PREDICTION USING ORTHOGONALIZED MODEL MIXING

by

Heather Denise DeSimone

Institute of Statistics and Decision Sciences
Duke University

Date: _____
Approved:

_____
Dr. Merlise Clyde, Supervisor

_____
Dr. Giovanni Parmigiani, Supervisor

_____
Dr. Donald Berry

_____
Dr. Victor Hasselblad

_____
Dr. Robert Wolpert

Dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Institute of Statistics and Decision Sciences
in the Graduate School of
Duke University

1996

ABSTRACT

(Statistics)

PREDICTION USING ORTHOGONALIZED MODEL MIXING

by

Heather Denise DeSimone

Institute of Statistics and Decision Sciences
Duke University

Date: _____
Approved:

_____
Dr. Merlise Clyde, Supervisor

_____
Dr. Giovanni Parmigiani, Supervisor

_____
Dr. Donald Berry

_____
Dr. Victor Hasselblad

_____
Dr. Robert Wolpert

An abstract of a dissertation submitted in partial
fulfillment of the requirements for the degree
of Doctor of Philosophy in the Institute of
Statistics and Decision Sciences in the Graduate School of
Duke University

1996

Abstract

This dissertation investigates modeling strategies and numerical methods for prediction under model averaging in normal linear models and in Poisson regression with extensions to other generalized linear models. The focus is on mixing over possible subsets of candidate predictors. For linear regression models a sampling approach which uses an importance sampling technique is developed. This technique is based on an approximation of the posterior model probabilities using an orthogonalized transformation of the variables. The posterior probability is approximated by a product of independent Bernoulli random variables, each indicating whether or not an element of the orthogonal basis is included. This leads to an efficient importance sampling algorithm. In extending this to Poisson regression, one difficulty is that we cannot analytically integrate out model specific parameters to obtain posterior model probabilities—a key step in obtaining the probabilities for sampling and model mixing in normal linear regression. Under regularity conditions, applying a variance stabilizing transformation to the response results in an approximately normal distribution with a known constant variance. A Taylor series expansion of the mean function results in a linear model, so in the approximate problem, the previous linear model results can be used to approximate the posterior

model probabilities for the Poisson problem. This allows for sampling directly from an approximation to the joint distribution over the model space.

To evaluate orthogonalized model mixing for normal linear models, it is applied to a set of crime data. The model space is small enough to allow for enumeration of all models for comparison and convergence checks. Furthermore, we demonstrate the feasibility of orthogonalized model mixing in a large size problem (88 variables) which is very difficult to attack by other methods. The large data set originates from an experiment designed to predict protein activity under various storage conditions.

To examine the approach for Poisson regression, orthogonalized model mixing is again applied first to a small data set for which enumeration of all models is available. Through comparison to a Gibbs sampler and a deterministic approach, we find that our method is fast in sampling models and that it supplies good approximations to the posterior model probabilities and predictive distributions. Our method for Poisson regression is then applied to a data set of 126 variables. This large data set was designed to examine the effect of particulate pollution on daily death counts and is difficult to analyze in terms of the original variables.

Acknowledgements

This dissertation would not have been possible without the help and advise of many people. I would especially like to acknowledge the contributions of my advisors, Merlise Clyde and Giovanni Parmigiani, for their countless hours of commitment and patience; the members of my committee, Don Berry, Victor Hasselblad, and Robert Wolpert, for their time and valuable suggestions; the faculty, staff, and students here at ISDS for creating a wonderful working environment; my family, for their love and support; and, most importantly, my husband Maciek Sasinowski, not only for his love and support, but also for his valuable help and dedication.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Regression explores the relationship between variables; specifically, between a response variable and some explanatory variables, or predictors, of the response. Selecting a subset of these predictor variables that "best" describes the response variable is called variable selection, and the subset is referred to as a model. Typically, statisticians use model selection techniques to choose a single model which is used to conduct their analyses. However, treating any selected model as the true model (the model that truly represents reality) ignores the uncertainty associated with that particular model. In other words, the analyses do not account for the fact that the chosen model may not be the true model. Ignoring this uncertainty, called model uncertainty, can often lead to unsatisfactory results such as underestimating prediction intervals. One alternative to single model selection which offers an effective and conceptually appealing treatment of model uncertainty is Bayesian model mixing. This is a technique that makes predictions based on averaging over all models where each model contributes to the prediction proportionally to the

support it receives from the observed data. One drawback to model mixing are computational difficulties when dealing with very large model spaces. In large problems it may be infeasible to make predictions based on all the models or even to calculate the posterior probabilities of all models in order to select a set of useful ones (models with high enough probability to contribute to the average). Therefore, an important research goal is to find a set of plausible models and their corresponding weights without enumerating the model space. Doing this efficiently is the challenge.

This dissertation presents an approach for finding a set of plausible models and their weights by implementing an orthogonalization technique. Orthogonalization eliminates collinearity and therefore reduces the number of competing plausible models. In addition, it allows for efficient methods to sample models. Before discussing our method of orthogonalized model mixing in detail, we review some recent Bayesian variable selection and related sampling techniques.

1.2 Literature Review

Various Bayesian model mixing and variable selection methods have been proposed in recent years. The original idea of model mixing dates back several decades (de Finetti 1937; Leamer 1978). It is a technique that makes predictions based on averaging over all models where each model contributes to the prediction proportionally to the support it receives from the observed data. In general, to predict a quantity of interest ϕ, set

$$\phi = \sum_{\gamma} \phi_{\gamma} \pi(\gamma|Y),$$

where ϕ_γ is the quantity of interest for the given model γ, $\pi(\gamma|Y)$ is the posterior probability of that model, and the summation is over the entire model space. This prediction may be difficult to obtain, however, because the summation over all models may be infeasible for very large model spaces and

$$\pi(\gamma|Y) = \frac{P(Y|\gamma)\pi(\gamma)}{\sum_{\gamma'} P(Y|\gamma')\pi(\gamma')}, \tag{1.1}$$

where $P(Y|\gamma) = \int P(Y|\beta,\gamma)P(\beta|\gamma)d\beta$ may not have an analytical solution in generalized linear models. Statisticians have only recently been able to attack those two difficulties through the development of new computational tools such as Markov chain Monte Carlo methods. Most recently, Draper (1995) discussed a Bayesian approach for the assessment of propagation of structural uncertainty, and Raftery, Madigan, and Volinsky (1996) discussed taking model uncertainty into account to strengthen prediction performance in survival analysis. Other variable selection methods based on sampling from the model space using Markov chains include the methods of George and McCulloch (1993,1994); George, McCulloch, and Tsay (1994); Madigan and Raftery (1994); Madigan and York (1995); Geweke (1994); Carlin and Chib (1995); Phillips and Smith (1995); and Green (1995).

Below is given a review of recent papers on the current literature describing approaches for linear regression and generalized linear models. For each paper we review the types of priors used in the respective setups, types of sampling techniques, and the applications to model mixing. We will begin by discussing an early Bayesian setup of variable selection by Mitchell and Beauchamp (1988). Using similar setups, but introducing sampling schemes, the ideas of George and McCulloch (1993, 1994); George, McCulloch, and Tsay (1994); Madigan and Raftery (1994); and Madigan and York (1995) are

described. Although Albert and Chib (1993) and Gamerman (1994) did not discuss model mixing, we review their techniques for sampling from posterior distributions. Then, an interesting sampler by Green (1995) is discussed. Finally, an approach by Kuo and Mallick (1995) that implements a sampling scheme which also uses a slightly different setup than Mitchell and Beauchamp is explored.

Mitchell and Beauchamp (1988) developed a Bayesian variable selection approach in linear regression. The basic goal of their paper is to select subsets of candidate variables for the purpose of predicting the dependent variable. The dependent variable is assigned a probability distribution using prior distributions on the unknown parameters for the regression model. The canonical regression setup with $Y|\beta,\sigma^2 \sim \mathrm{N}_n(X\beta,\sigma^2 I)$ is considered where Y is $n\times 1$, X is $n\times p$, β is $p\times 1$, and σ^2 is a scalar. β and σ^2 are unknown. Their setup includes a "spike and slab" prior distribution on each regression coefficient. This means the distribution is a mixture of a point mass at zero and a diffuse uniform distribution between two limits. This setup allows for a variable to be deleted by assigning a prior distribution with a probability mass concentrated at 0. Through enumeration of all submodels, the posterior probabilities of the submodels are calculated. When the number of submodels is large, the paper suggests to use a branch-and-bound method, developed earlier by Mitchell and Beauchamp (1986), which avoids the calculations for submodels with negligible posterior probabilities, determined through previous calculations. Model mixing can be implemented, although it was not discussed in the paper, using the calculated posterior probabilities and averaging over the available submodels.

George and McCulloch (1993) also developed a variable selection procedure that uses a similar setup but incorporates a sampling scheme to handle large

model spaces. The method, called stochastic search variable selection (SSVS), finds subsets of variables with high posterior probabilities without enumerating the model space. The setup is a normal regression with additional hierarchical parameters with a vector of indicator variables to identify the subsets of predictors. SSVS uses Gibbs sampling to sample from the multinomial posterior distribution on the set of all possible subsets. Frequent appearance of subsets identifies high probability subsets. The canonical regression setup described in Mitchell and Beauchamp (1988) is considered. The subset indicator variable, γ, is introduced into the problem through the prior distribution of β which is the normal mixture:

$$\beta_i|\gamma_i \ \sim \ (1-\gamma_i)\mathrm{N}(0,\tau_i^2) + \gamma_i \mathrm{N}(0, c_i^2\tau_i^2),$$

where $\gamma_i = 0$ or 1. George and McCulloch set τ_i positive and small so that if $\gamma_i = 0$, β_i could be estimated by 0, and set c_i large so that if $\gamma_i = 1$, β_i could be a non-zero estimate and included in the model. This is an important difference from Mitchell and Beauchamp (1988) because this approach does not put a probability mass on $\beta_i = 0$. Prior distributions are placed on σ^2 and γ. SSVS uses the Gibbs sampler to generate the sequence

$$\beta^0, \sigma^0, \gamma^0, \beta^1, \sigma^1, \gamma^1, \ldots, \beta^M, \sigma^M, \gamma^M,$$

where β^j is found by sampling from a normal distribution,

$$\beta^j \ \sim \ p(\beta^j|Y, \sigma^{j-1}, \gamma^{j-1}),$$

σ^j is found by sampling from an inverse gamma distribution,

$$\sigma^j \ \sim \ p(\sigma^j|Y, \beta^j, \gamma^{j-1}),$$

and γ_i^j is found by sampling from a Bernoulli distribution,

$$\gamma_i^j \;\sim\; p(\gamma_i^j|Y,\beta^j,\sigma^j,\gamma_{-i}^j) = p(\gamma_i^j|\beta^j,\sigma^j,\gamma_{-i}^j)$$

for $i = 1,\ldots,p$. Notice that, since γ only enters the problem through the prior on β, γ affects Y only through β, and therefore, the full conditional distribution of γ does not depend on Y. This reduces computational requirements and allows for faster convergence of the subset indicator variables. After the sequence $\gamma^0,\gamma^1,\ldots,\gamma^j,\ldots$ has converged, the most frequently sampled γ corresponds to the most likely subset of variables since it is the γ with the largest posterior probability under $p(\gamma|Y)$. Model mixing can then be implemented by weighting the models sampled by their frequencies and averaging over them. In George and McCulloch (1994) conjugate formulations which allow for analytical simplification (the β and σ can be eliminated from $\pi(\beta,\sigma,\gamma|Y)$ through integration) were discussed. This particular implementation of SSVS is discussed further in Chapter 2.

George, McCulloch, and Tsay (1994) extended the previously described linear regression SSVS method to generalized linear models where $Y_i \sim p(Y_i|\theta_i,\omega)$ with $\theta_i = X_i\beta$ and ω representing any additional parameters. The generalized SSVS computes the posterior via Gibbs sampling drawing from:

$$\begin{aligned} \beta &\mid \gamma,\omega,Y; \\ \omega &\mid \beta,Y; \\ \gamma_j &\mid \gamma_{-j},\beta. \end{aligned}$$

Feasibility of the approach is based on the ability to draw from the first two conditionals, since drawing γ_j is simply a Bernoulli draw, as is the case in the linear regression model. Applications of SSVS to designed experiments are

presented in Chipman (1996) and Clyde and Parmigiani (1994). Bennett and Wakefield (1994) applied SSVS to population models in pharmacokinetics.

Next, we review a method introduced by Madigan and Raftery (1994) that was also designed to sample models, but specifically intended to be utilized for model mixing. Unlike SSVS, this method uses a deterministic approach to select high probability models that can be used in model averaging. To account for model uncertainty, model averaging can be difficult to apply when the number of models is large; therefore, finding a smaller set of models makes the technique feasible. To find a smaller set of models to use in the averaging scheme, Madigan and Raftery proposed Occam's window. The algorithm first eliminates any models with small posterior model probability which is found exactly through enumeration. It then eliminates any models still considered that have smaller model probability than any of their submodels. The set of remaining models is used in the model averaging. By reducing the set of models to a reasonable number, model averaging becomes a feasible approach. Another approach which finds a subset of models, called the Markov chain Monte Carlo model composition (MC^3), is introduced in Madigan and York (1995). This approach directly approximates the exact solution using a stochastic process to move through the model space. The technique considers one variable at a time to determine whether to include it in the model. Raftery, Madigan, and Hoeting (1993) also used the standard normal-gamma conjugate class of priors, a similar setup to George and McCulloch (1994), and applied both of these methods for linear models to data on crime rates (Vandaele 1978); this data set is also used for comparison in Chapter 2. Proportional hazard models (Raftery, Madigan, Volinsky 1996) and graphical models (Madigan, Gavrin, and Raftery 1994; Madigan and Raftery 1994; Madigan and York 1995) are

other examples of applications of Occam's window and MC^3.

Another method which incorporates Occam's window is proposed by Volinsky et al. (1996) for generalized linear models. Similar to Madigan and Raftery (1994) and Madigan and York (1995), this method was specifically intended to be used in model mixing. It is based on an adaptation of Furnival and Wilson's (1974) leaps and bounds algorithm for linear regression and uses the Bayesian information criterion (BIC) approximation (Schwarz 1978; Raftery 1986) for the posterior model probabilities. The approximation is necessary because in generalized linear models the exact posterior model probabilities in Equation (1.1) cannot always be found. Lawless and Singhal (1978) developed a modification of the leaps and bounds algorithm for nonlinear regression models that Volinsky et al. modified for generalized linear models. The algorithm provides the top models for each model size and an approximate likelihood ratio test statistic that can be an approximation for the BIC value. The BIC value for the model γ is $\text{BIC}_\gamma = L^2_\gamma - df_\gamma \log n$ where L^2_γ is the deviance for model γ, df_γ is the corresponding number of degrees of freedom, and n is the number of observations. The best models are those with the largest BIC values. The approach then applies the idea of Occam's window with approximated model probabilities (found from the approximated BIC values) instead of the enumerated exact posterior probabilities as in Madigan and Raftery (1994). Occam's window is a set containing those models with the highest approximated posterior probabilities. For the smaller set of models in Occam's window the posterior model probabilities are approximated more accurately using the exact BIC calculations. The models contained in this set are the models chosen for model mixing. Also, for each model chosen, the regression coefficients are approximated with the maximum likelihood estimator. Other alterations on

the way to compare models using Occam's window using Bayes factors and applications are in Raftery (1995); Raftery, Madigan, and Volinsky (1996); Raftery (1996); and Raftery and Richardson (1996). Kass and Raftery (1995) offer a review of Bayes factors and its approximation using BIC.

Albert and Chib (1993) offer another approach for using a Gibbs sampler to handle the integration in Equation (1.1). This sampling approach can be used in connection with George, McCulloch, and Tsay (1994) to implement model mixing in generalized linear models. Albert and Chib use Bayesian methods and data augmentation to model categorical response data. Specifically, binary regression with the probit link is examined. The fact that the probit regression model can be represented as a normal regression model on continuous latent data is used. Using Gibbs sampling with this fact, simulating from the posterior distribution of β can be accomplished. The method proposed does not introduce a sampling scheme for the models. However, a separate Gibbs sampler step, similar to that in SSVS, could be applied to sample models, which would enable the method to be used in model mixing.

Another method for making Bayesian inference in generalized linear models, but not designed for model mixing, is described in Gamerman (1994). The method is designed for generalized linear models with random effects and focuses on obtaining samples from the posterior distribution when the model is assumed to be fixed. Gamerman combines the iterative weighted least squares algorithm with Markov chain Monte Carlo techniques, specifically the Metropolis-Hastings algorithm (see Subsection 3.2.1 for more details). The approach samples regression coefficients from a proposal density that combines the prior distribution and the weighted least squares estimator and its corresponding covariance matrix. The proposal density is a normal distribution

with moments that are based on the previous state. The acceptance probability is based on the ratio of the posterior densities evaluated at the proposed coefficients and at the previously accepted coefficients, and on the ratio of the proposal densities. The numerator of the proposal ratio is evaluated at the candidate coefficients with the moments evaluated at previously accepted coefficients, whereas the denominator is evaluated at the previously accepted coefficients with the moments evaluated using the candidate coefficients. The paper presents a method to find a sample of regression coefficients from the posterior distribution in generalized linear mixed models; however, it does not incorporate sampling a parameter for the models. A model parameter could be included. Then, for problems with a small number of variables, enumeration of the model space could be carried out, or, for many variables, a sampling scheme for selecting the models could be implemented. Applying these schemes enables the method to be used for model averaging.

Different than any method described previously, Green (1995) presents a Markov chain Monte Carlo method where the dimension of the parameter vector being sampled is not required to be fixed. The paper develops a reversible Markov chain sampler that jumps between parameter subspaces of different sizes. This "reversible jump MCMC" method extends Metropolis–Hastings methods and applies to many problems, such as variable selection, model selection and model averaging. Another dimension changing method is based on the jump-diffusion sampler of Grenander and Miller (1994) and elaborated on by Phillips and Smith (1995) for Bayesian model comparisons. Both of these approaches deal with changing dimensions in sampling and can be used to sample regression coefficients by changing dimensions for different models.

Unlike the setup of George and McCulloch (1993) who introduce the model

parameter through the prior on the coefficient, Kuo and Mallick (1995) incorporate the model parameter through the likelihood. Similar to the SSVS methods, model mixing can be implemented. Kuo and Mallick's method uses a MCMC algorithm to identify promising models in linear regression and generalized linear regression. Their proposed method selects a subset of independent variables. Their generalized linear model extension is described here. The goal is to find the posterior distribution of a model γ using a Markov chain Monte Carlo algorithm so that subsets of predictors with high posterior probability can be identified. The method uses a combination of the Gibbs sampler and Metropolis algorithm. The approximate posterior probability of a model γ is calculated from the frequency of γ appearing in the Gibbs sampler. Then sampling the regression coefficient β given Y and γ can be accomplished through a random walk proposal in the Metropolis algorithm. As an alternative, Kuo and Mallick also suggest adaptive rejection sampling similar to George, McCulloch and Tsay (1994) for updating β.

The brief summaries presented above demonstrate a wide variety of Bayesian approaches. Although many were not designed specifically for model averaging, through minor alterations all could become model mixing methods. Throughout the literature on linear and generalized linear models, many ideas with great potential have been presented and have inspired others to develop their own approaches. This dissertation is one such example.

1.3 Overview

This dissertation presents a new method to make predictions using model mixing to account for model uncertainty; in particular, the uncertainty associated

with selecting explanatory variables in large prediction problems. We will focus on normal linear regression and Poisson regression models for which we propose an approach that uses an orthogonalized transformation of the variables to approximate the posterior model probabilities. This approach supplies a simple and fast method to sample models.

Chapter 2 deals with linear regression models for which a sampling approach using an importance sampling technique is developed. This technique is based on an approximation to the posterior model probabilities and an orthogonal transformation of the variables. The posterior probability of a model is approximated by a product of independent Bernoulli random variables, each indicating whether or not an element of the orthogonal basis is included. This method is compared to and found more efficient than Gibbs samplers or other MCMC methods. Our comparison is based on alternative stochastic sampling methods such as the SSVS and MC^3 discussed in the previous subsection. Then, to show the method's ability to efficiently deal with very large data sets, it is applied to a data set from a designed experiment dealing with predicting protein activity under different storage conditions. The model space of this set is very large (the rank of the design matrix is 88) and very difficult to explore if expressed in terms of the original variables.

In Chapter 3 we consider Poisson regression models, where the mean of the response variable is related to a linear combination of the predictor variables through a function called the link function. In linear models, we find that using orthogonalized variables, in addition to centering and rescaling, leads to dramatic improvements in sampling high probability models. One difficulty in generalized linear models, in particular Poisson regression, is that one cannot analytically integrate out model specific parameters from the posterior model

probabilities —a key step in obtaining the probabilities for sampling and model mixing in Chapter 2. We deal with this difficulty by introducing a variance stabilizing transformation to obtain a constant variance and by converting the problem into a normal approximation which is valid for linear models. Then, through further approximation using a Taylor series expansion, the posterior model probabilities are approximated by a product form that is suitable for sampling. We also present a second approach to obtain a suitable function for sampling. This approach is based on using a Laplace approximation to integrate out the additional parameters and then use a Taylor series expansion to obtain the approximate posterior model probabilities. As in Chapter 2, to demonstrate the ability of the method to deal with large data sets, our method is applied to a data set of 126 variables that was designed to examine the effect of particulate pollution on daily death counts.

As a final introductory remark regarding Chapters 2 and 3, we caution that those chapters are intended to be self-contained articles, and therefore, may exhibit overlaps in the introductory sections.

Finally, Chapter 4 presents a scheme to generalize the proposed approach to other exponential family distributions. We then discuss possible directions for future work and present some concluding remarks.

Chapter 2

Linear Models[1]

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. Because 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.

[1]This chapter will appear in the September 1996 issue of the *Journal of the American Statistical Association* under the title Prediction via Orthogonalized Model Mixing.

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.

The focus of this chapter 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 standpoint 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 overfitting the data. The latter is achieved by a data-based shrinkage of the regression coefficients (see also George 1986a, 1986b). Here 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.

This chapter is organized as follows. In Section 2.2 we introduce orthogonalized model mixing by giving the basic notation and definitions. In Section 2.3 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 2.4 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 2.5 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 orthogonal-

ized 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 2.6, 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.

2.1 Orthogonalized Model Mixing

2.1.1 Prior Distributions and Orthogonalization

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 let X be the $n \times p$ design matrix including all candidate predictors. X can include transformations of the variables originally recorded. We begin by assuming that

$$Y|\beta, \sigma^2 \sim N_n(X\beta, \sigma^2 I_n), \tag{2.1}$$

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:

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

where B, b_0, ν and ξ are fixed hyperparameters. 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 (2.1) 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

$$\alpha|\sigma^2 \sim N_p(a_0, \sigma^2 A), \tag{2.2}$$

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 we take 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). However, this means introducing uncertainty due to sampling variation in the orthogonal basis, and, in this setup, data dependence in the prior distribution on α. Clyde and Parmigiani (1996) presented several alternative orthogonalization strategies in detail. Further discussion is also presented in Section 2.7.

The remainder of this chapter 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 nonorthogonal 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 by Clyde and Parmigiani (1996).

2.1.2 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 (2.2) 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 (1995), 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 2.3, we make the additional assumption that $\pi(\gamma)$ factors as

$$\pi(\gamma) = \prod_{i=1}^{p} \pi(\gamma_i) \equiv \prod_{i=1}^{p} \theta_i^{\gamma_i}(1-\theta_i)^{1-\gamma_i}. \tag{2.3}$$

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 2.2.3. In particular, taking θ_i's less

than 0.5 enforces a penalty for each additional term in the model (see Clyde and Parmigiani 1996).

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. Because our goal is prediction, the primary concern is selecting columns of Z, which identify interesting subspaces to represent the mean response. We find it attractive to give priority to this from both a modeling 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 $\mathrm{SSR}_i^2 = ||P_{\tilde{z}_i}\tilde{Y}||^2$, 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,

$$\pi(\gamma|Y) = \frac{p(Y|\gamma)\pi(\gamma)}{\sum_{\gamma'} p(Y|\gamma')\pi(\gamma')} = \frac{q_\gamma}{\sum_{\gamma'} q_{\gamma'}}. \tag{2.4}$$

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

$$\begin{aligned} \log(q_\gamma) &= \sum_{i=1}^{p} \gamma_i \left[\log\left(\frac{\theta_i}{1-\theta_i}\right) - \frac{1}{2}\log d_i \right] \\ &\quad -\frac{(n+\nu)}{2} \log\left(\nu\xi + \tilde{Y}'\tilde{Y} - \sum_{i=1}^{p} \gamma_i \, \mathrm{SSR}_i^2\right). \end{aligned} \tag{2.5}$$

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

$$\phi = \sum_{\gamma} \phi_\gamma \pi(\gamma|Y). \tag{2.6}$$

For example, Equation (2.6) 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 Equation (2.4) or ϕ in Equation (2.6) 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 2.3.

2.1.3 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 overfits the points. Regression on a subset of size k of the principal components based on the k largest eigenvalues is sometimes used to alleviate overfitting 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 drawback, as all possible subsets of principal components are incorporated in the regression.

To give the flavor of the shrinkage implication of model mixing, consider the form of the Bayes estimator of α. Under the full model, this is given by $\tilde{\alpha} = (\tilde{Z}'\tilde{Z})^{-1}\tilde{Z}\tilde{Y}$. Because 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 coordinatewise 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 ordinary least squares (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, then 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 and 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 Equation (2.5). Evaluating the amount of shrinkage for some key models can provide insight on the strength of the prior specification being used. (See George 1986a, 1986b for further discussion of multiple shrinkage estimators.)

2.2 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:

- 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.

- 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 Markov chain Monte Carlo (MCMC) methods.

- 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 Equation (2.4). This can be obtained by expressing Equation (2.5) as a linear function of γ. We achieve this via a Taylor series expansion of the last term in Equation (2.5). In particular, expanding around $\nu\xi + \tilde{Y}'\tilde{Y}$, we have

$$\begin{aligned} \log\left(\nu\xi + \tilde{Y}'\tilde{Y}\right) - \left(\sum_{i=1}^{p} \gamma_i \,\mathrm{SSR}_i^2\right) &= \log(\nu\xi + \tilde{Y}'\tilde{Y}) \\ &\quad - \sum_{j=1}^{k} \frac{1}{j}\left[\frac{\sum_{i=1}^{p} \gamma_i \,\mathrm{SSR}_i^2}{(\nu\xi + \tilde{Y}'\tilde{Y})}\right]^j + A_k \end{aligned} \tag{2.7}$$

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

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

Ignoring the remainder term, this expression would lead to a factorization of q_γ permitting independent importance sampling on the columns of Z. However, although 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

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

The calibrated factorizable expansion is then

$$\begin{aligned}
\log\left(\nu\xi + \tilde{Y}'\tilde{Y} \right. &- \left. \sum_{i=1}^{p} \gamma_i \,\mathrm{SSR}_i^2\right) \approx \frac{L_N - L_F}{\tilde{L}_N - \tilde{L}_F} \\
&\times \left[\log(\nu\xi + \tilde{Y}'\tilde{Y}) - \sum_{j=1}^{k} \frac{\sum_{i=1}^{p} \gamma_i \,\mathrm{SSR}_i^{2j}}{j(\nu\xi + \tilde{Y}'\tilde{Y})^j}\right] \qquad (2.8)
\end{aligned}$$

Replacing (2.8) into (2.5) we obtain:

$$\begin{aligned}
\log(q_\gamma) \approx\; & Q + \sum_{i=1}^{p} \gamma_i \left[\log\left(\frac{\theta_i}{1-\theta_i}\right) - \frac{1}{2}\log d_i \right. \\
& \left. + \frac{(n+\nu)}{2} \frac{L_N - L_F}{\tilde{L}_N - \tilde{L}_F} \sum_{j=1}^{k} \frac{\mathrm{SSR}_i^{2j}}{j(\nu\xi + \tilde{Y}'\tilde{Y})^j}\right] \qquad (2.9)
\end{aligned}$$

where Q is a quantity that does not depend on γ. Equation (2.9) 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:

$$\tilde{\pi}(\gamma|Y) = \prod_{i=1}^{p} p_i^{\gamma_i}(1-p_i)^{1-\gamma_i}, \tag{2.10}$$

where:

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

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 MCMC 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 p_i's. The easiest model space to sample is one where all the p_i's are 0 except for one. On the other hand, if $p_i \approx .5$, little is gained by sampling from $\tilde{\pi}$. The orthogonalization strategy can therefore be important in determining the efficiency of the sampler. Orthogonalizations that identify interesting target subspaces will lead to a mix of columns of Z with large p_i and small p_i. Thus summary measures of the variability of the p_i's provide an indirect way of assessing the efficiency of the adopted orthogonalization strategy.

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.

2.3 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

$$\phi = \frac{\sum_{\gamma\in D} q_\gamma\phi_\gamma + \sum_{\gamma\in\bar{D}} q_\gamma\phi_\gamma}{\sum_{\gamma\in D} q_\gamma + \sum_{\gamma\in\bar{D}} q_\gamma}, \tag{2.11}$$

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

$$\hat{\phi} = \sum_{\gamma \in D} w_\gamma \hat{Y}_\gamma, \tag{2.12}$$

where w_γ are normalized weights. The choices that we consider for empirical comparison in Section 2.5 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 (Carlin and Chib 1995; Geweke 1994). 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 unnormalized 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 1, so that this model dominates in the mixture estimator. Descriptive summaries of the weights can indicate when this is a problem.

2.4 Crime Data

2.4.1 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. Following Raftery, Madigan, and Hoeting (1993), we used the natural logs of all continuous variables in the analysis. We are interested in 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 zero (as in George and McCulloch 1993) and the variance chosen by assuming that $E(\sigma)c_{ii} = \delta/\mathrm{IQR}(X_i)$ is three standard deviations away from zero. For the one-dummy variable, we used the range instead of the interquartile range. In this way, the three-standard deviation interval on the marginal distribution of each coefficient includes "large" values of the coefficient. "Large" values refer to values that are 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: it is entirely in terms of observables (the response), and 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 $\xi = 0.5/3$. 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 1.

2.4.2 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 2.3.

2. Random Sampling. Columns of the orthogonal basis are included with probability 0.5, independently of each other. In our setting this is equivalent to using the prior distribution as an importance function for the model space.

3. Stochastic Search Variable Selection (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. Markov Chain Monte Carlo Model Composition (MC^3). 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 MC^3, as opposed to the Gibbs update in SSVS. The motivation for considering a hybrid algorithm of integrated SSVS and MC^3 is the following. In the MC^3 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 MC^3 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 MC^3 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.

2.4.3 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 mean squared error (MSE) of 0.0313 versus 0.0432 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.

Comparison of Stochastic Search Algorithms in the Orthogonalized Space. The five algorithms of Section 2.5.2 are compared in Figures 2.1 and 2.2. Figure 2.1, compares 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 iterations, 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 $\tilde{\pi}(\gamma|Y)$, 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 (2.6). 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 (2.12) to calculate $\hat{\mu}^A$ for all search methods. Then, the integrated squared error (ISE) 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 $p^A(y|Y, x_6)$,

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

Figure 2.2 shows the resulting comparisons using the ISE and the Kullback–Leibler divergence. 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 nine 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 MC^3 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 MC^3 $p = 15$ times longer than the alternatives, and taking every fifteenth model.

Comparison of Alternative Rules for Estimating the Predictive Mean. Figure 2.3 shows a comparison of the estimators of the predictive mean discussed in Section 2.4. As in previous comparison section, 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.

Comparison of Algorithms in the Original and Orthogonalized Space. Finally, Figure 2.4 compares orthogonalized model mixing using importance sampling, orthogonalized model mixing using a hybrid Markov chain, and 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. Because orthogonalized model mixing and standard model mixing generate a different set

of exact predictive means, each of the ISE's 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.

2.5 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. Many factors affect 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 was 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 four 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 $\theta_i = 0.8$. 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 analysis of variance (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 hyperparameters for the prior distribution of σ^2 were

taken as $\nu = 50$ and $\xi = 0.01$ 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 2.4. 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 23,000 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 $\tilde{Z}_\gamma$ 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. Because 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,

$$p(y^*|Y,Z^*,\gamma) \propto \left[(n+\nu)+(y^*-Z^*_\gamma\hat{\alpha}_\gamma)^T \times\frac{(I+Z^*_\gamma(\tilde{Z}_\gamma{}^T\tilde{Z}_\gamma)^{-1}Z^{*T}_\gamma)^{-1}}{(\xi\nu+S^2_\gamma)/(n+\nu)}(y^*-Z^*_\gamma\hat{\alpha}_\gamma)\right]^{-\frac{(n^*+n+\nu)}{2}} \quad (2.13)$$

with $S^2_\gamma = \tilde{Y}'\tilde{Y} - \sum_{i=1}^p \gamma_i \, \mathrm{SSR}_i^2$. Evaluation of the desired probabilities can proceed by pointwise 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 generation of an observed vector from (2.13). The first approach is preferable if marginal probabilities are of interest. The second approach, adopted here, is better suited to handle higher-dimensional integrals, such as the probabilities that each setting will have the highest protein activity.

Figure 2.5 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 2.6 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 (2.13).

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.

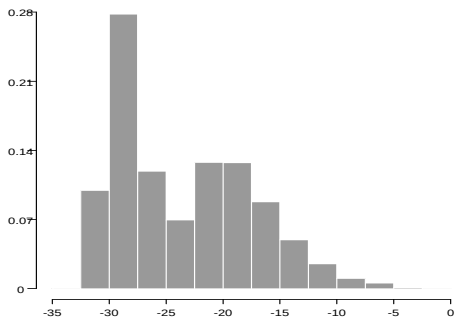
2.6 Discussion

Bayesian model mixing offers a fruitful theoretical framework for making predictions that account for uncertainty in the selection of predictor variables. In this chapter we introduced 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 the orthogonalization method and the prior probability distributions assigned to both the models and the coefficients. In earlier work, Clyde and Parmigiani (1996) looked at the predictive distributions resulting from model mixing under alternative orthogonalizations. The example considered 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.

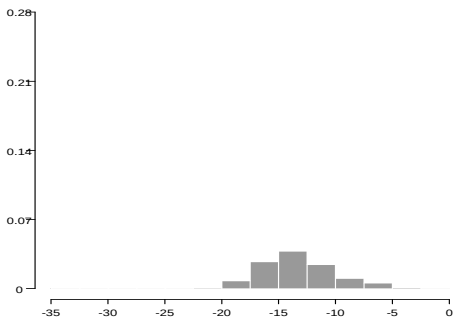
However, the choice of a prior and an orthogonal basis can affect the predictive distribution substantially. Experimental cases indicate that results seem to be more sensitive to the choice of prior parameters than to the choice of a basis. In particular, a key choice is that of the prior distribution 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, then 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.

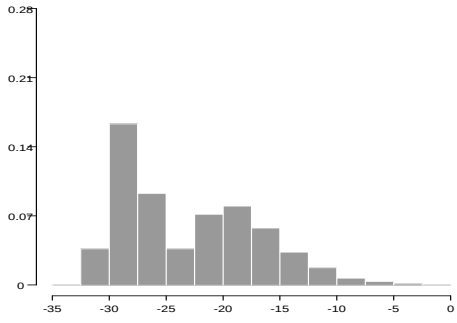
For these reasons, orthogonalized model mixing cannot be recommended as a black-box prediction method. However, advantages of careful implementations are both statistical and computational. Orthogonalization leads to a better behaved problem, as the number of competing models is reduced by eliminating 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.



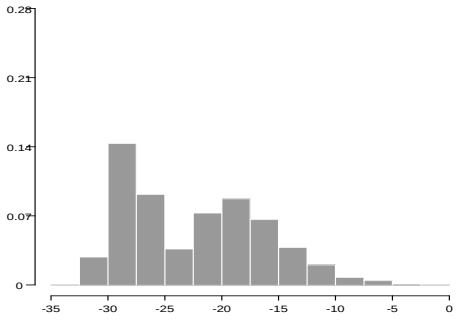
All Models
$\sum \pi(\gamma|Y) = 1$ for 32,768 models



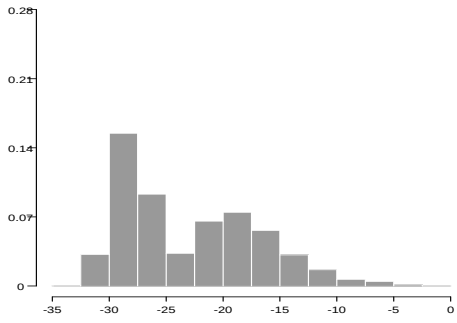
Importance Sampling
$\sum \pi(\gamma|Y) = 0.9990$ for 3,759 models



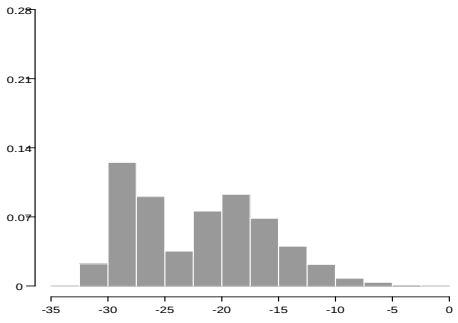
Random Sampling
$\sum \pi(\gamma|Y) = .63$ for 19,661 models



SSVS
$\sum \pi(\gamma|Y) = .57$ for 19,567 models



MC^3
$\sum \pi(\gamma|Y) = .58$ for 18,655 models



Hybrid
$\sum \pi(\gamma|Y) = .52$ for 19,171 models

Figure 2.1: Comparison of algorithms: Distributions of the logarithm of the posterior probabilities of the sampled models based on 30,000 iterations of each stochastic algorithm.

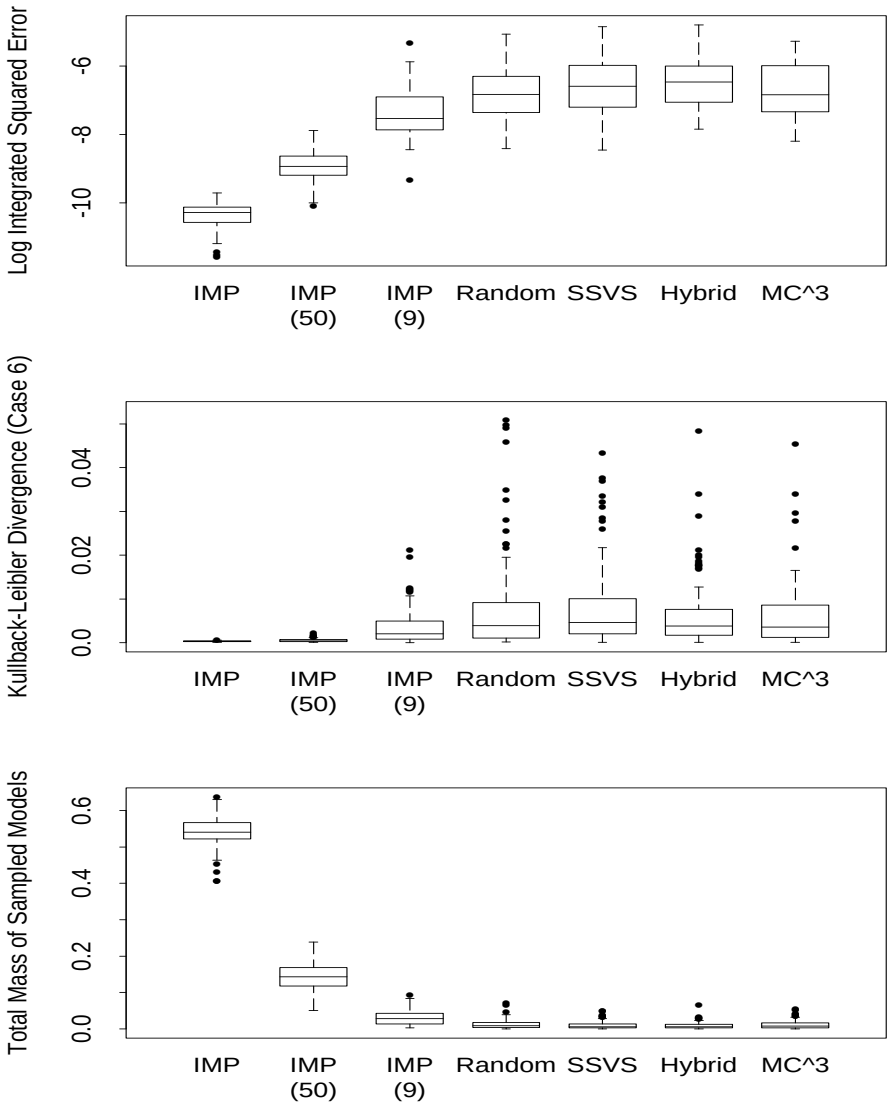


Figure 2.2: Comparisons of algorithms based on log ISE, Kullback–Leibler Divergence for the predictive distribution for case 6, and total posterior probability of sampled models. Boxplots are based on 100 samples of 300 iterations of each stochastic algorithm. Gains from importance sampling, labeled 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).

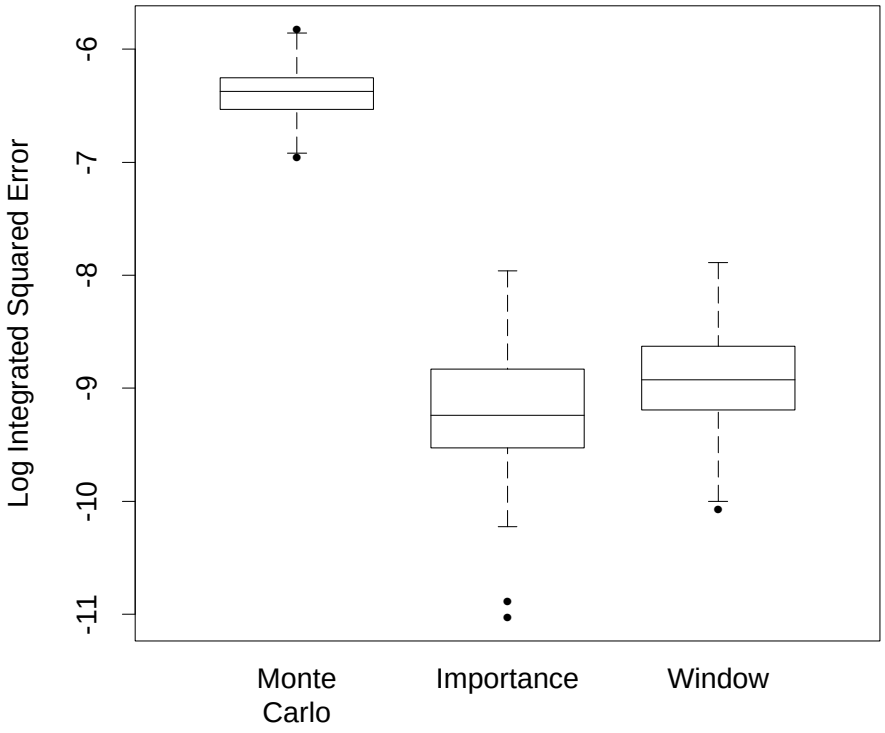


Figure 2.3: Comparison of log ISE's for alternative estimators of the exact predictive mean vector. Boxplots refer to, from left to right, the Monte Carlo estimator, the importance sampling estimator and the window estimator. Results are based on 100 replications using 50 samples from the importance sampler.

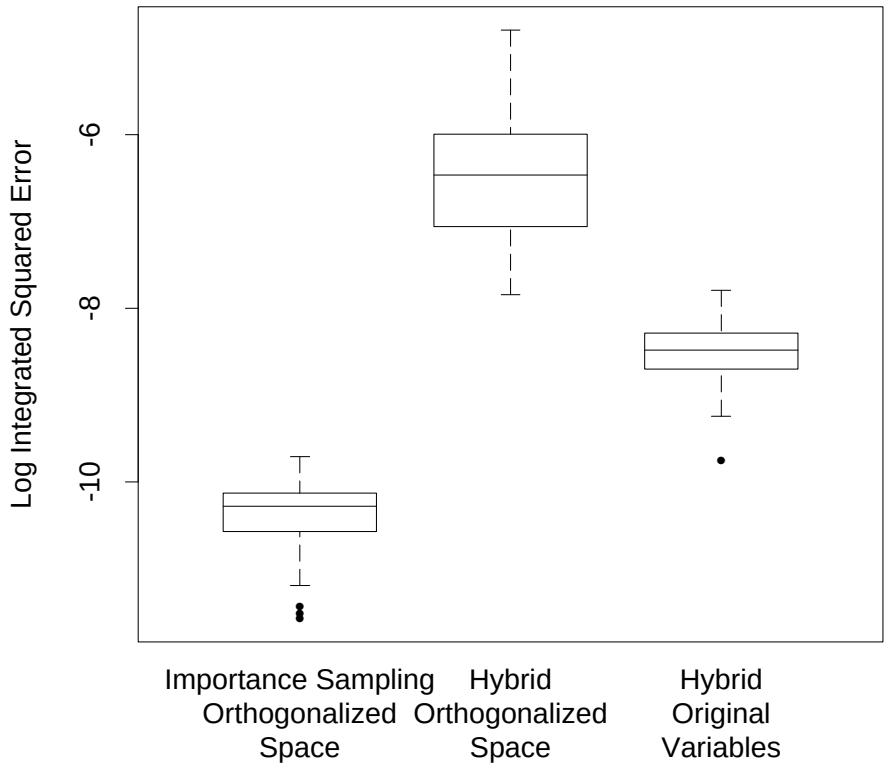


Figure 2.4: Comparison of log ISE for estimation of the exact predictive mean using orthogonalized model mixing with importance sampling and Markov Chains versus standard model mixing with Markov chains. Boxplots are based on 100 replications of 300 iterations of the stochastic search algorithms. In addition to the better performance displayed in the figure, orthogonalized model mixing is substantially faster.

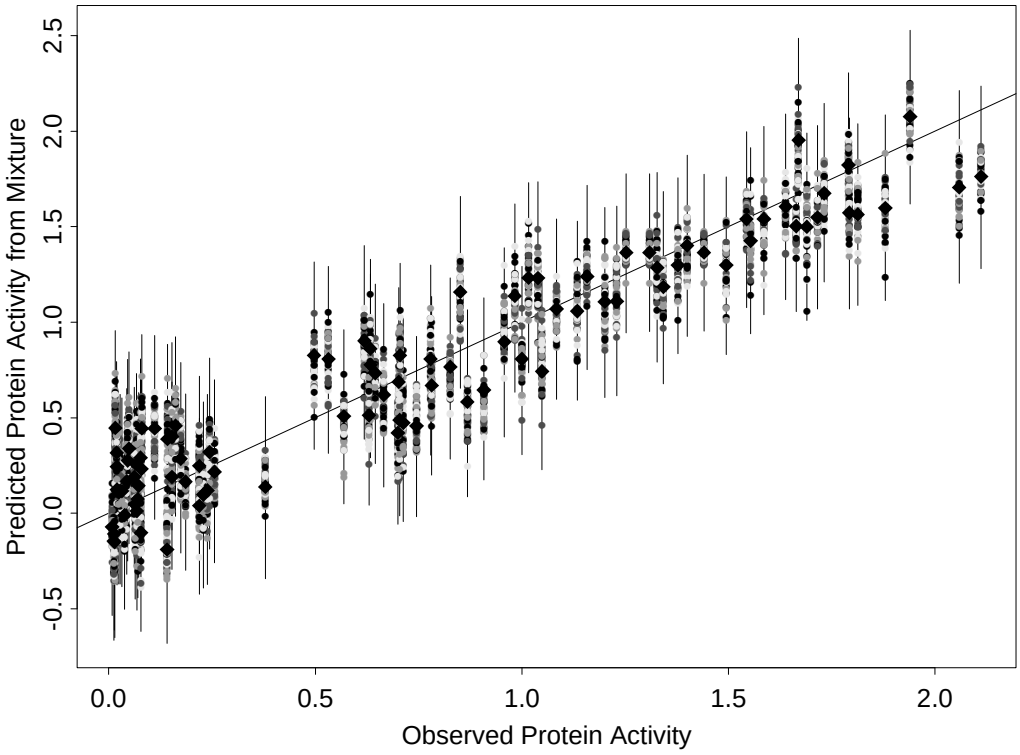


Figure 2.5: Mixture based predictions for the Protein Construct example. The solid diamonds represent the predictive means from model mixing. The vertical lines correspond to 95% probability intervals from model mixing. Also, stars represent the means of the model-specific predictive distributions for 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.

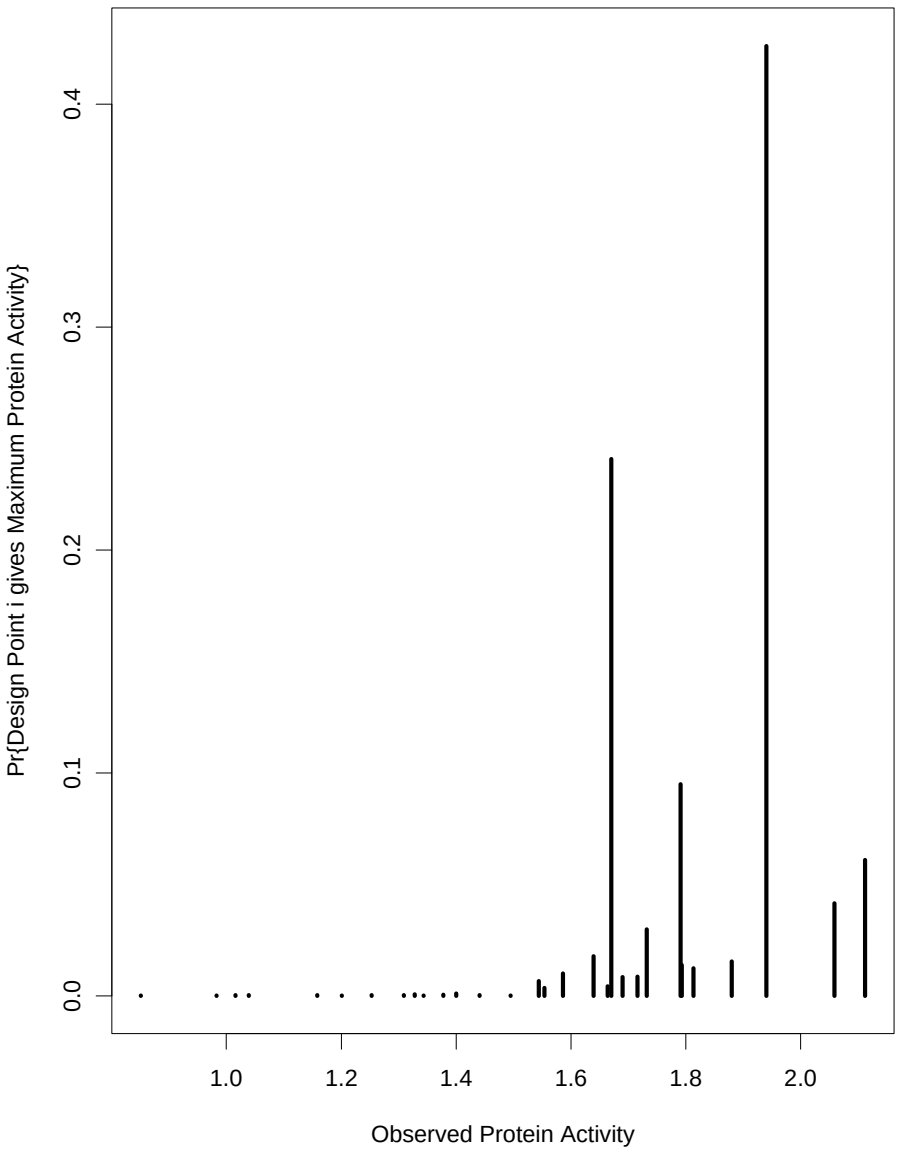


Figure 2.6: Probability of yielding the maximum for each of the experimental settings, by observed response level.

Chapter 3

Poisson Regression Models

We explore modeling strategies and numerical methods for prediction under model mixing in Poisson regression models. Bayesian model mixing offers an effective alternative to single model selection for making predictions. Unlike predictions based on a single model, which may lead to unsatisfactory predictions in complex models because model uncertainty is neglected, model mixing considers the uncertainty associated with the selection of a model by averaging over a set of models. In this alternative approach each model weights the prediction proportionally to the support it receives from the data.

Determining which models to include in the model mixing scheme is a challenging problem in statistical modeling because enumeration over all possible models is not feasible for problems containing a large number of predictor variables. In particular, for a problem containing p predictor variables, there exist 2^p possible models. Therefore, an important research goal is to find a set of plausible models and their corresponding weights without enumerating the model space. Doing this efficiently is the challenge.

Chapter 2 developed an importance sampling method for sampling model

space in linear regression models which appears to be more efficient than other MCMC methods in identifying models with high posterior probabilities. In this chapter we will further develop and extend this approach as outlined briefly below.

Section 3.1 introduces an approach for implementing model mixing in large prediction problems in generalized linear models, in particular, Poisson regression models. Our approach, which leads to both statistical and computational advantages, is based on an approximation of the posterior model probabilities using an orthogonalized transformation of the variables. Orthogonalization leads to a better identified problem as the number of competing plausible models is reduced as a result of eliminating high correlation. Also, based on orthogonalization of the space of candidate predictors, we can approximate the posterior probabilities of models by products of predictor-specific terms. This allows for direct sampling from an approximation to the joint distribution of the model space. The main appeal of orthogonalized model mixing is the ease of sampling. Orthogonalization makes it possible to draw elements of the orthogonal basis independently, with probability that closely approximates the actual posterior probability. Compared to standard Markov chain methods, orthogonalized model mixing is faster in sampling models and more efficient in finding models that contribute significantly to the prediction.

Approximate analysis of many generalized linear models can be based on applying a variance stabilizing transformation to the response and then using standard linear regression methods for the analysis. This can provide a basis for developing a sampling function for the posterior model probability. After applying the variance stabilizing transformation, the variance is a known constant. So, using the orthogonalization of the design matrix as in the linear re-

gression framework, the posterior model probability based on the approximate normal likelihood factors as p independent components. In the approximate problem we can integrate out the other parameters to obtain the probability of a model conditional on the response Y, up to a normalizing constant, which was not possible in the original problem. The normal approximation then defines a sampler that can be used to find models with high posterior probability. The models can then be used in a MCMC sampler to generate independent samples of the parameter β. The probability of a model given Y can be evaluated using the sampled output. These probabilities can then be used to weight predictions or other quantities of interest. Section 3.2 describes these possible sampling techniques.

Current literature offers other methods that may be used for comparisons. Using a small simulated data set, Section 3.4 compares two methods, the deterministic method of Volinsky, Madigan, Raftery, and Kronmal (1996) and the MCMC technique of Kuo and Mallick (1995) with the approach presented in Section 3.2. Another approach based on a Laplace approximation is introduced in Section 3.3 and is also compared to the method based on the variance stabilizing transformation in Section 3.4.

In Section 3.5 we use the model mixing methodology to analyze a data set on environmental hazards. The study examines if airborne particulates contribute to excess mortality. The data set is characterized by high correlation between explanatory variables. Traditional model selection methods have difficulties both in identifying important variables and in making predictions (Styer et al. 1994; Smith 1995). Because of this, predictions based on model mixing may be better than those based on one model. In addition, orthogonalizing the explanatory variables reduces correlation among some of the variables which

may strongly improve convergence of MCMC methods (Gelfand, Sahu, and Carlin 1996).

3.1 Orthogonalized Model Mixing

3.1.1 Prior and Posterior Distributions

We begin by giving the basic notation and definitions. Let Y be the response variable, an $n \times 1$ vector of observations, and X be the $n \times p$ design matrix. X contains all candidate predictors, including the intercept, and is assumed to have orthogonal columns. This holds in some designed experiments or contingency tables without missing cells. When the original predictors are not orthogonal, they may be transformed so that the variables are orthogonal to each other as well as to the intercept. There are many possible ways to construct a suitable orthogonalization. Clyde and Parmigiani (1996) discuss in detail the different approaches of principal components (Rao 1964), partial least squares (Wold et al. 1984), Gram-Schmidt, and sliced inverse regression (Li 1991) in which to find an orthogonal basis. Begin by assuming that the Y_j's are independent Poisson random variables with corresponding means λ_j so that the joint distribution of Y is given by

$$Y|\lambda \sim \prod_{j=1}^{n} \text{Pois}(\lambda_j).$$

Under the log link, $\lambda_j = \left(e^{x_j' \Gamma \beta}\right)$ where x_j' is the j^{th} row of the design matrix. As in Chapter 2, a model is represented by the vector γ, a $p \times 1$ vector of 0's and 1's indicating the inclusion of a variable, and Γ is the diagonal matrix with diagonal elements γ_i. In the previous chapter, the model γ was incorporated

through the prior distribution on the coefficients, which was equivalent to a mixture of a normal and a point mass at zero. However, for a nonconjugate setup, this may lead to a reducible Markov chain. Therefore, we now include the γ_i's through the likelihood.

Assume the following prior distributions. Let

$$\beta \sim \mathrm{N}(0, \Sigma)$$

where Σ is a diagonal covariance matrix with elements σ_i^2. As in Chapter 2, the covariance matrix for the coefficients must be diagonal for the factorization. The prior model probabilities are given by

$$\gamma \sim \prod_{i=1}^{p} \text{Bernoulli}(\rho_i)$$

where ρ_i is the prior probability that the i^{th} variable is included. The posterior probability of a model γ is given by

$$\pi(\gamma|Y) = \frac{P(Y|\gamma)\pi(\gamma)}{\sum_{\gamma'} P(Y|\gamma')\pi(\gamma')}. \tag{3.1}$$

This may be difficult to evaluate for two reasons. First, with p candidate predictor variables, the summation over all 2^p possible models $\gamma^{'}$ may be infeasible. The second reason is that $P(Y|\gamma)$ may not be easy to calculate, as is true in this case. In this setup we have that

$$P(Y|\gamma) = \int P(Y|\beta, \gamma)P(\beta|\gamma)d\beta \tag{3.2}$$

which does not have an analytical solution. However, by using a variance stabilizing transformation and a Taylor series expansion, we will derive an approximation for Equation (3.2), and therefore, an approximation for Equation

(3.1). This will be useful in finding models with approximately high posterior probability.

By applying a variance stabilizing transformation to Y, in this case the square root, we have the following approximate distribution for the transformed response:

$$Y^{1/2} \dot{\sim} \mathrm{N}_n \left(e^{\frac{1}{2}X\Gamma\beta}, \frac{1}{4}I \right),$$

where $e^{\frac{1}{2}X\Gamma\beta}$ is the vector with elements $e^{\frac{1}{2}x_j'\Gamma\beta}$, $j = 1, \ldots, n$ and I is the identity matrix. Therefore, by taking the square root transformation, for large mean, $Y^{1/2}$ is approximately distributed as a multivariate normal with an approximate mean that is the square root of the mean of Y and a known, constant variance of approximately $\frac{1}{4}I$. By taking the next term in an asymptotic expansion,

$$\mathrm{Var}(Y_i^{1/2}) \approx \frac{1}{4} + \frac{3}{32\lambda_i},$$

McCullagh and Nelder (1992) show that the variance is only approximately constant, but for large λ, the approximation holds. As Figure 3.1 shows, λ_i does not need to be too large for the constant approximation to be accurate. For example, a small mean such as $\lambda_i = 5$ yields a variance of 0.27. Through further approximation of the mean, the model becomes a linear model where, since the variance is known, only the parameter β needs to be integrated out. As the approximation stands, analytical integration is not possible because the mean is $e^{\frac{1}{2}X\Gamma\beta}$ and the parameter β cannot be integrated out. Using a Taylor series expansion, we can approximate the mean and derive an integrable transformation for Equation (3.2). Let $\mathbf{1}_n$ be the n-dimensional vector of 1's. To find an approximation for the mean, $e^{\frac{1}{2}X\Gamma\beta}$, we use a Taylor series expansion

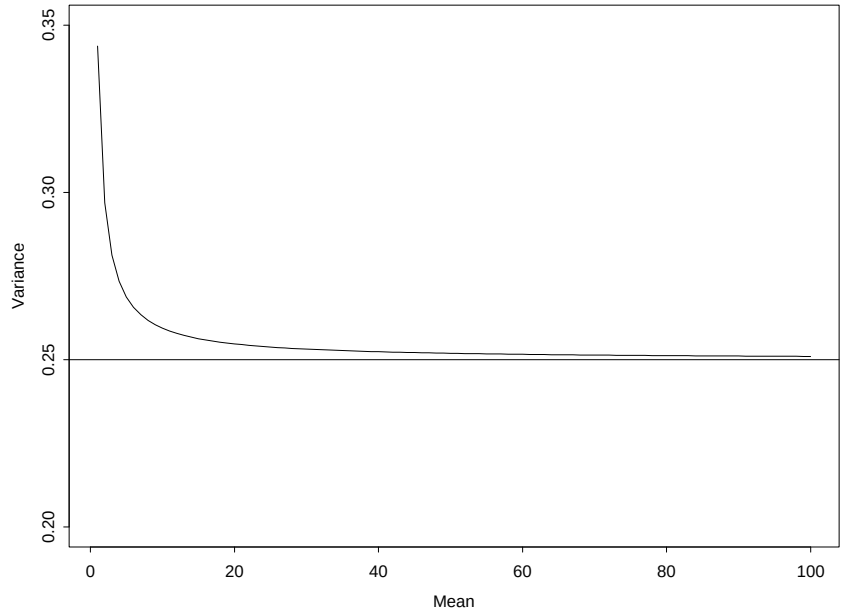


Figure 3.1: Variance vs. Mean. Note that the variance is only approximately constant.

around $\ln \bar{Y}$ where $\bar{Y}$ is the mean of the observations, leading to:

$$Y^{1/2} \dot{\sim} \mathrm{N}_n\left(e^{\ln \bar{Y}^{1/2}}\left(\mathbf{1}_n + \frac{1}{2}X\Gamma\beta - \frac{1}{2}\ln \bar{Y} \cdot \mathbf{1}_n\right), \frac{1}{4}I\right).$$

An additional transformation of Y, which yields $X\Gamma\beta$ as the mean of the normal distribution, is

$$W = 2\bar{Y}^{-1/2}\left(Y^{1/2} - \bar{Y}^{1/2} \cdot \mathbf{1}_n + \frac{1}{2}\bar{Y}^{1/2}\ln \bar{Y} \cdot \mathbf{1}_n\right) \dot{\sim} \mathrm{N}_n\left(X\Gamma\beta, \frac{1}{\bar{Y}}I\right). \quad (3.3)$$

This is useful because we can now approximate the integration in Equation (3.2). We expanded around $\ln \bar{Y}$; however, there are other possibilities. (See Section 3.3 for more details.)

Since W is a one-to-one function of Y, $\pi(\gamma|Y) = \pi(\gamma|W)$. Therefore, to calculate the model probabilities, we now need to derive $P(W|\gamma)$. The following

definitions will be useful: Let $\Sigma^{1/2}$ be a diagonal matrix with elements σ_i ($\Sigma = \Sigma^{1/2}\Sigma^{1/2}$) and

$$\tilde{X} = \begin{bmatrix} \bar{Y}^{1/2} X\Gamma \\ \Sigma^{-1/2} \end{bmatrix} \quad \text{and} \quad \tilde{W} = \begin{bmatrix} \bar{Y}^{1/2} W \\ 0 \end{bmatrix}.$$

$\tilde{X}$ and $\tilde{W}$ are the so-called augmented data, commonly used to perform Bayesian analysis of linear models using least squares techniques. Also let

$$\hat{\beta} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{W}.$$

Note that $\tilde{X}$ and $\tilde{W}$ are used only in the intermediate steps of the derivation, but not in the final approximation. They are transformed back into X and W in the last steps of the derivation. Since the columns of X are orthogonal, $X'X$ is a diagonal matrix. Let d_i be the i^{th} diagonal entry of $X'X$. Also use the notation that x_i is the i^{th} column of the matrix X. Finally, use the fact from the normal equations that $\tilde{X}(\tilde{W} - \tilde{X}\hat{\beta}) = 0$. The first step of the derivation requires the use of the approximation from Equation (3.3):

$$\begin{aligned}
P(W|\gamma) &= \int P(W|\beta,\gamma)P(\beta|\gamma)d\beta \\
&\dot{\propto} \int e^{-0.5(W-X\Gamma\beta)'\bar{Y}I(W-X\Gamma\beta)} \cdot e^{-0.5\beta'\Sigma^{-1}\beta} d\beta \\
&= \int e^{-0.5(\bar{Y}W'W - 2\bar{Y}W'X\Gamma\beta + \bar{Y}\beta'\Gamma X'X\Gamma\beta + \beta'\Sigma^{-1}\beta)} d\beta \\
&= \int e^{-0.5(\tilde{W}'\tilde{W} - 2\tilde{W}'\tilde{X}\beta + \beta'\tilde{X}'\tilde{X}\beta)} d\beta \\
&= \int e^{-0.5||\tilde{W}-\tilde{X}\beta||^2} d\beta \\
&= \int e^{-0.5||\tilde{W}-\tilde{X}\hat{\beta}+\tilde{X}\hat{\beta}-\tilde{X}\beta||^2} d\beta \\
&= \int e^{-0.5[(\tilde{W}-\tilde{X}\hat{\beta})'(\tilde{W}-\tilde{X}\hat{\beta}) + 2(\tilde{W}-\tilde{X}\hat{\beta})'(\tilde{X}\hat{\beta}-\tilde{X}\beta) + (\tilde{X}\hat{\beta}-\tilde{X}\beta)'(\tilde{X}\hat{\beta}-\tilde{X}\beta)]} d\beta
\end{aligned}$$

$$
\begin{aligned}
&= \int e^{-0.5[||\tilde{W}-\tilde{X}\hat{\beta}||^2+2(\hat{\beta}-\beta)\tilde{X}(\tilde{W}-\tilde{X}\hat{\beta})+||\tilde{X}\hat{\beta}-\tilde{X}\beta||^2]}d\beta \\
&= e^{-0.5||\tilde{W}-\tilde{X}\hat{\beta}||^2}\int e^{-0.5(\beta-\hat{\beta})'\tilde{X}'\tilde{X}(\beta-\hat{\beta})}d\beta \\
&= e^{-0.5(\tilde{W}'\tilde{W}-2\tilde{W}'\tilde{X}\hat{\beta}+\hat{\beta}'\tilde{X}'\tilde{X}\hat{\beta})}|2\pi(\tilde{X}'\tilde{X})^{-1}|^{1/2} \\
&\propto e^{0.5\tilde{W}'\tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{W}}|(\tilde{X}'\tilde{X})^{-1}|^{1/2} \\
&= e^{0.5\bar{Y}W'X\Gamma(\bar{Y}\Gamma X'X\Gamma+\Sigma^{-1})^{-1}\bar{Y}\Gamma X'W}|(\bar{Y}\Gamma X'X\Gamma+\Sigma^{-1})^{-1}|^{1/2} \\
&= e^{0.5\bar{Y}^2\sum_{i=1}^{p}\left(\frac{(W'x_i)^2\gamma_i}{\bar{Y}d_i\gamma_i+\sigma_i^{-2}}\right)}\cdot\prod_{i=1}^{p}\left(\bar{Y}d_i\gamma_i+\sigma_i^{-2}\right)^{-1/2} \\
&\propto \prod_{i=1}^{p}\left(e^{0.5\left[\frac{\bar{Y}^2(W'x_i)^2}{\bar{Y}d_i+\sigma_i^{-2}}\right]}\right)^{\gamma_i}\cdot\left(\frac{\bar{Y}d_i+\sigma_i^{-2}}{\sigma_i^{-2}}\right)^{-\gamma_i/2} \\
&= \prod_{i=1}^{p}\left(e^{0.5\left[\frac{\bar{Y}^2(W'x_i)^2}{\bar{Y}d_i+\sigma_i^{-2}}\right]}\right)^{\gamma_i}\cdot\left(\bar{Y}d_i\sigma_i^2+1\right)^{-\gamma_i/2}.
\end{aligned}
$$

By using a prior on γ that factors as $\pi(\gamma)=\prod_{i=1}^{p}\rho_i^{\gamma_i}(1-\rho_i)^{(1-\gamma_i)}$, we can now factor the approximation of Equation (3.1) as follows:

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

where

$$p_i=\frac{\rho_i\exp\left\{\frac{1}{2}\frac{\bar{Y}^2(W'x_i)^2}{\bar{Y}d_i+\sigma_i^{-2}}\right\}\cdot\left(\bar{Y}d_i\sigma_i^2+1\right)^{-1/2}}{1-\rho_i+\rho_i\exp\left\{\frac{1}{2}\frac{\bar{Y}^2(W'x_i)^2}{\bar{Y}d_i+\sigma_i^{-2}}\right\}\cdot\left(\bar{Y}d_i\sigma_i^2+1\right)^{-1/2}}. \tag{3.4}$$

Generating a sample of models from $\tilde{\pi}$ can be done by generating independent Bernoulli random variables, each referring to one variable in the orthogonal basis. Therefore, p_i represents the approximate posterior probability that the i^{th} variable is included in the model. The p_i's are calculated once at the beginning, so sampling takes very little computational time compared to samplers

that need to calculate probabilities of inclusion at every step for every variable such as in the Gibbs sampler of Kuo and Mallick (1995).

3.1.2 Prediction

With $\tilde{\pi}(\gamma|Y)$ factoring as given in Equation (3.4), we can easily do approximate model selection by choosing the model with the highest approximated posterior probability. This is given by setting $\gamma_i = 1$ for $p_i > 0.5$ and $\gamma_i = 0$ for $p_i < 0.5$. However, selecting one model for prediction ignores a major component of uncertainty. That uncertainty arises from whether or not the correct model has been selected. Bayesian model mixing accounts for model uncertainty, in that predictions are based on a set of plausible models, where each model contributes to the prediction proportionally to the support it receives from the observed data through the posterior probability, $\pi(\gamma|Y)$, or the approximate posterior probability, $\tilde{\pi}(\gamma|Y)$. For some quantity of interest, Δ, such as predictive density or predictive mean,

$$\Delta = \sum_{\gamma} \Delta_{\gamma} \pi(\gamma|Y) \approx \sum_{\gamma} \Delta_{\gamma} \tilde{\pi}(\gamma|Y)$$

where Δ_{γ} is the quantity of interest evaluated for the model γ. This, again, involves summing over all possible models. In general, exact model mixing can be a challenging computational problem since it requires enumeration of the model space and, in an earlier step, integrating out the regression coefficients. Therefore, instead of enumerating all possible models, finding a representable sample of models is a proposed solution. To evaluate Δ_{γ}, the quantity of interest for the sampled models, values for the coefficient parameter may be required. Therefore sampling over the parameter space is suggested for each model.

In recent literature several computational strategies that can be used for model mixing have been proposed for finding models with high posterior probabilities in generalized linear models. For example, George, McCulloch, and Tsay (1994) have extended the stochastic search variable selection (SSVS) method (George and McCulloch 1993) to the class of generalized linear models. It is based on sampling from the model space and the parameter space using Markov chains. Another method that samples over both the parameters and model space was developed by Kuo and Mallick (1995). (See Section 3.4.2 for more detail.) Raftery, Madigan, and Volinsky (1996) proposed a method that uses a Laplace approximation to integrate out the parameters. Further work has recently been done on this method that uses the Laplace approximation to derive an approximation to the Bayes factor that determines which models have high posterior probability, and therefore, should be included in the model mixing (Volinsky et al. 1996); software written by Volinsky is available through StatLib. This approach is discussed in detail in Section 3.4.2.

3.1.3 Prior Elicitation

In specifying the prior for the regression coefficient vector β, we chose to use a prior distribution with a precision matrix proportional to $X'X$ where X is the design matrix. Therefore, the prior covariance matrix is $\Sigma = c(X'X)^{-1}$ where c is a constant. Since the columns of X are orthogonal, $X'X$ is a diagonal matrix. This prior in linear models provides the same scale as the likelihood so the β_i's can be compared to each other. Under asymptotic conditions, the same holds for generalized linear models.

We assume that the prior on the model space should be independent and

suggest using a prior on γ that factors as a product of Bernoulli's,

$$\pi(\gamma) = \prod_{i=1}^{p} \rho_i^{\gamma_i}(1-\rho_i)^{1-\gamma_i}$$

where ρ_i is the prior probability that the i^{th} variable in included in the model. Choosing $\rho_i < 0.5$ creates a penalty for each additional term in the model. For our examples we use a uniform prior on the model space by setting $\rho_i = 0.5$ for all i since we have no prior information regarding the inclusion of certain variables.

3.2 The Algorithm for Sampling

Our goal is to generate a sample from the posterior distribution $P(\beta, \gamma|Y)$. We know that

$$P(\beta, \gamma|Y) \ \propto \ \text{likelihood} * \text{prior};$$

however, we cannot easily evaluate the normalizing constant, $\int$ likelihood $*$ prior. Since the exact posterior cannot be calculated analytically, samples from the posterior distribution cannot be obtained directly. One alternative is to construct a Markov chain to sample from the posterior distribution.

3.2.1 Metropolis–Hastings

One particular way of implementing a Markov chain is by using the Metropolis–Hastings algorithm (Hastings 1970) as follows. Let $\theta = (\beta, \gamma)$. Let π be a density and $q(\theta, \theta^*) > 0$ be an arbitrary transition probability function. The function q, also called the proposal distribution, can be a conditional distribution, $p(\theta^*|\theta)$, where θ^* depends on θ, the value from the previous step. A proposal distribution, $q(\theta, \theta^*)$, will eventually generate a sample from $\pi(\theta) =$

$P(\beta, \gamma|Y)$; however, the relationship between the functions π and q will affect the rate of convergence. The proposal distribution should be chosen such that it converges quickly to its target distribution and is easily sampled and evaluated. Then the Metropolis–Hastings algorithm is given by:

- If the chain is currently at $\theta_n = \theta$, then generate a candidate value θ^* from $q(\theta, \theta^*)$.
- With probability

$$\alpha(\theta, \theta^*) = \min\left\{\frac{\pi(\theta^*)q(\theta^*, \theta)}{\pi(\theta)q(\theta, \theta^*)}, 1\right\}$$

  accept the candidate value and move the chain to $\theta_{n+1} = \theta^*$.
- Otherwise reject and let $\theta_{n+1} = \theta$.

The algorithm requires only that π be known up to a normalizing constant since the algorithm depends only on π through the ratio $\pi(\theta^*)/\pi(\theta)$.

Tierney (1994) suggests a number of possible choices for $q(\theta, \theta^*)$. In the original Metropolis algorithm (Metropolis et al. 1953), $q(\theta, \theta^*) = q(\theta^*, \theta)$ so that the acceptance probability simplifies to

$$\alpha(\theta, \theta^*) = \min\left\{\frac{\pi(\theta^*)}{\pi(\theta)}, 1\right\}.$$

This method tends to reject candidate steps less often than other forms. In independence chains, θ^* is drawn from a fixed density f. Let $q(\theta, \theta^*) = f(\theta^*)$, then the acceptance probability becomes

$$\alpha(\theta, \theta^*) = \min\left\{\frac{w(\theta)}{w(\theta^*)}, 1\right\}$$

where $w(\theta) = \pi(\theta)/f(\theta)$. The independence Metropolis chain is similar to importance sampling; the function w could be considered the importance weight function that would be used in importance sampling if the observations were generated from the density f.

3.2.2 Independent Proposal

In this subsection, our goal is to generate a sample from $P(\beta, \gamma|Y)$. To do this, we will use the Metropolis-Hasting algorithm with an independent proposal q. We choose q such that generating θ^* does not depend on the current θ. Let $q(\theta, \theta^*) = \tilde{\pi}(\gamma^*|Y)\tilde{P}(\beta^*|Y, \gamma^*)$. A proposal for $\tilde{\pi}(\gamma|Y)$ based on a product of Bernoulli's was illustrated in the previous section [recall Equation (3.4)]. Thus, all we need is an approximate posterior distribution of β given γ. Since

$$P(\beta|Y, \gamma) \propto P(Y|\beta, \gamma)P(\beta|\gamma)$$

and using the same normal approximation for the likelihood used in Section 3.1.1, we can approximate the posterior distribution of β given γ by

$$\beta|Y, \gamma \dot{\sim} \mathrm{N}\left(\hat{\beta}_\gamma, \hat{\Sigma}_\gamma\right),$$

where

$$\hat{\beta}_\gamma = [\bar{Y}\Gamma X'X\Gamma + \Sigma^{-1}]^{-1}\left(\bar{Y}\Gamma X'W\right),$$

and

$$\hat{\Sigma}_\gamma = [\bar{Y}\Gamma X'X\Gamma + \Sigma^{-1}]^{-1}.$$

Now that we have obtained the approximate posterior distributions for γ and for β, we can run the Metropolis–Hastings algorithm with an independent proposal. The algorithm follows:

- $\gamma^* \sim \prod_{i=1}^{p} \tilde{\pi}(\gamma_i|Y) = \prod_{i=1}^{p} \text{Bernoulli}(p_i)$
- $\beta^*|\gamma^* \sim \tilde{P}(\beta|Y,\gamma) = N(\hat{\beta}_\gamma, \hat{\Sigma}_\gamma)$
- Accept (γ^*, β^*) with probability

$$\min\left\{\frac{P(\beta^*,\gamma^*|Y)/\tilde{P}(\beta^*|Y,\gamma^*)\tilde{\pi}(\gamma^*|Y)}{P(\beta,\gamma|Y)/\tilde{P}(\beta|Y,\gamma)\tilde{\pi}(\gamma|Y)}, 1\right\}$$

  then move chain to $\theta_{n+1} = (\gamma^*, \beta^*)$

- Otherwise reject and let $\theta_{n+1} = (\gamma, \beta)$.

The model probabilities can then be estimated by the Monte Carlo frequencies. Given a sample $\gamma^1, \ldots, \gamma^M$ from the Markov chain, where M is the number of models sampled, the estimated model probabilities are

$$\pi(\gamma = \gamma'|Y) = E(I(\gamma = \gamma')) \approx \frac{\sum_{j=1}^{M} I(\gamma^j = \gamma')}{M},$$

and the predictive distribution is estimated by

$$P(Y^*|Y) \approx \frac{\sum_{j=1}^{M} P(Y^*|\beta^j, \gamma^j)}{M}.$$

3.2.3 Random Walk

Sometimes the independence proposal discussed in the previous section is not the best choice, especially if there is high correlation between coefficients. If the coefficients are highly correlated and a value for β is accepted, then by assuming independence of the β_i's, the sampler could rapidly move out of the area of the density. In contrast, a random walk wanders away from the accepted β slowly and covers more of the posterior area. Therefore, instead of generating

β^* from $\mathrm{N}\left(\hat{\beta}_\gamma, \hat{\Sigma}_\gamma\right)$, we could use a random walk generation scheme that moves from one step to another based on where the chain currently resides. We still have that the candidate model, γ^*, is generated by the product, $\prod_{i=1}^{p} \tilde{\pi}(\gamma_i|Y)$; however, now set

$$\beta^* = \beta + kZ$$

where β is the current "link" in the Markov chain; Z is a multivariate normal generated from $\mathrm{N}(0, \Sigma)$, with Σ the prior covariance matrix; and k is a constant chosen to achieve a good acceptance rate. Perhaps, using an approximation to the posterior covariance matrix is a better choice than the prior covariance matrix. However, in order to keep the sampler fast, we prefer a diagonal covariance matrix. Roberts et al. (1994) and Gelman et al. (1995) suggest that a good acceptance rate is in the range of [0.15, 0.5] and provide some theoretical justification. Then, accept (γ^*, β^*) with probability

$$\min\left\{\frac{P(\beta^*, \gamma^*|Y)/\tilde{\pi}(\gamma^*|Y)}{P(\beta, \gamma|Y)/\tilde{\pi}(\gamma|Y)}, 1\right\}.$$

The model and predictive probabilities are calculated in the same manner as in the case with the independent proposal.

3.2.4 Importance Sampling

If we are interested in computing expectations, we also have the option to use importance sampling when we cannot sample directly from $P(\beta, \gamma)$. Similar to the previous subsections, we start by generating $\theta = (\gamma, \beta)$ from $IS(\theta)$, a density which approximates $P(\beta, \gamma)$ and from which is easy to sample. In particular, let

$$IS(\theta) = \tilde{\pi}(\gamma|Y)\tilde{P}(\beta|\gamma);$$

this is the same as the independent proposal distribution for the Metropolis–Hastings algorithm. However, instead of having an accept/reject step, importance sampling uses all values sampled in a weighting scheme. Set the importance weight to be

$$w_j = \frac{P(\beta, \gamma|Y)}{IS(\theta)}.$$

Therefore, the model probabilities can be estimated using the importance weights (Fosdick 1963; Hastings 1970) and the sample of size M, $\gamma^1, \ldots, \gamma^M$, from the Markov chain so

$$\pi(\gamma = \gamma'|Y) \approx \frac{\sum_{j=1}^{M} I(\gamma^j = \gamma')w_j}{\sum_{j=1}^{M} w_j},$$

and the predictive distribution can be estimated similarly with the importance sampling weights and the sample $\gamma^1, \beta^1, \ldots, \gamma^M, \beta^M$. Hence,

$$\begin{aligned} P(Y^*|Y) &= \int_{\gamma,\beta} P(Y^*|\beta, \gamma) \cdot P(\beta, \gamma|Y) \\ &\approx \frac{\sum_{j=1}^{M} P(Y^*|\beta^j, \gamma^j)w_j}{\sum_{j=1}^{M} w_j}. \end{aligned}$$

3.3 Other Approximations Explored

3.3.1 Other Expansions

In order to achieve the result in Equation (3.3), a Taylor series expansion around $\ln \bar{Y}$ was used to approximate $\exp(\frac{1}{2}X\Gamma\beta)$. Instead of expanding around $\ln \bar{Y}$, one could try expanding around $\ln \hat{Y}$, such that $\hat{Y} = \exp(\bar{X}\hat{\beta})$, where $\hat{\beta}$ is the maximum likelihood estimator (MLE) under the full Poisson model and where $\bar{X}$ is a row vector containing the column means. Therefore, $\hat{Y} = \exp(\hat{\beta}_0)$

where $\hat{\beta}_0$ is the MLE for the intercept term under the full model. This is a satisfactory expansion which produces similar results to that of expanding around $\ln \bar{Y}$. Another expansion was explored: expanding around $\ln \hat{Y}$ such that $\hat{Y} = \exp(X\hat{\beta})$ where $\hat{\beta}$ is the MLE under the full Poisson model. This introduces a new problem—the variance is no longer constant. This, in turn, causes a problem in developing a factorization. Specifically, the determinate of $(\Gamma X'((\hat{Y})'I)X\Gamma + \Sigma^{-1})^{-1}$, produced as a result of the integration in the Equation (3.2), no longer factors. With a constant, the variance term factors out allowing the orthogonalization to create a diagonal matrix, $X'X$. A possible solution is to use another approximation for the variance. One could approximate the variance with a constant, such as using the mean variance.

3.3.2 Laplace Approximation Method

Another method was explored which does not transform the observations like the variance stabilizing transformation. The Laplace approximation (Kass, Tierney, and Kadane 1988) provides an alternative to the variance stabilizing transformation for carrying out the integration in Equation (3.2) analytically without transforming Y. This approximation is very good and has been used frequently in literature to approximate the integral in Equation (3.2) (Raftery, Madigan, and Volinsky 1996). An outline of the derivation follows: To approximate $\int \exp f(u)du$ where u is a p-dimensional vector, the Laplace method uses a Taylor series expansion of $f(u)$. Let u^* represent the value of u in which $f(u)$ is at its maximum, we then have

$$f(u) = f(u^*) + \sum_i \frac{\partial f(u^*)}{\partial u_i} u_i + \frac{1}{2} \sum_{i,j} \frac{\partial^2 f(u^*)}{\partial u_i \partial u_j} u_i u_j + \cdots$$

$$\begin{aligned} &\approx \quad f(u^*) + \nabla f(u^*)u_i + \frac{1}{2}u'Hu \qquad \text{where} \quad H \equiv \frac{\partial^2 f(u^*)}{\partial u_i \partial u_j} \\ &= \quad f(u^*) + \frac{1}{2}u'Hu. \end{aligned}$$

Since $f(u)$ attains its maximum at u^*, $\nabla f(u^*) = 0$, thereby producing the last equation. Now approximating the integral is as simple as recognizing that it is the kernel of a normal distribution.

$$\begin{aligned} \int e^{f(u)}du &\approx \int e^{f(u^*)+\frac{1}{2}uHu}du \\ &= e^{f(u^*)}\sqrt{\det 2\pi(-H)^{-1}} \\ &= e^{f(u^*)}(2\pi)^{\frac{p}{2}}|-H|^{-\frac{1}{2}}. \end{aligned} \tag{3.5}$$

Using this approximation for $P(Y|\gamma)$, we have

$$\begin{aligned} P(Y|\gamma) &= \int \exp\log[P(Y|\beta,\gamma)P(\beta|\gamma)]d\beta \\ &\approx (2\pi)^{p/2}|\Psi|^{1/2}P(Y|\tilde{\beta}_\gamma,\gamma)P(\tilde{\beta}_\gamma|\gamma) \end{aligned} \tag{3.6}$$

where $\tilde{\beta}_\gamma$ is the posterior mode of β for the model γ, and Ψ is defined to be the negative inverse Hessian of $h(\beta) = \log[P(Y|\beta,\gamma)P(\beta|\gamma)]$ evaluated at $\beta = \tilde{\beta}_\gamma$.

With the likelihood and prior postulated in this chapter, we have:

$$h(\beta) = -\sum_{j=1}^{n} e^{x_j'\Gamma\beta} + \sum_{j=1}^{n}\log(e^{x_j'\Gamma\beta})^{Y_j} - \sum_{j=1}^{n}\log Y_j! - \frac{1}{2}\beta'\Sigma^{-1}\beta - \log\sqrt{\det 2\pi\Sigma},$$

$$\frac{\partial h(\beta)}{\partial\beta_i} = -\sum_{j=1}^{n} e^{x_j'\Gamma\beta}\gamma_i x_{ji} + \sum_{j=1}^{n} y_j\gamma_i x_{ji} - \frac{1}{2}\sum_{\ell=1}^{p} 2\beta_i\Sigma^{-1}_{i\ell},$$

$$\frac{\partial^2 h(\beta)}{\partial\beta_i\partial\beta_k} = -\sum_{j=1}^{n} e^{x_j'\Gamma\beta}\gamma_i x_{ji}\gamma_k x_{jk} - \Sigma^{-1}_{ik} = H_{ik}.$$

Therefore, since $\Psi = -H^{-1}$ with $\beta = \tilde{\beta}_\gamma$,

$$\Psi = (\Gamma X' \Lambda X \Gamma + \Sigma^{-1})^{-1}$$

where Λ is the diagonal matrix with diagonal elements $\lambda_j = e^{x_j' \Gamma \tilde{\beta}_\gamma}$. To increase computational speed, we further simplify the calculations in Equation (3.6) by setting $\tilde{\beta}_\gamma$ for all γ equal to $\tilde{\beta}$, the posterior mode of β for the full model, $\gamma = \mathbf{1}_p$. Then $P(Y|\gamma)$ can be approximated by

$$\begin{aligned}
\tilde{P}(Y|\gamma) &= (2\pi)^{p/2} |\Psi|^{1/2} P(Y|\tilde{\beta}, \gamma) P(\tilde{\beta}|\gamma) \\
&= (2\pi)^{\frac{p}{2}} |\Gamma X' \Lambda X \Gamma + \Sigma^{-1}|^{-\frac{1}{2}} \cdot \frac{e^{-\lambda' \cdot \mathbf{1}_n} \cdot e^{(X\Gamma\tilde{\beta})' Y}}{\prod_{j=1}^n Y_j!} \cdot \frac{e^{-\frac{1}{2}\tilde{\beta}' \Sigma^{-1} \tilde{\beta}}}{\sqrt{\det 2\pi\Sigma}}. \quad (3.7)
\end{aligned}$$

Approximating the diagonal entries of Λ by the mean of the observations, $\bar{Y}$, makes the determinant of $\Gamma X' \Lambda X \Gamma + \Sigma^{-1}$ easy to calculate and factor. Another simplification occurs by approximating $e^{x_j' \Gamma \tilde{\beta}}$ by a truncated Taylor series expansion, $e^{\frac{1}{2}\ln \bar{Y}}(1 + x_j' \Gamma \tilde{\beta} - \frac{1}{2} \ln \bar{Y})$. Then we have:

$$\begin{aligned}
\tilde{P}(Y|\gamma) &\propto |\bar{Y} \Gamma X' X \Gamma + \Sigma^{-1}|^{-\frac{1}{2}} \cdot e^{-\sum_{j=1}^n e^{\frac{1}{2}\ln \bar{Y}}(1 + x_j' \Gamma \tilde{\beta} - \frac{1}{2}\ln \bar{Y})} \cdot e^{(X\Gamma\tilde{\beta})' Y} \\
&= \left[\prod_{i=1}^p \left(\bar{Y} d_i + \sigma_i^{-2} \right)^{-\gamma_i/2} \left(\sigma_i^{-2} \right)^{-(1-\gamma_i)/2} \right] \\
&\qquad \cdot \left[\prod_{j=1}^n e^{-\bar{Y}^{1/2} - \bar{Y}^{1/2} X_j' \Gamma \tilde{\beta} + \frac{1}{2}\bar{Y}^{1/2} \ln \bar{Y}} \right] \cdot \left[e^{\sum_{j=1}^n \left(\sum_{i=1}^p x_{ji} \gamma_i \tilde{\beta}_i \right) Y_j} \right] \\
&\propto \left[\prod_{i=1}^p \left(\frac{\bar{Y} d_i + \sigma_i^{-2}}{\sigma_i^{-2}} \right)^{-\gamma_i/2} \right] \cdot \left[\prod_{i=1}^p \prod_{j=1}^n \left(e^{-\bar{Y}^{1/2} x_{ji} \tilde{\beta}_i} \right)^{\gamma_i} \right] \\
&\qquad \cdot \left[\prod_{i=1}^p \prod_{j=1}^n \left(e^{x_{ji} \tilde{\beta}_i Y_j} \right)^{\gamma_i} \right] \\
&= \prod_{i=1}^p \left(\frac{\bar{Y} d_i + \sigma_i^{-2}}{\sigma_i^{-2}} \right)^{-\gamma_i/2} \prod_{j=1}^n \left(e^{x_{ji} \tilde{\beta}_i Y_j - \bar{Y}^{1/2} x_{ji} \tilde{\beta}_i} \right)^{\gamma_i}
\end{aligned}$$

$$= \prod_{i=1}^{p} \left(\bar{Y} d_i \sigma_i^2 + 1\right)^{-\gamma_i/2} \prod_{j=1}^{n} \left(e^{x_{ji}\tilde{\beta}_i Y_j - \bar{Y}^{1/2} x_{ji}\tilde{\beta}_i}\right)^{\gamma_i}$$

where d_i is the i^{th} diagonal entry of the diagonal matrix $X'X$. So $\tilde{\pi}(\gamma|Y)$ can be factored as follows.

$$\tilde{\pi}(\gamma|Y) = \prod_{i=1}^{p} \tilde{p}_i{}^{\gamma_i} \left(1 - \tilde{p}_i\right)^{1-\gamma_i}$$

where

$$\tilde{p}_i = \frac{\rho_i \left(\bar{Y} d_i \sigma_i^2 + 1\right)^{-1/2} \cdot \prod_{j=1}^{n} e^{x_{ji}\tilde{\beta}_i Y_j - \bar{Y}^{1/2} x_{ji}\tilde{\beta}_i}}{1 - \rho_i + \rho_i \left(\bar{Y} d_i \sigma_i^2 + 1\right)^{-1/2} \cdot \prod_{j=1}^{n} e^{x_{ji}\tilde{\beta}_i Y_j - \bar{Y}^{1/2} x_{ji}\tilde{\beta}_i}}.$$

Using the Laplace approximation produces another $\tilde{\pi}$ from which models may be sampled. Since the approximation based on $\tilde{p}_i$ is derived using the posterior mode $\tilde{\beta}$ for one particular model (the full model $\gamma = \mathbf{1}_p$), it may not be as accurate as the approximation in Equation (3.4) based on p_i which does not use such restrictive values. Analytically, $\tilde{p}_i$ and p_i only differ in the exponential term. A comparison of $\tilde{p}_i$ and p_i is performed in Subsection 3.4.3.

3.4 Comparison Example

In this section we compare the performance of our methods with other methods in the literature using a small simulated example. The data and methods are outlined in detail. We are interested in the comparison of the coverage of model space, posterior model probabilities, and predictive distributions and their means.

3.4.1 Simulated Data

The data set used for the comparisons is simulated. The design matrix is that of a 2^3 factorial design excluding the three-way interaction term. It is possible to use our method on the saturated model; however, for one of our comparison methods it is not. Therefore, the three-way interaction term is excluded. This produces $p = 7$ variables and $n = 8$ observations. A problem of

Y	**Design Matrix,** X						
	X_1	X_2	X_3	X_4	X_5	X_6	X_7
0	1	1	1	1	1	1	1
8	1	-1	1	1	-1	-1	1
11	1	1	-1	1	-1	1	-1
1	1	-1	-1	1	1	-1	-1
0	1	1	1	-1	1	-1	-1
17	1	-1	1	-1	-1	1	-1
15	1	1	-1	-1	-1	-1	1
1	1	-1	-1	-1	1	1	1

Table 3.1: Simulated data.

this size allows for enumeration of the model space for exact calculation of the posterior probabilities of the models and predictive probability distribution. To generate the data, the model $\gamma = (1, 0, 0, 1, 1, 0, 0)$ was used, and the coefficients were set to $\beta = (1.10, 0, 0, -0.40, -1.50, 0, 0)$. Thus, the response variable was generated according to the model

$$\eta = \log E(Y) = 1.10X_1 - 0.40X_4 - 1.50X_5$$

where $Y \sim$ Poisson $(\exp(\eta))$. Because the full model is over-parameterized, running "glm" in S-PLUS on this data produces the MLE $\hat{\beta}$ = (-3.95, -5.18, -5.23, -0.27, -6.46, -0.078, -0.19) which is very different from β. According to

the S-PLUS output, there is high correlation between the coefficients $\beta_1, \beta_2, \beta_3$, and β_5. Therefore, because of this correlation, we chose to use a random walk proposal with the Metropolis–Hastings algorithm.

We applied our method using a normal prior on the β coefficients, $\mathrm{N}_p(0, \Sigma)$ where $\Sigma = c(X'X)^{-1} = cI$. We chose $c = 7$, a number based on a heuristic meant for linear models (as discussed by Ed George, 1995 ASA Meeting, collaborative work with D. Foster and R. McCulloch). This heuristic provides a prior when little or no prior information is available. We used the uniform prior on model spaces, with the intercept included with probability one.

3.4.2 Algorithms for Comparison

We consider two other methods for the purpose of comparison. To test how accurately the algorithms sample the model space and make predictions, we compare the method described in Section 3.2 to the methods described in Volinsky et al. (1996) and Kuo and Mallick (1995).

1. The BIC sampler proposed by Volinsky et al. (1996) is based on an adaptation of Furnival and Wilson's (1974) leaps and bounds algorithm for linear regression and uses the Bayesian information criterion (BIC) approximation (Schwarz 1978; Raftery 1986) for the posterior model probabilities. The leaps and bounds algorithm for linear regression is often used when the number of variables is large so that computing criteria for all possible subsets of predictors is infeasible. The technique uses information computed from regressions done on previous steps to bound the possible criterion values for regressions not yet computed. Therefore, the computation of most regressions is avoided. The BIC approximation for the posterior model probabilities, $P(\gamma|Y)$, is derived by

using the Laplace method for integrals on $P(Y|\gamma) = \int P(Y|\beta,\gamma)P(\beta|\gamma)d\beta$ and by using an approximation for Bayes factors. (See Raftery 1995 and Volinsky et al. 1996 for details.) A combination of these tools are used in the approach and described in this section. Software for implementing the method is written by Volinsky as a series of S-PLUS functions and is available from StatLib.

The BIC sampling program first reduces the model space by reducing the number of variables if that number exceeds 30, the limit of variables for implementing the leaps and bounds algorithm. It accomplishes this by using backward elimination. This is a stepwise regression technique that starts with the full model and removes one variable at each step. The variable with the smallest t or F value, calculated with the current number of variables in the equation, is removed. The backward elimination stops when only 30 variables are left. The BIC program then uses an adaptation of the leaps and bounds algorithm used for linear regression. Lawless and Singhal (1978) developed a modification of the leaps and bounds algorithm for nonlinear regression models that Volinsky et al. further adapted for generalized linear models. For each model k, the algorithm provides the top q models of each model size, the MLE $\hat{\theta}_k$, the $\text{Var}(\hat{\theta}_k)$, and R^2, as well as an approximate likelihood ratio test statistic and therefore, an approximate BIC value.

The BIC sampler then applies the idea of Occam's window (Madigan and Raftery 1994). Occam's window is a set containing those models with the highest posterior probabilities. The symmetric Occam's window is the set

$$\Omega : \left\{ \gamma : \frac{\text{Max}_{\gamma'}\{P(\gamma'|Y)\}}{P(\gamma|Y)} \leq C \right\},$$

where C is a chosen constant. The authors suggest that $C = 20$ is a reasonable choice. Only those models contained in Occam's window are included in

the calculation of the predictive distribution. The program uses a symmetric Occam's window by default, but it also allows for the option to use a strict Occam's window which reduces the set by eliminating the models with better submodels. As long as the number of models, q, returned from the adapted leaps and bounds algorithm is large, the models returned will be the ones contained in Occam's window, as well as many others. This procedure substantially reduces the model space by excluding those models that are unlikely to be in Ω. From here, the returned approximated likelihood ratio test statistics can be used to reduce the remaining subset of models to those most likely to be included in Occam's window. The remaining models are then fit using the "glm" function in S-PLUS. The program then calculates the exact BIC values (Raftery 1995) for those models. The BIC value for the k^{th} model γ_k is an approximation to $2 \log B_{sk}$ where B_{sk} is the Bayes factor for the s^{th} model γ_s against model γ_k. This is

$$\text{BIC}_k = L_k^2 - \text{df}_k \log n$$

where L_k^2 is the deviance for model γ_k and df_k is the corresponding number of degrees of freedom. Volinsky et al. define the best model as the model with the largest BIC. Using the BIC values to approximate $P(\gamma|Y)$, the program eliminates the models not belonging to Ω. The models contained in Ω are the models chosen for model mixing. The posterior model probabilities are calculated by normalizing over the model set. For each model in Ω, the regression coefficients are approximated by the maximum likelihood estimator.

2. The method based on the work of Kuo and Mallick (1995) uses a MCMC algorithm that identifies promising models. Their method selects a subset of independent variables in linear regression models. Their extension to generalized

linear models is described here. The goal is to find the posterior distribution of γ using a Markov chain Monte Carlo algorithm so that subsets of predictors with high posterior probability can be identified. Starting with initial choices for $\beta^{(0)}$ and $\gamma^{(0)}$, the idea behind the method is to generate MCMC samples of $\beta^{(1)}$, $\gamma^{(1)}$, $\beta^{(2)}$, $\gamma^{(2)}, \ldots$ using conditional densities. The approximate posterior probability of a model γ can be calculated from the frequency of γ appearing in the Gibbs sampler.

Sampling β, given Y and γ, can be performed by using the Metropolis algorithm. Let the current $\beta = \beta^{(i)}$ where i is the number of iterations in the Metropolis step. Set the candidate value $\beta^* = \beta^{(i)} + cZ$ where Z is a multivariate normal generated from $\mathrm{N}_p(0, V)$. V is a current estimate of the posterior covariance matrix of β. Set c to be a scalar that achieves a desirable acceptance rate. Then

$$\beta^{(i+1)} = \begin{cases} \beta^* \text{ with probability } p^{(i)} \\ \beta^{(i)} \text{ with probability } 1 - p^{(i)} \end{cases}$$

where

$$p^{(i)} = \min\left\{1, \frac{P(Y|\beta^*, \gamma)P(\beta^*)}{P(Y|\beta^{(i)}, \gamma)P(\beta^{(i)})}\right\}.$$

This step, or mini-chain, continues until equilibrium is reached. Kuo and Mallick suggest that it is sufficient to run between 20 and 50 iterations, using $\beta^{(I)}$ from the last iteration as the sampled β. (I is the number of iterations.)

Given the prior probability, ρ_j, of γ_j they update γ_j given γ_{-j}, β, and Y using the Bernoulli distribution $\mathrm{B}(1, c_j/(c_j + d_j))$, where

$$c_j = \rho_j \cdot P(Y|\beta, \gamma_{-j}, \gamma_j = 1)$$

and

$$d_j = (1 - \rho_j) \cdot P(Y|\beta, \gamma_{-j}, \gamma_j = 0).$$

In our implementation, we set V to be a diagonal estimate of the posterior covariance matrix of β (the diagonal entries of a second order approximation) for the full model, $\gamma = \mathbf{1}_p$. By not updating at every step, we increase computational speed. Using an estimate of the posterior covariance, $V = (\Gamma X' \Lambda X \Gamma + \Sigma^{-1})^{-1}$, at each iteration may improve convergence at increased computational expense. From this point, we will refer to this adapted Kuo and Mallick method as the Gibbs/Metropolis–Hastings (GMH) sampler. This approach was chosen for comparison because it incorporates models through the likelihood, like our approach, rather than through the prior. (See George, McCulloch, and Tsay 1994 for a similar method where the models are introduced through the prior.)

3.4.3 Comparisons

Since the size of the simulated problem is small, the posterior probabilities for each model can be calculated in a reasonable amount of time using Monte Carlo integration (for details see Tanner 1993). In this section four algorithms are compared. They include the two samplers of Section 3.4.2 and the two samplers introduced in Section 3.2, the independent Metropolis–Hastings using a random walk proposal (independent because the models are generated independently) and the importance samplers. The independent Metropolis–Hastings, importance, and Gibbs/Metropolis–Hastings samplers were programmed in C and were run for 5,000 iterations. Software, written in S-PLUS by Volinsky and available through StatLib, was used for implementing the BIC sampler.

This sampling program is deterministic, and therefore, no number of iterations is chosen.

Comparison of the Posterior Probabilities of the Models Sampled. Figure 3.2 compares the models selected by each of the four algorithms. The distribution of the logarithms of the probabilities of the models discovered is shown. The population distribution, shown at the top, is bimodal and can be easily divided into good and bad models. All of the samplers exclusively focus on the good models, and the total probability mass discovered by each sampler is approximately the same. All model probabilities in the histograms are the exact posterior model probabilities calculated through enumeration and not the approximate probabilities calculated by each sampler. The number of unique models sampled and total probability mass sampled are at the top of each histogram. Although the results appear similar, the times to run these programs are very different. The combination Gibbs/Metropolis–Hastings algorithm takes 14.6 times more CPU time than either the importance (IMP) sampler or the independent Metropolis–Hastings (IMH) algorithm. This is due to the extra iterations for the random walk. The BIC sampler uses 4.2 times more CPU time than IMP or IMH. However, because of the large degree of memory swapping required by the BIC S-PLUS program, the real time necessary to execute the program is longer than any of the others. The real and CPU times for running these programs on our small simulated problem are listed in the table below. The programs were run on a Digital Unix AlphaStation 250 4/266 system with 160 megabytes of memory, a machine that runs at 5.78 SPECfp95 (a System Performance Evaluation Corporation floating point 1995 benchmark), about twice the CPU speed for floating point operations of

Method	IMH	IMP	BIC	GMH
CPU Time (seconds)	1.3	1.3	5.5	19.0
Real Time (seconds)	2.6	3.5	23.8	20.1

Table 3.2: Time comparisons.

a Pentium100 using the same benchmark.

Exact model probabilities are not usually available for generalized linear models. Therefore, we need approximations for the model probabilities which the four samplers supply. For the models sampled using the IMH and GMH samplers, we obtain Monte Carlo frequencies to approximate the model probabilities. The importance weights supply approximate model probabilities for the IMP sampler. The BIC program lists in its output the posterior probabilities approximated by the BIC values for those models sampled. Figure 3.3 shows the results of the approximations of the four approaches. The exact log posterior model probabilities are plotted against the approximated log probabilities of the models sampled by each method. As the plots show, the IMH and IMP sampler better estimate the model probabilities than the other two methods. The BIC shows a strong correlation but also a strong bias which may be due to ignoring the prior distribution.

Predictive Distribution. Using model mixing and the results from each sampling technique, the predictive distributions for each algorithm and each design point are displayed in Figure 3.4. Model averaging was used for the predictive

distribution for BIC; however, for each model the regression coefficients were estimated by the maximum likelihood estimator. The horizontal bars represent the exact predictive distribution arrived at through the enumeration of the model space. Each vertical bar represent the predictive distribution based on one of the techniques compared. Our approximations using IMH and IMP sampling appear to be very close for the design points associated with means greater than 1 and not so close for those with means near 0. The plot for the predictive distribution for design point 3 shows a clear example of where the exact predictive distribution is well approximated by the models and weights found through the IMH and IMP samplers.

Kullback–Leibler Divergence. Figures for the Kullback–Leibler divergence are also listed for each predictive distribution in Figure 3.4. The Kullback–Leibler divergence provides a summary measure of how well the exact predictive distribution is approximated by each method. To calculate the Kullback–Leibler divergence, let $p(y^*|Y, X_i')$ and $p^A(y^*|Y, X_i')$ be the exact and approximate predictive distributions at design point X_i', then the Kullback–Leibler divergence is

$$\mathrm{KL}^A(X_i') = \sum_{y^*} \log\left(\frac{p(y^*|Y, X_i')}{p^A(y^*|Y, X_i')}\right) p(y^*|Y, X_i').$$

The smaller the Kullback–Leibler divergence, the closer method A approximates the exact predictive distribution. The figures are also summarized in Table 3.3. Figure 3.5 confirms that, except for the observations at 0 and 1, the predictive distributions based on our approximations using the independent Metropolis–Hastings algorithm better approximates the exact predictive distribution for our simulated data. It may be the case that, because the

normal approximation for the Poisson is only a good approximation for large observations, our predictive distribution is less accurate for smaller values than for larger ones. Also, except for the observations at 0 and 1, the importance sampler does very well approximating the exact predictive distributions.

Predicted Means. In Table 3.4 the exact predicted means based on model mixing in the enumerated model space are compared to the results obtained through the four sampling techniques. Also tabulated are the residuals (the difference between the exact and approximated predicted means), the mean squared error (the residuals squared and averaged), and the efficiency (the smallest mean squared error divided by the others). According to our definition of efficiency, our importance sampler is less than 20% as efficient as our independent Metropolis–Hastings sampler. For example, the importance sampler needs to run 5 times as long or needs 5 times the number of observations to be as efficient as the IMH sampler. BIC is a little more efficient than IMP, and GMH is even less efficient. All three are less efficient than IMH.

For each design point the predicted means are plotted in Figure 3.6 along with their approximate 95% probability intervals. The 95% probability intervals are only approximate because the distributions are discrete. Shown is the closest interval to 95%. The solid horizontal line is the exact predicted mean, and the dotted lines are the endpoints of its approximate 95% probability interval. In parentheses, above each plot, is the interval length in percent that is closest to the 95% probability interval for the exact predicted mean. Also plotted for each design point are the predicted means from Table 3.4 for each technique, along with the probability intervals. Listed below each algorithm label is the size of the probability interval that is closest to the 95% probability

interval. That interval is also graphed. There are three instances, for design points 2, 3, and 7, where BIC, which uses point estimates, underestimates the probability intervals. Notice that, because of the vertical scale on the plots for design points 1, 4, 5, and 8, it seems that the intervals are quite different; however, there is at most a difference of three units.

Variance Stabilizing Transformation versus Laplace Approximation. As discussed in Section 3.3.2, the Laplace approximation was explored to develop a different approximation for the posterior distribution of γ. This approximation was used to propose models for our sampler. Figure 3.7 compares this approximation (LAP) to that derived using the variance stabilizing transformation (VS). Points on the line indicate identical results from both approximations. The approximations roughly follow the line; however, the Kullback–Leibler divergence in Table 3.5 shows that the approximation based on the variance stabilizing transformation is more accurate than that based on the Laplace approximation. The $\tilde{p}_i$ values based on the Laplace approximation may be less accurate than the p_i values based on the variance stabilizing transformation because they are derived using the posterior mode $\tilde{\beta}$ for one particular model (the full model $\gamma = \mathbf{1}_p$). In obtaining values for the p_i's, on the other hand, we avoid having to use an approximation based on only one model (see Subsection 3.1.1).

Discussion. IMH and IMP use the same independent proposal for sampling the γ's, however, different proposals for the β's. Because of the high correlation among the coefficients, we used a random walk proposal instead of the independence proposal in the Metropolis–Hastings algorithm. However, the importance sampler used an independence proposal with a diagonal co-

variance matrix for β, and therefore did not account for the correlation. This may be one reason why the Metropolis–Hastings algorithm does better than the importance sampler. To allow for correlation, β could be updated using a multivariate normal distribution with a full covariance matrix. This approach, however, would be more computationally involved and would take longer to run.

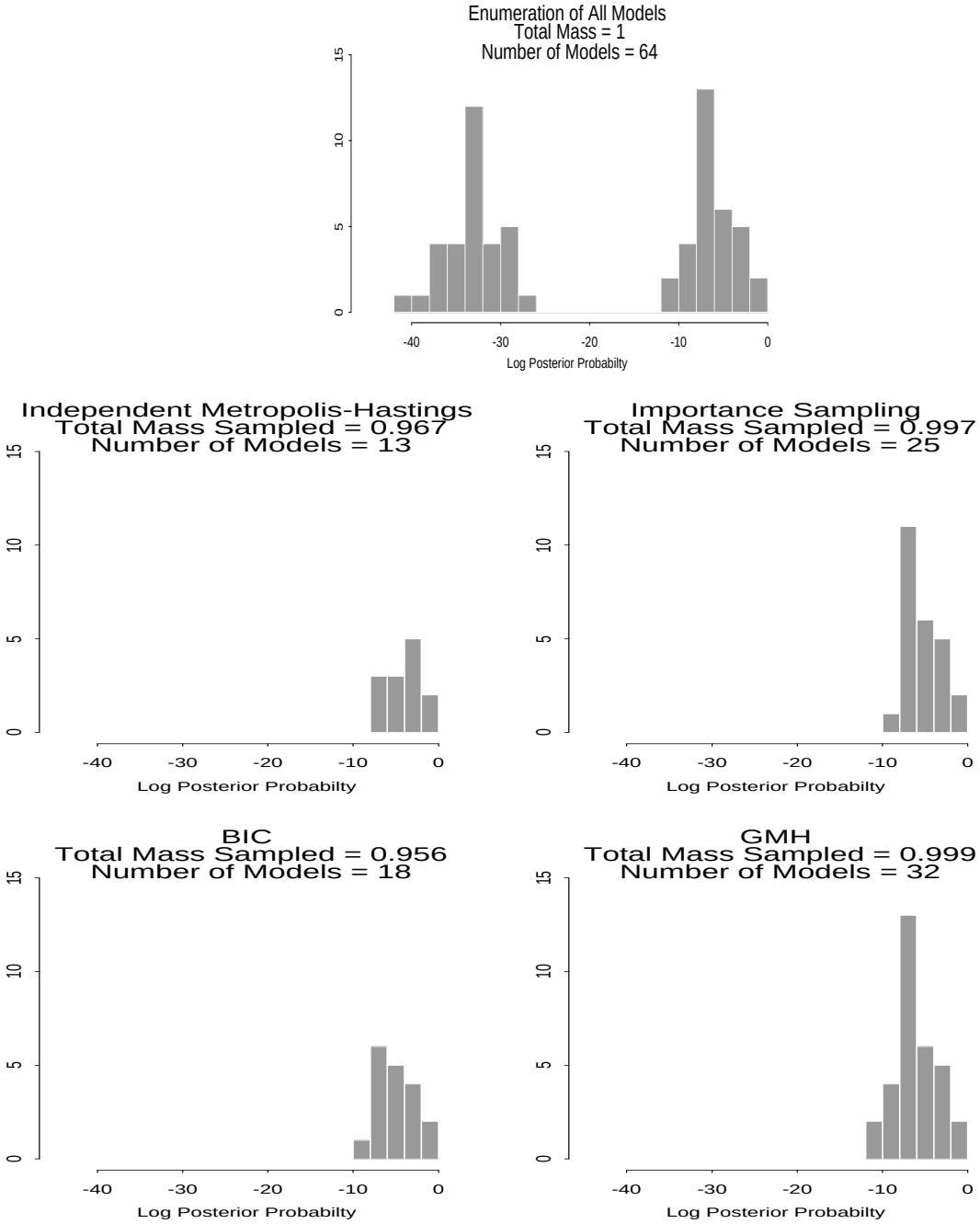


Figure 3.2: Comparison of the Algorithms: Distributions of the Logarithm of the Posterior Probabilities of the Sampled Models based on 5,000 iterations for the MCMC techniques. On the horizontal axis is the log-probability of the model; on the vertical axis is the number of models in that log-probability range. All model probabilities are the exact model probabilities calculated through enumeration and not the approximate probabilities calculated by each sampler.

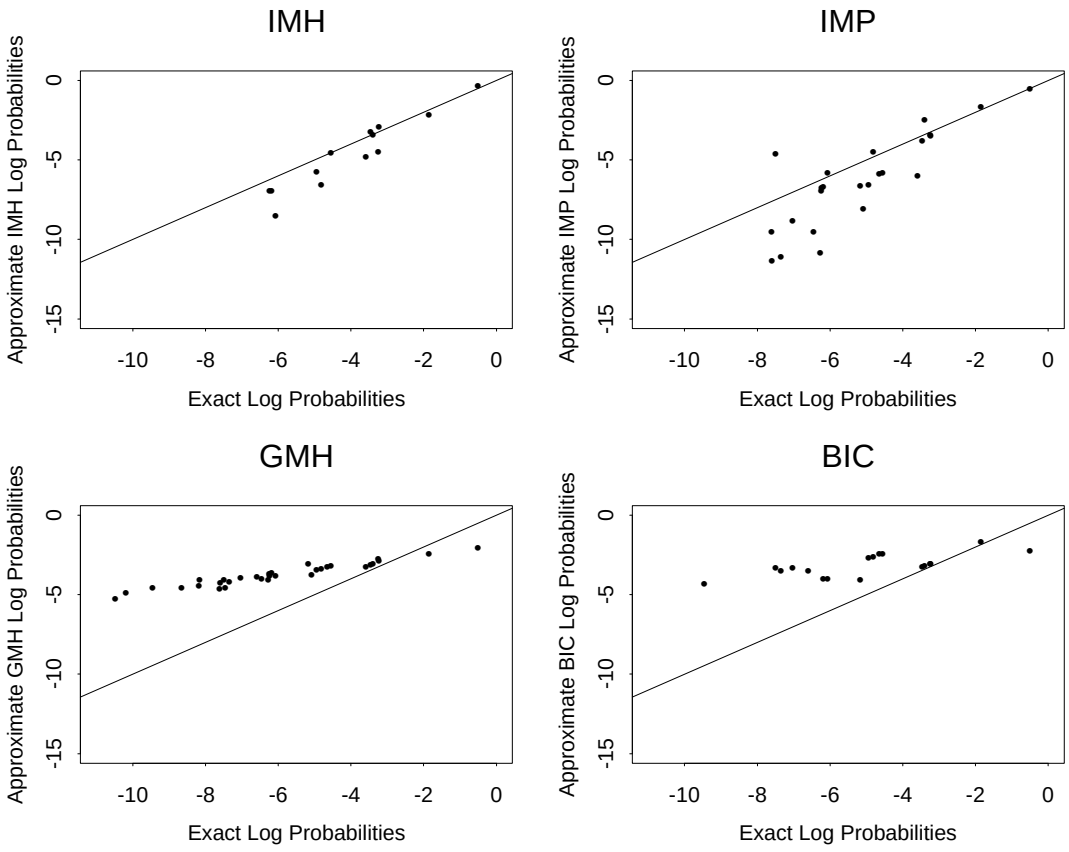
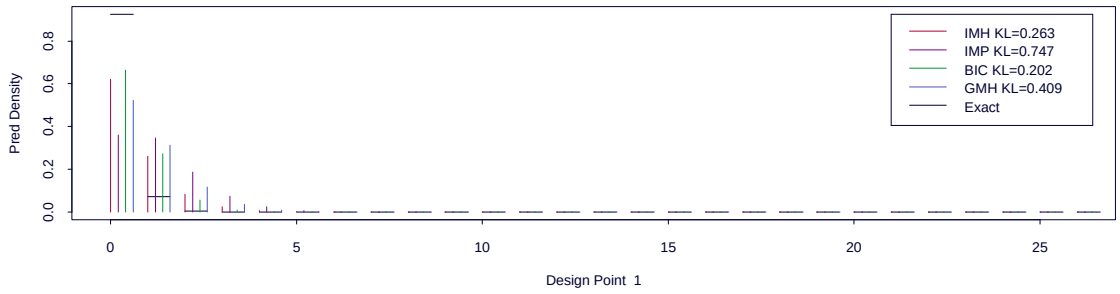
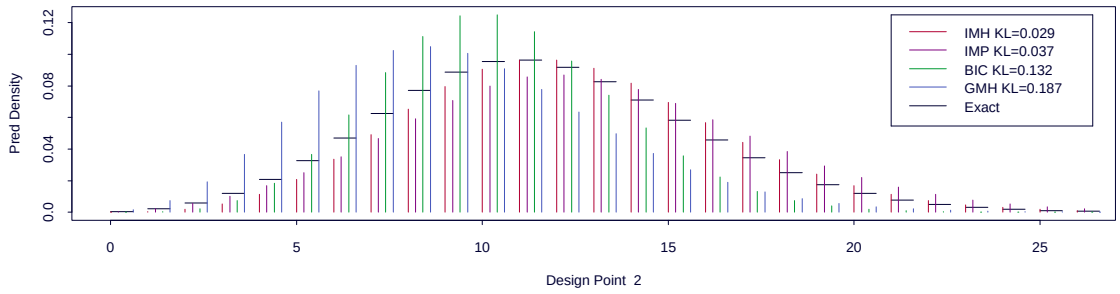
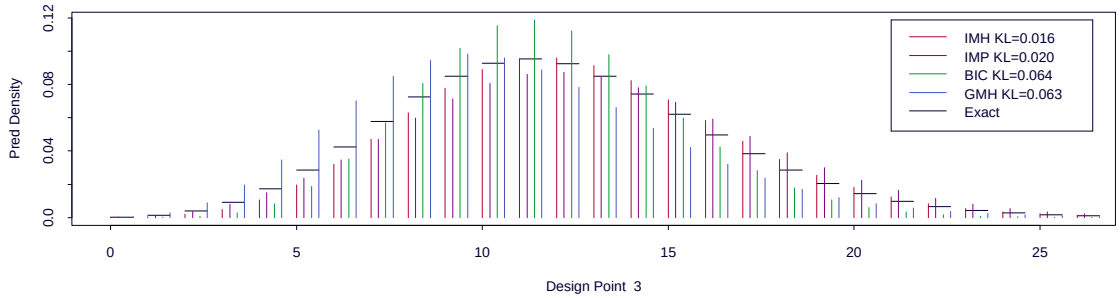


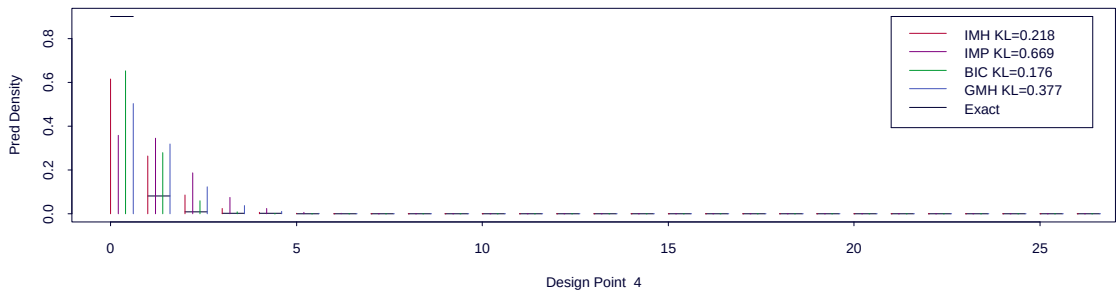
Figure 3.3: Comparison of the exact probabilities to the methods' approximations.

Prediction Distributions for each Design Point









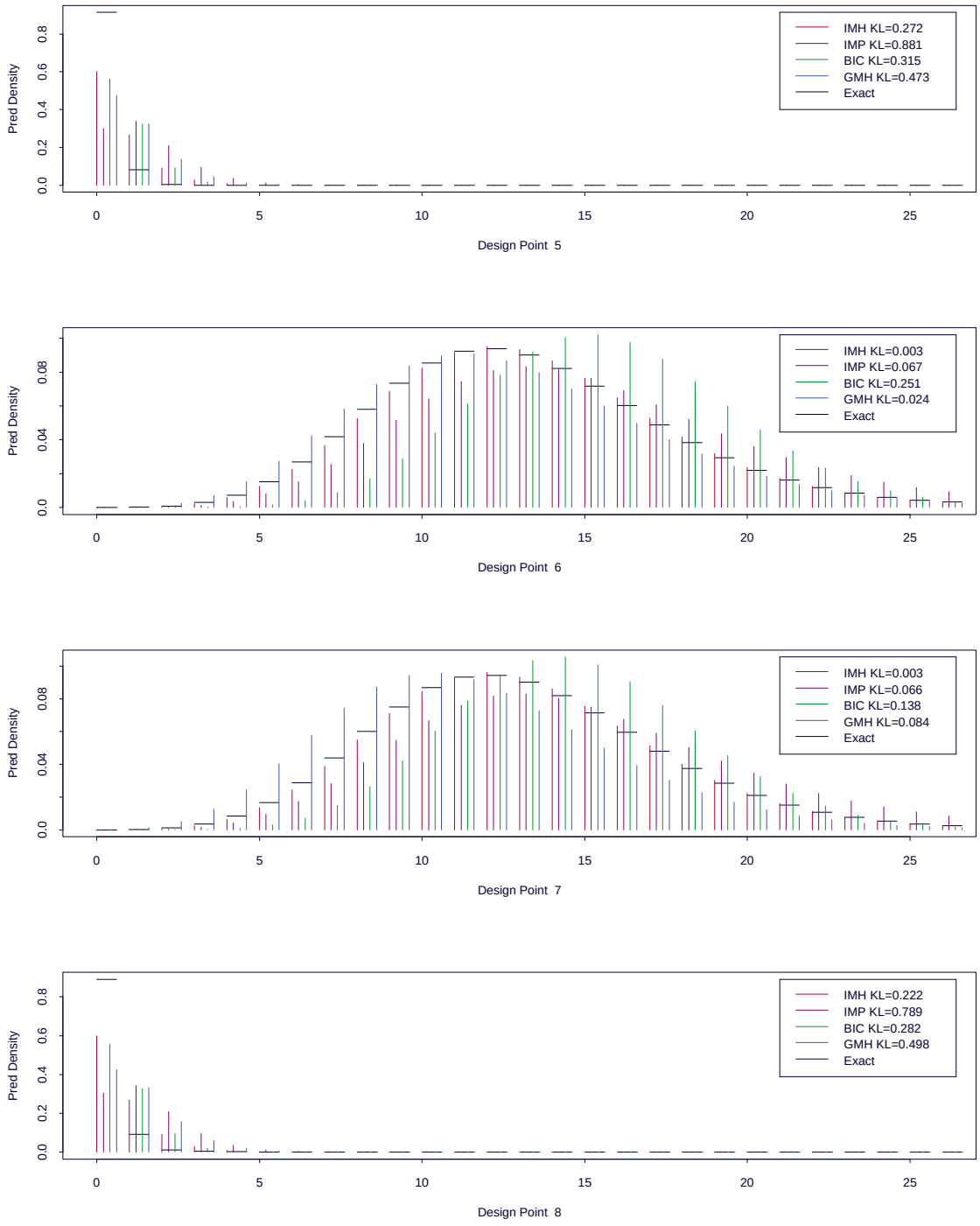


Figure 3.4: Predictive Distributions for each design point.

Method	IMH	IMP	BIC	GMH
DP 1 $Y_1 = 0$	0.263	0.747	0.202	0.409
DP 2 $Y_2 = 8$	0.029	0.037	0.132	0.187
DP 3 $Y_3 = 11$	0.016	0.020	0.064	0.063
DP 4 $Y_4 = 1$	0.218	0.669	0.176	0.377
DP 5 $Y_5 = 0$	0.272	0.881	0.315	0.473
DP 6 $Y_6 = 17$	0.003	0.067	0.251	0.024
DP 7 $Y_7 = 15$	0.003	0.066	0.138	0.084
DP 8 $Y_8 = 1$	0.222	0.789	0.282	0.498
Sum	1.025	3.27	1.560	2.115

Table 3.3: Kullback–Leibler Divergence for the predictive distribution.

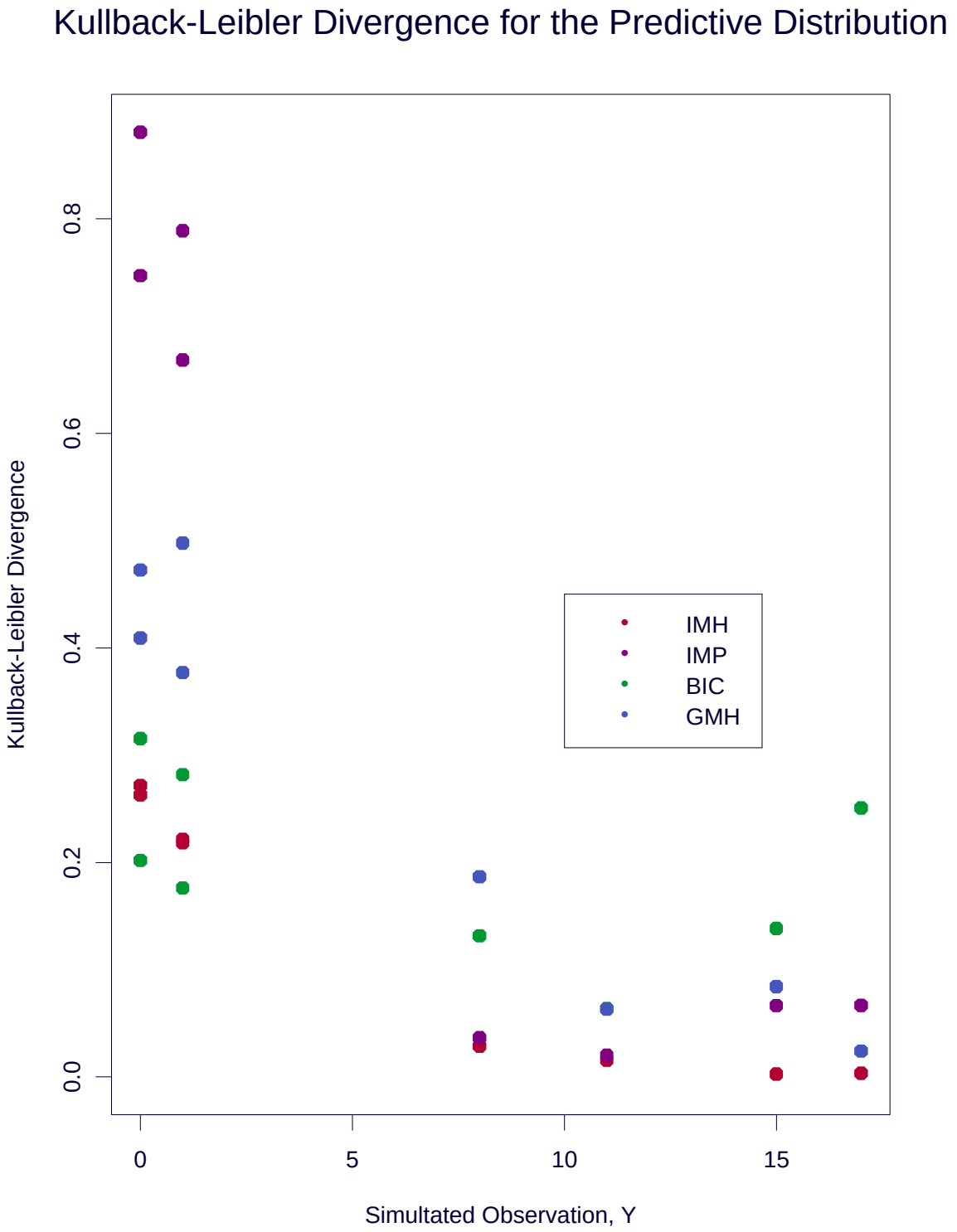


Figure 3.5: The Kullback–Leibler Divergence for the Predictive Distribution of each design point corresponding the the observation Y.

Method	Exact	IMH	IMP	BIC	GMH	Y
DP 1	0.0787	0.541 (-0.462)	1.092 (-1.013)	0.410 (-0.332)	0.704 (-0.626)	0
DP 2	11.290	12.208 (-0.918)	12.435 (-1.145)	10.058 (1.232)	8.846 (2.444)	8
DP 3	11.692	12.361 (-0.669)	12.551 (-0.859)	11.338 (0.354)	10.228 (1.464)	11
DP 4	0.123	0.554 (-0.430)	1.093 (-0.969)	0.428 (-0.304)	0.745 (-0.622)	1
DP 5	0.090	0.587 (-0.497)	1.287 (-1.197)	0.572 (-0.482)	0.821 (-0.730)	0
DP 6	12.957	13.182 (-0.224)	14.768 (-1.811)	15.287 (-2.329)	12.091 (0.867)	17
DP 7	12.743	12.992 (-0.249)	14.532 (-1.788)	14.317 (-1.574)	11.008 (1.735)	15
DP 8	0.141	0.592 (-0.451)	1.275 (-1.134)	0.589 (-0.448)	0.947 (-0.806)	1
MSE **Efficiency**		0.282 1.000	1.651 0.171	1.273 0.222	1.730 0.163	

Table 3.4: Predictive means with residuals and mean squared error.

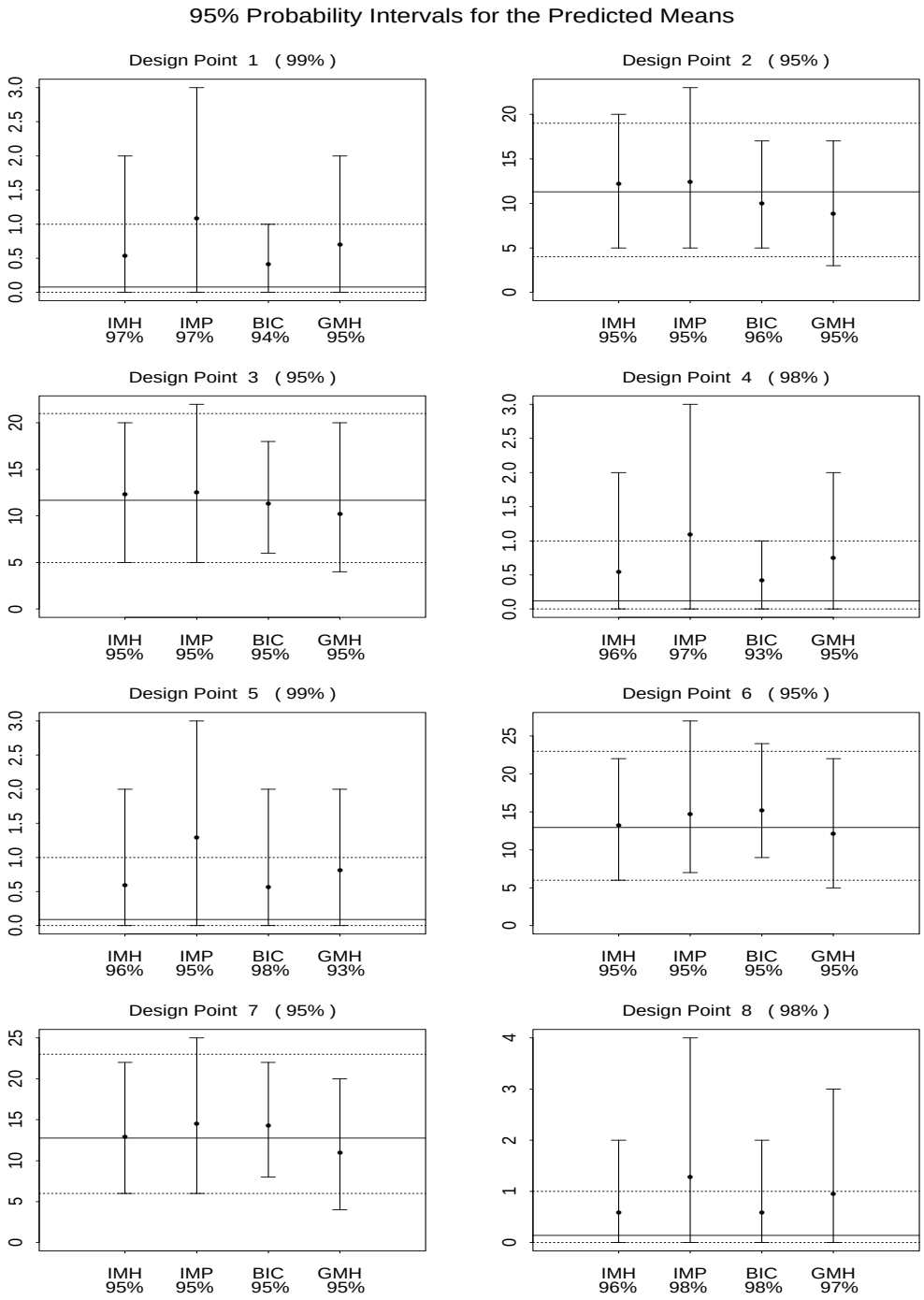


Figure 3.6: Each plot shows the predicted mean for the method with the approximate 95% probability intervals. Below each label is a more precise probability interval; the exact probability interval is in parentheses. When comparing intervals between different design points, notice the different vertical scales.

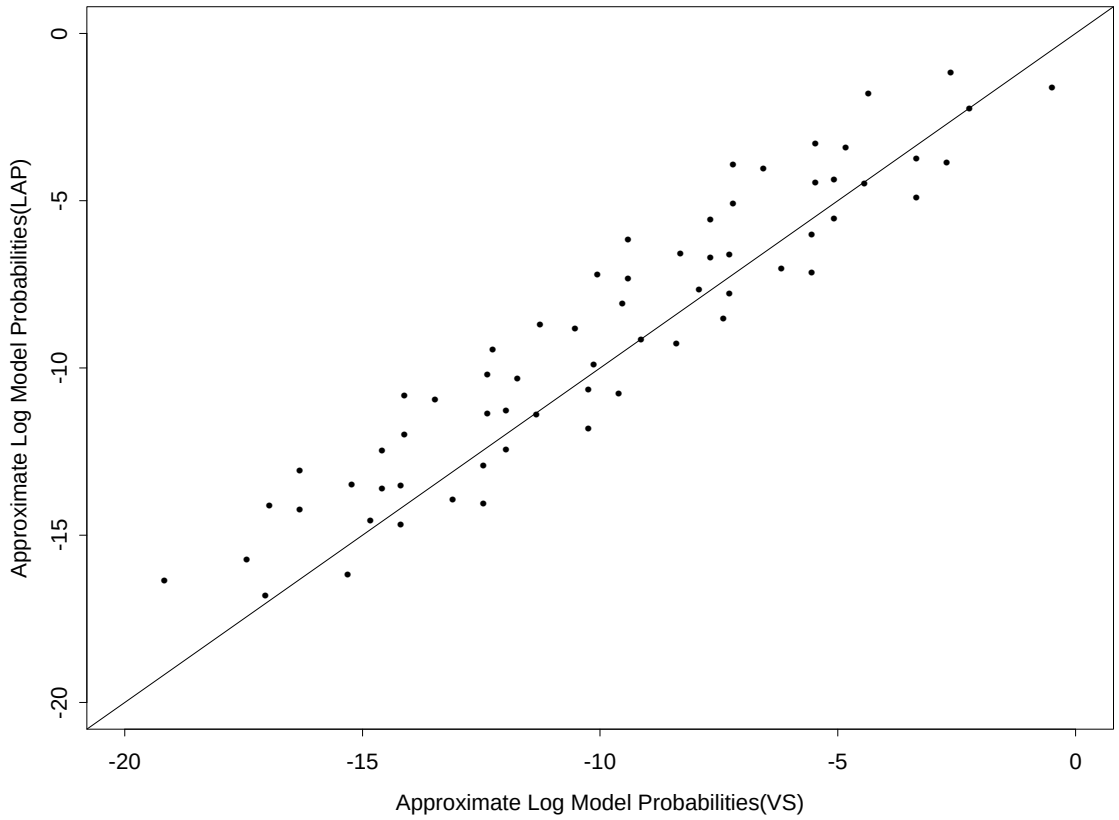


Figure 3.7: Comparison of the approximations using the variance stabilizing transformation and using the Laplace approximation.

Approximate Model Probabilities	Approximation Based on p_i (VS)	Approximation Based on $\tilde{p}_i$ (LAP)
Kullback–Leibler Divergence	0.115	0.608

Table 3.5: Kullback–Leibler Divergence to compare accuracy of the two approximations: Variance stabilizing transformation vs. Laplace approximation.

3.5 PM10 Example

We apply our approach to environmental data concerning the effect of particulate pollution on daily death counts. A front page headline in *The New York Times* on July 19, 1993, reads "Studies say soot kills up to 60,000 in U.S. each year." Schwartz (1993, 1994), a researcher at the Environmental Protection Agency (EPA), and his colleagues have produced calculations that suggest 50,000 to 60,000 deaths per year are caused by particle pollution at levels that fall within the legal limits (Stipp 1991). This number of deaths is larger than that caused by any other form of pollution. Most current efforts to remove particles from the air concentrate on catching only particles larger than 10 micrometers in diameter. Only a fraction of expenditures go to the removal of those smaller particles, called PM10 for particulate matter of size 10 micrometers or smaller. Physicians conjecture that the smaller particulates may be more dangerous because they can "penetrate further into the lung. Those below 2.5 micrometers across can reach the deepest recesses of the lungs and stick there" (Brown 1994.)

3.5.1 Background

We applied the methods of Sections 3.1 and 3.2 to data collected in Cook County (Chicago and its surrounding area), Illinois which has long records of monitoring particulate matter. The particulate measure used in this study is PM10.

Other known influences which can affect mortality, such as extreme temperatures, also need to be taken into consideration when determining the effect of particulate matter on daily death counts. The data consist of meteorological

conditions (such as temperature, specific humidity, and barometric pressure) and particulate levels. The response variable listed the daily death counts. The data set is characterized by high correlation between explanatory variables, and traditional model selection methods have difficulties both in identifying important variables and in making predictions (Styer et al. 1994; Smith 1995). Specifically, the significance of particulate matter for predicting mortality is extremely sensitive to the subset of meteorological variables included in the model, due in part to the strong correlations among the predictors. Because of this, predictions based on orthogonalized model mixing may offer a better solution than those previously tried. The response variable is death counts which lends itself naturally to a Poisson regression. Daily particulate matter, with lagged particulate matter variables and meteorological variables are considered and are highly correlated. This can have a great impact on the convergence of MCMC methods (Gelfand et al. 1996) and orthogonalizing the explanatory variables may be an improvement.

The mortality data consists of daily death counts from death certificate records for residents of Cook County for the period 1985 through 1990. The analysis was performed with total deaths from the population aged 65 or older excluding accidental causes of death as well as deaths of residents occurring outside the county.

Our analysis looked at two PM10 measures: same-day observations and three day averages. There were 20 separate monitoring stations in operation during the 1985 to 1990 period, although some of the stations only operated for a short period of time. One station had data collected on a daily basis. Not all stations measured PM10 daily; samples from the other sites were collected once every six days as required by the federal government. The same-day

observations were taken from the station taking daily observations, whereas the three-day averages were taken from all available monitoring data.

Data on meteorological variables were also considered in the study. These are based on hourly surface measures at O'Hare International Airport. The data was taken from the National Climate Data Center's National Solar and Meteorological Surface Observation Network (1961-1990) data base. It contains hourly surface observations as well as solar radiation data. We include the following meteorological variables in the design matrix X:

- the minimum and maximum *Temperature* for each day, along with the minimum and maximum for the previous two days;
- *Humidity*: 24 hour means along with one and two day lags;
- *Atmospheric pressure*: 24 hour means along with one and two day lags;
- *Wind chill* variable, using wind speed to represent the combined effect of temperature and wind.
- *Discomfort index*, function of temperature and humidity;
- *Solar radiation*, estimated from a different model and measures the amount of sunlight; and
- *Temporal* variables such as the day within the calendar year, day within each season, year, and season are also included.

Styer et al. (1994) described their analyses of the effect of PM10 on daily death counts. They used Poisson regression models along with a semi-parametric model. Through their analyses, they found potential nonlinear dependence of

mortality on meteorology. They also found that the effects of season are important. Specifically, small positive PM10 effects were found in the spring and fall, but no effect was found in the winter and summer seasons. They concluded that the "effects of particulates on mortality are not confirmed by the analyses; whether increases in PM10 cause increases in mortality remains unresolved."

The analysis discussed by Smith (1995) on "atmospheric particles and human mortality" is based on normal regression using log deaths as the dependent variable. He, too, found a strong seasonal effect, along with a strong day of the week effect. He also found the presence of serial correlations. The best method that he found used a separate 12-knot spline to model the seasonal effect for each year. This was the only model that Smith tried that removed all the serial correlation. This model produced a PM10 coefficient of 0.00036 and a standard error of 0.00021 which questions its importance as an effect through this analysis. In his conclusion, Smith admits that his results do not disprove the existence of a PM10 effect. His work does, however, show that the problem of determining it is very complicated. He concludes that there is no "clear-cut choice of the best model, yet the model selection has a crucial effect on the results obtained."

To summarize, these papers suggest that seasonal effects and nonlinear relations may be important factors. Therefore, we included second, third and fourth order polynomial terms along with terms for interaction with season. This leads to a total of 126 variables. The papers also suggest that the choice of the model used has significant effects on the results. Therefore, it is the perfect setting to use a model mixing technique to account for model uncertainty.

3.5.2 Orthogonalization

The method proposed in Section 3.1.1 assumes that the variables are orthogonal; however, the PM10 variables are not. Therefore, we need to orthogonalize the variables in such a way that the PM10 measures can be analyzed separately from the other meteorological variables. First, we center the variables; this makes all the variables orthogonal to the intercept. Then we separate the centered variables into two matrices: Meteorological = X_1 and PM10-related = X_2. To orthogonalize the meteorological variables, we let

$$Z_1 = X_1 U V^{-1/2}$$

$$\text{such that} \quad {X_1}' X_1 = UVU'.$$

Let U be a matrix containing the eigenvectors of ${X_1}' X_1$ and V be a diagonal matrix containing the corresponding eigenvalues. This yields ${Z_1}' Z_1 = I$. This orthogonalization is invariant to reordering. Then we adjust the PM10-related variables for the meteorological variables using a QR decomposition which keeps the order of the PM10 variables, allowing for separate analysis:

$$X_{2.1} = [1 - P_{Z_1}] X_2 = X_2 - Z_1 [V^{-1} {Z_1}' X_2] = QR$$

$$Z_{2.1} = X_{2.1} R^{-1} = Q$$

$$Z = [1, Z_1, Z_{2.1}].$$

Z is our new orthogonalized design matrix.

3.5.3 Results

We ran the independent Metropolis–Hastings sampler on the orthogonalized model space for 50,000 iterations after a burn-in period of 1000 iterations. The

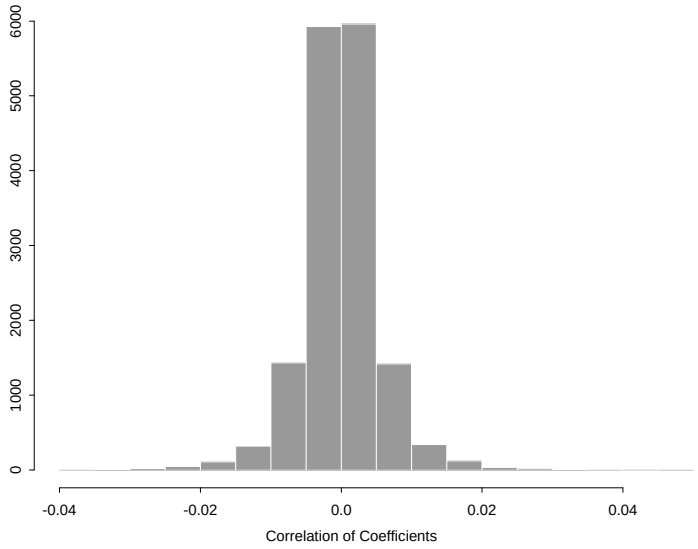


Figure 3.8: Correlation between coefficients.

independent proposal for generating candidate models and regression coefficients was used because we found very little correlation among the coefficients. See Figure 3.8 for a histogram of the correlations between coefficients. The runtime took 72 minutes in real time (71.2 minutes in CPU time) to complete and produced an acceptance rate of 0.53. A random walk for the regression coefficients was attempted; however, because of the large dimensionality of the problem, the acceptance rate was far too small to consider it a useful approach. A discussion of the results follow.

Figure 3.9 shows the 100 most often sampled models. Rows represent models and columns represent variables. Each row consists of black and/or colored boxes. A colored box signifies the inclusion of the variable represented in the corresponding column and a black box represents the exclusion of that variable. The various colors illustrate sampling frequencies of models with the most often sampled model drawn at the bottom of the plot (pale yellow) and

the least sampled models shown at the top of the plot (represented by dark purple). The last 14 columns are the PM10-related variables. Notice that there are almost as many colored boxed as there are black ones in these 14 columns, and that they are scattered about. This demonstrates the model uncertainty; and therefore, promotes the reason to use model mixing. Focusing on these 14 variables, Figure 3.10 presents boxplots of the conditional distribution of the coefficients given that its corresponding variable is included on the model; that is, the conditional distribution of the β_i given $\gamma_i = 1$. Because of the way the variables are orthogonalized in the design matrix, $Z'Z$ is the identity matrix except in the first position where the intercept cross products with itself. (In that position, the value is n.) Therefore, the variable coefficients are all on the same scale. This allows for the PM10-related coefficients to be compared in Figure 3.10. On the horizontal axis are the Monte Carlo frequencies of the variables being included. These are the approximate probabilities that $\gamma_i = 1$ for each of the 14 variables. The fourth variable from the end shows a probability of 0.58 of being included. This demonstrates the possible importance of this variable, and thereby, a possible effect of PM10 on daily death counts. To examine the effect of PM10, let

$$\begin{aligned} \lambda &= e^{Z\Gamma\beta} \\ &= e^{Z_{2.1}(\Gamma\beta)_{\mathrm{PM10}}} e^{Z_1(\Gamma\beta)_{\mathrm{Meteo}}} e^{\mathbf{1}_n \cdot (\Gamma\beta)_{\mathrm{Int}}} \\ &= \lambda_{\mathrm{PM10}} \cdot \lambda_{\mathrm{Meteo}} \cdot \lambda_{\mathrm{Int}}. \end{aligned}$$

Therefore, λ_{PM10} acts as a multiplier of the effect of PM10. That is, if $\lambda_{\mathrm{PM10}} = 1$, then there is no effect from PM10 on mortality. If $\lambda_{\mathrm{PM10}} > 1$, then there is a positive effect on death counts, and if $\lambda_{\mathrm{PM10}} < 1$, then there is a negative effect from PM10 on mortality. Figure 3.11 examines the posterior distributions of

λ_{PM10} over a 50 day period. The boxplots show as far as a 5–10% increase in deaths on three days (boxplots 20–22). We focus on November 13, 1988 (box plot 21) to examine policy making. Two possible policies were examined: thresholding the level of PM10 and reducing the PM10 level by 10%. In order to examine a thresholding policy, we took the original PM10 data and replaced any PM10 level greater than 50 micrograms per cubic meter with $50\mu g/m^3$. The EPA standard is set at 150, whereas a typical daily value in Cook County is about $40\mu g/m^3$. This thresholding implementation simulates what would happen if a law was set that limited the level of PM10 to $50\mu g/m^3$ per day, instead of the current limit of $150\mu g/m^3$. Another approach policy makers may try to implement is to find a way of reducing the level by a percentage. To examine this policy we reduced the original PM10 data by 10%. With these policies, predictions of the daily death count were compared to the prediction based on the observed data. Figure 3.12 shows the predictive distribution for each policy. There is only a small decrease in the average number of deaths with the 10% reduction; however, with the thresholded policy, there is a 1.7% change in the daily death count. The effect is difficult to detect, but the 1.7% may be important practically.

In conclusion, as demonstrated here and as noted by Styer et al. and Smith, model uncertainty dominates this problem. To account for this uncertainty, Bayesian model mixing was used in the choice of orthogonal covariates. The effect of PM10 is still difficult to detect. However, through individual examination using the marginal distributions and through policy examination using the predictive distributions, we found that the effect of PM10 may be important practically. In particular, depending on the day, PM10 can have an effect on mortality when accounting for model uncertainty.

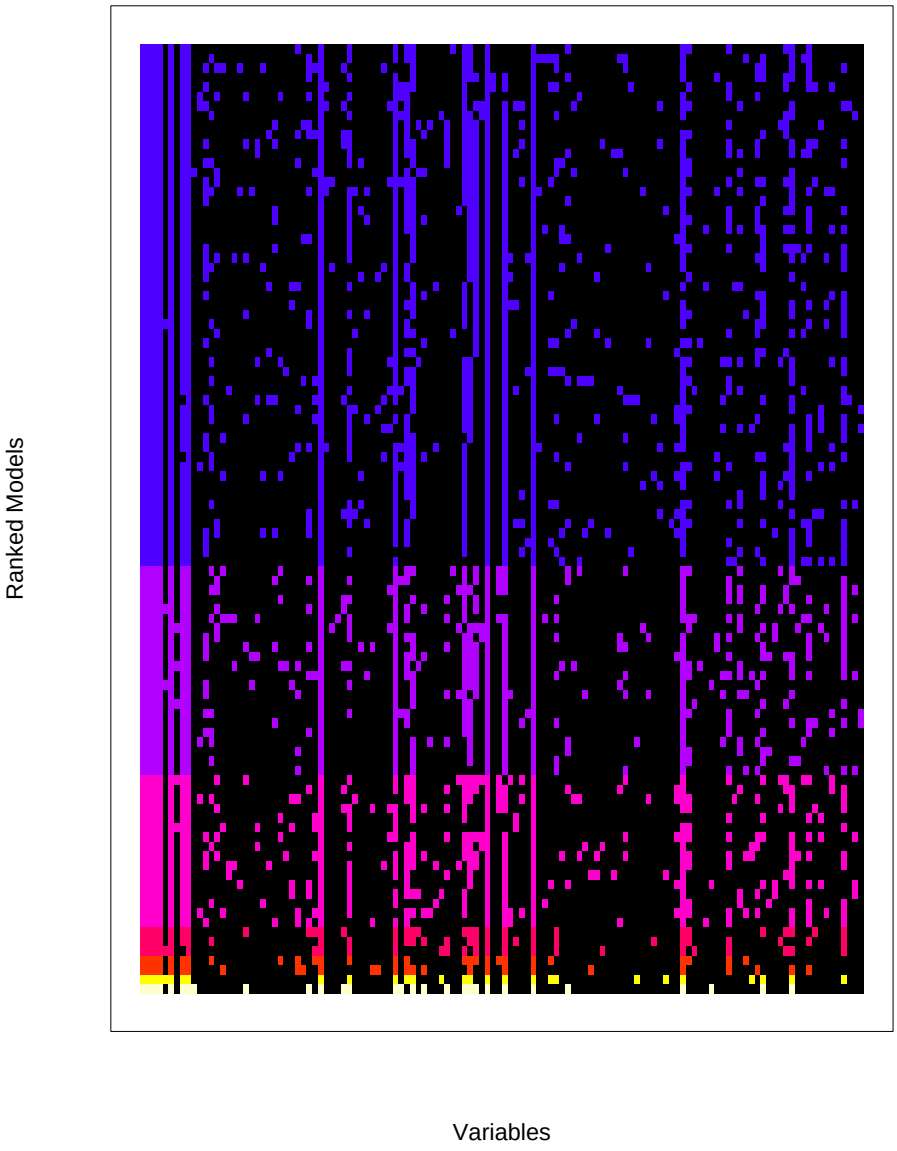


Figure 3.9: Top models sampled. Each row represents a model with the bottom row being the most frequently sampled model. A model is represented by colorful boxes for variable inclusion and black boxes for those excluded.

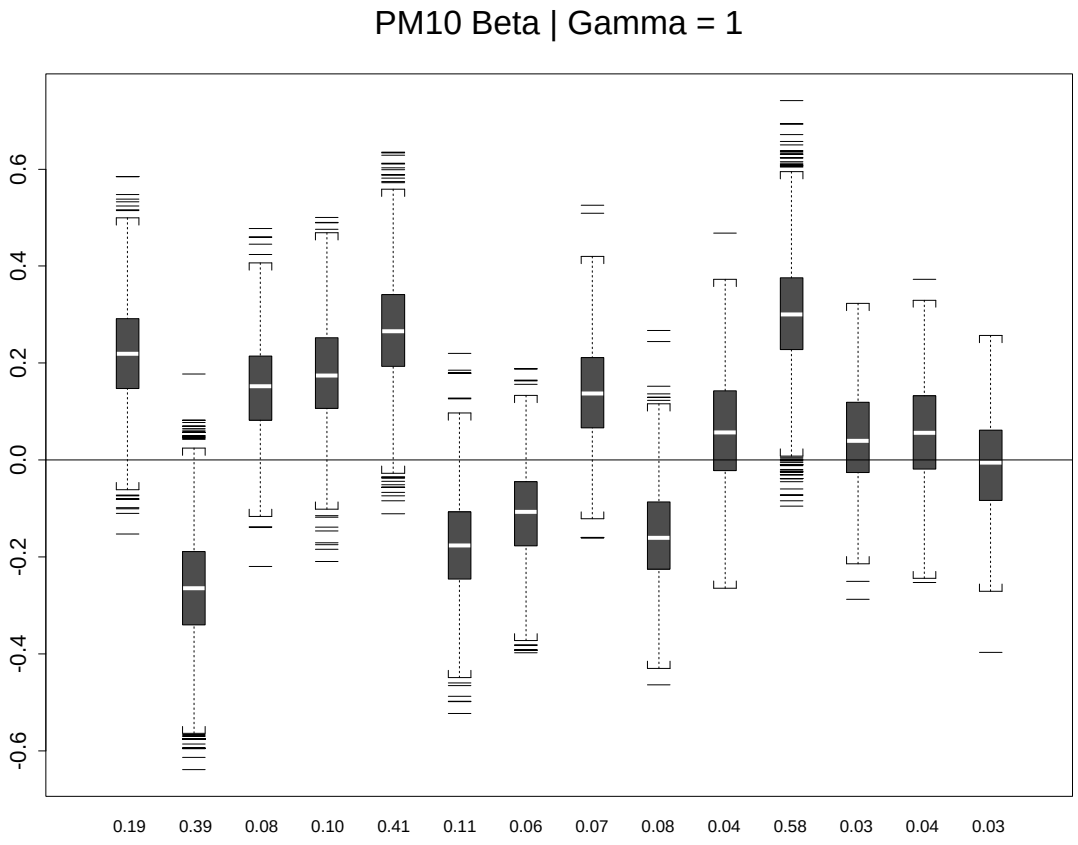


Figure 3.10: PM10 coefficients given that the variable is included. Because of the orthogonalization, the β's are in the same scale on the vertical axis and so can be compared to each other. Monte Carlo frequencies for the variable being included are on the horizontal axis.

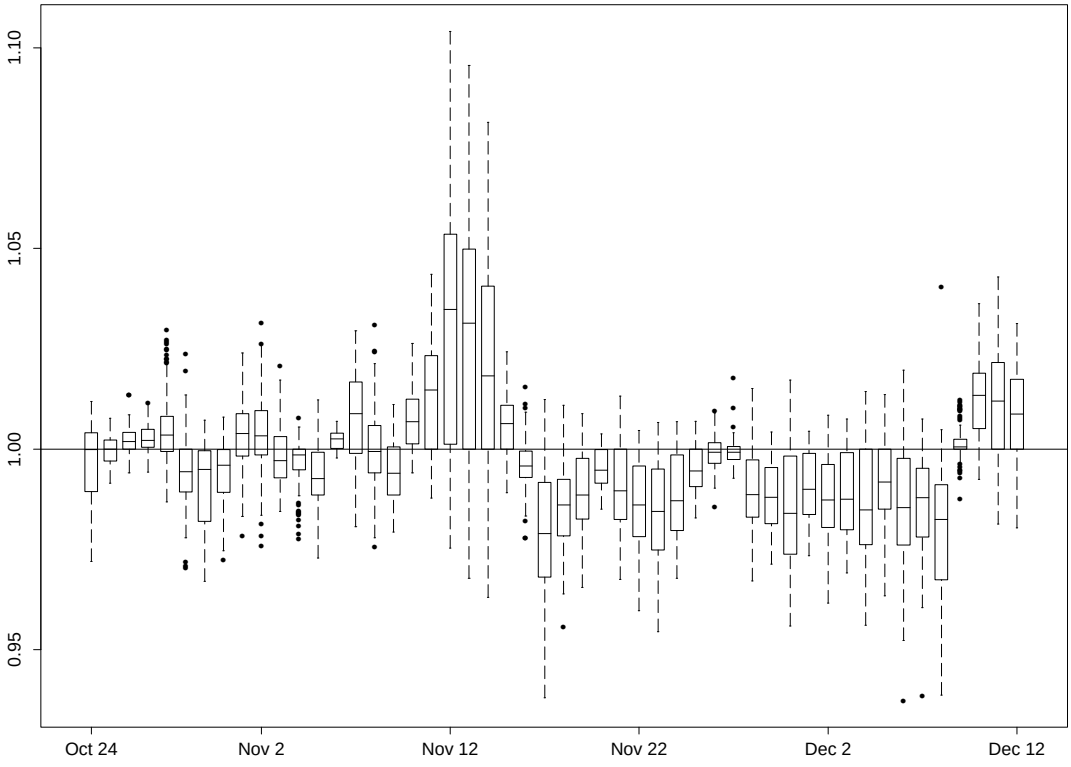


Figure 3.11: For each day over a 50 day period, posterior distributions of λ_{PM10}, the multiplier effect of PM10.

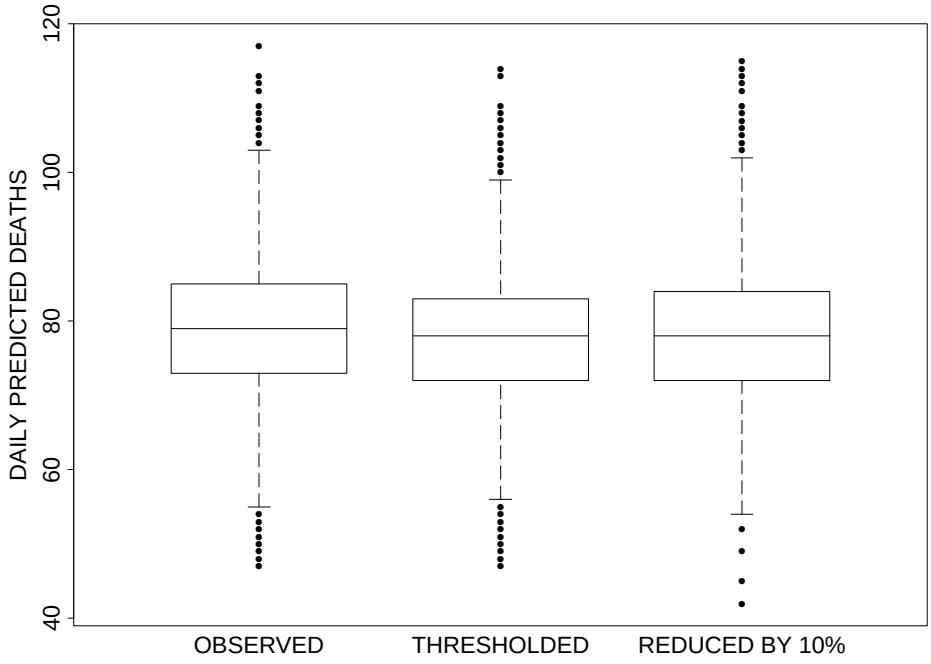


Figure 3.12: Predictive distributions of the observed PM10, the thresholded PM10, and the 10% reduction of PM10.

3.6 Conclusion

In this chapter, an approach designed for large prediction problems for implementing model mixing in Poisson regression was introduced. The implementation used an orthogonalized model space which offers statistical and computational advantages. Eliminating collinearity, and therefore, reducing the number of reasonable models, is one benefit of the orthogonalization. Another advantage due to the orthogonal candidate variables is that the posterior probabilities of models can be approximated through a factorization of predictor-specific terms. This allows for easy sampling of the model space. We showed the advantages of speed and accuracy using a small simulated example. We also demonstrated its feasibility on a large data set on environmental hazards and mortality, a problem of great difficulty and interest.

Chapter 4

Discussion

This dissertation presented modeling and computational strategies for prediction via model mixing. Our approach was based on transforming the space of predictors using an orthogonal transformation and on approximating posterior model probabilities by a product of Bernoulli random variables. We carried out this plan in both normal and Poisson regression. In the latter we applied a variance stabilizing transformation to the response variable. Under regularity conditions, this resulted in an approximately normal distribution with known constant variance. We then used a Taylor series expansion on the mean function to obtain a linear model. For this approximated problem, the normal linear results were used to approximate the posterior model probabilities for the Poisson problem. This allowed for sampling directly from an approximation to the joint distribution over the model space. In this chapter we discuss a possibility for generalizing this approach to general exponential family distributions and for improving the computational algorithm. In the final section we give some concluding remarks.

4.1 Generalizations to Other Exponential Family Distributions

Variance stabilizing transformations along with Taylor series expansions can be applied to other exponential family distributions such as the binomial and gamma distributions. As in the previous chapter, we begin by applying a variance stabilizing transformation $g(Y)$ to Y. The notation of this section is in terms of vectors and any function of a vector is defined as the function evaluated at each element of the vector, thereby producing a resultant vector. Assuming that the transformed response is approximately normal, we have

$$g(Y) \dot{\sim} \mathrm{N}(g(\lambda), cI), \tag{4.1}$$

where $\lambda = E(Y)$, $c = J_g(\lambda)\Sigma J_g'(\lambda)$ is a constant, Σ is the diagonal covariance matrix of Y, and $J_g(\lambda)$ is the Jacobian of g evaluated at λ. For example, applying the variance stabilizing transformation for the binomial distribution and assuming that the resulting response is approximately normal, we have

$$g(Y) = \arcsin\left(\frac{Y}{m}\right)^{1/2} \dot{\sim} \mathrm{N}\left(\arcsin(p^{1/2}), \frac{1}{4m}I\right).$$

It is desirable to have the mean of the approximation (4.1) equal to $X\beta$. Therefore, a Taylor series expansion may be needed to manipulate (4.1) into the form of the approximately linear model described in Chapter 3, where it is derived in detail how to use this approximation to calculate approximate posterior model probabilities that are products of predictor-specific terms.

4.2 An Alternate Implementation

One opportunity for improving the approach presented in this dissertation is to design a more efficient sampling algorithm for the regression coefficients. It is possible that $\tilde{P}(\beta|Y,\gamma) = N(\hat{\beta}_\gamma, \hat{\Sigma}_\gamma)$, where $\hat{\beta}_\gamma$ and $\hat{\Sigma}_\gamma$ are based on the least squares fit, does not approximate $P(\beta|Y,\gamma)$ sufficiently well since $\hat{\Sigma}_\gamma = [\bar{Y}\Gamma X'X\Gamma + \Sigma^{-1}]^{-1}$ is a diagonal matrix due to the orthogonality of X. Because $\hat{\Sigma}_\gamma$ implies independence of β in the proposal distribution, this proposal ignores possible coefficient correlation in the posterior. This could create a very low acceptance rate if the proposal distribution is a poor approximation to the posterior distribution using an independence chain. If the candidate for β is a better estimate for each model, then this should improve. A possible way to accomplish this is to run a mini-chain on β for each candidate model γ^* and take the values of the last iteration to be β^*. By running the mini-chain, we may find a better estimate for β^*. We propose an implementation that uses a separate accept/reject step for the models. The algorithm we propose follows.

Let $\theta_n = (\gamma^{(n)}, \beta^{(n)})$.

- $\gamma^* \sim \prod_{i=1}^{p} \tilde{\pi}(\gamma_i|Y) = \prod_{i=1}^{p} \text{Bernoulli}(p_i)$
- Accept γ^* with probability

$$\min\left\{\frac{P(\beta^{(n)},\gamma^*|Y)/\tilde{\pi}(\gamma^*|Y)}{P(\beta^{(n)},\gamma^{(n)}|Y)/\tilde{\pi}(\gamma^{(n)}|Y)}, 1\right\},$$

  then move chain to $\gamma^{(n+1)} = \gamma^*$. Otherwise reject and let $\gamma^{(n+1)} = \gamma^{(n)}$.

- Let $\beta^*_{(0)}|\gamma^{(n+1)} \sim \tilde{P}(\beta|Y,\gamma) = N(\hat{\beta}_\gamma, \tilde{\Sigma}_\gamma)$ to generate a starting value for

the mini-chain. Accept $\beta^*_{(0)}$ with probability

$$\min\left\{\frac{P(\beta^*_{(0)}|Y,\gamma^{(n+1)})/\tilde{P}(\beta^*_{(0)}|Y,\gamma^{(n+1)})}{P(\beta^{(n)}|Y,\gamma^{(n+1)})/\tilde{P}(\beta^{(n)}|Y,\gamma^{(n+1)})},1\right\};$$

else reject and set $\beta^*_{(0)} = \beta^{(n)}$. Choices for $\tilde{\Sigma}_\gamma$) can be based on on the least squares fit or on a first order approximation to the posterior covariance in the generalized linear model.

- $\beta^*_{(k)} \sim N(\beta_{(k-1)}, c\tilde{\Sigma})$ where c is a constant chosen for a good acceptance rate and $\tilde{\Sigma}$ is a covariance matrix that may be chosen from a variety of possibilities such as the posterior covariance, based on the least squares fit or a generalized linear models first order approximation, or the prior covariance. If $\tilde{\Sigma}$ is not diagonal, this will slow down the sampling algorithm due to the additional matrix calculations which are not necessary for diagonal matrices.

  Accept $\beta^*_{(k)}$ with probability

$$\min\left\{\frac{P(\beta^*_{(k)}|Y,\gamma^{(n+1)})}{P(\beta_{(k-1)}|Y,\gamma^{(n+1)})},1\right\}$$

  and set $\beta_{(k)} = \beta^*_{(k)}$; else reject and set $\beta_{(k)} = \beta_{(k-1)}$. Repeat this mini-chain for $k = 1,\ldots,K$ for some predetermined K.

- Set $\beta^{(n+1)} = \beta_{(K)}$.

Before running the mini-chain, we use an independent proposal to find a good starting place for the random walk. This step allows the random walk to start at a new value based on the current model, or if that value is rejected, continue with the $\beta^{(n)}$ from the previous step. If $\tilde{\Sigma}_\gamma$ is not diagonal, this

proposal is expensive to calculate, in which case we can use our proposal to "restart" the chain periodically rather than at each iteration.

Although this algorithm may take longer to implement, prediction accuracy may be improved by finding better estimates for β. This approach should be better than the Gibbs/Metropolis–Hastings (GMH) algorithm described in Chapter 3 because of the independent proposal that allows for a new starting value. The independent proposal for the model should also make this implementation more efficient than the Gibbs sampling in GMH.

4.3 Conclusion

An approach and algorithms for implementing model mixing in normal linear regression and generalized linear models are proposed in this dissertation. The methods are designed for making predictions in problems with a large number of correlated predictors. Through Bayesian model mixing, we can make predictions while accounting for the uncertainty in the selection of predictor variables. 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 assessments, as well as ways of incorporating information from all predictors without overfitting the data. Our approach is based on expressing the model space in terms of an orthogonalization of the design matrix. The orthogonalization eliminates correlations, and thus may reduce the number of high probability models allowing for a sampler to concentrate on the good models. This offers a simpler and faster way to sample models. A drawback of orthogonalization concerns prior elicitation: it may be more difficult to elicit the prior probability distribution

over the orthogonalized model space than over the original space.

For normal linear regression and Poisson regression, we approximated 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 were then used in sampling schemes to collect a sample of models from the orthogonalized model space.

For the linear model case, we compared orthogonalized model mixing with alternatives based on Markov chain approaches on the crime data of Vandaele (1978). We illustrated the fact that orthogonalized model mixing with importance sampling is faster in sampling models and that it tends to focus on models with high posterior probability. In the Poisson regression case, we used a small simulated data set to compare orthogonalized model mixing with a Markov chain approach and a deterministic approach. We demonstrated its speed in sampling and accuracy in prediction. We found if there is high correlation among coefficients, it may be better to use a random walk proposal over an independent proposal. However, a random walk for high dimensions may be very slow to converge. In this case, for problems with high coefficient correlation and high dimensions, an independent proposal using a full covariance matrix for β may be more appropriate. This, however, is computationally slower than using a diagonal covariance matrix. To demonstrate the practicality of the methods introduced in this dissertation, two very large data sets were analyzed. The first set originated from an experiment designed to predict protein activity under various storage conditions; the second data set came from a study which examined the effect of airborne particulates on daily death counts. Analyses of both of those sets presented the efficiency with which orthogonalized model mixing handles problems with very large dimensionality—problems

very difficult to deal with in terms of the original variables.

Bibliography

Albert, J.H. and Chib, S. 1993. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* 88 (June): 669–679.

Bennett, J. and Wakefield, J. 1994. Covariate modelling in population pharmacokinetics models. TR-94-19, Department of Mathematics, Imperial College, U.K.

Brown, W. 1994. Dying from too much dust. *New Scientist* 12 March: 12–13.

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* 24, 17–36.

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. 1996. Orthogonalizations and priors for orthogonalized model mixing. In *Modelling and Prediction: Honoring Seymour Geisser* ed. Jack C. Lee, Wesley O. Johnson and Arnold Zellner, Springer-Verlag.

de Finetti, B. 1937. La Prévision, ses lois logiques, ses sources subjectives. *Annales de l'Institut Poincaré* VIII-1, pp. 1–68.

Draper, D. 1995. Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society* 56, 45–98.

Fosdick, L.D. 1963. Monte Carlo calculations on the Ising lattice. *Methods in Computational Physics* 1, 245–280.

Foster, D., George, E., and McCulloch, G. 1995Γ Talk on *Calibrating Bayesian Variable Selection Procedures.*

Furnival, G.M. and Wilson, R.W. 1974. Regression by leaps and bounds. *Technometrics* 16, 499–511.

Gamerman, D. 1994. Efficient sampling from the posterior distribution in generalized linear mixed models. Universidade Federal do Rio de Janeiro.

Garthwaite, P.H. and Dickey, J.M. 1992. Elicitation of prior distributions for variable-selection problems in regression. *The Annals of Statistics* 20 (December): 1697–1719.

Gelfand, A., Sahu, S.K., and Carlin, B.P. 1996. Efficient parametrizations for generalized linear mixed models. In *Bayesian Statistics 5*, ed. J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith. Oxford: University Press, pp. 48–74.

Gelman, A., Roberts, G.O., and Gilks, W.R. 1995. Efficient Metropolis jumping rules. In *Bayesian Statistics 5*, ed. J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith. Oxford: University Press.

George, E.I. 1986a. Minimax multiple shrinkage estimation. *The Annals of Statistics* 14 (March): 188–205.

George, E.I. 1986b. Combining minimax shrinkage estimators. *Journal of the American Statistical Association* 81 (June): 437–445.

George, E.I. and McCulloch, R. 1993. Variable selection via Gibbs sampling. *Journal of the American Statistical Association* 88 (September): 881–889.

George, E.I. and McCulloch, R. 1994. Approaches to Bayesian 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. D.A. Berry, K.M. Chaloner, J.F Geweke. New York: Wiley, pp. 339–348

Geweke, J.F. 1994. Bayesian comparison of econometric models. Working Paper 532, Federal Reserve Bank of Minneapolis.

Grenander, U. and Miller, M.I. 1994. Representations of knowlegde in complex systems. *Journal of Royal Statistical Society - Series B* 56, 549–603.

Green, P.J. 1995. Reversible jump MCMC computation and Bayesian model determination. *Biometrika* 82, 711–732.

Hastings, W.K. 1970. Monte Carlo sampling methods using Markov Chains and their applications. *Biometrika* 57, 97–109.

Higdon, D. 1994. Spatial applications of Markov chain Monte Carlo for Bayesian inference. Ph.D. Thesis, Department of Statistics, University of Washington, Seattle.

Hilts, P. 1993. Studies say soot kills up to 60,000 in U.S. each year. *The New York Times*, 19 July.

Kadane, J.B., Dickey, J.M., Winkler, R.L., Smith, W.S., and Peters, S.C. 1980. Interactive elicitation of opinion for a normal linear model. *Journal of the American Statistical Association* 75, 845–885.

Kass, R.E., and Raftery, A.E. 1995. Bayes factors. *Journal of the American Statistical Association* 90, 773–795.

Kass, R.E., Tierney, L., and Kadane, B. 1988. Asymptotics in Bayesian computation. In *Bayesian Statistics 3*, ed. J.M. Bernardo, M.H. DeGroot, D.V. Lindley, and A.F.M. Smith. Oxford: University Press, pp. 261–278.

Kuo, L. and Mallick, B. 1995. Variable selection for regression models. University of Connecticut.

Lawless, J. and Singhal, K. 1978. Efficient screening of nonnormal regression models. *Biometrics* 34, 318–327.

Leamer, E.E. 1978. *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. New York: Wiley.

Li, K-C. 1991. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association* 86 (June): 316–327.

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, Part A—Theory and Methods* 24, 2271–2292.

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 (December): 1535–1546.

Madigan, D.M. and York, J. 1995. Bayesian graphical models for discrete data. *International Statistical Review* 63, 215–232.

McCullagh, P. and Nelder, J.A. 1992. *Generalized Linear Models*. London: Chapman and Hall.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. 1953. Equation of state calclulations by fast computing machines. *Journal of Chemical Physics* 21, 1087–1092.

Menius, A.J, Rocque, W., Emptage, M.R. and Young, S.S. 1994. Space filling experimental design for determining protein construct storage conditions. In *Proceedings of the 26th Symposium on the Interface*, pp. 106–110.

Mitchell, T.J. and Beauchamp, J.J. 1986. Algorithms for Bayesian variable selection in regression. In *Computer Science and Statistics: Proceedings of the 18th Symposium on the Interface*, ed. T.J. Boardman, Washington, DC: American Statistical Association, pp. 181–182.

Mitchell, T.J. and Beauchamp, J.J. 1988. Bayesian variable selection in linear regression. *Journal of the American Statistical Association* 83, 1023–1036.

Phillips, D.B. and Smith, A.F.M. 1995. Bayesian model comparison via jump diffusions. In *Markov Chain Monte Carlo in Practice*, ed. W.R. Gilks, S.T. Richardson, and D.J. Spiegelhalter. New York: Chapman and Hall, pp. 215–239.

Raftery, A.E. 1986. Choosing models for cross-classification. *American Sociological Review* 51, 145-146.

Raftery, A.E. 1995 Bayesian model selection in social research. *Sociological Methodology*, ed. Peter V. Marsden, Cambridge, Massachusetts: Blackwells.

Raftery, A.E. 1996. Approximate Bayes factors and accounting for model uncertainty in generalizzed linear models. *Biometika* ΠΤ.

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. 1996. 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. Oxford: University Press, pp.323–350.

Raftery, A.E. and Richardson, S. 1996. Model selection for generalized linear models via GLIB: Appication to nutrition and breast cancer. Chapter 12 of *Bayesian Biostatistics*, ed. D. Berry and D. Stangl. New York: Dekker.

Rao, C.R. 1964. The use and interpretation of principal components in applied research. *Sankhya, Series A* 26, 329–358.

Roberts, G.O., Gelman, A., Gilks, W.R. 1994. Weak convergence and optimal scaling of random walk Metropolis algorithms. *Research Report 94.16*, Statistical Laboratory, University of Cambridge.

Schwartz, J. 1993. Air pollution and daily mortality in Birmingham, Alabama. *American Journal of Epidomology*, 137, No. 10: 1136–1147.

Schwartz, J. 1994. Air pollution and hospital admission for the elderly in Birmingham, Alabama. *American Journal of Epidomology*, 130, No. 6: 589–599.

Schwarz, G. 1978. Estimating the dimension of a model. *Annals of Statistics* 6, 461–464.

Smith, R.L. 1995. Model selection in environmental statistics. Technical Report 32, National Institute of Statistical Sciences, Research Triangle Park, N.C.

Stipp, D. 1991. Bad things come insmall particles. *The Wall Street Journal* 24 April: B1, Eastern edition.

Styer, P., McMillan, N., Gao, F., Davis, J., and Sacks, J. 1994. The effect of airborne particulate matter on daily death counts. Technical Report 28, National Institute of Statistical Sciences, Research Triangle Park, N.C.

Tanner, M.A. 1993. *Tools for Statistical Inference.* Springer-Verlag: New York.

Tierney, L. 1994. Markov chains for exploring posterior distributions. *The Annals of Statistics* 22 (December): 1701–1762.

Vandaele, W. 1978. Participation in illegitimate activities: Ehrlich revisited. In *Deterrence and Incapacitation*, ed. Blumstein, A. Cohen, J. and Nagin, D., pp. 270–335, Washington, DC.

Volinsky, C., Madigan, D., Raftery, A.E., and Kronmal, R. 1996. Bayesian model averaging in proportional hazard models: Assesing stroke risk. Technical Report 302, Department of Statistics, University of Washington.

Weisberg, S. 1985. *Applied Linear Regression.* (2nd ed.) New York: Wiley.

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.

Biography

Heather Denise DeSimone was born in Pittsburgh, Pennsylvania, on March 9, 1969. She received her B.S. in Mathematics from Youngstown State University in Youngstown, Ohio, in June 1991, graduating *cum laude*. She received her M.S. in Operations Research from The College of William and Mary in Williamsburg, Virginia, in December 1992, and her M.S. in Statistics from Duke University, Durham, North Carolina, in May 1995.

She has authored or co-authored the following publications: Prediction via Orthogonalized Model Mixing, *Journal of the American Statistical Association* (1996); Comment to "Accounting for Model Uncertainty in Survival Analysis Improves Predictive Performance" by A.E. Raftery, D.M. Madigan and C.T. Volinski, in *Bayesian Statistics V* (1995); A comparison of algorithms for sampling models, *Proceedings of the 1994 Joint Statistical Meetings; Section on Bayesian Statistical Science* (1994); Change Ringing: Mathematical Music, *Pi Mu Epsilon Journal* (1992).