{p} V_j^T\beta_j.$$

The behavior of the empirical process in terms of the penalty depends of course heavily on the choice of the smoothness matrices B_j, $j = 1, \ldots, p$. One needs to show that for some λ_0 and μ_0, the set

$$\mathscr{T} := \left\{|\nu_n(\beta)| \le \lambda_0\|\beta\|_{2,1} + \lambda_0\mu_0\|B\beta\|_{2,1}, \, \forall \beta\right\}$$

has large probability. We consider two cases: Sobolev smoothness (where we only sketch the results) and diagonalized smoothness. Throughout, we assume that $\hat{\Sigma}^{(j)} = I$ for all j, i.e., that we have a *normalized* design.

8.4.2.1 Sobolev smoothness

Here, we give a brief indication of the behavior of the empirical process when $\|B_j\beta_j\|_2$ corresponds to the Sobolev semi-norm in the space of twice continuously differentiable functions. More details can be found in Section 14.11.

When $\|B_j\beta_j\|_2^2 = \int |f''_{j,\beta_j}|^2$, with $f_{j,\beta_j}(\cdot) = \sum b_{j,t}(\cdot)\beta_{j,t}$, it can be shown that for normalized design, $\mathbf{P}(\mathscr{T})$ is large for $\lambda_0 = \mu_0 = O(\log p/n)^{2/5}$. The idea to prove this is that, for $\alpha = 3/4$, the following result for the increments of the empirical process holds: for $n \to \infty$,

$$\sup_{\|B_j\beta_j\|_2 \le 1} \frac{\left|V_j^T\beta_j\right|}{\|\beta_j\|_2^\alpha} = O_\mathbf{P}(1), \ \forall \ j,$$

see Corollary 14.8. In fact, one can prove that the result holds uniformly in j at the cost of a $(\log p)$-factor:

$$\max_{1 \le j \le p} \sup_{\|B_j\beta_j\|_2 \le 1} \frac{\left|V_j^T\beta_j\right|}{\|\beta_j\|_2^\alpha} = O_\mathbf{P}(\sqrt{\log p}).$$

This follows from the exponential bound in Corollary 14.6. So then

$$|v_n(\beta)| \le O_\mathbf{P}\left(\sqrt{\frac{\log p}{n}}\right) \sum_{j=1}^p \|\beta_j\|_2^\alpha \|B_j\beta_j\|_2^{1-\alpha}.$$

Using the inequality (which holds for any $0 < \alpha < 1$)

$$a^\alpha b^{1-\alpha} \le a+b, \ a,b \ge 0,$$

this leads to the choice

$$\lambda_0 = O\left(\frac{\log p}{n}\right)^{\frac{1}{2(2-\alpha)}}, \ \mu_0 = O\left(\frac{\log p}{n}\right)^{\frac{1}{2(2-\alpha)}}.$$

So far, we formulated the results for the space of twice differentiable functions, with $\alpha = 3/4$. The results can be extended to spaces of m times differentiable functions. In that case, one takes $\alpha = 1 - 1/(2m)$.

8.4.2.2 Diagonalized smoothness

We consider now a rather explicit description of "smoothness", with as main purpose to be able to provide a simple derivation of the behavior of the empirical pro-

cess. This derivation is given in Lemma 8.4. It is based on Cauchy-Schwarz and Hölder inequalities, and on the behavior of the maximum of a finite number of Gaussian or chi-squared distributed random variables. The counterpart of Lemma 8.4, for Sobolev semi-norms as sketched above, is be based on entropy arguments (see Section 14.11).

Consider the penalty

$$\text{pen}(\beta) := \sum_{j=1}^{p} \left[\|\beta_j\|_2 + \mu \|D\beta_j\|_2 \right],$$

where μ is a regularization parameter, and

$$D := \text{diag}(d_1, d_2, \ldots),$$

i.e.,

$$\|D\beta_j\|_2^2 = \sum_t d_t^2 \beta_{j,t}^2, \ j = 1, \ldots, p.$$

Moreover, we assume that d_t is increasing in t, in fact, that

$$d_t = t^m,$$

for some $m > 1/2$[1].

We define

$$\chi_j^2 := \sum_{t=1}^{T_0} V_{j,t}^2,$$

where $T_0 \leq T$ is a *hidden* truncation level. We take

$$T_0 := \lfloor n^{\frac{1}{2m+1}} \rfloor + 1, \tag{8.1}$$

tacitly assuming that T, the number of functions $b_{j,t}$ as t varies, is at least this large. Note that, in the case of normalized design, for each j, the $\{V_{j,t}\}$ are i.i.d. $\mathcal{N}(0,1)$-distributed, and hence that χ_j^2 is $\chi_{T_0}^2$-distributed. For the random quantities, we derive bounds involving the following expressions (based on Lemmas 8.1 and 6.2). Let, for $x > 0$,

$$v_0^2 := \frac{2x + 2\log(pT)}{2m - 1},$$

and

$$\xi_0^2 := 1 + \sqrt{\frac{4x + 4\log p}{n^{\frac{1}{2m+1}}}} + \frac{4x + 4\log p}{n^{\frac{1}{2m+1}}}.$$

Take

$$\lambda_0^2 := 4n^{-1} + 4n^{-\frac{2m}{m+1}} \xi_0^2,$$

---

[1] One may, loosely, think of the $\{b_{j,t}\}_{t=1}^{T}$ as *eigenfunctions*, and the $\{d_t\}_{t=1}^{T}$ as *eigenvalues*, for the space of functions having m derivatives (say the Sobolev space $\{f_j : [0,1] \to \mathbb{R}, \int |f_j^{(m)}|^2 < \infty\}$).

and

$$\lambda_0^2 \mu_0^2 := 4n^{-\frac{4m}{2m+1}} v_0^2.$$

We begin with a technical intermediate result, that we state here for later reference.

Lemma 8.3. *For all j, it holds that*

$$\frac{1}{\sqrt{n}}|V_j^T \beta_j| \le \left(\chi_j / \sqrt{n} \right) \|\beta_j\|_2$$

$$+ \left(\max_t |V_{j,t}| \left[n(2m-1)T_0^{2m-1} \right]^{-1/2} \right) \|D\beta_j\|_2. \tag{8.2}$$

Proof. Apply the inequality of Cauchy-Schwarz and Hölder's inequality, respectively. This gives

$$\frac{1}{\sqrt{n}}|V_j^T \beta_j| \le \chi_j \sqrt{\frac{1}{n}\sum_{t=1}^{T_0} \beta_{j,t}^2} + \frac{1}{\sqrt{n}} \max_t |V_{j,t}| \sum_{t>T_0} |\beta_{j,t}|$$

$$\le \left(\chi_j / \sqrt{n} \right) \|\beta_j\|_2 + \max_t |V_{j,t}| \sqrt{\sum_{t>T_0} t^{-2m}/n} \|D\beta_j\|_2.$$

Moreover,

$$\sum_{t>T_0} t^{-2m} \le \left[(2m-1)T_0^{2m-1} \right]^{-1}.$$

$\square$

Lemma 8.4. *Consider normalized design, that is $\hat{\Sigma}_j = I$ for all j. We have with probability at least $1 - 3\exp[-x]$, simultaneously $\forall \beta$,*

$$|v_n(\beta)| \le \lambda_0 \left(\sum_{j=1}^{p} \|\beta_j\|_2 + \mu_0 \sum_{j=1}^{p} \|D\beta_j\|_2 \right).$$

Proof. Remember that

$$v_n(\beta) = \frac{2}{\sqrt{n}} \sum_{j=1}^{p} V_j^T \beta_j.$$

The choice

$$T_0 = \lfloor n^{\frac{1}{2m+1}} \rfloor + 1,$$

has $n^{\frac{1}{2m+1}} \le T_0 \le 1 + n^{\frac{1}{2m+1}}$ and hence

$$\frac{T_0}{n} \le n^{-1} + n^{-\frac{2m}{2m+1}},$$

and

$$\frac{1}{nT_0^{2m-1}} \leq n^{-\frac{4m}{2m+1}}.$$

So by Lemma 8.3

$$\frac{1}{\sqrt{n}}|V_j^T \beta_j| \leq \left(\sqrt{n^{-1} + n^{-\frac{2m}{2m+1}}} \chi_j/\sqrt{T_0} \right) \|\beta_j\|_2$$

$$+ \left(\max_t |V_{j,t}| n^{-\frac{2m}{2m+1}} /\sqrt{(2m-1)} \right) \|D\beta_j\|_2.$$

The result now follows from Lemma 6.2 and Lemma 8.1 which show that with probability at least $1 - 3\exp[-x]$, it holds that

$$\chi_j/\sqrt{T_0} \leq \xi_0, \ \max_t |V_{j,t}| \leq \nu_0 \sqrt{2m-1}, \forall \ j, t.$$

$\square$

8.4.3 The smoothed Lasso compatibility condition

We consider functions

$$f_{j,\beta_j}(\cdot) = \sum_{t=1}^{T} b_{j,t}(\cdot) \beta_{j,t}.$$

Without loss of generality, we may think of the X_i all in some common space $\mathcal{X}$ and all f_{j,β_j} as being defined on this space $\mathcal{X}$. We write

$$f_\beta := \sum_{j=1}^{p} f_{j,\beta_j}, \ \beta^T := (\beta_1^T, \dots, \beta_p^T).$$

Let $\|\cdot\|$ be some norm on the set of real-valued functions on $\mathcal{X}$. Write furthermore

$$\text{pen}_2(\beta) := \mu \|B\beta\|_{2,1} = \mu \sum_j \|B_j\beta_j\|_2.$$

We assume that the $L_2(Q_n)$-norm $\|\cdot\|_n$ can be approximated by $\|\cdot\|$, in the following sense:

Definition *The* approximation condition *holds for the additive model if there exists an* $\eta \geq 0$ *such that for all* β, *we have* [2]

$$\frac{\left| \|f_\beta\|_n - \|f_\beta\| \right|}{\sum_j \|f_{j,\beta_j}\| + \mathrm{pen}_2(\beta)} \leq \eta.$$

An illustration of the approximation condition is given in Section 8.7

With this approximation condition, the compatibility condition can be formulated in terms of $\|\cdot\|$ instead of $\|\cdot\|_n$. Let us recall the advantage of being able to switch to a different norm. In $L_2(Q_n)$, any collection of more than n vectors will be linearly dependent. With a different (Hilbert) norm, we may have linear independence, and even reasonably large eigenvalues. We in fact only need to control (a lower bound for) the *compatibility constant* (denoted by $\phi(S)$ below).

Definition *The* smoothed group Lasso compatibility condition *holds for the set S, with constant* $\phi(S) > 0$, *if for all* β *with*

$$\sum_{j \notin S} \|f_{j,\beta_j}\| + \mathrm{pen}_2(\beta)/3 \leq 5 \sum_{j \in S} \|f_{j,\beta_j}\|, \tag{8.3}$$

it holds that

$$\sum_{j \in S} \|f_{j,\beta_j}\|^2 \leq \|f_\beta\|^2/\phi^2(S).$$

We note that the collection of f_{j,β_j} that satisfy (8.3) can neither be too non-sparse (since $\sum_{j \notin S} \|f_{j,\beta_j}\| \leq 5 \sum_{j \in S} \|f_{j,\beta_j}\|$) nor too non-smooth (since $\mathrm{pen}_2(\beta)/3 \leq 5 \sum_{j \in S} \|f_{j,\beta_j}\|$). That is, we restrict the f_{j,β_j} substantially, which makes the compatibility condition true for a relatively large number of situations.

8.4.4 A smoothed group Lasso sparsity oracle inequality

We define

$$\mathrm{pen}_1(\beta) := \|\beta\|_{2,1} = \sum_j \|\beta_j\|_2, \ \mathrm{pen}_2(\beta) := \mu\|B\beta\|_{2,1} = \mu \sum_j \|B_j\beta_j\|_2,$$

and

---

[2] Compare with Lemma 6.17. Its proof employs the following observation. Let $\|f_\beta\|_n^2 := \beta^T \hat{\Sigma} \beta$ and $\|f_\beta\|^2 := \beta^T \Sigma \beta$. Then $\left| \|f_\beta\|_n^2 - \|f_\beta\|^2 \right| \leq \tilde{\lambda} \|\beta\|_1^2$, where $\tilde{\lambda} := \max_{j,k} |\hat{\sigma}_{j,k} - \sigma_{j,k}|$. So then $\left| \|f_\beta\|_n - \|f_\beta\| \right| / \|\beta\|_1 \leq \sqrt{\tilde{\lambda}} := \eta$.

$$\text{pen}(\beta) := \text{pen}_1(\beta) + \text{pen}_2(\beta).$$

Let $\mathscr{S}$ be a collection of index sets.

Definition of the oracle *Suppose the smoothed group Lasso compatibility condition holds for all S in $\mathscr{S}$. The oracle is*

$$\beta^* := \arg\min\left\{\frac{16\lambda^2|S_\beta|}{\phi^2(S_\beta)} + \|f_\beta - f^0\|_n^2 + 2\text{pen}_2(\beta) : \right.$$

$$\left. S_\beta \in \mathscr{S}, \sqrt{|S_\beta|}\eta/\phi(S_\beta) \le 1/16\right\}.$$

Moreover, we let $S_* := S_{\beta^*}$, $s_* := |S_*|$, $\phi_* := \phi(S_*)$, and $f^* := f_{\beta^*}$.

Observe that we minimize over a restricted set of β, tacitly assuming that this set is not empty.

We are now ready to formulate an oracle inequality for the smoothed group Lasso. Before doing so, we present the first steps of the proof, for later reference when considering alternative penalties (Subsection 8.4.5).

Let $\mathscr{T}$ be the set

$$\mathscr{T} := \{2|\varepsilon^T\mathbf{X}\beta|/n \le \lambda_0\|\beta\|_{2,1} + \lambda_0\mu_0\|B\beta\|_{2,1}, \,\forall\, \beta\}. \tag{8.4}$$

By Lemma 8.4, in the case of diagonalized smoothness, with the values λ_0 and μ_0 defined there, the probability of $\mathscr{T}$ is large (for x large):

$$\mathbf{P}(\mathscr{T}) \ge 1 - 3\exp[-x].$$

This can be accomplished by taking

$$\lambda_0 \asymp n^{-\frac{m}{2m+1}}, \mu_0 \asymp n^{-\frac{m}{2m+1}}\sqrt{\log(pn)}$$

(using $T \le n$, and assuming that $\log p/n^{\frac{1}{2m+1}} = O(1)$). Such order of magnitude is appropriate for more general smoothness semi-norms, with m being the smoothness, e.g, the number of derivatives a function possesses.

Let us write as usual

$$\hat{f} := \sum_{j=1}^p \hat{f}_j, \, \hat{f}_j := f_{j,\hat{\beta}_j},$$

$$f^* := \sum_{j=1}^p f_j^*, \, f_j^* := f_{j,\beta_j^*}.$$

Recall the Basic Inequality

$$\|\hat{f} - f^0\|_n^2 + \lambda\text{pen}(\hat{\beta}) \le 2(\varepsilon, \hat{f} - f^*)_n + \lambda\text{pen}(\beta^*) + \|f^* - f^0\|_n^2.$$

Hence, on $\mathscr{T}$, when $\lambda \ge 4\lambda_0$ and $\lambda\mu \ge 4\lambda_0\mu_0$,

$$\|\hat{f} - f^0\|_n^2 + \lambda \operatorname{pen}(\hat{\beta}) \le \frac{\lambda}{4}\operatorname{pen}(\hat{\beta} - \beta^*) + \lambda \operatorname{pen}(\beta^*) + \|f^* - f^0\|_n^2. \tag{8.5}$$

This inequality is the starting point of the proof of the next theorem. We remark that hence, the specific form of the penalty is mainly used to make sure that the set $\mathscr{T}$ has large probability, and that there are therefore many generalizations possible.

As a first step, we present a straightforward consequence of (8.5), which relies on the fact that $\operatorname{pen}(\beta) = \operatorname{pen}_1(\beta) + \operatorname{pen}_2(\beta)$. One easily checks that

$$4\|\hat{f} - f^0\|_n^2 + 3\lambda \operatorname{pen}(\hat{\beta}_{S^c_*}) + 3\lambda \operatorname{pen}_2(\hat{\beta}_{S_*} - \beta^*_{S_*})$$

$$\le 5\lambda \operatorname{pen}_1(\hat{\beta}_{S_*} - \beta^*_{S_*}) + 4\|f^* - f^0\|_n^2 + 8\lambda \operatorname{pen}_2(\beta^*). \tag{8.6}$$

Theorem 8.2. *Take the normalization $\hat{\Sigma}_j = I$ for all j. Consider the smoothed group Lasso*

$$\hat{\beta} := \arg\min_{\beta}\left\{\|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + \lambda\|\beta\|_{2,1} + \lambda\mu\|B\beta\|_{2,1}\right\}.$$

Suppose that $\lambda \ge 4\lambda_0$ and $\lambda\mu \ge 4\lambda_0\mu_0$. Then on

$$\mathscr{T} := \{2|\varepsilon^T\mathbf{X}\beta|/n \le \lambda_0\|\beta\|_{2,1} + \lambda_0\mu_0\|B\beta\|_{2,1}\},$$

we have

$$\|\hat{f} - f^0\|_n^2 + \lambda \operatorname{pen}(\hat{\beta} - \beta^*)/2 \le 3\left\{16\lambda^2 s_*/\phi_*^2 + \|f^* - f^0\|_n^2 + 2\lambda \operatorname{pen}_2(\beta^*)\right\}.$$

Asymptotics. As a consequence, say for the case of diagonalized smoothness, with $\lambda \asymp n^{-\frac{m}{2m+1}}$ and $\mu \asymp n^{-\frac{m}{2m+1}}\sqrt{\log(pn)}$ (using the bound $T \le n$, and assuming $\log p/n^{\frac{1}{2m+1}} = O(1)$), we get

$$\|\hat{f} - f^0\|_n^2 = O_{\mathbf{P}}\left(n^{-\frac{2m}{2m+1}} s_*/\phi_*^2 + \|f^* - f_0\|_n^2 + n^{-\frac{2m}{2m+1}}\sqrt{\log(pn)}\sum_{j\in S_*}\|D\beta_j^*\|_2\right). \tag{8.7}$$

Proof of Theorem 8.2. Throughout, we assume we are on $\mathscr{T}$.

Case i)

If

$$\lambda \operatorname{pen}_1(\hat{\beta}_{S_*} - \beta^*_{S_*}) \ge \|f^* - f^0\|_n^2 + 2\lambda \operatorname{pen}_2(\beta^*),$$

we get

$$4\|\hat{f} - f^0\|_n^2 + 3\lambda \operatorname{pen}(\hat{\beta}_{S^c_*}) + 3\lambda \operatorname{pen}_2(\hat{\beta}_{S_*} - \beta^*_{S_*}) \le 9\lambda \operatorname{pen}_1(\hat{\beta}_{S_*} - \beta^*_{S_*}). \tag{8.8}$$

So we then have

$$\operatorname{pen}(\hat{\beta}_{S^c_*}) + \operatorname{pen}_2(\hat{\beta}_{S_*} - \beta^*_{S_*}) \le 3\operatorname{pen}_1(\hat{\beta}_{S_*} - \beta^*_{S_*}).$$

We may apply the approximation condition to $\hat{\beta} - \beta^*$. This will allow us to switch to the norm $\|\cdot\|$, and then invoke the smoothed group Lasso compatibility condition. Recall that $\|\beta_j\|_2 = \|f_{j,\beta_j}\|_n$. By the approximation condition

$$\|\hat{\beta}_j - \beta_j^*\|_2 \le (1+\eta)\|\hat{f}_j - f_j^*\| + \eta\mu\|B_j(\hat{\beta}_j - \beta_j^*)\|_2,$$

and

$$\|\hat{\beta}_j - \beta_j^*\|_2 \ge (1-\eta)\|\hat{f}_j - f_j^*\| - \eta\mu\|B_j(\hat{\beta}_j - \beta_j^*)\|_2.$$

It follows that

$$3\text{pen}_1(\hat{\beta}_{S_*} - \beta_{S_*}^*) \le 3(1+\eta)\sum_{j\in S_*}\|\hat{f}_j - f_j^*\| + 3\eta\,\text{pen}_2(\hat{\beta}_{S_*} - \beta_{S_*}^*),$$

and

$$\text{pen}(\hat{\beta}_{S_*^c}) + \text{pen}_2(\hat{\beta}_{S_*} - \beta_{S_*}^*)$$
$$\ge (1-\eta)\sum_{j\notin S_*}\|\hat{f}_j\| + (1-\eta)\text{pen}_2(\hat{\beta}_{S_*^c}) + \text{pen}_2(\hat{\beta}_{S_*} - \beta_{S_*}^*).$$

Thus,

$$\sum_{j\notin S_*}\|\hat{f}_j\| + \text{pen}_2(\hat{\beta}_{S_*^c}) + \frac{1-3\eta}{1-\eta}\text{pen}_2(\hat{\beta}_{S_*} - \beta_{S_*}^*) \le 3\frac{1+\eta}{1-\eta}\sum_{j\in S_*}\|\hat{f}_j - f_j^*\|.$$

Since $\eta \le 1/4$, we conclude that

$$\sum_{j\notin S_*}\|\hat{f}_j\| + \text{pen}_2(\hat{\beta}_{S_*^c}) + \text{pen}_2(\hat{\beta}_{S_*} - \beta_{S_*}^*)/3 \le 5\sum_{j\in S_*}\|\hat{f}_j - f_j^*\|.$$

Hence, we may apply the smoothed group Lasso compatibility condition to $\hat{\beta} - \beta^*$. We now restart from (8.8), and add a term $3\lambda\,\text{pen}_1(\hat{\beta}_{S_*} - \beta_{S_*}^*)$ to the left- and right-hand side:

$$4\|\hat{f} - f^0\|_n^2 + 3\lambda\,\text{pen}(\hat{\beta}_{S_*^c}) + 3\lambda\,\text{pen}(\hat{\beta}_{S_*} - \beta_{S_*}^*) \le 12\lambda\,\text{pen}_1(\hat{\beta}_{S_*} - \beta_{S_*}^*)$$
$$\le 12\lambda(1+\eta)\sum_{j\in S_*}\|\hat{f}_j - f_j^*\| + 12\lambda\,\eta\,\text{pen}_2(\hat{\beta}_{S_*} - \beta_{S_*}^*),$$

which implies, using $\eta \le 1/16 \le 1/4$,

$$16\|\hat{f} - f^0\|_n^2 + 12\lambda\,\text{pen}(\hat{\beta}_{S_*^c}) + 9\lambda\,\text{pen}(\hat{\beta}_{S_*} - \beta_{S_*}^*) \le 60\lambda\sum_{j\in S_*}\|\hat{f}_j - f_j^*\|. \qquad (8.9)$$

Now

$$\lambda\sum_{j\in S_*}\|\hat{f}_j - f_j^*\| \le \lambda\sqrt{s_*}\sqrt{\sum_{j\in S_*}\|\hat{f}_j - f_j^*\|^2}$$
$$\le \lambda\sqrt{s_*}\|\hat{f} - f^*\|/\phi_*.$$

Moreover, switching back to the $\|\cdot\|_n$-norm,

$$\|\hat{f} - f^*\| \le \|\hat{f} - f^*\|_n + \eta \sum_j \|\hat{f}_j - f^*\| + \eta \, \text{pen}_2(\hat{f} - f^*)$$

$$\le \|\hat{f} - f^*\|_n + 6\eta \sum_{j \in S_*} \|\hat{f}_j - f^*\| + \frac{2}{3} \eta \, \text{pen}_2(\hat{f} - f^*).$$

We have thus established that

$$\lambda \sum_{j \in S_*} \|\hat{f}_j - f_j^*\| \le \frac{\lambda \sqrt{s_*}}{\phi_*} \left(\|\hat{f} - f^*\|_n + 6\eta \sum_{j \in S_*} \|\hat{f}_j - f^*\| + \frac{2}{3} \eta \, \text{pen}_2(\hat{\beta} - \beta^*) \right).$$

As $\sqrt{s_*} \eta / \phi_* \le 1/16$, we get by the approximation condition

$$\lambda \sum_{j \in S_*} \|\hat{f}_j - f_j^*\| \le \frac{8}{5} \frac{\lambda \sqrt{s_*}}{\phi_*} \left(\|\hat{f} - f^*\|_n + \frac{2}{3} \eta \, \text{pen}_2(\hat{\beta} - \beta^*) \right)$$

$$\le \frac{8}{5} \lambda \sqrt{s_*} \|\hat{f} - f^*\|_n / \phi^* + \text{pen}_2(\hat{\beta} - \beta^*)/15.$$

Returning to (8.9), we see that

$$16\|\hat{f} - f^0\|_n^2 + 12\lambda \, \text{pen}(\hat{\beta}_{S_*^c}) + 5\lambda \, \text{pen}(\hat{\beta}_{S_*} - \beta_{S_*}^*) \le 96\lambda \sqrt{s_*} \|\hat{f} - f^*\|_n / \phi_*.$$

Dividing by 8 and, to simplify the expression, inserting the bounds $12/8 \ge 5/8 \ge 1/2$ yields

$$2\|\hat{f} - f^0\|_n^2 + \lambda \, \text{pen}(\hat{\beta} - \beta^*)/2 \le 12\lambda \sqrt{s_*} \|\hat{f} - f^*\|_n / \phi_*.$$

Now, using the triangle inequality, and twice the bound $2ab \le a^2 + b^2$, we see that

$$12\lambda \sqrt{s_*} \|\hat{f} - f^*\|_n / \phi_* \le 12\lambda \sqrt{s_*} \|\hat{f} - f^0\|_n / \phi_* + 12\lambda \sqrt{s_*} \|f^* - f^0\|_n / \phi_*$$

$$\le \frac{36\lambda^2 s_*}{2\phi_*^2} + \|\hat{f} - f^0\|_n^2 + \frac{12\lambda^2 s_*}{2\phi_*^2} + 3\|f^* - f^0\|_n^2$$

$$= \|\hat{f} - f^0\|_n^2 + 3\left\{ \frac{16\lambda^2 s_*}{2\phi_*^2} + \|f^* - f^0\|_n^2 \right\}.$$

Case ii)

If

$$\lambda \, \text{pen}_1(\hat{\beta}_{S_*} - \beta_{S_*}^*) < \|f^* - f^0\|_n^2 + 2\lambda \, \text{pen}_2(\beta^*),$$

we obtain from (8.6),

$$4\|\hat{f} - f^0\|_n^2 + 3\lambda \, \text{pen}(\hat{\beta}_{S_*^c}) + 3\lambda \, \text{pen}_2(\hat{\beta}_{S_*} - \beta_{S_*}^*) \le 9\|f^* - f^0\|_n^2 + 18\lambda \, \text{pen}_2(\beta^*),$$

and hence

$$4\|\hat{f} - f^0\|_n^2 + 3\lambda\,\mathrm{pen}(\hat{\beta} - \beta^*) \leq 12\|f^* - f^0\|_n^2 + 24\lambda\,\mathrm{pen}_2(\beta^*).$$

$\square$

8.4.5 On the choice of the penalty

The penalty we proposed in the previous subsections is additive in β_j, with each term consisting of two parts. The first part $\|\beta_j\|_2$ is simply the ℓ_2-norm of β_j, and the second part $\|B_j\beta_j\|_2$ can be seen as representing the smoothness of f_{j,β_j}. We now consider some alternatives. Throughout, to simplify, we confine ourselves to the case of diagonalized smoothness, i.e., for all j, and for a given $m > 1/2$,

$$B_j = D,\ D\beta_j = \sum_t d_t \beta_{j,t},\ d_t = t^m.$$

8.4.5.1 Using only one of the two terms

One may ask the question: what happens if only one of the terms is included in the penalty? To address this, let us reconsider the empirical process $v_n(\beta)$. Lemma 8.4 shows that with our choice of penalty, and with a proper choice of the regularization parameters λ and μ, the empirical process is "overruled". In order to see what happens if we take a different penalty, we now leave the "hidden truncation level" T_0 to be specified, instead of fixing it at the value given in (8.1). The straightforward bound we applied throughout is

$$|v_n(\beta)| \leq \sum_{j=1}^{p} |V_j^T \beta_j| / \sqrt{n}.$$

We know from Lemma 8.1 and 6.2 respectively, that, uniformly in j, for $\chi_j^2 = \sum_{t=1}^{T_0} V_{j,t}^2$,

$$\chi_j / \sqrt{T_0} = O_{\mathrm{P}}(1)$$

(assuming that $\log p / T_0 = O(1)$). Furthermore, uniformly in j and t,

$$V_{j,t} = O_{\mathrm{P}}(\sqrt{\log(pT)}).$$

Thus, as in Lemma 8.3, for any given T_0, and taking care of the extreme values $\{0, T\}$ for T_0, we have

$$\frac{1}{\sqrt{n}}|V_j^T \beta_j| \leq O_{\mathrm{P}}\left(\sqrt{\frac{T_0}{n}}\right)\|\beta_j\|_2$$

$$+O_{\mathrm{P}}\left(\sqrt{\frac{\log(pT)}{n(T_0^{2m-1}\vee 1)}}\right)\|D\beta_j\|_2 1\{T_0 < T\}. \tag{8.10}$$

If the penalty only includes the terms $\|\beta_j\|_2$, this suggests taking the hidden truncation level T_0 equal to T. In that case, the second term vanishes, leaving as first term

$$\frac{1}{\sqrt{n}}|V_j^T\beta_j| \le O_{\mathrm{P}}\left(\sqrt{\frac{T}{n}}\right)\|\beta_j\|_2.$$

This leads to the choice

$$\lambda \asymp \sqrt{\frac{T}{n}}$$

for the regularization parameter. The rate of convergence (for the prediction error, say) is then of order

$$\|\hat{f} - f^0\|_n^2 = O_{\mathrm{P}}\left(\frac{T s_*}{n\phi_*^2} + \|f^* - f^0\|_n^2\right),$$

where $s_* = |S_{\beta^*}|$ is the sparsity index, and $f^* = f_{\beta^*}$ is a modified version of the oracle defined in a similar version as in Subsection 8.4.4, ϕ_* being an appropriate compatibility constant. Note that when taking the *actual* truncation level T of order $n^{\frac{1}{2m+1}}$, we then get

$$\lambda \asymp n^{-\frac{m}{2m+1}},$$

and hence

$$\|\hat{f} - f^0\|_n^2 = O_{\mathrm{P}}\left(n^{-\frac{2m}{2m+1}}s_*/\phi_*^2 + \|f^* - f^0\|_n^2\right).$$

The situation is then as in Ravikumar et al. (2009a), with T playing the role of a second regularization parameter. The estimator is then called the *SPAM* estimator (for Sparse Additive Modelling). Observe that there is no $\sqrt{\log(pT)}$-term in this case. However, the result is not directly comparable to (8.7) as the oracle β^*, and hence s_* and the approximation error $\|f^* - f^0\|_n^2$ are different entities. Let us spend a few thoughts on this matter here. Suppose that the true f^0 is additive: $f^0 = \sum_j f_j^0$. Suppose in addition that the f_j^0 are smooth, uniformly in j, in the sense that $f_j^0 = \sum_{t=1}^n \beta_{t,j}^0 b_{j,t}$ ($\{b_{j,t}\}_{t=1}^n$ being an orthonormal system in $L_2(Q_n)$ ($j = 1, \ldots, p$)), and $\sum_{t=1}^n t^{2m}|\beta_{j,t}^0|^2 \le 1$ for all j, and $\beta_{j,t}^0 = 0$ for all (j,t) with $j \notin S_0$. Let $s_0 = |S_0|$. Then an actual truncation level T of order $T \asymp n^{\frac{1}{2m+1}}$ allows for an approximation f_j^* with the first T basis functions, with approximation error $\|f_j^* - f_j^0\|_n^2 = O(n^{-\frac{2m}{2m+1}})$, uniformly in j. So then,

$$\sum_{j\in S_0}\|f_j^* - f_j^0\|_n^2 = O(n^{-\frac{2m}{2m+1}}s_0).$$

Let $f^* = \sum_j f_j^*$. Using the triangle inequality, the approximation error $\|f^* - f^0\|_n^2$ can be bounded by

$$\|f^* - f^0\|_n^2 = O(n^{-\frac{2m}{2m+1}} s_0^2),$$

i.e., we see the *squared* sparsity index s_0^2 appearing. To reduce this to s_0 may require the price of some log-factor, appearing say when using a switch of norms as in the approximation condition. We finally note that the above illustration assumes that a single truncation level T_0 provides a good approximation for all f_j^0. In general it will depend on the smoothness $\|D\beta_j^0\|_2$ of f_j^0. It is thus still to be clarified rigorously how SPAM of Ravikumar et al. (2009a) and the smooth group Lasso compare in various situations.

We now consider the other extreme case. If the penalty only includes the terms $\|D\beta_j\|_2$, then (8.10) suggests taking T_0 equal to 0. We then get

$$\frac{1}{\sqrt{n}}|V_j^T \beta_j| \le O_{\mathrm{P}}\left(\sqrt{\frac{\log(pT)}{n}}\right)\|D\beta_j\|_2.$$

This leads to the choice

$$\lambda\mu \asymp \sqrt{\frac{\log(pT)}{n}}.$$

This corresponds to the situation in Koltchinskii and Yuan (2008). We note that such an approach leads to the rate

$$\|\hat{f} - f^0\|_n^2 = O_{\mathrm{P}}\left(\sqrt{\frac{\log(pT)}{n}}\frac{s_*}{\phi_*^2} + \|f^* - f^0\|_n^2\right),$$

where the oracle f^*, the sparsity index s_* and the compatibility constant ϕ_* are defined in a analogous fashion as in the previous section. We conclude that the smoothness penalty alone possibly leads to a slower rate of convergence.

8.4.5.2 Taking a single square root

As explained in Section 5.4, one may prefer (from a computational point of view) to take the penalty

$$\mathrm{pen}(\beta) := \sum_{j=1}^p \sqrt{\|\beta_j\|_2 + \mu^2\|D\beta_j\|_2^2}.$$

However, our proof then leads to a, possibly substantial, loss in the bound for the rate of convergence. To explain this, let us, as before, write

$$\mathrm{pen}_1(\beta) := \sum_{j=1}^{p} \|\beta_j\|_2, \ \mathrm{pen}_2(\beta) := \mu \sum_{j=1}^{p} \|D\beta_j\|_2.$$

Inequality (8.5) now says that on the set $\mathscr{T}$ defined in (8.4),

$$\|\hat{f} - f^0\|_n^2 + \lambda \,\mathrm{pen}(\hat{\beta})$$

$$\leq \frac{\lambda}{4} \left(\mathrm{pen}_1(\hat{\beta} - \beta^*) + \mathrm{pen}_2(\hat{\beta} - \beta^*) \right) + \|f^* - f_0\|_n^2 + \lambda \,\mathrm{pen}(\beta^*).$$

Using the bound $\sqrt{a+b} \geq (\sqrt{a} + \sqrt{b})/\sqrt{2}$ we see that

$$\|\hat{f} - f^0\|_n^2 + \lambda \left(\frac{1}{\sqrt{2}} - \frac{1}{4} \right) \left(\mathrm{pen}_1(\hat{\beta}_{S^c_*}) + \mathrm{pen}_2(\hat{\beta}_{S^c_*}) \right) + \lambda \left(\frac{1}{\sqrt{2}} - \frac{1}{4} \right) \mathrm{pen}_2(\hat{\beta}_{S_*} - \beta^*_{S_*})$$

$$\leq \lambda \left(\frac{1}{4} + \frac{1}{\sqrt{2}} \right) \mathrm{pen}_1(\hat{\beta}_{S_*} - \beta^*_{S_*}) + \|f^* - f_0\|_n^2 + \lambda \left(1 + \frac{1}{\sqrt{2}} \right) \mathrm{pen}_2(\beta^*)$$

$$+ \lambda \left(1 - \frac{1}{\sqrt{2}} \right) \mathrm{pen}_1(\beta^*).$$

This is exactly inequality (8.6) above Theorem 8.2, except that in (8.6), $1/\sqrt{2}$ is equal to 1 instead. This means we have to adjust the constants, but what is worse, we are also confronted with an additional term $\lambda(1 - 1/\sqrt{2})\mathrm{pen}_1(\beta^*)$.

Alternatively, we may use the inequality

$$\|\hat{f} - f^0\|_n^2 + \lambda \left(1 - \frac{\sqrt{2}}{4} \right) \mathrm{pen}(\hat{\beta}_{S^c_*}) \leq \lambda \left(1 + \frac{\sqrt{2}}{4} \right) \mathrm{pen}(\hat{\beta}_{S_*} - \beta^*_{S_*}) + \|f^* - f^0\|_n^2$$

$$\leq \lambda \left(1 + \frac{\sqrt{2}}{4} \right) \left(\mathrm{pen}_1(\hat{\beta}_{S_*} - \beta^*_{S_*}) + \mathrm{pen}_2(\hat{\beta}_{S_*} - \beta^*_{S_*}) \right) + \|f^* - f^0\|_n^2.$$

But with this, we have no tools to bound the smoothness term

$$\mathrm{pen}_2(\hat{\beta}_{S_*}),$$

unless we impose additional assumptions, such as

$$\mathrm{pen}_2(\hat{\beta}_{S_*} - \beta^*_{S_*}) \leq \mathrm{const.} \ \mathrm{pen}_1(\hat{\beta}_{S_*} - \beta^*_{S_*}).$$

Problem 8.3 gives an example where indeed the separate square root penalty provides substantially smaller prediction error than the single square root penalty.

8.4.5.3 Taking the squared instead on the non-squared smoothness norm

Again, from a computational point of view, the following penalty may be easier:

$$\text{pen}(\beta) := \sum_{j=1}^{p} \|\beta_j\|_2 + \mu \sum_{j=1}^{p} \|D\beta_j\|_2^2.$$

However, then again our proof leads to a, possibly substantial, loss in the bound for the rate of convergence. To explain this, we write as before

$$\text{pen}_1(\beta) := \sum_{j=1}^{p} \|\beta_j\|_2.$$

Inequality (8.5) again gives that on the set $\mathscr{T}$ defined in (8.4),

$$\|\hat{f} - f^0\|_n^2 + \lambda \text{pen}(\hat{\beta})$$

$$\leq \frac{\lambda}{4}\text{pen}_1(\hat{\beta} - \beta^*) + \frac{\lambda\mu}{4} \sum_{j=1}^{p} \|D(\hat{\beta}_j - \beta_j^*)\|_2 + \|f^* - f_0\|_n^2 + \lambda\text{pen}(\beta^*).$$

This gives, as counterpart of inequality (8.6) presented above Theorem 8.2,

$$\|\hat{f} - f^0\|_n^2 + \frac{3}{4}\lambda\text{pen}_1(\hat{\beta}_{S_*^c}) + \lambda\mu \sum_{j \notin S_*} \left(\|D\hat{\beta}_j\|_2^2 - \|D\hat{\beta}_j\|_2/4 \right)$$

$$+ \lambda\mu \sum_{j \in S_*} \left(\|D(\hat{\beta}_j - \beta_j^*)\|_2^2/2 - \|D(\hat{\beta}_j - \beta_j^*)\|_2/4 \right)$$

$$\leq \frac{5}{4}\lambda\text{pen}_1(\hat{\beta}_{S_*} - \beta_{S_*}^*) + \|f^* - f_0\|_n^2 + 2\lambda\mu \sum_{j \in S_*} \|D\beta_j^*\|_2^2.$$

The problematic part is the term

$$\lambda\mu \sum_{j \notin S_*} \left(\|D\hat{\beta}_j\|_2^2 - \|D\hat{\beta}_j\|_2/4 \right).$$

If $\|D\hat{\beta}_j\|_2 \leq 1/4$, the j-th term will be negative, and hence needs to be transferred to the right-hand side. In the worst case, we therefore may get an additional term of order

$$\lambda\mu \sum_{j \notin S_*} \left(\|D\hat{\beta}_j\|_2^2 - \|D\hat{\beta}_j\|_2/4 \right),$$

which can be as large as $\asymp \lambda\mu(p-s_*)$, which is $\asymp \lambda^2(p-s_*)$ for $\lambda \asymp \mu$. So actually, this approach may give a huge prediction error.

8.4.5.4 Taking a single square root and adding a squared smoothness norm

A way to circumvent computational difficulties and keep nice theoretical properties
is using the penalty

$$\text{pen}(\beta) := \sum_{j=1}^{p} \sqrt{\|\beta_j\|_2^2 + \mu^2 \|D\beta_j\|_2^2} + \mu \sum_{j=1}^{p} \|D\beta_j\|_2^2.$$

As counterpart of inequality (8.6) presented above Theorem 8.2, we propose

$$\|\hat{f} - f^0\|_n^2 + \lambda \left(\frac{1}{\sqrt{2}} - \frac{1}{4} \right) \sum_{j \notin S_*} \left(\|\hat{\beta}_j\|_2 + \mu \|D\hat{\beta}_j\|_2 \right) + \lambda \mu \sum_{j \notin S_*} \|D\hat{\beta}_j\|_2^2$$

$$\leq \frac{5}{4} \lambda \sum_{j \in S_*} \|\hat{\beta}_j - \beta_j^*\|_2 + \|f^* - f^0\|_n^2 + \lambda \mu \sum_j (5\|D\beta_j^*\|_2/4 + \|D\beta_j^*\|_2^2)$$

$$+ \lambda \mu \sum_{j \in S_*} (5\|D\hat{\beta}_j\|_2/4 - \|D\hat{\beta}_j\|_2^2).$$

What we did here is move the term $\lambda \mu \sum_{j \in S_*} \|D\hat{\beta}_j\|_2$ to the right-hand side, i.e.,
it does not have a role anymore as penalty. It is simply overruled by its quadratic
version, as

$$\lambda \mu \sum_{j \in S_*} (5\|D\hat{\beta}_j\|_2/4 - \|D\hat{\beta}_j\|_2^2) \leq \frac{25}{64} \lambda \mu s_*.$$

One can therefore derive an oracle inequality of the same spirit as the one of Theo-
rem 8.2.

8.5 Linear model with time-varying coefficients

8.5.1 The loss function and penalty

The model is as in Section 5.8

$$Y_i(t) = \sum_{j=1}^{p} X_i^{(j)}(t)\beta_j^0(t) + \varepsilon_i(t), \ i = 1,\ldots,n, \ t = 1,\ldots,T.$$

The "truth" is denoted by

$$f^0(X_i(t),t) := \sum_{j=1}^{p} X_i^{(j)}(t)\beta_j^0(t), \ i = 1,\ldots,n, \ j = 1,\ldots,p, \ t = 1,\ldots,T.$$

We assume that $\beta_j^0(\cdot)$ is smooth, in the sense that there is an expansion of the form

$$\beta_j^0(t) = \sum_{r=1}^{R} b_{j,r}(t)\gamma_{j,r}, \ j = 1,\ldots,p, \ t = 1,\ldots,T.$$

For example, the $\{b_{j,r}(\cdot)\}_{r=1}^{R}$ may form a basis for a finite-dimensional space of cubic splines.

With the expansion, we have for each t a linear model: in matrix notation

$$\mathbf{Y}(t) = \mathbf{X}(t)\beta(t) + \varepsilon(t) = \left(\mathbf{X}^{(1)}(t)b^{(1)}(t), \cdots, \mathbf{X}^{(p)}(t)b^{(p)}(t) \right)\gamma + \varepsilon(t),$$

where
$$b^{(j)}(t) := (b_{j,1}(t),\ldots,b_{j,R}(t)), \ j = 1,\ldots,p,$$

and where
$$\mathbf{X}(t) := (\mathbf{X}^{(1)}(t),\ldots,\mathbf{X}^{(p)}(t)),$$

with $\mathbf{X}^{(j)}(t) := (X_1^{(j)}(t),\ldots,X_n^{(j)}(t))^T$, $j = 1,\ldots,p$. Furthermore,

$$\gamma := \begin{pmatrix} \gamma_1 \\ \vdots \\ \gamma_p \end{pmatrix} = \begin{pmatrix} \gamma_{1,1} \\ \vdots \\ \gamma_{1,R} \\ \vdots \\ \gamma_{p,1} \\ \vdots \\ \gamma_{p,R} \end{pmatrix}.$$

The time-varying coefficients are

$$\beta(t) = \beta_\gamma(t) := \begin{pmatrix} \beta_{1,\gamma_1}(t) \\ \vdots \\ \beta_{p,\gamma_p}(t) \end{pmatrix},$$

with
$$\beta_j(t) = \beta_{j,\gamma_j}(t) := b^{(j)}(t)\gamma_j, \ j = 1,\ldots,p.$$

The over-all design matrix is

$$\left(\mathbf{X}^{(1)}(t)b^{(1)}(t), \cdots, \mathbf{X}^{(p)}(t)b^{(p)}(t) \right).$$

The matrices $\mathbf{X}^{(j)}(t)b^{(j)}(t)$ are $n \times R$ matrices, which are clearly of rank 1. Hence, it is quite obvious that for exploiting the smoothness structure, one needs to analyze the T models simultaneously.

The squared error loss function is defined as the average squared error for each of the T linear models, i.e,

$$L_n(\gamma) := \frac{1}{nT} \sum_{t=1}^{T} \|\mathbf{Y}(t) - \mathbf{X}(t)\beta_\gamma(t)\|_2^2.$$

For a function $g : \mathbb{R} \to \mathbb{R}$, we use the notation

$$\|g\|_T^2 := \frac{1}{T} \sum_{t=1}^{T} g^2(t).$$

Then we have

$$\|\beta_{j,\gamma_j}\|_T = \|A_j\gamma_j\|_2,$$

with

$$A_j^T A_j = \frac{1}{T} \sum_{t=1}^{T} (b^{(j)}(t))^T (b^{(j)}(t)).$$

The penalty is chosen in the same spirit as in Section 8.4, namely,

$$\text{pen}(\beta_\gamma) := \sum_{j=1}^{p} \Big(\|\beta_j\|_T + \mu\|B_j\gamma_j\|_2 \Big)$$

$$:= \|A\gamma\|_{2,1} + \mu\|B\gamma\|_{2,1},$$

where, for each j, B_j is some given smoothness matrix.

The smoothed Lasso for the time-varying coefficients model is

$$\hat{\gamma} := \arg\min_\gamma \left\{ \frac{1}{nT} \sum_{t=1}^{T} \|\mathbf{Y}(t) - \mathbf{X}(t)\beta_\gamma(t)\|_2^2 + \lambda\|A\gamma\|_{2,1} + \lambda\mu\|B\gamma\|_{2,1} \right\}.$$

We let $\hat{\beta}(\cdot) = \beta_{\hat{\gamma}}(\cdot)$.

8.5.2 The empirical process

The empirical process is

$$v_n(\beta) := \frac{2}{nT} \sum_{t=1}^{T} \varepsilon^T(t)\mathbf{X}(t)\beta(t) = \frac{2}{\sqrt{n}} \sum_{j=1}^{p} (W_j, \beta_j)_T,$$

where, for $j = 1, \ldots, p$, $W_{j,t} := \varepsilon^T(t)\mathbf{X}^{(j)}(t)/\sqrt{n}$, $t = 1, \ldots, T$, and where W_j is the vector $W_j = (W_{j,1}, \ldots, W_{j,T})^T$. Moreover, we use the inner product notation

$$(W_j, \beta_j)_T = \frac{1}{T} \sum_{t=1}^T W_{j,t} \beta_j(t).$$

Recall that for each j, the $\{W_{j,t}\}_{t=1}^T$ are independent, and $\mathcal{N}(0, \hat{\sigma}_{j,t}^2)$-distributed, with variance $\hat{\sigma}_{j,t}^2 = (\mathbf{X}^{(j)}(t))^T (\mathbf{X}^{(j)}(t))/n$. These variables now formally play the role of "noise" variables. Consider the set

$$\mathcal{T} := \{2|(W_j, \beta_{j,\gamma_j})_T|/\sqrt{n} \le \lambda_0 \|A_j \gamma_j\|_2 + \lambda_0 \mu_0 \|B_j \gamma_j\|_2, \ \forall \ \gamma_j, \ \forall \ j\}.$$

When the co-variables $X_i^j(t)$ are properly normalized, say $\hat{\sigma}_{j,t}^2 = 1$ for all j and t, and if, say, the $\beta_j(\cdot)$ are in the Sobolev space of m times continuously differentiable functions, then it can be shown that $\mathcal{T}$ has large probability when both λ_0 and μ_0 are chosen of order $(\log p/(nT))^{\frac{m}{2m+1}}$, for $T \to \infty$, and n possibly remaining finite. This is in analogy to the additive model, with $W_{j,t}$ playing the role of ε_i.

8.5.3 The compatibility condition for the time-varying coefficients model

Remember our definitions of the empirical norm and the time-dependent norm respectively,

$$\|f\|_n := \sqrt{\frac{1}{n} \sum_{i=1}^n f^2(X_i)}, \ \|g\|_T := \sqrt{\frac{1}{T} \sum_{t=1}^T g^2(t)}.$$

(There is some ambiguity in the notation, which we believe to be acceptable). For a function $h(\cdot, \cdot)$ depending on both co-variables $X_i(t) \in \mathcal{X}$ as well as time $t \in \mathbb{R}$, we define the norm

$$\|h\|_{n,T} := \sqrt{\frac{1}{nT} \sum_{t=1}^T \sum_{i=1}^n h^2(X_i(t), t)}.$$

The compatibility condition is again a way to deal with linear dependencies when considering more than nT elements in an nT-dimensional space. We therefore first introduce some approximating norm on the space of functions $h : \mathcal{X} \times \mathbb{R} \to \mathbb{R}$, which we denote by $\|\cdot\|$.

Arriving now at a switch of norms, it is convenient to use the function notation: for $x = (x^{(1)}, \ldots, x^{(p)}) \in \mathcal{X}$,

$$f_{\beta(t)}(x) := x\beta(t).$$

We write short hand

$$f_\beta := f_{\beta(\cdot)}(\cdot).$$

Thus,

$$\|f_\beta\|_{n,T}^2 = \frac{1}{nT} \sum_{t=1}^{T} \sum_{i=1}^{n} \left(X_i(t)\beta(t) \right)^2 = \sum_{t=1}^{T} \|\mathbf{X}(t)\beta(t)\|_2^2/(nT).$$

As in the previous section, we further denote the two terms in the penalty as

$$\text{pen}_1(\beta_\gamma) := \|A\gamma\|_{2,1}, \ \text{pen}_2(\beta_\gamma) := \mu\|B\gamma\|_{2,1}.$$

The approximation condition is now as in Subsection 8.4.3.

Definition *The* approximation condition holds for the time-varying coefficients model, *if there exists an $\eta \geq 0$ such that for all $\beta = (\beta_1, \ldots, \beta_p)^T$, we have*

$$\frac{\left| \|f_\beta\|_{n,T} - \|f_\beta\| \right|}{\sum_j \|\beta_j\| + \text{pen}_2(\beta)} \leq \eta.$$

Definition *The* compatibility condition for the time-varying coefficients model holds for the set S, with constant $\phi(S) > 0$, if for all β with

$$\sum_{j \notin S} \|\beta_j\| + \text{pen}_2(\beta)/3 \leq 5 \sum_{j \in S} \|\beta_j\|, \tag{8.11}$$

one has the inequality

$$\sum_{j \in S} \|\beta_j\|^2 \leq \|f_\beta\|^2/\phi^2(S).$$

8.5.4 A sparsity oracle inequality for the time-varying coefficients model

Definition of the oracle *Let $\mathscr{S}$ be a collection of sets for which the compatibility condition for the time-varying coefficients model holds. The oracle is*

$$\beta^* = \beta_{\gamma^*} := \arg\min\left\{ \frac{16\lambda^2|S_\beta|}{\phi^2(S_\beta)} + \|f_\beta - f^0\|_{n,T}^2 + 2\text{pen}_2(\beta) : \right.$$

$$\left. S_\beta \in \mathscr{S}, \ \sqrt{|S_\beta|}\eta/\phi(S_\beta) \leq 1/16 \right\}.$$

Moreover, we let $\phi_* := \phi(S_*)$, $s_* := |S_*|$, $f^* := f_{\beta^*}$, and $\hat{f} := f_{\hat{\beta}}$.

Let us recall the set

$$\mathscr{T} := \{2|(W_j, \beta_{j,\gamma_j})_T|/(\sqrt{n}) \le \lambda_0 \|A_j \gamma_j\|_2 + \lambda_0 \mu_0 \|B_j \gamma_j\|_2, \ \forall \ \gamma_j, \ \forall \ j\}.$$

Theorem 8.3. *Consider the smoothed Lasso for the time-varying coefficients model*

$$\hat{\beta} = \beta_{\hat{\gamma}} := \arg\min_{\beta_\gamma} \left\{ \frac{1}{nT} \sum_{t=1}^{T} \|\mathbf{Y}(t) - \mathbf{X}(t)\beta_\gamma(t)\|_2^2 + \lambda \|A\gamma\|_{2,1} + \lambda\mu \|B\gamma\|_{2,1} \right\}.$$

Suppose that $\lambda \ge 4\lambda_0$ *and* $\lambda\mu \ge 4\lambda_0\mu_0$. *Then on* $\mathscr{T}$, *we have*

$$\|\hat{f} - f^0\|_{n,T}^2 + \lambda \operatorname{pen}(\hat{\beta} - \beta^*)/2$$

$$\le 3 \left\{ 16\lambda^2 s_* / \phi_*^2 + \|f^* - f^0\|_2^2 + 2\lambda \operatorname{pen}_2(\beta^*) \right\}.$$

Asymptotics. As a consequence, say for the case where $\lambda \asymp (\log p/(nT))^{\frac{m}{2m+1}}$ and $\mu \asymp (\log p/nT)^{\frac{m}{2m+1}}$ for $T \to \infty$, and assuming $\mathbf{P}(\mathscr{T}) \to 1$, we get

$$\|\hat{f} - f^0\|_{n,T}^2$$

$$= O_{\mathrm{P}} \left(\left(\frac{\log p}{nT} \right)^{\frac{2m}{2m+1}} \frac{s_*}{\phi_*^2} + \|f^* - f_0\|_{n,T}^2 + \left(\frac{\log p}{nT} \right)^{\frac{2m}{2m+1}} \sum_{j \in S_*} \|B_j \beta_j^*\|_2 \right).$$

One can clearly see the gain of exploiting the smoothness, over performing a Lasso on each of the T models separately. For a somewhat more detailed view on this matter, let us take $\mathscr{S} = \{S_0\}$, where S_0 is the active set of the truth $f^0(x(t), t) = \sum_{j \in S_0} x^{(j)}(t)\beta_j^0(t)$, where $\beta^0 = \beta_{\gamma^0}$. Let $s_0 := |S_0|$ and $\phi_{0,t}$ be the restricted eigenvalue for each single linear model. Define $\phi_0 = \min_t \phi_{0,t}$. With the smoothed group Lasso for the time varying-coefficients model, assuming the smoothness $\|B_j \gamma_j^0\|_2 \le 1$, we get a prediction error with order of magnitude

$$O_{\mathrm{P}} \left(\left(\frac{\log p}{nT} \right)^{\frac{2m}{2m+1}} \frac{s_0}{\phi_0^2} \right),$$

For the individual Lasso's, we have in the worst case (the case where $\phi_{0,t}$ is the smallest), a prediction error of order of magnitude

$$O_{\mathrm{P}} \left(\left(\frac{\log p}{n} \right) \frac{s_0}{\phi_0^2} \right).$$

Hence, when $T \gg (n/\log p)^{\frac{1}{2m}}$, the smoothed group Lasso gives, according to our bounds, better performance than individual Lasso's. It means that the gain in performance is higher when m, the amount of smoothness, is larger.

Proof of Theorem 8.3. The Basic Inequality is

$$\|\hat{f} - f^0\|_{n,T}^2 + \lambda \operatorname{pen}(\hat{\beta})$$

$$\le v_n(\hat{\beta} - \beta^*) + \|f^* - f^0\|_{n,T}^2 + \lambda \operatorname{pen}(\beta^*).$$

Hence, similar to (8.6), on the set $\mathscr{T}$, and defining $\beta_{j,S}(\cdot) = \beta_j(\cdot)\mathbb{1}\{j \in S\} = b^{(j)}(\cdot)\gamma_{j,S}$,

$$4\|\hat{f} - f^0\|_{n,T}^2 + 3\lambda \operatorname{pen}(\hat{\beta}_{S_*^c}) + 3\lambda \operatorname{pen}_2(\hat{\beta}_{S_*} - \beta_{S_*}^*)$$

$$\le 5\lambda \operatorname{pen}_1(\hat{\beta}_{S_*} - \beta_{S_*}^*) + 4\|f^* - f^0\|_{nT}^2 + 8\lambda \operatorname{pen}_2(\beta^*).$$

Thus, we can follow the line of reasoning as in the proof of Theorem 8.2.

$\square$

8.6 Multivariate linear model and multitask learning

The results of this section are mainly from Lounici et al. (2009).

8.6.1 The loss function and penalty

The model for multitask learning is

$$Y_{i,t} = \sum_{j=1}^{p} X_{i,t}^{(j)} \beta_{j,t}^0 + \varepsilon_{i,t}, \, i = 1, \ldots, n, \, t = 1, \ldots, T.$$

When, for all i and j, the covariate $X_{i,t}^{(j)} := X_i^{(j)}$ does not depend on t, this is the multivariate linear regression model. See also Section 5.9.

The "truth" is denoted by

$$\mathbf{f}_t^0 := \mathbb{E}\mathbf{Y}_t,$$

where $\mathbf{Y}_t := (Y_{1,t}, \ldots, Y_{n,t})^T$.

The squared error loss is averaged over the T regressions, giving

$$L_n(\beta) := \frac{1}{nT} \sum_{t=1}^{T} \|\mathbf{Y}_t - \mathbf{X}_t \beta(t)\|_2^2,$$

with $\beta(t) := (\beta_{1,t}, \ldots, \beta_{p,t})^T, t = 1, \ldots, T$. The penalty is

$$\operatorname{pen}(\beta) := \sum_{j=1}^{p} \|\beta_j\|_2/\sqrt{T} := \|\beta\|_{2,1}/\sqrt{T}.$$

Thus, the idea is that the regressions share the non-zero coefficients, i.e. (for all j), either $\beta^0_{j,t} = 0$ for all t, or $\beta^0_{j,t} \neq 0$ for all t. The Lasso for multitask learning is

$$\hat{\beta} = \arg\min_{\beta}\left\{\frac{1}{nT}\sum_{t=1}^{T}\|\mathbf{Y}_t - \mathbf{X}_t\beta(t)\|_2^2 + \lambda\|\beta\|_{2,1}/\sqrt{T}\right\}.$$

One easily sees that the situation is similar to the one of the previous section, except that the penalty does not have a smoothness part.

8.6.2 The empirical process

The empirical process is

$$v_n(\beta) := \frac{2}{nT}\sum_{t=1}^{T}\varepsilon_t^T\mathbf{X}_t\beta(t) = \frac{2}{T\sqrt{n}}\sum_{j=1}^{p}W_j^T\beta_j,$$

where $W_j := (W_{j,1},\ldots,W_{j,T})^T$ and $W_{j,t} := \varepsilon_t^T\mathbf{X}_t^{(j)}/\sqrt{n}$.

Lemma 8.5. *Suppose that $\hat{\sigma}_{j,t}^2 = 1$ for all j and t. Let, for some $x > 0$,*

$$\lambda_0^2 := \frac{4}{n}\left(1 + \sqrt{\frac{4x + 4\log p}{T}} + \frac{4x + 4\log p}{T}\right).$$

Then with probability at least $1 - e^{-x}$, it holds that simultaneously for all β,

$$|v_n(\beta)| \leq \lambda_0\|\beta\|_{2,1}/\sqrt{T}.$$

Proof of Lemma 8.5. We have

$$|\sum_{j=1}^{p}W_j^T\beta_j| \leq \sum_{j=1}^{p}\|W_j\|_2\|\beta_j\|_2$$

$$\leq \max_{1\leq j\leq p}\|W_j\|_2\|\beta\|_{2,1}.$$

The random variables $\|W_j\|_2^2$, $j = 1,\ldots,p$, all have a chi-squared distribution with T degrees of freedom. Hence, by Lemma 8.1,

$$\mathbf{P}\left(\max_{1\leq j\leq p}4\|W_j\|^2 \geq nT\lambda_0^2\right) \leq e^{-x}.$$

$\square$

8.6.3 The multitask compatibility condition

We let $(\mathbf{X}\beta)^T := ((\mathbf{X}_1\beta(1))^T, \ldots, (\mathbf{X}_T\beta(T))^T)$, and we write

$$\|\mathbf{X}\beta\|_{n,T}^2 := \frac{1}{nT}\sum_{t=1}^{T}\|\mathbf{X}_t\beta(t)\|_2^2.$$

Definition *We say that the* multitask compatibility condition *holds for the set S if for some $\phi(S) > 0$ and for all β with $\|\beta_{S^c}\|_{2,1} \leq 3\|\beta_S\|_{2,1}$, one has the inequality*

$$\|\beta_S\|_{2,1}^2 \leq \frac{\|\mathbf{X}\beta\|_{n,T}^2 Ts}{\phi^2(S)}.$$

We can compare the multitask compatibility condition with the adaptive version of the restricted eigenvalue condition we would use if we consider the T univariate regressions as one single regression with nT observations and design matrix

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 & 0 & \ldots & 0 \\ 0 & \mathbf{X}_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \mathbf{X}_T \end{pmatrix},$$

an $(nT \times pT)$-matrix. This model in matrix notation is

$$\mathbf{Y} = \mathbf{X}\beta^0 + \varepsilon,$$

where now

$$\mathbf{Y} := \begin{pmatrix} \mathbf{Y}_1 \\ \vdots \\ \mathbf{Y}_T \end{pmatrix}, \ \beta^0 := \begin{pmatrix} \beta^0(1) \\ \vdots \\ \beta^0(T) \end{pmatrix}, \ \varepsilon := \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_T \end{pmatrix}.$$

The univariate restricted eigenvalue condition is then as in Section 6.13.2:

Definition *The* univariate adaptive restricted eigenvalue condition *holds for the set $S \subset \{1, \ldots, p\}$, if for some constant $\phi_{\mathrm{adap}}(3, S_{\mathrm{full}}, Ts) > 0$, and for all β, with $\sum_{j \notin S}\|\beta_j\|_1 \leq 3\sqrt{Ts}\sqrt{\sum_{j \in S}\|\beta_j\|_2^2}$, one has*

$$\sum_{j \in S}\|\beta_j\|_2^2 \leq \left(\|\mathbf{X}\beta\|_2^2/n\right)/\phi_{\mathrm{adap}}^2(3, S_{\mathrm{full}}, Ts).$$

The next lemma is the counterpart of Lemma 8.2.

Lemma 8.6. *Suppose the univariate adaptive restricted eigenvalue condition is met for S, with constant $\phi_{\text{adap}}^2(3, S_{\text{full}}, Ts)$. Then the multitask compatibility condition holds, with constant $\phi(S) \geq \phi_{\text{adap}}(3, S_{\text{full}}, Ts)$.*

Proof of Lemma 8.6. First note that

$$\|\beta_S\|_{2,1} = \sum_{j \in S} \|\beta_j\|_2 \leq \sqrt{s} \sqrt{\sum_{j \in S} \|\beta_j\|_2^2}.$$

Suppose now that β satisfies

$$\|\beta_{S^c}\|_{2,1} \leq 3\|\beta_S\|_{2,1}.$$

Then

$$\sum_{j \notin S} \|\beta_j\|_1 \leq \sqrt{T} \sum_{j \notin S} \|\beta_j\|_2 = \sqrt{T}\|\beta_{S^c}\|_{2,1}$$

$$\leq 3\sqrt{T}\|\beta_S\|_{2,1} \leq 3\sqrt{Ts} \sqrt{\sum_{j \in S} \|\beta_j\|_2^2}.$$

Hence, by the univariate adaptive restricted eigenvalue condition,

$$\sum_{j \in S} \|\beta_j\|_2^2 \leq \left(\|\mathbf{X}\beta\|_2^2/n \right) / \phi_{\text{adap}}^2(3, S_{\text{full}}, Ts).$$

But then also

$$\|\beta_S\|_{2,1}^2 \leq s \sum_{j \in S} \|\beta_j\|_2^2 \leq \left(\|\mathbf{X}\beta\|_2^2/n \right) s / \phi_{\text{adap}}^2(3, S_{\text{full}}, Ts)$$

$$= \|\mathbf{X}\beta\|_{n,T}^2 \, sT / \phi_{\text{adap}}^2(3, S_{\text{full}}, Ts).$$

$\square$

8.6.4 A multitask sparsity oracle inequality

Definition of the oracle *Assume the multitask compatibility condition for the sets S in $\mathscr{S}$. The oracle β^* is*

$$\beta^* = \arg\min_{\beta: \, S_\beta \in \mathscr{S}} \left\{ \|\mathbf{X}\beta - \mathbf{f}^0\|_{n,T}^2 + \frac{4\lambda^2 s_\beta}{\phi^2(S_\beta)} \right\}.$$

We define $\phi_* := \phi(S_{\beta^*})$, $S_* := S_{\beta^*}$, $s_* := |S_*|$.

Theorem 8.4. *Let*

$$\mathscr{T} := \{|v_n(\beta)| \le \lambda_0 \|\beta\|_{2,1}/\sqrt{T}, \ \forall \ \beta\}.$$

Consider the multitask Lasso

$$\hat{\beta} := \arg\min\left\{\frac{1}{nT}\sum_{t=1}^{T}\|\mathbf{Y}_t - \mathbf{X}_t\beta(t)\|_2^2 + \lambda\|\beta\|_{2,1}/\sqrt{T}\right\}.$$

Take $\lambda \ge 4\lambda_0$. Then on $\mathscr{T}$

$$\|\mathbf{X}\hat{\beta} - \mathbf{f}^0\|_{n,T}^2 + \lambda\|\hat{\beta} - \beta^*\|_{2,1}/\sqrt{T} \le 6\|\mathbf{X}\beta^* - \mathbf{f}^0\|_{n,T}^2 + \frac{24\lambda^2 s_*}{\phi_*^2}.$$

Proof of Theorem 8.4. The Basic Inequality in this case is

$$\|\mathbf{X}\hat{\beta} - \mathbf{f}^0\|_{n,T}^2 + \lambda\|\hat{\beta}\|_{2,1}/\sqrt{T} \le v_n(\hat{\beta} - \beta^*) + \|\mathbf{X}\beta^* - \mathbf{f}^0\|_{n,T}^2 + \lambda\|\beta^*\|_{2,1}/\sqrt{T},$$

so on $\mathscr{T}$,

$$4\|\mathbf{X}\hat{\beta} - \mathbf{f}^0\|_{n,T}^2 + 4\lambda\|\hat{\beta}\|_{2,1}/\sqrt{T}$$

$$\le \lambda\|\hat{\beta} - \beta^*\|_{2,1}/\sqrt{T} + 4\|\mathbf{X}\beta^* - \mathbf{f}^0\|_{n,T}^2 + 4\lambda\|\beta^*\|_{2,1}/\sqrt{T},$$

and we can proceed as in the proof of Theorem 6.2. □

Asymptotics Suppose that $\log p/T = O(1)$ for $T \to \infty$. Then, with

$$\lambda_0^2 = \frac{4}{n}\left(1 + \sqrt{\frac{4x + 4\log p}{T}} + \frac{4x + 4\log p}{T}\right),$$

as given in Lemma 8.5, we see that we can take $\lambda = O(1/\sqrt{n})$, for $n \to \infty$. The prediction error is then of order

$$O_{\mathrm{P}}\left(\|\mathbf{X}\beta^* - \mathbf{f}^0\|_{n,T}^2 + \frac{s_*}{n\phi_*^2}\right).$$

In other words, up to a gain of a $(\log p)$-factor, and modulo compatibility, the prediction error for the multivariate model is about of the same order as the average prediction error for T single Lasso's.

To handle the set $\mathscr{T}$, one may insert Lemma 8.5, which relies on the assumption of normally distributed errors. An alternative route is to apply Lemma 8.7 below. We conclude that if the co-variables are bounded (say), it suffices to assume only bounded fourth moments for the errors,

Lemma 8.7. *Let $\{\varepsilon_{i,t} : i = 1,\ldots,n, t = 1,\ldots,T\}$ be independent random variables with $\mathbb{E}\varepsilon_{i,t} = 0$, $\mathbb{E}\varepsilon_{i,t}^2 = 1$, and $\mathbb{E}\varepsilon_{i,t}^4 \le \mu_4^4$ for all i and t. Moreover, let $\{\mathbf{X}_t^{(j)} : j = 1,\ldots,p, t = 1,\ldots,T\}$ be given n-vectors satisfying $(\mathbf{X}_t^{(j)})^T(\mathbf{X}_t^{(j)}) = n$ for all j and t. Define*

$$W_{j,t} := \varepsilon_t^T \mathbf{X}_t^{(j)}/\sqrt{n}, \ j = 1,\ldots,p, \ t = 1,\ldots,T,$$

and let $W_j := (W_{j,1}, \ldots, W_{j,t})^T$, $j = 1, \ldots, p$. Then for $p \geq e^3$, we have

$$\mathbb{E} \max_{1 \leq j \leq p} \|W_j\|_2^4$$

$$\leq \left\{ T + \sqrt{T} \left[8\log(2p) \right]^{3/2} \mu_4^2 \left[\frac{1}{nT} \sum_{t=1}^{T} \sum_{i=1}^{n} \max_{1 \leq j \leq p} (X_{i,t}^{(j)})^4 \right]^{1/2} \right\}^2.$$

The proof of Lemma 8.7 is given in Chapter 14 , Section 14.10.2.

8.7 The approximation condition for the smoothed group Lasso

Let us first recall the approximation condition for the additive model given in Section 8.4.3:

Definition *The* approximation condition *holds for the additive model if there exists an $\eta \geq 0$ such that for all β, we have*

$$\frac{\left| \|f_\beta\|_n - \|f_\beta\| \right|}{\sum_j \|f_{j,\beta_j}\| + \mathrm{pen}_2(\beta)} \leq \eta.$$

We will present illustrations of this condition for the cases considered, that is, for the case of Sobolev smoothness and for diagonalized smoothness. In the first case, we assume that the $\{X_i^{(j)}\}_{i=1}^n$ is a random sample from some distribution Q, and we show that the approximation condition is met with large probability when $\lambda = \mu \asymp n^{-\frac{m}{2m+1}}$, with $\eta \asymp \lambda^{\frac{2m-1}{4m}}$. For the case of diagonalized smoothness, we formulate the result directly in terms of an approximation condition on the Gram matrix.

8.7.1 Sobolev smoothness

We will present the result from Meier et al. (2009).

Let $\{X_i = (X_i^{(1)}, \ldots, X_i^{(p)})\}_{i=1}^n$ be i.i.d. copies of $X = (X^{(1)}, \ldots, X^{(p)}) \in [0,1]^p$. The distribution of X is denoted by Q, and we let $\| \cdot \|$ be the $L_2(Q)$-norm. The marginal distribution of $X^{(j)}$ is denoted by Q_j, $j = 1, \ldots, p$.

Let ν be Lebesgue measure on $[0,1]$, and define for $f_j : [0,1] \to \mathbb{R}$,

$$I^2(f_j) = \int |f_j^{(m)}|^2 d\nu = \|f_j^{(m)}\|_\nu^2,$$

where $\|\cdot\|_{\nu}$ denotes the $L_2(\nu)$-norm.

Theorem 8.5. *Assume that for all j, the densities $dQ_j/d\nu = q_j$ exist and that for some constant $\eta_0 > 0$,*

$$q_j \geq \eta_0^2.$$

Then there exists a constant A_m depending only on m, and some constant $C_{m,q}$ depending only on m and the lower bound η_0 for the marginal densities $\{q_j\}$, such that for $\lambda = \mu \geq A_m (\log p/n)^{\frac{m}{2m+1}}$, we have for all $t > 0$,

$$\mathbf{P}\left(\sup_f \frac{\left| \|f\|_n^2 - \|f\|^2 \right|}{\left(\sum_{j=1}^p \sqrt{\|f_j\|^2 + \mu^2 I^2(f_j)} \right)^2} \leq K_{m,q}(t) \lambda^{\frac{2m-1}{2m}} \right) \geq 1 - \exp(-n\lambda^2 t),$$

where

$$K_{m,q}(t) := C_{m,q}(1 + \sqrt{2t} + \lambda^{\frac{2m-1}{2m}} t),$$

Proof. This is shown in Meier et al. (2009), after translating the result in our notation. They use the parameters α and γ, which are

$$\alpha := 1 - \frac{1}{2m}.$$

and

$$\gamma := \frac{2(1-\alpha)}{2-\alpha}.$$

Moreover, they replace what we call λ by $\lambda^{\frac{2m}{2m+1}}$ (i.e., their λ is required to be suitably larger than $\sqrt{\log p/n}$). $\qquad\square$

The proof in Meier et al. (2009) is quite involved. We show in the next subsection that for diagonalized smoothness, one gets qualitatively the same result, but with a very simple proof.

8.7.2 Diagonalized smoothness

Recall that

$$\hat{\Sigma} := \mathbf{X}^T \mathbf{X}/n.$$

We will approximate $\hat{\Sigma}$ by a matrix Σ, which potentially is non-singular. For example, when the rows of $\mathbf{X}$ are normalized versions of n i.i.d. random vectors, the matrix Σ could be the population variant of $\hat{\Sigma}$. Write

$$\|\hat{\Sigma} - \Sigma\|_\infty := \max_{j,k} |\hat{\Sigma}_{j,k} - \Sigma_{j,k}|.$$

Take $D := \mathrm{diag}(d_1, d_2, \dots)$ and $d_t := t^m$, with $m > 1/2$ given.

Lemma 8.8. *Take* $T_0 := \lfloor n^{\frac{1}{2m+1}} \rfloor + 1$, *and* $\lambda \geq \sqrt{T_0/n}$, $\lambda\mu \geq T_0/n$. *Then for all* β

$$|\beta^T \hat{\Sigma} \beta - \beta^T \Sigma \beta| \leq n\|\hat{\Sigma} - \Sigma\|_\infty \lambda^2 \left(\sum_{j=1}^p \left(\|\beta_j\|_2 + \mu \|D\beta_j\|_2 \right) \right)^2.$$

Proof of Lemma 8.8. It holds that

$$|\beta^T \hat{\Sigma} \beta - \beta^T \Sigma \beta| \leq \|\hat{\Sigma} - \Sigma\|_\infty \|\beta\|_1^2,$$

and

$$\|\beta_j\|_1 \leq \sqrt{T_0}\|\beta_j\|_2 + T_0\|D\beta_j\|_2/\sqrt{n},$$

Hence

$$\|\beta\|_1 = \sum_{j=1}^p \|\beta_j\|_1 \leq \sum_{j=1}^p \left\{ \sqrt{T_0}\|\beta_j\|_2 + T_0/\sqrt{n}\|D\beta_j\|_2 \right\}.$$

$\square$

Problems

8.1. Consider the model
$$\mathbf{Y} = \mathbf{X}\beta^0 + \varepsilon,$$
where $\mathbf{X}^T\mathbf{X}/n = I$, and the estimator
$$\hat{\beta} := \arg\min_\beta \left\{ \|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + 2\lambda\|\beta\|_2 \right\}.$$

Let $\mathbf{Z} := \mathbf{X}^T\mathbf{Y}/n$.
(a) Show that
$$\|\hat{\beta}\|_2 = \begin{cases} \|\mathbf{Z}\|_2 - \lambda, & |\mathbf{Z}\|_2 > \lambda \\ 0, & \|\mathbf{Z}\|_2 \leq \lambda \end{cases}.$$
(b) Show that when $\|\hat{\beta}\|_2 \neq 0$,
$$\hat{\beta} = \frac{\mathbf{Z}}{1 + \lambda/\|\hat{\beta}\|_2}.$$

8.2. Consider the model
$$\mathbf{Y} = \mathbf{X}\beta^0 + \varepsilon,$$
where $\mathbf{X}^T\mathbf{X}/n = I$. Write $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$, where $\mathbf{X}_k$ is an $n \times p_k$-matrix, $k = 1, 2$. We use the estimator

$$\hat{\beta} := \arg \min_{\beta=(\beta_1,\beta_2)^T} \left\{ \|\mathbf{Y} - (\mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2)\|_2^2/n + 2\sqrt{\lambda_1^2\|\beta_1\|_2^2 + \lambda_2^2\|\beta_2\|_2^2} \right\}.$$

Define

$$x := \sqrt{\lambda_1^2\|\hat{\beta}_1\|_2^2 + \lambda_2^2\|\hat{\beta}_2\|_2^2}.$$

Show that when $x \neq 0$,

$$\hat{\beta}_1 = \frac{\mathbf{Z}_1}{1+\lambda_1^2/x}, \ \hat{\beta}_2 = \frac{\mathbf{Z}_2}{1+\lambda_2^2/x},$$

where $\mathbf{Z}_k = \mathbf{X}_k^T\mathbf{Y}/n$, $k = 1, 2$. Furthermore, then x is a solution of

$$(x+\lambda_1^2)^2(x+\lambda_2^2)^2 = \lambda_1^2\|\mathbf{Z}_1\|_2^2(x+\lambda_2^2)^2 + \lambda_2^2\|\mathbf{Z}_2\|_2^2(x+\lambda_1^2)^2.$$

8.3. Here is an example that shows that two separate square roots can yield better prediction error than a single square root. Consider two observations Y_1 and Y_2, where $Y_1 = \beta_1^0 + \varepsilon_1$, $\beta_1^0 = 1$, is signal, and $Y_2 = \varepsilon_2$ is only noise. We assume, for $i = 1, 2$, that $|Y_i| \leq \sigma_i$, and let $\lambda_i = \sigma_i$ be the tuning parameter. Moreover we assume $\sigma_2^2 = \sigma_1$ and σ_1 somewhat smaller than $1/4$:

$$\sigma_1^2 + 3\sigma_1 < 1/16.$$

Taking two separate square roots. Suppose we let $\hat{\beta}$ be the (weighted) Lasso

$$\hat{\beta} := \arg \min_{\beta=(\beta_1,\beta_2)^T} \left\{ (Y_1-\beta_1)^2 + (Y_2-\beta_2)^2 + 2\lambda_1|\beta_1| + 2\lambda_2|\beta_2| \right\}.$$

(a) Show that $\hat{\beta}_1 = Y_1 - \lambda_1$ and $\hat{\beta}_2 = 0$. The prediction error is thus

$$\|\hat{\beta} - \beta^0\|_2^2 = (\varepsilon_1 - \lambda_1)^2 \leq 4\sigma_1^2. \tag{8.12}$$

Taking a single square root. Consider now the group-type Lasso

$$\hat{\beta} := \arg\min_{\beta} \left\{ (Y_1-\beta_1)^2 + (Y_2-\beta_2)^2 + 2\sqrt{\lambda_1^2\beta_1^2 + \lambda_2^2\beta_2^2} \right\}.$$

(b) Verify that with $\beta_1 < 1/2$ the penalized loss is at least

$$(1/2 - \sigma_1)^2 \geq \frac{1}{16},$$

whereas with $\beta_1 = 1$ and $\beta_2 = 0$ the penalized loss is at most

$$\sigma_1^2 + \sigma_2^2 + 2\sigma_1 < \frac{1}{16}.$$

Conclude that $\hat{\beta}_1 \geq 1/2$. Therefore also $x := \sqrt{\lambda_1^2 \hat{\beta}_1^2 + \lambda_2^2 \hat{\beta}_2^2} \neq 0$, in fact

$$x^2 \geq \sigma_1^2/4.$$

(c) Derive that

$$\hat{\beta}_1 = \frac{Y_1}{1 + \lambda_1^2/x},$$

and

$$\hat{\beta}_2 = \frac{Y_2}{1 + \lambda_2^2/x}$$

(see also Problem 8.2).

(d) Show that for $Y_2 > \sigma_2/2$, the prediction error is at least

$$\frac{\varepsilon_2^2}{\left(1 + 2\sigma_2^2/\sigma_1\right)^2} \geq \frac{\sigma_2^2}{\left(2 + 4\sigma_2^2/\sigma_1\right)^2} \geq \frac{\sigma_1}{36}. \tag{8.13}$$

Conclusion in this example. Comparing (8.12) with (8.13), we see that when σ_1 is small ($\sigma_1 \ll 144$), the separate square roots penalty gives a much better prediction error than the single square root penalty.

8.4. Consider the model

$$Y_i = f^0(X_i) + \varepsilon_i, \ 1 = 1, \ldots, n.$$

Let $I(f)$ be some measure for the roughness of the function f. Assume that I is a semi-norm and suppose that for a given $0 < \alpha < 1$,

$$|(\varepsilon, f)_n| \leq C \|f\|_n^\alpha I^\alpha(f)/\sqrt{n}, \ \forall f.$$

The penalty used in Section 8.4 is based on the inequalities

$$a^\alpha b^{1-\alpha} \leq \sqrt{a^2 + b^2} \leq a + b,$$

which holds for all positive a and b. An alternative penalty is motived by the inequality

$$a^\alpha b^{1-\alpha} \leq a^2 + b^\gamma,$$

where

$$\gamma := \frac{2(1 - \alpha)}{2 - \alpha}.$$

This leads to the estimator

$$\hat{f} := \arg\min_f \left\{ \frac{1}{n} \sum_{i=1}^n (Y_i - f(X_i))^2 + \lambda^2 I^\gamma(f) \right\}.$$

Prove an oracle inequality for this estimator, where one takes $\lambda > Cn^{-\frac{2-\gamma}{2}}$ for a constant C depending on γ.

8.5. Consider the estimator

$$\hat{\beta} := \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - f_\beta(X_i))^2 + \lambda^2 \|\beta\|_2^\gamma \right\},$$

where $0 < \gamma < 2$ is given. Show that this problem can be numerically handled by solving for each $\mu > 0$,

$$\hat{\beta}(\mu) := \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - f_\beta(X_i))^2 + \mu^2 \|\beta\|_2^2 \right\},$$

and then solving for μ

$$\min_{\mu} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - f_{\hat{\beta}(\mu)}(X_i))^2 + c_1 \mu^2 \|\hat{\beta}(\mu)\|_2^2 + c_2 \mu^{-\frac{2(2-\gamma)}{\gamma}} \right\},$$

where c_1 and c_2 are appropriately chosen positive constants depending only on λ and γ.

8.6. In Theorem 8.1, we derived a bound for the ℓ_1/ℓ_2-estimation error

$$\|\hat{\beta} - \beta^*\|_{2,1}$$

of the group Lasso. Deduce a screening property from this (under beta-min conditions). Formulate group Lasso variants of the (minimal adaptive) restricted eigenvalue conditions that ensure appropriate bounds for the ℓ_2-error. Similarly for the Lasso for the time-varying coefficients model and for the Lasso for multitask learning. See Lounici et al. (2010).

Chapter 9

Non-convex loss functions and ℓ_1-regularization

Abstract Much of the theory and computational algorithms for ℓ_1-penalized methods in the high-dimensional context has been developed for convex loss functions, e.g., the squared error loss for linear models (Chapters 2 and 6) or the negative log-likelihood in a generalized linear model (Chapters 3 and 6). However, there are many models where the negative log-likelihood is a non-convex function. Important examples include mixture models or linear mixed effects models which we describe in more details. Both of them address in a different way the issue of modeling a grouping structure among the observations, a quite common feature in complex situations. We discuss in this chapter how to deal with non-convex but smooth ℓ_1-penalized likelihood problems. Regarding computation, we can typically find a local optimum of the corresponding non-convex optimization problem only whereas the theory is given for the estimator defined by a global optimum. Particularly in high-dimensional problems, it is difficult to compute a global optimum and it would be desirable to have some theoretical properties of estimators arising from "reasonable" local optima. However, this is largely an unanswered problem.

9.1 Organization of the chapter

The chapter is built upon describing two models. Section 9.2 discusses finite mixtures of regressions models, which builds on results from Städler et al. (2010), and Section 9.3 focuses on linear mixed effects models, based on work from Schelldorfer et al. (2011). Within Section 9.2 (about mixture models), we discuss methodology and computational issues while statistical theory is presented in Section 9.4.3. Section 9.3 (about linear mixed models) includes methodology and computational aspects and we outline some theory in Section 9.4.4. In Section 9.4, we present general mathematical theory for ℓ_1-penalization with smooth, non-convex negative log-likelihood functions which was developed by Städler et al. (2010). The framework encompasses mixture of regressions and linear mixed models as special cases.

9.2 Finite mixture of regressions model

Many applications deal with relating a response variable Y to a set of covariates $X^{(1)}, \ldots, X^{(p)}$ through a regression-type model. The homogeneity assumption that the regression coefficients are the same for all observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ can be inadequate. Parameters may change for different subgroups of observations. Such heterogeneity can be modeled with mixture models. Especially with high-dimensional data where the number of covariates p is much larger than sample size n, the homogeneity assumption seems rather restrictive: a fraction of covariates may exhibit a different influence on the response among subsets of observations, i.e., among different sub-populations. Hence, addressing the issue of heterogeneity in high-dimensional data is an important need in many practical applications. Besides methodology, computation and mathematical theory, we will illustrate on real data that substantial prediction improvements are possible by incorporating a heterogeneity structure to the model.

9.2.1 Finite mixture of Gaussian regressions model

We consider a continuous response Y and a p-dimensional covariate $X \in \mathcal{X} \subset \mathbb{R}^p$. Our primary focus is on the following mixture model involving Gaussian components:

$$
\begin{aligned}
&Y_i | X_i \text{ independent for } i = 1, \ldots, n, \\
&Y_i | X_i = x \sim h_\xi(y|x) dy \text{ for } i = 1, \ldots, n, \\
&h_\xi(y|x) = \sum_{r=1}^{k} \pi_r \frac{1}{\sqrt{2\pi}\sigma_r} \exp\left(-\frac{(y - x\beta_r)^2}{2\sigma_r^2} \right), \\
&\xi = (\beta_1, \ldots, \beta_k, \sigma_1, \ldots, \sigma_k, \pi_1, \ldots, \pi_{k-1}) \in \mathbb{R}^{kp} \times \mathbb{R}^k_{>0} \times \Pi, \\
&\Pi = \{\pi;\ \pi_r > 0 \text{ for } r = 1, \ldots, k-1 \text{ and } \sum_{r=1}^{k-1} \pi_r < 1\}.
\end{aligned}
\tag{9.1}
$$

Thereby, $X_i \in \mathcal{X} \subset \mathbb{R}^p$ are fixed or random covariates, ξ denotes the $(p+2) \cdot k - 1$ free parameters and $\pi_k = 1 - \sum_{r=1}^{k-1} \pi_r$. The model in (9.1) is a mixture of Gaussian regressions, where the rth component has its individual vector of regressions coefficients β_r and error variances σ_r^2. We sometimes denote it by FMR (Finite Mixture of Regressions) model.

9.2.1.1 Reparametrized mixture of regressions model

We will prefer to work with a reparametrized version of model (9.1) whose penalized maximum likelihood estimator is scale-invariant and easier to compute. The computational aspect will be discussed in greater detail in Sections 9.2.2.1 and 9.2.8. Define new parameters

$$\phi_r = \beta_r/\sigma_r, \ \ \rho_r = \sigma_r^{-1}, \ \ r = 1,\ldots,k.$$

This yields a one-to-one mapping from ξ in (9.1) to a new parameter vector $\theta = (\phi_1,\ldots,\phi_k,\rho_1,\ldots,\rho_k,\pi_1,\ldots,\pi_{k-1})$ and the model (9.1) in reparametrized form equals:

$$Y_i|X_i \text{ independent for } i = 1,\ldots,n,$$
$$Y_i|X_i = x \sim \mathrm{p}_\theta(y|x)dy \text{ for } i = 1,\ldots,n,$$
$$\mathrm{p}_\theta(y|x) = \sum_{r=1}^{k} \pi_r \frac{\rho_r}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(\rho_r y - x\phi_r)^2\right) \tag{9.2}$$
$$\theta = (\phi_1,\ldots,\phi_k,\rho_1,\ldots,\rho_k,\pi_1,\ldots,\pi_{k-1}) \in \mathbb{R}^{kp} \times \mathbb{R}_{>0}^{k} \times \Pi$$
$$\Pi = \{\pi; \ \pi_r > 0 \text{ for } r = 1,\ldots,k-1 \text{ and } \sum_{r=1}^{k-1} \pi_r < 1\}.$$

This is the main model we are analyzing and working with. We denote by θ^0 the true parameter.

The log-likelihood function in this model equals:

$$\ell(\theta;\mathbf{Y}) = \sum_{i=1}^{n} \log\left(\sum_{r=1}^{k} \pi_r \frac{\rho_r}{\sqrt{2\pi}} \exp(-\frac{1}{2}(\rho_r Y_i - X_i\phi_r)^2)\right). \tag{9.3}$$

Since we want to deal with the $p \gg n$ case, we have to regularize the maximum likelihood estimator (MLE) in order to obtain reasonably accurate estimates. We define below an ℓ_1-penalized MLE which is different from a naive ℓ_1-penalty for the MLE in the non-transformed model (9.1). Furthermore, it is well known that the (log-) likelihood function is generally unbounded. We will see in Section 9.2.2.2 that a suitable penalization will mitigate this problem.

9.2.2 ℓ_1-penalized maximum likelihood estimator

We first argue for the case of a (non-mixture) linear model why the reparametrization above in Section 9.2.1.1 is useful and quite natural.

9.2.2.1 ℓ_1-penalization for reparametrized linear models

Consider a Gaussian linear model

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon, \tag{9.4}$$

where $\varepsilon = \varepsilon_1, \ldots, \varepsilon_n$ i.i.d. $\sim \mathcal{N}(0, \sigma^2)$. The ℓ_1-penalized Lasso estimator is defined as

$$\hat{\beta}(\lambda) = \operatorname{argmin}_\beta \left(\|\mathbf{Y} - \mathbf{X}\beta\|_2^2 / n + \lambda \|\beta\|_1 \right).$$

The Gaussian assumption in model (9.4) is not crucial but it is useful to make connections to the likelihood framework. The Lasso estimator is equivalent to minimizing the penalized negative log-likelihood $n^{-1}\ell(\beta; Y_1, \ldots, Y_n)$ as a function of the regression coefficients β and using the ℓ_1-penalty $\|\beta\|_1 = \sum_{j=1}^p |\beta_j|$: equivalence means here that we obtain the same estimator for a potentially different tuning parameter. But the standard Lasso estimator above does not provide an estimate of the nuisance parameter σ^2.

In mixture models, it will be crucial to have a good estimator of σ^2 and the role of the scaling with the variance parameter is much more important than in homogeneous regression models. Hence, it is important to take σ^2 into the definition and optimization of the penalized maximum likelihood estimator. We could proceed with the following ℓ_1-regularized maximum likelihood estimator (see Section 3.2.1):

$$
\begin{aligned}
\hat{\beta}(\lambda), \hat{\sigma}^2(\lambda) &= \operatorname{argmin}_{\beta, \sigma^2} \left(-n^{-1}\ell(\beta, \sigma^2; Y_1, \ldots, Y_n) + \lambda \|\beta\|_1 \right) \\
&= \operatorname{argmin}_{\beta, \sigma^2} \left(\log(\sigma) + \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 / (2n\sigma^2) + \lambda \|\beta\|_1 \right). \tag{9.5}
\end{aligned}
$$

Note that we are penalizing only the β-parameter but the variance parameter σ^2 is influenced indirectly by the amount of shrinkage λ.

There is a severe drawback of the estimator in (9.5). The optimization in (9.5) is non-convex and hence, some of the major computational advantages of the Lasso for high-dimensional problems is lost. We address this issue by using the penalty term $\lambda \frac{\|\beta\|_1}{\sigma}$ leading to the following estimator:[1]

$$\hat{\beta}(\lambda), \hat{\sigma}^2(\lambda) = \operatorname{argmin}_{\beta, \sigma^2} \left(\log(\sigma) + \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 / (2n\sigma^2) + \lambda \frac{\|\beta\|_1}{\sigma} \right). \tag{9.6}$$

This estimator is equivariant under scaling of the response. More precisely, consider the transformation

---

[1] The penalty in (9.6) is a natural choice since the regularization parameter λ in the standard Lasso should be chosen as $const. \sigma \sqrt{\log(p)/n}$, see Lemma 6.2 which can be easily adapted to cover general error variance σ^2.

$$\mathbf{Y}' = b\mathbf{Y}; \ \beta' = b\beta; \ \sigma' = b\sigma \ (b > 0)$$

which leaves model (9.4) invariant. The estimator in (9.6) based on the transformed data $\{(X_i, Y_i'); \ i = 1, \ldots, n\}$ then yields $\hat{\beta}' = b\hat{\beta}$ where the latter estimate $\hat{\beta}$ is based on non-transformed data $\{(X_i, Y_i); \ i = 1, \ldots, n\}$. Furthermore, the estimator in (9.6) penalizes the ℓ^1-norm of the coefficients and small variances σ^2 simultaneously. Most importantly, we can reparametrize to achieve convexity of the optimization problem:

$$\phi_j = \beta_j / \sigma, \ \rho = \sigma^{-1}.$$

This then yields the following estimator which is invariant under scaling and whose computation involves convex optimization (Problem 9.1):

$$\hat{\phi}(\lambda), \hat{\rho}(\lambda) = \underset{\phi, \rho}{\arg\min} \left(-\log(\rho) + \frac{1}{2n} \|\rho\mathbf{Y} - \mathbf{X}\phi\|_2^2 + \lambda\|\phi\|_1 \right). \qquad (9.7)$$

The Karush-Kuhn-Tucker conditions (KKT) imply (Problem 9.2): every solution $(\hat{\phi}, \hat{\rho})$ of (9.7) satisfies,

$$| -\mathbf{X}_j^T(\rho\mathbf{Y} - \mathbf{X}\hat{\phi})/n| \leq \lambda \qquad \text{if} \qquad \hat{\phi}_j = 0 \ (j = 1, \ldots, p),$$
$$-\hat{\rho}\mathbf{X}_j^T(\mathbf{Y} - \mathbf{X}\hat{\phi}) + n\lambda \operatorname{sign}(\hat{\phi}_j) = 0 \qquad \text{if} \qquad \hat{\phi}_j \neq 0 \ (j = 1, \ldots, p),$$
$$\hat{\rho} = \frac{\mathbf{Y}^T\mathbf{X}\hat{\phi} + \sqrt{(\mathbf{Y}^T\mathbf{X}\hat{\phi})^2 + 4\|\mathbf{Y}\|_2^2 n}}{2\|Y\|_2^2}, \qquad (9.8)$$

where $\mathbf{X}_j$ denotes the $n \times 1$ vector $(X_1^{(j)}, \ldots, X_n^{(j)})^T$.

The estimator for σ^2 is then $\hat{\sigma}^2 = \hat{\rho}^{-2}$. We remark that for the case of a (non-mixture) linear model, Sun and Zhang (2010) propose an estimator for the error variance σ^2 which is less biased than $\hat{\sigma}^2$.

9.2.2.2 ℓ_1-penalized MLE for mixture of Gaussian regressions

Consider the mixture of Gaussian regressions model in (9.2), i.e., the FMR model. Define the following estimator for the unknown parameter $\theta = (\phi_1, \ldots, \phi_k, \rho_1, \ldots, \rho_k, \pi_1, \ldots, \pi_{k-1})$:

$$\hat{\theta}(\lambda) = \underset{\theta \in \Theta}{\arg\min} -n^{-1}\ell_{\text{pen},\lambda}(\theta), \qquad (9.9)$$

with

$$-n^{-1}\ell_{\text{pen},\lambda}(\theta) = -n^{-1}\sum_{i=1}^{n}\log\left(\sum_{r=1}^{k}\pi_r\frac{\rho_r}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}(\rho_r Y_i - X_i\phi_r)^2\right)\right)$$

$$+\lambda\sum_{r=1}^{k}\|\phi_r\|_1, \tag{9.10}$$

$$\Theta = \mathbb{R}^{kp} \times \mathbb{R}^{k}_{>0} \times \Pi, \tag{9.11}$$

where $\Pi = \{\pi; \pi_r > 0 \text{ for } r = 1,\ldots,k-1 \text{ and } \sum_{r=1}^{k-1}\pi_r < 1\}$.

One can also use a modified penalty term of the form

$$\lambda\sum_{r=1}^{k}\pi_r^{\gamma}\|\phi_r\|_1 \quad (\gamma \geq 0), \tag{9.12}$$

for e.g. $\gamma = 1/2$ or $\gamma = 1$, and we then denote the negative penalized log-likelihood by $-n^{-1}\ell_{\text{pen},\lambda}^{(\gamma)}$. The choice $\gamma = 0$ yields the proposal in (9.10). Using $\gamma \neq 0$ is more appropriate for unbalanced cases where the true probabilities for the mixture components π^0 differ substantially. We sometimes refer to the estimator in (9.9) as the FMRLasso. We note that it involves optimization of a non-convex negative log-likelihood, due to the appearance of several mixture components.

The penalty function in (9.10) involves the ℓ_1-norm of the component specific ratio's $\phi_r = \frac{\beta_r}{\sigma_r}$ and hence small variances are penalized. As shown next (for the case with $\gamma = 0$), the penalized likelihood stays finite when $\sigma_r \to 0$: this is in sharp contrast to the unpenalized MLE where the likelihood tends to infinity if $\sigma_r \to 0$, see for example McLachlan and Peel (2000).

Proposition 9.1. *Assume that $Y_i \neq 0$ for all $i = 1,\ldots,n$. Then the penalized negative likelihood $-n^{-1}\ell_{\text{pen},\lambda}(\theta)$ in (9.10) with $\lambda > 0$ is bounded from below for all values $\theta \in \Theta$ from (9.11).*

Proof. We restrict ourselves to a two class mixture with $k = 2$. Consider the function $u(\xi)$, ξ as in (9.1), defined as

$$u(\xi) = \exp(\ell_{\text{pen}}(\xi))$$

$$\propto \prod_{i=1}^{n}\left(\frac{\pi}{\sigma_1}e^{\frac{-(Y_i-X_i\beta_1)^2}{2\sigma_1^2}} + \frac{(1-\pi)}{\sigma_2}e^{\frac{-(Y_i-X_i\beta_2)^2}{2\sigma_2^2}}\right)e^{-\frac{\lambda}{n}\frac{\|\beta_1\|_1}{\sigma_1}}e^{-\frac{\lambda}{n}\frac{\|\beta_2\|_1}{\sigma_2}}. \tag{9.13}$$

We will show that $u(\xi)$ is bounded from above as a function of $\xi = (\sigma_1,\sigma_2,\beta_1,\beta_2,\pi) \in \Xi = \mathbb{R}^2_{>0} \times \mathbb{R}^{2p} \times (0,1)$. Then clearly $-n^{-1}\ell_{\text{pen},\lambda}(\theta)$ is bounded from below on $\theta = (\rho_1,\rho_2,\phi_1,\phi_2,\pi) \in \Theta = \mathbb{R}^2_{>0} \times \mathbb{R}^{2p} \times (0,1)$.

Before giving a rigorous proof, we remark that the critical point for unboundedness is if we choose for an arbitrary sample point $i \in 1,\ldots,n$ a β_1^* such that $Y_i - X_i\beta_1^* = 0$

and letting $\sigma_1 \to 0$. Without the penalty term $\exp(-\frac{\lambda}{n} \frac{\|\beta_1^*\|_1}{\sigma_1})$ in (9.13) the function would tend to infinity as $\sigma_1 \to 0$. But as $Y_i \neq 0$ for all $i \in 1, \ldots, n$, β_1^* cannot be zero and as a consequence, $\exp(-\frac{\lambda}{n} \frac{\|\beta_1^*\|_1}{\sigma_1})$ forces $u(\xi)$ to tend to 0 as $\sigma_1 \to 0$.

We give now a more formal proof for boundedness of $u(\xi)$. Choose a small $0 < \varepsilon_1 < \min Y_i^2$ and $\varepsilon_2 > 0$. Since $Y_i \neq 0$, $i = 1 \ldots n$, there exists a small constant $m > 0$ such that

$$0 < \min Y_i^2 - \varepsilon_1 \leq (Y_i - X_i \beta_1)^2 \tag{9.14}$$

holds for all $i = 1 \ldots n$ as long as $\|\beta_1\|_1 < m$, and

$$0 < \min Y_i^2 - \varepsilon_1 \leq (Y_i - X_i \beta_2)^2 \tag{9.15}$$

holds for all $i = 1 \ldots n$ as long as $\|\beta_2\|_1 < m$. Furthermore, there exists a small constant $\delta > 0$ such that

$$\frac{1}{\sigma_1} e^{-\frac{(\min Y_i^2 - \varepsilon_1)}{2\sigma_1^2}} < \varepsilon_2 \quad \text{and} \quad \frac{1}{\sigma_1} e^{-\frac{\lambda}{n} \frac{m}{\sigma_1}} < \varepsilon_2 \tag{9.16}$$

hold for all $0 < \sigma_1 < \delta$, and

$$\frac{1}{\sigma_2} e^{-\frac{(\min Y_i^2 - \varepsilon_1)}{2\sigma_2^2}} < \varepsilon_2 \quad \text{and} \quad \frac{1}{\sigma_2} e^{-\frac{\lambda}{n} \frac{m}{\sigma_2}} < \varepsilon_2 \tag{9.17}$$

hold for all $0 < \sigma_2 < \delta$.

Define the set $B = \{(\sigma_1, \sigma_2, \beta_1, \beta_2, \pi) \in \Xi;\ \delta \leq \sigma_1, \sigma_2\}$. Now $u(\xi)$ is trivially bounded on B. From the construction of B and equations (9.14)-(9.17) we easily see that $u(\xi)$ is also bounded on B^c and therefore bounded on Ξ. $\qquad \square$

9.2.3 Properties of the ℓ_1-penalized maximum likelihood estimator

As with the standard Lasso, due to the ℓ_1-penalty, the estimator in (9.9) is shrinking some of the coefficients of $\phi_1, \ldots, \phi_k$ exactly to zero, depending on the magnitude of the regularization parameter λ. Thus, we can do variable selection as follows. Denote by

$$\hat{S} = \hat{S}(\lambda) = \{(r, j);\ \hat{\phi}_{r,j}(\lambda) \neq 0,\ r = 1, \ldots, k,\ j = 1, \ldots, p\}. \tag{9.18}$$

The set $\hat{S}$ denotes the collection of non-zero estimated, i.e., selected, regression coefficients among the k mixture components.

We present in Section 9.4.3 an oracle inequality for the estimator in (9.9) describing prediction optimality and a result on estimating the high-dimensional regression parameters in terms of $\|\hat{\phi} - \phi^0\|_1$, where ϕ^0 denotes the true parameter. From the

latter, a variable screening result can be derived exactly along the lines as in Section 2.5 in Chapter 2 (e.g. formula (2.13)), assuming sufficiently large (in absolute value) non-zero coefficients (the analogue of the beta-min condition in formula (2.23), see also Sections 7.4 and 7.8.5), saying that with high probability

$$\hat{S} \supseteq S_0 = \{(r,j);\ \phi_{r,j}^0 \neq 0,\ r = 1,\ldots,k,\ j = 1,\ldots,p\}. \tag{9.19}$$

9.2.4 Selection of the tuning parameters

The regularization parameters to be selected are the number of mixture components k and the penalty parameter λ. In addition, we may also want to select the type of the penalty function, i.e., selection of γ in (9.12) among a few different candidate values.

We can use a cross-validation scheme for tuning parameter selection minimizing the cross-validated negative log-likelihood. Alternatively, a simple and computationally very convenient approach is to use the BIC criterion which minimizes

$$\mathrm{BIC} = -2\ell(\hat{\theta}_{\lambda,k}^{(\gamma)}) + \log(n)\mathrm{df}, \tag{9.20}$$

over a grid of candidate values for k, λ and maybe also γ. Here, $\hat{\theta}_{\lambda,k}^{(\gamma)}$ denotes the estimator in (9.9) using the parameters λ, k, γ in (9.10) or (9.12), respectively, and $-\ell(\cdot)$ is the negative log-likelihood. Furthermore, $\mathrm{df} = k + (k-1) + \sum_{j=1}^{p}\sum_{r=1}^{k} 1(\hat{\phi}_{r,j} \neq 0)$ are the number of non-zero estimated parameters. A motivation for defining the degrees of freedom in this way is described in Section 2.11 from Chapter 2 for the ordinary Lasso in linear models, see also (2.35). However, there is as yet no rigorous theoretical argument justifying the use of the BIC criterion above for the ℓ_1-penalized MLE estimator in high-dimensional mixture models.

Regarding the grid of candidate values for λ, we consider $0 \leq \lambda_{\mathrm{grid},1} < \ldots < \lambda_{\mathrm{grid},g} = \lambda_{\max}$, where $\lambda_{\max}$ is given by

$$\lambda_{\max} = \max_{j=1,\ldots,p} \left| \frac{\mathbf{X}_j^T \mathbf{Y}}{\sqrt{n}\|\mathbf{Y}\|_2} \right|, \tag{9.21}$$

and $\mathbf{X}_j$ denotes the $n \times 1$ vector $(X_1^{(j)},\ldots,X_n^{(j)})^T$. At $\lambda_{\max}$, all coefficients $\hat{\phi}_j$, $(j = 1,\ldots,p)$ of the one-component model are exactly zero. This fact easily follows from (9.8).

9.2.5 Adaptive ℓ_1-penalization

An adaptive Lasso as described in Section 2.8 in Chapter 2 (see also Chapter 7) can also be used here to effectively address some bias problems of the (one-stage) Lasso-type estimator. If the underlying truth is very sparse, we expect a better variable selection and prediction accuracy with the adaptive procedure.

The two-stage adaptive ℓ_1-penalized estimator for a mixture of Gaussian regressions is defined as follows. Consider an initial estimate $\hat{\theta}_{\text{init}}$, for example from the estimator in (9.9). The adaptive criterion to be minimized involves a re-weighted ℓ_1-penalty term:

$$-n^{-1}\ell_{\text{adapt}}^{(\gamma)}(\theta) = -n^{-1}\sum_{i=1}^{n}\log\left(\sum_{r=1}^{k}\pi_r\frac{\rho_r}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}(\rho_r Y_i - X_i\phi_r)^2\right)\right)$$

$$+ \lambda\sum_{r=1}^{k}\pi_r^{\gamma}\sum_{j=1}^{p}w_{r,j}|\phi_{r,j}|,$$

$$w_{r,j} = \frac{1}{|\hat{\phi}_{\text{init};r,j}|}, \quad \theta = (\rho_1,\ldots,\rho_k,\phi_1,\ldots,\phi_k,\pi_1,\ldots,\pi_{k-1}), \quad (9.22)$$

where $\gamma \in \{0,1/2,1\}$. The estimator is then defined as

$$\hat{\theta}_{\text{adapt};\lambda}^{(\gamma)} = \arg\min_{\theta\in\Theta} -n^{-1}\ell_{\text{adapt}}^{(\gamma)}(\theta),$$

with Θ is as in (9.11). We refer to this estimator as the FMRAdaptLasso.

9.2.6 Riboflavin production with bacillus subtilis

We apply the Lasso-type estimator for FMR models to a data set about riboflavin (vitamin B_2) production by bacillus subtilis. The data has been kindly provided by DSM (Switzerland). The real-valued response variable is the logarithm of the riboflavin production rate and there are $p = 4088$ covariates (genes) measuring the logarithm of the expression level of 4088 genes. These measurements are from $n = 146$ samples of genetically engineered mutants of bacillus subtilis. The population seems to be rather heterogeneous as there are different strains of bacillus subtilis which are cultured under different fermentation conditions. We do not know the different homogeneous subgroups. For this reason, an FMR model with more than one component might be more appropriate than a single linear regression model.

We compute the FMRLasso estimator from (9.9) for $k = 1,\ldots,5$ components. To keep the computational effort reasonable we use only the 100 covariates (genes) ex-
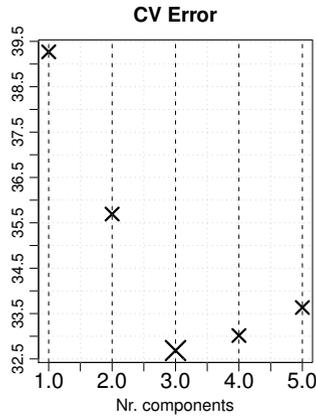
Fig. 9.1 Riboflavin production data. Cross-validated negative log-likelihood loss (*CV Error*) for the FMRLasso estimator when varying over different numbers of components. The figure is taken from Städler et al. (2010).
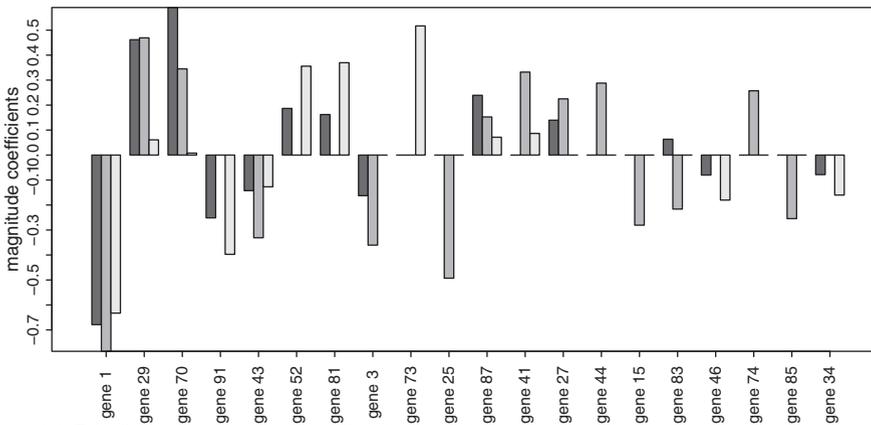
Fig. 9.2 Riboflavin production data. Coefficients of the twenty most important genes, ordered according to $\sum_{r=1}^{3} |\hat{\beta}_{r,j}|$, for the prediction optimal model with three components. The figure is taken from Städler et al. (2010).

hibiting the highest empirical variances.[2] We choose the tuning parameter λ by 10-fold cross-validation (using the log-likelihood loss). As a result we get five different estimators which we compare according to their cross-validated log-likelihood loss (*CV Error*). These numbers are plotted in Figure 9.1. The estimator with three components performs clearly best, resulting in a 17% improvement in prediction over a (non-mixture) linear model, and it selects 51 variables (genes). In Figure 9.2 the coefficients of the twenty most important genes, ordered according to $\sum_{r=1}^{3} |\hat{\beta}_{r,j}|$,

---

[2] We first select the 100 covariates having largest empirical variances. We then normalize these 100 variables to mean zero and empirical variance one.

are shown (we back-transform $\hat{\beta}_{r,j} = \hat{\phi}_{r,j}/\hat{\rho}_r$). From the important variables, only variable (gene) 83 exhibits an opposite sign of the estimated regression coefficients among the three different mixture components. However, it happens that some co-variates (genes) exhibit a strong effect in one or two mixture components but none in the remaining other components. Finally, for comparison, the one-component (non-mixture) model selects 26 genes where 24 of them are also selected in the three-component model.

9.2.7 Simulated example

We consider a scenario, with successively growing number of covariates where we compare the performance of the unpenalized MLE (Flexmix, according to the name of the R-package) with the estimators from Section 9.2.2.2 (FMRLasso) and Section 9.2.5 (FMRAdapt). For the two latter methods, we use the penalty function in (9.12) with $\gamma = 1$.

The simulation is from a Gaussian FMR model as in (9.2): the coefficients π_r, β_r, σ_r are as follows,

$$\beta_1 = (3,3,3,3,3,0,\ldots,0), \ \beta_2 = (-1,-1,-1,-1,-1,0,\ldots,0),$$
$$\sigma_1 = \sigma_2 = 2, \ \pi_1 = \pi_2 = 1/2.$$

The covariate X is generated from a multivariate normal distribution with mean 0 and covariance matrix I. This results in a signal to noise ratio of 12.1. Finally, we use sample size $n = 100$ and vary p from 5 to 125 by adding up to 120 noise covariates.

We use training-, validation- and test data of equal size n. The estimators are computed on the training data, with the tuning parameter λ selected by minimizing twice the negative log-likelihood (log-likelihood loss) on the validation data. As performance measure, the predictive log-likelihood loss (twice the negative log-likelihood) of the estimated model is computed on the test data.

Regarding variable selection, we count a covariable $X^{(j)}$ as selected if $\hat{\beta}_{r,j} \neq 0$ for at least one $r \in \{1,\ldots,k\}$. To assess the performance of FMRLasso on recovering the sparsity structure, we report the number of truly selected covariates (True Positives) and falsely selected covariates (False Positives).

The boxplots in Figures 9.3 of the predictive log-likelihood loss (*Error*), the True Positives (*TP*) and the False Positives (*FP*) over 100 simulation runs summarize the results for the different models. We see from this figure that the MLE performs very badly when adding noise covariates. On the other hand, the penalized estimators remain stable. There is also a substantial gain of the FMRAdaptLasso over FMRLasso in terms of log-likelihood loss and false positives.
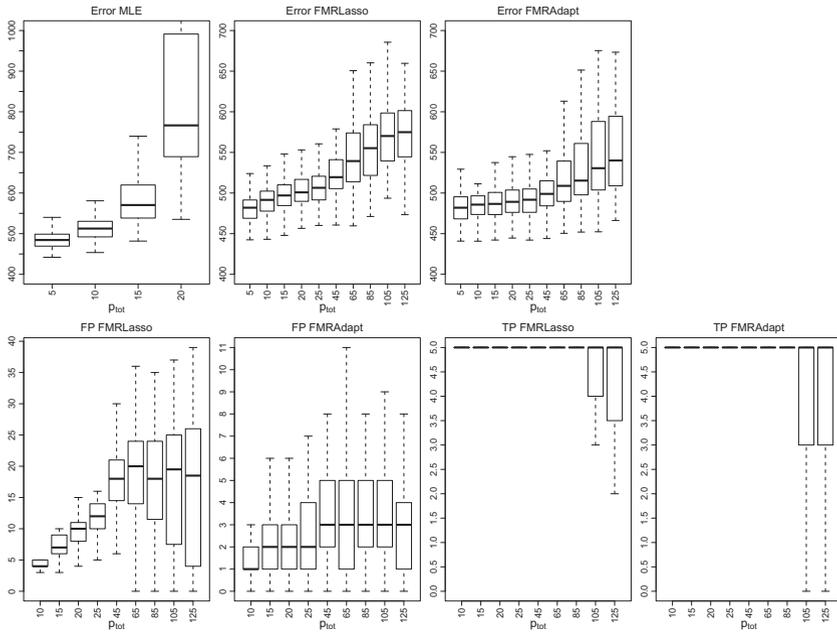
Fig. 9.3 Simulation when varying the dimension p (denoted by p_{tot}). Top: negative log-likelihood loss (*Error*) for MLE, FMRLasso, FMRAdaptLasso. Bottom: False Positives (*FP*) and True Positives (*TP*) for FMRLasso and FMRAdapt. The figure is taken from Städler et al. (2010).

9.2.8 Numerical optimization

We present here an EM and generalized EM (GEM) algorithm for optimizing the criterion in (9.10). The GEM modification is used for dealing with the penalty function in (9.12) with $\gamma \neq 0$. In Section 9.2.9.1 we will discuss numerical convergence properties of the algorithm.

9.2.9 GEM algorithm for optimization

Maximization of the log-likelihood of a mixture density is often done using the traditional EM algorithm of Dempster et al. (1977). We closely follow Städler et al. (2010) who describe an efficient adaptation for high-dimensional FMR models.

Consider the complete log-likelihood:

$$\ell_c(\theta; Y, \Delta) = \sum_{i=1}^{n} \sum_{r=1}^{k} \Delta_{i,r} \log\left(\frac{\rho_r}{\sqrt{2\pi}} e^{-\frac{1}{2}(\rho_r Y_i - X_i \phi_r)^2} \right) + \Delta_{i,r} \log(\pi_r).$$

Here, $(\Delta_{i,1}, \ldots, \Delta_{i,k})$ $(i = 1, \ldots, n)$ are i.i.d unobserved multinomial variables show-
ing the component-membership of the ith observation in the FMR model: $\Delta_{i,r} = 1$ if
observation i belongs to component r and $\Delta_{i,r} = 0$ otherwise. The expected complete
(scaled) negative log-likelihood is then:

$$Q(\theta|\theta') = -n^{-1}\mathbb{E}[\ell_c(\theta; Y, \Delta)|Y, \theta'],$$

and the expected complete penalized negative log-likelihood (scaled) is

$$Q_{\text{pen}}(\theta|\theta') = Q(\theta|\theta') + \lambda \sum_{r=1}^{k} \pi_r^\gamma \|\phi_r\|_1.$$

The EM-algorithm works by iterating between the E- and M-step. Denote the pa-
rameter value at iteration m by $\theta^{[m]}$ $(m = 0, 1, 2, \ldots)$, where $\theta^{[0]}$ is a vector of starting
values.

E-Step: Compute $Q(\theta|\theta^{[m]})$ or equivalently

$$\hat{\gamma}_{i,r} = \mathbb{E}[\Delta_{i,r}|Y, \theta^{[m]}] = \frac{\pi_r^{[m]} \rho_r^{[m]} e^{-\frac{1}{2}(\rho_r^{[m]}Y_i - X_i\phi_r^{[m]})^2}}{\sum_{r=1}^{k} \pi_r^{[m]} \rho_r^{[m]} e^{-\frac{1}{2}(\rho_r^{[m]}Y_i - X_i\phi_r^{[m]})^2}} \quad r = 1, \ldots, k, \quad i = 1, \ldots, n,$$

(see also the generalized M-step below how the $\hat{\gamma}_{i,r}$'s are used).

Generalized M-Step: Improve $Q_{\text{pen}}(\theta|\theta^{[m]})$ w.r.t. $\theta \in \Theta$.

a) *Improvement with respect to π:*
   Fix ϕ at the present value $\phi^{[m]}$.

   If $\gamma = 0$ in the penalty function in (9.12):

   $$\pi^{[m+1]} = \frac{\sum_{i=1}^{n} \hat{\gamma}_i}{n}, \quad \hat{\gamma}_i = (\hat{\gamma}_{i,1}, \ldots, \hat{\gamma}_{i,k})^T$$

   which is an explicit simple up-date minimizing $Q_{\text{pen}}(\theta|\theta^{[m]})$ with respect to π
   while keeping the other parameters ρ and ϕ fixed.

   If $\gamma \neq 0$, improve

   $$-n^{-1} \sum_{i=1}^{n} \sum_{r=1}^{k} \hat{\gamma}_{i,r} \log(\pi_r) + \lambda \sum_{r=1}^{k} \pi_r^\gamma \|\phi_r^{[m]}\|_1 \qquad (9.23)$$

   with respect to the probability simplex

   $$\{\pi; \pi_r > 0 \text{ for } r = 1, \ldots, k \text{ and } \sum_{r=1}^{k} \pi_r = 1\}.$$

Denote by $\bar{\pi}^{[m+1]} = \frac{\sum_{i=1}^{n} \hat{\gamma}_i}{n}$ which is a feasible point of the simplex. We can update π as

$$\pi^{[m+1]} = \pi^{[m]} + t^{[m]}(\bar{\pi}^{[m+1]} - \pi^{[m]}),$$

where $t^{[m]} \in (0, 1]$. In practice $t^{[m]}$ is chosen to be the largest value in the grid $\{\delta^u; u = 0, 1, 2, \ldots\}$ $(0 < \delta < 1)$ such that (9.23) is decreased. A typical choice is $\delta = 0.1$.

b) *Coordinate descent improvement with respect to ϕ and ρ:*

A simple calculation shows (Problem 9.3), that the M-Step decouples for each component into k distinct optimization problems of the form

$$-\log(\rho_r) + \frac{1}{2n_r}\|\rho_r \tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\phi_r\|^2 + \frac{n\lambda}{n_r}\left(\pi_r^{[m+1]}\right)^{\gamma}\|\phi_r\|_1, \qquad r = 1, \ldots, k \quad (9.24)$$

with

$$n_r = \sum_{i=1}^{n} \hat{\gamma}_{i,r}, \quad (\tilde{Y}_i, \tilde{X}_i) = \sqrt{\hat{\gamma}_{i,r}}(Y_i, X_i), \qquad r = 1, \ldots, k.$$

We denote by $\tilde{\mathbf{Y}}$ and $\tilde{\mathbf{X}}$ the vector or matrix including all sample indices $i = 1, \ldots, n$, respectively. The expression in (9.24) is of the same form as (9.7): in particular, it is convex in $(\rho_r, \phi_{r,1}, \ldots, \phi_{r,p})$. Instead of fully optimizing (9.24) we only minimize with respect to each of the coordinates, holding the other coordinates at their current value. Closed-form coordinate updates can easily be computed for each component r $(r = 1, \ldots, k)$ using (9.8):

$$\rho_r^{[m+1]} = \frac{\tilde{\mathbf{Y}}^T \tilde{\mathbf{X}} \phi_r^{[m]} + \sqrt{(\tilde{\mathbf{Y}}^T \tilde{\mathbf{X}} \phi_r^{[m]})^2 + 4\|\tilde{\mathbf{Y}}\|^2 n_r}}{2\|\tilde{\mathbf{Y}}\|^2},$$

$$\phi_{r,j}^{[m+1]} = \begin{cases} 0 & \text{if } |S_j| \leq n\lambda \left(\pi_r^{[m+1]}\right)^{\gamma}, \\ \left(n\lambda \left(\pi_r^{[m+1]}\right)^{\gamma} - S_j\right)/\|\tilde{\mathbf{X}}_j\|^2 & \text{if } S_j > n\lambda \left(\pi_r^{[m+1]}\right)^{\gamma}, \\ -\left(n\lambda \left(\pi_r^{[m+1]}\right)^{\gamma} + S_j\right)/\|\tilde{\mathbf{X}}_j\|^2 & \text{if } S_j < -n\lambda \left(\pi_r^{[m+1]}\right)^{\gamma}, \end{cases}$$

where S_j is defined as

$$S_j = -\rho_r^{[m+1]} \tilde{\mathbf{X}}_j^T \tilde{\mathbf{X}} + \sum_{s<j} \phi_{r,s}^{[m+1]} \tilde{\mathbf{X}}_j^T \tilde{\mathbf{X}}_s + \sum_{s>j} \phi_{r,s}^{[m]} \tilde{\mathbf{X}}_j^T \tilde{\mathbf{X}}_s$$

and $j = 1, \ldots, p$.

Because we only improve $Q_{\text{pen}}(\theta|\theta^{[m]})$ instead of a full minimization, see M-step a) and b), this is a generalized EM (GEM) algorithm. We call it the block coordinate descent generalized EM algorithm (BCD-GEM); the word block refers to the fact that we are up-dating all components of π at once. Its numerical properties are discussed next.

9.2.9.1 Numerical Convergence of the BCD-GEM algorithm

We are addressing here convergence properties of the BCD-GEM algorithm described in Section 9.2.9 for the case with $\gamma = 0$. A detailed account of the convergence properties of the EM algorithm in a general setting has been given by Wu (1983). Under regularity conditions including differentiability and continuity of the objective function, he proves convergence to stationary points for the EM algorithm. For the GEM algorithm, similar statements are true under conditions which are often hard to verify. We are a bit less ambitious and ask only whether a cluster point of an iterative algorithm equals a stationary point of the objective function (see Proposition 9.2 below), that is, we do not address the issue whether the algorithm actually converges (the latter can be checked on a given example up to numerical errors).

As a GEM algorithm, the BCD-GEM algorithm has the descent property which means, that the criterion function is reduced in each iteration,

$$-n^{-1}\ell_{\text{pen},\lambda}(\theta^{[m+1]}) \leq -n^{-1}\ell_{\text{pen},\lambda}(\theta^{[m]}). \tag{9.25}$$

Since $-n^{-1}\ell_{\text{pen},\lambda}(\theta)$ is bounded from below as discussed in Proposition 9.1, the following result holds. For the BCD-GEM algorithm, $-n^{-1}\ell_{\text{pen},\lambda}(\theta^{[m]})$ decreases monotonically to some value $\bar{\ell} > -\infty$.

Furthermore, we show here convergence to a stationary point for the convex penalty function in (9.10) (which is (9.12) with $\gamma = 0$).

Proposition 9.2. *Consider the BCD-GEM algorithm for the objective function in (9.10) (i.e. $\gamma = 0$) and denote by $\hat{\theta}^{[m]}$ the parameter vector after m iterations. Then, every cluster point of the sequence $\{\hat{\theta}^{[m]}; m = 0, 1, 2, \ldots\}$ is a stationary point of the objective function in (9.10).*

A proof is given in the next Section 9.2.10. It uses the crucial facts that $Q_{\text{pen}}(\theta|\theta')$ is a convex function in θ and that it is strictly convex in each coordinate of θ.

9.2.10 Proof of Proposition 9.2

First, we give a definition of a stationary point for non-differentiable functions (see also Tseng (2001)). Let u be a function defined on a open set $U \subset \mathbb{R}^D$: $x \in U$ is called a stationary point if $u'(x;d) = \lim_{\alpha \downarrow 0} \frac{u(x+\alpha d) - u(x)}{\alpha} \geq 0 \quad \forall d \in \mathbb{R}^D$.

The density of the complete data is given by

$$f_c(Y, \Delta | \theta) = \prod_{i=1}^{n} \prod_{r=1}^{k} \pi_r^{\Delta_{i,r}} \left(\frac{\rho_r}{\sqrt{2\pi}} e^{-\frac{1}{2}(\rho_r Y_i - X_i \phi_r)^2} \right)^{\Delta_{i,r}},$$

whereas the density of the observed data is given by

$$f_{obs}(Y|\theta) = \prod_{i=1}^{n} \sum_{r=1}^{k} \pi_r \frac{\rho_r}{\sqrt{2\pi}} e^{-\frac{1}{2}(\rho_r Y_i - X_i \phi_r)^2},$$

$$\theta = (\phi_1, \ldots, \phi_k, \rho_1, \ldots, \rho_k, \pi_1, \ldots, \pi_{k-1}) \in \Theta.$$

Furthermore, the conditional density of the complete data given the observed data is given by $g(Y, \Delta | Y, \theta) = f_c(Y, \Delta | \theta) / f_{obs}(Y|\theta)$. Then, the penalized negative log-likelihood fulfills the equation

$$v_{\text{pen}}(\theta) = -n^{-1} \ell_{\text{pen}, \lambda}^{[0]}(\theta) = -n^{-1} \log f_{obs}(Y|\theta) + \lambda \sum_{r=1}^{k} \|\phi_r\|_1$$

$$= Q_{\text{pen}}(\theta | \theta') - H(\theta | \theta'), \tag{9.26}$$

where $Q_{\text{pen}}(\theta | \theta') = -n^{-1} \mathbb{E}[\log f_c(Y, \Delta | \theta) | Y, \theta'] + \lambda \sum_{r=1}^{k} \|\phi_r\|_1$ (compare Section 9.2.9) and $H(\theta | \theta') = -n^{-1} \mathbb{E}[\log g(Y, \Delta | Y, \theta) | Y, \theta']$.

By Jensen's inequality we get the following well-known relationship:

$$H(\theta | \theta') \geq H(\theta' | \theta') \quad \forall \quad \theta \in \Theta. \tag{9.27}$$

We leave the derivation of (9.27) as Problem 9.4. We note that $Q_{\text{pen}}(\theta | \theta')$ and $H(\theta | \theta')$ are continuous functions in θ and θ'. If we think of them as functions in θ with fixed θ' we write also $Q_{\text{pen}, \theta'}(\theta)$ and $H_{\theta'}(\theta)$. Furthermore $Q_{\text{pen}, \theta'}(\theta)$ is a convex function in θ and strictly convex in each coordinate of θ.

We are now ready to start with the main parts of the proof which is inspired by Bertsekas (1995). Let $\theta^{[m]}$ be the sequence generated by the BCD-GEM algorithm; note that we drop the hat-notation in this proof. We need to show for a converging subsequence $\theta^{[m_j]} \to \bar{\theta} \in \Theta$ that $\bar{\theta}$ is a stationary point of $v_{\text{pen}}(\theta)$. Taking directional derivatives in equation (9.26) yields

$$v_{\text{pen}}'(\bar{\theta}; d) = Q_{\text{pen}, \bar{\theta}}'(\bar{\theta}; d) - \nabla H_{\bar{\theta}}(\bar{\theta})^T d.$$

Note that $\nabla H_{\bar{\theta}}(\bar{\theta}) = 0$ as $H_{\bar{\theta}}(x)$ is minimized for $x = \bar{\theta}$ (equation (9.27)). Therefore it remains to show that $Q'_{\text{pen},\bar{\theta}}(\bar{\theta};d) \geq 0$ for all directions d. Let

$$z_i^{[m]} = (\theta_1^{[m+1]}, \ldots, \theta_i^{[m+1]}, \theta_{i+1}^{[m]}, \ldots, \theta_D^{[m]}),$$

where $D = \dim(\theta) = (k+1)p + k - 1$. Using the definition of the algorithm we have:

$$Q_{\text{pen},\theta^{[m]}}(\theta^{[m]}) \geq Q_{\text{pen},\theta^{[m]}}(z_1^{[m]}) \geq \ldots \geq Q_{\text{pen},\theta^{[m]}}(z_{D-1}^{[m]}) \geq Q_{\text{pen},\theta^{[m]}}(\theta^{[m+1]}). \quad (9.28)$$

Additionally, from the properties of GEM (equation (9.26) and (9.27)) we have:

$$v_{\text{pen}}(\theta^{[0]}) \geq v_{\text{pen}}(\theta^{[1]}) \geq \ldots \geq v_{\text{pen}}(\theta^{[m]}) \geq v_{\text{pen}}(\theta^{[m+1]}). \quad (9.29)$$

Equation (9.29) and the converging subsequence imply that the sequence $\{v_{\text{pen}}(\theta^{[m]}); m = 0, 1, 2, \ldots\}$ converges to $v_{\text{pen}}(\bar{\theta})$. Thus we have:

$$0 \leq Q_{\text{pen},\theta^{[m]}}(\theta^{[m]}) - Q_{\text{pen},\theta^{[m]}}(\theta^{[m+1]})$$
$$= v_{\text{pen}}(\theta^{[m]}) - v_{\text{pen}}(\theta^{[m+1]}) + H_{\theta^{[m]}}(\theta^{[m]}) - H_{\theta^{[m]}}(\theta^{[m+1]})$$
$$\leq v_{\text{pen}}(\theta^{[m]}) - v_{\text{pen}}(\theta^{[m+1]}), \quad (9.30)$$

where we use in the last inequality that $H_{\theta^{[m]}}(\theta^{[m]}) - H_{\theta^{[m]}}(\theta^{[m+1]}) \leq 0$ due to (9.27). The right-hand side converges to $v_{\text{pen}}(\bar{\theta}) - v_{\text{pen}}(\bar{\theta}) = 0$ and we conclude that the sequence $\{Q_{\text{pen},\theta^{[m]}}(\theta^{[m]}) - Q_{\text{pen},\theta^{[m]}}(\theta^{[m+1]}); m = 0, 1, 2, \ldots\}$ converges to zero.

We now show that $\{\theta_1^{[m_j+1]} - \theta_1^{[m_j]}\}$ converges to zero for the subsequence m_j ($j \to \infty$). Assume the contrary, in particular that $\{z_1^{[m_j]} - \theta^{[m_j]}\}$ does not converge to 0. Let $\delta^{[m_j]} = \|z_1^{[m_j]} - \theta^{[m_j]}\|_2$. Without loss of generality (by restricting to a subsequence) we may assume that there exists some $\bar{\delta} > 0$ such that $\delta^{[m_j]} > \bar{\delta}$ for all j. Let $s_1^{[m_j]} = \frac{z_1^{[m_j]} - \theta^{[m_j]}}{\delta^{[m_j]}}$ where $s_1^{[m_j]}$ differs from zero only for the first component. As $s_1^{[m_j]}$ belongs to a compact set ($\|s_1^{[m_j]}\|_2 = 1$) we may assume that $s_1^{[m_j]}$ converges to $\bar{s}_1$. Let us fix some $\varepsilon \in [0, 1]$. Notice that $0 \leq \varepsilon\bar{\delta} \leq \delta^{[m_j]}$. Therefore, $\theta^{[m_j]} + \varepsilon\bar{\delta}s_1^{[m_j]}$ lies on the segment joining $\theta^{[m_j]}$ and $z_1^{[m_j]}$, and belongs to Θ, because Θ is convex. As $Q_{\text{pen},\theta^{[m_j]}}(.)$ is convex and $z_1^{[m_j]}$ minimizes this function over all values that differ from $\theta^{[m_j]}$ along the first coordinate, we obtain

$$Q_{\text{pen},\theta^{[m_j]}}(z_1^{[m_j]}) = Q_{\text{pen},\theta^{[m_j]}}(\theta^{[m_j]} + \delta^{[m_j]}s_1^{[m_j]}) \leq Q_{\text{pen},\theta^{[m_j]}}(\theta^{[m_j]} + \varepsilon\bar{\delta}s_1^{[m_j]})$$
$$\leq Q_{\text{pen},\theta^{[m_j]}}(\theta^{[m_j]}). \quad (9.31)$$

We conclude, using (9.31) in the second and (9.28) in the last inequality,

$$0 \leq Q_{\text{pen},\theta^{[m_j]}}(\theta^{[m_j]}) - Q_{\text{pen},\theta^{[m_j]}}(\theta^{[m_j]} + \varepsilon\bar{\delta}s_1^{[m_j]})$$

$$\leq Q_{\text{pen},\theta^{[m_j]}}(\theta^{[m_j]}) - Q_{\text{pen},\theta^{[m_j]}}(z_1^{[m_j]}) \leq Q_{\text{pen},\theta^{[m_j]}}(\theta^{[m_j]}) - Q_{\text{pen},\theta^{[m_j]}}(\theta^{[m_j+1]}).$$

Using (9.30) and continuity of $Q_{\text{pen},x}(y)$ in both arguments x and y we conclude by taking the limit $j \to \infty$:

$$Q_{\text{pen},\bar{\theta}}(\bar{\theta} + \varepsilon\bar{\delta}\bar{s}_1) = Q_{\text{pen},\bar{\theta}}(\bar{\theta}) \quad \forall \varepsilon \in [0,1].$$

Since $\bar{\delta}\bar{s}_1 \neq 0$, this contradicts the strict convexity of $Q_{\text{pen},\bar{\theta}}(x_1, \bar{\theta}_2, \ldots, \bar{\theta}_D)$ as a function of the first coordinate. Thus, this contradiction establishes that $z_1^{[m_j]}$ converges to $\bar{\theta}$.

From the definition of the algorithm we have:

$$Q_{\text{pen}}(z_1^{[m_j]}|\theta^{[m_j]}) \leq Q_{\text{pen}}(x_1, \theta_2^{[m_j]}, \ldots, \theta_D^{[m_j]}|\theta^{[m_j]}) \qquad \forall x_1.$$

By continuity and taking the limit $j \to \infty$ we then obtain:

$$Q_{\text{pen},\bar{\theta}}(\bar{\theta}) \leq Q_{\text{pen},\bar{\theta}}(x_1, \bar{\theta}_2, \ldots, \bar{\theta}_D) \qquad \forall x_1.$$

Repeating the argument for the other coordinates we conclude that $\bar{\theta}$ is a coordinatewise minimum. Therefore, following Tseng (2001), $\bar{\theta}$ is easily seen to be a stationary point of $Q_{\text{pen},\bar{\theta}}(.)$, in particular $Q'_{\text{pen},\bar{\theta}}(\bar{\theta};d) \geq 0$ for all directions d (see the definition of a stationary point at the beginning of the proof). $\qquad \square$

9.3 Linear mixed effects models

Unlike for mixture models as in Section 9.2, a grouping structure among observations may be known. Such a structure can then be incorporated using mixed effects models which extend linear models by including random effects in addition to fixed effects. The maximum likelihood approach leads to a problem with a non-convex loss function. From an algorithmic point of view, we develop a coordinate gradient descent method and show its numerical convergence to a stationary point. Regarding statistical properties, oracle results can be established using the more general theory in Section 9.4. Besides methodology and theory, we will also empirically illustrate that there may be a striking improvement if we take the cluster or grouping structure in the data into account.

9.3.1 The model and ℓ_1-penalized estimation

We assume N different groups with corresponding grouping index $i = 1, \ldots, N$. There are n_i observations within the ith group with corresponding index $j = 1, \ldots, n_i$. Denote by $N_T = \sum_{i=1}^{N} n_i$ the total number of observations. For each observation, we observe a univariate response variable Y_{ij}, a p-dimensional fixed effects covariate X_{ij} and a q-dimensional random effects covariate Z_{ij}. We consider the following model:

$$Y_{ij} = X_{ij}\beta + Z_{ij}b_i + \varepsilon_{ij}, \quad i = 1, \ldots, N, \quad j = 1, \ldots, n_i. \tag{9.32}$$

assuming that $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ independent for $i = 1, \ldots, N$ and $j = 1, \ldots, n_i$, $b_i \sim \mathcal{N}_q(0, \Gamma)$ independent for $i = 1, \ldots, N$ and independent of $\varepsilon_{11}, \ldots, \varepsilon_{Nn_N}$. Here we denote by $\beta \in \mathbb{R}^p$ the vector of the unknown fixed effects regression coefficients and by $b_i \in \mathbb{R}^q$ $(i = 1, \ldots, N)$ the random effects regression coefficients. All observations have the coefficient vector β in common whereas the value of b_i depends on the group that the observation belongs to. In other words, for each group there are group-specific deviations b_i from the overall effects β. We assume in the sequel that the design variables X_{ij} and Z_{ij} are deterministic, i.e., fixed design. Furthermore, we assume that $\Gamma = \Gamma_\tau$ is a covariance matrix where τ is a set of parameters of dimension q^* such that Γ is positive definite. Possible structures for Γ include a multiple of the identity, $\Gamma = \tau^2 I$ with $q^* = 1$, a diagonal matrix $\Gamma = \mathrm{diag}(\tau_1^2, \ldots, \tau_q^2)$ with $q^* = q$ or a general positive definite matrix with $q^* = q(q+1)/2$.

We re-write model (9.32) using the standard notation in mixed effects models (Pinheiro and Bates, 2000):

$$\mathbf{Y}_i = \mathbf{X}_i\beta + \mathbf{Z}_ib_i + \varepsilon_i, \quad i = 1, \ldots, N, \tag{9.33}$$

where $\mathbf{Y}_i$ is an $n_i \times 1$ vector of responses of the ith group, $\mathbf{X}_i$ is an $n_i \times p$ fixed effects design matrix, $\mathbf{Z}_i$ an $n_i \times q$ random effects design matrix and ε_i an $n_i \times 1$ error vector. We allow that the number p of fixed effects regression coefficients may be much larger than the total number of observations, i.e., $p \gg N_T$. Conceptually, the number q of random effects may be very large if the covariance matrix $\Gamma = \Gamma_\tau$ is of low dimension q^*. However, in the sequel we will restrict ourselves to the case where a covariate is modeled with a random effect only if it allows for a fixed effect in the model as well, that is $\mathbf{Z}_i \subseteq \mathbf{X}_i$. The aim is to estimate the fixed effects parameter vector β, the random effects b_i and the covariance parameters σ^2 and Γ.

From model (9.33) we derive that $\mathbf{Y}_1, \ldots, \mathbf{Y}_N$ are independent with distributions,

$$\mathbf{Y}_i \sim \mathcal{N}_{n_i}(\mathbf{X}_i\beta, V_i(\tau, \sigma^2)),$$
$$V_i(\tau, \sigma^2) = \mathbf{Z}_i\Gamma_\tau\mathbf{Z}_i^T + \sigma^2 I_{n_i \times n_i}.$$

Hence, the negative log-likelihood function of $\mathbf{Y}_1, \ldots, \mathbf{Y}_N$ is given by (Problem 9.5)

$$-\ell(\beta,\tau,\sigma^2) = \frac{1}{2}\sum_{i=1}^{N}\left(n_i\log(2\pi) + \log\det(V_i) + (\mathbf{Y}_i - \mathbf{X}_i\beta)^T V_i^{-1}(\mathbf{Y}_i - \mathbf{X}_i\beta)\right).$$

$$(9.34)$$

9.3.2 The Lasso in linear mixed effects models

Due to the possibly large number of covariates, i.e., the $p \gg N_T$ setting, we regularize by ℓ_1-penalization for the fixed regression coefficients and thus achieve a sparse solution with respect to the fixed effects. Consider the objective function

$$Q_\lambda(\beta,\tau,\sigma^2) = \frac{1}{2}\sum_{i=1}^{N}\left(\log\det(V_i) + (\mathbf{Y}_i - \mathbf{X}_i\beta)^T V_i^{-1}(\mathbf{Y}_i - \mathbf{X}_i\beta)\right) + \lambda\|\beta\|_1,$$

$$(9.35)$$

where $\lambda \geq 0$ is the regularization parameter and $V_i = V_i(\tau)$. Consequently, we estimate the fixed regression coefficient vector β and the covariance parameters τ and σ^2 by

$$\hat{\beta}(\lambda), \hat{\tau}^2(\lambda), \hat{\sigma}^2(\lambda) = \underset{\beta,\tau\in\mathscr{R}^{q^*},\sigma^2}{\arg\min}\ Q_\lambda(\beta,\tau,\sigma^2), \qquad (9.36)$$

where $\mathscr{R}^{q^*} = \{\tau \in \mathbb{R}^{q^*};\ \Gamma_\tau \text{ positive definite}\}$. We refer to the estimator as the LMM-Lasso (Linear Mixed effects Model Lasso). For fixed covariance parameters τ, σ^2, the minimization with respect to β is a convex optimization problem. However, over all parameters, we have a non-convex objective function and hence, we have to deal with a non-convex problem, see Problem 9.5.

9.3.3 Estimation of the random effects coefficients

The random regression coefficients b_i $(i = 1,\ldots,N)$ can be estimated using the maximum a-posteriori (MAP) principle. Denoting by $p(\cdot)$ the density of the corresponding Gaussian random variable, we consider

$$
\begin{aligned}
b_i^* &= \underset{b_i}{\arg\max}\, p(b_i|\mathbf{Y}_1,\ldots,\mathbf{Y}_N;\beta,\tau,\sigma^2) = \underset{b_i}{\arg\max}\, p(b_i|\mathbf{Y}_i;\beta,\tau,\sigma^2) \\
&= \underset{b_i}{\arg\max}\, \frac{p(\mathbf{Y}_i|b_i;\beta,\sigma^2)\cdot p(b_i|\tau)}{p(\mathbf{Y}_i|\beta,\tau,\sigma^2)} \\
&= \underset{b_i}{\arg\min}\left\{\frac{1}{\sigma^2}\|\mathbf{Y}_i - \mathbf{X}_i\beta - \mathbf{Z}_i b_i\|_2^2 + b_i^T\Gamma_\tau^{-1}b_i\right\}.
\end{aligned}
$$

Thus, we obtain

$$b_i^* = \left(\mathbf{Z}_i^T \mathbf{Z}_i + \sigma^2 \Gamma_\tau^{-1}\right)^{-1} \mathbf{Z}_i^T \mathbf{r}_i, \quad \mathbf{r}_i = (\mathbf{Y}_i - \mathbf{X}_i \beta)$$

which corresponds to a generalized Ridge Regression. Since the true values of β, τ and σ^2 are unknown, the b_i's are estimated by

$$\hat{b}_i = \left(\mathbf{Z}_i^T \mathbf{Z}_i + \hat{\sigma}^2 \Gamma_{\hat{\tau}}^{-1}\right)^{-1} \mathbf{Z}_i^T \hat{\mathbf{r}}_i,$$

where $\hat{\mathbf{r}}_i = (\mathbf{Y}_i - \mathbf{X}_i \hat{\beta})$, using the estimates from (9.36).

9.3.4 Selection of the regularization parameter

The estimation method requires to choose a regularization parameter λ. We can either use a cross-validation scheme for evaluating the out-of-sample negative log-likelihood or employ the Bayesian Information Criterion (BIC) defined by

$$-2\ell(\hat{\beta}, \hat{\tau}, \hat{\sigma}^2) + \log N_T \mathrm{df}, \tag{9.37}$$

where $\mathrm{df} = |\{j; \ \hat{\beta}_j \neq 0\}| + q^* + 1$ is the number of nonzero estimated parameters. The use of df as a measure of the degrees of freedom is motivated by the results described in Section 2.11 in Chapter 2 for the ordinary Lasso in linear models. A more rigorous theoretical argument justifying the use of the BIC criterion for the ℓ_1-penalized MLE in high-dimensional linear mixed effects models is missing: the BIC has been empirically found to perform reasonably well (Schelldorfer et al., 2011).

9.3.5 Properties of the Lasso in linear mixed effects models

Like the Lasso in linear models, the estimator in (9.36) is shrinking some of the coefficients $\beta_1, \ldots, \beta_p$ exactly to zero, depending on the value of the regularization parameter λ. Therefore, we can do variable selection (for fixed effects) as discussed before. Consider

$$\hat{S} = \hat{S}(\lambda) = \{j; \ \hat{\beta}_j(\lambda) \neq 0, \ j = 1, \ldots, p\}$$

as an estimator of the true underlying active set $S_0 = \{j; \ \beta_j^0 \neq 0, \ j = 1, \ldots, p\}$, where β^0 denotes the true parameter vector.

We discuss in Section 9.4.4 an oracle inequality for the estimator in (9.36). It implies optimality of the estimator for prediction and an ℓ_1-estimation error bound for $\|\hat{\beta} - \beta^0\|_1$. As usual, we can then derive a result for variable screening (see Section 2.5,

Corollary 7.6 and Section 7.8.5): with high probability, $\hat{S} \supseteq S_0$, assuming a beta-min condition ensuring sufficiently large (in absolute value) non-zero coefficients.

9.3.6 Adaptive ℓ_1-penalized maximum likelihood estimator

As we have discussed in Section 2.8 in Chapter 2 for linear models, the adaptive Lasso is an effective way to address the bias problems of the Lasso. The adaptive ℓ_1-penalized maximum likelihood estimator uses the following objective function instead of (9.35):

$$
\begin{aligned}
& Q_{\text{adapt},\lambda}(\beta, \tau, \sigma^2) \\
& = \frac{1}{2} \sum_{i=1}^{N} \left(\log \det(V_i) + (\mathbf{Y}_i - \mathbf{X}_i \beta)^T V_i^{-1} (\mathbf{Y}_i - \mathbf{X}_i \beta) \right) + \lambda \sum_{k=1}^{p} w_k |\beta_k|,
\end{aligned}
$$

where the weights $w_1, \ldots, w_p$ are derived from an initial estimation in (9.36) with $w_k = 1/|\hat{\beta}_{\text{init},k}(\lambda)|$ for $k = 1, \ldots, p$. The adaptive estimator is then defined by:

$$
\hat{\beta}_{\text{adapt}}(\lambda), \hat{\tau}_{\text{adapt}}(\lambda), \hat{\sigma}^2_{\text{adapt}}(\lambda) = \arg\min_{\beta, \tau \in \mathscr{R}^{q^*}, \sigma^2} Q_{\text{adapt},\lambda}(\beta, \tau, \sigma^2), \qquad (9.38)
$$

where $\mathscr{R}^{q^*}$ is as in (9.36). We indicate at the end of Section 9.4.4 that an oracle inequality applies to the adaptive estimator, implying optimality for prediction and results on estimation error and variable screening.

9.3.7 Computational algorithm

The estimation of the fixed regression parameters and the covariance parameters can be computed using a Coordinate Gradient Descent (CGD) algorithm. Such an algorithm has been described in Section 4.7.2 in Chapter 4 for the group Lasso with non-quadratic loss functions (where we used a block CGD algorithm whereas here, we do not have to deal with blocks). The main idea is to cycle through the coordinates and minimize the objective function with respect to only one coordinate while keeping the other parameters fixed, i.e., a Gauss-Seidel algorithm.

For computation (and also statistical theory), it is more convenient to work with a reparametrization. Define the parameter $\theta^T = (\beta^T, \tau^T, \log(\sigma)) = (\beta^T, \eta^T) \in \mathbb{R}^{p+q^*+1}$, where $\eta^T = (\tau^T, \log(\sigma)) := (\eta_1, \eta_2)$. Consider the functions

$$
\text{pen}(\beta) = \|\beta\|_1 = \sum_{k=1}^{p} |\beta_k|,
$$

$$g(\theta) = \left(\frac{1}{2} \sum_{i=1}^{N} \left(\log \det(V_i) + (\mathbf{Y}_i - \mathbf{X}_i \beta)^T V_i^{-1} (\mathbf{Y}_i - \mathbf{X}_i \beta) \right) \right),$$

$$V_i = \mathbf{Z}_i \Gamma_{\eta_1} \mathbf{Z}_i^T + \exp(\eta_2) I_{n_i \times n_i},$$

$$Q_\lambda(\theta) = g(\theta) + \mathrm{pen}(\beta).$$

The estimator in (9.36) is then re-written as

$$\hat{\theta}(\lambda) = \arg\min_{\theta} Q_\lambda(\theta), \tag{9.39}$$

where the optimization is with respect to $\theta = (\beta^T, \eta_1^T, \eta_2^T)$ with Γ_{η_1} positive definite.

In each step, we approximate $Q_\lambda(.)$ by a strictly convex quadratic function. Then we calculate a descent direction and we employ an inexact line search to ensure a decrease in the objective function. The ordinary Lasso in a generalized linear model has only regression coefficients to cycle through. This is in contrast to the problem here involving two kinds of parameters: fixed effects regression and covariance parameters.

Let $I(\theta)$ be the Fisher information of the model and e_j be the jth unit vector. The computational algorithm is summarized in Algorithm 5, and we turn now to some

---

Algorithm 5 Coordinate Gradient Descent (CGD) Algorithm

1: Let $\hat{\theta}^{[0]} \in \mathbb{R}^{p+q^*+1}$ be an initial parameter vector. Set $m = 0$.
2: **repeat**
3:     Increase m by one: $m \leftarrow m+1$.
    Denote by $\mathscr{S}^{[m]}$ the index cycling through the coordinates $\{1, \ldots, p, p+1, \ldots, p+q^*+1\}$: $\mathscr{S}^{[m]} = \mathscr{S}^{[m-1]} + 1 \mod (p+q^*+1)$. Abbreviate by $j = \mathscr{S}^{[m]}$ the value of $\mathscr{S}^{[m]}$.
4:     Choose an approximate 1×1 Hessian $H^{[m]} > 0$.
5:     Compute
    $d^{[m]} = \arg\min_d \left\{ g(\hat{\theta}^{[m-1]}) + \frac{\partial}{\partial \theta_j} g(\theta)|_{\hat{\theta}^{[m-1]}} d + \frac{1}{2} d^2 H^{[m]} + \lambda \mathrm{pen}(\hat{\theta}^{[m-1]} + d e_j) \right\}$. See (9.40)
    below.
6:     Choose a stepsize $\alpha^{[m]} > 0$ by the Armijo rule and set $\hat{\theta}^{[m]} = \hat{\theta}^{[m-1]} + \alpha^{[m]} d^{[m]} e_j$. The step length $\alpha^{[m]}$ is chosen in such a way that in each step, there is an improvement in the objective function $Q_\lambda(.)$. The Armijo rule itself is defined as follows:
    Armijo Rule: Choose $\alpha_0 > 0$ and let $\alpha^{[m]}$ be the largest element of $\{\alpha_0 \delta^\ell\}_{\ell=0,1,2,..}$ satisfying

$$Q_\lambda(\hat{\theta}^{[m]} + \alpha^{[m]} d^{[m]} e_j) \le Q_\lambda(\hat{\theta}^{[m]}) + \alpha^{[m]} \rho \triangle^{[m]},$$

    where $\triangle^{[m]} = \partial/\partial \theta_j g(\theta)|_{\hat{\theta}^{[m-1]}} d^{[m]} + \nu (d^{[m]})^2 H^{[m]} + \lambda \mathrm{pen}(\hat{\theta}^{[m]} + d^{[m]} e_j) - \lambda \mathrm{pen}(\hat{\theta}^{[m]})$.
    A reasonable choice of the constants are $\delta = 0.1, \rho = 0.001, \nu = 0$ and $\alpha_0 = 1$, see Bertsekas (1995) (and hence with $\nu = 0$, the quadratic term above is irrelevant). For a computational short-cut, see (9.41) below.
7: **until** numerical convergence

---

details of it.

The initial value $\theta^{[0]}$ matters since we are pursuing a non-convex optimization exhibiting local minima. A pragmatic but useful approach is to choose (a CV-tuned) ordinary Lasso solution for $\beta^{[0]}$, ignoring the grouping structure among the observations. By doing so, we ensure that we are at least as good (with respect to the objective function in (9.39)) as an ordinary Lasso in a linear model. With $\beta^{[0]}$ at hand, we then determine the covariance parameters $\eta^{[0]}$ by pursuing the iterations in Algorithm 5 for the coordinates $p+1,\dots p+q^*+1$.

Choice of an approximate Hessian $H^{[m]}$. The choice of an approximate 1×1 Hessian $H^{[m]}$ evaluated at the previous iteration $\theta^{[m-1]}$ (for the true Hessian which is the second partial derivative $\frac{\partial}{\partial \mathscr{S}^{[m]}} g(\theta)$ evaluated at $\theta^{[m-1]}$) in Algorithm 5 is also driven by considering computational efficiency. For numerical convergence (see Theorem 9.3 below), it is necessary that $H^{[m]}$ is positive and bounded. We base the choice on the Fisher information $I(\theta^{[m-1]})$, as described already in Section 4.7.2 in Chapter 4 (there, the approximate Hessian is a matrix due to the block up-date structure of the algorithm): denoting by $j = \mathscr{S}^{[m]}$,

$$H^{[m]} = \min(\max(I(\theta^{[m-1]})_{jj}, c_{min}), c_{max})$$

for some constants $0 < c_{min} < c_{max} < \infty$, e.g., $c_{min} = 10^{-6}$ and $c_{max} = 10^{8}$.

Computation of the direction $d^{[m]}$. Regarding the computation of the direction $d^{[m]}$, we have to distinguish whether the index $j = \mathscr{S}^{[m]}$ appears in the penalty $\mathrm{pen}(\theta)$ or not:

$$d^{[m]} = \begin{cases} \mathrm{median}\left(\frac{\lambda - \frac{\partial}{\partial \theta_j} g(\theta)|_{\theta^{[m-1]}}}{H^{[m]}}, -\hat{\beta}_j^{[m-1]}, \frac{-\lambda - \frac{\partial}{\partial \theta_j} g(\theta)|_{\theta^{[m-1]}}}{H^{[m]}} \right), \\ \qquad\qquad\qquad\qquad\qquad\qquad \text{if } j \in \{1,\dots,p\}, \\ \\ -\frac{\partial}{\partial \theta_j} g(\theta)|_{\theta^{[m-1]}} / H^{[m]}, \\ \qquad\qquad\qquad\qquad\qquad\qquad \text{if } j \in \{p+1,\dots,p+q^*+1\}. \end{cases}$$

$$(9.40)$$

Simplification of the direction and Armijo rule for the β parameter. If $H^{[m]}$ is not truncated, i.e., the numerical value equals $H^{[m]} = I(\theta^{[m-1]})_{jj}$ ($j = \mathscr{S}^{[m]}$), the update for the parameter vector β is explicit by taking advantage that $g(\theta)$ is quadratic with respect to β. Using $\alpha_0 = 1$, the stepsize $\alpha^{[m]} = \alpha_0 = 1$ chosen by the Armijo rule (with $\ell = 0$) leads to the minimum of $g(\theta)$ with respect to the component β_j. The update $\hat{\beta}_j^{[m]}$ is then given analytically by

$$\hat{\beta}_j^{[m]}(\lambda) = \mathrm{sign}\left(\sum_{i=1}^{N} (\mathbf{Y}_i - \tilde{\mathbf{Y}}_i) V_i^{-1} x_{\mathscr{S}^{[m]}}^{(i)} \right) \frac{\left(|\sum_{i=1}^{N} (\mathbf{Y}_i - \tilde{\mathbf{Y}}_i) V_i^{-1} x_{\mathscr{S}^{[m]}}^{(i)}| - \lambda \right)_+}{\sum_{i=1}^{N} x_{\mathscr{S}^{[m]}}^{(i)T} V_i^{-1} x_{\mathscr{S}^{[m]}}^{(i)}},$$

$$(9.41)$$

where the $n_i \times p$ matrix $\mathbf{X}_i = (x_1^{(i)}, \ldots, x_p^{(i)})$ and $\tilde{\mathbf{Y}}_i = \mathbf{X}_i^{(-j)} \beta_{(-j)}^{[m-1]}$ (leaving out the jth variable). Most often, $H^{[m]} = I(\theta^{[m-1]})_{jj}$ is not truncated and hence, the analytical formula (9.41) can be used (without the need to compute a direction and performing a line search). This simplification reduces the computational cost considerably, especially in the high-dimensional setup.

Due to the non-convexity of the objective function, we are not pursuing some warm-start initial values but instead, we use for all λ (from a grid of possible values) the same initial value from e.g. an ordinary Lasso solution. Computational speed-up can be achieved if the solution is sparse: as in described in Section 4.7.1, an active set strategy is very effective here as well. Instead of cycling through all coordinates, we can restrict ourselves to the current active set $S(\hat{\beta})$ and update all coordinates of $\hat{\beta}$ only once a while, e.g., every 10th or 20th iteration.

Using the general theory from Tseng and Yun (2009) on block coordinate gradient descent algorithms, one can establish the following result.

Proposition 9.3. *If $\hat{\theta}^{[m]}$ is chosen according to Algorithm 5, then every cluster point of $\{\hat{\theta}^{[m]}\}_{m \geq 0}$ is a stationary point of the objective function in (9.39).*

We refer the reader for a proof to Schelldorfer et al. (2011).

Due to the non-convexity of the optimization problem, the CGD Algorithm 5 is not finding a global optimum. However, the non-convexity is only due to the covariance parameters. If q^* is small, we could, in principle, compute a global optimum over all the parameters by using convex optimization for fixed τ, σ^2 and varying these parameters over a (q^*+1)-dimensional grid.[3]

9.3.8 Numerical results

We illustrate the performance of the ℓ_1-penalized maximum likelihood estimator (9.36) (LMMLasso) where we choose the regularization parameter via BIC in (9.37). We compare it with the Lasso and adaptive Lasso for linear models, using BIC for selecting the regularization parameter, as described in Chapter 2; and thus, with the latter two methods, the grouping structure is neglected. We mainly focus here on predicting new observations whose group-membership is known, e.g., so-called within-group prediction.

We consider the following scenario: $N = 25$ groups, $n_i \equiv 6$ for $i = 1, \ldots, N$ observations per group, $q = 3$ random effects and $s_0 = 5$ active fixed effects variables with $\beta = (1, 1.5, 1.2, 1, 2, 0, \ldots, 0)^T$, $\sigma = 1$. The covariates are generated as $(X_{ij})^{(-1)} \sim \mathcal{N}_{p-1}(0, \Sigma)$ with $\Sigma_{rs} = 0.2^{|r-s|}$ for $r, s \in \{1, \ldots, p-1\}$ whereas the first

---

[3] For the case where $q^* = 1$ and $\tau \in \mathbb{R}_{>0}$, the non-convexity arises due to a one-dimensional ratio of the variance parameters.

component of X_{ij} corresponds to an intercept, i.e., $X_{ij}^{(1)} \equiv 1$. We only alter the number of fixed covariates p and the variance parameter τ^2, where $\Gamma_\tau = \tau^2 I_{q \times q}$. The three models considered are

$$\text{M1: } p = 10, \quad \text{M2: } p = 100, \quad \text{M3: } p = 500.$$

For measuring the quality of prediction, we generate a test set with 50 observations per group and calculate the mean squared prediction error. The results are shown in Table 9.1. We see very clearly that with higher degree of grouping structure, i.e.,

Model	τ^2	LMMLasso	Lasso	adaptive Lasso
M1	0	1.01	1.00	1.01
($p = 10$)	0.25	1.33	1.76	1.84
	1	1.66	3.74	3.74
	2	1.67	5.92	6.25
M2	0	1.12	1.26	1.09
($p = 100$)	0.25	1.51	1.75	1.75
	1	1.94	4.35	4.53
	2	2.49	7.04	7.02
M3	0	1.22	1.18	1.26
($p = 500$)	0.25	1.83	2.63	2.67
	1	2.00	4.35	3.78
	2	2.54	10.30	8.26

Table 9.1 Mean squared prediction error for three simulation examples. LMMLasso uses information about which variables have random effects and the structure $\Gamma = \tau^2 I$, whereas Lasso and adaptive Lasso ignore the grouping structure in the data. Regularization parameters are chosen via BIC.

with larger value of τ^2, the prediction is markedly improved by taking the grouping structure into account: that is, Lasso and adaptive Lasso in linear models perform substantially worse than LMMLasso. We do not show here the error when treating all groups as separate datasets: such an approach would also be substantially worse than LMMLasso which borrows strength from other groups via the fixed effects (which are the same across groups).

9.3.8.1 Application: Riboflavin production data

We apply the ℓ_1-penalty procedure for linear mixed effects models, the LMMLasso from (9.36), on real data about riboflavin production with bacillus subtilis. A version of the data-set has been introduced in Section 9.2.6. The response variable is the logarithm of the riboflavin production rate of Bacillus subtilis. Here, there are $p = 4088$ covariates measuring the gene expression levels and $N = 28$ samples (groups) with $n_i \in \{2, \ldots, 6\}$ and $N_T = 111$ observations. Observations in the same group arise from repeated measurements of the same strain of (genetically engineered) bacillus

subtilis while different groups correspond to different strains. We standardize all covariates to have overall mean zero and variance one. For this example, we include an unpenalized fixed effect intercept term (the first fixed effect covariable), that is, the penalty is of the form $\lambda \sum_{j=2}^{4089} |\beta_j|$.

First, we address the issue of determining the covariates which have both a fixed and a random regression coefficient, that is, we have to find the matrix $\mathbf{Z}_i \subset \mathbf{X}_i$. We first use the (ordinary) Lasso for fixed effects only, using 10-fold cross-validation for tuning parameter selection. This yields a first active set $\hat{S}_{\text{init}}$. Then, for each variable $j \in \hat{S}_{\text{init}}$, we fit a mixed effects model where only the jth variable has a random effect, that is, we fit different mixed effects models with a single random effect only. From these models, we obtain estimates of the corresponding variances of random effects:

$$\{\hat{\tau}_j; \ j \in \hat{S}_{\text{init}}\}.$$

We then include random effects for covariates $j \in \hat{S}_{\text{init}}$ where $\hat{\tau}_j > \kappa$ for some threshold κ. Following this strategy (and using $\kappa = 0.05$), it seems reasonable to fit a model where two covariates have an additional random effect. Denoting these variables as k_1 and k_2, the model can be written as

$$Y_{ij} = X_{ij}\beta + Z_{ij,k_1}b_{i,k_1} + Z_{ij,k_2}b_{ik_2} + \varepsilon_{ij}, \ \ i = 1, \ldots, N, \ \ j = 1, \ldots, n_i. \quad (9.42)$$

For the covariance structure, we assume independent random effects with different variances.

We compare the results of LMMLasso and the adaptive LMMLasso with the plain Lasso and plain adaptive Lasso; the latter two methods ignore the grouping structure in the observations. Table 9.2 describes the results. We see that the error variance of

Estimates	LMMLasso	adaptive LMMLasso	Lasso	adaptive Lasso
$\hat{\sigma}^2$	0.18	0.15	0.30	0.20
$\hat{\tau}_{k_1}^2$	0.17	0.08	–	–
$\hat{\tau}_{k_2}^2$	0.03	0.03	–	–
$|\hat{S}|$	18	14	21	20

Table 9.2 Riboflavin production data. Estimates of the error variance, of variances of two random effects and size of the estimated active set. LMMLasso and its adaptive version, in comparison to the Lasso and the adaptive Lasso ignoring the grouping structure in the data.

the Lasso can be considerably reduced using the LMMLasso, and likewise for the corresponding adaptive versions. For the LMMLasso, 53% (= (0.17 + 0.03)/(0.18 + 0.17 + 0.03)) of the total variability is due to the between-group effect. This clearly indicates that there is indeed a substantial variation between the groups. The estimated active sets of LMMLasso or adaptive LMMLasso are a bit smaller than using Lasso or the adaptive Lasso, respectively.

In Section 9.2.6, we used a finite mixture of regressions (FMR) model to account for possible inhomogeneity among different (mixture) components in the riboflavin production data. There, we ignored the fact that we actually have information about grouping of different observations (different groups correspond to different strains of bacillus subtilis). We simply used a blind mixture modeling approach to incorporate inhomogeneity into the model. In contrast, a linear mixed effects model is based on a known grouping structure among the observations. The two modeling approaches should not be compared to each other.

9.4 Theory for ℓ_1-penalization with non-convex negative log-likelihood

We present here some theory for ℓ_1-penalized smooth likelihood problems which are generally non-convex: ℓ_1-penalized likelihood estimation in mixture of regressions models or mixed effects models discussed in the previous sections are then special cases thereof.

9.4.1 The setting and notation

Consider a parametrized family of densities $\{f_\psi;\ \psi \in \Psi\}$ with respect to Lebesgue measure μ on $\mathbb{R}$ (i.e. the range for the response variable). The parameter space Ψ is assumed to be a bounded subset of some finite-dimensional space, say

$$\Psi \subset \{\psi \in \mathbb{R}^d;\ \|\psi\|_\infty \leq K\},$$

where we have equipped (quite arbitrarily) the space $\mathbb{R}^d$ with the sup-norm $\|\psi\|_\infty = \max_{1 \leq j \leq d} |\psi_j|$. In our setup, the dimension d will be regarded as a fixed constant (which still covers high-dimensionality of the covariates, as we will see).

We assume a setting with a covariate $X \in \mathscr{X} \subseteq \mathbb{R}^p$ and a response variable $Y \in \mathbb{R}$. The true conditional density of Y given $X = x$ is assumed to be equal to

$$f_{\psi^0}(\cdot|x) = f_{\psi^0(x)}(\cdot),$$

where

$$\psi^0(x) \in \Psi,\ \forall\, x \in \mathscr{X}.$$

That is, we assume that the true conditional density of Y given x is depending on x only through some parameter function $\psi^0(x)$. Of course, the introduced notation also applies to fixed instead of random covariates.

The parameter $\{\psi^0(x); x \in \mathcal{X}\}$ is assumed to have a nonparametric part of interest $\{g^0(x); x \in \mathcal{X}\}$ and a low-dimensional nuisance part η^0, i.e.,

$$\psi^0(\cdot) = ((g^0(\cdot))^T, (\eta^0)^T)^T,$$

with

$$g^0(x) \in \mathbb{R}^k, \ \forall x \in \mathcal{X}, \ \eta^0 \in \mathbb{R}^m, \ k+m = d.$$

In case of finite mixture of regressions (FMR) models from Section 9.2, $g(x)^T = (\phi_1^T x, \phi_2^T x, \ldots, \phi_k^T x)$ and η involves the parameters $\rho_1, \ldots, \rho_k, \pi_1, \ldots, \pi_{k-1}$. (In previous chapters we used the notation $x\phi_1$, assuming that x is a $1 \times p$ vector). More details are given in Section 9.4.3. For linear mixed effects models from Section 9.3, $g(x) = \beta^T x$ and $\eta = (\tau, \log(\sigma))^T$, see also Section 9.4.4.

With minus the log-likelihood as loss function, the so-called excess risk

$$\mathcal{E}(\psi|\psi^0) = -\int \log\left(\frac{f_\psi(y)}{f_{\psi^0}(y)}\right) f_{\psi^0}(y)\mu(dy)$$

is the Kullback-Leibler information. For fixed covariates $X_1, \ldots, X_n$, we define the average excess risk

$$\bar{\mathcal{E}}(\psi|\psi^0) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{E}\left(\psi(X_i)\Big|\psi^0(X_i)\right),$$

and for random design, we take the expectation $\mathbb{E}(\mathcal{E}(\psi(X)|\psi^0(X)))$.

9.4.1.1 The margin

As in Section 6.4 from Chapter 6, we call the behavior of the excess risk $\mathcal{E}(\psi|\psi^0)$ near ψ^0 the margin. We will show in Lemma 9.1 that the margin is quadratic.

Denote by

$$\ell_\psi(\cdot) = \log f_\psi(\cdot)$$

the log-density. Assuming the derivatives exist, we define the score function

$$s_\psi(\cdot) = \frac{\partial \ell_\psi(\cdot)}{\partial \psi},$$

and the Fisher information

$$I(\psi) = \int s_\psi(y)s_\psi^T(y)f_\psi(y)\mu(dy) = -\int \frac{\partial^2 \ell_\psi(y)}{\partial \psi \partial \psi^T} f_\psi(y)\mu(dy).$$

Of course, we can then also look at $I(\psi(x))$ using the parameter function $\psi(x)$.

In the sequel, we introduce some conditions (Conditions 1 - 5). First, we will assume boundedness of third derivatives.

Condition 1 *It holds that*

$$\sup_{\psi \in \Psi} \max_{(j_1,j_2,j_3) \in \{1,\dots,d\}^3} \left| \frac{\partial^3}{\partial \psi_{j_1} \partial \psi_{j_2} \partial \psi_{j_3}} \ell_\psi(\cdot) \right| \le G_3(\cdot),$$

where

$$\sup_x \int G_3(y) f_{\psi^0}(y|x) d\mu(y) \le C_3 < \infty.$$

For a symmetric, positive semi-definite matrix A, we denote by $\Lambda^2_{\min}(A)$ be its smallest eigenvalue.

Condition 2 *For all x, the Fisher information matrix $I(\psi^0(x))$ is positive definite, and in fact*

$$\Lambda_{\min} = \inf_x \Lambda_{\min}(I(\psi^0(x))) > 0.$$

Furthermore, we will need the following identifiability condition.

Condition 3 *For all $\varepsilon > 0$, there exists an $\alpha_\varepsilon > 0$, such that*

$$\inf_x \inf_{\substack{\psi \in \Psi \\ \|\psi - \psi^0(x)\|_2 > \varepsilon}} \mathscr{E}(\psi | \psi^0(x)) \ge \alpha_\varepsilon.$$

Based on these three conditions we have the following result:

Lemma 9.1. *Assume Conditions 1, 2, and 3. Then*

$$\inf_x \frac{\mathscr{E}(\psi | \psi^0(x))}{\|\psi - \psi^0(x)\|_2^2} \ge \frac{1}{c_0^2},$$

where

$$c_0^2 = \max\left[\frac{1}{\varepsilon_0}, \frac{dK^2}{\alpha_{\varepsilon_0}} \right], \quad \varepsilon_0 = \frac{3\Lambda^2_{\min}}{2d^{3/2}}.$$

A proof is given in Section 9.5.

9.4.1.2 The empirical process

We now specialize to the case where

$$\psi(x)^T = (g_\phi(x)^T, \eta^T),$$

$$g_\phi(x)^T = (g_1(x), \ldots, g_k(x)),$$
$$g_r(x) = g_{\phi_r}(x) = \phi_r^T x, \ x \in \mathbb{R}^p, \ \phi_r \in \mathbb{R}^p, \ r = 1, \ldots, k.$$

Thus, we focus on $g(\cdot)$ functions which are linear in some parameters ϕ. We also write

$$\psi_\vartheta(x)^T = (g_\phi(x)^T, \eta^T), \qquad \vartheta^T = (\phi_1^T, \ldots, \phi_k^T, \eta^T)$$

to make the dependence of the parameter function $\psi(x)$ on ϑ more explicit.

We will assume that

$$\sup_x \|\phi^T x\|_\infty = \sup_x \max_{1 \le r \le k} |\phi_r^T x| \le K.$$

This can be viewed as a combined condition on $\mathscr{X}$ and ϕ. For example, if $\mathscr{X}$ is bounded by a fixed constant, this supremum (for fixed ϕ) is finite.

Our parameter space is now

$$\tilde{\Theta} \subset \{\vartheta = (\phi_1^T, \ldots, \phi_k^T, \eta^T)^T; \ \sup_x \|\phi^T x\|_\infty \le K, \ \|\eta\|_\infty \le K\}. \qquad (9.43)$$

Note that $\tilde{\Theta}$ is in principle $(pk+m)$-dimensional. The true parameter ϑ^0 is assumed to be an element of $\tilde{\Theta}$.

Let us define

$$\ell_\vartheta(x, \cdot) = \log f_{\psi(x)}(\cdot), \ \ \psi(x)^T = \psi_\vartheta(x)^T = (g_\phi(x)^T, \eta^T), \ \vartheta^T = (\phi_1^T, \ldots, \phi_k^T, \eta^T),$$

and consider the empirical process for fixed covariates $X_1, \ldots, X_n$:

$$V_n(\vartheta) = \frac{1}{n} \sum_{i=1}^n \left(\ell_\vartheta(X_i, Y_i) - \mathbb{E}[\ell_\vartheta(X_i, Y)|X_i] \right).$$

We now fix some $T \ge 1$ and $\lambda_0 \ge 0$ and define the set

$$\mathscr{T} = \left\{ \sup_{\vartheta = (\phi^T, \eta^T)^T \in \tilde{\Theta}} \frac{\left| V_n(\vartheta) - V_n(\vartheta^0) \right|}{(\|\phi - \phi^0\|_1 + \|\eta - \eta^0\|_2) \vee \lambda_0} \le T\lambda_0 \right\}. \qquad (9.44)$$

9.4.2 Oracle inequality for the Lasso for non-convex loss functions

For an optimality result, we need some condition on the design. Denote the active set, i.e., the set of non-zero coefficients, by

$$S_0 = \{(r,j);\ \phi_{r,j}^0 \neq 0\},\ s_0 = |S_0|,$$

and let

$$\phi_S = \{\phi_{r,j};\ (r,j) \in S\},\ S \subseteq \{1,\ldots,k\} \times \{1,\ldots,p\}.$$

Furthermore, let

$$\hat{\Sigma}_X = \frac{1}{n} \sum_{i=1}^n X_i^T X_i,$$

(where here, we denote by $X_i = (X_i^{(1)},\ldots,X_i^{(p)})$ a $(1 \times p)$-vector).

Condition 4 *(Restricted eigenvalue condition; see also Section 6.13.7). There exists a constant $\kappa \geq 1$, such that for all $\phi \in \mathbb{R}^{pk}$ satisfying*

$$\|\phi_{S_0^c}\|_1 \leq 6\|\phi_{S_0}\|_1,$$

it holds that

$$\|\phi_{S_0}\|_2^2 \leq \kappa^2 \sum_{r=1}^k \phi_r^T \hat{\Sigma}_X \phi_r.$$

For $\psi(\cdot)^T = (g(\cdot)^T, \eta^T)$, we use the notation

$$\|\psi\|_{Q_n}^2 = \frac{1}{n} \sum_{i=1}^n \sum_{r=1}^k g_r^2(X_i) + \sum_{j=1}^m \eta_j^2.$$

We also write for $g(\cdot) = (g_1(\cdot),\ldots,g_k(\cdot))^T$,

$$\|g\|_{Q_n}^2 = \frac{1}{n} \sum_{i=1}^n \sum_{r=1}^k g_r^2(X_i).$$

Thus

$$\|g_\phi\|_{Q_n}^2 = \sum_{r=1}^k \phi_r^T \hat{\Sigma}_X \phi_r,$$

and the bound in the restricted eigenvalue condition then reads

$$\|\phi_{S_0}\|_2^2 \leq \kappa^2 \|g_\phi\|_{Q_n}^2.$$

We employ the ℓ_1-penalized estimator

$$\hat{\vartheta}(\lambda) = (\hat{\phi}^T(\lambda), \hat{\eta}^T(\lambda))^T$$

$$= \underset{\vartheta^T = (\phi^T, \eta^T) \in \tilde{\Theta}}{\arg\min} \left\{ -\frac{1}{n} \sum_{i=1}^n \ell_\vartheta(X_i, Y_i) + \lambda \sum_{r=1}^k \|\phi_r\|_1 \right\}. \tag{9.45}$$

We omit in the sequel the dependence of $\hat{\vartheta}$ on λ. Note that we consider here a global minimizer: it may be difficult to compute if the empirical risk $-n^{-1} \sum_{i=1}^n \ell_\vartheta(X_i, Y_i)$

is non-convex in ϑ. We write $\|\phi\|_1 = \sum_{r=1}^{k} \|\phi_r\|_1$ and denote by

$$\hat{\psi}(x) = (g_{\hat{\phi}}(x)^T, \hat{\eta}^T)^T.$$

Theorem 9.1. *(Oracle result for fixed design). Assume fixed covariates $X_1, \ldots, X_n$, Conditions 1-3 and 4, and that $\lambda \geq 2T\lambda_0$ for the estimator in (9.45) with T and λ_0 as in (9.44). Then on $\mathscr{T}$, defined in (9.44), for the average excess risk (average Kullback-Leibler loss),*

$$\bar{\mathscr{E}}(\hat{\psi}|\psi^0) + 2(\lambda - T\lambda_0)\|\hat{\phi} - \phi^0\|_1 \leq 9(\lambda + T\lambda_0)^2 c_0^2 \kappa^2 s_0,$$

where c_0 and κ are defined in Lemma 9.1 and Condition 4, respectively.

A proof is given in Section 9.5. The oracle inequality of Theorem 9.1 has the usual interpretation, see Chapter 6. The rate for λ_0 is

$$\lambda_0 \asymp M_n \log(n) \sqrt{\frac{\log(p \vee n)}{n}},$$

as described by the definition in (9.46), where the rate for M_n is depending on the model of interest. For example, for finite mixture of regressions (FMR) models, the rate is $M_n \asymp \sqrt{\log(n)}$ as used in Lemma 9.2 and 9.3 below. On the other hand, for linear mixed effects models, we have $M_n \asymp \log(n)$ as indicated after formula (9.53). We then obtain

$$\bar{\mathscr{E}}(\hat{\psi}(\lambda)|\psi^0) \leq 9(\lambda + T\lambda_0)^2 c_0^2 \kappa^2 s_0 = O(\kappa^2 s_0 M_n^2 \log(n)^2 \log(p \vee n)/n),$$

saying that the average Kullback-Leibler (excess) risk achieves the optimal convergence rate, up to the factor $\log(n)^2 M_n^2 \log(p \vee n)$ as if one knew the s_0 non-zero coefficients.

As a second implication, we obtain an estimation error bound

$$\|\hat{\phi} - \phi^0\|_1 \leq 9(\lambda + T\lambda_0) c_0^2 \kappa^2 s_0/2.$$

From such a result, and assuming a beta-min condition requiring that the non-zero coefficients are sufficiently large, one can derive a variable screening result (see Section 2.5, Corollary 7.6 and Section 7.8.5): on $\mathscr{T}$,

$$\hat{S} = \{(r, j); \ \hat{\phi}_{r,j} \neq 0\} \supseteq S_0.$$

We will show in Section 9.4.3, for finite mixture of regressions (FMR) models, that the probability of the set $\mathscr{T}$ is large, as established by Corollary 9.1. The proof relies on more general results, here formulated as Lemma 9.2. We make the following assumption.

Condition 5 *For the score function $s_\vartheta(\cdot) = s_{\psi_\vartheta}(\cdot)$ we have:*

$$\sup_{\vartheta \in \bar{\Theta}} \|s_\vartheta(\cdot)\|_\infty \leq G_1(\cdot).$$

Condition 5 primarily has notational character. Later, in Lemma 9.2 and particularly in Lemma 9.3, the function $G_1(\cdot)$ needs to be sufficiently regular.

Define

$$\lambda_0 = M_n \log n \sqrt{\frac{\log(p \vee n)}{n}}, \tag{9.46}$$

(often, $M_n \asymp \sqrt{\log(n)}$ or $M_n \asymp \log(n)$, depending on the model under consideration). Let $\mathbf{P_x}$ denote the probability for a fixed design $\mathbf{x} = (X_1, \ldots, X_n)$. With the expression $1\{\cdot\}$ we denote the indicator function.

Lemma 9.2. *Assume Condition 5. Then, for constants c_1, c_2 and c_3 depending on k and K, and for all $T \geq 1$,*

$$\sup_{\vartheta^T = (\phi^T, \eta^T) \in \bar{\Theta}} \frac{\left| V_n(\vartheta) - V_n(\vartheta^0) \right|}{(\|\phi - \phi^0\|_1 + \|\eta - \eta^0\|_2) \vee \lambda_0} \leq c_1 T \lambda_0,$$

with $\mathbf{P_x}$ probability at least

$$1 - c_2 \exp\left[-\frac{T^2 \log(n)^2 n \log(p \vee n)}{c_3^2} \right] - \mathbf{P_x}\left(\frac{1}{n} \sum_{i=1}^n F(Y_i) > T \lambda_0^2/(dK) \right),$$

where (for $i = 1, \ldots, n$)

$$F(Y_i) = G(Y_i) 1\{G_1(Y_i) > M_n\} + \mathbb{E}\left[G_1(Y) 1\{G_1(Y) > M_n\} \Big| X = x_i \right].$$

Regarding the constants λ_0 and K, see (9.46) and (9.43), respectively.

A proof is given in Section 9.5.

9.4.3 Theory for finite mixture of regressions models

In the finite mixture of regressions (FMR) model from (9.2) with k components, the parameter is $\vartheta^T = (\phi^T, \eta^T) = (\phi_1^T, \ldots, \phi_k^T, \log \rho_1, \ldots, \log \rho_k, \log \pi_1, \ldots, \log \pi_{k-1})$, where $\rho_r = \sigma_r^{-1}$ is the inverse standard deviation in mixture component r and the

π_r's are the mixture probabilities. For mathematical convenience and simpler notation, we consider here the log-transformed ρ and π parameters in order to have lower and upper bounds for ρ and π. Obviously, there is a one-to-one correspondence between ϑ and θ from Section 9.2.1.1.

Let the parameter space be

$$\tilde{\Theta} \subset \{\vartheta; \ \sup_x \|\phi^T x\|_\infty \leq K, \|\log\rho\|_\infty \leq K, -K \leq \log\pi_1 \leq 0, \ldots, -K \leq \log\pi_{k-1} \leq 0,$$

$$\sum_{r=1}^{k-1} \pi_r < 1\}, \tag{9.47}$$

and $\pi_k = 1 - \sum_{r=1}^{k-1} \pi_r$.

As in (9.45), we consider

$$\hat{\vartheta}(\lambda) = \underset{\vartheta \in \tilde{\Theta}}{\arg\min} \quad -n^{-1} \sum_{i=1}^n \log\left(\sum_{r=1}^k \pi_r \frac{\rho_r}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(\rho_r Y_i - X_i \phi_r)^2\right)\right)$$

$$+ \lambda \sum_{r=1}^k \|\phi_r\|_1. \tag{9.48}$$

This is the estimator from Section 9.2.2.2 with $\gamma = 0$. We emphasize the boundedness of the parameter space by using the notation $\tilde{\Theta}$. We focus here on any global minimizer of the penalized negative log-likelihood which is very difficult (or virtually impossible) to compute.

In the following, when writing $\hat{\theta}(\lambda)$, we mean the estimator transformed from $\hat{\vartheta}(\lambda)$ to $\hat{\theta}(\lambda)$ in the parametrization θ from Section 9.2.1.1. As before, we denote the average excess risk by $\bar{\mathscr{E}}(\hat{\theta}(\lambda)|\theta^0)$.

9.4.3.1 Oracle result for FMR models

We specialize now our results from Section 9.4.2 to FMR models.

Proposition 9.4. *For fixed design FMR models as in (9.2) with $\tilde{\Theta}$ in (9.47), Conditions 1,2 and 3 are met, for appropriate C_3, $\Lambda_{\min}$ and $\{\alpha_\varepsilon\}$, depending on k and K. Also Condition 5 holds, with*

$$G_1(y) = e^K |y| + K.$$

Proposition 9.4 follows from straightforward calculations and we leave the derivation as Problem 9.7.

In order to show that the probability for the set $\mathscr{T}$ is large, we invoke Lemma 9.2 and the following result.

Lemma 9.3. *For fixed design FMR models as in (9.2) with $\tilde{\Theta}$ in (9.47): for some constants c_4, c_5 and c_6, depending on k, and K, and for $M_n = c_4\sqrt{\log n}$ and $n \geq c_6$, the following holds:*

$$\mathbf{P}_{\mathbf{x}}\left(\frac{1}{n}\sum_{i=1}^n F(Y_i) > c_5\frac{\log n}{n}\right) \leq \frac{1}{n},$$

where, for $i = 1,\ldots,n$,

$$F(Y_i) = G_1(Y_i)\mathrm{l}\{G_1(Y_i) > M_n\} + \mathbb{E}[G_1(Y)\mathrm{l}\{G_1(Y) > M_n\}|X = X_i],$$

and $G_1(\cdot)$ is as in Proposition 9.4.

A proof is given in Section 9.5.

Hence, the oracle result in Theorem 9.1 for the ℓ_1-penalized estimator in the FMR model holds on a set $\mathscr{T}$, summarized in Theorem 9.2, and this set $\mathscr{T}$ has large probability due to Lemma 9.2 and Lemma 9.3 as described in the following corollary.

Corollary 9.1. *For fixed design FMR models as in (9.2) with $\tilde{\Theta}$ in (9.47), we have for constants c_2, c_4, c_7, c_8 depending on k, and K,*

$$\mathbf{P}[\mathscr{T}] \geq 1 - c_2\exp\left[-\frac{T^2\log(n)^2\log(p \vee n)}{c_7^2}\right] - n^{-1} \text{ for all } n \geq c_8,$$

where $\mathscr{T}$ is defined with $\lambda_0 = M_n\log(n)\sqrt{\log(p \vee n)/n}$ and $M_n = c_4\sqrt{\log(n)}$.

Theorem 9.2. *(Oracle result for FMR models). Consider a fixed design FMR model as in (9.2) with $\tilde{\Theta}$ in (9.47). Assume Condition 4 (restricted eigenvalue condition) and that $\lambda \geq 2T\lambda_0$ for the estimator in (9.48). Then on $\mathscr{T}$, which has large probability as stated in Corollary 9.1, for the average excess risk (average Kullback-Leibler loss),*

$$\bar{\mathscr{E}}(\hat{\theta}(\lambda)|\theta^0) + 2(\lambda - T\lambda_0)\|\hat{\phi}_{S_0^c}\|_1 \leq 9(\lambda + T\lambda_0)^2 c_0^2\kappa^2 s_0,$$

where c_0 and κ are defined in Lemma 9.1 and Condition 4, respectively.

Note that the Conditions 1, 2, 3 and 5 hold automatically for FMR models, as described in Proposition 9.4. We still require a restricted eigenvalue condition on the design, here Condition 4.

The interpretation of Theorem 9.2 is as follows. One can choose $\lambda = 2T\lambda_0 \asymp \sqrt{\log(n)^3\log(p \vee n)/n}$, using λ_0 as in Corollary 9.1. For the average excess risk (Kullback-Leibler divergence) we have a convergence rate of

$$\bar{\mathscr{E}}(\hat{\theta}(\lambda)|\theta_0) = O_P\left(\kappa^2 s_0\log(n)^3\log(p \vee n)/n\right)$$

which is up to the log-factors the rate if one knew which of the s_0 variables were active. Furthermore, the oracle inequality implies

$$\|\hat{\phi} - \phi^0\|_1 = O_P\left(\kappa^2 s_0 \sqrt{\log(n)^3 \log(p \vee n)/n}\right).$$

This allows to derive a variable screening property analogous to Section 2.5 (e.g. formula (2.13)), and we also refer to Corollary 7.6. If the non-zero coefficients of ϕ^0 are sufficiently large,

$$\min_{(r,j)\in S_0} |\phi^0_{r,j}| \gg O\left(\kappa^2 s_0 \sqrt{\log(n)^3 \log(p \vee n)/n}\right),$$

(the analogue of the beta-min condition in formula (2.23)) then, with high probability,

$$\hat{S} = \{(r,j); \ \hat{\phi}_{r,j} \neq 0, \ r = 1,\ldots,k, \ j = 1,\ldots,p\} \supseteq S_0$$

using the notation as in (9.19).

Without Condition 4 about restricted eigenvalues, one can still derive a high-dimensional consistency result:

$$\bar{\mathscr{E}}(\hat{\theta}(\lambda)|\theta^0) = o_P(1) \ (n \to \infty), \tag{9.49}$$

requiring $\|\phi^0\|_1 = \sum_{r=1}^k \|\phi^0_r\|_1 = o(\sqrt{n/(\log(n)^3 \log(p \vee n))}) \ (n \to \infty)$ and choosing $\lambda = C\sqrt{\log(n)^3 \log(p \vee n)/n}$ for some $C > 0$ sufficiently large. We leave the derivation of (9.49) as Problem 9.8.

9.4.4 Theory for linear mixed effects models

We can establish an oracle inequality for linear mixed effects models using the general theory presented in Sections 9.4.1-9.4.2.

As before, consider the group index $i = 1,\ldots,N$ and let $n_i \equiv n$ denote the number of observations within a group. Furthermore, denote by $\mathbf{Y}_i \in \mathscr{Y} \subset \mathbb{R}^n$ the response variable, $\mathbf{X}_i$ the fixed covariates in some space $\mathscr{X}^n \subset \mathbb{R}^{n \times p}$ and $\mathbf{Z}_i \subset \mathbf{X}_i$. Define the parameter $\theta^T = (\beta^T, \tau^T, \log(\sigma)) = (\beta^T, \eta^T) \in \mathbb{R}^{p+q^*+1}$ as in Section 9.3.7 and denote by θ^0 the true parameter vector. For a constant $0 < K < \infty$, define the parameter space to be

$$\tilde{\Theta} = \{\theta = (\beta^T, \eta^T)^T; \sup_{\mathbf{x} \in \mathscr{X}} |\beta^T x| \leq K, \|\eta\|_\infty \leq K, \Gamma_{\eta_1} \text{ positive definite}\}$$
$$\subset \mathbb{R}^{p+q^*+1}, \tag{9.50}$$

where $\|\eta\|_\infty = \max_{l=1,\ldots,q^*+1} |\eta_l|$ and Γ_{η_1} is as in (9.39). We modify the estimator in (9.36) by restricting the solution to be in the compact parameter space $\tilde{\Theta}$:

$$\hat{\theta}(\lambda) = \underset{\theta \in \tilde{\Theta}}{\arg\min} \left(\frac{1}{2} \sum_{i=1}^{N} (\log(\det(V_i)) + (\mathbf{Y}_i - \mathbf{X}_i\beta)^T V_i^{-1} (\mathbf{Y}_i - \mathbf{X}_i\beta)) + \lambda \|\beta\|_1 \right).$$

$$(9.51)$$

Denote by $f_{\theta, \mathbf{X}_i, \mathbf{Z}_i}$ ($\theta \in \Theta$) the Gaussian density for $\mathbf{Y}_i$ with respect to the above parametrization. The excess risk is

$$\mathscr{E}_{\mathbf{X}, \mathbf{Z}}(\theta | \theta^0) = \int \log \left(\frac{f_{\theta^0, \mathbf{X}, \mathbf{Z}}(y)}{f_{\theta, \mathbf{X}, \mathbf{Z}}(y)} \right) f_{\theta^0, \mathbf{X}, \mathbf{Z}}(y) \mu(dy), \qquad (9.52)$$

and we define the average excess risk as

$$\overline{\mathscr{E}}_{\mathbf{X}_1, \dots, \mathbf{X}_N, \mathbf{Z}_1, \dots, \mathbf{Z}_N}(\theta | \theta^0) = \frac{1}{N} \sum_{i=1}^{N} \mathscr{E}_{\mathbf{X}_i, \mathbf{Z}_i}(\theta | \theta^0).$$

We now state some assumptions.

(A1)   The eigenvalues of $\mathbf{Z}_i^T \mathbf{Z}_i$, denoted by $(v_j^{(i)})_{j=1}^{q}$ ($i = 1, \dots, N$), are bounded: $v_j^{(i)} \leq K < \infty$ for all i and j, with K from (9.50).

(A2)

(a)   Let $(\omega_j^{(i)})_{j=1}^{n}$ be the eigenvalues of $\mathbf{Z}_i \Gamma_{\tau^0} \mathbf{Z}_i^T$ for $i = 1, \dots, N$. At least two eigenvalues are different, i.e., for all i there exist $j_1 \neq j_2 \in \{1, \dots, n\}$ such that $\omega_{j_1}^{(i)} \neq \omega_{j_2}^{(i)}$.

(b)   For $i = 1, \dots, N$, the matrices Ω_i defined by

$$(\Omega_i)_{r,s} = \text{trace} \left(V_i^{-1} \frac{\partial V_i}{\partial \theta_{p+r}} V_i^{-1} \frac{\partial V_i}{\partial \theta_{p+s}} \right) \Bigg|_{\theta = \theta^0}, \quad r, s = 1, \dots, q^* + 1$$

are strictly positive definite.

(A3)   There exists a constant $\kappa \geq 1$, such that for all $\beta \in \mathbb{R}^p$ satisfying

$$\|\beta_{S_0^c}\|_1 \leq 6\|\beta_{S_0}\|_1$$

it holds that

$$\|\beta_{S_0}\|_2^2 \leq \kappa^2 \beta^T \hat{\Sigma}_{X,n} \beta,$$

where $\hat{\Sigma}_{X,n} = (Nn)^{-1} \sum_{i=1}^{N} \sum_{j=1}^{n} X_{ij}^T X_{ij}$ with $(1 \times p)$-vector $X_{ij} = (X_{ij}^{(1)}, \dots, X_{ij}^{(p)})$, $S_0 = \{j; \beta_j^0 \neq 0\}$ and $s_0 = |S_0|$.

Assumption (A2)(a) automatically holds if the covariance matrix of the random effects is of the form $\Gamma_\tau = \tau^2 I_{q \times q}$ (Problem 9.6). Assumption (A3) is a restricted

eigenvalue assumption as in Condition 4 in Section 9.4.2. Following the general theory in Section 9.4, we outline now an oracle inequality for linear mixed effects models. Define

$$\lambda_0 = M_N \log(N) \sqrt{\frac{\log(p \vee N)}{N}} \quad , \tag{9.53}$$

where the constant $M_N \asymp \log(N)$. For any $T \geq 1$, let $\mathscr{T}$ be a set defined by the underlying empirical process. It can be shown that this set $\mathscr{T}$ has large probability, thereby using assumption (A1).

Proposition 9.5. *Consider the estimator (9.51). Under Assumptions (A1)-(A3), and for $\lambda \geq 2T\lambda_0$, then, on $\mathscr{T}$, for the average excess risk,*

$$\bar{\mathscr{E}}(\hat{\theta}(\lambda)|\theta_0) + 2(\lambda - T\lambda_0)\|\hat{\beta} - \beta^0\|_1 \leq 9(\lambda + T\lambda_0)^2 c_0^2 \kappa^2 s_0$$

for a constant c_0 (which is independent of N, n, p and the design).

We do not provide a proof but point out that the result can be established using the theory from Section 9.4. The details are given in Schelldorfer et al. (2011). We remark again that assumption (A1) is used to show that $\mathbf{P}[\mathscr{T}]$ is large, but we also need it to ensure quadratic behavior of the margin (see also Lemma 9.1).

The interpretation of Proposition 9.5 is again as for FMR models discussed after Theorem 9.2 in Section 9.4.3.1. Choose $\lambda = 2T\lambda_0 \asymp \sqrt{\log(N)^4 \log(p \vee N)/N}$, using (9.53) with $M_N \asymp \log(N)$, and thus, for the average excess risk (Kullback-Leibler divergence), the estimator has convergence rate

$$\bar{\mathscr{E}}(\hat{\theta}(\lambda)|\theta_0) = O_P\left(\kappa^2 s_0 \log(N)^4 \log(p \vee N)/N\right)$$

which is up to the log-factors the rate if one knew which of the s_0 variables were active. The oracle inequality also implies an ℓ_1-norm estimation error bound:

$$\|\hat{\beta} - \beta^0\|_1 = O_P\left(\kappa^2 s_0 \sqrt{\log(N)^4 \log(p \vee N)/N}\right).$$

Using this, we can then derive a variable screening property as in Section 2.5, see also Corollary 7.6: if the non-zero coefficients of β^0 are sufficiently large,

$$\min_{j \in S_0} |\beta_j^0| \gg O\left(\kappa^2 s_0 \sqrt{\log(N)^4 \log(p \vee N)/N}\right)$$

(the analogue of the beta-min condition in formula 2.23) then, with high probability,

$$\hat{S} = \{j; \hat{\beta}_j \neq 0, \ j = 1, \ldots, p\} \supseteq S_0.$$

An oracle inequality and an ℓ_1-estimation error bound can be established for the adaptive LMMLasso estimator in (9.38) as well. Assuming a beta-min condition, such a result is given in Schelldorfer et al. (2011).

9.5 Proofs for Section 9.4

9.5.1 Proof of Lemma 9.1

It is clear that

$$\mathscr{E}(\psi|\psi^0) = (\psi - \psi^0)^T I(\psi^0)(\psi - \psi^0)/2 + r_\psi,$$

where

$$|r_\psi| \leq \frac{\|\psi - \psi^0\|_1^3}{6} \int \sup_{\psi \in \Psi} \max_{j_1,j_2,j_3} \left| \frac{\partial^3 l_\psi}{\partial \psi_{j_1} \partial \psi_{j_2} \partial \psi_{j_3}} \right| f_{\psi^0} d\mu$$

$$\leq \frac{d^{3/2} C_3}{6} \|\psi - \psi^0\|_2^3.$$

Hence

$$\mathscr{E}(\psi|\psi^0(x)) \geq \|\psi - \psi^0(x)\|_2^2 \Lambda_{\min}^2 / 2 - d^{3/2} C_3 \|\psi - \psi^0(x)\|_2^3 / 6.$$

Now, apply the auxiliary lemma below, with $K_0^2 = dK^2$, $\Lambda^2 = \Lambda_{\min}^2/2$, and $C = d^{3/2} C_3 / 6$.   □

Auxiliary Lemma. *Let* $h : [-K_0, K_0] \to [0, \infty)$ *have the following properties:*

(i) $\forall \, \varepsilon > 0 \, \exists \, \alpha_\varepsilon > 0$ *such that* $\inf_{\varepsilon < |z| \leq K_0} h(z) \geq \alpha_\varepsilon$,

(ii) $\exists \, \Lambda > 0$, $C > 0$, *such that* $\forall \, |z| \leq K_0$, $h(z) \geq \Lambda^2 z^2 - C|z|^3$.

Then $\forall \, |z| \leq K_0$,

$$h(z) \geq z^2 / C_0^2,$$

where

$$C_0^2 = \max \left[\frac{1}{\varepsilon_0}, \frac{K_0^2}{\alpha_{\varepsilon_0}} \right], \quad \varepsilon_0 = \frac{\Lambda^2}{2C}.$$

Proof (Auxiliary Lemma).
If $\varepsilon_0 > K_0$, we have $h(z) \geq \Lambda^2 z^2 / 2$ for all $|z| \leq K_0$.

If $\varepsilon_0 \leq K_0$ and $|z| \leq \varepsilon_0$, we also have $h(z) \geq (\Lambda^2 - \varepsilon_0 C) z^2 \geq \Lambda^2 z^2 / 2$.

If $\varepsilon_0 \leq K_0$ and $\varepsilon_0 < |z| \leq K_0$, we have $h(z) \geq \alpha_{\varepsilon_0} = K_0^2 \alpha_{\varepsilon_0} / K_0^2 \geq |z|^2 \alpha_{\varepsilon_0} / K_0^2$.   □

9.5.2 Proof of Lemma 9.2

In order to prove Lemma 9.2, we first state and proof a suitable entropy bound:

We introduce the norm

$$\|h(\cdot,\cdot)\|_{P_n} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}h^2(x_i,Y_i)}.$$

For a collection $\mathscr{H}$ of functions on $\mathscr{X}\times\mathscr{Y}$, we let $H(\cdot,\mathscr{H},\|\cdot\|_{P_n})$ be the entropy of $\mathscr{H}$ equipped with the metric induced by the norm $\|\cdot\|_{P_n}$ (for a definition of the entropy of a metric space see Section 14.12 in Chapter 14).

Define for $\varepsilon > 0$,

$$\tilde{\Theta}(\varepsilon) = \{\vartheta^T = (\phi_1^T,\ldots,\phi_k^T,\eta^T)\in\tilde{\Theta}: \|\phi-\phi_0\|_1 + \|\eta-\eta_0\|_2 \le \varepsilon\}.$$

Entropy Lemma. *For a constant C_0 depending on k and m (the dimensions of the parts of the parameter ψ), we have for all $u > 0$ and $M_n > 0$,*

$$H\left(u,\left\{(\ell_\vartheta - \ell_{\vartheta^*})1\{G_1 \le M_n\}: \vartheta\in\tilde{\Theta}(\varepsilon)\right\},\|\cdot\|_{P_n}\right) \le C_0\frac{\varepsilon^2 M_n^2}{u^2}\log\left(\frac{\varepsilon M_n}{u}\right).$$

Proof (Entropy Lemma).
We have

$$|\ell_\vartheta(x,y) - \ell_{\tilde{\vartheta}}(x,y)|^2 \le G_1^2(y)\left[\sum_{r=1}^{k}|(\phi_r - \tilde{\phi}_r)^T x| + \|\eta - \tilde{\eta}\|_1\right]^2$$

$$\le dG_1^2(y)\left[\sum_{r=1}^{k}|(\phi_r - \tilde{\phi}_r)^T x|^2 + \|\eta - \tilde{\eta}\|_2^2\right].$$

It follows that

$$\|(\ell_\vartheta - \ell_{\tilde{\vartheta}})1\{G_1 \le M_n\}\|_{P_n}^2 \le dM_n^2\left[\sum_{r=1}^{k}\frac{1}{n}\sum_{i=1}^{n}|(\phi_r - \tilde{\phi}_r)^T x_i|^2 + \|\eta - \tilde{\eta}\|_2^2\right].$$

Let $N(\cdot,\Lambda,d)$ denote the covering number of a metric space (Λ,d) with metric (induced by the norm) d, and $H(\cdot,\Lambda,d) = \log N(\cdot,\Lambda,d)$ be its entropy (for a definition of the covering number of a metric space see Section 14.12 in Chapter 14. If Λ is a ball with radius ε in Euclidean space $\mathbb{R}^N$, one has by Lemma 14.27,

$$H(u,\Lambda,d) \le N\log\left(\frac{3\varepsilon}{u}\right), \forall u > 0.$$

Thus $H(u, \{\eta \in \mathbb{R}^m : \|\eta - \eta_0\|_2 \leq \varepsilon\}, \|\cdot\|_2) \leq m \log\left(\frac{5\varepsilon}{u}\right)$, $\forall u > 0$. Moreover, applying a bound as in Lemma 2.6.11 of van der Vaart and Wellner (1996), see Lemma 14.29 in Chapter 14, gives

$$H\left(2u, \left\{\sum_{r=1}^{k}(\phi_r - \phi_{0,r})^T x_r : \|\phi - \phi_0\|_1 \leq \varepsilon\right\}, \|\cdot\|_{P_n}\right) \leq \left(\frac{\varepsilon^2}{u^2} + 1\right) \log(1 + kp).$$

We can therefore conclude that

$$H\left(3\sqrt{d}M_n u, \left\{(\ell_\vartheta - \ell_{\vartheta_0}) 1\{G_1 \leq M_n\} : \vartheta \in \tilde{\Theta}(\varepsilon)\right\}, \|\cdot\|_{P_n}\right)$$
$$\leq \left(\frac{\varepsilon^2}{u^2} + m + 1\right)\left(\log\left(\frac{3\varepsilon}{u}\right) + \log(1 + kp)\right).$$

$\square$

Let us now turn to the main proof of Lemma 9.2.

In what follows, $\{c_t\}$ are constants depending on k, m and K. The truncated version of the empirical process is

$$V_n^{\text{trunc}}(\vartheta) = \frac{1}{n}\sum_{i=1}^{n}\left(\ell_\vartheta(x_i, Y_i)1\{G_1(Y_i) \leq M_n\} - \mathbb{E}\left[\ell_\vartheta(x_i, Y)1\{G_1(Y) \leq M_n\}\Big| X = x_i\right]\right).$$

Let $\varepsilon > 0$ be arbitrary. We apply Corollary 14.4 (Section 14.7) to the class

$$\left\{(\ell_\vartheta - \ell_{\vartheta_0})1\{G_1 \leq M_n\} : \vartheta \in \tilde{\Theta}(\varepsilon)\right\}.$$

The result (14.10) then gives

$$\mathbf{P_x}\left(\sup_{\vartheta \in \tilde{\Theta}(\varepsilon)} |V_n^{\text{trunc}}(\vartheta) - V_n^{\text{trunc}}(\vartheta_0)| \geq c_6 \varepsilon T M_n \log n \sqrt{\frac{\log(p \vee n)}{n}}\right)$$
$$\leq c_7 \exp\left[-\frac{T^2 \log^2 n \log(p \vee n)(\varepsilon^2 \vee 1)}{c_8^2}\right].$$

We then invoke the peeling device (see van de Geer (2000)): split the set $\tilde{\Theta}$ into sets

$$\{\vartheta \in \tilde{\Theta} : 2^{-(j+1)} \leq \|\phi - \phi_0\|_1 + \|\eta - \eta_0\|_2 \leq 2^{-j}\},$$

where $j \in \mathbb{Z}$, and $\leq 2^{-j+1} \geq \lambda_0$. There are no more than $c_9 \log n$ indices $j \leq 0$ with $2^{-j+1} \geq \lambda_0$. Hence, we get

$$\sup_{\vartheta^T = (\phi^T, \eta^T) \in \tilde{\Theta}} \frac{\left|V_n^{\text{trunc}}(\vartheta) - V_n^{\text{trunc}}(\vartheta_0)\right|}{(\|\phi - \phi^*\|_1 + \|\eta - \eta^*\|_2) \vee \lambda_0} \leq 2c_6 T M_n \log n \sqrt{\frac{\log(p \vee n)}{n}},$$

with $\mathbf{P_x}$ probability at least

$$1 - c_7 [c_9 \log n] \exp \left[-\frac{T^2 \log^2 n \log(p \vee n)}{c_8^2} \right]$$

$$\geq 1 - c_2 \exp \left[-\frac{T^2 \log^2 n \log(p \vee n)}{c_{10}^2} \right].$$

Finally, to remove the truncation, we use

$$|(\ell_\vartheta(x,y) - \ell_{\vartheta_0}(x,y))\mathrm{l}\{G_1(y) > M_n\}| \leq dKG_1(y)\mathrm{l}\{G_1(y) > M_n\}.$$

Hence

$$\frac{\left| (V_n^{\text{trunc}}(\vartheta) - V_n^{\text{trunc}}(\vartheta_0)) - (V_n(\vartheta) - V_n(\vartheta_0)) \right|}{(\|\phi - \phi^*\|_1 + \|\eta - \eta^*\|_2) \vee \lambda_0}$$
$$\leq \frac{dK}{n\lambda_0} \sum_{i=1}^n \left(G_1(Y_i)\mathrm{l}\{G_1(Y_i) > M_n\} + \mathbb{E}\left[G_1(Y)\mathrm{l}\{G_1(Y) > M_n\} \Big| X = x_i \right] \right).$$

$\square$

9.5.3 Proof of Theorem 9.1

Using the definition of $\hat{\psi}$, and on $\mathscr{T}$ defined in (9.44), we have the basic inequality

$$\bar{\mathscr{E}}(\hat{\psi}|\psi^0) + \lambda \|\hat{\phi}\|_1 \leq T\lambda_0 \left[(\|\hat{\phi} - \phi^0\|_1 + \|\hat{\eta} - \eta^0\|_2) \vee \lambda_0 \right] + \lambda \|\phi^0\|_1.$$

By Lemma 9.1,
$$\bar{\mathscr{E}}(\hat{\psi}|\psi^0) \geq \|\hat{\psi} - \psi^0\|_{Q_n}^2 / c_0^2.$$

Case 1 Suppose that
$$\|\hat{\phi} - \phi^0\|_1 + \|\hat{\eta} - \eta^0\|_2 \leq \lambda_0.$$

Then we find, using the triangle inequality,
$$\bar{\mathscr{E}}(\hat{\psi}|\psi^0) \leq T\lambda_0^2 + \lambda \|\hat{\phi} - \phi^0\|_1,$$

and hence

$$\bar{\mathscr{E}}(\hat{\psi}|\psi^0) + 2\lambda \|\hat{\phi} - \phi^0\|_1 \leq T\lambda_0^2 + 3\lambda \|\hat{\phi} - \phi^0\|_1$$
$$\leq (3\lambda + T\lambda_0)\lambda_0.$$

Case 2 Suppose that
$$\|\hat{\phi} - \phi^0\|_1 + \|\hat{\eta} - \eta^0\|_2 \geq \lambda_0,$$

and that
$$T\lambda_0\|\hat{\eta} - \eta^0\|_2 \geq (\lambda + T\lambda_0)\|\hat{\phi}_{S_0} - (\phi^0)_{S_0}\|_1.$$

Then we get, using now the triangle inequality $\|\phi^0\|_1 - \|\hat{\phi}_{S_0}\|_1 \leq \|\hat{\phi}_{S_0} - (\phi^0)_{S_0}\|_1$, and adding $(\lambda - \lambda_0 T)\|\hat{\phi}_{S_0} - (\phi^0)_{S_0}\|_1$ to left- and right-hand side,

$$\begin{aligned}
\bar{\mathscr{E}}(\hat{\psi}|\psi^0) + (\lambda - T\lambda_0)\|\hat{\phi} - \phi^0\|_1 &\leq (\lambda + T\lambda_0)\|\hat{\eta} - \eta^0\|_2 \\
&\leq (\lambda + T\lambda_0)^2 c_0^2/2 + \|\hat{\eta} - \eta^0\|_2^2/(2c_0^2) \\
&\leq (\lambda + T\lambda_0)^2 c_0^2/2 + \bar{\mathscr{E}}(\hat{\psi}|\psi^0)/2.
\end{aligned}$$

So then
$$\bar{\mathscr{E}}(\hat{\psi}|\psi^0) + 2(\lambda - T\lambda_0)\|\hat{\phi} - \phi^0\|_1 \leq (\lambda + T\lambda_0)^2 c_0^2.$$

Case 3 Suppose that
$$\|\hat{\phi} - \phi^0\|_1 + \|\hat{\eta} - \eta^0\|_2 \geq \lambda_0,$$

and that
$$T\lambda_0\|\hat{\eta} - \eta^0\|_2 \leq (\lambda + T\lambda_0)\|\hat{\phi}_{S_0} - (\phi^0)_{S_0}\|_1.$$

Then we have
$$\bar{\mathscr{E}}(\hat{\psi}|\psi^0) + (\lambda - T\lambda_0)\|\hat{\phi}_{S_0^c}\|_1 \leq 2(\lambda + T\lambda_0)\|\hat{\phi}_{S_0} - (\phi^0)_{S_0}\|_1.$$

So then
$$\|\hat{\phi}_{S_0^c}\|_1 \leq 6\|\hat{\phi}_{S_0} - (\phi^0)_{S_0}\|_1.$$

We can then apply the restricted eigenvalue condition to $\hat{\phi} - \phi^0$. But first, add $(\lambda - \lambda_0 T)\|\hat{\phi}_{S_0} - (\phi^0)_{S_0}\|_1$ to left- and right-hand side. The restricted eigenvalue condition then gives (invoking $2(\lambda + T\lambda_0) + (\lambda - T\lambda_0) \leq 3(\lambda + T\lambda_0)$)

$$\begin{aligned}
\bar{\mathscr{E}}(\hat{\psi}|\psi^0) + (\lambda - T\lambda_0)\|\hat{\phi} - \phi^0\|_1 &\leq 3(\lambda + T\lambda_0)\sqrt{s_0}\|\hat{\phi}_{S_0} - \phi^0\|_2 \\
&\leq 3(\lambda + T\lambda_0)\sqrt{s_0}\kappa\|\hat{g} - g^0\|_{Q_n} \\
&\leq 9(\lambda + T\lambda_0)^2 c_0^2 \kappa^2 s_0/2 + \bar{\mathscr{E}}(\hat{\psi}|\psi^0)/2.
\end{aligned}$$

So we arrive at
$$\bar{\mathscr{E}}(\hat{\psi}|\psi^0) + 2(\lambda - T\lambda_0)\|\hat{\phi} - \phi^0\|_1 \leq 9(\lambda + T\lambda_0)^2 c_0^2 \kappa^2 s_0.$$

$\square$

9.5.4 Proof of Lemma 9.3

Let Z be a standard normal random variable. Then by straightforward computations, for all $M > 0$,

$$\mathbb{E}[|Z|\mathbb{1}\{|Z| > M\}] \leq 2\exp[-M^2/2],$$

and

$$\mathbb{E}[|Z|^2\mathbb{1}\{|Z| > M\}] \leq (M+2)\exp[-M^2/2].$$

Thus, for n independent copies $Z_1, \ldots, Z_n$ of Z, and $M = 2\sqrt{\log n}$,

$$\mathbf{P}\left(\frac{1}{n}\sum_{i=1}^{n}|Z_i|\mathbb{1}\{|Z_i| > M\} > \frac{4\log n}{n}\right)$$

$$\leq \mathbf{P}\left(\frac{1}{n}\sum_{i=1}^{n}|Z_i|\mathbb{1}\{|Z_i| > M\} - \mathbb{E}[|Z|\mathbb{1}\{|Z| > M\}] > \frac{2\log n}{n}\right)$$

$$\leq \frac{n\mathbb{E}[|Z|^2\mathbb{1}\{|Z| > M\}]}{4(\log n)^2} \leq \frac{2}{n}.$$

The result follows from this, as

$$G_1(Y) = e^K|Y| + K,$$

and Y has a Gaussian mixture distribution. $\qquad\square$

Problems

9.1. Show that the objective function in the optimization in (9.7) is convex in ϕ, ρ.

9.2. Derive formula (9.8).

9.3. Generalized M-step in GEM algorithm for fitting FMR models
Prove formula (9.24) and derive the up-dates for ρ_r and $\phi_{r,j}$ at the end of Section 9.2.9 (before Section 9.2.9.1).

9.4. Prove (9.27) by using Jensen's inequality, saying that

$$\mathbb{E}[g(X)] \geq g(\mathbb{E}[X])$$

for any convex function $g(\cdot)$.

9.5. Linear mixed effects models
(a) Derive the negative log-likelihood (9.34) for linear mixed effects models.

(b) Using a simple example of a linear mixed effects model, show that the negative log-likelihood in (9.34) is a non-convex function in β, τ, σ^2 (when β, τ, σ^2 range over the whole parameter $\mathbb{R}^p \times \mathbb{R}_{>0} \times \mathbb{R}_{>0}$, assuming $\Gamma_\tau = \tau^2 I$).

9.6. Show that assumption (A2)(a) in Section 9.4.4 automatically holds for the co-variance model $\Gamma_\tau = \tau^2 I_{q \times q}$.

9.7. Show that Proposition 9.4 holds true.

9.8. Consistency of FMRLasso
Derive the result in (9.49), using the first equation from the proof of Theorem 9.1 in Section 9.5.3. Thereby, use the fact that Conditions 1, 2, 3, and 5 hold automatically for FMR models as shown in Proposition 9.4.

Chapter 10
Stable solutions

Abstract Estimation of discrete structure such as in variable selection or graphical modeling is notoriously difficult, especially for high-dimensional data. Subsampling or bootstrapping have the potential to substantially increase the stability of high-dimensional selection algorithms and to quantify their uncertainties. Stability via subsampling or bootstrapping has been introduced by Breiman (1996) in the context of prediction. Here, the focus is different: the resampling scheme can provide finite sample control for certain error rates of false discoveries and hence a transparent principle to choose a proper amount of regularization for structure estimation. We discuss methodology and theory for very general settings which include variable selection in linear or generalized linear models or graphical modeling from Chapter 13. For the special case of variable selection in linear models, the theoretical properties (developed here) for consistent selection using stable solutions based on subsampling or bootstrapping require slightly stronger assumptions and are less refined than say for the adaptive or thresholded Lasso.

10.1 Organization of the chapter

After an introduction, we present in Section 10.2 some examples motivating the need for stability. The definition of so-called "Stability Selection", originally proposed and analyzed in Meinshausen and Bühlmann (2010), is given in Section 10.3. There, we also include the main theorem on error control for the expected number of false positive selections. The following sections discuss further numerical examples and extensions, making also a brief comparison with the theory for the (adaptive) Lasso. The proof of the main theorem is presented in Section 10.7.

10.2 Introduction, stability and subsampling

We have discussed in Chapter 2, Section 2.6, and in Chapters 3, 4 and 7 the problem of variable (or group of variables) selection. In particular, we have argued that two-stage procedures like the adaptive or thresholded (Group)Lasso (Sections 2.8, 2.9, 4.6 and 7.8) or the relaxed Lasso (Section 2.10) have better potential and properties for selection than a single-stage Lasso procedure.

The results for the Lasso and its adaptive or thresholded version say that consistent variable selection is possible under (fairly restrictive) conditions on the design and on the size of the non-zero regression coefficients (i.e. the beta-min condition as discussed in Section 7.4), and if the regularization is chosen appropriately. Two questions which arise in this context are as follows. First, how "stable" (under re- or subsampling) is such a selection and can we do better with another "more stable" procedure? Secondly, can we achieve some type-I error control of falsely selecting an irrelevant variable? We will show that subsampling or bootstrapping and multiple sample splitting are simple but effective techniques for increased "stability" and for assigning p-values; for the latter see Chapter 11. Stability via subsampling or bootstrapping has been introduced by Breiman (1996) but only in the context of prediction and mainly for decision tree methods.

We assume that the data are of the form

$$Z_1, \ldots, Z_n \text{ i.i.d.}$$

Important examples include the case of generalized regression where $Z_i = (X_i, Y_i)$ with univariate response Y_i and p-dimensional covariate X_i, or $Z_i = X_i$ could be a p-dimensional variable as appearing in graphical modeling (see Chapter 13) or in cluster analysis.

Most concrete is a random-design linear model as discussed in Chapter 2, formula (2.1). We consider here the matrix- and vector-notation

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon, \tag{10.1}$$

where $\mathbf{Y}$, $\mathbf{X}$ and ε are $n \times 1$, $n \times p$ and $n \times 1$ vectors or matrices, respectively (where ε is independent of $\mathbf{X}$, with i.i.d. components and having mean zero). The goal is variable selection, i.e. estimation of

$$S_0 = \{j; \beta_j^0 \neq 0\}.$$

If $p \gg n$, we need to regularize the estimation procedure. Choosing the amount of regularization for variable selection can be more difficult and challenging than for prediction where a cross-validation scheme can be used. We refer to Chapter 7 where we discuss under which circumstances it is possible to simultaneously achieve accurate prediction and reasonable variable selection.

Here, we address the problem of proper regularization with a very generic subsampling approach (and bootstrapping would behave similarly). We show that subsampling can be used to determine the amount of regularization such that a certain familywise type-I error rate in multiple testing can be conservatively controlled, even for finite sample size. Beyond the issue of choosing the amount of regularization, the subsampling approach yields a new structure estimation or variable selection scheme. It is found empirically that it is often substantially better (and never really worse) than approaches without using the additional subsampling procedure.

In the sequel of this chapter, we consider the following setting and notation. For a generic structure estimation or variable selection technique, we assume that we have a tuning parameter $\lambda \in \Lambda \subseteq \mathbb{R}^+$ that determines the amount of regularization. A prime example is the penalty parameter in the Lasso, see (2.2) in Chapter 2, for a linear model as in (10.1). Alternatively, for such a linear model, we could also use some forward selection or boosting method and the parameter λ would then be the number of steps in these algorithms, see Chapter 12. (We note the difference that a large number of steps of iterations would have a meaning opposite to a large penalty parameter. However, this does not cause conceptual problems.) For every value $\lambda \in \Lambda$, we obtain a structure estimate $\hat{S}(\lambda) \subseteq \{1, \ldots, p\}$, where the latter enumerates the p features where each of these can be present or absent.

10.2.1 Stability paths for linear models

We motivate the concept of stability paths first for linear models

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

as in (10.1). Stability paths are derived from the concept of regularization paths. A regularization path is given by the coefficient value of each variable over all regularization parameters: $\{\hat{\beta}_j(\lambda); \ \lambda \in \Lambda, \ j = 1, \ldots, p\}$. For any given regularization parameter $\lambda \in \Lambda$, the selected set of variables

$$\hat{S}(\lambda) = \{j; \ \hat{\beta}_j(\lambda) \neq 0\}$$

is implicitly a function of the samples $I = \{1, \ldots, n\}$. We write $\hat{S}(\lambda) = \hat{S}_\lambda(I)$ where necessary to express this dependence on the sample.

Let I^* now be a random subsample of $\{1, \ldots, n\}$ of size $\lfloor n/2 \rfloor$, drawn without replacement. For every set $K \subseteq \{1, \ldots, p\}$, the subsampling-probability of being in the selected set $\hat{S}_\lambda(\cdot)$ is

$$\hat{\Pi}_K(\lambda) = \mathbf{P}^*[K \subseteq \hat{S}_\lambda(I^*)]. \tag{10.2}$$

The probability $\mathbf{P}^*$ in (10.2) is with respect to the random subsampling and it equals the relative frequency for $K \subseteq \hat{S}_\lambda(I_b)$ over all $\binom{n}{m}$ subsets I_b ($b = 1 \ldots, \binom{n}{m}$) of size $m = \lfloor n/2 \rfloor$, which itself is a U-statistic of order $m = \lfloor n/2 \rfloor$. The expression in (10.2)

can be approximated by B random subsamples $I^{*1}, \ldots, I^{*B}$ (B large)

$$B^{-1} \sum_{b=1}^{B} 1(K \subseteq \hat{S}_\lambda(I^{*b})).$$

The subsample size of $\lfloor n/2 \rfloor$ is chosen as it resembles most closely the bootstrap (Freedman, 1977; Bühlmann and Yu, 2002) (and it allows for slightly faster computation than bootstrapping since the estimator is based and computed many times on subsample size $\lfloor n/2 \rfloor$ only).

For every variable $j = 1, \ldots, p$, the stability path is given by the selection probabilities

$$\{\hat{\Pi}_j(\lambda); \ j = 1, \ldots, p, \ \lambda \in \Lambda\}.$$

It is complementing the usual regularization path plots that show the coefficients of all variables

$$\{\hat{\beta}_j(\lambda); \ j = 1, \ldots, p, \ \lambda \in \Lambda\}$$

as a function of the regularization parameter λ. Figure 10.1 illustrates the two different paths.

10.2.1.1 Riboflavin production with bacillus subtilis

We have introduced in Section 9.2.6 a data set about riboflavin (vitamin B2) production with bacillus subtilis. The real-valued response variable Y is the logarithm of the riboflavin production rate and there is a $p = 4088$-dimensional covariate measuring the logarithm of the expression level of 4088 genes. Here, we consider a smaller subset of the data with (sub-)sample size $n = 115$ which should be more homogeneous than the larger data set in Section 9.2.6 with $n = 146$ samples. Certain mutations of genes are thought to lead to higher vitamin concentrations and the challenge is to identify those relevant genes via a linear regression analysis. That is, we consider a linear model as in (10.1) and the goal is to infer the set of active variables $S_0 = \{j; \ \beta_j^0 \neq 0\}$.

We use the Lasso for variable selection (see Section 2.6), or at least for variable screening (see Section 2.5), to infer the active set S_0. To see how the Lasso and the related stability path cope with noise variables, we randomly permute all but 6 of the $p = 4088$ gene expression variables across the $n = 115$ samples, using the same permutation to keep the dependence structure between the permuted gene expression variables intact. The set of 6 non-permuted genes has been chosen randomly (once) among the 200 genes with the highest marginal empirical correlation with the response Y. The Lasso regularization path $\{\hat{\beta}(\lambda); \ \lambda \in \Lambda\}$ is shown in the left panel of Figure 10.1, as a function of the regularization parameter λ but rescaled so

that $\lambda = 1$ corresponds to the minimal λ-value for which the null model is selected, usually denoted as $\lambda_{\max}$ (see e.g Section 2.12) and $\lambda = 0$ amounts to the so-called Basis Pursuit solution which includes $\min(n, p)$ variables. Three (among the six) of the "relevant" (non-permuted) genes stand out, but the remaining other three "relevant" variables are hidden within the paths of noise covariates (permuted variables). The right panel of Figure 10.1 shows the stability path. At least four (among the six) "relevant" variables stand out much clearer now than they did in the Lasso regularization path plot.
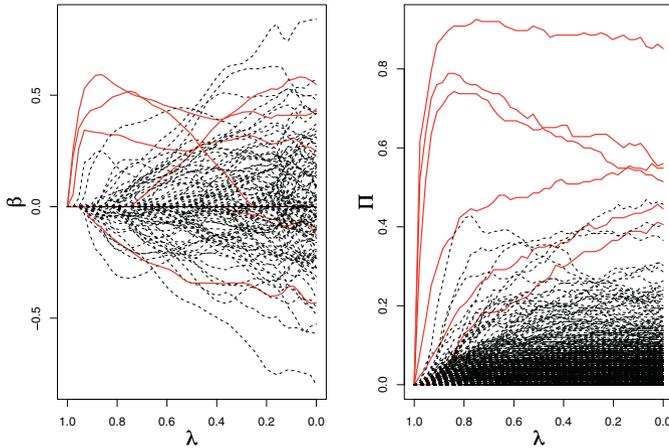
Fig. 10.1 Left: the Lasso regularization path for the riboflavin production data set with $n = 115$ and $p = 4088$. The paths of the 6 non-permuted variables (genes) are plotted as solid, red lines, while the paths of the 4082 permuted genes are shown as broken, black lines. Selecting a model including all 6 non-permuted variables (genes) invariably means selecting a large number of irrelevant noise variables. Right: the stability path of the Lasso. The first 4 variables chosen with stability selection are truly non-permuted variables. The figure is taken from Meinshausen and Bühlmann (2010).

Choosing the right regularization parameter is very difficult for the original Lasso path. The prediction optimal, cross-validated choice often includes false positive selections, as outlined in Section 2.5.1 in Chapter 2: this can be observed in this example as well, where 14 permuted noise variables are included in the model chosen by 10-fold cross-validation. Figure 10.1 motivates that choosing the right regularization parameter is much less critical for the stability path and that we have a better chance to select truly relevant variables.

10.2.1.2 Motif regression in computational biology

We present here another example illustrating that scoring the relevance of a variable in terms of the subsampling selection probabilities $\hat{\Pi}_j(\lambda)$ is often much better than in terms of the absolute values of regression coefficients $|\hat{\beta}_j(\lambda)|$.

We consider a real data set about motif regression. We refer to Sections 2.5.2 and 2.8.5.1 for a brief motivation and description of motif regression. Here, we consider a subset of the data in Section 2.8.5.1. We have a univariate real-valued response variable Y_i, measuring the expression of gene i and we have a p-dimensional covariate $X_i \in \mathbb{R}^p$ where $X_i^{(j)}$ is an abundance score of a short candidate motif j in the DNA segment around gene i. The latter is based on DNA sequence data only. There are $n = 1200$ samples and $p = 660$ covariates arising from a heat-shock experiment with yeast. We relate the centered response Y and centered covariates X using a linear model

$$Y_i = \sum_{j=1}^{p} \beta_j X_i^{(j)} + \varepsilon_i \ (i = 1, \ldots, n).$$

The goal is variable selection based on the idea that the relevant covariates in the linear model correspond to the relevant motifs (binding sites) of a particular transcription factor.

We use the Lasso in (2.2) with regularization parameter $\hat{\lambda}_{CV}$ from 10-fold cross-validation. This procedure selects 20 variables having corresponding estimated regression coefficients different from zero. We then run the subsampling and compute the subsampling selection probabilities for Lasso with $\hat{\lambda}_{CV}$. Table 10.1 describes the results for the 9 most promising variables (corresponding to motifs) ordered with respect to $|\hat{\beta}_j(\hat{\lambda}_{CV})|$. Questions which arise include the following. Should we report

motif j	41	29	635	19	34	603	618	596	30
$|\hat{\beta}_j|$	1.42	1.27	0.81	0.61	0.57	0.49	0.33	0.3	0.3
$\hat{\Pi}_j$	100%	100%	100%	74%	98%	32%	81%	80%	97%

Table 10.1 Lasso with $\hat{\lambda}_{CV}$ from 10-fold cross-validation for a motif regression problem with $n = 1200$ and $p = 660$. The number of non-zero estimated coefficients is 20: the first row shows the variables with the nine largest coefficients in absolute values. Second row: subsampling selection probabilities for these nine variables.

the relevance of the variables in terms of $\hat{\beta}_j(\hat{\lambda}_{CV})$ or according to a different ordering? How many of these 20 variables are relevant? When looking at the subsampling selection probabilities in Table 10.1, we see that subsampling assigns another ordering.

Next, we generate some semi-synthetic data. We select 5 variables at random, say $j_1, \ldots, j_5 \in \{1, \ldots, p\}$, among all $p = 660$ covariates and set

$$Y_i = \sum_{k=1}^{5} \beta_{j_k}^0 X_i^{(j_k)} + \varepsilon_i \ (i = 1, \ldots, n = 1200),$$

where $\varepsilon_1, \ldots, \varepsilon_n$ i.i.d. $\sim \mathcal{N}(0, \sigma^2)$. We choose $\beta_{j_1}^0 = \ldots = \beta_{j_5}^0$ and σ^2 such that the signal to noise ratio is very low with SNR = 0.1. For estimation, we still use all the $p = 660$ covariates although 655 of them are noise variables. Now we know the true active set $S_0 = \{j; \beta_j^0 \neq 0\}$ (whose cardinality is 5). Figure 10.2 illustrates that using the subsampling selection probabilities $\hat{\Pi}_j(\lambda)$, we can perfectly distinguish between active and noise covariates if we were able to choose an appropriate threshold value for $\hat{\Pi}_j(\hat{\lambda}_{CV})$. This is in contrast to using $\hat{\beta}_j(\hat{\lambda}_{CV})$ where we cannot make a perfect distinction between active and noise covariates using any threshold value for the estimated regression coefficients.
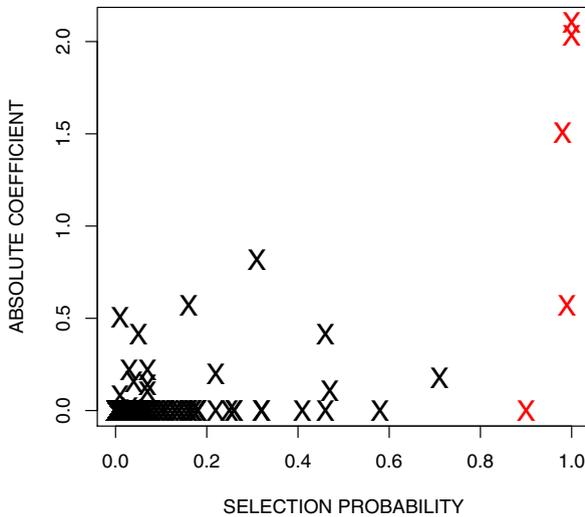

Fig. 10.2 Semi-synthetic motif regression problem with real data design and artificial sparse regression coefficients and Gaussian noise. We use the Lasso with $\hat{\lambda}_{CV}$ from cross-validation. x-axis: $\hat{\Pi}_j(\hat{\lambda}_{CV})$ and y-axis: $|\hat{\beta}_j(\hat{\lambda}_{CV})|$. Red crosses: true active variables; black crosses: noise variables.

10.3 Stability selection

In a traditional setting, variable selection would amount to choosing one element of the set of models

$$\{\hat{S}(\lambda); \ \lambda \in \Lambda\}, \tag{10.3}$$

where Λ is again the set of considered regularization parameters, which can be either continuous or discrete. The set in (10.3) is the set of all variable selection subsets that arise when varying the regularization parameter $\lambda \in \Lambda$. There are typically two problems: first, the true active set of variables S_0 might not be a member of (10.3). Second, even if it is a member, it is typically very hard with high-dimensional data to determine the right amount of regularization λ to select exactly S_0, or at least a close approximation. When $\hat{S}(\lambda)$ is from the Lasso, the first issue is characterized by the irrepresentable condition which is sufficient and (essentially) necessary that S_0 is in the set in (10.3), see Section 2.6.1 in Chapter 2 and Section 7.5.1 in Chapter 7.

With stability selection, we do not simply select one set of variables in the list (10.3). Instead, the data are perturbed (by subsampling) many times and we choose all structures or variables that occur in a large fraction of the resulting selection sets. We use the following definition for stable variables.

For a cutoff π_{thr} with $0 < \pi_{\text{thr}} < 1$ and a set of regularization parameters Λ, the set of stable variables is defined as

$$\hat{S}_{\text{stable}} = \{j; \ \max_{\lambda \in \Lambda} \hat{\Pi}_j(\lambda) \geq \pi_{\text{thr}}\}. \tag{10.4}$$

Here, $\hat{\Pi}$ is as defined in (10.2). We keep variables with a high selection probability and disregard those with low selection probabilities. Of course, the problem has now shifted to choose a good cutoff value $0 < \pi_{\text{thr}} < 1$ which is a tuning parameter of the stability selection procedure. We will discuss its choice below in Section 10.3.1. It is worthwhile to emphasize that empirical results do not depend very much on the choice of the initial regularization λ (if $\Lambda = \lambda$ is a singleton) or the initial region Λ for the regularization parameter: loosely speaking, as long as λ or Λ contain values leading to overestimation of the true active set of variables S_0, the results after the stability selection procedure as described in (10.4) are stable. See for example Figure 10.1.

10.3.1 Choice of regularization and error control

We focus here on the problem how to choose the regularization parameter π_{thr} in the stability selection procedure in (10.4). We address it by controlling the expected number of false positives (false selections), i.e. type I error control.

For such an error control, we introduce first some additional notation. Let $\hat{S}_\Lambda = \cup_{\lambda \in \Lambda} \hat{S}(\lambda)$ be the set of selected variables when varying the regularization parameter $\lambda \in \Lambda$. Let q_Λ be the expected number of selected variables $q_\Lambda = \mathbb{E}|\hat{S}_\Lambda(I)|$. (Note the slight change of notation where we emphasize with $\hat{S}_\Lambda(I)$ the dependence on the sample I). Define V to be the number of falsely selected variables (false positives) with stability selection,

$$V = \left| S_0^c \cap \hat{S}_{\text{stable}} \right|.$$

The goal is to achieve control or an upper bound for $\mathbb{E}[V]$, the expected number of false positives.

Since the distribution of the underlying estimator $\hat{S}(\lambda)$ depends on many unknown quantities, exact finite-sample control of $\mathbb{E}[V]$ is difficult in general. But we provide a simple answer under some simplifying assumptions.

Theorem 10.1. *Assume that the distribution of $\{1(j \in \hat{S}(\lambda))\}, j \in S_0^c\}$ is exchangeable for all $\lambda \in \Lambda$. Also, assume that the original selection procedure is not worse than random guessing, i.e.,*

$$\frac{\mathbb{E}|S_0 \cap \hat{S}_\Lambda|)}{\mathbb{E}(|S_0^c \cap \hat{S}_\Lambda|)} \geq \frac{|S_0|}{|S_0^c|}. \tag{10.5}$$

Then, the expected number V of falsely selected variables is bounded for $\pi_{\text{thr}} \in (1/2, 1)$ by

$$\mathbb{E}[V] \quad \leq \quad \frac{1}{2\pi_{\text{thr}} - 1} \frac{q_\Lambda^2}{p}. \tag{10.6}$$

A proof is given in Section 10.7. The exchangeability condition is a restrictive assumption and we discuss it in more detail in Section 10.3.1.3. The expected number of falsely selected variables is sometimes called the per-family error rate (PFER) or, if divided by p, $\mathbb{E}[V]/p$ is the per-comparison error rate (PCER) in multiple testing (Dudoit et al., 2003). Note that Theorem 10.1 does not require a beta-min condition (see Section 7.4) because the theorem only makes a statement about false positive selections (while a beta-min condition is needed to avoid false negatives).

For fixed Λ, the threshold value π_{thr} is the tuning parameter for stability selection. We propose to fix this value via the PFER control $\mathbb{E}[V] \leq \nu$ where ν is specified a-priori. Note that this fits into the commonly used framework of fixing type-I error control beforehand. For example, when choosing $\nu = \alpha$, a small number such as $\alpha = 0.05$, then $\mathbb{E}[V] \leq \alpha$ implies

$$\mathbf{P}[V > 0] \leq \alpha \tag{10.7}$$

which says that the familywise error rate, i.e. the probability of at least one false positive selection, is conservatively controlled at level α. We leave the derivation of (10.7) as Problem 10.1.

Given ν, we can then solve for the tuning parameter:

$$\text{if } q_\Lambda^2 \le pv : \ \pi_{\text{thr}} = (1 + \frac{q_\Lambda^2}{pv})/2, \tag{10.8}$$

and if $q_\Lambda^2 > pv$, we cannot control the error $\mathbb{E}[V]$ with the formula appearing in Theorem 10.1. To use (10.8), we need knowledge about q_Λ. This can be easily achieved by regularizing the selection procedure in terms of the number q of selected variables: we then write $\hat{S}_q$ and obviously, $|\hat{S}_q| = q$. For example, with the Lasso in (2.2), the number q may be given by the variables which enter first in the regularization path when varying from a maximal value $\lambda_{\max}$ to some minimal value $\lambda_{\min}$. Other examples are described below. The choice of the value q is not very important as long as we select it within a reasonable range. In absence of any idea how to choose q, we can use formula (10.8) in the other direction. We can take a default stability threshold parameter, say $\pi_{\text{thr}} = 0.9$, and then solve for q

$$q = \lfloor \sqrt{vp(2\pi_{\text{thr}} - 1)} \rfloor.$$

As discussed above, we can either fix the regularization region Λ and then choose π_{thr} such that $E(V)$ is controlled at the desired level; or vice-versa, we fix the stability threshold π_{thr} and choose Λ.

Choosing less variables (reducing q_Λ) or increasing the threshold π_{thr} for stability selection will, unsurprisingly, reduce the expected number of falsely selected variables, with an achievable non-trivial and rather "minimal" value for the PFER $\mathbb{E}[V] \le 1/p$ when using $\pi_{\text{thr}} = 1$ and having $q_\Lambda = 1$. This seems low enough for all practical purposes as long as say $p > 10$.

Without stability selection, the regularization parameter λ depends on the unknown noise level of the observations. The advantage of stability selection is that exact error control is possible, and the method works fine even though the noise level is unknown.

10.3.1.1 Pointwise control

For some applications, in particular beyond variable selection in linear models, evaluation of subsampling replicates of $\hat{S}(\lambda)$ can be computationally very demanding for a single value of λ. If this single value λ is chosen such that some overfitting occurs and the set $\hat{S}(\lambda)$ is too large, in the sense that $\hat{S}(\lambda) \supseteq S_0$ with high probability (see Section 2.5), the approach as above can be used with $\Lambda = \{\lambda\}$ being a singleton. Results typically do not depend strongly on the utilized regularization λ. See the example for graphical modeling in Section 13.4.1. Setting $\Lambda = \{\lambda\}$, one can immediately transfer all results above to the case of what we call here pointwise control. For methods which select structures incrementally, i.e. for which $\hat{S}(\lambda) \subseteq \hat{S}(\lambda')$ for all $\lambda \ge \lambda'$, pointwise control with λ and control with $\Lambda = [\lambda, \infty)$ are equivalent

since $\hat{\Pi}_j(\lambda)$ is then monotonically increasing with decreasing λ for all $j = 1, \ldots, p$. See Problem 10.2.

10.3.1.2 Examples of procedures choosing q variables

We have discussed above that the error control using Theorem 10.1 requires knowledge of the value $q_\Lambda = \mathbb{E}|\hat{S}_\Lambda(I)|$, where $\hat{S}_\Lambda(I) = \cup_{\lambda \in \Lambda} \hat{S}_\lambda(I)$. Even with pointwise control where $\Lambda = \{\lambda\}$, the value q_Λ may be unknown. Trivially, q_Λ is known for variable selection procedures which select q variables: then $q_\Lambda = q$. We describe now examples of procedures which select q variables.

Consider the Lasso in (2.2) for a range $\Lambda = [\lambda_{\min}, \lambda_{\max}]$ of regularization parameters. Define the Lasso-based procedure $\hat{S}_q$ selecting the q variables which enter first in the regularization path when varying from the maximal value $\lambda_{\max}$ to the minimal value $\lambda_{\min}$. Note that if there would be less than q active variables in the regularization path over the range Λ, we would select all active variables and this number is bounded by q which is sufficient for the error control in Theorem 10.1.

Alternatively, consider the Lasso in (2.2) for a singleton $\Lambda = \{\lambda\}$. We then have an estimated active set $\hat{S}(\lambda)$. Define $\hat{S}_q$ as the procedure selecting the q variables from $\hat{S}(\lambda)$ whose regression coefficients are largest in absolute values. Typically, we would choose λ such that $|\hat{S}(\lambda)| \geq q$. If there would be less than q active variables in $\hat{S}(\lambda)$, we would select all active variables and this number is bounded by q which again is sufficient for the error control in Theorem 10.1.

Consider the L_2Boosting procedure as described in Section 12.4.4 in Chapter 12. We define $\hat{S}_q$ to include the first q variables which arise during the boosting iterations. Similarly, we may use a forward selection algorithm, see Section 12.7. Then, $\hat{S}_q$ contains the first q selected variables.

10.3.1.3 The exchangeability condition and the generality of Theorem 10.1

We remark that Theorem 10.1 applies to a very general range of discrete structure estimation problems where inclusion or exclusion for individual features is possible. Variable selection in linear and other models is of this type: a feature is then a variable and each feature can be included (selected) or excluded (non-selected). Another example is graphical modeling, as discussed in Chapter 13: a feature is an edge between variables and each feature can be included or excluded. A third example is clustering (Problem 10.3). Theorem 10.1 is valid for all these problems.

Due to the generality of Theorem 10.1, the involved exchangeability assumption is perhaps stronger than one would wish, but there does not seem to be an easy way of getting error control in the same generality without making similar assumptions.

For example, Fan et al. (2009b) make use of a similar condition for error control in regression.

In regression and classification, the exchangeability assumption is fulfilled for all reasonable procedures $\hat{S}$ (which do not depend on the ordering of covariables) if the design is random and the distribution of $(Y, X^{(S_0)}, X^{(S_0^c)})$ is invariant under permutations of variables in S_0^c. The simplest example is independence between each $X^{(j)}$ and all other variables $\{X^{(k)}; \ k \in S_0\} \cup Y$ for each $j \in S_0^c$. Another example is as follows.

Example 10.1. Consider a linear model as in (10.1) with random design and active set $S_0 = \{j; \ \beta_j \neq 0\}$:

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon, \ \varepsilon \sim \mathcal{N}_n(0, \sigma^2 I_{n \times n}),$$
$$\text{the rows of } \mathbf{X} \text{ are i.i.d. } \sim \mathcal{N}_p(0, \Sigma),$$
$$\Sigma_{j,j} = \sigma^2 \text{ for all } j \in \{1, \ldots, p\},$$
$$\Sigma_{j,\ell} = \Sigma_{k,\ell} \text{ for all } j, k \in S_0^c, \ \ell \in S_0.$$

Then, the distributions of $(Y, X^{(S_0)}, \{X^{(j)}; \ j \in S_0^c\})$ and of $(Y, X^{(S_0)}, \{X^{(\pi(j))}; \ j \in S_0^c\})$ are the same for any permutation $\pi : S_0^c \to S_0^c$ (Problem 10.4). Therefore, the exchangeability condition holds for any procedure $\hat{S}$ which is invariant under reordering of the covariates.

A special case is equicorrelation with $\Sigma_{j,k} = \rho \ (j \neq k)$ and $\Sigma_{j,j} = 1 \ (j = 1, \ldots, p)$. For $0 < \rho < 1$, the irrepresentable condition (see Sections 2.6 and 7.5.1) holds (for some $0 < \theta < 1$ depending on ρ and s_0), see Problem 2.4 and Problem 6.14. For $\rho < 0$, the irrepresentable condition can fail: an example is with $p \geq 3$ and $-1/(2s_0 - 1) \geq \rho > -1/(p-1)$ (Problem 10.4). Thus, this is a very special example where the exchangeability condition holds but the irrepresentable condition fails. However, we then must have $s_0 > p/2$ which excludes all high-dimensional scenarios due to failure of sparsity.

It appears that the exchangeability condition (for say the Lasso) exhibits at least the same degree of restrictiveness as the irrepresentable condition described in Section 2.6 and Section 7.5.1. For example, in case of no noise and if the problem is identified, the exchangeability condition (roughly) implies that either none or all of the noise variables are selected ("roughly" means here that the randomness of the covariates is neglected). In the former case and for the Lasso, this implies that the irrepresentable condition must hold (necessity of the condition, see Theorem 7.1, part 2). In the latter case, when all noise variables are selected, the "not worse than random guessing" condition (10.5) may not hold and also formula (10.6) becomes extremely loose since $q^2 = |S_0^c|^2 = (p - s_0)^2$ is (typically) large relative to p. For real data, we have no guarantee that the exchangeability assumption is fulfilled but some numerical examples in Section 10.4 show that the bound from Theorem 10.1 holds up very well.

10.4 Numerical results

We consider here the performance of stability selection on semi-synthetic data sets. We use real data design matrices $\mathbf{X}$ and build synthetic data, according to a linear model, by generating regression coefficients β_j $(j = 1, \ldots, p)$ and random errors $\varepsilon_1, \ldots, \varepsilon_n$ i.i.d. $\mathcal{N}(0, \sigma^2)$. Since the design is from real data, we call it semi-synthetic data.

We use the following real data-sets for the design. The first one is from motif regression with $p = 660$ and $n = 2587$. For a brief description and motivation about motif regression, we refer to Sections 2.5.2 and 2.8.5.1. The real-valued predictor covariates are abundance scores for p candidate motifs (for each of the genes). Our data set is from a heat-shock experiment with yeast. In addition, we consider the riboflavin production data set, described in Section 9.2.6: here, we use data with $p = 4088$ and $n = 158$. We generate sparse regression coefficients β_j i.i.d. Uniform$([0, 1])$ and the size of the active set is varied with s_0 taking 16 different values between 4 and 50. We choose error variances σ^2 to achieve signal to noise ratios (SNRs) in $\{0.5, 2\}$. In total, there are 64 scenarios.

We then test how well the error control of Theorem 10.1 holds up for these semi-synthetic data-sets. We are interested in the comparison between the cross-validated solution for the Lasso (without stability selection) and stability selection using the Lasso. For stability selection, we chose $q = \sqrt{0.8p}$ (the first q variables entering the regularization path, see Section 10.3.1) and a threshold of $\pi_{\mathrm{thr}} = 0.6$, corresponding to a control of $\mathbb{E}[V] \leq 2.5$, where V is the number of wrongly selected variables. The control is mathematically derived under the assumption of exchangeability as described in Theorem 10.1. This assumption is most likely not fulfilled for the given real data designs and it is of interest to see how well the error bound holds up for our semi-synthetic data. The results are shown in Figure 10.3. In comparison to the Lasso with cross-validated $\hat{\lambda}_{CV}$, stability selection reduces the number of falsely selected variables dramatically, while maintaining almost the same power to detect relevant variables. It is no surprise that the Lasso selects too many noise covariates, as discussed in Section 2.5.1 (see also Chapter 7, Section 7.8.3). The number of falsely chosen variables is remarkably well controlled at the desired level, giving empirical evidence that the derived error control is useful beyond the discussed setting of exchangeability. Stability selection thus helps to select an appropriate amount of regularization such that false positive selections are under control.

The variables selected from stability selection can be refitted by e.g. least squares estimation, and we can then compare its prediction error with the Lasso. So far, there is no systematic study (empirical or theoretical) for this. Shawe-Taylor and Sun (2010) present a real-data medical example consisting of $n = 1842$ subjects and $p = 793$ covariates (six classical risk factors and 787 single-nucleotide polymorphism genotype features): they report that stability selection with refitting (using the Lasso) is only slightly worse in terms of prediction than the Lasso while the former yields a much sparser model fit.
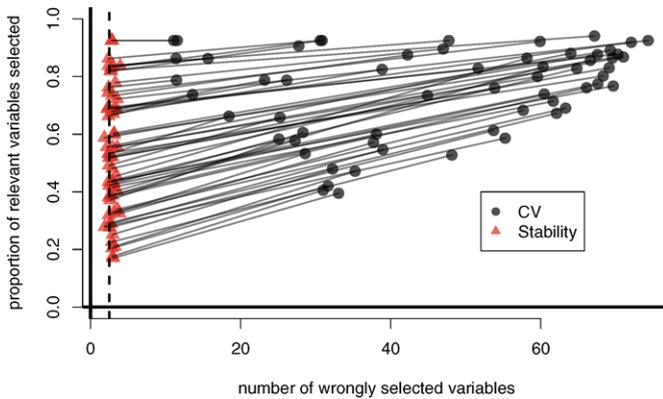
Fig. 10.3 Comparison of stability selection with cross-validation. The cross-validated solution (for standard Lasso) is indicated by a dot and the corresponding stability selection by a red triangle, showing the average proportion of correctly identified relevant variables versus the average number of falsely selected variables. Each pair consisting of a dot and triangle corresponds to a simulation setting (some specified SNR and s_0). The broken vertical line indicates the value at which the number of wrongly selected variables is controlled, namely $\mathbb{E}(V) \leq 2.5$. Looking at stability selection, the proportion of correctly identified relevant variables is very close to the CV-solution, while the number of falsely selected variables is reduced dramatically. The figure is taken from Meinshausen and Bühlmann (2010).

10.5 Extensions

As written in Section 10.3.1.3, stability selection can be applied in many other discrete structure estimation problems. We demonstrate its use also in Section 13.4.1 in Chapter 13, in the context of estimating an undirected conditional independence graph.

10.5.1 Randomized Lasso

An interesting aspect is that stability selection with the original procedure alone yields often substantial improvements already in terms of false positive selections while maintaining power for detection of true structures. Moreover, adding some extra sort of randomness can lead to further gains.

Randomized Lasso is a generalization of the Lasso which uses a weighted ℓ_1-norm penalty

$$\sum_{j=1}^{p} W_j |\beta_j|,$$

where W_j are positive weights discussed in more detail below. The randomized Lasso estimator is then defined as

$$\hat{\beta}_{\text{random}}(\lambda) = \arg\min_{\beta}(\|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + \lambda \sum_{j=1}^{p} W_j \beta_j). \tag{10.9}$$

Thus, this is a weighted Lasso procedure, like e.g. the adaptive Lasso in Section 2.8 in Chapter 2, and it is of the form appearing also in Section 6.9 in Chapter 6 and in Chapter 7. The weights W_j here, however, are chosen at random with no relation to the data. We generate

$$W_1, \ldots, W_p \text{ i.i.d. with values in the range } [\gamma, 1] \ (0 < \gamma \le 1),$$

and $W_1, \ldots, W_p$ are independent of the data $(X_1, Y_1), \ldots, (X_n, Y_n)$. Concretely, we can generate the weights from a two-point distribution

$$W_j \in \{\gamma, 1\}, \ \mathbf{P}[W_j = \gamma] = 1/2,$$

and $\gamma \in (0, 1)$ is called the weakness parameter. The word "weakness" is borrowed from the terminology of weak greedy algorithms (Temlyakov, 2000) which are loosely related to the randomized Lasso. In practice, choosing γ in the range of $(0.2, 0.8)$ gives very useful results.

The penalty weights are simply chosen at random. Thus, the implementation is very straightforward by appropriate re-scaling of the predictor variables. An exact description is given in Section 2.8.4.

However, purely random weights in the penalty seem to be nonsensical at first sight since one cannot expect any improvement from such a random perturbation. If applied only with one random perturbation, randomized Lasso is not very useful. But applying randomized Lasso many times and looking for variables that are chosen often, analogously as with stability selection in (10.4) is a very powerful procedure. In particular, randomized Lasso can be combined with subsampling observations.

Meinshausen and Bühlmann (2010) show that randomized Lasso, in conjunction with stability selection, can bring further empirical improvements over stability selection with the plain Lasso procedure. We remark that this finding is loosely related to empirical results about Random Forest (Breiman, 2001).

10.6 Improvements from a theoretical perspective

It is shown in Meinshausen and Bühlmann (2010) that randomization is beneficial
to achieve variable selection consistency under weaker conditions than what the
non-randomized procedure necessarily needs, at least for the case of Lasso and for
orthogonal matching pursuit which is briefly described in Section 12.7.1 in Chapter
12.

When using the Lasso in (2.2), Section 2.6 discusses that the neighborhood stability
or irrepresentable condition, a rather restrictive assumption on the design, is neces-
sary and sufficient for variable selection consistency, saying that

$$\mathbf{P}[\hat{S}(\lambda) = S_0] \to 1 \; (n \to \infty).$$

Meinshausen and Bühlmann (2010) prove that the randomized Lasso, applying
(10.9) many times and computing relative selection frequencies as in (10.4), is
consistent for variable selection under a sparse eigenvalue condition on the design
which is weaker than the irrepresentable condition. The details are as follows.

Definition 10.1. (Sparse eigenvalues) *For $S \subseteq \{1,\dots,p\}$, $\beta \in \mathbb{R}^p$, let β_S be defined
as $\beta_{j,S} := \beta_j 1(j \in S)$ $(j = 1,\dots,p)$. Denote by $\hat{\Sigma} = n^{-1}\mathbf{X}^T\mathbf{X}$. The minimal sparse
eigenvalues are then defined for $s \le p$ as*

$$\Lambda_{\min}^2(s) = \inf_{|S|=s; \beta \in \mathbb{R}^p} \frac{\beta_S^T \hat{\Sigma} \beta_S}{\|\beta_S\|_2^2}.$$

Analogously, for the maximal sparse eigenvalues:

$$\Lambda_{\max}^2(s) = \sup_{|S|=s; \beta \in \mathbb{R}^p} \frac{\beta_S^T \hat{\Sigma} \beta_S}{\|\beta_S\|_2^2}.$$

Sparse eigenvalues are related to the compatibility condition and restricted eigen-
values as discussed in Sections 6.13.5 and 6.13 in Chapter 6, and see also Section
7.5.4 in Chapter 7. The assumption in Meinshausen and Bühlmann (2010) is: there
exists some $C > 1$ and some $\kappa \ge 9$ such that

$$\frac{\Lambda_{\max}(Cs_0^2)}{\Lambda_{\min}^{3/2}(Cs_0^2)} < \sqrt{C}/\kappa, \qquad s_0 = |S^0|. \tag{10.10}$$

Furthermore, the non-zero regression coefficients need to be sufficiently large, i.e.
fulfill a beta-min condition (see also Section 7.4),

$$\min\{|\beta_j^0|; \; 1 \le j \le p, \; \beta_j^0 \ne 0\} > D\sigma s_0^{5/2} \sqrt{\log(p)/n}, \tag{10.11}$$

where $D > 0$ is a constant depending on C in the sparse eigenvalue condition (10.10)
and σ^2 denotes the error variance.

When comparing to variable selection consistency with for example the adaptive or thresholded Lasso as discussed in Section 7.9 from Chapter 7, we see that the conditions in (10.10) and (10.11) are more restrictive (bound on sparse eigenvalues imply bounds on restricted eigenvalues, see Sections 6.13.5 and 6.13 in Chapter 6). In addition, under the conditions above, the Lasso has reasonable variable selection properties with $|\hat{S}_{\text{Lasso}} \setminus S_0| = O(s_0)$, see Lemma 7.3 in Section 7.8.3. We also note that (10.11) might be an unrealistic assumption: without such a condition, we cannot detect small regression coefficients (but e.g. the Lasso will still select all variables whose regression coefficients are "large" in absolute value), see also at the end of Section 2.6 in Chapter 2.

We conclude that from a theory perspective, the randomized Lasso or stability selection are not doing any better than say the adaptive Lasso or even the Lasso itself (when taking the view that $O(s_0)$ false positive selections are acceptable). In fact, the theory for variable selection with the adaptive Lasso is based on weaker assumptions than what is described above. This may be due to the fact that the analysis in Meinshausen and Bühlmann (2010) is for a randomized algorithm and much more coarse than for the adaptive Lasso. However, the result on consistent variable selection with the randomized Lasso is of interest as it indicates that randomization can actually improve the procedure in the sense that it is valid (i.e. consistent) for a wider class of problems than the non-randomized method.

10.7 Proofs

10.7.1 Sample splitting

An alternative to subsampling is sample splitting. Instead of observing whether a given variable is selected for a random subsample, one can look at a random split of the data into two non-overlapping samples of equal size $\lfloor n/2 \rfloor$ and see whether the variable is chosen in both sets simultaneously.

Let I_1 and I_2 be two random subsets of $\{1, \ldots, n\}$ with $|I_1| = |I_2| = \lfloor n/2 \rfloor$ and $I_1 \cap I_2 = \emptyset$ (and when n is even, $I_1 \cup I_2 = \{1, \ldots, n\}$). Define the simultaneously selected set as the intersection of $\hat{S}_\lambda(I_1)$ and $\hat{S}_\lambda(I_2)$,

$$\hat{S}^{\text{simult},\lambda} = \hat{S}_\lambda(I_1) \cap \hat{S}_\lambda(I_2).$$

Define the simultaneous selection probabilities for any set $K \subseteq \{1, \ldots, p\}$ as

$$\hat{\Pi}_K^{\text{simult}}(\lambda) = \mathbf{P}^*[K \subseteq \hat{S}^{\text{simult},\lambda}], \tag{10.12}$$

where the probability $\mathbf{P}^*$ is with respect to the random sample splitting.

Stability selection as defined in Section 10.3 works with the selection probabilities based on subsampling but the following lemma lets us convert these probabilities into simultaneous selection probabilities based on sample splitting. The latter is used for the proof of Theorem 10.1. The bound is rather tight for selection probabilities close to 1.

Lemma 10.1. *For any set $K \subseteq \{1,\ldots,p\}$, a lower bound for the simultaneous selection probabilities is given by*

$$\hat{\Pi}_K^{\text{simult}}(\lambda) \geq 2\hat{\Pi}_K(\lambda) - 1. \tag{10.13}$$

The inequality holds for every realization ω (of the n original data points) in the underlying probability space Ω.

Proof. Let I_1 and I_2 be the two random subsets from a sample split of $\{1,\ldots,n\}$ with $|I_1| = |I_2| = \lfloor n/2 \rfloor$ and $I_1 \cap I_2 = \emptyset$. Denote by $s_K(\{1,1\})$ the probability

$$\mathbf{P}^*[\{K \subseteq \hat{S}_\lambda(I_1)\} \cap \{K \subseteq \hat{S}_\lambda(I_2)\}].$$

Note that the two events are not independent as the probability is only with respect to a random split of the fixed samples $\{1,\ldots,n\}$ into I_1 and I_2. The probabilities $s_K(\{1,0\}), s_K(\{0,1\}), s_K(\{0,0\})$ are defined equivalently by $\mathbf{P}^*[\{K \subseteq \hat{S}_\lambda(I_1)\} \cap \{K \not\subseteq \hat{S}_\lambda(I_2)\}]$, $\mathbf{P}^*[\{K \not\subseteq \hat{S}_\lambda(I_1)\} \cap \{K \subseteq \hat{S}_\lambda(I_2)\}]$, and $\mathbf{P}^*[\{K \not\subseteq \hat{S}_\lambda(I_1)\} \cap \{K \not\subseteq \hat{S}_\lambda(I_2)\}]$, respectively. Note that $\hat{\Pi}_K^{\text{simult}}(\lambda) = s_K(\{1,1\})$, $s_K(\{1,0\}) = s_K(\{0,1\})$, and hence

$$\hat{\Pi}_K(\lambda) = s_K(\{1,0\}) + s_K(\{1,1\}) = s_K(\{0,1\}) + s_K(\{1,1\})$$
$$1 - \hat{\Pi}_K(\lambda) = s_K(\{0,1\}) + s_K(\{0,0\}) = s_K(\{1,0\}) + s_K(\{0,0\}).$$

As $s_K(\{0,0\}) \geq 0$, it also follows that $s_K(\{1,0\}) \leq 1 - \hat{\Pi}_K(\lambda)$. Hence

$$\hat{\Pi}_K^{\text{simult}}(\lambda) = s_K(\{1,1\}) = \hat{\Pi}_K(\lambda) - s_K(\{1,0\}) \geq 2\hat{\Pi}_K(\lambda) - 1,$$

which completes the proof. □

10.7.2 Proof of Theorem 10.1

The proof uses mainly Lemma 10.2 below. We first show that $\mathbf{P}(j \in \hat{S}_\Lambda) \leq q_\Lambda/p$ for all $j \in S_0^c$, using the made definitions $\hat{S}_\Lambda = \cup_{\lambda \in \Lambda} \hat{S}_\lambda$ and $q_\Lambda = \mathbb{E}[|\hat{S}_\Lambda|]$. Define furthermore $S_0^c(\Lambda) = S_0^c \cap \hat{S}_\Lambda$ to be the set of noise variables (in S_0^c) which appear in $\hat{S}_\Lambda$ and analogously $U_\Lambda = S \cap \hat{S}_\Lambda$. The expected number of falsely selected variables can be written as

$$\mathbb{E}[|S_0^c(\Lambda)|] = \mathbb{E}[|\hat{S}_\Lambda|] - \mathbb{E}[|U_\Lambda|] = q_\Lambda - \mathbb{E}[|U_\Lambda|].$$

Using the assumption (10.5) (which asserts that the method is not worse than random guessing), it follows that $\mathbb{E}[|U_\Lambda|] \geq \mathbb{E}[|S_0^c(\Lambda)|]|S|/|S_0^c|$. Putting together,

$$(1 + |S|/|S_0^c|)\mathbb{E}[|S_0^c(\Lambda)|] \leq q_\Lambda$$

and hence $|S_0^c|^{-1}\mathbb{E}[|S_0^c(\Lambda)|] \leq q_\Lambda/p$. Using the exchangeability assumption, we have $\mathbb{P}[j \in \hat{S}_\Lambda] = \mathbb{E}[|S_0^c(\Lambda)|]/|S_0^c|$ for all $j \in S_0^c$ and hence, for $j \in S_0^c$, it holds that $\mathbb{P}(j \in \hat{S}_\Lambda) \leq q_\Lambda/p$, as desired. Note that this result is independent of the sample size used in the construction of $\hat{S}_\lambda$, $\lambda \in \Lambda$. Now using Lemma 10.2 below, it follows that $\mathbb{P}[\max_{\lambda \in \Lambda} \hat{\Pi}_j^{\text{simult}}(\lambda) \geq \xi] \leq (q_\Lambda/p)^2/\xi$ for all $0 < \xi < 1$ and $j \in S_0^c$. Using Lemma 10.1, it follows that

$$\mathbb{P}[\max_{\lambda \in \Lambda} \hat{\Pi}_j(\lambda) \geq \pi_{\text{thr}}] \leq \mathbb{P}[(\max_{\lambda \in \Lambda} \hat{\Pi}_j^{\text{simult}}(\lambda) + 1)/2 \geq \pi_{\text{thr}}] \leq (q_\Lambda/p)^2/(2\pi_{\text{thr}} - 1).$$

Hence

$$\mathbb{E}[V] = \sum_{j \in S_0^c} \mathbb{P}[\max_{\lambda \in \Lambda} \hat{\Pi}_j(\lambda) \geq \pi_{\text{thr}}] \leq q_\Lambda^2/(p(2\pi_{\text{thr}} - 1)),$$

which completes the proof. □

Lemma 10.2. *Let $K \subseteq \{1,\ldots,p\}$ and $\hat{S}_\lambda$ the set of selected variables based on a sample size of $\lfloor n/2 \rfloor$.*
If $\mathbb{P}[K \subseteq \hat{S}_\lambda] \leq \varepsilon$, then $\mathbb{P}[\hat{\Pi}_K^{\text{simult}} \geq \xi] \leq \varepsilon^2/\xi$.
If $\mathbb{P}[K \subseteq \cup_{\lambda \in \Lambda}\hat{S}_\lambda] \leq \varepsilon$ for some $\Lambda \subseteq \mathbb{R}^+$, then $\mathbb{P}[\max_{\lambda \in \Lambda} \hat{\Pi}_K^{\text{simult}}(\lambda) \geq \xi] \leq \varepsilon^2/\xi$.

Proof. Let $I_1, I_2 \subseteq \{1,\ldots,n\}$ be, as above, the random split of the samples $\{1,\ldots,n\}$ into two disjoint subsets, where both $|I_i| = \lfloor n/2 \rfloor$ for $i = 1, 2$. Define the binary random variable H_K^λ for all subsets $K \subseteq \{1,\ldots,p\}$ as

$$H_K^\lambda = I\{K \subseteq \{\hat{S}_\lambda(I_1) \cap \hat{S}_\lambda(I_2)\}\},$$

where $1(\cdot)$ denotes the indicator function. Denote the data (the n samples) by Z. The simultaneous selection probability $\hat{\Pi}_K^{\text{simult}}(\lambda)$, as defined in (10.12), is then $\hat{\Pi}_K^{\text{simult}}(\lambda) = \mathbb{E}^*(H_K^\lambda) = \mathbb{E}(H_K^\lambda|Z)$, where the expectation $\mathbb{E}^*$ is with respect to the random split of the n samples into sets I_1 and I_2 (and additional randomness if $\hat{S}_\lambda$ is a randomized algorithm). To prove the first part, the inequality $\mathbb{P}[K \subseteq \hat{S}_\lambda(I_1)] \leq \varepsilon$ (for a sample size $\lfloor n/2 \rfloor$), implies that $\mathbb{P}[H_K^\lambda = 1] \leq \mathbb{P}[K \subseteq \hat{S}_\lambda(I_1)]^2 \leq \varepsilon^2$ and hence $\mathbb{E}[H_K^\lambda] \leq \varepsilon^2$. Therefore, $\mathbb{E}[H_K^\lambda] = \mathbb{E}[\mathbb{E}[H_K^\lambda|Z]] = \mathbb{E}[\hat{\Pi}_K^{\text{simult}}(\lambda)] \leq \varepsilon^2$. Using a Markov-type inequality, $\xi\mathbb{P}[\hat{\Pi}_K^{\text{simult}}(\lambda) \geq \xi] \leq \mathbb{E}[\hat{\Pi}_K^{\text{simult}}(\lambda)] \leq \varepsilon^2$. Thus $\mathbb{P}[\hat{\Pi}_K^{\text{simult}}(\lambda) \geq \xi] \leq \varepsilon^2/\xi$, completing the proof of the first claim. The proof of the second part follows analogously. □

Problems

10.1. Prove formula (10.7) by showing

$$\mathbf{P}[V > 0] \leq \mathbb{E}[V].$$

In addition, describe in words why the inequality above may be very coarse.

10.2. Consider a variable selection method (e.g. in a linear model) having the monotonicity property $\hat{S}(\lambda) \subseteq \hat{S}(\lambda')$ for all $\lambda \geq \lambda'$. Show that pointwise control with λ and control with $\Lambda = [\lambda, \infty)$ are equivalent.

10.3. Consider the problem of clustering n observations $X_1, \ldots, X_n$ into k different clusters $\mathscr{C}_1, \ldots, \mathscr{C}_k$ with $\mathscr{C}_j \subseteq \{1, \ldots, n\}$, $\cup_{j=1}^k \mathscr{C}_j = \{1, \ldots, n\}$ and $\mathscr{C}_i \cap \mathscr{C}_j = \emptyset$ for $i \neq j$. Show that membership of observations to the same cluster can be encoded by a graph (define the nodes and edges) and formulate selection and stability selection for the problem of clustering.

10.4. Exchangeability condition
Consider Example 10.1.
(a) Show that the distributions of

$$(Y, X^{(S_0)}, \{X^{(j)}; \ j \in S_0^c\}) \text{ and of } (Y, X^{(S_0)}, \{X^{(\pi(j))}; \ j \in S_0^c\})$$

are the same for any permutation $\pi : S_0^c \to S_0^c$.
(b) Consider the case of equicorrelation. Show that the irrepresentable condition fails for $p \geq 3$, $s_0 > p/2$ with $-1/(2s_0 - 1) > \rho > -1/(p-1)$.
Hint: Use that for equicorrelation with $\Sigma_{j,j} = 1$ $(j = 1, \ldots, p)$, $\Sigma_{j,k} = \rho$ $(j \neq k)$, the inverse is given by

$$\Sigma^{-1} = \frac{1}{1-\rho}(I_{p \times p} - \frac{\rho}{1+(p-1)\rho}\tau\tau^T), \ \tau = \tau_{p \times 1} = (1, 1, \ldots, 1).$$

See also Problem 2.4 and Problem 6.14.

10.5. Instead of stability selection as in (10.4), we can use the simultaneous selection probabilities from (10.12) and define

$$\hat{S}_{\text{stable}}^{\text{simult}} = \{j; \ \max_{\lambda \in \Lambda} \hat{\Pi}_j^{\text{simult}}(\lambda) \geq \pi_{\text{thr}}^{\text{simult}}\}.$$

Derive an error control in terms of $\mathbb{E}[V]$ as in Theorem 10.1, but now for $\hat{S}_{\text{stable}}^{\text{simult}}$ and using $\pi_{\text{thr}}^{\text{simult}}$ (i.e. the analogue of (10.6)).
Hint: use Lemma 10.1 and adapt the proof of Theorem 10.1 in Section 10.7.2.

Chapter 11
P-values for linear models and beyond

Abstract In the classical low-dimensional setup, error control of false selections based on p-values is a widely used standard in many areas of sciences. In the high-dimensional setting, assigning significance is challenging. Most computationally efficient selection algorithms cannot guard against inclusion of noise variables and hence, some thresholding operation on the basis of p-values for individual regression coefficients is desirable. We describe in this chapter a rather generic way to achieve this goal. Using multiple sample splitting, which is very simple to implement and bears similarities to subsampling and stability selection from Chapter 10, we show that such a random sampling scheme yields asymptotically valid p-values for controlling the familywise error or false discovery rate in high-dimensional linear or generalized linear models.

11.1 Organization of the chapter

We largely focus on assigning p-values for regression coefficients in linear models. After introducing the sample splitting method, we discuss in detail in Section 11.3 the much better multi sample splitting approach which has been originally proposed and studied in Meinshausen et al. (2009). Thereby, we describe p-values which achieve asymptotic error control for either the familywise error or the false discovery rate. Numerical illustrations are shown in Section 11.5 and consistent variable selection based on thresholding with p-values is discussed in Section 11.6. Extensions to other models and other error measures are discussed in Section 11.7. The technical proofs are collected in Section 11.8.

11.2 Introduction, sample splitting and high-dimensional variable selection

Constructing p-values belongs to the problem of quantifying uncertainty of estimators. In Chapter 10, we described an approach based on subsampling to assign uncertainty. In contrast to Theorem 10.1 for stability selection we show here that p-values for high-dimensional regression problems can be constructed under much weaker assumptions on the design than the exchangeability condition. While the main application of the p-value procedure is high-dimensional data, where the number p of variables can greatly exceed sample size n, we illustrate empirically that the method is also quite competitive in comparison to more standard error control for $n > p$ settings, indeed often giving a better detection power in the presence of highly correlated variables.

We consider the high-dimensional linear model similar to (10.1) from Chapter 10:

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon,$$

but here with fixed design and Gaussian errors $\varepsilon_1, \ldots, \varepsilon_n$ i.i.d. $\sim \mathcal{N}(0, \sigma^2)$. Extensions to other models are outlined in Section 11.7.

Denote by

$$S_0 = \{j; \ \beta_j^0 \neq 0\}$$

the set of active variables, where β^0 denotes the true parameter vector, and similarly by $S_0^c = \{j; \ \beta_j = 0\}$ the set of noise variables. Our goal is to assign p-values for the

$$\text{null-hypotheses } H_{0,j} : \beta_j = 0, \text{ versus the}$$
$$\text{alternatives } H_{A,j} : \beta_j \neq 0,$$

for all $j = 1, \ldots, p$. The modification to one-sided alternatives is straightforward but not treated in this chapter.

An approach proposed by Wasserman and Roeder (2009) is to split the data into two parts, reducing the dimensionality to a manageable size of predictors (keeping the important variables with high probability) using the first half of the data, and then to assign p-values and making a final selection using classical least squares estimation based on the second part of the data. The procedure of Wasserman and Roeder (2009) attempts to control the familywise error rate (FWER) which is defined as $\mathbf{P}[V > 0]$, the probability of making at least one false rejection where V denotes the number of false selections (i.e. false positives).

The data are split randomly into two disjoint sets $I_1, I_2 \subset \{1, \ldots, n\}$ with $|I_1| = \lfloor n/2 \rfloor$, $I_2 = n - \lfloor n/2 \rfloor$, $I_1 \cap I_2 = \emptyset$ and hence $I_1 \cup I_2 = \{1, \ldots, n\}$. Thus, the corresponding data sub-samples are $(\mathbf{X}_{I_1}, \mathbf{Y}_{I_1})$ and $(\mathbf{X}_{I_2}, \mathbf{Y}_{I_2})$. Let $\hat{S}$ be a variable selection or screening procedure. We denote by $\hat{S}(I_1)$ the set of selected predictors

based on $(\mathbf{X}_{I_1}, \mathbf{Y}_{I_1})$ which may include the choice of potential tuning or regularization parameters. A prime example for variable selection or screening is the Lasso in (2.2) but other methods such as boosting or forward selection from Chapter 12 could be considered as well. The regression coefficients and the corresponding p-values $\tilde{P}_1, \ldots, \tilde{P}_p$ of the selected predictors are determined based on the other half of the data $(\mathbf{X}_{I_2}, \mathbf{Y}_{I_2})$ by using ordinary least squares estimation and the corresponding t-tests on the set of variables from $\hat{S}(I_1)$ and setting $\tilde{P}_j = 1$ for all $j \notin \hat{S}(I_1)$, i.e.,

$$\tilde{P}_j = \begin{cases} P_{\text{raw},j} \text{ based on } \mathbf{Y}_{I_2}, \mathbf{X}_{I_2, \hat{S}(I_1)} & , \text{ if } j \in \hat{S}(I_1), \\ 1 & , \text{ if } j \notin \hat{S}(I_1), \end{cases} \qquad (11.1)$$

where $P_{\text{raw},j}$ is the p-value from the two-sided t-test, using least squares estimation, for $H_{0,j}$ (based on the second half of the data I_2 and using only the variables in $\hat{S}(I_1)$). If the selected set of variables contains the true model S_0, i.e.

$$\hat{S}(I_1) \supseteq S_0,$$

the p-values $\tilde{P}_j$ are controlling the (single testing) type I error, assuming Gaussian errors ε_i and $\text{rank}(\mathbf{X}_{I_2, \hat{S}(I_1)}) = |\hat{S}(I_1)|$. Regarding the latter, $\mathbf{X}_{I_2, \hat{S}(I_1)}$ is the design sub-matrix with rows corresponding to I_2 and columns corresponding to $\hat{S}(I_1)$. The assumption on the rank is most often fulfilled if $|\hat{S}(I_1)| < n/2$. Finally, each p-value $\tilde{P}_j$ is adjusted by a factor $|\hat{S}(I_1)|$ to correct for the multiplicity of the testing problem:

$$\tilde{P}_{\text{corr},j} = \min(\tilde{P}_j \cdot |\hat{S}(I_1)|, 1) \ (j = 1, \ldots, p). \qquad (11.2)$$

Assuming the conditions (11.3) and (11.4), the p-values in (11.2) control the familywise error rate (Problem 11.2). The conditions we need are:

$$\lim_{n \to \infty} \mathbf{P}[\hat{S}_{\lfloor n/2 \rfloor} \supseteq S_0] = 1. \qquad (11.3)$$

Furthermore, we assume

$$\lim_{n \to \infty} \mathbf{P}[|\hat{S}_{\lfloor n/2 \rfloor}| < n/2] = 1. \qquad (11.4)$$

Here, $\hat{S}_m$ denotes any variable selection procedure based on m observations. Some discussion about these conditions is given below. Of course, if (11.3) and (11.4) hold, they also hold for $\hat{S}(I_1)$. For any subset $I_{(m)} \subset \{1, \ldots, n\}$ with $|I_{(m)}| = m = n - \lfloor n/2 \rfloor$, let $\hat{\Sigma}(I_{(m)}) = m^{-1} \mathbf{X}_{I_{(m)}}^T \mathbf{X}_{I_{(m)}}$. Furthermore, we denote by $\hat{\Sigma}(I_{(m)})_{S,S}$ the $|S| \times |S|$ sub-matrix corresponding to rows and columns of the subset $S \subset \{1, \ldots, p\}$. We assume:

$$\Lambda_{\min}(\hat{\Sigma}(I_{(m)})_{S,S}) > 0 \quad \text{for all } S \text{ with } |S| < n/2,$$
$$\text{for all } I_{(m)} \text{ with } |I_{(m)}| = m = n - \lfloor n/2 \rfloor, \qquad (11.5)$$

where $\Lambda^2_{\min}(A)$ denotes the minimal eigenvalue of a symmetric matrix A. The condition in (11.5) is related to a sparse eigenvalue assumption, see Sections 6.13.5 and 10.6, but we do not require here a uniform positive lower bound.

For the Lasso, the screening property in (11.3) holds assuming a compatibility condition on the design, sparsity and a beta-min condition requiring sufficiently large (in absolute value) non-zero regression coefficients, see Corollary 7.6 in Chapter 7, and (11.4) is always true as described in Chapter 2, Section 2.5. The beta-min assumption about large non-zero regression coefficients might be unrealistic in practice: but without such a condition, we cannot detect small regression coefficients, see also at the end of Section 2.6 in Chapter 2, Section 7.4, and we also refer to Leeb and Pötscher (2005). Relaxation of (11.3) to some degree should be possible in order that the methods in this chapter for constructing p-values would still work. Other examples where (11.3) holds, assuming suitable conditions on the design and sufficiently large (in absolute value) non-zero regression coefficients, include the adaptive Lasso from Section 2.8 in Chapter 2, L_2 Boosting described in Section 12.4.4 in Chapter 12, orthogonal matching pursuit (Section 12.7.1 in Chapter 12) or Sure Independence Screening (Fan and Lv, 2008) (Section 13.9.5 in Chapter 13). The assumption in (11.4) is a sparsity property: if the true active set S_0 has cardinality less than $n/2$, fulfillment of (11.3) and (11.4) is possible (assuming in addition a beta-min and a design condition). Finally, (11.5) is a mild condition on the design since $m = |I_{(m)}| \geq n/2$ and the cardinality of the sets $|S| < n/2$: that is, the subsample size m is larger than the number of variables in S and thus, requiring positive definiteness (not with a uniform lower bound for the minimal eigenvalue) is not very restrictive. A violation of the sparsity property (11.4) or (11.5) would make it impossible to apply classical t-tests on the retained variables from $\hat{S}(I_1)$.

To formalize the control of the familywise error rate, we introduce some notation. The selected model is given by all variables for which the adjusted p-value in (11.2) is below a significance-cutoff $\alpha \in (0,1)$:

$$\hat{S}_{\text{single-split FWER}}(\alpha) = \{j; \ \tilde{P}_{\text{corr},j} \leq \alpha\}. \tag{11.6}$$

We denote the number of false positives (false selections) by

$$V_{\text{single-split FWER}}(\alpha) = |\hat{S}_{\text{single-split FWER}}(\alpha) \cap S_0^c|.$$

Then, we have the following results.

Lemma 11.1. *Consider the linear model as in (10.1) with fixed design and Gaussian errors* $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. *Assume that (11.3), (11.4) and (11.5) hold. Then, for any* $0 < \alpha < 1$:

$$\limsup_{n \to \infty} \mathbf{P}[V_{\text{single-split FWER}}(\alpha) > 0] \leq \alpha.$$

Proof. Consider the event $A_n = \{\hat{S}_{I_1} \supseteq S_0\} \cap \{|\hat{S}_{I_1}| < n/2\}$. By (11.3) and (11.4) we know that $\mathbf{P}[A_n] \to 1$ $(n \to \infty)$. Furthermore, on A_n, the p-values $\tilde{P}_{\text{corr},j}$ in (11.2)

control the familywise error rate, due to the fact that the data from I_2 are independent from the data in I_1, and because of the Gaussian assumption and (11.5) which ensure validity of the classical t-tests (Problem 11.2). Note that the Bonferroni correction factor in (11.2) is sufficient on A_n because we know that all variables which are not elements of $\hat{S}_{I_1}$ must be noise variables from S_0^c, due to (11.3). □

The single data-splitting method for the p-values in (11.2) is easy to implement. It relies, however, on an arbitrary split of the data into I_1 and I_2. Results, at least for finite samples, can change drastically if this split is chosen differently. This in itself is unsatisfactory since results are not reproducible, as illustrated in Figure 11.1, see also Problem 11.1.
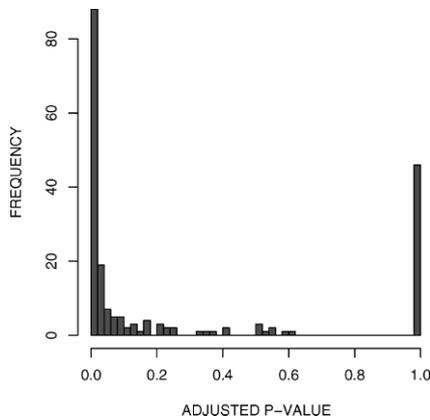
Fig. 11.1 Histogram of adjusted p-values $\tilde{P}_{\text{corr},j}$ for a single variable in the motif regression example of Section 2.5.2 with $n = 287$ and $p = 195$: the different p-values correspond to different random splits of the data into I_1 and I_2. Due to the high variability, we call the phenomenon a "p-value lottery". The figure is taken from Meinshausen et al. (2009).

11.3 Multi sample splitting and familywise error control

An obvious alternative and improvement to a single arbitrary sample split is to divide the sample repeatedly. For each split we end up with a set of p-values as in (11.2). It is not obvious, though, how to combine and aggregate these p-values from multiple sample splits. A possible solution has been worked out in Meinshausen et al. (2009).

For each hypothesis $H_{0,j}$, a distribution of p-values is obtained for multiple random sample splitting. We will show that error control can be based on the quantiles

of this distribution. In contrast to the "p-value lottery" phenomenon illustrated in
Figure 11.1, the multi sample split method makes results reproducible, at least ap-
proximately if the number of random splits is chosen to be sufficiently large (in
our presented theory, the number of random splits is fixed though). Moreover, we
will show empirically that, maybe unsurprisingly, the resulting procedure is more
powerful than the single-split method, see Section 11.5.

The multi sample split method is defined as follows:
For $b = 1, \ldots, B$:

1. Randomly split the original data into two disjoint groups $I_1^{(b)}$ and $I_2^{[b]}$ of (almost)
   equal size.

2. Using only $I_1^{[b]}$, estimate the set of active predictors $\hat{S}^{[b]} = \hat{S}(I_1^{[b]})$.

3. Using only $I_2^{[b]}$, compute the adjusted (non-aggregated) p-values as in (11.2), i.e.,

$$\tilde{P}_{\text{corr},j}^{[b]} = \min(\tilde{P}_j^{[b]} \cdot |\hat{S}^{[b]}|, 1) \ (j = 1, \ldots, p)$$

   where $\tilde{P}_j^{[b]}$ is based on the two-sided t-test, as in (11.1), based on $I_2^{[b]}$ and $\hat{S}^{[b]} = \hat{S}(I_1^{[b]})$.

Finally, we aggregate over the B p-values $\tilde{P}_{\text{corr},j}^{[b]}$, as discussed next.

11.3.1 Aggregation over multiple p-values

The procedure described above leads to a total of B p-values for each covariate
$j = 1, \ldots, p$. For each $j = 1, \ldots, p$, the goal is to aggregate the p-values $\tilde{P}_{\text{corr},j}^{[b]}$ over
the indices $b = 1, \ldots, B$. This can be done using quantiles. For $\gamma \in (0,1)$ define

$$Q_j(\gamma) = \min\left\{ q_\gamma(\{\tilde{P}_{\text{corr},j}^{[b]}/\gamma; b = 1, \ldots, B\}), 1 \right\}, \tag{11.7}$$

where $q_\gamma(\cdot)$ is the (empirical) γ-quantile function.

A p-value for each variable $j = 1, \ldots, p$ is then given by $Q_j(\gamma)$, for any fixed $0 < \gamma <
1$. We will describe in Section 11.3.2 that this is an asymptotically correct p-value
for controlling the familywise error rate.

A proper selection of γ may be difficult. Error control is not guaranteed anymore if
we search for the best value of γ. But we can use instead an adaptive version which
selects a suitable value of the quantile based on the data. Let $\gamma_{\min} \in (0,1)$ be a lower
bound for γ, typically 0.05, and define

$$P_j = \min \left\{ \left(1 - \log \gamma_{\min} \right) \inf_{\gamma \in (\gamma_{\min}, 1)} Q_j(\gamma), 1 \right\} \quad (j = 1, \ldots, p). \qquad (11.8)$$

The extra correction factor $1 - \log \gamma_{\min}$ ensures that the familywise error rate remains controlled despite of the adaptive search for the best quantile, as described in Theorem 11.1 in Section 11.3.2. For the recommended choice of $\gamma_{\min} = 0.05$, this factor is upper bounded by 4; in fact, $1 - \log(0.05) \approx 3.996$.

Figure 11.2 takes up again the example from Figure 11.1 to illustrate the difference between the single sample split and the multi sample split method. The left panel contains the histogram of the adjusted p-values $\tilde{P}^{[b]}_{\text{corr}, j}$ for $b = 1, \ldots, B$ of a particular (selected) variable j of a motif regression problem using a linear model with $n = 287$ and $p = 195$, as described in Section 2.5.2. The single sample split method is equivalent to picking one of these p-values randomly and selecting the variable if this randomly picked p-value is sufficiently small, say less than a significance level α. To avoid this "p-value lottery", the multi sample split method computes the empirical distribution of all p-values $\tilde{P}^{[b]}_{\text{corr}, j}$ for $b = 1, \ldots, B$ and rejects the null hypothesis $H_{0,j} : \beta_j = 0$ if the empirical distribution crosses from above the broken line in the right panel of Figure 11.2. A short derivation of the latter is as follows. Variable j is selected if and only if $P_j \le \alpha$, see (11.8). Using $\gamma_{\min} = 0.05$, this happens if and only if there exists some $\gamma \in (0.05, 1)$ such that $Q_j(\gamma) \le \alpha/(1 - \log 0.05) = \alpha/3.996$. Equivalently, using definition (11.7), the γ-quantile of the adjusted p-values, $q_\gamma(\{\tilde{P}^{[b]}_{\text{corr}, j}\}_b)$, has to be smaller than or equal to $\alpha\gamma/3.996$. This in turn is equivalent to the event that the empirical distribution of the adjusted p-values $\{\tilde{P}^{[b]}_{\text{corr}, j}; b = 1, \ldots, B\}$ is crossing from above the bound $f(p) = \max\{0.05, (3.996/\alpha)p\}$ for some $p \in (0, 1)$. This bound is shown as a broken line in the right panel of Figure 11.2.

11.3.2 Control of familywise error

The resulting adjusted p-values P_j $(j = 1, \ldots, p)$ from (11.8) can be used for both familywise error (FWER) and false discovery rate (FDR) control. For FWER control at level $\alpha \in (0, 1)$, simply all p-values below α are rejected and the selected subset is

$$\hat{S}_{\text{multi-split FWER}}(\alpha) = \{j : P_j \le \alpha\}. \qquad (11.9)$$

Denote by $V_{\text{multi-split FWER}}(\alpha) = |\hat{S}_{\text{multi-split FWER}}(\alpha) \cap S_0^c|$ the number of false positive selections.

Theorem 11.1. *Consider a linear model as in (10.1) with fixed design and Gaussian errors and assume that (11.3), (11.4) and (11.5) hold. Furthermore, the number B of random splits in the multi-split method is fixed. Then, for any $\gamma_{\min} \in (0, 1)$ (see (11.8)),*
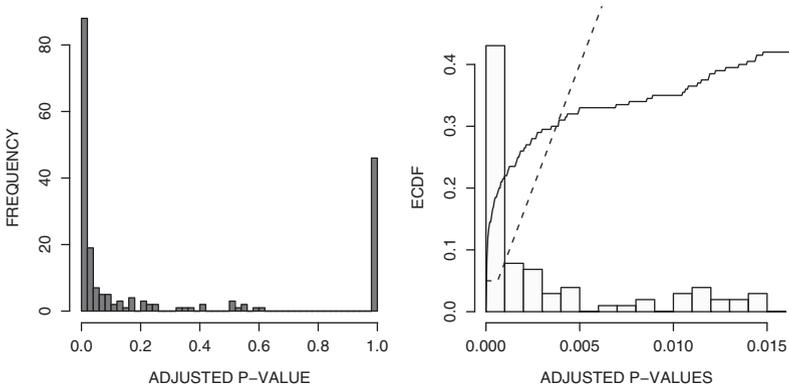
Fig. 11.2 Left: a histogram of adjusted p-values $\tilde{P}_{\text{corr},j}^{[b]}$ for a single variable in the motif regression example of Section 2.5.2 with $n = 287$ and $p = 195$. This is the same plot as in Figure 11.1. The single data split method picks randomly one of these p-values (a "p-value lottery") and rejects $H_{0,j}$ if it is below α. For the multi data split method, we reject $H_{0,j}$ if and only if the empirical distribution function of the adjusted p-values crosses from above the broken line (which is $f(p) = \max\{0.05, (3.996/\alpha)p\}$) for some value $p \in (0,1)$ on the x-axis. This bound is shown as a broken line for $\alpha = 0.05$. For the given example, the bound is indeed exceeded and the variable is thus selected. The figure is taken from Meinshausen et al. (2009).

$$\limsup_{n \to \infty} \mathbf{P}[V_{\text{multi–split FWER}}(\alpha) > 0] \leq \alpha.$$

A proof is given in Section 11.8. Instead of working with the adaptive p-values in (11.8), we could use an empirical quantile $Q_j(\gamma)$ from (11.7) for fixed $0 < \gamma < 1$. We then define

$$\hat{S}_{\text{multi–split FWER}}(\alpha|\gamma) = \{j;\ Q_j(\gamma) \leq \alpha\},$$

and define $V_{\text{multi–split FWER}}(\alpha|\gamma) = \hat{S}_{\text{multi–split FWER}}(\alpha|\gamma) \cap S_0^c$. Then, the following error control holds.

Proposition 11.1. *Consider a linear model as in (10.1) with fixed design and Gaussian errors and assume that (11.3), (11.4) and (11.5) hold. Furthermore, the number B of random splits in the multi-split method is fixed. Then, for $0 < \gamma < 1$, and where*

$$\limsup_{n \to \infty} \mathbf{P}[V_{\text{multi–split FWER}}(\alpha|\gamma) > 0] \leq \alpha.$$

A proof is given in Section 11.8. Besides better reproducibility, the multi sample split version is empirically found to be more powerful than the single split selection method (Section 11.5). Regarding asymptotic familywise error control, the presented theory does not allow for a distinction between the single split method (with fixed $B = 1$) and the multi sample split procedure (with fixed $B > 1$).

11.4 Multi sample splitting and false discovery rate

Control of the familywise error rate is often considered as too conservative. If many rejections are made, Benjamini and Hochberg (1995) proposed to control instead the expected proportion of false rejections, the false discovery rate (FDR). Let $V = |\hat{S} \cap S_0^c|$ be the number of false rejections for a selection method $\hat{S}$ and denote by $R = |\hat{S}|$ the total number of rejections. The false discovery rate is defined as the expected proportion of false rejections

$$\mathbb{E}[Q] \text{ with } Q = V/\max\{1, R\}. \tag{11.10}$$

For no rejections with $R = 0$, the denominator ensures that the false discovery proportion Q is 0, conforming with the definition in Benjamini and Hochberg (1995).

The original FDR controlling procedure in Benjamini and Hochberg (1995) first orders the observed raw p-values as $P_{\text{raw},(1)} \leq P_{\text{raw},(2)} \leq \ldots \leq P_{\text{raw},(p)}$ and defines

$$k = \max\{i : P_{\text{raw},(i)} \leq \frac{i}{p} q\} \ (0 < q < 1). \tag{11.11}$$

Then all variables or hypotheses with the smallest k p-values are rejected and no rejection is made if the set in (11.11) is empty. The FDR is controlled this way at level q under the condition that all p-values are independent. It has been shown in Benjamini and Yekutieli (2001) that the procedure is conservative under a wider range of dependencies among p-values; see also Blanchard and Roquain (2008) for related work. It would, however, require a big leap of faith to assume any such assumption for our setting of high-dimensional regression. For general dependencies, Benjamini and Yekutieli (2001) showed that control is guaranteed at level $q \sum_{i=1}^{p} i^{-1} \approx q(1/2 + \log(p))$, see also Theorem 11.2.

The standard FDR procedure is working with the raw p-values, which are assumed to be uniformly distributed on $[0, 1]$ for true null hypotheses. The division by p in (11.11) is an effective correction for multiplicity. The multi-split method in Section 11.3.1, however, is producing already adjusted p-values as in (11.8). Working with multiplicity-corrected p-values, the division by p in (11.11) turns out to be superfluous. Instead, we can order the corrected p-values P_j ($j = 1, \ldots, p$) from (11.8) in increasing order $P_{(1)} \leq P_{(2)} \leq \ldots \leq P_{(p)}$ and select the h variables with the smallest p-values, where

$$h = h(q) = \max\{i : P_{(i)} \leq iq\}. \tag{11.12}$$

The selected set of variables is denoted, with the value of $h = h(q)$ given in (11.12), by

$$\hat{S}_{\text{multi-split FDR}}(q) = \{j : P_j \leq P_{(h(q))}\},$$

with no rejections and $\hat{S}_{\text{multi-split FDR}}(q) = \emptyset$, if $P_{(i)} > iq$ for all $i = 1, \ldots, p$.

This procedure will achieve FDR control at level

$$q\sum_{i=1}^{p} i^{-1} \approx q(1/2 + \log p),$$

as discussed in Section 11.8.3. Thus, to get FDR control at level $\alpha \in (0,1)$ we define:

$$\hat{S}_{\text{multi−split FDR}}(\alpha) = \{j : P_j \leq P_{(h(q(\alpha)))}\},$$
$$h = h(q(\alpha)) \text{ as in (11.12)}, \ q(\alpha) = \frac{\alpha}{\sum_{i=1}^{p} i^{-1}}. \tag{11.13}$$

We will show error control in the following section and demonstrate empirically in Section 11.5 the advantages of the multi split version over both the single split and standard FDR controlling procedures.

11.4.1 Control of false discovery rate

The adjusted p-values can be used for FDR control, as described above in Section 11.4. Let $\hat{S}_{\text{multi−split FDR}}(\alpha)$ be the set of selected variables, as defined in (11.13). Denote by $V_{\text{multi−split FDR}}(\alpha)) = |\hat{S}_{\text{multi−split FDR}}(\alpha) \cap S_0^c|$ and $R_{\text{multi−split FDR}}(\alpha) = |\hat{S}_{\text{multi−split FDR}}(\alpha)|$. We then have the following result.

Theorem 11.2. *Consider a linear model as in (10.1) with fixed design and Gaussian errors and assume that (11.3), (11.4) and (11.5) hold. Furthermore, the number B of random splits in the multi-split method is fixed. Then, for any $\gamma_{\min} \in (0,1)$ (see (11.8)),*

$$\limsup_{n \to \infty} \mathbb{E}[Q(\alpha)] \leq \alpha,$$
$$Q(\alpha) = V_{\text{multi−split FDR}}(\alpha)/\max\{1, R_{\text{multi−split FDR}}(\alpha)\}.$$

A proof is given in Section 11.8. As with FWER-control in Section 11.3.2, we could be using, for any fixed value of γ, the values $Q_j(\gamma)$ $(j = 1, \ldots, p)$ in (11.7) instead of P_j $(j = 1, \ldots, n)$ in (11.8). Then, the analogue of Proposition 11.1 also holds for the FDR controlling procedure described above.

11.5 Numerical results

In this section we consider the empirical performance of the p-value method in conjunction with $\hat{S}$ from (versions of) the Lasso as a variable screening method. We use a default significance value of $\alpha = 0.05$ everywhere.

11.5.1 Simulations and familywise error control

We simulate data from a linear model,

$$Y_i = \sum_{j=1}^{p} \beta_j X_i^{(j)} + \varepsilon_i \ (i = 1, \ldots, n),$$

where $\varepsilon_1, \ldots, \varepsilon_n$ i.i.d. $\sim \mathcal{N}(0, \sigma^2)$, with random design $X_1, \ldots, X_n$ i.i.d $\mathcal{N}_p(0, \Sigma)$ and a sparse β-vector with active set S_0 and $s_0 = |S_0|$. In each simulation run, a new parameter vector β is created by either "uniform" or "varying-strength" sampling. Under "uniform" sampling, s_0 randomly chosen components of β are set to 1 and the remaining $p - s_0$ components to 0. Under "varying-strength" sampling, s_0 randomly chosen components of β are set to values $1, \ldots, s_0$. The error variance σ^2 is adjusted such that the signal to noise ratio (SNR) is maintained at a desired level at each simulation run (see Table 11.1). We consider two classes of scenarios with:

(A) $n = 100$, $p = 100$ and a Toeplitz covariance matrix Σ with $\Sigma_{j,k} = \rho^{|j-k|}$.

(B)   As (A) but with $p = 1000$.

For (A) and (B), we vary in each setting the SNR to 0.25, 1, 4 and 16, and the number s_0 of active variables is either 5 or 10; we fix the value of $\rho = 0.5$. Thus, we consider for each setting (A) and (B) 16 different scenarios (4 different SNRs, 2 different sparsity values s_0 and 2 different sampling schemes for β). We perform 50 simulations for each scenario.

As initial variable selection or screening method $\hat{S}$ we use three approaches which are all based on the Lasso. The first one, denoted by $\hat{S}_{\text{fixed}}$, uses the Lasso and selects those $\lfloor n/6 \rfloor$ variables which appear most often in the regularization path when varying the penalty parameter. The constant number of $\lfloor n/6 \rfloor$ variables is chosen, somewhat arbitrarily, to ensure a reasonably large set of selected coefficients on the one hand and to ensure, on the other hand, that least squares estimation will work reasonably well on the second half of the data with sample size $n - \lfloor n/2 \rfloor$. The second method $\hat{S}_{CV}$ is more data-driven: it uses the Lasso with penalty parameter chosen by 10-fold cross-validation and selecting the variables whose corresponding estimated regression coefficients are different from zero. The third method, $\hat{S}_{\text{adapt}}$ is the adaptive Lasso, with the Lasso solution used as initial estimator for the adaptive

Lasso, and where the regularization parameters are chosen based on 10-fold cross-validation. The selected variables are again the ones whose corresponding estimated regression parameters are different from zero. The number of random splits in the multi sample split method is always chosen as $B = 100$.

The average number of true positives and the familywise error rate (FWER) for the single and multi sample split methods are considered. Results are shown in Figure 11.3 with the default value $\gamma_{min} = 0.05$ in (11.8). Using the multi sample split
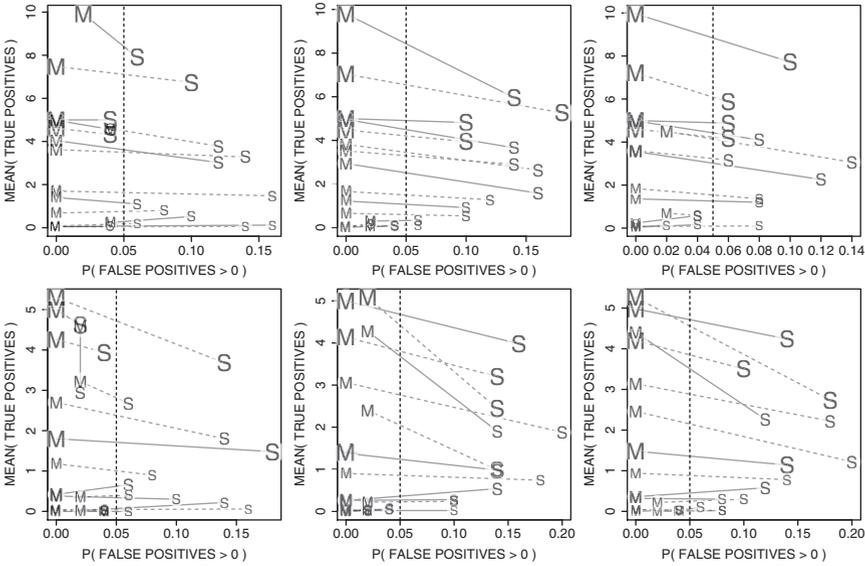

Fig. 11.3 Simulation results for setting (A) in the top and (B) in the bottom row. Average number of true positives vs. the familywise error rate (FWER) for the single split method ("S") against the multi-split version ("M"). FWER is controlled (asymptotically) at $\alpha = 0.05$ for both methods and this value is indicated by a broken vertical line. From left to right are results for $\hat{S}_{fixed}$, $\hat{S}_{CV}$ and $\hat{S}_{adapt}$. Results of a single scenario with specific values of (n, p), SNR and sparsity are joined by a line, which is solid if the regression coefficients follow the "uniform" sampling and broken otherwise. Increasing SNR is indicated by increasing symbol size. The figure is taken from Meinshausen et al. (2009).

method, the average number of true positives (the variables in S_0 which are selected) is typically slightly increased while the FWER (the probability of selecting variables in S_0^c) is reduced sharply. The single sample split method has often a FWER above the level $\alpha = 0.05$ at which it is asymptotically controlled while for the multi sample split method, the FWER is above the nominal level only in few scenarios. The asymptotic control seems to give a good control in finite sample settings with the multi sample split method. The single sample split method, in contrast, selects in nearly all cases too many noise variables, exceeding the desired FWER sometimes substantially. This suggests that the error control for finite sample sizes works much better for the multi sample split method yet with a larger number of true discoveries.

11.5.1.1 Comparisons with adaptive Lasso

Next, we compare the multi sample split method with the adaptive Lasso described in Section 2.8 in (2.8). We have used the adaptive Lasso previously as a variable selection or screening method in the multi sample split procedure. But the adaptive Lasso is often employed on its own and we use the same choices as previously: the initial estimator is obtained as the Lasso solution with a 10-fold CV-choice of the penalty parameter, and the regularization parameter in the adaptive Lasso penalty is also obtained by 10-fold CV. Table 11.1 shows the simulation results for the multi sample split method using $\hat{S}_{\text{adapt}}$ and the adaptive Lasso on its own, side by side for a simulation setting with $n = 100$, $p = 200$ and the same settings as in (A) and (B) otherwise. The adaptive Lasso selects roughly 20 noise variables (out of $p = 200$ variables), even though the number of truly relevant variables is just 5 or 10. The average number of false positives is at most 0.04 and often simply 0 with the multi sample split method. There is clearly a price to pay for controlling the

Uniform			E(True Positives)		E(False Positives)		P(False Positives > 0)	
			Multi	Adaptive	Multi	Adaptive	Multi	Adaptive
Sampling	$|S_0|$	SNR	Split	Lasso	Split	Lasso	Split	Lasso
NO	10	0.25	0.00	2.30	0	9.78	0	0.76
NO	10	1	0.58	6.32	0	20.00	0	1
NO	10	4	4.14	8.30	0	25.58	0	1
NO	10	16	7.20	9.42	0.02	30.10	0.02	1
YES	10	0.25	0.02	2.52	0	10.30	0	0.72
YES	10	1	0.10	7.46	0.02	21.70	0.02	1
YES	10	4	2.14	9.96	0	28.46	0	1
YES	10	16	9.92	10.00	0.04	30.66	0.04	1
NO	5	0.25	0.06	1.94	0	11.58	0	0.84
NO	5	1	1.50	3.86	0.02	19.86	0.02	1
NO	5	4	3.52	4.58	0.02	23.56	0.02	1
NO	5	16	4.40	4.98	0	27.26	0	1
YES	5	0.25	0.02	2.22	0	12.16	0	0.8
YES	5	1	0.82	4.64	0.02	22.18	0.02	1
YES	5	4	4.90	5.00	0	24.48	0	1
YES	5	16	5.00	5.00	0	28.06	0	1

Table 11.1 Comparing the multi sample split method using the adaptive Lasso $\hat{S}_{\text{adapt}}$ (Multi Split) with the variable selection made by the plain adaptive Lasso with a CV-choice of the involved penalty parameters (Adaptive Lasso) for settings as in (A) or (B) but with $n = 100$ and $p = 200$.

familywise error rate: the multi sample split method detects on average less truly relevant variables than the adaptive Lasso. For very low SNR, the difference is most pronounced. The multi sample split method selects in general neither correct nor wrong variables for SNR $= 0.25$, while the adaptive Lasso averages between 2 to 3 correct selections, among 9-12 wrong selections. Depending on the objectives of the study, one would prefer either of the outcomes. For larger SNR, the multi sample

split method detects almost as many truly important variables as the adaptive Lasso, while still reducing the number of falsely selected variables from 20 or above to roughly 0.

The multi sample split method seems most beneficial in settings where the cost of making an erroneous selection is rather high. For example, expensive follow-up experiments in biomedical applications are usually required for scientific validation, and a stricter error control will place more of the available resources into experiments which are likely to be successful.

11.5.2 Familywise error control for motif regression in computational biology

We apply the multi sample split method to real data about motif regression to find binding sites for the HIF1α transcription factor. More details about the problem and the data are given in Section 2.5.2. Here, the motif regression problem amounts to variable selection in a linear model with sample size $n = 287$ and $p = 195$ covariates.

We use the multi sample split method with the adaptive Lasso $\hat{S}_{\text{adapt}}$ as described in Section 11.5.1. The multi sample split method identifies one variable at the 5% significance level with an adjusted p-value of 0.0059, see (11.8). The single sample split method is not able to identify a single significant predictor. In view of the asymptotic error control in Theorem 11.1 and the empirical results in Section 11.5.1, there is substantial evidence that the single selected variable is a truly relevant variable. For this specific application it seems desirable to pursue a conservative approach with FWER control.

### 11.5.3 Simulations and false discovery rate control

We now look empirically at the behavior of the multi sample split method for FDR control, discussed in Section 11.4, and its power to detect truly interesting variables. We use the Lasso with cross-validation for choosing the regularization parameter, denoted by $\hat{S}_{CV}$, for the variable selection or screening step. Turning again to the simulation setting (A) in Section 11.5.1, we vary the sample size n, the number of variables p, the signal to noise ratio SNR, the correlation ρ in the Toeplitz design covariance matrix and the number $s_0 = |S_0|$ of active variables.

We have empirically illustrated above that the multi sample split method is preferable to the single-split method for familywise error control. Here, we are more interested in a comparison to well understood traditional FDR controlling procedures. For $p < n$, the standard approach would be to compute the least squares estima-

tor once for the full dataset. For each variable, a p-value is obtained and the FDR controlling procedure as in (11.11) can be applied. This approach obviously breaks down for $p > n$ whereas the multi sample split method can be applied both to low-dimensional ($p < n$) and high-dimensional ($p \geq n$) settings.

In all settings, the true FDR of the multi sample split method is often close to zero and always below the controlled value of $\alpha = 0.05$. Results regarding power are shown in Figure 11.4 for control at $\alpha = 0.05$. The multi sample split method tracks the power of the standard FDR controlling procedure quite closely for low-dimensional data with $p < n$. In fact, the multi data split method is doing considerably better if n/p is below, say, 1.5 or the correlation among the variables (and hence among the tests) is large. An intuitive explanation for this behavior is that, as p approaches n, the variance of all OLS components $\hat{\beta}_j$ ($j = 1, \ldots, p$) is increasing and it reduces the ability to select the truly important variables. The multi sample split method, in contrast, trims the total number of variables to a substantially smaller number on one half of the samples and then suffers less from an increased variance in the estimated coefficients on the second half of the samples. Repeating this over multiple splits leads to a surprisingly powerful variable selection procedure even for low-dimensional data.
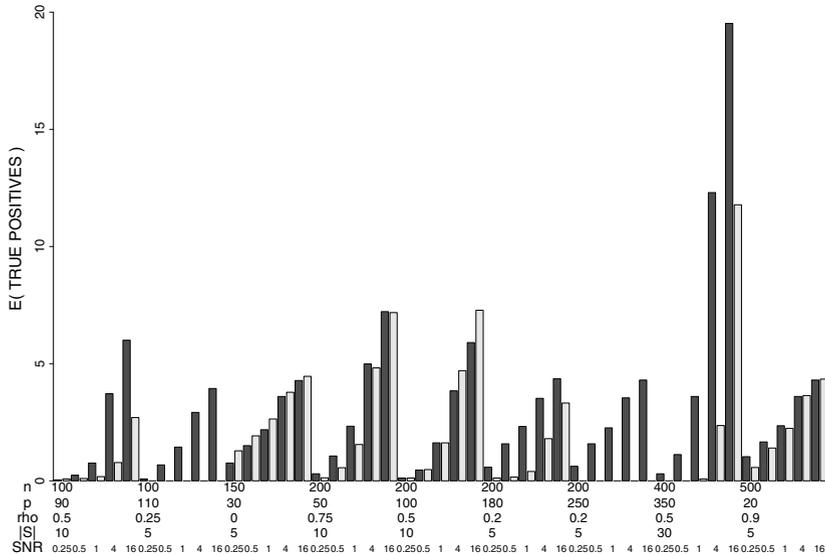
Fig. 11.4 Power of FDR controlling procedures: the multi sample split method (dark bar) and standard FDR control (light bar). The settings of n, p, ρ, s_0 and SNR are given for nine different scenarios. The height of the bars corresponds to the average number of selected relevant variables. For $p > n$, the standard method breaks down and the corresponding bars are set to height 0. The figure is taken from Meinshausen et al. (2009).

11.6 Consistent variable selection

We will show here that the sample splitting procedure, in particular the multi sample split version, lead to consistent variable selection for a high-dimensional linear model. The single sample split method (which is not advised to be used) can be viewed as some sort of two-stage procedure: any reasonable variable selection or screening method $\hat{S}$ is followed by least squares estimation on the selected variables. In contrast to the Lasso-OLS hybrid or the more general relaxed Lasso, see Section 2.10, or also in contrast to the adaptive or thresholded Lasso, see Sections 2.8 and 2.9, the two-stage procedure here is based on independent half-samples.

If we let the significance level $\alpha = \alpha_n \to 0$ for $n \to \infty$, the probability of falsely including a noise variable vanishes because of Proposition 11.1 or Theorem 11.1 on familywise error control. In order to get consistent variable selection, in the sense that

$$\mathbf{P}[\hat{S} = S_0] \to 1 \ (n \to \infty),$$

where $\hat{S} = \hat{S}_{\text{single-split FWER}}$ as in (11.6) or $\hat{S}_{\text{multi-split FWER}}$ as in (11.9), we have to analyze the asymptotic behavior of the power. This will be discussed next.

11.6.1 Single sample split method

We present now sufficient conditions for consistent variable selection with the single sample split method. Although the method is not advised to be used, see the "p-value lottery" in Figure 11.1, it is easier to understand the mathematical properties for this procedure first and then analyze the multi sample splitting method afterwards. We assume that the splitting into I_1 and I_2 is fixed (e.g. a fixed realization of a random splitting mechanism). We make the following assumptions. First, we need a slightly stronger sparsity property:

$$\mathbf{P}[|\hat{S}_{\lfloor n/2 \rfloor}| \le a_n] \to 1, \ \ a_n = o(n). \tag{11.14}$$

Note that together with the screening property (11.3), this implies

$$s_0 = |S_0| \le a_n = o(n). \tag{11.15}$$

We also refer to Corollary 7.10 and Corollary 7.14 in Chapter 7 showing that the Lasso has no more than $O(s_0)$ false positive selections under strong conditions while the adaptive Lasso achieves $O(s_0)$ false positives under more relaxed assumptions, see Section 7.11.4. Furthermore, we need an assumption regarding the power of the individual tests. Denote by $T_{I_2,S;j}$ the t-statistics for variable j based on the second half of the sample I_2 and based on a linear model with covariates from the set $S \subseteq$

$\{1,\ldots,p\}$ (assuming implicitly that $j \in S$). In other words,

$$T_{I_2,S;j} = \frac{\hat{\beta}_{I_2;j|S}^{\text{OLS}}}{\hat{\sigma}_{I_2;S}^{\text{OLS}} \sqrt{(\mathbf{X}_{I_2,S}^T \mathbf{X}_{I_2,S})_{j,j}^{-1}}}, \tag{11.16}$$

where $\hat{\beta}_{I_2;j|S}^{\text{OLS}}$ and $\hat{\sigma}_{I_2;S}^{\text{OLS}}$ are the standard ordinary least squares estimators based on sub-sample I_2 and covariates from S. We then assume

$$\sup_{S \in \mathscr{S}} \sum_{j \in S_0} \mathbf{P}[|T_{I_2,S;j}| < t(1 - \alpha_n/2, |I_2| - |S|)] \to 0 \ (n \to \infty),$$
$$\mathscr{S} = \{S; \ S_0 \subseteq S, \ |S| \le a_n\}, \tag{11.17}$$

where a_n is as in (11.14). Thereby, $t(\gamma, m)$ denotes the γ-quantiles of a t_m-distribution. We note that the degrees of freedom above are lower bounded by

$$|I_2| - |S| \ge n - \lfloor n/2 \rfloor - a_n \to \infty \ (n \to \infty),$$

since $a_n = o(n)$. We will show in Lemma 11.3 that (11.17) holds assuming that the true absolute coefficients $|\beta_j|$ are sufficiently large for all $j \in S_0$ and some rather weak regularity condition on the design.

Lemma 11.2. *Consider the linear model as in (10.1) with fixed design and Gaussian errors $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. Assume the screening property (11.3), the sparsity property (11.14) and (11.17). Then, $\mathbf{P}[\hat{S}_{\text{single-split FWER}}(\alpha_n) \supseteq S_0] \to 1 \ (n \to \infty)$.*

Proof.

$$\mathbf{P}[\cap_{j \in S_0^c} \{j \in \hat{S}_{\text{single-split FWER}}(\alpha_n)\}] = 1 - \mathbf{P}[\cup_{j \in S_0^c} \{j \notin \hat{S}_{\text{single-split FWER}}(\alpha_n)\}]$$
$$\ge 1 - \sum_{j \in S_0^c} \mathbf{P}[j \notin \hat{S}_{\text{single-split FWER}}(\alpha_n)]$$
$$= 1 - \sum_{j \in S_0^c} \mathbf{P}[|T_{I_2,\hat{S}_{I_1};j}| < t(1 - \alpha_n/2, |I_2| - |\hat{S}_{I_1}|)] \to 1 \ (n \to \infty),$$

where the last convergence is due to assumption (11.17). $\qquad\qquad\square$

11.6.1.1 Verification of condition (11.17)

We give here sufficient conditions such that (11.17) holds. Assume that for $I_2 = I_{2,n}$ as n is growing,

$$\inf_{j \in S_0} |\beta_j| \ge C_n n^{-1/2} \sqrt{\log(|S_0|) \sup_{S \in \mathscr{S}} (|I_{2,n}|^{-1} \mathbf{X}_{I_{2,n},S}^T \mathbf{X}_{I_{2,n},S})_{j,j}^{-1}}$$

for some $C_n \to \infty$ growing arbitrarily slowly. $\tag{11.18}$

This condition holds as follows: assume for $I_2 = I_{2,n}$ and for some $M < \infty$,

$$\sup_{S \in \mathscr{S}, n \in \mathbb{N}} (|I_{2,n}|^{-1} \mathbf{X}_{I_{2,n},S}^T \mathbf{X}_{I_{2,n},S})_{j,j}^{-1} \leq M < \infty, \tag{11.19}$$

then (11.18) is implied by

$$\inf_{j \in S_0} |\beta_j| \geq C_n n^{-1/2} \sqrt{\log(|S_0|)}, \tag{11.20}$$

where the constant M is absorbed into C_n. Furthermore, assumption (11.19) holds if for some $D > 0$,

$$\inf_{S \in \mathscr{S}, n \in \mathbb{N}} \Lambda_{\min}(|I_{2,n}|^{-1} \mathbf{X}_{I_{2,n},S}^T \mathbf{X}_{I_{2,n},S}) \geq D > 0, \tag{11.21}$$

where $\Lambda_{\min}^2(A)$ denotes the minimal eigenvalue of a symmetric matrix A. We leave the derivation as Problem 11.3. Assumption (11.21) is a sparse eigenvalue assumption as used in Section 10.6 and discussed in Section 6.13.5. (We note that the beta-min condition below in (11.22), needed for variable screening, is sufficient for (11.20) if $|S_0| \sqrt{\log(p)/\log(|S_0|)}$ is growing with n).

Lemma 11.3. *Consider the linear model as in (10.1) with fixed design and Gaussian errors $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. Assume that (11.18) holds. Then there exists a sequence $\alpha_n \to 0$ ($n \to \infty$) such that condition (11.17) holds (for such an α_n).*

A proof is given in Section 11.8. The sequence α_n could be made more explicit as a function of the value a_n in (11.14) and the lower bound for the absolute regression coefficients in (11.18).

We now give some sufficient conditions for consistent variable selection with the single sample split method.

Corollary 11.1. *Consider the linear model as in (10.1) with fixed design and Gaussian errors. Assume the screening property (11.3), the sparsity property from (11.4), the assumption on the non-zero regression coefficients from (11.18) and a design condition as in (11.21). Then, there exists a sequence $\alpha_n \to 0$ ($n \to \infty$) such that the single sample split FWER procedure satisfies:*

$$\mathbf{P}[\hat{S}_{\text{single}-\text{split FWER}}(\alpha_n) = S_0] \to 1 \ (n \to \infty).$$

The corollary follows from Lemma 11.1 (which holds also for $\alpha = \alpha_n$ depending on n in an arbitrary way; see the proof), Lemma 11.2 and 11.3.

Using the Lasso as variable screening procedure $\hat{S}$, the screening property in (11.3) follows from a compatibility condition on the design and a beta-min condition on the non-zero regression coefficients of the form

$$\inf_{j \in S_0} |\beta_{0,j}| \geq C|S_0| \sqrt{\frac{\log(p)}{n}}. \tag{11.22}$$

For more details, see Section 2.5 and Section 6.2.2. Condition (11.22) together with (11.21) implies (11.18). We note that (11.22) might be overly unrealistic in practice: this problem is briefly discussed at the end of Section 2.6 in Chapter 2 and in Section 7.4.

11.6.2 Multi sample split method

It turns out that variable selection consistency of the single sample split method implies the same property for the multi sample split procedure, as described next.

Proposition 11.2. *Let $\hat{S}_{\text{single−split FWER}}(\alpha)$ be the selected model of the single sample split method. Assume that $\alpha = \alpha_n \to 0$ can be chosen for $n \to \infty$ such that $\lim_{n\to\infty} \mathbf{P}[\hat{S}_{\text{single−split FWER}}(\alpha_n) = S_0] = 1$, for every single sample split into I_1 and I_2. Then, for any $\gamma_{\min} \in (0,1)$ (see (11.8)), the multi sample split method, with fixed number B of random splits, is also consistent for variable selection, i.e.,*

$$\lim_{n\to\infty} \mathbf{P}[\hat{S}_{\text{multi−split FWER}}(\alpha_n) = S_0] = 1.$$

A proof is given in Section 11.8. Proposition 11.2 is no surprise as the multi sample split method should be at least as good as the single sample split analogue. This is intuitively clear and is also illustrated with empirical results in Section 11.5.

Sufficient conditions for variable selection consistency with the single sample split method are summarized in Corollary 11.1. We remark that we have to strengthen condition (11.21) to hold for every (single) sample split, i.e.

$$\inf_{S\in\mathscr{S},I_{(m)},n\in\mathbb{N}} \Lambda_{\min}(m^{-1}\mathbf{X}^T_{I_{(m)},S}\mathbf{X}_{I_{(m)},S}) \geq D > 0,$$

where the infimum also runs over all subsets $I_{(m)} \subset \{1,\ldots,n\}$ of cardinality $m = n - \lfloor n/2 \rfloor$.

11.7 Extensions

We briefly discuss how the sample-splitting methodology can be used for other models and for controlling other error measures.

11.7.1 Other models

Due to the generic nature of the sample splitting methodology, extensions are straightforward to situations where (asymptotically valid) p-values P_{raw_j} for hypotheses $H_{0,j}$ $(j = 1, \ldots, p)$, based on the second half of the data I_2 as in (11.1), are available.

An important class of examples are generalized linear models (GLMs), described in Chapter 3. The null-hypotheses are $H_{0,j} : \beta_j = 0$ $(j = 1, \ldots, p)$ as for linear models. We could use the Lasso for GLMs, see Section 3.2.1, for variable screening or selection based on the first half of the sample I_1. The p-values based on the second half of the data I_2 rely on classical (e.g. likelihood ratio) tests applied to the selected submodel, analogous to the methodology proposed for linear models.

Another example are (generalized) additive models, described in Chapter 5. The null-hypotheses are then $H_{0,j} : f_j(\cdot) \equiv 0$ $(j = 1, \ldots, p)$, saying that additive functions are zero. We could use the estimator in (5.5) using the sparsity smoothness penalty for variable screening based on the first half of the sample I_1. Approximate p-values based on the second half of the data I_2 could be obtained by likelihood ratio tests, see e.g. Wood (2006).

A third example are Gaussian Graphical Models, described in Chapter 13, Section 13.4. There, the dimensionality reduction on the first half of the sample I_1 can be based on an ℓ_1-penalization scheme such as the GLasso, see Section 13.4.1. The p-values based on the second half of the data I_2 rely again on classical likelihood ratio tests, see e.g. Lauritzen (1996).

11.7.2 Control of expected false positive selections

In some settings, control of FWER, say at $\alpha = 0.05$, is too conservative. One can either resort to control of FDR, as described in Section 11.4. Alternatively, the FWER control procedure can easily be adjusted to control the expected number of false positives. Define, for $M > 1$ and $0 < \alpha < 1$,

$$\hat{S}_{\mathrm{multi-split\ FP}}(\alpha, M) = \{j;\ P_{\mathrm{non-capp},j}/M \leq \alpha\},$$

where $P_{\mathrm{non-capp},j}$ are adjusted p-values as in (11.8), but not capped at 1. Furthermore, let

$$V_{\mathrm{multi-split\ FP}}(\alpha, M) = |\hat{S}_{\mathrm{multi-split\ FP}}(\alpha, M) \cap S_0^c|.$$

Then, the expected number of false positives is controlled:

$$\limsup_{n \to \infty} \mathbb{E}[V_{\mathrm{multi-split\ FP}}(\alpha, M)] \leq \alpha M. \tag{11.23}$$

A proof of this follows directly from the proof of Theorem 11.1. We leave it as Problem 11.4. For example, setting $M = \alpha^{-1}\delta$ ($\delta > 0$ small) offers a much less conservative error control than using the familywise error. Note that control of the expected number of false positives could also be achieved with the stability selection method, albeit under a strong exchangeability assumption as described in Theorem 10.1.

11.8 Proofs

11.8.1 Proof of Proposition 11.1

For technical reasons we define

$$K_j^{[b]} = \tilde{P}_{\text{corr},j}^{[b]} 1(S_0 \subseteq \hat{S}^{[b]}) + 1(S_0 \not\subseteq \hat{S}^{[b]}). \tag{11.24}$$

Thus, $K_j^{[b]}$ is the adjusted p-value if the estimated active set contains the true active set, and it equals 1 otherwise. Consider the set

$$A_n = \{K_j^{[b]} = \tilde{P}_{\text{corr},j}^{[b]} \text{ for all } b = 1,\ldots,B\}.$$

Because of assumption (11.3), and since B is fixed, we have

$$\mathbf{P}[A_n] \to 1 \ (n \to \infty).$$

Therefore, we define all the quantities involving $\tilde{P}_{\text{corr},j}^{[b]}$ also with $K_j^{[b]}$, and it is suffi-cient to show, under this slightly altered (theoretical) procedure that

$$\mathbf{P}[\min_{j \in S_0^c} Q_j(\gamma) \le \alpha] \le \alpha.$$

In particular we can omit here the limes superior since on the set A_n, we can show finite sample error control.

We also omit for the proof the truncation function $\min\{1,\cdot\}$ from the definitions of $Q_j(\gamma)$ and P_j in (11.7) and (11.8), respectively. The selected sets of variables are clearly unaffected and the notation simplifies considerably.

Define for $u \in (0,1)$ the quantity $\pi_j(u)$ as the fraction of split samples that yield $K_j^{[b]}$ less than or equal to u,

$$\pi_j(u) = \frac{1}{B} \sum_{b=1}^{B} 1(K_j^{[b]} \le u).$$

Note that the events $\{Q_j(\gamma) \le \alpha\}$ and $\{\pi_j(\alpha\gamma) \ge \gamma\}$ are equivalent. Hence,

$$\mathbf{P}[\min_{j \in S_0^c} Q_j(\gamma) \le \alpha] \le \sum_{j \in S_0^c} \mathbb{E}[\mathbb{1}(Q_j(\gamma) \le \alpha)] = \sum_{j \in S_0^c} \mathbb{E}[\mathbb{1}(\pi_j(\alpha\gamma) \ge \gamma)]. \qquad (11.25)$$

Using a Markov inequality,

$$\sum_{j \in S_0^c} \mathbb{E}[\mathbb{1}(\pi_j(\alpha\gamma) \ge \gamma)] \le \frac{1}{\gamma} \sum_{j \in S_0^c} \mathbb{E}[\pi_j(\alpha\gamma)]. \qquad (11.26)$$

By definition of $\pi_j(\cdot)$,

$$\frac{1}{\gamma} \sum_{j \in S_0^c} \mathbb{E}[\pi_j(\alpha\gamma)] = \frac{1}{\gamma} \frac{1}{B} \sum_{b=1}^{B} \sum_{j \in S_0^c \cap \hat{S}^{[b]}} \mathbb{E}[\mathbb{1}(K_j^{[b]} \le \alpha\gamma)]. \qquad (11.27)$$

Here we have used that $K_j^{[b]} = 1$ for $j \notin \hat{S}^{[b]}$. Moreover, using the definition of $K_j^{[b]}$ in (11.24),

$$\mathbb{E}[\mathbb{1}(K_j^{[b]} \le \alpha\gamma)] \le \mathbf{P}[\tilde{P}_{\text{corr},j}^{[b]} \le \alpha\gamma | S_0 \subseteq \hat{S}^{[b]}] = \frac{\alpha\gamma}{|\hat{S}^{[b]}|}. \qquad (11.28)$$

This is a consequence of the uniform distribution of $\tilde{P}_j^{[b]}$ given $S_0 \subseteq \hat{S}^{[b]}$, for $j \in S_0^c$. Summarizing, using (11.25)-(11.28) we get

$$\mathbf{P}\left(\min_{j \in S_0^c} Q_j(\gamma) \le \alpha\right) \le \frac{1}{\gamma} \frac{1}{B} \sum_{b=1}^{B} \mathbb{E}\left(\sum_{j \in S_0^c \cap \hat{S}^{[b]}} \frac{\alpha\gamma}{|\hat{S}^{[b]}|}\right) \le \alpha,$$

which completes the proof. □

11.8.2 Proof of Theorem 11.1

As in the proof of Proposition 11.1, we will work with $K_j^{[b]}$ instead of $\tilde{P}_{\text{corr},j}^{[b]}$. Analogously, instead of the non-adjusted $\tilde{P}_j^{[b]}$, we work with

$$\tilde{K}_j^{[b]} = \tilde{P}_j^{[b]} \mathbb{1}(S_0 \subseteq \hat{S}^{[b]}) + \mathbb{1}(S_0 \not\subseteq \hat{S}^{[b]}).$$

For any $\tilde{K}_j^{[b]}$ with $j \in S_0^c$ and $\alpha \in (0,1)$,

$$\mathbb{E}\left(\frac{\mathbb{1}(\tilde{K}_j^{[b]} \le \alpha\gamma)}{\gamma}\right) \le \alpha. \qquad (11.29)$$

Furthermore,

$$\mathbb{E}\left(\max_{j\in S_0^c}\frac{1(K_j^{[b]}\leq\alpha\gamma)}{\gamma}\right)\leq\mathbb{E}\left(\sum_{j\in S_0^c}\frac{1(K_j^{[b]}\leq\alpha\gamma)}{\gamma}\right)\leq\mathbb{E}\left(\sum_{j\in S_0^c\cap\hat{S}^{[b]}}\frac{1(K_j^{[b]}\leq\alpha\gamma)}{\gamma}\right),$$

where we use that $K_j^{[b]}=1$ for $j\notin\hat{S}^{[b]}$, and hence, with (11.29) and using the definition (11.24) of $K_j^{[b]}$,

$$\mathbb{E}\left(\max_{j\in S_0^c}\frac{1(K_j^{[b]}\leq\alpha\gamma)}{\gamma}\right)\leq\mathbb{E}\left(\sum_{j\in S_0^c\cap\hat{S}^{[b]}}\frac{\alpha}{|\hat{S}^{[b]}|}\right)\leq\alpha.\qquad(11.30)$$

For a random variable U taking values in $[0,1]$,

$$\sup_{\gamma\in(\gamma_{\min},1)}\frac{1(U\leq\alpha\gamma)}{\gamma}=\begin{cases}0 & U\geq\alpha,\\ \alpha/U & \alpha\gamma_{\min}\leq U<\alpha,\\ 1/\gamma_{\min} & U<\alpha\gamma_{\min}.\end{cases}$$

Moreover, if U has a uniform distribution on $[0,1]$,

$$\mathbb{E}\left(\sup_{\gamma\in(\gamma_{\min},1)}\frac{1(U\leq\alpha\gamma)}{\gamma}\right)=\int_0^{\alpha\gamma_{\min}}\gamma_{\min}^{-1}dx+\int_{\alpha\gamma_{\min}}^{\alpha}\alpha x^{-1}dx=\alpha(1-\log\gamma_{\min}).$$

Hence, by using that $\tilde{K}_j^{[b]}$ has a uniform distribution on $[0,1]$ for all $j\in S_0^c$, conditional on the event $S_0\subseteq\hat{S}^{[b]}$,

$$\mathbb{E}\left(\sup_{\gamma\in(\gamma_{\min},1)}\frac{1(\tilde{K}_j^{[b]}\leq\alpha\gamma)}{\gamma}\right)\leq\mathbb{E}\left(\sup_{\gamma\in(\gamma_{\min},1)}\frac{1(\tilde{K}_j^{[b]}\leq\alpha\gamma)}{\gamma}\Big|S\subseteq\hat{S}^{[b]}\right)=\alpha(1-\log\gamma_{\min}),$$

where we used that $1(\tilde{K}_j^{[b]}\leq\alpha\gamma)=0$ if $S_0\not\subseteq\hat{S}^{[b]}$. Analogously to (11.30), we can then deduce that

$$\sum_{j\in S_0^c}\mathbb{E}\left(\sup_{\gamma\in(\gamma_{\min},1)}\frac{1(K_j^{[b]}\leq\alpha\gamma)}{\gamma}\right)\leq\alpha(1-\log\gamma_{\min}).$$

Averaging over all split samples yields

$$\sum_{j\in N}\mathbb{E}\left(\sup_{\gamma\in(\gamma_{\min},1)}\frac{\frac{1}{B}\sum_{b=1}^B 1(K_j^{[b]}/\gamma\leq\alpha)}{\gamma}\right)\leq\alpha(1-\log\gamma_{\min}).$$

Using again a Markov inequality,

$$\sum_{j \in S_0^c} \mathbb{E}[\sup_{\gamma \in (\gamma_{\min}, 1)} 1(\pi_j(\alpha \gamma) \geq \gamma)] \leq \alpha(1 - \log \gamma_{\min}),$$

where we have used the same definition for $\pi_j(\cdot)$ as in the proof of Proposition 11.1. Due to the equivalence $\{Q_j(\gamma) \leq \alpha\} = \{\pi_j(\alpha \gamma) \geq \gamma\}$, it follows that

$$\sum_{j \in S_0^c} \mathbf{P}[\inf_{\gamma \in (\gamma_{\min}, 1)} Q_j(\gamma) \leq \alpha] \leq \alpha(1 - \log \gamma_{\min}),$$

implying that

$$\sum_{j \in S_0^c} \mathbf{P}[\inf_{\gamma \in (\gamma_{\min}, 1)} Q_j(\gamma)(1 - \log \gamma_{\min}) \leq \alpha] \leq \alpha.$$

Using the definition of P_j in (11.8),

$$\sum_{j \in S_0^c} \mathbf{P}[P_j \leq \alpha] \leq \alpha, \tag{11.31}$$

and thus, by the union bound,

$$\mathbf{P}[\min_{j \in S_0^c} P_j \leq \alpha] \leq \alpha,$$

which completes the proof. $\square$

11.8.3 Proof of Theorem 11.2

As in the proofs of Proposition 11.1 and Theorem 11.1, we use implicitly for all p-values a version as in (11.24). We denote by

$$q = \alpha \sum_{i=1}^{p} i^{-1}, \tag{11.32}$$

as in (11.13). Let

$$p_{ijk} = \mathbf{P}[\{P_i \in [(j-1)q, jq]\} \cap C_k^{(i)}] \ (i, j = 1, \ldots, p),$$

where $C_k^{(i)}$ is the event:

$$C_k^{(i)} = \{\text{variable } i \text{ rejected} \Rightarrow k - 1 \text{ other variables are rejected}\}.$$

Here, the word rejected means that the variable is not in the set $\hat{S}_{\text{multi-split FDR}}(\alpha)$. This notation is as in Benjamini and Yekutieli (2001, proof of Theorem 1.3).

Now, as shown in equation (10) and then again in (28) in Benjamini and Yekutieli (2001),

$$\mathbb{E}[Q(\alpha)] = \sum_{i \in S_0^c} \sum_{k=1}^p \frac{1}{k} \sum_{j=1}^k p_{ijk}.$$

Using this result, we use in the beginning of the next (in-)equalities a similar argument to Benjamini and Yekutieli (2001),

$$\mathbb{E}[Q(\alpha)] = \sum_{i \in S_0^c} \sum_{k=1}^p \frac{1}{k} \sum_{j=1}^k p_{ijk} = \sum_{i \in S_0^c} \sum_{j=1}^p \sum_{k=j}^p \frac{1}{k} p_{ijk}$$

$$\leq \sum_{i \in S_0^c} \sum_{j=1}^p \sum_{k=j}^p \frac{1}{j} p_{ijk} \leq \sum_{i \in S_0^c} \sum_{j=1}^p \frac{1}{j} \sum_{k=1}^p p_{ijk} = \sum_{j=1}^p \frac{1}{j} \sum_{i \in S_0^c} \sum_{k=1}^p p_{ijk}. \quad (11.33)$$

Let us denote by

$$f(j) := \sum_{i \in S_0^c} \sum_{k=1}^p p_{ijk}, \quad j = 1, \dots, p.$$

The last inequality in (11.33) can then be rewritten as

$$\mathbb{E}[Q(\alpha)] \leq \sum_{j=1}^p \frac{1}{j} f(j) = f(1) + \sum_{j=2}^p \frac{1}{j} \left(\sum_{j'=1}^j f(j') - \sum_{j'=1}^{j-1} f(j') \right)$$

$$= \sum_{j=1}^{p-1} \left(\frac{1}{j} - \frac{1}{j+1} \right) \sum_{j'=1}^j f(j') + \frac{1}{p} \sum_{j'=1}^p f(j'). \quad (11.34)$$

Note that, in analogy to (27) in Benjamini and Yekutieli (2001),

$$\sum_{k=1}^p p_{ijk} = \mathbf{P}\left(\{ P_i \in [(j-1)q, jq] \} \cap \bigcup_k C_k^{(i)} \right) = \mathbf{P}\left(P_i \in [(j-1)q, jq] \right),$$

and hence

$$f(j) = \sum_{i \in S_0^c} \sum_{k=1}^p p_{ijk} = \sum_{i \in S_0^c} \mathbf{P}\left(P_i \in [(j-1)q, jq] \right).$$

Thus, it follows that

$$\sum_{j'=1}^j f(j') = \sum_{i \in S_0^c} \mathbf{P}[P_i \leq jq] \leq jq,$$

where the last inequality is due to (11.31) in the proof of Theorem 11.1. Using this bound in (11.34), we obtain

$$\mathbb{E}[Q(\alpha)] \leq \sum_{j=1}^{p-1}(\frac{1}{j} - \frac{1}{j+1})jq + \frac{1}{p}pq = (\sum_{j=1}^{p-1}\frac{1}{j(j+1)}j+1)q = q\sum_{j=1}^{p}\frac{1}{j} = \alpha,$$

due to the definition in (11.32). This completes the proof. □

11.8.4 Proof of Proposition 11.2

Because the single sample split method is model selection consistent, it must hold that $\mathbf{P}[\max_{j\in S_0} \tilde{P}_j|\hat{S}_{I_1}| \leq \alpha_n] \to 1$ for $n \to \infty$. Using multiple sample splits, this property holds for each of the (fixed number of) B splits and hence

$$\mathbf{P}[\max_{j\in S_0}\max_{b=1,\dots,B} \tilde{P}_j^{[b]}|\hat{S}^{[b]}| \leq \alpha_n] \to 1.$$

This implies that the quantile $\max_{j\in S_0} Q_j(1)$ is bounded from above by α_n, with probability converging to 1 as $n \to \infty$. The maximum over all $j \in S_0$ of the adjusted p-values $P_j = (1 - \log\gamma_{\min})\inf_{\gamma\in(\gamma_{\min},1)} Q_j(\gamma)$, see (11.8), is thus bounded from above by $(1 - \log\gamma_{\min})\alpha_n$, again with probability converging to 1 for $n \to \infty$. Thus, all variables in S_0 are selected with probability tending to 1 which completes the proof. □

11.8.5 Proof of Lemma 11.3

We write (11.16) as:

$$T_{I_2,S,j} = \frac{\hat{\beta}_{j|S}^{\mathrm{OLS}}}{\sigma\sqrt{(\mathbf{X}_{I_2,S}^T\mathbf{X}_{I_2,S})_{j,j}^{-1}}} \cdot \frac{\sigma}{\hat{\sigma}_S^{\mathrm{OLS}}} =: U_{j|S} \cdot V_S.$$

For the first variable, since $S \supseteq S_0$,

$$U_{j|S} = \frac{\hat{\beta}_{j|S}^{\mathrm{OLS}}}{\sigma\sqrt{(\mathbf{X}_{I_2,S}^T\mathbf{X}_{I_2,S})_{j,j}^{-1}}} \sim \mathscr{N}(\mu_j, 1),$$

$$\mu_j = \frac{\beta_j}{\sigma\sqrt{(\mathbf{X}_{I_2,S}^T\mathbf{X}_{I_2,S})_{j,j}^{-1}}}. \tag{11.35}$$

For the second variable, we use that

$$(\hat{\sigma}_S^{\mathrm{OLS}})^2/\sigma^2 \sim \chi_{n-|S|}^2/(n-|S|).$$

Thus, since $|S| \leq a_n = o(n)$, $|S_0| \leq a_n = o(n)$ (see (11.15)), and the exponentially fast concentration of the χ_n^2/n distribution around 1, see Example 14.1 in Chapter 14, we have for any $\delta > 0$,

$$\sup_{S \in \mathscr{S}} |S_0| \mathbf{P}\left(\left|\frac{\sigma}{\hat{\sigma}_S^{\text{OLS}}} - 1\right| \leq \delta\right) \to 1 \ (n \to \infty). \tag{11.36}$$

Next, we use

$$\mathbf{P}\left(|U_{j|S}|V_S > c\right) \geq \mathbf{P}\left(|U_{j|S}|V_S > c, V_S \in [1 - \delta, 1 + \delta]\right)$$
$$\geq \mathbf{P}\left(|U_{j|S}| > \frac{c}{1 - \delta}, V_S \in [1 - \delta, 1 + \delta]\right).$$

where the last inequality holds due to the implication that $|U_{j|S}| > \frac{c}{1-\delta}$ and $V_S \in [1 - \delta, 1 + \delta]$ imply that $|U_{j|S}|V_S > c$. Now, we use that $\mathbf{P}[A \cap B] \geq \mathbf{P}[A] - \mathbf{P}[B^c]$ for any two events A and B: we leave the derivation as Problem 11.5. Therefore,

$$\mathbf{P}\left(|U_{j|S}|V_S > c\right) > \mathbf{P}\left(|U_{j|S}| > \frac{c}{1 - \delta}\right) - \mathbf{P}(V_s \notin [1 - \delta, 1 + \delta]). \tag{11.37}$$

Now,

$$\mathbf{P}\left(|U_{j|S}| > \frac{c}{1 - \delta}\right) \geq \begin{cases} \mathbf{P}\left(Z > \frac{c}{1-\delta} - \dfrac{\beta_j |I_2|^{1/2}}{\sigma\sqrt{(|I_2|^{-1}\mathbf{X}_{I_2,S}^T \mathbf{X}_{I_2,S})_{j,j}^{-1}}}\right) & \text{if } \beta_j > 0, \\[3ex] \mathbf{P}\left(Z < -\dfrac{c}{1-\delta} + \dfrac{\beta_j |I_2|^{1/2}}{\sigma\sqrt{(|I_2|^{-1}\mathbf{X}_{I_2,S}^T \mathbf{X}_{I_2,S})_{j,j}^{-1}}}\right) & \text{if } \beta_j < 0, \end{cases}$$

where $Z \sim \mathscr{N}(0, 1)$, see (11.35).

Now choose $c = t(1 - \alpha_n/2, |I_2| - a_n)$. Due to (11.18), the dominating terms are $\dfrac{\beta_j |I_2|^{1/2}}{\sigma\sqrt{(|I_2|^{-1}\mathbf{X}_{I_2,S}^T \mathbf{X}_{I_2,S})_{j,j}^{-1}}}$, and because of (11.18) we obtain

$$\sup_{S \in \mathscr{S}} \sum_{j \in S_0} \mathbf{P}\left(|U_{j|S}| \leq \frac{t(1 - \alpha_n/2, |I_2| - a_n)}{1 - \delta}\right) \to 0 \ (n \to \infty), \tag{11.38}$$

where α_n can be chosen to converge to 0 sufficiently slowly. The exact derivation is left as Problem 11.6.

Using (11.37), (11.38) and (11.36), we complete the proof. $\qquad\square$

Problems

11.1. Describe why the "p-value lottery" in Figure 11.1 arises. Which conditions ensuring variable screening, see (11.3), appear to be violated ? (See also Section 2.5 in Chapter 2, describing the screening properties of the Lasso). Are there other plausible reasons?

11.2. Show that the p-values in (11.2) control the familywise error rate, assuming (11.3), (11.4) and Gaussian errors.

11.3. Condition (11.21) is sufficient for condition (11.19) by using the following fact. For a symmetric $m \times m$ matrix A with positive minimal eigenvalue $\Lambda_{\min}^2(A) > 0$:

$$A_{j,j}^{-1} \le \frac{1}{\Lambda_{\min}^2(A)}.$$

Prove that the latter holds.
Hint: Use the spectral decomposition $A = UDU^T$, where $U^T U = UU^T$ and $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_m)$ containing the eigenvalues $\lambda_1 \le \ldots \le \lambda_m$. The inverse is then $A^{-1} = UD^{-1}U^T$.

11.4. Prove formula (11.23).

11.5. Prove that for any two events A and B:

$$\mathbf{P}[A \cap B] \ge \mathbf{P}[A] - \mathbf{P}[B^c].$$

11.6. Prove the bound in (11.38) by showing

$$\sup_{S \in \mathscr{S}, j \in S_0} |S_0| \mathbf{P}[|U_{j|S}| \le \frac{t(1 - \alpha_n/2, |I_2| - a_n)}{1 - \delta}] \to 0.$$

Note that the sequence α_n can be chosen to converge to 0 sufficiently slowly.

Chapter 12
Boosting and greedy algorithms

Abstract Boosting algorithms or greedy methods are computationally fast and often powerful for high-dimensional data problems. They have been mainly developed for classification and regression. Regularization arises in form of algorithmic constraints rather than explicit penalty terms. Interestingly, both of these regularization concepts are sometimes close to being equivalent, as shown for linear models by Efron et al. (2004). We present boosting and related algorithms, including a brief discussion about forward variable selection and orthogonal matching pursuit. The exposition in this chapter is mainly focusing on methodology and describing simple computational ideas. Mathematical theory is developed for special cases like one-dimensional function estimation or high-dimensional linear models.

12.1 Organization of the chapter

After an introduction, we present in Section 12.3 the gradient boosting algorithm representing the core methodological idea of boosting. Different loss functions for different settings, including regression and classification, and their corresponding boosting algorithms are described in Section 12.4. The choice of the weak learner or base procedure is treated in Section 12.5. Section 12.4.4.1 contains a more detailed description of L_2Boosting based on the squared error loss for regression. It covers asymptotic optimality and high-dimensional consistency as well as methodological connections to the Lasso. Forward variable selection and orthogonal matching pursuit are discussed in Section 12.7. Finally, all proofs are collected in Section 12.8.

12.2 Introduction and preliminaries

Freund and Schapire's AdaBoost algorithm for classification (Freund and Schapire, 1996, 1997), the first successful boosting algorithm, has attracted much attention in the machine learning community (cf. Schapire, 2002, and the references therein) as well as in related areas in statistics (Breiman, 1998, 1999; Friedman et al., 2000; Friedman, 2001; Bühlmann and Hothorn, 2007). AdaBoost and its various versions have been empirically found to be very competitive in a variety of applications. Furthermore, there is a striking similarity between gradient based boosting and the Lasso in linear or generalized linear models. Thus, despite substantial conceptual differences, boosting-type algorithms are implicitly related to ℓ_1-regularization.

12.2.1 Ensemble methods: multiple prediction and aggregation

Boosting algorithms have been originally proposed as ensemble methods. They rely on the principle of generating multiple predictions from re-weighted data which are then aggregated using linear (or convex) combination or majority voting for building the final estimator or prediction.

First, we specify a base procedure, sometimes also called weak learner, which constructs a function estimate or a prediction $\hat{g}(\cdot)$ based on some input data $(X_1, Y_1), \ldots,$ (X_n, Y_n) with covariates X_i and responses Y_i:

$$(X_1, Y_1), \ldots, (X_n, Y_n) \quad \xrightarrow{\text{base procedure}} \quad \hat{g}(\cdot).$$

A very popular base procedure is a regression tree. Other examples will be described in Section 12.5.

Generating an ensemble from the base procedure, i.e., an ensemble of function estimates or predictions, works generally as follows:

$$\text{re-weighted data 1} \quad \xrightarrow{\text{base procedure}} \quad \hat{g}^{[1]}(\cdot)$$

$$\text{re-weighted data 2} \quad \xrightarrow{\text{base procedure}} \quad \hat{g}^{[2]}(\cdot)$$

$$\ldots \qquad\qquad\qquad\qquad \ldots$$

$$\ldots \qquad\qquad\qquad\qquad \ldots$$

$$\text{re-weighted data } M \quad \xrightarrow{\text{base procedure}} \quad \hat{g}^{[M]}(\cdot)$$

$$\text{aggregation: } \hat{f}_A(\cdot) = \sum_{m=1}^{M} \alpha_m \hat{g}^{[m]}(\cdot),$$

using suitable coefficients $\alpha_1, \ldots, \alpha_M$. What is termed here "re-weighted data" means that we assign individual data weights to every of the n sample points. Thereby, we implicitly assume that the base procedure allows to do some weighted fitting, i.e., estimation is based on a weighted sample. Throughout this chapter (except in Section 12.2.2), we assume that a base procedure estimate $\hat{g}(\cdot)$ is real-valued, i.e., a regression-type procedure.

The above description of an ensemble scheme is too general to be of any direct use. The specification of the data re-weighting mechanism as well as the form of the linear combination coefficients $\{\alpha_m\}_{m=1}^M$ are crucial, and various choices characterize different ensemble schemes. Boosting methods are special kinds of sequential ensemble schemes, where the data weights in iteration m depend on the results from the previous iteration $m-1$ only.

12.2.2 AdaBoost

The AdaBoost algorithm for binary classification (Freund and Schapire, 1997) is the most well known boosting algorithm. Consider data $(X_1, Y_1), \ldots, (X_n, Y_n)$ with $Y_i \in \{0, 1\}$ and p-dimensional covariates X_i. The base procedure is a classifier with values in $\{0, 1\}$, for example a classification tree (slightly different from a real-valued function estimator as assumed above). The digression from using a real-valued base procedure is only for this subsection as we will not consider the Ada-Boost algorithm any further. AdaBoost is briefly presented in Algorithm 6: the intention of the description is for the sake of completeness since it has been the first successful boosting algorithm. Afterwards, in the following sections, we will exclusively focus on so-called gradient boosting algorithms.

By using the terminology m_{stop} (instead of M as in the general description of ensemble schemes), we emphasize here and later that the iteration process should be stopped to avoid overfitting. It is a tuning parameter of AdaBoost and it is typically selected using some cross-validation scheme.

12.3 Gradient boosting: a functional gradient descent algorithm

Breiman (1998, 1999) showed that the AdaBoost algorithm can be represented as a steepest descent algorithm in function space which we call functional gradient descent (FGD). Friedman et al. (2000) and Friedman (2001) developed later a more general, statistical framework which yields a direct interpretation of boosting as a method for function estimation. In their terminology, it is a "stagewise, additive modeling" approach (but the word "additive" doesn't imply a model fit which is additive in the covariates, see Section 12.5).

Algorithm 6 AdaBoost algorithm

1: Initialize weights for individual sample points: $w_i^{[0]} = 1/n$ for $i = 1, \ldots, n$. Set $m = 0$.
2: Increase m by one: $m \leftarrow m + 1$.
   Fit the base procedure to the weighted data, i.e., do a weighted fitting using the weights $w_i^{[m-1]}$, yielding the classifier $\hat{g}^{[m]}(\cdot)$.
3: Compute the weighted in-sample misclassification rate

$$\text{err}^{[m]} = \sum_{i=1}^{n} w_i^{[m-1]} 1\left(Y_i \neq \hat{g}^{[m]}(X_i)\right) / \sum_{i=1}^{n} w_i^{[m-1]},$$

$$\alpha^{[m]} = \log\left(\frac{1 - \text{err}^{[m]}}{\text{err}^{[m]}}\right),$$

   and up-date the weights

$$\tilde{w}_i = w_i^{[m-1]} \exp\left(\alpha^{[m]} 1\left(Y_i \neq \hat{g}^{[m]}(X_i)\right)\right),$$

$$w_i^{[m]} = \tilde{w}_i / \sum_{j=1}^{n} \tilde{w}_j.$$

4: Iterate steps 2 and 3 until $m = m_{\text{stop}}$ and build the aggregated classifier by weighted majority voting:

$$\hat{f}_{\text{AdaBoost}}(x) = \arg\max_{y \in \{0,1\}} \sum_{m=1}^{m_{\text{stop}}} \alpha^{[m]} 1\left(\hat{g}^{[m]}(x) = y\right).$$

   The classification rule is then given by

$$\mathscr{C}_{\text{AdaBoost}}(x) = 1\left(\hat{f}_{\text{AdaBoost}}(x) > 0\right).$$

Consider the problem of estimating a real-valued function

$$f^0(\cdot) = \arg\min_{f(\cdot)} \mathbb{E}[\rho(f(X), Y)], \tag{12.1}$$

where $\rho(\cdot, \cdot)$ is a loss function which is typically assumed to be differentiable and convex with respect to the first argument and minimization is over all (measurable) functions $f(\cdot)$. For example, the squared error loss $\rho(f, y) = |y - f|^2$ yields the well-known population minimizer $f^0(x) = \mathbb{E}[Y | X = x]$.

12.3.1 The generic FGD algorithm

In the following, FGD and boosting are used as equivalent terminology for the same method or algorithm. Estimation of $f^0(\cdot)$ in (12.1) with boosting can be done by considering the empirical risk $n^{-1} \sum_{i=1}^{n} \rho(f(X_i), Y_i)$ and pursuing iterative steepest

descent in function space. We note that there is no explicit regularization term when minimizing the empirical risk: as we will describe below, regularization with boosting is done via algorithmic constraints, namely the number of iterations in the optimization process. The following algorithm has been given by Friedman (2001).

Algorithm 7 Generic FGD algorithm

1: Initialize $\hat{f}^{[0]}(\cdot)$ with an offset value. Common choices are

$$\hat{f}^{[0]}(\cdot) \equiv \arg\min_{c} n^{-1} \sum_{i=1}^{n} \rho(c, Y_i)$$

or $\hat{f}^{[0]}(\cdot) \equiv 0$. Set $m = 0$.

2: Increase m by one: $m \leftarrow m + 1$.
   Compute the negative gradient $-\frac{\partial}{\partial f}\rho(f, Y)$ and evaluate at $\hat{f}^{[m-1]}(X_i)$:

$$U_i = -\frac{\partial}{\partial f}\rho(f, Y_i)\big|_{f=\hat{f}^{[m-1]}(X_i)}, \quad i = 1, \dots, n.$$

3: Fit the negative gradient vector $U_1, \dots, U_n$ to $X_1, \dots, X_n$ by the real-valued base procedure (e.g. regression)

$$(X_i, U_i)_{i=1}^{n} \xrightarrow{\text{base procedure}} \hat{g}^{[m]}(\cdot).$$

Thus, $\hat{g}^{[m]}(\cdot)$ can be viewed as an approximation of the negative gradient vector.

4: Up-date $\hat{f}^{[m]}(\cdot) = \hat{f}^{[m-1]}(\cdot) + \nu \cdot \hat{g}^{[m]}(\cdot)$, where $0 < \nu \leq 1$ is a step-length factor (see below), i.e., proceed along an estimate of the negative gradient vector.

5: Iterate steps 2 to 4 until $m = m_{\text{stop}}$ for a pre-specified stopping iteration m_{stop}.

The stopping iteration m_{stop}, which is the main tuning and regularization parameter, can be determined via cross-validation. The choice of the step-length factor ν in step 4 is of minor importance, as long as it is "small" such as $\nu = 0.1$. A smaller value of ν typically requires a larger number of boosting iterations and thus more computing time, while the predictive accuracy has been empirically found to be potentially better and almost never worse when choosing ν "sufficiently small" like $\nu = 0.1$ (Friedman, 2001). We remark that Friedman (2001) suggests to use an additional line search between steps 3 and 4 (in case of other loss functions $\rho(\cdot, \cdot)$ than squared error): it yields a slightly different algorithm but the additional line search seems unnecessary for achieving a good estimator $\hat{f}^{[m_{\text{stop}}]}$.

12.3.1.1 Alternative formulation in function space

In steps 2 and 3 of the generic FGD Algorithm 7, we associated with $U_1, \dots, U_n$ a negative gradient vector. A reason for this can be seen from the following formulation in function space.

Consider the empirical risk functional $C(f) = n^{-1}\sum_{i=1}^n \rho(f(X_i), Y_i)$, that is, we regard $C(\cdot)$ as a mapping from $f \in L_2(Q_n)$ to $\mathbb{R}$, where $Q_n = n^{-1}\sum_{i=1}^n \delta_{X_i}$ is the empirical measure of the X_i's. The associated inner product is $(f, g)_n = n^{-1}\sum_{i=1}^n f(X_i)g(X_i)$ $(f, g \in L_2(Q_n))$. We can then calculate the negative (functional) Gâteaux derivative $dC(\cdot)$ of the functional $C(\cdot)$,

$$-dC(f)(x) = -\frac{\partial}{\partial \alpha}C(f + \alpha\delta_x)|_{\alpha=0}, \quad f \in L_2(Q_n),\ x \in \mathbb{R}^p,$$

where δ_x denotes the delta- (or indicator-) function at $x \in \mathbb{R}^p$. In particular, when evaluating the derivative $-dC$ at $\hat{f}^{[m-1]}$ and X_i, we get

$$-dC(\hat{f}^{[m-1]})(X_i) = n^{-1}U_i, \tag{12.2}$$

with $U_1, ..., U_n$ exactly as in Steps 2 and 3 of the generic FGD Algorithm 7 (Problem 12.1). Thus, the negative gradient vector $U_1, ..., U_n$ can be interpreted as a functional (Gâteaux) derivative evaluated at the data points.

12.4 Some loss functions and boosting algorithms

Various boosting algorithms can be defined by specifying different loss functions $\rho(\cdot, \cdot)$.

12.4.1 Regression

For regression with response $Y \in \mathbb{R}$, the most popular choice is the squared error loss (scaled by the factor $1/2$ such that the negative gradient vector equals the residuals, see Section 12.4.4 below):

$$\rho_{L_2}(f, y) = \frac{1}{2}|y - f|^2 \tag{12.3}$$

with population minimizer

$$f_{L_2}^0(x) = \mathbb{E}[Y|X = x].$$

The corresponding boosting algorithm is L_2Boosting, described in more detail in Section 12.4.4.

An alternative loss function which has some robustness properties (with respect to the error distribution, i.e., in "Y-space") is the L_1-loss

$$\rho_{L_1}(f,y) = |y - f|$$

with population minimizer

$$f^0(x) = \text{median}(Y|X = x).$$

Although the L_1-loss is not differentiable at the point $y = f$, we can compute partial derivatives since for fixed f, the single point $Y_i = f(X_i)$ (usually) has probability zero to be realized by the data. Alternatively, as a compromise between the L_1- and L_2-loss, we may use the Huber-loss function from robust statistics: for $\delta > 0$,

$$\rho_{\text{Huber}}(f,y) = \begin{cases} |y - f|^2/2, & \text{if } |y - f| \le \delta \\ \delta(|y - f| - \delta/2), & \text{if } |y - f| > \delta. \end{cases}$$

A strategy for choosing (a changing) δ adaptively in iteration m has been proposed by Friedman (2001):

$$\delta_m = \text{median}(\{|Y_i - \hat{f}^{[m-1]}(X_i)|; \ i = 1, \dots, n\}),$$

where the previous fit $\hat{f}^{[m-1]}(\cdot)$ is used.

12.4.2 Binary classification

For binary classification, the response variable is $Y \in \{0, 1\}$ with $\mathbf{P}[Y = 1|X = x] = \pi(x)$. Often, it is notationally more convenient to encode the response by $\tilde{Y} = 2Y - 1 \in \{-1, +1\}$.

We consider the negative binomial log-likelihood as loss function. As described in Section 3.3.1, formula (3.5), the loss function (using some scaling) is

$$\rho_{\text{log-lik}}(f, \tilde{y}) = \log_2(1 + \exp(-2\tilde{y}f)), \tag{12.4}$$

which then becomes an upper bound of the misclassification error, see Figure 12.1.

The population minimizer can be shown to be (Problem 12.2)

$$f^0_{\text{log-lik}}(x) = \frac{1}{2} \log\left(\frac{\pi(x)}{1 - \pi(x)}\right), \quad \pi(x) = \mathbf{P}[Y = 1|X = x], \tag{12.5}$$

see also Example 6.4 in Section 6.6.

We now briefly repeat the argument at the end of Section 3.3.1. The loss function in (12.4) is a function of $\tilde{y}f$, the so-called margin value (bearing some remote relations to the margin condition from statistical theory, see Section 6.4), where the function f induces the following classifier for Y:

$$\mathscr{C}(x) = \begin{cases} 1 \text{ if } f(x) > 0 \\ 0 \text{ if } f(x) \le 0. \end{cases}$$

Therefore, a misclassification happens if $\tilde{Y}f(X) < 0$. Hence, the misclassification loss is

$$\rho_{\text{misclass}}(f,y) = 1(\tilde{y}f < 0), \tag{12.6}$$

whose population minimizer is equivalent to the Bayes classifier (for $\tilde{Y} \in \{-1,+1\}$)

$$f^0_{\text{misclass}}(x) = \begin{cases} +1 \text{ if } \pi(x) > 1/2 \\ -1 \text{ if } \pi(x) \le 1/2, \end{cases}$$

where $\pi(x) = \mathbf{P}[Y = 1|X = x]$. Note that the misclassification loss in (12.6) cannot be used for boosting or FGD (Algorithm 7): it is discontinuous and also non-convex as a function of the margin value $\tilde{y}f$ or as a function of f. The negative log-likelihood loss in (3.5) can be viewed as a convex upper approximation of the (computationally intractable) non-convex misclassification loss, see Figure 12.1. We will describe in Section 12.4.4 the BinomialBoosting algorithm (similar to LogitBoost (Friedman et al., 2000)) which uses the negative log-likelihood as loss function.

Another upper convex approximation of the misclassification loss function in (12.6) is the exponential loss

$$\rho_{\text{exp}}(f,y) = \exp(-\tilde{y}f). \tag{12.7}$$

The population minimizer can be shown to be the same as for the log-likelihood loss (Problem 12.2):

$$f^*_{\text{exp}}(x) = \frac{1}{2}\log\left(\frac{\pi(x)}{1-\pi(x)}\right), \quad \pi(x) = \mathbf{P}[Y = 1|X = x]. \tag{12.8}$$

Using functional gradient descent (FGD) with different loss functions yields different boosting algorithms. When using the log-likelihood loss in (12.4), we obtain LogitBoost (Friedman et al., 2000) or BinomialBoosting from Section 12.4.4; and with the exponential loss in (12.7), we essentially get AdaBoost from Section 12.2.2.

We interpret the boosting estimate $\hat{f}^{[m]}(\cdot)$ as an estimate of the population minimizer $f^0(\cdot)$. Thus, the output from AdaBoost, Logit- or BinomialBoosting are estimates of the (log-odds ratio)/2. In particular, we define probability estimates via

$$\hat{p}^{[m]}(x) = \frac{\exp(\hat{f}^{[m]}(x))}{\exp(\hat{f}^{[m]}(x)) + \exp(-\hat{f}^{[m]}(x))}.$$

The standard loss function for support vector machines is the hinge loss:

$$\rho_{\text{hinge}}(f,y) = [1 - \tilde{y}f]_+,$$

where $[x]_+ = x1_{\{x>0\}}$ denotes the positive part of x. It is also an upper convex bound of the misclassification error, see Figure 12.1. Its population minimizer is again the Bayes classifier

$$f^0_{\text{hinge}}(x) = \text{sign}(\pi(x) - 1/2)$$

for $\tilde{Y} \in \{-1, +1\}$, see also Problem 6.8 in Chapter 6. Since $f^0_{\text{hinge}}(\cdot)$ is a classifier and non-invertible function of $\pi(x)$, there is no direct way to obtain conditional class probability estimates (that is, there is no unique solution in the equation above to solve for $\pi(x)$ as a function of $f^0_{\text{hinge}}(x)$). Motivated from the population point of

Loss functions for binary classification
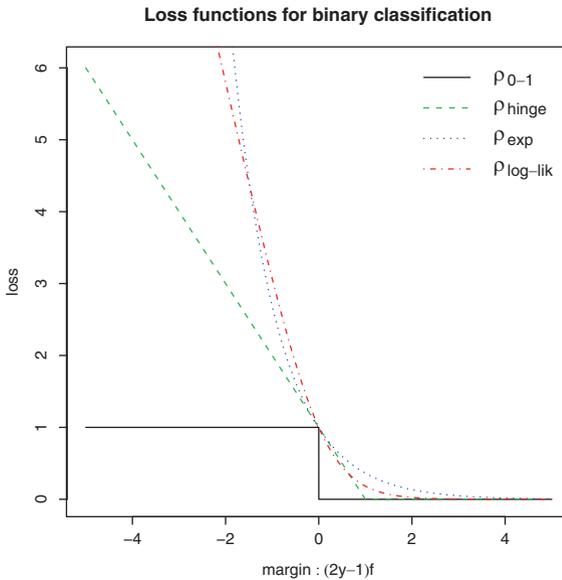
Fig. 12.1 Losses, as functions of the margin $\tilde{y}f = (2y-1)f$, for binary classification: misclassification (0-1), hinge, negative log-likelihood and exponential loss.

view, the L_2- or L_1-loss can also be used for binary classification. For $Y \in \{0, 1\}$, the population minimizers are

$$f^0_{L_2}(x) = \mathbb{E}[Y|X = x] = \pi(x) = \mathbf{P}[Y = 1|X = x],$$

$$f^0_{L_1}(x) = \text{median}(Y|X = x) = \begin{cases} 1 \text{ if } \pi(x) > 1/2 \\ 0 \text{ if } \pi(x) \le 1/2. \end{cases}$$

Thus, a population minimizer of the L_1-loss is the Bayes classifier. Moreover, both the L_1- and L_2-loss functions can be parametrized as functions of the margin value $\tilde{y}f$ ($\tilde{y} \in \{-1, +1\}$):

$$|\tilde{y} - f| = |1 - \tilde{y}f|,$$
$$|\tilde{y} - f|^2 = |1 - \tilde{y}f|^2 = 1 - 2\tilde{y}f + (\tilde{y}f)^2. \tag{12.9}$$

The L_1- and L_2-loss functions are non-monotone functions of the margin value $\tilde{y}f$. This is an undesirable property, implying that they assign a large loss if $\tilde{y}f$ takes on large values (greater than 1).

The negative log-likelihood loss in (3.5)

$$\rho_{\log-\text{lik}}(f, y) = \log_2(1 + \exp(-\tilde{y}f)),$$

has three nice properties: (i) it yields probability estimates; (ii) it is a monotone loss function of the margin value $\tilde{y}f$; (iii) it grows linearly as the margin value $\tilde{y}f$ tends to $-\infty$, unlike the exponential loss in (12.7). The third point reflects a robustness aspect: it is similar to Huber's loss function which also penalizes large values linearly (instead of quadratically as with the L_2-loss).

12.4.3 Poisson regression

For count data with $Y \in \{0, 1, 2, \ldots\}$, we can use Poisson regression: we assume that $Y|X = x$ has a Poisson$(\lambda(x))$ distribution and the goal is to estimate the function $f(x) = \log(\lambda(x))$. The negative log-likelihood yields the loss function

$$\rho(y, f) = -yf + \exp(f), \quad f = \log(\lambda),$$

which can be used in the functional gradient descent algorithm in Section 12.3.1.

12.4.4 Two important boosting algorithms

Table 12.1 summarizes the most popular loss functions and their corresponding boosting algorithms. The two algorithms appearing in the last two rows of the table are described next in more detail.

range spaces	$\rho(f, y)$	$f^0(x)$	algorithm
$y \in \{0,1\}, f \in \mathbb{R}$	$\exp(-(2y-1)f)$	$\frac{1}{2}\log\left(\frac{\pi(x)}{1-\pi(x)}\right)$	AdaBoost
$y \in \{0,1\}, f \in \mathbb{R}$	$\log_2(1 + e^{-2(2y-1)f})$	$\frac{1}{2}\log\left(\frac{\pi(x)}{1-\pi(x)}\right)$	LogitBoost / BinomialBoosting
$y \in \mathbb{R}, f \in \mathbb{R}$	$\frac{1}{2}|y - f|^2$	$\mathbb{E}[Y|X = x]$	L_2Boosting

Table 12.1 Various loss functions $\rho(y, f)$, population minimizers $f^0(x)$ and names of corresponding boosting algorithms; $\pi(x) = \mathbf{P}[Y = 1|X = x]$.

12.4.4.1 L_2Boosting

L_2Boosting is the simplest and perhaps most instructive boosting algorithm. It is very useful for (high-dimensional) regression. Applying the general description of the FGD Algorithm 7 from Section 12.3.1 to the squared error loss function $\rho_{L_2}(y,f) = |y - f|^2/2,$[1] we obtain the following algorithm.

Algorithm 8 L_2Boosting algorithm

1: Initialize $\hat{f}^{[0]}(\cdot)$ with an offset value. The default value is $\hat{f}^{[0]}(\cdot) \equiv \overline{Y}$. Set $m = 0$.
2: Increase m by one: $m \leftarrow m + 1$.
   Compute the residuals $U_i = Y_i - \hat{f}^{[m-1]}(X_i)$ for $i = 1, \ldots, n$.
3: Fit the residual vector $U_1, \ldots, U_n$ to $X_1, \ldots, X_n$ by the real-valued base procedure (e.g. regression)

$$(X_i, U_i)_{i=1}^n \xrightarrow{\text{base procedure}} \hat{g}^{[m]}(\cdot).$$

4: Up-date $\hat{f}^{[m]}(\cdot) = \hat{f}^{[m-1]}(\cdot) + \nu \cdot \hat{g}^{[m]}(\cdot)$, where $0 < \nu \leq 1$ is a step-length factor (as in the general FGD Algorithm 7).
5: Iterate steps 2 to 4 until $m = m_{\text{stop}}$ for some stopping iteration m_{stop}.

The stopping iteration m_{stop} is the main tuning parameter which can be selected using cross-validation.

The derivation from the generic FGD Algorithm 7 in Section 12.3.1 is straightforward by noting that the negative gradient vector becomes the standard residual vector. Thus, L_2Boosting amounts to refitting residuals multiple times. Tukey (1977) recognized this to be useful and proposed "twicing" which is L_2Boosting using $m_{\text{stop}} = 2$ and $\nu = 1$.

12.4.4.2 BinomialBoosting: the FGD version of LogitBoost

We already gave some reasons at the end of Section 12.4.2 why the negative log-likelihood loss function in (12.4) is very useful for binary classification problems. Friedman et al. (2000) were first in advocating this, and they proposed LogitBoost which is very similar to the generic FGD Algorithm 7 when using the loss from (12.4): the deviation from the generic FGD is due to using Newton's method involving the Hessian matrix (instead of a step-length for the gradient).

For the sake of coherence with the generic FGD algorithm in Section 12.3.1, we describe here BinomialBoosting (Bühlmann and Hothorn, 2007) which is a version of LogitBoost.

---

[1] The factor $1/2$ leads to a convenient notation where the evaluated negative gradient of the loss function becomes the standard residual vector.

Algorithm 9 BinomialBoosting algorithm

Apply the generic FGD Algorithm 7 from Section 12.3.1 using the loss function $\rho_{\log-\text{lik}}$ from (12.4). The default offset value is $\hat{f}^{[0]}(\cdot) \equiv \log(\hat{p}/(1-\hat{p}))/2$, where $\hat{p}$ is the relative frequency of $Y = 1$.

With BinomialBoosting, there is no need that the base procedure is able to do weighted fitting: this constitutes a slight difference to the requirement for Logit-Boost (Friedman et al., 2000).

12.4.5 Other data structures and models

Due to the generic nature of boosting or functional gradient descent, we can use the technique in very many other settings. For data with univariate responses and loss functions which are differentiable (almost everywhere) with respect to the first argument (the function f), the boosting algorithm is described in Section 12.3.1.

A slightly less standard example than regression or classification is survival analysis. The negative logarithm of the Cox' partial likelihood can be used as a loss function for fitting proportional hazards models to censored response variables with boosting algorithms (Ridgeway, 1999).

12.5 Choosing the base procedure

Every boosting algorithm requires the specification of a base procedure or weak learner. This choice can be driven by the goal of optimizing the predictive capacity only or by considering some structural properties of the boosting estimate in addition. The latter has the advantage that it allows for better interpretation of the resulting fitted model.

We recall that the generic boosting estimator is a sum of base procedure estimates

$$\hat{f}^{[m]}(\cdot) = \nu \sum_{k=1}^{m} \hat{g}^{[k]}(\cdot) + \hat{f}^{[0]}(\cdot).$$

Therefore, structural properties of the boosting function estimator are induced by a linear combination of structural characteristics of the base procedure.

We discuss next some important examples of base procedures yielding useful structures for the boosting estimator $\hat{f}^{[m]}(\cdot)$. The notation is as follows: $\hat{g}(\cdot)$ is an estimate from a base procedure which is based on data $(X_1, U_1), \ldots, (X_n, U_n)$ where $(U_1, \ldots, U_n)$ denotes the current negative gradient vector (of the loss).

12.5.1 Componentwise linear least squares for generalized linear models

Boosting can be very useful for fitting potentially high-dimensional generalized linear models

$$Y_1, \ldots, Y_n \text{ independent}$$

$$h(\mathbb{E}[Y_i | X_i = x]) = \sum_{j=1}^{p} \beta_j x^{(j)},$$

as described in (3.1) in Chapter 3 (but we denote here the link function by $h(\cdot)$ instead of $g(\cdot)$). We have dropped an intercept term μ which we implicitly incorporate into the covariate x.

Consider the base procedure

$$\hat{g}(x) = \hat{\gamma}_{\hat{j}} x^{(\hat{j})},$$

$$\hat{\gamma}_j = \sum_{i=1}^{n} X_i^{(j)} U_i / \sum_{i=1}^{n} \left(X_i^{(j)} \right)^2, \ \hat{j} = \arg\min_{1 \le j \le p} \sum_{i=1}^{n} \left(U_i - \hat{\gamma}_j X_i^{(j)} \right)^2. \quad (12.10)$$

It selects the best variable in a simple linear model in the sense of ordinary least squares fitting and uses the corresponding estimate $\hat{\gamma}_{\hat{j}}$. It is instructive to give the equivalent formulation (Problem 12.3):

$$\hat{j} = \arg\max_{j=1,\ldots,p} \frac{|\sum_{i=1}^{n} U_i X_i^{(j)}|^2}{\sum_{i=1}^{n} (X_i^{(j)})^2}. \quad (12.11)$$

In case of centered predictor variables with $n^{-1} \sum_{i=1}^{n} X_i^{(j)} = 0$, this is saying that $\hat{j}$ chooses the variable which maximizes the absolute correlation with the residual vector.

When using the FGD Algorithm 7 (i.e. any of the gradient boosting methods) with this base procedure, we select in every iteration one predictor variable, not necessarily a different one, and we up-date the function linearly:

$$\hat{f}^{[m]}(x) = \hat{f}^{[m-1]}(x) + \nu \hat{\gamma}_{\hat{j}_m} x^{(\hat{j}_m)},$$

where $\hat{j}_m$ denotes the index of the selected predictor variable in iteration m. Alternatively, the up-date of the coefficient estimates is

$$\hat{\beta}_j^{[m]} = \begin{cases} \hat{\beta}_j^{[m-1]} + \nu \cdot \hat{\gamma}_j & \text{if } j = \hat{j}_m, \\ \hat{\beta}_j^{[m-1]} & \text{if } j \ne \hat{j}_m. \end{cases}$$

Thus, only the $\hat{j}_m$th component of the coefficient estimate $\hat{\beta}^{[m]}$ is up-dated. We summarize that we obtain a linear fit $\hat{f}^{[m]}(\cdot)$ for the population minimizer $f^0(\cdot)$ of the loss function.

With L_2Boosting (Algorithm 8) and the componentwise linear least squares base procedure in (12.10) we obtain a linear model fit for every iteration m. As m tends to infinity, $\hat{f}^{[m]}(\cdot)$ converges to a least squares solution. The method is also known as matching pursuit in signal processing (Mallat and Zhang, 1993), weak greedy algorithm in computational mathematics (Temlyakov, 2000), and it is a Gauss-Southwell algorithm (Southwell, 1946) for solving a linear system of equations. We will discuss statistical properties of L_2Boosting (Algorithm 8) with componentwise linear least squares in Section 12.6.2.

Using BinomialBoosting with componentwise linear least squares from (12.10), we obtain a fit, including variable selection, of a linear logistic regression model. The reason is as follows. The loss function from (12.4) has $1/2$ times the log odds ratio as population minimizer: $f^0(x) = \log\{\pi(x)/(1-\pi(x))\}/2$, see (12.5). Furthermore, $\hat{f}^{[m]}(x)$ is linear in x, as discussed above. And hence, since $\hat{f}^{[m]}(x)$ is an estimate of $f^0(x) = \log\{\pi(x)/(1-\pi(x))\}/2$, we conclude that BinomialBoosting is fitting a linear logistic regression model (whose fitted regression coefficients are to be multiplied by a factor of 2 when using the standard logit link $\log(\pi(x)/(1-\pi(x))$ without the factor $1/2$).

As will be discussed in more detail in Section 12.6.2, boosting typically shrinks the (generalized) regression coefficients towards zero. Usually, we do not want to shrink the intercept term. In addition, we advocate to use boosting on mean centered predictor variables $\tilde{X}_i^{(j)} = X_i^{(j)} - \overline{X}^{(j)}$. In case of a linear model, when centering also the response $\tilde{Y}_i = Y_i - \overline{Y}$, this becomes

$$\tilde{Y}_i = \sum_{j=1}^{p} \beta^{(j)} \tilde{X}_i^{(j)} + \text{noise}_i,$$

forcing the regression surface through the center $(\overline{\tilde{x}}^{(1)}, \ldots, \overline{\tilde{x}}^{(p)}, \overline{\tilde{y}}) = (0, 0, \ldots, 0)$ as with ordinary least squares and avoiding the use of an intercept term. Conceptually, it is not necessary to center the response variables when using the default offset value $\hat{f}^{[0]} = \overline{Y}$ in the L_2Boosting Algorithm 8 (for e.g. BinomialBoosting, we would center the predictor variables only but never the response, and we would use $\hat{f}^{[0]} \equiv \arg\min_c n^{-1} \sum_{i=1}^{n} \rho(Y_i, c)$).

12.5.2 Componentwise smoothing spline for additive models

Additive and generalized additive models have become very popular, allowing for more flexibility than the linear structure in generalized linear models (Hastie and

Tibshirani, 1990). We discussed in Chapter 5 how penalty-based methods can be used to fit such models in high-dimensional problems. Here we show that boosting can be used as an alternative procedure.

We use a nonparametric base procedure for function estimation. To fix ideas, we consider the case where

$\hat{g}_j(\cdot)$ is a least squares cubic smoothing spline estimate based on

$U_1,\ldots,U_n$ against $X_1^{(j)},\ldots,X_n^{(j)}$ with fixed degrees of freedom df. (12.12)

That is,

$$\hat{g}_j(\cdot) = \underset{f(\cdot)}{\arg\min} \left(n^{-1} \sum_{i=1}^{n} \left(U_i - f(X_i^{(j)}) \right)^2 + \lambda \int (f''(x))^2 dx \right), \quad (12.13)$$

where $\lambda > 0$ is a tuning parameter such that the trace of the corresponding hat matrix equals df. For further details, we refer to Green and Silverman (1994).

The base procedure is then

$$\hat{g}(x) = \hat{g}_{\hat{j}}(x^{(\hat{j})}),$$

$$\hat{g}_j(\cdot) \text{ as above and } \hat{j} = \underset{1 \le j \le p}{\arg\min} \sum_{i=1}^{n} \left(U_i - \hat{g}_j(X_i^{(j)}) \right)^2,$$

where the degrees of freedom df are the same for all $\hat{g}_j(\cdot)$.

L_2Boosting (Algorithm 8) with componentwise smoothing splines yields an additive model, including variable selection, i.e., a fit which is additive in the predictor variables. This can be seen immediately since L_2Boosting proceeds additively for up-dating the function $\hat{f}^{[m]}(\cdot)$, see Section 12.4.4. We can finally normalize to obtain the following additive model estimator:

$$\hat{f}^{[m]}(x) = \hat{\mu} + \sum_{j=1}^{p} \hat{f}_j^{[m]}(x^{(j)}),$$

$$n^{-1} \sum_{i=1}^{n} \hat{f}_j^{[m]}(X_i^{(j)}) = 0 \text{ for all } j = 1,\ldots,p.$$

Figure 12.2 illustrates such additive model fitting with componentwise smoothing splines where the stopping iteration m has been selected with an AIC-type criterion as proposed in Bühlmann (2004, 2006). The underlying model is

$$Y_i = \sum_{j=1}^{10} f_j(X_i^{(j)}) + \varepsilon_i, \; i = 1,\ldots,n = 200,$$

$$X_1,\ldots,X_n \sim \text{ Uniform}([0,1]^p), \; p = 100,$$

$$\varepsilon_1,\ldots,\varepsilon_n \text{ i.i.d. } \sim \mathcal{N}(0,\sigma^2), \; \sigma^2 = 0.5,$$
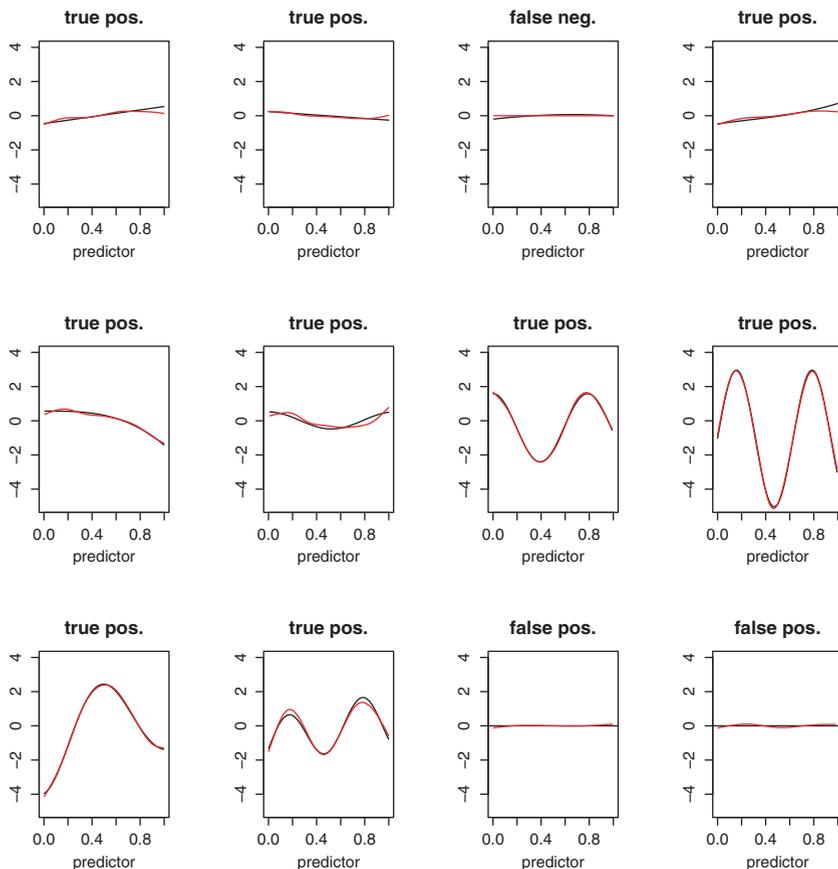
Fig. 12.2 Additive model: $n = 200$, $p = 100$. Black: true functions; Red: estimated functions using L_2Boosting with componentwise smoothing splines. Number of active variables is 10: there are one false negative, two false positive, 9 true positive and 88 true negative (not shown) selections. The figure is essentially taken from Bühlmann (2004).

and $\varepsilon_1, \ldots, \varepsilon_n$ are independent of $X_1, \ldots, X_n$. The different functions $f_j(\cdot)$, having different "complexity", are displayed in Figure 12.2. The model contains 90 noise covariates. Based on one representative realization from this model, Figure 12.2 shows the function estimates $\hat{f}_j$ for $j = 1, \ldots, 10$, i.e., for the true active variables, and also for two other variables which correspond to false positive selections where $\hat{f}_j(\cdot) \not\equiv 0$ but $f_j(\cdot) \equiv 0$. In total (among the $p = 100$ variables), there are 2 false positive and one false negative selections. We also see from Figure 12.2 that the boosting additive model estimate yields a very reasonable fit, given the high-dimensional na-

ture of the problem.[2] The three false selections are close to the true underlying functions. Alternative estimators for high-dimensional additive models are discussed in Chapter 5 and in Section 8.4: the theoretical properties are much better developed for such ℓ_1/ℓ_2-penalized estimators than for boosting.

The degrees of freedom in the smoothing spline base procedure should be chosen "small" such as df $= 2.5$ which we used in Figure 12.2. This yields low variance but typically large bias of the base procedure. The bias can then be reduced by additional boosting iterations. This choice of low variance but high bias has been analyzed in Bühlmann and Yu (2003) and we discuss it also in Section 12.5.4.

Componentwise smoothing splines can be generalized to pairwise smoothing splines which searches for and fits the best pairs of predictor variables such that a smooth of $U_1, \ldots, U_n$ against this pair of predictors reduces the residual sum of squares most. With L_2Boosting (Algorithm 8), this yields a nonparametric model fit with first order interaction terms (since we fit an additive combination of smooth functions in two covariables). The procedure has been empirically demonstrated to work rather well in high-dimensional problems (Bühlmann and Yu, 2006).

As with the componentwise linear least squares base procedure, we can use componentwise smoothing splines also in BinomialBoosting (Algorithm 9), yielding an additive logistic regression fit. Conceptually, there is nothing special about choosing a smoothing spline estimator in (12.13). One could use other (smooth) function estimators, such as e.g. regression or P-splines (Schmid and Hothorn, 2008) which are computationally more efficient than smoothing splines.

12.5.3 Trees

In the machine learning community, regression trees are the most popular base procedures. They have the advantage to be invariant under monotone transformations of predictor variables, i.e., we do not need to search for good data transformations. Moreover, regression trees handle covariates measured at different scales (continuous, ordinal or nominal variables) in a unified way.

When using stumps, i.e., a tree with two terminal nodes only, the boosting estimate will be an additive model in the original predictor variables, because every stump-estimate is a function of a single predictor variable only. Similarly, boosting trees with (at most) d terminal nodes results in a nonparametric function estimate having at most interactions of order $d - 2$ (in our terminology, an additive function has interaction degree equal to zero). Therefore, if we want to constrain the degree of interactions, we can easily do this by constraining the (maximal) number of nodes in the base procedure.

---

[2] We should take into account that the covariates are generated independently and the noise variance $\sigma^2 = 0.5$ is rather small.

12.5.4 The low-variance principle

We have seen above that the structural properties of a boosting estimate are determined by the choice of a base procedure. The structure specification should come first. After having made a choice, the question becomes how "complex" the base procedure should be. For example, how should we choose the degrees of freedom for the componentwise smoothing spline in (12.12)? A general answer is to choose the base procedure (having the desired structure) with low variance at the price of larger estimation bias. For the componentwise smoothing splines, this would imply a low number of degrees of freedom, e.g., $df = 4$.

We give some reasons for the low-variance principle in Section 12.6.1. Moreover, it has been demonstrated in Friedman (2001) that a small step-size factor ν in Step 4 of the generic FGD Algorithm 7 (or specialized versions thereof) can be often beneficial and almost never yields substantially worse predictive performance of boosting estimates. Note that a small step-size factor can be seen as a shrinkage of the base procedure by the factor ν, implying low variance but potentially large estimation bias.

12.5.5 Initialization of boosting

We have briefly described in Sections 12.3.1 and 12.5.1 the issue of choosing an initial value $\hat{f}^{[0]}(\cdot)$ for boosting. This can be quite important for applications where we would like to estimate parts of a model in an unpenalized (non-regularized) fashion and others being subject to regularization.

For example, we may think of a parametric form of $\hat{f}^{[0]}(\cdot)$, estimated by maximum likelihood, and deviations from the parametric model would be built in by pursuing boosting iterations (with a nonparametric base procedure). A concrete example would be: $\hat{f}^{[0]}(\cdot)$ is the maximum likelihood estimate in a generalized linear model and boosting would be done with componentwise smoothing splines to model additive deviations from a generalized linear model. A related strategy has been used in Audrino and Bühlmann (2003) for modeling multivariate volatility in financial time series.

Another example would be a linear model $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ as in (12.19) where some of the covariates, say the first q predictor variables $X^{(1)}, \ldots, X^{(q)}$, enter the estimated linear model in an unpenalized way. We propose to do ordinary least squares regression on $X^{(1)}, \ldots, X^{(q)}$: consider the projection P_q onto the linear span of $X^{(1)}, \ldots, X^{(q)}$ and use L_2Boosting (Algorithm 8) with componentwise linear least squares on the new response $(I - P_q)\mathbf{Y}$ and the new $(p - q)$-dimensional predictor $(I - P_q)\mathbf{X}$. The final model estimate is then

$$\hat{f}^{[m]}(x) = \sum_{j=1}^{q} \hat{\beta}_{\mathrm{OLS},j} x^{(j)} + \sum_{j=q+1}^{p} \hat{\beta}_{j}^{[m_{\mathrm{stop}}]} \tilde{x}^{(j)},$$

where the latter part is from L_2Boosting and $\tilde{x}^{(j)}$ is the residual when linearly re-gressing $x^{(j)}$ to $x^{(1)}, \ldots, x^{(q)}$. A special case which is used in most applications is with $q = 1$ and $X^{(1)} \equiv 1$ encoding for an intercept. Then, $(I - P_1)Y = Y - \overline{Y}$ and $(I - P_1)X^{(j)} = X^{(j)} - n^{-1}\sum_{i=1}^{n} X_i^{(j)}$. This is exactly the proposal at the end of Section 12.5.1. For generalized linear models, analogous concepts can be used.

12.6 L_2Boosting

As described in Section 12.4.4.1, the L_2Boosting Algorithm 8 is a functional gradient descent using the squared error loss which amounts to repeated fitting of ordinary residuals. Here, we aim at better understanding of such a simple L_2Boosting algorithm. We first start with a toy problem of curve estimation and we will then describe a result for high-dimensional linear models.

12.6.1 Nonparametric curve estimation: some basic insights about boosting

We study the toy problem of estimating a regression function $\mathbb{E}[Y|X = x]$ with one-dimensional predictor $X \in \mathscr{X} \subseteq \mathbb{R}$ and a continuous response $Y \in \mathbb{R}$ in the following model:

$$Y_i = f^0(X_i) + \varepsilon_i, \ i = 1, \ldots, n,$$
$$\varepsilon_1, \ldots, \varepsilon_n \text{ i.i.d. with } \mathbb{E}[\varepsilon_i] = 0, \ \mathrm{Var}(\varepsilon_i) = \sigma^2, \qquad (12.14)$$

where $f^0(\cdot)$ is a real-valued, typically nonlinear function, and the predictors $X_i \in \mathscr{X} \subseteq \mathbb{R}$ are deterministic.

Consider the case with a linear base procedure having a hat matrix $\mathscr{H} : \mathbb{R}^n \to \mathbb{R}^n$ (not necessarily symmetric), mapping the response variables $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ to their fitted values $(\hat{f}(X_1), \ldots, \hat{f}(X_n))^T$. Examples include nonparametric kernel smoothers or the smoothing spline in (12.13). Then, the hat matrix of the L_2Boosting fit (for simplicity, with $\hat{f}^{[0]} \equiv 0$ and $\nu = 1$) in iteration m equals (Problem 12.4):

$$\mathscr{B}_m = \mathscr{B}_{m-1} + \mathscr{H}(I - \mathscr{B}_{m-1}) = I - (I - \mathscr{H})^m. \qquad (12.15)$$

Formula (12.15) allows for several insights. First, if the base procedure satisfies $\|I - \mathscr{H}\| < 1$ for a suitable norm, i.e., it has a "learning capacity" such that the

residual vector is shorter than the input-response vector, we see that $\mathscr{B}_m$ converges to the identity I as $m \to \infty$, and $\mathscr{B}_m \mathbf{Y}$ converges to the fully saturated model $\mathbf{Y}$, interpolating the response variables exactly. Thus, we exploit here explicitly that we have to stop early with the boosting iterations in order to prevent over-fitting.

When specializing to the case of a cubic smoothing spline base procedure, see (12.13), it is useful to invoke some eigen-analysis (the generalization to a smoothing spline of order r is treated in Theorem 12.1 below). The spectral representation is

$$\mathscr{H} = UDU^T, \;\; U^TU = UU^T = I, \; D = \mathrm{diag}(\lambda_1, \ldots, \lambda_n),$$

where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ denote the (ordered) eigenvalues of $\mathscr{H}$. It then follows with (12.15) that

$$\begin{aligned} \mathscr{B}_m &= UD_mU^\top, \\ D_m &= \mathrm{diag}(d_{1,m}, \ldots, d_{n,m}), \quad d_{i,m} = 1 - (1 - \lambda_i)^m. \end{aligned}$$

It is well known (Green and Silverman, 1994) that a cubic smoothing spline satisfies:

$$\lambda_1 = \lambda_2 = 1, \;\; 0 < \lambda_i < 1 \; (i = 3, \ldots, n).$$

Therefore, the eigenvalues of the boosting hat operator (matrix) in iteration m satisfy:

$$\begin{aligned} &d_{1,m} \equiv d_{2,m} \equiv 1 \text{ for all m}, \\ &0 < d_{i,m} = 1 - (1 - \lambda_i)^m < 1 \; (i = 3, \ldots, n), \quad d_{i,m} \to 1 \; (m \to \infty). \end{aligned}$$

When comparing the spectrum, i.e., the set of eigenvalues, of a smoothing spline with its boosted version, we observe the following. For both cases, the largest two eigenvalues are equal to 1. Moreover, all other eigenvalues can be changed by either varying the degrees of freedom df $= \sum_{i=1}^{n} \lambda_i$ in a single smoothing spline, or by varying the boosting iteration m when using a fixed (low-variance) smoothing spline base procedure having fixed (low) values λ_i. In Figure 12.3 we demonstrate the difference between the two approaches for changing "complexity" of the estimated curve in this toy example (first shown in Bühlmann and Yu (2003)). Both methods have about the same minimum mean squared error but L_2Boosting overfits much more slowly than a single smoothing spline.

By careful inspection of the eigen-analysis for this simple case of boosting a smoothing spline, Bühlmann and Yu (2003) proved an asymptotic minimax rate for the mean squared error

$$n^{-1} \sum_{i=1}^{n} \mathbb{E}[(\hat{f}^{[m]}(X_i) - f^0(X_i))^2]. \tag{12.16}$$

We make the following assumption on the design points:
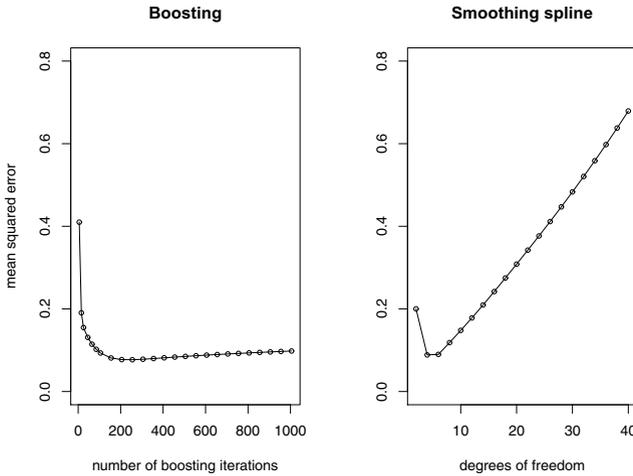
Fig. 12.3 Mean squared prediction error $\mathbb{E}[(\hat{f}(X) - f^0(X))^2]$ for the regression model $Y_i = 0.8X_i + \sin(6X_i) + \varepsilon_i$ ($i = 1,\ldots,n = 100$), with $\varepsilon \sim \mathcal{N}(0,2), X_i \sim$ Uniform$(-1/2,1/2)$, averaged over 100 simulation runs. Left: L_2Boosting with smoothing spline base procedure (having fixed degrees of freedom df = 4) and using $\nu = 0.1$, for varying number of boosting iterations. Right: single smoothing spline with varying degrees of freedom. The figure is taken from Bühlmann and Hothorn (2007).

(A)   The predictor variables $X_1,\ldots,X_n \in [a,b]$ ($-\infty < a < b < \infty$) are deterministic and satisfy: there exists a positive constant B such that for every n,

$$\frac{\sup_{x \in [a,b]} \inf_{1 \leq i \leq n} |x - X_i|}{\inf_{1 \leq i \neq j \leq n} |X_i - X_j|} \leq B < \infty.$$

Assumption (A) holds for the equidistant design and is almost surely fulfilled for i.i.d. realizations from a suitably regular probability distribution on $[a,b]$. Denote the Sobolev space by

$$\mathscr{F}^{(r)} = \{f : f \ (r-1)\text{-times continuously differentiable and } \int_a^b (f^{(r)}(x))^2 dx < \infty\}.$$

The smoothing spline corresponding to smoothness r is then defined as

$$\hat{g}_{\lambda,r}(\cdot) = \underset{g(\cdot) \in \mathscr{F}^{(r)}}{\arg\min} \ n^{-1} \sum_{i=1}^{n} (U_i - g(X_i))^2 + \lambda \int (g^{(r)}(x))^2 dx, \qquad (12.17)$$

for data $(X_1,U_1),\ldots,(X_n,U_n)$.

Theorem 12.1. *Consider the model in (12.14) with $X_i \in [a,b]$ satisfying (A). Suppose $\mathscr{H}$ is a smoothing spline $\hat{g}_{\lambda_0,r}(\cdot)$ corresponding to a fixed smoothing parame-*

ter λ_0 and smoothness $r \in \mathbb{N}$, as in (12.17). If the true function f^0 is in $\mathscr{F}^{(\xi)}$ with $\xi \geq r$ $(\xi,\ r \in \mathbb{N})$, then there is a boosting iteration $m = m(n) = O(n^{2r/(2\xi+1)}) \to \infty$ $(n \to \infty)$ such that $\hat{f}^{[m]}(\cdot)$ achieves the optimal minimax rate $n^{-2\xi/(2\xi+1)}$ of the function class $\mathscr{F}^{(r)}$ in terms of MSE as defined in (12.16).

A proof is given in Section 12.8.1. Two items are interesting. First, minimax rates are achieved by using a base procedure with fixed degrees of freedom which means low variance from an asymptotic perspective. Secondly, L_2Boosting with cubic smoothing splines has the capability to capture higher order smoothness of the true underlying function (without the need of choosing a higher order spline base procedure). Asymptotic convergence and minimax rate results have been established for early-stopped boosting in much more general settings by Yao et al. (2007) and Bissantz et al. (2007).

12.6.1.1 L_2Boosting using kernel estimators

As pointed out above, L_2Boosting of smoothing splines can achieve faster mean squared error convergence rates than the classical $O(n^{-4/5})$, assuming that the true underlying function is sufficiently smooth (corresponding to $r = 2$ in (12.17)). We illustrate here a related phenomenon with kernel estimators.

We consider fixed, univariate design points $X_i = i/n$ $(i = 1, \ldots, n)$ and the Nadaraya-Watson kernel estimator for the nonparametric regression function $\mathbb{E}[Y|X = x]$:

$$\hat{g}(x;h) = (nh)^{-1} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) Y_i = n^{-1} \sum_{i=1}^{n} K_h(x - X_i) Y_i,$$

where $h > 0$ is the bandwidth, $K(\cdot)$ a kernel in the form of a probability density which is symmetric around zero and $K_h(x) = h^{-1}K(x/h)$. It is straightforward to derive the form of L_2Boosting using $m = 2$ iterations (with $\hat{f}^{[0]} \equiv 0$ and $\nu = 1$), i.e., twicing (Tukey, 1977) with the Nadaraya-Watson kernel estimator:

$$\hat{f}^{[2]}(x) = (nh)^{-1} \sum_{i=1}^{n} K_h^{\text{tw}}(x - X_i) Y_i, \ K_h^{\text{tw}}(u) = 2K_h(u) - K_h * K_h(u), \quad (12.18)$$

where $K_h * K_h(u) = n^{-1} \sum_{r=1}^{n} K_h(u - X_r) K_h(X_r)$ (Problem 12.5). For fixed design points $X_i = i/n$, the kernel $K_h^{\text{tw}}(\cdot)$ is asymptotically equivalent to a higher-order kernel (which can take negative values) yielding a squared bias term of order $O(h^8)$, assuming that the true regression function is four times continuously differentiable. Thus, twicing or L_2Boosting with $m = 2$ iterations amounts to be a Nadaraya-Watson kernel estimator with a higher-order kernel. This explains from another angle why boosting is able to improve the mean squared error rate of the base procedure. More details including also non-equispaced designs are given in DiMarzio and Taylor (2008).

12.6.2 L_2Boosting for high-dimensional linear models

We look here at the problem of fitting a high-dimensional linear model which we have extensively treated in Chapters 2 and 6 when using the Lasso. Consider the linear model as in e.g. formula (2.1) in Chapter 2:

$$Y_i = \sum_{j=1}^p \beta_j X_i^{(j)} + \varepsilon_i, \ i = 1,\ldots,n, \tag{12.19}$$

where $\varepsilon_1,\ldots,\varepsilon_n$ are i.i.d. with $\mathbb{E}[\varepsilon_i] = 0$ and independent from all X_i's, and we allow the number of predictors p to be much larger than the sample size n.

Estimating the model (12.19) can be done using L_2Boosting with the componentwise linear least squares base procedure from Section 12.5.1 which fits in every iteration the best predictor variable reducing the residual sum of squares most. This method has some basic properties which are shared by the Lasso as well (see Section 12.6.2.1): when stopping early which is needed to avoid over-fitting, the L_2Boosting method (often) does variable selection, and the coefficient estimates $\hat{\beta}^{[m]}$ are (typically) shrunken versions of a least squares estimate $\hat{\beta}_{\mathrm{OLS}}$.

12.6.2.1 Connections to the Lasso

Hastie et al. (2001) pointed out an intriguing connection between L_2Boosting with componentwise linear least squares and the Lasso. Efron et al. (2004) made the connection rigorous and explicit: they consider a version of L_2Boosting, called forward stagewise linear regression (FSLR), and they show that FSLR with infinitesimally small step-sizes (i.e., the value ν in step 4 of the L_2Boosting algorithm in Section 12.4.4.1) produces a set of solutions which is equivalent (as step-sizes tend to zero) to the set of Lasso solutions when varying the regularization parameter λ in the Lasso

$$\hat{\beta}(\lambda) = \arg\min_\beta \left(\|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + \lambda\|\beta\|_1 \right).$$

However, such an equivalence only holds if the design matrix $\mathbf{X}$ satisfies a restrictive "positive cone condition" (Efron et al., 2004).

We briefly illustrate the similarity between L_2Boosting and Lasso for the riboflavin data ($p = 4088$), described in Section 9.2.6, but using here a more homogeneous (sub-)sample of size $n = 71$. We estimate a high-dimensional linear model, once with L_2Boosting using componentwise linear least squares and step-size $\nu = 0.1$, and once with the Lasso. We regularize both methods such that 32 variables among the $p = 4088$ variables are selected, by appropriately choosing the stopping iteration or the penalty parameter, respectively. Among these 32 variables from each method,

22 are selected by both L_2Boosting and Lasso: thus, the overlap of jointly selected variables is quite remarkable. Furthermore, Figure 12.4 shows the estimated coeffi-
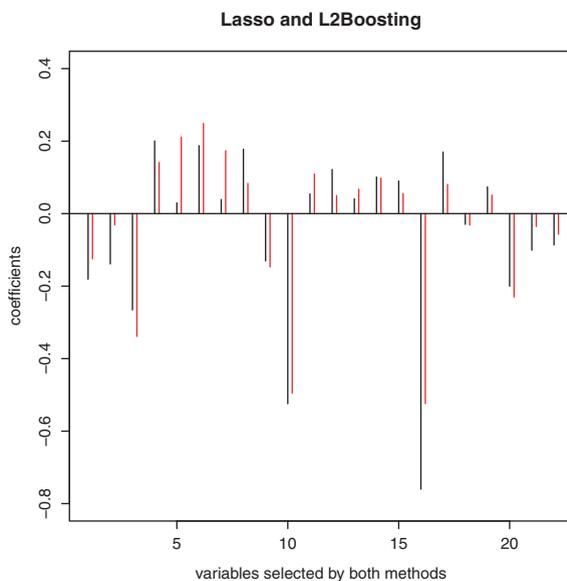

Fig. 12.4 Linear model coefficients for riboflavin data ($n = 71$, $p = 4088$). Estimated coefficients from Lasso (black) and L_2Boosting (red) of 22 variables which are selected by both methods, where each method alone chooses 32 variables. The 10 estimated coefficients corresponding to variables which are selected by one of the methods only have values in the range $[-0.07, 0.18]$.

cients $\hat{\beta}_j$ of these 22 variables: for each among the 22 variables, the estimates from L_2Boosting and Lasso exhibit consistently the same sign and in fact, their numerical values are quite close to each other. The coefficients corresponding to variables chosen by one of the methods only are relatively small, in a range of $[-0.07, 0.18]$.

Despite the fact that L_2Boosting and Lasso are not equivalent methods in general, it may be useful to interpret boosting as being "related" to ℓ_1-penalty based methods.

12.6.2.2 Asymptotic consistency in high dimensions

We present here a result on asymptotic consistency for prediction in the high-dimensional but sparse linear model as in (12.19). To capture the notion of high-dimensionality, we use a triangular array of observations as described in formula (2.6) in Section 2.4:

$$Y_{n;i} = \sum_{j=1}^{p_n} \beta_{n;j}^0 X_{n;i}^{(j)} + \varepsilon_{n;i}, \ i = 1,\ldots,n; \ \ n = 1,2,\ldots \quad (12.20)$$

and we use the short hand notation $\mathbf{Y} = \mathbf{X}\beta_n^0 + \varepsilon$.

We make the following assumptions.

(A1)   The number of covariates p_n in model (12.20) satisfies

$$\log(p_n)/n \to 0 \ \ (n \to \infty).$$

(A2)   The covariates are deterministic (fixed design) and scaled

$$n^{-1}\sum_{i=1}^{n}(X_{n;i}^{(j)})^2 \equiv 1 \text{ for all } j = 1,\ldots,p_n,$$

and the model is sparse with respect to the ℓ_1-norm

$$\|\beta_n^0\|_1 = \sum_{j=1}^{p} |\beta_{n;j}^0| = o\left(\sqrt{\frac{n}{\log(p_n)}}\right) \ \ (n \to \infty).$$

and the regression function satisfies

$$\|\mathbf{X}\beta_n^0\|_2^2/n = n^{-1}\sum_{i=1}^{n} \left(f_n^0(X_{n;i})\right)^2 \le C < \infty \text{ for all } n \in \mathbb{N},$$

where $f_n^0(x) = x\beta_n^0$.

(A3)   The errors satisfy

$$\varepsilon_{n;1},\ldots\varepsilon_{n;n} \text{ i.i.d. } \sim \mathcal{N}(0,\sigma^2) \text{ for all } n \in \mathbb{N}.$$

with $0 < \sigma^2 < \infty$.

Assumption (A1) is standard in high-dimensional asymptotics, as discussed in Chapter 6. The scaling in assumption (A2) is without loss of generality and the ℓ_1-norm sparsity assumption is as for the Lasso (see below). The third assumption in (A2) says, together with assumption (A3), that the signal to noise ratio is not increasing as $n \to \infty$. Assumption (A3) is made to simplify the mathematical analysis but it could be relaxed to assuming 4th moments $\mathbb{E}|\varepsilon_{n;i}|^4 \le M < \infty$ only, see Section 14.5.3 in Chapter 14, and under additional assumptions on the design we can further relax to assume second moments only, see Problem 14.6 in Chapter 14. The following theorem is a refinement of a result in Bühlmann (2006).

Theorem 12.2. *Consider the model (12.20) satisfying (A1)-(A3). Then the L Boosting estimator $\hat{f}_n^{[m]}(\cdot)$ with the componentwise linear least squares base procedure satisfies: for*

$$m_n \to \infty, \ m_n = o(\sqrt{n/\log(p_n)}) \ (n \to \infty),$$

we have,

$$\|\mathbf{X}(\hat{\beta}_n^{[m_n]} - \beta_n^0)\|_2^2/n = o_P(1) \ (n \to \infty).$$

Furthermore, for $m_n = K\sqrt{\frac{n}{\log(p_n)}}^{\frac{3-\delta}{4-2\delta}}$ with any constant $0 < K < \infty$ and $0 < \delta <$
5/8, we have

$$\|\mathbf{X}(\hat{\beta}_n^{[m_n]} - \beta_n^0)\|_2^2/n = O_P\left(\max\left\{ \sqrt{\frac{\log(p_n)}{n}}^{\frac{1-\delta}{2(2-\delta)}}, \|\beta_n^0\|_1\sqrt{\frac{\log(p_n)}{n}} \right\} \right) \ (n \to \infty).$$

A proof is given in Section 12.8.2. We note that the assertion of the theorem is about
prediction in terms of $n^{-1}\sum_{i=1}^n (\hat{f}^{[m_n]}(X_{n;i}) - f_n^0(X_{n;i}))^2$: the presented rate is best for
$\delta > 0$ close to zero resulting in

$$O_P\left(\max\left\{ \sqrt{\frac{\log(p_n)}{n}}^{\frac{1}{4}-\delta'}, \|\beta_n^0\|_1\sqrt{\frac{\log(p_n)}{n}} \right\} \right)$$

for any $0 < \delta'(< 5/44)$ (and the upper bounds for δ and δ' are not of special inter-
est). Such a consistency result for prediction also holds for the Lasso, as described
in formula (2.7) in Chapter 2 or with more details given in Corollary 6.1 in Section
6.2.2. For consistency of the Lasso, we also need $\|\beta_n\|_1 = o(\sqrt{n/\log(p_n)}) \ (n \to \infty)$
(see Corollary 6.1), as in the second part of assumption (A2). Furthermore, both
of the consistency results for L_2Boosting with componentwise linear least squares
and for the Lasso hold for arbitrary designs without any compatibility or restricted
eigenvalue conditions. In terms of rate of convergence, without assuming compati-
bility conditions, the Lasso achieves

$$\|\mathbf{X}(\hat{\beta}_{\mathrm{Lasso}}(\lambda) - \beta_n^0)\|_2^2/n = O_P\left(\|\beta_n^0\|_1\sqrt{\frac{\log(p_n)}{n}} \right),$$

when choosing $\lambda \asymp \sqrt{\log(p_n)/n}$, as described in Corollary 6.1. This rate is also
achieved with L_2Boosting if $\|\beta_n^0\|_1$ is sufficiently large, i.e., for

$$\sqrt{n/\log(p_n)}^{\frac{3-\delta}{2(2-\delta)}} \ll \|\beta_n^0\|_1 \ll \sqrt{n/\log(p_n)},$$

(the left-hand side is small for $\delta > 0$ close to zero) while for small $\|\beta_n^0\|_1 = O(1)$,
we achieve a convergence rate $O(\sqrt{\log(p_n)/n}^{\frac{1}{4}-\delta'})$ for any $0 < \delta'(< 5/44)$ when
choosing m_n as in Theorem 12.2 with $\delta = 8\delta'/(4\delta'+1)$.

12.7 Forward selection and orthogonal matching pursuit

For high-dimensional generalized linear models, it is straightforward to pursue forward selection. At first sight, this resembles the FGD boosting Algorithm 7 in Section 12.3.1. Consider the generalized linear model as in formula (3.1) in Chapter 3:

$$Y_1, \ldots, Y_n \text{ independent}$$

$$g(\mathbb{E}[Y_i|X_i = x]) = \sum_{j=1}^{p} \beta_j x^{(j)},$$

where $g(\cdot)$ is a real-valued, known link function and for simplicity, we absorb an intercept term into the right hand side of the equation above. Associated with this model, we have a loss function $\rho(\cdot, \cdot)$ and an empirical risk

$$\rho(\beta) = n^{-1} \sum_{i=1}^{n} \rho_\beta(X_i, Y_i),$$

as described in Section 3.2.1; the notation for the empirical risk is just an abbreviation indicating the dependence on the loss ρ and the parameter β. Forward variable selection proceeds as follows. In every iteration $m = 1, 2, \ldots$, we have a previous active set of variables

$$S^{[m-1]} \subseteq \{1, \ldots, p\}.$$

We are then looking for a single additional variable reducing the empirical risk most when refitting all the previous coefficients. This can be formalized as follows. For a subset $S \subseteq \{1, \ldots, p\}$, $\beta_S \in \mathbb{R}^p$ is defined as

$$\beta_{j,S} = \begin{cases} \beta_j, & j \in S, \\ 0, & j \notin S. \end{cases}$$

We estimate the coefficients corresponding to S by

$$\hat{\beta}_S = \arg\min_{\beta_S} \rho(\beta_S), \tag{12.21}$$

where the minimization is done only over the components corresponding to S. Forward variable selection then searches in every iteration m for the best single variable with index $\hat{j}_m$, in conjunction with the previous active set $S^{[m-1]}$, for reducing the empirical risk most:

$$\hat{j}_m = \arg\min_{j \in \{1, \ldots, p\} \setminus S^{[m-1]}} \rho(\hat{\beta}_{S^{[m-1]} \cup \{j\}}) \tag{12.22}$$

and the new active set is then

$$S^{[m]} = S^{[m-1]} \cup \{\hat{j}_m\}.$$

The algorithm is summarized in Algorithm 10. The main difference to boosting and

---

Algorithm 10 Forward variable selection

1: Initialize the active set of variables $S^{[0]} = \emptyset$.
2: **repeat**
3:     Increase m by one: $m \leftarrow m + 1$.
4:     Search for the best variable reducing the empirical risk most:

$$\hat{j}_m \text{ as in } (12.22).$$

5:     Update $S^{[m]} = S^{[m-1]} \cup \{\hat{j}_m\}$ and the corresponding estimator is denoted by

$$\hat{f}_{\text{FWD}}^{[m]} = x\hat{\beta}_{S^{[m]}}$$

       as defined in (12.21).
6: **until** m reaches a stopping value m_{stop}.

---

the FGD algorithm in Section 12.3.1 is that all coefficients in the active set $S^{[m]}$ are re-fitted. From an approximation point of view without noise, this is desirable as the numerical convergence to a minimum is (typically) much faster. For example, when considering the squared error loss and in absence of noise, $\hat{f}_{\text{FWD}}^{[m]}(x) = x\hat{\beta}_{\text{FWD}}^{[m]}$ converges to the true underlying $f^0(x)$ at rate $m^{-1/2}$ whereas with L_2Boosting using componentwise linear least squares, the corresponding convergence rate is $m^{-1/6}$ (Temlyakov, 2000). However, when having substantial noise with low signal to noise ratio, the more slowly proceeding boosting algorithms have been empirically found to perform substantially better, see also Table 2.1 for the example in Section 2.4.1. So far, this empirical phenomenon is not well understood from a theoretical point of view.

12.7.1 Linear models and squared error loss

For linear models with squared error loss and empirical risk

$$\rho(\beta) = n^{-1} \sum_{i=1}^{n} (Y_i - X_i\beta)^2,$$

formula (12.22) is rather explicit and can be computed recursively based on the fitted model in the previous iteration (see also Problem 12.7). A closely related version is described next.

12.7.1.1 Orthogonal matching pursuit

We have seen that L_2Boosting with componentwise linear least squares, also called matching pursuit, chooses the variable with index $\hat{j}$ satisfying

$$\hat{j} = \arg\max_{j} \left(\frac{\sum_{i=1}^{n} U_i X_i^{(j)}}{\sum_{i=1}^{n} (X_i^{(j)})^2} \right), \tag{12.23}$$

where U_i is the current ith residuum $U_i = Y_i - \hat{f}^{[m-1]}(X_i)$, see (12.11). The idea of orthogonal matching pursuit, summarized in Algorithm 11, is to use the same selector $\hat{j}$ which reduces residual sum of squares most: but then, we re-estimate all coefficients using least squares and hence the residuals are $U_i = Y_i - \hat{f}_{\mathrm{OMP}}^{[m-1]}(X_i)$ where $\hat{f}_{\mathrm{OMP}}^{[m-1]}(\cdot)$ is based on least squares estimation. The difference to the forward

---

Algorithm 11 Orthogonal matching pursuit

1: Initialize the active set of variables $S^{[0]} = \emptyset$.
2: **repeat**
3:     Increase m by one: $m \leftarrow m+1$.
4:     Search for the best variable

$$\hat{j}_m \text{ as in (12.23).}$$

5:     Update $S^{[m]} = S^{[m-1]} \cup \{\hat{j}_m\}$.
6:     Estimate the coefficients by least squares

$$\hat{\beta}_{S^{[m]}} = \arg\min_{\beta_{S^{[m]}}} \|\mathbf{Y} - \mathbf{X}_{S^{[m]}} \beta_{S^{[m]}}\|^2 / n,$$

where $\mathbf{X}_S$ is the $n \times |S|$ sub-design matrix corresponding to the variables from $S \subseteq \{1,\ldots,p\}$. Denote the estimate by

$$\hat{\beta}_{\mathrm{OMP}}^{[m]} = \hat{\beta}_{S^{[m]}}.$$

7: **until** m reaches a stopping value m_{stop}.

---

selection Algorithm 10 is that the selector $\hat{j}_m$ may not be exactly the same: in the orthogonal matching pursuit Algorithm 11 we select the variable before refitting all others from the previous active set $S^{[m]}$ while the forward selection Algorithm 10 selects the variable which is best when having refitted all other coefficients from $S^{[m]}$. The difference between the methods is usually small and orthogonal matching pursuit is computationally faster. See also Problem 12.7. Comparing the orthogonal matching pursuit Algorithm 11 with L_2Boosting (i.e. matching pursuit) we see that the latter is not re-estimating the coefficients from the previous active set. This is a major difference, and the connection to the Lasso, see Section 12.6.2.1, only holds when dropping the refitting of the coefficients. Thus, what seems a-priori naive,

i.e., not re-estimating coefficients, turns out to have an interesting connection to ℓ_1-penalized methods.

From a theoretical perspective, orthogonal matching pursuit (Algorithm 11) has desirable properties (Tropp, 2004; Tropp and Gilbert, 2007) which are comparable to ℓ_1-penalization. We develop here a consistency result for prediction in a high-dimensional sparse linear model. To capture the notion of high-dimensionality, we consider the triangular array model as in (12.20):

$$Y_{n;i} = \sum_{j=1}^{p_n} \beta_{n;j}^0 X_{n;i}^{(j)} + \varepsilon_{n;i}, \ i = 1,\ldots,n; \ n = 1,2,\ldots \tag{12.24}$$

and we use the short hand notation $\mathbf{Y} = \mathbf{X}\beta_n^0 + \varepsilon$.

We make exactly the same assumptions as for for prediction consistency of L_2Boosting in a high-dimensional linear model in Section 12.6.2.

(B1)   The number of covariates p_n in model (12.24) satisfies

$$\log(p_n)/n \to 0 \ (n \to \infty)$$

(B2)   The covariates are deterministic (fixed design) and scaled

$$n^{-1} \sum_{i=1}^{n} (X_{n;i}^{(j)})^2 \equiv 1 \text{ for all } j = 1,\ldots,p,$$

and the model is sparse with respect to the ℓ_1-norm

$$\|\beta_n^0\|_1 = \sum_{j=1}^{p} |\beta_{n;j}^0| = o\left(\sqrt{\frac{n}{\log(p_n)}}\right).$$

and the regression function satisfies

$$\|\mathbf{X}\beta_n^0\|_2^2/n = n^{-1}\sum_{i=1}^{n} \left(f^0(X_{n;i})\right)^2 \leq C < \infty \text{ for all } n \in \mathbb{N},$$

where $f^0(x) = x\beta^0$.

(B3)   The errors satisfy

$$\varepsilon_{n;1},\ldots \varepsilon_{n;n} \text{ i.i.d. } \sim \mathcal{N}(0,\sigma^2)$$

with $0 < \sigma^2 < \infty$.

Theorem 12.3. *Consider the model (12.24) satisfying (B1)-(B3). Then the orthogonal matching pursuit estimate $\hat{\beta}_{\text{OMP};n}^{[m]}$ satisfies: for $m_n \to \infty$, $m_n = O(\log(p_n))$ $(n \to \infty)$,*

$$\|\mathbf{X}(\hat{\beta}_{\text{OMP};n}^{[m_n]} - \beta_n^0)\|_2^2/n = o_P(1) \ (n \to \infty).$$

Alternatively, the same result holds for $m_n \to \infty$, $m_n = o(n(\max\{1, \|\beta_n^0\|_1^2\})^{-1})$.

Furthermore, assuming that $\|\beta_n^0\|_1 \geq B > 0$ for all $n \in \mathbb{N}$, and choosing $m_n = K\|\beta_n^0\|_1^{-2/3}n^{1/3}$ with any constant $0 < K < \infty$, we have,

$$\|\mathbf{X}(\hat{\beta}_{\text{OMP};n}^{[m_n]} - \beta_n^0)\|_2^2/n = O_P\left(\max\left\{\|\beta_n^0\|^{2/3}n^{-1/3}, \|\beta_n^0\|_1\sqrt{\frac{\log(p_n)}{n}}\right\}\right) \ (n \to \infty).$$

A proof is given in Section 12.8.3. The theorem is about prediction and as such, it should be compared to analogous results for L_2Boosting in Theorem 12.2 and to the Lasso as described in formula (2.7) from Chapter 2 or in more details in Corollary 6.1 in Section 6.2.2. All of the methods are consistent when assuming the same condition on $\|\beta_n^0\|_1 = o(\sqrt{n/\log(p_n)})$ and making no assumptions on the design.

For the Lasso without additional assumptions on the design, we have the convergence rate

$$\|\mathbf{X}(\hat{\beta}_{\text{Lasso}}(\lambda) - \beta_n^0)\|_2^2/n = O_P\left(\|\beta_n^0\|_1\sqrt{\frac{\log(p_n)}{n}}\right),$$

when choosing $\lambda \asymp \sqrt{\log(p_n)/n}$, as described in Corollary 6.1. This rate is also achieved for orthogonal matching pursuit if $\|\beta_n^0\|_1$ is large, namely in the (small) range

$$\sqrt{\frac{n}{\log(p_n)}}\log(p_n)^{-1} \ll \|\beta_n^0\|_1 \ll \sqrt{\frac{n}{\log(p_n)}}.$$

In comparison, L_2Boosting achieves the same convergence rate in a typically larger range of values for $\|\beta^0\|_1$, see end of Section 12.6.2.2. Finally, it is worth emphasizing that all the mentioned results for consistent high-dimensional prediction with the Lasso, L_2Boosting or orthogonal matching pursuit do not make any (restrictive) assumptions on the fixed design. When making additional assumptions on the design, we have shown in Chapter 6 that Lasso achieves a much faster convergence rate. We do not give an analysis of orthogonal matching pursuit under additional design conditions but we refer to Zhang (2009a) for further results.

Our theoretical viewpoints on consistent high-dimensional prediction yield little insights about the finer differences of the methods for finite sample sizes. In high noise problems, L_2Boosting or ℓ_1-penalization seem to have better empirical predictive power, see also Table 2.1 for the example in Section 2.4.1.

12.8 Proofs

12.8.1 Proof of Theorem 12.1

The proof is taken from Bühlmann and Yu (2003). Let $\mathscr{H}$ be the smoothing spline operator corresponding to smoothness r and with smoothing parameter $c = \lambda_0$ (to avoid notational confusion with eigenvalues). It is well-known (cf. Wahba (1990, p.61)) that the eigenvalues of $\mathscr{H}$ take the form in decreasing order

$$\lambda_1 = ... = \lambda_r = 1, \;\; \lambda_k = \frac{nq_{k,n}}{n\lambda_0 + nq_{k,n}} \;\; \text{for } k = r+1,...,n.$$

Moreover, for n large, $q_{k,n} \sim Ak^{-2r} := Aq_k$ where A is universal and depends on the density of the design points X_i. Let $c_0 = c/A$, then

$$\lambda_k \sim \frac{q_k}{c_0 + q_k} \;\; \text{for } k = r+1,...,n.$$

From (12.15) we know that the boosting operator in iteration m has the spectral representation $\mathscr{B}_m = UD_mU^T$ with eigenvalues $D_m = \text{diag}(d_{1,m},...,d_{n,m})$ where $d_{i,m} = 1 - (1 - \lambda_i)^m$. Denote by

$$\mu = U^T f^0.$$

For the true function $f \in \mathscr{F}^{(\nu)}$,

$$\frac{1}{n} \sum_{k=r+1}^{n} \mu_k^2 k^{2\nu} \leq M < \infty.$$

First, consider the bias term:

$$\text{bias}^2(m;f) = n^{-1} \sum_{i=1}^{n} (\mathbb{E}[\hat{f}_n^{[m]}(X_i)] - f^0(X_i))^2$$

$$= n^{-1}(U^T f^0)^T (D_m - I)^2 U^T f^0 = n^{-1} \sum_{k=r+1}^{n} (1 - \lambda_k)^{2m} \mu_k^2.$$

We can bound it as follows.

$$\text{bias}^2(m;f) = \frac{1}{n} \sum_{k=r+1}^{n} (1 - \lambda_k)^{2m} \mu_k^2$$

$$\sim \frac{1}{n} \sum_{k=r+1}^{n} (1 - q_k/(c_0 + q_k))^{2m} k^{-2\nu} \mu_k^2 k^{2\nu}$$

$$\leq \max_{k=r+1,\dots,n} (1 - q_k/(c_0 + q_k))^{2m} k^{-2\nu} \times \frac{1}{n} \sum_{k=r+1}^{n} \mu_k^2 k^{2\nu}$$

$$= \max_{k=r+1,\dots,n} \exp(h(k)) \times \frac{1}{n} \sum_{k=r+1}^{n} \mu_k^2 k^{2\nu},$$

where

$$h(x) = \log[(1 - x^{-2r}/(c_0 + x^{-2r}))^{2m} x^{-2\nu}]$$
$$= 2m \log(1 - 1/(c_0 x^{2r} + 1)) - 2\nu \log(x).$$

Taking derivative gives

$$h'(x) = \frac{2r}{x} \frac{1}{c_0 x^{2r} + 1} [2m - \frac{\nu}{r}(c_0 x^{2r} + 1)].$$

Hence for any given positive integer n_1, if $x \leq n_1$ and $m \geq \frac{\nu}{2r}(c_0 n_1^{2r} + 1)$, $h(x)$ is increasing and so is $\exp(h(x))$, and

$$\exp(h(x)) \leq \exp((h(n_1)) = (1 - 1/(c_0 n_1^{2r} + 1))^{2m} n_1^{-2\nu}.$$

On $[n_1 + 1, n]$,

$$\exp(h(x)) \leq (1 - 1/(c_0 n^{2r} + 1))^{2m} n_1^{-2\nu}.$$

Putting them together we get for growing n_1 and $m \geq \frac{\nu}{2r}(c_0 n_1^{2r} + 1)$,

$$\text{bias}^2(m; f) \leq O(M n_1^{-2\nu} [2(1 - 1/(c_0 n^{2r} + 1))^{2m+2}])$$

which is of the order $O(n_1^{-2\nu})$ for $n_1 \to \infty$ and $n_1 \leq n$.

Now let's deal with the variance term. For any $n_1 > r$,

$$\text{variance}(m; \sigma^2) = n^{-1} \sum_{i=1}^{n} \text{Var}(\hat{f}_n^{(m)}(X_i)) = n^{-1} \text{trace}(\text{Cov}(\mathscr{B}_m Y))$$

$$= \frac{\sigma^2}{n} \text{trace}(D_m^2) = \frac{\sigma^2}{n} \left(r + \sum_{k=r+1}^{n} (1 - (1 - \lambda_k)^m)^2 \right)$$

$$\leq \frac{\sigma^2 n_1}{n} + \frac{\sigma^2}{n} \sum_{k=n_1+1}^{n} (1 - (1 - \lambda_k)^m)^2 := I_1 + I_2.$$

Because $(1 - x)^a \geq 1 - ax$ for any $x \in [0, 1]$ and $a \geq 1$,

$$1 - (1 - \lambda_k)^m \leq 1 - (1 - m\lambda_k) = m\lambda_k.$$

It follows that

$$I_2 \le \frac{\sigma^2}{n} \sum_{k=n_1+1}^{n} m^2 \lambda_k^2 \sim \frac{\sigma^2 m^2}{n} \sum_{k=n_1+1}^{n} \frac{1}{(c_0 k^{2r}+2)^2}$$

$$\le \frac{\sigma^2 m^2}{n} \sum_{k=n_1+1}^{n} \frac{1}{(c_0 k^{2r})^2} \le \frac{\sigma^2 m^2}{n} \int_{n_1}^{\infty} \frac{1}{(c_0 x^{2r})^2} dx$$

$$= \frac{\sigma^2 m^2}{c_0^2 (4r-1)n} n_1 / n_1^{4r} \le O(n_1/n),$$

if we take $m = m(n_1) = \frac{v}{2r}(c_0 n_1^{2r}+1) = O(n_1^{2r})$. Hence for this choice of $m(n_1)$,

$$\text{variance}(m(n_1); \sigma^2) \le O(n_1/n).$$

Together with the bound for the bias we get

$$n^{-1} \sum_{i=1}^{n} \mathbb{E}[(\hat{f}^{[m]}(X_i) - f^0(X_i))^2] \le O(n_1/n) + O(n_1^{-2v}).$$

For minimizing this expression, we take $n_1 = O(n^{1/(2v+1)})$ and for $m(n) = m(n_1) = O(n^{2r/(2v+1)})$: the minimized MSE has the minimax optimal rate $O(n^{-2v/(2v+1)})$ of the smoother function class $\mathscr{F}^{(v)}$. $\square$

12.8.2 Proof of Theorem 12.2

We closely follow ideas from Temlyakov (2000) for the analysis of the "weak greedy algorithm" in the noise-free case. The proof here is self-contained, and simpler and more general than in Bühlmann (2006).

For vectors $u, v \in \mathbb{R}^n$, denote by

$$(u, v)_n = n^{-1} \sum_{i=1}^{n} u_i v_i$$

the (scaled) Euclidean inner product in $\mathbb{R}^n$. We also use the notation $f = (f(X_1), \dots, f(X_n))^T$ for the $n \times 1$ vector for a function $f(\cdot)$ defined on the covariate space. Furthermore, define by

$$\psi_j = (X_1^{(j)}, \dots, X_n^{(j)})^T, \ j = 1, \dots, p,$$

the vector of the jth covariable.

Without loss of generality, we assume that the constant C in assumption (A2) satisfies:

$$\|\mathbf{X}\beta^0\|_n^2 = \|f^0\|_n^2 \le 1.$$

This can be achieved by scaling for the case $1 < C < \infty$ and noting that due to $C < \infty$, the scaled version has still bounded error variances from below.

Denote the remainder term (residuals) in iteration k by

$$\hat{R}^k f^0 = f^0 - \hat{f}^{[k]}.$$

Here and in the following, we suppress the index n in $\hat{R}^k f^0$, $\hat{f}^{[k]}$ and the true regression function f^0. A straightforward calculation leads to:

$$\|\hat{R}^k f^0\|_n^2 = \|\hat{R}^{k-1} f^0\|_n^2 - |(f^0 - \hat{f}^{[k-1]}, \psi_{\hat{j}_k})_n|^2 + |(\varepsilon, \psi_{\hat{j}_k})_n|^2. \qquad (12.25)$$

We use the short-hand notation

$$a_k = \|\hat{R}^k f^0\|_n^2.$$

As described in Lemma 6.2 in Section 6.2.2 from Chapter 6:

$$\Delta_n := \max_{j=1,\dots,p} |(\varepsilon, \psi_j)_n| = O_P(\sqrt{\log(p_n)/n}). \qquad (12.26)$$

Thus, (12.25) and (12.26) lead to the bound

$$a_k \le a_{k-1} - |(\hat{R}^{k-1} f^0, \psi_{\hat{j}_k})_n|^2 + \Delta_n^2. \qquad (12.27)$$

Next, we want to relate $|(\hat{R}^{k-1} f^0, \psi_{\hat{j}_k})_n|$ and its "noisy" version $|(\mathbf{Y} - \hat{f}^{[k-1]}, \psi_{\hat{j}_k})_n|$ with $\max_j |(\hat{R}^{k-1} f^0, \psi_j)_n|$; note that the selector $\hat{j}_k$ is constructed from

$$\hat{j}_k = \arg\max_{j=1,\dots,p} |(\mathbf{Y} - \hat{f}^{[k-1]}, \psi_j)_n|.$$

Lemma 12.1. *If for some* $0 < \kappa < 1/2$

$$\max_{j=1,\dots,p} |(\hat{R}^{k-1} f^0, \psi_j)_n| \ge 2\Delta_n/\kappa,$$

then

$$|(\hat{R}^{k-1} f^0, \psi_{\hat{j}_k})_n| \ge (1 - \kappa) \max_{j=1,\dots,p} |(\hat{R}^{k-1} f^0, \psi_j)_n|,$$
$$|(\mathbf{Y} - \hat{f}^{[k-1]}, \psi_{\hat{j}_k})_n| \ge (1 - \kappa/2) \max_{j=1,\dots,p} |(\hat{R}^{k-1} f^0, \psi_j)_n|.$$

Proof. We have

$$|(\hat{R}^{k-1} f^0, \psi_{\hat{j}_k})_n| = |(\mathbf{Y} - \hat{f}^{[k-1]}, \psi_{\hat{j}_k})_n - (\varepsilon, \psi_{\hat{j}_k})_n|$$
$$\ge |(\mathbf{Y} - \hat{f}^{[k-1]}, \psi_{\hat{j}_k})_n| - \Delta_n = \max_j |(\mathbf{Y} - \hat{f}^{[k-1]}, \psi_j)_n| - \Delta_n$$
$$= \max_j |(\hat{R}^{k-1} f^0, \psi_j)_n + (\varepsilon, \psi_j)_n| - \Delta_n \ge \max_j |(\hat{R}^{k-1} f^0, \psi_j)_n| - 2\Delta_n.$$

Hence, since $-2\Delta_n \geq -\kappa \max_j |(\hat{R}^{k-1} f^0, \psi_j)_n|$ (assumption of the lemma),

$$|(\hat{R}^{k-1} f^0, \psi_{\hat{j}_k})_n| \geq (1-\kappa) \max_j |(\hat{R}^{k-1} f^0, \psi_j)_n.$$

Furthermore, analogously as above,

$$|(\mathbf{Y} - \hat{f}^{[k-1]}, \psi_{\hat{j}_k})_n| = \max_j |(\mathbf{Y} - \hat{f}^{[k-1]}, \psi_j)_n|$$

$$\geq \max_j |(\hat{R}^{k-1} f^0, \psi_j)_n| - \Delta_n \geq (1-\kappa/2) \max_j |(\hat{R}^{k-1} f^0, \psi_j)_n|.$$

<div style="text-align:right">□</div>

Denote by

$$d_k = |(\mathbf{Y} - \hat{f}^{[k-1]}, \psi_{\hat{j}_k})_n|.$$

We then have:

$$a_k = \|\hat{R}^k f^0\|_n^2 = \|\hat{R}^{k-1} f^0 - (\mathbf{Y} - \hat{f}^{[k-1]}, \psi_{\hat{j}_k})_n \psi_{\hat{j}_k}\|_n^2$$

$$= a_{k-1} + d_k^2 - 2(\mathbf{Y} - \hat{f}^{[k-1]}, \psi_{\hat{j}_k})_n (\hat{R}^{k-1} f^0, \psi_{\hat{j}_k})_n$$

$$\leq a_{k-1} - d_k^2 + 2d_k(\varepsilon, \psi_{\hat{j}_k})_n.$$

Thus, instead of (12.27), we obtain

$$a_k \leq a_{k-1} - d_k^2 + 2d_k \Delta_n. \tag{12.28}$$

We now establish a decay of a_k in terms of a fraction of d_k^2.

Lemma 12.2. *If for some $0 < \kappa < 1/2$*

$$\max_{j=1,\dots,p} |(\hat{R}^{k-1} f^0, \psi_j)_n| \geq \kappa^{-1}(1-\kappa/2)^{-1} 2\Delta_n,$$

then

$$a_k \leq a_{k-1} - (1-\kappa)d_k^2.$$

Proof. The assumption of the lemma implies:

$$2d_k \Delta_n \leq \kappa(1-\kappa/2) d_k \max_j |(\hat{R}^{k-1} f^0, \psi_j)_n| \leq \kappa(1-\kappa/2) \frac{d_k^2}{(1-\kappa/2)} = \kappa d_k^2,$$

where in the second inequality, we have used the second assertion in Lemma 12.1 (and the assumption of the lemma implies the assumption of Lemma 12.1). Together with (12.28), this completes the proof. □

We now follow Temlyakov (2000, proof of Th. 5.1). Consider the recursion

$$b_0 = \|\beta_n^0\|_1, \quad b_k = b_{k-1} + d_k. \tag{12.29}$$

Using that $f^0 = \sum_{j=1}^p \beta_j^0 \psi_j$, we can write

$$\hat{R}^{k-1} f^0 = \sum_{j=1}^p \gamma_j \psi_j, \quad \|\gamma\|_1 \le b_{k-1}. \tag{12.30}$$

Furthermore, for any vector $g = \sum_{j=1}^p c_j \psi_j$:

$$\|g\|_n^2 \le \|c\|_1 \max_{j=1,\dots,p} |(g, \psi_j)_n|. \tag{12.31}$$

The derivation is left as Problem 12.6.

Therefore, using (12.30):

$$\max_{j=1,\dots,p} |(\hat{R}^{k-1} f^0, \psi_j)_n| \ge a_{k-1}/b_{k-1}. \tag{12.32}$$

Then, under the conditions of and using the second assertion in Lemma 12.1:

$$d_k \ge (1 - \kappa/2) \max_j |(\hat{R}^{k-1} f^0, \psi_j)_n| \ge (1 - \kappa/2) a_{k-1}/b_{k-1},$$

and hence

$$d_k \ge \frac{(1 - \kappa/2) a_{k-1}}{b_{k-1}}. \tag{12.33}$$

From Lemma 12.2 and (12.33) we obtain (under the assumption of Lemma 12.2):

$$a_k \le a_{k-1} \left(1 - (1 - \kappa)(1 - \kappa/2)^2 \frac{a_{k-1}}{b_{k-1}^2} \right).$$

In the sequel, we abbreviate by

$$C_\kappa = \sqrt{1 - \kappa}(1 - \kappa/2) \quad (0 < \kappa < 1/2).$$

Since $b_{k-1} \le b_k$ we get:

$$a_k b_k^{-2} \le a_{k-1} b_{k-1}^{-2} \left(1 - C_\kappa^2 a_{k-1} b_{k-1}^{-2} \right). \tag{12.34}$$

Furthermore,

$$a_0 b_0^{-2} \le \|\beta_n^0\|_1^{-2}, \tag{12.35}$$

using w.l.o.g. the scaling $\|f^0\|_n^2 \le 1$, see above.

We now assume that the condition of Lemma 12.2 holds for $k = 1, \dots, m$: that is, we consider the event

$$B_n(m) = \cap_{k=1}^m \{ \max_{j=1,\dots,p} |(\hat{R}^{k-1}f^0, \psi_j)_n| \geq \kappa^{-1}(1-\kappa/2)^{-1}2\Delta_n \}. \quad (12.36)$$

Then, on $B_n(m)$, and using (12.34) and (12.35), we can invoke Lemma 12.3 below to obtain

$$a_m b_m^{-2} \leq \|\beta_n^0\|_1^{-2}(1+C_\kappa^2 m)^{-1}. \qquad (12.37)$$

Lemma 12.3. *(DeVore and Temlyakov (1996)) Let $\{c_m\}_{m\in\mathbb{N}_0}$ be a sequence of non-negative numbers such that*

$$c_0 \leq D, \;\; c_m \leq c_{m-1}(1 - \alpha c_{m-1}) \;\; (0 < \alpha < 1).$$

Then,

$$c_m \leq D(1+\alpha m)^{-1}.$$

We refer to DeVore and Temlyakov (1996) for a proof.

Furthermore, Lemma 12.2 and (12.33) imply on $B_n(m)$:

$$a_m \leq a_{m-1} - (1-\kappa)d_m^2 \leq a_{m-1} - (1-\kappa)(1-\kappa/2)d_m a_{m-1}/b_{m-1}$$
$$= a_{m-1}\left(1 - \frac{D_\kappa d_m}{b_{m-1}}\right), \qquad (12.38)$$

where we denote by $D_\kappa = (1-\kappa)(1-\kappa/2)$. The recursion for b_m can be written as

$$b_m = b_{m-1}(1 + d_m/b_{m-1}).$$

This, together with (12.38) and the inequality

$$(1+u)^\alpha \leq 1 + \alpha u, \;\; 0 \leq \alpha \leq 1, \; u \geq 0,$$

leads to

$$a_m b_m^{D_\kappa} \leq a_{m-1}(1 - \frac{D_\kappa d_m}{b_{m-1}})b_{m-1}^{D_\kappa}(1 + D_\kappa \frac{d_m}{b_{m-1}}) \leq a_{m-1}b_{m-1}^{D_\kappa}.$$

Thus, using that $b_0 = \|\beta_n^0\|_1$ and $a_0 \leq 1$

$$a_m b_m^{D_\kappa} \leq a_{m-1}b_{m-1}^{D_\kappa} \leq \dots \leq a_0 b_0^{D_\kappa} \leq \|\beta_n^0\|_1^{D_\kappa}.$$

Combining this with (12.37) we obtain:

$$a_m^{2+D_\kappa} = (a_m b_m^{-2})^{D_\kappa}(a_m b_m^{D_\kappa})^2$$
$$\leq \|\beta_n^0\|_1^{-2D_\kappa}(1+C_\kappa^2 m)^{-D_\kappa}\|\beta_n^0\|_1^{2D_\kappa} = (1+C_\kappa^2 m)^{-D_\kappa}.$$

Since for $0 < \kappa < 1/2$,

$$C_\kappa^{-2D\kappa} = \left((1-\kappa)(1-\kappa/2)^2\right)^{-(1-\kappa)(1-\kappa/2)} \leq 2,$$

we obtain the bound

$$a_m^{2+D\kappa} \leq 2m^{-D\kappa}.$$

Thus,

$$\text{on } B_n(m): \quad \|\hat{R}^m f^0\|_n^2 = a_m \leq 2^{\frac{1}{2+D\kappa}} m^{-\frac{D\kappa}{2+D\kappa}} \leq 2m^{-\frac{D\kappa}{2+D\kappa}}. \qquad (12.39)$$

Now we analyze the behavior of a_m on

$$B_n(m)^c = \{ \text{ there exists } k^* \ (1 \leq k^* \leq m) \text{ such that}$$
$$\max_{j=1,\ldots,p} |(\hat{R}^{k^*-1} f^0, \psi_j)_n| < \kappa^{-1}(1-\kappa/2)^{-1} 2\Delta_n \}.$$

First, due to (12.25):

$$\|\hat{R}^m f^0\|_n^2 \leq \|\hat{R}^{m-1} f^0\|_n^2 + \Delta_n \leq \ldots \leq \|\hat{R}^{k^*-1} f^0\|_n^2 + (m-k^*)\Delta_n.$$

Furthermore, using (12.32) and the definition of the sequence b_k $(k = 0, 1, \ldots)$ in (12.29),

$$\|\hat{R}^{k^*-1} f^0\|_n^2 \leq \max_j |(\hat{R}^{k^*-1} f^0, \psi_j)_n| b_{k^*-1}$$

$$\leq \max_j |(\hat{R}^{k^*-1} f^0, \psi_j)_n| (\|\beta_n^0\|_1 + \sum_{r=0}^{k^*-1} \|\mathbf{Y} - \hat{f}^{[r]}\|_n)$$

$$\leq \max_j |(\hat{R}^{k^*-1} f^0, \psi_j)_n| (\|\beta_n^0\|_1 + k^*\|\mathbf{Y}\|_n).$$

Using the third assumption in (A2) (and using w.l.o.g. that $C = 1$, as stated at the beginning of the proof), we obtain

$$\|Y\|_n^2 \leq (1 + \sigma^2) + v_n =: \gamma_n = O_P(1),$$

where $v_n = o_P(1)$ and hence $\gamma_n = O_P(1)$. Moreover, using $k^* \leq m$, the bound above becomes:

$$\text{on } B_n(m)^c :$$
$$\|\hat{R}^m f^0\|_n^2 \leq \max_j |(\hat{R}^{k^*-1} f^0, \psi_j)_n| (\|\beta_n^0\|_1 + m\gamma_n) + m\Delta_n$$
$$\leq \kappa^{-1}(1-\kappa/2)^{-1} 2\Delta_n (\|\beta_n^0\|_1 + m\gamma_n) + m\Delta_n.$$

Hence, together with (12.39):

$$\|\hat{R}^m f^0\|_n^2 \le \max\{2m^{-\frac{D_\kappa}{2+D_\kappa}}, \kappa^{-1}(1-\kappa/2)^{-1}2\Delta_n(\|\beta_n^0\|_1 + m\gamma_n)\} + m\Delta_n.$$

We note that $0 < \kappa < 1/2$ and hence D_κ is a fixed number and recall $\Delta_n = O_P(\sqrt{\log(p_n)/n})$ and $\gamma_n = O_P(1)$. Thus, for $m_n \to \infty$, $m_n = o(\sqrt{n/\log(p_n)})$ and using that $\|\beta_n^0\|_1 = o(\sqrt{n/\log(p_n)})$, it follows that $\|\hat{R}^m f^0\|_n^2 = o_P(1)$.

Regarding the convergence rate, we optimize m and choose

$$m \asymp \sqrt{n/\log(p_n)}^{\frac{2+D_\kappa}{2(1+D_\kappa)}} \quad (0 < \kappa < 1/2).$$

Setting $\delta = 1 - D_\kappa$ resulting in $0 < \delta < 5/8$, we hence choose

$$m \asymp \sqrt{n/\log(p_n)}^{\frac{3-\delta}{4-2\delta}} \quad (0 < \delta < 5/8)$$

and we obtain the claimed convergence rate. $\square$

12.8.3 Proof of Theorem 12.3

We use a similar notation as in the proof of Theorem 12.2. For vectors $u, v \in \mathbb{R}^n$, denote by

$$(u, v)_n = n^{-1}\sum_{i=1}^n u_i v_i$$

the (scaled) Euclidean inner product in $\mathbb{R}^n$. Furthermore, we denote by $f = (f(X_1), \dots, f(X_n))^T$ the $n \times 1$ vector for a function $f(\cdot)$ defined on the covariate space, and by $\psi_j = (X_1^{(j)}, \dots, X_n^{(j)})^T$ $(j = 1, \dots, p)$ the vector of the jth covariable.

For $k = 1, \dots, m$, denote the remainder term (residuals) in iteration k by

$$\hat{R}^k f^0 = f^0 - \hat{f}_{\text{OMP}}^{[k]},$$

where $\hat{f}_{\text{OMP}}^{[k]} = \mathbf{X}\hat{\beta}_{\text{OMP}}^{[k]}$. Furthermore, we consider a noise-free version which uses the same (estimated) selected variables with indices $\hat{j}_k$, $k = 1, 2, \dots$. Denote by $S^{[k]}$ the active set of variables after k steps, as described in Algorithm 11, let $P_{S^{[k]}}$ be the projection from $\mathbb{R}^n$ onto the linear span of $S^{[k]}$, and denote by

$$f_{\text{OMP}}^{[k]} = P_{S^{[k]}} f^0,$$
$$R^k f^0 = f^0 - f_{\text{OMP}}^{[k]} \tag{12.40}$$

the projection of the true underlying function (vector) f^0 and the corresponding remainder term. We note that $\hat{f}_{\text{OMP}}^{[k]} = P_{S^{[k]}} \mathbf{Y}$. Finally, we denote by

$$\hat{R}^0 f^0 = R^0 f^0 \equiv 0.$$

Here and in the following, we suppress the index n in expressions like $\hat{R}^k f^0$, $\hat{f}_{\text{OMP}}^{[k]}$ and the true regression function f^0.

Without loss of generality, we assume that the constant C in assumption (B2) satisfies:

$$\|\mathbf{X}\beta^0\|_n^2 = \|f^0\|_n^2 \le 1.$$

This can be achieved by scaling for the case $1 < C < \infty$ and noting that due to $C < \infty$, the scaled version has still bounded error variances from below.

First, we show that the noise-free version $f_{\text{OMP}}^{[k]}$, with the estimated selected variables $\hat{j}_1, \ldots, \hat{j}_k$ is converging to zero as $k \to \infty$. To do so, we prove that $|(R^{k-1} f^0, \psi_{\hat{j}_k})_n|$ and $\max_j |(R^{k-1} f^0, \psi_j)_n|$ are close.

The stochastic part of the analysis boils down to the upper bounds of

$$\Delta_n := \max_{j=1,\ldots,p} |(\varepsilon, \psi_j)_n| = O_P(\sqrt{\log(p_n)/n}),$$

see (12.26), and the following.

Lemma 12.4. *For* $m = m_n \to \infty$ $(n \to \infty)$,

$$\Gamma_n(m) := \max_{k=1,\ldots,m_n} \max_{j=1,\ldots,p} |(\hat{R}^{k-1} f^0, \psi_j)_n - (R^{k-1} f^0, \psi_j)_n| \le O_P(\sqrt{m_n/n}).$$

Proof. Using the Cauchy-Schwarz inequality,

$$\max_{j=1,\ldots,p} |(\hat{R}^{k-1} f^0, \psi_j)_n - (R^{k-1} f^0, \psi_j)_n| \le \|\hat{R}^{k-1} f^0 - R^{k-1} f^0\|_n. \quad (12.41)$$

The (square of the) right-hand side can be written as

$$\|\hat{R}^{k-1} f^0 - R^{k-1} f^0\|_n^2 = \|P_{S^{[k-1]}} (\mathbf{Y} - f^0)\|_n^2 = \|P_{S^{[k-1]}} \varepsilon\|_n^2$$

and clearly,

$$\|P_{S^{[k-1]}} \varepsilon\|_n^2 \le \|P_{S^{[k'-1]}} \varepsilon\|_n^2 \text{ for } k' \ge k.$$

Thus,

$$\mathbb{E}\big[\max_{k=1,\ldots,m} \|\hat{R}^{k-1} f^0 - R^{k-1} f^0\|_n^2\big] \le \mathbb{E}[\|P_{S^{[m-1]}} \varepsilon\|_n^2] \le \sigma^2(m-1)/n. \quad (12.42)$$

Hence, together with (12.41), the claim follows. $\qquad\square$

Lemma 12.5. *Let $k \in \mathbb{N}$. If*

$$\max_{j=1,\dots,p} |(R^{k-1}f^0, \psi_j)_n| > 2(\Gamma_n(m) + \Delta_n)/\kappa,$$

then

$$|(R^{k-1}f^0, \psi_{\hat{j}_k})_n| \geq (1 - \kappa) \max_j |(R^{k-1}f^0, \psi_j)_n|.$$

Proof. Consider first the bound

$$\max_j |(\mathbf{Y} - \hat{f}_{\mathrm{OMP}}^{[k-1]}, \psi_j)_n - (f^0 - \hat{f}_{\mathrm{OMP}}^{[k-1]}, \psi_j)_n| \leq \max_j (\varepsilon, \psi_j)_n| = \Delta_n.$$

This, together with Lemma 12.4, implies:

$$\max_j |(\mathbf{Y} - \hat{f}_{\mathrm{OMP}}^{[k-1]}, \psi_j)_n - (R^{[k-1]}f^0, \psi_j)_n| \leq \Delta_n + \Gamma_n(m).$$

Thus,

$$
\begin{aligned}
|(R^{k-1}f^0, \psi_{\hat{j}_k})_n| &\geq |(\mathbf{Y} - \hat{f}_{\mathrm{OMP}}^{[k-1]}, \psi_{\hat{j}_k})_n| - \Gamma_n(m) - \Delta_n \\
&= \max_j |(\mathbf{Y} - \hat{f}_{\mathrm{OMP}}^{[k-1]}, \psi_j)_n| - \Gamma_n(m) - \Delta_n \\
&\geq \max_j |(R^{k-1}f^0, \psi_j)_n| - 2\Gamma_n(m) - 2\Delta_n \\
&\geq (1 - \kappa) \max_j |(R^{k-1}f^0, \psi_j)_n|,
\end{aligned}
$$

where for the last inequality, we have used the assumption of the lemma. $\square$

Our proof is now following ideas from Temlyakov (2000, Th. 3). We assume that the condition of Lemma 12.5 holds for $k = 1, \dots, m$: that is, we consider the event

$$B_n(m) = \cap_{k=1}^m \{ \max_{j=1,\dots,p} |(R^{k-1}f^0, \psi_j)_n| \geq \kappa^{-1} 2(\Gamma_n(m) + \Delta_n) \}. \qquad (12.43)$$

Then, for $k = 1, \dots, m$ and using Lemma 12.5:

$$
\begin{aligned}
\text{on } B_n(m): \quad \|R^k f^0\|_n^2 &\leq \|R^{k-1}f^0 - (R^{k-1}f^0, \psi_{\hat{j}_k})_n \psi_{\hat{j}_k}\|_n^2 \\
&= \|R^{k-1}f^0\|_n^2 - |(R^{k-1}f^0, \psi_{\hat{j}_k})_n|^2 \\
&\leq \|R^{k-1}f^0\|_n^2 - (1 - \kappa)^2 \max_j |(R^{k-1}f^0, \psi_j)_n|^2. \qquad (12.44)
\end{aligned}
$$

We now use the following inequality (Problem 12.8): for $n \times 1$ vectors $g = \sum_j \gamma_j \psi_j$ and any h

$$|(h, g)_n| \leq \|\gamma\|_1 \max_j |(h, \psi_j)_n|. \qquad (12.45)$$

Hence, using $g = f^0$, $h = R^{k-1} f^0$:

$$\max_j |(R^{k-1} f^0, \psi_j)_n|^2 \geq \|\beta_n^0\|_1^{-2} |(R^{k-1} f^0, f^0)_n|^2 = \|\beta_n^0\|_1^{-2} \|R^{k-1} f^0\|_n^4.$$

(12.46)

Thus, (12.44) becomes:

$$\text{on } B_n(m): \quad \|R^k f^0\|_n^2 \leq \|R^{k-1} f^0\|_n^2 \left(1 - (1-\kappa)^2 \|\beta_n^0\|_1^{-2} \|R^{k-1} f^0\|_n^2 \right),$$

and hence also

$$\|\beta_n^0\|_1^{-2} \|R^k f^0\|_n^2 \leq \|\beta_n^0\|_1^{-2} \|R^{k-1} f^0\|_n^2 \left(1 - (1-\kappa)^2 \|\beta_n^0\|_1^{-2} \|R^{k-1} f^0\|_n^2 \right).$$

Furthermore,

$$\|\beta_n^0\|_1^{-2} \|R^0 f^0\|_n^2 = \|\beta_n^0\|_1^{-2} \|f^0\|_n^2 \leq \|\beta_n^0\|_1^{-2}$$

using w.l.o.g. the scaling $\|f^0\|_n^2 \leq 1$, see above. Thus, using Lemma 12.3, applied to $c_m = \|\beta_n^0\|_1^{-2} \|R^m f^0\|_n^2$, we obtain:

$$\text{on } B_n(m): \quad \|\beta_n^0\|_1^{-2} \|R^m f^0\|_n^2 \leq \|\beta_n^0\|_1^{-2} (1 + (1-\kappa)^2 m)^{-1},$$

and hence

$$\text{on } B_n(m): \quad \|R^m f^0\|_n^2 \leq (1 + (1-\kappa)^2 m)^{-1} \leq (1-\kappa)^{-2} m^{-1}. \quad (12.47)$$

On the complementary event $B_n(m)^c$, we argue as follows. Write

$$B_n(m)^c = \{ \text{ there exists } k^* \ (1 \leq k^* \leq m) \text{ such that}$$
$$\max_{j=1,\dots,p} |(R^{k^*-1} f^0, \psi_j)_n| < \kappa^{-1} 2(\Gamma_n(m) + \Delta_n)\}.$$

Using the norm-reducing property,

$$\|R^m f^0\|_n^2 \leq \|R^{k^*-1} f^0\|_n^2,$$

and using (12.46),

$$\|R^{k^*-1} f^0\|_n^2 \leq \|\beta_n^0\|_1 \max_j |(R^{k^*-1} f^0, \psi_j)_n| \leq \|\beta_n^0\|_1 \kappa^{-1} 2(\Gamma_n(m) + \Delta_n),$$

we have:

$$\text{on } B_n(m)^c: \quad \|R^m f^0\|_n^2 \leq \|\beta_n^0\|_1 \kappa^{-1} 2(\Gamma_n(m) + \Delta_n).$$

Together with (12.47), we obtain:

$$\|R^m f^0\|_n^2 \leq \max\{(1-\kappa)^{-2} m^{-1}, \|\beta_n^0\|_1 \kappa^{-1} 2(\Gamma_n(m) + \Delta_n)\}. \quad (12.48)$$

From (12.42) we obtain

$$\Xi_n(m) = \|\hat{R}^m f^0 - R^m f^0\|_n^2 = O_P(m/n),$$

and hence, together with (12.48):

$$\|\hat{R}^m f^0\|_n^2 \le 4\max\{(1-\kappa)^{-2}m^{-1}, \|\beta_n^0\|_1\kappa^{-1}2(\Gamma_n(m)+\Delta_n), \Xi_n(m)\}.$$

Recall that $\Delta_n = O_P(\sqrt{\log(p_n)/n})$, $\Gamma_n(m) = O_P(\sqrt{m/n})$ and $\Xi_n(m) = O_P(m/n)$. Choosing $m = m_n \to \infty$ $m_n = O(\log(p_n))$ together with the assumption $\|\beta_n^0\|_1 = o(\sqrt{n/\log(p_n)})$ leads to $\|\hat{R}^{m_n} f^0\| = o_P(1)$. Obviously, the choice $m_n \to \infty$, $m_n = o(n(\max\{1, \|\beta_n^0\|_1^2\})^{-1})$ also leads to $\hat{R}^{m_n} f^0 = o_P(1)$.

Regarding the convergence rate, choosing

$$m_n \asymp \|\beta_n^0\|_1^{-2/3} n^{1/3}$$

and assuming that $\|\beta_n^0\|_1 \ge B > 0$ for all n, the term $\Xi_n(m)$ is asymptotically of lower order and we obtain the claimed convergence rate. $\square$

Problems

12.1. Gradient descent in function spaces
Prove formula (12.2).

12.2. Population minimizers of loss functions
For the "logit"- and exponential loss, derive formulae (12.5) and (12.8). See also Friedman et al. (2000).

12.3. Prove formula (12.11) and show that in case of centered predictor variables, the selected variable with index $\hat{j}$ maximizes the absolute correlation of $X^{(\hat{j})}$ with the residual vector.

12.4. Consider the L_2Boosting hat matrix for linear base procedures. Derive formula (12.15).

12.5. Twicing
Consider twicing with the Nadaraya-Watson kernel estimator. Derive formula (12.18) which indicates a relation to higher-order kernel estimators.

12.6. Consistency of L_2Boosting for high-dimensional linear models: for upper-bounding the squared ℓ_2-norm, prove formula (12.31).

12.7. Consider a linear model with orthonormal design $n^{-1}\mathbf{X}^T\mathbf{X} = I$.
(a) Show that forward variable selection corresponds to hard-thresholding described in Section 2.3 in Chapter 2.
(b) Show that orthogonal matching pursuit also corresponds to hard-thresholding.

12.8. Consistency of orthogonal matching pursuit for high-dimensional linear models: for upper-bounding the absolute value of the inner product, prove formula (12.45).

Chapter 13
Graphical modeling

Abstract Graphical models are very useful to describe conditional independences among a set of random variables. We focus on undirected graphs only and their interpretation as conditional independence graphs. For undirected Gaussian graphical models, ℓ_1-regularization methods can be used in a similar fashion as for linear models. Closely linked to the estimation of the underlying undirected graph is the problem of covariance estimation which we briefly discuss as well. Besides ℓ_1-penalization, we describe a completely different approach using the framework of so-called faithful distributions which also has implications on variable selection for regression. The chapter contains methodology, algorithms and mathematical theory.

13.1 Organization of the chapter

After an introduction with the basic definitions, we largely focus on undirected Gaussian graphical models in Section 13.4. Estimation with ℓ_1-penalization is either based on the joint Gaussian likelihood, as described in Section 13.4.1 where the GLasso (Graphical Lasso) is defined, or we rely on the regression formulation discussed in Section 13.4.2. Covariance estimation based on undirected graphs is briefly outlined in Section 13.4.3 and the Ising model for binary variables is introduced in Section 13.5. All these sections have methodological character only. The different approach for undirected graphical modeling and variable selection in linear models, based on so-called faithful distributions, is described in Sections 13.6 - 13.9. For this part, we discuss methodology, computational algorithms and mathematical theory.

13.2 Preliminaries about graphical models

A graph G consists of a set of vertices V and a set of edges E. The set of edges E is a subset of $V \times V$ consisting of ordered pairs of distinct vertices. An edge is undirected if $(j,k) \in E$ and $(k,j) \in E$ whereas an edge is directed from vertex j to vertex k if $(j,k) \in E$ and $(k,j) \notin E$. The neighbors or adjacency set of a node j in the undirected graph G is denoted by $\mathrm{adj}(G,j) = \{k \in V; \ (j,k) \in E \text{ and } (k,j) \in E\}$.

In a graphical model, the vertices of a graph, i.e., the set V, correspond to a collection of random variables

$$X = (X^{(1)}, \ldots, X^{(p)}) \sim P, \tag{13.1}$$

where, throughout the whole chapter, we index the set $V = \{1, \ldots, p\}$ with $|V| = p$, and P is the probability distribution of X. The pair (G,P) is referred to as a graphical model. Among the variety of graphical models, we will describe only a few concepts. In particular, we will only consider models with undirected graphs. We refer to Lauritzen (1996) or Edwards (2000) for a detailed and broad description.

13.3 Undirected graphical models

In an undirected graph, all edges are undirected. Assuming a Markov property of the distribution P with respect to the graph G, we can infer some (but maybe not all) conditional independences among the random variables $X^{(1)}, \ldots, X^{(p)}$.

13.3.1 Markov properties for undirected graphs

Consider a graphical model (G,P).

Definition 13.1. *We say that P satisfies the* pairwise Markov property *with respect to the undirected graph G if for any pair of unconnected vertices $(j,k) \notin E$ ($j \neq k$),*

$$X^{(j)} \perp X^{(k)} | X^{(V \setminus \{j,k\})}.$$

Here and in the sequel $X^{(A)} \perp X^{(B)} | X^{(C)}$ denotes that $X^{(A)} = \{X^{(j)}; \ j \in A\}$ is (mutually) conditionally independent of $X^{(B)} = \{X^{(j)}; \ j \in B\}$ given $X^{(C)} = \{X^{(j)}; \ j \in C\}$ where $A, B, C \subseteq \{1, \ldots, p\}$.

A stronger notion is the global Markov property. For its definition, we introduce the following. A path is a sequence of vertices $\{j_1, \ldots, j_\ell\}$ such that $(j_i, j_{i+1}) \in E$ for

$i = 1, \ldots, \ell - 1$. Consider a triple of disjoint sets $A, B, C \subseteq V$. We say that C separates A and B if every path from $j \in A$ to $k \in B$ contains a vertex in C.

Definition 13.2. *We say that P satisfies the* global Markov property *with respect to the undirected graph G if for any triple of disjoint sets A, B, C such that C separates A and B,*

$$X^{(A)} \perp X^{(B)} | X^{(C)}.$$

In general, the global Markov property implies the local Markov property. The converse holds for a large class of models, as described next.

Proposition 13.1. *If the distribution P has a positive and continuous density with respect to Lebesgue measure, the global and local Markov properties are equivalent.*

The statement of Proposition 13.1 is proved in Lauritzen (1996, p.35). Within the class of undirected graphical models, the following is the most popular.

Definition 13.3. *A* conditional independence graph (CIG) *is a graphical model (G, P), with undirected graph G, where the pairwise Markov property holds.*

Thus, a CIG has the property: if $(j, k) \notin E$ $(j \neq k)$, then $X^{(j)} \perp X^{(k)} | X^{(V \setminus \{j, k\})}$. For special cases, e.g. when P is multivariate Gaussian, the converse relation is true as well, i.e., if $X^{(j)} \perp X^{(k)} | X^{(V \setminus \{j, k\})}$, then $(j, k) \notin E$.

13.4 Gaussian graphical models

We specify (13.1) to the assumption

$$X = (X^{(1)}, \ldots, X^{(p)}) \sim \mathcal{N}_p(0, \Sigma) \tag{13.2}$$

with positive definite $p \times p$ covariance matrix Σ. The mean zero assumption is mainly for simplifying the notation.

A Gaussian graphical model (GGM) is a conditional independence graph with a multivariate Gaussian distribution as in (13.2). The required pairwise Markov property in the CIG, see Definition 13.3, is equivalent to the global Markov property due to the Gaussian assumption, see Proposition 13.1.

The edges in a GGM are given by the inverse of the covariance matrix:

$$(j, k) \text{ and } (k, j) \in E \Longleftrightarrow X^{(j)} \perp X^{(k} | X^{(V \setminus \{j, k\})} \Longleftrightarrow \Sigma^{-1}_{j,k} \neq 0. \tag{13.3}$$

This is a well-known result, see for example Lauritzen (1996). Thus, for a GGM, we have an "if and only if" interpretation of an edge which is in general stronger than in a CIG (see Definition 13.3).

Furthermore, the inverse of the covariance matrix corresponds to partial correlations. Denote by $K = \Sigma^{-1}$ and its scaled version by C where

$$C_{j,k} = \frac{K_{j,k}}{\sqrt{K_{j,j}K_{k,k}}}.$$

Then, the partial correlation between $X^{(j)}$ and $X^{(k)}$ given $X^{(V \setminus \{j,k\})}$ equals

$$\rho_{jk|V \setminus \{j,k\}} = -C_{j,k}.$$

This is also a well-known result which can be found in Lauritzen (1996). In particular, we have the following relation for a GGM:

$$(j,k) \text{ and } (k,j) \in E \Longleftrightarrow \Sigma^{-1}_{j,k} \neq 0 \Longleftrightarrow \rho_{jk|V \setminus \{j,k\}} \neq 0.$$

Finally, partial correlations are directly related to regression coefficients. Consider a regression

$$X^{(j)} = \beta_k^{(j)} X^{(k)} + \sum_{r \in V \setminus \{j,k\}} \beta_r^{(j)} X^{(r)} + \varepsilon^{(j)}, \tag{13.4}$$

where $\mathbb{E}[\varepsilon^{(j)}] = 0$ and due to the Gaussian assumption, $\varepsilon^{(j)}$ is independent from $\{X^{(r)}; \, r \in V \setminus \{j\}\}$ (Problem 13.1). Then,

$$\beta_k^{(j)} = -K_{j,k}/K_{j,j}, \quad \beta_j^{(k)} = -K_{j,k}/K_{k,k}.$$

Thus, we also have for a GGM:

$$(j,k) \text{ and } (k,j) \in E \Longleftrightarrow \Sigma^{-1}_{j,k} \neq 0 \Longleftrightarrow \rho_{jk|V \setminus \{j,k\}} \neq 0 \Longleftrightarrow \beta_k^{(j)} \neq 0 \text{ and } \beta_j^{(k)} \neq 0. \tag{13.5}$$

Formula (13.5) links a GGM to the variable selection problem in regression. In particular for the high-dimensional case, this is a fruitful connection: more details are given in Section 13.4.2.

13.4.1 Penalized estimation for covariance matrix and edge set

Estimation with an ℓ_1-penalty can be used for inferring the structure of a GGM and its underlying covariance matrix Σ. The negative Gaussian log-likelihood (scaled with n^{-1}) for data

$$X_1, \ldots, X_n \text{ i.i.d. } \mathcal{N}_p(\mu, \Sigma),$$

and when plugging in the maximum likelihood solution $\hat{\mu} = n^{-1} \sum_{i=1}^{n} X_i$ for μ, equals (Problem 13.2)

$$-n^{-1}\ell(\Sigma^{-1}; X_1, \ldots, X_n) = -\log(\det \Sigma^{-1}) + \text{trace}(\hat{\Sigma}_{\text{MLE}} \Sigma^{-1}) + D, \quad (13.6)$$

where $\hat{\Sigma}_{\text{MLE}} = n^{-1} \sum_{i=1}^{n} (X_i - \hat{\mu})(X_i - \hat{\mu})^T$ is the empirical covariance matrix, and D is a constant with respect to Σ^{-1}. Note that we allow here for $\mu \neq 0$ (compare with (13.2)). As for the Lasso (for general likelihood problems), we add an ℓ_1-penalty and consider the estimator:

$$\hat{\Sigma}^{-1}(\lambda) = \arg\min_{\Sigma^{-1} \succ 0} (-\log(\det \Sigma^{-1}) + \text{trace}(\hat{\Sigma}_{\text{MLE}} \Sigma^{-1}) + \lambda \|\Sigma^{-1}\|_1),$$

$$\|\Sigma^{-1}\|_1 = \sum_{j<k} |\Sigma_{j,k}^{-1}|, \quad (13.7)$$

where the minimization is over positive definite matrices. Since $\hat{\Sigma}^{-1}(\lambda)$ is positive definite, its inverse $\hat{\Sigma}(\lambda)$ exists and is an estimate of the covariance matrix Σ. The minimization in (13.7) amounts to a convex optimization problem and fast algorithms have been proposed (Friedman et al., 2007b; Banerjee et al., 2008). The procedure is sometimes referred to as the Graphical Lasso (GLasso) (Friedman et al., 2007b) because of the use of the ℓ_1-penalty. Typically, it shrinks some of the non-diagonal elements exactly to zero, i.e. $\hat{\Sigma}_{j,k}^{-1}(\lambda) = 0$ for some (j,k) (depending on the size of λ). Another version of the penalty function is $\lambda \sum_{j \leq k} |\Sigma_{j,k}^{-1}|$, where the diagonal elements of Σ^{-1} are penalized as well (Friedman et al., 2007b). Note that $1/\Sigma_{j,j}^{-1} = \text{Var}(\varepsilon^{(j)})$, where $\varepsilon^{(j)}$ is the error in the regression in (13.4): hence, penalizing $|\Sigma_{j,j}^{-1}|$ encourages large values for the error variance $\text{Var}(\varepsilon^{(j)})$ which seems unnatural and we prefer the definition where the diagonal entries of Σ^{-1} are not penalized, as also proposed in Rothman et al. (2008).

Selection of the regularization parameter λ can be done using cross-validation for the negative Gaussian log-likelihood loss. When having a training and validation set, the validated negative Gaussian log-likelihood loss can be derived from (13.6) up to a constant:

$$-\log(\det \hat{\Sigma}_{\text{train}}^{-1}(\lambda)) + \text{trace}(\hat{\Sigma}_{\text{valid,MLE}} \hat{\Sigma}_{\text{train}}^{-1}(\lambda)),$$

where $\hat{\Sigma}_{\text{valid,MLE}}$ is the empirical covariance matrix from the validation sample. A cross-validation scheme then repeats this operation and averages the resulting scores.

The accuracy of $\hat{\Sigma}^{-1}$ or $\hat{\Sigma}$ can be measured by a prediction-type loss such as the Kullback-Leibler divergence

$$\rho_{KL}(\hat{\Sigma}^{-1}, \Sigma^{-1}) = \text{trace}(\Sigma \hat{\Sigma}^{-1}) - \log |\Sigma \hat{\Sigma}^{-1}| - p,$$

or considering an estimation error in terms of e.g. the Frobenius-norm

$$\|\hat{\Sigma}^{-1} - \Sigma^{-1}\|_F = \sum_{j,k} (\hat{\Sigma}_{j,k}^{-1} - \Sigma_{j,k}^{-1})^2,$$

or the operator matrix ℓ_q-norm $\|\cdot\|_{q,q}$ ($1 \le q \le \infty$).

For estimating the edge set in a Gaussian graphical model, we can proceed in the spirit of variable selection with the Lasso in (generalized) linear models. Without any significance testing, we can use the estimator

$$\hat{E}(\lambda) = \{(j,k) \in V \times V; \ \hat{\Sigma}_{j,k}^{-1}(\lambda) \ne 0\}. \tag{13.8}$$

This estimator is motivated by the population version

$$E_0 = \{(j,k) \in V \times V; \ (\Sigma^0)_{j,k}^{-1} \ne 0\},$$

see formula (13.3).

Such an estimator is computationally feasible in high dimensions and consistent for inferring the true underlying edge set E_0 if we require a rather restrictive form of an irrepresentable condition (in terms of $(\Sigma^0)^{-1}$ and Σ^0). Such a condition involves the Fisher-information of the multivariate Gaussian distribution: we refer to Ravikumar et al. (2008) for more details. However, it is argued in Meinshausen (2008) that the nodewise regression pursuit discussed below in Section 13.4.2 is asymptotically consistent for estimating the edge set E_0 under less restrictive conditions on the covariance Σ^0 or its inverse; also Ravikumar et al. (2008) discuss this issue in detail.

Practically relevant is the following analogy to the variable selection property of the Lasso in regression: when choosing the regularization parameter λ via cross-validation for the likelihood loss, we typically obtain the screening property (assuming implicitly that a certain compatibility or restricted eigenvalue condition holds and that the non-zero elements of $(\Sigma^0)^{-1}$ are sufficiently large):

$$\hat{E}(\hat{\lambda}_{CV}) \supseteq E_0,$$

saying that the estimated graph contains the true underlying edge set E_0. Furthermore, we can use a two stage adaptive GLasso estimator:

$$\hat{\Sigma}^{-1}(\lambda) = \underset{\Sigma^{-1} \succ 0}{\arg\min}(-\log(\det \Sigma^{-1}) + \text{trace}(\hat{\Sigma}_{\text{MLE}}\Sigma^{-1}) + \lambda \sum_{j<k} w_{jk}|\Sigma_{j,k}^{-1}|),$$

$$w_{jk} = 1/|\hat{\Sigma}_{\text{init};j,k}^{-1}|,$$

where $\hat{\Sigma}_{\text{init}}^{-1}$ is the GLasso estimator from the first initial stage. Fan et al. (2009a) relate this procedure to a one-step approximation for a non-convex penalty function, similar to our discussion of formula (2.30) in Section 2.8.5, and see also Section 6.11 and Section 7.13. We do not provide any mathematical results regarding the estimator in (13.7), and we refer the reader to Rothman et al. (2008) and Ravikumar et al. (2008). Their mathematical techniques are rather different than our theoretical

framework from Chapters 6 and 7 which - in principle - could be used as well to establish consistent estimation and optimality via oracle inequalities. Roughly speaking, if $p \gg n$, a suitable sparsity condition and an irrepresentable assumption (in terms of $(\Sigma^0)^{-1}$ and Σ^0) are required.

We summarize that the GLasso estimator in (13.7) exploits and enjoys positive definiteness for $\hat{\Sigma}^{-1}$ and $\hat{\Sigma}$ and hence, the estimate is often very good in terms of the predictive Kullback-Leibler divergence or e.g. the Frobenius or operator norm. For estimation of the edge set in a Gaussian graphical model, however, asymptotic consistency of the GLasso requires sufficiently large non-zero entries of $(\Sigma^0)^{-1}$ and strong coherence or irrepresentable conditions which are more restrictive than what is needed for the nodewise regression pursuit described below in Section 13.4.2.

13.4.1.1 Stability selection with the GLasso

We present empirical results when using the GLasso estimator from (13.7) for a sub-problem of the dataset about riboflavin production with bacillus subtilis, see Section 9.2.6: here, we consider $p = 160$ gene expression variables and $n = 115$ samples. The response variable, measuring the riboflavin production rate, is not of interest. We show (Figure 13.1) the estimated zero-pattern of Σ^{-1} which we display in terms of a Gaussian conditional independence graph with estimated edge set as in (13.8). In addition, we also present the solution when using stability selection with pointwise control as described in Section 10.3.1 in Chapter 10, i.e., we do stability selection for single regularization parameters λ of the GLasso estimator. Thereby, we run stability selection by controlling the number of expected false positives $\mathbb{E}[V] \leq 30$, that is we expect fewer than 30 wrong (false positive) among the 12'720 possible edges in the graph.

Figure 13.1 illustrates the results for different tuning parameters λ. The most striking feature is that with stability selection, the estimated edge set changes very little as we vary the regularization parameter λ of the original GLasso procedure.

Next, we permute the variables (gene expression values) randomly, using a different permutation for each variable (gene). Thus, all variables are independent of each other and the underlying conditional independence graph is the empty graph. Then, the exchangeability condition from Theorem 10.1 for stability selection holds. The results are displayed in Figure 13.2. Even though the original GLasso procedure selects way too many edges (since the regularization parameters is of too small order of magnitude), most of them turn out to be unstable and stability selection yields essentially the true underlying empty graph.
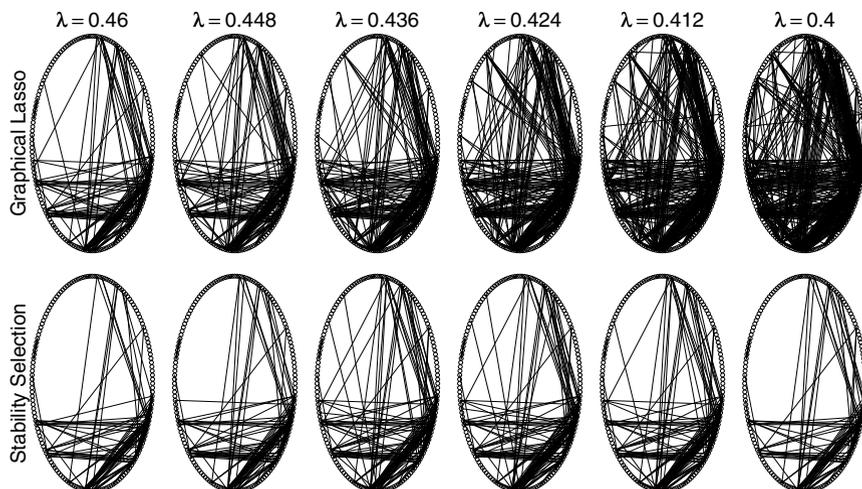
Fig. 13.1 Riboflavin production dataset with $p = 160$ and $n = 115$. Part of the regularization path of the GLasso (top row) and the corresponding point-wise stability selected models (bottom row). The figure is taken from Meinshausen and Bühlmann (2010).
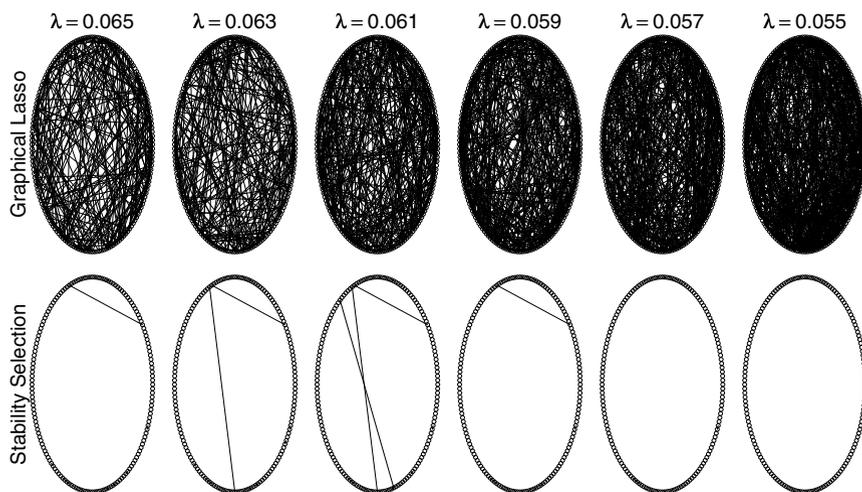
Fig. 13.2 Permuted variables in riboflavin production dataset with $p = 160$ and $n = 115$. Part of the regularization path of the GLasso (top row) and the corresponding point-wise stability selected models (bottom row). The figure is taken from Meinshausen and Bühlmann (2010).

13.4.2 Nodewise regression

Formula (13.5) leads to inferring the edges in a Gaussian graphical model by pursuing many regressions, as originally proposed by Meinshausen and Bühlmann

(2006):

$$X^{(j)} = \sum_{r \neq j} \beta_r^{0(j)} X^{(r)} + \varepsilon^{(j)}, \quad j = 1, \ldots, p, \tag{13.9}$$

where $\beta^{0(j)}$ denotes the true parameter vector. This is called nodewise regression. We assume that we have a variable selection procedure for each of the p regressions above. That is, we have estimates

$$\hat{S}^{(j)} \text{ for } S_0^{(j)} = \{r; \beta_r^{0(j)} \neq 0, r = 1, \ldots, p, r \neq j\}, \quad j = 1, \ldots, p.$$

For example, the Lasso yields

$$\hat{S}^{(j)} = \{r; \hat{\beta}_r^{(j)}(\lambda) \neq 0\}, \tag{13.10}$$

where $\hat{\beta}^{(j)}(\lambda)$ are the estimated regression coefficients from the Lasso with tuning parameter λ (when regressing $X^{(j)}$ versus $\{X^{(r)}; r \neq j\}$). See Section 2.6 and Chapter 7 for more details. An alternative variable selection procedure is given by the adaptive or thresholded Lasso, analogously as above but using $\hat{\beta}^{(j)}$ from the adaptive or thresholded two-stage procedures. For details about these methods, see Sections 2.8-2.9 and Chapter 7.

Based on $\hat{S}^{(j)}$, we build an estimate of the graph structure as follows. We can use the "or"-rule and define:

estimate an edge between nodes j and k $\Longleftrightarrow k \in \hat{S}^{(j)}$ or $j \in \hat{S}^{(k)}$.

A more conservative approach is based on the "and"-rule:

estimate an edge between nodes j and k $\Longleftrightarrow k \in \hat{S}^{(j)}$ and $j \in \hat{S}^{(k)}$.

We note that for the population analogue as in (13.5), the "and"- and "or"-rule coincide.

From the viewpoint of asymptotic consistency for the edges in a GGM, it is sufficient that

$$\sum_{j=1}^{p} \mathbf{P}[\hat{S}^{(j)} \neq S_0^{(j)}] \to 0 \ (n \to \infty), \tag{13.11}$$

for the "and"- or "or"-rule. This follows directly from (13.5) and the Bonferroni bound for $\mathbf{P}[\hat{S}^{(j)} \neq S_0^{(j)}$ for some $j = 1, \ldots, p]$.

Assuming for every regression in (13.9) the irrepresentable condition (for random designs) and a beta-min condition on the size of the minimal absolute value of the non-zero regression coefficients, the Lasso fulfills (13.11). This follows from Section 2.6 and with more rigorous statements given in Chapter 7. Meinshausen and Bühlmann (2006) who proposed this approach formulate the conditions in terms of

the true underlying covariance matrix Σ^0 and $(\Sigma^0)^{-1}$. Similarly, (13.11) also holds for the adaptive or thresholded Lasso under more relaxed conditions on the designs in the many regressions, see also Chapter 7.

Nodewise regression seems, at first sight, less powerful for inferring the edge set E_0 than a simultaneous approach considering all nodes in the graph at once as with the GLasso approach in Section 13.4.1. However, the sufficient and essentially necessary conditions for consistent estimation of E_0 (in terms of $(\Sigma^0)^{-1}$ or Σ^0) are weaker for nodewise regression than for GLasso. More discussion on this issue is given in Meinshausen (2008) and Ravikumar et al. (2008).

13.4.3 Covariance estimation based on undirected graph

A powerful way to estimate a high-dimensional covariance matrix and its inverse can be based on the structure of a graph. Consider an undirected conditional independence graph (CIG) G. We can then infer the covariance matrix Σ via maximum likelihood estimation with a constraint that the zero-elements of Σ^{-1} correspond to the non-edges in the CIG G. We abbreviate this constraint by $C(\Sigma^{-1} \leftrightarrow G)$:

$$\hat{\Sigma}_G^{-1} = \underset{\Sigma^{-1} \succ 0, C(\Sigma^{-1} \leftrightarrow G)}{\arg\min} \left(-\log(\det \Sigma^{-1}) + \operatorname{trace}(\hat{\Sigma}_{\mathrm{MLE}} \Sigma^{-1}) \right), \quad (13.12)$$

and by matrix inversion we also obtain a covariance estimate $\hat{\Sigma}_G$. If the CIG G is unknown, we can base the estimator in (13.12) on an estimate $\hat{G}$ to obtain $\hat{\Sigma}_{\hat{G}}^{-1}$ and $\hat{\Sigma}_{\hat{G}}$. The computation of the estimator in (13.12) is as for the GLasso in (13.7), involving convex optimization over positive definite matrices but without penalty term in (13.12).

The estimate $\hat{G}$ could be from the GLasso in (13.7). Then, the estimator in (13.12) is a GLasso-MLE hybrid estimator, analogous to the Lasso-OLS estimator briefly described in Chapter 2, Section 2.10.

Alternatively, we can use the nodewise regression approach from the previous Section 13.4.2 for an estimate $\hat{G}$. This has the advantage that the estimator $\hat{G}$ is consistent for a broader range of scenarios than the GLasso, as briefly discussed in Section 13.4.2. In particular, using the estimator in (13.12) in conjunction with the powerful nodewise regression approach addresses the drawback that nodewise regression alone does not yield an estimate of the covariance Σ or its inverse Σ^{-1}. Zhou et al. (2010) present some mathematical analysis of such a two-stage procedure. There, the nodewise regression estimates $\hat{\beta}^{(j)} = \hat{\beta}^{(j)}(\lambda)$ in (13.10) are thresholded Lasso estimators:

$$\hat{\beta}_{\mathrm{thres},r}^{(j)} = \hat{\beta}_r^{(j)} 1(|\hat{\beta}_r^{(j)}| > \lambda_{\mathrm{thres}}), \quad (13.13)$$

where $\lambda_{\text{thres}} > 0$ is a threshold parameter, analogous to Section 2.9 and Section 7.6.

When using the GLasso or the nodewise regression approach for $\hat{G}$ and then employing the estimator in (13.12) in a second stage, we typically get (much) improved performance for estimation of Σ and Σ^{-1} if the true underlying inverse covariance matrix is sparse with a few non-zero entries having large absolute values, in comparison to a single-stage GLasso estimate. This finding is in line with what we discussed for linear models in Sections 2.8, 2.9 and 2.10, and we also refer to Sections 6.10 and 7.7.

13.4.3.1 Gene expressions from isoprenoid biosynthesis pathway in arabidopsis thaliana

We illustrate the methods on gene expression data from the isoprenoid biosynthesis pathway in arabidopsis thaliana given in Wille et al. (2004). In plants, isoprenoids play important roles in a variety of processes such as photosynthesis, respiration, regulation of growth and development. The data set consists of $p = 39$ isoprenoid genes for which we have $n = 118$ gene expression measurements under various experimental conditions. As performance measure we use the 10-fold cross-validated negative Gaussian log-likelihood for centered data. Figure 13.3 presents some re-
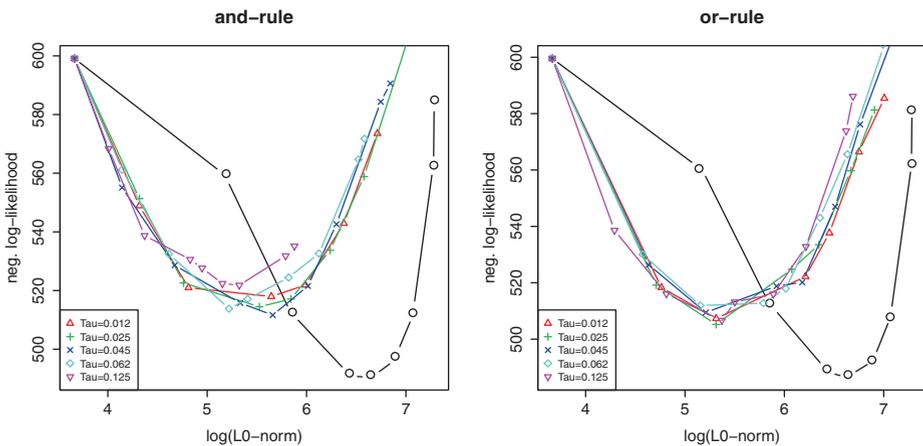
Fig. 13.3 Covariance estimation for arabidopsis gene expression data with $n = 118$ and $p = 39$. x-axis: $\log$(no. non-zero elements of $\hat{\Sigma}^{-1}$); y-axis: 10-fold cross-validated negative Gaussian. GLasso (black line) and MLE in (13.12) based on nodewise regression estimated graph using the thresholded Lasso for various threshold parameters denoted by τ (colored lines). Left: using the "and"-rule for nodewise regression graph estimate; Right: using the "or"-rule for nodewise regression graph estimate (see Section 13.4.2).

sults for the GLasso and the covariance estimator in (13.12) based on an estimated undirected graph from nodewise Lasso regression with thresholding as indicated in (13.13). Regarding the latter, different threshold parameters are displayed with different curves, and varying the regularization parameter in GLasso or in the nodewise Lasso regressions yield the curves in the graphs. We see that the GLasso exhibits an approximately 5% improved performance in terms of the cross-validated negative log-likelihood. However, when requiring more sparse solutions, estimation based on an undirected graph from nodewise regression is better. This finding fits into the typical picture that for highly sparse inverse covariance matrices, the two-stage procedure using MLE based on an estimated graph performs better.

13.5 Ising model for binary random variables

We consider here the situation where all the variables $X^{(1)}, \ldots, X^{(p)} \in \{0,1\}$ are binary. An interesting model for such binary variables is the Ising model with the joint distribution

$$p(x^{(1)}, \ldots, x^{(p)}) = \frac{1}{Z(\gamma)} \exp\left(\sum_{j,k=1,\ldots,p} \gamma_k^{(j)} x^{(k)} x^{(j)}\right), \qquad (13.14)$$

where all $\gamma_k^{(j)} \in \mathbb{R}$ and $Z(\gamma)$ is a normalization factor ensuring that the probabilities sum up to one. The conditional independence graph is given from Definition 13.3. For the Ising model (13.14), the following is true (Problem 13.3):

$$X^{(j)} \perp X^{(k)} | X^{(V \setminus \{j,k\})} \iff \gamma_k^{(j)} \neq 0 \text{ and } \gamma_j^{(k)} \neq 0. \qquad (13.15)$$

This is analogous to formula (13.5), due to the structure of the Ising model in (13.14). However, this structure does not necessarily hold for more general distributions of binary variables: that is, there could be higher-order interaction terms whereas this is not possible in the multivariate Gaussian case.

Formula (13.15) can be re-written in terms of logistic regression. Consider

$$\text{logit}(\mathbf{P}[X^{(j)} = 1 | X^{(V \setminus \{j\})}]) = \sum_{k \neq j} \beta_k^{(j)} X^{(k)} \quad (j = 1, \ldots, p),$$

where $\text{logit}(p) = \log(p/(1-p))$ for $0 < p < 1$. It then holds that

$$\beta_k^{(j)} = 2\gamma_k^{(j)} \text{ and hence } \gamma_k^{(j)} \neq 0 \iff \beta_k^{(j)} \neq 0. \qquad (13.16)$$

We leave the derivation of (13.16) as Problem 13.4. Therefore, because of (13.15), we can infer the conditional independence graph for binary random variables from

an Ising model in (13.14) via nodewise logistic regression, analogous to the method described in Section 13.4.2.

We can use any reasonable variable selection procedure for each of the p logistic regressions above with estimates

$$\hat{S}^{(j)} \text{ for } S_0^{(j)} = \{r; \ \beta_r^{0(j)} \neq 0, \ r = 1, \ldots, p, \ r \neq j\}, \ \ j = 1, \ldots, p,$$

where we denote by $\beta^{0(j)}$ the true underlying logistic regression parameters.

For example, the Lasso for logistic regression with tuning parameter λ, as described in Chapter 3, yields

$$\hat{S}^{(j)} = \{r; \ \hat{\beta}_r^{(j)}(\lambda) \neq 0\}.$$

Based on $\hat{S}^{(j)}$, we can estimate the conditional independence graph with edge set E_0 analogously as with nodewise regression in the Gaussian case (Section 13.4.2). When using the "or"-rule, we define $\hat{E}_{\text{or}}$:

$$\{(j,k),(k,j)\} \in \hat{E}_{\text{or}} \iff k \in \hat{S}^{(j)} \text{ or } j \in \hat{S}^{(k)}.$$

A more conservative approach is based on the "and"-rule with the estimator $\hat{E}_{\text{and}}$:

$$\{(j,k),(k,j)\} \in \hat{E}_{\text{and}} \iff k \in \hat{S}^{(j)} \text{ and } j \in \hat{S}^{(k)}.$$

We note that both rules coincide for the population analogue in (13.15). The method is consistent for finding the true underlying edge set, assuming sparsity, restrictive conditions in terms of the corresponding Fisher-information matrix (i.e. analogous to the irrepresentable condition for Gaussian graphical models) and a beta-min condition saying that the non-zero coefficients $\gamma_k^{(j)}$ in the Ising model are sufficiently large. For detailed mathematical arguments, we refer the interested reader to Ravikumar et al. (2009b).

13.6 Faithfulness assumption

Consider a graphical model (G, P) where P satisfies a Markov condition (pairwise or global) relative to G. Thanks to the Markov condition, we can infer from the graph G some conditional independences for the distribution P. In general, the distribution P may include other conditional independence relations than those entailed by the Markov condition.

Definition 13.4. *We say that the probability distribution P is* faithful *to the graph G if the following equivalences hold: for every triple of disjoint sets $A, B, C \subseteq V$,*

$$C \text{ separates } A \text{ and } B \Longleftrightarrow X^{(A)} \perp X^{(B)} | X^{(C)}.$$

Faithfulness says that we can read of *all* conditional independences from the graphical concept of separation. The implication "$\Longrightarrow$" follows from assuming the global Markov property, and faithfulness requires that the other implication "$\Longleftarrow$" holds as well.

As a direct implication of the faithfulness assumption we obtain (Problem 13.5):

$$X^{(j)} \perp X^{(k)} | X^{(C_1)} \Longrightarrow X^{(j)} \perp X^{(k)} | X^{(C_2)} \text{ for all } C_2 \supseteq C_1 \text{ with } j, k \notin C_2. \quad (13.17)$$

13.6.1 Failure of faithfulness

We give two examples of non-faithful distributions.

13.6.1.1 Cancellation of regression coefficients

Consider an undirected conditional independence graph G with vertices $V = \{1, 2, 3\}$ and $E = \{(1,2),(2,1),(2,3),(3,2),(1,3),(3,1)\}$. That is, all vertices are connected by an edge. Assume the global Markov property. Via this property, we can read off
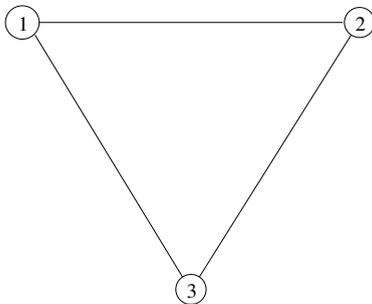
Fig. 13.4 Conditional independence graph G where all $p = 3$ vertices are fully connected.

some (conditional) independence relations from the graph: for this case, however, no (conditional) independences can be inferred (since none of the pairs of nodes can be separated by another node). Nevertheless, it may happen that two variables are marginally independent, i.e., an independence relation which cannot be read off from the graph. A concrete construction for this phenomenon is as follows:

$$X^{(1)} = \varepsilon^{(1)},$$

$$X^{(2)} = \alpha X^{(1)} + \varepsilon^{(2)},$$
$$X^{(3)} = \beta X^{(1)} + \gamma X^{(2)} + \varepsilon^{(3)},$$

where $\varepsilon^{(1)}, \varepsilon^{(2)}, \varepsilon^{(3)}$ i.i.d. $\sim \mathcal{N}(0,1)$, $\varepsilon^{(3)}$ independent of $\{X^{(2)}, X^{(1)}\}$ and $\varepsilon^{(2)}$ independent of $X^{(1)}$. Then,

$$(X^{(1)}, X^{(2)}, X^{(3)}) \sim \mathcal{N}_3(0, \Sigma),$$
$$\Sigma = \begin{pmatrix} 1 & \alpha & \beta + \alpha\gamma \\ \alpha & \alpha^2 + 1 & \alpha\beta + \gamma(\alpha^2 + 1) \\ \beta + \alpha\gamma & \alpha\beta + \gamma(\alpha^2 + 1) & \beta^2 + \gamma^2(\alpha^2 + 1) + 1 + 2\alpha\beta\gamma \end{pmatrix}.$$

We can enforce marginal independence of $X^{(1)}$ and $X^{(3)}$ by choosing α, β, γ such that $\beta + \alpha\gamma = 0$, that is, cancellation of regression coefficients takes place. For example, take $\alpha = \beta = 1$, $\gamma = -1$. Then,

$$\Sigma = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}, \quad \Sigma^{-1} = \begin{pmatrix} 3 & -2 & -1 \\ -2 & 2 & 1 \\ -1 & 1 & 1 \end{pmatrix}.$$

Thus, since $\text{Cov}(X^{(1)}, X^{(3)}) = 0$ and because of joint Gaussianity, we conclude that $X^{(1)} \perp X^{(3)}$. Moreover, as described in Section 13.4, zero partial correlations correspond to zeroes in Σ^{-1}: since there are no zeroes in Σ^{-1}, all partial correlations are non-zero and hence, $X^{(i)} \not\perp X^{(j)}|X^{(k)}$ for all combinations of distinct indices $i \neq j \neq k \in \{1,2,3\}$. This, of course, is compatible with the graph in Figure 13.4.

Thus, this is an example with a distribution which is not faithful: for example $X^{(1)} \perp X^{(3)}$ does not imply $X^{(1)} \perp X^{(3)}|X^{(2)}$, compare with (13.17). It happens because the regression coefficients cancel in a very specific way. In a certain sense, this is "unlikely": Section 13.9.1 describes more details about this. But clearly, requiring faithfulness is restricting the class of probability distributions.

13.6.1.2 Moving-average model

Consider a moving average model of order q for stationary time series:

$$X_t = \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t, \; t = 1, 2, \ldots,$$
$$\varepsilon_1, \varepsilon_2, \ldots \text{ i.i.d. with } \mathbb{E}[\varepsilon_t] = 0, \; \mathbb{E}[\varepsilon_t^2] = \sigma^2 < \infty.$$

Then, the autocorrelation function satisfies

$$\text{Cor}(X_t, X_{t+k}) = 0 \text{ for } k > q.$$

On the other hand, if $\theta \neq 0$, the partial autocorrelation function satisfies

$$\text{Parcor}(X_t, X_{t+k}|X_{t+1}, \ldots, X_{t+k-1}) \neq 0 \text{ for all } k \geq 1,$$

cf. Brockwell and Davis (1991) (for $k = 1$, the partial correlation is defined as the marginal correlation). For example, for $q = 1$:

$$\text{Parcor}(X_t, X_{t+k}|X_{t+1}, \ldots, X_{t+k-1}) = -\frac{(-\theta_1)^k(1 - \theta_1^2)}{1 - \theta_1^{2(k+1)}}, \quad k \geq 2.$$

Thus, these models have the property that a correlation can be zero while the partial correlation is non-zero. In the Gaussian case where $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$ and partial correlations describe conditional dependences, this property is in conflict with the faithfulness assumption, see formula (13.17). Hence, Gaussian moving-average models correspond to non-faithful distributions (with respect to conditional independence graphs).

13.6.2 Faithfulness and Gaussian graphical models

In the sequel, we denote partial correlations by

$$\rho_{jk|C} = \text{Parcor}(X^{(j)}, X^{(k)}|X^{(C)}),$$

where $X^{(C)} = \{X^{(r)}; r \in C\}$ for some subset $C \subseteq V \setminus \{j, k\}$.

Proposition 13.2. *Consider a GGM where P is faithful with respect to the graph G. We then have:*

$$\rho_{jk|C_1} = 0 \implies \rho_{jk|C_2} = 0 \text{ for all } C_2 \supseteq C_1 \text{ with } j, k \notin C_2.$$

Proposition 13.2 immediately follows from (13.17). A weaker notion of so-called partial faithfulness is discussed in Section 13.9.1: there, we require the assertion of Proposition 13.2 only for the set $C_2 = V \setminus \{j, k\}$. In case of faithfulness, we have a stronger result than in formula (13.5).

Proposition 13.3. *Consider a GGM where P is faithful with respect to the graph G. We then have:*

$$(j, k) \text{ and } (k, j) \in E \iff \rho_{jk|C} \neq 0 \text{ for all subsets } C \subseteq V \setminus \{j, k\}.$$

We leave the proof as Problem 13.6. We see from Proposition 13.3 that the faithfulness assumption yields a graph with fewer edges than without requiring faithfulness. In case of the latter, we have no edge between vertices j and k if and only if $\rho_{jk|V \setminus \{j,k\}} = 0$, whereas with faithfulness, we only need that $\rho_{jk|C} = 0$ for some $C \subseteq V \setminus \{j, k\}$.

Proposition 13.3 has algorithmic implications. We can hierarchically screen marginal correlations ρ_{jk} and then low-order partial correlations $\rho_{jk|C}$ with $|C|$ small: if one of them is zero, we know that there is no edge between j and k. Of course, it is much easier to estimate marginal or low-order partial correlations than a higher-order partial correlation. We will see in Section 13.7 how the PC-algorithm exploits this property in a hierarchical way, assuming sparsity which restricts the size of the adjacency set $\mathrm{adj}(G, j) = \{k \in V; (j, k) \text{ and } (k, j) \in E\}$ for every vertex $j \in V$ in the graph G.

13.7 The PC-algorithm: an iterative estimation method

The PC-algorithm (Spirtes et al., 2000), where "PC" stands for the first names of the inventors **P**eter Spirtes and **C**larke Glymour, is a clever iterative multiple testing procedure for inferring zero partial correlations. Roughly speaking, we exploit the fact from Proposition 13.2 in the following way. If a marginal correlation $\rho_{jk} = 0$, there is no need to consider partial correlations $\rho_{jk|C}$ of higher order with $|C| \geq 1$. Analogously, if a first order partial correlation $\rho_{jk|m} = 0$, we do not need to consider higher order partial correlations $\rho_{jk|C}$ with $m \in C$ and $|C| \geq 2$; and so on. Thus, faithfulness allows to test partial correlations in a hierarchical way, from marginal to first- and then to higher-order partial correlations.

13.7.1 Population version of the PC-algorithm

In the population version of the PC-algorithm ($\mathrm{PC}_{\mathrm{pop}}$), we assume that perfect knowledge about all necessary conditional independence relations is available. What we refer here to as the PC-algorithm is what others call the first part of the PC-algorithm; the second part infers directions for some edges in a directed acyclic graph (DAG). Most often, the PC-algorithm is used in connection with estimating a DAG (or its equivalence class), but we do not cover this topic in the book.

The maximal value of ℓ of the order of the partial correlations in Algorithm 12 is denoted by

$$m_{\mathrm{reach}} = \text{ maximal value of } \ell \text{ reached,} \qquad (13.18)$$

which depends on the underlying distribution.

We explain now the principles of the $\mathrm{PC}_{\mathrm{pop}}$ algorithm in words. Thereby, we focus on Gaussian distributions where we can characterize conditional independences by corresponding partial correlations being equal to zero. The algorithm then starts with a full graph where all nodes are connected with edges to each other. In Step 6, we

---

Algorithm 12 The PC$_{\text{pop}}$-algorithm

 1: **INPUT:** Vertex set V, conditional independence information
 2: **OUTPUT:** Conditional independence graph G
 3: Form the complete undirected graph $\tilde{G}$ on the vertex set V
 4: Set $\ell = -1$; set $G = \tilde{G}$
 5: **repeat**
 6:     Increase $\ell \leftarrow \ell + 1$
 7:     **repeat**
 8:         Select a (new) ordered pair of nodes (j, k) that are adjacent in G such that
             $|\text{adj}(G, j) \setminus \{k\}| \geq \ell$
 9:         **repeat**
10:             Choose (new) $C \subseteq \text{adj}(G, j) \setminus \{k\}$ with $|C| = \ell$.
11:             **if** j and k are conditionally independent given C **then**
12:                 Delete edge $j - k$
13:                 Denote this new graph by G
14:             **end if**
15:         **until** edge $j - k$ is deleted or all $C \subseteq \text{adj}(G, j) \setminus \{k\}$ with $|C| = \ell$ have been chosen
16:     **until** all ordered pairs of adjacent nodes (j, k) such that $|\text{adj}(G, j) \setminus \{k\}| \geq \ell$ and $C \subseteq$
         $\text{adj}(G, j) \setminus \{k\}$ with $|C| = \ell$ have been tested for conditional independence
17: **until** for each ordered pair of adjacent nodes (j, k): $|\text{adj}(G, j) \setminus \{k\}| < \ell$.

---

start with $\ell = 0$, and we first consider marginal correlations (Step 10 with $|C| = 0$) among all pairs of variables (Step 8 and noting that the current G is the full graph): we then delete the edges whose corresponding marginal correlations are zero (Steps 11 and 12). After this first round considering marginal correlations only, we obtain a smaller graph (Step 13). We then increase the index ℓ to $\ell = 1$ which means that we will look at partial correlations of order $\ell = 1$. We select a pair of variables with nodes j and k which are connected by an edge and we consider the partial correlation of order $\ell = 1$ by conditioning on some variable whose node C is connected to at least one of the nodes j or k: if the partial correlation is zero, we delete the edge between j and k, and if it is non-zero, we try other conditioning nodes C which are connected to at least one of the nodes j or k. If none of the conditioning nodes C yields zero partial correlation, we keep the edge between j and k; and vice-versa, we delete the edge if there is a conditioning node C which leads to zero partial correlation. We then do this partial correlation screening of order one for all other pairs of connected nodes: in every of these screens, the current graph may become less dense with fewer edges. Then, we increase the index ℓ to $\ell = 2$ and consider partial correlations of order $\ell = 2$ and so on. Since $\ell \leq m_{\text{reach}}$, we will only consider partial correlations of order less or equal to m_{reach}: if the true underlying graph is sparse, m_{reach} is small (see Proposition 13.4) and hence, estimation of such lower-order partial correlations is not too much ill-posed. Another important algorithmic feature is that the conditioning variables with nodes C (Step 10) are only from the neighborhoods of the current graph and hence, assuming a sparse underlying true graph, it is feasible to screen among all such conditioning nodes C. In fact the, algorithm is computationally feasible for p in the thousands but assuming that the true graph is reasonably sparse. We will discuss in Section 13.7.2 how to modify the PC$_{\text{pop}}$ algorithm when estimating partial correlations from data.

A proof that Algorithm 12 produces the correct CIG can be deduced from the proof of Proposition 13.6 in Section 13.9.2 which treats the simpler problem of variable selection in a linear model. From this, it follows that the PC$_{pop}$ algorithm could be slightly simplified: in Step 8, instead of going through all pairs of ordered random variables, we could consider an asymmetric version with pairs of ordered random variables $\{(X^{(j)}, X^{(k)}); \ j < k\}$ which - as a disadvantage - depends on the ordering when it involves estimation (but not for the population version). We summarize the property of the PC$_{pop}$ as follows.

Proposition 13.4. *Consider a Gaussian graphical model with conditional independence graph (CIG) G and distribution P, and assume that P is faithful to G. Denote the maximal number of neighbors by $q = \max_{1 \le j \le p} |\text{adj}(G, j)|$. Then, the PC$_{pop}$-algorithm constructs the true underlying graph G. Moreover, for the reached level: $m_{\text{reach}} \in \{q - 1, q\}$.*

We note that the PC$_{pop}$-Algorithm 12 also works for non-Gaussian, faithful distributions. However, inferring conditional independences in non-Gaussian distributions from finitely many data is generally much more difficult than what we discuss next.

13.7.2 Sample version for the PC-algorithm

For finite samples, we need to estimate conditional independences. We limit ourselves to the Gaussian case, where all nodes correspond to random variables with a multivariate normal distribution as in (13.2). Then, conditional independences can be inferred from partial correlations equaling zero: $\rho_{jk|C} = \text{Parcor}(X^{(j)}, X^{(k)} | X^{(C)}) = 0$ if and only if $X^{(j)} \perp X^{(k)} | X^{(C)}$, see (13.5) (thereby using $V = C \cup \{j, k\}$).

We can thus estimate partial correlations to obtain estimates of conditional independences. The sample partial correlation $\hat{\rho}_{jk|C}$ can be calculated via regression, inversion of parts of the empirical covariance matrix or recursively by using the following identity: for any $h \in C$,

$$\hat{\rho}_{jk|C} = \frac{\hat{\rho}_{jk|C \setminus h} - \hat{\rho}_{jh|C \setminus h} \hat{\rho}_{kh|C \setminus h}}{\sqrt{(1 - \hat{\rho}_{jh|C \setminus h}^2)(1 - \hat{\rho}_{kh|C \setminus h}^2)}}. \tag{13.19}$$

Using this recursion, we start with the sample covariance estimate $\hat{\Sigma}_{\text{MLE}} = n^{-1} \sum_{i=1}^{n} (X_i - \overline{X})(X_i - \overline{X})^T$, calculate from it the sample correlation matrix and recursively compute sample partial correlations according to formula (13.19).

For testing whether a partial correlation is zero or not, we apply Fisher's z-transform

$$Z(jk|C) = \frac{1}{2} \log \left(\frac{1 + \hat{\rho}_{jk|C}}{1 - \hat{\rho}_{jk|C}} \right). \tag{13.20}$$

Under the null-hypothesis that $\rho_{jk|C} = 0$ and assuming a multivariate Gaussian distribution, the distribution of Fisher's z-transform $\sqrt{n - |C| - 3}Z(jk|C) \approx \mathcal{N}(0,1)$ is asymptotically standard Normal, see for example Anderson (1984, Sec. 4.3.3). Hence, we use the following decision- or thresholding-rule: reject the null-hypothesis $H_0(jk|C) : \rho_{jk|C} = 0$ against the two-sided alternative $H_A(jk|C) : \rho_{jk|C} \neq 0$ if

$$\sqrt{n - |C| - 3}|Z(jk|C)| > \Phi^{-1}(1 - \alpha/2),$$

where $\Phi(\cdot)$ denotes the cdf of $\mathcal{N}(0,1)$.

The sample version of the PC-algorithm is almost identical to the population version in Section 13.7.1.

Algorithm 13 The PC-algorithm

1: **INPUT:** Vertex set V, sample covariance estimate $\hat{\Sigma}_{\text{MLE}}$
2: **OUTPUT:** Estimated conditional independence graph $\hat{G}$
3: Run the PC_{pop}-Algorithm 12 from Section 13.7.1 but replace in line 11 of Algorithm 12 the if-statement by
   if $\sqrt{n - |C| - 3}|Z(jk|C)| \leq \Phi^{-1}(1 - \alpha/2)$ **then**

The PC-Algorithm 13 yields a data-dependent value $\hat{m}_{\text{reach},n}$ which is the sample version of (13.18). The only tuning parameter of the PC-algorithm is α, which is a significance level for testing individual partial correlations.

As we will see in Section 13.8, the PC-algorithm is asymptotically consistent even if p is much larger than n but assuming that the underlying graph is sparse and that the non-zero partial correlations are sufficiently large.

13.7.2.1 Computational complexity

The computational complexity of the PC-algorithm is difficult to evaluate exactly, but the worst case is bounded by

$$O(p^{2+\hat{m}_{\text{reach}}}) \text{ bounded with high probability by } O(p^{2+q}) \ (p \to \infty), \quad (13.21)$$

where $q = \max_{1 \leq j \leq p} |\text{adj}(G, j)|$ is the maximal size of the neighborhoods. We note that the bound may be very loose for many distributions. In fact, we can do the computations for fairly dense graphs, for example some nodes j having neighborhoods of size $|\text{adj}(G, j)|$ equal to 10 or 20.

We provide a small example of the processor time for estimating the graph using the PC-algorithm. The runtime analysis was done on a dual core processor 2.6 GHz and 4 GB RAM and using the R-package pcalg. The number of variables varied between $p = 10$ and $p = 1000$ while the number of samples was fixed at $n = 1000$.

The sparseness is measured in terms of average size of the adjacency sets $\text{adj}(G, j)$, over all nodes j (and also over different random graph structures): the corresponding theoretical quantity (expectation) is denoted by $\mathbb{E}[N]$ and we consider $\mathbb{E}[N] = 2$ and $\mathbb{E}[N] = 8$. For each parameter setting, 10 replicates were used. In each case, the tuning parameter used in the PC-algorithm was $\alpha = 0.01$. The average processor time together with its standard deviation for estimating the graph is given in Table 13.1. Graphs of $p = 1000$ nodes and 8 neighbors on average could be estimated in about 25 minutes, while graphs with up to $p = 100$ nodes could be estimated in about a second.

p	$\mathbb{E}[N]$	CPU
10	2	0.028 (0.003)
10	8	0.072 (0.004)
30	2	0.10 (0.01)
30	8	0.56 (0.03)
50	2	0.19 (0.01)
50	8	1.29 (0.04)
100	2	0.54 (0.03)
100	8	4.68 (0.16)
300	2	4.0 (0.05)
300	8	42 (1.43)
1000	2	50 (0.22)
1000	8	565 (26.1)

Table 13.1 Average processor time in seconds (CPU) for estimating simulated undirected graphs (GGMs) using the PC-algorithm with $\alpha = 0.01$, with standard errors in brackets. Various values of p and expected neighborhood size $\mathbb{E}[N]$, sample size $n = 1000$.

13.8 Consistency for high-dimensional data

This section is based on results from Kalisch and Bühlmann (2007). We assume that the data are realizations of i.i.d. random vectors $X_1, \ldots, X_n$ with $X_i \in \mathbb{R}^p$ having a probability distribution P. To capture high-dimensional behavior, we will let the dimension grow as a function of sample size: thus, $p = p_n$ and also the distribution $P = P^{(n)}$ and the graph $G = G^{(n)}$ are depending on n. That is, we consider a triangular scheme of observations (see also (2.6)):

$$X_{n,1}, \ldots, X_{n,n} \text{ i.i.d. } \sim P^{(n)}, \ n = 1, 2, 3, \ldots$$

Our assumptions are as follows.

(A1) The distribution $P^{(n)}$ is multivariate Gaussian and faithful to an undirected graph $G^{(n)}$ for all $n \in \mathbb{N}$.

(A2)   The dimension $p_n = O(n^a)$ for some $0 \le a < \infty$.

(A3)   The maximal number of neighbors in the undirected graph $G^{(n)}$, denoted by
$q_n = \max_{1 \le j \le p_n} |\mathrm{adj}(G, j)|$, satisfies $q_n = O(n^{1-b})$ for some $0 < b \le 1$.

(A4)   The partial correlations satisfy:

$$\inf\{|\rho_{jk|C}|;\ \rho_{jk|C} \ne 0,$$
$$j, k = 1, \ldots, p_n \ (j \ne k), C \subseteq \{1, \ldots, p_n\} \setminus \{j, k\},\ |C| \le q_n\} \ge c_n,$$

where $c_n^{-1} = O(n^d)$ $(n \to \infty)$ for some $0 < d < b/2$ and $0 < b \le 1$ as in (A3);

$$\sup_n \{|\rho_{jk|C}|;$$
$$j, k = 1, \ldots, p_n \ (j \ne k), C \subseteq \{1, \ldots, p_n\} \setminus \{j, k\}, |C| \le q_n\} \le M < 1.$$

Condition (A1) is an often used assumption in graphical modeling, although the
faithfulness assumption does restrict the class of probability distributions. Assump-
tion (A2) allows for an arbitrary polynomial growth of dimension as a function of
sample size. It could be relaxed to $p_n = O(\exp(n^\delta))$ for some sufficiently small
$0 < \delta < 1$. (A3) is a sparseness assumption. (A4) ensures detectability of non-zero
partial correlations: it could be reformulated to $c_n = \sqrt{q_n/n^{1-\kappa}}$ for $\kappa > 0$ arbitrarily
small which is slightly more restrictive (if p_n is polynomial in n) than the detectabil-
ity bound $\sqrt{q_n \log(p_n)/n}$, see for example the beta-min condition for regression in
(2.23) and in Section 7.4. Furthermore, (A4) restricts the linear dependence by re-
quiring an upper bound $M < 1$ for partial correlations. We note that (A4) involves
condition sets C with $|C| \le q_n$ and hence, it is a kind of "sparse partial correlation"
condition related to the discussion on sparse eigenvalues in Section 6.13.5 from
Chapter 6. The following result then holds for the PC-algorithm.

Theorem 13.1. *Consider a Gaussian graphical model with distribution $P^{(n)}$ and
underlying conditional independence graph $G^{(n)}$. Assume (A1)-(A4). Denote by
$\hat{G}_n(\alpha_n)$ the estimate from the PC-Algorithm 13 in Section 13.7.2. Then, there ex-
ists $\alpha_n \to 0$ $(n \to \infty)$, see below, such that*

$$\mathbf{P}[\hat{G}_n(\alpha_n) = G^{(n)}]$$
$$= 1 - O(\exp(-Kn^{1-2d})) \to 1 \ (n \to \infty),$$

for some $0 < K < \infty$ depending on M in (A4), and $d > 0$ is as in (A4).

A proof is given in the Section 13.8.2. A choice for the value of the tuning parameter
is $\alpha_n = 2(1 - \Phi(n^{1/2}c_n/2))$ which depends on the unknown lower bound of partial
correlations in (A4).

13.8.1 An illustration

We show the behavior of various error rates in a high-dimensional setting where the number of variables increases almost exponentially, the number of samples increases linearly and the expected neighborhood size increases sub-linearly. The theoretically expected neighborhood size of the adjacency sets $\text{adj}(G, j)$ over all nodes j (and also over different random graph structures) is denoted by $\mathbb{E}[N]$. By inspecting the theory, we would expect the false positive error rate (FPR) and true positive error rate (TPR),

$$\text{FPR} = \frac{\text{number of (estimated) false positives}}{\text{number of (true) non-positives}} = \frac{|\hat{E} \cap E_0^c|}{|E_0^c|}, \quad (13.22)$$

$$\text{TPR} = \frac{\text{number of (estimated) true positives}}{\text{number of (true) positives}} = \frac{|\hat{E} \cap E_0|}{|E_0|}, \quad (13.23)$$

to stay constant or even decrease. Table 13.2 shows the results of a small numerical study addressing this question; for details about the data-generating model see Kalisch and Bühlmann (2007). We used the PC-Algorithm 13 with $\alpha = 0.05$, and the results are based on 20 simulation runs. Indeed, because the expected neighbor-

p	n	$\mathbb{E}[N]$	TPR	FPR
9	50	1.4	0.61 (0.03)	0.023 (0.005)
27	100	2.0	0.70 (0.02)	0.011 (0.001)
81	150	2.4	0.753 (0.007)	0.0065 (0.0003)
243	200	2.8	0.774 (0.004)	0.0040 (0.0001)
729	250	3.2	0.794 (0.004)	0.0022 (0.00004)
2187	300	3.5	0.805 (0.002)	0.0012 (0.00002)

Table 13.2 The number of variables p increases almost exponentially, the sample size n increases linearly and the expected neighborhood size $E[N]$ increases sub-linearly. The results are based on using $\alpha = 0.05$, 20 simulation runs, and standard deviations are given in brackets.

hood size $\mathbb{E}[N] = 0.2\sqrt{n}$ increases sub-linearly in n, despite that p increases almost exponentially in n, both the FPR and TPR error rates improve as n gets larger, as supported by the consistency result in Theorem 13.1, see also assumption (A3). The ratio $\log(p)/n$ equals for the various scenarios: $0.044, 0.033, 0.029, 0.027, 0.026$ and 0.026, respectively.

13.8.2 Theoretical analysis of the PC-algorithm

The statistical properties of iterative algorithms can be studied by first considering the population version and then showing that the accumulation of estimation errors is under control. We use this strategy here to prove Theorem 13.1.

13.8.2.1 Analysis of partial correlations

We first establish uniform consistency of estimated partial correlations. Denote by $\hat{\rho}_{n;jk}$ and $\rho_{n;jk}$ the sample and population correlations between $X^{(j)}$ and $X^{(k)}$, respectively. Note that also the population parameters $\rho_{n;jk}$ depend on n due to the triangular array asymptotic framework as described at the beginning of Section 13.8. Likewise, $\hat{\rho}_{n;jk|C}$ and $\rho_{n;jk|C}$ denote the sample and population partial correlation between $X^{(j)}$ and $X^{(k)}$ given $X^{(C)} = \{X^{(r)}; r \in C\}$, where $C \subseteq \{1,\dots,p_n\} \setminus \{j,k\}$. In fact, it turns out that we only need to consider conditioning sets from

$$K_{jk}^{m_n} = \{C \subseteq \{1,\dots,p_n\} \setminus \{j,k\}; \ |C| \le m_n\}$$

whose cardinality is bounded by

$$|K_{jk}^{m_n}| \le B p_n^{m_n} \text{ for some } 0 < B < \infty. \tag{13.24}$$

Here, $m_n \to \infty$ $(n \to \infty)$ is growing sufficiently slowly (we will use later that $m_n = O(n^{1-b})$ as the rate of q_n in Assumption (A3)).

Lemma 13.1. *Assume (A1) (without requiring faithfulness) and $\sup_{n, j \ne k} |\rho_{n;jk}| \le M < 1$ (compare with (A4)). Then, for any $0 < \gamma < 2$,*

$$\sup_{j,k \in \{1,\dots,p_n\}} \mathbf{P}[|\hat{\rho}_{n;jk} - \rho_{n;jk}| > \gamma] \le C_1(n-2)\exp\left((n-4)\log(\frac{4-\gamma^2}{4+\gamma^2})\right),$$

for some constant $0 < C_1 < \infty$ depending on M in (A4) only.

We note that Problem 14.4 describes a related result.

Proof. The statement of Lemma 13.1 is no surprise, due to the Gaussian assumption.

We make substantial use of Hotelling (1953)'s work. Denote by $f_n(r, \rho)$ the probability density function of the sample correlation $\hat{\rho} = \hat{\rho}_{n+1;jk}$ based on $n+1$ observations and by $\rho = \rho_{n+1;jk}$ the population correlation. (It is notationally easier to work with sample size $n+1$; and we use the abbreviated notations with $\hat{\rho}$ and ρ). For $0 < \gamma \le 2$,

$$\mathbf{P}[|\hat{\rho} - \rho| > \gamma] = \mathbf{P}[\hat{\rho} < \rho - \gamma] + \mathbf{P}[\hat{\rho} > \rho + \gamma].$$

It can be shown, that $f_n(r,\rho) = f_n(-r,-\rho)$, see Hotelling (1953, p.201). This symmetry implies,

$$\mathbf{P}_\rho[\hat\rho < \rho - \gamma] = \mathbf{P}_{\tilde\rho}[\hat\rho > \tilde\rho + \gamma] \text{ with } \tilde\rho = -\rho. \tag{13.25}$$

Thus, it suffices to show that $\mathbf{P}[\hat\rho > \rho + \gamma] = \mathbf{P}_\rho[\hat\rho > \rho + \gamma]$ decays exponentially in n, uniformly for all ρ.

It has been shown (Hotelling, 1953, p.201, formula (29)), that for $-1 < \rho < 1$,

$$\mathbf{P}[\hat\rho > \rho + \gamma] \le \frac{(n-1)\Gamma(n)}{\sqrt{2\pi}\Gamma(n+\frac{1}{2})} M_0(\rho + \gamma)(1 + \frac{2}{1-|\rho|}) \tag{13.26}$$

with

$$M_0(\rho + \gamma) = \int_{\rho+\gamma}^1 (1-\rho^2)^{\frac{n}{2}}(1-x^2)^{\frac{n-3}{2}}(1-\rho x)^{-n+\frac{1}{2}}dx$$

$$= \int_{\rho+\gamma}^1 (1-\rho^2)^{\frac{\tilde n+3}{2}}(1-x^2)^{\frac{\tilde n}{2}}(1-\rho x)^{-\tilde n-\frac{5}{2}}dx \quad (\text{using } \tilde n = n-3)$$

$$\le \frac{(1-\rho^2)^{\frac{3}{2}}}{(1-|\rho|)^{\frac{5}{2}}} \int_{\rho+\gamma}^1 (\frac{\sqrt{1-\rho^2}\sqrt{1-x^2}}{1-\rho x})^{\tilde n}dx$$

$$\le \frac{(1-\rho^2)^{\frac{3}{2}}}{(1-|\rho|)^{\frac{5}{2}}} 2 \max_{\rho+\gamma \le x \le 1}(\frac{\sqrt{1-\rho^2}\sqrt{1-x^2}}{1-\rho x})^{\tilde n}. \tag{13.27}$$

We will show now that $g_\rho(x) = \frac{\sqrt{1-\rho^2}\sqrt{1-x^2}}{1-\rho x} < 1$ for all $\rho + \gamma \le x \le 1$ and $-1 < \rho < 1$ (in fact, $\rho \le 1 - \gamma$ due to the first restriction). Consider

$$\sup_{-1<\rho<1; \rho+\gamma\le x\le 1} g_\rho(x) = \sup_{-1<\rho\le 1-\gamma} \frac{\sqrt{1-\rho^2}\sqrt{1-(\rho+\gamma)^2}}{1-\rho(\rho+\gamma)}$$

$$= \frac{\sqrt{1-\frac{\gamma^2}{4}}\sqrt{1-\frac{\gamma^2}{4}}}{1-(\frac{-\gamma}{2})(\frac{\gamma}{2})} = \frac{4-\gamma^2}{4+\gamma^2} < 1 \text{ for all } 0 < \gamma \le 2. \tag{13.28}$$

Therefore, for $-1 < -M \le \rho \le M < 1$ (see assumption (A4)) and using (13.26)-(13.28) together with the fact that $\frac{\Gamma(n)}{\Gamma(n+\frac{1}{2})} \le const.$ with respect to n, we have

$$\mathbf{P}[\hat\rho > \rho + \gamma]$$

$$\le \frac{(n-1)\Gamma(n)}{\sqrt{2\pi}\Gamma(n+\frac{1}{2})} \frac{(1-\rho^2)^{\frac{3}{2}}}{(1-|\rho|)^{\frac{5}{2}}} 2(\frac{4-\gamma^2}{4+\gamma^2})^{\tilde n}(1 + \frac{2}{1-|\rho|})$$

$$\le \frac{(n-1)\Gamma(n)}{\sqrt{2\pi}\Gamma(n+\frac{1}{2})} \frac{1}{(1-M)^{\frac{5}{2}}} 2(\frac{4-\gamma^2}{4+\gamma^2})^{\tilde n}(1 + \frac{2}{1-M}) \le$$

$$\leq C_1(n-1)(\frac{4-\gamma^2}{4+\gamma^2})^{\tilde{n}} = C_1(n-1)\exp\left((n-3)\log(\frac{4-\gamma^2}{4+\gamma^2})\right),$$

where $0 < C_1 < \infty$ depends on M only, but not on ρ or γ. By invoking (13.25), the proof is complete (note that the number n in the proof corresponds to the actual sample size $n-1$). □

Lemma 13.1 can be easily extended to partial correlations, as shown by Fisher (1924), using projections for Gaussian distributions.

Proposition 13.5. *(Fisher, 1924)*
Assume (A1) (without requiring faithfulness). If the cumulative distribution function (cdf) of $\hat{\rho}_{n;j,k}$ is denoted by $F(\cdot|n,\rho_{n;j,k})$, then the cdf of the sample partial correlation $\hat{\rho}_{n;jk|C}$ with $|C| = m < n-1$ is $F[\cdot|n-m,\rho_{n;jk|C}]$. That is, the effective sample size is reduced by m.

A proof can be found in Fisher (1924); see also Anderson (1984). □

Lemma 13.1 and Proposition 13.5 yield then the following.

Corollary 13.1. *Assume (A1) (without requiring faithfulness) and the upper bound in the second requirement of (A4). Then, for $m_n < n-4$, any $0 < \gamma < 2$,*

$$\sup_{j,k\in\{1,\dots,p_n\},C\in K_{jk}^{m_n}} \mathbf{P}[|\hat{\rho}_{n;jk|C} - \rho_{n;jk|C}| > \gamma]$$

$$\leq C_1(n-2-m_n)\exp\left((n-4-m_n)\log(\frac{4-\gamma^2}{4+\gamma^2})\right),$$

for some constant $0 < C_1 < \infty$ depending on M in (A4) only.

The PC-algorithm is testing partial correlations after the z-transform $g(\rho) = 0.5\log((1+\rho)/(1-\rho))$. Denote by $Z_{n;jk|C} = g(\hat{\rho}_{n;jk|C})$ and by $z_{n;jk|C} = g(\rho_{n;jk|C})$.

Lemma 13.2. *Assume the conditions from Corollary 13.1. Define $L = 1/(1-(1+M)^2/4)$, with M as in assumption (A4). Then, for $m_n \to \infty$ $(n \to \infty)$, $m_n < n-4$, and for any $0 < \gamma < 2L$,*

$$\sup_{j,k\in\{1,\dots,p_n\},C\in K_{jk}^{m_n}} \mathbf{P}[|Z_{n;jk|C} - z_{n;jk|C}| > \gamma]$$

$$\leq O(n-m_n)\left(\exp\left\{(n-4-m_n)\log(\frac{4-(\gamma/L)^2}{4+(\gamma/L)^2})\right\} + \exp\{-C_2(n-m_n)\}\right)$$

for some constant $0 < C_2 < \infty$ depending on M in (A4) only.

Proof. A Taylor expansion of the z-transform $g(\rho) = 0.5\log((1+\rho)/(1-\rho))$ yields:

$$Z_{n;jk|C} - z_{n;jk|C} = g'(\tilde{\rho}_{n;jk|C})(\hat{\rho}_{n;jk|C} - \rho_{n;jk|C}), \tag{13.29}$$

where $|\tilde{\rho}_{n;jk|C} - \rho_{n;jk|C}| \leq |\hat{\rho}_{n;jk|C} - \rho_{n;jk|C}|$. Moreover, $g'(\rho) = 1/(1-\rho^2)$. By applying Corollary 13.1 with $\gamma = \kappa_M = (1-M)/2$ we have

$$\inf_{j,k,C \in K_{jk}^{m_n}} \mathbf{P}[|\tilde{\rho}_{n;jk|C} - \rho_{n;jk|C}| \leq \kappa_M]$$
$$> 1 - C_1(n - 2 - m_n)\exp(-C_2(n - m_n)), \tag{13.30}$$

where C_1, C_2 depend on M. Since

$$g'(\tilde{\rho}_{n;jk|C}) = \frac{1}{1 - \tilde{\rho}_{n;jk|C}^2} = \frac{1}{1 - (\rho_{n;jk|C} + (\tilde{\rho}_{n;jk|C} - \rho_{n;jk|C}))^2}$$
$$\leq \frac{1}{1 - (M + \kappa_M)^2} \text{ if } |\tilde{\rho}_{n;jk|C} - \rho_{n;jk|C}| \leq \kappa_M,$$

where we also invoke (the second part of) assumption (A4) for the last inequality. Therefore, since $\kappa_M = (1-M)/2$ yielding $1/(1 - (M + \kappa_M)^2) = L$, and using (13.30), we get

$$\inf_{j,k,C \in K_{jk}^{m_n}} \mathbf{P}[|g'(\tilde{\rho}_{n;jk|C})| \leq L]$$
$$\geq 1 - C_1(n - 2 - m_n)\exp(-C_2(n - m_n)). \tag{13.31}$$

Since $|g'(\rho)| \geq 1$ for all ρ, we obtain with (13.29):

$$\sup_{j,k,C \in K_{jk}^{m_n}} \mathbf{P}[|Z_{n;jk|C} - z_{n;jk|C}| > \gamma] \tag{13.32}$$
$$\leq \sup_{j,k,C \in K_{jk}^{m_n}} \mathbf{P}[|g'(\tilde{\rho}_{n;jk|C})| > L] + \sup_{j,k,C \in K_{jk}^{m_n}} \mathbf{P}[|\hat{\rho}_{n;jk|C} - \rho_{n;jk|C}| > \gamma/L].$$

Formula (13.32) follows from elementary probability calculations: for two random variables U, V with $|U| \geq 1$ ($|U|$ corresponding to $|g'(\tilde{\rho})|$ and $|V|$ to the difference $|\hat{\rho} - \rho|$),

$$\mathbf{P}(|UV| > \gamma) = \mathbf{P}(|UV| > \gamma, |U| > L) + \mathbf{P}(|UV| > \gamma, 1 \leq |U| \leq L)$$
$$\leq \mathbf{P}(|U| > L) + \mathbf{P}(|V| > \gamma/L).$$

The statement then follows from (13.32), (13.31) and Corollary 13.1. $\square$

13.8.2.2 Proof of Theorem 13.1

For the analysis of the PC-algorithm, it is useful to consider a more general version, the so-called $\mathrm{PC}_{\mathrm{pop}}(m)$, as described in Algorithm 14 below. By definition, the PC-

---

Algorithm 14 The $\mathrm{PC}_{\mathrm{pop}}(m)$-algorithm

1: **INPUT:** Stopping level m, vertex set V, conditional independence information
2: **OUTPUT:** Conditional independence graph G
3: Form the complete undirected graph $\tilde{G}$ on the vertex set V
4: Set $\ell = -1$; set $G = \tilde{G}$
5: **repeat**
6:     Increase $\ell \leftarrow \ell + 1$
7:     **repeat**
8:         Select a (new) ordered pair of nodes (j,k) that are adjacent in G such that $|\mathrm{adj}(G,j) \setminus \{k\}| \geq \ell$
9:         **repeat**
10:            Choose (new) $C \subseteq \mathrm{adj}(G,j) \setminus \{k\}$ with $|C| = \ell$.
11:            **if** j and k are conditionally independent given C **then**
12:                Delete edge $j - k$
13:                Denote this new graph by G.
14:            **end if**
15:        **until** edge $j - k$ is deleted or all $C \subseteq \mathrm{adj}(G,j) \setminus \{k\}$ with $|C| = \ell$ have been chosen
16:     **until** all ordered pairs of adjacent variables (j,k) such that $|\mathrm{adj}(G,j) \setminus \{k\}| \geq \ell$ and $C \subseteq \mathrm{adj}(G,j) \setminus \{k\}$ with $|C| = \ell$ have been tested for conditional independence
17: **until** $\ell = m$ or for each ordered pair of adjacent nodes (j,k): $|\mathrm{adj}(G,j) \setminus \{k\}| < \ell$.

---

algorithm in Section 13.7.1 equals the $\mathrm{PC}_{\mathrm{pop}}(m_{\mathrm{reach}})$-algorithm. There is the obvious sample version, the $\mathrm{PC}(m)$-algorithm, and the PC-algorithm in Section 13.7.2 is the same as the $\mathrm{PC}(\hat{m}_{\mathrm{reach}})$-algorithm, where $\hat{m}_{\mathrm{reach}}$ is the sample version of (13.18).

The population version $\mathrm{PC}_{\mathrm{pop}}(m_n)$-algorithm when stopped at level $m_n = m_{\mathrm{reach},n}$ constructs the true skeleton according to Proposition 13.4. Moreover, the $\mathrm{PC}_{\mathrm{pop}}(m)$-algorithm remains to be correct when using $m \geq m_{\mathrm{reach},n}$. The following Lemma extends this result to the sample $\mathrm{PC}(m)$-algorithm.

Lemma 13.3. *Assume (A1), (A2), (A3) where $0 < b \leq 1$ and (A4) where $0 < d < b/2$. Denote by $\hat{G}_n(\alpha_n, m_n)$ the estimate from the $PC(m_n)$-algorithm and by $G^{(n)}$ the true underlying conditional independence graph. Moreover, denote by $m_{\mathrm{reach},n}$ the value described in (13.18). Then, for $m_n \geq m_{\mathrm{reach},n}$, $m_n = O(n^{1-b})$ $(n \to \infty)$, there exists $\alpha_n \to 0$ $(n \to \infty)$ such that*

$$\mathbf{P}[\hat{G}_n(\alpha_n, m_n) = G^{(n)}]$$
$$= 1 - O(\exp(-Kn^{1-2d})) \to 1 \ (n \to \infty),$$

where $d > 0$ is as in (A4) and $0 < K < \infty$ depending on M in (A4).

Proof. An error occurs in the sample PC-algorithm if there is a pair of nodes j, k and a conditioning set $C \in K_{jk}^{m_n}$ (although the algorithm is typically only going through a random subset of $K_{jk}^{m_n}$) where an error event $E_{jk|C}$ occurs; $E_{jk|C}$ denotes that "an error occurred when testing partial correlation for zero at nodes j, k with conditioning set C", and the two kind of errors (false positives and false negatives) are described precisely in (13.34) below. Thus,

$$\mathbf{P}[\text{an error occurs in the PC}(m_n)\text{-algorithm}]$$

$$\leq P\left[\bigcup_{j \neq k, C \in K_{jk}^{m_n}} E_{jk|C}\right] \leq O(p_n^{m_n+2}) \sup_{j \neq k, C \in K_{jk}^{m_n}} \mathbf{P}[E_{jk|C}], \tag{13.33}$$

using that the cardinality of the set $|\{j, k, C \in K_{jk}^{m_n}\}| = O(p_n^{m_n+2})$, see also formula (13.24). Now

$$E_{jk|C} = E_{jk|C}^{I} \cup E_{jk|C}^{II}, \tag{13.34}$$

where

$$\text{type I error } E_{jk|C}^{I}: \ \sqrt{n - |k| - 3}|Z_{kl|C}| > \Phi^{-1}(1 - \alpha/2) \text{ and } z_{jk|C} = 0,$$

$$\text{type II error } E_{jk|C}^{II}: \ \sqrt{n - |k| - 3}|Z_{jk|C}| \leq \Phi^{-1}(1 - \alpha/2) \text{ and } z_{jk|C} \neq 0.$$

Choose $\alpha = \alpha_n = 2(1 - \Phi(n^{1/2}c_n/2))$, where c_n is from (A4). Then,

$$\sup_{j,k,C \in K_{jk}^{m_n}} \mathbf{P}[E_{jk|C}^{I}] = \sup_{j,k,C \in K_{jk}^{m_n}} \mathbf{P}[|Z_{jk|C} - z_{jk|C}| > (n/(n - |C| - 3))^{1/2}c_n/2]$$

$$\leq O(n - m_n)\exp(-C_3(n - m_n)c_n^2), \tag{13.35}$$

for some $0 < C_3 < \infty$ (depending on M in (A4)) using Lemma 13.2 and the fact that $\log(\frac{4-\delta^2}{4+\delta^2}) \sim -\delta^2/2$ as $\delta \to 0$. Furthermore, with the choice of $\alpha = \alpha_n$ above,

$$\sup_{j,k,C \in K_{jk}^{m_n}} \mathbf{P}[E_{jk|C}^{II}] = \sup_{j,k,C \in K_{jk}^{m_n}} \mathbf{P}[|Z_{jk|C}| \leq \sqrt{n/(n - |C| - 3)}c_n/2]$$

$$\leq \sup_{j,k,C \in K_{jk}^{m_n}} \mathbf{P}[|Z_{jk|C} - z_{jk|C}| > c_n(1 - \sqrt{n/(n - |C| - 3)}/2)],$$

because $\inf_{j,k,C \in K_{jk}^{m_n}} |z_{jk|C}| \geq c_n$ since $|g(\rho)| \geq |\rho|$ for all ρ and using assumption (A4). By invoking Lemma 13.2 we then obtain:

$$\sup_{j,k,C \in K_{jk}^{m_n}} \mathbf{P}[E_{jk|C}^{II}] \leq O(n - m_n)\exp(-C_4(n - m_n)c_n^2) \tag{13.36}$$

for some $0 < C_4 < \infty$ (depending on M in (A4)). Now, by (13.33)-(13.36) we get

$$\mathbf{P}[\text{an error occurs in the PC}(m_n)\text{-algorithm}]$$

$$\leq O(p_n^{m_n+2}(n-m_m)\exp(-C_5(n-m_n)c_n^2))$$
$$\leq O(n^{a(m_n+2)+1}\exp(-C_5(n-m_n)n^{-2d}))$$
$$= O\left(\exp\left(a(m_n+2)\log(n)+\log(n)-C_5(n^{1-2d}-m_n n^{-2d})\right)\right) = o(1),$$

because n^{1-2d} dominates all other terms in the argument of the exp-function due to the assumption in (A4) that $d < b/2$. This completes the proof. $\qquad\square$

Lemma 13.3 leaves some flexibility for choosing m_n. The PC-algorithm yields a data-dependent reached stopping level $\hat{m}_{\text{reach},n}$, that is, the sample version of (13.18).

Lemma 13.4. *Assume (A1)-(A4). Then,*

$$\mathbf{P}[\hat{m}_{\text{reach},n} = m_{\text{reach},n}] = 1 - O(\exp(-Kn^{1-2d})) \to 1 \ (n \to \infty),$$

where $d > 0$ is as in (A4) and $0 < K < \infty$ depending on M in (A4).

Proof. Consider the population algorithm $\text{PC}_{\text{pop}}(m)$: the reached stopping level satisfies $m_{\text{reach}} \in \{q_n-1, q_n\}$, see Proposition 13.4. The sample $\text{PC}(m_n)$-algorithm with stopping level in the range of $m_{\text{reach}} \leq m_n = O(n^{1-b})$ (we can choose such an m_n since $m_{\text{reach}} \in \{q_n-1, q_n\}$ and $q_n = O(n^{1-b})$) coincides with the population version on a set A having probability $P[A] = 1 - O(\exp(-Kn^{1-2d}))$, see the last formula in the proof of Lemma 13.3. Hence, on the set A, $\hat{m}_{\text{reach},n} = m_{\text{reach}} \in \{q_n-1, q_n\}$. The claim then follows from Lemma 13.3. $\qquad\square$

Lemma 13.3 and 13.4 together complete the proof of Theorem 13.1.

13.9 Back to linear models

We have seen in (13.5) a direct relation between Gaussian graphical models and regression coefficients. We will argue here that a weaker form of the faithfulness condition from Section 13.6 turns out to be very useful for variable selection in a linear model. The presentation in this section is largely based on results from Bühlmann et al. (2010).

Let $X = (X^{(1)}, \dots, X^{(p)}) \in \mathcal{X}$ be a vector of covariates with $\mathbb{E}[X] = 0$ and $\text{Cov}(X) = \Sigma_X$ and let $\varepsilon \in \mathbb{R}$ with $\mathbb{E}[\varepsilon] = 0$ and $\text{Var}(\varepsilon) = \sigma^2 > 0$ such that ε is independent of $X^{(1)}, \dots, X^{(p)}$. Consider a response $Y \in \mathbb{R}$ defined by the following random design linear model:

$$Y = \mu + \sum_{j=1}^p \beta_j^0 X^{(j)} + \varepsilon, \tag{13.37}$$

for some parameters $\mu \in \mathbb{R}$ and true parameter vector $\beta^0 = (\beta_1^0, \ldots, \beta_p^0)^T \in \mathbb{R}^p$ (we do not emphasize the "truth" of the other parameters with the notation "^{0}"). We assume:

(B1) Σ_X is strictly positive definite.

Note that assumption (B1) implies identifiability of the regression parameters from the joint distribution of (X, Y) since

$$\beta^0 = \Sigma_X^{-1} (\text{Cov}(Y, X^{(1)}) \ldots, \text{Cov}(Y, X^{(p)}))^T,$$

implicitly assuming that second moments of $X^{(j)}$'s exist. More discussion about assuming a positive definite population covariance for the covariates X is also given in Section 6.12. Furthermore, we assume that $\mathbb{E}[Y^2] < \infty$.

We consider sparse linear models where some (or most) of the β_j^0's are equal to zero. The goal is variable selection, that is to identify the active set

$$S_0 = \{j;\ \beta_j^0 \neq 0,\ j = 1, \ldots, p\}$$

based on a sample of independent observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ which are distributed as (X, Y). We denote the number of nonzero β_j^0's by $s_0 = |S_0|$.

13.9.1 Partial faithfulness

We now introduce the concept of partial faithfulness which is weaker than requiring a faithful distribution for the random variables $Y, X^{(1)}, \ldots, X^{(p)}$ in the linear model in (13.37). Partial faithfulness will allow us to identify the active set S_0 of covariates using a simplified version of the PC-Algorithm 13.

Definition 13.5. (Partial faithfulness) *Let $X \in \mathbb{R}^p$ be a random vector (e.g. covariates), and let $Y \in \mathbb{R}$ be a random variable (e.g. response). The distribution of (X, Y) is said to be (X, Y)-partially faithful if for every $j \in \{1, \ldots, p\}$:*

$$\text{Parcor}(Y, X^{(j)} | X^{(C)}) = 0 \text{ for some } C \subseteq \{1, \ldots, p\} \setminus j$$
$$\Longrightarrow \text{Parcor}(Y, X^{(j)} | X^{(\{1, \ldots, p\} \setminus j)}) = 0.$$

We remark that partial faithfulness has no direct relation to a graphical model and its definition does not require a graphical concept. It can be shown that partial faithfulness is generally a weaker condition than requiring faithfulness of a distribution in a directed acyclic graph model which relates to a linear model as in (13.37), see Bühlmann et al. (2010). Furthermore, we note that for the linear model (13.37) with

assumption (B1), $\beta_j^0 = 0$ if and only if $\mathrm{Parcor}(Y, X^{(j)} | X^{(\{1,\dots,p\} \setminus j)}) = 0$. Hence, such a model satisfies the partial faithfulness assumption if for every $j \in \{1, \dots, p\}$:

$$\mathrm{Parcor}(Y, X^{(j)} | X^{(C)}) = 0 \text{ for some } C \subseteq \{1, \dots, p\} \setminus j \implies \beta_j^0 = 0. \quad (13.38)$$

A direct consequence of partial faithfulness is as follows (Problem 13.7).

Corollary 13.2. *Consider the linear model (13.37) satisfying the* (X, Y)*-partial faithfulness condition. Then the following holds for every* $j \in \{1, \dots, p\}$:

$$\mathrm{Parcor}(Y, X^{(j)} | X^{(C)}) \neq 0 \text{ for all } C \subseteq \{1, \dots, p\} \setminus j \iff \beta_j^0 \neq 0.$$

Corollary 13.2 shows that a variable from the active set S_0 has a strong interpretation in the sense that all corresponding partial correlations are different from zero when conditioning on any subset $C \subseteq \{1, \dots, p\} \setminus j$.

13.9.1.1 Sufficient condition for partial faithfulness

We consider the linear model in (13.37) and we assume that (B1) holds. Moreover, we make a condition on the structure of β_j^0 $(j = 1, \dots, p)$: to do so, we will use the framework where the non-zero coefficients are fixed realizations from a probability distribution. A sufficient condition for partial faithfulness is (see Theorem 13.2):

(B2)   The true regression coefficients satisfy:

$$\{\beta_j^0; \ j \in S_0\} \sim f(b)db,$$

where $f(\cdot)$ denotes a density on (a subset of) $\mathbb{R}^{s_0}$ of an absolutely continuous distribution with respect to Lebesgue measure.

Assumption (B2) says that the non-zero regression coefficients are (fixed) realizations from an absolutely continuous distribution with respect to Lebesgue measure. Once the β_j^0's are realized, we fix them such that they can be considered as deterministic in the linear model (13.37). This framework is loosely related to a Bayesian formulation treating the β_j^0's as independent and identically distributed random variables from a prior distribution which is a mixture of point mass at zero (for β_j^0's with $j \notin S_0$) and a density with respect to Lebesgue measure (for β_j^0's with $j \in S_0$). Assumption (B2) is rather mild in the following sense: the regression coefficients having values zero can arise in an arbitrary way and only the non-zero coefficients are restricted to exclude adversarial cases.

Theorem 13.2. *Consider the linear model (13.37) satisfying assumptions (B1) and (B2). Then* (X, Y)*-partial faithfulness holds almost surely (with respect to the distribution generating the non-zero regression coefficients, see (B2)).*

A proof is given in Section 13.9.6. Theorem 13.2 says that failure of (X,Y)-partial faithfulness has probability zero (i.e., Lebesgue measure zero).

13.9.2 The PC-simple algorithm

We present here a slightly simplified version of the PC-Algorithm 13 in Section 13.7 for estimating the active set $S_0 = \{j;\ \beta_j^0 \neq 0\}$ in a linear model. We closely follow Section 13.7.

13.9.2.1 Population version of the PC-simple algorithm

We explore how partial faithfulness can be used for variable selection. In order to show the key ideas of the algorithm, we first assume that the population partial correlations are known. Afterwards, we consider the more realistic situation where they need to be estimated from data.

Recall that partial faithfulness for the linear model in (13.37) says:

$$\mathrm{Parcor}(Y, X^{(j)}|X^{(C)}) = 0 \text{ for some } C \subseteq \{1,\ldots,p\} \setminus \{j\} \implies \beta_j^0 = 0.$$

The easiest relation is with $C = \emptyset$:

$$\mathrm{Cor}(Y, X^{(j)}) = 0 \implies \beta_j^0 = 0, \tag{13.39}$$

showing that the active set S_0 cannot contain any j for which $\mathrm{Cor}(Y, X^{(j)}) = 0$. Hence, we can screen all marginal correlations between pairs $(Y, X^{(j)})$, $j = 1,\ldots,p$, and build a first set of candidate active variables

$$S^{[1]} = \{j;\ \mathrm{Cor}(Y, X^{(j)}) \neq 0,\ j = 1,\ldots,p\}.$$

We call this the step_1 active set or the correlation screening active set, and we know by (13.39) that

$$S_0 \subseteq S^{[1]}. \tag{13.40}$$

Such correlation screening may reduce the dimensionality of the problem by a substantial or even huge amount.

Furthermore, we can screen partial correlations of order one by using the following relation: for $j \in S^{[1]}$,

$$\mathrm{Parcor}(Y, X^{(j)}|X^{(k)}) = 0 \text{ for some } k \in S^{[1]} \setminus \{j\} \implies \beta_j^0 = 0. \tag{13.41}$$

That is, for checking whether the jth covariate remains in the model, we can additionally screen all partial correlations of order one. Note that we only consider partial correlations given variables in the step$_1$ active set $S^{[1]}$. This is similar to what is done in the PC algorithm, and it yields an important computational reduction while still allowing to eventually identify the true active set S_0 (see Algorithm 15 and Theorem 13.6). Thus, screening partial correlations of order one using (13.41) leads to a smaller set

$$S^{[2]} = \{j \in S^{[1]}; \; \text{Parcor}(Y, X^{(j)}|X^{(k)}) \neq 0 \text{ for all } k \in S^{[1]} \setminus \{j\}\} \subseteq S^{[1]}.$$

This new step$_2$ active set $S^{[2]}$ has further reduced the dimensionality of the candidate active set, and because of (13.41) we still have that $S^{[2]} \supseteq S_0$.

We can continue screening of higher-order partial correlations, resulting in a nested sequence of step$_m$ active sets

$$S^{[1]} \supseteq S^{[2]} \supseteq \ldots \supseteq S^{[m]} \supseteq \ldots \supseteq S_0. \tag{13.42}$$

A step$_m$ active set $S^{[m]}$ can be used as dimensionality reduction and any favored variable selection method could then be used for the reduced linear model with covariates corresponding to indices in $S^{[m]}$. Alternatively, we can continue the algorithm until the candidate active set does not change anymore. This leads to the PC-simple algorithm described in Algorithm 15.

Algorithm 15 The population version of the PC-simple algorithm.

1: Set $m = 1$. Do correlation screening, see (13.39), and build the step$_1$ active set
  $S^{[1]} = \{j; \; \text{Cor}(Y, X^{(j)}) \neq 0, \; j = 1, \ldots, p\}$.
2: **repeat**
3:   Increase $m \leftarrow m + 1$.
    Construct the step$_m$ active set:

$$S^{[m]} = \{j \in S^{[m-1]}; \; \text{Parcor}(Y, X^{(j)}|X^{(C)}) \neq 0$$
$$\text{for all } C \subseteq S^{[m-1]} \setminus \{j\} \text{ with } |C| = m - 1\}.$$

4: **until** $|S^{[m]}| \leq m$.

In analogy to (13.18), we denote the value m that is reached in Algorithm 15 by m_{reach}:

$$m_{\text{reach}} = \min\{m; \; |S^{[m]}| \leq m\}. \tag{13.43}$$

The PC-simple algorithm is similar to the PC-Algorithm 13 in Section 13.7. But the PC-algorithm considers all ordered pairs of variables in $(X^{(1)}, \ldots, X^{(p)}, Y)$, while we only consider pairs $(Y, X^{(j)})$, $j \in \{1, \ldots, p\}$. The reason for not considering pairs $(X^{(j)}, X^{(k)})$ is that we are only interested in associations between Y and $X^{(j)}$: we can then restrict ourselves to consider conditioning sets in the neighborhood of Y only

(instead of both neighborhoods of Y and $X^{(j)}$ as in the PC-algorithm; see also the comment appearing before Proposition 13.4).

The following result describes correctness of the PC-simple algorithm.

Proposition 13.6. *For the linear model (13.37) satisfying (B1) and (X,Y)-partial faithfulness, the population version of the PC-simple algorithm identifies the true underlying active set, i.e., $S^{[m_{\text{reach}}]} = S_0 = \{j;\ \beta_j^0 \neq 0,\ j = 1,\ldots,p\}$.*

A proof is given in Section 13.9.6.

13.9.2.2 Sample version of the PC-simple algorithm

For finite samples, we need to estimate partial correlations. We use the following shorthand notation:

$$\rho(Y,j|C) = \text{Parcor}(Y,X^{(j)}|X^{(C)}), \qquad \hat{\rho}(Y,j|C) = \widehat{\text{Parcor}}(Y,X^{(j)}|X^{(C)}),$$
$$\rho(j,k|C) = \text{Parcor}(X^{(i)},X^{(j)}|X^{(C)}), \qquad \hat{\rho}(j,k|C) = \widehat{\text{Parcor}}(X^{(i)},X^{(j)}|X^{(C)}),$$

where the "hat-versions" denote sample partial correlations. These sample partial correlations can be calculated recursively, as in (13.19): for any $k \in C$ we have

$$\hat{\rho}(Y,j|C) = \frac{\hat{\rho}(Y,j|C\setminus\{k\}) - \hat{\rho}(Y,k|C\setminus\{k\})\hat{\rho}(j,k|C\setminus\{k\})}{\sqrt{\{1 - \hat{\rho}^2(Y,k|C\setminus\{k\})\}\{1 - \hat{\rho}^2(j,k|C\setminus\{k\})\}}}.$$

As in (13.20), we test whether a partial correlation is zero using Fisher's z-transform,

$$Z(Y,j|C) = \frac{1}{2}\log\left(\frac{1 + \hat{\rho}(Y,j|C)}{1 - \hat{\rho}(Y,j|C)}\right). \tag{13.44}$$

Furthermore, exactly as in Section 13.7.2, we use the following decision rule: reject the null-hypothesis $H_0(Y,j|C):\ \rho(Y,j|C) = 0$ against the two-sided alternative $H_A(Y,j|C):\ \rho(Y,j|C) \neq 0$ if

$$\sqrt{n - |C| - 3}|Z(Y,j|C)| > \Phi^{-1}(1 - \alpha/2),$$

where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function and α is a (single testing) significance level. The Gaussian distribution serves as a reference: even in absence of a Gaussian distribution, the rule above can be seen as a thresholding operation.

The sample version of the PC-simple algorithm (Algorithm 16) is obtained by replacing the statements about $\text{Parcor}(Y,X^{(j)}|X^{(C)}) \neq 0$ (including $C = \emptyset$) in Step 3 in Algorithm 15 by

$$\sqrt{n - |C| - 3}|Z(Y,j|C)| > \Phi^{-1}(1 - \alpha/2).$$

Algorithm 16 The PC-simple algorithm

1: Run the population version of the PC-simple Algorithm 15 from Section 13.9.2.1 but replace
   in line 3 of Algorithm 15 the statement about partial correlations being non-zero by
   $\sqrt{n-|C|-3}|Z(Y,j|C)| > \Phi^{-1}(1-\alpha/2)$

The resulting estimated set of variables is denoted by $\hat{S}(\alpha) = \hat{S}^{[\hat{m}_{\text{reach}}]}(\alpha)$, where $\hat{m}_{\text{reach}}$ is the estimated version of the quantity in (13.43). The only tuning parameter α of the PC-simple algorithm is the (single testing) significance level for testing partial correlations. Asymptotic properties of the PC-simple Algorithm 16 are discussed in Section 13.9.4.

We note that the PC-simple algorithm is very different from a greedy forward (or backward) scheme: it screens many correlations or partial correlations at once and may delete many variables at once. Furthermore, it is a more sophisticated pursuit of variable screening than the marginal correlation approach in Fan and Lv (2008), described in Section 13.9.5 or its extension to low-order partial correlation screening (Wille and Bühlmann, 2006).

Since the PC-simple algorithm is a simplified version of the PC algorithm, its computational complexity is bounded above by the worst-case polynomial runtime of the PC algorithm, see (13.21) for a crude bound. In fact, we can easily use the PC-simple algorithm for sparse problems where $p \approx 100 - 5'000$, as demonstrated in Section 13.9.3.

13.9.3 Numerical results

13.9.3.1 Real data: riboflavin production by bacillus subtilis

We consider the high-dimensional real data set about riboflavin (vitamin B2) production by the Bacillus subtilis which we introduced in Section 9.2.6. There is a continuous response variable Y which measures the logarithm of the production rate of riboflavin, and there are $p = 4088$ covariates corresponding to the logarithms of expression levels of genes. We consider a genetically homogeneous sample of $n = 71$ individuals. One of the main goals is variable selection, in order to find genes which are most relevant for the production rate. We pursue this step using a linear model.

We use the PC-simple Algorithm 16 and we compare it with the variable selectors based on the Lasso, as described in Sections 2.5 and 2.6 and formula (2.10), and based on the Elastic Net discussed in Section 2.13 in Chapter 2. For the latter, we

vary the ℓ_1-penalty parameter only while keeping the ℓ_2-penalty parameter fixed at the default value from the R-package `elasticnet`. We run the PC-simple algorithm on the full data set, with various values of the tuning parameter α. Then, we compute the Lasso and Elastic Net and choose the regularization parameters such that the same number of selected variables arise as for the PC-simple method.

Table 13.3 indicates that the variable selection results of the Lasso and Elastic Net are more similar than the ones from the PC-simple method. Although the PC-simple algorithm seems to select variables in a "slightly different" way than the penalized Lasso and Elastic Net methods, we find a remarkable overlap of the few selected

α for PC-simple	selected var.	PC∩Lasso	PC∩ENet	Lasso∩ENet
0.001	3	0	0	2
0.01	4	2	1	3
0.05	5	2	1	3
0.15	6	3	2	3

Table 13.3 Variable selection for riboflavin production dataset. The columns show the number of selected variables (selected var.), the number of variables that were selected by both PC-simple and Lasso (PC∩Lasso), the number of variables that were selected by both PC-simple and Elastic Net (PC∩ENet), and the number of variables that were selected by both Lasso and Elastic Net (Lasso∩ENet).

genes among $p = 4088$ candidates.

13.9.3.2 Simulated data

We simulate data according to a Gaussian linear model as in (13.37) having p covariates with covariance matrix $\Sigma_{X;j,k} = \rho^{|j-k|}$. In order to generate values for β^0, we follow (B2): a certain number s_0 of coefficients β_j^0 have a value different from zero. The values of the nonzero β_j^0's are sampled independently from a standard normal distribution and the indices of the nonzero β_j^0's are evenly spaced between 1 and p. We consider a low- and a high-dimensional setting:

Low-dimensional: $p = 19, s_0 = 3, n = 100; \rho \in \{0, 0.3, 0.6\}$ with 1000 replicates;

High-dimensional: $p = 499, s_0 = 10, n = 100; \rho \in \{0, 0.3, 0.6\}$ with 300 replicates.

We evaluate the performance of the methods using ROC curves which measure some performance aspect of variable selection independently from the issue of choosing good tuning parameters. We compare the PC-simple algorithm with the Lasso and with the Elastic Net, and the latter two are used as described in the previous Subsection 13.9.3.1. In our PC-simple algorithm, the proposed default value for the tuning parameter is $\alpha = 0.05$: its performance is indicated by the intersection of the vertical lines and the ROC curves in Figure 13.5.

(a) Low dimensional, $\rho = 0$.

(b) High dimensional, $\rho = 0$.

(c) Low dimensional, $\rho = 0.3$.

(d) High dimensional, $\rho = 0.3$.

(e) Low dimensional, $\rho = 0.6$.

(f) High dimensional, $\rho = 0.6$.
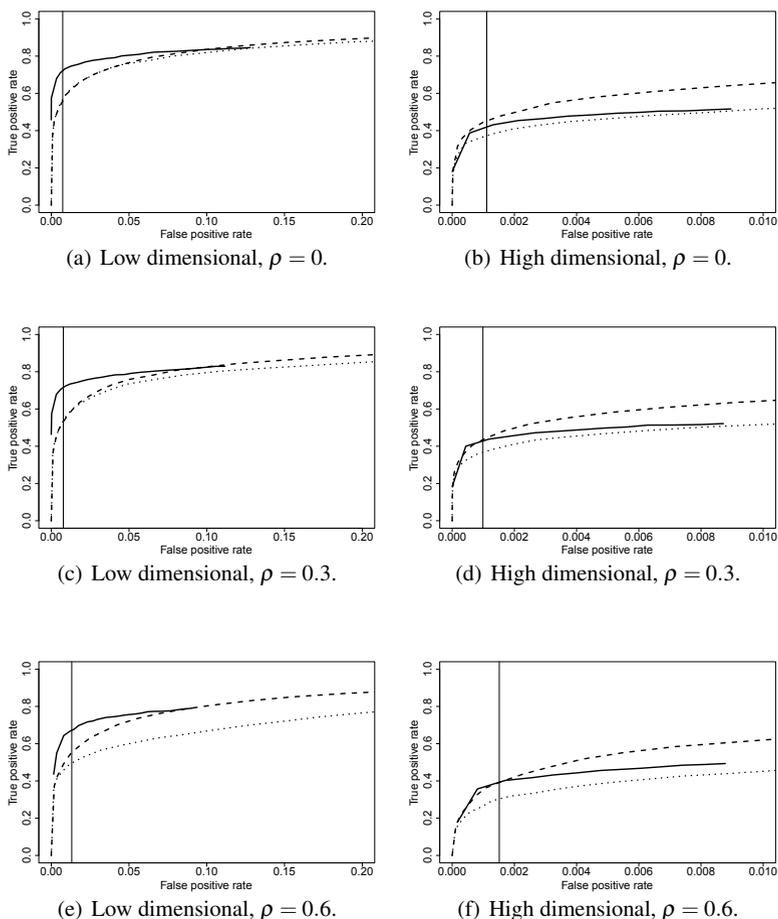
Fig. 13.5 ROC curves for the simulation study in Section 13.9.3.2. PC-simple algorithm (solid line), Lasso (dashed line) and Elastic Net (dotted line). The solid vertical lines indicate the performance of the PC-simple algorithm using the default $\alpha = 0.05$. The figure is taken from Bühlmann et al. (2010).

We first discuss the results for the low-dimensional settings (Figures 13.5(a), 13.5(c), 13.5(e)). For small false positive rates (FPRs, see equation (13.22) for the definition), the PC-simple algorithm is clearly better than Lasso or Elastic Net. If the correlation among the covariates increases, the performance of Elastic Net deteriorates, whereas the performances of PC-simple and Lasso do not vary much. When focusing on the FPR arising from the default value for α in the PC-simple method, PC-simple outperforms Lasso and Elastic Net by a large margin.

For the high-dimensional settings (Figures 13.5(b), 13.5(d), 13.5(f)), we see that for small FPRs, the difference between the methods is small. Lasso performs best, while Elastic Net is worst and PC-simple is somewhere in between. For larger FPRs, this effect becomes stronger.

13.9.4 Asymptotic results in high dimensions

13.9.4.1 Consistency of the PC-simple algorithm

We show that the PC-simple Algorithm 16 from Section 13.9.2.2 is asymptotically consistent for variable selection in high-dimensional situations with $p \gg n$, assuming certain assumptions.

We consider the linear model in (13.37). In order to simplify the asymptotic calculations, we assume a joint Gaussian distribution (see (C1) below). To capture high-dimensional behavior, we let the dimension grow as a function of sample size. That is, we consider a triangular array of observations (e.g. see (2.6)) where $p = p_n$, the regression coefficients $\beta_j^0 = \beta_{n;j}^0$ and hence the active set $S_0 = S_{0,n}$ with $s_0 = s_{0,n} = |S_{0,n}|$ and also the distribution of (X, Y) (e.g. partial correlations $\rho(\cdot, \cdot | \cdot) = \rho_n(\cdot, \cdot | \cdot)$) change with n. In the following, we denote by $\{j\}^c = \{1, \ldots, p\} \setminus j$. We make the following assumptions.

(C1)   The distribution in model (13.37)

$$(X, Y) \sim P^{(n)} = \mathcal{N}_{p_n+1}(\mu_{X,Y;n}, \Sigma_{X,Y;n})$$

is Gaussian and $P^{(n)}$ satisfies assumption (B1) and the partial faithfulness condition for all n.

(C2)   The dimension $p_n = O(n^a)$ for some $0 \le a < \infty$.

(C3)   The cardinality of the active set $s_{0,n} = |S_{0,n}| = |\{j;\ \beta_{n;j}^0 \ne 0,\ j = 1, \ldots, p_n\}|$ satisfies: $s_{0,n} = O(n^{1-b})$ for some $0 < b \le 1$.

(C4)   The partial correlations $\rho_n(Y, j | C)$ satisfy:

$$\inf \left\{ |\rho_n(Y, j | C)|;\ \rho_n(Y, j | C) \ne 0,\ j = 1, \ldots, p_n,\ C \subseteq \{j\}^c,\ |C| \le s_{0,n} \right\} \ge c_n,$$

where $c_n^{-1} = O(n^d)$ for some $0 \le d < b/2$, and b is as in (C3).

(C5)   The partial correlations $\rho_n(Y, j | C)$ and $\rho_n(j, k | C)$ satisfy:

$$(i) \quad \sup_{n, j, C \subseteq \{j\}^c, |C| \le s_{0,n}} |\rho_n(Y, j | C)| \le M < 1,$$

$$(ii) \qquad \sup_{n,j\neq k,C\subseteq\{j,k\}^c,|C|\leq s_{0,n}} |\rho_n(j,k|C)| \leq M < 1.$$

The Gaussian assumption in (C1) is not crucial: Theorem 13.6 shows that the population case does not require a Gaussian assumption and (C1) is only made to simplify asymptotic calculations. A more detailed discussion of assumptions (C1)-(C5) is given below.

Denote by $\hat{S}_n(\alpha)$ the estimated set of variables from the PC-simple Algorithm 16 in Section 13.9.2.2 with tuning parameter α.

Theorem 13.3. *Consider the linear model (13.37) and assume (C1)-(C5). Then there exists a sequence $\alpha_n \to 0$ $(n \to \infty)$ such that the PC-simple algorithm satisfies:*

$$\mathbf{P}[\hat{S}_n(\alpha_n) = S_{0,n}] = 1 - O(\exp(-Kn^{1-2d})) \to 1 \ (n \to \infty),$$

for some $0 < K < \infty$ depending on M in (C5), and $d > 0$ is as in (C4).

A proof is given in Section 13.9.6. A choice for the value of the tuning parameter leading to consistency is $\alpha_n = 2(1 - \Phi(n^{1/2}c_n/2))$. Note that this choice depends on the unknown lower bound of the partial correlations in (C4). This value α_n, although introduced as a significance level of a single test, is a tuning parameter which allows to control type I and II errors over the many tests which are pursued in the PC-simple algorithm.

13.9.4.2 Discussion of the conditions of Theorem 13.3

We have discussed in previous chapters the Lasso and versions of it for high-dimensional and computationally tractable variable selection in linear models. None of them exploit partial faithfulness and thus, it is interesting to discuss the conditions from this section with a view towards other established results.

For the Lasso, the neighborhood stability or irrepresentable condition is sufficient and essentially necessary for consistent variable selection (assuming in addition a beta-min condition on the regression coefficients), as described in Sections 2.6, 2.6.1 and also in Section 7.5.1. We recall that the neighborhood stability or the irrepresentable condition can quite easily fail to hold (e.g. in Example 13.1 below) which, due to the necessity of the condition, implies inconsistency of the Lasso for variable selection. The adaptive Lasso, described in Section 2.8, or other two-stage Lasso and thresholding procedures yield consistent variable selection under compatibility or restricted eigenvalue conditions which are weaker than the neighborhood stability or irrepresentable condition, see also Sections 7.8-7.9 and Example 13.1 below. These conditions are not directly comparable with our assumptions (C1)-(C5).

Somewhat unfortunately, assumptions (C1)-(C5) cannot be separated in terms of design and signal strength. Nevertheless, we make an attempt for an interpretation. We assume a random design where assumptions (B1) and (C5,(ii)) hold. Requiring (B1) is rather weak since we do not require explicitly any behavior of the covariance matrix $\Sigma_X = \Sigma_{X;n}$ in the sequence of distributions $P^{(n)}$ ($n \in \mathbb{N}$), except for strict positive definiteness for all n (but not an explicit bound on the minimal eigenvalue). On the other hand, (C5,(ii)) excludes perfect collinearity where the fixed upper bound on partial correlations is placing some additional restrictions on the design. Furthermore, the linear model of $Y|X$ involves regression coefficients which are required to satisfy certain conditions as follows. The partial faithfulness condition follows from Theorem 13.2 if we assume (B2) from Section 13.9.1 for every n. Dependence relations among covariates enter implicitly in restricting the regression coefficients via assumptions (C4) and (C5,(i)). Assumption (C4) imposes a constraint regarding the detectability of small non-zero partial correlations. The condition is slightly more restrictive than requiring the order of the detection limit $\sqrt{s_n \log(p_n)/n}$ for the Lasso given in e.g. (2.23) (see also the discussion of (A1)-(A4) in Section 13.8). Assumption (C2) allows for an arbitrary polynomial growth of dimension as a function of sample size, i.e., high-dimensionality, while (C3) is a sparseness assumption in terms of the number of effective variables. Both (C2) and (C3) are standard assumptions in high-dimensional asymptotics. If the dimension p is fixed (with fixed distribution P in the linear model), (C2), (C3) and (C4) hold automatically, and (C1) and (C5) remain as the only conditions. For the high-dimensional case, a simple modification of Example 13.1 is presented below where conditions (C1)-(C5) hold.

It is easy to construct examples where the Lasso fails to be consistent while the PC-simple algorithm recovers the true set of variables, as shown by the following example.

Example 13.1. Consider a Gaussian linear model as in (13.37) with

$$p = 4,\ s_0 = 3,\ \sigma^2 = 1,\ \mu = (0,\dots,0)^T,\ \mathbb{E}[X^{(j)}] = 0 \text{ for all } j,$$

$$\Sigma_X = \begin{pmatrix} 1 & \rho_1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_1 & \rho_2 \\ \rho_1 & \rho_1 & 1 & \rho_2 \\ \rho_2 & \rho_2 & \rho_2 & 1 \end{pmatrix},\ \rho_1 = -0.4,\ \rho_2 = 0.2,$$

$$\beta_1, \beta_2, \beta_3 \text{ fixed i.i.d. realizations from } \mathcal{N}(0,1),\ \beta_4 = 0.$$

It can be shown that the irrepresentable condition, see formula (2.20), fails to hold (Problem 13.8). Hence, the Lasso is inconsistent for variable selection in this model. On the other hand, (C1) holds because of Theorem 13.2, and also (C5) is true. These are all the conditions for the PC-simple algorithm for a fixed distribution P. Hence, the PC-simple algorithm is consistent for variable selection. It should be noted though that the adaptive Lasso is also consistent for this example.

For the high-dimensional case, we can modify the example as follows. Consider $s_0 = 3$ active variables, with design and coefficients as in Example 13.1, and $p - s_0$ noise covariates which are independent from the 3 active variables (and p satisfies (C2) and the design satisfies (B1) and (C5,(ii))): then, all our assumptions (C1)-(C5) hold while the Lasso is not consistent for this modification of Example 13.1.

13.9.5 Correlation screening (sure independence screening)

A very simple variable screening procedure for a linear model is given by selecting all variables having sufficiently large absolute marginal correlation with the response:

$$\hat{S}_{\mathrm{corr-scr}}(\tau) = \{j; \, |\hat{\rho}(Y, j)| > \tau\}, \qquad (13.45)$$

where $0 < \tau \leq 1$ is a threshold parameter. The procedure is useful if

$$\mathbf{P}[\hat{S}_{\mathrm{corr-scr}}(\tau) \supseteq S_0] \to 1 \ (n \to \infty).$$

Fan and Lv (2008) use the terminology "sure independence screening" (SIS) for the correlation procedure in (13.45). They argue that SIS can achieve the screening property in high-dimensional scenarios. Here, we will derive another justification, based on partial faithfulness, why correlation screening works at least asymptotically. Thereby, we note that the estimated version of correlation screening in (13.40) and (13.45) are equivalent.

13.9.5.1 Asymptotic behavior of correlation screening

For correlation screening, see formula (13.40), we do not require any sparsity. We define:

(D1)   The marginal correlations $\rho_n(Y, j)$ satisfy:

$$\inf\left\{|\rho_n(Y, j)|; \, \rho_n(Y, j) \neq 0, \, j = 1, \ldots, p_n\right\} \geq c_n,$$

where $c_n^{-1} = O(n^d)$ with $0 \leq d < 1/2$.

(D2)   The marginal correlations $\rho_n(Y, j)$ satisfy:

$$\sup_{n, j} |\rho_n(Y, j)| \leq M < 1.$$

Denote by $\hat{S}_n^{[1]}(\alpha)$ the correlation screening active set estimated from data using significance level α, i.e., the first step in the sample version of the PC-simple Algorithm 16.

Theorem 13.4. *Consider the linear model (13.37) and assume (C1), (C2), (D1) and (D2). Then there exists a sequence $\alpha_n \to 0$ $(n \to \infty)$ such that:*

$$\mathbf{P}[\hat{S}_n^{[1]}(\alpha_n) \supseteq S_{0,n}] = 1 - O(\exp(-Kn^{1-2d})) \to 1 \ (n \to \infty),$$

for some $0 < K < \infty$ depending on M in (D2), and $d > 0$ is as in (D1).

A proof is given in Section 13.9.6. A possible choice of the regularization parameter is $\alpha_n = 2(1 - \Phi(n^{1/2}c_n/2))$ where c_n is the unknown lower bound of non-zero correlations in assumption (D1). As pointed out above, we do not make any assumptions on sparsity. However, for non-sparse problems, many correlations may be non-zero, preventing an effective dimension reduction. In such problems, $\hat{S}^{[1]}$ can still be large, for example almost as large as the full set $\{1, 2, \ldots, p\}$.

Fan and Lv (2008) have shown that correlation (or sure independence) screening is overestimating the active set S_0, as stated in Theorem 13.4, assuming rather restrictive conditions on the covariance Σ_X. Theorem 13.4 demonstrates that this result also holds under very different assumptions on Σ_X when partial faithfulness is assumed in addition.

13.9.6 Proofs

13.9.6.1 Proof of Theorem 13.2

Consider the linear model (13.37) satisfying assumptions (B1) and (B2). In order to prove that the partial faithfulness assumption holds almost surely, it suffices to show that the following is true for all $j \in \{1, \ldots, p\}$:

$$\beta_j \neq 0 \implies \mathrm{Parcor}(Y, X^{(j)} | X^{(C)}) \neq 0 \ \ a.s. \text{ for all } C \subseteq \{j\}^c$$

(a.s. is with respect to the distribution generating the β_j's).

Thus, let $j \in \{1, \ldots, p\}$ such that $\beta_j \neq 0$, and let $C \subseteq \{j\}^c$. Recall that $\mathrm{Parcor}(Y, X^{(j)} | X^{(C)}) = 0$ if and only if the partial covariance $\mathrm{Parcov}(Y, X^{(j)} | X^{(C)})$ between Y and $X^{(j)}$ given $X^{(C)}$ equals zero (cf. Anderson (1984, page 37, definition 2.5.2)). Partial covariances can be computed using a recursive formula given in e.g. Anderson (1984, page 43, equation (26)). This formula shows that the partial covariance is linear in its arguments, and that $\mathrm{Parcov}(\varepsilon, X^{(j)} | X^{(C)}) = 0$ for all $j \in \{1, \ldots, p\}$ and $C \subseteq \{j\}^c$. Hence,

$$\text{Parcov}(Y, X^{(j)} | X^{(C)}) = \text{Parcov}(\mu + \sum_{r=1}^{p} \beta_r X^{(r)} + \varepsilon, X^{(j)} | X^{(C)})$$

$$= \sum_{r=1}^{p} \beta_r \text{Parcov}(X^{(r)}, X^{(j)} | X^{(C)})$$

$$= \beta_j \text{Parcov}(X^{(j)}, X^{(j)} | X^{(C)}) + \sum_{r=1, r \neq j}^{p} \beta_r \text{Parcov}(X^{(r)}, X^{(j)} | X^{(C)}).$$

Since $\beta_j \neq 0$ by assumption, and since $\text{Parcov}(X^{(j)}, X^{(j)} | X^{(C)}) \neq 0$ by assumption (B1), the only way for $\text{Parcov}(Y, X^{(j)} | X^{(C)})$ to equal zero is if there is a special parameter constellation of the β_r's, such that

$$\sum_{r=1, r \neq j}^{p} \beta_r \text{Parcov}(X^{(r)}, X^{(j)} | X^{(C)}) = -\beta_j \text{Parcov}(X^{(j)}, X^{(j)} | X^{(C)}). \quad (13.46)$$

But such a parameter constellation has Lebesgue measure zero under assumption (B2) (for fixed distribution of X, the probability that the sampled β_r's fulfill (13.46) is zero). □

13.9.6.2 Proof of Proposition 13.6

By partial faithfulness and equation (13.42), $S_0 \subseteq S^{[m_{\text{reach}}]}$. Hence, we only need to show that S_0 is not a strict subset of $S^{[m_{\text{reach}}]}$. We do this using contraposition. Thus, suppose that $S_0 \subset S^{[m_{\text{reach}}]}$ strictly. Then there exists a $j \in S^{[m_{\text{reach}}]}$ such that $j \notin S_0$. Fix such an index j. Since $j \in S^{[m_{\text{reach}}]}$, we know that

$$\text{Parcor}(Y, X^{(j)} | X^{(C)}) \neq 0 \text{ for all } C \subseteq S^{[m_{\text{reach}}-1]} \setminus \{j\} \text{ with } |C| \leq m_{\text{reach}} - 1. \ (13.47)$$

The argument for (13.47) is as follows. The statement for sets C with $|C| = m_{\text{reach}} - 1$ is due to the definition of iteration m_{reach} of the PC-simple algorithm. Sets C with lower cardinality are considered in previous iterations of the algorithm, and since $S^{[1]} \supseteq S^{[2]} \supseteq \ldots$, all subsets $C \subseteq S^{[m_{\text{reach}}-1]}$ with $|C| \leq m_{\text{reach}} - 1$ are considered and hence, (13.47) holds.

We now show that we can take $C = S_0$ in (13.47). First, note that the supposition $S_0 \subset S^{[m_{\text{reach}}]}$ and our choice of j imply that

$$S_0 \subseteq S^{[m_{\text{reach}}]} \setminus \{j\} \subseteq S^{[m_{\text{reach}}-1]} \setminus \{j\}.$$

Moreover, $S_0 \subset S^{[m_{\text{reach}}]}$ (strictly) implies that $|S_0| \leq |S^{[m_{\text{reach}}]}| - 1$. Combining this with $|S^{[m_{\text{reach}}]}| \leq m_{\text{reach}}$ (see the definition of m_{reach} in (13.43)), yields that $|S_0| \leq m_{\text{reach}} - 1$. Hence, we can indeed take $C = S_0$ in (13.47), yielding that $\text{Parcor}(Y, X^{(j)} | X^{(S_0)}) \neq 0$.

On the other hand, $j \notin S_0$ implies that $\beta_j = 0$, and hence $\text{Parcor}(Y, X^{(j)}|X^{(S_0)}) = 0$. This is a contradiction, and hence S_0 cannot be a strict subset of $S^{[m_{\text{reach}}]}$. $\square$

13.9.6.3 Proof of Theorem 13.3

A first main step is to show that the population version of the PC-simple algorithm infers the true underlying active set $S_{0,n}$, assuming partial faithfulness. We formulated this step in Proposition 13.6 as a separate result, and its proof is given above.

The arguments for controlling the estimation error due to a finite sample size are very similar to the ones used in the proof of Theorem 13.1. We proceed in two steps.

Analysis of partial correlations.
An exponential inequality for estimating partial correlations up to order $m_n = o(n)$ follows from Lemma 13.2. We use the following notation: $K_j^{m_n} = \{C \subseteq \{1, \ldots, p_n\} \setminus \{j\}; |C| \leq m_n\}$ $(j = 1, \ldots, p_n)$, $Z_n(Y, j|C) = g(\hat{\rho}_n(Y, j|C))$ and $z_n(Y, j|C) = g(\rho_n(Y, j|C))$, where $g(\rho) = \frac{1}{2}\log(\frac{1+\rho}{1-\rho})$ is Fisher's Z-transform. Lemma 13.2 now reads as follows.

Lemma 13.5. *Suppose that the Gaussian assumption from (C1) and condition (C5) hold. Define $L = 1/(1 - (1+M)^2/4)$, with M as in assumption (C5). Then, for $m_n < n - 4$ and $0 < \gamma < 2L$,*

$$\sup_{C \in K_j^{m_n}, 1 \leq j \leq p_n} \mathbf{P}[|Z_n(Y, j|C) - z_n(Y, j|C)| > \gamma]$$

$$\leq O(n) \left(\exp\left\{(n - 4 - m_n)\log(\frac{4 - (\gamma/L)^2}{4 + (\gamma/L)^2})\right\} + \exp\{-C_2(n - m_n)\}\right)$$

for some constant $0 < C_2 < \infty$ depending on M in (C5).

Analysis of the PC-simple algorithm.
First, we consider a version of the PC-simple algorithm that stops after a fixed (i.e., non-random) number of m iterations (and if $m \geq \hat{m}_{\text{reach}}$, where $\hat{m}_{\text{reach}}$ is the estimation analogue of (13.43), we set $\hat{S}^{[m]} = \hat{S}^{[\hat{m}_{\text{reach}}]}$). We denote this version by PC-simple(m) and the resulting estimate of the active set by $\hat{S}(\alpha, m)$. This construction is analogous to the PC(m)-algorithm whose population version is described in Algorithm 14.

Lemma 13.6. *Assume (C1)-(C5). Then, for m_n satisfying $m_n \geq m_{\text{reach},n}$ (see (13.43)) and $m_n = O(n^{1-b})$ (with b as in (C3)), there exists a sequence $\alpha_n \to 0$ such that*

$$\mathbf{P}[\hat{S}_n(\alpha_n, m_n) = S_{0,n}] = 1 - O(\exp(-Kn^{1-2d})) \to 1 \ (n \to \infty),$$

for some $0 < K < \infty$ depending on M in (C5), and $d > 0$ is as in (C4).

A concrete choice of α_n is $\alpha_n = 2(1 - \Phi(n^{1/2}c_n/2))$, where c_n is the lower bound from (C4) (which is typically unknown).

Proof. Obviously, the population version of the PC-simple(m_n) algorithm is correct for $m_n \geq m_{reach,n}$, see Theorem 13.6. An error can occur in the PC-simple(m_n) algorithm if there exists a covariate $X^{(j)}$ and a conditioning set $C \in K_j^{m_n}$ (although the algorithm is typically only going through random subsets of $K_j^{m_n}$) where an error event $E_{j|C}$ occurs; $E_{j|C}$ denotes the event that "an error occurred when testing $\rho_n(Y, j|C) = 0$" (either a false positive or false negative decision, see (13.49) and afterwards). Thus,

$$\mathbf{P}[\text{an error occurs in the PC-simple}(m_n)\text{-algorithm}]$$
$$\leq \mathbf{P}[\bigcup_{C \in K_j^{m_n}, 1 \leq j \leq p_n} E_{j|C}] \leq O(p_n^{m_n+1}) \sup_{C \in K_j^{m_n}, j} \mathbf{P}[E_{j|C}], \qquad (13.48)$$

using that the cardinality of the set $\{C; \ C \in K_j^{m_n}, \ j = 1, \ldots, p_n\}$, indexing the union in (13.48), is bounded by $O(p_n^{m_n+1})$. Now, let

$$E_{j|C} = E_{j|C}^I \cup E_{j|C}^{II}, \qquad (13.49)$$

where

type I error $E_{j|C}^I$: $\ \sqrt{n - |C| - 3}|Z_n(Y, j|C)| > \Phi^{-1}(1 - \alpha/2)$ and $z_n(Y, j|C) = 0$,

type II error $E_{j|C}^{II}$: $\ \sqrt{n - |C| - 3}|Z_n(Y, j|C)| \leq \Phi^{-1}(1 - \alpha/2)$ and $z_n(Y, j|C) \neq 0$.

Choose $\alpha = \alpha_n = 2(1 - \Phi(n^{1/2}c_n/2))$, where c_n is from (C4). Then,

$$\sup_{C \in K_j^{m_n}, 1 \leq j \leq p_n} \mathbf{P}[E_{j|C}^I]$$
$$= \sup_{C \in K_j^{m_n}, j} \mathbf{P}\left(|Z_n(Y, j|C) - z_n(Y, j|C)| > \sqrt{n/(n - |C| - 3)}c_n/2\right)$$
$$\leq O(n) \exp(-C_3(n - m_n)c_n^2), \qquad (13.50)$$

for some $0 < C_3 < \infty$ depending on M in (C5), using Lemma 13.5 and the fact that $\log(\frac{4-\delta^2}{4+\delta^2}) \leq -\delta^2/2$ for $0 < \delta < 2$. Furthermore, with the choice of $\alpha = \alpha_n$ above,

$$\sup_{C \in K_j^{m_n}, 1 \leq j \leq p_n} \mathbf{P}[E_{j|C}^{II}] = \sup_{C \in K_j^{m_n}, j} \mathbf{P}\left(|Z_n(Y, j|C)| \leq \sqrt{n/(n - |C| - 3)}c_n/2\right)$$
$$\leq \sup_{C \in K_j^{m_n}, j} \mathbf{P}\left(|Z_n(Y, j|C) - z_n(Y, j|C)| > c_n(1 - \sqrt{n/(n - |C| - 3)}/2)\right),$$

because $\inf_{C \in K_j^{m_n}, j} \{|z_n(Y, j|C)|; \ z_n(Y, j|C) \neq 0\} \geq c_n$ since $|g(\rho)| = |\frac{1}{2} \log(\frac{1+\rho}{1-\rho})| \geq$ $|\rho|$ for all ρ and using assumption (C4). This shows the crucial role of assumption (C4) in controlling the type II error. By invoking Lemma 13.5 we then obtain:

$$\sup_{C \in K_j^{m_n}, j} \mathbf{P}[E_{j|C}^{II}] \leq O(n) \exp(-C_4(n - m_n)c_n^2) \tag{13.51}$$

for some $0 < C_4 < \infty$ depending on M in (C5). Now, by (13.48)-(13.51) we get

$$\mathbf{P}[\text{an error occurs in the PC-simple}(m_n)\text{-algorithm}]$$
$$\leq O(p_n^{m_n+1} n \exp(-C_5(n - m_n)c_n^2))$$
$$\leq O(n^{a(m_n+1)+1} \exp(-C_5(n - m_n)n^{-2d}))$$
$$= O\left(\exp\left(a(m_n + 1)\log(n) + \log(n) - C_5(n^{1-2d} - m_n n^{-2d})\right)\right)$$
$$= O(\exp(-Kn^{1-2d})),$$

for some $0 < K < \infty$ depending on M in (C5), because n^{1-2d} dominates all other terms in the argument of the exp-function, due to $m_n = O(n^{1-b})$ and the assumption in (C4) that $d < b/2$. This completes the proof.   □

Lemma 13.6 leaves some flexibility for choosing m_n. The PC-simple algorithm yields a data-dependent stopping level $\hat{m}_{\text{reach},n}$, that is, the sample version of (13.43).

Lemma 13.7. *Assume (C1)-(C5). Then,*

$$\mathbf{P}[\hat{m}_{\text{reach},n} = m_{\text{reach},n}] = 1 - O(\exp(-Kn^{1-2d})) \to 1 \ (n \to \infty)$$

for some $0 < K < \infty$ depending on M in (C5), and $d > 0$ is as in (C4).

Proof. Consider the population version of the PC-simple algorithm, with stopping level m_{reach} as defined in (13.43). Note that $m_{\text{reach}} = m_{\text{reach},n} = O(n^{1-b})$ under assumption (C3). The sample PC-simple(m_n) algorithm with stopping level in the range of $m_{\text{reach}} \leq m_n = O(n^{1-b})$, coincides with the population version on a set A having probability $P[A] = 1 - O(\exp(-Kn^{1-2d}))$, see the last formula in the proof of Lemma 13.6. Hence, on the set A, $\hat{m}_{\text{reach},n} = m_{\text{reach}}$.   □

Lemma 13.6 and 13.7 together complete the proof of Theorem 13.3.

13.9.6.4 Proof of Theorem 13.4

By definition, $S_{0,n} \subseteq S^{[1]}$, where $S^{[1]}$ is the set of variables from correlation screening.

Denote by $Z_n(Y, j)$ the quantity as in (13.20) with $C = \emptyset$ and by $z_n(Y, j)$ its population analogue, i.e., the z-transformed population correlation. An error occurs

when screening the jth variable if $Z_n(Y, j)$ has been tested to be zero but in fact $z_n(Y, j) \neq 0$. We denote such an error event by E_j^{II}. Note that

$$\sup_{1 \leq j \leq p_n} \mathbf{P}[E_j^{II}] \leq O(n) \exp(-C_1 n c_n^2),$$

for some $0 < C_1 < \infty$ depending on M in (D2), see formula (13.51) above (we do not use any sparsity assumption for this derivation; we do invoke (D1) which requires a lower bound on non-zero marginal correlations). Thus, the probability of an error occurring in the correlation screening procedure is bounded by

$$\mathbf{P}[\cup_{1 \leq j \leq p_n} E_j^{II}] = O(p_n n) \exp(-C_1 n c_n^2) = O(\exp((1+a) \log(n) - C_1 n^{1-2d}))$$
$$= O(\exp(-C_2 n^{1-2d}))$$

for some $0 < C_2 < \infty$ depending on M in (D2) (since $0 \leq d < 1/2$). This completes the proof.   □

Problems

13.1. Consider the regressions in (13.4). Discuss that it is always possible to find such a representation, assuming that $X^{(1)}, \ldots, X^{(p)}$ have a multivariate Gaussian distribution. Some textbooks (e.g. Section 2.5 in Anderson (1984)) may be used to recall the basic properties of conditional distributions from a multivariate Gaussian.

13.2. For the negative log-likelihood of a multivariate Gaussian distribution, derive formula (13.6).

13.3. Ising model
Show that (13.15) holds.

13.4. Show that (13.16) holds, establishing the equivalence of non-zero parameters in the Ising model and non-zero coefficients in logistic regressions.

13.5. Faithfulness
Show that formula (13.17) holds, assuming the faithfulness condition.

13.6. Give a proof of Proposition 13.3.

13.7. Derive the statement in Corollary 13.2.

13.8. Show that the irrepresentable condition fails to hold in Example 13.1.

Chapter 14
Probability and moment inequalities

Abstract The chapter is a mini-guide to empirical process theory, in particular probability and moment inequalities for the empirical process indexed by functions. We give an essentially complete (but often rather compact) treatment of the results we need in the previous chapters.

14.1 Organization of this chapter

This chapter contains some important probability and moment inequalities for the maximal difference between averages and expectations. We start out with summarizing a few elementary but useful results. We then reprove Bernstein's inequality as it was given in Bennet (1962). Section 14.4 reproves Hoeffding's inequality (Hoeffding, 1963). In Section 14.5, we provide the moment and probability inequalities for the maximum of a finite number, say p, averages. Section 14.6 extends these inequalities to the case of arbitrary many averages: it presents a concentration inequality of Bousquet (2002), of Massart (2000a), and concentration inequalities for sub-Gaussian random variables. Section 14.7 reviews some symmetrization and contraction inequalities. These are applied in Section 14.8 to obtain concentration results for Lipschitz loss functions. In Section 14.9, we discuss some issues related to least squares estimation with random design. Symmetrization moreover allows one to derive bounds assuming only lower order moments, as we shall show in Section 14.10. Section 14.11 invokes entropy arguments to arrive at the increments of the empirical process for the sub-Gaussian case. In Section 14.12, we present some concrete entropy results for illustrating the bounds of the preceding section.

14.2 Some simple results for a single random variable

14.2.1 Sub-exponential random variables

A random variable X is called sub-exponential if, for some constants K and σ_0,

$$2K^2 (\mathbb{E} e^{|X|/K} - 1 - \mathbb{E}|X|/K) \leq \sigma_0^2. \tag{14.1}$$

Our first result is a preliminary lemma that can be used to prove Bernstein's inequality (see Lemma 14.9), and that moreover will help the reader to appreciate Bousquet's inequality (see Theorem 14.1).

Lemma 14.1. *Let $X \in \mathbb{R}$ be a random variable with $\mathbb{E}X = 0$. Then it holds that*

$$\log \mathbb{E}\exp[X] \leq \mathbb{E} e^{|X|} - 1 - \mathbb{E}|X|.$$

Proof. For any $c > 0$,

$$\exp[X - c] - 1 \leq \frac{\exp[X]}{1+c} - 1 = \frac{e^X - 1 - X + X - c}{1+c} \leq \frac{e^{|X|} - 1 - |X| + X - c}{1+c}.$$

Let

$$c = \mathbb{E} e^{|X|} - 1 - \mathbb{E}|X|.$$

Hence, since $\mathbb{E}X = 0$,

$$\mathbb{E}\exp[X - c] - 1 \leq \frac{\mathbb{E} e^{|X|} - 1 - \mathbb{E}|X| - c}{1+c} = 0.$$

$\square$

Note that (14.1) implies that for all $m \geq 2$,

$$\mathbb{E}|X|^m \leq \frac{m!}{2} K^{m-2} \sigma_0^2.$$

Conversely, we have:

Lemma 14.2. *Suppose that for some positive constants K and σ_0, and for all $m \in \{2, 3, \ldots\}$,*

$$\mathbb{E}|X|^m \leq \frac{m!}{2} K^{m-2} \sigma_0^2.$$

Then

$$2[2K]^2 \left(\mathbb{E}\exp[|X|/[2K]] - 1 - \mathbb{E}|X|/[2K] \right) \leq 2\sigma_0^2.$$

Proof. Let $L = 2K$. Then

$$2L^2(\mathbb{E}\exp[|X|/L] - 1 - \mathbb{E}|X|/L) = \sum_{m=2}^{\infty} \frac{2}{m!} \frac{\mathbb{E}|X|^m}{L^{m-2}}$$

$$\leq \sum_{m=2}^{\infty} \frac{K^{m-2}\sigma_0^2}{L^{m-2}} = \frac{\sigma_0^2}{1 - K/L} = 2\sigma_0^2.$$

$\square$

14.2.2 Sub-Gaussian random variables

A random variable X is called sub-Gaussian if, for some constants K and σ_0,

$$K^2(\mathbb{E}e^{|X|^2/K^2} - 1) \leq \sigma_0^2. \tag{14.2}$$

We relate (14.2) to the behavior of moment generating function of X.

Lemma 14.3. *Let $X \in \mathbb{R}$ be a random variable with $\mathbb{E}X = 0$ and with, for certain positive constants K and σ_0,*

$$K^2(\mathbb{E}e^{|X|^2/K^2} - 1) \leq \sigma_0^2.$$

Then for all L,
$$\mathbb{E}e^{X/L} \leq \exp[2(K^2 + \sigma_0^2)/L^2].$$

Proof. Take $\beta := 1 + \frac{\sigma_0^2}{K^2}$. By Chebyshev's inequality, we have for all $t > 0$,

$$\mathbf{P}(|X| \geq t) \leq \beta \exp[-t^2/K^2].$$

Hence, for $m \in \{2, 3, \ldots\}$,

$$\mathbb{E}|X|^m \leq \beta \int_0^{\infty} \exp[-t^{\frac{2}{m}}/K^2] dt$$

$$= \beta K^m \Gamma\left(\frac{m}{2} + 1\right) \leq \beta^m K^m \Gamma\left(\frac{m}{2} + 1\right),$$

where in the last inequality, we used $\beta \geq 1$. So, for $L \geq 0$,

$$\mathbb{E}e^{X/L} = 1 + \sum_{m=2}^{\infty} \frac{1}{m!} \mathbb{E}X^m/L^m$$

$$\leq 1 + \sum_{m=2}^{\infty} \frac{\Gamma(\frac{m}{2}+1)}{\Gamma(m+1)} \beta^m K^m/L^m \leq 1 + \sum_{m=2}^{\infty} \frac{\beta^m K^m/L^m}{\Gamma(\frac{m}{2}+1)}$$

$$= 1 + \sum_{m=1}^{\infty} \frac{(\beta K^2/L^2)^m}{\Gamma(m+1)} + \sum_{m=1}^{\infty} \frac{(\beta K^2/L^2)^{m+\frac{1}{2}}}{\Gamma(m+\frac{3}{2})}$$

$$\leq 1 + \sum_{m=1}^{\infty} \frac{(\beta K^2/L^2)^m(1+\beta K/L)}{\Gamma(m+1)} = 1 + (1+\beta K/L)(\exp[\beta K^2/L^2]-1).$$

Now, invoke that for all $x > 0$, $x \leq e^{x^2}$, which implies

$$1 + (1+x)(e^{x^2}-1) \leq e^{2x^2}.$$

The result for $L \leq 0$ follows by replacing X by $-X$. □

A random variable X has sub-Gaussian tails if for some non-negative constants a and K,

$$\mathbf{P}(|X| \geq t) \leq 2\exp[-t^2/K^2], \ t > a.$$

By Chebyshev's inequality, sub-Gaussian random variables have sub-Gaussian tail behavior. The next lemma shows that the converse is also true.

We invoke the notation

$$x_+ := x\mathbb{l}\{x > 0\},$$

and stress here that $(x_+)^2 = x^2\mathbb{l}\{x > 0\} \neq x^2$.

Lemma 14.4. *Suppose that for some constants $K > 0$ and $a \geq 0$ and for all $t > a$,*

$$\mathbf{P}(|X| \geq t) \leq 2\exp[-2t^2/K^2].$$

Then

$$\mathbb{E}\exp\left[\frac{(|X|-a)_+}{K}\right]^2 \leq 1 + 2\exp[-2a^2/K^2].$$

Proof. This follows from

$$\mathbb{E}\exp\left[\frac{(|X|-a)_+}{K}\right]^2 = \int_0^{\infty} \mathbf{P}\left(\exp\left[\frac{(|X|-a)_+}{K}\right]^2 \geq u\right) du$$

$$\leq 1 + \int_1^{\infty} \mathbf{P}\left(\exp\left[\frac{(|X|-a)_+}{K}\right]^2 \geq u\right) du$$

$$\leq 1 + \int_1^{\infty} \mathbf{P}(|X| \geq a + \sqrt{K^2\log u}) du \leq 1 + 2\int_1^{\infty} \exp[-2(a+\sqrt{K^2\log u})^2/K^2]du$$

$$\leq 1 + 2\exp[-2a^2/K^2]\int_1^{\infty} \exp[-2\log u] = 1 + 2\exp[-2a^2/K^2].$$

□

It is quite clear that if X is sub-Gaussian, then $Y := X^2$ is sub-exponential. We present the next result for later reference.

Lemma 14.5. *Suppose that for some positive constants K and σ_0,*

$$K^2(\mathbb{E}\exp[X^2/K^2] - 1) \le \sigma_0^2.$$

Let $Y := X^2$. Then for all $m \ge 2$,

$$\mathbb{E}|Y - \mathbb{E}Y|^m \le \frac{m!}{2}(2K^2)^{m-2}(2K\sigma_0)^2.$$

Proof. Clearly,

$$\mathbb{E}\exp[Y/K^2] - 1 \le \sigma_0^2/K^2.$$

Therefore

$$\mathbb{E}|Y|^m/K^{2m} \le m!(\mathbb{E}\exp[Y^2/K^2] - 1) \le m!\sigma_0^2/K^2.$$

Hence

$$\mathbb{E}|Y|^m \le m!\sigma_0^2 K^{2m-2}.$$

But then

$$\mathbb{E}|Y - \mathbb{E}Y|^m \le 2^{m-1}\mathbb{E}|Y|^m \le m!\sigma_0^2 2^{m-1}K^{2m-2}$$

$$= \frac{m!}{2}(2K\sigma_0)^2(2K^2)^{m-2}.$$

$\square$

14.2.3 Jensen's inequality for partly concave functions

We will derive m-th moment inequalities ($m \ge 1$) from exponential moments (see e.g. Lemma 14.7, Lemma 14.12, Corollary 14.1, and Lemma 14.14). For this purpose, we present the following lemma.

Lemma 14.6. *(Jensen's inequality for partly concave functions) Let X be a real-valued random variable, and let g be an increasing function on $[0,\infty)$, which is concave on $[c,\infty)$ for some $c \ge 0$. Then*

$$\mathbb{E}g(|X|) \le g\left[\mathbb{E}|X| + c\mathbf{P}(|X| < c)\right].$$

Proof. We have

$$\mathbb{E}g(|X|) = \mathbb{E}g(|X|)\mathrm{l}\{|X| \ge c\} + \mathbb{E}g(|X|)\mathrm{l}\{|X| < c\}$$

$$\leq \mathbb{E}g(|X|)\mathrm{l}\{|X| \geq c\} + g(c)\mathbf{P}(|X| < c)$$

$$= \mathbb{E}\left[g(|X|)\middle||X| \geq c\right]\mathbf{P}(|X| \geq c) + g(c)\mathbf{P}(|X| < c).$$

We now apply Jensen's inequality to the term on the left, and also use the concavity on $[c, \infty)$ to incorporate the term on the right:

$$\mathbb{E}g(|X|) \leq g\left[\mathbb{E}\left(|X|\middle||X| \geq c\right)\right]\mathbf{P}(|X| \geq c) + g(c)\mathbf{P}(|X| < c)$$

$$\leq g\left[\mathbb{E}|X| + c\mathbf{P}(|X| < c)\right].$$

$\square$

Application of the previous lemma gives

Lemma 14.7. *We have for any real-valued random variable X, and all $m \geq 1$,*

$$\mathbb{E}|X|^m \leq \log^m\left[\mathbb{E}e^{|X|} - 1 + e^{m-1}\right].$$

Proof. The function
$$g(x) = \log^m(1+x), \ x \geq 0,$$
is concave for all $x \geq e^{m-1} - 1$. Hence, by Lemma 14.6,

$$\mathbb{E}|X|^m = \mathbb{E}\log^m\left[e^{|X|} - 1 + 1\right]$$

$$\leq \log^m\left[\mathbb{E}(e^{|X|} - 1) + 1 + (e^{m-1} - 1)\right].$$

$\square$

14.3 Bernstein's inequality

In this section, we let $Z_1, \ldots, Z_n$ be independent random variables with values in some space $\mathscr{Z}$ and we let γ be a real-valued function on $\mathscr{Z}$, satisfying for a positive constant K,

$$\mathbb{E}\gamma(Z_i) = 0, \ \forall \ i, \ \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}|\gamma(Z_i)|^m \leq \frac{m!}{2}K^{m-2}, \ m = 2, 3, \ldots. \tag{14.3}$$

First, we derive a bound for the exponential moment of an average.

Lemma 14.8. *Assume (14.3). For all $L > K$, we have*

$$\mathbb{E}\exp\left[\sum_{i=1}^{n}\gamma(Z_i)/L\right] \leq \exp\left[\frac{n}{2(L^2 - LK)}\right].$$

Proof. This follows from Lemma 14.1, after noting that

$$\sum_{i=1}^{n}\mathbb{E}\left\{e^{|\gamma(Z_i)|/L} - 1 - |\gamma(Z_i)|/L\right\} \leq \sum_{m=2}^{\infty}\sum_{i=1}^{n}\frac{\mathbb{E}|\gamma(Z_i)|^m}{L^m m!}$$

$$\leq \frac{n}{2L^2}\sum_{m=2}^{\infty}(K/L)^{m-2} = \frac{n}{2L^2(1 - K/L)}.$$

$\square$

The bound of Lemma 14.8 leads to Bernstein's inequality, as given in Bennet (1962).

Lemma 14.9. *(Bernstein's inequality) Assume (14.3). Let $t > 0$ be arbitrary. Then*

$$\mathbf{P}\left(\frac{1}{n}\sum_{i=1}^{n}\gamma(Z_i) \geq Kt + \sqrt{2t}\right) \leq \exp[-nt].$$

Proof. Let $a > 0$ be arbitrary. In Lemma 14.8, take

$$K/L = 1 - (1 + 2aK/n)^{-1/2},$$

and apply Chebyshev's inequality to obtain

$$\mathbf{P}\left(\sum_{i=1}^{n}\gamma(Z_i) \geq a\right) \leq \exp\left[-\frac{a^2}{aK + n + \sqrt{2aKn + n^2}}\right].$$

Now, choose $a/n = Kt + \sqrt{2t}$.

$\square$

14.4 Hoeffding's inequality

This section presents Hoeffding's inequality as obtained by Hoeffding (1963). Let $Z_1, \ldots, Z_n$ be independent random variables with values in some space $\mathscr{Z}$ and let γ be a real-valued function on $\mathscr{Z}$, satisfying

$$\mathbb{E}\gamma(Z_i) = 0, \ |\gamma(Z_i)| \leq c_i \ \forall \ i. \tag{14.4}$$

Lemma 14.10. *Assume (14.4). For all $L > 0$,*

$$\mathbb{E}\exp\left[\sum_{i=1}^{n}\gamma(Z_i)/L\right]\leq\exp\left[\frac{\sum_{i=1}^{n}c_i^2}{2L^2}\right].$$

Proof. Write $X_i:=\gamma(Z_i)$, $i=1,\ldots,n$. By the convexity of the exponential function $\exp[x/L]$, we know that for any $0\leq\alpha\leq 1$,

$$\exp[\alpha x/L+(1-\alpha)y/L]\leq\alpha\exp[x/L]+(1-\alpha)\exp[y/L].$$

Define now

$$\alpha_i=\frac{c_i-X_i}{2c_i}.$$

Then

$$X_i=\alpha_i(-c_i)+(1-\alpha_i)c_i,$$

so

$$\exp[X_i/L]\leq\alpha_i\exp[-c_i/L]+(1-\alpha_i)\exp[c_i/L].$$

But then, since $\mathbb{E}\alpha_i=1/2$, we find

$$\mathbb{E}\exp[X_i/L]\leq\frac{1}{2}\exp[-c_i/L]+\frac{1}{2}\exp[c_i/L].$$

Now, for all x,

$$\exp[-x]+\exp[x]=2\sum_{k=0}^{\infty}\frac{x^{2k}}{(2k)!},$$

whereas

$$\exp[x^2/2]=\sum_{k=0}^{\infty}\frac{x^{2k}}{2^k k!}.$$

Since

$$(2k)!\geq 2^k k!,$$

we thus know that

$$\exp[-x]+\exp[x]\leq 2\exp[x^2/2],$$

and hence

$$\mathbb{E}\exp[X_i/L]\leq\exp\left[\frac{c_i^2}{2L^2}\right].$$

Therefore,

$$\mathbb{E}\exp\left[\sum_{i=1}^{n}X_i/L\right]\leq\exp\left[\frac{\sum_{i=1}^{n}c_i^2}{2L^2}\right].$$

$\square$

Consequently, using Chebyshev's inequality, we have sub-Gaussian tails for averages.

Lemma 14.11. *Assume (14.4). We have for all $t>0$*

$$\mathbf{P}\left(\left|\sum_{i=1}^{n}\gamma(Z_i)\right|\ge t)\right)\le 2\exp\left[-\frac{t^2}{2\sum_{i=1}^{n}c_i^2}\right].$$

Proof. It follows from Chebyshev's inequality and Lemma 14.10 that for all $L > 0$

$$\mathbf{P}\left(\sum_{i=1}^{n}\gamma(Z_i)\ge t\right)\le\exp\left[\frac{\sum_{i=1}^{n}c_i^2}{2L^2}-\frac{t}{L}\right].$$

Take $L = (\sum_{i=1}^{n}c_i^2)/t$ to complete the proof. $\qquad\square$

14.5 The maximum of p averages

14.5.1 Using Bernstein's inequality

Let $Z_1,\ldots,Z_n$ be independent $\mathscr{Z}$-valued random variables, and $\gamma_1,\ldots,\gamma_p$ be real-valued functions on $\mathscr{Z}$, satisfying for $j = 1,\ldots,p$,

$$\mathbb{E}\gamma_j(Z_i) = 0,\ \forall\,i,\ \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}|\gamma_j(Z_i)|^m\le\frac{m!}{2}K^{m-2},\ m = 2,3,\ldots. \qquad (14.5)$$

We start out with an inequality for the m-th moment of the maximum of p averages.

Lemma 14.12. *Assume (14.5). We have for $m = 1,2,\ldots$*

$$\mathbb{E}\left(\max_{1\le j\le p}|\frac{1}{n}\sum_{i=1}^{n}\gamma_j(Z_i)|^m\right)$$

$$\le\left(\frac{K\log(2p+\mathrm{e}^{m-1}-p)}{n}+\sqrt{\frac{2\log(2p+\mathrm{e}^{m-1}-p)}{n}}\right)^m.$$

Proof. By Lemma 14.7, for all $L > 0$, and all m

$$\mathbb{E}\left(\max_j|\sum_{i=1}^{n}\gamma_j(Z_i)|^m\right)\le L^m\log^m\left[\mathbb{E}\exp[\max_j|\sum_{i=1}^{n}\gamma_j(Z_i)|/L]-1+\mathrm{e}^{m-1}\right].$$

From Lemma 14.8, and invoking $\mathrm{e}^{|x|}\le\mathrm{e}^x+\mathrm{e}^{-x}$, we obtain for $L > K$,

$$L^m\log^m\left[\mathbb{E}\exp[\max_j|\sum_{i=1}^{n}\gamma_j(Z_i)|/L]-1+\mathrm{e}^{m-1}\right]$$

$$\leq L^m \log^m \left[p \left\{ 2\exp\left[\frac{n}{2(L^2-LK)}\right] -1 \right\} + e^{m-1} \right]$$

$$\leq L^m \log^m \left[(2p+e^{m-1}-p)\exp\left[\frac{n}{2(L^2-LK)}\right] \right]$$

$$= \left(L\log(2p+e^{m-1}-p) + \left[\frac{n}{2(L-K)}\right] \right)^m .$$

Now, take

$$L = K + \sqrt{\frac{n}{2\log(2p+e^{m-1}-p)}}.$$

$\square$

We define

$$\lambda(K,n,p) := \sqrt{\frac{2\log(2p)}{n}} + \frac{K\log(2p)}{n}. \qquad (14.6)$$

In what follows, the quantity $\lambda(K,n,p)$ plays an important role: it will appear in several places. From an asymptotic point of view, that is, when n is large and K is not too large, the first (square-root) term in the definition of $\lambda(K,n,p)$ is the leading term.

In most of our applications, p will be large, in particular, it will be much larger than m. To simplify the expressions for this case, we have, as corollary of Lemma 14.12

Corollary 14.1. *Assume (14.5). For all $m \leq 1 + \log p$, it holds that*

$$\mathbb{E}\left(\max_{1\leq j\leq p} \left| \frac{1}{n}\sum_{i=1}^{n} \gamma_j(Z_i) \right|^m \right) \leq \lambda^m(K,n,p).$$

We also put forward a probability inequality involving $\lambda(K,n,p)$.

Lemma 14.13. *Assume (14.5). We have for all $t > 0$,*

$$\mathbf{P}\left(\max_{1\leq j\leq p} \left| \frac{1}{n}\sum_{i=1}^{n} \gamma_j(Z_i) \right| \geq Kt + \sqrt{2t} + \lambda(K,n,p) \right) \leq \exp[-nt].$$

Proof. Bernstein's inequality says that for each j,

$$\mathbf{P}\left(\frac{1}{n}\sum_{i=1}^{n} \gamma_j(Z_i) \geq Kt + \sqrt{2t} \right) \leq \exp[-nt].$$

The result now follows immediately from

$$Kt + \sqrt{2t} + \lambda(K,n,p) = Kt + \sqrt{2t} + \sqrt{\frac{2\log 2p}{n}} + K\frac{\log(2p)}{n}$$

$$\geq K\left(t+\frac{\log(2p)}{n}\right)+\sqrt{2\left(t+\frac{\log(2p)}{n}\right)}.$$

$\square$

Example 14.1. Consider independent random variables $X_1,\ldots,X_n$ in $\mathbb{R}^p$. Define

$$\sigma_j^2 := \frac{1}{n}\sum_{i=1}^n \mathbb{E}|X_i^{(j)}|^2, \ j=1,\ldots,p.$$

Suppose that the $X_i^{(j)}$ are, uniformly in j, sub-Gaussian: for some constants K and σ_0^2,

$$K^2(\mathbb{E}\exp[|X_i^{(j)}|^2/K^2]-1)\leq \sigma_0^2, \ j=1,\ldots,p.$$

Then by Lemma 14.5, the $|X_i^{(j)}|^2$ are sub-exponential, with for all $m\geq 2$, and all j,

$$\mathbb{E}\left||X_i^{(j)}|^2-\mathbb{E}|X^{(j)}|^2\right|^m \leq \frac{m!}{2}(2K^2)^{m-2}(2K\sigma_0)^2.$$

We may normalize this to

$$\mathbb{E}\left|\frac{|X_i^{(j)}|^2}{2K\sigma_0}-\frac{\mathbb{E}|X_i^{(j)}|^2}{2K\sigma_0}\right|^m \leq \frac{m!}{2}\left(\frac{K}{\sigma_0}\right)^{m-2}.$$

Let

$$\hat{\sigma}_j^2 := \frac{1}{n}\sum_{i=1}^n |X_i^{(j)}|^2, \ j=1,\ldots,p.$$

Then, from Corollary 14.1, for $m\leq 1+\log p$,

$$\mathbb{E}\left(\max_{1\leq j\leq p}\left|\hat{\sigma}_j^2-\sigma_j^2\right|^m\right) \leq (2K\sigma_0)^m\lambda^m(K/\sigma_0,n,p).$$

It follows moreover from Lemma 14.13 that for all t,

$$\mathbf{P}\left(\max_{1\leq j\leq p}|\hat{\sigma}_j^2-\sigma_j^2|\geq 2K^2t+2K\sigma_0\sqrt{2t}+2K\sigma_0\lambda\left(K/\sigma_0,n,p\right)\right)\leq \exp[-nt].$$

14.5.2 Using Hoeffding's inequality

Let $Z_1,\ldots,Z_n$ be independent $\mathscr{Z}$-valued random variables, and $\gamma_1,\ldots,\gamma_p$ be real-valued functions on $\mathscr{Z}$, satisfying for $j=1,\ldots,p$,

$$\mathbb{E}\gamma_j(Z_i)=0, \ |\gamma(Z_i)|\leq c_{i,j}, \ \forall \, i. \tag{14.7}$$

Lemma 14.14. *Assume (14.7). For $m \geq 1$ and $p \geq e^{m-1}$, we have*

$$\mathbb{E} \max_{1 \leq j \leq p} \left| \sum_{i=1}^{n} \gamma_j(Z_i) \right|^m \leq \left[2\log(2p) \right]^{m/2} \max_{1 \leq j \leq p} \left[\sum_{i=1}^{n} c_{i,j}^2 \right]^{m/2}.$$

Proof. By Hoeffding's inequality (see Lemma 14.10), for all $L > 0$ and all j,

$$\mathbb{E} \exp \left[\sum_{i=1}^{n} \gamma_j(Z_i)/L \right] \leq \exp \left[\frac{\sum_{i=1}^{n} c_{i,j}^2}{2L^2} \right].$$

In view of Jensen's inequality

$$\mathbb{E} \max_{1 \leq j \leq p} \left| \sum_{i=1}^{n} \gamma_j(Z_i) \right|^m$$

$$\leq L^m \log^m \left\{ \mathbb{E} \exp \left[\max_{1 \leq j \leq p} \left| \sum_{i=1}^{n} \gamma_j(Z_i) \right| / L \right] - 1 + e^{m-1} \right\}$$

$$\leq L^m \log^m \left\{ p \max_{1 \leq j \leq p} \left(\mathbb{E} \exp \left[\left| \sum_{i=1}^{n} \gamma_j(Z_i) \right| / L \right] - 1 \right) + e^{m-1} \right\}$$

$$\leq L^m \log^m \left\{ p \max_{1 \leq j \leq p} \left(2 \exp \left[\sum_{i=1}^{n} c_{i,j}^2 /(2L^2) \right] - 1 \right) + e^{m-1} \right\}$$

$$\leq L^m \left\{ \log(2p) + \max_{1 \leq j \leq p} \frac{\sum_{i=1}^{n} c_{i,j}^2}{2L^2} \right\}^m.$$

Choosing

$$L := \max_{1 \leq j \leq p} \sqrt{\frac{\sum_{i=1}^{n} c_{i,j}^2}{2\log(2p)}}$$

gives the result.   □

The extension of Hoeffding's probability inequality to p variables reads as follows.

Lemma 14.15. *Let*

$$\|c\|_n^2 := \sum_{i=1}^{n} c_i^2 /n.$$

We have for all $t > 0$,

$$\mathbf{P} \left(\max_{1 \leq j \leq p} \left| \frac{1}{n} \sum_{i=1}^{n} \gamma_j(Z_i) \right| \geq \|c\|_n \sqrt{2 \left(t^2 + \frac{\log(2p)}{n} \right)} \right) \leq \exp[-nt^2].$$

Proof. By Lemma 14.11, we have for all $t > 0$ and all j,

$$\mathbf{P}\left(\left|\sum_{i=1}^{n}\gamma_{j}(Z_{i})\right|\geq t\right)\leq 2\exp\left[-\frac{nt^{2}}{2\|c\|_{n}^{2}}\right].$$

Hence, using the union bound,

$$\mathbf{P}\left(\max_{1\leq j\leq p}\left|\frac{1}{n}\sum_{i=1}^{n}\gamma_{j}(Z_{i})\right|\geq\|c\|_{n}\sqrt{2\left(t^{2}+\frac{\log(2p)}{n}\right)}\right)\leq 2p\exp[-nt^{2}-\log(2p)].$$

□

14.5.3 Having sub-Gaussian random variables

Let $Z_{1},\ldots,Z_{n}$ be independent $\mathscr{Z}$-valued random variables, and $\gamma_{1},\ldots,\gamma_{p}$ be real-valued functions on $\mathscr{Z}$, satisfying for $j=1,\ldots,p$,

$$\mathbb{E}\gamma_{j}(Z_{i})=0,\ K_{i}^{2}(\mathbb{E}\exp[\gamma_{j}^{2}(Z_{i})/K_{i}^{2}]-1)\leq\sigma_{0,i}^{2}\ \forall\ i. \tag{14.8}$$

Lemma 14.16. *Assume (14.8). Define*

$$R^{2}:=\sum_{i=1}^{n}(K_{i}^{2}+\sigma_{0,i}^{2})/n.$$

Then

$$\mathbf{P}\left(\max_{1\leq j\leq p}\left|\sum_{i=1}^{n}\gamma_{j}(Z_{i})/n\right|\geq R\sqrt{8\left(t^{2}+\frac{\log(2p)}{n}\right)}\right)\leq\exp[-nt^{2}],$$

and

$$\mathbb{E}\exp\left[\max_{1\leq j\leq p}\frac{|\sum_{i=1}^{n}\gamma_{j}(Z_{i})|^{2}}{n16R^{2}}\right]\leq 2p+1.$$

Proof. By Lemma 14.3, for all j and L,

$$\mathbb{E}\exp[\gamma_{j}(Z_{i})/L]\leq\exp[2(K_{i}^{2}+\sigma_{0,i}^{2})/L^{2}],\ i=1,\ldots,n.$$

Hence for all j and L,

$$\mathbb{E}\exp\left[\sum_{i=1}^{n}\gamma_{j}(Z_{i})/(nL)\right]\leq\exp[2R^{2}/nL^{2}].$$

It follows from Chebyshev's inequality that

$$\mathbf{P}\left(|\sum_{i=1}^{n}\gamma(Z_{i})/n|\geq t\right)\leq 2\exp\left[\frac{2R^{2}}{nL^{2}}-\frac{t}{L}\right].$$

Take $L = \frac{4R^2}{nt}$ to get

$$\mathbf{P}\left(\left|\sum_{i=1}^{n}\gamma(Z_i)/n\right| \geq t\right) \leq 2\exp\left[-\frac{nt^2}{8R^2}\right].$$

One easily sees that this implies the probability inequality for the maximum over j. Moreover, by Lemma 14.4 (where we take $a = 0$),

$$\mathbb{E}\exp\left[\frac{|\sum_{i=1}^{n}\gamma_j(X_i)|^2}{n16R^2}\right] \leq 3.$$

Thus

$$\mathbb{E}\exp\left[\max_{1\leq j\leq p}\frac{|\sum_{i=1}^{n}\gamma_j(Z_i)|^2}{n16R^2}\right] \leq \mathbb{E}\sum_{j=1}^{p}\left(\exp\left[\frac{|\sum_{i=1}^{n}\gamma_j(X_i)|^2}{n16R^2}\right] - 1\right) + 1$$

$$\leq 2p + 1.$$

$\square$

14.6 Concentration inequalities

Concentration inequalities have been derived by Talagrand (e.g. Talagrand (1995)) and further studied by Ledoux (e.g. Ledoux (1996)), Massart (Massart, 2000a) and Bousquet (2002) and others. Massart (2000b) has highlighted the importance of these results for statistical theory.

Let $Z_1, \ldots, Z_n$ be independent random variables with values in some space $\mathscr{Z}$ and let Γ be a class of real-valued functions on $\mathscr{Z}$. Define

$$\mathbf{Z} := \sup_{\gamma\in\Gamma}\left|\frac{1}{n}\sum_{i=1}^{n}\left(\gamma(Z_i) - \mathbb{E}\gamma(Z_i)\right)\right|.$$

14.6.1 Bousquet's inequality

Theorem 14.1 below is a result on the amount of concentration of $\mathbf{Z}$ around its mean, which has the flavor of a Bernstein inequality.

Theorem 14.1. *(Concentration Theorem (Bousquet, 2002)) Suppose*

$$\mathbb{E}\gamma(Z_i) = 0 \ \forall \ \gamma \in \Gamma, \forall \ i, \ \frac{1}{n}\sum_{i=1}^{n}\sup_{\gamma \in \Gamma}\mathbb{E}\gamma^2(Z_i) \leq 1,$$

and moreover, for some positive constant K,

$$\|\gamma\|_\infty \leq K, \ \forall \ \gamma \in \Gamma.$$

Then for all L > 0,
$$\log \mathbb{E}\exp[n(\mathbf{Z} - \mathbb{E}\mathbf{Z})/L]$$
$$\leq n[e^{K/L} - 1 - K/L][2\mathbb{E}\mathbf{Z}/K + 1/K^2].$$

To appreciate Theorem 14.1 one may compare it with Lemma 14.1.

The probability inequality given in the following corollary can now be derived in the same way as in Lemma 14.8 and Lemma 14.9.

Corollary 14.2. *Suppose the conditions of Theorem 14.1 hold. We have for all t > 0,*

$$\mathbf{P}\left(\mathbf{Z} \geq \mathbb{E}\mathbf{Z} + \frac{tK}{3} + \sqrt{2}t\sqrt{1 + 2K\mathbb{E}\mathbf{Z}}\right) \leq \exp[-nt].$$

Bousquet's inequality indeed hardly pays a price for its generality. Let us make a comparison with Lemma 14.13, which is about the special case where $|\Gamma| := p$ is finite. We assume now that the functions are bounded by K, which implies (14.5) with K replaced by $K/3$.

Corollary 14.3. *Suppose the conditions of Theorem 14.1 hold and that $|\Gamma| = p$. Then for all t > 0,*

$$\mathbf{P}\left(\max_{1 \leq j \leq p}\left|\frac{1}{n}\sum_{i=1}^{n}\gamma_j(Z_i)\right| \geq A(t)\right) \leq \exp[-nt],$$

where, from Corollary 14.1 (with m = 1) and Corollary 14.2, we may take

$$A(t) = \frac{tK}{3} + \lambda\left(\frac{K}{3}, n, p\right) + \sqrt{2}t\sqrt{1 + 2K\lambda\left(\frac{K}{3}, n, p\right)}.$$

In fact, Lemma 14.13 tells us that we may take

$$A(t) = \frac{tK}{3} + \lambda\left(\frac{K}{3}, n, p\right) + \sqrt{2}t.$$

14.6.2 *Massart's inequality*

The next inequality from Massart (2000a) has the flavor of a Hoeffding inequality.

Theorem 14.2. *(Concentration Theorem (Massart, 2000a)) Suppose* $\mathbb{E}\gamma(Z_i) = 0$ *and* $|\gamma(Z_i)| \leq c_{i,\gamma}$ *for some real numbers* $c_{i,\gamma}$ *and for all* $1 \leq i \leq n$ *and* $\gamma \in \Gamma$, *where*

$$\sup_{\gamma \in \Gamma} \sum_{i=1}^{n} c_{i,\gamma}^2 / n \leq 1.$$

Then for any positive t,

$$\mathbf{P}(\mathbf{Z} \geq \mathbb{E}\mathbf{Z} + t) \leq \exp\left[-\frac{nt^2}{8}\right].$$

14.6.3 *Sub-Gaussian random variables*

Bousquet's and Massart's concentration inequalities require that the random variables involved are bounded. In this section, we consider the sub-Gaussian case.

The next lemma will be applied with, for all s, the random variables X_s of the form

$$X_s := \max_{1 \leq j \leq N_s} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \gamma_j^s(Z_i) \right|,$$

where for $s \in \{1, \ldots, S\}$ (where $S \in \mathbb{N} \cup \infty$), and for $j = 1, \ldots, N_s$, γ_j^s are given functions on $\mathscr{Z}$ indexed by j and s, which satisfy the assumption (14.8). We will need the entropy results of Section 14.12 to complete the picture: see Theorem 14.7.

Lemma 14.17. *Let* $X_1, \ldots, X_S$ *be positive random variables, that satisfy for some positive constants* $\{\delta_s, N_s\}_{s=1}^{S}$,

$$\mathbb{E}\exp\left[X_s^2 / \delta_s^2\right] \leq 1 + 2N_s, \ s = 1, \ldots, S.$$

Define

$$a := 2\sqrt{2} \sum_{s=1}^{S} \delta_s \sqrt{\log(1 + 2N_s) \vee s},$$

and

$$K := 4 \sum_{s=1}^{S} \delta_s \sqrt{s}.$$

Then for all $t \geq a$,

$$\mathbf{P}\left(\sum_{s=1}^{S} X_s \geq t\right) \leq 2\exp\left[-\frac{2t^2}{K^2}\right],$$

and moreover

$$\mathbb{E}\exp\left[\frac{(\sum_{s=1}^{S} X_s - a)_+}{K}\right]^2 \leq 1 + 2\exp[-2a^2/K^2].$$

Proof. Define, for $t \geq a$,

$$\eta_s = \frac{\sqrt{2}\delta_s\sqrt{\log(1+2N_s)}}{t} \vee \frac{\delta_s\sqrt{s}}{2\sum_{s=1}^{S}\delta_s\sqrt{s}}, \quad s = 1, \ldots, S.$$

Then

$$\sum_{s=1}^{S} \eta_s \leq 1,$$

and hence

$$\mathbf{P}\left(\sum_{s=1}^{S} X_s \geq t\right) \leq \sum_{s=1}^{S} \mathbf{P}(X_s \geq t\eta_s).$$

Fix now some s. By Chebyshev's inequality,

$$\mathbf{P}(X_s \geq t\eta_s) \leq (1+2N_s)\exp[-t^2\eta_s^2/\delta_s^2]$$

$$\leq \exp\left[-\frac{t^2\eta_s^2}{2\delta_s^2}\right] \leq \exp\left[-\frac{2t^2s}{K^2}\right].$$

Hence

$$\mathbf{P}\left(\sum_{s=1}^{S} X_s \geq t\right) \leq \sum_{s=1}^{S} \exp\left[-\frac{2t^2s}{K^2}\right] \leq 2\exp\left[-\frac{2t^2}{K^2}\right].$$

Insert Lemma 14.4 to finish the proof. □

14.7 Symmetrization and contraction

In order to be able to apply Bousquet's or Massart's inequality, one needs a bound for the mean $\mathbb{E}\mathbf{Z}$. The results in this section are often useful to obtain such a bound.

A *Rademacher sequence* is a sequence $\varepsilon_1, \ldots, \varepsilon_n$ of i.i.d. copies of a random variable ε taking values in $\{\pm 1\}$, with $\mathbf{P}(\varepsilon = +1) = \mathbf{P}(\varepsilon = -1) = 1/2$.

Theorem 14.3. *(Symmetrization Theorem (van der Vaart and Wellner, 1996)) Let* $\varepsilon_1, \ldots, \varepsilon_n$ *be a Rademacher sequence independent of* $Z_1, \ldots, Z_n$. *Then for any* $m \geq 1$,

$$\mathbb{E}\left(\sup_{\gamma\in\Gamma}\left|\sum_{i=1}^{n}\{\gamma(Z_i)-E\gamma(Z_i)\}\right|^m\right)\le 2^m\mathbb{E}\left(\sup_{\gamma\in\Gamma}\left|\sum_{i=1}^{n}\varepsilon_i\gamma(Z_i)\right|^m\right). \qquad (14.9)$$

One may also formulate a symmetrization result for probabilities: see Problem 14.5.

Let us now find a bound for the right hand side of (14.9). For simplicity, we consider only the case $m = 1$. Section 14.11 considers sub-Gaussian "moments". We will apply the chaining argument developed by Kolmogorov, see e.g. van der Vaart and Wellner (1996), van de Geer (2000), and the references therein, and see also Talagrand (2005) for the refinement to so-called *generic* chaining. We will use the entropy of a class of functions (see also Section 14.11).

Endow Γ with the (random) norm

$$\|\gamma\|_n := \sqrt{\frac{1}{n}\sum_{i=1}^{n}\gamma^2(Z_i)}.$$

Definition *For $\delta > 0$, the δ-covering number $N(\delta,\Gamma,\|\cdot\|_n)$ of $(\Gamma,\|\cdot\|_n)$ is the minimum number of balls with radius δ necessary to cover the class Γ. The* entropy *is $H(\cdot,\Gamma,\|\cdot\|_n) := \log N(\cdot,\Gamma,\|\cdot\|_n)$.*

Lemma 14.18. *Suppose $(Z_1,\ldots,Z_n) = (z_1,\ldots,z_n)$ is fixed. Define*

$$R_n := \sup_{\gamma\in\Gamma}\|\gamma\|_n,$$

and $N_s = N(2^{-s}R_n,\Gamma,\|\cdot\|_n)$, $s = 1,\ldots,S$, where

$$S := \min\{s \ge 1 : 2^{-s} \le 4/\sqrt{n}\}.$$

Let $\varepsilon_1,\ldots,\varepsilon_n$ be a Rademacher sequence. We have

$$\mathbb{E}\left(\sup_{\gamma\in\Gamma}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\gamma(z_i)\right|\right) \le 6\sum_{s=1}^{S}\sqrt{\frac{\log(1+N_s)}{n}}2^{-s}R_n + \frac{4R_n}{\sqrt{n}}.$$

Proof. Let for $s = 0,1,2,\ldots$, $\{\gamma_j^s\}_{j=1}^{N_s}$ be a minimal $2^{-s}R_n$-covering set of $(\Gamma,\|\cdot\|_n)$, with $N_s = N(2^{-s}R_n,\Gamma,\|\cdot\|_n)$, and for each γ, there exists a $\gamma^s \in \{\gamma_1^s,\ldots,\gamma_{N_s}^s\}$ such that $\|\gamma-\gamma^s\|_n \le 2^{-s}R_n$. Choose $\gamma^0 \equiv 0$. Now we insert the chaining argument. We get

$$\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\gamma(z_i)\right| \le \sum_{s=1}^{S}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(\gamma^s-\gamma^{s-1})(z_i)\right| + \left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(\gamma-\gamma^S)(z_i)\right|,$$

where in the above expression, each of the γ^s are taken as $2^{-s}R_n$-approximation of γ. Clearly,

$$\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(\gamma-\gamma^S)(z_i)\right| \le \|\gamma-\gamma^S\|_n \le 2^{-S}R_n \le 4R_n/\sqrt{n}.$$

Then, by Lemma 14.14 (with a slight improvement for the case $m = 1$), and using $\|\gamma^s - \gamma^{s-1}\|_n \le 3 \times 2^{-s}R_n$ for all s,

$$\mathbf{E}\left(\sup_{\gamma\in\Gamma}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(\gamma-\gamma^S)(z_i)\right|\right) \le \sum_{s=1}^{S}3\sqrt{\frac{2\log(1+N_sN_{s-1})}{n}}2^{-s}R_n + \frac{4R_n}{\sqrt{n}}$$

$$\le 6\sum_{s=1}^{S}\sqrt{\frac{\log(1+N_s)}{n}}2^{-s}R_n + \frac{4R_n}{\sqrt{n}}.$$

$\square$

The next corollary can be invoked in Section 9.5. The bound on the covering number that we use here is based on entropy calculations for convex hulls, see Lemma 14.28.

Corollary 14.4. *Let $Z_1,\ldots,Z_n$ be independent random variables with values in $\mathcal{Z}$ and Γ be a class of real-valued functions on $\mathcal{Z}$. Suppose that for some non-random constants $R_n < \infty$ and $A > 0$,*

$$\|\gamma\|_n \le R_n,$$

and

$$\log(1+N(2^{-s}R_n,\Gamma,\|\cdot\|_n)) \le A2^{2s}, \forall\, 0 \le s \le S,$$

where

$$S := \min\{s \ge 1:\ 2^{-s} \le 4/\sqrt{n}\}.$$

Then, using the bounds in Lemma 14.18, and the bound $S \le \log_2 n/2$, we get, for a Rademacher sequence $\varepsilon_1,\ldots,\varepsilon_n$ independent of $Z_1,\ldots,Z_n$,

$$\mathbb{E}\left(\sup_{\gamma\in\Gamma}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\gamma(Z_i)\right|\right) \le \frac{R_n}{\sqrt{n}}\left(3\sqrt{A}\log_2 n + 4\right).$$

Massart's Inequality (Theorem 14.2) then yields

$$\mathbf{P}\left(\sup_{\gamma\in\Gamma}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\gamma(Z_i)\right| > \frac{R_n}{\sqrt{n}}\left(3\sqrt{A}\log_2 n + 4 + t\right)\right) \le \exp\left[-\frac{nt^2}{8}\right].$$

Now, desymmetrizing (see Problem 14.5) gives that for $n \ge 8K^2$,

$$\mathbf{P}\left(\sup_{\gamma\in\Gamma}\left|\frac{1}{n}\sum_{i=1}^{n}\gamma(Z_i)\right| > \frac{4R_n}{\sqrt{n}}\left(3\sqrt{A}\log_2 n + 4 + 4t\right)\right) \le 4\exp\left[-\frac{nt^2}{8}\right].$$

Tailoring the result for the demands in Section 9.5, we have as a consequence, for all $T \ge 1$, and for $R_n^2 \le (\varepsilon^2 \wedge 1)M_n^2$, where $\varepsilon > 0$ and $M_n > 0$ are certain constants

$$\mathbf{P}\left(\sup_{\gamma \in \Gamma}\left|\frac{1}{n}\sum_{i=1}^{n}\gamma(Z_i)\right| > \frac{8\varepsilon M_n T}{\sqrt{n}}\left(3\sqrt{A}\log_2 n + 4\right)\right)$$

$$\leq 4\exp\left[-\frac{T^2(3\sqrt{A}\log_2 n + 4)^2(\varepsilon^2 \vee 1)}{8}\right]. \tag{14.10}$$

When Γ is a collection of linear functions $\{f_\beta(\cdot) = \sum_{j=1}^{p}\beta_j\psi_j(\cdot) : \beta \in \mathbb{R}^p\}$, one can apply Hölder's inequality

$$\mathbb{E}\sup_{\beta}\left|\sum_{i=1}^{n}\varepsilon_i f_\beta(Z_i)\right| \leq \sup_{\beta}\|\beta\|_q \mathbb{E}\left(\sum_{j=1}^{p}\left|\sum_{i=1}^{n}\varepsilon_i\psi_j(Z_i)\right|^r\right)^{1/r},$$

where $1 \leq q \leq \infty$ and $1/q + 1/r = 1$. Of special importance here is the case $q = 1$ ($r = \infty$): it yields a bound in terms of the ℓ_1-norm of β.

In the case where the functions in Γ are not linear, but some Lipschitz function of a linear function, the following contraction inequality is very useful.

Theorem 14.4. *(Contraction Theorem (Ledoux and Talagrand, 1991)) Let $z_1,\ldots,z_n$ be non-random elements of $\mathcal{Z}$ and let $\mathcal{F}$ be a class of real-valued functions on $\mathcal{Z}$. Consider Lipschitz functions $\gamma_i : \mathbb{R} \to \mathbb{R}$, i.e.*

$$|\gamma_i(s) - \gamma_i(\tilde{s})| \leq |s - \tilde{s}|, \ \forall \, s,\tilde{s} \in \mathbb{R}.$$

Let $\varepsilon_1,\ldots,\varepsilon_n$ be a Rademacher sequence. Then for any function $f^* : \mathcal{Z} \to \mathbb{R}$, we have

$$\mathbb{E}\left(\sup_{f \in \mathcal{F}}\left|\sum_{i=1}^{n}\varepsilon_i\{\gamma_i(f(z_i)) - \gamma_i(f^*(z_i))\}\right|\right)$$

$$\leq 2\mathbb{E}\left(\sup_{f \in \mathcal{F}}\left|\sum_{i=1}^{n}\varepsilon_i(f(z_i) - f^*(z_i))\right|\right).$$

14.8 Concentration inequalities for Lipschitz loss functions

We provide probability inequalities for the set $\mathcal{T}$ as given in Section 6.6 and Section 6.7, for the case of Lipschitz loss.

Consider independent random variables $\{Z_i\}_{i=1}^{n}$, with (for $i = 1,\ldots,n$) Z_i in some space $\mathcal{Z}$. For $f : \mathcal{Z} \to \mathbb{R}$, define

$$\|f\|_n^2 := \frac{1}{n}\sum_{i=1}^{n}f^2(Z_i), \ \|f\|^2 := \mathbb{E}\|f\|_n^2.$$

Let, for all i, $\gamma_i : \mathbb{R} \to \mathbb{R}$ be given functions. Let $\{\psi_1, \ldots, \psi_p\}$ be a given dictionary on $\mathscr{Z}$, and define

$$f_\beta := \sum_{j=1}^{p} \beta_j \psi_j, \ \beta \in \mathbb{R}^p.$$

Assume γ_i is a Lipschitz function, with Lipschitz constant L not depending on i:

$$|\gamma_i(s) - \gamma_i(\tilde{s})| \le L|s - \tilde{s}| \ \forall s, \tilde{s} \in \mathbb{R}. \tag{14.11}$$

Define the empirical process

$$v_n(\beta) := \frac{1}{n} \sum_{i=1}^{n} \left[\gamma_i(f_\beta(Z_i)) - \mathbb{E}\gamma_i(f_\beta(Z_i)) \right], \ \beta \in \mathbb{R}^p.$$

Let β^* be fixed. For all $\delta > 0$, we define the random variable

$$\bar{\mathbf{Z}}_\delta := \sup_{\|f_\beta - f_{\beta^*}\| \le \delta} |v_n(\beta) - v_n(\beta^*)|/\sqrt{p},$$

and, for all $M > 0$, we define

$$\mathbf{Z}_M := \sup_{\|\beta - \beta^*\|_1 \le M} |v_n(\beta) - v_n(\beta^*)|.$$

Lemma 14.19. *Assume (14.11). We have (for all $\delta > 0$)*

$$\mathbb{E}\bar{\mathbf{Z}}_\delta \le 4\delta L/\sqrt{n}.$$

Proof. Without loss of generality, we may assume that $\sum_{i=1}^{n} \mathbb{E}\psi^T(X_i)\psi(X_i)/n = I$, so that $\|f_\beta - f_{\beta^*}\| = \|\beta - \beta^*\|$, and $\|\psi_j\| = 1$ for all j. Let $\varepsilon_1, \ldots, \varepsilon_n$ be a Rademacher sequence, independent of $\mathbf{z} := Z_1, \ldots, Z_n$. By the Symmetrization Theorem (Theorem 14.3) combined with the Contraction Theorem (Theorem 14.4), we have

$$\sqrt{p}\mathbb{E}\bar{\mathbf{Z}}_\delta \le 2\mathbb{E} \sup_{\|f_\beta - f_{\beta^*}\| \le \delta} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \left(\gamma_i(f_\beta(Z_i)) - \gamma_i(f_{\beta^*}(Z_i)) \right) \right|$$

$$\le 4L\mathbb{E} \sup_{\|f_\beta - f_{\beta^*}\| \le \delta} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \left(f_\beta(Z_i) - f_{\beta^*}(Z_i) \right) \right|.$$

But, in view of the Cauchy-Schwarz inequality,

$$\left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \left(f_\beta(Z_i) - f_{\beta^*}(Z_i) \right) \right| \le \|\beta - \beta^*\|_2 \sqrt{\sum_{j=1}^{p} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \psi_j(Z_i) \right|^2},$$

and

$$\mathbb{E} \sum_{j=1}^{p} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \psi_j(Z_i) \right|^2 = \frac{p}{n}.$$

$\square$

Lemma 14.20. *Assume (14.11). We have (for all $M > 0$)*

$$\mathbb{E}\mathbf{Z}_M \leq 4ML\sqrt{\frac{2\log(2p)}{n}} \mathbb{E}\left(\max_{1 \leq j \leq p} \|\psi_j\|_n \right).$$

If we assume that for all $m \geq 2$ and all j, and for some constant K,

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}|\psi_j(Z_i)|^m \leq \frac{m!}{2} K^{m-2},$$

we furthermore have

$$\mathbb{E}\mathbf{Z}_M \leq 4ML\lambda(K,n,p),$$

where

$$\lambda(K,n,p) := \left(\sqrt{\frac{2\log(2p)}{n}} + \frac{K\log(2p)}{n} \right).$$

Proof. Let $\varepsilon_1, \ldots, \varepsilon_n$ be a Rademacher sequence, independent of $\mathbf{z} := Z_1, \ldots, Z_n$. By the Symmetrization Theorem (Theorem 14.3) combined with the Contraction Theorem (Theorem 14.4), we have

$$\mathbb{E}\mathbf{Z}_M \leq 2\mathbb{E} \sup_{\|\beta - \beta^*\|_1 \leq M} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \left(\gamma_i(f_\beta(Z_i)) - \gamma_i(f_{\beta^*}(Z_i)) \right) \right|$$

$$\leq 4L\mathbb{E} \sup_{\|\beta - \beta^*\|_1 \leq M} \left| \varepsilon_i \left(f_\beta(Z_i) - f_{\beta^*}(Z_i) \right) \right|.$$

At this point we invoke Hölder's inequality:

$$\sup_{\|\beta - \beta^*\|_1 \leq M} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \left(f_\beta(Z_i) - f_{\beta^*}(Z_i) \right) \right| \leq M \max_{1 \leq j \leq p} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \psi_j(Z_i) \right|.$$

By Hoeffding's moment inequality (Lemma 14.14),

$$\mathbb{E} \max_{1 \leq j \leq p} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \psi_j(Z_i) \right| \leq \sqrt{\frac{2\log(2p)}{n}} \mathbb{E}\left(\max_{1 \leq j \leq p} \|\psi_j\|_n \right).$$

Under the assumption

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}|\psi_j(Z_i)|^m \leq \frac{m!}{2} K^{m-2},$$

it furthermore holds, using Bernstein's moment inequality (Lemma 14.12),

$$\mathbb{E} \max_{1 \le j \le p} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \psi_j(Z_i) \right| \le \lambda(K, n, p),$$

with

$$\lambda(K, n, p) := \sqrt{\frac{2 \log(2p)}{n}} + \frac{K \log(2p)}{n}.$$

<div align="right">□</div>

Example 14.2. Let $Y_1, \ldots, Y_n$ be independent $\mathscr{Y}$ valued response variables, $\mathscr{Y} \subset \mathbb{R}$, and $x_1, \ldots, x_n$ fixed co-variables in some space $\mathscr{X}$. Suppose that the ψ_j are functions on $\mathscr{X}$, with $\|\psi_j\|_n \le 1$ for all j. We assume that the loss function ρ_{f_β} is of the form

$$\rho_{f_\beta}(x_i, Y_i) = \gamma(Y_i, f_\beta(x_i)) + c_i(f_\beta),$$

where $c_i(f_\beta)$ is a constant depending (possibly) on f_β and i. We further assume for all $s, \tilde{s} \in \mathbb{R}$ and all $y \in \mathscr{Y}$,

$$|\gamma(y, s) - \gamma(y, \tilde{s})| \le L|s - \tilde{s}| \tag{14.12}$$

Then

$$\mathbb{E}\mathbf{Z}_M \le 8ML\sqrt{\frac{2 \log(2p)}{n}} \max_{1 \le j \le p} \|\psi_j\|_n \le 8ML\sqrt{\frac{2 \log(2p)}{n}}.$$

Massart's concentration inequality (Theorem 14.2) yields

$$\mathbf{P}\left(\mathbf{Z}_M \ge 2ML\left(4\sqrt{\frac{2 \log(2p)}{n}} + \sqrt{\frac{2t}{n}}\right)\right) \le \exp[-t].$$

Note that the result does not require any higher order moments for the response variables. This is due to the Lipschitz condition, i.e., to the fact that ρ_f is a robust loss function.

If we assume a uniform L_∞-bound K for the dictionary $\{\psi_j\}$, we can apply Bousquet's concentration inequality. As illustrated in the previous example however, such a bound K is not a sine qua non.

Theorem 14.5. *Assume (14.12) and that*

$$\frac{1}{n} \sum_{i=1}^{n} \max_{1 \le j \le p} \mathbb{E}\psi_j^2(Z_i) \le 1,$$

and that for some constant K,

$$\max_{1 \le j \le p} \|\psi_j\|_\infty \le K.$$

Let

$$\mathscr{T} := \left\{ \mathbf{Z}_M \leq ML\left[4\lambda\left(\frac{K}{3},n,p\right) + \frac{tK}{3n} + \sqrt{\frac{2t}{n}}\sqrt{1 + 8\lambda\left(\frac{K}{3},n,p\right)}\right]\right\},$$

where

$$\lambda\left(\frac{K}{3},n,p\right) := \sqrt{\frac{2\log(2p)}{n}} + \frac{K\log(2p)}{3n}.$$

Then

$$\mathbf{P}(\mathscr{T}) \geq 1 - \exp[-t].$$

Proof. This follows from combining Lemma 14.20 with Bousquet's probability inequality (Corollary 14.2). $\qquad\square$

14.9 Concentration for squared error loss with random design

This section completes the results of Section 6.6 and Section 6.7 for the case of least squares estimation with random design. We prove probability inequalities for the empirical process associated with squared error loss, where random design gives an additional term to handle as compared to fixed design. Indeed, the empirical process consists of two terms, a linear term involving the noise, and a quadratic term involving only the regression functions. Subsection 14.9.1 handles the first term, Subsection 14.9.2 handles the second term, and Subsection 14.9.3 combines the two.

Let $\{(\varepsilon_i, X_i)\}_{i=1}^n$ be independent, and satisfy for $i = 1, \ldots, n$, that $\varepsilon_i \in \mathbb{R}$ and $X_i \in \mathscr{X}$ are independent, and moreover that

$$\mathbb{E}\varepsilon_i = 0, \ \mathbb{E}|\varepsilon_i|^m \leq \frac{m!}{2}K_\varepsilon^{m-2}\sigma^2, \ m = 2, 3, \ldots. \tag{14.13}$$

Let $\mathscr{F} = \{f_\beta = \sum_{j=1}^p \beta_j \psi_j : \ \beta \in \mathbb{R}^p\}$ be a class of linear functions on $\mathscr{X}$, and let $f^* = f_{\beta^*}$ be a fixed function in $\mathscr{F}$. Suppose that

$$\frac{1}{n}\sum_{i=1}^n \max_{1 \leq j \leq p} \mathbb{E}\psi_j^2(X_i) \leq 1, \ \max_{1 \leq j \leq p}\|\psi_j\|_\infty \leq K. \tag{14.14}$$

Let $K_0 := K_\varepsilon K$.

Let us furthermore recall that $\lambda(K, n, p)$ was defined in (14.6).

The distribution of X_i is denoted by $Q^{(i)}$, and we let $Q = \sum_{i=1}^n Q^{(i)}/n$. Moreover, we let Q_n be the empirical distribution based on $X_1, \ldots, X_n$. Thus

$$Q_n f = \frac{1}{n}\sum_{i=1}^n f(X_i),$$

and

$$Qf := \mathbb{E}Q_n f.$$

14.9.1 The inner product of noise and linear functions

Lemma 14.21. *Assume (14.13) and (14.14). For all $t > 0$,*

$$\mathbf{P}\left(\sup_{f_\beta \in \mathscr{F}} \frac{\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i (f_\beta - f^*)(X_i)\right|}{\|\beta - \beta^*\|_1} \geq \alpha_\varepsilon(t)\right)$$

$$\leq \exp[-nt],$$

where

$$\alpha_\varepsilon(t) := K_0 t + \sigma\lambda\left(\frac{K_0}{\sigma}, n, p\right) + \sigma\sqrt{2t}.$$

Proof. It holds that

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}|\varepsilon_i \psi_j(X_i)|^m \leq \frac{m!}{2}K_0^{m-2}\sigma^2.$$

Hence by Lemma 14.13, for all $t > 0$,

$$\mathbf{P}\left(\max_{1\leq j\leq p}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \psi_j(X_i)\right| \geq tK_0 + \sigma\lambda\left(\frac{K_0}{\sigma}, n, p\right) + \sigma\sqrt{2t}\right) \leq \exp[-nt].$$

The result now follows from Hölder's inequality:

$$\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i (f_\beta - f^*)(X_i)\right| \leq \|\beta - \beta^*\|_1 \max_{1\leq j\leq p}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \psi_j(X_i)\right|.$$

$\square$

14.9.2 Squared linear functions

Lemma 14.22. *Assume (14.14). Let $\mathscr{F}_M := \{f_\beta = \sum_{j=1}^p \beta_j\psi_j : \|\beta - \beta^*\|_1 \leq M\}$ be a class of linear functions on $\mathscr{X}$, and let f^0 be a fixed function, with possibly $f^0 \notin \mathscr{F}_M$. Suppose that for some $\eta > 0$,*

$$\|f - f^0\|_\infty \leq \eta, \ \forall f \in \mathscr{F}_M.$$

Define

$$\mathbf{Z}_M := \sup_{f \in \mathscr{F}_M} \left| \frac{1}{2}(Q_n - Q)\left((f - f^0)^2 - (f^* - f^0)^2\right) \right|.$$

Then for $t > 0$,

$$\mathbf{P}(\mathbf{Z}_M \geq M\alpha_X(t)) \leq \exp[-nt],$$

where

$$\alpha_X(t)/(2\eta) := \frac{2tK}{3} + 4\lambda_X + \sqrt{2t}\sqrt{1 + 8K\lambda_X},$$

with

$$\lambda_X := \lambda\left(\frac{K}{3}, n, p\right).$$

Proof. Consider the mapping

$$\gamma_f(x) := (f - f^0)^2(x)/2\eta, \ f \in \mathscr{F}_M.$$

Then

$$|\gamma_f - \gamma_{\tilde{f}}| \leq \left(\frac{|f - f^0 + \tilde{f} - f^0|}{2\eta}\right)|f - \tilde{f}| \leq |f - \tilde{f}|, \ \forall \ f, \tilde{f} \in \mathscr{F}_M.$$

Thus, by Symmetrization Theorem 14.3 and the Contraction Theorem 14.4,

$$\mathbb{E}\mathbf{Z}_M \leq 4\eta\mathbb{E}\left(\sup_{f \in \mathscr{F}_M} \left| \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i(f(X_i) - f^*(X_i)) \right|\right)$$

$$\leq 8\eta M\mathbb{E}\left(\max_{1 \leq j \leq p} \left| \frac{1}{n}\sum_{i=1}^{n} \psi_j(X_i) \right|\right)$$

$$\leq 8\eta M\left(\sqrt{\frac{2\log(2p)}{n}} + \frac{K\log(2p)}{3n}\right) = 8\eta M\lambda_X,$$

where in the last step we invoked Corollary 14.1.

Next, we use Bousquet's inequality. We first note that for all $f \in \mathscr{F}_M$,

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(\gamma_f - \gamma_{f^*})^2(X_i) \leq \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(f - f^*)^2(X_i) \leq M^2,$$

and

$$|\gamma_f - \gamma_{f^*}| \leq |f - f^*| \leq MK.$$

So for $t > 0$,

$$\mathbf{P}\left(\mathbf{Z}_M \geq \mathbb{E}\mathbf{Z}_M + \sqrt{2t}\sqrt{4\eta^2 M^2 + 4\eta MK\mathbb{E}\mathbf{Z}_M} + \frac{4\eta tMK}{3}\right) \leq \exp[-nt].$$

Now, insert $\mathbb{E}\mathbf{Z}_M \leq 8\eta M\lambda_X$. □

The above theorem can be improved in certain settings. Note that it makes use of the contraction inequality, which in turn is why we need the assumption that the class of functions is uniformly bounded (by η). We may use that when

$$\|\beta - \beta^*\|_1 \leq M, \ \max_{1 \leq j \leq p} \|\psi_j\|_\infty \leq K,$$

then

$$\|f_\beta - f_{\beta^*}\|_\infty \leq MK.$$

In the case of the Lasso with random design, one applies this with $K = O(\lambda s_*/\phi_*^2)$, so that the above approach leads to restrictions on the magnitude of s_* (modulo K, η and ϕ_*, and with $\lambda \asymp \sqrt{\log p/n}$, the sparsity s_* should generally be of small order $\sqrt{n/\log p}$).

We will quote a result of Guédon et al. (2007), which can be used to relax the conditions on the rate of growth of the sparsity. In Section 14.12, we define the δ-entropy $H(\cdot, \Lambda, d)$ of a subset of a metric space (Λ, d). The *entropy integral* is then

$$\mathscr{D}(\Lambda, d) := \int_0^{\text{Diam}(\Lambda)} \sqrt{H(u, \Lambda, d)} du,$$

whenever the integral is finite (and using a continuous majorant of the entropy if necessary). The entropy integral plays an important role in empirical process theory. We refer to Lemma 14.18 for an example. In the latter lemma, we use an entropy sum instead of an integral, but it is a common custom to replace the sum by the perhaps more elegant integral.

A collection $\mathscr{F}$ of real-valued functions on $\mathscr{X}$ can be equipped with the metric induced by the norm

$$\|f\|_{\infty,n} := \max_{1 \leq i \leq n} |f(X_i)|.$$

We then define

$$U_n^2(\mathscr{F}) := \mathbb{E}\mathscr{D}^2(\mathscr{F}, \|\cdot\|_{\infty,n}).$$

Theorem 14.6. *(Guédon et al., 2007).* *Let*

$$R := \sup_{f \in \mathscr{F}} \|f\|,$$

where $\|\cdot\|$ denotes the $L_2(Q)$-norm. Then for a universal constant c,

$$\mathbb{E} \sup_{f \in \mathscr{F}} \left| (Q_n - Q)f^2 \right| \leq c \max\left\{ RU_n(\mathscr{F})/\sqrt{n}, U_n^2(\mathscr{F})/n \right\}.$$

The above result is invoked in Bartlett et al. (2009) to answer a question posed in Greenshtein and Ritov (2004): see also the discussion in Section 2.4.

14.9.3 Squared error loss

Let $Y_i = f^0(X_i) + \varepsilon_i$, $i = 1, \ldots, n$. Define

$$\gamma_f(x, y) = (y - f(x))^2, \ x \in \mathscr{X}, \ y \in \mathbb{R}.$$

Lemma 14.23. *Let $\mathscr{F}_M := \{f_\beta = \sum_{j=1}^p \beta_j \psi_j : \ \|\beta - \beta^*\|_1 \leq M\}$ be a class of linear functions on $\mathscr{X}$, and let f^0 be a fixed function, with possibly $f^0 \notin \mathscr{F}_M$. Suppose that*

$$\|f - f^0\|_\infty \leq \eta, \ \forall \ f \in \mathscr{F}_M.$$

Then for $t > 0$,

$$\mathbf{P}\left(\sup_{f_\beta \in \mathscr{F}_M} |(P_n - P)(\gamma_{f_\beta} - \gamma_{f^*})| \geq M(\alpha_\varepsilon(t) + \alpha_X(t)) \right) \leq 2 \exp[-nt],$$

where

$$\alpha_\varepsilon(t) := K_0 t + \lambda_\varepsilon + \sigma\sqrt{2t},$$

with

$$\lambda_\varepsilon := \sigma\lambda\left(\frac{K_0}{\sigma}, n, p\right),$$

and where

$$\alpha_X(t)/(2\eta) := \frac{2tK}{3} + 4\lambda_X + \sqrt{2t}\sqrt{1 + 8K\lambda_X},$$

with

$$\lambda_X := \lambda\left(\frac{K}{3}, n, p\right).$$

Proof. This follows from combining the previous two subsections. □

14.10 Assuming only lower order moments

We show that by symmetrization, one can prove moment and probability inequalities using only second moments. We restrict ourselves to a finite class of functions $\{\gamma_1, \ldots, \gamma_p\}$ (infinite classes can often be treated using e.g. entropy calculations, see Section 14.11 for an illustration).

14.10.1 Nemirovski moment inequality

We prove an inequality for the m-th moment of maxima of sums of independent random variables. The case $m = 2$ is considered in Dümbgen et al. (2010). It is - modulo constants - Nemirovski's inequality (the latter actually concerns the second moment of ℓ_q-norms ($1 \leq q \leq \infty$) of sums of independent random variables in $\mathbb{R}^p$, whereas we only consider the case $q = \infty$).

Lemma 14.24. (Nemirovski moment inequality) *For $m \geq 1$ and $p \geq \mathrm{e}^{m-1}$, we have*

$$\mathbb{E} \max_{1 \leq j \leq p} \left| \sum_{i=1}^{n} \left(\gamma_j(Z_i) - \mathbb{E}\gamma_j(Z_i) \right) \right|^m \leq \left[8 \log(2p) \right]^{m/2} \mathbb{E} \left[\max_{1 \leq j \leq p} \sum_{i=1}^{n} \gamma_j^2(Z_i) \right]^{m/2}.$$

Proof. Let $(\varepsilon_1, \ldots, \varepsilon_n)$ be a Rademacher sequence independent of $\mathbf{z} := (Z_1, \ldots, Z_n)$. Let $\mathbb{E}_{\mathbf{z}}$ denote conditional expectation given $\mathbf{z}$. By Hoeffding's moment inequality (Lemma 14.14), conditionally on $\mathbf{z}$,

$$\mathbb{E}_{\mathbf{z}} \max_{1 \leq j \leq p} \left| \sum_{i=1}^{n} \left(\varepsilon_i \gamma_j(Z_i) \right) \right|^m \leq \left[2 \log(2p) \right]^{m/2} \max_{1 \leq j \leq p} \left[\sum_{i=1}^{n} \gamma_j^2(Z_i) \right]^{m/2}.$$

Hence,

$$\mathbb{E} \max_{1 \leq j \leq p} \left| \sum_{i=1}^{n} \varepsilon_i \gamma_j(Z_i) \right|^m \leq \left[2 \log(2p) \right]^{m/2} \mathbb{E} \left[\max_{1 \leq j \leq p} \sum_{i=1}^{n} \gamma_j^2(Z_i) \right]^{m/2}.$$

Finally, we desymmetrize (see Theorem 14.3):

$$\left(\mathbb{E} \max_{1 \leq j \leq p} \left| \sum_{i=1}^{n} (\gamma_j(Z_i) - \mathbb{E}\gamma_j(Z_i)) \right|^m \right)^{1/m} \leq 2 \left(\mathbb{E} \max_{1 \leq j \leq p} \left| \sum_{i=1}^{n} \varepsilon_i \gamma_j(Z_i) \right|^m \right)^{1/m}.$$

$\square$

Example 14.3. Consider independent centered random variables $\varepsilon_1, \ldots, \varepsilon_n$, with variance $\mathbb{E}\varepsilon_i^2 \leq \sigma^2$ for all i. Moreover, let $\{x_{i,j} : i = 1, \ldots, n, \ j = 1, \ldots, p\}$ be given constants. Define for $i = 1, \ldots, n$,

$$K_i := \max_{1 \leq j \leq p} |x_{i,j}|.$$

Then clearly

$$\mathbb{E} \left[\max_{1 \leq j \leq p} \sum_{i=1}^{n} \varepsilon_i^2 x_{i,j}^2 \right] \leq \sigma^2 \sum_{i=1}^{n} K_i^2.$$

Lemma 14.24 shows therefore that for $p \geq \mathrm{e}$,

$$\mathbb{E} \max_{1 \leq j \leq p} \left| \sum_{i=1}^{n} \varepsilon_i x_{i,j}/n \right|^2 \leq \sigma^2 \left[\frac{8 \log(2p)}{n} \right] \left[\frac{\sum_{i=1}^{n} K_i^2}{n} \right].$$

14.10.2 A uniform inequality for quadratic forms

The next result can be used for example in the context of multivariate regression with group Lasso, as considered in Section 8.6.

Lemma 14.25. *Let* $\{\varepsilon_{i,t} : i = 1, \ldots, n, \ t = 1, \ldots, T\}$ *be independent random variables with* $\mathbb{E}\varepsilon_{i,t} = 0$, $\mathbb{E}\varepsilon_{i,t}^2 = 1$, *and* $\mathbb{E}\varepsilon_{i,t}^4 \leq \mu_4^4$ *for all i and t. Moreover. let* $\{x_{i,j,t} : i = 1, \ldots, n, \ j = 1, \ldots, p, \ t = 1, \ldots, T\}$ *be given constants satisfying* $\sum_{i=1}^{n} x_{i,j,t}^2 = n$ *for all j and t. Then for* $p \geq e^3$, *we have*

$$\mathbb{E} \max_{1 \leq j \leq p} \left[\sum_{t=1}^{T} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_{i,j,t} \varepsilon_{i,t} \right)^2 \right]^2$$

$$\leq \left\{ T + \sqrt{T} \left[8 \log(2p) \right]^{3/2} \mu_4^2 \left[\frac{1}{nT} \sum_{t=1}^{T} \sum_{i=1}^{n} \max_{1 \leq j \leq p} x_{i,j,t}^4 \right]^{1/2} \right\}^2.$$

Proof. Define for all j and t,

$$V_{j,t} := \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_{i,j,t} \varepsilon_{i,t} \right)^2,$$

and let $V_t := (V_{1,t}, \ldots, V_{p,t}), t = 1, \ldots T$. Then $V_1, \ldots, V_T$ are independent, and $\mathbb{E}V_{j,t} = 1$ for all t and j. Hence, by the Nemirovski moment inequality given in the previous section (Lemma 14.24), for $p \geq e$

$$\mathbb{E} \max_{1 \leq j \leq p} \left| \sum_{t=1}^{T} (V_{j,t} - 1) \right|^2 \leq 8 \log(2p) \sum_{t=1}^{T} \mathbb{E} \left[\max_{1 \leq j \leq p} V_{j,t}^2 \right].$$

Also, applying again the Nemirovski moment inequality, we obtain for $p \geq e^3$, and for all t,

$$\mathbb{E} \max_{1 \leq j \leq p} V_{j,t}^2 = \mathbb{E} \max_{1 \leq j \leq p} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_{i,j,t} \varepsilon_{i,t} \right)^4$$

$$\leq \left[8 \log(2p) \right]^2 \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^{n} \left(\max_{1 \leq j \leq p} x_{i,j,t}^2 \varepsilon_{i,t}^2 \right) \right]^2$$

$$\leq \left[8\log(2p)\right]^2 \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\max_{1\leq j\leq p} x_{i,j,t}^4 \varepsilon_{i,t}^4\right)\right]$$

$$\leq \left[8\log(2p)\right]^2 \mu_4^4\left[\frac{1}{n}\sum_{i=1}^{n}\max_{1\leq j\leq p} x_{i,j,t}^4\right].$$

$\square$

14.11 Using entropy for concentration in the sub-Gaussian case

Consider independent centered random variables $\varepsilon_1, \ldots, \varepsilon_n$, that are sub-Gaussian: for some constants K and σ_0^2,

$$K^2(\mathbb{E}\exp[\varepsilon_i^2/K^2] - 1) \leq \sigma_0^2, \ i = 1, \ldots, n. \tag{14.15}$$

Later, we will use the short-hand notation

$$K_0/3 := 2^5\sqrt{K^2 + \sigma_0^2}.$$

Let $\{x_1, \ldots, x_n\} \subset \mathcal{X}$ be fixed. Write

$$Q_n := \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i},$$

and let $\mathcal{F} \subset L_2(Q_n)$. Denote the $L_2(Q_n)$-norm by $\|\cdot\|_n$. Throughout this section, we assume that

$$\sup_{f\in\mathcal{F}}\|f\|_n \leq 1. \tag{14.16}$$

We consider the process

$$\{(\varepsilon, f)_n : f \in \mathcal{F}\},$$

where $(\varepsilon, f)_n := \sum_{i=1}^{n}\varepsilon_i f(x_i)/n$.

We note that $\varepsilon_i f(x_i)$ is a sub-Gaussian random variable: for $i = 1, \ldots, n$,

$$K_i^2(\mathbb{E}\exp[\varepsilon_i^2 f^2(x_i)/K_i^2] - 1) \leq \sigma_{0,i}^2,$$

where $K_i^2 = K^2 f^2(x_i)$ and $\sigma_{0,i}^2 = \sigma_0^2 f^2(x_i)$.

In order to be able to use the result of Section 14.5.3, which is for finite classes of functions, we approximate $\mathcal{F}$ by a finite sub-class.

Definition *For $\delta > 0$, the δ-covering number $N(\delta, \mathcal{F}, \|\cdot\|_n)$ of $(\mathcal{F}, \|\cdot\|_n)$ is the minimum number of balls with radius δ necessary to cover the class $\mathcal{F}$. The* entropy *is $H(\cdot, \mathcal{F}, \|\cdot\|_n) := \log N(\cdot, \mathcal{F}, \|\cdot\|_n)$.*

We will again apply the chaining argument. We combine chaining with Lemma 14.17. In Lemma 14.17, we choose $S = \infty$, which is allowed if the infinite sum

$$\sum_{s=1}^{\infty} 2^{-s} \sqrt{\log(1 + 2N(2^{-s}, \mathscr{F}, \|\cdot\|_n))},$$

converges, which we will assume implicitly. This implicit assumption can actually be avoided using the Cauchy-Schwarz inequality

$$|(\varepsilon, f)_n| \leq \|\varepsilon\|_n \|f\|_n,$$

where (with some abuse of notation)

$$\|\varepsilon\|_n^2 := \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i^2.$$

One can show that the chaining argument only needs a finite number S of links, with $2^{-S}R \asymp 1/\sqrt{n}$. To avoid digressions, we do not treat this in detail here, see also Lemma 14.18.

Theorem 14.7. *Assume (14.15) and (14.16). Let, for each $s \in \{0, 1, \ldots\}$, $\mathscr{F}_s := \{f_j^s\}_{j=1}^{N_s} \subset \mathscr{F}$ be a minimal 2^{-s}-covering set of $(\mathscr{F}, \|\cdot\|_n)$. Define*

$$a_0 := \sum_{s=1}^{\infty} 2^{-s-1} \sqrt{\log(1 + 2N_s) \vee s},$$

and

$$\mathbf{Z} := \sup_{f \in \mathscr{F}} |(\varepsilon, f)_n|.$$

It holds that

$$\mathbb{E} \exp\left[\left(\sqrt{n}\mathbf{Z}/K_0 - a_0 \right)_+ \right]^2 \leq 1 + 2\exp[-2a_0^2]. \tag{14.17}$$

Theorem 14.7 is of the same spirit as the results in Viens and Vizcarra (2007). The latter paper is however more general, as it considers more general tail behavior. Moreover, it replaces the bound a_0 in the left-hand side of (14.17) by $\mathbb{E}\mathbf{Z}$. In applications, one then needs to derive a bound for $\mathbb{E}\mathbf{Z}$ by separate means.

Proof of Theorem 14.7. For all $f \in \mathscr{F}$ and all s, there is a $g_f^s \in \mathscr{F}_s$ such that $\|f - g_f^s\|_n \leq 2^{-s}$. Hence we have for all $s \in \{1, 2, \ldots\}$,

$$(\varepsilon, f)_n = \sum_{s=1}^{\infty} (\varepsilon, g_f^s - g_f^{s-1})_n + (\varepsilon, f - g_f^S)_n,$$

where we take $g_f^0 \equiv 0$. Define

$$X_s := \max_{f \in \mathscr{F}} \sqrt{n} |(\varepsilon, g_f^s - g_f^{s-1})_n|.$$

Note that
$$\|g_f^s - g_f^{s-1}\|_n \leq \|g_f^s - f\|_n + \|f - g_f^{s-1}\|_n \leq 3 \times 2^{-s}.$$

Moreover card($\{g_f^s - g_f^{s-1}\}$) $\leq N_s N_{s-1} \leq N_s^2$. Lemma 14.16 therefore tells us that, for $\delta_s := 12\sqrt{K^2 + \sigma_0^2} 2^{-s}$,

$$\mathbb{E}\exp[X_s^2/\delta_s^2] \leq 1 + 2N_s^2.$$

Combine this with Lemma 14.17, and the bounds $\log(1 + 2N_s^2) \leq 2\log(1 + 2N_s)$, and

$$\sum_{s=1}^{S} 2^{-s}\sqrt{s} \leq 2,$$

to obtain

$$\mathbb{E}\exp\left[\left(\sum_{s=1}^{S} X_s/K_0 - a_0\right)_+\right]^2 \leq 1 + 2\exp[-2a_0^2].$$

$\square$

Let for all $R > 0$,
$$\mathscr{F}(R) := \{f \in \mathscr{F} : \|f\|_n \leq R\}.$$

The following corollary shows how the empirical process behaves as a function of the radius R.

Define
$$a_0(R) := \sum_{s=1}^{\infty} 2^{-s-1}R\sqrt{\log(1 + 2N(2^{-s}R, \mathscr{F}(R), \|\cdot\|_n) \vee s},$$

and
$$\mathbf{Z}(R) := \sup_{f \in \mathscr{F}(R)} |(\varepsilon, f)_n|.$$

Corollary 14.5. *Assume (14.15). For all R, it holds that*

$$\mathbb{E}\exp\left[\frac{(\sqrt{n}\mathbf{Z}(R)/K_0 - a_0(R))_+}{R}\right]^2 \leq 1 + 2\exp\left[-\frac{2a_0^2(R)}{R^2}\right].$$

We will now turn to the weighted empirical process.

Lemma 14.26. *Assume (14.15) and (14.16). Suppose that $a(R) \geq a_0(R)$ is such that $a(R)$ in non-decreasing in R and $a(R)/R$ is non-increasing in R ($0 < R \leq 1$), and that*
$$B := \sum_{s=1}^{\infty} \exp[-a^2(2^{-s})/2^{-2s}] < \infty.$$

Then

$$\mathbb{E}\exp\left[\sup_{f\in\mathscr{F}}\left[\frac{(\sqrt{n}|(\varepsilon,f)_n|/K_0-2a(\|f\|_n))_+}{2\|f\|_n}\right]^2\right]\leq 1+2B$$

Proof. We can assume $\|f\|_n > 0$ for all $f \in \mathscr{F}$. Let $s \in \{0,1,\dots\}$. Because $a(R)/R$ is non-increasing in R, it holds that

$$a(2^{-s}) = \frac{a(2^{-s})}{2^{-s}}2^{-s} \leq \frac{a(2^{-s-1})}{2^{-s-1}}2^{-s} = 2a(2^{-s-1}).$$

Hence

$$\sup_{f\in\mathscr{F},\,2^{-s-1}\leq\|f\|_n\leq 2^{-s}}\left[\frac{(\sqrt{n}|(\varepsilon,f)_n|/K_0-2a(\|f\|_n))_+}{2\|f\|_n}\right]$$

$$\leq \sup_{f\in\mathscr{F},\,\|f\|_n\leq 2^{-s}}\left[\frac{(\sqrt{n}|(\varepsilon,f)_n|/K_0-2a(2^{-s-1}))_+}{2^{-s}}\right]$$

$$\leq \sup_{f\in\mathscr{F},\,\|f\|_n\leq 2^{-s}}\left[\frac{(\sqrt{n}|(\varepsilon,f)_n|/K_0-a(2^{-s}))_+}{2^{-s}}\right].$$

We now use the so-called peeling device (see van de Geer (2000), and the references therein). In view of Corollary 14.5 we then get

$$\mathbb{E}\exp\left[\sup_{f\in\mathscr{F}}\left[\frac{(\sqrt{n}|(\varepsilon,f)_n|/K_0-2a(\|f\|_n))_+}{2\|f\|_n}\right]^2\right]-1$$

$$\leq \sum_{s=0}^{\infty}\left\{\mathbb{E}\exp\left[\sup_{f\in\mathscr{F},\,2^{-s-1}<\|f\|_n\leq 2^{-s}}\left[\frac{(\sqrt{n}|(\varepsilon,f)_n|/K_0-2a(\|f\|_n))_+}{2\|f\|_n}\right]^2\right]-1\right\}$$

$$\leq \sum_{s=0}^{\infty}\left\{\mathbb{E}\exp\left[\sup_{f\in\mathscr{F},\,\|f\|_n\leq 2^{-s}}\left[\frac{(\sqrt{n}|(\varepsilon,f)_n|/K_0-2a(2^{-s}))_+}{2^{-s}}\right]^2\right]-1\right\}$$

$$\leq 2\sum_{s=0}^{\infty}\exp\left[-\frac{2a^2(2^{-s})}{2^{-2s}}\right] = 2B.$$

$\square$

Corollary 14.6. *Assume (14.15). Let $\mathscr{F}$ be a class of functions with $\|f\|_n \leq 1$ for all $f \in \mathscr{F}$, and with, for some $0 < v < 1$ and some constant A_v,*

$$\log(1+2N(\delta,\mathscr{F},\|\cdot\|_n)) \leq \left(\frac{A_v}{\delta}\right)^{2v},\ 0 < \delta \leq 1.$$

Then

$$a_0(R) \leq \frac{1}{2}A_v^v R^{1-v}\sum_{s=1}^{\infty}2^{-s(1-v)} = \frac{1}{2}A_v^v R^{1-v}(2^{1-v}-1)^{-1} := a(R).$$

Moreover

$$\sum_{s=0}^{\infty} \exp\left[-\frac{2a^2(2^{-s})}{2^{-2s}}\right]$$

$$= \sum_{s=0}^{\infty} \exp\left[-\frac{A_v^{2v}2^{2vs}}{2(2^{1-v}-1)^2}\right] \le \sum_{s=0}^{\infty} \exp\left[-\frac{A_v^{2v}v(s+1)}{2(2^{1-v}-1)^2}\right]$$

$$= \left(\exp\left[\frac{A_v^{2v}v}{2(2^{1-v}-1)^2}\right]-1\right)^{-1} := B_v.$$

Lemma 14.26 for this case gives

$$\mathbb{E}\exp\left[\sup_{f\in\mathscr{F}}\left[\left(\frac{(\sqrt{n}|(\varepsilon,f)_n|}{\|f\|_n K_0} - \frac{A_v^v\|f\|_n^{-v}}{2^{1-v}-1}\right)_+\right]^2\right]$$

$$\le 1 + 2B_v.$$

Chebyshev's inequality shows that for all $t > 0$,

$$\mathbf{P}\left(\exists f \in \mathscr{F}: \sqrt{n}|(\varepsilon,f)_n| \ge K_0 A_v^v\|f\|_n^{1-v}(2^{1-v}-1)^{-1} + K_0\|f\|_n t\right)$$

$$\le \exp[-t^2/K_0^2](1+2B_v).$$

Corollary 14.7. *The result of the present corollary can be used in Subsection 6.11.1, which concerns the estimator with ℓ_r-penalty, $0 < r < 1$. Assume (14.15). Let, for a given $0 < r < 1$, and a given dictionary $\{\psi_j\}_{j=1}^p$ with $\|\psi_j\|_n \le 1$,*

$$\mathscr{F} := \{f_\beta = \sum\beta_j\psi_j : \|\beta\|_r \le 1\}.$$

Then from Lemma 14.32, for

$$\alpha := \frac{r}{2(1-r)}.$$

and for $2p^{-(1+\alpha)}/\sqrt{\alpha} \le \delta \le 2(\sqrt{\alpha})^{\frac{1}{\alpha}}$, we have

$$\log(1 + 2N(\delta,\mathscr{F},\|\cdot\|_n)) \le 4(1+\alpha)(\alpha\delta^2/4)^{-\frac{r}{2-r}}\log(1+2p)$$

$$:= \left(\frac{c_r^{\frac{2-r}{r}}\log^{\frac{2-r}{2r}}(1+2p)}{\delta}\right)^{\frac{2r}{2-r}}.$$

Values of δ outside the above range can be handled easily. So, in the notation of Corollary 14.6, the constants are $v = r/(2-r)$, and $A_v^v = c_r\sqrt{\log(1+2p)}$. Define

$$b_{r,p} := \left(\exp\left[\frac{c_r^2\log(1+2p)\frac{r}{2-r}}{2(2^{\frac{2(1-r)}{2-r}}-1)^2}\right]-1\right)^{-1},$$

and let

$$\mathscr{T} := \left\{ \forall\, f \in \mathscr{F} : \ \sqrt{n}|(\varepsilon,f)_n| \leq \frac{K_0\sqrt{\log(1+2p)}c_r}{2^{\frac{2(1-r)}{2-r}} - 1}\|f\|_n^{\frac{2(1-r)}{2-r}} + K_0\|f\|_n t \right\}.$$

Then from Corollary 14.6,

$$\mathbf{P}(\mathscr{T}) \geq 1 - \exp[-t^2/K_0^2](1+2b_{r,p}).$$

Note that $b_{r,p} \to 0$ as $p \to \infty$. The probability of $\mathscr{T}$ is large for large t.

Corollary 14.8. *Assume (14.15). Fix an $m \in \mathbb{N}$. Let*

$$\mathscr{F} := \{f : [0,1] \to \mathbb{R} : \ \|f\|_n \leq 1, \ \mathrm{TV}(f^{(m-1)}) \leq 1\},$$

where $\mathrm{TV}(g) = \int |g|$ is the total variation of the function g. We refer to Section 14.12.6 for more details. The following result can be invoked in Subsection 8.4.2, where the high-dimensional additive model is studied. There, for a function $f : [0,1] \to \mathbb{R}$, the squared Sobolev semi-norm $\int |f^{(m)}(x)|^2 dx$ is used as a measure of smoothness. It is clear that

$$\left\{ f : \ \|f\|_n \leq 1, \ \int |f^{(m)}(x)|^2 dx \leq 1 \right\} \subset \mathscr{F}.$$

Theorem 14.10 shows that for some constant c_m,

$$\log(1 + 2N(\delta, \mathscr{F}, \|\cdot\|_n)) \leq 4\left(\frac{c_m}{\delta}\right)^{\frac{1}{m}}, \ \delta > 0,$$

i.e., in the notation of Corollary 14.6, the constants are $\nu = 1/(2m)$ and $A_\nu = 2^{\frac{1}{m}}c_m$. In the same way as in the previous corollary, one may conclude that

$$\sup_{f \in \mathscr{F}} \frac{\sqrt{n}|(\varepsilon,f)_n|}{\|f\|_n^{1-\frac{1}{2m}}} = O_\mathbf{P}(1).$$

14.12 Some entropy results

We derive entropies for classes of functions relevant for our work. These can be used in the maximal inequalities given in Section 14.6.3 to arrive at concrete bounds for special cases. Our main focus will be on linear classes of functions, with an ℓ_1-bound on the coefficients.

In Subsection 14.12.1, we give an entropy bound for a ball in finite-dimensional space endowed with ℓ_q-norm ($1 \leq q \leq \infty$). We also present there the entropy of

the convex hull of a given set $\{\psi_j\} \subset L_2(Q)$ of functions. Subsection 14.12.2 refines this to the case where the coefficients are restricted to lie in some smaller set, such as an ℓ_r-ball with $0 < r < 1$. Subsections 14.12.3 and 14.12.4 study the case where the functions $\{\psi_j\}$ are (highly) correlated, and in fact can be approximated by a smaller δ-covering. Subsection 14.12.3 gives a relatively simple bound for the entropy involving a logarithmic term, whereas Subsection 14.12.4 shows that the logarithmic term can be removed in certain cases, but that one then possibly pays a price in the constants. Subsection 14.12.5 provides some important refinements. Subsection 14.12.6 considers an example and Subsection 14.12.7 has the proofs for this section.

The definition of entropy, and related concepts, is as follows.

Let (Λ, d) be a subset of a metric space.

Definition *For $\delta > 0$, the δ-capacity $\mathscr{C}(\delta, \Lambda, d)$ of Λ is the maximal number of elements of a δ-packing set, that is, of a subset Λ_δ of Λ having each pair of distinct elements λ_j and λ_k at least δ apart (i.e., $d(\lambda_j, \lambda_k) > \delta$).*

The δ-covering number $N(\delta, \Lambda, d)$ of Λ is the smallest number of closed balls with radius δ, that covers the space Λ. We call the centers of the balls a δ-covering set. The entropy *of Λ is $H(\cdot, \Lambda, d) := \log N(\cdot, \Lambda, d)$.*

It is sometimes useful to require that the covering sets are within Λ. This can be always be accomplished: from a general δ-covering set we can construct a 2δ-covering set within Λ. Note moreover that a maximal δ-packing set is also a δ-covering set and that 2δ-packing set is never larger than a δ-covering set.

We generally consider sets Λ for which there indeed exist finite coverings (such sets are called *totally bounded*).

In what follows, Q is a probability measure on the space $\mathscr{X}$, $\|\cdot\|$ is the $L_2(Q)$-norm, and $\{\psi_j\}_{j=1}^p \subset L_2(Q)$ are given functions with norm $\|\psi_j\| \leq 1$, for all j. For $\beta \in \mathbb{R}^p$, we write

$$f_\beta = \sum_{j=1}^p \beta_j \psi_j.$$

Entropy results for the class

$$\{f_\beta : \|\beta\|_1 = 1\}$$

can be deduced from those for the convex hull

$$\mathrm{conv}(\{\psi_j\}) := \{f_\beta : \beta_j \geq 0, \|\beta\|_1 = 1\}$$

by separating the positive and negative coefficients:

$$f_\beta = \sum_{\beta_j > 0} \beta_j \psi_j + \sum_{\beta_j < 0} \beta_j \psi_j.$$

Moreover, the class

$$\{f_\beta : \|\beta\|_1 \le 1\}$$

(i.e., the coefficients are in the ℓ_1-ball, including the interior) can be handled by adding the additional function $\psi_0 \equiv 0$ to the set $\{\psi_j\}_{j=1}^p$.

We use the notation: for $x > 0$

$\lceil x \rceil$ is the smallest integer larger than or equal to x

$\lfloor x \rfloor$ is the largest integer (including 0) smaller than or equal to x.

14.12.1 Entropy of finite-dimensional spaces and general convex hulls

Lemma 14.27. *Endow $\mathbb{R}^p$ with the metric corresponding to the ℓ_q-norm, where $1 \le q \le \infty$. Consider the ℓ_q-ball*

$$\Theta := \{\theta \in \mathbb{R}^p : \|\theta\|_q \le R\}.$$

Then

$$\mathcal{C}(\delta, \Theta, \|\cdot\|_q) \le \left(\frac{2R+\delta}{\delta}\right)^p, \quad 0 < \delta \le R.$$

We next consider a lemma from Pollard (1990): Theorem 6.2, and van der Vaart and Wellner (1996): Lemma 2.6.11.

Lemma 14.28. *For all $\delta > 0$, we have*

$$H\left(\delta, \text{conv}(\{\psi_j\}_{j=1}^p), \|\cdot\|\right)$$

$$\le \left\lceil \frac{1}{\delta^2} \right\rceil \left(1 + \log\left(1 + p\delta^2\right)\right) \wedge \left\lceil \frac{1}{\delta^2} \right\rceil \log p.$$

14.12.2 Sets with restrictions on the coefficients

Lemma 14.29. *Let $\mathcal{B} \subset \mathbb{R}^p$. Then for all $\delta > 0$,*

$$H\left(\delta, \{f_\beta : \beta \in \mathcal{B}\}, \|\cdot\|\right) \le \min_{u>0} \left\{ H(u, \mathcal{B}, \|\cdot\|_1) + \left\lceil \frac{u^2}{\delta^2} \right\rceil \log(1 + 2p) \right\}.$$

As a special case, we consider coefficients that are required to lie within an ℓ_r-ball, $0 < r < 1$.

Lemma 14.30. *For a given $0 < r < 1$, let $\mathcal{B}$ be the set $\mathcal{B} := \{\beta \in \mathbb{R}^p : \|\beta\|_r^r \leq 1\}$. Then for all $u > 0$,*

$$H(u, \mathcal{B}, \|\cdot\|_1) \leq \left(\frac{2}{u}\right)^{\frac{r}{1-r}} \left(\log p + \log\left(\frac{4+u}{u}\right)\right).$$

We now give a bound for the entropy which approximately does the minimization as given in Lemma 14.29. To facilitate these derivations, we present (without proof) the following straightforward result.

Lemma 14.31. *Define for $x > 0$, and for positive constants c and α, the function*

$$f(x) := x^{-\alpha} + x/c.$$

Then

$$\min_{x > 0} f(x) = (\alpha + 1)(\alpha c)^{-\frac{\alpha}{1+\alpha}},$$

and

$$\arg\min_{x > 0} f(x) = (\alpha c)^{\frac{1}{1+\alpha}}.$$

The combination of Lemma 14.29 and Lemma 14.30 results in the following bound.

Lemma 14.32. *For a given $0 < r < 1$, let $\mathcal{B}$ be the set $\mathcal{B} := \{\beta \in \mathbb{R}^p : \|\beta\|_r^r \leq 1\}$. Define*

$$\alpha := \frac{r}{2(1-r)}.$$

Then, for $2p^{-(1+\alpha)}/\sqrt{\alpha} \leq \delta \leq 2(\sqrt{\alpha})^{\frac{1}{\alpha}}$, we have

$$H\left(\delta, \{f_\beta : \beta \in \mathcal{B}\}, \|\cdot\|\right) \leq 4(1+\alpha)(\alpha\delta^2/4)^{-\frac{r}{2-r}} \log(1+2p).$$

14.12.3 Convex hulls of small sets: entropy with log-term

We consider sets of functions $\{\psi_j\}$ with relatively small covering number $N(\cdot, \{\psi_j\}, \|\cdot\|)$, and examine the entropy for the convex hull $\text{conv}(\{\psi_j\})$ of $\{\psi_j\}$. Dudley (1987) gives a bound for this entropy for the case where the u-covering number of $\{\psi_j\}$ is a polynomial in $1/u$. Its derivation, as given in Pollard (1990) is less complicated than the one of the next subsection, but the result may involve a redundant logarithmic term. We reprove the result of Pollard (1990) here, using the same technique of proof, and extending it to general covering numbers (i.e., not only polynomial ones).

The next lemma is the core of the result, as in Pollard (1990): Theorem 6.2.

Lemma 14.33. *Consider functions* $g_a = (g_{a_1}, \ldots g_{a_N})^T$, *of the form*

$$g_{a_k} = \sum_j \alpha_{j,k} \phi_{j,k},$$

where $a_k = \{\alpha_{j,k}\}$, *with* $\alpha_{j,k} \geq 0$ *for all* j *and* k, *and* $\sum_j \alpha_{j,k} = 1$ *for all* k. *Moreover, the* $\phi_{j,k}$ *are given functions with* $\|\phi_{j,k}\| \leq u$ *for all* j *and* k. *The total number as* j *and* k *vary is denoted by* $p := \mathrm{card}(\{\phi_{j,k}\})$. *Fix a* $\theta \in \mathbb{R}^N$ *satisfying* $\theta_k \geq 0$ *for all* k, *and* $\sum_{k=1}^N \theta_k = 1$. *Let* $\mathcal{G}_\theta := \{g = \sum_{k=1}^N \theta_k g_{a_k}\}$. *Then*

$$H(\delta, \mathcal{G}_\theta, \|\cdot\|) \leq \left(\frac{u^2}{\delta^2} + N \right) \log p.$$

As a consequence of Lemma 14.33, we get a bound for the entropy, again (as in Lemma 14.29) by trading off the u-covering number against the squared radius u^2.

Lemma 14.34. *Define* $N(u) := N(u, \{\psi_j\}, \|\cdot\|)$, $u > 0$. *We have*

$$H\left(3\delta, \mathrm{conv}(\{\psi_j\}), \|\cdot\| \right)$$

$$\leq \min_{u > 0} \left\{ \left(3N(u) + \frac{4u^2}{\delta^2} \right) \log \left(\left(\frac{8+\delta}{\delta} \right) N(\delta) \right) \right\}.$$

14.12.4 Convex hulls of small sets: entropy without log-term

This subsection considers the same problem as the previous one. We now do our calculations in such a way to possibly remove the logarithmic term in certain cases. This allows one to recover known results from approximation theory for a large class of function spaces. An example will be the space of $(m-1)$-times differentiable functions, with the $(m-1)$-th derivative being of bounded variation, see Subsection 14.12.6.

The results of this section are inspired by Ball and Pajor (1990) and van der Vaart and Wellner (1996). For the proofs, the first ingredient is along the lines of Lemma 2.6.11 in van der Vaart and Wellner (1996).

In what follows, we consider a non-increasing sequence of positive numbers $\{u_s\}_{s=0}^\infty$, and let $\{\psi_k^s\}$ be u_s-covering sets of $\{\psi_j\}$, with cardinality $N_s := |\{\psi_k^s\}|$. Then for each j, we can consider the closest neighbour of ψ_j in the u_s-covering set, say $\psi_{k_j^s}^s$.

It is clear that then, for fixed $s \geq 1$, the number of functions $\{\psi_{k_j^s}^s - \psi_{k_j^s}^{s-1}\}$ as j varies is at most $N_s N_{s-1}$. We now show that this number can brought down to N_s, by agreeing upon a *tree structure* for the sequence of covering sets.

Definition *Let* $\{u_s\}_{s=0}^{\infty}$ *be a non-increasing sequence of positive numbers,* $\{\psi_k^s\}$ *be* u_s-*covering sets of* $\{\psi_j\}$, *with cardinality* $N_s := |\{\psi_k^s\}|$. *For any* $s \geq 1$, *and any* $k \in \{1, \ldots, N_{s-1}\}$, *we define the* off-spring V_k^s *of* k *at generation* s *as the indices of the set of functions in* $\{\psi_j^s\}$ *which are closest to* ψ_k^{s-1}, *i.e.*

$$V_k^s := \left\{ j : \|\psi_j^s - \psi_k^{s-1}\| = \min_{l \in \{1, \ldots, N_{s-1}\}} \|\psi_j^s - \psi_l^{s-1}\| \right\}.$$

The generation tree *is the collection of mappings*

$$g^s : \{1, \ldots, N_s\} \rightarrow \{1, \ldots, N_{s-1}\}, \ s = 1, 2, \ldots,$$

of off-spring to parent, defined as

$$g^s(j) = k \text{ if } j \in V_k^s.$$

Lemma 14.35. *(Generation Tree Lemma) Let* $\{u_s\}_{s=0}^{\infty}$ *be a non-increasing sequence of positive numbers, with* $u_0 = 1$, *and let* $\{\psi_k^s\} \subset \{\psi_j\}$ *be* u_s-*covering sets of* $\{\psi_j\}$, *with cardinality* $N_s := |\{\psi_k^s\}|$. *Take* $\psi_1^0 \equiv 0$. *Then*

$$\|\psi_k^s - \psi_{g^s(k)}^{s-1}\| \leq u_{s-1}, \ s = 1, 2, \ldots.$$

Moreover, for each generation T, ψ_j *can be decomposed as*

$$\psi_j = \sum_{s=1}^{T} (\psi_{k_j^s}^s - \psi_{k_j^{s-1}}^{s-1}) + (\psi_j - \psi_{k_j^T}^T),$$

where $\{k_j^s\}_{s=1}^{T-1}$ *follows the branches of the tree (i.e.* $k_j^{s-1} = g^s(k_j^s)$, $s = 1, \ldots, T-1$), *and where*

$$\|\psi_j - \psi_{k_j^T}^T\| \leq u_T.$$

The Generation Tree Lemma plays an important role in the proof of the next lemma.

Lemma 14.36. *Define (for* $s \in \{0, 1, \ldots\}$) $\delta_s = 2^{-s}$, *and let* $\{\psi_j^s\}_{j=1}^{N_s} \subset \{\psi_j\}$ *be a* u_s-*covering set of* $\{\psi_j\}$, *where* u_s *is non-increasing in* s *and where*

$$2u_s^2/\delta_s^2 \leq N_s \leq 4u_s^2/\delta_s^2.$$

Define for all s,

$$\mathscr{F}_s := \mathrm{conv}(\{\psi_j^s\}).$$

Then

$$H(\delta_s, \mathscr{F}_s, \|\cdot\|) \leq \log\left(9[2e]^9\right) \sum_{t=1}^{s} N_{t-1}.$$

Our second ingredient uses a so-called "chaining" argument to derive entropies.

Lemma 14.37. *Define (for $s \in \{0, 1, \ldots\}$) $\delta_s = 2^{-s}$, and let $\{\psi_j^s\}_{j=1}^{N_s} \subset \{\psi_j\}$ be a u_s-covering set of $\{\psi_j\}$, where u_s is non-increasing in s and where*

$$2u_s^2/\delta_s^2 \le N_s \le 4u_s^2/\delta_s^2.$$

Fix an s and let, for all j, $\phi_j^t := \psi_{k_j^t}^t$, $t = 1, \ldots, T$, where k_j^t is chosen as in Lemma 14.35 and T is the smallest integer such that $u_T \le \delta_s$. Let moreover $\eta_t > 0$, $\sum_{t=1}^T \eta_t \le 1$. Then

$$H(2\delta_s, \text{conv}(\{\psi_j - \phi_j^s\}), \|\cdot\|) \le \sum_{t=1}^{T-s} \left(\frac{2^{-2t+1}N_{s+t-1}}{\eta_t^2} + 1 \right) \left(1 + \log\left(1 + 2^{2(t+1)} \right) \right).$$

The next theorem follows directly from combining Lemma 14.36 and Lemma 14.37.

Theorem 14.8. *Define (for $s \in \{0, 1, \ldots\}$) $\delta_s = 2^{-s}$, and let $\{\psi_j^s\}_{j=1}^{N_s} \subset \{\psi_j\}$ be a u_s-covering set of $\{\psi_j\}$, where*

$$2u_s^2/\delta_s^2 \le N_s \le 4u_s^2/\delta_s^2.$$

Let $\eta_t > 0$, $\sum_{t=1}^T \eta_t = 1$, where T is the smallest integer such that $u_T \le \delta_s$. Then

$$H(3\delta_s, \text{conv}(\{\psi_j\}), \|\cdot\|) \le \log\left(9\,[2e]^9 \right) \sum_{t=1}^s N_{t-1}$$

$$+ \sum_{t=1}^{T-s} \left(\frac{2^{-2t+1}N_{s+t-1}}{\eta_t^2} + 1 \right) \left(1 + \log\left(1 + 2^{2(t+1)} \right) \right).$$

The next theorem considers an important special case.

Theorem 14.9. *Suppose that for some positive constants A and W with $A^W \ge 2$,*

$$N(u, \{\psi_j\}, \|\cdot\|) \le \left(\frac{A}{u} \right)^W, \quad u > 0.$$

Then for all $s \in \{0, 1, 2, \ldots\}$ and for $\delta_s = 2^{-s}$,

$$H(\delta_s, \text{conv}(\{\psi_j\}), \|\cdot\|) \le C_W A^{\frac{2W}{2+W}} \delta_s^{-\frac{2W}{2+W}},$$

where

$$C_W 4^{-\frac{2W}{2+W}} := 2\log\left(9\,[2e]^9 \right) \left(2^{\frac{W}{2+W}} - 2^{-\frac{W}{2+W}} \right)^{-1} + 2\log(5e) \left(\frac{3(2+W)}{2\log 2} \right)^4 2^{\frac{2}{2+W}}.$$

14.12.5 Further refinements

It is clear that entropy bounds for $\operatorname{conv}\{\psi_j\}$ based only on the covering number of $\{\psi_j\}$ can be sup-optimal. For example, suppose that

$$\psi_j = \psi_{1,k} + \psi_{2,l},$$

with the $\psi_{1,k}$ and $\psi_{2,l}$ varying in certain collections $\{\psi_{1,k}\}$ and $\{\psi_{2,l}\}$ respectively. Clearly,

$$N(u, \{\psi_j\}, \|\cdot\|) \leq N(u, \{\psi_{1,k}\}, \|\cdot\|) N(u, \{\psi_{2,l}\}, \|\cdot\|),$$

and in many cases, this bound cannot be substantially improved. It means e.g. that if $N(u, \{\psi_{1,k}\}, \|\cdot\|) \asymp u^{-W}$ as well as $N(u, \{\psi_{2,l}\}, \|\cdot\|) \asymp u^{-W}$, then typically $N(u, \{\psi_j\}, \|\cdot\|) \asymp u^{-2W}$. Nevertheless, $H(\delta, \operatorname{conv}(\{\psi_j\}), \|\cdot\|) \asymp \delta^{-\frac{2W}{2+W}}$, which can be easily seen by applying Theorem 14.9 to the two sets $\{\psi_{1,k}\}$ and $\{\psi_{2,l}\}$ separately.

The following lemma presents a further refinement.

Lemma 14.38. *Define (for $s \in \{0, 1, \ldots\}$) $\delta_s = 2^{-s}$. Suppose that for all s, there is a set $\{\psi_j^s\}_{j=1}^{N_s} \subset \{\psi_j\}$, such that each ψ_j is assigned to a $\phi_j^s := \psi_k^s$, and such that*

$$H(\delta_s, \operatorname{conv}(\{\psi_j - \phi_j^s\}), \|\cdot\|) \leq H_s.$$

Assume that $H_0 = 0$ and $1 = N_0 \leq N_2 \leq \ldots$. Then

$$H(\delta_s, \operatorname{conv}(\{\psi_j^s\}), \|\cdot\|) \leq \sum_{t=1}^{s} \left(H_{t-1} + \log((1+2^3)(1+2^4)) N_t \right).$$

14.12.6 An example: functions with $(m-1)$-th derivative of bounded variation

The total variation of a function $f : \mathbb{R} \to \mathbb{R}$ is defined as

$$\mathrm{TV}(f) = \sup_{N \geq 1} \sup_{z_0 < z_1 < \cdots < z_{N-1} < z_N} \sum_{j=1}^{N} |f(z_j) - f(z_{j-1})|.$$

The function f is called of bounded variation if $\mathrm{TV}(f) < \infty$. One easily checks that $\mathrm{TV}(f) \leq M$ if and only if f can be written as $f = c + f_+ - f_-$, where c is a constant, and f_+ and f_- are non-decreasing, with values in $[-M, M]$. We will use the notation $\uparrow$ for "non-decreasing".

We consider the entropy of a collection of functions whose $(m-1)$-th derivative is of bounded variation. As noted above, this problem can be reduced to considering a collection of functions with $(m-1)$-th derivative monotone and bounded.

Let

$$\mathscr{F}^{(m)} := \left\{ f : [0,1] \to \mathbb{R} : f^{(k)}(0) = 0, \; k = 0,1,\ldots m-1, \; f^{(m-1)} \uparrow, 0 \le f^{(m-1)} \le 1 \right\}.$$

One easily verifies that

$$\mathscr{F}^{(m)} = \mathrm{conv}(\{ \psi_v^{(m)} : v \in [0,1] \}),$$

where

$$\psi_v^{(m)}(x) = \frac{(x-v)^{m-1}}{(m-1)!} 1\{x \ge v\}.$$

We first consider the case $m = 1$. Then $\mathscr{F}^{(1)}$ is the class of non-decreasing functions (on $[0,1]$) with values in $[0,1]$.

Lemma 14.39. *We have*

$$N(u, \{ \psi_v^{(1)} : v \in [0,1] \}, \| \cdot \|) \le \lceil u^{-2} \rceil \le 2u^{-2}, \; 0 < u \le 1.$$

It is easy to see that the above lemma and its corollary below remain true if $\mathscr{F}^{(1)}$ is the class of bounded non-decreasing functions on the whole real line.

Corollary 14.9. *Application of Theorem 14.9 gives that for all s, and with $\delta_s = 2^{-s}$,*

$$H(\delta_s, \mathscr{F}^{(1)}, \| \cdot \|) \le \sqrt{2} C_2 \delta_s^{-1}.$$

Next, we consider general m. Here we need the refinements of Theorem 14.9.

Theorem 14.10. *For all s, and for $\delta_s = 2^{-s}$,*

$$H(\delta_s, \mathscr{F}^{(m)}, \| \cdot \|) \le c_m^{1/m} \delta_s^{-1/m},$$

where

$$c_m^{1/m} := \sqrt{2} C_2 \prod_{l=2}^{m} 2^{1/l}$$

$$+ 2\log((1+2^3)(1+2^4)) \sum_{k=0}^{m-2} 2^{\frac{2}{m-k}} / (2^{\frac{1}{m-k}} - 1) \prod_{l=m-k+1}^{m} 2^{\frac{1}{l}},$$

with C_2 given in Theorem 14.9.

14.12.7 Proofs for this section (Section 14.12)

Proof of Lemma 14.27. Let $\{\theta_j\}_{j=1}^{\mathcal{N}} \subset \Theta$, with $\mathcal{N} := \mathscr{C}(\delta, \Theta, \|\cdot\|_q)$, be all at least δ apart. If $\|\theta - \theta_j\|_q \le \delta/2$ and $\|\theta_j - \theta_k\|_q > \delta$, we must have (by the triangle inequality) $\|\theta - \theta_k\|_q > \|\theta_j - \theta_k\|_q - \|\theta - \theta_j\|_q > \delta/2$. Hence, the balls $B_j := \{\theta : \|\theta - \theta_j\|_q \le \delta/2\}$, $j = 1, \ldots, \mathcal{N}$ are mutually disjoint. The (Lebesgue) size of an ℓ_q-ball with radius r is $C_{p,q} r^p$, where $C_{p,q}$ is some constant depending on p and q. It follows that the size of $\cup_j B_j$ is at least

$$\mathcal{N} C_{p,q} (\delta/2)^p.$$

For all $\theta \in \cup_j B_j$,

$$\|\theta\|_q \le R + \delta/2.$$

The size of $\cup B_j$ is therefore at most

$$C_{p,q}(R + \delta/2)^p.$$

Thus we must have

$$\mathcal{N} C_{p,q} (\delta/2)^p \le C_{p,q}(R + \delta/2)^p.$$

$\square$

Proof of Lemma 14.28. Let $\beta \in \mathbb{R}^p$ be given, and satisfy $\beta_j \ge 0$ and $\sum_{j=1}^p \beta_j = 1$. Consider the random variable $\psi^\beta \in L_2(Q)$, with distribution

$$\mathbf{P}(\psi^\beta = \psi_j) = \beta_j, \; j = 1, \ldots, p.$$

Then clearly

$$\mathbb{E}\psi^\beta = f_\beta.$$

Let, for some $m \in \mathbb{N}$ to be specified, $\psi_1^\beta, \ldots, \psi_m^\beta$ be i.i.d. copies of ψ^β. Define their average

$$\bar{\psi}^\beta := \frac{1}{m} \sum_{i=1}^m \psi_i^\beta.$$

Then

$$\mathbb{E}\|\bar{\psi}^\beta - f_\beta\|^2 = \mathbb{E} \int (\bar{\psi}^\beta - f_\beta)^2 dQ$$

$$= \int \mathbb{E}(\bar{\psi}^\beta - f_\beta)^2 dQ = \int \text{var}(\bar{\psi}^\beta - f_\beta) dQ = \frac{1}{m} \int \text{var}(\psi^\beta - f_\beta) dQ$$

$$\le \frac{1}{m} \int \mathbb{E}(\psi^\beta)^2 dQ = \frac{1}{m} \int \sum_{j=1}^p \beta_j \psi_j^2 dQ \le \frac{1}{m},$$

where in the last inequality, we invoked $\|\psi_j\| \le 1$ for all j. Hence, with $m \ge 1/\delta^2$, there must be (by Chebyshev's inequality) a realization $\bar{\psi}^\beta$ with $\|\bar{\psi}^\beta - f_\beta\|^2 \le \delta^2$.

Now, we vary β over the simplex $\{\beta : \beta_j \geq 0, \sum_{j=1}^{p} \beta_j = 1\}$. Then the $\bar{\psi}^\beta$ vary over the class $\sum_{i=1}^{m} \psi_{j_i}/m$, where $(j_1, \ldots, j_m) \in \{1, \ldots, p\}^m$. The number of such $\bar{\psi}^\beta$ is (at most)

$$\binom{p+m-1}{m} \leq e^m \left(1 + \frac{p}{m}\right)^m \wedge p^m,$$

where in the last inequality we applied Stirling's formula. $\square$

Proof of Lemma 14.29. Cover the set $\mathscr{B}$ by ℓ_1-balls with radius u, say $B_1, \ldots, B_N$, where $B_j = \{\beta : \|\beta - \beta_j\|_1 \leq u\}$, and $N = N(u, \mathscr{B}, \|\cdot\|_1)$.

The total covering number is at most

$$N\left(\delta, \{f_\beta : \beta \in \mathscr{B}\}, \|\cdot\|\right) \leq \sum_{j=1}^{N} N(\delta, \{f_\beta - f_{\beta_j} : \beta \in B_j\}, \|\cdot\|).$$

The result follows by inserting the bound of Lemma 14.28:

$$H\left(\delta, \{f_\beta - f_{\beta_j} : \beta \in B_j\}, \|\cdot\|\right) \leq \left\lceil \frac{u^2}{\delta^2} \right\rceil \log(1 + 2p).$$

$\square$

Proof of Lemma 14.30. Let $u > 0$ be arbitrary. Then

$$\sum_{|\beta_j| \leq u^{\frac{1}{1-r}}} |\beta_j| \leq u \|\beta\|_r^r \leq u.$$

Moreover, defining $N_u := \#\{j : |\beta_j| > u^{\frac{1}{1-r}}\}$, we obtain

$$1 \geq \sum_{|\beta_j| > u^{\frac{1}{1-r}}} |\beta_j|^r \geq N_u u^{\frac{r}{1-r}},$$

or

$$N_u \leq u^{-\frac{r}{1-r}}.$$

Application of Lemma 14.27 with $q = 1$, and noting that $\|\beta\|_1 \leq \|\beta\|_r$, gives

$$N\left(u, \{\beta : \|\beta\|_0^0 \leq N_u, \|\beta\|_r \leq 1\}, \|\cdot\|_1\right) \leq \binom{p}{N_u} \left(\frac{2+u}{u}\right)^{N_u}.$$

Collecting the small coefficients $|\beta_j| \leq u^{\frac{1}{1-r}}$ as well, and taking logarithms, gives

$$H\left(2u, \{\|\beta\|_r \leq 1\}, \|\cdot\|_1\right) \leq \log\binom{p}{N_u} + N_u \log\left(\frac{2+u}{u}\right)$$

$$\leq u^{-\frac{r}{1-r}} \left(\log p + \log\left(\frac{2+u}{u}\right)\right).$$

$\square$

Proof of Lemma 14.32. If $u \geq 2/p$, we have

$$\frac{4+u}{u} \leq 1+2p.$$

Hence, by Lemma 14.30, for such u,

$$H(u, \mathcal{B}, \|\cdot\|_1) \leq \left(\frac{2}{u}\right)^{\frac{r}{1-r}} (\log p + \log(1+2p))$$

$$\leq 2\left(\frac{2}{u}\right)^{\frac{r}{1-r}} \log(1+2p)$$

$$= 2\left(\frac{4}{u^2}\right)^{\alpha} \log(1+2p).$$

In view of Lemma 14.29, we get for all u satisfying $u \geq 2/p$, $u \geq \delta$,

$$H\left(\delta, \{f_\beta : \beta \in \mathcal{B}\}, \|\cdot\|\right) \leq 2\log(1+2p)\left[\left(\frac{4}{u^2}\right)^{\alpha} + \frac{u^2}{\delta^2}\right].$$

We now insert the value $u^2/4 = (\alpha\delta^2/4)^{\frac{1}{1+\alpha}}$. $\square$

Proof of Lemma 14.33.

Consider the random vector $\phi^a := (\phi^{a_1}, \ldots, \phi^{a_N})$, where $\phi^{a_1}, \ldots, \phi^{a_N}$ are independent, and for $k = 1, \ldots, N$,

$$\mathbf{P}(\phi^{a_k} = \phi_{j,k}) = \alpha_{j,k}, \ j = 1, \ldots N_k.$$

Let $\{\phi_i^a\}_{i\geq 1}$ be i.i.d. copies of ϕ^a.

For $k = 1, \ldots, N$, define

$$m_k := \left\lceil \frac{\theta_k u^2}{\delta^2} \right\rceil,$$

and

$$\bar{\phi}^{a_k} := \frac{1}{m_k} \sum_{k=1}^{m_k} \phi_i^{a_k}.$$

Then we have

$$\mathbb{E}\bar{\phi}^{a_k} = g_{a_k}, \ k = 1, \ldots, N,$$

and hence

$$\mathbb{E}\left\| \sum_{k=1}^{N} \theta_k(\bar{\phi}^{a_k} - g_{a_k}) \right\|^2 = \int \mathbb{E}\left(\sum_{k=1}^{N} \theta_k(\bar{\phi}^{a_k} - g_{a_k}) \right)^2 dQ$$

$$= \int \text{var}\left(\sum_{k=1}^{N} \theta_k\bar{\phi}^{a_k} \right) dQ = \int \sum_{k=1}^{N} \theta_k^2 \text{var}(\bar{\phi}^{a_k}) dQ = \int \sum_{k=1}^{N} \theta_k^2 \text{var}(\phi^{a_k})/m_k dQ$$

$$\leq \int \sum_{k=1}^{N} \theta_k^2 \mathbb{E}(\phi^{a_k})^2 / m_k dQ = \sum_{k=1}^{N} \theta_k^2 \int \mathbb{E}(\phi^{a_k})^2 dQ / m_k$$

$$= \sum_{k=1}^{N} \theta_k^2 \sum_j \alpha_{j,k} \|\phi_{j,k}\|^2 / m_k \leq \sum_{k=1}^{N} \theta_k^2 u^2 / m_k \leq \delta^2 \sum_{k=1}^{N} \theta_k = \delta^2.$$

Hence, there exists a realization $\bar{\phi}^a = (\bar{\phi}^{a_1}, \ldots, \bar{\phi}^{a_N})$, such that

$$\left\| \sum_{k=1}^{N} \theta_k (\bar{\phi}^{a_k} - g_{a_k}) \right\| \leq \delta.$$

We now count the number of functions in this covering. We bound the number of choices for $\bar{\phi}^{a_k}$ by

$$p^{m_k}.$$

The number m_k can in turn be bounded by $\theta_k u^2 / \delta^2 + 1$. The total number of choices for $\bar{\phi}^a$ is therefore at most

$$\prod_{k=1}^{N} p^{\theta_k u^2 / \delta^2 + 1} = p^{u^2 / \delta^2 + N}.$$

Taking logarithms, we find

$$H(\delta, \mathcal{G}_\theta, \|\cdot\|) \leq \left(\frac{u^2}{\delta^2} + N \right) \log p.$$

$\square$

Proof of Lemma 14.34. Fix some $\delta > 0$. Let $\{\psi_1, \ldots, \psi_{N(\delta)}\} \subset \{\psi_j\}$ be a 2δ-covering set of $\{\psi_j\}$, assuming without loss of generality that the covering is formed by the first $N(\delta)$ functions in $\{\psi_j\}$. We will construct a δ-covering of $\{\sum_{j=1}^{N(\delta)} \beta_j \psi_j : \beta_j \geq 0, \|\beta\|_1 = 1\}$. This is then a 3δ covering of conv$(\{\psi_j\})$. Let $\{\phi_k\}_{k=1}^{N(u)}$ be a u-covering set of $\{\psi_j\}$. For all k, define

$$V_k := \left\{ j : \|\psi_j - \phi_k\| = \min_l \|\psi_j - \phi_l\| \right\}.$$

Let for $\beta_j \geq 0, \sum_j \beta_j = 1$,

$$\theta_k := \sum_{j \in V_k} \beta_j.$$

Let $\{\theta_k^l\}$ be a $\delta/4$-covering set of $\{\{\theta : \theta_k \geq 0, \sum_{k=1}^{N(u)} \theta_k = 1\}, \|\cdot\|_1\}$. Define

$$g_{a_k} := \sum_{j \in V_k} \alpha_{j,k}(\psi_j - \phi_k), \ \alpha_{j,k} := \beta_j / \theta_k.$$

Then

$$\sum_j \beta_j \psi_j = \sum_k \theta_k \phi_k + \sum_k \sum_{j \in V_k} \beta_j (\psi_j - \phi_k)$$

$$= \sum_k \theta_k \phi_k + \sum_k \theta_k g_{a_k}$$

$$= \sum_k \theta_k \phi_k + \sum_k (\theta_k - \theta_k^l) g_{a_k} + \sum_k \theta_k^l g_{a_k}.$$

By Lemma 14.27

$$H\left(\delta/4, \{\sum_k \theta_k \phi_k\}, \|\cdot\|\right) \le N(u) \log\left(\frac{8+\delta}{\delta}\right),$$

and

$$H\left(\delta/4, \{\theta \in \mathbb{R}^{N(u)} : \|\theta\|_1 \le 1\}, \|\cdot\|_1\right) \le N(u) \log\left(\frac{8+\delta}{\delta}\right)$$

and also, by Lemma 14.33, for all l,

$$H(\delta/2, \mathscr{G}_{\theta^l}, \|\cdot\|) \le \left(\frac{4u^2}{\delta^2} + N(u)\right) \log N(\delta).$$

Collecting the terms gives

$$H\left(\delta, \text{conv}(\{\psi_j\}_{j=1}^{N(\delta)}), \|\cdot\|\right) \le \left(3N(u) + \frac{4u^2}{\delta^2}\right) \log\left(\left(\frac{8+\delta}{\delta}\right) N(\delta)\right).$$

$\square$

Proof of Lemma 14.35. The first result follows from the definition of covering number. For the second result, we define k_j^T by

$$\|\psi_j - \psi_{k_j^T}^T\| = \min_{k=1,\dots,N_T} \|\psi_j - \psi_k^T\|.$$

Then again the result follows from the definition of covering number. $\square$

Proof of Lemma 14.36. Let us write for all s,

$$M_s := N(\delta_s, \mathscr{F}_s, \|\cdot\|).$$

Fix some $s \ge 1$. Let for $k = 1, \dots, N_{s-1}$,

$$V_k^s := \{j : \|\psi_j^s - \psi_k^{s-1}\| = \min_l \|\psi_j^s - \psi_l^{s-1}\|\},$$

be the off-spring of ψ_k^{s-1}. Then for all k and all $j \in V_k^s$, $\|\psi_j^s - \psi_k^{s-1}\| \le u_{s-1}$. For any β with $\beta_j \ge 0$, $\sum_j \beta_j = 1$, we have

$$\sum_j \beta_j \psi_j^s = \sum_k \theta_k \psi_k^{s-1} + \sum_k \sum_{j \in V_k^s} \beta_j (\psi_j^s - \psi_k^{s-1}),$$

where $\theta_k := \sum_{j \in V_k^s} \beta_j$, $k = 1, \ldots, N_{s-1}$.

Let $\{f_l^{s-1}\}_{l=1}^{M_{s-1}}$ be the centers of a δ_{s-1}-covering of $\mathscr{F}_{s-1}$. The ball

$$\left\{ f_\theta := \sum_{k=1}^{N_{s-1}} \theta_k \psi_k^{s-1} : \|f_\theta - f_l^{s-1}\| \le \delta_{s-1}, \ \theta \in \mathbb{R}^{N_{s-1}} \right\}$$

can be covered by

$$\left(\frac{4\delta_{s-1} + \delta_s}{\delta_s} \right)^{N_{s-1}} = 9^{N_{s-1}}$$

balls with radius $\delta_s/2$ (see Lemma 14.27). Hence, we get

$$N(\delta_s/2, \mathscr{F}_{s-1}, \|\cdot\|) \le M_{s-1} 9^{N_{s-1}}.$$

Consider now the class of functions $\mathrm{conv}(\{\psi_j^s - \psi_k^{s-1} : \ j \in V_k^s, \ k = 1, \ldots, N_{s-1}\})$. Since this is the convex hull of a class of functions, with cardinality N_s because of the tree structure (see the Generation Tree Lemma 14.35), each function having $L_2(Q)$-norm $\|\psi_j^s - \psi_k^{s-1}\| \le u_{s-1}$, we get from Lemma 14.28, and using $u_{s-1} \ge u_s$ and $\delta_s = \delta_{s-1}/2$,

$$N\left(\delta_s/2, \mathrm{conv}(\{\psi_j^s - \psi_k^{s-1}\}), \|\cdot\| \right) \le \left[e\left(1 + N_s \frac{\delta_s^2}{4u_{s-1}^2} \right) \right]^{4u_{s-1}^2/\delta_s^2 + 1}$$

$$\le [2e]^{16u_{s-1}^2/\delta_{s-1}^2 + 1} \le [2e]^{8N_{s-1} + 1} \le [2e]^{9N_{s-1}}.$$

Hence

$$M_s \le M_{s-1} 9^{N_{s-1}} [2e]^{9N_{s-1}},$$

or, taking logarithms,

$$\log M_s \le \log M_{s-1} + \log\left(9 [2e]^9 \right) N_{s-1}.$$

But then

$$\log M_s \le \log\left(9 [2e]^9 \right) \sum_{t=0}^{s-1} N_t + \log M_0.$$

The result follows, as $M_0 = 1$.   □

Proof of Lemma 14.37. Let $\beta_j \ge 0$, $\sum_j \beta_j = 1$. Then

$$\left\| \sum_j \beta_j (\psi_j - \phi_j^T) \right\| \le \max_j \|\psi_j - \phi_j^T\| \le u_T \le \delta_s.$$

We can write

$$\sum_j \beta_j(\phi_j^T - \phi_j^s) = \sum_{t=1}^{T-s} \sum_j \beta_j(\phi_j^{s+t} - \phi_j^{s+t-1}),$$

where we follow the branches of the generation tree. Define

$$H_{s,t} := H\left(\eta_t \delta_s, \text{conv}(\{\phi_j^{s+t} - \phi_j^{s+t-1}\}), \|\cdot\|\right).$$

Since by the Generation Tree Lemma 14.35, $\text{card}(\{\phi_j^{s+t} - \phi_j^{s+t-1}\}) \le N_{s+t}$, and since $\|\phi_j^{s+t} - \phi_j^{s+t-1}\| \le u_{s+t-1}$, by Lemma 14.28,

$$H_{s,t} \le \left(\frac{u_{s+t-1}^2}{\delta_s^2 \eta_t^2} + 1\right)\left(1 + \log\left(1 + N_{s+t}\delta_s^2\eta_t^2/u_{s+t-1}^2\right)\right).$$

Now

$$\frac{u_{s+t-1}^2}{\delta_s^2} = 2^{-2(t-1)}\frac{u_{s+t-1}^2}{\delta_{s+t-1}^2} \le 2^{-2t+1}N_{s+t-1},$$

and

$$\frac{N_{s+t}\delta_s^2}{u_{s+t-1}^2} \le \frac{N_{s+t}\delta_s^2}{u_{s+t}^2} = \frac{N_{s+t}\delta_{s+t}^2 2^{2t}}{u_{s+t}^2} \le 2^{2(t+1)}.$$

Thus, inserting $\eta_t \le 1$,

$$H_{s,t} \le \left(\frac{2 2^{-2t}N_{s+t-1}}{\eta_t^2} + 1\right)\left(1 + \log\left(1 + 2^{2(t+1)}\right)\right).$$

Hence,

$$H(\delta_s, \text{conv}(\{\phi_j^T - \phi_j^s\}), \|\cdot\|) \le \sum_{t=1}^{T-s} H_{s,t}$$

$$\le \sum_{t=1}^{T-s}\left(\frac{9 2^{-2t+1}N_{s+t-1}}{\eta_t^2} + 1\right)\left(1 + \log\left(1 + 2^{2(t+1)}\right)\right).$$

$\square$

Proof of Theorem 14.9. Define (for $s \in \{0, 1, \dots\}$),

$$u_s := (A^{\frac{W}{2}}\delta_s/\sqrt{2})^{\frac{2}{2+W}},$$

and

$$N_s := \left\lceil A^{\frac{2W}{2+W}} 2^{\frac{W}{2+W}} \delta_s^{-\frac{2W}{2+W}} \right\rceil.$$

Then

$$N(u_s, \{\psi_j\}, \|\cdot\|) \le A^W u_s^{-W}$$

$$= 2\frac{u_s^2}{\delta_s^2} = A^{\frac{2W}{2+W}} 2^{\frac{W}{2+W}} \delta_s^{-\frac{2W}{2+W}} \le N_s$$

$$\leq 2A^{\frac{2W}{2+W}} 2^{\frac{W}{2+W}} \delta_s^{-\frac{2W}{2+W}} = 4\frac{u_s^2}{\delta_s^2}.$$

We get

$$\sum_{t=1}^{s-1} N_{t-1} \leq 2A^{\frac{2W}{2+W}} 2^{\frac{W}{2+W}} \sum_{t=1}^{s-1} \delta_t^{-\frac{2W}{2+W}},$$

and

$$\sum_{t=1}^{s-1} \delta_t^{-\frac{2W}{2+W}} \leq \frac{\delta_s^{-\frac{2W}{2+W}} - 1}{2^{\frac{2W}{2+W}} - 1} \leq \frac{\delta_s^{-\frac{2W}{2+W}}}{2^{\frac{2W}{2+W}} - 1}.$$

Hence,

$$\sum_{t=1}^{s-1} N_{t-1} \leq 2A^{\frac{2W}{2+W}} (2^{\frac{W}{2+W}} - 2^{-\frac{W}{2+W}})^{-1} \delta_s^{-\frac{2W}{2+W}}.$$

Moreover,

$$2^{-2t} N_{s+t-1} \leq 2A^{\frac{2W}{2+W}} 2^{\frac{W}{2+W}} 2^{-2t} \delta_{s+t-1}^{-\frac{2W}{2+W}}$$

$$= 2^{\frac{2}{2+W}} A^{\frac{2W}{2+W}} 2^{-\frac{4t}{2+W}} \delta_s^{-\frac{2W}{2+W}}.$$

Using

$$\left(1 + \log\left(1 + 2^{2(t+1)}\right)\right) \leq t\log(5e),$$

it follows that

$$\sum_{t=1}^{T-s} \frac{2^{-2t} N_{s+t-1}}{\eta_t^2} \left(1 + \log\left(1 + 2^{2(t+1)}\right)\right)$$

$$\leq \log(5e) A^{\frac{2W}{2+W}} \delta_s^{-\frac{2W}{2+W}} \sum_{t=1}^{\infty} \frac{2^{-\frac{4t}{2+W}} t}{\eta_t^2} 2^{\frac{2}{2+W}}.$$

Take (for $t = 1, \ldots$)

$$\eta_t = \frac{(2^{-\frac{4t}{2+W}} t)^{1/3}}{\sum_{t=1}^{\infty} (2^{-\frac{4t}{2+W}} t)^{1/3}},$$

and employ the bound

$$\sum_{t=1}^{\infty} (2^{-\frac{4t}{2+W}} t)^{1/3} \leq 1 + \int_0^{\infty} (2^{-\frac{4t}{2+W}} t)^{1/3} dt$$

$$\leq 1 + \left(\frac{3(2+W)}{4\log 2}\right)^{4/3} \Gamma(4/3) \leq \left(\frac{3(2+W)}{2\log 2}\right)^{4/3},$$

to find

$$\sum_{t=1}^{T-s} \frac{2^{-2t} N_{s+t-1}}{\eta_t^2} \left(1 + \log\left(1 + 2^{2(t+1)}\right)\right)$$

$$\leq \log(5e) \left(\frac{3(2+W)}{2\log 2}\right)^4 A^{\frac{2W}{2+W}} \delta_s^{-\frac{2W}{2+W}} 2^{\frac{2}{2+W}}.$$

Now, insert Theorem 14.8 to obtain

$$H(4\delta_s, \text{conv}(\{\psi_j\}), \|\cdot\|) \le H(3\delta_s, \text{conv}(\{\psi_j\}), \|\cdot\|)$$

$$\le 2\log\left(9\,[2e]^9\right) A^{\frac{2W}{2+W}} (2^{\frac{W}{2+W}} - 2^{-\frac{W}{2+W}})^{-1} \delta_s^{-\frac{2W}{2+W}}$$

$$+ 2\log(5e) \left(\frac{3(2+W)}{2\log 2}\right)^4 A^{\frac{2W}{2+W}} \delta_s^{-\frac{2W}{2+W}} 2^{\frac{W}{2+W}}.$$

$\square$

Proof of Lemma 14.38. Let $s \ge 1$ be arbitrary. For each ψ_j^s, there is a $\phi_j^{s-1} \in \{\psi_j^{s-1}\}$ assigned to it, such that

$$H(\delta_{s-1}, \text{conv}(\{\psi_j^s - \phi_j^{s-1}\}), \|\cdot\|) \le H_{s-1}.$$

So $\text{conv}(\{\psi_j^s - \phi_j^{s-1}\})$ is covered by $\exp[H_{s-1}]$ balls with radius δ_{s-1}. Since $\text{card}(\{\psi_j^s - \phi_j^{s-1}\}) = N_s$, we can (by Lemma 14.27) cover each of these balls using at most 9^{N_s} balls with radius $\delta_s/2$. So in total, we can cover $\text{conv}(\{\psi_j^s - \phi_j^{s-1}\})$ by $9^{N_s} e^{H_s}$ balls with radius $\delta_s/2$. Define

$$M_s := \exp[H(\delta_s, \text{conv}(\{\psi_j^s\}), \|\cdot\|)].$$

Then, using similar arguments, we can cover $\text{conv}(\{\psi_j^{s-1}\})$ by $(1+2^4)^{N_{s-1}} M_{s-1}$ balls with radius $\delta_s/2$.

Hence,

$$\log M_s \le \log(1+2^3)N_s + H_{s-1} + \log(1+2^4)N_{s-1} + \log M_{s-1}$$

$$= \log((1+2^3)(1+2^4))N_s + H_{s-1} + \log M_{s-1}.$$

Repeating the argument gives

$$\log M_s \le \log((1+2^3)(1+2^4)) \sum_{t=1}^{s} N_s + \sum_{t=0}^{1} H_{t-1}.$$

$\square$

Proof of Lemma 14.39. Without loss of generality, we may assume that Q is the uniform distribution (Problem 14.8). Let $0 < u \le 1$ be arbitrary. Take $v_j = ju^2$, $j = 0, 1, \ldots, N$, where $N := \lceil 1/u^2 \rceil$. Then for $v \in (v_{j-1}, v_j]$,

$$\|1\{\cdot \ge v\} - 1\{\cdot \ge v_j\}\|^2 = (v - v_j)^2 \le u^2.$$

$\square$

Proof of Theorem 14.10. Fix some $\delta_s > 0$. Take $v_j := j\delta_s^{\frac{1}{m}}$, $j = 0, 1, \ldots, N_s$, where $N_s := \lceil 1/\delta_s^{1/m} \rceil$. Define $\psi_j^s := \psi_{v_j}^{(m)}$. For $v \in (v_{j-1}, v_j]$, assign $\phi_v^{(m)}$ to $\psi_{v_j}^{(m)} = \psi_j^s$. One then easily verifies that

$$\frac{\psi_v^{(m)} - \phi_v^s}{\delta_s^{1/m}} \in \mathscr{F}^{(m-1)}.$$

Further,

$$H(\delta_t, \text{conv}(\{\psi_v - \phi_v^s\}), \|\cdot\|) \leq H(\delta_t/\delta_s^{\frac{1}{m}}, \mathscr{F}^{(m-1)}, \|\cdot\|).$$

Hence, for $m = 2$, and using Corollary 14.9

$$H(\delta_t, \text{conv}(\{\psi_v - \phi_v^s\}, \|\cdot\|) \leq H(\delta_t/\delta_s^{\frac{1}{2}}, \mathscr{F}^{(1)}, \|\cdot\|)$$

$$\leq \sqrt{2}C_2\delta_s^{1/2}\delta_t^{-1},$$

so that

$$\sum_{t=0}^{s} H(\delta_t, \text{conv}(\{\psi_v - \phi_v^s\}, \|\cdot\|) \leq \sqrt{2}C_2\delta_s^{-1/2}.$$

We moreover have, again for $m = 2$,

$$\sum_{t=0}^{s-1} N_t \leq 2\delta_s^{-1/2}\sqrt{2}/(\sqrt{2}-1).$$

Application of Lemma 14.38 gives

$$H(\delta_s, \mathscr{F}^{(2)}, \|\cdot\|) \leq \sqrt{2}[\sqrt{2}C_2 + 2\log((1+2^3)(1+2^4))\sqrt{2}/(\sqrt{2}-1)]\delta_s^{-1/2}.$$

Invoking this, we see that for $m = 3$,

$$H(\delta_t, \text{conv}(\{\psi_v - \phi_v^s\}, \|\cdot\|) \leq H(\delta_t/\delta_s^{\frac{1}{3}}, \mathscr{F}^{(2)}, \|\cdot\|)$$

$$\leq \sqrt{2}[\sqrt{2}C_2 + \log((1+2^3)(1+2^4))\sqrt{2}/(\sqrt{2}-1)]\delta_s^{1/6}\delta_t^{-1/2},$$

so that

$$\sum_{t=0}^{s-1} H(\delta_t, \text{conv}(\{\psi_v - \phi_v^s\}, \|\cdot\|)$$

$$\leq \sqrt{2}[\sqrt{2}C_2 + 2\log((1+2^3)(1+2^4))\sqrt{2}/(\sqrt{2}-1)]\sqrt{2}/(\sqrt{2}-1)\delta_s^{-1/3}.$$

For $m = 3$,

$$\sum_{t=1}^{s} N_t \leq 2\delta_s^{-1/3}2^{1/3}(2^{1/3}-1).$$

Again, application of Lemma 14.38 gives

$$H(\delta_s, \mathscr{F}^{(3)}, \|\cdot\|) \le 2^{1/3}[\sqrt{2}[\sqrt{2}C_2$$

$$+2\log((1+2^3)(1+2^4))\sqrt{2}/(\sqrt{2}-1)]\sqrt{2}/(\sqrt{2}-1)$$

$$+2\log((1+2^3)(1+2^4))2^{1/3}/(2^{1/3}-1)]\delta_s^{-1/3}.$$

The proof is finished by repeating the argument (Problem 14.9).   □

Problems

14.1. Suppose that for some constant σ^2 and for all L,

$$\mathbb{E}\exp[X/L] \le \exp[\sigma^2/(2L^2)].$$

Show that X is sub-Gaussian.

14.2. Show that

$$\frac{K^2}{2}\left(\mathbb{E}e^{|X|/K} - 1 - \mathbb{E}|X|/K\right) \le K^2\left(\mathbb{E}e^{X^2/K^2} - 1\right).$$

14.3. Consider independent random variables $X_1, \ldots, X_n$, where $X_i = (X_i^{(1)}, \ldots, X_i^{(p)})$. Define $\mathbf{X}^T := (X_1^T, \ldots, X_n^T)$, and

$$\hat{\Sigma} := \mathbf{X}^T\mathbf{X}/n,$$

and

$$\Sigma := \mathbb{E}\hat{\Sigma}.$$

Let

$$\|\hat{\Sigma} - \Sigma\|_\infty := \sup_{j,k}|\hat{\Sigma}_{j,k} - \Sigma_{j,k}|.$$

Suppose that the $X_i^{(j)}$ are, uniformly in i and j, sub-Gaussian: for some constants K and σ_0^2,

$$K^2(\mathbb{E}\exp[|X_i^{(j)}|^2/K^2] - 1) \le \sigma_0^2.$$

for all $i = 1, \ldots, n$, $j = 1, \ldots, p$. Use similar arguments as in Example 14.1 to show that for all t,

$$\mathbf{P}\left(\|\hat{\Sigma} - \Sigma\|_\infty \ge 2K^2 t + 2K\sigma_0\sqrt{2t} + 2K\sigma_0\lambda\left(K/\sigma_0, n, \binom{p}{2}\right)\right) \le \exp[-nt],$$

where

$$\lambda\left(K/\sigma_0,n,\binom{p}{2}\right) := \sqrt{\frac{2\log(p(p-1))}{n}} + \frac{K\log(p(p-1))}{n}.$$

14.4. This exercise is a variant of Problem 14.3. For simplicity, let us look at the case of independent copies $X_1,\ldots,X_n$ of a random variable $X = (X^{(1)},\ldots,X^{(p)}) \in \mathbb{R}^p$. Define

$$\mu_j := \mathbb{E}X^{(j)}, \ \sigma_j^2 := \mathbb{E}|X^{(j)} - \mu_j|^2, \ j = 1,\ldots,p.$$

Let, for $j = 1,\ldots,p$,

$$\hat{\sigma}_j^2 := \frac{1}{n}\sum_{i=1}^n |X_i^j - \bar{X}^{(j)}|^2,$$

where

$$\bar{X}^{(j)} = \frac{1}{n}\sum_{i=1}^n X_i^{(j)}.$$

Suppose that the $X^{(j)}$ are, uniformly in j, sub-Gaussian: for some constants K and σ_0^2,

$$K^2(\mathbb{E}\exp[|X^{(j)}|^2/K^2] - 1) \le \sigma_0^2.$$

for all $j = 1,\ldots,p$. Use similar arguments as in Example 14.1 to derive an exponential probability inequality for $\max_{1 \le j \le p} |\hat{\sigma}_j^2 - \sigma_j^2|$.

14.5. In this problem, we derive a symmetrization inequality for probabilities (see Pollard (1984), van de Geer (2000)). Let P_n be the empirical distribution of n independent random variables $Z_1,\ldots,Z_n$, defined on a space $\mathscr{Z}$, and let P_n' be an independent copy of P_n (i.e. P_n' is the empirical distribution of an independent copy $(Z_1',\ldots,Z_n')$ of $(Z_1,\ldots,Z_n)$). Let furthermore $\{\gamma_j\}_{j=1}^p$ be a collection of real-valued functions on $\mathscr{Z}$, that satisfy

$$\mathbb{E}\gamma_j(Z_i) = 0, \ \forall i, \ j.$$

Define for any function $\gamma : \mathscr{Z} \to \mathbb{R}$,

$$\|\gamma\|^2 := \frac{1}{n}\sum_{i=1}^n \mathbb{E}\gamma^2(Z_i),$$

whenever it exists.
(a) Suppose that for some R,

$$\max_j \|\gamma_j\| \le R.$$

Show that for any $t > 0$ and any $0 < \eta < 1$, and for $n > R^2/(\eta^2\delta^2)$,

$$\mathbf{P}(\max_j |P_n\gamma_j| > t) \le \frac{\mathbf{P}(\max_j |(P_n - P_n')\gamma_j| > (1-\eta)t)}{1 - R^2/(n\eta^2 t^2)}.$$

(b) Let $\varepsilon_1,\ldots,\varepsilon_n$ be a Rademacher sequence, independent of $Z_1,\ldots,Z_n,Z_1',\ldots,Z_n'$.

Show that for all $t > 0$,

$$\mathbf{P}\left(\max_j |(P_n - P_n')\gamma_j| > t\right) \le 2\mathbf{P}\left(\max_j \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \gamma_j(Z_i)\right| > t/2\right).$$

14.6. This exercise considers a similar situation as in Lemma 14.25, but now for the case where the sums $\sum_{i=1}^n x_{i,j,t}\varepsilon_i$ share the same ε_i for different t. The random variables $\{\varepsilon_i : i = 1,\ldots,n\}$ are assumed to be independent with $\mathbb{E}\varepsilon_i = 0$, and $\mathbb{E}|\varepsilon_i|^m \le \mu_m^m$ for all i. Moreover. $\{x_{i,j,t} : i = 1,\ldots,n,\ j = 1,\ldots,p,\ t = 1,\ldots,T_j\}$ is a given collection of constants. Let

$$\bar{T} := \frac{1}{p}\sum_{j=1}^p T_j.$$

Then for $p\bar{T} \ge e^{2m-1}$,

$$\mathbb{E}\max_{1\le j\le p}\left[\frac{1}{T_j}\sum_{t=1}^{T_j}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n x_{i,j,t}\varepsilon_i\right)^2\right]^m$$

$$\le \left[8\log(2p)\right]^m \mu_m^m \left[\frac{1}{n}\sum_{i=1}^n \max_{1\le j\le p}\max_{1\le t\le T_j} x_{i,j,t}^{2m}\right].$$

14.7. Consider independent real-valued random variables $\{Y_i\}_{i=1}^n$ and fixed $x_i \in \mathcal{X}$, $i = 1,\ldots,n$. For $f : \mathcal{X} \to \mathbb{R}$, define

$$\|f\|_n^2 := \frac{1}{n}\sum_{i=1}^n f^2(x_i).$$

Let $\psi_1,\ldots,\psi_p$ be given functions on $\mathcal{X}$, and define

$$f_\beta := \sum_{j=1}^p \beta_j \psi_j,\ \beta \in \mathbb{R}^p.$$

Let $\rho : \mathbb{R} \to \mathbb{R}$ be a given loss function. Assume that ρ is a Lipschitz function, with Lipschitz constant L:

$$|\rho(s) - \rho(\tilde{s})| \le L|s - \tilde{s}|\ \forall s, \tilde{s} \in \mathbb{R}.$$

Let β^* be fixed. Define

$$v_n(\beta) := \frac{1}{n}\sum_{i=1}^n \left(\rho(Y_i - f_\beta(x_i)) - \mathbb{E}\rho(Y_i - f_\beta(x_i))\right).$$

For all $\delta > 0$, we define the random variable

$$\bar{\mathbf{Z}}_\delta := \sup_{\|f_\beta - f_{\beta^*}\|_n \leq \delta} |v_n(\beta) - v_n(\beta^*)|/\sqrt{p}.$$

Apply Lemma 14.19 and Massart's concentration inequality (Theorem 14.2) to find a probability inequality for the set $\mathscr{T} = \{\bar{\mathbf{Z}}_\delta \leq \delta L(4/\sqrt{n} + t/\sqrt{np})\}$.

14.8. Let Q be a probability measure on $\mathbb{R}$ which has a strictly increasing continuous distribution function and let

$$\mathscr{F} := \{f : \mathbb{R} \to [0,1], \ f \uparrow\}.$$

(a) Check that the entropy of $\mathscr{F}$ for the $L_q(Q)$ norm does not depend on Q ($1 \leq q < \infty$).
(b) Verify that the entropy is not finite for the L_∞-norm.
(c) Verify that these entropies for possibly non-strictly increasing or non-continuous Q are not larger than for the strictly increasing continuous case.

14.9. This is applied in the proof of Theorem 14.10. Consider the recursion

$$x_m = b_m x_{m-1} + c_m, \ m = 1, \ldots,$$

where the b_m and c_m are given constants. Verify that

$$x_m = x_1 \prod_{l=2}^{m} b_l + \sum_{k=1}^{m-2} c_{m-k} \prod_{l=m-k+1}^{m} b_l.$$

Author Index

Anderson, T.W., 452, 458, 475, 480
Audrino, F., 404

Bach, F., 250
Ball, K., 520
Banerjee, O., 437
Barron, A. , 208
Bartlett, P., 14, 104, 507
Bates, D.M., 311
Beer, M.A., 31
Benjamini, Y., 367, 382, 383
Bennet, G., 481
Berkhof J., 12
Bertsekas, D.P., 1, 9, 15, 16, 39, 68, 71, 73, 76, 308, 315
Bickel, P.J., 100, 106, 136, 158, 162, 170, 185, 437, 438
Birgé, L., 208
Bissantz, N., 408
Blanchard, G., 367
Boldrick, J.C., 347
Bousquet, O., 481, 494
Boyd, S., 51
Breiman, L., 1, 13, 20, 28, 339, 340, 353, 388, 389
Brockwell, P.J., 448
Brutlag, D.L., 18, 90
Bühlmann, P., 18, 21, 22, 30, 52, 53, 61, 72, 75, 89–91, 156, 172, 244, 250, 252, 258, 287, 293, 301, 303, 304, 313, 317, 331, 339, 342, 343, 351, 353–355, 359, 363, 365, 370, 373, 388, 397, 401, 403, 404, 406, 411, 418, 420, 439, 441, 442, 453, 455, 462, 463, 468, 469
Bunea, F., 100, 106, 166, 173, 180
Burge, C.B., 59–61

Candès, E., 100, 106, 136, 158, 169, 174, 185, 247
Caponnetto, A., 408
Caruana, R., 96
Chesneau, C., 250, 252
Conlon, E.M., 18, 31

Dahinden, C., 52, 53
d'Aspremont, 437
Davis, R.A., 448
Dempster, A.P., 304
DeVore, R.A., 424
DiMarzio, M., 408
Donoho, D.L., 2, 7, 100
Dudley, R.M., 519
Dudoit, S., 347
Dümbgen, L., 108, 178, 509

Edwards, D., 434
Efron, B., 17, 34–36, 38, 387, 409
El Ghaoui, L., 437
Elad, M., 2
Emerick, M.C., 52

Fan, J., 32, 350, 362, 438, 468, 474, 475
Feng, Y., 438
Fisher, R.A., 458
Freedman, D., 342
Freund, Y., 388, 389
Friedman, J.H., 1, 39, 48, 97, 388, 389, 391, 393, 394, 397, 398, 404, 409, 430, 437
Fu, W.J., 39
Fuchs, J., 2

Gatu, C., 20
Gilbert, A.C., 416
Glymour, C., 449

Index

References

ANDERSON, T. (1984). *An Introduction to Multivariate Statistical Analysis.* 2nd ed. Wiley.

AUDRINO, F. and BÜHLMANN, P. (2003). Volatility estimation with functional gradient descent for very high-dimensional financial time series. *Journal of Computational Finance* **6** 65–89.

BACH, F. (2008). Consistency of the group Lasso and multiple kernel learning. *Journal of Machine Learning Research* **9** 1179–1225.

BALL, K. and PAJOR, A. (1990). The entropy of convex bodies with few extreme points. *London Mathematical Society Lecture Note Series* **158** 25–32.

BANERJEE, O., EL GHAOUI, L. and D'ASPREMONT, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research* **9** 485–516.

BARRON, A., BIRGÉ, L. and MASSART, P. (1999). Risk bounds for model selection via penalization. *Probability Theory and Related Fields* **113** 301–413.

BARTLETT, P., MENDELSON, S. and NEEMAN, J. (2009). ℓ_1-regularized regression: persistence and oracle inequalities. Manuscript.

BEER, M. and TAVAZOIE, S. (2004). Predicting gene expression from sequence. *Cell* **117** 185–198.

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B* **57** 289–300.

BENJAMINI, Y. and YEKUTIELI, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* **29** 1165–1188.

BENNET, G. (1962). Probability inequalities for sums of independent random variables. *Journal of the American Statistical Association* **57** 33–45.

BERTSEKAS, D. (1995). *Nonlinear Programming.* Athena Scientific, Belmont, MA.

BICKEL, P., RITOV, Y. and TSYBAKOV, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics* **37** 1705–1732.

BISSANTZ, N., HOHAGE, T., MUNK, A. and RUYMGAART, F. (2007). Convergence rates of general regularization methods for statistical inverse problems and applications. *SIAM Journal of Numerical Analysis* **45** 2610–2636.

BLANCHARD, G. and ROQUAIN, E. (2008). Two simple sufficient conditions for FDR control. *Electronic Journal of Statistics* **2** 963–992.

BOUSQUET, O. (2002). A Bennet concentration inequality and its application to suprema of empirical processes. *Comptes Rendus de l'Académie des Sciences, Paris* **334** 495–550.

BOYD, S. and VANDENBERGHE, L. (2004). *Convex Optimization*. Cambridge University Press, New York.

BREIMAN, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics* **37** 373–384.

BREIMAN, L. (1996). Bagging predictors. *Machine Learning* **24** 123–140.

BREIMAN, L. (1998). Arcing classifiers (with discussion). *Annals of Statistics* **26** 801–849.

BREIMAN, L. (1999). Prediction games and arcing algorithms. *Neural Computation* **11** 1493–1517.

BREIMAN, L. (2001). Random forests. *Machine Learning* **45** 5–32.

BREIMAN, L., FRIEDMAN, J., OLSHEN, R. and STONE, C. (1984). *Classification and Regression Trees*. Wadsworth.

BROCKWELL, P. and DAVIS, R. (1991). *Time Series: Theory and Methods*. 2nd ed. Springer.

BÜHLMANN, P. (2004). Bagging, boosting and ensemble methods. In *Handbook of Computational Statistics: Concepts and Methods* (J. Gentle, W. Härdle and Y. Mori, eds.). Springer, 877–907.

BÜHLMANN, P. (2006). Boosting for high-dimensional linear models. *Annals of Statistics* **34** 559–583.

BÜHLMANN, P. and HOTHORN, T. (2007). Boosting algorithms: regularization, prediction and model fitting (with discussion). *Statistical Science* **22** 477–505.

BÜHLMANN, P., KALISCH, M. and MAATHUIS, M. (2010). Variable selection in high-dimensional linear models: partially faithful distributions and the PC-simple algorithm. *Biometrika* **97** 261–278.

BÜHLMANN, P. and MEIER, L. (2008). Discussion of "One-step sparse estimates in nonconcave penalized likelihood models" (auths H. Zou and R. Li). *Annals of Statistics* **36** 1534–1541.

BÜHLMANN, P. and YU, B. (2002). Analyzing bagging. *Annals of Statistics* **30** 927–961.

BÜHLMANN, P. and YU, B. (2003). Boosting with the L_2 loss: regression and classification. *Journal of the American Statistical Association* **98** 324–339.

BÜHLMANN, P. and YU, B. (2006). Sparse boosting. *Journal of Machine Learning Research* **7** 1001–1024.

BUNEA, F. (2008). Honest variable selection in linear and logistic regression models via ℓ_1 and $\ell_1 + \ell_2$ penalization. *Electronic Journal of Statistics* **2** 1153–1194.

BUNEA, F., TSYBAKOV, A. and WEGKAMP, M. (2006). Aggregation and sparsity via ℓ_1-penalized least squares. In *Proceedings of 19th Annual Conference on Learning Theory, COLT 2006. Lecture Notes in Artificial Intelligence*. Springer Verlag.

BUNEA, F., TSYBAKOV, A. and WEGKAMP, M. (2007a). Aggregation for Gaussian regression. *Annals of Statistics* **35** 1674.

BUNEA, F., TSYBAKOV, A. and WEGKAMP, M. (2007b). Sparse density estimation with ℓ_1 penalties. In *Proceedings of 20th Annual Conference on Learning Theory, COLT 2007. Lecture Notes in Artificial Intelligence*. Springer.

BUNEA, F., TSYBAKOV, A. and WEGKAMP, M. (2007c). Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics* **1** 169–194.

BURGE, C. (1998). Modeling dependencies in pre-mRNA splicing signals. In *Computational Methods in Molecular Biology* (S. Salzberg, D. Searls and S. Kasif, eds.), chap. 8. Elsevier Science, 129–164.

BURGE, C. and KARLIN, S. (1997). Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology* **268** 78–94.

CANDÈS, E., ROMBERG, J. and TAO, T. (2006). Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics* **59** 1207–1223.

CANDÈS, E. and TAO, T. (2005). Decoding by linear programming. *IEEE Transactions on Information Theory* **51** 4203–4215.

CANDÈS, E. and TAO, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics* **35** 2313–2351.

CARUANA, R. (1997). Multitask learning. *Machine Learning* **28** 41–75.

CHESNEAU, C. and HEBIRI, M. (2008). Some theoretical results on the grouped variable Lasso. *Mathematical Methods of Statistics* **17** 317–326.

CONLON, E., LIU, X., LIEB, J. and LIU, J. (2003). Integrating regulatory motif discovery and genome-wide expression analysis. *Proceedings of the National Academy of Sciences* **100** 3339–3344.

DAHINDEN, C., KALISCH, M. and BÜHLMANN, P. (2010). Decomposition and model selection for large contingency tables. *Biometrical Journal* **52** 233–252.

DAHINDEN, C., PARMIGIANI, G., EMERICK, M. and BÜHLMANN, P. (2007). Penalized likelihood for sparse contingency tables with an application to full-length cDNA libraries. *BMC Bioinformatics* **8** 1–11.

DEMPSTER, A., LAIRD, N. and RUBIN, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B* **39** 1–38.

DEVORE, R. and TEMLYAKOV, V. (1996). Some remarks on greedy algorithms. *Advances in Computational Mathematics* **5** 173–187.

DIMARZIO, M. and TAYLOR, C. (2008). On boosting kernel regression. *Journal of Statistical Planning and Inference* **138** 2483–2498.

DONOHO, D. (2006). For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics* **59** 797–829.

DONOHO, D. and ELAD, M. (2003). Uncertainty principles and ideal atomic decomposition. *Proceedings of the National Academy of Sciences* **100** 2197–2202.

DONOHO, D. and HUO, X. (2001). Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory* **47** 2845–2862.

DONOHO, D. and JOHNSTONE, I. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425–455.

DUDLEY, R. (1987). Universal Donsker classes and metric entropy. *Annals of Probability* **15** 1306–1326.

DUDOIT, S., SHAFFER, J. and BOLDRICK, J. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science* **18** 71–103.

DÜMBGEN, L., VAN DE GEER, S., VERAAR, M. and WELLNER, J. (2010). Nemirovski's inequalities revisited. *The American Mathematical Monthly* **117** 138–160.

EDWARDS, D. (2000). *Introduction to Graphical Modelling*. 2nd ed. Springer Verlag.

EFRON, B. (2004). The estimation of prediction error: covariance penalties and cross-validation. *Journal of the American Statistical Association* **99** 619–632.

EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression (with discussion). *Annals of Statistics* **32** 407–451.

FAN, J., FENG, Y. and WU, Y. (2009a). Network exploration via the adaptive Lasso and SCAD penalties. *Annals of Applied Statistics* **3** 521–541.

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96** 1348–1360.

FAN, J. and LV, J. (2008). Sure independence screening for ultra-high dimensional feature space (with discussion). *Journal of the Royal Statistical Society Series B* **70** 849–911.

FAN, J., SAMWORTH, R. and WU, Y. (2009b). Ultrahigh dimensional variable selection: beyond the linear model. *Journal of Machine Learning Research* **10** 1989–2014.

FISHER, R. (1924). The distribution of the partial correlation coefficient. *Metron* **3** 329–332.

FREEDMAN, D. (1977). A remark on the difference between sampling with and without replacement. *Journal of the American Statistical Association* **72** 681.

FREUND, Y. and SCHAPIRE, R. (1996). Experiments with a new boosting algorithm. In *Proceedings of the Thirteenth International Conference on Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA.

FREUND, Y. and SCHAPIRE, R. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences* **55** 119–139.

FRIEDMAN, J. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics* **29** 1189–1232.

FRIEDMAN, J., HASTIE, T., HÖFLING, H. and TIBSHIRANI, R. (2007a). Pathwise coordinate optimization. *Annals of Applied Statistics* **1** 302–332.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion). *Annals of Statistics* **28** 337–407.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2007b). Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics* **9** 432–441.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularized paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33** 1–22.

FU, W. (1998). Penalized regressions: the Bridge versus the Lasso. *Journal of Computational and Graphical Statistics* **7** 397–416.

FUCHS, J. (2004). On sparse representations in arbitrary redundant bases. *IEEE Transactions on Information Theory* **50** 1341–1344.

GATU, C., YANEV, P. and KONTOGHIORGHES, E. (2007). A graph approach to generate all possible regression submodels. *Computational Statistics & Data Analysis* **52** 799–815.

GREEN, P. and SILVERMAN, B. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach.* Chapman & Hall, New York.

GREENSHTEIN, E. (2006). Best subset selection, persistence in high dimensional statistical learning and optimization under ℓ_1 constraint. *Annals of Statistics* **34** 2367–2386.

GREENSHTEIN, E. and RITOV, Y. (2004). Persistence in high-dimensional predictor selection and the virtue of over-parametrization. *Bernoulli* **10** 971–988.

GUÉDON, O., MENDELSON, S., PAJOR, A. and TOMCZAK-JAEGERMANN, N. (2007). Subspaces and orthogonal decompositions generated by bounded orthogonal systems. *Positivity* **11** 269–283.

HASTIE, T. and TIBSHIRANI, R. (1990). *Generalized Additive Models.* Chapman & Hall, London.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2001). *The Elements of Statistical Learning; Data Mining, Inference and Prediction.* Springer, New York.

HEBIRI, M. and VAN DE GEER, S. (2010). The smooth Lasso and other $\ell_1 + \ell_2$-penalized methods. ArXiv:1003.4885.

HOEFFDING, W. (1963). Probability inequalities for sums of bounded variables. *Journal of the American Statistical Association* **58** 13–30.

HOFMANN, M., GATU, C. and KONTOGHIORGHES, E. (2007). Efficient algorithms for computing the best subset regression models for large-scale problems. *Computational Statistics & Data Analysis* **52** 16–29.

HOTELLING, H. (1953). New light on the correlation coefficient and its transforms. *Journal of the Royal Statistical Society Series B* **15** 193–232.

HUANG, J., MA, S. and ZHANG, C.-H. (2008). Adaptive Lasso for sparse high-dimensional regression models. *Statistica Sinica* **18** 1603–1618.

HUANG, J. and ZHANG, T. (2010). The benefit of group sparsity. *Annals of Statistics* **38** 1978–2004.

KALISCH, M. and BÜHLMANN, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research* **8** 613–636.

KING, G. and ZENG, L. (2001). Logistic regression in rare events data. *Political Analysis* **9** 137–163.

KOLTCHINSKII, V. (2009a). Sparsity in penalized empirical risk minimization. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* **45** 7–57.

KOLTCHINSKII, V. (2009b). The Dantzig selector and sparsity oracle inequalities. *Bernoulli* **15** 799–828.

KOLTCHINSKII, V., TSYBAKOV, A. and LOUNICI, K. (2010b). Nuclear norm penalization and optimal rates for noisy low rank matrix completion. ArXiv:1011.6256.

KOLTCHINSKII, V. and YUAN, M. (2008). Sparse recovery in large ensembles of kernel machines. In *Proceedings of the 21st Annual Conference on Learning Theory, COLT 2008. Lecture Notes in Artificial Intelligence*. Springer.

KOLTCHINSKII, V. and YUAN, M. (2010). Sparsity in multiple kernel learning. *Annals of Statistics* **38** 3660–3695.

LAURITZEN, S. (1996). *Graphical Models*. Oxford University Press.

LEDOUX, M. (1996). Talagrand deviation inequalities for product measures. *ESAIM: Probability and Statistics* **1** 63–87.

LEDOUX, M. and TALAGRAND, M. (1991). *Probability in Banach Spaces: Isoperimetry and Processes*. Springer Verlag, New York.

LEEB, H. and PÖTSCHER, B. (2003). The finite-sample distribution of post-model-selection estimators and uniform versus nonuniform approximations. *Econometric Theory* **19** 100–142.

LEEB, H. and PÖTSCHER, B. (2005). Model selection and inference: facts and fiction. *Econometric Theory* **21** 21–59.

LIU, X., BRUTLAG, D. and LIU, J. (2002). An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nature Biotechnology* **20** 835–839.

LOUBES, J.-M. and VAN DE GEER, S. (2002). Adaptive estimation in regression, using soft thresholding type penalties. *Statistica Neerlandica* **56** 453–478.

LOUNICI, K. (2008). Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electronic Journal of Statistics* **2** 90–102.

LOUNICI, K., PONTIL, M., TSYBAKOV, A. and VAN DE GEER, S. (2009). Taking advantage of sparsity in multi-task learning. In *Proceedings of 22th Annual Conference on Learning Theory, COLT 2009. Lecture Notes in Artificial Intelligence*. Springer.

LOUNICI, K., PONTIL, M., VAN DE GEER, S. and TSYBAKOV, A. (2010). Oracle inequalities and optimal inference under group sparsity. ArXiv:1007.1771.

MALLAT, S. and ZHANG, Z. (1993). Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing* **41** 3397–3415.

MASSART, P. (2000a). About the constants in Talagrand's concentration inequalities for empirical processes. *Annals of Probability* **28** 863–884.

MASSART, P. (2000b). Some applications of concentration inequalities to statistics. *Annales de la Faculté des Sciences de Toulouse* **9** 245–303.

MCCULLAGH, P. and NELDER, J. (1989). *Generalized Linear Models*. 2nd ed. Chapman & Hall, London.

MCLACHLAN, G. and PEEL, D. (2000). *Finite Mixture Models*. Wiley, New York.

MEIER, L. and BÜHLMANN, P. (2007). Smoothing ℓ_1-penalized estimators for highdimensional time-course data. *Electronic Journal of Statistics* **1** 597–615.

MEIER, L., VAN DE GEER, S. and BÜHLMANN, P. (2008). The group Lasso for logistic regression. *Journal of the Royal Statistical Society Series B* **70** 53–71.

MEIER, L., VAN DE GEER, S. and BÜHLMANN, P. (2009). High-dimensional additive modeling. *Annals of Statistics* **37** 3779–3821.

MEINSHAUSEN, N. (2007). Relaxed Lasso. *Computational Statistics & Data Analysis* **52** 374–393.

MEINSHAUSEN, N. (2008). A note on the Lasso for graphical Gaussian model selection. *Statistics and Probability Letters* **78** 880–884.

MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics* **34** 1436–1462.

MEINSHAUSEN, N. and BÜHLMANN, P. (2010). Stability selection (with discussion). *Journal of the Royal Statistical Society Series B* **72** 417–473.

MEINSHAUSEN, N., MEIER, L. and BÜHLMANN, P. (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association* **104** 1671–1681.

MEINSHAUSEN, N. and YU, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics* **37** 246–270.

NARDI, Y. and RINALDO, A. (2008). On the asymptotic properties of the group Lasso estimator for linear models. *Electronic Journal of Statistics* **2** 605–633.

OSBORNE, M., PRESNELL, B. and TURLACH, B. (2000). A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis* **20** 389–403.

PINHEIRO, J. and BATES, D. (2000). *Mixed-Effects Models in S and S-Plus*. Springer, New York.

POLLARD, D. (1984). *Convergence of Stochastic Processes*. Springer.

POLLARD, D. (1990). *Empirical Processes: Theory and Applications*. IMS Lecture Notes.

RAVIKUMAR, P., LAFFERTY, J., LIU, H. and WASSERMAN, L. (2009a). Sparse additive models. *Journal of the Royal Statistical Society Series B* **71** 1009–1030.

RAVIKUMAR, P., WAINWRIGHT, M. and LAFFERTY, J. (2009b). High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Annals of Statistics* **38** 1287–1319.

RAVIKUMAR, P., WAINWRIGHT, M., RASKUTTI, G. and YU, B. (2008). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. ArXiv:0811.3628.

RIDGEWAY, G. (1999). The state of boosting. *Computing Science and Statistics* **31** 172–181.

ROSSET, S. and ZHU, J. (2007). Piecewise linear regularized solution paths. *Annals of Statistics* **35** 1012–1030.

ROTHMAN, A., BICKEL, P., LEVINA, E. and ZHU, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics* **2** 494–515.

SCHAPIRE, R. (2002). The boosting approach to machine learning: an overview. In *MSRI Workshop on Nonlinear Estimation and Classification* (D. Denison, M. Hansen, C. Holmes, B. Mallick and B. Yu, eds.). Springer.

SCHELLDORFER, J., BÜHLMANN, P. and VAN DE GEER, S. (2011). Estimation for high-dimensional linear mixed-effects models using ℓ_1-penalization. *Scandinavian Journal of Statistics* (to appear).

SCHMID, M. and HOTHORN, T. (2008). Boosting additive models using component-wise p-splines as base-learners. *Computational Statistics & Data Analysis* **53** 298–311.

SCHÖLKOPF, B. and SMOLA, A. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge.

SHAWE-TAYLOR, J. and SUN, S. (2010). Discussion of "Stability selection" (auths N. Meinshausen and P. Bühlmann). *Journal of the Royal Statistical Society Series B* **72** 451–453.

SOUTHWELL, R. (1946). *Relaxation Methods in Theoretical Physics*. Oxford University Press.

SPIRTES, P., GLYMOUR, C. and SCHEINES, R. (2000). *Causation, Prediction, and Search*. 2nd ed. MIT Press.

STÄDLER, N., BÜHLMANN, P. and VAN DE GEER, S. (2010). ℓ_1-penalization for mixture regression models (with discussion). *Test* **19** 209–285.

SUN, T. and ZHANG, C.-H. (2010). Discussion of "ℓ_1-penalization for mixture regression models" (auths N. Städler, P. Bühlmann and S. van de Geer). *Test* **19** 270–275.

TALAGRAND, M. (1995). Concentration of measure and isoperimetric inequalities in product spaces. *Publications Mathématiques de l'IHES* **81** 73–205.

TALAGRAND, M. (2005). *The generic chaining: upper and lower bounds of stochastic processes*. Springer Verlag.

TARIGAN, B. and VAN DE GEER, S. (2006). Classifiers of support vector machine type with ℓ_1 complexity regularization. *Bernoulli* **12** 1045–1076.

TEMLYAKOV, V. (2000). Weak greedy algorithms. *Advances in Computational Mathematics* **12** 213–227.

TEMLYAKOV, V. (2008). Nonlinear methods of approximation. *Foundations of Computational Mathematics* **3** 33–107.

TIBSHIRANI, R. (1996). Regression analysis and selection via the Lasso. *Journal of the Royal Statistical Society Series B* **58** 267–288.

TROPP, J. (2004). Greed is good: algorithmic results for sparse approximation. *IEEE Transactions on Information Theory* **50** 2231–2242.

TROPP, J. and GILBERT, A. (2007). Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on Information Theory* **53** 4655–4666.

TSENG, P. (2001). Convergence of a block coordinate descent method for nonsmooth separable minimization. *Journal of Optimization Theory and Applications* **109** 475–494.

TSENG, P. and YUN, S. (2009). A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming, Series B* **117** 387–423.

TSYBAKOV, A. (2004). Optimal aggregation of classifiers in statistical learning. *Annals of Statistics* **32** 135–166.

TUKEY, J. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading, MA.

VAN DE GEER, S. (2000). *Empirical Processes in M-Estimation*. Cambridge University Press.

VAN DE GEER, S. (2001). Least squares estimation with complexity penalties. *Mathematical Methods of Statistics* **10** 355–374.

VAN DE GEER, S. (2002). M-estimation using penalties or sieves. *Journal of Statistical Planning and Inference* **32** 55–69.

VAN DE GEER, S. (2007). The deterministic Lasso. In *JSM proceedings, 2007, 140*. American Statistical Association.

VAN DE GEER, S. (2008). High-dimensional generalized linear models and the Lasso. *Annals of Statistics* **36** 614–645.

VAN DE GEER, S. (2010). The Lasso with within group structure. In *IMS Collections: Festschrift for Jana Jurečková*. IMS, 235–244.

VAN DE GEER, S. and BÜHLMANN, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics* **3** 1360–1392.

VAN DE GEER, S., BÜHLMANN, P. and ZHOU, S. (2010). The adaptive and thresholded Lasso for potentially misspecified models. ArXiv:1001.5176.

VAN DE WIEL, M., BERKHOF, J. and VAN WIERINGEN, W. (2009). Testing the prediction error difference between two predictors. *Biostatistics* **10** 550–560.

VAN DER VAART, A. and WELLNER, J. (1996). *Weak Convergence and Empirical Processes*. Springer Series in Statistics, Springer-Verlag, New York.

VAPNIK, V. (2000). *The Nature of Statistical Learning Theory*. 2nd ed. Statistics for Engineering and Information Science, Springer-Verlag, New York.

VIENS, F. and VIZCARRA, A. (2007). Supremum concentration inequality and modulus of continuity for sub-nth chaos processes. *Journal of Functional Analysis* **248** 1–26.

WAHBA, G. (1990). *Spline models for observational data*. CBMS-NSF Regional Conference Series in Applied Mathematics, 59, SIAM.

WAINWRIGHT, M. (2007). Information-theoretic limitations on sparsity recovery in the high-dimensional and noisy setting. *IEEE Transactions on Information Theory* **55** 5728–5741.

WAINWRIGHT, M. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory* **55** 2183–2202.

WALLACE, D. (1959). Bounds for normal approximations of Student's t and the chi-square distributions. *Annals of Mathematical Statistics* **30** 1121–1130.

WASSERMAN, L. and ROEDER, K. (2009). High dimensional variable selection. *Annals of Statistics* **37** 2178–2201.

WEST, M., BLANCHETTE, C., DRESSMAN, H., HUANG, E., ISHIDA, S., SPANG, R., ZUZAN, H., OLSON, J., MARKS, J. and NEVINS, J. (2001). Predicting the clinical status of human breast cancer by using gene expression profiles. *Proceedings of the National Academy of Sciences* **98** 11462–11467.

WILLE, A. and BÜHLMANN, P. (2006). Low-order conditional independence graphs for inferring genetic networks. *Statistical Applications in Genetics and Molecular Biology* **5** 1–32.

WILLE, A., ZIMMERMANN, P., VRANOVÁ, E., FÜRHOLZ, A., LAULE, O., BLEULER, S., HENNIG, L., PRELIĆ, A., VON ROHR, P., THIELE, L., ZITZLER, E., GRUISSEM, W. and BÜHLMANN, P. (2004). Sparse graphical Gaussian modeling for genetic regulatory network inference. *Genome Biology* **5** R92, 1–13.

WOOD, S. (2006). *Generalized Additive Models: An Introduction with R*. Chapman & Hall.

WU, C. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics* **11**.

YAO, Y., ROSASCO, L. and CAPONNETTO, A. (2007). On early stopping in gradient descent learning. *Constructive Approximation* **26** 289–315.

YEO, G. and BURGE, C. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *Journal of Computational Biology* **11** 475–494.

YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B* **68** 49.

ZHANG, C. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics* **38** 894–942.

ZHANG, C. and HUANG, J. (2008). The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Annals of Statistics* **36** 1567–1594.

ZHANG, T. (2009a). On the consistency of feature selection using greedy least squares regression. *Journal of Machine Learning Research* **10** 555–568.

ZHANG, T. (2009b). Some sharp performance bounds for least squares regression with L1 regularization. *Annals of Statistics* **37** 2109–2144.

ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research* **7** 2541–2563.

ZHOU, S. (2009a). Restricted eigenvalue conditions on subgaussian random matrices. ArXiv:0912.4045.

ZHOU, S. (2009b). Thresholding procedures for high dimensional variable selection and statistical estimation. In *Advances in Neural Information Processing Systems 22*. MIT Press.

ZHOU, S. (2010). Thresholded Lasso for high dimensional variable selection and statistical estimation. ArXiv:1002.1583.

ZHOU, S., RÜTIMANN, P., XU, M. and BÜHLMANN, P. (2010). High-dimensional covariance estimation based on Gaussian graphical models. ArXiv:1009.0530.

ZOU, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association* **101** 1418–1429.

ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the Elastic Net. *Journal of the Royal Statistical Society Series B* **67** 301–320.

ZOU, H., HASTIE, T. and TIBSHIRANI, R. (2007). On the "degrees of freedom" of the Lasso. *Annals of Statistics* **35** 2173–2192.

ZOU, H. and LI, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models (with discussion). *Annals of Statistics* **36** 1509–1566.