Discover Temporal Dynamics of Biomarkers in Predictive Modeling with Longitudinal Data

Jiayu Zhou* Jimeng Sun† Fei Wang† Jianying Hu† Shahram Ebadollahi† Jieping Ye*

Abstract

In the longitudinal study measurements of various biomarkers are taken from the same set of patients repeatedly over long periods. The longitudinal data provides important temporal information about the development of diseases. When applying standard data mining techniques to perform biomarker analysis and build predictive models to study diseases, typically the important temporal relationship among the measurements of the same biomarker is not considered. In this paper we present a novel method for predictive modeling leveraging the temporal information in the longitudinal patients data. Specifically, we propose a temporal dynamics model featured by a structural sparsity regularization designed according to the temporal structures of the longitudinal data. The model simultaneously identifies important biomarkers and discovers their temporal dynamics. We design experiments to show that the proposed method can better capture the temporal dynamics of the biomarkers and discuss related issues when applying the proposed method in healthcare analysis.

1 Introduction

Longitudinal study is widely used in medical researches of many diseases especially in dementia [1, 26] and heart disease [25, 8]. In longitudinal studies, measurements of biomarkers are taken from the same set of patients repeatedly over long periods of time from years to decades. The availability of longitudinal data has attracted intensive research efforts towards methodologies of longitudinal data analysis [16, 5], which have led to many important medical findings.

Currently many types of biomarkers are widely used in the longitudinal study, including biomedical images such as magnetic resonance imaging (MRI) [27] and positron emission tomography (PET) [6], lab tests from plasma [9] and cerebrospinal fluid (CSF) [14], vital signs and gene expression information from microarray [31]. Notably, biomarkers from imaging and gene information are typically high dimensional. Since the cost of many biomarkers (e.g., MRI imaging) is very high, in some longitudinal studies very limited patients that are available. Thus commonly the longitudinal data analysis involves the high-dimensional data with only a few subjects, suffering the so-called *curse of dimensionality* where we may get low-quality predictive models due to severe overfitting.

To tackle the high-dimensional issue there are mainly two approaches: dimension reduction and feature selection. The former approach seeks to project the data into a lower dimensional space that best represents the original data [13]. One disadvantage of this approach is that after the projection, the new features are linear combination of all original features and are thus no longer interpretable. The latter approach can be used to identify a set of features that are relative to the predictive modeling task and in building predictive models we only use these related features (e.g., [7, 15, 17, 23]). In healthcare analysis, identifying meaningful and important biomarkers is a critical step in the study of many diseases and drug development, and therefore feature selection techniques are preferred.

Traditionally, biomarker selection and predictive modeling are two processes that are usually performed independently. To identify the relevant biomarkers, supervised or unsupervised feature selection techniques can be applied to the dataset, and then predictive models are built based only on the selected biomarkers. In high-dimensional space, however, the performance of the aforementioned two-stage approach may be suboptimal. Alternatively, over the last decade has been witness to the rise of simultaneously feature selection and high dimensional predictive modeling via sparse learning [28, 3]. Extensive researches are done in the field of predictive modeling involving sparse learning: study the theories of the sparse learning [2, 4]; enable structure sparsity and hierarchy sparsity [19, 12]; enhance the stability of the research [20, 33]; improve the performance of the computation [29, 30].

Not only an accurate set of relevant biomarkers yields high quality predictive models, it also gives a

[*] Computer Science and Engineering, Arizona State University

[†] IBM T.J. Watson Research Lab

deeper understanding of the nature of the disease. In the longitudinal studies, it is well believed that a small subset of biomarkers are related to the disease progression, and biomarkers involved at different stages may be different (i.e. the dynamics of biomarkers) [11]. This property is called *temporal dynamics* of the biomarkers. The temporal dynamics is related to nature of the disease, and provides important information towards better understanding of the disease. In the development of Alzheimer's disease, for example, CSF biomarkers of Aβ-plaque are early markers before changes of brain structure happens, which is followed by symptoms of memory declination. In the later stage of Alzheimer's the loss of cognitive functionality become obvious [11].

In longitudinal data analysis, the temporal dynamics should be considered in both biomarker selection and predictive modeling. Currently, few existing feature selection method is able to incorporate the temporal information, and no existing solution can be used to explore the temporal patterns of the biomarkers during predictive modeling. In this paper we propose a novel predictive modeling method that is capable of simultaneously building prediction models, selecting the relevant biomarkers, and identifying the temporal dynamic of those biomarkers. Specifically, we propose a convex optimization formulation with structural sparsity that simultaneously identifies a small set of biomarkers that are relevant to the prediction task, identifies the temporal dynamics of relevant biomarkers, allowing different biomarkers to be selected at different time periods, and also incorporates temporal smoothness in the longitudinal study. In this short paper we present experimental results in a synthetical data; more experiments on real-data will be included in a longer version of the paper.

The rest of this paper is organized as follows: Section 2 introduces our approach for biomarker analysis that leverages temporal dynamics; Section 3 presents the experiments showing the effectiveness of our approach that captures temporal dynamics; Section 4 firstly gives analysis of our methods that gives practical guides in real-world applications, and then concludes the paper.

2 Discover Temporal Dynamic of Biomarkers via Structural Regularization

2.1 Simultaneously Predictive Modeling and Biomarker Identification via Sparse Learning In longitudinal studies, measurements of biomarkers are taken from the patients repeatedly at multiple time points (e.g., monthly, semiannually or annually). For example in the 5-year dementia study of Alzheimer's Disease Neuronimaging Initiative (ADNI) [10], each subject takes brain imaging scans (MRI and/or PET) every year. During the data analysis, there are MRI biomarkers, such as the volume of hippocampus, collected from the same subject at each year during the study.

Table 1: The math notations used in this paper.

Symbol	Size	Meaning
X	$n \times d \cdot t$	Patient matrix
w	$d \cdot t \times 1$	Predictive models
y	$n \times 1$	Patient labels
$X_{(i)}$	$n \times d$	Measurements at time point i

A simple way to build predictive models using the longitudinal data is to concatenate all features available for each patient, ignoring the structures inside the longitudinal data. In high dimensional scenarios, we want to only include a small set of biomarkers that are relevant to our modeling and thus sparse inducing norms can be used in the learning to performance simultaneously predictive modeling and biomarker selection. We assume in the following context that there are n patient in the longitudinal study and all of them have been enrolled and followed up for t years, and we assume each year the same type of d measurement is take from all patients. We collectively denote the input data as $X \in \mathbb{R}^{n\times(d\cdot t)} = \{X_{(1)}, X_{(2)}, \ldots, X_{(t)}\}$, where $X_{(i)} \in \mathbb{R}^{n\times d}$ is the measurement data for the time point i, and the corresponding output label is given by y. We focus on linear predictors, i.e., given sample $x \in \mathbb{R}^{dt}$ the predictor is given by $f(x; w) = w^T x$, where w is the model parameter. We summarize the notations used in this paper in Table 1. To build a predictive model on $\{X, y\}$ using sparse learning is to solve the following optimization problem:

$$\min_w L(w, X, y) + \lambda_1 \|w\|_1, \tag{2.1}$$

where L is a convex and smooth loss function. For regression, the least squared loss function can be used:

$$L(w, X, y) = \|Xw - y\|_2^2/n,$$

and for binary classification logistic loss can be used:

$$L(w, b, X, y) = \frac{1}{n}\sum_{i=1}^{n} \log(1 + \exp(-y_i(w^T X_i + b))),$$

where b is the bias term of the model, and X_i is the data of the i-th sample.

2.2 Formulation for Temporal Dynamic Discovery It is well believed that as the disease progresses, biomarkers involved at different stages may be

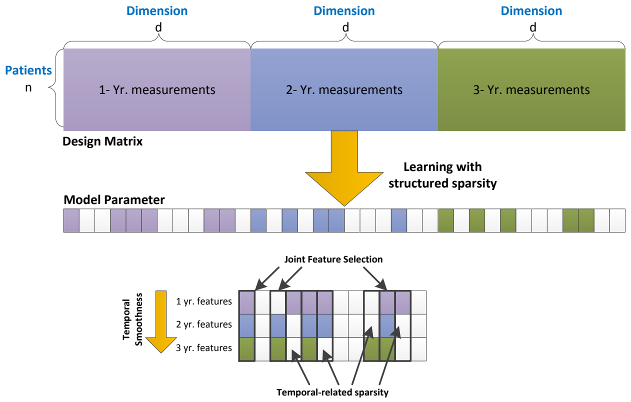

Figure 1: Discover temporal dynamics of the biomarkers using structural sparsity. By stacking the sub-vectors of the predictive model into a matrix, the temporal dynamics of the biomarkers can be shown in a straight forward way. The proposed formulation gives three kinds of sparsity: the first type of sparsity gives elementwise sparsity; the temporal sparsity encourages that the selection status of the same type of feature at two immediate time points should be similar; the group Lasso guarantees that only a small subset of features will be selected at all time points.

different. A hypothetical model of dynamic biomarkers has been proposed in Alzheimer's disease (AD), which suggests that the biomarkers do not reach abnormal levels or peak simultaneously, but do so in an ordered manner, and in order to effectively use the biomarkers, the time-dependent ordering of biomarkers must be thoroughly understood [11]. These statements have posted several requirements to our analysis: due to the dynamics of the biomarker, at different stage the biomarkers should be treated differently, while we also cannot ignore the temporal continuous of the same biomarker at different stages, in the sense that the changes of the same biomarker between immediate stages should not be too large. In this paper we assume that the patients are aligned according to a known anchor time such as diagnosis date or disease start date in the retrospective study. Therefore such temporal relationship among disease stages can be considered as that among different time points in the longitudinal data.

In each time point of the longitudinal study, the same set of measurement types are used. It is intuitive to assume that model parameters of the same type of feature are related to each other. Building model using Lasso does not consider such relatedness among features. We propose to use fused Lasso term [18] to penalize the difference of measurements between immediate time points of the same biomarker to capture such relatedness:

$$\sum_{i=1}^{d} \sum_{j=2}^{t} \|w_{(j-2)d+i} - w_{(j-1)d+i}\|_1,$$

According to elementary linear algebra, the fused term can be succinctly represented in a structure matrix $R \in \mathbb{R}^{d(t-1)\times dt}$ such that each row of R consists $1, -1$ at a pair of immediate time points of one biomarker, and 0 at all other places. The formulation is thus given by:

$$\min_w L(w, X, y) + \lambda_1 \|w\|_1 + \lambda_2 \|Rw\|_1, \tag{2.2}$$

where λ_1 and λ_2 are parameters controlling elementwise sparsity and fused sparsity, respectively. Besides the temporal smoothness, we also want to perform joint feature selection, which selects only a small subset of features at all time points. In other words, the values of the model for some biomarkers at all time points are zero. To achieve this, we incorporate $\ell_{2,1}$-norm regularization to achieve the joint feature selection. Let the vector $\hat{w}_i = [w_i \ w_{d+i} \ \dots \ w_{(t-1)d+i}]^T$ collectively denote the different time point of the i-th biomarker, our formulation is given by:

$$\min_w L(w, X, y) + \lambda_1 \|w\|_1 + \lambda_2 \|Rw\|_1 + \lambda_3 \sum_{i=1}^{d} \|\hat{w}_i\|_2, \tag{2.3}$$

where λ_3 is the parameter used to control the number of biomarkers involved in the model. The illustration of building model using this approach is given in Figure 1. The proposed formulation gives several kinds of sparsity. The first type of sparsity is similar to Lasso that gives 'random' sparsity. The second type of sparsity is temporal sparsity, where the selection status of the same type of feature at two immediate time points should be similar (they tend to both be selected by the algorithm or both not selected). The third type of sparsity is introduced by the $\ell_{2,1}$, which guarantees that a small subset of features will be selected at all time points.

2.3 Optimization Algorithms

The formulation in Eq. (2.3) is convex optimization problem with smooth and non-smooth parts. We propose to solve the problem by the accelerated gradient method (AGM) [22, 21]. One of the key steps in using AGM is the computation of the proximal operator associated with the composite of non-smooth penalties defined as follows:

$$\begin{aligned} \pi(v) = \operatorname*{argmin}_w \ & \frac{1}{2}\|w - v\|_2^2 + \lambda_1 \|w\|_1 \\ & + \lambda_2 \|Rw\|_1 + \lambda_3 \sum_{i=1}^{d} \|\hat{w}_i\|_2. \end{aligned} \tag{2.4}$$

Denote $\hat{v}_i = [v_i \; v_{d+i} \; \dots \; v_{(t-1)d+i}]^T$, we notice that th problem in Eq. (2.4) is decoupled for each w_i and v_i:

$$\pi(\hat{v}_i) = \operatorname*{argmin}_{\hat{w}_i} \frac{1}{2}\|\hat{w}_i - \hat{v}_i\|_2^2 + \lambda_1\|\hat{w}_i\|_1 + \lambda_2\|R\hat{w}_i\|_1 + \lambda_3\|\hat{w}_i\|_2. \tag{2.5}$$

The subproblem in Eq. (2.5) can be decompose into two steps and solved efficiently [32]. To make th paper self-contained we summarize the result in th following lemma.

Lemma 2.1. *[32] Define*

$$\pi_{FL}(\hat{v}_i) = \operatorname*{argmin}_{\hat{w}_i} \frac{1}{2}\|\hat{w}_i - \hat{v}_i\|_2^2 + \lambda_1\|\hat{w}_i\|_1 + \lambda_2\|R\hat{w}_i\|_1$$

$$\pi_{GL}(\hat{u}_i) = \operatorname*{argmin}_{\hat{w}_i} \frac{1}{2}\|\hat{w}_i - \hat{u}_i\|_2^2 + \lambda_3\|\hat{w}_i\|_2.$$

Then the following holds:

$$\pi(\hat{v}_i) = \pi_{GL}(\pi_{FL}(\hat{w}_i)).$$

The fused Lasso projection π_{FL} can be efficiently solved using [18] and the ℓ_1 projection has a closed solution via soft-thresholding [28]. We summarize the overall APM algorithm for solving 2.3 in Algorithm 1.

Algorithm 1 Optimization Algorithm for Solving 2.3

```
Input: $w_0$, $\gamma_0 \in \mathbb{R}$, regularization parameters $\lambda_1$, $\lambda_2$, $\lambda_3$ and max iteration number $q$.
Output: $w$.
Set $w_1 = w_0$, $t_{-1} = 0$, and $t_0 = 1$.
for $i = 1$ to $q$ do
    Set $\beta_i = (t_{i-2} - 1)/t_{i-1}$,
    $s_i = w_i + \beta_i(w_i - w_{i-1})$.
    while (true)
        Compute $w^* = \mathrm{argmin}_w \mathcal{P}_{\gamma,s}(s_i + \nabla_{s_i}\mathcal{L}/\gamma_i)$
        if $f(w^*) \leq \mathcal{P}_{\gamma,s_i}(w^*)$ then break the while loop
            else set $\gamma_i = \gamma_i \times 2$.
        end-if
    end-while
    Set $w_{i+1} = w^*$ and $\gamma_{i+1} = \gamma_i$.
    if stopping criterion is satisfied then break the for loop.
    Set $t_i = (1 + \sqrt{1 + 4t_{i-1}^2})/2$.
end-for
Set $w = w_{i+1}$.
```

where $\mathcal{P}_{\gamma,s}(w)$ denotes the proximal operator that can be show to solve Eq. (2.4):

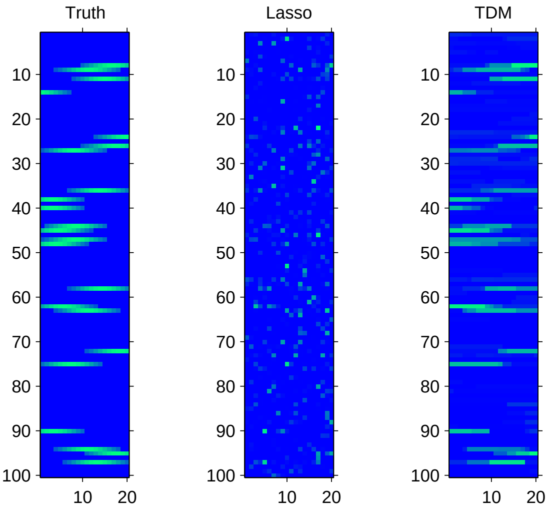

Figure 2: The stacked view of the ground truth model (left), and models recovered by Lasso (middle) and our proposed temporal dynamic model (right). Each row represents a biomarker and each column is a time point. The sparsity obtained using Lasso is close to random, while the TDM is able to identify all relevant biomarkers (and a few false positives) and their temporal patterns.

$$\mathcal{P}_{\gamma,s}(w) = f(s) + \langle \nabla_w f(s), w - s\rangle + \frac{\gamma}{2}\|w - s\|_F^2 + \lambda_1\|w\|_1 + \lambda_2\|Rw\|_1 + \lambda_3\sum_{i=1}^{d}\|\hat{w}_i\|_2$$

The APM algorithm in Algorithm 1 has convergence speed $O(\frac{1}{k^2})$, where k is the iteration number, which is proven to be the optimal first order convergence rate.

3 Experiments

In this short paper we only present the experimental results of an synthetical data; experiments on real-data will be included in a longer version of this paper.

We design experiment to show that the proposed method has the capability of capturing temporal dynamics in the sense that it gives desired sparsity when patterns of such temporal dynamics presents. We firstly construct our ground truth model w that consists of 20 time points and 100 biomarkers (this in total gives 2000 features), in which only 23 biomarkers are relevant to the prediction and have non-zero weights at a subset of time points. Among the 23 biomarkers, we assume that the weights of each biomarkers are normally distributed such that the weight of one biomarker is peaked randomly at one time point and has a standard deviation of 5, and we set the values that are 3 times less than the peak value to be zeros. Therefore the model w simulates the temporal dynamics and the stacked view (each row represents a biomarker and each column is a time point) of W is shown in Figure 2 (left). We then randomly sample 200 data points to form the data matrix $X \in \mathbb{R}^{200\times 1000} \sim \mathcal{N}(0,1)$, and generate the label $y = Xw + \varepsilon$ where is element-wise Gaussian noise $\varepsilon \sim \mathcal{N}(0, 0.1)$.

We apply Lasso and the proposed temporal dynamic model (TDM) to learn the model, and the parameters are estimated using cross validation. We present the learnt models in the stacked view in Figure 2. We see that the sparsity obtained using Lasso is close to random, while the TDM is able to identify all relevant biomarkers (and a few false positives) and their temporal patterns. We observe an obvious advantage of TDM that better recovers the underlying structure of the model than Lasso.

4 Discussion and Conclusion

Longitudinal study is widely used in medical researches where measurements of biomarkers are taken from the same set of patients repeatedly over long periods of time. In the longitudinal data analysis, few existing method is able to incorporate the temporal information, and no existing solution can be used to discover the temporal dynamics of the biomarkers during predictive modeling. In this paper we present a novel temporal dynamic model for predictive modeling on longitudinal data that simultaneously identifies relevant biomarkers and discovers the temporal dynamics of these biomarkers.

The proposed model is based on sparse inducing norms and its structure sparsity is designed according to the temporal structures of the longitudinal data. Specifically we consider temporal relation of the measurements of the same biomarker at the different stages of diseases. If the biomarker is not relevant to the prediction task, measurements at all time points should have zero weights. It also encourages the temporal smoothness that requires the change of measurements at different time point to be small for the same biomarker. The formulation includes Lasso [28], fused Lasso [18] and group Lasso [19] as special cases.

The temporal dynamic model requires the longitudinal data to fit into regular time frames or time points. In some retrospective cohorts, however, patient visit record may not follow a certain time frames. For example, a patient may only visit the hospital when certain symptoms occurs. In this case our purposed model cannot be directly applied. However, as long as the patients are aligned to an anchor time point (the onset of certain disease), one can always select a proper granularity to aggregate the visit data in a certain time window to form time points. For instance, we can select a time window of one year and within each window we average the visit data into a single measurement to represent this window.

The pathology of some diseases may be very complex, e.g., the Alzheimer's disease where the stages of the disease may have clear boundaries and therefore it is hard to give the exact stage for a certain patient [24]. Even if we have certain criteria to classify a patient into categories of Alzheimer's, mild cognitive impairment and normal control, they are too coarse to be an anchor for aligning patients. This then raise the problem that how to find an anchor time point and how we can align the patients when such an anchor cannot be obtained. Indeed domain expert may be able to provide finer stage information about a patient, but it is too time consuming for complex diseases such as Alzheimer's, even in a small cohort. One promising direction is to design algorithms that automatically align the patients during the predictive modeling according the biomarker measurements, which is our plan for future works.

In some studies, due to some reasons the measurement of each time point is not exactly the same. For example, the image data may not be available at the 2nd time point but still available in the following time points. In such cases, the proposed formulation can not be directly applied, since the lengths of the model are not exactly the same at different time points. But it is easy to extend the model to handle such 'missing' features by adjusting the groups in the group Lasso, i.e., allowing groups of different sizes.

Acknowledgments

This work was supported in part by NSF (IIS-0953662, CCF-1025177) and NIH (R01LM010730).

References

[1] A. Baddeley, S. Bressi, S. Della Sala, R. Logie, and H. Spinnler. The decline of working memory in alzheimer's disease a longitudinal study. *Brain*, 114(6):2521–2542, 1991.

[2] R. Baraniuk. Compressive sensing [lecture notes]. *Signal Processing Magazine, IEEE*, 24(4):118–121, 2007.

[3] P. Bühlmann and S. Van De Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, 2011.

[4] E. Candès. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(9):589–592, 2008.

[5] P. Diggle, P. Heagerty, K. Liang, and S. Zeger. *Analysis of longitudinal data*, volume 25. Oxford University Press, USA, 2002.

[6] A. Drzezga, N. Lautenschlager, H. Siebner, M. Riemenschneider, F. Willoch, S. Minoshima, M. Schwaiger, and A. Kurz. Cerebral metabolic changes accompanying conversion of mild cognitive impairment into alzheimer's disease: a pet follow-up study. *European journal of nuclear medicine and molecular imaging*, 30(8):1104–1113, 2003.

[7] R. Duda, P. Hart, and D. Stork. Pattern classification and scene analysis 2nd ed. 1995.

[8] J. Eriksson, T. Forsen, J. Tuomilehto, P. Winter, C. Osmond, and D. Barker. Catch-up growth in childhood and death from coronary heart disease: longitudinal study. *Bmj*, 318(7181):427–431, 1999.

[9] A. Hye, S. Lynham, M. Thambisetty, M. Causevic, J. Campbell, H. Byers, C. Hooper, F. Rijsdijk, S. Tabrizi, S. Banner, et al. Proteome-based plasma biomarkers for alzheimer's disease. *Brain*, 129(11):3042–3050, 2006.

[10] C. Jack, M. Bernstein, N. Fox, P. Thompson, G. Alexander, D. Harvey, B. Borowski, P. Britson, J. L Whitwell, C. Ward, et al. The alzheimer's disease neuroimaging initiative (adni): Mri methods. *Journal of Magnetic Resonance Imaging*, 27(4):685–691, 2008.

[11] C. Jack Jr, D. Knopman, W. Jagust, L. Shaw, P. Aisen, M. Weiner, R. Petersen, and J. Trojanowski. Hypothetical model of dynamic biomarkers of the alzheimers pathological cascade. *Lancet neurology*, 9(1):119, 2010.

[12] R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. Proximal methods for sparse hierarchical dictionary learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2010.

[13] I. Jolliffe. *Principal component analysis*. Wiley Online Library, 2005.

[14] M. Kanai, E. Matsubara, K. Isoe, K. Urakami, K. Nakashima, H. Arai, H. Sasaki, K. Abe, T. Iwatsubo, T. Kosaka, et al. Longitudinal study of cerebrospinal fluid levels of tau, aβ1–40, and aβ1–42 (43) in alzheimer's disease: a study in japan. *Annals of neurology*, 44(1):17–26, 2004.

[15] K. Kira and L. Rendell. A practical approach to feature selection. In *Proceedings of the ninth international workshop on Machine learning*, pages 249–256. Morgan Kaufmann Publishers Inc., 1992.

[16] K. Liang and S. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.

[17] H. Liu and R. Setiono. Chi2: Feature selection and discretization of numeric attributes. In *Tools with Artificial Intelligence, 1995. Proceedings., Seventh International Conference on*, pages 388–391. IEEE, 1995.

[18] J. Liu, L. Yuan, and J. Ye. An efficient algorithm for a class of fused lasso problems. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 323–332. ACM, 2010.

[19] L. Meier, S. Van De Geer, and P. Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008.

[20] N. Meinshausen and P. Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

[21] A. Nemirovski. Efficient methods in convex programming. 2005.

[22] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*. Springer Netherlands, 2004.

[23] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(8):1226–1238, 2005.

[24] J. Petrella, R. Coleman, and P. Doraiswamy. Neuroimaging and early diagnosis of alzheimer disease: A look to the future1. *Radiology*, 226(2):315–336, 2003.

[25] B. Saltin, G. Blomqvist, J. Mitchell, R. Johnson Jr, K. Wildenthal, C. Chapman, E. Frenkel, W. Norton, M. Siperstein, W. Suki, et al. A longitudinal study of adaptive changes in oxygen transport and body composition. *Circulation*, 38(5S7):VII–1, 1968.

[26] I. Skoog, L. Nilsson, G. Persson, B. Lernfelt, S. Landahl, B. Palmertz, L. Andreasson, A. Oden, and A. Svanborg. 15-year longitudinal study of blood pressure and dementia. *The Lancet*, 347(9009):1141–1145, 1996.

[27] T. Stoub, M. Bulgakova, S. Leurgans, D. Bennett, D. Fleischman, D. Turner, et al. Mri predictors of risk of incident alzheimer disease a longitudinal study. *Neurology*, 64(9):1520–1524, 2005.

[28] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[29] R. Tibshirani, J. Bien, J. Friedman, T. Hastie, N. Simon, J. Taylor, and R. Tibshirani. Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2012.

[30] J. Wang, B. Lin, P. Gong, P. Wonka, and J. Ye. Lasso screening rules via dual polytope projection. *arXiv preprint arXiv:1211.3966*, 2012.

[31] W. Yao, Z. Cheng, C. Busse, A. Pham, M. Nakamura, and N. Lane. Glucocorticoid excess in mice results in early activation of osteoclastogenesis and adipogenesis and prolonged suppression of osteogenesis: A longitudinal study of gene expression in bone tissue from glucocorticoid-treated mice. *Arthritis & Rheumatism*, 58(6):1674–1686, 2008.

[32] J. Zhou, J. Liu, V. Narayan, and J. Ye. Modeling disease progression via fused sparse group lasso. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1095–1103. ACM, 2012.

[33] J. Zhou, J. Sun, Y. Liu, J. Hu, and J. Ye. Patient risk prediction model via top-k stability selection. In *Proceedings of the 13th SIAM International Conference on Data Mining*, 2013.