

Prediction via Orthogonalized Model Mixing

MERLISE CLYDE, HEATHER DESIMONE AND GIOVANNI PARMIGIANI

Merlise Clyde and Giovanni Parmigiani are Assistant Professors, and Heather DeSimone is PhD candidate, Institute of Statistics and Decision Sciences, Duke University, Durham, NC 27708-0251. Work partially supported by the National Science Foundation under grant DMS-9305699 and by the Arts and Sciences Research Council of Duke University. Stan Young and Alan Menius of Glaxo Research Institute provided the data set discussed in Section 6. The associate editor and reviewers provided insightful comments and very helpful suggestions.

Abstract

In this paper we introduce an approach and algorithms for model mixing in large prediction problems with correlated predictors. We focus on the choice of predictors in linear models, and mix over possible subsets of candidate predictors. Our approach is based on expressing the space of models in terms of an orthogonalization of the design matrix. Advantages are both statistical and computational. Statistically, orthogonalization often leads to a reduction in the number of competing models by eliminating correlations. Computationally, large model spaces cannot be enumerated; recent approaches are based on sampling models with high posterior probability via Markov chains. Based on orthogonalization of the space of candidate predictors, we can approximate the posterior probabilities of models by products of predictor-specific terms. This leads to an importance sampling function for sampling directly from the joint distribution over the model space, without resorting to Markov chains. Compared to the latter, orthogonalized model mixing by importance sampling is faster in sampling models, and is also more efficient in finding models that contribute significantly to the prediction. Further advantages are in the speed of convergence and the availability of more reliable convergence diagnostic tools.

We illustrate these in practice, using a data set on prediction of crime rates. The model space is small enough that enumeration of all models is available for comparison and convergence checks. Also, we demonstrate the feasibility of orthogonalized model mixing in a large size problem, which is very difficult to attack by other methods. The data set is from a designed experiment dealing with predicting protein activity under different storage conditions. The model space is large (the rank of the design matrix is 88) and very difficult to explore if expressed in terms of the original variables. We obtain prediction intervals and a probability distribution of the setting that produces the highest response.

KEY WORDS: Variable Selection, Model Uncertainty, Importance Sampling, Bayesian Linear Models.

1. INTRODUCTION

Advances in statistical methodology and computing have made available powerful modeling tools in a variety of areas. Along with the added modeling flexibility, increasing attention needs to be directed to assessing the consequences of modeling assumptions, and to propagating model uncertainty to conclusions. Debates on the effect of modeling assumptions on crucial scientific and policy prediction, such as global warming and the health impact of toxic waste, have reached the mass media. In such complex modeling problems, predictions based on choosing a single model are often not satisfactory, a fact that has been long recognized in the literature (see for example Weisberg, 1985).

Bayesian methods offer a very effective and conceptually appealing alternative: predictions can be based on a set of plausible models rather than a single model; each model contributes to the prediction proportionally to the support it receives from the observed data and prior knowledge. As the predictive distribution is a mixture distribution, we refer to this approach as model mixing, as opposed to the more conventional approach of model selection. Model mixing has a long history (de Finetti 1937, Leamer 1978), but its use in applications is only recently being explored, following the recent advances in Bayesian computing, and, in particular, in Markov chain Monte Carlo methods. Recent articles include Draper (1994) and Raftery, Madigan and Volinsky (1995), both with discussion and extensive references.

One aspect of statistical modeling that is typically difficult, and also crucial in influencing predictions, is the selection of the predictor variables to be included in a model. In realistic formulations, the list of candidate predictors includes transformations of the variables originally recorded and interactions between them, and is necessarily large. Even for problems of moderate size, it may be computationally infeasible to make predictions based on all models, or even to select useful models based on complete enumeration of all models. The challenge is, therefore, that of finding efficient ways of exploring the space of models, selecting plausible ones, and attributing to each of them a weight (approximating the posterior probability) for the mixing-based prediction.

In recent years, several very effective methods have been proposed for finding models with high posterior probability, without enumerating the model space. For example, variable selection methods based on sampling from the model space using Markov chains are the Stochastic Search Variable Selection, or SSVS (George and McCulloch 1993, 1994),

the Markov chain Monte Carlo Model Composition, or MC³ (Madigan and York 1993) and the methods of Carlin and Chib (1995), Geweke (1994), Phillips and Smith (1994) and Green (1995). A deterministic algorithm is Occam's window (Madigan and Raftery 1994). The resulting collection of models, or sometimes a further subset, can then be used for model mixing. Examples include applications of Occam's Window and MC³ to linear models (Raftery, Madigan and Hoeting 1993), proportional hazard models (Raftery, Madigan and Volinsky 1995) and graphical models (Madigan and York 1993, Madigan and Raftery 1994, and Madigan, Gavrin and Raftery 1994) and applications of SSVS to designed experiments (Chipman, 1994 and Clyde and Parmigiani, 1994) generalized linear models (George, McCulloch and Tsay 1994) and population models in pharmacokinetics (Bennett and Wakefield, 1994); related analyses are also discussed by Draper (1994) and Higdon (1994).

The focus of this paper is on model mixing for prediction. In this context, there are opportunities for constructing models and algorithms that can be substantially more effective than those originally designed for variable selection. From the point of view of prediction, mixing over models with different predictor sets can be seen as a more general and powerful model. Practically, the added generality offers more realistic uncertainty assessment, as well as ways of incorporating information from all predictors without over fitting the data. The latter is achieved by a data-based shrinkage of the regression coefficients (see also George, 1986a, 1986b). In this paper we propose to approach model mixing by expressing the model space in terms of an orthogonal transformation of the matrix of predictors. This strategy defines a new class of mixture models: the orthogonalized model mixing class. Advantages of this over mixing in the original variable space occur in at least two fundamental ways: First, it is simpler and substantially faster to sample models; Second, the number of competing plausible models is usually smaller as a result of eliminating near-multicollinearity. A drawback is the more difficult elicitation of the prior probability distribution over the model space.

The plan of the paper is the following. In section 2, we introduce orthogonalized model mixing by giving the basic notation and definitions. In section 3, we propose an algorithm for sampling models. We approximate the posterior probability of a model by a product of independent Bernoulli random variables, each indicating whether an element of the orthogonal basis is included or not. Such probabilities are then used as an

importance sampling function over the new model space. Independence allows for efficient coverage of the model space. In particular, a crucial advantage of this approach is that one can sample directly from the approximate posterior distribution, so that many of the problems associated with Markov chains (Clyde and Parmigiani, 1994) are substantially mitigated. These include large time requirements for the adequate simulation of large chains, difficulty in traversing the model space because of correlation between variables and between successive draws in MCMC, and difficulty in assessing convergence.

Quantities of interest, such as predictive distributions and expected utilities, depend on all models, but need to be evaluated based on the subset of sampled models. In section 4, we review and compare simple alternative estimation strategies, based on viewing the problem as discovery sampling.

Next we examine the performance of orthogonalized model mixing in two applications. The first application, presented in Section 5, is to the crime data of Vandaele (1978). The model space is relatively small ($2^{15} = 32,768$ models), so that enumeration of all models is available for comparison and convergence checks. We compare orthogonalized model mixing with alternatives based on Markov chain approaches. We illustrate the fact that orthogonalized model mixing with importance sampling is faster in sampling models, and in addition it tends to focus on models with high posterior probability. We also include a brief comparison of prediction estimators based on the sample of discovered models. The second data set, presented in Section 6, is from a designed experiment dealing with predicting protein activity under different storage conditions. The model space is large (the rank of the design matrix is 88) and very difficult to explore if expressed in terms of the original variables. We use this example to illustrate the feasibility of orthogonalized model mixing in problems with very large dimensionality. We obtain prediction intervals and a probability distribution of the design setting that produces the highest response.

2. ORTHOGONALIZED MODEL MIXING

2.1. *Prior Distributions and Orthogonalization for the Full Model*

We begin by giving the basic notation and definitions. Let Y be the $n \times 1$ vector of observed values of the response variable, and X the $n \times p$ design matrix including all candidate predictors. X can include transformations of the variables originally recorded.

We begin by assuming that

$$Y|\beta, \sigma^2 \sim N_n(X\beta, \sigma^2 I_n), \quad (1)$$

where β is $p \times 1$, σ^2 is a scalar, and I_n is the $n \times n$ identity matrix. We term this the full model, as it includes all candidate predictors. We take the prior distribution for the model parameters to be the natural conjugate prior:

$$\begin{aligned} \beta|\sigma^2 &\sim N_p(b_0, \sigma^2 B) \\ \nu\xi/\sigma^2 &\sim \chi_\nu^2, \end{aligned}$$

where B , b_0 , ν and ξ are fixed hyper-parameters. Elicitation of these parameters is discussed by Kadane *et al.* (1980) and by Garthwaite and Dickey (1992) in the context of variable selection.

Consider now a transformation, $Z = XW$, of the design matrix, such that the columns of Z are orthogonal. The mean space is represented as a collection of subspaces spanned by the columns of Z . The linear model in (1) can be rewritten in terms of the orthogonalized variables as

$$Y = Z\alpha + e,$$

where $\alpha = W^{-1}\beta$, and $e \sim N(0, \sigma^2 I_n)$. The transformed prior distribution on α , conditional on inclusion of all predictors is:

$$\alpha|\sigma^2 \sim N_p(a_0, \sigma^2 A), \quad (2)$$

where $a_0 = W^{-1}b_0$ and $A = W^{-1}B(W^{-1})'$. The prior distribution on σ is unchanged by the transformation to orthogonal variables. In implementing an importance sampler, we will exploit the orthogonality of Z . Further simplifications are obtained when A is diagonal, and we will require this throughout. When B is specified a priori based on expert opinion, a diagonal A results from suitably choosing the orthogonalization strategy. This is the approach taken here. When the orthogonalization needs to be arbitrary, a diagonal A can be achieved by restricting the choices for B .

In model mixing, different orthogonal bases can lead to different predictive distributions, and specification of the basis is an open ended modeling problem. In some cases the basis may be driven by the problem, as is the case with some designed experiments. In other cases, specification can be guided by features of the basis, such as smoothness in wavelet-based curve fitting. In general, different orthogonalizations may result in different degrees of parsimony in the representation of the mean space of Y . One would like the orthogonalization to achieve closeness to “target” or “optimal” subspaces. This may be better achieved by an orthogonalization based on Y , such as partial least squares (Wold *et al.* 1984) or sliced inverse regression (Li 1991). This, however, means introducing uncertainty due to sampling variation in the orthogonal basis, and, in this setup, data dependence in the prior distribution on α . In Clyde and Parmigiani (1995) we present several alternative orthogonalization strategies in detail. Further discussion is also in Section 7.

The remainder of this paper is based on constructing W via generalized principal components (Rao 1964), a well understood basis, for which computing routines are readily available. The resulting orthogonal variables are invariant under reordering and rescaling the original predictors, and do not require knowledge of the response Y . This feature is appealing in non-orthogonal designed experiments, as it frees the orthogonalization process from sampling variation.

In standard principal components analysis, the matrix W is given by the eigenvectors of $X'X$. In generalized principal components, W is orthonormal with respect to an inner product determined by a given $p \times p$ positive definite matrix. In our context, it is convenient to choose this matrix to be B^{-1} , where B is the covariance matrix of β . Let $B = CC'$, where C is a square root of B . The requirement that A is diagonal can always be achieved by taking W to be CU , where U corresponds to the eigenvectors of $C'X'XC$. This amounts to first rotating X to XC and β to $C^{-1}\beta$, so that the rotated parameters are now independent, and then determining Z based on standard principal components in the rotated variables. The resulting prior covariance for α is $\sigma^2 I_p$. This provides a simple way of accommodating an arbitrarily specified B .

In summary, our preferred strategy for the specification of the prior on the full model is based on assigning a prior distribution on β , which determines both the inner product in the generalized principal components and the prior distribution for α . Alternatives based

on specifying the prior distribution directly on α are discussed in Clyde and Parmigiani (1995).

2.2. Model Mixing

The process of selecting columns of Z for prediction is modeled via a further hierarchical level in the prior distribution. In particular, define the $p \times 1$ vector γ to be a sequence of binary random variables, each indicating whether the corresponding column of Z is included in the model. The set of all possible γ 's will be referred to as the orthogonalized model space when ambiguity with the original model space may occur. This specification is equivalent to assuming that the prior distribution on α is a mixture of (2) and a point mass at zero. Similar priors are used for variable selection on the original model space by Mitchell and Beauchamp (1988) and Madigan and York (1993), among others, and are a limiting case of the more general formulation of George and McCulloch (1993, 1994).

The prior distribution on γ is denoted by $\pi(\gamma)$. In section 3, we will make the additional assumption that $\pi(\gamma)$ factors as

$$\pi(\gamma) = \prod_{i=1}^p \pi(\gamma_i) \equiv \prod_{i=1}^p \theta_i^{\gamma_i} (1 - \theta_i)^{1-\gamma_i}. \quad (3)$$

Elicitation of the θ_i 's can be guided by the degree of parsimony in representing the target subspaces for the mean response or by the resulting amount of shrinkage, as discussed further in section 2.3. In particular, taking θ_i 's less than .5 enforces a penalty for each additional term in the model (see Clyde and Parmigiani, 1995).

Our prior specification identifies prior distributions for the coefficients α and β given any γ . Importantly, even if, as we suggest, one first assigns the prior distribution on β given $\gamma = \mathbf{1}$, and then derives the prior distribution on α from it, the implied prior distribution on β given $\gamma \neq \mathbf{1}$ will depend on the representation of the model space via the columns of Z . When the orthogonalization is constructed based on contrasts of interest, it is appealing to assign a point mass to some of the α_i 's being 0, rather than to β_i 's being 0. In general, however, it may be hard to interpret conditional prior distributions in the original space. As our goal is prediction, the primary concern is the selection of columns of Z , which identify interesting subspaces to represent the mean response. We

find it attractive to give priority to this, both from a modelling and a computational perspective. In variable selection, other strategies may be preferable.

Under this choice of orthogonalization and the conjugate prior distributions, computation of posterior and predictive distributions can be carried out using standard least squares regression techniques, by augmenting the Y and Z matrices as follows:

$$\tilde{Y} = \begin{bmatrix} Y \\ a_0 \end{bmatrix} \quad \text{and} \quad \tilde{Z} = \begin{bmatrix} Z \\ I_p \end{bmatrix}$$

where \tilde{Y} is $(n+p) \times 1$ and \tilde{Z} is $(n+p) \times p$. Next, let \tilde{z}_i be the i -th column of \tilde{Z} . Define SSR_i^2 as the regression sum of squares from the regression of \tilde{Y} on \tilde{z}_i . In particular, then $\text{SSR}_i^2 = \|P_{\tilde{z}_i}\tilde{Y}\|^2$, where $P_{\tilde{z}_i} = \tilde{z}_i\tilde{z}_i'/(\tilde{z}_i'\tilde{z}_i)$ is the projection operator on the column \tilde{z}_i . Also, the matrix $\tilde{Z}'\tilde{Z}$ is diagonal with generic element d_i . The posterior probabilities of models are available in closed form up to a normalizing constant, that is:

$$\pi(\gamma|Y) = \frac{p(Y|\gamma)\pi(\gamma)}{\sum_{\gamma'} p(Y|\gamma')\pi(\gamma')} = \frac{q_\gamma}{\sum_{\gamma'} q_{\gamma'}}. \quad (4)$$

Conjugate updating and straightforward manipulations lead to the following convenient expression for q_γ :

$$\log(q_\gamma) = \sum_{i=1}^p \gamma_i \left[\log \left(\frac{\theta_i}{1-\theta_i} \right) - \frac{1}{2} \log d_i \right] - \frac{(n+\nu)}{2} \log \left(\nu\xi + \tilde{Y}'\tilde{Y} - \sum_{i=1}^p \gamma_i \text{SSR}_i^2 \right). \quad (5)$$

We now have all the necessary elements for addressing predictive problems, such as finding the multivariate predictive distribution $f(\cdot|Y, X^*)$, where X^* is a specified matrix, the mean $\hat{Y}(X^*)$ of this distribution, the expected utility U associated with a decision δ whose outcome depends on future values of Y , or other quantities of interest. Denote the quantity of interest by ϕ , possibly a vector. In many cases, computations can proceed by

determining ϕ_γ (the quantity of interest conditional on γ) for each γ , and then evaluating

$$\phi = \sum_{\gamma} \phi_{\gamma} \pi(\gamma|Y). \quad (6)$$

For example, (6) can be used to determine the predictive distribution for future observations, which is a mixture of the predictive distributions based on the individual models. Computing $\pi(\gamma|Y)$ from (4) or ϕ in (6), involves summing over all possible models, which is computationally infeasible for relatively large p . This motivates interest for stochastic searches of the model space, discussed in Section 3.

2.3. Multiple Shrinkage

In model mixing, all columns of Z contribute to some extent to the prediction. The prior specification affects the smoothness of the predicted response. Prior distributions that put large weights on models with a small number of columns encourage more shrinkage. On the other extreme, a prior distribution concentrating on $\gamma = \mathbf{1}$ corresponds to the full model, which often over fits the points. Regression on a subset of size k of the principal components based on the k largest eigenvalues is sometimes used to alleviate over fitting and multicollinearity. One potential drawback is that it can lead to exclusion of important directions that have small eigenvalues but are highly correlated with Y . See Jolliffe (1982) for examples. On the contrary, model mixing does not suffer from this, as all possible subsets of principal components are incorporated in the regression.

To give the flavor of the shrinkage implication of model mixing, it is interesting to consider the form of the Bayes estimator of α . Under the full model, this is given by: $\tilde{\alpha} = (\tilde{Z}'\tilde{Z})^{-1}\tilde{Z}'\tilde{Y}$. Since the columns of Z are orthogonal, the matrix $\Lambda \equiv Z'Z$ is diagonal and the Bayes estimate of α under the full model can be computed coordinate-wise as

$$\tilde{\alpha}_i = \frac{\lambda_i}{\lambda_i + 1} \hat{\alpha}_i + \frac{1}{\lambda_i + 1} a_{0i},$$

where $\hat{\alpha}$ is the OLS estimator of α , λ_i is the i -th diagonal element of Λ , and a_{0i} is the i -th element of a_0 . There is differential shrinkage of each component to the prior mean.

If $z^* = x^*CU$, where x^* is p -dimensional row vector, the predictive mean of y at x^* under model γ is $\sum z_i^* \gamma_i \tilde{\alpha}_i$. Under model mixing, $E(\gamma_i \tilde{\alpha}_i | Y) = \pi(\gamma_i = 1 | Y) \tilde{\alpha}_i$, and the predictive mean is $\sum z_i^* \pi(\gamma_i = 1 | Y) \tilde{\alpha}_i$. This emphasizes the multiple shrinkage nature of model mixing. Some shrinkage is incorporated in $\tilde{\alpha}_i$, which depends on Λ and B . Further shrinkage is determined by the posterior model probabilities, that depend on Λ , B , and also on the model specific regression sum of squares SSR_i , the prior model probabilities θ_i , and the prior hyperparameters ν and ξ for the error variance. The role of these parameters in determining the amount of shrinkage from the model probability can be understood from formula (5). Evaluating the amount of shrinkage for some key models can provide insight on the strength of the prior specification being used. See George (1986a and 1986b) for further discussion of multiple shrinkage estimators.

3. AN IMPORTANCE SAMPLING FUNCTION FOR THE MODEL SPACE

For moderate to large p , enumeration of the model space is impossible in practice. In this section we discuss a stochastic search algorithm based on importance sampling for the model space. In summary, exploiting the orthogonalization, the importance sampler draws elements of the orthogonal basis independently, with probability that approximates very closely the actual posterior model probability. Importance sampling results in three main advantages over conventional Markov chains for stochastic search:

a) speed: it is faster to sample each model, as one QR decomposition is necessary overall for the entire sampler. In the original model space, sampling a new model requires updating the regression at each step of the chain;

b) convergence: we are interested in convergence of the sample-based predictive distribution to the exact predictive distribution based on enumeration in the space of models. This typically occurs earlier with orthogonalized model mixing, especially if the original model space is difficult to traverse because of high correlations which may induce slow mixing of MCMC methods;

c) diagnostics: the importance sampling probabilities are available exactly (rather than up to a normalizing constant); therefore at any point in the sampler one can compute an estimate of the total mass sampled by adding the importance sampling probabilities for the sampled models. This can be used for inference and possibly for convergence diagnostics.

Importance sampling by drawing elements of the orthogonal basis independently requires an approximate product-form representation for the q_γ in (4). This can be obtained by expressing (5) as a linear function of γ . We achieve this via a Taylor series expansion of the last term in (5). In particular, expanding around $\nu\xi + \tilde{Y}'\tilde{Y}$, we have

$$\log\left(\nu\xi + \tilde{Y}'\tilde{Y} - \sum_{i=1}^p \gamma_i \text{SSR}_i^2\right) = \log(\nu\xi + \tilde{Y}'\tilde{Y}) - \sum_{j=1}^k \frac{1}{j} \left[\frac{\sum_{i=1}^p \gamma_i \text{SSR}_i^{2j}}{(\nu\xi + \tilde{Y}'\tilde{Y})^j} \right] + A_k \quad (7)$$

where A_k is a remainder term. Ignoring the cross product terms after expanding the expression in square brackets,

$$\log\left(\nu\xi + \tilde{Y}'\tilde{Y} - \sum_{i=1}^p \gamma_i \text{SSR}_i^2\right) = \log(\nu\xi + \tilde{Y}'\tilde{Y}) - \sum_{j=1}^k \frac{\sum_{i=1}^p \gamma_i \text{SSR}_i^{2j}}{j(\nu\xi + \tilde{Y}'\tilde{Y})^j} + A_k.$$

Ignoring the remainder term, this expression would lead to a factorization of q_γ permitting independent importance sampling on the columns of Z . However, while this approximation would work well with models with low dimension (it is exact for the model including only the intercept), it worsens as the dimensionality increases, inducing an undesirable systematic bias. One simple alternative is to calibrate the expansion by making it exact for the full model as well. This can be done as follows. Let SSR_1^2 be the regression sum of squares for the regression of \tilde{Y} on the intercept only. Also, let

$$\begin{aligned} L_N &= \log(\nu\xi + \tilde{Y}'\tilde{Y} - \text{SSR}_1^2) \\ L_F &= \log\left(\nu\xi + \tilde{Y}'\tilde{Y} - \sum_{i=1}^p \text{SSR}_i^2\right) \\ \tilde{L}_N &= \sum_{j=1}^k \frac{\text{SSR}_1^{2j}}{j(\nu\xi + \tilde{Y}'\tilde{Y})^j} \\ \tilde{L}_F &= \sum_{j=1}^k \frac{\sum_{i=1}^p \text{SSR}_i^{2j}}{j(\nu\xi + \tilde{Y}'\tilde{Y})^j}. \end{aligned}$$

The calibrated factorizable expansion is then

$$\log \left(\nu\xi + \tilde{Y}'\tilde{Y} - \sum_{i=1}^p \gamma_i \text{SSR}_i^2 \right) \approx \frac{L_N - L_F}{\tilde{L}_N - \tilde{L}_F} \left[\log(\nu\xi + \tilde{Y}'\tilde{Y}) - \sum_{j=1}^k \frac{\sum_{i=1}^p \gamma_i \text{SSR}_i^{2j}}{j(\nu\xi + \tilde{Y}'\tilde{Y})^j} \right] \quad (8)$$

Replacing (8) into (5) we obtain:

$$\log(q_\gamma) \approx Q + \sum_{i=1}^p \gamma_i \left[\log \left(\frac{\theta_i}{1 - \theta_i} \right) - \frac{1}{2} \log d_i + \frac{(n + \nu) L_N - L_F}{2 \tilde{L}_N - \tilde{L}_F} \sum_{j=1}^k \frac{\text{SSR}_i^{2j}}{j(\nu\xi + \tilde{Y}'\tilde{Y})^j} \right] \quad (9)$$

where Q is a quantity that does not depend on γ . Equation (9) defines the approximate posterior distribution up to a proportionality constant. Exploiting the product form of the approximate posterior distribution, it is straightforward to work out the normalizing constant. This leads to the following approximate posterior model probability:

$$\tilde{\pi}(\gamma|Y) = \prod_{i=1}^p p_i^{\gamma_i} (1 - p_i)^{1 - \gamma_i} \quad (10)$$

where:

$$p_i = \frac{\theta_i \exp \left\{ -\frac{1}{2} \log d_i + \frac{L_N - L_F}{\tilde{L}_N - \tilde{L}_F} \sum_{j=1}^k \frac{(n + \nu) \text{SSR}_i^{2j}}{2j(\nu\xi + \tilde{Y}'\tilde{Y})^j} \right\}}{1 - \theta_i + \theta_i \exp \left\{ -\frac{1}{2} \log d_i + \frac{L_N - L_F}{\tilde{L}_N - \tilde{L}_F} \sum_{j=1}^k \frac{(n + \nu) \text{SSR}_i^{2j}}{2j(\nu\xi + \tilde{Y}'\tilde{Y})^j} \right\}}$$

Generating a sample of models from $\tilde{\pi}$ is straightforward. It can be done independently on each of the elements of the orthogonal basis by generating Bernoulli random variables and does not require Markov chain Monte Carlo methods. After the initial computation of the orthogonal basis and the SSR_i 's, sampling is done directly from $\tilde{\pi}$ and is very fast.

The sampling efficiency of the importance sampler will increase with the variability of the $\tilde{\pi}_i$'s. The easiest model space to sample is one where all the $\tilde{\pi}_i$'s are 0 except for one. On the other hand, if $\tilde{\pi}_i \approx .5$, little is gained by sampling from $\tilde{\pi}$. The orthogonalization strategy can therefore be important in determining the efficiency of the sampler. Orthog-

orthogonalizations that identify interesting target subspaces will lead to a mix of columns of Z with large $\tilde{\pi}_i$ and small $\tilde{\pi}_i$. Thus, summary measures of the variability of the $\tilde{\pi}_i$'s provide an indirect way of assessing the efficiency of the orthogonalization strategy adopted.

Also, when the prediction problem can be cast in terms of a one dimensional quantity of interest, such as the predicted Y at some specified predictor vector x_0 , one may be interested in searching for models that contribute highly to this particular prediction. This is related to, but not identical to, having a high posterior probability. The importance sampling function could be tailored more specifically to this situation.

In developing the importance sampling probabilities, σ was integrated out to obtain $P(\gamma|Y)$. In the orthogonal variable model space, an alternative can be developed by sampling from $P(\gamma|Y, \sigma)$, which factors exactly into the product of p independent Bernoulli random variables. The full conditional distributions for the Gibbs sampler for σ and γ are easy to generate from, and permit sampling the whole vector γ jointly, as is the case in the importance sampler. This strategy can also be expected to provide rapid mixing in the model space.

4. ANALYSIS OF SIMULATION OUTPUT

After N draws from $\tilde{\pi}$, there are m discovered models available for analysis. Call this set D . For each model γ in D , we have available the unnormalized posterior probability q_γ , and often the quantities ϕ_γ relevant to prediction. We can also keep count of the frequency f_γ of model γ in the N draws. Based on this information we want to estimate

$$\phi = \frac{\sum_{\gamma \in D} q_\gamma \phi_\gamma + \sum_{\gamma \in \bar{D}} q_\gamma \phi_\gamma}{\sum_{\gamma \in D} q_\gamma + \sum_{\gamma \in \bar{D}} q_\gamma} \quad (11)$$

which depends on the undiscovered models.

The set D can be thought of as a sample without replacement from a finite population, with sampling proportional to the size of $\tilde{\pi}$. Methods for analyzing similar data are discussed in the literature (see for example West, 1994) and can be used to derive posterior distributions for ϕ based on the discovered models. These typically require additional simulation to make inference about ϕ . In this context we seek approaches that require

a minimal amount of computation, as computing time can be more efficiently used to obtain a larger set D .

For convenience of exposition, we focus on the predictive mean vector at the observed design matrix, so that $\phi_\gamma = \hat{Y}_\gamma$. In particular, we consider estimators of ϕ of the form

$$\hat{\phi} = \sum_{\gamma \in D} w_\gamma \hat{Y}_\gamma \tag{12}$$

where w_γ are normalized weights. The choices that we consider for empirical comparison in Section 5 are as follows:

1. Monte Carlo estimator. The weight w_γ is the relative frequency f_γ/N of model γ in the N draws. This approach is appropriate for Markov chain output when q_γ are not available (Geweke, 1994, Carlin and Chib, 1995). In our formulation, using a simple Monte-Carlo average ignores the information contained in q_γ and in the sampling mechanism. As a result, one can construct more efficient estimators.

2. Window estimator. A simple but effective alternative is renormalization of the un-normalized posterior probabilities within the set D . Formally:

$$w_\gamma = \frac{q_\gamma}{\sum_{\gamma' \in D} q_{\gamma'}}.$$

This has precedents in Madigan and Raftery (1994), for example in Occam's Window.

3. Importance Sampling estimator. A further alternative is to adopt the standard importance sampling weights. Then the weight of model γ is given by

$$w_\gamma = \frac{f_\gamma q_\gamma / \tilde{\pi}(\gamma)}{\sum_{\gamma' \in D} f_{\gamma'} q_{\gamma'} / \tilde{\pi}(\gamma')}.$$

This choice can be troublesome when the importance sampling function works poorly. In particular, when a $\tilde{\pi}(\gamma)$ is very small, but the corresponding $\pi(\gamma)$ is large, the resulting weight will be close to one, so that this model dominates in the mixture estimator. Descriptive summaries of the weights can indicate when this is a problem.

5. CRIME DATA

5.1. *Background*

The crime data of Vandaele (1978) is commonly used as a test case in variable selection problems (see also Raftery, Madigan and Hoeting, 1993). There are 15 candidate predictors, leading to 32,768 models. Enumeration of all models is available for comparison and convergence checks. As in Raftery, Madigan, and Hoeting (1994), we used the natural logs of all continuous variables in the analysis. We are interested in the following: a) comparison of alternative algorithms for stochastic search of orthogonalized model spaces, b) comparison of alternative rules for estimating the predictive mean based on the sampled models, and c) comparison of stochastic search algorithms in the original and orthogonalized space in terms of efficiency in estimating the respective predictive means for the two spaces.

The prior on the β coefficients is proper but dispersed. The matrix C is diagonal; we chose the elements c_{ii} as follows. We first selected a large interval of size δ in the scale of the response variable. Then for each of the predictors, the prior mean was set to 0, as in George and McCulloch (1993), and the variance chosen by assuming that $E(\sigma)c_{ii} = \delta/\text{IQR}(X_i)$ is 3 standard deviations away from 0. For the one dummy variable, we used the range instead of the interquartile range. In this way the 3-SD interval on the marginal distribution of each coefficient includes “large” values of the coefficient. By this we mean values that would be sufficient to explain completely the variation of the response, based on one unit of typical variation in X_i . This procedure for prior elicitation of c_{ii} ’s is appealing to us for two reasons: first it is done entirely in terms of observables (the response); second it reduces the elicitation of p quantities to just one, handling all candidate predictors homogeneously. The prior hyperparameters for the distribution on σ^2 are $\nu = 3$ and $\xi = .5/3$. The degrees of freedom ν are kept small to reflect lack of information, while having a distribution with a finite mean and variance. The value of ξ was selected based on the anticipated range of the response, and was designed to allow for options ranging from good to very poor fit. The same prior distribution for β and σ^2 was used in the original and orthogonalized model space. In both cases we used the uniform prior on model spaces, with the intercept being included with probability one.

5.2. *Algorithms*

We considered five stochastic search schemes in the orthogonalized space. The first two are based on importance sampling. The remaining three are Markov chains. Other choices of Markov chain could be constructed: we used some that are successful based on the current literature. The five schemes are:

1. Importance sampling as described in Section 3.
2. Random Sampling. Columns of the orthogonal basis are included with probability .5, independently of each other. In our setting, this is equivalent to using the prior distribution as an importance function for the model space.
3. SSVS. This is based on George and McCulloch (1994). In this implementation of SSVS, the chain moves from a current model γ to the next model γ' by first selecting a random permutation of the variables. Then, for variable j

$$P(\gamma'_j = 1 - \gamma_j | \gamma_{(j)}) = \frac{q_{\gamma_{[j]}}}{q_{\gamma_{[j]}} + q_\gamma}$$

where $\gamma_{(j)}$ is γ with the j -th element deleted, and $\gamma_{[j]}$ is γ with γ_j replaced by $1 - \gamma_j$. Repeating for all other coordinates in the randomly selected order gives γ' .

4. MC³. This is based on Madigan and York (1995). The chain moves from a current model γ to the next model γ' by selecting one coordinate j at random and updating based on:

$$P(\gamma'_j = 1 - \gamma_j | \gamma_{(j)}) = \min\left(1, \frac{q_{\gamma_{[j]}}}{q_\gamma}\right).$$

5. Hybrid. This combines elements of the two previous algorithms. The chain moves from a current model γ to the next model γ' by first selecting a random permutation of the variables. Then, for variable j , updating is done as in MC³, as opposed to the Gibbs update in SSVS. The motivation for considering a hybrid algorithm of integrated SSVS and MC³ is the following. In the MC³ the transition probability of moving to the new model, γ' is $\min(1, q_{\gamma'}/q_\gamma)$ which is always greater than or equal to the transition probabilities of the SSVS algorithm. Using this transition probability in the SSVS algorithm might result in better mixing over the space of models. The MC³ algorithm selects a random coordinate to change at each step while the SSVS approach goes through all p coordinates in either a deterministic or random order. The probability that a coordinate

is visited in p steps for the MC³ algorithm is

$$1 - \left(\frac{p-1}{p}\right)^p$$

which in the limit as p goes to infinity is $1 - e^{-1}$. This means that some coordinates are changed several times within the p steps. This may not be the most efficient way to cover the model space, and going through all p coordinates may result in better mixing.

The importance sampler defined here is limited to orthogonalized spaces. Random sampling and the three MCMC algorithms provide a method for sampling from nonorthogonal situations as well.

5.3. Results

All computations in this section were done on a DEC workstation AXP3000/400 and programmed in XLISP-STAT. Enumeration was done both in the original and orthogonalized model space. Enumeration times were 78 and 13 minutes respectively. This comparison depends on a number of specific factors, but we expect a similarly substantial advantage to apply to most problems. Orthogonalized model mixing provided a closer fit to the observed data with a MSE of .0313 versus .0432 for model mixing in the original space. This comparison depends crucially on the data set, and possibly on the specific orthogonal basis and cannot be generalized.

a) Comparison of stochastic search algorithms in the orthogonalized space.

The 5 algorithms of Section 5.2 are compared in Figures 1 and 2. In Figure 1, we compare efficiency in discovering models with high posterior probability by considering the distribution of the logarithms of the probabilities of the models discovered by the various algorithms. The number of models sampled is 30,000. The population distribution, given at the top left, is bimodal, and can be roughly divided into good and bad models. Importance sampling is the only method that focuses exclusively on the good models. The figures on total probability mass discovered, added at the bottom of each histogram, emphasize that importance sampling centers in on the models with higher posterior probability. After 30,000 iteration all algorithms have similar predictive accuracy, with a small

advantage in favor of importance sampling. The total number of models is 32,768 and we found that the enumeration time is similar to the running time for importance sampling, both being much smaller than the time necessary to run the Markov chains.

With large p , the number of runs that one can afford is a small fraction of the total number of models. To reproduce such situations we analyzed runs of $N = 300$ iterations, replicated 100 times to obtain distributions. To bypass burn-in time in the Markov chain algorithms, we used random draws from $\hat{\pi}(\gamma|Y)$, an approximation of the ergodic distribution of the chain, as a starting point. We compared algorithms based on: integrated squared error on the predictive means, Kullback-Leibler divergence of the predictive distributions, and total mass sampled.

Let μ denote the exact predictive mean vector at the design matrix X . By exact, we mean obtained by complete enumeration of the model space as in calculation of ϕ in equation (6). The conditional posterior predictive mean is $\mu_\gamma = \hat{Y}_\gamma$. Let $\hat{\mu}^A$ be the approximation based on a stochastic search sample, with the superscript A indexing the algorithms. We use the window weights in equation (12) to calculate $\hat{\mu}^A$ for all search methods. Then the integrated squared error is: $\text{ISE}^A = \sum_{j=1}^n (\hat{\mu}_j^A - \mu_j)^2/n$.

The ISE comparison depends only on the means of the exact and approximate predictive distributions. To assess the quality of the approximation of the whole distribution, we computed the Kullback-Leibler divergence between the exact distribution and the approximations based on stochastic search. For simplicity we focused on observation 6, chosen because of the variation in the model specific predictive distributions. If x_6 is the row vector of X corresponding to observation 6, and $p(y|Y, x_6)$ and $p^A(y|Y, x_6)$ are the exact and approximate predictive distributions at x_6 , then the appropriate divergence is:

$$\int \log \left(\frac{p(y|Y, x_6)}{p^A(y|Y, x_6)} \right) p(y|Y, x_6) dy,$$

which was evaluated based on a trapezoidal rule. As in the ISE comparison, the window weights were used to weight the model specific predictive distributions in calculating $p^A(y|Y, x_6)$,

$$p^A(y|Y, x_6) = \sum_{\gamma \in D} w_\gamma p^A(y|Y, x_6, \gamma).$$

The resulting comparisons using the ISE and the Kullback–Leibler divergence are shown in Figure 2. This also includes boxplots of the total posterior mass sampled under the different sampling methods. The orthogonalized importance sampling algorithm outperforms the other alternatives. At least in orthogonalized spaces, random sampling leads to an algorithm that performs similarly to Markov chains. Results from importance sampling with smaller sample sizes have been also included. These underscore the advantages of the fast convergence granted by the independent importance sampling scheme. Both the ISE and the Kullback–Leibler divergence are smaller with 9 iterations of importance sampling than with 300 Markov chain iterations. Similar results for the Kullback–Leibler divergence were obtained using other design points.

One technical point in comparing the Markov chain algorithms is that the definition of one iteration for MC³ is different from that for SSVS and the Hybrid. We opted for keeping the same number of model transitions over all MCMC methods. In this case, this is achieved by running MC³ $p = 15$ times longer than the alternatives, and taking every 15-th model.

b) Comparison of alternative rules for estimating the predictive mean.

Figure 3 shows a comparison of the estimators of the predictive mean discussed in Section 4. As in comparison a), we use the ISE to compare the estimator to the exact predictive mean under enumeration of the orthogonalized model space. Estimators are based on the same samples of models generated from the importance sampler. The window estimator and the importance sampling estimator appear to perform better than the Monte Carlo average.

c) Comparison of algorithms in the original and orthogonalized space.

Finally, Figure 4 compares 1) orthogonalized model mixing using importance sampling, 2) orthogonalized model mixing using a hybrid Markov chain, and 3) standard model mixing using a hybrid Markov chain in the original space. Again, the goal of the comparison is accuracy in recovering the exact predictive distribution, obtained by enumeration in the respective model spaces. As orthogonalized model mixing and standard model mixing generate a different set of exact predictive means, each of the integrated squared errors is computed with respect to the respective exact predictive mean. Boxplots are based on 100 replications of samples of 300 models. In this example, orthogonalized model mixing

shows the best performance. Also, the MCMC algorithm displays a better performance when it is applied to the original variables than when it is applied to the orthogonalized variables. However, it should be noted that importance sampling and MCMC applied to the orthogonalized variables are substantially faster than MCMC in the original variables, so that one can typically afford a larger Monte Carlo sample. It would be inappropriate to conclude from this comparison that orthogonalized model mixing gives a better representation of the mean response compared to model mixing in the original variable. The appropriate comparison to evaluate fit is that based on enumeration mentioned earlier in this section.

6. PROTEIN CONSTRUCT DATA

The next application is to a data set with a large model space, and is included to illustrate the feasibility of orthogonalized model mixing in complex problems. The goal of the experiment, designed and performed by Glaxo Research Institute, is to determine optimal conditions for storing proteins while maintaining a high level of protein activity. There are many factors affecting storage conditions and eight variables were identified as having a potentially important impact on storage conditions. A complete factorial experiment would have required 18,144 runs. The 96 runs actually performed had been selected based on a space filling design algorithm, as implemented by the SAS procedure OPTEX. One of the design goals had been to achieve identifiability of main effects and two-way interactions. The 96 storage conditions were reproduced in the laboratory, and an aliquot of purified protein was added to each of the storage conditions. Protein activity was determined after 4 weeks using an enzymatic assay. Further details can be found in Menius et al. (1994).

Coding the categorical variables as indicator variables, and including interaction terms and second order terms for the continuous independent variables, the total number of candidate predictors is $p = 88$. The resulting data set is challenging and seems to defy most of the standard model selection techniques. Clyde and Parmigiani (1994) discuss Markov chain methods. Here we apply the orthogonalized model mixing approach to address one of Glaxo's main questions: the selection of optimal storage conditions. We answer by building a predictive distribution with orthogonalized model mixing and using it to evaluate the probability that each one of a set of candidate storage conditions is optimal. The general strategy is similar to Higdon (1994).

The prior on γ is $\theta_i = .8$. This reflects the prior knowledge that the variables chosen for the experiments were considered important by chemists, so that the response curve in the orthogonalized space is likely *a priori* to require a high number of terms. A standard ANOVA analysis would treat the dummy variables as grouped variables, and have all variables that represent a factor enter the model together. We are interested in avoiding sensitivity to variable selection in prediction and in the ranking of settings, and we do not need to impose strong parsimony via the prior on model space.

One design point was replicated. This gave an estimate of pure error of approximately 0.01. The hyper-parameters for the prior distribution of σ^2 were taken as $\nu = 50$ and $\xi = .01$ so that the mean was roughly the same as the pure error estimate and the degrees of freedom gave a reasonable range of values for σ^2 , based on the chemists opinion. The elicitation of the prior on the regression coefficients proceeded along the same lines of Section 4. Since we dealt with a designed experiment, we used the range of the design variables, rather the interquartile range.

Our goal is to determine predictive means, predictive probability intervals and probabilities of yielding the highest protein activity for each of the settings. We sampled 23000 models using the importance sampler. The predictive distribution of the future response vector Y^* at the $n^* \times p$ design matrix X^* can be estimated from the sample of models as

$$\hat{p}(Y^*|Y) = \sum_{\gamma \in D} w_{\gamma} p(Y^*|Y, \gamma),$$

where w_{γ} are the window weights.

In particular, for a given model γ , let \tilde{Z}_{γ} be the matrix obtained by selecting the columns of \tilde{Z} that correspond to a 1 in the vector γ . Also, let $\tilde{\alpha}_{\gamma}$ be the vector obtained by selecting the elements of $\tilde{\alpha}$ that correspond to a 1 in the vector γ . From orthogonality, $\tilde{\alpha}_{\gamma} = (\tilde{Z}'_{\gamma} \tilde{Z}_{\gamma})^{-1} \tilde{Z}'_{\gamma} \tilde{Y}$. Let Z^* denote the design matrix in the transformed space, $Z^* = X^*CU$ and Z^*_{γ} denote the matrix obtained by selecting the columns of Z^* that correspond to a 1 in the vector γ . Since we are interested in evaluating the settings used in the original experiment, we will take $X^* = X$ and $n^* = n$.

The random vector $Y^*|Y, \gamma$ has a n^* -dimensional multivariate t distribution, that is

$$p(y^*|Y, Z^*, \gamma) \propto \left[(n + \nu) + (y^* - Z_\gamma^* \hat{\alpha}_\gamma)^T \frac{(I + Z_\gamma^* (\tilde{Z}_\gamma^T \tilde{Z}_\gamma)^{-1} Z_\gamma^{*T})^{-1}}{(\xi \nu + S_\gamma^2)/(n + \nu)} (y^* - Z_\gamma^* \hat{\alpha}_\gamma) \right]^{-\frac{(n^* + n + \nu)}{2}} \quad (13)$$

with $S_\gamma^2 = \tilde{Y}'\tilde{Y} - \sum_{i=1}^p \gamma_i \text{SSR}_i^2$. Evaluation of the desired probabilities can proceed by point-wise evaluation of \hat{p} followed by numerical integration, or by a new simulation based on resampling models according to their weight w_γ and then generating an observed vector from (13). The first approach is preferable if marginal probabilities are of interest. The second, adopted here, is better suited to handle higher dimensional integrals, such as the probabilities that each setting will have the highest protein activity.

Figure 5 shows the predictions based on the mixture of models, together with centered 95% posterior probability regions, obtained by simulation. It also illustrates the sensitivity of the predicted values to the choice of the model. Figure 6 gives the estimated probability that each of the settings used in the present experiment generates the maximum protein activity level. A similar technique can be applied to extrapolate for other potentially interesting settings based on (13).

We performed various diagnostic checks. In particular, we monitored the residual vector and the ISE of the predictions based on model mixing, the total mass sampled and the relationship between the $\tilde{\pi}(\gamma)$ and the q_γ 's. These stabilize satisfactorily. Interestingly, the predictions based on model mixing stabilize earlier compared to the total mass of model space actually sampled, which we found applies in many other examples.

7. DISCUSSION

Bayesian model mixing offers a fruitful theoretical framework for making predictions that account for uncertainty in the selection of predictor variables. In this paper we introduce an approach and algorithms for implementing model mixing in large prediction problems with correlated predictors. Our approach is based on expressing the space of models in terms of an orthogonalization of the design matrix. Two key elements of this approach are: a) the orthogonalization method and b) the prior probability distributions assigned to both the models and the coefficients. In Clyde and Parmigiani (1995) we

look at the predictive distributions resulting from model mixing under alternative orthogonalizations. The example we consider is simulated, but interesting and realistic. In particular, the true model used for generating the data does not belong to the model space, and there is a wide range of correlations among the original variables. The predictive distributions are quite close, and remain close under a range of reasonable priors and simulated data sets.

However, both the choice of prior and orthogonal basis can affect the predictive distribution substantially. Experimental cases indicate that results seem to be more sensitive to the choice of prior parameters, than to the choice of basis. In particular, a key choice is that of the prior distribution on σ , because a tight control on the amount of noise in the model results in a control over the parsimony of the curve used. Also, the effects of the prior and the orthogonalization can be very strongly related, as might be expected. If the prior distribution on β is fixed, say based on shrinkage considerations, and the priors on both model spaces are uniform, different orthogonalizations can lead to widely different amounts of shrinkage of the predictions. One example arises when the “true” model is a subset of the original variables. The fitted values can be recovered for most orthogonalizations under the full model. However, if the prior on the model space is assigned to favor parsimony in terms of the number of predictors, then orthogonalization can lead to a worse fit.

For these reasons, orthogonalized model mixing cannot be recommended as a black-box prediction method. However, advantages of careful implementations are both statistical and computational. Orthogonalization leads to a better behaved problem, as the number of competing models is reduced by eliminating correlations. Also, compared to Markov chains methods, orthogonalized model mixing by importance sampling is faster in sampling models, and is also more efficient in finding models that contribute significantly to the prediction. Further advantages over standard Markov chain methods are related to the speed of convergence and the availability of more reliable convergence diagnostic tools. We illustrated these points using the crime data. We also demonstrated the feasibility of orthogonalized model mixing in a large problem which is very difficult to attack by other methods.

REFERENCES

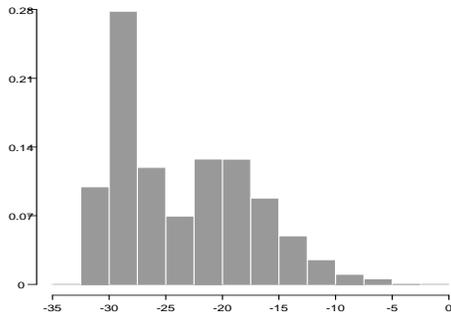
Bennett, J. and Wakefield, J. (1994). Covariate modelling in population pharmacokinetics

- models. TR-94-19, Department of Mathematics, Imperial College, U.K.
- Carlin, B.P. and Chib, S. (1995). Bayesian Model Choice via Markov chain Monte Carlo. *Journal of Royal Statistical Society - Series B*, 57, 473–484.
- Chipman, H. (1996). Bayesian Variable Selection with Related Predictors. *Canadian Journal of Statistics*, to appear
- Clyde, M.A. and Parmigiani, G. (1994). Protein Construct Storage: Bayesian Variable Selection and Prediction With Mixtures. ISDS DP 94-14, Duke University.
- Clyde, M.A. and Parmigiani, G. (1995). Orthogonalizations and Priors for Orthogonalized Model Mixing. ISDS DP 95-07, Duke University.
- de Finetti, B. (1937). La Prévision, ses lois logiques, ses sources subjectives. *Annales de l'Institut Poincaré* VIII-1, pp. 1–68.
- Draper, D. (1994). Assessment and propagation of model uncertainty (with Discussion). *Journal of the Royal Statistical Society* 56.
- Garthwaite, P.H. and Dickey, J.M. (1992). Elicitation of prior distributions for variable-selection problems in regression. *Ann. of Statistics* 20, pp. 1697-1719.
- George, E.I. (1986a). Minimax multiple shrinkage estimation. *Annals of Statistics* 14, pp. 188-205.
- George, E.I. (1986b). Combining minimax shrinkage estimators. *Journal of the American Statistical Association* 81, pp. 437-445.
- George, E.I. and McCulloch, R. (1993). Variable Selection via Gibbs Sampling. *Journal of the American Statistical Association*, 88, pp. 881–889.
- George, E.I. and McCulloch, R. (1994). Fast Bayes Variable Selection. Graduate School of Business, University of Chicago.
- George, E.I., McCulloch, R., and Tsay, R. (1994). Two Approaches to Bayesian Model Selection with Applications. In *Bayesian Statistics and Econometrics: Essays in Honor of A. Zellner*, ed. Berry D.A., Chaloner K.M., Geweke J.F, New York.
- Geweke, J.F. (1994). Bayesian comparison of econometric models. Working Paper 532, Federal Reserve Bank of Minneapolis.
- Green, P.J. (1995) Reversible Jump MCMC Computation and Bayesian Model Determination. Technical Report 94-19, University of Bristol.
- Higdon, D. (1994). Spatial Applications of Markov chain Monte Carlo for Bayesian Inference. PhD Thesis, Department of Statistics, University of Washington, Seattle.

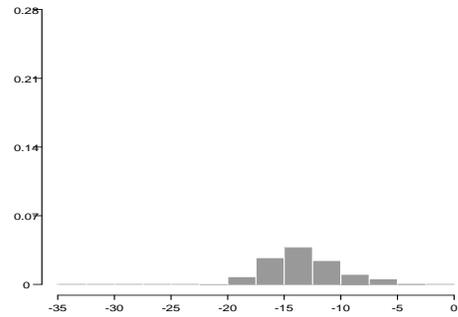
- Leamer, E.E. (1978). *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. Wiley, New York.
- Li, K-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association* 86, pp. 316–327.
- Madigan, D.M. and York, J. (1995). Bayesian Graphical Models for Discrete Data. *International Statistical Review*, 63, 215-232.
- Madigan, D.M. and Raftery, A.E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam’s window. *Journal of the American Statistical Association*. 89, 1535–1546.
- Madigan, D.M., Gavrin, J., and Raftery, A.E. (1994). Eliciting prior information to enhance the predictive performance of Bayesian graphical models. *Communications in Statistics - Theory and Methods*, 24, 2271-2292.
- Menius, A.J, Rocque, W., Emptage, M.R. and Young, S.S. (1994). Space Filling Experimental Design for Determining Protein Construct Storage Conditions. *Proceedings of the 26th Symposium on the Interface*, 106–110.
- Mitchell, T.J. and Beauchamp, J.J. (1988). Bayesian Variable Selection in Linear Regression. *Journal of the American Statistical Association* 83 , pp. 1023–1036.
- Phillips, D.B. and Smith, A.F.M. (1994). Bayesian Model Comparison via Jump Diffusions. TR-94-20, Department of Mathematics, Imperial College, U.K.
- Raftery, A.E., Madigan, D.M., and Hoeting, J. (1993). Model selection and accounting for model uncertainty in linear regression models. TR 262, Department of Statistics, University of Washington.
- Raftery, A.E., Madigan, D.M., and Volinsky C.T. (1995). Accounting for Model Uncertainty in Survival Analysis Improves Predictive Performance (with discussion). In *Bayesian Statistics 5*, ed. J.M. Bernardo, J.O. Berger, A.P. Dawid and Smith, A.F.M.
- Rao, C.R. (1964). The Use and Interpretation of Principal Components in Applied Research. *Sankhya A* 26, pp. 329–358.
- Vandaele, W. (1978). Participation in Illegitimate Activities: Ehrlich Revisited. In *Deviance and Incapacitation*, ed. Blumstein, A. Cohen, J. and Nagin, D., pp. 270–335, Washington, DC.
- Weisberg, S. (1985). *Applied Linear Regression. 2nd Edition*. Wiley, New York.
- West, M. (1994). Discovery Sampling and Selection Models. In *Decision Theory and*

Related Topics, ed. J.O. Berger, S.S. Gupta, New York. 221-235.

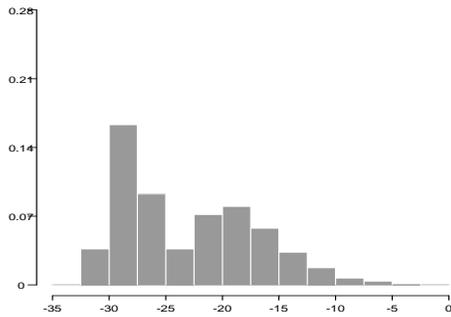
Wold, S., Ruhe, A., Wold, H., and Dunn, W.J.III (1984). The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing* 5, pp. 735-743.



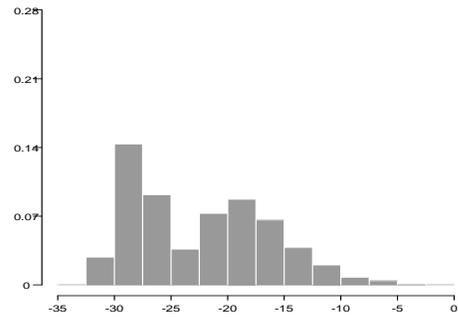
All Models
 $\sum \pi(\gamma|Y) = 1$ for 32768 models



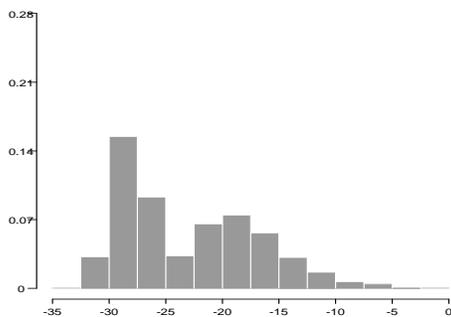
Importance Sampling
 $\sum \pi(\gamma|Y) = 0.9990$ for 3759 models



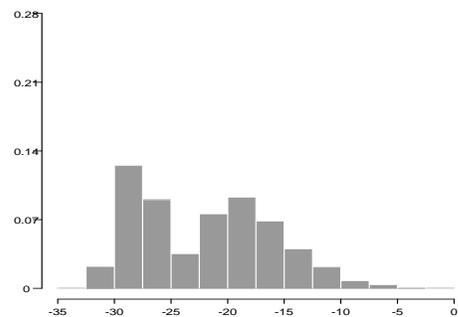
Random Sampling
 $\sum \pi(\gamma|Y) = .63$ for 19661 models



SSVS
 $\sum \pi(\gamma|Y) = .57$ for 19567 models



MC³
 $\sum \pi(\gamma|Y) = .58$ for 18655 models



Hybrid
 $\sum \pi(\gamma|Y) = .52$ for 19171 models

Figure 1: Comparison of algorithms: Distributions of the logarithm of the posterior probabilities of the sampled models based on 30000 iterations of each stochastic algorithm.

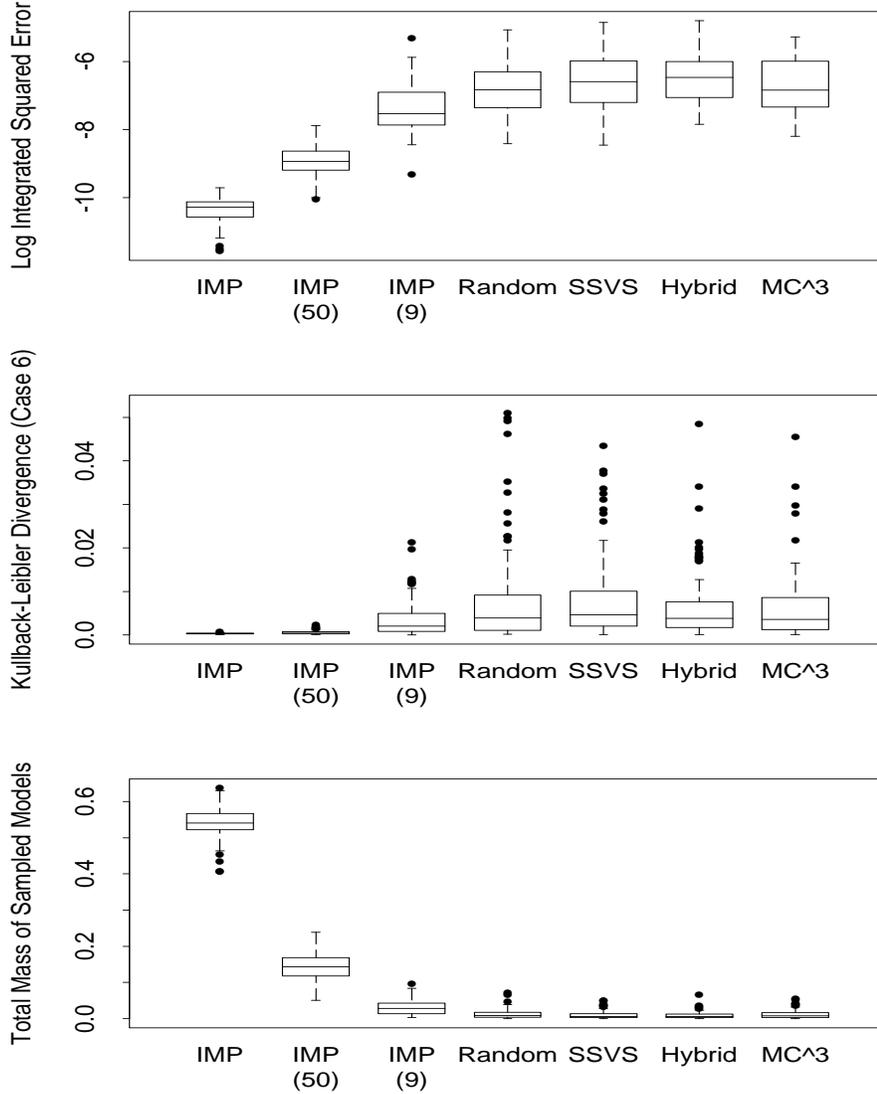


Figure 2: Comparisons of algorithms based on log ISE, Kullback–Leibler Divergence for the predictive distribution for case 6, and total posterior probability of sampled models. Boxplots are based on 100 samples of 300 iterations of each stochastic algorithm. Gains from importance sampling, labeled IMP, are substantial. To underscore this, we have also included results from importance sampling based on 9 iterations, IMP (9), and 50 iterations, IMP (50).

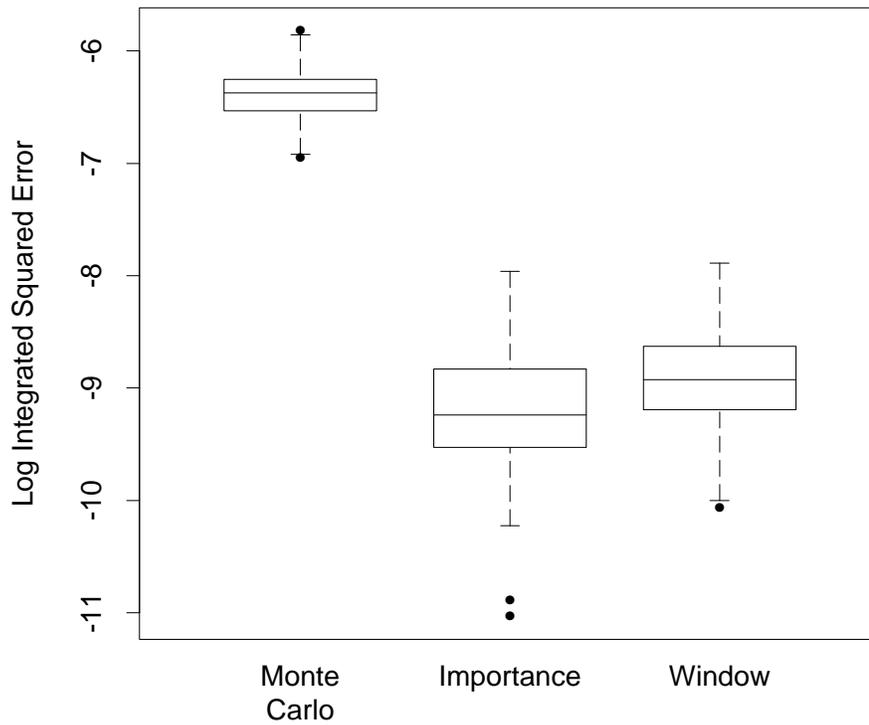


Figure 3: Comparison of log ISE's for alternative estimators of the exact predictive mean vector. Boxplots refer, from left to right, to the Monte Carlo estimator, the importance sampling estimator and the window estimator. Results are based on 100 replications using 50 samples from the importance sampler.

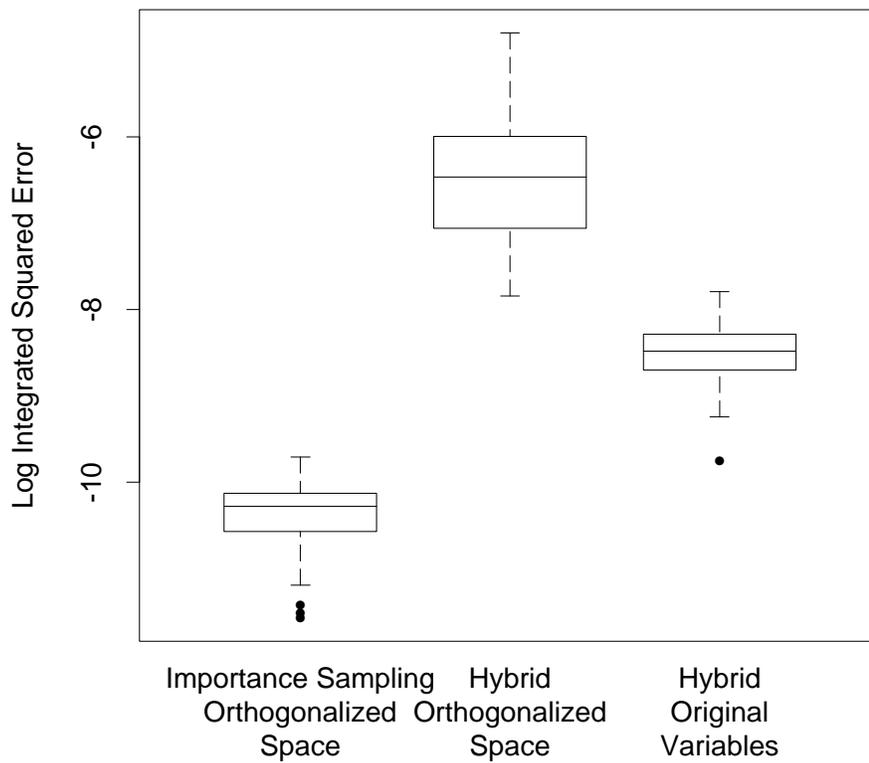


Figure 4: Comparison of log ISE for estimation of the exact predictive mean using orthogonalized model mixing with importance sampling and Markov Chains versus standard model mixing with Markov chains. Boxplots are based on 100 replications of 300 iterations of the stochastic search algorithms. In addition to the better performance displayed in the figure, orthogonalized model mixing is substantially faster.

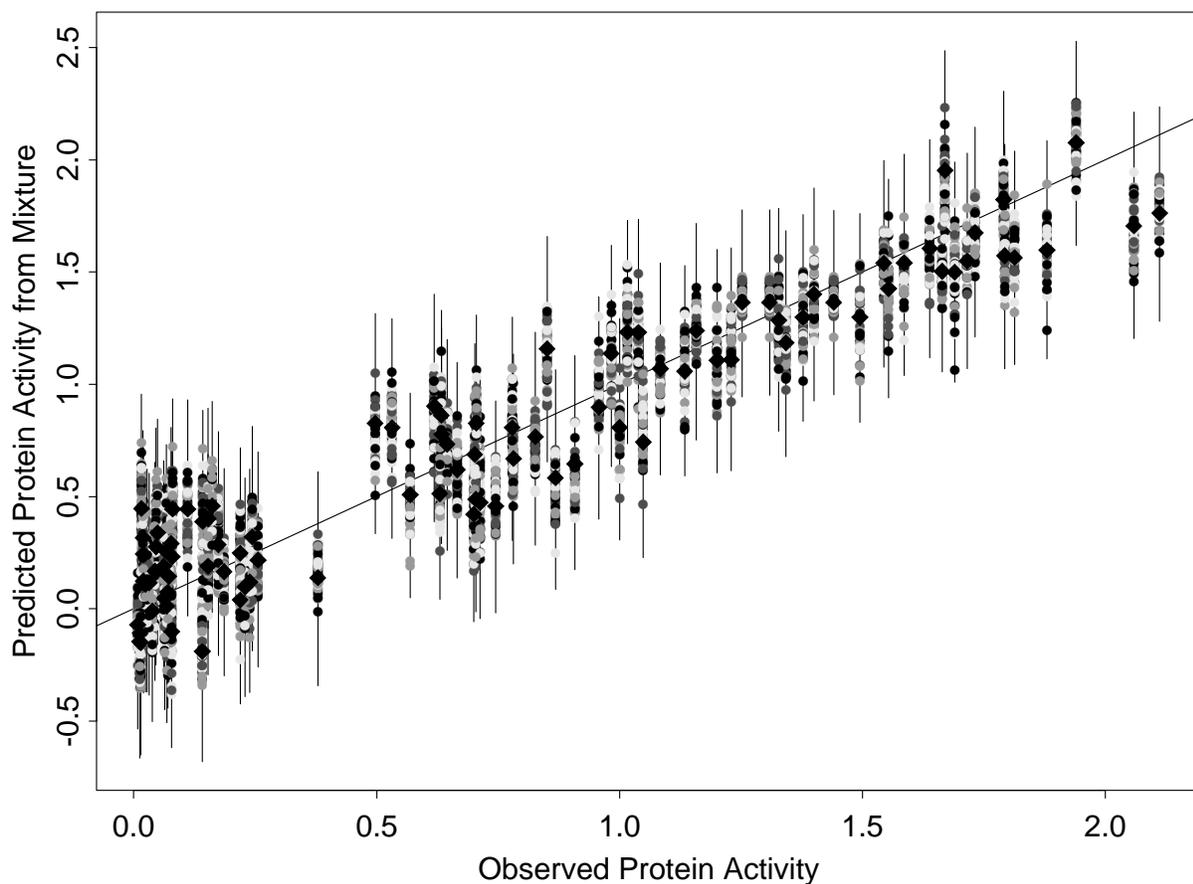


Figure 5: Mixture based predictions for the Protein Construct example. The solid diamonds represent the predictive means from model mixing. The vertical lines correspond to 95% probability intervals from model mixing. Also, stars represent the means of the model-specific predictive distributions for the 50 most probable models discovered. The variability induced by model uncertainty on each individual prediction amounts to a substantial fraction of the variability of the response.

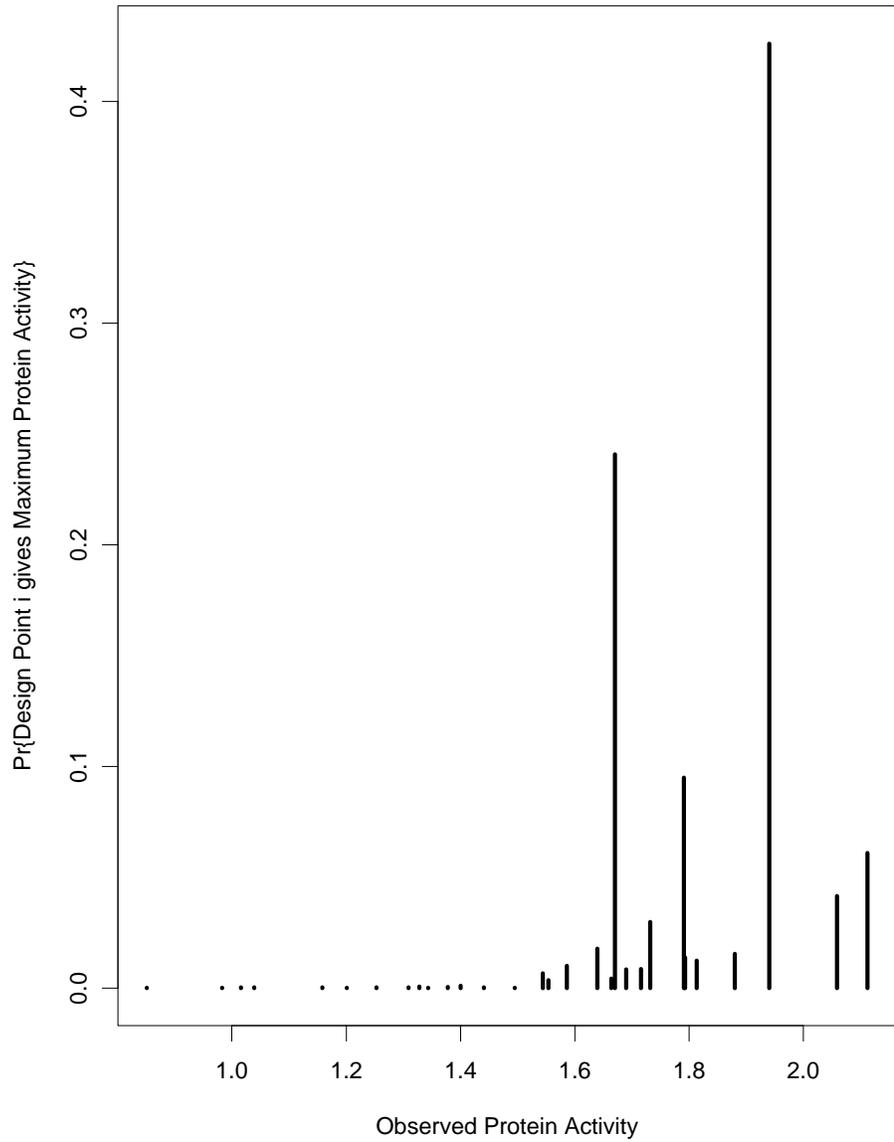


Figure 6: Probability of yielding the maximum for each of the experimental settings, by observed response level.