REGRESSION MODEL SEARCH AND UNCERTAINTY WITH MANY PREDICTORS

by

Christopher Mark Hans

Institute of Statistics and Decision Sciences
Duke University

Date: _____

Approved:

_____
Dr. Mike West, Supervisor

_____
Dr. James O. Berger

_____
Dr. Merlise Clyde

_____
Dr. Sayan Mukherjee

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

2005

ABSTRACT

(Statistics)

REGRESSION MODEL SEARCH AND UNCERTAINTY WITH MANY PREDICTORS

by

Christopher Mark Hans

Institute of Statistics and Decision Sciences
Duke University

Date: _____

Approved:

_____
Dr. Mike West, Supervisor

_____
Dr. James O. Berger

_____
Dr. Merlise Clyde

_____
Dr. Sayan Mukherjee

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

2005

Abstract

In problems of variable selection and model uncertainty, as well as in multivariate structure assessment, our ability to coherently model and analyze data when faced with increasing variable dimension is challenged by questions of model structuring, theoretical specification and computation. This dissertation addresses each of these issues, primarily in the contexts of regression and prediction, and demonstrates how coherent Bayesian models can be developed and applied in problems in high dimension.

Chapter 1 sets the context for the dissertation, describing the theoretical and computational issues that arise as a result of increased variable dimension. The idea of "sparsity" in high dimensional multivariate models is introduced.

Chapter 2 introduces a novel stochastic search algorithm for exploring large regression model spaces. Contrasts are made with existing Markov chain Monte Carlo methods. A simulation study is used to validate the method, and analytic evaluation of the method's properties is described.

Chapter 3 gives an overview of regression model selection and averaging from a Bayesian perspective using the search methods described in Chapter 2. Particular prior distributions and their advantages for use in linear regression modeling with many variables are described, with emphasis on coherency and aspects of sparsity.

Chapter 4 illustrates high dimensional linear regression model search using gene expression data from a survival study in brain cancer.

Chapter 5 introduces useful results regarding the marginal likelihood under a particular probability model. A lower bound on the marginal likelihood for models of a common dimension is established and related to sparsity and Bayesian

regularization. Reasonable assumptions about the distribution of the predictor variables allow for Bayesian learning about the sparsity inducing prior parameter.

Chapter 6 contains two examples of regression modeling and prediction in high dimension outside of the context of the linear model from clinical genomics studies in breast and lung cancer.

Finally, Chapter 7 concludes the dissertation by summarizing coherent Bayesian regression modeling in high dimensions. Generalizations of the stochastic search method are described, and future work in complex high dimensional multivariate modeling is set forth.

Acknowledgements

I would like to acknowledge with gratitude those friends and colleagues at ISDS who have given me help and support over the past four years. My primary thanks go to my advisor, Mike West, for his encouragement and guidance during the development of this research and throughout my graduate school experience. Mike's infectious energy and eager accessibility, along with the support and respect he affords his students, have made Mike an ideal mentor and serve as a model as I continue in academia.

I also would like to thank Jim Berger for his support at the beginning of my time at ISDS and for giving me several research opportunities that helped lead into this work. Merlise Clyde has always been willing to engage in lengthy discussion about any stumbling block and has cheerfully pointed me in the right direction on more than one occasion.

Beatrix Jones and Adrian Dobra played an important role in the initial stages of this work and I would like to express my appreciation for those early interactions. Thanks go as well to Quanli Wang and Eric van Gyzen for their help with computational matters.

I am especially grateful to Carlos Carvalho for his friendship and role as sounding board and collaborator. Thank you to Liz Essary for her patience while I was writing, and to Leanna House for her patience in the office. I would also like to acknowledge my family, who gave me the opportunity to persue my interests.

Finally, I would like to acknowledge funding support from the W.M. Keck Foundation and from the National Science Foundation.

Contents

List of Tables

List of Figures

Chapter 1

High Dimensional Regression Modeling

The desire to stretch the limits of statistical modeling and associated inference problems has greatly increased in recent years due to a variety of factors. Perhaps the two most relevant factors are related to increases in technology that have affected (i) the way we *collect* data, and hence the type of data we collect, and (ii) the way we *analyze* data. Technological advances have allowed us to design "high throughput" experiments where (tens or hundreds of) thousands of quantities of interest can be quickly collected, stored and recalled for analysis. Rather than collecting only the information needed to address specific, predetermined questions, in many cases it has become standard to collect as much data as possible with the view that more is better, or at least not any worse; this assumes, of course, that we can actually extract meaningful information from the data, i.e. that we can separate interpretable signal from noise.

This assumption relies heavily on the second factor: the way we analyze data. Here, too, technology has pushed us along, coupled of course with methodological advances. Evermore available computing resources have enticed statisticians to work in parameter spaces of larger and larger dimension, testing the limits of both

methodology and computation.

This dissertation addresses both methodological and computational issues arising from regression modeling in high dimension: the "large p" scenario where we have many possible predictor variables. An often encountered subclass of problems is the "large p, small n" situation, where one has many more predictor variables than cases with which to work (e.g. West, 2003), as is typically the case with analysis of expensive high-throughput experiments such as gene expression microarrays. Herein, these settings are viewed within the Bayesian paradigm and addressed as model uncertainty problems. The remainder of this chapter introduces the problem in the context of the variable selection problem. The following chapters, in turn, describe the regression modeling uncertainty problem, address computation for the problem, introduce new methods of inference for the problem and provide analysis of several high dimensional datasets.

1.1 Challenges of High Dimensional Problems

Challenges in high dimensional problems are typically inferential or computational in nature, and sometimes both. Inference problems often center around constructing estimators that maintain desirable properties as the number of dimensions increases, e.g. the case of estimating the mean of a multivariate normal distribution in greater than two dimensions. The associated computational problems deal with calculating estimators, e.g. finding maximum likelihood or Bayesian estimates in combinatorial optimization situations. The following example highlights both of these issues and provides the context needed for the types of high dimensional modeling I address in this thesis.

1.1.1 Example: Variable Selection

Building predictive models often requires the modeler to choose relevant explanatory factors from a set of possible predictor variables. When one particular set of variables is identified as providing adequate predictive ability and all future inference is based on this model, the process of identifying the set of variables is sometimes referred to as "variable selection." An alternative approach, termed "model averaging", regards the true predictive model as a mixture, where each component consists of a model that uses some subset of the predictor variables. Under this approach, the choice of variables is not explicit, as all the possible predictors can be incorporated in the model; rather the choice is implicit in the sense that the weights for the mixture components must be computed; these weights are themselves typically functions of subsets of the predictors. A good review of both topics can be found in Clyde and George (2004) and the references therein; I discuss Bayesian model averaging for linear regression models in Section 3.

Both model selection and averaging share the often problematic requirement that – when faced with many candidate predictor variables – a search of the model space must be performed to either identify the models (subsets of variables) to be selected or to compute the weights needed for averaging. In small dimensional problems, variants on common forward-backward stepwise variable selection strategies can often quickly find "good" models, although they are prone to becoming "stuck" in local maxima and often do not provide an adequate representation of the model space when there is collinearity in the data, an increasingly common occurence in high dimensional problems. The "leaps and bounds" algorithm of Furnvial and Wilson (1974), another traditional search method, does not suffer from the same local mode problem, but instead can take so long to identify

the best models when p is large that the approach becomes less useful. These methods also lack explicit measures of model uncertainty, as inference typically focuses on only one model (although there has been work on post selection uncertainty from a frequentist perspective, see e.g. Pötscher (1991); Kabaila (1995); Hjort and Claeskens (2003); Leeb (2005)). From a Bayesian perspective, there are many Markov chain Monte Carlo (MCMC) algorithms designed to explore the posterior distribution of a regression model space (e.g., George and McCulloch, 1993, 1997; Green, 1995; Madigan and York, 1995; Geweke, 1996; Raftery *et al.*, 1997), using Gibbs sampling (Gelfand and Smith, 1990) or some variant of the Metropolis-Hastings algorithm. When the number of possible predictors is not large, these methods are able to effectively identify the models that have the most predictive power and, simultaneously, offer summaries of the importance of each individual variable.

When the possible number of predictor variables is large, the above methods tend to be ineffective at providing a comprehensive summary of the model space: stepwise algorithms become entrenched in local maxima with the possibility of missing key variables due to collinearity, and MCMC approaches are slow to converge – if they do at all in a practical amount of time – due to the enormous size of the model space. In this case, characterizing regression model uncertainty is difficult as it entails performing a search on a space of models which may be so vast that only a small portion of it may be visited in a reasonable amount of time. Given this computational constraint, it is necessary to design search algorithms that can quickly identify the important predictor variables and intelligently explore the neighborhoods of models that are most relevant from a predictive perspective.

Here I consider the case of the normal linear regression model as an example of the complications raised by consideration of large numbers of candidate predictor variables. Let $Y \sim \mathrm{N}(X\boldsymbol{\beta}, \sigma^2 I_n)$, where Y is an $n \times 1$ response vector, $X = (x_1, \ldots, x_n)'$ is an $n \times p$ design matrix for the n samples, the x_i are $p \times 1$ vectors of covariate information, $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients and σ^2 is the variance of the error term. Unless otherwise specified, throughout the dissertation I consider all data to be standardized in the sense that the columns of X have been transformed to have sample mean zero and unit sample variance. The same is true for the response variable y in the case of the normal linear model, and hence an intercept term is omitted.

When p is "large", the variable selection problem amounts to choosing some subset of the p variables to include in the linear model on which inference will be based. The model averaging problem would ideally average over models consisting of all subsets of the p variables, or perhaps over all subsets up to a maximum size determined *a priori*. The term "large" is, of course, subjective; here I focus on scenarios where p is in the hundreds, thousands or tens of thousands.

Whatever the purpose of considering different subsets of models, it is convenient to have representative notation, and throughout I adopt common notation from the variable selection literature: let $\boldsymbol{\gamma}$ be a $p \times 1$ indicator vector indexing a particular model, where $\gamma_j = 1$ if variable j is included in the model and $\gamma_j = 0$ otherwise. I define the "dimension" or "size" of a model $\boldsymbol{\gamma}$ to be $|\boldsymbol{\gamma}| = \sum_{j=1}^{p} \gamma_j$, the total number of variables in the model. Throughout I assume that models are full rank.

The variable selection problem, then, amounts to determining which elements of $\boldsymbol{\gamma}$ are zero and which elements are one. This binary representation makes it

Table 1.1: Order of magnitude of the number of regression models in regression model spaces for various values of p, the total number of regressors, and K, the maximum model size allowed.

K \ p	500	1000	2500	5000	7500	10000
4	10^9	10^{10}	10^{12}	10^{13}	10^{14}	10^{14}
5	10^{11}	10^{12}	10^{14}	10^{16}	10^{17}	10^{17}
6	10^{13}	10^{15}	10^{17}	10^{19}	10^{20}	10^{21}
7	10^{15}	10^{17}	10^{20}	10^{22}	10^{23}	10^{24}

clear that the number of total possible models is 2^p, which is of course quite large even for not-so-large values of p. Even when focus is placed only on models of lower dimension, i.e. when we allow for models with only a few predictor variables, perhaps up to a maximum size K, the total possible number of models can be unwieldy from both computational and inferential perspectives. This is illustrated in Table 1.1 for various values of p and K.

It is clearly impossible, in a practical sense, to enumerate even these reduced spaces. From a model selection perspective, where the goal is to identify the "best" (set of) model(s), this obstacle is not necessarily insurmountable so long as a given model space is "well behaved" in the sense that the best models are all contained in a relatively small neighborhood that can be easily reached from any point in the space. This is rarely the case, however, as complicated datasets present complicated collinearity structure that leads to complicated model spaces. Rather, it is more realistic that pockets of "good" models are scattered throughout the space, and the computational goal is to find these pockets and quickly explore their neighborhoods.

A key aspect in modeling as we scale to higher dimensions is the idea of

sparsity. In general this can be viewed as imposing constraints that force objects of interest to lie in lower-dimensional spaces; for the linear regression problem it can be viewed as preferring small sized regression models. This coincides with the general scientific view of parsimony: if two competing models are equally well supported by the data, the less complicated model should be preferred. A similar view has been adopted in a model uncertainty context by Raftery *et al.* (1997) and called "Occam's Window". Enforcing sparsity in the case of the linear regression model, i.e., focusing attention on small sized regression models, will be seen to both simplify computation and provide good predictive performance in cases where the possibility of overfitting is evident. From a Bayesian perspective, sparsity is introduced – at least in main part – through prior distributions; such "sparsity inducing" priors are introduced in Chapter 3 and explored in more detail in Chapter 5.

1.2 Parallel Computing

As will be discussed in Chapter 2, many of the key aspects of this work involve the use of parallel (distributed, cluster) computing. All examples given below were run on the Computational Science, Engineering and Medicine (CSEM) cluster at Duke University, a collection of Intel x86 based machines running Linux. The nodes used were a subset of forty dual-processing, 3.1 GHz machines. All programs were written in C++ using MPI to coordinate communication between processing elements.

Chapter 2

Model Space Exploration

Even when the focus is on "sparse" regression models, i.e. models with only a few predictor variables, searching through model spaces derived from large datasets is very difficult. As even this reduced space cannot be enumerated, we require stochastic search methods that can capture the interesting features of the model space, Γ, while compiling a set of the best models found, Γ^*, to be used to address uncertainty in the modeling process when inference is performed.

In this chapter I review several existing MCMC methods for model space exploration, present a novel stochastic search approach designed with high dimensional datasets in mind, and provide analysis of its performance.

2.1 Markov Chain Monte Carlo Approaches

MCMC approaches for model search/variable selection can generally be classified as Gibbs sampling or Metropolis-Hastings (M-H) based algorithms, with the latter group including reversible jump based methods (Green, 1995). I refer to Gibbs algorithms as those that sample from full conditional distributions, and

M-H algorithms as those that explicitly use a non-Gibbs proposal distribution in sampling. Here I review the Gibbs sampler and M-H algorithms used for model space exploration for the normal linear model; Clyde and George (2004), Godsill (2001) and George and McCulloch (1997) provide thorough reviews.

MCMC methods simulate a Markov chain

$$\boldsymbol{\gamma}^{[1]}, \boldsymbol{\gamma}^{[2]}, \cdots, \boldsymbol{\gamma}^{[T]} \tag{2.1}$$

that converges in distribution to $p(\boldsymbol{\gamma}|y)$, the posterior distribution over models. However when p is large, the chain will typically not converge in any reasonable amount of time. Though it is not, perhaps, appropriate to use such chains for Monte Carlo integration, the resulting sequence can be thought of as a stochastic search of the model space and the values $\boldsymbol{\gamma}^{[t]}$ used to identify high probability models.

I focus below on MCMC for "conjugate" models, where the marginal likelihood $p(y|\boldsymbol{\gamma}) = \int p(y|\boldsymbol{\theta}, \boldsymbol{\gamma}) p(\boldsymbol{\theta}|\boldsymbol{\gamma})\ d\boldsymbol{\theta}$ is available in closed form, as this is the type of probability model I consider throughout. MCMC for the nonconjugate case is discussed in George and McCulloch (1993, 1996), Geweke (1996) and Kuo and Mallick (1998), each of which uses prior distributions related to those of Mitchell and Beauchamp (1988).

2.1.1 Model Space Exploration via Gibbs Sampling

The simple one-at-a-time, fixed scan Gibbs sampler for variable selection, described by George and McCulloch (1997), Smith and Kohn (1996, 1997) and Brown *et al.* (1998b), creates the sequence (2.1) by updating the components of $\boldsymbol{\gamma}$ by sampling from

$$p(\gamma_j|\boldsymbol{\gamma}_{-j}, y) \propto p(y|\boldsymbol{\gamma}) p(\gamma_j|\boldsymbol{\gamma}_{-j}) \tag{2.2}$$

for $j = 1, \ldots, p$ at each iteration, where $\boldsymbol{\gamma}_{-j} = (\gamma_1, \ldots, \gamma_{j-1}, \gamma_{j+1}, \ldots, \gamma_p)'$. Common priors have $p(\gamma_j|\boldsymbol{\gamma}_{-j}) = p(\gamma_j)$, and so implementing (2.2) requires sampling a Bernoulli random variable with probability a function of

$$\frac{p(y|\gamma_j = 1, \boldsymbol{\gamma}_{-j})}{p(y|\gamma_j = 0, \boldsymbol{\gamma}_{-j})} \cdot \frac{p(\gamma_j = 1)}{p(\gamma_j = 0)}. \tag{2.3}$$

The conditional probability of $\gamma_j = 1$ is an increasing function of (2.3). Other Gibbs updating strategies besides one-at-a-time fixed scan are available, such as blocked and random scan Gibbs; details can be found in George and McCulloch (1997).

If the sequence (2.1) generated by Gibbs based on (2.3) is to be used to identify high probability models, then Gibbs needs to spend a large percentage of its iterations in high posterior regions of the model space. Consider the following scenario: the current Gibbs model at iteration t has k predictor variables. Assume that the remaining $p-k$ variables are exchangeable and unrelated to the outcome variable, a common situation when there are many possible predictors. As Gibbs cycles through the elements with $\gamma_j^{[t]} = 0$, if the ratio of the first term in (2.3) is approximately one for each of these, Gibbs will try to add each of these variables with probability $p(\gamma_j = 1)$. This amounts to $p - k$ independent Bernoulli trials, and so on average $(p - k) \times \Pr(\gamma_j = 1)$ variables will be added. Of course, Gibbs will also try to remove the original k variables, which may or may not happen depending on how much fit to the data they provide the model. Hence due to the nature in which Gibbs moves from $\boldsymbol{\gamma}^{[t]}$ to $\boldsymbol{\gamma}^{[t+1]}$, extra variables will tend to be added to the model even when this moves $\boldsymbol{\gamma}^{[t+1]}$ away from high probability regions of the posterior. This tendency can add many extra, extraneous variables.

The behavior described above is dependent in part on the equivalence of

marginal likelihoods for many models. If all 2^p marginal likelihoods are equivalent, then Gibbs will simply construct a chain converging in distribution to the prior, $p(\boldsymbol{\gamma})$. One might expect that the inclusion of information from the likelihood would stop Gibbs from wandering for too long in low posterior regions as described above; however, this is not the case, as in order for the Gibbs sampler to sample from the correct stationary distribution it must spend time in these low posterior regions. This is demonstrated in Section 4.7.3 in a real data example.

2.1.2 Model Space Exploration via Metropolis-Hastings

If Gibbs sampling can be problematic for examples with large p due to the nature of the transition from $\boldsymbol{\gamma}^{[t]}$ to $\boldsymbol{\gamma}^{[t+1]}$, then MCMC methods using different transition kernels are of interest. Metropolis-Hastings algorithms provide a flexible framework for specifying which types of moves are allowed at each iteration. Given a current model, $\boldsymbol{\gamma}^{[t]}$, a candidate model $\boldsymbol{\gamma}'$ is sampled from a proposal distribution, $T(\boldsymbol{\gamma}'; \boldsymbol{\gamma}^{[t]})$. The next model, $\boldsymbol{\gamma}^{[t+1]}$, is set to be $\boldsymbol{\gamma}'$ with probability

$$\alpha = \min\left\{ \frac{p(\boldsymbol{\gamma}'|y)}{p(\boldsymbol{\gamma}^{[t]}|y)} \frac{T(\boldsymbol{\gamma}^{[t]}; \boldsymbol{\gamma}')}{T(\boldsymbol{\gamma}'; \boldsymbol{\gamma}^{[t]})}, 1 \right\},$$

and set to be $\boldsymbol{\gamma}^{[t]}$ otherwise.

Madigan and York (1995) and Raftery *et al.* (1997) propose a M-H algorithm, Markov chain Monte Carlo Model Composition (MC^3), using a symmetric proposal distribution, $T(\boldsymbol{\gamma}'; \boldsymbol{\gamma}^{[t]}) = T(\boldsymbol{\gamma}^{[t]}; \boldsymbol{\gamma}')$. First they define a neighborhood, $\text{nbd}(\boldsymbol{\gamma})$, containing $\boldsymbol{\gamma}$, all models obtained by deleting a variable from $\boldsymbol{\gamma}$, and all models obtained by adding a variable to $\boldsymbol{\gamma}$. They then set $T(\boldsymbol{\gamma}'; \boldsymbol{\gamma}^{[t]}) = 0$ for all $\boldsymbol{\gamma}' \notin \text{nbd}(\boldsymbol{\gamma}^{[t]})$ and set $T(\boldsymbol{\gamma}'; \boldsymbol{\gamma}^{[t]})$ constant for all $\boldsymbol{\gamma}' \in \text{nbd}(\boldsymbol{\gamma}^{[t]})$. As this proposal

distribution is symmetric, the acceptance probability reduces to

$$\alpha = \min\left\{\frac{p(y|\boldsymbol{\gamma}')}{p(y|\boldsymbol{\gamma}^{[t]})}\frac{p(\boldsymbol{\gamma}')}{p(\boldsymbol{\gamma}^{[t]})}, 1\right\}. \tag{2.4}$$

MC^3 is equivalent to randomly choosing an index j and proposing $\gamma_j' = 1-\gamma_j^{[t]}$. George and McCulloch (1997) note that this is equivalent to a random scan Gibbs sampler. However, the probability of accepting the change under Gibbs is strictly less than the probability of accepting the change based on (2.4), indicating that the M-H algorithm is more likely to move at each step.

Brown *et al.* (1998a) develop a M-H approach using a slightly richer proposal distribution. They generate proposals based on a current model $\boldsymbol{\gamma}^{[t]}$ by making one of two possible moves, with probabilities ϕ and $1-\phi$ respectively:

1. **Add/Delete:** Randomly choose a variable and propose $\gamma_j' = 1 - \gamma_j^{[t]}$,

2. **Swap:** Randomly choose a j such that $\gamma_j^{[t]} = 1$ and an l such that $\gamma_l^{[t]} = 0$ and propose $\gamma_j' = 0$ and $\gamma_l' = 1$.

The (symmetric) proposal distribution based on a current model $\boldsymbol{\gamma}^{[t]}$ of dimension k is therefore

$$T(\boldsymbol{\gamma}';\boldsymbol{\gamma}^{[t]}) = \begin{cases} \frac{\phi}{p}, & \text{if } |\boldsymbol{\gamma}'| = k-1 \text{ or } k+1, \\ \frac{1-\phi}{k(p-k)}, & \text{if } |\boldsymbol{\gamma}'| = k, \\ 0, & \text{otherwise.} \end{cases}$$

Denison *et al.* (1998a) propose a similar MCMC method for Bayesian curve fitting via piecewise polynomials. They allow for an add, delete or swap-like move at each iteration, where the elements of interest are "knot" locations for basis

functions. A modified version of their prior distribution on the number of knots is used as a proposal distribution for the add, delete and swap moves.

Metropolis-Hastings methods relying on symmetric, near-neighbor based proposal distributions such as described above can be problematic when p is large relative to the model dimensions of interest. In this case, the probability of proposing a model of dimension $k+1$ can be so much larger than proposing a model of dimension $k-1$ that M-H tends to keep adding variables to the model and rarely makes a move to a lower dimension. This is especially relevant in cases where $n \leq p$ and the largest number of allowed variables in a model is capped at $n-1$; in this case there will always be an ample supply of variables to try adding to the model.

For example, in MC^3, if the current model is of size k, the probability of proposing models of dimension $k-1$ and k are $k/(p+1)$ and $1/(p+1)$, respectively, whereas the probability of proposing a model of dimension $k+1$ is $(p-k)/(p+1)$. When p is large relative to values of k of interest, M-H will therefore nearly always try to add variables, creating a chain that does not move much around the model space.

In the approach of Brown *et al.* (1998a), the probabilities of proposing models of dimension $k-1$ and k are $\phi k/p$ and $1-\phi$, respectively, while the probability of proposing a model of dimension $k+1$ is $\phi(p-k)/p$. Here the ascent to higher dimensions will be slower due to $\phi < 1$; however, again, the resulting chain will not move around the space freely enough.

George and McCulloch (1997) discuss alternative proposal distributions that avoid this problem. Rather than sampling uniformly from a set of neighboring models, they suggest first sampling a dimension to which to move with prespecified

probabilities and then sampling uniformly from a set within that dimension. This type of asymmetric proposal distribution can focus the M-H search on models of smaller dimension; however, they require the user to prespecify the probabilities for sampling models of a given dimension, which can be inefficient if the posterior distribution of interest is concentrated on dimensions different to those favored by the proposal.

A new method for model space exploration via MCMC has recently been proposed by Nott and Green (2004) that improves on the above approaches when there is multicollinearity in the predictor variables. The approach, which uses auxiliary variable methods related to those of Higdon (1998), is only demonstrated for examples with up to 62 predictor variables, leaving open the question of scalability in p.

2.2 Shotgun Stochastic Search

In the following sections I propose and explore a *Shotgun Stochastic Search* (SSS) approach – and related methods – for regression model space exploration. SSS is inspired by MCMC, but addresses the problems raised above for model search in high dimensions. A key element of the approach is that it is parallelizable in a distributed computing environment, allowing access to novel search methods that would otherwise be too slow to be practicable.

In general, a shotgun stochastic search of a model space is a sequential, local-move, neighborhood-based procedure, where the following are performed at each step:

Step 1 Use the current model to define a neighborhood of proposal models,

Step 2 Evaluate each proposal model in this neighborhood *in parallel*,

Step 3 Choose a new current model from the proposals.

Three keys to this method, depicted graphically in Figure 2.1, are the choice of neighborhood, the parallelization of Step 2 and the choice of the criterion used to distinguish between models. The main idea is that, for any particular regression model, there will be many other regression models with similar fit to the data, e.g. models with overlapping sets of covariates; quickly identifying and evaluating these models provides both a rich description of the (local) model space and a new set of competitive models from which to choose the next move. In essence, SSS generates multiple candidate models and "shoots out" proposed moves in various directions, like a shotgun.

In order for the approach to be effective, the neighborhood of the current model must be defined to be comprehensive enough to allow the search to move easily throughout the model space. As I describe in Section 2.3, this is accomplished by ensuring that SSS considers incorporating each possible predictor variable in one of the proposal models at each iteration. This approach has the added benefit that, over the course of the search, every variable is evaluated in the context of many different regression models. The potential computational burden of considering every possible predictor variable at each iteration is lessened by the fact that Step 2 can be parallelized using distributed computing, as each of the proposal models can be evaluated independently on separate processing elements. The results can then be combined into a list which, besides serving as the proposed moves for the next iteration, can be used to update a running list of top models to be used for inference (see Section 3.3). This is a clear advantage of SSS procedures over many Metroplis-based MCMC algorithms, where typically only one proposal

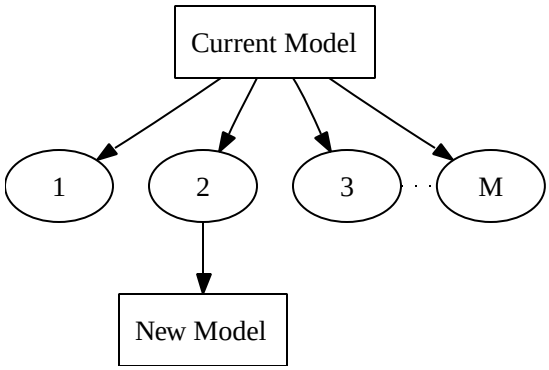

Figure 2.1: Graphical depiction of one step in the general shotgun stochastic search described in Section 2.2. STEP 2, depicted by the ovals, is the key step: M candidate models are evaluated in parallel.

model is evaluated and recorded at each iteration, thereby reducing the number of interesting models evaluated. SSS catalogues many models as it proceeds.

I have yet to describe the third key to the search, the criterion used to compare models. This tends to be a problem-specific choice, although from a Bayesian perspective the straightforward answer is to use the relative posterior probabilities of the models being considered when those (relative) probabilities are available in closed form, or when they can be easily approximated numerically. Details of how models are compared from a Bayesian perspective are given in Chapter 3. Other model scores, such as AIC or BIC, may of course be substituted.

2.3 Regression Model SSS

The two major components of SSS that need to be specified are the neighborhood component (how proposal models are "shot out" at each iteration) and the sampling component (how we choose from among the proposals). I have already stated that the neighborhood component should be broad enough to include each of the possible predictor variables in some way, and now add the requirement that

the neighborhood should also include regression models of various dimensions to allow the search to move freely across model size.

Informally, take the neighborhood to be every regression model having a one variable difference with the current model. For a given regression model of dimension k (i.e., having k predictor variables) the neighborhood is broken down into a set with three elements, $\text{nbd}(\boldsymbol{\gamma}) = \{\boldsymbol{\gamma}^+, \boldsymbol{\gamma}^\circ, \boldsymbol{\gamma}^-\}$, where $\boldsymbol{\gamma}^+$ is a set containing neighboring models of dimension $k+1$, called the "addition" moves, $\boldsymbol{\gamma}^\circ$ is a set containing neighboring models of dimension k, called the "replacement" moves, and $\boldsymbol{\gamma}^-$ is a set containing neighboring models of dimension $k-1$, called the "deletion" moves. Together, these three sets make up the set of proposal models that "shoot out" at each iteration. Their names are derived from the models from which they are comprised, e.g. $\boldsymbol{\gamma}^+$ contains all of the models obtained by adding one at a time each of the $p-k$ remaining predictor variables to the current model and $\boldsymbol{\gamma}^-$ contains the k models obtained by deleting one at a time each of the variables currently in the model. The replacement model set, $\boldsymbol{\gamma}^\circ$, contains all of the models obtained by replacing one at a time each of the variables in the current model with the $p-k$ remaining predictor variables.

For example, suppose $p = 5$ and the current regession model is $\{x_1, x_3, x_4\}$, corresponding to $\boldsymbol{\gamma} = (1, 0, 1, 1, 0)'$. Then

$$\begin{aligned}
\boldsymbol{\gamma}^- &= \Big\{\{x_3, x_4\}, \{x_1, x_4\}, \{x_1, x_3\}\Big\}, \\
\boldsymbol{\gamma}^\circ &= \bigcup_{j\in\{2,5\}} \Big\{\{x_1, x_3, x_j\}, \{x_1, x_j, x_4\}, \{x_j, x_3, x_4\}\Big\}, \\
\boldsymbol{\gamma}^+ &= \bigcup_{j\in\{2,5\}} \{x_1, x_3, x_4, x_j\}.
\end{aligned}$$

It should be noted that $|\boldsymbol{\gamma}^+| = p - k$, $|\boldsymbol{\gamma}^\circ| = k(p-k)$ and $|\boldsymbol{\gamma}^-| = k$ if $2 \leq k < p$, with the convention that $\boldsymbol{\gamma}^+ = \emptyset$ when $k = p$. SSS evaluates the null model, $\boldsymbol{\gamma} = \mathbf{0}$, and all possible one variable models before starting the search, hence it only considers models of dimension at least $k = 2$ as the search progresses.

As p is typically large, $|\boldsymbol{\gamma}^\circ| \gg |\boldsymbol{\gamma}^+| \gg |\boldsymbol{\gamma}^-|$, which can be problematic for sampling. If all of the models have equal weight and we sample one model directly from $\text{nbd}(\boldsymbol{\gamma})$, then as $p \to \infty$ the probability of staying in the same dimension goes to $k/(k+1)$, the probability of increasing dimension goes to $1/(k+1)$ and the probability of decreasing dimension goes to zero. This is a problem similar to that with symmetric proposal distributions for M-H discussed in Section 2.1.2. This imbalance clearly needs to be addressed for SSS to move across dimension effectively; this is accomplish by breaking the sampling step into two parts. First, three models, $\boldsymbol{\gamma}_*^+$, $\boldsymbol{\gamma}_*^\circ$ and $\boldsymbol{\gamma}_*^-$ are sampled from $\boldsymbol{\gamma}^+$, $\boldsymbol{\gamma}^\circ$ and $\boldsymbol{\gamma}^-$, respectively. The new model is then sampled from the set of three proposals. Breaking down the sampling into these two steps balances out the dimensional disparity in the neighborhood and encourages movement across dimension.

Implicit in the approach is the requirement of a measure of goodness of fit that can be used to score each regression model. Common scores used in the regression model context are the set of information critera (A/BIC), the simple (or adjusted) R^2 statistic, and variants of the likelihood ratio statistic. In a Bayesian approach, the key score is the unnormalized posterior probability of each model, $p(\boldsymbol{\gamma}|y) \propto p(y|\boldsymbol{\gamma})p(\boldsymbol{\gamma})$, making the sampling mechanism interpretable and allowing for comparisons with MCMC algorithms. The Bayesian Information Criterion (BIC) can be viewed as an approximation to the marginal likelihood of a given model, $p(y|\boldsymbol{\gamma})$, under a flat prior distribution (Raftery, 1995) and so could be used

in similar fashion. Other scores such as R^2 and AIC can be used, but the user would have to decide how to use these scores to move from model to model across iterations, i.e. how to normalize the scores into a probability vector from which to sample. In general I refer to a score for a model $\boldsymbol{\gamma}$ that can be normalized within a set of scores to become a probability as $S(\boldsymbol{\gamma})$.

Regression Model Shotgun Stochastic Search:

Let $\boldsymbol{\gamma}$ be a regression model and let $S(\boldsymbol{\gamma})$ be its corresponding (unnormalized) score. Initialize an empty list, Γ^*, that will contain the best B regression models evaluated. Given a starting model $\boldsymbol{\gamma}^{[0]}$, iterate in $t = 1, \ldots, T$ the following steps:

Step 1 *In parallel*, compute $S(\boldsymbol{\gamma})$ for all $\boldsymbol{\gamma} \in \text{nbd}(\boldsymbol{\gamma}^{[t]})$, constructing $\boldsymbol{\gamma}^+$, $\boldsymbol{\gamma}^\circ$ and $\boldsymbol{\gamma}^-$. Update the list of the overall best models evaluated, Γ^*.

Step 2 Sample $\boldsymbol{\gamma}_*^+$, $\boldsymbol{\gamma}_*^\circ$ and $\boldsymbol{\gamma}_*^-$, from $\boldsymbol{\gamma}^+$, $\boldsymbol{\gamma}^\circ$ and $\boldsymbol{\gamma}^-$, respectively, with probabilities proportional to $S(\boldsymbol{\gamma})^{\alpha_1}$, normalized within each set.

Step 3 Sample $\boldsymbol{\gamma}^{[t+1]}$ from $\{\boldsymbol{\gamma}_*^+, \boldsymbol{\gamma}_*^\circ, \boldsymbol{\gamma}_*^-\}$ with probability proportional to $S(\boldsymbol{\gamma})^{\alpha_2}$, renormalized within this set.

One iteration is depicted graphically in Figure 2.2. The annealing parameters α_1 and α_2 are positive numbers set by the user to control the greediness of the search. Values less than one flatten out the proposal distribution, allowing the search to wander around more freely. The second parameter, α_2, can be adjusted to control how readily the search moves across dimension, ensuring that a representative piece of the model space will be explored. Values of α_1 and α_2 much greater than one turn the search into a hill climbing algorithm, moving deterministically to the best neighboring model and eventually resulting in oscillation between two models

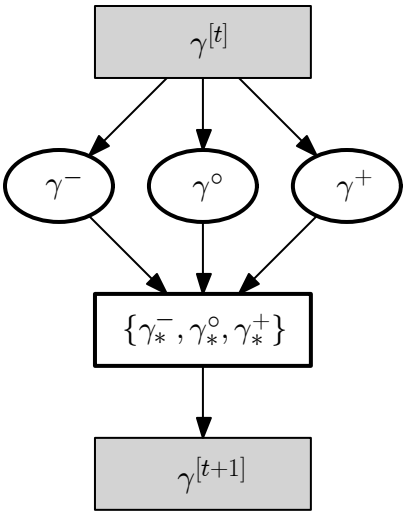

Figure 2.2: One iteration of regression model SSS as described in Section 2.3. The ovals represent the parallelizable step and the unshaded box is the dimension-balanced collection of models from which $\boldsymbol{\gamma}^{[t+1]}$ is sampled.

at a local mode. If it is desired, a separate value of α_1 can be used for each of the sets $\boldsymbol{\gamma}^+$, $\boldsymbol{\gamma}^\circ$ and $\boldsymbol{\gamma}^-$. Throughout, unless otherwise specified I take $\alpha_1 = \alpha_2 = 1$.

As the search progresses, SSS maintains a list of the best models evaluated, Γ^*, according to their scores, $S(\boldsymbol{\gamma})$. Because this list is constructed based on every model evaluated in Step 1 and not solely based on the models sampled in Step 3, we end up with a list of models that is representative of the high posterior regions of the model space explored.

2.4 Relationship to MCMC

In cases of high dimensional parameter spaces, MCMC approaches are often used as stochastic search tools to identify regions of high posterior probability (or in the context of model selection, to identify the "best" models), rather than with the aim of performing Monte Carlo integration to summarize the posterior distribution. In

these cases, the Markov chains created by the MCMC algorithms are not expected to converge to a stationary distribution in a reasonable amount of time; rather they are expected to hone-in on high posterior regions. In this section I show that small changes to SSS result in an MCMC algorithm whose particular form has advantages over common MCMC approaches.

Say we wish to use a Metropolis-Hastings algorithm to sample from a discrete distribution, $P(x)$, where we can evaluate $P(x)$ up to a normalizing constant, $P(x) = Q(x)/Z$. Consider proposal distributions that sample from $P(x)$ restricted to a neighborhood $B(\cdot)$:

$$\begin{aligned} T(x_{t+1}; x_t) &= \frac{P(x_{t+1})\mathbf{1}(x_{t+1} \in B(x_t))}{\sum_{s \in B(x_t)} P(s)} \\ &= \frac{Q(x_{t+1})\mathbf{1}(x_{t+1} \in B(x_t))}{\sum_{s \in B(x_t)} Q(s)}. \end{aligned}$$

As long as the chain is started in a region of nonzero probability, the acceptance probability at each iteration is

$$\alpha = \min \left\{ 1, \frac{Q\Big(B(x_{t+1})\Big)}{Q\Big(B(x_t)\Big)} \right\}. \tag{2.5}$$

In other words, if the distribution we wish to sample from, restricted to a neighborhood, is used as a proposal distribution, then the new state will always be accepted if $P(B(x_{t+1})) \geq P(B(x_t))$ (if the new neighborhood contains more mass than the old neighborhood); otherwise the new state will be accepted with probability equal to the relative size of the smaller set.

The SSS approach described in Section 2.3 can be easily adapted to become a Metropolis-Hastings algorithm using the proposal distribution described above.

Relating notation, $P(x_t)$ is $p(\boldsymbol{\gamma}^{[t]}|y)$, $Q(x_t)$ is $S(\boldsymbol{\gamma}^{[t]}) = p(y|\boldsymbol{\gamma}^{[t]})p(\boldsymbol{\gamma}^{[t]})$, and $B(x_t)$ is $\text{nbd}(\boldsymbol{\gamma}^{[t]})$. After performing Step 1 at iteration t in SSS, sample a proposal $\boldsymbol{\gamma}'$ from the discrete distribution $S(\cdot)$ normalized within $\text{nbd}(\boldsymbol{\gamma}^{[t]})$, and set $\boldsymbol{\gamma}^{[t+1]} = \boldsymbol{\gamma}'$ with probability α from (2.5) (otherwise, set $\boldsymbol{\gamma}^{[t+1]} = \boldsymbol{\gamma}^{[t]}$). Step 2 and Step 3, which are related to the two stage sampling process that corrects the dimension imbalance, are ignored; however, this does not necessarily have the same effect as in the M-H algorithm described in Section 2.1.2, because SSS samples from the restricted posterior distribution rather than randomly proposing models.

Compared to using common proposal distributions for Metropolis-Hastings algorithms that are specified independently of the posterior, using the restricted posterior as the proposal distribution is advantageous in that the resulting chain can move more quickly to regions of high posterior probability. This is illustrated in Figure 2.3 for the continuous random variable case. The open circle represents the location of the current state in the chain and is plotted at its corresponding height on the posterior. Typical proposal distributions are commonly symmetric in the sense that $T(x_{t+1}; x_t) = T(x_t; x_{t+1})$, and are often distance based, weighting points closer to x_t more heavily than those far away. Under these conditions, proposal values near the solid circle are likely to be drawn; however, the resulting acceptance probability will be less than one as the current value is at a higher point on the density. In fact, for any proposal that happens to be drawn falling in the neighborhood defined by the thin dashed lines, the acceptance probability will be less than one as the current value is at a local mode.

When the restricted posterior, constrained to the neighborhood defined by the thin dashed lines, is used as the proposal distribution, proposals near the solid circle are again likely to be drawn as that is a region of relative high posterior

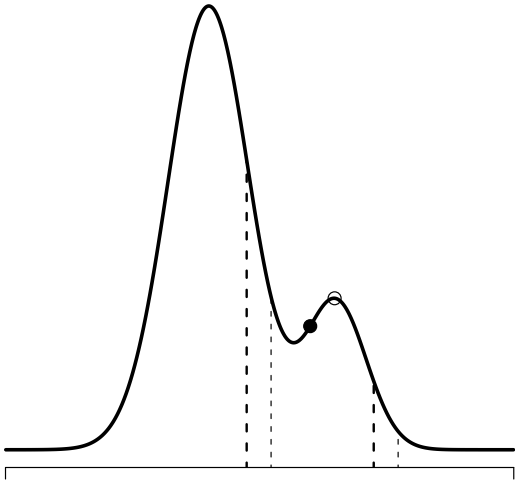

Figure 2.3: Illustration for the discussion of the Metropolis-Hastings algorithm in Section 2.4.

probability. In this case, however, the resulting acceptance probability will be one, as the neighborhood around the proposal, defined by the thick dashed line, has more posterior mass than the neighborhood around the current value. Hence using the restricted posterior distribution as a proposal will make it easier to move away from local modes if there are other modes nearby.

In this sense, MCMC based on SSS behaves differently than previous MCMC approaches such as the Markov chain Monte Carlo Model Composition (MC^3) algorithm of Madigan and York (1995) and Raftery *et al.* (1997) described in Section 2.1.2. As MC^3 proceeds, if the chain is in state $\boldsymbol{\gamma}^{[t]}$, a proposal move $\boldsymbol{\gamma}'$ is drawn from $T_*(\boldsymbol{\gamma}'; \boldsymbol{\gamma}^{[t]})$, a discrete uniform distribution over $\text{nbd}(\boldsymbol{\gamma}^{[t]})$, and is accepted with probability

$$\alpha_* = \min\left\{1, \frac{p(y|\boldsymbol{\gamma}')p(\boldsymbol{\gamma}')}{p(y|\boldsymbol{\gamma}^{[t]})p(\boldsymbol{\gamma}^{[t]})}\right\},$$

which favors rejecting moves away from local modes as illustrated above even

when the moves away would take the chain to more likely neighborhoods. In addition to this advantage, we expect SSS to perform better due to the fact that the neighborhood around the current model is much richer than in MC^3, as it includes all of the "replacement" models, $\boldsymbol{\gamma}^\circ$, allowing SSS to quickly explore larger regions of the posterior. We also expect SSS to perform better than the approach of Brown *et al.* (1998a), who consider a similar neighborhood structure, as they only look at one model at random in the neighborhood.

The above discussion of SSS as MCMC is based on a modified version of SSS. Directly converting SSS into a Metropolis-Hastings algorithm is complicated by the two stage sampling process used to balance dimension in the proposal distribution. The acceptance probability,

$$\alpha = \min\left\{1, \frac{p(\boldsymbol{\gamma}'|y)}{p(\boldsymbol{\gamma}|y)}\frac{T(\boldsymbol{\gamma}^{[t]};\boldsymbol{\gamma}')}{T(\boldsymbol{\gamma}';\boldsymbol{\gamma}^{[t]})}\right\},$$

requires calculation of the transition probabilities, which in turn requires marginalizing over the two dimensions not sampled in the second stage. For example, if the sampled proposal $\boldsymbol{\gamma}'$ is from the addition set $\boldsymbol{\gamma}^+$, then the required forward transition probability is

$$\begin{aligned} T(\boldsymbol{\gamma}';\boldsymbol{\gamma}^{[t]}) &= \sum_{\boldsymbol{\gamma}_*^\circ \in \boldsymbol{\gamma}^\circ} \sum_{\boldsymbol{\gamma}_*^- \in \boldsymbol{\gamma}^-} \left[\frac{p(\boldsymbol{\gamma}'|y)}{p(\boldsymbol{\gamma}'|y) + p(\boldsymbol{\gamma}_*^\circ|y) + p(\boldsymbol{\gamma}_*^-|y)}\right] \times \\ &\qquad \left[\frac{p(\boldsymbol{\gamma}'|y)}{\sum_{\mathbf{u}\in\boldsymbol{\gamma}^+} p(\mathbf{u}|y)} \cdot \frac{p(\boldsymbol{\gamma}_*^\circ|y)}{\sum_{\mathbf{v}\in\boldsymbol{\gamma}^\circ} p(\mathbf{v}|y)} \cdot \frac{p(\boldsymbol{\gamma}_*^-|y)}{\sum_{\mathbf{w}\in\boldsymbol{\gamma}^-} p(\mathbf{w}|y)}\right], \end{aligned} \quad (2.6)$$

where $p(\boldsymbol{\gamma}|y)$ can be replaced by $S(\boldsymbol{\gamma})$ as the normalizing constants cancel. The large summation in (2.6) makes computation of both the forward and backward proposal probabilities undesirable. This should not be viewed as problematic,

though, as computation of these probabilities is only needed to compute the acceptance probability, α. Because SSS samples proposals based on the restricted posterior distribution, and because the resulting chain will not be used for Monte Carlo integration, it seems inefficient to reject a move. SSS can thus be treated as a stochastic search tool that is similar to a Metropolis-Hastings algorithm and used to thoroughly explore regions of high posterior probability.

2.5 Analysis of SSS Performance

2.5.1 Random Walk SSS

Consider the case of a *fixed dimensional* SSS, one where we condition on a particular number of variables k. Assume that the true model $\boldsymbol{\gamma}^*$ is of dimension k, and only allow moves within this dimension, effectively setting $\text{nbd}(\boldsymbol{\gamma}) = \boldsymbol{\gamma}^\circ$. A fixed dimensional SSS creates a Markov chain $\{\boldsymbol{\gamma}_t\}$ over the state space of models restricted to size k, $\Gamma^{(k)}$, which contains $\binom{p}{k}$ elements. As I have conditioned on a particular model size, k, any model $\boldsymbol{\gamma}$ can be categorized as belonging to one of $k+1$ classes: the class where $\boldsymbol{\gamma}$ shares *none* of the same variables as the true model, $\boldsymbol{\gamma}^*$, the class where $\boldsymbol{\gamma}$ shares *one* of the same variables as $\boldsymbol{\gamma}^*$, up to the class where $\boldsymbol{\gamma}$ contains *all* k of the same variables as $\boldsymbol{\gamma}^*$. Thus we can define the map $\psi(\boldsymbol{\gamma}_t) = Z_t$, where $Z_t \in \{0, \ldots, k\}$ and indicates how many of the variables in $\boldsymbol{\gamma}_t$ are shared by $\boldsymbol{\gamma}^*$. As interest is on the expected time to find the true model, I analyze the induced chain $\{Z_t\}$ which is defined on a much smaller state space. Reformulated, the problem is now to find the expected time for the chain $\{Z_t\}$ to reach state k.

Let $T(p, k) = \min\{t \geq 0 : Z_t = k\}$ be a random variable representing the

time to reach the true model when there are p possible predictors and the true model is of dimension k. Interest is on the quantities $v_i(p,k) = \mathbb{E}[T(p,k)|Z_0 = i]$, $i = 0, \ldots, k$, noting that $v_k(p,k) = 0$. As a technical note, the state space for the chain is $\{\max\{0, 2k-p\}, \ldots, k\}$, which is $\{0, \ldots, k\}$ for relevant values of p and k ($p \gg k$). The reduced state space in certain situations is due to the fact that, if you have say $p = 4$ and $k = 3$, there are no models with zero variables in common with $\boldsymbol{\gamma}^*$. For simplicity of presentation I only consider cases here where $2k - p \leq 0$ so that $v_0(p,k)$ is meaningful.

In order to analyze the chain we must specify the transition matrix $\mathrm{P}_{p,k}$ for a case with p predictor variables conditioned on the true model being of size k. $\mathrm{P}_{p,k}$ is hence a $(k+1) \times (k+1)$ stochastic matrix with entries $\mathrm{P}_{p,k}(i+1, j+1) = \Pr(Z_{t+1} = j | Z_t = i)$ for $i, j = 0, \ldots, k$. The state $k+1$ is treated as an absorbing state, implying $\mathrm{P}_{p,k}(k+1, k+1) = 1$ and $\mathrm{P}_{p,k}(k+1, l) = 0$ for all $l \neq k+1$. Due to the Markovian nature of the chain, we have

$$v_i = 1 + \sum_{j=0}^{k-1} \Pr(Z_{t+1} = j | Z_t = i) v_j, \qquad i = 0, \ldots, k-1. \tag{2.7}$$

Define the substochastic matrix $\mathrm{Q}_{p,k}$ to be the matrix $\mathrm{P}_{p,k}$ with the final row and column removed, i.e., where the absorbing state has been removed. We can then write (2.7) as

$$\mathbf{v} = \mathbf{1} + \mathrm{Q}_{p,k}\mathbf{v},$$

and hence the vector of expected times to find $\boldsymbol{\gamma}^*$ is

$$\mathbf{v}(p,k) = (I_k - \mathrm{Q}_{p,k})^{-1}\mathbf{1}.$$

Primary interest is on $v_0(p,k)$, the expected time to find $\boldsymbol{\gamma}^*$ starting from a model

with no variables in common with $\boldsymbol{\gamma}^*$, because when $p \gg k$ randomly choosing a starting point will put us in this situation with high probability.

As a baseline, consider a *random walk shotgun stochastic search* (RWSSS). This is a SSS where we set $S(\boldsymbol{\gamma}) = |\text{nbd}(\boldsymbol{\gamma})|^{-1}$, i.e. where we sample uniformly from the neighborhood around the current model. From this we can specify the elements of $\text{P}_{p,k}$:

$$\Pr(Z_{t+1} = j | Z_t = i) = \begin{cases} \frac{i(p-2k+i)}{k(p-k)} & \text{if } j = i-1,\ 0 < i < k, \\ \frac{(k-i)(p-2(k-i))}{k(p-k)} & \text{if } j = i \neq k, \\ \frac{(k-i)^2}{k(p-k)} & \text{if } j = i+1,\ i < k, \\ 1 & \text{if } j = i = k, \\ 0 & \text{o.w.} \end{cases} \tag{2.8}$$

Values of $v_0(p, k)$ are shown in the first panel of Figure 2.4 for the RWSSS. For $p = 500$, $\boldsymbol{\gamma}^*$ is found on average after about 125,000 steps when the true model has two variables. As the model space grows larger, say to when the true model has six variables, the number of expected steps increases to about 20 trillion. Including distinguishing information about the models, namely by sampling based on their relative posterior probabilities, will reduce these expected times dramatically.

2.5.2 SSS for Orthogonal Designs

In the special case of an orthogonal design, when $x_i' x_j = 0$ for all $i \neq j$, we can extend the results of the fixed dimensional RWSSS in Section 2.5.1 to the case of a fixed dimensional SSS. Consider two models of size k, $\boldsymbol{\gamma}_a$ and $\boldsymbol{\gamma}_b$, that differ by only one variable. Let $X_a = (X_1\ X_2)$ and $X_b = (X_1\ X_3)$ be the the design

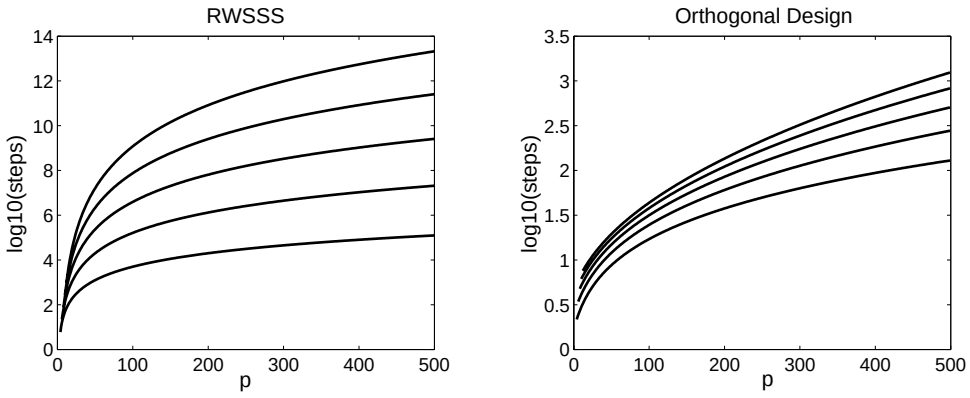

Figure 2.4: Expected steps (log base 10 scale) to find the true model for RWSSS and SSS under an orthogonal design for various values of p and k. The five lines represent $k = 2, \ldots, 6$, with the lowest line being $k = 2$ in each plot. Details are given in Sections 2.5.1 and 2.5.2.

matrices for these two models, where X_1 is a set of $k-1$ common variables. Under the model specified in Section 3.1, the ratio of marginal likelihoods for these two models under orthogonality is

$$\frac{p(y|\boldsymbol{\gamma}_a)}{p(y|\boldsymbol{\gamma}_b)} = \left(\frac{a - y'X_1X_1'y - y'X_3X_3'y}{a - y'X_1X_1'y - y'X_2X_2'y}\right)^{\nu},$$

where $\nu = (n + \delta + k)/2$ and $a = (\tau + n - 1)^2$. The numerator and denominator differ only by the last term, a scaled version of the least squares estimate $\hat{\beta}_j = (n-1)^{-1}x_j'y$. Therefore, the amount by which the ratio differs from unity will depend on the difference between the two regression coefficients $\hat{\beta}_2$ and $\hat{\beta}_3$. These in turn should be related to whether or not the corresponding variables are shared by the true model: $\hat{\beta}_j$ should be relatively large if x_j is in the true model and relatively small otherwise.

As above, consider a Markov chain $\{Z_t\}$ on the state space $\{0, \ldots, k\}$. In order to compute the required transition probabilities under an orthogonal design, I make the simplifying assumption that all variables not in the true model have

the same (relatively small) scaled regression coefficient $\epsilon = x_j'y$, and that all of the variables that are in the true model have the same (relatively large) scaled regression coefficient $\lambda = x_j'y$. After specifying these two values the relevant transition probabilites $\mathrm{P}_{p,k}(i,j)$ can be derived:

$$\Pr(Z_{t+1}=j|Z_t=i) = \begin{cases} (2.9) & \text{if } j=i-1,\ 0<i<k, \\ (2.10) & \text{if } j=i\neq k, \\ (2.11) & \text{if } j=i+1,\ i<k, \\ 1 & \text{if } j=i=k, \\ 0 & \text{o.w.}, \end{cases}$$

where the referenced equations are

$$\frac{1}{1+\left(\frac{k-i}{i}\right)\left(\frac{p-2k+2i}{p-2k+i}\right)\left(1-\frac{\epsilon^2-\lambda^2}{b_i}\right)^{\nu}+\left(\frac{(k-i)^2}{i(p-2k+i)}\right)\left(\frac{b_i-(\epsilon^2-\lambda^2)}{b_i+(\epsilon^2-\lambda^2)}\right)^{\nu}}, \tag{2.9}$$

$$\frac{1}{1+\left(\frac{i}{k-i}\right)\left(\frac{p-2k+i}{p-2k+2i}\right)\left(1-\frac{\epsilon^2-\lambda^2}{b_i}\right)^{-\nu}+\left(\frac{k-i}{p-2k+2i}\right)\left(1+\frac{\epsilon^2-\lambda^2}{b_i}\right)^{-\nu}}, \tag{2.10}$$

$$\frac{1}{1+\left(\frac{i(p-2k+i)}{(k-i)^2}\right)\left(\frac{b_i+(\epsilon^2-\lambda^2)}{b_i-(\epsilon^2-\lambda^2)}\right)^{\nu}+\left(\frac{p-2k+2i}{k-i}\right)\left(1+\frac{\epsilon^2-\lambda^2}{b_i}\right)^{\nu}}, \tag{2.11}$$

and $b_i = a-(k-i)\epsilon^2 - i\lambda^2$.

The second panel of Figure 2.4 shows expected hitting times as a function of p for values $k=2,\ldots,6$ under an orthogonal design. Here $n=500$, $\epsilon=(n-1)0.005$, $\lambda=(n-1)0.1$, $\tau=1$ and $\delta=3$. As seen in Figure 2.4, under an orthogonal design the expected number of steps required to hit the true model is drastically smaller under SSS than RWSSS: the expected time to find the true model for

model spaces with p around 500 is on the order of several thousand steps. Note that this is the expected time for the chain to achieve $Z_t = k$, however in SSS the true model would be evaluated the step after the chain achieved $Z_t = k - 1$, because one of the models in $\text{nbd}(\boldsymbol{\gamma}_t) = \boldsymbol{\gamma}^\circ$ would be the true model. Hence we could take state $k-1$ as the absorbing state in the analysis and find the expected time until the true model is *evaluated*, which of course will be smaller than the results presented here.

2.5.3 Simulation Study

In this section I report a simulation study based on a real dataset to demonstrate the effectiveness of SSS as the number of possible predictor variables, p, increases. I do not restrict the analysis to a fixed dimensional SSS as was considered in the previous two sections, but instead allow SSS to move across dimension fully as described in Section 2.3. The data on which the simulation is based is a gene expression dataset from a survival study in brain cancer based at the Keck Center for Neuro-Oncology at Duke University. A description of the data and an analysis using SSS can be found in Chapter 4.

The study consists of $n = 41$ patients, and for each patient there is gene expression data consisting of $p = 8,408$ genes from a tumor specimen. I selected four genes from the dataset as the variables comprising the "true" model $\boldsymbol{\gamma}^*$ and simulated $m = 1, \ldots, 50$ outcomes using the actual gene expression values x_{ij} for the $j = 1, \ldots, 4$ "true" variables according to the regression model

$$y_i^{(m)} = 1.3x_{i1} + 0.3x_{i2} - 1.2x_{i3} - 0.5x_{i4} + \varepsilon_i^{(m)}, \tag{2.12}$$

for $i = 1, \ldots, 41$ where the $\varepsilon_i^{(m)}$ are i.i.d. mean zero normal random variables with

variance 0.5. The simulated outcomes were then standardized to have mean zero and unit variance within each of the 50 simulations.

To assess the performance of SSS as the size of the dataset increases, I ran SSS for the 50 simulated responses using 12 datasets with increasing values of p, as shown in Table 2.1. The datasets were constructed by first reordering the observed $41 \times 8,408$ data matrix X so that the four variables used in the simulation are labeled as variables $1, 2, 3$ and 4. To construct a data matrix $X^{(m,p)}$ for a particular simulation, when $p \leq 8,408$ I extracted the first p columns of X to form $X^{(m,p)}$ and then randomly permuted the columns. Hence all 50 datasets $X^{(m,p)}$ for a given $p \leq 8,408$ contain the same variables and differ only by a column permutation. For the datasets with $p > 8,408$, before permuting the columns I added $p - 8,408$ columns of random draws from a $\mathrm{N}(0, I_{41})$ distribution (after centering and scaling the random draws), effectively adding random noise to the dataset. Note that, for a given $p > 8,408$, different random draws are used for each of the 50 simulated $X^{(m,p)}$.

The prior distributions over the parameter space used in the simulation study are consistent with those used in the analysis in Chapter 4, with $\tau = 1$ and $\delta = 3$. For the prior distribution on the model space, I used the prior (3.7) with $\pi = 4/p$ in order to maintain focus on sparse models as the size of the model space increased.

For a given run of SSS, it was declared that SSS had found the true model when the true model was evaluated by SSS, i.e. when $\boldsymbol{\gamma}^* \in \mathrm{nbd}(\boldsymbol{\gamma}^{[t]})$. For each value (m, p), if SSS found the true model within 10,000 iterations I recorded the number of iterations required to find the model and the elapsed time. If the model was not found within 10,000 iterations I recorded the time required for the

10,000 iteration run. Note that the "true model" was not necessarily the highest posterior probability model, but simply the generating model.

Computation was done using 21 processing elements (one master node and 20 compute nodes) on a cluster of dual-processing, 3.1 GHz Intel x86 based machines running Linux. SSS was run for one value of (m, p) at a time using the 21 processors, and the resulting run-time for the simulation was less than fifteen days.

Results from the simulation study are shown in Table 2.1 and Figure 2.5. SSS found over 96% of the models, as seen in the column labeled "missed" in Table 2.1, which shows the number of models not found within 10,000 iterations. Additionally, SSS found over 94% of models for datasets with $p \geq 5,000$. Increasing the number of irrelevant variables in the dataset resulted in an increase in the number of iterations needed to find the true model, however the true model was still found by SSS a large percentage of the time.

I used the simulation study to obtain an estimate of the run-time of SSS as a function of p for datasets with a similar number of observations. The top panel of Figure 2.6 displays the run times for datasets with $p \geq 1000$. The lines connect output from SSS runs with the same value of p. A straight line would indicate that the load on the computer cluster was constant over runs for a given p; as the lines are fairly straight our results should not have been greatly affected by other processes running on the cluster. The first plot in the bottom panel shows the estimated number of iterations per second as a function of p, where the point estimates are the slope coefficients from linear regressions of iterations on seconds for each of the empirical lines shown in the top panel. The second plot on the bottom panel is the inverse of the first. For each additional 1000 variables added

Table 2.1: Results from the simulation study described in Section 2.5.3.

		iterations			seconds		
p	missed	average	min	max	average	min	max
10		3.38	2	5	0.02	0	1
100		6.66	4	23	0.1	0	1
500		29.68	4	412	1.5	0	22
1000		65.88	3	646	6.64	0	66
2500		181.84	4	1877	46.26	1	492
5000	1	846.88	5	7925	437.02	3	3997
7500	4	790.26	4	9269	601.41	3	7031
8408	1	1001.33	4	8309	878.02	4	7466
10000	3	1542.87	5	9706	1599.51	4	10553
12500	2	1228.35	4	9575	1528.50	5	11937
15000	5	973.07	5	6044	1454.69	7	9403
17500	3	1030.32	5	6938	1785.17	10	12236
20000	3	1259.04	4	9791	2578.11	7	19609
22500	3	1080.89	4	8639	2455.43	8	18371
25000	5	1384.18	6	9850	3522.18	15	24842

to the dataset, SSS took an extra 0.1 seconds per iteration to run.

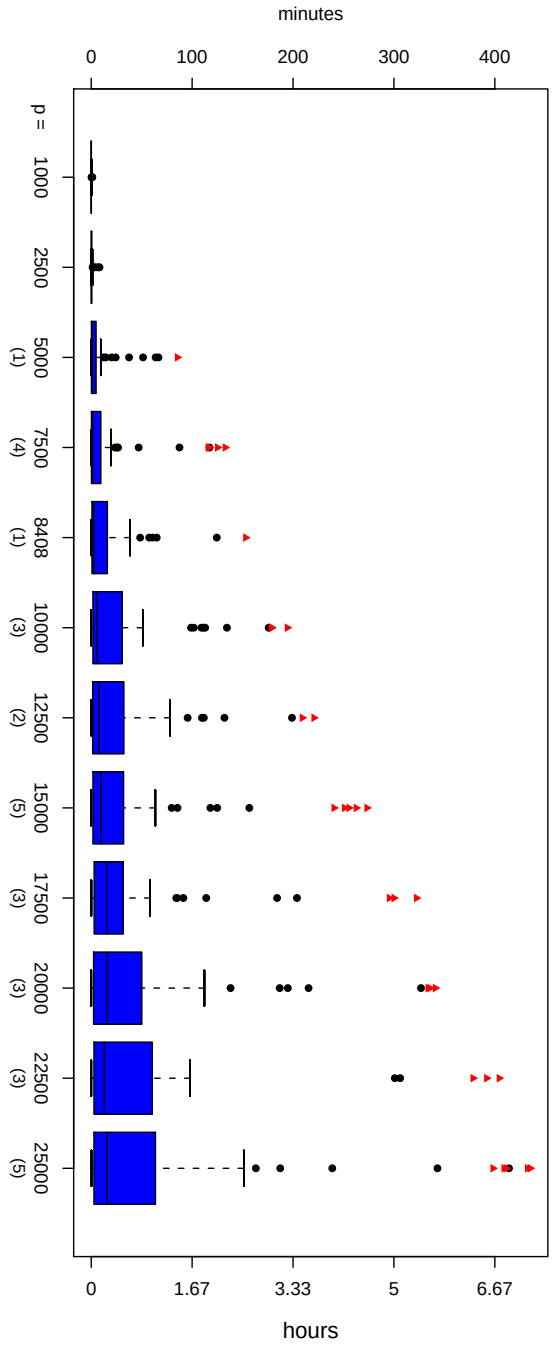

Figure 2.5: The time required to find the true model for the simulation study described in Section 2.5.3. The numbers in parentheses indicate the number of models not found by SSS in 10,000 iterations for a given dataset size p; these runs are indicated by red triangles. The boxplots are constructed based on only those runs where SSS found $\boldsymbol{\gamma}^*$.

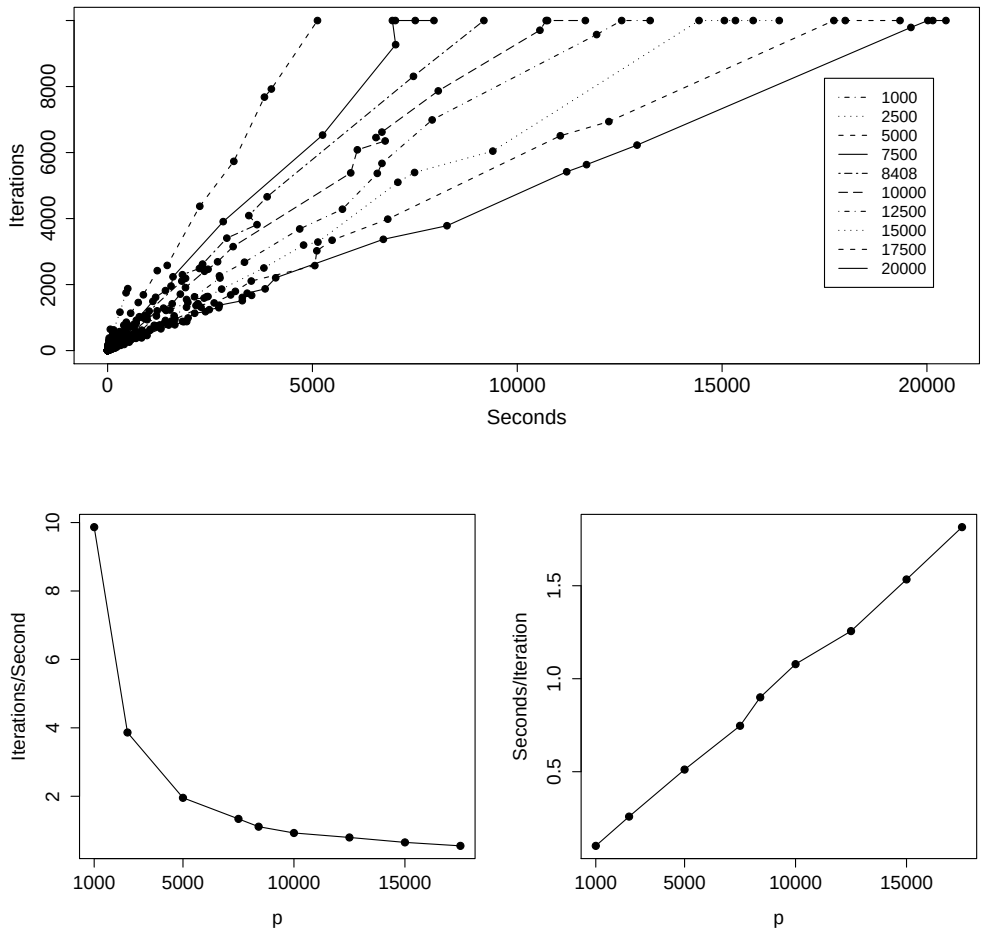

Figure 2.6: Run times for SSS from the simulation study in Section 2.5.3. The legend in the first plot refers to the total number of possible predictors, p.

Chapter 3

Bayesian Model Averaging via SSS

In this chapter I describe the special case of the normal linear model, showing how to obtain $p(\boldsymbol{\gamma}|y)$ under a class of conjugate priors, including specification of priors over model space in terms of variable inclusion probabilities. I subsequently describe how to use the output from SSS to perform approximate Bayesian model averaging for prediction and identification of key variables.

3.1 Normal Linear Regression

As introduced in Section 1.1.1, consider the normal linear regression model $Y = \mathrm{N}(X\boldsymbol{\beta}, \sigma^2 I_n)$, where Y is an $n \times 1$ response variable, $X = (x_1, \ldots, x_n)'$ is an $n \times p$ design matrix for the n samples, the x_i are $p \times 1$ vectors of covariate information, $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients, and σ^2 is the variance of the error term. Throughout I assume the data have been standardized in the sense that both the observed value of Y, y, and the columns of X have been scaled to have unit variance and zero mean, and therefore do not include an intercept term in the model.

3.1.1 Parameter Space Priors

I place priors on $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma^2)'$ that are consistent across models in the sense that they are derived from an encompassing model via conditioning. This follows Dobra *et al.* (2004) in assuming each observation $(y_i, x_i')'$ to have joint normality, $\mathrm{N}(0, \Sigma)$, with $(p+1) \times (p+1)$ covariance matrix Σ and a corresponding precision matrix $\Omega = \Sigma^{-1}$. For any given regression model, $\boldsymbol{\theta}$ can be written as a transformation of elements of Σ, and so placing a prior on Σ induces consistent priors across regression models.

Let $\Sigma \sim \mathrm{IW}(\delta, \tau I)$, with δ degrees of freedom and scale matrix τI_{p+1} (see Appendix A for notational details). Consider a regression model $\boldsymbol{\gamma}$. The vector $(y_i, x_{i,\gamma}')'$, consisting of y_i and those variables in x_i that are included in the regression model, has covariance matrix $\Sigma_{y,\gamma} \sim \mathrm{IW}(\delta, \tau I_{k+1})$, where $k = \mathbf{1}'\boldsymbol{\gamma}$. Partition $\Sigma_{y,\gamma}$ as

$$\Sigma_{y,\gamma} = \begin{bmatrix} \sigma_{yy} & \kappa_\gamma' \\ \kappa_\gamma & \Sigma_\gamma \end{bmatrix},$$

where σ_{yy} is a scalar and κ_γ is a $k \times 1$ vector. Standard normal theory (see, e.g., West and Harrison, 1997, Chapter 17) gives the conditional distribution of y_i given those x_i implied by $\boldsymbol{\gamma}$ as

$$p(y_i | x_{i,\gamma}, \Sigma_{y,\gamma}) = \mathrm{N}(x_{i,\gamma}'\boldsymbol{\beta}_\gamma \ , \ \sigma_\gamma^2),$$

where $\boldsymbol{\beta}_\gamma = \Sigma_\gamma^{-1}\kappa_\gamma$ and $\sigma_\gamma^2 = \sigma_{yy} - \kappa_\gamma'\Sigma_\gamma^{-1}\kappa_\gamma$. For brevity I write $\boldsymbol{\beta}$ for $\boldsymbol{\beta}_\gamma$ and σ^2 for σ_γ^2 when it is clear from the context that the parameters are defined with respect to a particular model. As derived in Appendix A, the priors implied by

the inverse Wishart specified above for a particular regression with k variables are

$$\sigma^2|\boldsymbol{\gamma} \sim \text{IG}\left(\frac{\delta+k}{2}, \frac{\tau}{2}\right), \tag{3.1}$$

$$\boldsymbol{\beta}|\sigma^2, \boldsymbol{\gamma} \sim \text{N}(0, \tau^{-1}\sigma^2 I_k). \tag{3.2}$$

3.1.2 Parameter Space Posteriors

Under the priors specified above, the posterior distribution of $\boldsymbol{\theta}$ for any given model can be factored as $p(\boldsymbol{\theta}|y, \boldsymbol{\gamma}) = p(\boldsymbol{\beta}|\sigma^2, y, \boldsymbol{\gamma})p(\sigma^2|y, \boldsymbol{\gamma})$, where routine calculations give

$$\sigma^2|y, \boldsymbol{\gamma} \sim \text{IG}\left(\frac{n+\delta+k}{2}, \frac{\tau+q_\gamma}{2}\right), \tag{3.3}$$

$$\boldsymbol{\beta}|\sigma^2, y, \boldsymbol{\gamma} \sim \text{N}(M_\gamma^{-1}X_\gamma' y\,,\, \sigma^2 M_\gamma^{-1}), \tag{3.4}$$

$M_\gamma = \tau I_k + X_\gamma' X_\gamma$ and $q_\gamma = y'y - y'X_\gamma M_\gamma^{-1} X_\gamma' y$. Routine calculations also give the marginal posterior distribution of $\boldsymbol{\beta}$:

$$\boldsymbol{\beta}|y, \boldsymbol{\gamma} \sim \text{T}_{n+\delta+k}\left(M_\gamma^{-1}X_\gamma' y\,,\, \left(\frac{\tau+q_\gamma}{n+\delta+k}\right) M_\gamma^{-1}\right), \tag{3.5}$$

a multivariate T distribution, where I write the density function of a $\text{T}_\nu(\mu, \Sigma)$ distribution of dimension d as

$$p(x) = \frac{\Gamma\left(\frac{\nu+d}{2}\right)/\Gamma\left(\frac{\nu}{2}\right)}{\nu^{d/2}\pi^{d/2}|\Sigma|^{1/2}}\left(1 + \frac{1}{\nu}(x-\mu)'\Sigma^{-1}(x-\mu)\right)^{-(\nu+d)/2}.$$

These posterior distributions will be used in Section 3.3 for computing model averaged predictions.

3.1.3 Marginal Likelihood

The marginal likelihood for a given model is

$$\begin{aligned} p(y|\boldsymbol{\gamma}) &= \int p(y|\boldsymbol{\theta},\boldsymbol{\gamma})p(\boldsymbol{\theta}|\boldsymbol{\gamma})\, d\boldsymbol{\theta} \\ &= \frac{\Gamma\left(\frac{n+\delta+k}{2}\right)/\Gamma\left(\frac{\delta+k}{2}\right)}{\pi^{n/2}\tau^{(n-k)/2}|M_\gamma|^{1/2}(1+q_\gamma/\tau)^{(n+\delta+k)/2}}. \end{aligned} \tag{3.6}$$

For the case of the null model,

$$p(y|\boldsymbol{\gamma}=\mathbf{0}) = \frac{\Gamma\left(\frac{n+\delta}{2}\right)/\Gamma\left(\frac{\delta}{2}\right)}{\pi^{n/2}\tau^{n/2}(1+y'y/\tau)^{(n+\delta)/2}}.$$

By Bayes theorem, the posterior probability of any model is $p(\boldsymbol{\gamma}|y) \propto p(y|\boldsymbol{\gamma})p(\boldsymbol{\gamma})$. The marginal likelihood will be used in Sections 3.3 and 3.4 in forming model averaged predictions and estimates of variable importance.

3.2 Model Space Prior Distributions

As discussed in Chapter 1, prior beliefs for problems with large p typically focus on sparsity, indicating that prior distributions over model space should place most of their mass on models with few predictor variables. Perhaps the most common Bayesian variable selection prior is the independent Bernoulli prior (George and McCulloch, 1993, 1997; Raftery *et al.*, 1997), namely

$$p(\boldsymbol{\gamma}) = \pi^k(1-\pi)^{p-k}, \tag{3.7}$$

where conditioning on π is supressed and $k = |\boldsymbol{\gamma}|$. Under this prior, the induced distribution over model dimension, k, is $\text{Bin}(p,\pi)$,

$$\Pr(|\boldsymbol{\gamma}| = k) = \binom{p}{k}\pi^k(1-\pi)^{p-k}. \tag{3.8}$$

"Default" Bayesian analyses often take $\pi = 1/2$ (e.g., Smith and Kohn, 1996), representing prior ignorance on possible inclusion for each variable separately. In problems with large p, the induced binomial distribution on model size indicates that when $\pi = 1/2$, on average $p/2$ variables will be included in the model *a priori*, removing focus from sparse areas of model space. As noted in Brown *et al.* (1998a) and Dobra *et al.* (2004), the prior expected model size is πp, and so priors of the form $\pi = k'/p$, where k' is small, maintain focus on sparse models as p increases. Experience has shown that values of k' between 2 and 100 are reasonable for examples with p on the order of 10^3 or 10^4. Unless otherwise specified, throughout I use the prior (3.7) with k' set sufficiently small.

3.2.1 Alternative Model Space Priors

Alternate priors based on (3.7) have been proposed (see, e.g., Chipman *et al.*, 2001). Separate values π_j could be used for each variable, however this requires either specifiying these p values *a priori* or including them in the model, a difficult task for large p problems. Other approaches involve placing a prior distribution on π, say Beta(α, β), resulting in

$$p(\boldsymbol{\gamma}) = \frac{B(\alpha + k, \beta + p - k)}{B(\alpha, \beta)},$$

where $B(\alpha, \beta)$ is the beta function. This approach was used by Kohn *et al.* (2001), who chose α and β by specifying the first two moments of π; Hans and Dunson (2005) adopted a similar approach in the context of a model selection problem related to piece-wise regression.

Other formulations first place a prior distribution on model size, $\Pr(|\boldsymbol{\gamma}| = k)$,

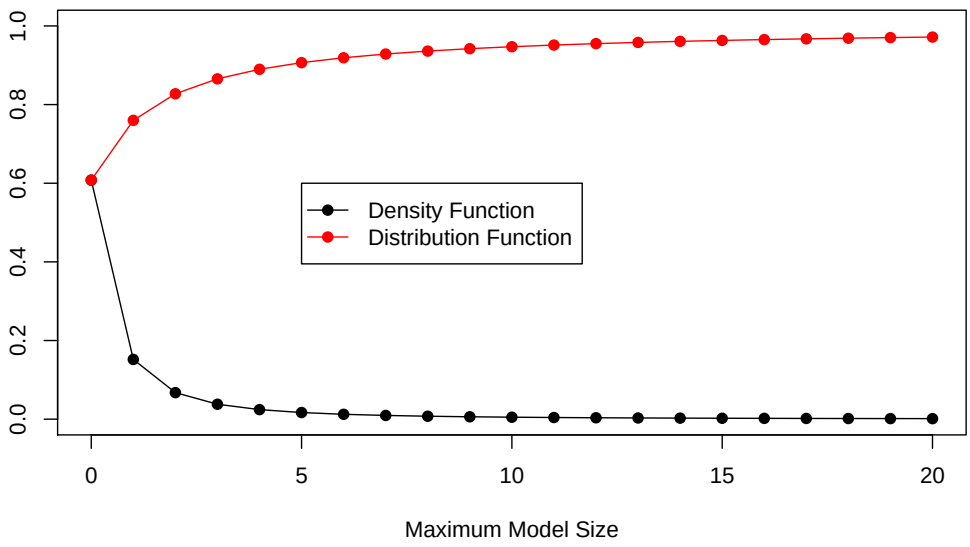

Figure 3.1: Density and cumulative distribution functions for the discrete Cauchy distribution (3.10).

and then conditionally on model size assume a uniform prior:

$$p(\boldsymbol{\gamma}) = \binom{p}{k}^{-1} \Pr(|\boldsymbol{\gamma}| = k). \tag{3.9}$$

Denison *et al.* (1998a,b) took such an approach, using a truncated Poisson distribution for the prior on model size.

Priors such as (3.7) have the tendency to stack up probability on a few dimensions. Priors on sparse models that do not distinguish as harshly between dimension can be constructed via mixtures. Hierarchically, given a maximum allowable model size k^*, consider specifying a discrete uniform distribution on the number of predictor variables k:

$$p(k = r|k^*) = \frac{\mathbf{1}(r \in \{0, \ldots, k^*\})}{k^* + 1}.$$

Under this formulation, all models up to a particular size k^* receive the same prior probability, $1/\sum_{j=0}^{k^*}\binom{p}{j}$. To balance the prior probabiliy on model size, a marginal prior for k^* should be used that tails off appropriately as model size increases. One example is a discrete version of the Cauchy distribution

$$p(k^*) = \frac{6}{\pi^2(k^*+1)^2}, \quad k^* = 0, 1, \ldots. \tag{3.10}$$

As shown in Figure 3.1, this distribution favors small values of k^* but has relatively fat tails: about 95% of the distribution's mass falls between zero and ten, but less than 99% of the mass falls between zero and 50. This coincides with prior belief in sparsity but allows for the possibility of larger sized models.

The marginal distribution on model size can be computed as

$$\begin{aligned}
p(k=r) &= \sum_{k^*=0}^{\infty} p(k=r|k^*)p(k^*) \\
&= \frac{6}{\pi^2}\sum_{k^*=0}^{\infty}\frac{\mathbf{1}(r\le k^*)}{k^*+1}\frac{1}{(k^*+1)^2} \\
&= \frac{6}{\pi^2}\sum_{k^*=r}^{\infty}\frac{1}{(k^*+1)^3} \\
&= \frac{6}{\pi^2}\left(\sum_{k^*=0}^{\infty}\frac{1}{(k^*+1)^3} - \sum_{k^*=0}^{r-1}\frac{1}{(k^*+1)^3}\right) \\
&= \frac{6}{\pi^2}\left(\sum_{x=1}^{\infty}\frac{1}{x^3} - \sum_{x=1}^{r}\frac{1}{x^3}\right) \\
&= \frac{6}{\pi^2}\left(\zeta(3) - \sum_{x=1}^{r}\frac{1}{x^3}\right).
\end{aligned} \tag{3.11}$$

$\zeta(\cdot)$ is the Riemann zeta function (Abramowitz and Stegun, 1972, Section 23.2) and $\zeta(3)$, sometimes refered to as Apéry's constant, is an irrational number

(Apéry, 1979) approximately equal to 1.2021. The corresponding probability for any given model is

$$p(\boldsymbol{\gamma}) = \binom{p}{k}^{-1} \frac{6}{\pi} \left(\zeta(3) - \sum_{x=1}^{k} \frac{1}{x^3} \right).$$

Values of (3.11) are shown in Figure 3.2, along with values for the standard prior (3.8). The mixture prior places much more mass on the null model, however, the probabilities on the other dimensions are relatively more diffuse for the mixture. This suggests the approach of placing a prior probability directly on the null model, and then normalizing the rest of the distribution with respect to this value.

A similar approach is to replace the discrete Cauchy distribution (3.10) with a Poisson distribution, $k^* \sim \text{Pois}(\lambda)$, with λ fixed *a priori*. The resulting marginal distribution on model size is

$$\begin{aligned}
p(k = r) &= \sum_{m=0}^{\infty} p(k = r|m)p(m) \\
&= \sum_{m=0}^{\infty} \frac{\lambda^m e^{-\lambda}}{m!} \frac{\mathbf{1}(r \in \{0, \ldots, m\})}{m+1} \\
&= \lambda^{-1} \sum_{m=0}^{\infty} \mathbf{1}(r \in \{0, \ldots, m\}) \frac{\lambda^{m+1} e^{-\lambda}}{(m+1)!} \\
&= \lambda^{-1} \sum_{m=r+1}^{\infty} \frac{\lambda^m e^{-\lambda}}{m!} \\
&= \lambda^{-1} \left(1 - \sum_{m=0}^{r} \frac{\lambda^m e^{-\lambda}}{m!} \right) \\
&= \lambda^{-1} \left(1 - \Pr(m \leq r|\lambda) \right), \qquad (3.12)
\end{aligned}$$

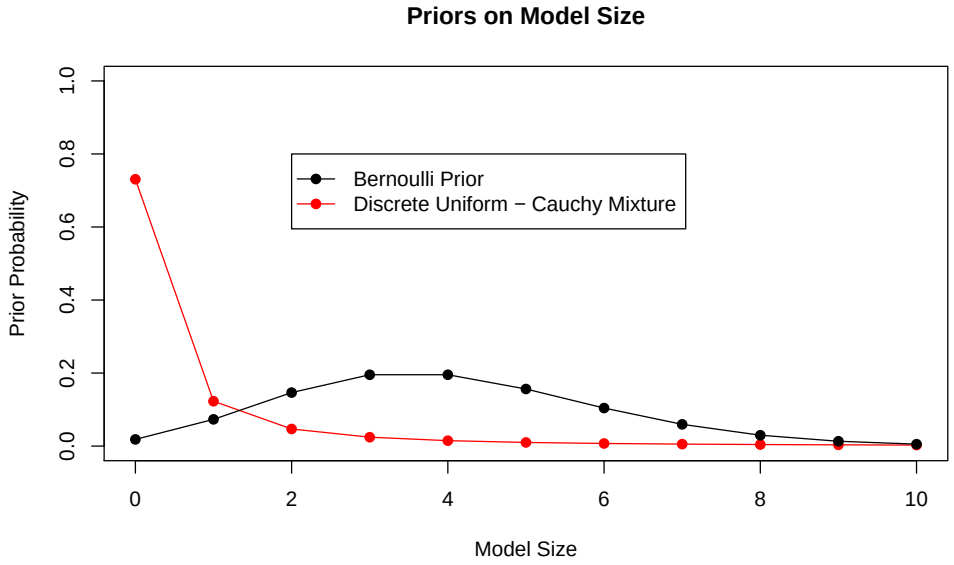

Figure 3.2: Comparison of the independent Bernoulli prior with the discrete uniform-Cauchy mixture. For the Bernoulli prior, $p = 8,408$ and $k' = 4$.

where $\Pr(\cdot|\lambda)$ is the Poisson cumulative distribution function with parameter λ. This marginal distribution is plotted in Figure 3.3 for several values of λ. Values of λ between three and seven appear to provide reasonable penalty on dimension.

3.2.2 Reformulating the Model Space

Rather than placing prior distributions on individual models, it is worthwhile to consider placing prior distributions on *sets* of models. The idea is to form a collection of models that are similar in some sense, and then write the likelihood conditioned on one of these sets as a mixture of the models in the set.

Let m^k by any model of size k. Define a metamodel, M^*, to be a collection of models,

$$M^* = \{\emptyset, m^1, m^2, \ldots, m^{k^*}\},$$

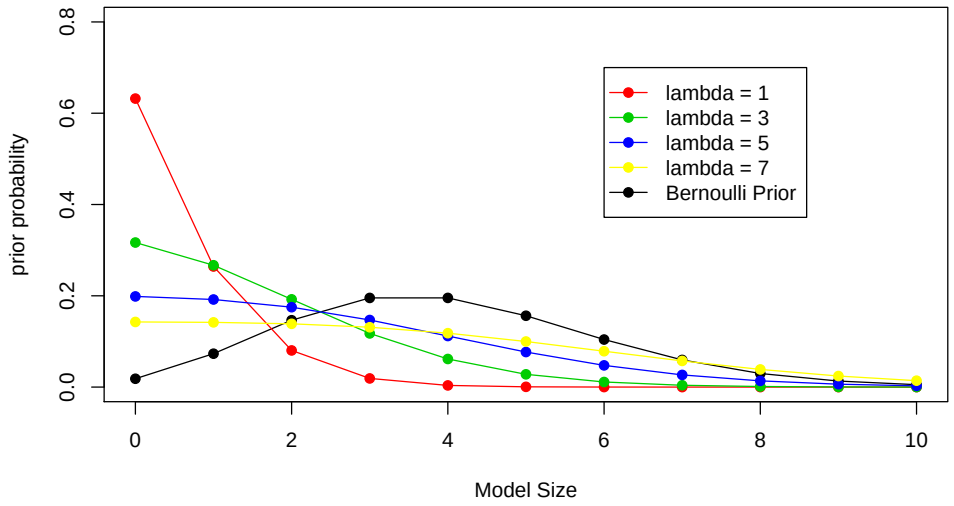

Figure 3.3: Comparison of the prior on model size induced by both the discrete uniform-Poisson mixture distribution and the Bernoulli prior (with $p = 8,408$ and $k' = 4$).

for some fixed value $k^* \leq p$. Conditionally on a metamodel write

$$p(y|M^*) = \sum_{j=0}^{k^*} \alpha_j p(y|m^j),$$

where $m^0 = \emptyset$ and $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_{k^*})'$ is a probability vector.

Impose structure on M^* in the following way: for any model $m^i \in M^*$, if $\gamma_j = 1$ for the model $\boldsymbol{\gamma}$ corresponding to m^i, then $\gamma_j = 1$ for each of the models $\boldsymbol{\gamma}$ corresponding to $m^l \in M^*$, $l > i$. In other words, a metamodel is a set of nested models. Under this structure, the size of the metamodel space $\mathcal{M}^*$ is

$$|\mathcal{M}^*| = \frac{p!}{(p-k^*)!} = k^*! \binom{p}{k^*}.$$

If $k^* = p$, the metamodel space is much larger than the standard model space Γ.

Conditioning on this structure, *a priori* assume a discrete uniform prior distribution over the metamodel space:

$$p(M^*) = \frac{(p-k^*)!}{p!} = \frac{1}{k^*!}\binom{p}{k^*}^{-1}, \quad \forall M^* \in \mathcal{M}^*.$$

Given a metamodel, M^*, the prior probability of each of its components is given by the mixing weights defined by $\boldsymbol{\alpha}$:

$$p(m^k|M^*) = \alpha_k.$$

Accordingly, marginal priors can be computed:

$$\begin{aligned}
p(\emptyset) &= \sum_{M^* \in \mathcal{M}^*} p(\emptyset|M^*)p(M^*) \\
&= \sum_{M^* \in \mathcal{M}^*} [\alpha_0 (p-k^*)^!/p!] \\
&= [p!/(p-k^*)!][\alpha_0 (p-k^*)!/p!] \\
&= \alpha_0 \\
&= \alpha_0 \binom{p}{0}^{-1},
\end{aligned}$$

$$\begin{aligned}
p(m^1) &= \sum_{M^* \in \mathcal{M}^* : m^1 \in M^*} p(m^1|M^*)p(M^*) \\
&= \sum_{M^* \in \mathcal{M}^* : m^1 \in M^*} [\alpha_1 (p-k^*)!/p!] \\
&= [p!/(p(p-k^*)!][\alpha_1 (p-k^*)!/p!] \\
&= \alpha_1/p \\
&= \alpha_1 \binom{p}{1}^{-1},
\end{aligned}$$

$$
\begin{aligned}
p(m^2) &= \sum_{M^* \in \mathcal{M}^* : m^2 \in M^*} p(m^2|M^*)p(M^*) \\
&= \sum_{M^* \in \mathcal{M}^* : m^2 \in M^*} [\alpha_2 (p - k^*)!/p!] \\
&= \alpha_2 \binom{p}{2}^{-1}.
\end{aligned}
$$

Note that if $k^* = p$, the marginal model probabilities above correspond to the formulation (3.9), where $\boldsymbol{\alpha}$ corresponds to the prior distribution on model size. Hence we can recover our original independent Bernoulli prior (3.7) by setting $\boldsymbol{\alpha}$ based on a $\text{Bin}(p, \pi)$ distribution. Specification (3.7) can then be thought of either as treating each variable as an independent Bernoulli random variable, or as placing a uniform distribution on structured sets of similar models, where the elements of each set are weighted appropriately.

3.3 Variable Identification and Prediction

Having fully specified the probability model, models discovered by SSS can now be compared and combined by letting the score introduced in Section 2.3 be the unnormalized posterior probability of a model, $p(\boldsymbol{\gamma}|Y) \propto S(\boldsymbol{\gamma}) = p(y|\boldsymbol{\gamma})p(\boldsymbol{\gamma})$. Using the list of top models discovered by SSS, Γ^*, the relative importance of each predictor variable x_j is measured by computing

$$
\tilde{p}(\gamma_j = 1|y) = C^{-1} \sum_{\boldsymbol{\gamma} \in \Gamma^*} \mathbf{1}(\gamma_j = 1) S(\boldsymbol{\gamma}), \tag{3.13}
$$

where the normalizing constant is the posterior mass contained in Γ^*, $C = \sum_{\boldsymbol{\gamma} \in \Gamma^*} S(\boldsymbol{\gamma})$. If we could have explored the entire space (so that $\Gamma^* = \Gamma$), then (3.13) would represent the posterior probability of variable inclusion for variable

x_j. Rather, as we have only explored and recorded some part of the model space, (3.13) represents the posterior probability of variable inclusion *conditioned* on the set Γ^*. Of course, if the set of models not discovered by SSS had near zero probability relative to Γ^*, then (3.13) would indeed be the posterior probability desired. Due to the size of the model space for large p examples, this is unlikely to be the case, and so $\tilde{p}(\gamma_j = 1|y)$ should be viewed as a measure of the relative importance of variables x_j in the context of the top predictive models found. Similarly, measure the relative importance of individual models discovered by computing

$$\tilde{p}(\boldsymbol{\gamma}|y) = C^{-1}S(\boldsymbol{\gamma}). \tag{3.14}$$

A measure of posterior importance of model size (a measure of how much sparsity the data support with respect to the prior) can be computed as

$$\tilde{p}(|\boldsymbol{\gamma}| = k) = C^{-1} \sum_{\boldsymbol{\gamma} \in \Gamma^*} \mathbf{1}(|\boldsymbol{\gamma}| = k) S(\boldsymbol{\gamma}). \tag{3.15}$$

If desired, higher-order variable inclusion probabilities can be computed, for example

$$\tilde{p}(\gamma_i = \gamma_j = 1|y) = C^{-1} \sum_{\boldsymbol{\gamma} \in \Gamma^*} \mathbf{1}(\gamma_i = \gamma_j = 1) S(\boldsymbol{\gamma}) \tag{3.16}$$

for pairwise inlusion.

Having conditioned on the list of top models, Γ^*, the distribution of model averaged fitted values can be simluated from, allowing for Monte Carlo integration. To obtain the samples needed first draw $f = 1, \ldots, F$ times from the discrete distribution implied by (3.14) over Γ^*, providing samples $\boldsymbol{\gamma}^{(f)}$. For each sampled model, draw $\sigma^{2(f)}$ from (3.3), followed by a draw of $\boldsymbol{\beta}^{(f)}$ from (3.4) conditionally on $\sigma^{2(f)}$. The fitted value $\mu^{(f)} = X_{\gamma^{(f)}}\boldsymbol{\beta}^{(f)}$ is then stored. Summaries of this

distribution can then be computed via Monte Carlo integration. See Figure 4.7 for an example. Alternately one could sample $\boldsymbol{\beta}^{(f)}$ marginally from (3.5), bypassing the draw of $\sigma^{2(f)}$. The distribution of model averaged fitted values can also be computed analytically as a mixture over all the models in Γ^*, where the mixing weights are the values $\tilde{p}(\boldsymbol{\gamma}|y)$, avoiding the introduction of Monte Carlo error.

3.4 Cross-Validated Prediction

Leave-one-out cross-validation (LOOCV) provides an often useful predictive model evaluation tool. Ideally, one would run SSS n times, leaving out the ith observation on run i, compiling lists Γ_i^*, and then simulating from the predictive distributions for the left-out observation in a manner similar to that described above for the fitted values. For even small values of n, this becomes computationally infeasible as SSS may take hours to run for each hold out observation.[1] Instead, I take the view that, had SSS been rerun holding out each observation in turn, it is not the *elements* in the lists Γ_i^* that would be different from the elements in Γ^*, but that it is the *weights* $p(y|\boldsymbol{\gamma})$ that would be different. I discuss two methods for using Γ^* for LOOCV prediction: importance sampling based methods and a brute-force calculation method.

3.4.1 Importance Sampling LOOCV

Smith and Gelfand (1992) and Gelfand, Dey, and Chang (1992) introduced an importance sampling technique for obtaining samples from $p(\boldsymbol{\theta}|\boldsymbol{\gamma}, y_{-i})$ based on a

[1] It should be noted, though, that in principle this is not problematic. Given access to enough processors, the SSS runs can be made independently in parallel.

sample $\{\boldsymbol{\theta}^{(j)}\}$ from $p(\boldsymbol{\theta}|\boldsymbol{\gamma}, y)$ by resampling from the latter using the weights

$$w_i^{(j)} = \frac{1}{p(y_i|\boldsymbol{\theta}^{(j)}, \boldsymbol{\gamma})}, \tag{3.17}$$

which is derived from Bayes theorem

$$p(\boldsymbol{\theta}|\boldsymbol{\gamma}, y_{-i}) \propto \frac{p(\boldsymbol{\theta}|\boldsymbol{\gamma}, y)}{p(y_i|\boldsymbol{\theta}, \boldsymbol{\gamma})}. \tag{3.18}$$

A key to this approach is the conditional independence of y_i given $(\boldsymbol{\gamma}, \boldsymbol{\theta})$. This importance sampling method will be accurate to the extent that the two distributions of interest are similar, i.e. the more dissimilar $p(\boldsymbol{\theta}|y, \boldsymbol{\gamma})$ and $p(\boldsymbol{\theta}|y_{-i}, \boldsymbol{\gamma})$ are, the worse the approximation. Peruggia (1997) gives conditions under which the resampling weights will have finite variance for Bayesian linear models.

Of interest in the context of SSS is sampling from $p(\boldsymbol{\gamma}, \boldsymbol{\theta}|y_{-i})$ to measure predictive accuracy as in Gelfand (1996). Extending the above approach, we can use Bayes theorem

$$p(\boldsymbol{\gamma}, \boldsymbol{\theta}|y_{-i}) \propto \frac{p(\boldsymbol{\gamma}, \boldsymbol{\theta}|y)}{p(y_i|\boldsymbol{\gamma}, \boldsymbol{\theta})} \tag{3.19}$$

to find the weights

$$w_i^{(j)} = \frac{1}{p(y_i|\boldsymbol{\gamma}^{(j)}, \boldsymbol{\theta}^{(j)})} \tag{3.20}$$

needed to resample from an original sample $\{\boldsymbol{\gamma}^{(j)}, \boldsymbol{\theta}^{(j)}\}$ from $p(\boldsymbol{\gamma}, \boldsymbol{\theta}|y)$. Again, this will only be accurate to the extent that the two distributions are fairly similar. From experience with SSS output from large datasets, it is often the case that there are only a few nontrivial weights on the sampled values $\{\boldsymbol{\gamma}^{(j)}, \boldsymbol{\theta}^{(j)}\}$, indicating that the approximation over both models and parameters may not be good.

It might be the case that while the distribution $p(\boldsymbol{\gamma}, \boldsymbol{\theta}|y)$ is not a good approximation to $p(\boldsymbol{\gamma}, \boldsymbol{\theta}|y_{-i})$ for some observations, the conditional distributions $p(\boldsymbol{\theta}|\boldsymbol{\gamma}, y)$ may be "closer" to the conditional distributions $p(\boldsymbol{\theta}|\boldsymbol{\gamma}, y_{-i})$ for those observations. If this is the case, the importance sampling procedure might be more accurate if a two step approach is taken. For each observation i:

Step 1 Obtain a sample $\{\boldsymbol{\gamma}_{-i}^{(j)}\}$ from $p(\boldsymbol{\gamma}|y_{-i})$ by resampling draws from $p(\boldsymbol{\gamma}|y)$;

Step 2 For each $\boldsymbol{\gamma}_{-i}^{(j)}$ use the importance sampling technique based on (3.17) and (3.18) to draw a sample $\{\boldsymbol{\theta}_{-i}^{(j)}|\boldsymbol{\gamma}_{-i}^{(j)}\}$ from $p(\boldsymbol{\theta}|\boldsymbol{\gamma}_{-i}^{(j)}, y_{-i})$.

The resulting draws will be approximately from $p(\boldsymbol{\gamma}, \boldsymbol{\theta}|y_{-i})$.

To accomplish Step 1, draws need to be made from

$$p(\boldsymbol{\gamma}|y_{-i}) \propto \frac{p(\boldsymbol{\gamma}|y)}{p(y_i|\boldsymbol{\gamma}, y_{-i})}.$$

Draws can be made directly from the numerator, and hence if the denominator can be approximated the the draws can be weighted by $w_i = 1/p(y_i|\boldsymbol{\gamma}, y_{-i})$ and resampled. The denominator can be written as

$$\begin{aligned} p(y_i|\boldsymbol{\gamma}, y_{-i}) &= \int p(y_i|\boldsymbol{\theta}, \boldsymbol{\gamma}, y_{-i}) p(\boldsymbol{\theta}|\boldsymbol{\gamma}, y_{-i})\ d\boldsymbol{\theta} \\ &= \int p(y_i|\boldsymbol{\theta}, \boldsymbol{\gamma}) p(\boldsymbol{\theta}|\boldsymbol{\gamma}, y_{-i})\ d\boldsymbol{\theta}. \end{aligned} \tag{3.21}$$

Noting that we can obtain samples from $p(\boldsymbol{\theta}|\boldsymbol{\gamma}, y_{-i})$ using the importance sampling techniques of (3.17) and (3.18), a Monte Carlo estimate of (3.21) can by computed by

$$p(y_i|\boldsymbol{\gamma}, y_{-i}) = \mathbb{E}_{p(\theta|\gamma, y_{-i})}\Big[p(y_i|\boldsymbol{\theta}, \boldsymbol{\gamma})\Big] \approx \frac{1}{J}\sum_{j=1}^{J} p(y_i|\boldsymbol{\theta}^{(j)}, \boldsymbol{\gamma}),$$

where the $\boldsymbol{\theta}^{(j)}$ are those drawn from $p(\boldsymbol{\theta}|\boldsymbol{\gamma}, y_{-i})$ via importance sampling, and the $\boldsymbol{\gamma}$ are those drawn originally from $p(\boldsymbol{\gamma}|y)$. The original draws $\{\boldsymbol{\gamma}_{-i}^{(j)}\}$ can now be weighted and resampled to obtain draws from $p(\boldsymbol{\gamma}|y_{-i})$ for each observation as required in Step 1.

To perform Step 2 for a single observation i, draws need to be made from $p(\boldsymbol{\theta}|\boldsymbol{\gamma}, y_{-i})$ for each of the sampled models. This can be done using importance sampling via (3.17) and (3.18).

This approach, while it may provide a better approximation to $p(\boldsymbol{\gamma}, \boldsymbol{\theta}|y_{-i})$ than (3.19) and (3.20), is computationally intensive, requiring several importance sampling steps and a separate Monte Carlo integration, all of which may need many draws in order to be stable. Further, the computation must be done separately for each observation (although note again that the computations may be done independently in parallel), leading to potentially long computation times for approximate results.

3.4.2 Direct LOOCV Calculations

Rather than using an importance sampling approach to approximate the case-deletion distributions, a direct approach is to recompute the marginal likelihoods for each model in Γ^* n times, each time holding out a different sample, giving us $p_i(y|\boldsymbol{\gamma})$ for each $\boldsymbol{\gamma} \in \Gamma^*$ and the corresponding n lists of top models Γ_i^*. This can be done independently in parallel, requiring little computational burden so long as a parallel distributed computing environment is available. The approximation to the posterior distribution of the model space (3.14) can be replaced with

$$\tilde{p}_i(\boldsymbol{\gamma}|y) = C_i^{-1} S_i(\boldsymbol{\gamma})$$

for each hold out observation, where $S_i(\boldsymbol{\gamma}) = p_i(y|\boldsymbol{\gamma})p(\boldsymbol{\gamma})$ and $C_i = \sum_{\boldsymbol{\gamma}\in\Gamma_i^*} S_i(\boldsymbol{\gamma})$. The predictive distributions for the hold-out observations can then be sampled as described above. As computation for each hold-out sample is done independently of the others, the leave-one-out cross-validated predictive distributions can be simulated in parallel, requiring little computation even when Γ^* contains many models, assuming that computation of the marginal likelihood can be done quickly, as is required when performing SSS.

Chapter 4

Example: Glioblastoma Survival Study

A contemporary example that demonstrates the usefulness of high dimensional model search is the analysis of data from gene expression arrays. In this section I present an analysis using gene expression data from a survival study in brain cancer based at the W.M. Keck Center for Neuro-Oncology at Duke University.

4.1 Description of the Data

The study consists of $n = 41$ patients over age 50 diagnosed with glioblastoma, a particularly lethal form of brain cancer associated with relatively short survival times. In the general population the median survival time is about 10 to 12 months (Legler *et al.*, 1999), although significant variability is observed. It is of interest to explore possible biological explanations for such variability through the analysis of gene expression data in an attempt to identify sets of genes that may serve as indicators of survival time.

Collected from each patient was their survival time in days (measured from initial diagnosis), along with a set of clinical variables including age, sex, race and

type of treatment, and a tumor specimen. All of the patients in the sample are deceased, and hence there is no censoring information. From the tumor specimen, gene expression data is available on Affymetrix human U133A microrarrays, processed using the current standard RMA method (Irizarry *et al.*, 2003a,b), to generate summary estimates of expression levels of each gene in each sample, implemented using the Bioconductor software suite[1].

The U133A GeneChip provides expression information for over $20,000$ probe sets, where a probe set is identified with a particular gene and there may be multiple (slightly different) probe sets that correspond to the same gene. An initial screening of the probe sets was performed to (i) exclude probes whose estimated expression levels did not vary appreciably across samples and (ii) exclude probe sets whose estimated expression levels are "in the noise". The first was accomplished by excluding probes whose levels did not vary at least four-fold across samples, and the second was accomplished by excluding probes whose maximum level did not exceed seven (on the $\log_2$ scale). This provides a total of $p = 8,408$ genes/probe sets to serve as the potential predictors of survival time. A more detailed description of the data can be found in Rich *et al.* (2005).

4.2 Exploratory Analysis

Table 4.1 provides summary statistics for survival time and the age clinical variable, and Figure 4.1 shows box plots of the survival times. The analysis below is performed on the natural logarithm of the survival times, and so summaries of this transformation are also presented. Figure 4.2 is a scatter plot of log survival

[1]Bioconductor is open source software for bioinformatics data analysis, available at http://www.bioconductor.org.

Table 4.1: Summary statistics for for the Keck dataset. "Survival" is the survival time in days, with statistics in years also presented.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Survival	40	207	398	465.7	606	1696
(years)	0.11	0.57	1.09	1.28	1.66	4.65
Log Survival	3.689	5.333	5.986	5.836	6.407	7.436
Age	50	57	62	63.05	68	79

Table 4.2: Genes with the five largest values of R^2 for univariate regressions with log survival time as the dependent variable.

Gene	ENSA	CRABP1	TF	TFa	MTMR2
R^2	0.245	0.231	0.227	0.227	0.218

versus age, with a least squares line added. The p-value for the coefficient on age is 0.097, presenting some evidence of decreased survival for older patients. The survival times are fairly evenly distributed between zero and three years, with one long term survivor of over four years.

It is common for cancer clinicians to identify potentially interesting genes by looking at the R^2 value for the univariate regression of each gene on survival time. A histogram of these values is shown in Figure 4.3; no models appear to be particularly noteworthy. Table 4.2 lists the genes with the largest five values of R^2.

4.3 Small Subsets Regression Analysis

A log-normal survival model was used to assess the association between the genetic data and survival time,

$$\log y = \boldsymbol{\beta}_\gamma X_\gamma + \boldsymbol{\epsilon}, \tag{4.1}$$

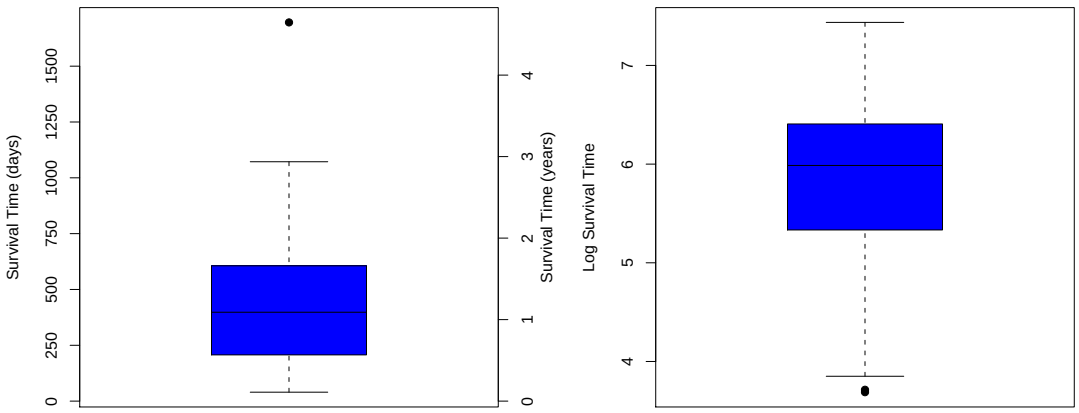

Figure 4.1: Boxplots of survival time for the Keck data.

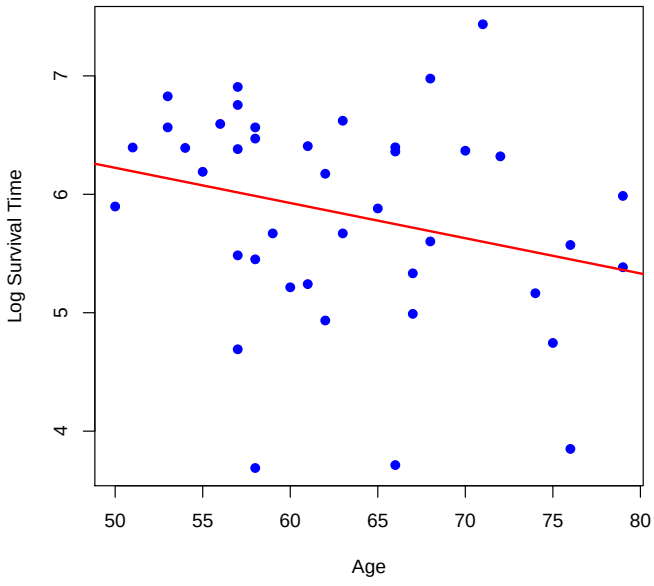

Figure 4.2: Log survival vs. age, with least squares line added.

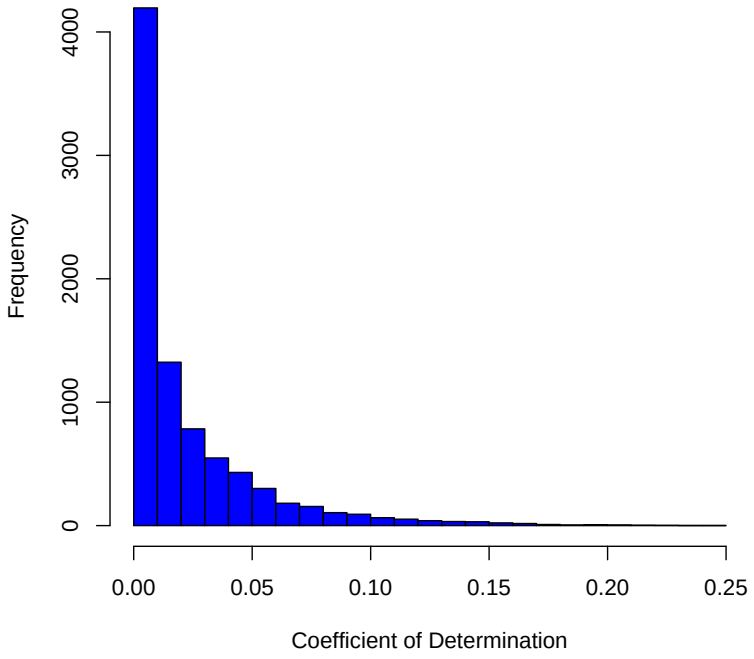

Figure 4.3: Values of R^2 for the $8,408$ univariate regressions of log survival time on each gene.

where $\boldsymbol{\gamma}$ indexes some subset of the genes and $\boldsymbol{\epsilon} \sim \mathrm{N}(0, \sigma^2 I_n)$. The $41 \times 8,408$ data matrix X has been centered and scaled, and hence there is no intercept term in (4.1).

As the focus is on sparse models, I take the prior distribution to be as in (3.7), with $\pi = 10/p$. For the prior distribution on the parameter space, I take $\tau = 1$ and $\delta = 3$, as described in Section 3.1.1. The value $\tau = 1$ is chosen due to the common scale of the predictor variables, and $\delta = 3$ is chosen for compatability with related work in constructing Gaussian graphical models for this data (see Jones *et al.*, 2005; Rich *et al.*, 2005).

To explore the model space and build a list of small subset regression models, I ran SSS for 40,000 iterations, saving the top one million models. The annealing parameters used were $\alpha_{1\circ} = 0.6$, $\alpha_{1-} = 1$, $\alpha_{1+} = 0.8$ and $\alpha_2 = 0.4$, which were chosen based on experience. Twenty six processors were used (one main processor with 25 compute nodes), and the resulting run time was just under 12 hours.

4.4 SSS Results

The top one million models contain a mix of zero through six variable models, as shown in Table 4.3. Using (3.15), I computed the estimated posterior probability of model size $|\boldsymbol{\gamma}|$, a relative measure of the importance of model size based on the set Γ^*. As seen in Table 4.3, the data give most support under the model to regressions on four, five and three variables, in that order. No model of size greater than six was found by SSS to belong in the top one million models. The null model has relative posterior probability of 0.03, which is a non-trivial amount of mass for a single model in this set; however, as described below, the data give large support to sets of models containing particular variables, indicating that model *selection* methods that focus on choosing one model may be less appropriate than model *averaging* methods that combine information across models.

Figure 4.4 gives a sense of how the prior is penalizing dimension. It appears as though that starting at dimension three, overall the prior is penalizing about the same amount as the gain in marginal likelihood from adding an extra variable. Sensitivity to the model space prior distribution is discussed in Section 4.5.

Contionally on Γ^*, four genes were found to have posterior inclusion probability (3.13) greather than 0.1, as shown in the diagonal entries of Table 4.4. Two of these variables, DCX and DCXa, are two probesets representing one gene, DCX.

The sample correlation between the expression levels for these two probes is 0.956, and in fact there are many "replicate" models in Γ^*, where DCXa replaces DCX. They rarely entered into a model together; their pairwise inclusion probability is less than 0.001, as see in the appropriate off-diagonal entry of Table 4.4. It appears as though DCX and DCXa were only found to be interesting in models also containing SPARC, as their pairwise inclusion probabilities are the same as their marginal inclusion probabilities. Of the 273,248 (95,824) models found containing DCX (DCXa), all but 134 (123) of them also contained SPARC.

The key gene that emerges from the SSS analysis is SPARC (Osteonectin), which dominates the list of models. Rich *et al.* (2005) give a description of SPARC:

> Osteonectin/SPARC was originally discovered as an important component of bone (Termine *et al.*, 1981) but is also expressed in epithelia exhibiting high rates of turnover (gut, skin, and glandular tissue), as well as vascular smooth muscle cells and endothelial cells. In addition to its normal physiological role, Osteonectin/SPARC is abnormally expressed in cancers. Many cancers, including cancers of the gastrointestinal tract, breast, lung, kidney, adrenal cortex, prostate, bladder and meninges (Porter *et al.*, 1995; Rempel *et al.*, 1998, 1999; Thomas *et al.*, 2000; Bellahcene and Castronovo, 1995), express increased SPARC levels that are associated with a conversion to invasive and metastatic tumors.

In terms of the other important genes, DCX is a gene that is related to both smoothness in brains and the development of abnormal amounts of gray matter; SEMA3B, along with DCX, is known to regulate neuronal migration (Rich *et al.*, 2005).

The bottom row of Table 4.4 gives the rank of the absolute value of the sample correlation of the top seven genes with log survival time. While a few of these genes, notably SPARC and NELL1, may have been found by looking at the top univariate regressions, genes such as DCX, KIAA0831 and HMP19, which appear in many of the best models, simply would not have been found in this manner.

Table 4.5 lists the top fifteen models and their comprising genes. The genes

are listed in order of their posterior inclusion probabilities and the corresponding model probabilities are given in the final row. All but four of the top 15 models include both SPARC and SEMA3B. Their pair-wise inclusion probability is 0.633, indicating that models consisting in part of these two genes play a dominant role. The highest probability model consists of SPARC, SEMA3B and DCX, who have a three-way inclusion probability of 0.345 which rises to 0.475 if models with DCXa replacing DCX are also included. Of interest is model 11, which is the only model in the top 15 that does not contain any of the the top five variables. This model contains the gene RAC1, a botulinum toxin related to RAS, a well-known oncogene. None of these eight genes (SPARC, SEMA3B, DCX, KIAA0831, NELL1, FABP5, RAC1 and MCAM) are appreciably correlated with each other; the largest absolute correlation between any pair does not exceed 0.4.

As stated above, SSS was run for 40,000 iterations, which based on experience is a reasonable run time. Figure 4.5 shows the rate at which posterior mass accumulated in Γ^* as a function of both time (in hours) and iterations. A large percentage of the mass is found very quickly, with 80% discovered in less than two hours. Small amounts of mass continue to accumulate slowly for the remainder of the run, which indicates that had we run SSS longer it would likely keep finding new models to swap into Γ^*. It is very difficult to say whether or not any of these models would be among the best in Γ^* or not. Figure 4.6 shows the log posterior values for models in Γ^* as a function of the iteration in which they were found, with the best 1000 models highlighted. Most of these were found very quickly, although there are quite a few that were found after the search had run for many iterations. The eleventh best model, containing NELL1, FABP5, RAC1 and MCAM, was found at iteration 32,786. An appropriate rubric for determining when to stop

Table 4.3: Posterior probability of model size, k, conditioned on the top 1,000,000 models found by SSS. The highest posterior probability dimension is highlighted. The character $*$ indicates a value < 0.001.

$k:$	0	1	2	3	4	5	6
# of models	1	8,408	120,614	116,545	275,203	477,880	1,349
$\tilde{p}(\lvert\gamma\rvert = k \lvert y)$	0.030	0.084	0.062	0.165	0.421	0.237	$*$

Table 4.4: Genewise and pairwise inclusion probabilities for the top seven genes. The diagonal elements are the quantities $\tilde{p}(\gamma_j = 1|y)$, and the off-diagonal elements are the quantities $\tilde{p}(\gamma_i = \gamma_j = 1|y)$. The character $*$ indicates a value < 0.001. The final row gives the rank of the absolute value of the correlation of the gene with log survival.

	SPARC	SEMA3B	DCX	DCXa	KIAA0831	HMP19	NELL1
SPARC	0.797	0.633	0.348	0.133	0.062	0.048	0.019
SEMA3B		0.634	0.345	0.130	0.062	0.047	0.002
DCX			0.348	$*$	0.059	0.002	0.002
DCXa				0.133	0.002	$*$	$*$
KIAA0831					0.062	$*$	$*$
HMP19						0.048	$*$
NELL1							0.034
Corr. Rank	15	59	159	214	5500	424	20

Table 4.5: Top 15 models from the Keck example. Values of $\tilde{p}(\boldsymbol{\gamma}|y)$ for these models are given below. The genes are in decreasing order by $\tilde{p}(\gamma_j = 1|y)$, and their ranks are given in the leftmost column.

		Model														
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	SPARC	•		•	•	•	•	•	•	•	•		•	•	•	•
2	SEMA3B	•		•	•	•		•		•	•		•	•	•	•
3	DCX	•			•					•	•				•	
4	DCXa			•												
5	KIAA0831				•						•					
6	HMP19												•			
7	NELL1								•			•				
8	C13orf7					•										
10	FABP5											•				
11	ZNF217					•										
12	L1CAM						•							•		
13	SOX4								•							
14	STMN2							•								
16	RAC1											•				
17	TRIM9									•						
20	MCAM											•				
21	IPO4														•	
25	SNX1										•					
26	HRI															•
29	GSTA4															•
Model Size		3	0	3	4	4	2	3	3	4	5	4	3	3	4	4
Post. Prob.		0.043	0.030	0.024	0.013	0.006	0.003	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.001	0.001

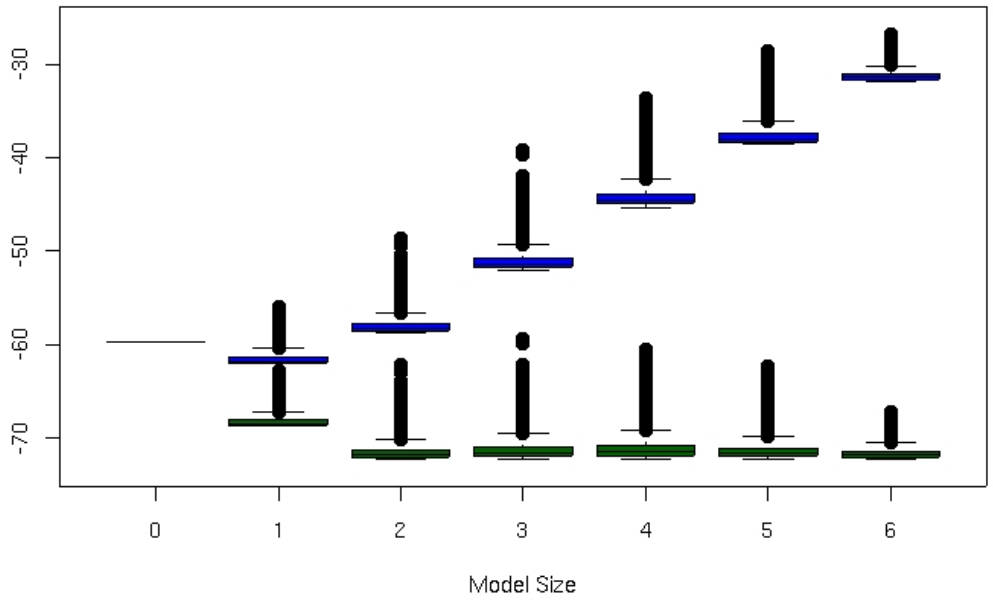

Figure 4.4: The top one million models found by SSS. The blue boxplots are values of the log marginal likelihoods, and the green boxplots are values of the unnormalized log posterior, $\log p(y|\boldsymbol{\gamma}) + \log p(\boldsymbol{\gamma})$, which were the values used to construct Γ^*.

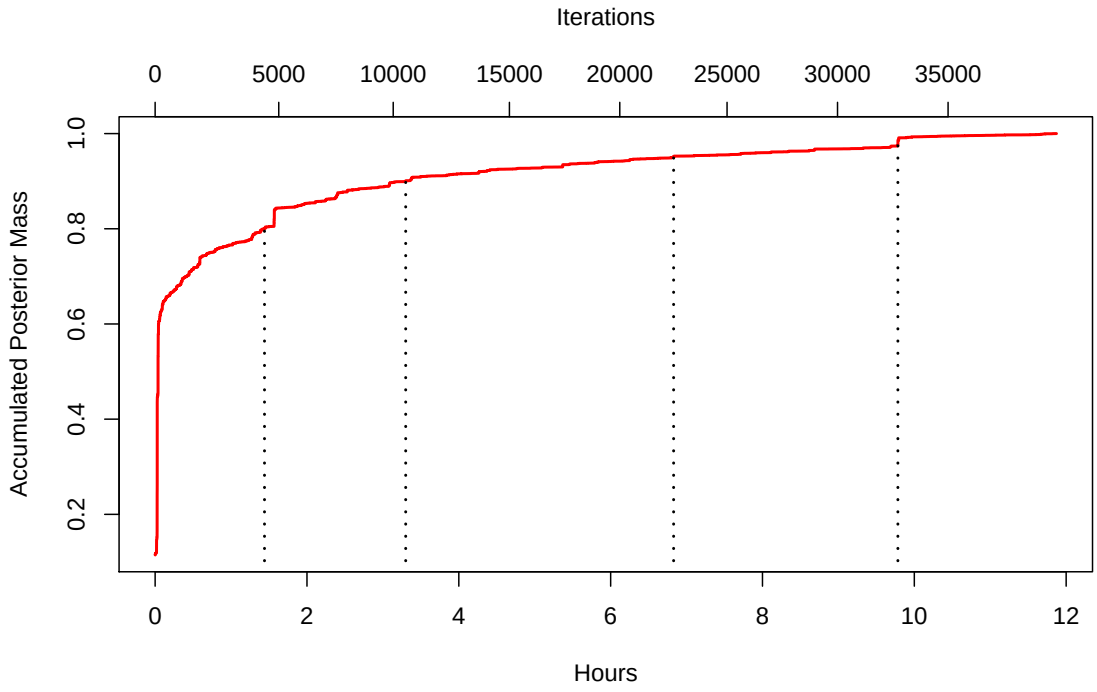

Figure 4.5: Accumulation of posterior mass in Γ^* for the example with $\pi = 10/p$. The dotted lines denote 80%, 90%, 95% and 97.5% of the mass.

the search would perhaps involve both the rate at which mass is accumulating in Γ^* and also how many iterations have passed since a model belonging to the set of the top x has been found. Of course, there would be no guarantee that all the best models had been found.

4.4.1 Assessing Model Fit

To assess the fit of the model to the data, I first sampled from the posterior distribution of β, $p(\beta|y)$, as described in Section 3.3, by first sampling $\boldsymbol{\gamma}^{(f)}$ models, $f = 1, \ldots, F$, from $\tilde{p}(\boldsymbol{\gamma}|y)$, constructed based on Γ^*, and then for each model by

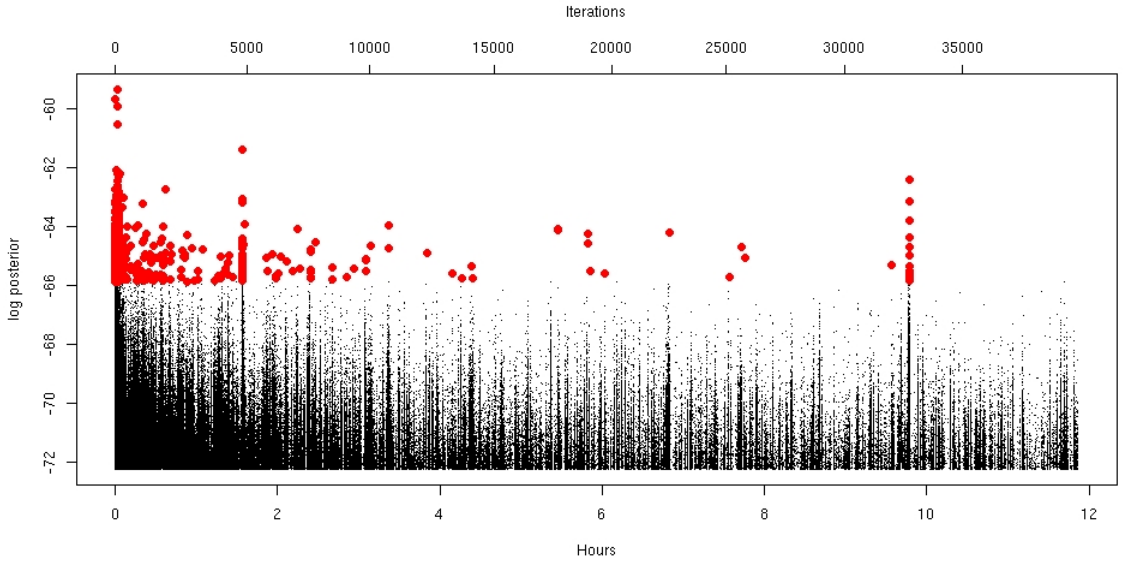

Figure 4.6: The values $\log p(y|\boldsymbol{\gamma}) + \log p(\boldsymbol{\gamma})$ for the models in Γ^* as a function of the iteration in which they were found. The top 1000 models are plotted as red circles.

sampling a value $\beta^{(f)}$ from

$$p(\beta|y, \boldsymbol{\gamma}^{(f)}) = \mathrm{T}_{n+\delta+k}\left(M_\gamma^{-1}X_\gamma' y \ , \ \left(\frac{\tau + q_\gamma}{n+\delta+k}\right) M_\gamma^{-1}\right),$$

a multivariate t distribution. Fitted values $\mu^{(f)} = X_{\gamma^{(f)}}\beta^{(f)}$ and associated 95% intervals are plotted in Figure 4.7. Shrinkage towards the empirical mean, 5.836, is somewhat evident, however there appears to some association between the gene expression data and log survival.

To examine the extent of this association, I constructed a "metagene" to serve as a survival index by taking the first principal component from a singular value decomposition of the three genes comprising the top model, SPARC, SEMA3B and DCX. Small values of the metagene should represent high-risk, or low survival, cases whereas large values should represent low-risk, or longer survival, cases. Figure 4.8 displays the a scatterplot of the survival index with both SPARC and

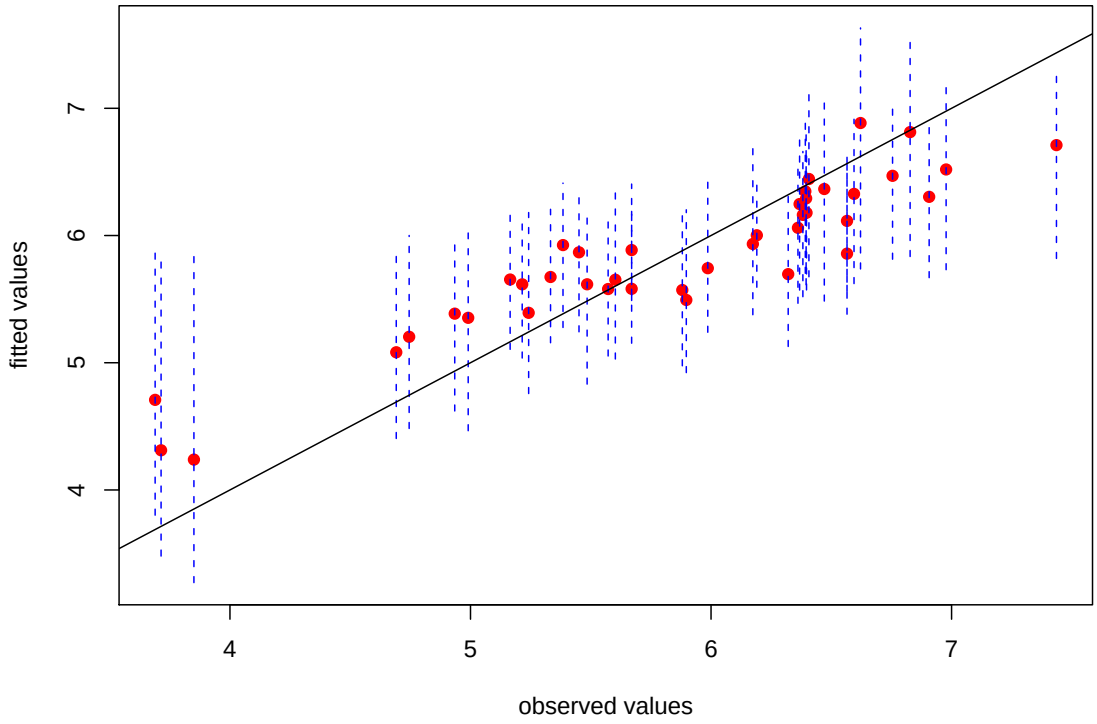

Figure 4.7: Model averaged fitted means, μ_i, for the Keck example with 95% intervals. "Observed values" refers to the observed y_i.

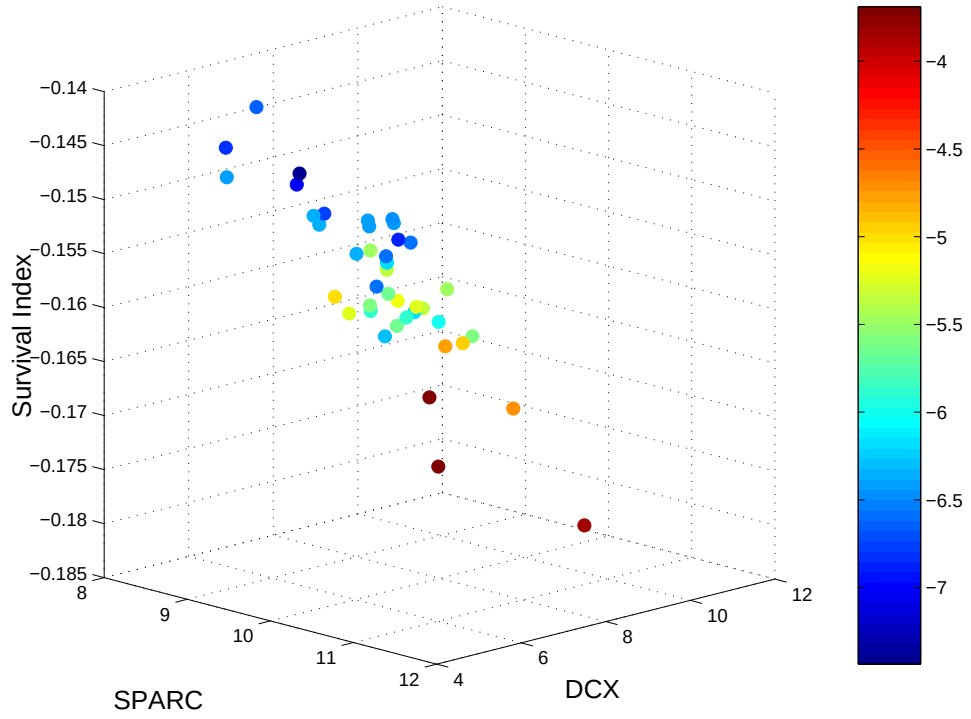

Figure 4.8: Scatterplot of the survival index metagene with SPARC and DCX. The cases are color coded by their observed survival times.

DCX. The cases are color coded by observed log survival time, with low risk cases having darker shades of blue high risk cases having darker shades of red. The metagene is able to separate the samples fairly well into high and low risk, indicating that there is indeed information relating to survival in the genetic data.

4.4.2 Assessing Aspects of Predictive Fit

To assess aspects of the predictive fit of the overall model, I performed a leave-one-out cross-validation as described in Section 3.4.2. After leaving out each observation one at a time and compiling the reordered lists of top models, Γ_i^*, I computed model averaged predicted means for each observation based on the

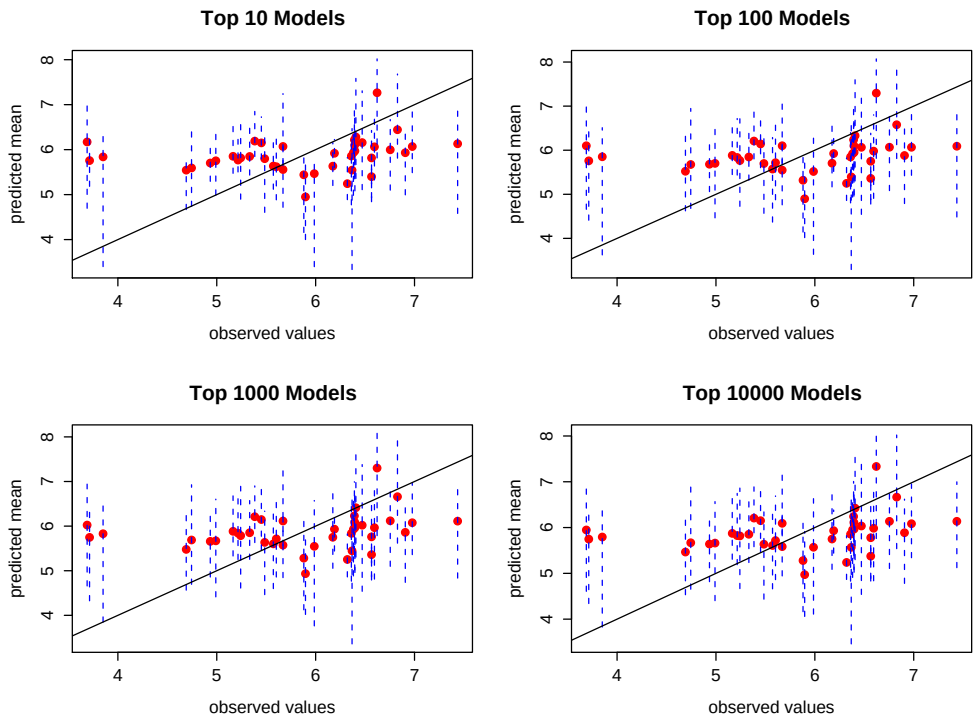

Figure 4.9: Leave one out cross validated predicted means using various numbers of top models for the analysis with $\pi = 10/p$. Point estimates and intervals are based on 10,000 draws for each observation from the posterior distributions $p(\boldsymbol{\beta}, \boldsymbol{\gamma}|y_{-i})$.

top models in each list. Figure 4.9 displays these predictions, along with 95% intervals, having used the lists of top 10, 100, 1000 and 10000 models separately for prediction. In the leave-one-out cross-validated predictive context, there is considerable shrinkage towards the mean for each observation. Either there is not much predictive power in the data or else dimension has been overpenalized, resulting in the highest posterior probability models not being rich enough from a predictive perspective. The latter possibility is investigated in the next section.

4.5 Alternate Analyses

The choice $\pi = 10/p$ is somewhat arbitrary. In general it reflects prior belief in sparsity, as the penalty for adding each extra variable, on the log scale, is $\log(\pi/(1-\pi)) = -6.733$. Alternatively, using $\pi = 2/p$ or $\pi = 100/p$ induces log-penalties of -8.344 and -4.432, respectively. Which of these truly reflects prior belief? There is perhaps no definitive answer to the question, although one approach for specifying π is described in Section 5.3. In this section I compare the results from the previous section with results from two additional runs of SSS, using the two alternate priors above, to demonstrate the sensitivity of the analysis to the choice of prior, $p(\boldsymbol{\gamma})$.

4.5.1 Effect of π on Posterior Model Size

Table 4.6 displays the values $\tilde{p}(|\boldsymbol{\gamma}| = k|y)$ for each of the three analyses. Clearly, the set Γ^* changes dramatically as π increases, favoring richer models and pushing down the probability on the null model. This will certainly affect predictive performance as there is a vast difference in the amount of shrinkage toward the null model for the three analyses. The sensitivity of $\tilde{p}(\boldsymbol{\gamma}|y)$ to the choice of π indicates that one should be careful about interpreting the amount of support the data give to the null model.

4.5.2 Effect of π on Variable Inclusion Probabilities

Table 4.7 displays the top fifteen genes for each of the three analyses, listed in decreasing order by $\tilde{p}(\gamma_j = 1|y)$. The most obvious difference is that values of $\tilde{p}(\gamma_j = 1|y)$ are much smaller for small values of π, due to the fact that the

Table 4.6: Comparison of posterior probability of model size for the three analyses, conditioned on the top 1,000,000 models for each run of SSS. The highest posterior probability dimension is highlighted for each analysis. The first line for each run is the number of models found, and the second line is $\tilde{p}(|\boldsymbol{\gamma}| = k|y)$. The character $*$ indicates a value < 0.001.

π, k:	0	1	2	3	4	5	6	7
$2/p$	1	8,408	798,128	121,998	61.441	10,024	0	0
	0.576	0.320	0.068	0.025	0.011	*	0	0
$10/p$	1	8,408	120,614	116,545	275,203	477,880	1,349	0
	0.030	0.084	0.062	0.165	0.421	0.237	*	0
$100/p$	1	2	40	841	40,774	794,870	159,849	3,623
	*	*	*	0.003	0.083	0.698	0.214	0.002

null model (and one variable models) are absorbing much more posterior mass in these cases. There is fairly high concordance, however, in the genes that make up the three lists. The top three genes, SPARC, SEMA3B and DCX, are the same across the three lists; only one gene in the top fifteen for the $10/p$ analysis was not included in either of the other two lists.

4.5.3 Effect of π on Predictive Fit

Figures 4.10 and 4.11 display leave one out cross validated predicted means for the analysis with $\pi = 2/p$ and $\pi = 100/p$, respectively. There is clear shrinkage for the former, while the latter plots seem to indicate slightly better prediction than for the $\pi = 10/p$ case, due to the fact that the high posterior models have more variables in the $100/p$ case. The intervals in both cases are 95% regions, and cover the observed value for most cases in the latter figure.

Table 4.7: Comparison of variable inclusion probabilities for the top fifteen variables from each of the three analyses, conditioned on the top 1,000,000 models for each run. Variables in blue are common to all three lists, while those in green and red are common to two.

$\pi = 2/p$		$\pi = 10/p$		$\pi = 100/p$	
SPARC	0.047	SPARC	0.797	SPARC	0.896
SEMA3B	0.023	SEMA3B	0.634	SEMA3B	0.763
DCX	0.013	DCX	0.348	DCX	0.433
ENSA	0.009	DCXa	0.133	NELL1	0.186
DCXa	0.006	KIAA0831	0.062	FABP5	0.172
CRABP5	0.005	HMP19	0.048	KIAA0831	0.150
TF	0.005	NELL1	0.062	DCXa	0.107
CNGA3	0.005	C13orf7	0.037	C13orf7	0.105
TFa	0.005	CNGA3	0.036	RAC1	0.102
L1CAM	0.004	FABP5	0.027	HMP19	0.086
MTMR2	0.004	ZNF217	0.027	SOX4	0.081
NKX2-2	0.004	L1CAM	0.024	ZNF217	0.070
NAP1L1	0.004	SOX4	0.021	MCAM	0.065
AK3	0.004	STMN2	0.020	HAN11	0.053
FUBP1	0.004	FUBP1	0.020	HSPA9B	0.038

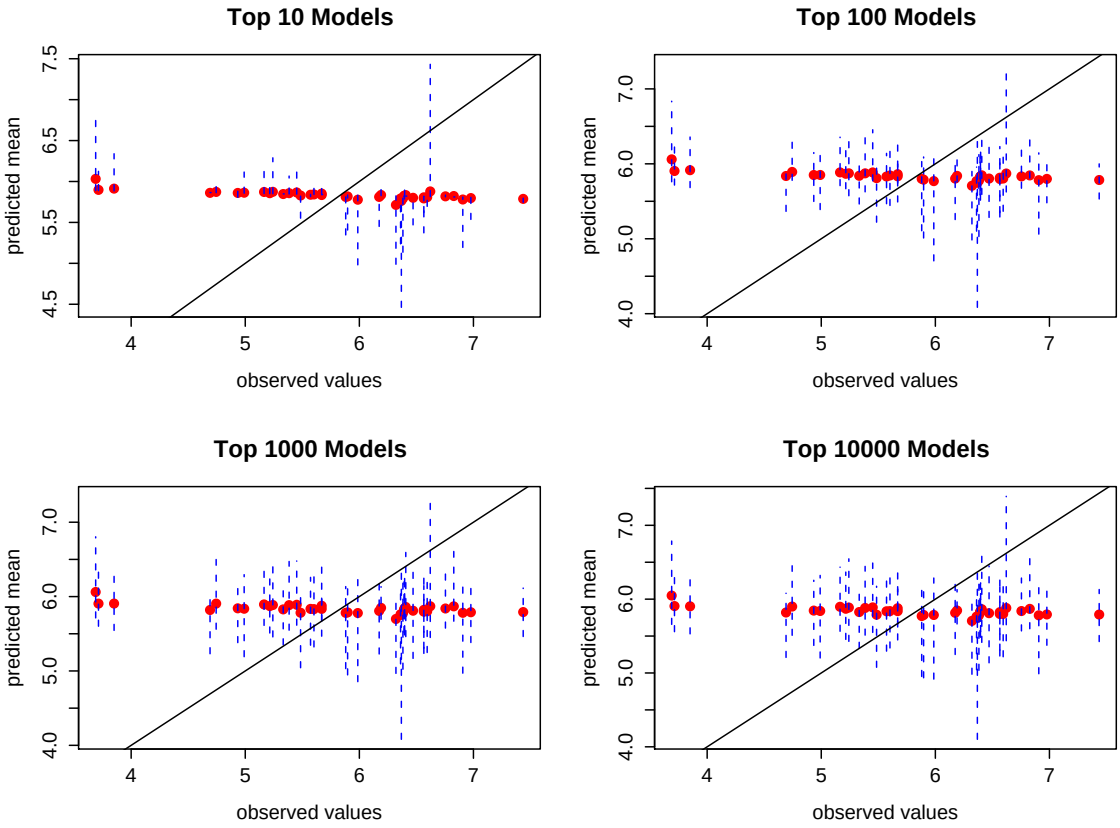

Figure 4.10: Leave one out cross validated predicted means using various numbers of top models for the analysis with $\pi = 2/p$. Point estimates and intervals are based on 10,000 draws for each observation from the posterior distributions $p(\boldsymbol{\beta}, \boldsymbol{\gamma}|y_{-i})$.

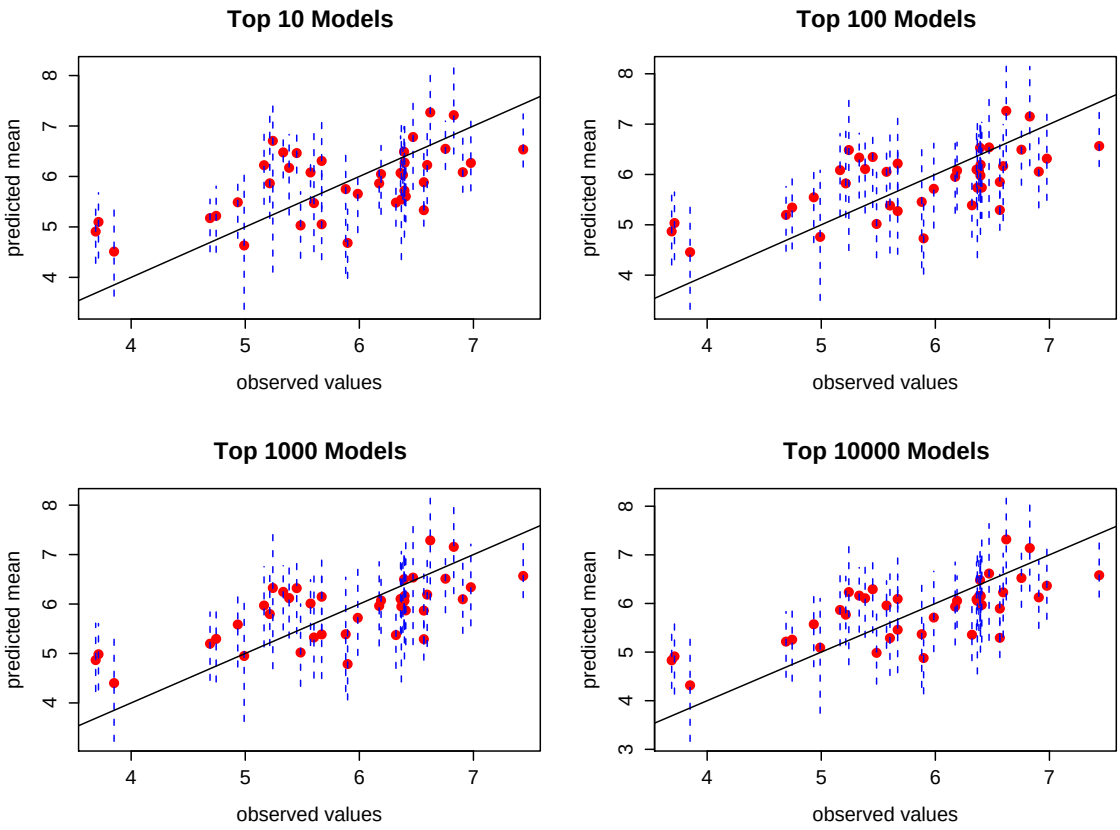

Figure 4.11: Leave one out cross validated predicted means using various numbers of top models for the analysis with $\pi = 100/p$. Point estimates and intervals are based on 10,000 draws for each observation from the posterior distributions $p(\boldsymbol{\beta}, \boldsymbol{\gamma}|y_{-i})$.

4.6 Effect of the Size of Γ^*

Figures 4.9, 4.10 and 4.11 give some insight into how predictive performance changes as the number of top models conditioned on increases: we typically expect slightly more shrinkage toward the overall mean, as the infusion of many more (relatively) "bad" models should just add random noise. We can also explore how the number of top models saved by SSS and subsequently used for averaging affects our posterior estimates of model size and variable importance.

Figures 4.12 through 4.14 show the estimated posterior probabilities of model size as a function of B, the number of top models stored in Γ^* for the three analyses. One million top models seems to be enough to have a stable estimate for this example, although one can see in all three figures that posterior mass is still slowly accumulating.

Figures 4.15 through 4.17 show cumulative posterior probability plots for variable importance measures $\tilde{p}(\gamma_j = 1|y)$ for SPARC, SEMA3B, DCX, NELL1 and CNGA3 for the three analyses. The values stabilize fairly quickly, except perhaps for NELL1 in the $100/p$ analysis, which is still decreasing after summing over the top million models.

4.7 Comparison with MCMC Methods

Important comparisons between SSS and MCMC methods include: (i) the amount of posterior mass found, (ii) cumulative posterior mass found as a function of both time and model evaluations, (iii) the best models found and (iv) the amount of time/model evaluations needed to find the best models. Here I compare SSS with Gibbs sampling as described in Section 2.1.1, in both cases using $\pi = 10/p$.

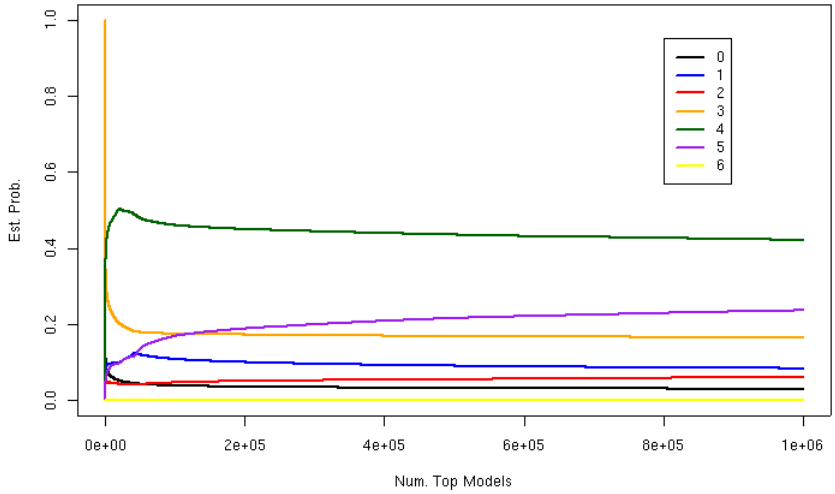

Figure 4.12: Values of $\tilde{p}(|\boldsymbol{\gamma}| = k|y)$ as a function of the size of Γ^* for the analysis with $\pi = 10/p$.

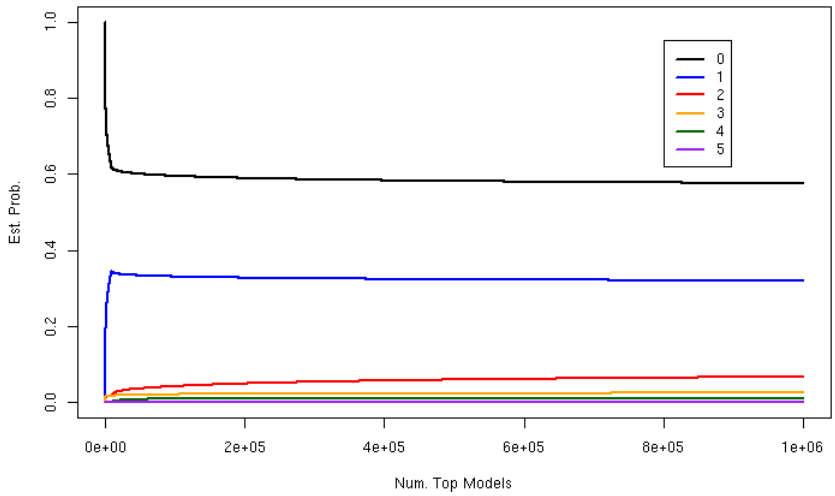

Figure 4.13: Values of $\tilde{p}(|\boldsymbol{\gamma}| = k|y)$ as a function of the size of Γ^* for the analysis with $\pi = 2/p$.

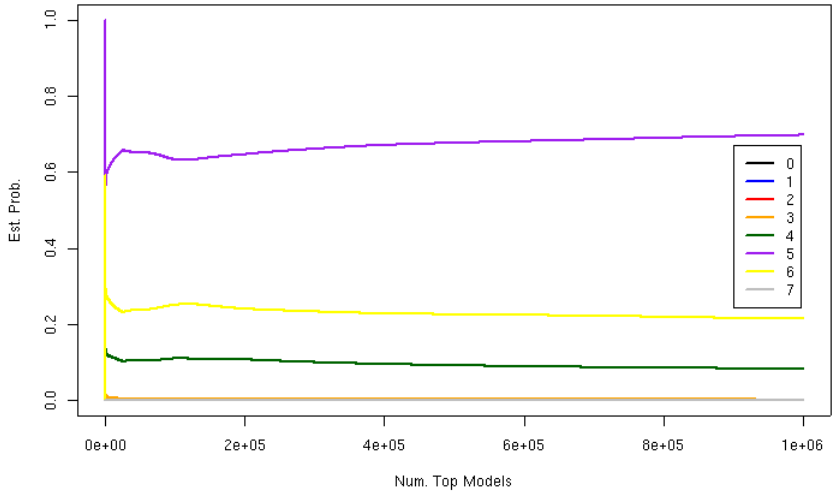

Figure 4.14: Values of $\tilde{p}(|\boldsymbol{\gamma}| = k|y)$ as a function of the size of Γ^* for the analysis with $\pi = 100/p$.

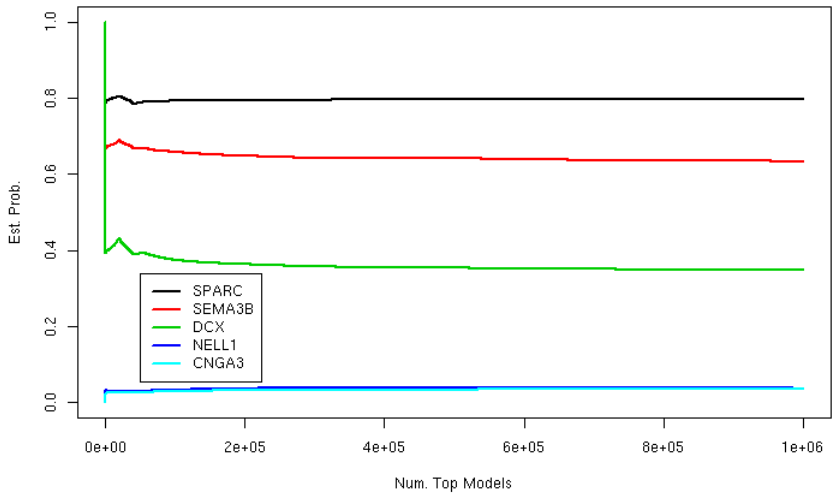

Figure 4.15: Values of $\tilde{p}(\gamma_j = 1|y)$ for several genes as a function of the size of Γ^* for the analysis with $\pi = 10/p$.

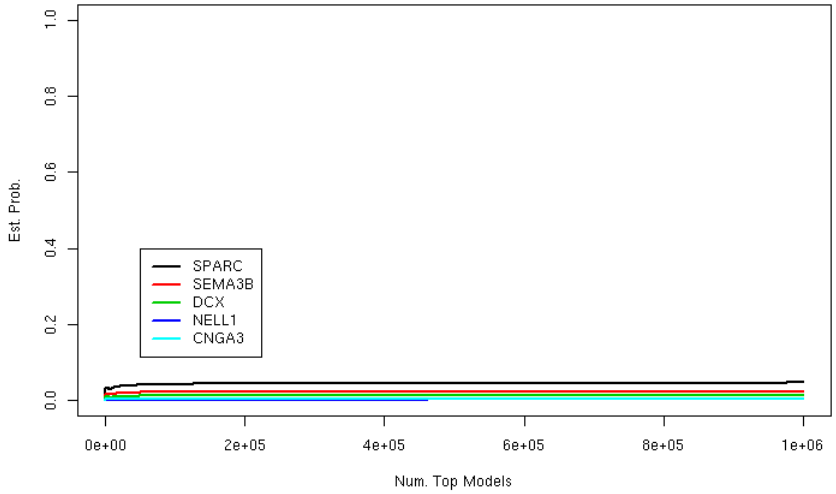

Figure 4.16: Values of $\tilde{p}(\gamma_j = 1|y)$ for several genes as a function of the size of Γ^* for the analysis with $\pi = 2/p$.

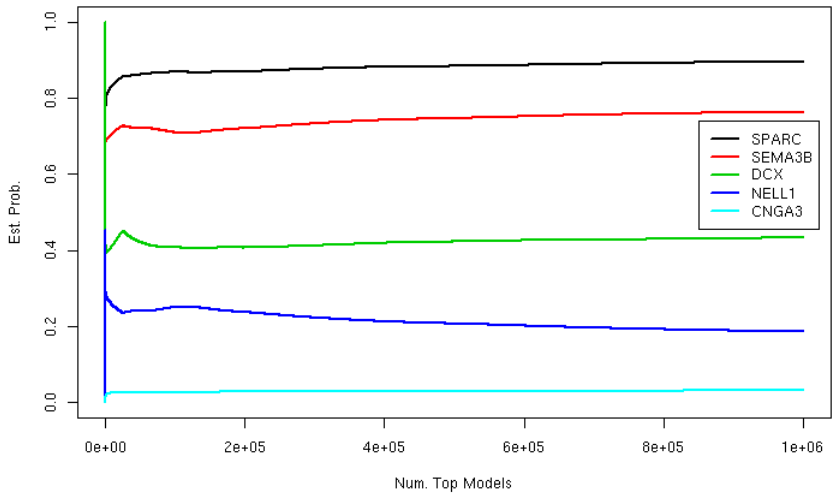

Figure 4.17: Values of $\tilde{p}(\gamma_j = 1|y)$ for several genes as a function of the size of Γ^* for the analysis with $\pi = 100/p$.

I used two baselines to compare the two methods: elapsed time and number of model evaluations. SSS ran for 40,000 iterations, taking 11 hours and 53 minutes. The number of model evaluations performed by SSS for a fixed number of iterations depends on the current model at each iteration: if the current model is of size k, then $p + k(p - k)$ model evaluations are performed. For this example, $1,137,195,208$ model evaluations were performed in 40,000 iterations. The Gibbs sampler makes p model evaluations at each iteration, and so Gibbs must be run for $135,252$ iterations in order to make comparisons based on the number of models evaluated.

To compare the two methods based on run-time, a separate Gibbs run was performed that was stopped after 11 hours and 53 minutes had elapsed. This resulted in 29,163 iterations (245,202,504 model evaluations). In the sections below, it is assumed that when comparisons are made based on run-time, the 11 hour and 53 minute Gibbs run is being used, and when comparisons are made based on model evaluations, the $135,252$ iteration Gibbs run is being used.

4.7.1 Run-time Comparisons

After running for 11 hours and 53 minutes, the posterior mass of the top million models found by Gibbs accounted for 75.41% of the total mass of the models found by SSS. Figure 4.18 shows the accumulated posterior mass by time for both methods, where the plot is normalized by the total mass found by SSS. SSS accumulates more mass for a fixed run time compared to Gibbs, and its rate of accumulation is much greater than Gibbs early on, indicating that SSS is finding the best models faster.

Looking at only the very best models found by both methods, for the most

part SSS and Gibbs find the same models, however SSS finds them much faster. Figure 4.19 displays the best 60 models found by both SSS and Gibbs as function of the time taken to find them. The character "X" indicates a model found by only one of the two methods. Of the top 60 models, SSS found five models not found by Gibbs, and Gibbs found three models not found by SSS. The three models found by Gibbs but not SSS were of size five (two models) and six. This is likely due to the fact that Gibbs generally wanders around low probability regions of the model space (i.e., regions with larger numbers of variables) and happened to stumble upon a few competitive models. SSS on the other hand spends more time in regions of generally higher posterior probability, and hence did not explore the spaces $\Gamma^{(5)}$ and $\Gamma^{(6)}$ as well.

Overall, SSS found models in the set of the top 60 much faster in terms of run-time than Gibbs. Of the top 60 models found by both SSS and Gibbs, all but two were found first by SSS (indicated by the black/red lines).

4.7.2 Model Evaluation Comparisons

After running for 135,252 iterations, the posterior mass of the top million models found by Gibbs accounted for 97.49% of the total mass of the models found by SSS. The search ran for 55 hours and 13 minutes compared to 11 hours and 53 minutes for SSS. Figure 4.20 shows the accumulated posterior mass by model evaluations for both methods, where the plot is normalized by the total mass found by SSS. SSS accumulated slightly more mass for a fixed number of model evaluations compared to Gibbs, and its rate of accumulation is much greater than Gibbs early on, indicating that SSS is finding the best models using fewer model evaluations.

Looking at only the very best models found by both methods, Gibbs found all of the top 54 models found by SSS plus six others not found by SSS. Figure 4.21 shows that in all but six cases, SSS found the top models using fewer model evaluations than Gibbs.

4.7.3 Nature of the Gibbs Sampler in Large p Problems

As discussed in Section 2.1.1, the Gibbs sampler has a tendency to wander around low posterior regions of the model space. Figure 4.22 displays the sizes of the model at the end of each cycle through the variables for the first 50,000 iterations of Gibbs. The distribution of model sizes is nearly $\text{Bin}(p, \pi)$, as seen in the Q-Q plot in Figure 4.23. There is of course variability of model size within each cycle through the variables. Figure 4.24 is a plot of the value of the log posterior, $\log p(y|\boldsymbol{\gamma}) + \log p(\boldsymbol{\gamma})$, for the first 50,000 iterations. The red line is the cut off point for the top million models found. It is clear that Gibbs spends a significant portion of its time in low probability regions.

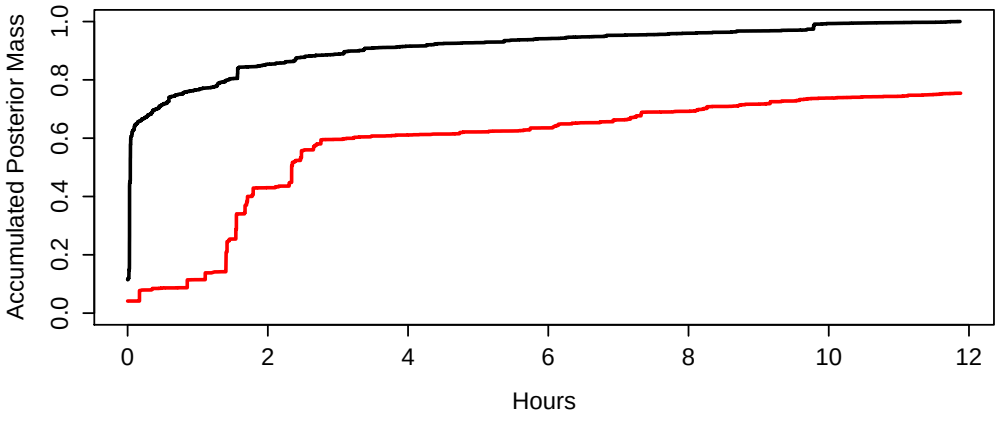

Figure 4.18: Comparison of accumulated posterior mass as a function of run-time for SSS (black) and Gibbs (red), based on the top million models for a run of 11 hours and 53 minutes.

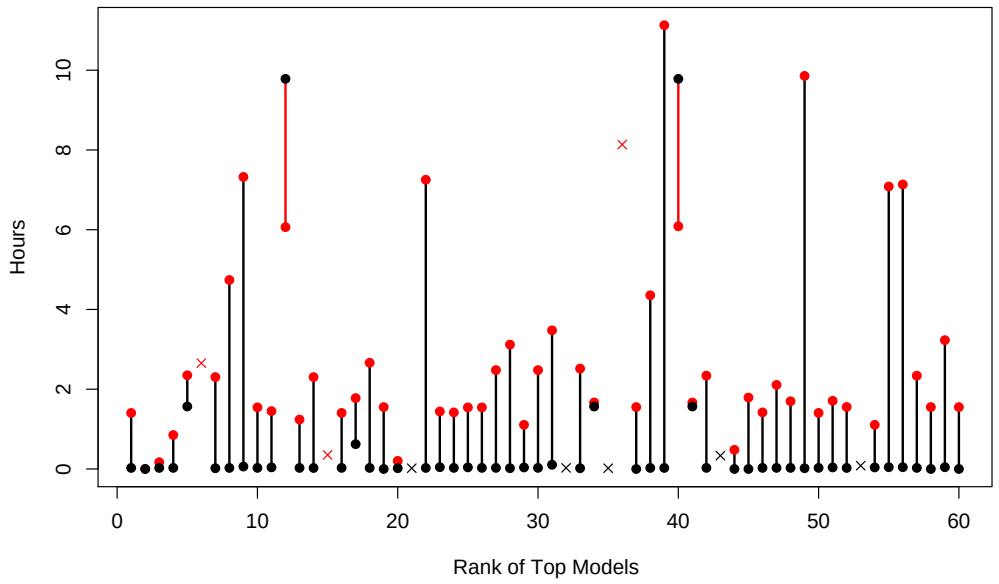

Figure 4.19: Comparison of time to find the top 60 models for SSS (black) and Gibbs (red). The character "X" means that this model was only found by the method indicated by color.

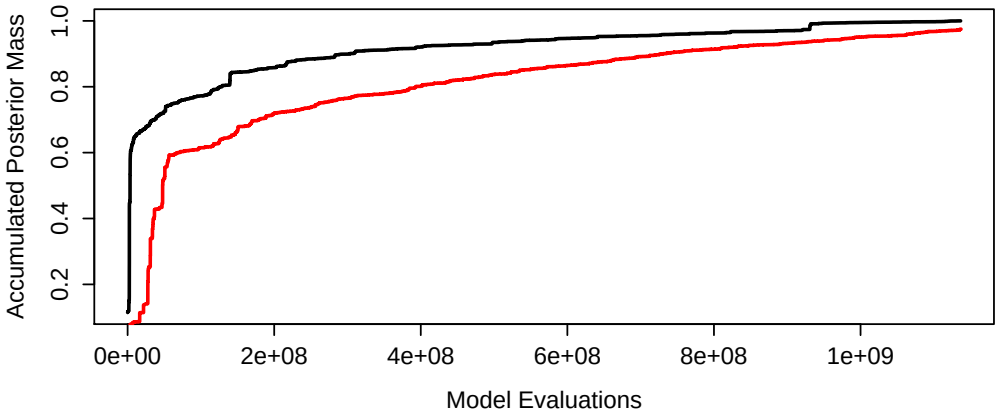

Figure 4.20: Comparison of accumulated posterior mass as a function of model evaluations for SSS (black) and Gibbs (red), based on the top million models for a run consisting of 1,137,195,208 model evaluations.

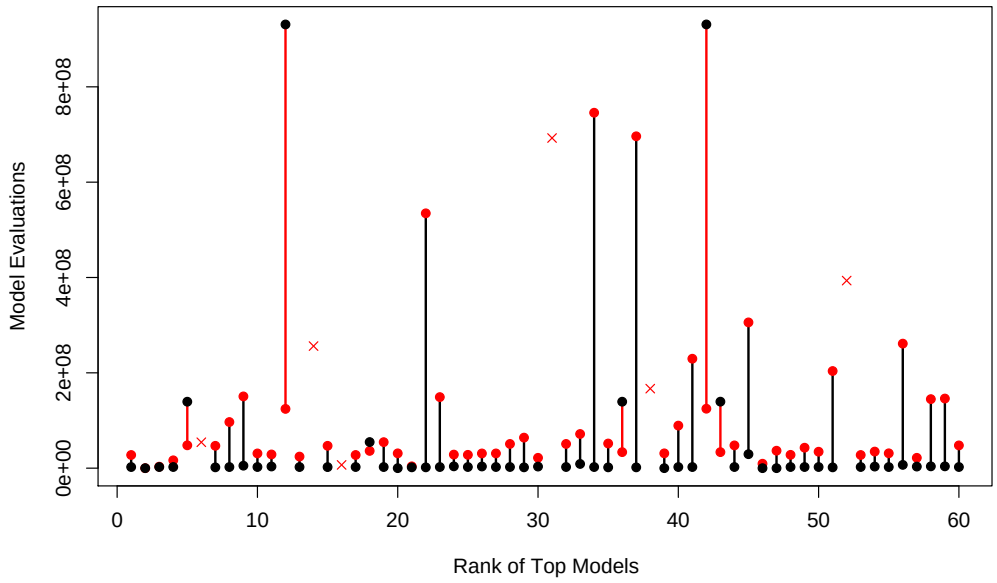

Figure 4.21: Comparison of model evaluations required to find the top 60 models for SSS (black) and Gibbs (red). The character "X" means that this model was only found by the method indicated by color.

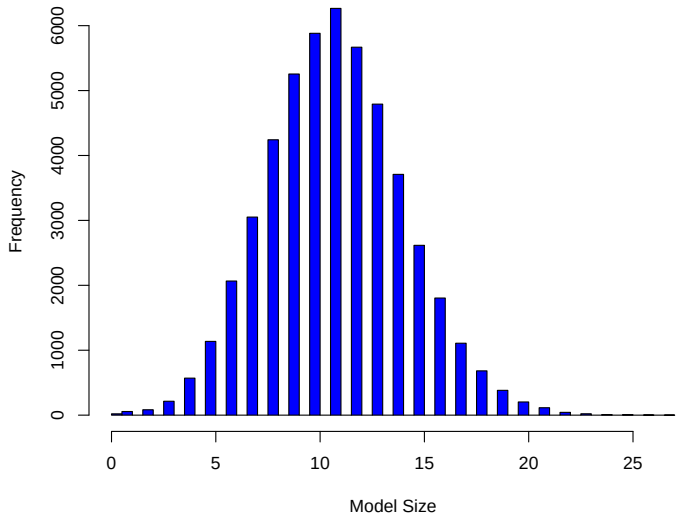

Figure 4.22: The size of the model at the end of each cycle through the variables for the first 50,000 iterations of Gibbs.

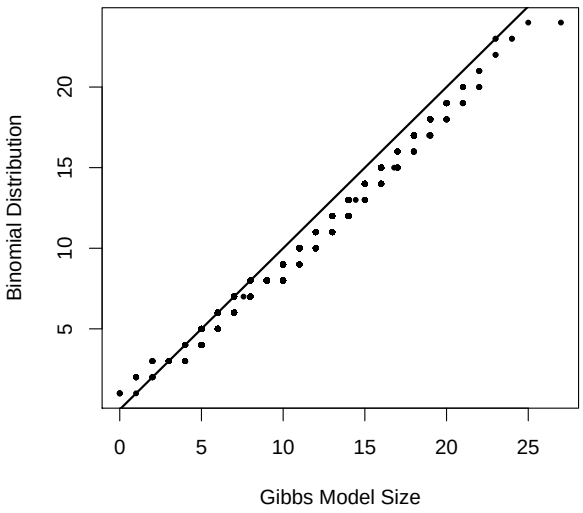

Figure 4.23: Q-Q plot of the model size at the end of each Gibbs iteration for the first 50,000 iterations and a Bin(p, π) distribution.

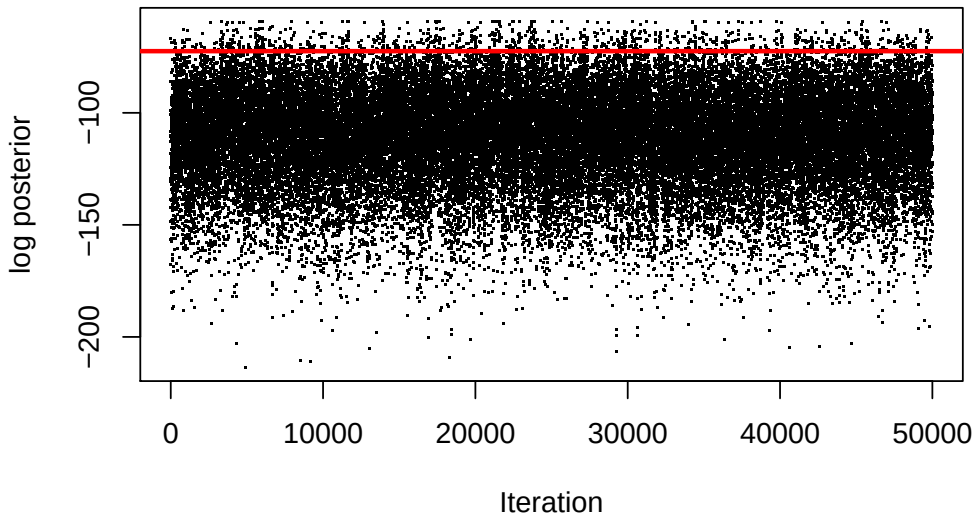

Figure 4.24: The log posterior of the model at the end of each cycle through the variables for the first 50,000 iterations of Gibbs. The red line is at -72.48, the cut off for the top million models.

Chapter 5

Sparsity in the Normal Linear Model

In this chapter I introduce methods for approximate inference on the parameter controlling the amount of sparsity in my specification of the Bayesian normal linear model. This requires investigation of the marginal likelihood $p(y|\boldsymbol{\gamma})$ and I begin with a necessary result.

5.1 A Lower Bound on the Marginal Likelihood

Recall that for each response variable y_i, $i = 1, \ldots, n$, we have covariate information x_{ij}, $j = 1, \ldots, p$. For the discussion below, as throughout, assume that the data have been centered and scaled to have mean zero and unit variance: $\mathbf{1}'x_j = \mathbf{1}'y = 0$ and $x_j'x_j = y'y = n - 1$.

Allowing for a slight abuse of notation, let X be the design matrix for a model consisting of some subset of the predictor variables, and let the number of variables in the model be k, as the notation X_γ becomes cumbersome. As

described in Section 3.1, the probability model is

$$\begin{aligned} p(y|\boldsymbol{\beta}, \psi, X) &= \mathrm{N}(X\boldsymbol{\beta}\ ,\ \sigma^2 I_n), \\ p(\boldsymbol{\beta}|\sigma^2, k) &= \mathrm{N}(0\ ,\ \tau^{-1}\sigma^2 I_k), \\ p(\sigma^2|k) &= \mathrm{IG}\left(\frac{\delta+k}{2}\ ,\ \frac{\tau}{2}\right). \end{aligned}$$

From this, the marginal likelihood (3.6) is

$$p(y|X) = \frac{\Gamma\left(\frac{n+\delta+k}{2}\right)/\Gamma\left(\frac{\delta+k}{2}\right)}{\pi^{n/2}\tau^{(n-k)/2}|M|^{1/2}\{1+q/\tau\}^{(n+\delta+k)/2}}, \tag{5.1}$$

where $M = \tau I_k + X'X$ and $q = y'y - y'XM^{-1}X'y$. The null model is defined to be the model with no predictor variables, formally $k = 0$, $q = y'y$ and $|M| = 1$.

Theorem 5.1. *Assume that y and the x_j have been centered and scaled to have zero mean and unit variance, and that for any $k < n$ dimensional model defined by X, $\mathrm{rank}(X) = k$. Then for fixed values of $\delta > 0$, $\tau > 0$ and y, the marginal likelihood (5.1) for any model X of dimension $k < n$ is lower bounded:*

$$p(y|X) \geq \frac{\Gamma\left(\frac{n+\delta+k}{2}\right)/\Gamma\left(\frac{\delta+k}{2}\right)}{\pi^{n/2}\ \tau^{-(\frac{\delta}{2}+k)}\ (\tau+n-1)^{\frac{n+\delta}{2}+k}} \equiv p^*(y|X), \tag{5.2}$$

with equality when $X'X = (n-1)I_k$, i.e. when X is an orthogonal matrix, and when $X'y = 0$, i.e. when y is orthogonal to the column space of X.

Proof. The components of (5.1) that involve X are q and M. To complete the proof, I first establish that $q \leq n-1$ and secondly that $|M| \leq (\tau+n-1)^k$. Substituting these upper bounds into the denominator of (5.1) yields the lower bound in (5.2).

Regarding q, from Appendix B.1, we have that $XM^{-1}X'$ is a positive semidefinite matrix. Hence $\forall y \in \mathbb{R}^n$, $y'XM^{-1}X'y \geq 0$, resulting in $q \leq y'y = n - 1$.

To establish that $|M| \leq (\tau + n - 1)^k$, I use the result proved in Appendix B.1 that $|M| \leq \prod_{i=1}^{k} m_{ii}$. As the data are standardized so that $x_j'x_j = n - 1$ for $j = 1, \ldots, p$, we have $m_{ii} = \tau + n - 1$ and thus have $|M| \leq (\tau + n - 1)^k$.

Substituting these two bounds, $q \leq n - 1$ and $|M| \leq (\tau + n - 1)^k$ into the denominator of (5.1) yields the lower bound in (5.2).

When $X'X = (n-1)I_k$ and $X'y = 0$, we have $q = n-1$ and $M = (\tau+n-1)I_k$, and direct substitution into (5.1) gives equality in (5.2). □

The key to the existence of the lower bound is the constraint that the data have been standardized; otherwise, probability statements would have to be made about likely values of $|M|$ and q in order to gauge likely values of the marginal likelihood, which is a more complicated task.

The correspondence of equality in (5.2) with the response vector being an element of the null space of X can be interpreted as the lower bound being the "worst case scenario" – when X has no predictive power with respect to the outcome. All other models can then be interpreted with this baseline value in mind. Of course, if y is random noise that is uncorrelated with X, the marginal likelihood will be greater than the lower bound with probability one. Thus the lower bound provides calibration in a sense, and the task is to determine what models fall outside of the "noise" region near the lower bound.

5.1.1 Sparsity and Bayesian Shrinkage

In traditional model selection, the tendency to overfit is usually balanced by the inclusion of a penalty term in the selection criterion. There are many off-the-shelf criteria for comparing models, including AIC (Akaike, 1970), BIC (Schwartz, 1978), RIC (Foster and George, 1994) and C_p (Mallows, 1973), along with variants on them, several of which are considered in Shao (1997). The idea is that more complex models will incur larger penalties, allowing the addition of model components only if these components dramatically improve the model fit. For example, for normal linear models, $\text{BIC} = n\log(1-R^2) + k\log n$, and $\text{AIC} = n\log(1-R^2) + 2k$, where R^2 is the coefficient of determintation for a particular model of size k. Smaller values of BIC and AIC indicate "better" models.

Alternative ways to avoid overfitting not involving selection criteria fall into the category of shrinkage methods (see, e.g. Hastie *et al.*, 2001, Chapters 3 and 5). Full models are typically fit with constraints on the regression coefficients that shrink their values toward zero, effectively lowering the degrees of freedom. For a fixed value of σ^2, the prior distribution (3.2) corresponds to a shrinkage procedure known as ridge regression, where the regression coefficients are estimated under the constraint that $\boldsymbol{\beta}'\boldsymbol{\beta}$ is less than a specified constant related to $\tau^{-1}\sigma^2$.

It is well known that Bayesian model selection methods act as automatic "Occam's razors" without the need for inclusion of a separate penalty term (see, e.g. Smith and Spiegelhalter, 1980; Jefferys and Berger, 1992; Berger and Pericchi, 2001). In general this is accomplished through the specification of the probability model, i.e. through the likelihood and prior components. In the variable selection framework, shrinkage priors weight values for the regression coefficients that are nearer to zero more highly. So, when choosing between models of differing

dimension with similar fits to the data, the more parsimonious model is favored.

This is encoded mathematically in the marginal likelihood, $p(y|\boldsymbol{\gamma})$. After integrating the likelihood over the prior distribution, the marginal likelihood represents the balance between the fit of the data to the model and the penalty on model size imposed through the shrinkage prior. Hence we can think about breaking down the marginal likelihood into components

$$\text{marginal likelihood } = \text{constant } \times (\text{model fit } \circ \text{ dimension penalty}).$$

The open circle indicates that the marginal likelihood cannot simply be factored into a product of three separate components, but that the model fit and dimension penalty parts are often interrelated. Indeed, looking at (5.1), the components can not be separated out easily: the portion related to model fit, $(1 + q/\tau)^{(n+\delta+k)/2}$, is also explicitly related to the dimension of the model.

Consequently, to make comparisons between the amount of penalty on model dimension provided by different classes of model selection criteria and/or shrinkage priors, one must attempt to make comparisons between two models with "similar" model fit. This is not difficult for a criterion such as BIC, which is conveniently partitioned into two components: the model fit, $n \log(1 - R^2)$, and the dimension penalty, $k \log n$. Hence one can simply specify a value of R^2 (or equivalenty, for standardized data, a value of the residual sum of squares, RSS) to represent the model fit.

Comparing this directly to (5.1) is difficult, as (5.1) is not a nice function of

R^2. We could perhaps define a "ridge regression R^2" as

$$\begin{aligned} 1 - R^2_\tau &= (n-1)^{-1} y'(I_n - XM^{-1}X')y \\ &= (n-1)^{-1} RSS_\tau \\ &= (n-1)^{-1} q, \end{aligned}$$

rather than the traditional $1 - R^2 = (n-1)^{-1} y'(I_n - X(X'X)^{-1}X')y$ for standardized data. Even so, in order to make a direct comparison, assumptions would still need to be made about X, as $|M|$ is implicitly a function of dimension.

The lower bound formulation provides a natural way to assess the dimension penalty component, as the model fit has effectively been removed due to the orthogonality between X and y. We can write the lower bound as a function of the remaining two components (constant and dimension penalty)

$$p^*(y|\boldsymbol{\gamma}) = \left[(\tau\pi)^{-n/2} \left(\frac{\tau + n - 1}{\tau} \right)^{-\frac{n+\delta}{2}} \right] \cdot \left[\left(\frac{\tau + n - 1}{\tau} \right)^{-k} \frac{\Gamma\left(\frac{n+\delta+k}{2}\right)}{\Gamma\left(\frac{\delta+k}{2}\right)} \right].$$

Focusing on the case $\tau = 1$ and setting $g(n, \delta, k) \equiv \log\Gamma((n+\delta+k)/2) - \log\Gamma((\delta + k)/2)$,

$$\log p^*(y|\boldsymbol{\gamma}) = -\left(\frac{n}{2} \log \pi + \frac{n+\delta}{2} \log n \right) - k \log n + g(n, \delta, k). \tag{5.3}$$

As shown in Appendix C.2, for even values of n the final term can be written as

$$g(n, \delta, k) = -n \log 2 + \sum_{i=1}^{n/2} \log(n + \delta + k - 2i).$$

As seen in Figure 5.1, the function $g(n, \delta, k)$ is nearly linear for values of k, n and δ that we are concerned with. As can be seen from values of the log marginal

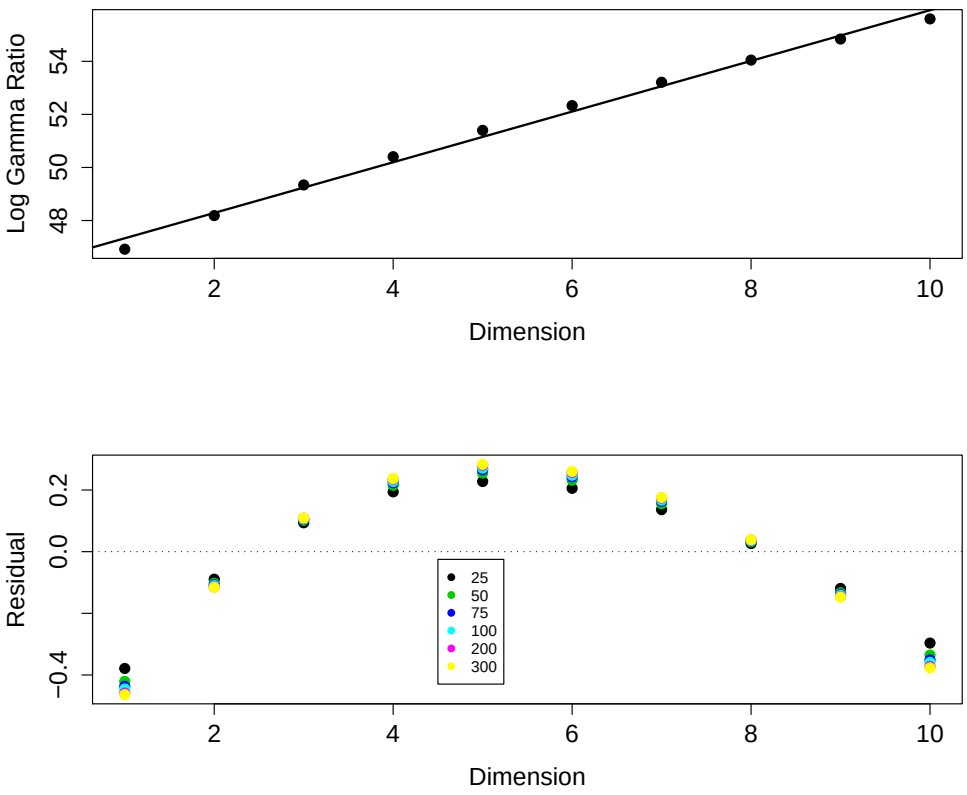

Figure 5.1: The top panel shows the function $g(n, \delta, k)$ for fixed $\delta = 3$ and $n = 41$. The line is a least squares fit. The bottom panel plots the residuals of least squares fits for $g(n, \delta, k)$ vs. k for fixed values of $\delta = 3$ and values of n as indicated in the legend.

likelihood in Figure 5.2, the departures from linearity of $g(n, \delta, k)$ as a function of k (as shown in the residual plot in Figure 5.1) are small enough relative to changes in $\log p^*(y|\boldsymbol{\gamma})$ as a function of k to make the term "nearly linear" meaningful.

The orthogonality between y and X is equivalent to $R^2 = 0$, and hence the corresponding BIC penalty is $k \log n$ and the corresponding AIC penalty is $2k$, which are both exactly linear in k. Figure 5.2 shows values of these penalties, along with $\log p(y|\boldsymbol{\gamma})$ as a function of k. The red line is $-\text{BIC} + \log p(y|\boldsymbol{\gamma} = \mathbf{0})$;

the negative multiplier means that larger (more positive) values of BIC represent more favored models, and the shift by the log marginal likelihood for the null model is intended to facilitate comparison. The same is true for the green line, which represents AIC. The figure was made with the Keck example in mind (see Chapter 4), with $n = 41$, $p = 8,408$, $\delta = 3$ and $\tau = 1$.

Care should be taken in interpreting the figure. The two relevant comparisons are the red and green lines, and the solid black and blue lines. The red and green lines, BIC and AIC, are used similarly as model selection criteria; we see that having removed model fit, the penalty incurred from adding an extra irrelevant variable is greater in BIC than in AIC (which will always be true for $n > 7$). Raftery (1995) shows that BIC is an approximation to $-2\log p(y|\boldsymbol{\gamma})$, and hence the solid black line can be compared to the blue line, $-\frac{1}{2}\text{BIC} + \log p(y|\boldsymbol{\gamma} = \mathbf{0})$. If this approximation were used in lieu of the marginal likelihood, then the BIC based criteria would penalize the addition of irrelevant variables less harshly than the proper Bayesian analysis.

Directly comparing the dimension penalty from $p(y|\boldsymbol{\gamma})$ to BIC (the red line) or AIC is potentially problematic, as the black line implicitly assumes a uniform prior distribution over the model space, $p(\boldsymbol{\gamma}) = |\Gamma|^{-1}$. In many model/variable selection type problems, such a prior is viewed as unrealistic, especially in cases where p is large, as the prior mean on model dimension is $p/2$. Using a sparsity inducing prior as discussed in Section 3.2 will penalize dimension *a priori* more reasonably: under the Bernoulli prior (3.7) the prior probability of a model with k variables is $(1-\pi)/\pi$ times greater than a model with $k+1$ variables, which, for small values of π, greatly favors parsimonious models. Comparing BIC or AIC to the log unnormalized posterior probability $\log p(y|\boldsymbol{\gamma}) + \log p(\boldsymbol{\gamma})$ (the dashed black

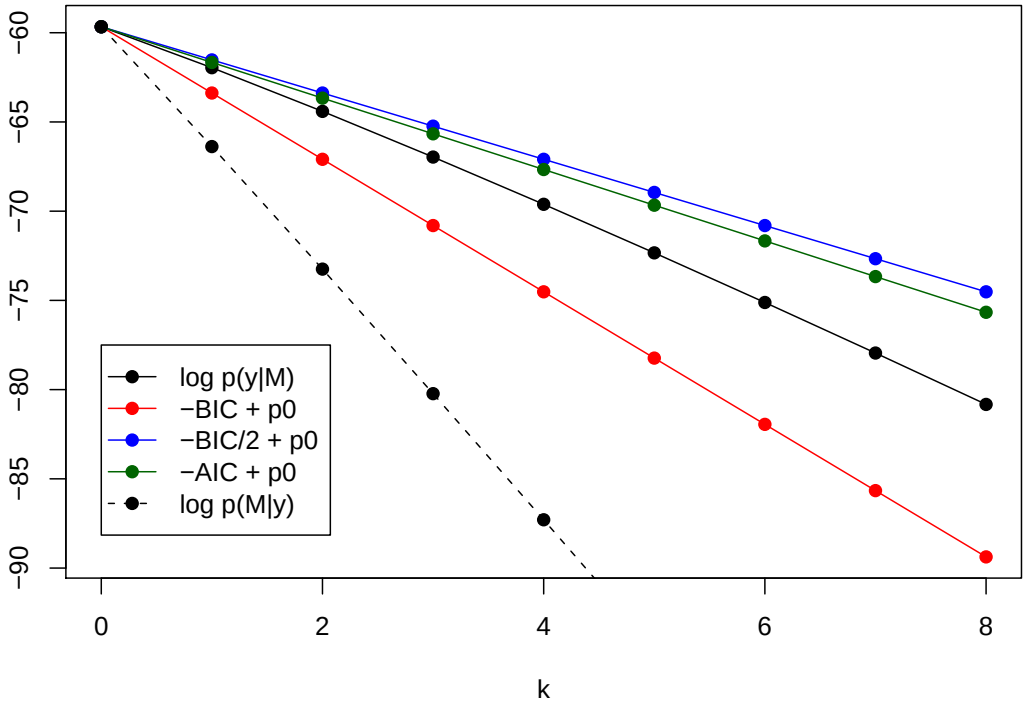

Figure 5.2: Comparison of dimension penalties for several selection criteria for models with no fit to the data.

line in Figure 5.2), we see that the proper Bayesian model selection criterion penalizes dimension much more heavily. The example here uses $\pi = 100/p$. It is not until $\pi > 2150/p$ that BIC penalizes uniformly more harshly on the range $k \in \{1, \ldots, 8\}$

Of course, these comparisons are conditional on there being no fit of the model to the data ($R^2 = 0$) and the predictor variables being orthogonal to each other. However they do serve as a baseline for understanding the way in which different selection criteria penalize the addition of irrelevant variables to a model, which gives us an idea about how much extra model "fit" is needed in order to favor the

addition of a variable that does indeed have some predictive power.

5.2 Characterizing the Marginal Likelihood

Section 5.1 introduced a lower bound on the marginal likelihood for models of a given dimension that corresponds to the case where y and the variables X in a given model are mutually orthogonal. Here I present a stochastic version of the lower bound result, where y and the x_j are no longer assumed to be orthogonal, but rather distributionally independent. In essence, the result can be viewed as providing a "distribution" of the marginal likelihood in the case where we are presented with random models, i.e., where the variables comprising the model have been drawn from a distribution that is independent of the outcome. Such a result is particularly relevant in settings where p is large and it is thought that many of the variables will be unrelated to the outcome, as will be underscored in Section 5.3.

5.2.1 Stochastic Version of the Lower Bound

For a fixed, observed value y and a given model comprised of some set of k predictor variables, X, we are interested in the marginal likelihood as a function of a random X. I make the assumption that y and X are independent as stated above, but here do not presume them to be orthogonal.

Of interest is the quantity

$$f(X; y) = \frac{p(y|X)}{p^*(y|X)},$$

where I have scaled by the lower bound so that the function is defined on the

interval $[1, \infty)$. Using (5.1) and (5.2), this can be rewritten as

$$f(X; y) = \frac{\tau^{-(n+\delta+k)}(\tau + n - 1)^{\frac{n+\delta}{2}+k}}{|M|^{1/2}(1 + q/\tau)^{(n+\delta+k)/2}}.$$

In the discussion that follows, I assume that $(y_i, x_i')' \sim \mathrm{N}(0, I)$ and constrain the observed values so that $X'X = \mathrm{diag}_k(n-1)$ and $y'y = n - 1$. Accordingly,

$$|M|^{1/2} = |\tau + X'X|^{1/2} = (\tau + n - 1)^{k/2},$$

and

$$\begin{aligned} f(X; y) &= \frac{\tau^{-(n+\delta+k)/2}(\tau + n - 1)^{(n+\delta)/2+k}}{(\tau + n - 1)^{k/2}(1 + q/\tau)^{(n+\delta+k)/2}} \\ &= \frac{\left(\frac{\tau+n-1}{\tau}\right)^{(n+\delta+k)/2}}{(1 + q/\tau)^{(n+\delta+k)/2}}. \end{aligned}$$

For ease of notation, I let $\mu = (n + \delta + k)/2$, leaving us with

$$f(X; y) = \left(\frac{\tau + n - 1}{\tau + q}\right)^{\mu}.$$

Focusing attention on q,

$$\begin{aligned} q &= y'y - y'XM^{-1}X'y \\ &= n - 1 - (\tau + n - 1)^{-1}y'XX'y. \end{aligned}$$

Letting $z = X'y$ we have

$$z \mid y \sim \mathrm{N}(0, (n-1)I),$$

and hence

$$(n-1)^{-1/2}z \mid y \sim \mathrm{N}(0, I).$$

Letting $u = (n-1)^{-1/2} z$,

$$q = n - 1 - \frac{n-1}{\tau + n - 1} u'u.$$

Hence

$$\tau + q = (\tau + n - 1)\left(1 - \frac{n-1}{(\tau+n-1)^2} u'u\right),$$

and

$$f(X; y) = (1 - w)^{-\mu}, \tag{5.4}$$

where $w = \nu u'u$ and $\nu = \frac{n-1}{(\tau+n-1)^2}$. Noting that $u'u \mid y \sim \chi^2_k$, we have

$$w \mid y \sim \text{Gamma}\left(\frac{k}{2}, \frac{1}{2\nu}\right). \tag{5.5}$$

As in the previous section, focus is on the log marginal likelihood, and so I consider the function

$$\log f(X; y) = -\mu \log(1 - w). \tag{5.6}$$

Recall that I originally constrained the observed values of X to lie on the cone of matrices such that $X'X$ is diagonal, and hence the distribution of $w|y$ stated above is only an approximation to the actual case where $X'X$ is only "close" to diagonal. It is only under the orthogonality constraint that (5.6) is guaranteed to exist (i.e., that $1 - w > 0$), as $|M|$ was calculated under this constraint. Under this condition, the approximate distribution of $\log f(X; y)$ can be derived. Let $v = -\mu \log(1-w)$, with corresponding inverse transformation $w = 1 - \exp\{-v/\mu\}$. Using $\alpha = k/2$ and $\beta = 1/(2\nu)$,

$$p(v) = \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{1}{\mu} (1 - e^{-v/\mu})^{\alpha - 1} \exp\{-v/\mu - \beta(1 - e^{-v/\mu})\}.$$

Rewriting $\mu^{-1} = \mu^{-\alpha}\mu^{\alpha-1}$, we have

$$\begin{aligned} p(v) &= \frac{(\beta/\mu)^\alpha}{\Gamma(\alpha)}\Big(\mu(1-e^{-v/\mu})\Big)^{\alpha-1}\exp\{-v/\mu - \beta(1-e^{-v/\mu})\} \\ &= \frac{(\beta/\mu)^\alpha}{\Gamma(\alpha)}\Big(\mu(1-e^{-v/\mu})\Big)^{\alpha-1}\exp\{-(\beta/\mu)(v/\beta + \mu(1-e^{-v/\mu}))\}. \quad (5.7) \end{aligned}$$

If the quantities $\mu(1-e^{-v/\mu})$ and $v/\beta + \mu(1-e^{-v/\mu})$ are both good approximations of v, then (5.7) is approximately a gamma density. The first order Taylor polynomial around zero for these two quantities are v and $v(1-2/n)$, respectively, which are both good approximations even for modest n (see Figure 5.3). Hence we have the approximation

$$\log f(X;y) \overset{d}{\approx} \text{Gamma}\left(\frac{k}{2}\ ,\ \frac{(\tau+n-1)^2}{(n-1)(n+\delta+k)}\right), \quad (5.8)$$

which in the limit as $n \to \infty$ for fixed k is

$$\log f(X;y) \overset{d}{\approx} \text{Gamma}\left(\frac{k}{2}\ ,\ 1\right).$$

Rather than computing the transformation explicitly as above, the approximate result (5.8) can also be arrived at using a Taylor expansion of $g(w) = -\mu\log(1-w)$ around zero. The linear approximation is

$$\begin{aligned} g(w) &\approx g(0) + wg'(0) \\ &= w\mu, \end{aligned}$$

and thus if w has a Gamma $\left(\frac{k}{2}, \frac{1}{2}\right)$ distribution, then $w \cdot \mu$ has a Gamma $\left(\frac{k}{2}, \frac{1}{2\nu\mu}\right)$ distribution, which is exactly the distribution in (5.8).

Table 5.1 gives the shape and scale parameters in (5.8) as a function of k for $\delta = 3$ and $\tau = 1$ under the column "Approx." The values under "Keck" and

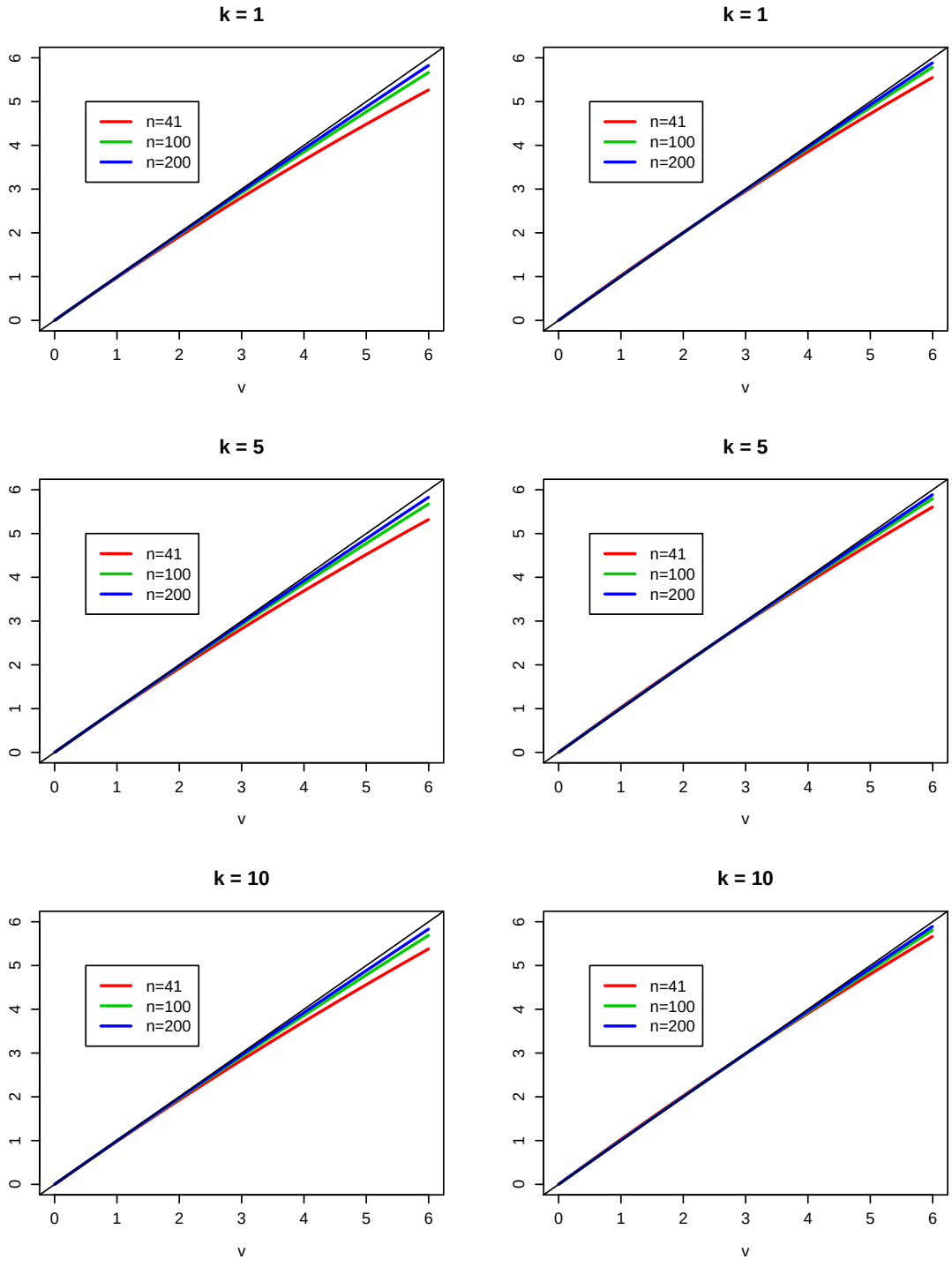

Figure 5.3: The left panel corresponds to $\mu(1 - e^{-v/\mu})$ and the right panel corresponds to $v/\beta + \mu(1 - e^{-v/\mu})$ for $\tau = 1$ and $\delta = 3$.

Table 5.1: Values of the shape and scale parameters for the gamma distribution discussed in Section 5.2.1. The columns under "Approx." are from (5.8) for $n = 41$, $\tau = 1$ and $\delta = 3$. The other two columns are described in Section 5.2.1.

	Approx.		Keck		Random	
k	Shape	Scale	Shape	Scale	Shape	Scale
1	0.5	0.934	0.500	1.024	0.497	0.884
2	1.0	0.914	1.106	1.028	1.016	0.870
3	1.5	0.894	1.788	1.109	1.547	0.841
4	2.0	0.876	2.552	1.014	2.134	0.831
5	2.5	0.858	3.408	1.010	2.718	0.813
6	3.0	0.841	4.333	1.005	3.330	0.796
7	3.5	0.824	5.281	0.994	3.906	0.770

"Random" are method of moments estimates[1] of the shape and scale parameters based on 100,000 samples of models (for each dimension) drawn independently from (i) the Keck dataset and (ii) standard normal distributions, respectively. The data in both cases have been scaled to have zero mean and unit variance, and both have a total of 8,408 possible predictors and 41 observations. The response variable y for the "Keck" example is the actual (normalized) log-survival times; for the "Random" example it is a vector whose elements were drawn from n independent standard normal distributions (independently from the sampled data matrix X), and then normalized. These values are also plotted in Figure 5.4.

The method of moments estimates for the models drawn randomly from independent normal distributions appear to match the moments of the approximation (5.8) better than the randomly drawn Keck models; this is likely due to collinearity present in the Keck data beyond the amount that appears randomly in the normal draws. It is clear in Figure 5.4 that the method of moments estimates

[1]The method of moments estimates are calculated as $\hat{\alpha} = \bar{X}^2/\hat{\sigma}^2$ and $\hat{\beta} = \bar{X}/\hat{\sigma}^2$, where $\bar{X}$ is the sample mean and $\hat{\sigma}^2$ is the sample variance.

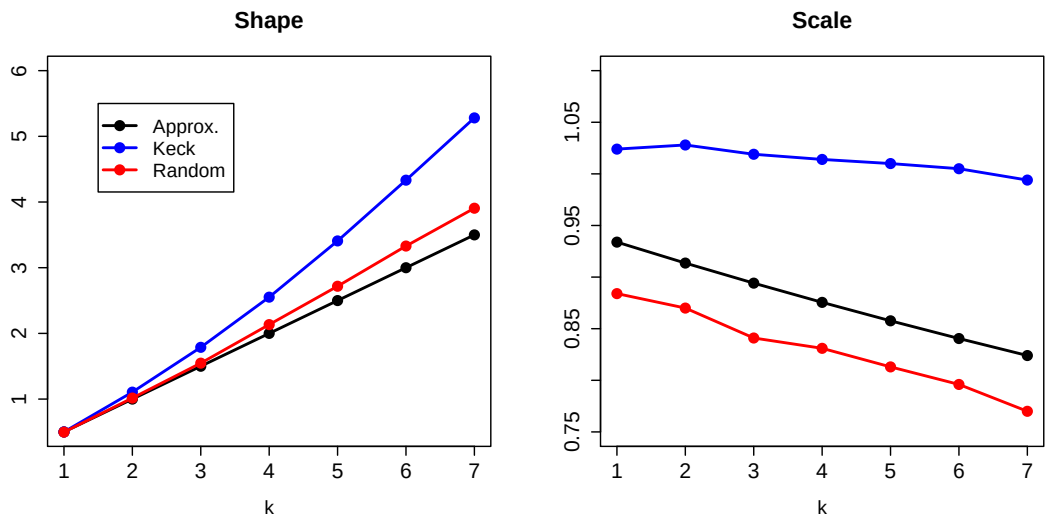

Figure 5.4: Values of the shape and scale parameters for the gamma distribution (5.8). The colors correspond to the columns in Table 5.1.

for the shape parameter for both the Keck and Random models becomes less like the approximation (5.8) as k increases, which we would expect, as it becomes less likely that $X'X$ is "close" to orthogonal as k increases.

5.2.2 Assessing the Approximation

The previous section established that under certain conditions, the shifted log marginal likelihood for models of a given dimension has approximately a gamma distribution (as a function of models). The approximation relies on the design matrix being orthogonal, which of course is uncommon in practice, and so we should check how good the approximate gamma distribution given by (5.8) is for both simulated and real data of interest.

Figures 5.5 and 5.6 give empirical evidence that the approximation is close to reality. Figure 5.5 contains Q-Q plots, where the vertical axes represent ran-

dom draws from gamma distributions with shape and scale parameters defined by (5.8) for model sizes one through six. The horizontal axes represent the 100,000 randomly sampled models used to construct the column "Random" in Table 5.1. Figure 5.6 was similarly constructed, with the horizontal axes representing the 100,000 randomly sampled models used to construct the column "Keck" in Table 5.1. The approximate gamma distribution seems to be a close match to the empirical distribution in both cases, however there is a general tendency for the empirical distributions to have mass farther to the right than the approximate distributions ($k = 2, \ldots, 6$ in Figure 5.5 and $k = 5, 6$ in Figure 5.6).

This suggests that perhaps a gamma distribution is a good approximation to values of $\log f(X; y)$ observed in practice, however that the shape and scale parameters given by (5.8) are not quite right due to the orthogonality restriction. This is displayed in Figures 5.7 and 5.8. Here, the vertical axes represent draws from gamma distributions where the shape and scale parameters are given not by (5.8), but rather by the method of moments estimates from Table 5.1. Figure 5.7 corresponds to the models drawn from random normal distributions, and Figure 5.8 corresponds to the models drawn randomly from the Keck dataset. In both cases, the fitted gamma distributions are a very good match to the observed data, indicating that slight adjustments to the shape and scale parameters given in (5.8) to account for observed collinearity in the data allow for a better approximating gamma distribution.

5.2.3 SVD Representation

To see precisely why the result from the previous section is only an approximation, it is convenient to first transform the representation of the linear model. Notation-

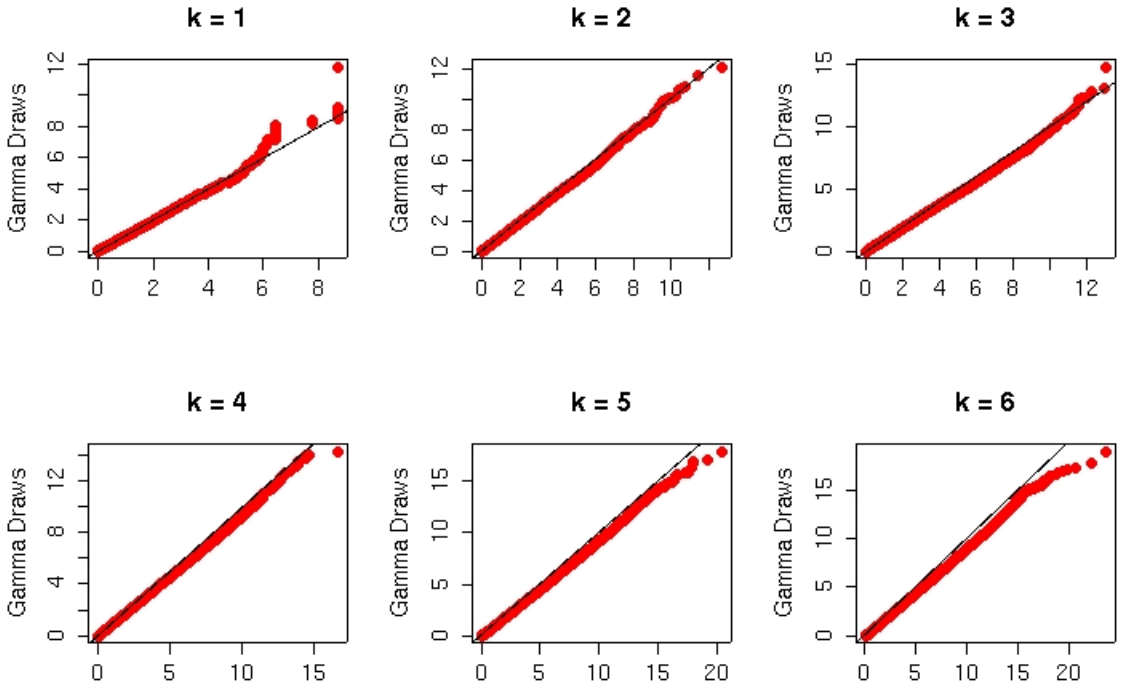

Figure 5.5: Q-Q plots to assess the approximation (5.8). The vertical axes represent random draws from gamma distributions with shape and scale parameters defined by the values under the column "Approx." in Table 5.1. The horizontal axes are the values $\log p(y|\boldsymbol{\gamma}) + \log p^*(y|\boldsymbol{\gamma})$ for the 100,000 randomly sampled models corresponding to the column "Random" in Table 5.1.

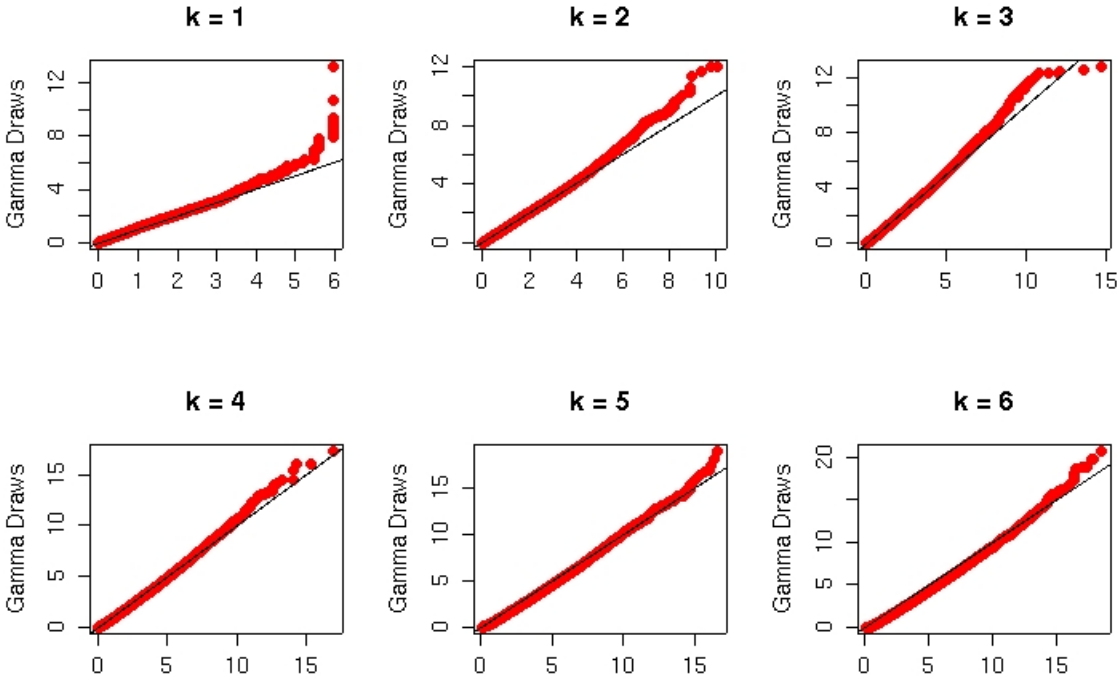

Figure 5.6: Q-Q plots to assess the approximation (5.8). The vertical axes represent random draws from gamma distributions with shape and scale parameters defined by the values under the column "Approx." in Table 5.1. The horizontal axes are the values $\log p(y|\boldsymbol{\gamma}) + \log p^*(y|\boldsymbol{\gamma})$ for the 100,000 randomly sampled Keck models corresponding to the column "Keck" in Table 5.1.

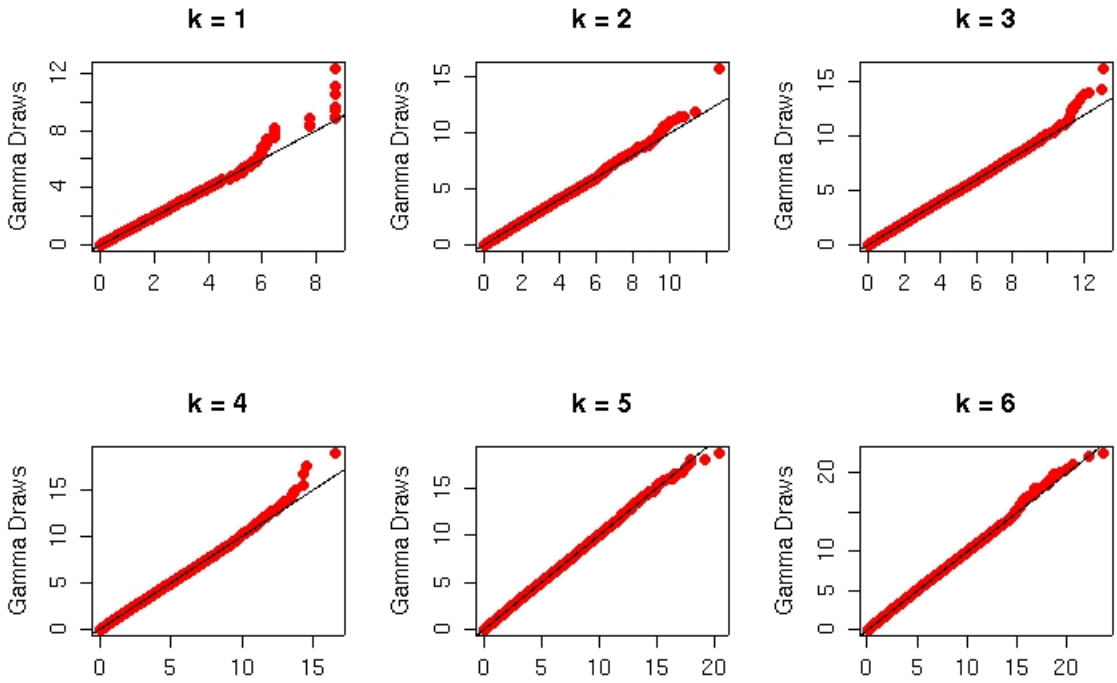

Figure 5.7: Q-Q plots to assess a general gamma distribution fit to $\log f(X; y)$. The vertical axes represent random draws from gamma distributions with shape and scale parameters defined by the method of moments estimates in column "Random" from Table 5.1. The horizontal axes are the values $\log p(y|\boldsymbol{\gamma}) + \log p^*(y|\boldsymbol{\gamma})$ for the 100,000 randomly sampled models corresponding to the column "Random" in Table 5.1.

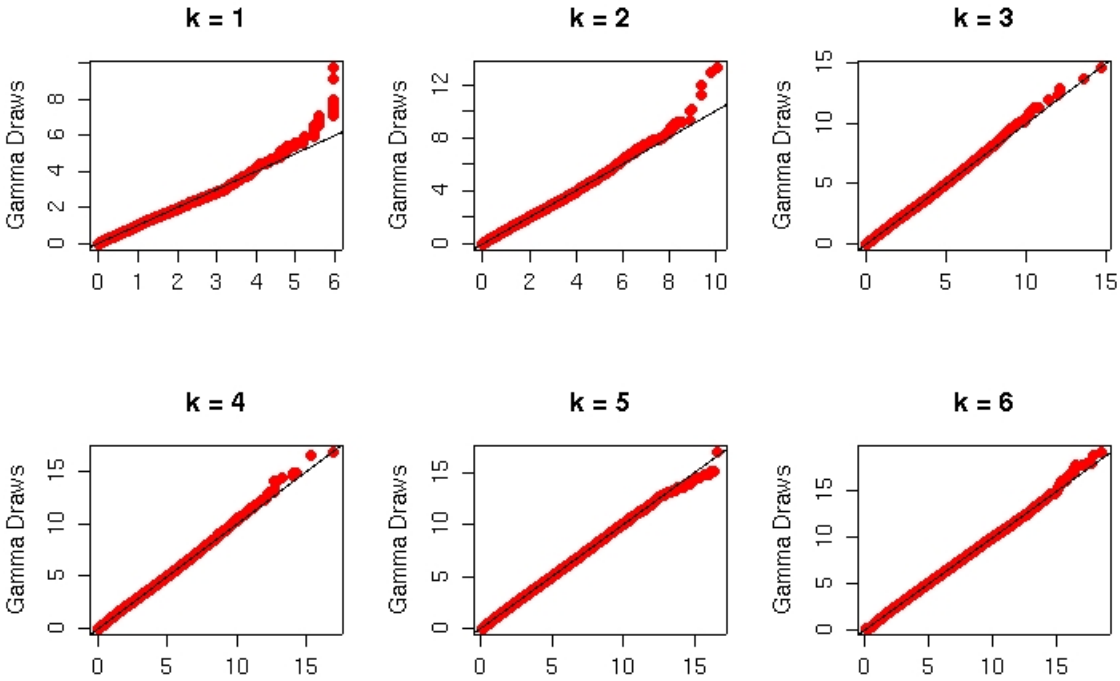

Figure 5.8: Q-Q plots to assess a general gamma distribution fit to $\log f(X; y)$. The vertical axes represent random draws from gamma distributions with shape and scale parameters defined by the method of moments estimates in column "Keck" from Table 5.1. The horizontal axes are the values $\log p(y|\boldsymbol{\gamma}) + \log p^*(y|\boldsymbol{\gamma})$ for the 100,000 randomly sampled models corresponding to the column "Keck" in Table 5.1.

wise, define the *singular value decomposition* of a centered and scaled $n \times k$ matrix X as

$$X = UDV',$$

where $U'U = I_k$ the columns of U contain the left singular vectors of X, D is a diagonal matrix whose elements are the k singular values of X, and $V'V = VV' = I_k$. Rewrite this decomposition as $X = AF$, where

$$\begin{aligned} A &= (n-1)^{1/2}U, \\ F &= (n-1)^{-1/2}DV'. \end{aligned}$$

Then $A'A = (n-1)I_k$, i.e. the columns of A are scaled to have mean zero and unit variance, and $FF' = (n-1)^{-1}D^2$. We can then reparameterize the model as

$$\begin{aligned} y &= X\boldsymbol{\beta} + \epsilon \\ &= AF\boldsymbol{\beta} + \epsilon \\ &= A\boldsymbol{\mu} + \epsilon, \end{aligned}$$

where $\epsilon \sim \mathrm{N}(0, \sigma^2 I_n)$. The induced prior on $\boldsymbol{\mu}$ is

$$p(\boldsymbol{\mu}|\sigma^2, \boldsymbol{\gamma}) = \mathrm{N}\left(0\ ,\ \frac{\sigma^2}{(n-1)\tau}D^2\right).$$

Under this parameterization, the marginal likelihood is

$$p(y|X) = \frac{\Gamma\left(\frac{n+\delta+k}{2}\right)/\Gamma\left(\frac{\delta+k}{2}\right)}{\pi^{n/2}\tau^{(n-k)/2}|\tau I_k + D^2|^{1/2}\{1 + q_*/\tau\}^{(n+\delta+k)/2}}, \tag{5.9}$$

where $q_* = y'y - y'AM_*^{-1}A'y$, and $M_* = (n-1)(I_k + \tau D^{-2})$. Note also that

$$|\tau I_k + D^2| = \prod_{j=1}^{k}(\tau + d_j^2),$$

and

$$q_* = y'y - y'U \begin{pmatrix} \frac{d_1^2}{\tau+d_1^2} & & \\ & \ddots & \\ & & \frac{d_k^2}{\tau+d_k^2} \end{pmatrix} U'y.$$

Using the notation of Dawid (1981), $X \sim \mathrm{N}(I_n, I_k)$ means that the $n \times k$ matrix X has a matrix normal distribution, where the rows are independent and identically distribution $\mathrm{N}(0, I_k)$ random variables. We can write $U = XVD^{-1}$, and so under the matrix normality assumption on X, $U \sim \mathrm{N}(I_n, D^{-2})$. As D^{-2} is a diagonal matrix, we have that the rows of U are independent and identically distribution $\mathrm{N}(0, D^{-2})$ random variables, implying also that the columns are independent of each other.

Let

$$B = \begin{pmatrix} \frac{d_1}{\sqrt{\tau+d_1^2}} & & \\ & \ddots & \\ & & \frac{d_k}{\sqrt{\tau+d_k^2}} \end{pmatrix}.$$

Setting $U^* = UB$, then $U^* \sim \mathrm{N}(I_n, C)$, where

$$C = \begin{pmatrix} \frac{1}{\tau+d_1^2} & & \\ & \ddots & \\ & & \frac{1}{\tau+d_k^2} \end{pmatrix}.$$

The quantity q_* can now be written as $y'y - y'U^*U^{*\prime}y$. For fixed y, we have $y'U^* \sim \mathrm{N}(I_1, (n-1)C)$, i.e., $U^{*\prime}y \sim \mathrm{N}(0, (n-1)C)$. Let $z = (n-1)^{-1/2}U^{*\prime}y$. Then $q_* = (n-1)(1 - z'z)$, where $z \sim \mathrm{N}(0, C)$. Also,

$$z'z = \sum_{j=1}^{k} \frac{1}{\tau + d_j^2} \chi_1^2,$$

a mixture of standard chi-squared random variables with unequal mixing weights. We can write

$$\begin{aligned} f(X;y) &= \frac{p(y|X)}{p^*(y|X)} \\ &= \frac{(\tau+n-1)^{\frac{k}{2}}}{\left[\prod_{j=1}^{k}(\tau+d_j^2)^{1/2}\right]\left(1-\frac{n-1}{\tau+n-1}\sum_{j=1}^{k}\frac{1}{\tau+d_j^2}\chi_1^2\right)^{\mu}}. \end{aligned}$$

The condition that X is an orthogonal (scaled) matrix corresponds to $D^2 = (n-1)I_k$. If this is the case, then

$$f(X;y) = \left(1-\frac{n-1}{(\tau+n-1)^2}\chi_k^2\right)^{-\mu},$$

which is the same as (5.4). So the "closer" that X is to orthogonal, the closer the d_j^2 will be to $n-1$, and hence the closer the (scaled) mixture of χ_1^2 random variables will be to the gamma distribution in (5.5).

5.3 Bayesian Inference on Sparsity

The probability model specified in Chapter 3 required the parameter controlling sparcity, π, to be specified *a priori*. This choice is quite arbitrary; in general, small values serve to enforce sparsity in the model space posterior distribution, however the posterior can be sensitive to changes in π, as demonstrated in Chapter 4. As seen in Figure 5.2, setting $\pi = 0.5$ imposed much less of a penalty on dimension than $\pi = 2/p$. These two values of π may be sufficiently different enough that, *a priori*, we may be able to choose between them, however it is much less clear whether $\pi = 2/p$ or $\pi = 10/p$ truly reflects our prior beliefs.

Two possible ways to avoid the problem of having to arbitrarily specify π are to (i) include π as a parameter in the model by assigning it a prior distribution (e.g., Chipman *et al.*, 2001), and (ii) estimate π using an empirical Bayes-type approach (e.g., George and Foster (2000), Cui and George (2004) and Johnstone and Silverman (2004, 2005) in a related setting). I first explore the latter by describing how to construct a marginal likelihood for π using output from a SSS and the previous results from this chapter.

5.3.1 Marginal Likelihood for π

The marginal likelihood $p(y|\pi)$ is computed by marginalizing over the model space:

$$\begin{aligned} p(y|\pi) &= \sum_{\gamma \in \Gamma} p(y, \boldsymbol{\gamma}|\pi) \\ &= \sum_{\gamma \in \Gamma} p(y|\boldsymbol{\gamma}, \pi) p(\boldsymbol{\gamma}|\pi) \\ &= \sum_{\gamma \in \Gamma} p(y|\boldsymbol{\gamma}) p(\boldsymbol{\gamma}|\pi), \end{aligned} \tag{5.10}$$

where $p(\boldsymbol{\gamma}|\pi)$ is the prior distribution over the model space described in Section 3.2,

$$p(\boldsymbol{\gamma}|\pi) = \pi^k (1-\pi)^{p-k}.$$

For each dimension k, define the corresponding subset of Γ to be

$$\Gamma^{(k)} = \{\boldsymbol{\gamma} \in \Gamma \ : \ |\boldsymbol{\gamma}| = k\},$$

$k \in \{0, \ldots, p\}$, noting that $\Gamma = \bigcup_{k=0}^{p} \Gamma^{(k)}$. We can then rewrite (5.10) as

$$\begin{aligned} p(y|\pi) &= \sum_{k=0}^{p} \sum_{\gamma \in \Gamma^{(k)}} p(y|\boldsymbol{\gamma}) p(\boldsymbol{\gamma}|\pi) \\ &= \sum_{k=0}^{p} \pi^k (1-\pi)^{p-k} \sum_{\gamma \in \Gamma^{(k)}} p(y|\boldsymbol{\gamma}). \end{aligned} \tag{5.11}$$

We can evaluate the marginal likelihood $p(y|\pi)$ over a grid if we can compute the inner sums in (5.11); this is practically intractable, however, as there are too many models to evaluate. If we could instead estimate the average value of the marginal likelihoods within each dimension,

$$\overline{m}_k = \binom{p}{k}^{-1} \sum_{\gamma \in \Gamma^{(k)}} p(y|\boldsymbol{\gamma}), \tag{5.12}$$

we could then estimate the marginal likelihood,

$$p(y|\pi) = \sum_{k=0}^{p} \overline{m}_k \binom{p}{k} \pi^k (1-\pi)^{p-k}. \tag{5.13}$$

5.3.2 Bounding the Model Space

In the context of the examples considered througout, prior belief places little support on models with many variables, and in general focus is on mixing over models with only a few predictor variables. This prior belief in sparsity is well-encoded in the model by placing an upper bound on the number of variables allowed in any model. For the rest of this chapter, I entertain a restricted model space

$$\Gamma_{k^*} = \{\boldsymbol{\gamma} \in \Gamma \ : \ |\boldsymbol{\gamma}| \leq k^*\},$$

and update the prior distribution over the model space as

$$p(\boldsymbol{\gamma}|\pi) = \pi^k(1-\pi)^{p-k}C_{k^*}(\pi)^{-1}\mathbf{1}(k < k^*),$$

where

$$C_{k^*}(\pi) = \sum_{l=0}^{k^*} \binom{p}{l} \pi^l(1-\pi)^{p-l}.$$

Using this restricted model space, the marginal likelihood $p(y|\pi)$ is now

$$p(y|\pi) = C_{k^*}(\pi)^{-1} \sum_{k=0}^{k^*} \overline{m}_k \binom{p}{k} \pi^k(1-\pi)^{p-k}. \tag{5.14}$$

5.3.3 Estimating $\overline{m}_k$

We can use the output from SSS along with the results from Section 5.2.1 to construct a conservative estimate of $\overline{m}_k$. First, rather than allowing SSS to move across dimension, a fixed dimensional SSS should be run on dimension k separately for $k = 2, \ldots, k^*$, providing lists of B best models evaluated, $\Gamma^{*(k)}$. The model space for each dimension can be broken down into two components, models we have recorded and models we have not recorded:

$$\Gamma^{(k)} = \Gamma^{*(k)} \cup \Gamma^{-(k)},$$

where $\Gamma^{-(k)}$ is the latter set. The quantity $\overline{m}_k$ can then be rewritten as

$$\overline{m}_k = \binom{p}{k}^{-1} \left[\sum_{\gamma \in \Gamma^{*(k)}} p(y|\boldsymbol{\gamma}) \; + \sum_{\gamma \in \Gamma^{-(k)}} p(y|\boldsymbol{\gamma}) \right]. \tag{5.15}$$

The first sum in (5.15) can be computed using the output from the fixed dimensional SSS search, while the second sum needs to be estimated.

The second sum represents those models of dimension k that were *not* in the list of B top models found by SSS. If we had taken B just large enough to capture all the "interesting" models, then then set $\Gamma^{-(k)}$ would represent those models that are uninformative about the outcome, i.e., those models having y uncorrelated with the comprising variables. If this were the case, we could then use the results from Section 5.2.1 to estimate the average value of the marginal likelihood for the models in $\Gamma^{-(k)}$ and then multiply this quantity by $\binom{p}{k} - B$ to obtain an estimate of the second sum in (5.15).

The parametric approximation (5.8) is to the (shifted) log marginal likelihood, and so to compute the average value of the marginal likelihood for the models in $\Gamma^{-(k)}$, we need to transform $W = \exp\{Z + c_k\}$, where Z is distributed according to (5.8) and $c_k = \log p^*(y|\boldsymbol{\gamma})$ for a model of size k. The resulting density function is

$$p(w) = \frac{\beta^\alpha e^{\beta c_k}}{\Gamma(\alpha)} w^{-(\beta+1)} (\log w - c_k)^{\alpha-1}, \quad \exp(c_k) \leq w < \infty, \tag{5.16}$$

where $\alpha = k/2$ and $\beta = (\tau + n - 1)^2/[(n-1)(n+\delta+k)]$. Note that when $\tau = 1$, $\beta < 1$ for all $n > (\delta + k)/(\delta + k - 1)$, which is effectively $n > 1$ when $\delta = 3$ and $k \geq 0$. The expected value is

$$\begin{aligned}
\mathbb{E}[W] &\propto \int_{e^{c_k}}^{\infty} w^{-\beta} (\log w - c_k)^{\alpha-1} \, dw \\
&\geq \int_{1}^{\infty} w^{-\beta} (\log w - c_k)^{\alpha-1} \, dw \qquad (5.17) \\
&\geq \text{const.} \int_{1}^{\infty} w^{-\beta} \, dw. \qquad (5.18)
\end{aligned}$$

Line (5.17) follows because $c_k < 0$ for values of n, δ, τ and k of interest, and line (5.18) follows because $\log w - c_k > 0$. The integral (5.18) diverges when $\beta < 1$,

and so $\mathbb{E}[W]$ does not exist. Hence we cannot simply take $\mathbb{E}[W]$, multiply it by $\binom{p}{k} - B$ and substitute it for the second sum in (5.15).

Alternatively, we can substitute $\exp\{\mathbb{E}[Z] + c_k\}$ for the second sum in (5.15); by Jensen's inequality, if $\mathbb{E}[W]$ did exist, then

$$\exp\{\mathbb{E}[Z] + c_k\} \leq \mathbb{E}[W], \tag{5.19}$$

and so the estimate would be conservative in that it would be closer to the lower bound.

5.3.4 Keck Data Example

Fixed dimensional SSS was run on the Keck dataset as described in Chapter 4 for dimensions $k = 2, \ldots, k^*$, where k^* is taken to be six. Additionally the marginal likelihoods for the null model and all one variable models were computed. The lists $\Gamma^{*(k)}$ for $k = 2, \ldots, 6$ each contain one million models; boxplots are shown in Figure 5.9. The values $\sum_{\gamma \in \Gamma^{*(k)}} p(y|\boldsymbol{\gamma})$ and the estimates of $\sum_{\gamma \in \Gamma^{-(k)}} p(y|\boldsymbol{\gamma})$ are given in Table 5.2.

Figure 5.10 displays the estimate of the resulting marginal likelihood, $p(y|\pi)$. The modal value is zero, indicating that the data support very small models. In an empirical Bayes type approach, $\arg\max p(y|\pi) = 0$ would be the estimate for π, indicating the null model is preferred. If a fully Bayesian approach were taken, and $p(y|\pi)$ treated as an unormalized posterior under a flat prior $p(\pi) = \mathbf{1}(0 < \pi < 1)$, an estimate of π might be the posterior mean. This can be computed numerically to be 2.625e-3, as indicated in Figure 5.10, corresponding to a model with on average 22 predictors. The flat prior corresponds to an *a priori* average value of $\pi = 0.5$, which is much different than our actual beliefs. If a fully Bayesian approach is to be taken, priors on π that represent belief in sparsity must be used.

Table 5.2: Sums of marginal likelihoods for models in $\Gamma^{*(k)}$ for the Keck example. The lower panel displays estimates of the mass in $\Gamma^{-(k)}$ using the method based on (5.19).

k :	0	1	2	3	4	5	6
$\Gamma^{*(k)}$	1.22e-26	2.86e-23	2.65e-20	3.78e-17	6.94e-14	1.81e-11	3.77e-10
$\Gamma^{-(k)}$	0	0	1.10e-20	4.38e-18	1.19e-15	2.48e-13	4.15e-11
Total	1.22e-26	2.86e-23	3.75e-20	4.22e-17	7.06e-14	1.83e-11	4.19e-10

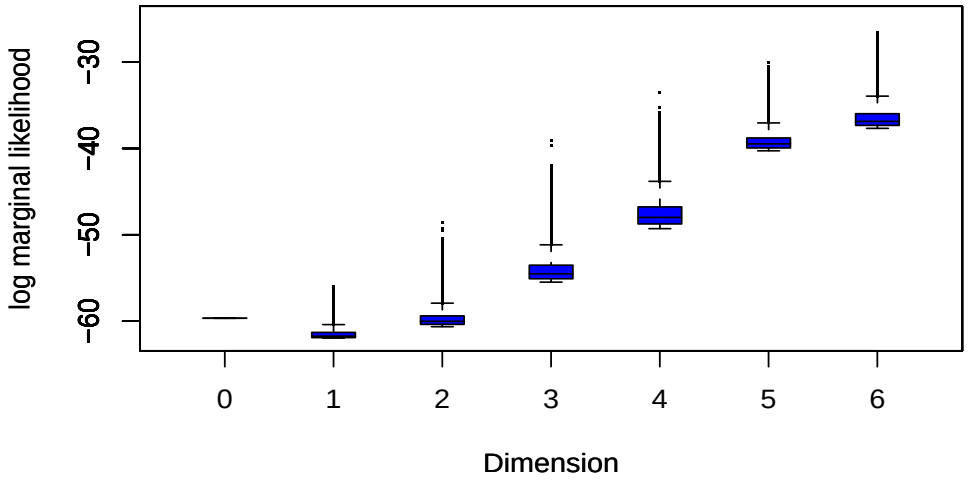

Figure 5.9: Marginal likelihood values from fixed dimensional runs of SSS.

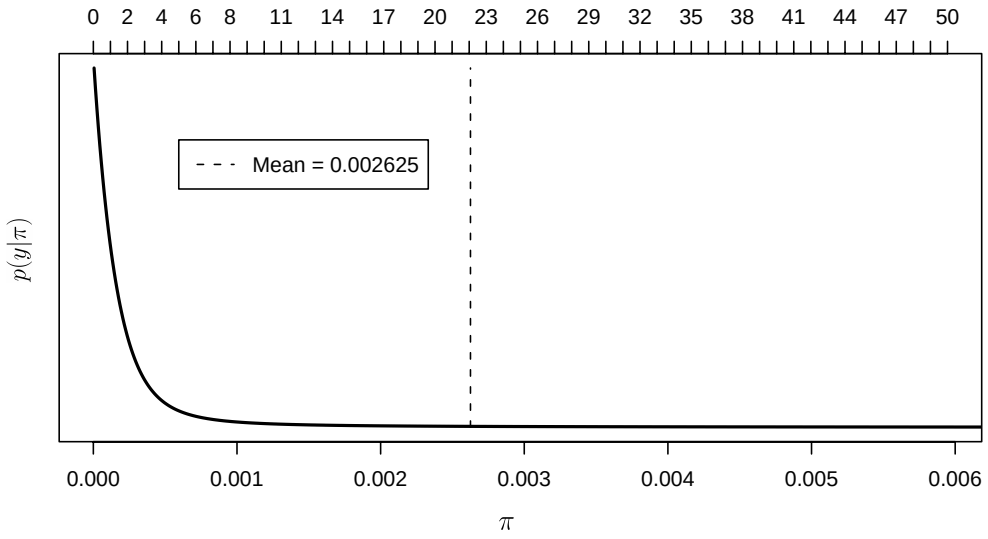

Figure 5.10: Marginal likelihood for π for the Keck data, with $k^* = 6$. The axis on top of the figure denotes the values $p\pi$.

5.3.5 Marginal Posterior Distribution for π

Rather than taking an empirical Bayes type approach where $p(y|\pi)$ is used to formulate an estimate of π, we can instead use the approximation to the marginal likelihood to do full Bayesian learning about the sparsity controlling parameter π. If we assign a prior distribution to π, the resulting marginal posterior is

$$p(\pi|y) \propto p(y|\pi)p(\pi). \tag{5.20}$$

If a uniform prior, $p(\pi) = 1$, is assigned as discussed above, then (5.13) and (5.14) are proportional to (5.20), and the marginal posterior can be evaluated over a grid on the interval $[0, 1]$.

More flexibly, and more concordantly with prior belief, we can assign a beta prior distribution,

$$p(\pi) = \text{Beta}(k', p), \tag{5.21}$$

where p is the total number of predictors and k' is specified *a priori*. If we take $k' \ll p$, then

$$\begin{aligned} \mathbb{E}[\pi] &= \frac{k'}{k'+p} \approx \frac{k'}{p} \\ \mathbb{V}[\pi] &= \frac{k'p}{(k'+p)^2(k'+p+1)} \approx \frac{k'}{p^2}. \end{aligned}$$

Note that this beta distribution is well-approximated by a gamma distribution, as seen above by the first two moments, in a manner similar to the Poisson approximation to the $\text{Bin}(n,p)$ distribution as $n \to \infty$ for fixed np.

Under formulation (5.21), the prior mean of π approximately corresponds to an expected model size of k' if we did not impose an upper bound k^*. Hence this prior is, in a sense, a stochastic version of the independent Bernoulli prior (3.7), where now the prior variance of π is reasonable for the size of regression models considered here. Using prior (5.21) and setting an upper bound on model size, the resulting posterior distribution is

$$\begin{aligned} p(\pi|y) &\propto C_{k^*}(\pi)^{-1} \left[\sum_{l=0}^{k^*} \overline{m}_k \binom{p}{l} \pi^l (1-\pi)^{p-l} \right] \pi^{k'-1}(1-\pi)^{p-1} \\ &= C_{k^*}(\pi)^{-1} \sum_{l=0}^{k^*} \overline{m}_k \binom{p}{l} \pi^{l+k'-1}(1-\pi)^{2p-l-1}. \end{aligned} \tag{5.22}$$

Again, this can be computed on a grid of values for π on the interval $[0,1]$. Note that if $k^* = p$, (5.22) would be a mixture of beta distributions. Priors and posteriors for various k' in the Keck example are shown in Figure 5.11. Posterior means, modes and highest posterior density regions can then be used to gauge the amount of sparsity supported by the data.

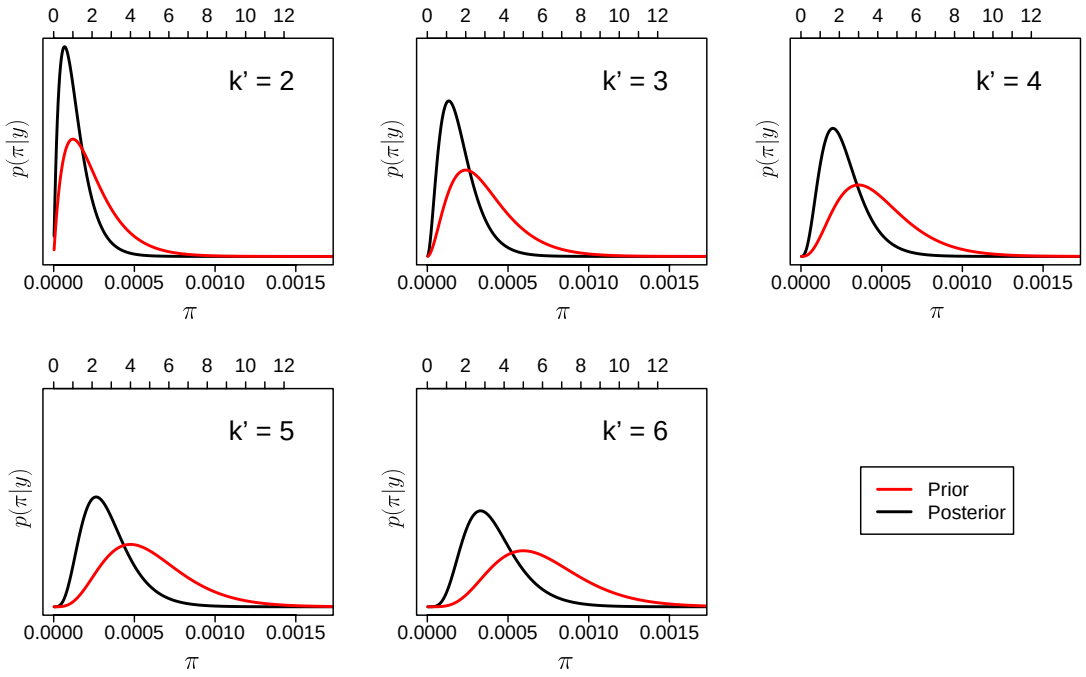

Figure 5.11: Marginal posterior distributions $p(\pi|y)$ for the Keck data, with $k^* = 6$, for several values of k'. The axes on top of the figures denote the values $p\pi$.

The resulting posteriors are sensitive to the choice of k' just as the model space posterior was sensitive to the choice of k' when it was used to set a fixed value $\pi = k'/p$ for the independent Bernoulli prior. The fully Bayes formulation is less rigid, however, as it allows for variability in π and in turn allows the data to inform more strongly about reasonable model sizes.

Chapter 6

Further Examples in Clinico-Genomics

Regression modeling is not limited to the case of the normal linear model, and resulting methodology is needed that addresses model search in large p scenarios. A key to the search methodology developed in Chapter 3 was that the regression model scores, $p(y|\boldsymbol{\gamma})p(\boldsymbol{\gamma})$, could be computed quickly in closed form. This is typically not the case for generalized linear models, where marginal likelihoods must be approximated numerically. In this chapter I extend the SSS search method to the cases of binary regression and survival time modeling via Weibull regression.

6.1 Binary Regression Models

In the case of independent binary outcomes, y_i, assume the logistic regression model framework and set $p(y|\boldsymbol{\beta}, \boldsymbol{\gamma}) = \prod_{i=1}^{n} \phi_i^{y_i}(1-\phi_i)^{1-y_i}$, where $\phi_i = 1/(1 + \exp\{-(\beta_0 + \mathbf{x}_i'\boldsymbol{\beta}_\gamma)\})$. Here $\mathbf{x}_i$ is taken to mean the dependent variable vector for respondent i containing only those variables indicated by the model $\boldsymbol{\gamma}$. Even though the data matrix X is assumed to be standardized, the inclusion of the intercept term β_0 is necessary to account for the baseline response probability.

While the methods below are derived for logistic regression, any link function could be used.

6.1.1 Prior Distributions

For a given model, the parameter space prior distribution is taken to be

$$[\beta_0, \boldsymbol{\beta}_\gamma | \boldsymbol{\gamma}] \sim \mathrm{N}(0, \tau I_{k+1}), \tag{6.1}$$

where k is the number of variables in model $\boldsymbol{\gamma}$. As the predictor variables have been standardized to have common (unit) variance, τ is taken to be one. Figure 6.1 shows the implied prior on ϕ_i under the logistic link function for models of sizes $k = 0, \ldots, 3$ when $\tau = 1$. In each case with $k > 0$, the value of the predictor variable is set to be an "extreme" value, 1.96, which is the 97.5th quantile of the standard normal distribution. Prior mass accumulates at the boundaries as k increases. A smaller prior variance would slow this accumulation, as shown in Figure 6.2, where $\tau = 0.5$.

The prior distribution for the model space is the same as for the linear model, a product of indepdent Bernoulli random variables as in (3.7) with $\pi = k'/p$. In the context of binary regrssion models p represents the total number of possible predictor variables excluding the intercept, as the intercept is included in every model. Similarly, the "dimension" or "size" of a model will not include the intercept parameter, hence $k = 0$ refers to the model with only β_0 and $\boldsymbol{\gamma}$ is taken to be a $p \times 1$ vector.

6.1.2 Marginal Likelihood

The marginal likelihood $p(y|\boldsymbol{\gamma})$ is not available in closed form, however a numerical approximation can be made using a Laplace approximation. Following DiCiccio

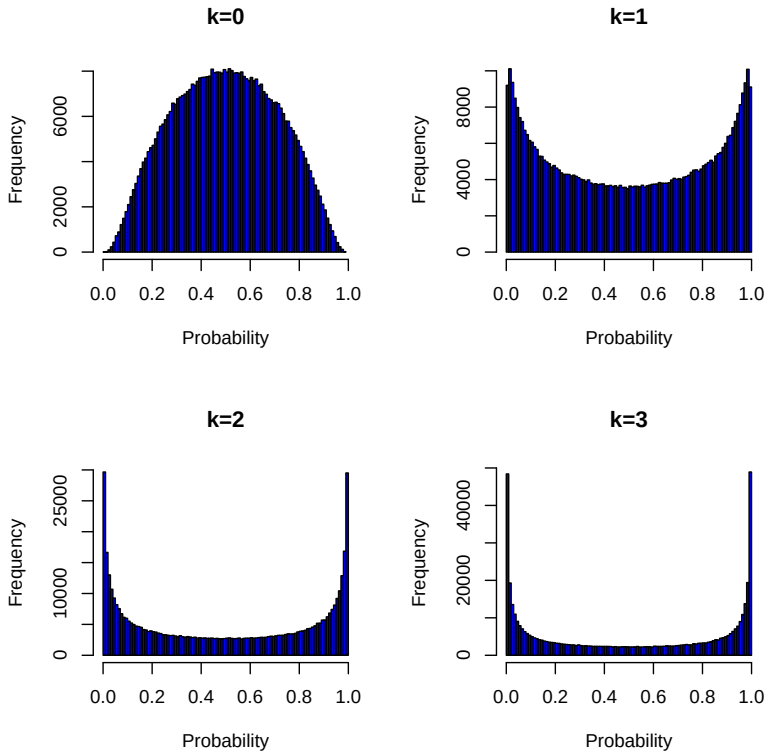

Figure 6.1: Implied prior for ϕ_i when $\tau = 1$ for models with k predictor variables, each of which is set to be $x = 1.96$, the 97.5th quantile of the standard normal distribution.

et al. (1997), letting $h(\boldsymbol{\beta}|\boldsymbol{\gamma}) = p(y|\boldsymbol{\beta}, \boldsymbol{\gamma})p(\boldsymbol{\beta}|\boldsymbol{\gamma})$ we can estimate

$$\begin{aligned} \hat{p}(y|\boldsymbol{\gamma}) &= (2\pi)^{(k+1)/2}|\hat{\Sigma}|^{1/2}h(\hat{\boldsymbol{\beta}}|\boldsymbol{\gamma}) \qquad (6.2) \\ &= |\hat{\Sigma}|^{1/2}\tau^{-(k+1)/2}\exp\left\{-\frac{\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}}{2\tau}\right\}\prod_{i=1}^{n}\phi_i(\hat{\boldsymbol{\beta}})^{y_i}(1-\phi_i(\hat{\boldsymbol{\beta}})^{1-y_i}, \end{aligned}$$

where

$$\hat{\Sigma} = -\left(\frac{\partial^2 \log h(\hat{\boldsymbol{\beta}}|\boldsymbol{\gamma})}{\partial\hat{\beta}_i\partial\hat{\beta}_j}\right)^{-1}. \qquad (6.3)$$

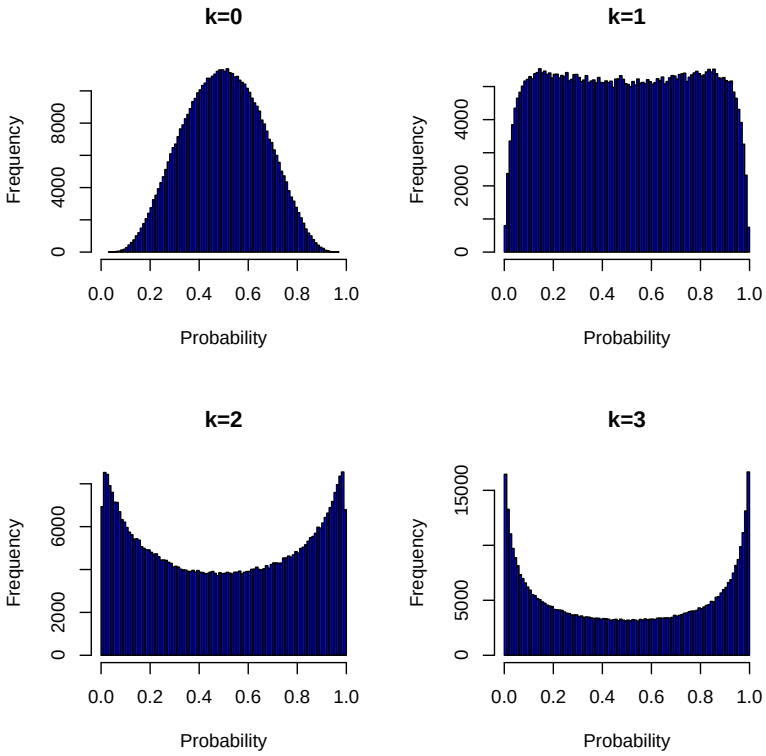

Figure 6.2: Implied prior for ϕ_i when $\tau = 0.5$ for models with k predictor variables, each of which is set to be $x = 1.96$, the 97.5th quantile of the standard normal distribution.

The posterior mode, $\hat{\boldsymbol{\beta}} = \arg\max_\beta p(y|\boldsymbol{\beta}, \boldsymbol{\gamma})p(\boldsymbol{\beta}|\boldsymbol{\gamma})$, can be found via Newton's method; details are given in Appendix D.

6.1.3 Posterior Summarization

Posterior quantities of interest, such as $\tilde{p}(\boldsymbol{\gamma}|y)$ can be found as before using SSS output with the marginal likelihood estimates $\hat{p}(y|\boldsymbol{\gamma})$. To simulate from the model averaged distribution of fitted values, as was described for linear models in Section 3.3, F models are first sampled from the approximation to the model space posterior based on the values $\tilde{p}(\boldsymbol{\gamma}|y)$ using Γ^*. For each sample f, a draw of the re-

gression coefficients needs to be made from their posterior distribution, $p(\boldsymbol{\beta}|\boldsymbol{\gamma}, y)$, which is not available in closed form. One possible method of obtaining a sample is to use an MCMC method such as a Metropolis-Hastings algorithm with a simple proposal distribution, retaining the last value as the needed sample. Alternatively, adaptive rejection sampling (Gilks and Wild, 1992) could be used in a Gibbs sampler to draw from the full conditional distributions, as they are log concave.

A fast sampling method is to approximate the posterior with a normal distribution, centered at the posterior mode $\hat{\boldsymbol{\beta}}$ with covariance matrix Σ^{-1} as in (6.3), which is the correct distribution asymptotically. No additional computation is required, as these quantities are calculated in order to estimate the marginal likelihood and can be stored with the model. For each sampled model, f, a draw of the regression coefficients $\boldsymbol{\beta}^{(f)}$ is then made from the approximate posterior distribution, $\mathrm{N}(\hat{\boldsymbol{\beta}}, \hat{\Sigma})$. For each $\boldsymbol{\beta}^{(f)}$ the value $\phi_i^{(f)} = 1/(1 + \exp\{-\mathbf{x}_i'\boldsymbol{\beta}^{(f)}\})$ is computed. Summaries of the posterior distribution of each ϕ_i can then be plotted (e.g. Figure 6.3).

Cross-validated prediction can be performed as in Section 3.4, where the marginal likelihoods for the models in Γ^* are recomputed n times, each time holding out a different sample. This is now more computationally demanding as the marginal likelihoods must be approximated numerically, however the calculations can be done in parallel. The cross-validated predictive distributions can be sampled from in parallel using the method described above, and resulting estimates and intervals for the ϕ_i can be computed.

6.2 Example: Predicting Lymph Node Status

The data are coupled gene expression and lymph node positivity status in human breast cancers. From a data base of 348 cases, individuals were identified who were clinically defined as low risk for disease recurrence, or death from disease, in terms of lymph node negativity (no evidence of cancer metastasis in the axillary lymph nodes) at the point of surgery; these patients are compared to those that are in a generally far higher risk group, i.e. those with at least nine nodes in the axillary regions showing evidence of cancer metastasis. This analysis follows previous work and relates to the general interest in the potential for tumor derived gene expression profiles to aid in prognosis – in this case, improved prediction of low versus high risk based on genomic information could feed into decisions about post-surgical treatments (West *et al.*, 2001; Huang *et al.*, 2002; Nevins *et al.*, 2003; Huang *et al.*, 2003; Pittman *et al.*, 2004). Prediction of lymph node status based on gene expression profiles is a challenging problem, due to the complex heterogeneity of the disease in terms of genetic and environmental factors, and also as a result of the levels of experimental and technical noise in gene expression data. Advances in the ability to better predict lymph node status would be of substantial interest in clinical cancer genomics.

The data consist of $n = 148$ samples with $n_0 = 100$ low risk (node negative) and $n_1 = 48$ high risk (high node positive) cases. Gene expression data is available on Affymetrix HU95aV2 oligonucleotide microarrays, which were processed using the current standard RMA method (Irizarry *et al.*, 2003a,b) to generate summary estimates of expression levels of each gene in the sample. This primary RMA data was then further screened and normalized, and a total of $4,512$ genes showing

evidence of more than trivial variation above the noise level were selected to include in the analysis. In addition to these candidate predictors, each patient has a number of traditional clinical factors available, including an estimate of tumor size in centimeters and protein assay-based estrogen receptor (ER) status, coded as a binary covariate. Using the gene expression data together with these two clinical factors thus provides $p = 4,514$ candidate predictors in total, in addition to the intercept term occurring in all models.

6.2.1 Small Subset Regression Analysis

Binary regression models are used as described above, with $y_i = 0$ denoting the node negative cases, and $y_i = 1$ denoting the (high) node positive cases. As the focus is on sparse models, the prior distribution on the model space is taken to be the independent Bernoulli prior (3.7) with $\pi = 10/p$. For the prior distribution on the regression coefficients, (6.1) was used with $\tau = 1$. SSS was run for 20,000 iterations saving the top 100,000 models evaluated.

6.2.2 Results

The top 100,000 models evaluated contain a mix of one through seven variable models, as shown in Table 6.1 in the row labeled $\tau = 1$. Recall that $k = 1$ corresponds to a model with an intercept plus one predictor variable. Under this model specification, the data give most support to regression models of size five, six and four, in that order. No models of size eight or greater were found by SSS to belong in the list of top models. The null model has a (relative) log posterior value, $\log p(y|\boldsymbol{\gamma}) + \log p(\boldsymbol{\gamma})$, of -95.273, while the worst model in Γ^* has a value of -85.896.

Table 6.1: Posterior probability of model size conditioned on the top 100,000 models evaluated by SSS for the lymph node status example. The character * represents a value < 0.001.

	k:	1	2	3	4	5	6	7
$\tau = 1$	# of models	1	54	1,311	11,838	54,597	30,619	1,580
	$\tilde{p}(\|\gamma\| = k\|y)$	*	0.001	0.020	0.184	0.534	0.253	0.007
$\tau = \frac{1}{2}$	# of models	1	162	5,576	35,863	55,900	2,498	0
	$\tilde{p}(\|\gamma\| = k\|y)$	*	0.006	0.089	0.422	0.466	0.017	0

Table 6.2: Genewise and pairwise inclusion probabilities for the top seven genes. The diagonal entries are the quantities $\tilde{p}(\gamma_j = 1|y)$, and the off-diagonal entries are the quantities $\tilde{p}(\gamma_i = \gamma_j = 1|y)$. The character * indicates a value < 0.001.

	RGS3	DXYS155E	ATP6V1F	MGC8721	VDAC1	GEM	WSB1
RGS3	0.991	0.716	0.495	0.351	0.169	0.133	0.125
DXYS155E		0.716	0.454	0.319	0.069	0.069	0.121
ATP6V1F			0.498	0.250	0.010	0.045	0.108
MGC8721				0.352	0.016	0.042	0.054
VDAC1					0.171	0.037	0.001
GEM						0.134	*
WSB1							0.125

Conditionally on Γ^*, eight genes were found to have posterior inclusion probability greater than 0.10; the top seven are given in Table 6.2. The most important gene, RGS3, occurs in almost all of the models. Pairwise importance measures (3.16) are displayed in the off diagonal entries of the table and confirm that models consisting of the top four variables dominate the list. Indeed, the four-way inclusion probability of the top four genes is 0.244, just less than a third of the total mass for five, six and seven variable models.

To asses the fit of the model, model averaged mean probabilities ϕ_i and associated 80% intervals were computed using the top ten models. Figure 6.3 plots

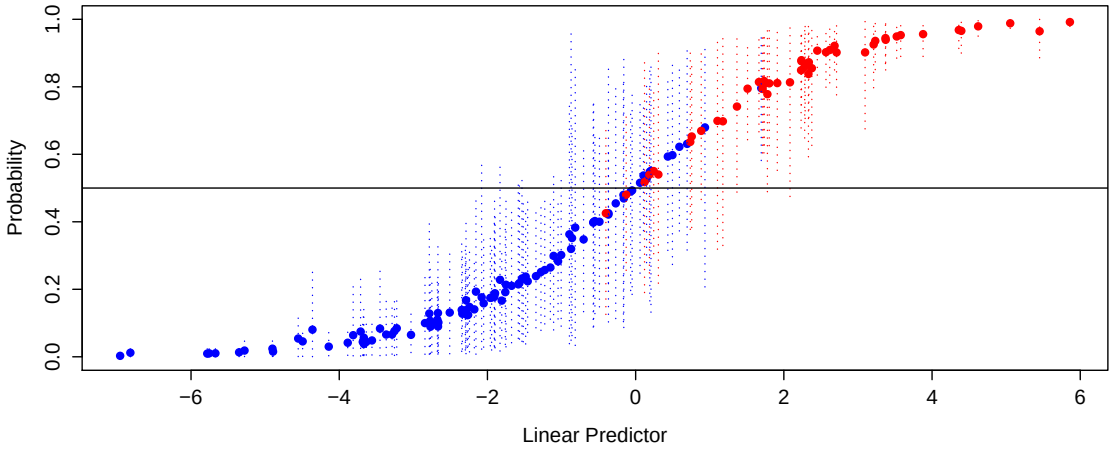

Figure 6.3: Model averaged fitted response probabilities based on the top ten models with associated 80% intervals. The red points indicate $y_i = 1$ and the blue indicate $y_i = 0$.

these model averaged fitted values vs. the linear predictor $\log(\phi_i/(1-\phi_i))$, which serves as a linear risk index. The fitted values have been corrected for the baseline incidence rate of 32.4%, making $\phi_i = 0.5$ the reference point. The model fit is quite good; 95.8% of the red points (true positives) are above 0.5, and 89% of the blue points (true negatives) are below 0.5. Further, 70.8% of the lower red quantiles are above 0.5, while 66.7% of the upper blue quantiles are below. These are patients whom we have fit properly with high probability.

To assess the way in which the top genes combine across models in a predictive context, two "metagenes" were constructed from the genes that comprise the top ten models (a total of 18 genes). The metagenes here are defined as the first two principal components from a singular value decomposition of the 18 genes. If the singular value decomposition of the data submatrix is $X = UDV'$ as in Section 5.2.3, then the two metagenes are the first two columns of V. Figure 6.4

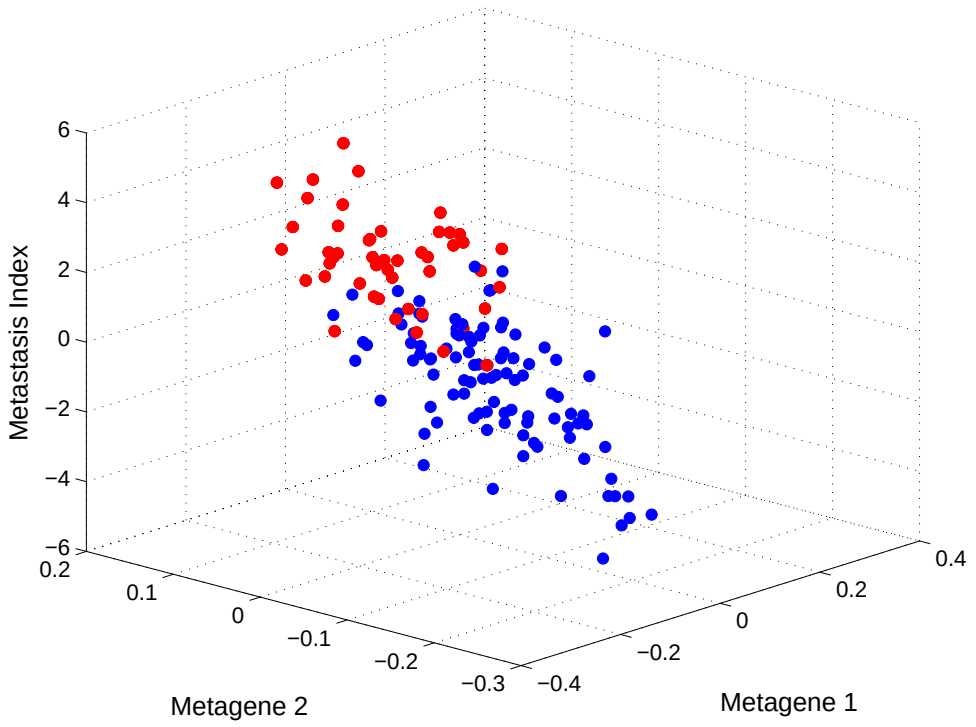

Figure 6.4: Association of the metagenes with the model averaged metastasis index (linear predictor) based on the 18 genes comprising the top ten models. Red points denote $y_i = 1$ and blue denote $y_i = 0$.

shows the association between these two metagenes and the model averaged linear predictor computed above. A concordance between the empirical metagenes and the averaged predictions is expected, but it is evident from variation in the scatter plot that the complex, data-weighted mixing over the set of regression models is generating predictions that are not captured by a single linear fit – the metagene – to the selected set of most interesting predictors.

To assess aspects of the predictive fit of the overall model, a leave-one-out cross-validation was performed as described above. After leaving out each observation one at a time and compiling the reordered lists of top models Γ_i^*, model averaged predictions of response probability for each observation were computed, using

the model averaged posterior mean values of ϕ_i based on the top 10,000 models in each list. A histogram of predicted risk index $\log(\phi_i/(1-\phi_i))$ is shown in Figure 6.5. Values to the right of zero are predicted to be node positive, while those to the left of zero are predicted to be node negative. The histogram is broken down and colored by the true values of positive or negative. On the basis of simple thresholding of these point estimates at zero, corresponding to a simple thresholding of the corresponding point predictions of metastasis, the analysis indicates an approximate sensitivity of 79.2% (for true positives) and a specificity of 76% (for true negatives). This level of predictive discrimination is quite high and suggests promise for the approach relative to prior analyses on much smaller and selected subsets of patients (West *et al.*, 2001; Huang *et al.*, 2003).

6.2.3 Alternate Analysis

The analysis was also performed for $\tau = 0.5$ to assess the sensitivity of the results to the choice of prior. Posterior probability of model size is given in the second row of Table 6.1. Five variable models still receive the most posterior support, but four variable models are now much more likely. The shift toward smaller models corresponds to the larger amount of shrinkage induced by the smaller prior variance. Under this new prior, the null model has a (relative) log posterior value of -95.190, while the worst model in Γ^* has a value of -88.399.

A comparison of the top twenty variables for the analysis with $\tau = 1$ with those from the analysis with $\tau = 0.5$ is given in Table 6.3. Besides differences in the values of $\tilde{p}(\gamma_j = 1|y)$, the two lists are fairly concordant, with most of the variables occuring in only one list having rank less than 35 on the other list. Model averaged fitted values of ϕ_i are not shown, however they are very similar.

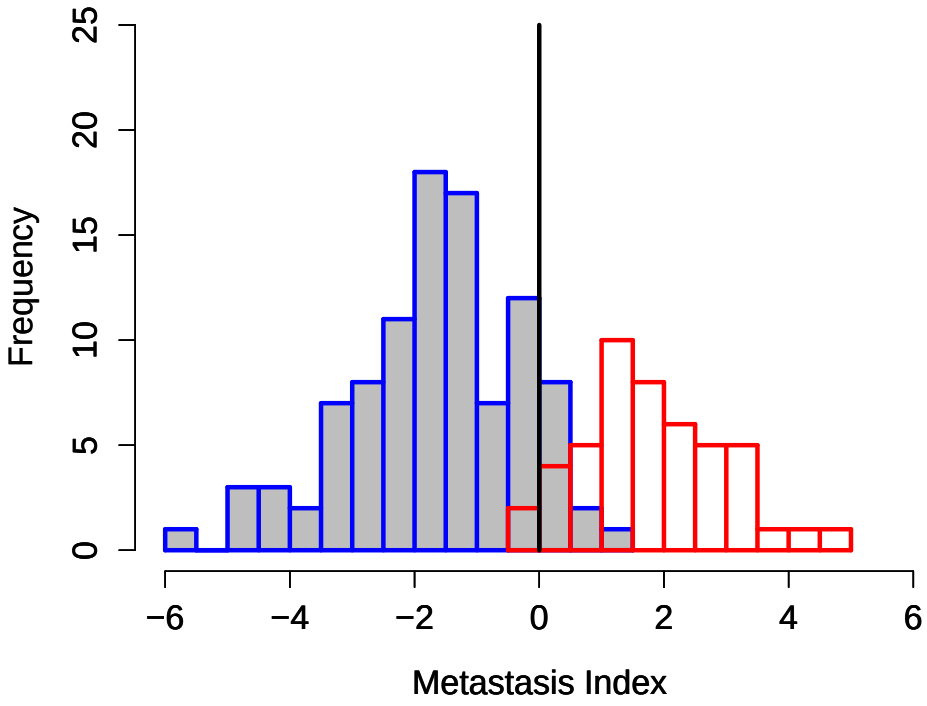

Figure 6.5: Histograms of the leave-one-out cross-validated predictions on the linear predictor scale. The blue, shaded histogram is for the true negatives and the red histogram is for the true positives.

Table 6.3: Comparison of variable inclusion probabilities for the top 20 genes from the analyses with $\tau = 1$ and $\tau = 0.5$. Red entries are those occuring in the top 20 on only one of the lists, and their ranks on the opposite list are given parenthetically.

$\tau = 1$			$\tau = 0.5$		
RGS3	0.991		RGS3	0.969	
DXYS155E	0.716		DXYS155E	0.468	
ATP6V1F	0.498		ATP6V1F	0.351	
MGC8721	0.352		VDAC1	0.256	
VDAC1	0.171		GEM	0.164	
GEM	0.134		PRRG1	0.134	
WSB1	0.125		MGC8721	0.098	
PRRG1	0.110		WSB1	0.086	
UBC	0.075	(22)	LOC283970	0.069	
KIF13B	0.065		SULT2B1	0.063	
HSPA9B	0.062		TOMM40	0.062	
OGT	0.054		KIFI3B	0.053	
PJA2	0.054	(98)	FAM38A	0.051	(31)
LOC283970	0.053		HSPA9B	0.049	
SULT2B1	0.051		GNAS	0.047	(33)
TOMM40	0.050		OGT	0.047	
MGC5508	0.047	(23)	GNAS′	0.045	(32)
—	0.041	(32)	HSPC1111	0.043	(24)
ZNF364	0.039		tumor size	0.039	(46)
DVL3	0.031	(26)	ZNF364	0.038	

Overall, the results from both analyses are similar enough to not greatly affect inference.

6.3 Survival Modeling via Weibull Regression

The survival time model used in Chapter 4 was the log-normal model, which resulted in closed form calculations of the marginal likelihood. Here I analyze a second set of survival time data that now includes censoring information using a more flexible family of models based on the Weibull distribution (see, e.g. Ibrahim

et al., 2001, Chapter 2).

Suppose we have survival times y_i for $i = 1, \ldots, n$ subjects, and suppose that the survival times follow a Weibull distribution, Weib(α, λ), with density function

$$f(y_i|\alpha, \lambda) = \alpha y_i^{\alpha-1} \exp\{\lambda - \exp(\lambda) y_i^{\alpha}\}.$$

Assume censoring information is available, where $\nu_i = 0(1)$ if subject i is alive (dead), and denote $d = \sum_{i=1}^{n} \nu_i$. We can then write the likelihood as

$$\begin{aligned} p(y|\alpha, \lambda) &= \prod_{i=1}^{n} p(y_i|\alpha, \lambda)^{\nu_i} S(y_i|\alpha, \lambda)^{1-\nu_i} \\ &= \alpha^d \exp\left\{ d\lambda + \sum_{i=1}^{n} (\nu_i(\alpha - 1) \log y_i - \exp(\lambda) y_i^{\alpha}) \right\}, \end{aligned} \tag{6.4}$$

where $S(y_i|\alpha, \lambda)$ is the survival function

$$\begin{aligned} \Pr(Y \geq y|\alpha, \lambda) &= S(y|\alpha, \lambda) \\ &= \exp(-\exp(\lambda) y^{\alpha}). \end{aligned} \tag{6.5}$$

A regression model is constructed by parameterizing λ via the predictor variables: $\lambda_i = \mathbf{x}_i' \boldsymbol{\beta}$, where the first value in $\mathbf{x}_i$ is a constant, 1, and hence the intercept parameter β_0 is taken to be the first element of $\boldsymbol{\beta}$.

6.3.1 Prior Distributions

As for binary regression models, the parameter space prior distribution for the regression coefficients is taken to be

$$[\beta_0, \boldsymbol{\beta}_\gamma | \boldsymbol{\gamma}] \sim \mathrm{N}(0, \tau I_{k+1}),$$

where k is the number of variables in model $\boldsymbol{\gamma}$. I take $\tau = 1$ throughout. The prior distribution for α is taken, independently of $\boldsymbol{\beta}$, to be

$$\alpha \sim \text{Gamma}(\alpha_0, \kappa_0).$$

I set $(\alpha_0, \kappa_0) = (1, 0.8)$ throughout, which is reasonably vague and has prior location around 1.25 (Pittman *et al.*, 2004).

The prior distribution for the model space is the same as for binary regression (see Section 6.1.1).

6.3.2 Marginal Likelihood

The marginal likelihood $p(y|\boldsymbol{\gamma})$ is not available in closed form and must be estimated, as is the case for binary regression. A Laplace approximation is again used, where the posterior mode and corresponding Hessian matrix are computed via Newton-Raphson as described in Appendix D, leading to the estimate

$$\begin{aligned}\hat{p}(y|\boldsymbol{\gamma}) &= (2\pi)^{1/2}|\hat{\Sigma}|^{1/2}\tau^{-k/2}\frac{\kappa_0^{\alpha_0}}{\Gamma(\alpha_0)}\hat{\alpha}^{\alpha_0+d-1} \times \\ &\qquad \exp\left\{\sum_i(\nu_i\mathbf{x}_i'\hat{\boldsymbol{\beta}} + \nu_i(\hat{\alpha}-1)\log y_i - y_i^{\hat{\alpha}}\hat{\lambda}_i) - \kappa_0\hat{\alpha} - \frac{1}{2\tau}\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}\right\}.\end{aligned}$$

6.3.3 Posterior Summarization

Summarization of the posterior distribution is done in the same manner as for binary regression models (see Section 6.1.3) where a normal approximation to the posterior distribution $p(\alpha, \boldsymbol{\beta}|y, \boldsymbol{\gamma})$ is constructed using the posterior mode and estimated covariance from Newton-Raphson. After sampling F models based on $\tilde{p}(\boldsymbol{\gamma}|y)$, a draw is made from the approximation to $p(\alpha, \boldsymbol{\beta}|y, \boldsymbol{\gamma})$. Associated with

each draw $\boldsymbol{\beta}^{(f)}$ is a linear predictor for each individual, $\lambda_i^{(f)} = \mathbf{x}_i' \boldsymbol{\beta}^{(f)}$ where $\mathbf{x}_i$ corresponds to the model $\boldsymbol{\gamma}^{(f)}$. A key quantity of interest for each individual is the survival distribution (6.5), draws from which can be obtained using the draws from $p(\alpha, \boldsymbol{\beta}|y)$ obtained by sampling over the model space above. Monte Carlo integration for each individual can be performed on the draws

$$S_i^{(f)}(t) = \exp(-\exp(\lambda_i^{(f)})\, t^{\alpha^{(f)}}),$$

where t is a particular survival time of interest.

Leave-one-out cross-validated predictive distributions can be constructed in a manner similar to that for binary regression. After recomputing the marginal likelihoods holding out each observation, the predictive survival distributions can be sampled from, just as the fitted survival distributions were sampled from above.

6.4 Example: Lung Cancer Survival

The data are from a study at the Duke Cancer Center. Patients in the study consist of a mix of gender, age, and race. All subjects have been diagnosed with carcinoma of the lung, and their resulting tumors are a mix of adenocarinoma and squamous cell carcinoma. Of interest is survival time for the patients, of which we have $d = 45$ observed times out of $n = 91$ patients; 46 patients have censored outcomes. Expression data is available on 54,613 probe sets from Affymetrix HU95aV2 (50%) and HU95b (50%) chips, with values estimated via the RMA method as in the example in the previous section. The 14,592 probe sets with the most non-trivial expression value distributions across samples were used in an initial analysis, and the resulting 2,717 probe sets most associated with survival time were then used as follows.

Table 6.4: Posterior probability of model size conditioned on the top 100,000 models evaluated by SSS for the lung cancer survival example. The character * represents a value < 0.001.

k:	1	2	3	4	5
# of models	12	1,295	24,560	69,004	5,129
$\tilde{p}(\lvert\gamma\rvert = k \lvert y)$	*	0.014	0.254	0.690	0.041

6.4.1 Results

SSS was run for 10,000 iterations saving the top 100,000 models, with the prior distribution on the model space as $\pi = 100/p$, where $p = 2,717$, to avoid overpenalization and allow for potentially richer models. The models in Γ^* are mostly three and four variable models, as seen in Table 6.4. The null model has a (relative) log posterior value, $\log p(y|\boldsymbol{\gamma}) + \log p(\boldsymbol{\gamma})$, of -249.777, while the worst model in Γ^* has -248.802.

Using the models in Γ^*, fitted values for the survival function were estimated for survival times of 12, 18, 24 and 36 months. These results can be seen in Figures 6.7 and 6.8. Red points, those individuals who actually died before the time indicated, should ideally fall below the line at 0.5, while blue points, those individuals who have survived at least until the time indicated, should ideally fall above. One would expect the intervals to be larger nearer to the indicated survival time and smaller farther away, which can be seen for those individuals with large survival times. Overall, the fitted values coincide well with the observed times. The marginal posterior distribution for the Weibull shape parameter, α, is shown in Figure 6.6. Approximately 73% of the posterior mass is above one, the value of α corresponding to an exponential distribution.

Leave-one-out cross-validated predictions of the survival function were also

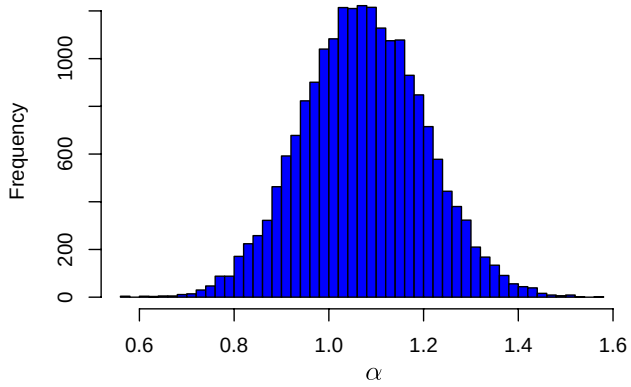

Figure 6.6: Samples from the model averaged approximation to $p(\alpha|y)$.

made for the same time points. Figures 6.9 and 6.10 show these predictions, which were based on the top 1,000 models for each hold-out respondent. The predictive accuracy is quite high: a simple thresholding of estimated survival probability (posterior mean) at 0.5 corresponds to the sensitivities and specificities shown in Table 6.5 for times of 12, 18, 24 and 36 months, and in Figure 6.11 over a range of times from 12 to 60 months. The results indicate that long term survival survival (over a range of about three to five years) can be predicted with sensitivity and specificity greater than 80%.

Figure 6.11 is based on a thresholding rule for the posterior mean survival probability at 0.5. As seen in Figure 6.9, several of the 90% intervals cover the line 0.5 and so use of the posterior mean may overstate the predictive accuracy. Rather than using the posterior mean, Figure 6.12 shows results for two different rules

Table 6.5: Sensitivity and specificity for predictions for the lung cancer data based on simple thresholding of the posterior mean at 0.5. Survival times are given in months.

Survival Time:	12	18	24	36
Sensitivity	1.000	0.926	0.885	0.975
Specificity	0.500	0.667	0.760	0.784

representing more conservative thresholding. Individuals having the 60th (70th) quantile of their survival distribution less than 0.5 are predicted as non-survivors, and those having the 40th (30th) quantile of their survival distribution greater than 0.5 are predicted as survivors. Such prediction rules lead to some individuals not being predicted as either survivors or non-survivors; these are "borderline" cases that are hard to predict, and perhaps require more investigation. Figure 6.12 also shows the number of unclassified cases, broken down by true survivors and nonsurvivors. The highest number of unclassified cases across all the survival times considered was 24. Even under these more conservative thresholding rules, where the sensitivity and specificity will be lowered due to the fact that some cases are not classified, we still see a high level of predictive accuracy with respect to long term survival.

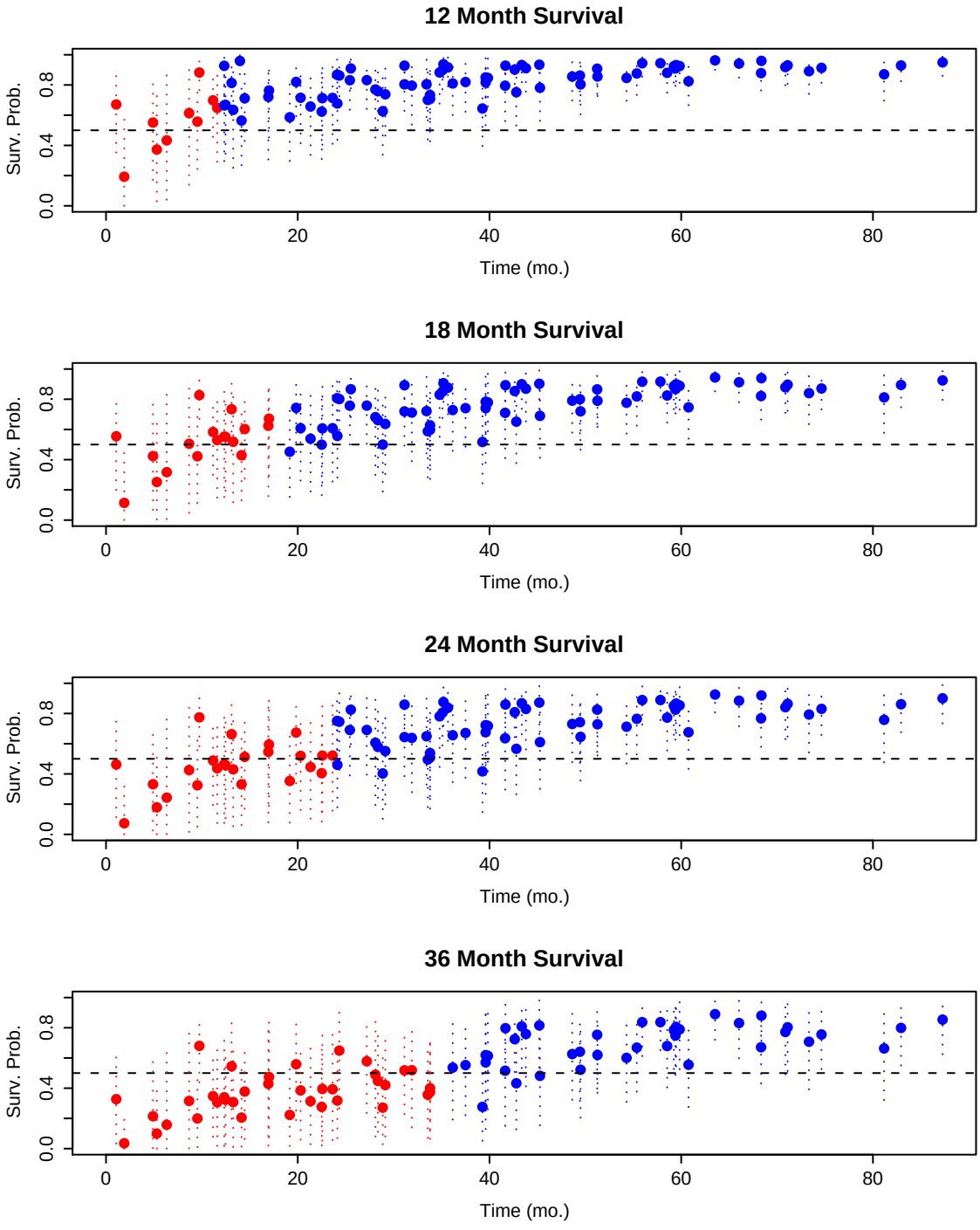

Figure 6.7: Fitted survival probabilities for the lung cancer example. Red points denote true negatives and blue points denote true positives. 90% intervals are given.

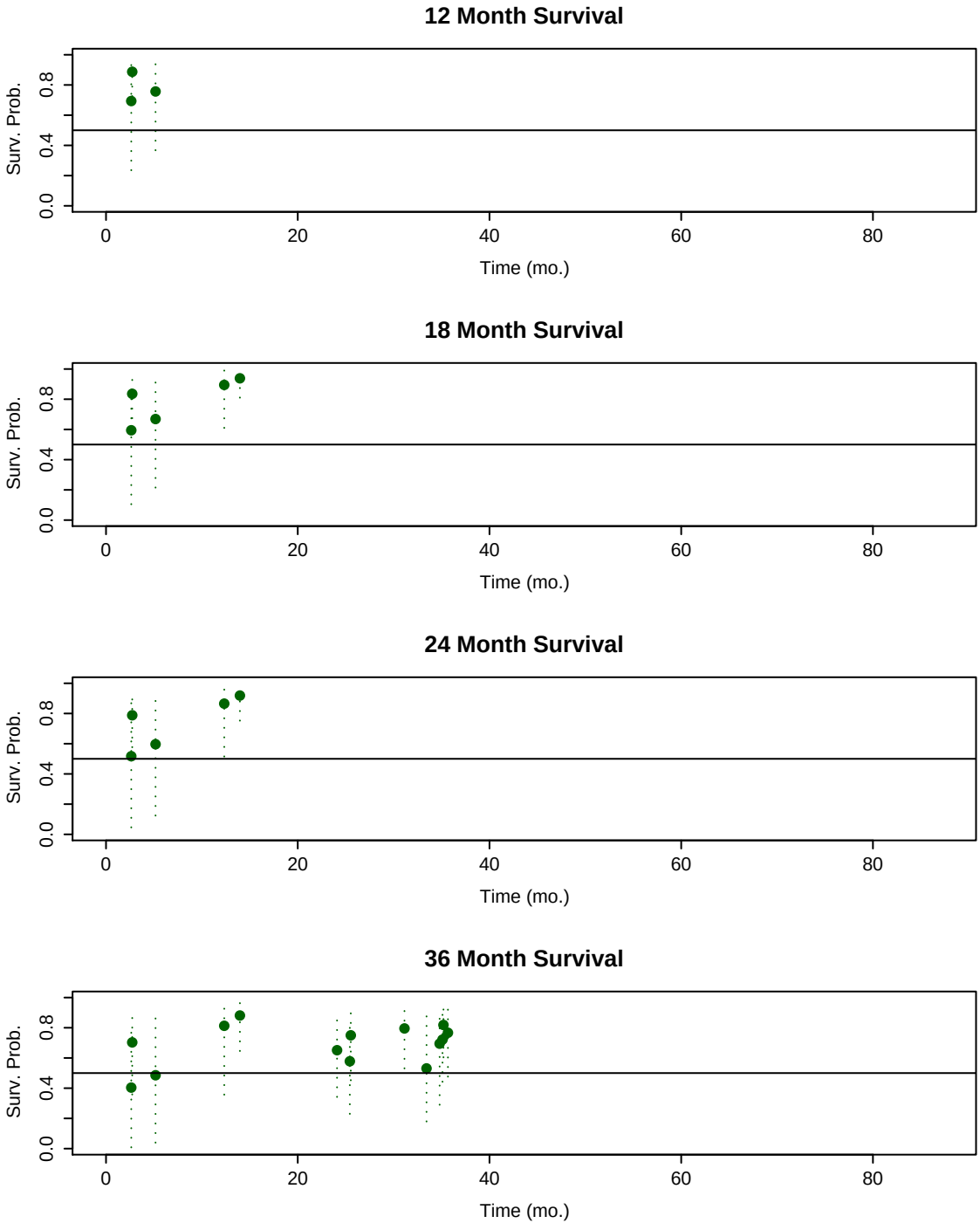

Figure 6.8: Fitted probabilities for the lung cancer example for subjects with censoring times less than the survival times of interest. 90% intervals are given.

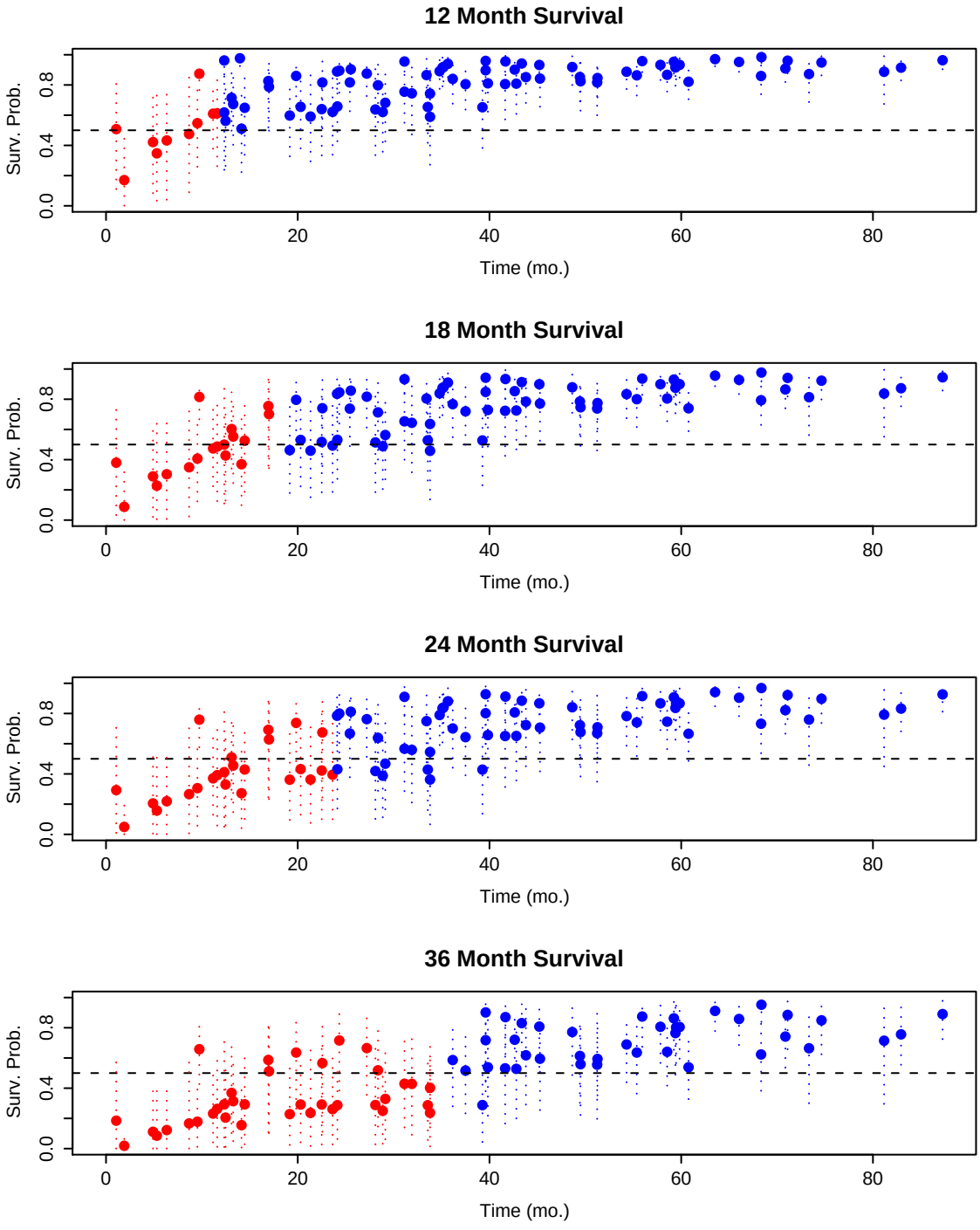

Figure 6.9: LOOCV predictions for the lung cancer example. Red points denote true negatives and blue points denote true positives. 90% intervals are given.

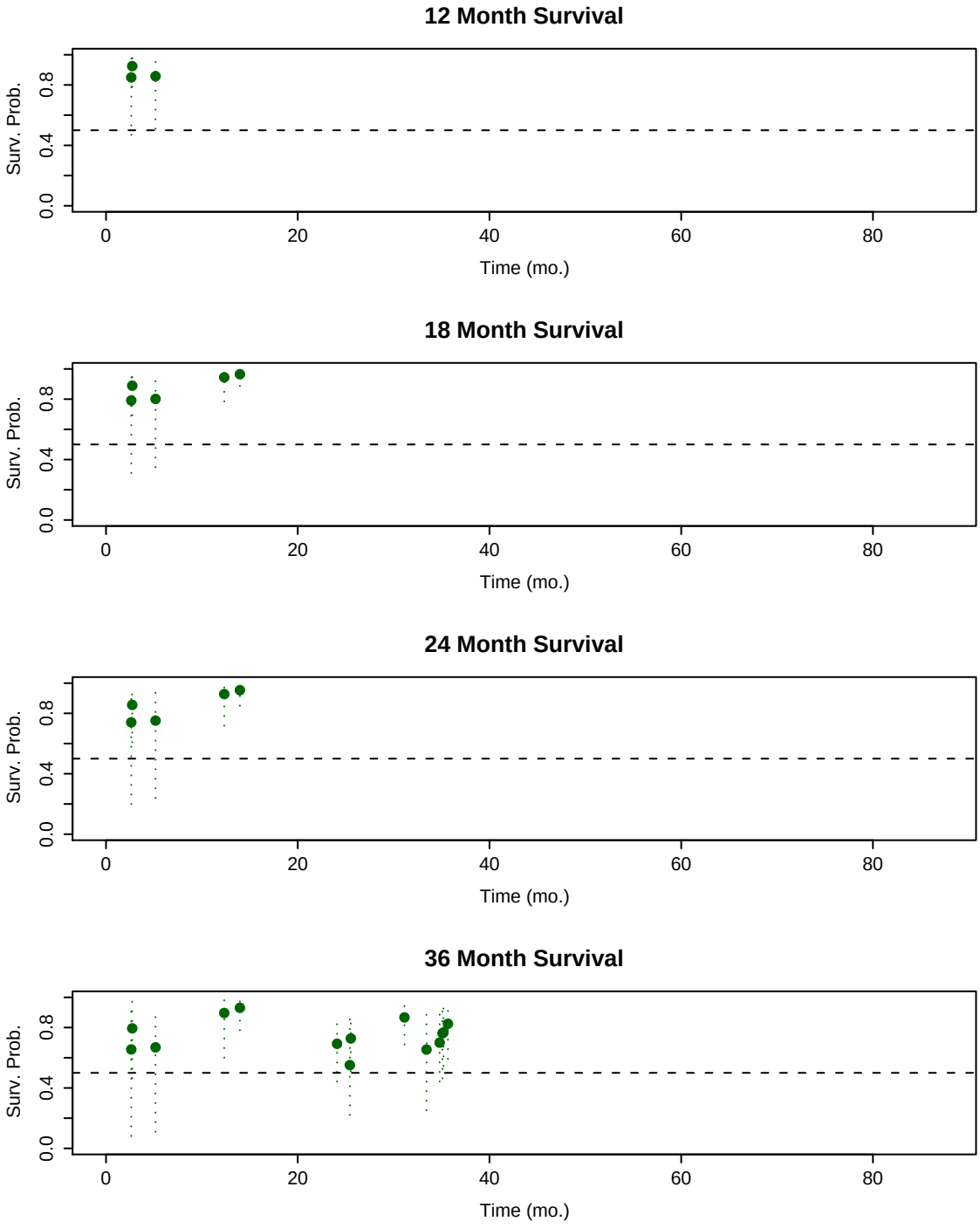

Figure 6.10: LOOCV predictions for the lung cancer example for subjects with censoring times less than the survival times of interest. 90% intervals are given.

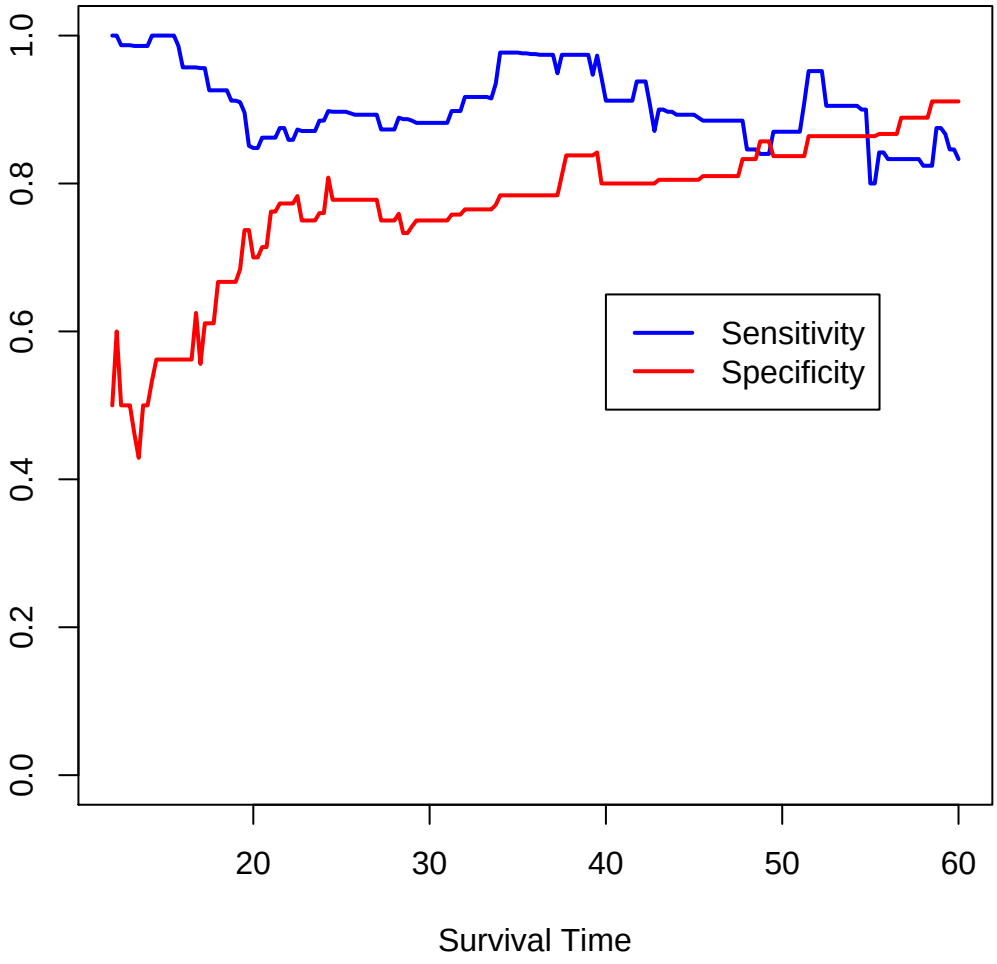

Figure 6.11: Sensitivity and specificity of survival time predictions for the lung cancer data based on simple thresholding of the posterior mean at 0.5.

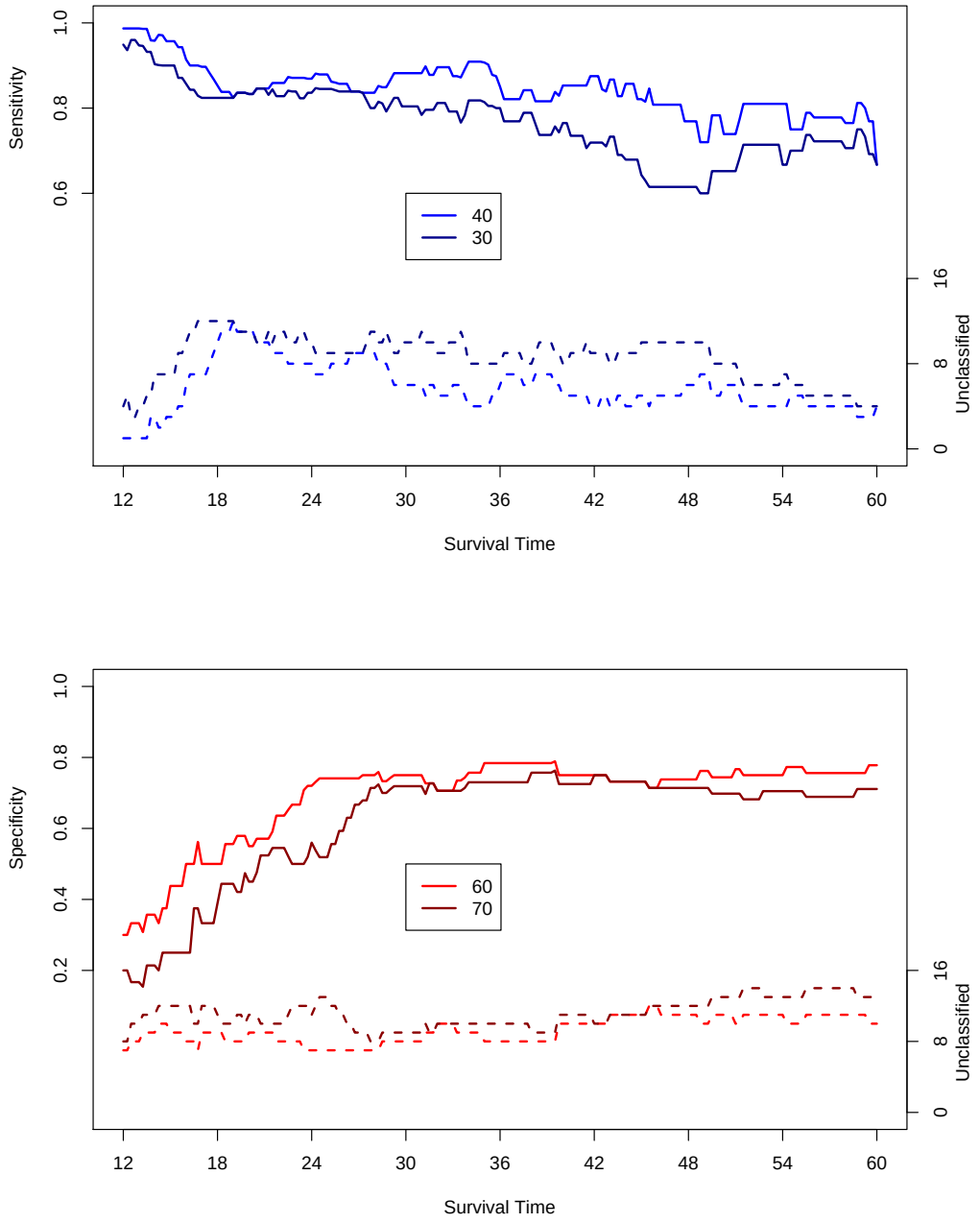

Figure 6.12: Sensitivity and specificity of survival time predictions for the lung cancer data based on two different thresholding rules. The dashed lines are the number of cases not classified.

Chapter 7

Conclusions and Future Work

7.1 Summary

This work addressed several key issues related to Bayesian model search and averaging in high dimensions. I presented the "shotgun stochastic search", a novel approach for high dimensional model space exploration that quickly identifies high posterior regions and creates a list of high probability models which can be used for inference and prediction. It was shown through a simulation study that SSS outperforms standard MCMC methods with respect to several criteria. Theoretical connections were made between SSS and the Metropolis-Hastings algorithm, demonstrating how SSS can be adapted to become an MCMC algorithm.

I presented results relating to the marginal likelihood for models of the same dimension, leading to methods for approximate Bayesian learning about sparsity inducing parameters. Comparisons between model selection criteria were made based on these result.

Three examples of datasets with thousands of predictor variables were given, demonstrating the effectiveness of SSS and Bayesian model averaging in high di-

mensions. Three separate modeling contexts, those of linear, binary and Weibull survival regression models, were introduced for use with SSS, along with methodology for assessing model uncertainty in inference and prediction in high dimension.

I note that although I have focused throughout on the particular parameter space prior distributions described in Chapter 3, there are many other formulations that do not rely on this ridge regression type formulation. One popular example is Zellner's g-prior (Zellner, 1986), where the prior distribution on the regression coefficients for a given model is

$$p(\boldsymbol{\beta}_\gamma|\boldsymbol{\gamma}, \sigma^2, g) = \mathrm{N}(0, g\sigma^2(X'_\gamma X_\gamma)^{-1}), \tag{7.1}$$

where g is a positive hyperparameter. This class of prior distributions has particular computational advantages (George and McCulloch, 1997; Smith and Kohn, 1996; Clyde and George, 2004), but has undesirable asymptotic properties related to the Bayes factor for comparing any model $\boldsymbol{\gamma}$ to the null model (Berger and Pericchi, 2001). In the special case of an orthogonal design, the priors I describe in Chapter 3 are themselves g-priors and so results relating to g-priors can be applied to the setting considered here. West (2003) shows that the g-prior is a special, limiting case of of factor regression with a formal latent factor model on the predictor variables, hence providing a Bayesian justification for this particular design-dependent approach.

Comparisons to other parameter space prior distributions are also of interest, especially to those intended to induce sparsity. One such class of priors is the class of double exponential priors (e.g. Johnstone and Silverman, 2005; Yuan and Lin, 2005), which places more mass near zero and has fatter tails than Gaussian priors. For discussions of sparsity induced by parameter space priors, see also Wolfe *et al.* (2004) and Tipping (2001).

7.2 Connections to Other Modeling Frameworks

Regression modeling can be thought of as a special case of the broader class of Gaussian graphical models. A Gaussian graphical model allows for inference on conditional independence relationships for elements of a random vector from a multivariate normal distribution. Let $\mathbf{x}$ be a p vector from a $\mathrm{N}(0, \Omega^{-1})$ distribution. Then the (full) conditional distributions are of the form

$$p(x_j|\mathbf{x}_{-j}) = \mathrm{N}(-\Omega'_{-j}\omega_{jj}^{-1} \ , \ \omega_{jj}^{-1})$$

where Ω_{-j} is a $p-1$ vector representing the jth column of Ω having removed the element on the diagonal, ω_{jj} (c.f. with the results in Appendix A, where different notation is used). It is clear from this formulation that x_j is conditionally independent of x_i $(i \neq j)$ when the element of Ω_{-j} corresponding to the ith row of Ω is zero. Gaussian graphical models and corresponding methodology (Dempster, 1972; Whittaker, 1990; Dawid and Lauritzen, 1993; Lauritzen, 1996; Giudici, 1996; Giudici and Green, 1999; Cowell *et al.*, 1999) allow for structural learning about such zero values, and hence allow for learning about conditional independence structure. Setting $(y, \mathbf{x}')' \sim \mathrm{N}(0, \Omega^{-1})$, it is clear the question of variable selection for linear regression models is related to learning about conditional independence relationships for the distribution $p(y|\mathbf{x})$.

Shotgun stochastic search methods have been developed for exploring Gaussian graphical model spaces (Jones *et al.*, 2005), and constructive methods proposed for building graphs by sequentially learning about conditional independence relationships via structured regression modeling (Dobra *et al.*, 2004). Of future research interest is the development of Gaussian graphical model techniques that simultaneously fit regressions of y on subsets of $\mathbf{x}$ and model the covariance struc-

ture of $\mathbf{x}$. Initial experiences in the development of Rich *et al.* (2005) have shown that simply including y in a Gaussian graphical model along with $\mathbf{x}$ can be problematic when p is large. In this case the resulting graph space is so large that the (needed) sparsity inducing priors tend to supress edges (first order dependence) between y and possible predictor variables. Additionally, using current SSS methods, there is no special focus on (i.e., exploration of) graphs with edges involving y; as the graph space is very large, important regions of the regression model space of interest may not be visited.

Two immediate areas of future work emerge from this. New prior distributions are needed that treat graphical structure between y and the elements of $\mathbf{x}$ differently than graphical structure within $\mathbf{x}$ in order to prevent sparsity inducing priors over the graph space based on $\mathbf{x}$ to overwhelm important graphical structure between y and the elements of $\mathbf{x}$. Secondly, hybrid SSS methodologies can be developed that model graphical structure within $\mathbf{x}$ and graphical structure between y and $\mathbf{x}$ in an iterative manner: runs of SSS can be made on $\mathbf{x}$ *conditionally* on structure between y and $\mathbf{x}$, and then runs can be made on the regression edges of interest *conditionally* on a structure for $\mathbf{x}$. Such methods force focus on the regression models of interest and should yield more relevant results.

7.3 Future Work in Large p Regression

7.3.1 Model Space Priors

There is still much work to be done in the area of Bayesian model search and uncertainty with many predictors. Of primary interest is the development of prior distributions over model space that do not treat variables exchangeably. It

has been noted (George, 1999; Chipman *et al.*, 2001) that when there is nontrivial multicollinearity in the predictor variables, independent Bernoulli priors place too much prior mass on clusters of similar models. Some work has been done relating to prior distributions that can handle "dependence by design" situations such as for interaction terms and polynomials (Chipman, 1996), however emphasis needs to be placed on situtations where p is large as the patterns of collinearity can be become more complicated in large spaces.

A recently proposed prior distribution over model space that conditions on the observed covariates is due to Yuan and Lin (2005), who take

$$p(\boldsymbol{\gamma}|X) \propto \pi^k (1-\pi)^{p-k} |X_\gamma' X_\gamma|^{1/2}. \tag{7.2}$$

As the authors note, as any two variables become increasingly collinear, (7.2) converges to a prior placing zero mass on models containing both variables. It is of interest to examine how this prior penalizes dimension with respect to different patterns of multicollinearity in the observed data matrix X.

This prior, while having the desirable property regarding multicollinearity described above, is *ad hoc* and dependent on the particular value of X that is observed, rather than on parameters describing the distribution of X. Ideally one would formulate a model, say $p(\mathbf{x}_i|\Sigma) = \mathrm{N}(0, \Sigma)$ independently for each $\mathbf{x}_i$, and then specify a prior $p(\boldsymbol{\gamma}|\Sigma)$ rather than $p(\boldsymbol{\gamma}|X)$. Of key interest in future work regarding such priors is discovering links between (7.2) and $p(\mathbf{x}|\Sigma)$, perhaps directly via marginalization: $p(\boldsymbol{\gamma}|X) = \int p(\boldsymbol{\gamma}|\Sigma) p(\Sigma|\mathbf{x})\, d\Sigma$, assuming the conditional independence of $\boldsymbol{\gamma}$ and X given Σ. Such links, similar to those between the g-prior and factor regression described by West (2003), will give (7.2) a theoretically sound footing and obviate the problems associated with specifying a prior conditionally on the observed data.

One possible approach in abstracting (7.2) to a more general class of priors is to view the problem from an information theoretic perspective. Assuming that $\mathbf{x} \sim \mathrm{N}(0, \Sigma)$ independently for each individual, define the *entropy* of a model $\boldsymbol{\gamma}$ to be

$$\begin{aligned} H(\mathbf{x}_\gamma) &= -\int p(\mathbf{x}_\gamma) \log p(\mathbf{x}_\gamma) d\mathbf{x}_\gamma \\ &= \log\left(|\Sigma|^{1/2}(2\pi e)^{k/2}\right) \end{aligned} \tag{7.3}$$

(c.f. Shannon, 1948; Lindley, 1956; Kullback, 1959; Bernardo and Smith, 1994, pp. 157–160). Attempts have been made to cast (7.3) as an absolute measure of information in $\mathbf{x}$ about $p(\cdot)$, although this is not generally accepted as meaningful, in part due to the lack of invariance of (7.3) under a transformation of $\mathbf{x}$. Viewing (7.3) simply as an integrated score function, a possible prior formulation sets

$$\begin{aligned} p(\boldsymbol{\gamma}|\Sigma) &\propto f(k) \exp(H(\mathbf{x}_\gamma)) \\ &= f(k)(2\pi e)^{k/2}|\Sigma_\gamma|^{1/2}, \end{aligned} \tag{7.4}$$

where $f(k)$ is perhaps some term to penalize dimension. While this formulation is perhaps equally as *ad hoc*, (7.4) gives a possible interpretation to (7.2) when $X'_\gamma X_\gamma$ is used as an estimate of Σ_γ.

An alternate approach, more closely related to (7.2), is to consider the *directed divergence*

$$I(2:1) = -\int p_2(\mathbf{x}) \log \frac{p_1(\mathbf{x})}{p_2(\mathbf{x})} d\mathbf{x},$$

where p_1 and p_2 are two possible density functions. $I(2:1)$ has the interpretation as "the mean information per observation from p_2 for discrimination in favor of H_1 against H_2" (Kullback, 1959, Chapter 1), where H_i is the hypothesis that the

data are distributed according to p_i. If we take $p_1(\mathbf{x}_\gamma)$ to be the model where the x_j are jointly independent, $\mathbf{x}_\gamma \sim \mathrm{N}(0, I_k)$, and take $p_2(\mathbf{x}_\gamma)$ to be the model where the x_j are possibly correlated, $\mathbf{x}_\gamma \sim \mathrm{N}(0, \Sigma_\gamma)$, with each x_j having variance one ($\Sigma_{jj} = 1$), then the directed divergence for a given model is

$$I(\gamma) = -\log\left(|\Sigma_\gamma|^{1/2}\right).$$

Setting

$$\begin{aligned} p(\boldsymbol{\gamma}|\Sigma) &\propto f(k)\exp(-I(\gamma)) \\ &= f(k)|\Sigma_\gamma|^{1/2}, \end{aligned}$$

we have a "population" version of (7.2) when $f(k) = \pi^k(1-\pi)^{p-k}$. The null model has no particular interpretation under this construction; setting $|\Sigma| = 1$ for the null model corresponds to the approach taken by Yuan and Lin (2005). A more flexible family of prior distributions is obtained by taking

$$p(\boldsymbol{\gamma}|\Sigma) \propto f(k)|\Sigma_\gamma|^{\beta},$$

where β controls how tolerable we are to multicollinear models. Other possible formulations related to the *divergence*, $J(1,2) = I(1:2) + I(2:1)$, are also available. Careful investigation must be made to understand how such priors penalize dimension for large "p" scenarios with complex covariance patterns.

These information theoretic views provide a generalization of (7.2), allowing extension to other classes of models; however, they do not explicitly link (7.2) to a proper modeling framework for the observed values X, as the substitution of $X_\gamma' X_\gamma$ for Σ_γ must be made. Casting (7.2) as the consequence of an overarching model is of key interest for future work.

7.3.2 Further Comparisons of Model Size Penalization

The results of Section 5 can be extended to other probability models. For example, under the g-prior formulation described in Berger and Pericchi (2001) using prior (7.1) and $p(\sigma^2) \propto \sigma^{-2}$, the marginal likelihood is

$$p_g(y|\boldsymbol{\gamma}) = \frac{\Gamma\left(\frac{n}{2}\right)}{2\pi^{n/2}(1+g)^{k/2}} \left(y'y - \frac{g}{(1+g)} y'X(X'X)^{-1}X'y \right)^{-n/2},$$

which has a corresponding lower bound

$$p_g^*(y|\boldsymbol{\gamma}) = \frac{\Gamma\left(\frac{n}{2}\right)}{2\pi^{n/2}(1+g)^{k/2}(n-1)^{n/2}}$$

when the data are standardized. Comparisons to the results in Chapter 5 are relevant, along with comparisons to other commonly used parameter space prior distributions, to assess the implicit penalties on dimension.

7.3.3 Extended Analysis of SSS

The results in Section 2.5 regarding the expected time for SSS to find the true model under an orthogonal design were derived for a case of fixed-dimensional SSS. These results can be extended to the case where SSS is allowed to move freely across dimension. It is of interest to compare these expected hitting times to those for MC^3 under an orthogonal design, which can be similarly computed. A direct analytic comparison may provide more insight into how the nature of the neighborhood used by SSS provides improved performance over MCMC algorithms with simpler proposal distributions.

Appendix A

Wishart Distribution

I use notation for the Wishart/inverse Wishart distributions as in Dawid (1981).

Wishart Distribution: Denote by $\mathrm{W}(\nu, \Sigma)$ the Wishart distribution with ν degrees of freedom and non-negative definite symmetric scale matrix Σ of order $p \times p$. Define the *standard Wishart* as $\mathrm{W}(\nu, I_p)$.

Inverse Wishart Distribution: If $\Omega \sim \mathrm{W}(\nu, \tau^{-1} I_p)$ with $\nu > p - 1$, then $\Sigma = \Omega^{-1} \sim \mathrm{IW}(\delta, \tau I_p)$, where $\delta = \nu - p + 1$. The constraint on the degrees of freedom parameter is $\delta > 0$.

A.1 Implied Regression Priors

Let $\Sigma \sim \mathrm{IW}(\delta, \tau I)$, with δ degrees of freedom and scale matrix τI_{p+1}, where Σ is the covariance matrix for a vector of normally distributed, zero mean data: $(y, x')' \sim \mathrm{N}(0, \Sigma)$, where y is a scalar. Let x_γ be some subset of the vector x with k components. The covariance matrix for $(y, x_\gamma')'$ has marginal distribution

$\Sigma_{y,\gamma} \sim \mathrm{IW}(\delta, \tau I_{k+1})$. Partition $\Sigma_{y,\gamma}$ as

$$\Sigma_{y,\gamma} = \begin{array}{c} 1 \\ k \end{array} \begin{bmatrix} \sigma_{11} & \kappa_\gamma' \\ \kappa_\gamma & \Sigma_\gamma \end{bmatrix}.$$

Standard normal theory gives us the conditional distribution of y given x_γ as

$$p(y|x_\gamma, \Sigma) = \mathrm{N}(y;\, x_\gamma'\boldsymbol{\beta}_\gamma\ ,\ \sigma_\gamma^2),$$

where $\boldsymbol{\beta}_\gamma = \Sigma_\gamma^{-1}\kappa_\gamma$ and $\sigma_\gamma^2 = \sigma_{11} - \kappa_\gamma'\Sigma_\gamma^{-1}\kappa_\gamma$. For brevity, write β for β_γ and σ^2 for σ_γ^2, and note that σ^2 is the Schur complement of Σ_γ.

Let $\Omega = \Sigma_{y,\gamma}^{-1}$. Then $\Omega \sim \mathrm{W}(\delta + k, \tau^{-1} I_{k+1})$. Partition

$$\Omega = \begin{array}{c} 1 \\ k \end{array} \begin{bmatrix} \omega_{11} & \Omega_{21}' \\ \Omega_{21} & \Omega_{22} \end{bmatrix}. \tag{A.1}$$

Then $\omega_{11} \sim \mathrm{W}(\delta + k, \tau^{-1} I_1)$, and $\omega_{11}^{-1} \sim \mathrm{IW}(\delta + k, \tau I_1)$.

The following results are needed (from Harville, 1997):

Corollary A.1 (Harville (1997), 8.5.12). *Let Ω be a non-singular matrix partitioned as in (A.1), and define $\Sigma = \Omega^{-1}$. Partition Σ as*

$$\Sigma = \begin{array}{c} 1 \\ k \end{array} \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}.$$

Then

$$\omega_{11}^{-1} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \tag{A.2}$$

(i.e., ω_{11}^{-1} equals the Schur complement of Σ_{22}) and

$$\Omega_{21}\omega_{11}^{-1} = -\Sigma_{22}^{-1}\Sigma_{21}. \tag{A.3}$$

Using (A.2), $\omega_{11}^{-1} = \sigma^2$, and hence $\sigma^2 \sim \text{IW}(\delta + k, \tau I_1)$, which is the same as $\sigma^2 \sim \text{IG}((\delta + k)/2, \tau/2)$.

Using (A.3), we have $\beta = -\Omega_{21}/\omega_{11}$. We need the conditional distribution of β given σ^2, which is equivalent to the conditonal distribution $-\Omega_{21}/\omega_{11}$ given ω_{11}^{-1}, as $\sigma^2 = \omega_{11}^{-1}$. The following result is needed (from Muirhead (1982, 3.2.14), due to Bartlett (1933)):

Theorem A.1 (Bartlett's Decomposition). *Let A be $\text{W}(n, I_m)$, where $n \geq m$ is an integer, and put $A = TT'$, where T is a lower-triangular $m \times m$ matrix with positive diagonal elements. Then the elements t_{ij} $(1 \leq j \leq i \leq m)$ of T are all independent, t_{ii}^2 is χ_{n-i+1}^2 $(i = 1, \ldots, m)$, and t_{ij} is $\text{N}(0, 1)$ $(1 \leq j < i \leq m)$.*

Following Odell and Feiveson (1966) and McCulloch and Rossi (1994), and using the result that if $A \sim \text{W}(\nu, C_{p\times p})$ and B is a $p \times q$ matrix, then $B'AB \sim \text{W}(\nu, B'CB)$ (Mardia *et al.*, 1979, Theorem 3.4.1), we can decompose

$$\Omega = \tau^{-1} TT',$$

where TT' is the Bartlett decomposition for a standard Wishart with $\delta + k$ degrees of freedom. The ith element of Ω_{21} can be written as

$$(\Omega_{21})_i = \tau^{-1}\, t_{11}\, t_{i+1,1}, \quad i = 1, \ldots, k$$

Due to the equivalence of ω_{11} and t_{11}^2 and the joint independence of the t_{ij} for $i \geq j$, conditioning on ω_{11} implies that

$$(\Omega_{21})_i | \omega_{11} \overset{\text{ind.}}{\sim} \text{N}(0, \tau^{-1}\omega_{11}), \quad i = 1, \ldots, k.$$

Because $\sigma^2 = \omega_{11}^{-1}$,

$$\beta = (-\Omega_{21}/\omega_{11}) | \sigma^2 \sim \text{N}(0, \tau^{-1}\sigma^2 I_k).$$

Appendix B

Linear Algebra Results

Useful linear algebra results follow. Theorems and corollaries cited below as "H" are from Harville (1997, Chapters 13 and 14).

B.1 Results for Lower Bound Theorem

The first needed result is that, under the conditions of Theorem 5.1, $XM^{-1}X'$ is a positive semidefinite matrix. From Corollary H.14.2.14 we have that $X'X$ is positive definite because $\text{rank}(X) = k$. Due to the positive definiteness of τI_k, Corollary H.14.2.5 gives us positive definitiness of M, and we have positive definiteness of M^{-1} from Corollary H.14.2.11. Theorem H.14.2.9 gives us the desired result, that $XM^{-1}X'$ is positive semidefinite due to the fact that $\text{rank}(X) = k < n$.

The second result needed is that $|M| \leq \prod_{i=1}^{k} m_{ii}$. First, as the diagonal elements of a positive definite matrix are positive (Corollary H.14.2.13), we have $m_{ii} > 0$ for $k = 1, \ldots, k$. Second, I use the result that

$$\begin{vmatrix} \mathbf{T} & \mathbf{U} \\ \mathbf{V} & \mathbf{W} \end{vmatrix} = |\mathbf{T}||\mathbf{W} - \mathbf{V}\mathbf{T}^{-1}\mathbf{U}|,$$

for a nonsingular, $m \times m$ matrix $\mathbf{T}$, an $m \times n$ matrix $\mathbf{U}$, an $n \times m$ matrix $\mathbf{V}$ and an $n \times n$ matrix $\mathbf{W}$ (Theorem H.13.3.8). Now partition M as

$$M = \begin{bmatrix} M_{k-1} & u_{k-1} \\ u'_{k-1} & m_{kk} \end{bmatrix},$$

where M_{k-1} is the principal submatrix obtained by removing the kth row and column of M (corresponding to $\mathbf{T}$) and u_{k-1} is a $(k-1) \times 1$ vector (corresponding to $\mathbf{U}$). We thus have

$$|M| = |M_{k-1}|(m_{kk} - u'_{k-1}M_{k-1}^{-1}u_{k-1}).$$

By Corollary H.14.2.12, any principal submatrix of a positive definite matrix is positive definite (and hence nonnegative definite), so $x'M_{k-1}^{-1}x \geq 0$ for every $x \in \mathbb{R}^{k-1}$. As $m_{kk} > 0$, we have

$$|M| \leq |M_{k-1}|m_{kk}.$$

Repeating the argument $k-1$ more times, we have

$$|M| \leq \prod_{i=1}^{k} m_{ii}.$$

Appendix C

Gamma Function Results

The gamma function for real arguments is defined as

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t}\, dt.$$

For integer arguments $n > 0$, the closed form expression is

$$\Gamma(n) = (n-1)!.$$

C.1 Gamma Functions of Half-Integer Arguments

For half-integer values, the gamma function has a closed form solution (Abramowitz and Stegun, 1972, chapt. 6):

$$\Gamma\left(\frac{n}{2}\right) = \frac{(n-2)!!\sqrt{\pi}}{2^{(n-1)/2}}, \tag{C.1}$$

where $n!!$ is a double factorial:

$$n!! \equiv \begin{cases} n \cdot (n-2) \cdots 5 \cdot 3 \cdot 1 & n > 0 \text{ odd}, \\ n \cdot (n-2) \cdots 6 \cdot 4 \cdot 2 & n > 0 \text{ even}, \\ 1 & n = -1, 0. \end{cases}$$

C.2 Ratios of Gamma Functions

Equation 5.3 requires the calculation of a ratio of gamma functions of half integer arguments,

$$G(n, \delta, k) = \frac{\Gamma\left(\frac{n+\delta+k}{2}\right)}{\Gamma\left(\frac{\delta+k}{2}\right)}.$$

Using (C.1), we can write this as

$$G(n, \delta, k) = 2^{-n} \cdot \frac{(n+\delta+k-2)!!}{(\delta+k-2)!!}.$$

For n even, this reduces to

$$G(n, \delta, k) = 2^{-n} \cdot \prod_{i=1}^{n/2} (n+\delta+k-2i).$$

For n odd there is no cancellation.

Appendix D

Newton-Raphson Algorithm

The Newton-Raphson algorithm (see, e.g., Lange, 1999, Chapter 11) is used to maximize an objective function, say $f(\boldsymbol{\theta})$, by iterating

$$\boldsymbol{\theta}^{[t+1]} = \boldsymbol{\theta}^{[t]} - \left(\frac{\partial^2 f(\boldsymbol{\theta}^{[t]})}{\partial \theta_i^{[t]} \partial \theta_j^{[t]}} \right)^{-1} \frac{d\, f(\boldsymbol{\theta}^{[t]})}{d\boldsymbol{\theta}^{[t]}}.$$

When the objective function is the posterior distribution of $\boldsymbol{\theta}$, we have $f(\boldsymbol{\theta}) = \log h(\boldsymbol{\theta})$ and

$$h(\boldsymbol{\theta}) = p(y|\boldsymbol{\theta})p(\boldsymbol{\theta}).$$

Define the score function and the Hessian matrix to be

$$g(\boldsymbol{\theta}) = \frac{d \log h(\boldsymbol{\theta})}{d\boldsymbol{\theta}} \quad \text{and} \quad G(\boldsymbol{\theta}) = \left(\frac{\partial^2 \log h(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right),$$

respectively, and let $H(\boldsymbol{\theta}) = -G(\boldsymbol{\theta})$. To find the posterior mode, Newton-Raphson iterates

$$\boldsymbol{\theta}^{[t+1]} = \boldsymbol{\theta}^{[t]} + H(\boldsymbol{\theta}^{[t]})^{-1} g(\boldsymbol{\theta}^{[t]})$$

in t until convergence.

D.1 Logistic Regression

Under a logistic regression model we have the likelihood function

$$p(y|\boldsymbol{\theta}) = \prod_{i=1}^{n} \phi_i^{y_i}(1-\phi_i)^{1-y_i},$$

where $\phi_i = 1/(1+\exp\{-\mathbf{x}_i'\boldsymbol{\theta}\})$. Let the prior distribution on $\boldsymbol{\theta}$ be specified as in Section 6.1.1 for a model with k predictor variables plus an intercept:

$$p(\boldsymbol{\theta}) = (2\pi\tau)^{-(k+1)/2} \exp\left\{\frac{-\boldsymbol{\theta}'\boldsymbol{\theta}}{2\tau}\right\}.$$

Using the notation above,

$$\log h(\boldsymbol{\theta}) = \text{cons.} - \frac{\boldsymbol{\theta}'\boldsymbol{\theta}}{2\tau} + \sum_{i=1}^{n} y_i \log\phi_i - (1-y_i)\log(1-\phi_i).$$

Noting that

$$\frac{d\log\phi_i}{d\boldsymbol{\theta}} = \frac{\mathbf{x}_i}{1+e^{\mathbf{x}_i'\boldsymbol{\theta}}} \quad \text{and} \quad \frac{d\log(1-\phi_i)}{d\boldsymbol{\theta}} = -\phi_i\mathbf{x}_i,$$

we have

$$g(\boldsymbol{\theta}) = -\frac{\boldsymbol{\theta}}{\tau} + \sum_{i=1}^{n}(y_i-\phi_i)\mathbf{x}_i$$

and

$$G(\boldsymbol{\theta}) = -\frac{1}{\tau}I_{k+1} - \sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'\phi_i(1-\phi_i).$$

D.2 Weibull Survival Regression

Under a Weibull regression model the likelihood function is given by (6.4). The prior distribution on the paramters $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}')'$ is

$$p(\boldsymbol{\theta}) = p(\boldsymbol{\beta})p(\alpha) \propto \exp\left\{-\frac{\boldsymbol{\beta}'\boldsymbol{\beta}}{2\tau}\right\} \alpha^{\alpha_0 - 1} \exp\{-\alpha\kappa_0\},$$

and so we have

$$h(\boldsymbol{\theta}) = (\alpha_0 + d - 1)\log\alpha + \sum_i [\nu_i \mathbf{x}_i'\boldsymbol{\beta} + \nu_i(\alpha - 1)\log y_i - y_i^\alpha \exp(\mathbf{x}_i'\boldsymbol{\beta})] - \kappa_0\alpha - \frac{1}{2\tau}\boldsymbol{\beta}'\boldsymbol{\beta}.$$

Write

$$g(\boldsymbol{\theta}) = (g_1(\boldsymbol{\theta}), g_2(\boldsymbol{\theta})')',$$

where

$$g_1(\boldsymbol{\theta}) = \frac{\partial \log h(\boldsymbol{\theta})}{\partial\alpha} \quad \text{and} \quad g_2(\boldsymbol{\theta}) = \frac{\partial \log h(\boldsymbol{\theta})}{\partial\boldsymbol{\beta}}.$$

We then have

$$\begin{aligned} g_1(\boldsymbol{\theta}) &= \frac{\alpha_0 + d - 1}{\alpha} + \sum_i [\nu_i \log y_i - y_i^\alpha (\log y_i)\exp(\mathbf{x}_i'\boldsymbol{\beta})] - \kappa_0, \\ g_2(\boldsymbol{\theta}) &= \sum_i [\nu_i \mathbf{x}_i - y_i^\alpha \exp(\mathbf{x}_i'\boldsymbol{\beta})\mathbf{x}_i] - \frac{1}{\tau}\boldsymbol{\beta}. \end{aligned}$$

Continuing with notation, let

$$G(\boldsymbol{\theta}) = \begin{pmatrix} G_{11}(\boldsymbol{\theta}) & G_{12}(\boldsymbol{\theta})' \\ G_{12}(\boldsymbol{\theta}) & G_{22}(\boldsymbol{\theta}) \end{pmatrix} = \begin{pmatrix} \frac{\partial g_1(\theta)}{\partial\alpha} & \frac{\partial g_2(\theta)'}{\partial\alpha} \\ \frac{\partial g_2(\theta)}{\partial\alpha} & \frac{\partial g_2(\theta)}{\partial\beta} \end{pmatrix}.$$

Then

$$
\begin{aligned}
G_{11}(\boldsymbol{\theta}) &= \frac{1-d-\alpha_0}{\alpha^2} - \sum_i y_i^{\alpha} (\log y_i)^2 \exp(\mathbf{x}_i'\boldsymbol{\beta}), \\
G_{12}(\boldsymbol{\theta}) &= -\sum_i y_i^{\alpha} \exp(\mathbf{x}_i'\boldsymbol{\beta}) (\log y_i) \mathbf{x}_i, \\
G_{22}(\boldsymbol{\theta}) &= -\frac{1}{\tau} - \sum_i y_i^{\alpha} \exp(\mathbf{x}_i'\boldsymbol{\beta}) \mathbf{x}_i \mathbf{x}_i'.
\end{aligned}
$$

Bibliography

Abramowitz, M. and Stegun, I. (1972). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. John Wiley & Sons, New York.

Akaike, H. (1970). Statistical predictor inference. *Annals of the Institute of Statistical Mathematics* **22**, 203–217.

Apéry, R. (1979). Irrationalité de $\zeta(2)$ et $\zeta(3)$. *Astérisque* **61**, 11–13.

Bartlett, M. S. (1933). On the theory of statistical regression. *Proceedings of the Royal Society of Edinburgh* **53**, 260–283.

Bellahcene, A. and Castronovo, V. (1995). Increased expression of Osteonectin and Osteopontin, two bone matrix proteins, in human breast cancer. *American Journal of Pathology* **146**, 95–100.

Berger, J. O. and Pericchi, L. R. (2001). Objective Bayesian methods for model selection: Introduction and comparison (with discussion). In P. Lahiri, ed., *Model Selection*, 135–207. IMS, Beachwood, OH.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. John Wiley & Sons, New York; Chichester.

Brown, P. J., Vannucci, M., and Fearn, T. (1998a). Bayesian wavelength selection in multicomponent analysis. *Journal of Chemometrics* **12**, 173–182.

Brown, P. J., Vannucci, M., and Fearn, T. (1998b). Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society,* Series B **60**, 672–641.

Chipman, H. A. (1996). Bayesian variable selection with related predictors. *Canadian Journal of Statistics* **24**, 17–36.

Chipman, H. A., George, E. I., and McCulloch, R. E. (2001). The practical implementation of Bayesian model selection (with discussion). In P. Lahiri, ed., *Model Selection*, 65–134. IMS, Beachwood, OH.

Clyde, M. and George, E. I. (2004). Model uncertainty. *Statistical Science* **19**, 81–94.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L., and Spiegelhalter, D. J. (1999). *Probabilistic Networks and Expert Systems*. Springer–Verlag, New York.

Cui, W. and George, E. I. (2004). Empirical Bayes vs. fully Bayes variable selection. Tech. rep., The Wharton School, University of Pennsylvania.

Dawid, A. P. (1981). Some matrix-variate distribution theory: Notational considerations and a Bayesian application. *Biometrika* **68**, 265–274.

Dawid, A. P. and Lauritzen, S. L. (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models. *The Annals of Statistics* **3**, 1272–1317.

Dempster, A. P. (1972). Covariance selection. *Biometrics* **28**, 157–175.

Denison, D. G. T., Mallick, B. K., and Smith, A. F. M. (1998a). Automatic Bayesian curve fitting. *Journal of the Royal Statistical Society,* Series B **60**, 333–350.

Denison, D. G. T., Mallick, B. K., and Smith, A. F. M. (1998b). A Bayesian CART algorithm. *Biometrika* **85**, 363–377.

DiCiccio, T. J., Kass, R. E., and Wasserman, L. (1997). Computing bayes factors by combining simulation and asymptotic approximations. *Journal of the American Statistical Association* **92**, 903–915.

Dobra, A., Hans, C., Jones, B., Nevins, J., Yao, G., and West, M. (2004). Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis* **90**, 196–212.

Foster, D. P. and George, E. I. (1994). The risk inflation criterion for multiple regression. *Annals of Statistics* **22**, 1947–1975.

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

Gelfand, A. E. (1996). Model determination using sampling-based methods. In

W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, eds., *Markov chain Monte Carlo in Practice*, 145–161. Chapman & Hall/CRC, London.

Gelfand, A. E., Dey, D. K., and Chang, H. (1992). Model determination using predictive distributions with implementation via sampling based methods. In J. M. Bernardo, J. O. Berger, P. A. Dawid, and A. F. M. Smith, eds., *Bayesian Statistics 4*, 147–167. Clarendon Press, Oxford, U.K.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistics Association* **85**, 398–409.

George, E. I. (1999). Discussion of "Bayesian model averaging and model search strategies" by M.A. Clyde. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, eds., *Bayesian Statistics 6*, 175–177. Oxford University Press.

George, E. I. and Foster, D. P. (2000). Calibration and empirical Bayes variable selection. *Biometrika* **87**, 731–747.

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

George, E. I. and McCulloch, R. E. (1996). Stochastic search variable selection. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, eds., *Markov chain Monte Carlo in Practice*, 203–214. Chapman & Hall/CRC, London.

George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica* **7**, 339–373.

Geweke, J. (1996). Variable selection and model comparison in regression. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, eds., *Bayesian Statistics 5*, 609–620. Oxford University Press.

Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics* **41**, 337–348.

Giudici, P. (1996). Learning in graphical Gaussian models. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, eds., *Bayesian Statistics 5*, 621–628. Oxford University Press.

Giudici, P. and Green, P. J. (1999). Decomposable graphical Gaussian model determination. *Biometrika* **86**, 785–801.

Godsill, S. J. (2001). On the relationship between Markov chain Monte Carlo methods for model uncertainty. *Journal of Computational and Graphical Statistics* **10**, 230–248.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 4, 711–732.

Hans, C. and Dunson, D. B. (2005). Bayesian inferences on umbrella orderings. *Biometrics,* to appear.

Harville, D. A. (1997). *Matrix Algebra From a Statistician's Perspective.* Springer, New York.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning.* Springer, New York.

Higdon, D. M. (1998). Auxiliary variable methods for Markov chain Monte Carlo with applications. *Journal of the American Statistical Association* **93**, 585–595.

Hjort, N. L. and Claeskens, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Society* **98**, 879–899.

Huang, E., Chen, S., Dressman, H., Pittman, J., Tsou, M. H., Horng, C. F., Bild, A., Iversen, E. S., Liao, M., Chen, C. M., West, M., Nevins, J. R., and Huang, A. T. (2003). Gene expression predictors of breast cancer outcomes. *The Lancet* **361**, 1590–1596.

Huang, E., West, M., and Nevins, J. R. (2002). Gene expression profiles and predicting clinical characteristics of breast cancer. *Hormone Research* **58**, 55–73.

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2001). *Bayesian Survival Analysis.* Springer, New York.

Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., and Speed, T. P. (2003a). Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research* **31**, e15.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003b). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **2**, 249–264.

Jefferys, W. and Berger, J. O. (1992). Ockham's razor and Bayesian analysis. *American Scientist* **80**, 64–72.

Johnstone, I. M. and Silverman, B. W. (2004). Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *The Annals of Statistics* **32**, 1594–1649.

Johnstone, I. M. and Silverman, B. W. (2005). Empirical Bayes selection of wavelet thresholds. *The Annals of Statistics* **33**.

Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C., and West, M. (2005). Experiments in stochastic computation for high-dimensional graphical models. *Statistical Science,* to appear.

Kabaila, P. (1995). The effect of model selection on confidence regions and prediction regions. *Journal of Econometric Theory* **11**, 537–549.

Kohn, R., Smith, M., and Chan, D. (2001). Nonparametric regression using linear combinations of basis functions. *Statistics and Computing* **11**, 313–322.

Kullback, S. (1959). *Information Theory and Statistics*. Dover, Mineola, New York.

Kuo, L. and Mallick, B. (1998). Variable selection for regression models. *Sankhya,* Series B **60**, 65–81.

Lange, K. (1999). *Numerical Analysis for Statisticians*. Springer, New York.

Lauritzen, S. L. (1996). *Graphical Models*. Clarendon Press, Oxford.

Leeb, H. (2005). The distribution of a linear predictor after model selection: Conditional finite-sample distributions and asymptotic approximation. *Journal of Statistical Planning and Inference* **134**, 64–89.

Legler, J. M., Ries, L. A. G., Smith, M. A., Warren, J. L., Heineman, E. F., Kaplan, R. S., and Linet, M. S. (1999). Brain and other central nervous system

cancers: Recent trends in incidence and mortality. *Journal of the National Cancer Institute* **91**, 1382–1390.

Lindley, D. V. (1956). On a measure of information provided by an experiment. *Annals of Mathematical Statistics* **27**, 986–1005.

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

Mallows, C. L. (1973). Some comments on C_p. *Technometrics* **15**, 661–675.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate Analysis*. Academic Press, London.

McCulloch, R. and Rossi, P. E. (1994). An exact likelihood analysis of the multinomial probit model. *Journal of Econometrics* **64**, 207–240.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association* **83**, 1023–1032.

Muirhead, R. J. (1982). *Aspects of multivariate statistical theory*. John Wiley & Sons, New York.

Nevins, J. R., Huang, E. S., Dressman, H., Pittman, J., Huang, A. T., and West, M. (2003). Towards integrated clinico-genomic models for personalized medicine: Combining gene expression signatures and clinical factors in breast cancer outcomes prediction. *Human Molecular Genetics* **12**, 153–157.

Nott, D. J. and Green, P. J. (2004). Bayesian variable selection and the Swendson-Wang algorithm. *Journal of Computational and Graphical Statistics* **13**, 1–17.

Odell, P. L. and Feiveson, A. H. (1966). A numerical procedure to generate a sample covariance matrix. *Journal of the American Statistical Association* **61**, 199–203.

Peruggia, M. (1997). On the variability of case-deletion importance sampling weights in the Bayesian linear model. *Journal of the American Statistical Association* **92**, 199–207.

Pittman, J., Huang, E., Dressman, H., Horng, C. F., Cheng, S. H., Tsou, M. H., Chen, C. M., Bild, A., Iversen, E. S., Huang, A. T., Nevins, J. R., and West, M. (2004). Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes. *Proceedings of the National Academy of Sciences* **101**, 8431–8436.

Porter, P. I., Sage, E. H., Lane, T. F., Funk, S. E., and Gown, A. M. (1995). Distribution of SPARC in normal and neoplastic human tissue. *Journal of Histochemistry and Cytochemistry* **43**, 791–800.

Pötscher, B. (1991). Effects of model selection on inference. *Econometric Theory* **7**, 163–185.

Raftery, A. E. (1995). Bayesian model selection in social research (with discussion). In P. Marsden, ed., *Sociological Methodology*, 111–196. Blackwells, Cambridge, Mass.

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association* **92**, 1197–1208.

Rempel, S. A., Ge, S., and Gutierrez, J. A. (1999). SPARC: A potential diagnostic marker of invasive meningiomas. *Clinical Cancer Research* **5**, 237–241.

Rempel, S. A., Golembieski, W. A., Ge, S., Lemke, N., Elisevich, K., Mikkelsen, T., and Gutierrez, J. A. (1998). SPARC: A signal of astrocytic neoplastic transformation and reactive response in human primary and xenograft gliomas. *Journal of Neuropathology and Experimental Neurology* **57**, 1112–1121.

Rich, J. N., Hans, C., Jones, B., Iversen, E. S., McClendon, R. E., Rasheed, B. K. A., Dobra, A., Dressman, H. K., Bigner, D. D., Nevins, J. R., and West, M. (2005). Gene expression profiling and genetic markers in glioblastoma survival. *Cancer Research* **65**, 10.

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

Shannon, C. (1948). A mathematical theory of communication. *Bell System Tech. Journal* **27**, 379–423; 623–656.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica* **7**, 221–264.

Smith, A. F. M. and Gelfand, A. E. (1992). Bayesian statistics without tears: A sampling-resampling perpsective. *The American Statistician* **46**, 84–88.

Smith, A. F. M. and Spiegelhalter, D. J. (1980). Bayes factors as choice criteria. *Journal of the Royal Statistical Society,* Series B **42**, 213–220.

Smith, M. and Kohn, R. (1996). Nonparametric regression using Bayesian variable selection. *Journal of Econometrics* **75**, 317–343.

Smith, M. and Kohn, R. (1997). A Bayesian approach to nonparametric bivariate regression. *Journal of the American Statistical Association* **92**, 1522–1535.

Termine, J. D., Kleinman, H. K., Whitson, S. W., Conn, K. M., McGarvey, M. I., and Martin, G. R. (1981). Osteonectin, a bone-specific protein linking mineral to collagen. *Cell* **26**, 99–105.

Thomas, R., True, L. D., Bassuk, J. A., Lange, P. H., and Vessella, R. I. (2000). Differential expression of Osteonectin/SPARC during human prostate cancer progression. *Clinical Cancer Research* **6**, 1140–1149.

Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research* **1**, 211–244.

West, M. (2003). Bayesian factor regression models in the "large p, small n" paradigm. In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, eds., *Bayesian Statistics 7*, 723–732. Oxford University Press.

West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., Marks, J. R., and Nevins, J. R. (2001). Predicting the clinical status of human breast cancer utilizing gene expression profiles. *Proceedings of the National Academy of Sciences* **98**, 11462–11467.

West, M. and Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models*. Springer–Verlag, New York, 2nd edn.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. John Wiley and Sons, Chichester, United Kingdon.

Wolfe, P. J., Godsill, S. J., and Ng, W.-J. (2004). Bayesian variable selection and regularisation for time-frequency surface estimation. *Journal of the Royal Statistical Society,* Series B **66**, 575–589.

Yuan, M. and Lin, Y. (2005). Efficient empirical Bayes variable selection and estimation in linear models. *Journal of the American Statistical Association,* to appear.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In P. K. Goel and A. Zellner, eds., *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, 233–243. North-Holland, Amsterdam.

Biography

Christopher Mark Hans was born on December 19, 1978 in Amherst, New York. He received his A.B. degree *cum laude* in statistics on June 7, 2001 from Harvard University in Cambridge, Massachusetts and received an M.S. in statistics on September 1, 2003 from Duke University in Durham, North Carolina. He has co-authored the following articles:

1. Dobra, A., Hans, C., Jones, B., Nevins, J.R., Yao, G. and West, M. (2004). Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*, **90**, 196–212.

2. Hans, C., Dobra, A. and West, M. (2005). Shotgun stochastic search for regression with many candidate predictors. *ISDS Discussion Paper 05-10*.

3. Hans, C. and Dunson, D.B. (2005). Bayesian inferences on umbrella orderings. *Biometrics*, in press.

4. Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C. and West, M. (2005). Experiments in stochastic computation for high-dimensional graphical models. *Statistical Science*, in press.

5. Rich, J.N., Hans, C., Jones, B., Iversen, E.S., McClendon, R.E., Rasheed, B.K.A., Dobra, A., Dressman, H.K., Bigner, D.D., Nevins, J.R. and West, M. (2005). Gene expression profiling and genetic markers in glioblastoma survival. *Cancer Research*, **65** (10).

6. van Dyk, D.A. and Hans, C. (2002). Accounting for absorption lines in images obtained with the Chandra X-ray observatory. In A. Lawson and D. Denison, eds., *Spatial Cluster Modelling*, 175–198. Chapman & Hall/CRC.