ABSTRACT

CURTIS, STEVEN M. Variable Selection Methods with Applications to Shape Restricted Regression. (Under the direction of Professors S. Ghosal and S. K. Ghosh).

This dissertation consists of four major projects. Two of these projects develop and extend Bayesian variable selection methods. The remaining projects apply existing variable selection and quadratic programming methods to the problem of fitting a shape-restricted regression curve. We summarize each of these projects below.

High correlation among predictors has long been an annoyance in regression analysis. The crux of the problem is that the linear regression model assumes each predictor has an independent effect on the response that can be encapsulated in the predictor's regression coefficient. When predictors are highly correlated, the data do not contain much information on the independent effects of each predictor. The high correlation among predictors can result in large standard errors for the regression coefficients and coefficients with signs opposite of what is expected based on *a priori*, subject-matter theory. We propose a Bayesian model that accounts for correlation among the predictors by simultaneously performing selection and clustering of the predictors. Our model combines a Dirichlet process prior and a variable selection prior for the regression coefficients. In our model, highly correlated predictors can be grouped together by setting their corresponding coefficients exactly equal. Similarly, redundant predictors can be removed from the model through the variable selection component of our prior. We demonstrate the competitiveness of our method through simulation studies and analysis of real data.

The literature is replete with variable selection techniques for the classical linear regression model. It is only relatively recently that authors have begun to explore variable selection in fully nonparametric and additive regression models. One such variable selection technique is a generalization of the LASSO called the group LASSO. In this work, we demonstrate a connection between the group LASSO and Bayesian inference in additive models with a multivariate Laplace prior for model parameters similar to the connection between the LASSO and Bayesian inference in the linear model with a univariate Laplace prior for regression coefficients. We use this connec-

tion to derive approximate posterior model probabilities for additive models. We use the concept of regular and nonregular models to reduce the size of the model space and avoid costly computations.

The simple regression problem in statistics consists of determining the relationship between a response variable and a single predictor variable through a regression function. Prior information is often available that suggests the regression function should have a certain shape (e.g. monotonically increasing or concave) but not necessarily a specific parametric form. Recently, Bernstein polynomials have been used to impose certain shape restrictions on regression functions. In this work, we demonstrate a connection between the monotonic regression problem and the variable selection problem in the linear model. We develop a Bayesian procedure for fitting the monotonic regression model by adapting the variable selection procedure of previous authors. We demonstrate the effectiveness of our method through simulations and the analysis of real data.

The workhorse in statistical inference is the linear regression model. However, in empirical research, the assumptions implicit in the linear regression model are often too restrictive. The literature contains many nonparametric methods for fitting regression curves that only make minimal smoothness assumptions about the regression curve. However, in many situations substantive subject-matter information exists on the general shape of the regression function—nondecreasing or concave, for example. We use Bernstein polynomials to fit shape-restricted regression curves. We fit the Bernstein regression curves to data using quadratic programming to impose the necessary shape restrictions. We choose tuning parameters using Schwarz's information criterion. Our proposed method has advantages over other methods in that it is straightforward to implement in existing software and generalize to other shape restrictions.

Variable Selection Methods with Applications to Shape Restricted Regression

by
Steven McKay Curtis

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fullfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina

2008

APPROVED BY:

Dr. Helen Zhang	Dr. Howard Bondell
Dr. Subhashis Ghosal Co-Chair of Advisory Committee	Dr. Sujit K. Ghosh Co-chair of Advisory Committee

DEDICATION

To Steph.

BIOGRAPHY

In his sophomore year of high school, Steven McKay Curtis (or simply "McKay," as his parents have always called him) decided that school was "the pits" and decided to invest as little effort into schooling as possible. He dropped out of all of his advanced courses and coasted through the rest of high school content to learn as little as possible. So it is not without some irony that he has written this dissertation to fulfill the requirements of a doctoral degree in statistics.

At some time between his sophomore year of high school and his freshman year of college, McKay had a change of heart regarding education. He has since received a Bachelor of Arts degree in economics (in 2002) and a Master of Science degree in statistics (in 2004) from Brigham Young University, and he has nearly completed all the requirements for the attainment of Doctor of Philosophy in statistics at North Carolina State University.

Internships at companies like Nestlé and Intel have helped convince McKay that the worst day in school is still better than the best day at work in a "real" job. Thus, he hopes to remain in academia (in some capacity) until he is kicked out or kicks the bucket.

The many nonacademic milestones in McKay's life include his completing a two-year religious mission to Ontario, Canada (for which he is still bitter that he did not get sent to a country where he could learn a foreign language—like his brother Ryan, who now speaks fluent Spanish, and his brother Shane, who now speaks fluent Russian); his marriage to Stephanie Perry on August 31, 2001 (whom he met as a Teaching Assistant in a statistics class); and the birth of his three children—Max in 2003, Renée in 2005, and Logan in 2007.

McKay's deepest political desire is that one day the US government will remove the onerous trade restrictions on sugar imports, so that his favorite soft drink—Dr. Pepper—will again be made with pure cane sugar instead of high-fructose corn syrup (outside the state of Texas).

ACKNOWLEDGEMENTS

I would like to first thank my two advisors—Dr. Subhashis Ghosal and Dr. Sujit Ghosh—for without their direction, encouragement and wealth of good ideas, this dissertation would never have been written. Dr. Ghosal has always been very generous with his good ideas and with his time. In addition to meeting with me on a weekly basis, Dr. Ghosal has been willing to accommodate me when I have dropped by his office—even before I was his student!

Dr. Ghosh has also been generous with his time by meeting with me on a weekly basis. Dr. Ghosh also exhibits a limitless supply of optimism in the face of research set-backs—a quality I hope to emulate.

I would like to thank my "boss" at SAS, Fang K. Chen. Fang has a wealth of practical experience in Bayesian computation and has been willing to share that with me by giving me several instructive projects to work on. Fang has also been a great friend, and I have enjoyed many personal conversations with him.

My advisors at BYU have also been influential on my development as a person and a statistician. Dr. (or "Coach") Gil Felligham was and is a good motivator, and Dr. Shane Reese has given me good advice, both personal and professional, on many occasions.

Drs. John Monahan, Helen Zhang, and Howard Bondell have all been friendly and helpful when I have stopped by their offices (on multiple occasions!) to ask questions. Alvin Van Orden's perceptive questions forced me to think more carefully about statistical practice and philosophy. (I am still hoping to convert him to Bayesianism, however.) Conversations with Jaren Pope about economics and political philosophy were always enlightening and challenging. Jake Bartlett has patiently endured my many computer-science questions about programming in C, C++, Fortran, Python, and Perl, and has given me much sound computing advice. Arun Krishna has been a good sounding board when I hit snags in my research. Ian Fiske has given me great advice and tips on computing in R, LaTeX, and Emacs. Anthony Franklin has been generous in sharing his newly discovered tips for R and LaTeX.

A few members of my local church have been particularly supportive during these

past four years. Bishop Charles Anderson was exceptionally understanding with my religious concerns. Paul Nelson epitomizes a good "home teacher" and deserves many thanks for bailing me out of some home-improvement pratfalls. Jaren Pope also deserves thanks as an understanding home teacher.

Members of my family also deserve thanks. My mother, Betty Ann Curtis, has always been supportive of my academic endeavors (although she still may wonder why I bothered with a PhD!). My father, Steven Curtis, has given more than just moral support. He sacrificed three and a half days of his time to help me drive a moving truck from Tucson to Raleigh, so I could begin the PhD program at NC State. In the coming weeks, he will spend four and a half days helping me drive the nearly 3,000 miles from Raleigh to Seattle, so I can start a postdoc at the University of Washington. I thank my older sister Rachelle for introducing me to statistics and my younger brothers and sister—Colby, Ryan, Shane, and Tiare—for making me feel like I know something when they occasionally ask me questions about statistics.

Finally, I would like to thank my wife, Stephanie. She has been extremely patient with my "education obsession." We will be married seven years at the end of this coming August, and for all of those years I have been a student—a profession which requires too much time and pays too little money! Through it all, Stephanie has been my biggest supporter and best friend. She has endured long hours at home raising our three (wonderful but exhausting) children—Max, Renée, and Logan. To help supplement our income and to ensure she could be at home for our children, Stephanie has spent many hours at home toiling at jobs that have not been enjoyable. Stephanie has always been encouraging when the vagaries of graduate study have taken disappointing directions. I feel very lucky to share life's journey with her.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

Most problems in statistics involve determining the relationship of a particular quantity of interest, or response variable (usually denoted with a Y), and one or more predictor variables or covariates (usually denoted with $x_1, \ldots, x_p$). Most often, the response variable Y is assumed to vary randomly around its mean, where the mean is some function of the predictor variables. More precisely, the response variable is assumed to have the following relationship

$$Y_i = f(x_{1i}, \ldots, x_{pi}) + \epsilon_i, \tag{1.1}$$

where ϵ_i is typically assumed to have a symmetric distribution (often a normal distribution with mean zero and variance σ^2) around zero. We will call this model the general regression model.

Many forms of data analysis rely on some version of the general regression model. The most basic form of the model assumes that the regression function $f(\cdot)$ can be represented as a linear function of coefficients, i.e.

$$f(x_1, \ldots, x_p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p = \beta_0 + \mathbf{x}^T \boldsymbol{\beta} \tag{1.2}$$

The linear regression model provides a nice interpretation for the effect of each predictor on the mean of the response, which makes this model popular in applied research. According the linear regression model, a one-unit increase in the predictor x_j gives a β_j increase in the mean of the response variable.

The linear regression model embodies very strict assumptions on the form of the mean of the response variable. These strict assumptions may not be valid in many practical applications. Thus, researchers have derived alternative models with less stringent assumptions. One popular approach presented by Stone (1985) assumes that

$$f(x_1, \ldots, x_p) = \beta_0 + f_1(x_1) + \cdots + f_p(x_p), \tag{1.3}$$

where the $f_j(\cdot)$ are arbitrary smooth functions (see also Hastie and Tibshirani, 1990)). Regression models of the form (1.3) are called additive models. Additive models retain some of the simplicity of the linear regression model because each predictor has a separate additive effect on the mean of the response.

The additive regression model can further be generalized by adding functional interaction components to (1.3). Functional ANOVA models were first introduced by Barry (1986) and involve "partitioning" the regression function into main effects ($f_1(x_1)$,...,$f_p(x_p)$), as in the additive model, two-way interaction effects ($f_{1,2}(x_1, x_2)$, ...,$f_{1,p}(x_1, x_p)$, $f_{2,3}(x_2, x_3)$, ...), and other higher order interactions. (See also Wahba (1990); Gu (2002).)

Fully nonparametric regression tries to estimate the regression function $f(x_1, \ldots, x_p)$ without assuming any of the previous structure. Estimating the regression function nonparametrically is a notoriously difficult problem because of the curse of dimensionality. The curse of dimensionality was succinctly stated by Wasserman (2005, page 101), who said

> To maintain a given degree of accuracy of an estimator, the sample size must increase exponentially with the dimension.

When p is a large, a researcher may wish to confine a study to some number of predictors less than the total number of predictors p. Reducing the number of predictors under consideration can help avoid the curse of dimensionality, but it may also improve prediction of future variables or improve understanding of the phenomenon under study (Cox and Snell, 1974).

In this dissertation, we develop new variable selection methodology and apply existing variable selection methods to other statistical problems (e.g., shape-restricted

regression). In the remainder of this introduction, we review the relevant background literature for the methods that are developed in the remaining chapters of this dissertation. We begin by reviewing variable selection in the general regression model by first considering variable selection in linear regression, additive regression, and fully nonparametric regression. We then review the literature on other topics that we encounter in our study of variable selection. These other topics include shape-restricted regression and clustering. Finally, we end this chapter with a discussion of what follows in the rest of this dissertation.

1.1 Variable Selection in the Classical Linear Model

The classical linear regression model assumes

$$\mathbf{Y} \sim \mathsf{N}_n\big(\mathbf{1}\beta_0 + \mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}\big) \tag{1.4}$$

where $\mathbf{Y}$ is an $n \times 1$ vector of random responses, $\mathbf{X}$ is an $n \times p$ matrix of known predictors, β_0 is an unknown scalar, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a $p \times 1$ vector of unknown regression coefficients, σ^2 is an unknown scalar variance term, and $\mathbf{I}$ is an identity matrix of the appropriate size. In a variable selection problem, a researcher's goal is to determine which of the p predictors are zero or small enough to ignore. Often, a $p \times 1$ vector of binary variables $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)^T$ is used to represent which predictors are deemed to have zero regression coefficients, where $\gamma_j = 1$ indicates $\beta_j \neq 0$ and $\gamma_j = 0$ indicates $\beta_j = 0$. The notation $\mathbf{X}_{\boldsymbol{\gamma}}$ is often used to denote an $n \times q_{\boldsymbol{\gamma}}$ matrix (where $q_{\boldsymbol{\gamma}}$ is the sum of elements in $\boldsymbol{\gamma}$) with columns corresponding to the nonzero elements of $\boldsymbol{\gamma}$. Similarly, $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ denotes a $q_{\boldsymbol{\gamma}} \times 1$ vector of regression coefficients corresponding to the nonzero elements of $\boldsymbol{\gamma}$. In this section, we review both non-Bayesian and Bayesian methods of variable selection.

1.1.1 Non Bayesian Variable Selection

Over the years, a plethora of variable selection methods have been developed (Miller, 2002). Greedy algorithms like forward and backward stepwise selection attempt to find the best subset by proceeding in steps where at each step the best predictor (according to some criterion like an F-test statistic) is included in the regression model or the worst predictor is removed from the model (Neter, Kutner, Wasserman, and Nachtsheim, 1996).

Other approaches, like Mallows C_p (which we denote $C_{\boldsymbol{\gamma}}$), involve estimating the out-of-sample predictive error of a each model $\boldsymbol{\gamma}$ (Mallows, 1973; Hastie, Tibshirani, and Friedman, 2001, Section 7.5). The $C_{\boldsymbol{\gamma}}$ statistic was originally derived by considering the the estimation of σ^2. The residual sum of squares in the linear regression model can be used to estimate σ^2 by noting

$$\begin{aligned} \mathrm{E}(\mathrm{RSS}_{\boldsymbol{\gamma}}) &= \mathrm{E}[(\mathbf{y} - \mathbf{X}_{\boldsymbol{\gamma}}\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}})^T(\mathbf{y} - \mathbf{X}_{\boldsymbol{\gamma}}\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}})] \\ &= (n - p - 1)\sigma^2 + B_{\boldsymbol{\gamma}} \end{aligned} \tag{1.5}$$

where $\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ is the least-squares estimator for predictors $\mathbf{X}_{\boldsymbol{\gamma}}$ and $B_{\boldsymbol{\gamma}}$ is a nonnegative bias term that is zero if all predictors not included in the model $\boldsymbol{\gamma}$ have zero coefficients. When the regression model contains all nonzero coefficients, the term (1.5) can be used to unbiasedly estimate σ^2 by dividing $\mathrm{RSS}_{\boldsymbol{\gamma}}$ by $n-p-1$. Thus, Mallows' criterion

$$C_{\boldsymbol{\gamma}} = \frac{\mathrm{RSS}_{\boldsymbol{\gamma}}}{\hat{\sigma^2}} - n + 2q_{\boldsymbol{\gamma}} \tag{1.6}$$

has expectation (asymptotically) equal to $q_{\boldsymbol{\gamma}}$. Models which exclude important predictors will have a $C_{\boldsymbol{\gamma}}$ statistic greater than $q_{\boldsymbol{\gamma}}$, and models which include all relevant predictors (and possibly some irrelevant predictors) will have $C_{\boldsymbol{\gamma}}$ statistics that are close to $q_{\boldsymbol{\gamma}}$.

The approach of Akaike (1973) attempts to find the model with the smallest Kullback-Leibler (Kullback and Leibler, 1951) divergence from the true, but unknown, data generating process. The general form of AIC is

$$\mathrm{AIC} = -2\log(f(\mathbf{y}|\boldsymbol{\theta})) + 2\dim(\boldsymbol{\theta}),$$

where $f(\mathbf{y}|\boldsymbol{\theta})$ is the model for the data conditional on parameters $\boldsymbol{\theta}$ and $\dim(\boldsymbol{\theta})$ is the dimension of $\boldsymbol{\theta}$. In the context of linear regression with normal errors the AIC for model $\boldsymbol{\gamma}$ is

$$\text{AIC}_{\boldsymbol{\gamma}} = -\frac{(\mathbf{y} - \mathbf{X}_{\boldsymbol{\gamma}}\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}})^T(\mathbf{y} - \mathbf{X}_{\boldsymbol{\gamma}}\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}})}{\hat{\sigma}^2_{\boldsymbol{\gamma}}} + 2q_{\boldsymbol{\gamma}},$$

where $\hat{\sigma}^2_{\boldsymbol{\gamma}}$ is the estimate of the error variance for model $\boldsymbol{\gamma}$. In this form, the $\text{AIC}_{\boldsymbol{\gamma}}$ is equivalent to Mallows' $C_{\boldsymbol{\gamma}}$ (Hastie et al., 2001, page 204).

Variable selection proceeds by computing the $C_{\boldsymbol{\gamma}}$ statistic or $\text{AIC}_{\boldsymbol{\gamma}}$ statistic for all possible models and choosing the most parsimonious models with $C_{\boldsymbol{\gamma}}$ statistics close to $q_{\boldsymbol{\gamma}}$ or the models with the lowest values of $\text{AIC}_{\boldsymbol{\gamma}}$ for further scientific investigation.

Penalized Least Squares

Many variable selection techniques rely on penalized, least-squares techniques. Penalized least-squares problems are solved by finding the values of the regression coefficients that minimize

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \boldsymbol{\lambda}^T\mathbf{J}(\boldsymbol{\beta}), \tag{1.7}$$

where $\mathbf{J}(\cdot) = (J_1(\cdot), \ldots, J_g(\cdot))$ is a vector-valued function that penalizes large coefficients and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_g)$ is a vector of tuning parameters that controls the balance between model fit and the penalty term. Different forms of the penalty term in (1.7) give estimators with different behavior. Some of the penalties that have appeared in the literature are summarized in this section and in Table 1.1.

Penalized regression in the context of the linear model was introduced by Hoerl and Kennard (1970) as a way to correct for highly collinear predictors in regression. Hoerl and Kennard (1970) use $\mathbf{J}(\boldsymbol{\beta}) = \sum_{j=1}^{p} \beta_j^2$ and call this penalty the ridge-regression penalty. This penalty introduces some bias into the estimates by shrinking the standard least-squares estimates toward zero; however, under certain conditions, the ridge regression estimates are improvement over least squares because they have smaller variance. Although this penalty does not set coefficients exactly equal to zero (i.e., this penalty does not perform variable selection), it is often used in conjunction with other penalties that do perform variable selection.

Breiman (1995) develops a variable selection method that is very similar to the form in (1.7). Breiman (1995) gives the method the morbid but descriptive name "nonnegative garrote." Webster's dictionary defines a garrote as "the apparatus used...for execution by strangulation." The statistical method in Breiman (1995) minimizes

$$\underset{\mathbf{c}}{\text{argmin}} = \left\| \mathbf{y} - \sum_{j=1}^{p} \mathbf{x}_j \hat{\beta}_j^{LS} c_j \right\|^2 + \lambda \sum_{j=1}^{p} c_j \tag{1.8}$$

subject to the constraint $c_j \geq 0$, where $\hat{\beta}_j^{LS}$ denotes the typical least-squares estimate of the j^{th} regression coefficient. The aptness of the name "nonnegative garrote" is apparent from the form of (1.8). Small values of c_j constrict (strangle?!) their respective least-squares estimates by shrinking them toward zero. Also, the solution to (1.8) may set some of the c_j to zero, thereby eliminating ("executing") their corresponding least-squares estimates. The c_j are restricted to be nonnegative, which removes the possibility that the nonnegative garrote estimates have a different sign than their least squares estimates.

The Least Absolute Shrinkage and Selection Operator (LASSO) of Tibshirani (1996) uses $\mathbf{J}(\boldsymbol{\beta}) = \sum_{j=1}^{p} |\beta_j|$ as a penalty. Tibshirani shows that because of the unique geometry of the penalty term $\sum_{j=1}^{p} |\beta_j|$ the solution to

$$\underset{\boldsymbol{\beta}}{\text{argmin}} \left\| \mathbf{y} - \beta_0 - \mathbf{X}\boldsymbol{\beta} \right\|^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{1.9}$$

includes $\beta_j = 0$ for some subset of the p predictors when λ is large enough.

Many other penalized regression techniques alter the LASSO penalty to obtain other desirable variable selection features. The "elastic-net" penalty $\boldsymbol{\lambda}^T \mathbf{J}(\boldsymbol{\beta}) = \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2$ of Zou and Hastie (2005) uses a convex combination of the ridge-regression penalty and the LASSO penalty. Because the elastic-net penalty includes the LASSO penalty, the elastic-net solution can set regression coefficients to zero. Additionally, Zou and Hastie (2005) show that for certain values of the tuning parameters the elastic net will select highly correlated predictors as a group rather than eliminating some of them from the model arbitrarily.

The Octogonal Shrinkage and Clustering Algorithm for Regression (OS-

CAR) method of Bondell and Reich (2008) uses $\boldsymbol{\lambda}^T\mathbf{J}(\boldsymbol{\beta}) = \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j<k} \max\{\beta_j, \beta_k\}$, which is a convex combination of the LASSO penalty and the sum of pairwise maximums of regression coefficients. Like the elastic net, the LASSO part of the OSCAR penalty can set regression coefficients to zero. The OSCAR can also include highly correlated predictors in the model as a group by its ability to set regression coefficients equal to each other. This makes the OSCAR a viable, predictor-clustering algorithm (unlike the elastic net), because researchers can identify clusters by noting which regression coefficients are set equal.

Fan and Li (2001) develop the Smoothly Clipped Absolute Deviation (SCAD) penalty that is defined by its derivative $[\nabla J(\boldsymbol{\beta})]_j = \lambda \left[\mathbb{1}_{(-\infty,\lambda]}(\beta_j) + \frac{(a\lambda - \beta_j)_+}{(a-1)\lambda} \mathbb{1}_{(\lambda,\infty)}(\beta_j) \right]$, where $\mathbb{1}_A(\cdot)$ is the indicator function that equals one if its argument is in the set A and equals zero otherwise. This derivative corresponds to a penalty of $J(\boldsymbol{\beta}) = \sum_{j=1}^{p} h_\lambda(|\beta_j|)$, where $h_\lambda(z)$

$$h_\lambda(z) = \begin{cases} z & \text{for } 0 < z \leq \lambda \\ -\frac{1}{2(a-1)}(z - 2a\lambda z + \lambda^2) & \text{for } \lambda < z \leq a\lambda \\ \frac{1}{2}(a+1)\lambda^2 & \text{for } a\lambda < z \end{cases}$$

and a is a tuning parameter. Based on results of simulation studies, Fan and Li recommend setting $a = 3.7$. For values of $\beta_j \leq \lambda$, the SCAD penalty is the same as the LASSO. Thus, the point of nondifferentiability at zero allows the SCAD to set coefficients to zero. When $\beta_j \geq a\lambda$, the SCAD penalty is constant as a function of β_j which avoids unnecessary bias in the estimates for large parameters.

Fan and Li (2001) show that estimates using the SCAD penalty have the "oracle" properties. These two properties are defined as follows (see also Zou, 2006). Let $T = \{j \in \{1, \ldots, p\} | \beta_j \neq 0\}$—that is, the set T contains the indices of the true regression coefficients that are nonzero. Thus, the set T^c contains the indices of regression coefficients that are zero. Let $\tilde{\boldsymbol{\beta}}$ be the penalized regression estimates and let $\tilde{T} = \{j \in \{1, \ldots, p\} | \tilde{\beta}_j \neq 0\}$—that is, $\tilde{T}$ contains the indices of all penalized regression estimates that are nonzero. The first oracle property is $\lim_{n\to\infty} \Pr(\tilde{T}^c = T^c) = 1$. In other words, as the sample size increases, the penalized regression procedure will eliminate the irrelevant predictors with probability tending to one.

Table 1.1: Penalized regression techniques.

Regression Penalties	
Method	Penalty
Ridge regression	$\lambda \sum_{j=1}^{p} \beta_j^2$
LASSO	$\lambda \sum_{j=1}^{p} \lvert\beta_j\rvert$
Elastic net	$\lambda_1 \sum_{j=1}^{p} \lvert\beta_j\rvert + \lambda_2 \sum_{j=1}^{2} \beta_j^2$
OSCAR	$\lambda_1 \sum_{j=1}^{p} \lvert\beta_j\rvert + \lambda_2 \sum_{j<k} \max(\lvert\beta_j\rvert, \lvert\beta_k\rvert)$
SCAD	$\sum_{j=1}^{p} h_\lambda(\lvert\beta_j\rvert)$ where $h_\lambda(z) = \begin{cases} z & \text{for } 0 < z \leq \lambda \\ -\frac{1}{2(a-1)}(z - 2a\lambda z + \lambda^2) & \text{for } \lambda < z \leq a\lambda \\ \frac{1}{2}(a+1)\lambda^2 & \text{for } a\lambda < z \end{cases}$

The second oracle property is $\sqrt{n}(\tilde{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_T) \rightarrow \mathsf{N}(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is the asymptotic covariance matrix of the nonzero regression coefficients if the subset of true predictors were known. In other words, a procedure with this property behaves (asymptotically) as though it knows the true model.

Zou (2006) presents a penalized regression method that also possesses the oracle properties. This method is called the Adaptive LASSO and imposes the penalty $J(\boldsymbol{\beta}) = \sum_{j=1}^{p} w_j |\boldsymbol{\beta}_j|$. The w_j terms are derived from a root-n consistent estimator $\hat{\boldsymbol{\beta}}$ of the regression coefficients (such as the least-squares estimates) and are $w_j = 1/|\hat{\beta}_j|^\eta$, where η is another tuning parameter. Zou demonstrates that the adaptive LASSO can be estimated using the highly efficient Least Angle Regression (LARS) algorithm of Efron, Hastie, Johnstone, and Tibshirani (2004). The speed at which the Adaptive LASSO can be calculated is an advantage over other methods such as SCAD, especially when data sets are large.

Tuning Parameter Selection

Most classical variable selection methods require the selection of one or more tuning parameters. How to choose the values of these tuning parameters remains an active area of research. One clever approach by Wu, Boos, and Stefanski (2007) is to

add "false" predictors (i.e. randomly generated predictors) to the linear regression model to obtain an estimate of how often a variable selection procedure selects noise variables. Tuning parameters can then be selected to control for the "false-selection" rate.

1.1.2 Bayesian Variable Selection Methods

In this dissertation, we take a Bayesian approach to variable selection. The Bayesian approach to statistical inference involves modeling all unknown quantities with probability distributions (O'Hagan, 2004), then updating these probabilities by conditioning on the observed data. In typical estimation problems, a parametric distribution for the observed data is assumed and all parameters in this distribution are assigned probability distributions. All statistical inference is based on the posterior distribution of the parameters that can be obtained using Bayes theorem. In variable selection, the unknown quantities are not simply model parameters, but also the number of regression coefficients to be included in the regression function. In a Bayesian approach, a prior distribution must be assigned to all possible models $\boldsymbol{\gamma}$ and all other model parameters $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ and σ^2. Priors are usually assigned in a hierarchical fashion as in

$$p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2) = p(\boldsymbol{\beta}|\boldsymbol{\gamma}, \sigma^2)p(\boldsymbol{\gamma})p(\sigma^2). \tag{1.10}$$

Posterior model probabilities can then be used in a decision-theoretic setting (Berger, 1985) to pick the best subset of predictors or can be used in model averaging (Raftery, Madigan, and Hoeting, 1997).

Bayesian variable selection techniques can be broadly categorized by whether they assign an independent prior $p(\boldsymbol{\beta}|\boldsymbol{\gamma}) = p(\beta_1|\boldsymbol{\gamma}) \times \cdots \times p(\beta_p|\boldsymbol{\gamma})$ or a dependent prior for $\boldsymbol{\beta}$ in model $\boldsymbol{\gamma}$. We summarize some of these approaches below.

Independent Priors for β_j

If a method uses an independent prior for the regression coefficients, then it is usually specified as a mixture of two components—which we denote $p^{(0)}(\cdot)$ and $p^{(1)}(\cdot)$—conditioned on γ_j that correspond to the prior when $\gamma_j = 0$ and $\gamma_j = 1$ respectively.

More explicitly, the prior for each β_j is

$$p(\beta_j|\gamma_j) = (1-\gamma_j)p^{(0)}(\beta_j) + \gamma_j p^{(1)}(\beta_j) \tag{1.11}$$

The remaining parameters—$\boldsymbol{\gamma}$, β_0, and σ^2—are given independent prior distributions $p(\boldsymbol{\gamma})$, $p(\beta_0)$, and $p(\sigma^2)$ to complete the model specification.

Mitchell and Beauchamp (1988) propose the "spike and slab" prior for $p(\beta_j|\gamma_j)$. They set $p^{(0)}(\beta_j) = \mathbb{1}_{\{0\}}(\beta_j)$ (the "spike") and $p^{(1)}(\beta_j) = \frac{1}{2a}\mathbb{1}_{[-a,a]}(\beta_j)$ (the "slab"). Each γ_j is given an independent Bernoulli distribution with parameter w_j. They derive an explicit form for the posterior probability of each model and choose values of the hyperparameters (specifically, the value of the ratio $\frac{2(1-w_j)a}{w_j}$) using a Bayesian cross validation approach.

George and McCulloch (1993) and George and McCulloch (1997) use a very general form of prior for $\boldsymbol{\beta}$ but examine specifically a special case of the form in (1.11). The general prior they use is of the form

$$\boldsymbol{\beta}|\boldsymbol{\gamma} \sim \mathsf{N}_p(\mathbf{0}, \mathbf{D}_{\boldsymbol{\gamma}}\mathbf{R}_{\boldsymbol{\gamma}}\mathbf{D}_{\boldsymbol{\gamma}}) \tag{1.12}$$

where $\mathbf{R}_{\boldsymbol{\gamma}}$ is a correlation matrix, $\mathbf{D}_{\boldsymbol{\gamma}}$ is a diagonal matrix with the i^{th} diagonal element equal to $\sqrt{v_{0\boldsymbol{\gamma}(i)}}$ if $\boldsymbol{\gamma}_j = 0$ and $\sqrt{v_{1\boldsymbol{\gamma}(i)}}$ if $\boldsymbol{\gamma}_j = 1$, and $\boldsymbol{\gamma}(i) = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_{j-1}, \boldsymbol{\gamma}_{j+1}, \ldots, \boldsymbol{\gamma}_p)$. The notation $v_{0\boldsymbol{\gamma}(i)}$ and $v_{1\boldsymbol{\gamma}(i)}$ implies that the variance of β_j can vary depending on which of the other predictors are included in the model.

If $\mathbf{R}_{\boldsymbol{\gamma}} = \mathbf{I}_p$ and the diagonal elements of $\mathbf{D}_{\boldsymbol{\gamma}}$ are t_j if $\boldsymbol{\gamma}_j = 0$ and $c_j t_j$ if $\boldsymbol{\gamma}_j = 1$, where $c_j > 1$, then conditioned on $\boldsymbol{\gamma}$ the prior for $\boldsymbol{\beta}$ is of the form in (1.11), where $p^{(0)}(\beta_j) = \phi(\beta_j; 0, t_j^2)$, $p^{(1)}(\beta_j) = \phi(\beta_j; 0, c_j^2 t_j^2)$, and $\phi(\cdot; m, s^2)$ denotes the density of a normal distribution with mean m and variance s^2. Thus, when $\gamma_j = 0$, the prior for β_j has a variance small enough to render it practically insignificant.

George and McCulloch (1993) explore different distributions for $p(\boldsymbol{\gamma})$ such as independent Bernoulli priors for each γ_j and a prior on the size of the model, $p(\boldsymbol{\gamma}) = w_{q_{\boldsymbol{\gamma}}}\binom{p}{q_{\boldsymbol{\gamma}}}^{-1}$, where $w_{q_{\boldsymbol{\gamma}}}$ is the probability of a model of size $q_{\boldsymbol{\gamma}}$ and is specified for each model of size $q_{\boldsymbol{\gamma}}$.

Perhaps the most significant contribution of George and McCulloch (1993) is that they introduce the idea of searching the models space stochastically using Gibbs

sampling (Geman and Geman, 1984; Gelfand and Smith, 1990). Because p is large in many applications, enumerating all possible 2^p models is not feasible. The Gibbs sampling approach of George and McCulloch (1993) avoids this problem by stochastically searching for good models.

Geweke (1996) uses the same form for $p^{(0)}(\beta_j)$ as Mitchell and Beauchamp (1988) but a different form for $p^{(1)}(\beta_j)$. He uses a truncated normal distribution for $p^{(1)}(\beta_j)$. Geweke also uses independent Bernoulli prior for each γ_j. The form of $p^{(1)}(\beta_j)$ combined with the normal distribution for the response variable make the model amenable to a Gibbs sampling scheme, which Geweke derives.

Yuan and Lin (2005) also use $p^{(0)}(\beta_j) = \mathbb{1}_{\{0\}}(\beta_j)$ as in Mitchell and Beauchamp (1988). However, they use a double exponential distribution for $p^{(1)}(\beta_j)$, where the density for a random variable X with a double exponential distribution is $p_X(x) = \tau \exp(-\tau x)/2$. In certain situations, double exponential priors have some desirable theoretical advantages over the commonly used normal distribution (Johnstone and Silverman, 2005). For each model $\boldsymbol{\gamma}$, Yuan and Lin (2005) use the prior $p(\boldsymbol{\gamma}) \propto p_\gamma^{q_\gamma}(1-p_\gamma)^{(p-q_\gamma)} \det(\mathbf{X}_{\boldsymbol{\gamma}}^T \mathbf{X}_{\boldsymbol{\gamma}})$. This is similar to the independent Bernoulli prior except for the additional term $\det(\mathbf{X}_{\boldsymbol{\gamma}}^T \mathbf{X}_{\boldsymbol{\gamma}})$ that penalizes models with highly correlated predictors. Yuan and Lin (2005) show that by selecting certain values of p_γ and τ the model with the maximum posterior probability is also the model that would be selected under the LASSO. Yuan and Lin (2005) then derive empirical Bayes estimates for the hyper (tuning) parameters of their model. They show with a simulation study, that the empirical Bayes estimates of the tuning parameters in the LASSO outperform selection of tuning parameters through other methods (e.g., cross-validation).

Although we have presented the prior from George and McCulloch (1993) as a mixture prior on the regression coefficients, the prior could be viewed as a mixture prior on the variance term of a normal distribution. The prior for β_j can be written as $\beta_j \sim \mathsf{N}\left(0, \nu_j \tau_j^2\right)$ where ν_j has a discrete distribution that takes on some large value v_1 (greater than one) with probability ω_j and a value of 1 with probability $(1-\omega_j)$. Ishwaran and Rao (2005) generalize this prior by adding hyperpriors on τ_j^2 and on ω_j. Ishwaran and Rao (2005) let $\tau^{-2} \sim \mathsf{Gam}(a, b)$ and $\omega_j \sim \mathsf{Unif}(0, 1)$. Ishwaran and Rao (2005) also reverse the role of ν_j from the role of the prior proposed by George

and McCulloch (1993). Instead of being a constant that "blows up" the variance for regression parameters with $\gamma_j = 1$, the parameter ν_j is a random variable that takes on the value 1 with probability ω_j (when the j^{th} predictor is "in" the model) and a value close to zero with probability $(1 - \omega_j)$ (when the j^{th} predictor is "out" of the model). Additionally, Ishwaran and Rao (2005) rescale the response variable and add a variance penalty parameter λ_n to the variance of the response. Their model can be summarized by

$$\begin{aligned}
(Y_i^*|\boldsymbol{\beta}, \mathbf{x}_i, \sigma^2, \lambda_n) &\sim \mathsf{N}\left(\mathbf{x}_i^T\boldsymbol{\beta}, \sigma^2\lambda_n\right) && \text{for } i = 1, \ldots, n \\
(\beta_j|\nu_j, \tau_j^2) &\sim \mathsf{N}\left(0, \nu_j\tau_j^2\right) && \text{for } j = 1, \ldots, p \\
(\nu_j|w) &\sim (1-w)\delta_{v_0}(\cdot) + w\delta_1(\cdot) \\
\tau_j^{-2} &\sim \mathsf{Gam}(a_1, b_1) \\
w &\sim \mathsf{Unif}(0, 1)\,,
\end{aligned}$$

where $Y_i^* = \sqrt{\hat{\sigma}^2 n}Y_i$.

The above model formulation has some advantages over other variable selection methods. First, the effect of the variable-selection priors vanishes as the sample size increases. In other words, for large sample sizes, previous variable selection methods fail to set very many regression coefficients to zero. In the formulation of Ishwaran and Rao (2005), however, the rescaling of the response corrects for this and allows for the desired variable selection behavior even in large samples. Second, with a properly chosen value of λ_n the shrinkage effect of the posterior mean for $\boldsymbol{\beta}$ vanishes as the sample size increases; therefore, estimators based on the posterior mean are asymptotically unbiased for the true values of the regression parameters.

Dependent Priors for β_j

As an alternative to the independent priors in (1.11), many authors specify a joint distribution for $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ that allows for *a priori* dependence among the regression coefficients.

Table 1.2: Independent priors for regression coefficients in the variable selection problem.

Independent Priors for β_j			
Author(s)	$p^{(0)}(\beta)$	$p^{(0)}(\beta)$	$p(\boldsymbol{\gamma})$
Mitchell and Beauchamp (1988)	$\mathbb{1}_{\{0\}}(\beta_j)$	$\mathsf{Unif}(-a, a)$	$\prod w_j^{\gamma_j}(1-w_j)^{1-\gamma_j}$
George and McCulloch (1993)	$\mathsf{N}\left(0, \tau_j^2\right)$	$\mathsf{N}\left(0, c_j\tau_j^2\right)$	$\binom{p}{q_{\boldsymbol{\gamma}}}^{-1} w_{q_{\boldsymbol{\gamma}}}$
Geweke (1996)	$\mathbb{1}_{\{0\}}(\beta_j)$	$\mathsf{N}\left(0, \tau_j^2\right)\mathbb{1}(a, b)$	$\prod w_j^{\gamma_j}(1-w_j)^{1-\gamma_j}$
Yuan and Lin (2005)	$\mathbb{1}_{\{0\}}(\beta_j)$	$\mathsf{DoubExp}(\tau)$	$\det(\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{X}_{\boldsymbol{\gamma}})\times$ $p^{q_{\boldsymbol{\gamma}}}(1-p)^{p-q_{\boldsymbol{\gamma}}}$

George and McCulloch (1993) and George and McCulloch (1997) use the prior specified in (1.12) with $\mathbf{R}_{\boldsymbol{\gamma}} = (\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{X}_{\boldsymbol{\gamma}})^{-1}$. George and McCulloch (1997) also specify a slight variation of this prior that allows for analytic solutions because of the conjugate nature of the likelihood and the prior

$$\boldsymbol{\beta}|\boldsymbol{\gamma} \sim \mathsf{N}_p\left(\mathbf{0}, \sigma^2\mathbf{D}_{\boldsymbol{\gamma}}\mathbf{R}_{\boldsymbol{\gamma}}\mathbf{D}_{\boldsymbol{\gamma}}\right), \tag{1.13}$$

where $\mathbf{R}_{\boldsymbol{\gamma}}$ can again be chosen as $(\mathbf{X}^T\mathbf{X})^{-1}$ to mimic the correlation structure of the predictors.

The prior in (1.13) is a variation of Zellner's "g-prior" (Zellner, 1986). Zellner proposed

$$\boldsymbol{\beta}|\sigma^2, g \sim \mathsf{N}_p\left(\mathbf{0}, g\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\right) \tag{1.14}$$

as a prior for Bayesian regression. Because the prior for $\boldsymbol{\beta}$ is conditioned on σ^2, posterior model probabilities have an analytic form. The constant g provides a way to control the uncertainty in the prior relative to the variance of the observations around the mean, and the $(\mathbf{X}^T\mathbf{X})^{-1}$ term provides a prior covariance structure for $\boldsymbol{\beta}$ that mimics the covariance found in the data. When used in variable selection, the g-prior is typically conditioned on $\boldsymbol{\gamma}$ to give

$$\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\sigma^2, g \sim \mathsf{N}_p\left(\mathbf{0}, g\sigma^2(\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{X}_{\boldsymbol{\gamma}})^{-1}\right) \tag{1.15}$$

(see also Marin and Robert, 2007, Chapter 3).

George and Foster (2000) use the g-prior to derive a connection between model selection criteria such as Mallows' $C_{\boldsymbol{\gamma}}$ (Mallows, 1973), AIC (Akaike, 1973), BIC (Schwarz, 1978), and RIC (Foster and George, 1994). They demonstrate that when the g-prior is used in conjunction with $p(\boldsymbol{\gamma}) = w^{q_{\boldsymbol{\gamma}}}(1-w)^{p-q_{\boldsymbol{\gamma}}}$, then the model ordering induced by posterior probabilities is the same as the model ordering induced by model selection criteria, such as those mentioned above. The approach of George and Foster (2000) requires the selection or estimation of tuning parameters g in the prior for $\boldsymbol{\beta}$ and w in the prior for $\boldsymbol{\gamma}$. Rather than select values for these parameters based on *a priori* knowledge or estimate them via a prior a and Bayes theorem, George and Foster (2000) derive empirical estimates of these parameters based on the marginal likelihood of w and g.

Recently, Krishna, Bondell, and Ghosh (2008) have proposed a variation of Zellner's prior. The regression coefficients $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ of the $\boldsymbol{\gamma}$ model are given the prior

$$\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\sigma^2, g, \lambda \sim \mathsf{N}_p\big(\mathbf{0}, g\sigma^2 k(\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{X}_{\boldsymbol{\gamma}})^{\lambda}\big)\,, \tag{1.16}$$

where $k = \frac{\mathrm{tr}[(\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{X}_{\boldsymbol{\gamma}})^{-1}]}{\mathrm{tr}[(\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{X}_{\boldsymbol{\gamma}})^{\lambda}]}$ is scale factor, and the model $\boldsymbol{\gamma}$ is given the prior

$$p(\boldsymbol{\gamma}) = w^{q_{\boldsymbol{\gamma}}}(1-w)^{p-q_{\boldsymbol{\gamma}}}\det(\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{X}_{\boldsymbol{\gamma}})^{\lambda/2}$$

Varying the parameter λ will adjust the prior mass associated with correlated predictors. A value of $\lambda = -1$ (the value for Zellner's prior) penalizes highly correlated predictors. A large positive value for λ places higher prior mass on correlated predictors and increases the probability that correlated predictors will be selected together in the final model. Krishna et al. (2008) use an empirical Bayes approach to select the parameters λ and w.

Liang, Paulo, Molina, Clyde, and Berger (2008) note that choosing a value for g in Zellner's g-prior can be problematic and lead to undesirable behavior of model-selection criteria. Specifically, they note that Zellner's g-prior exhibits two "paradoxes." The first, called Bartlett's paradox (Bartlett, 1957), is that as g increases to ∞ the Bayes factor for comparing model $\boldsymbol{\gamma}$ to the null model (i.e., the model with only an intercept term) goes to zero for any value of the data. Most practitioners choose large values of g in an effort to be "noninformative" (i.e., let the data speak).

Unfortunately, choosing g to be large arbitrarily biases model selection in favor of null model.

Liang et al. (2008) call the second paradox the information paradox. In situations where support for the model $\boldsymbol{\gamma}$ increases without bound (e.g., when the coefficient of determination for $\boldsymbol{\gamma}$ $R^2_{\boldsymbol{\gamma}}$ increases to one), it is reasonable to expect that the Bayes factor of $\boldsymbol{\gamma}$ to the null model should increase without bound. However, this is not the case; the Bayes factor is bounded above by a function of the parameter g.

Liang et al. (2008) recommend using mixtures of g-priors for variable selection. They recommend putting a hyperprior on the parameter g and integrating over this parameter to induce a prior on the regression coefficients. The first prior they recommend is the Cauchy prior of Zellner and Siow (1980), which can be obtained as a scale mixture of normal distributions (Andrews and Mallows, 1974). The second prior they recommend is

$$p(g) = \frac{a-2}{2}(1+g)^{-a/2}\mathbb{1}_{\{0\}}(g)\,,$$

which was used by Strawderman (1971) to estimate multivariate normal means. For values of $a > 2$ this prior for g is proper. Liang et al. call this prior the "hyper-g prior."

Liang et al. (2008) demonstrate under what conditions the Zellner-Siow prior and the hyper-g prior resolve the information paradox. They also demonstrate that the two previous priors are consistent for model selection when the true model is not the null model (the posterior probability of the true model converges to one as n goes to infinity) and that the priors are prediction consistent (the model averaged estimate of a new response value at new values of the predictors converges to the expected value of the response at those predictor values).

Data-Based Priors

Most Bayesian variable-selection methods require the elicitation of several parameter values. Researchers often use variable-selection methods in situations when there is little prior information about parameters, which makes elicitation of parameters difficult. Unfortunately, standard noninformative priors for estimation cannot be

Table 1.3: Dependent priors for regression coefficients in the variable selection problem.

Dependent Priors for β_j		
Author(s)	$p(\boldsymbol{\beta}\|\boldsymbol{\gamma})$	$p(\boldsymbol{\gamma})$
George and McCulloch (1993)	$\mathsf{N}_p(\mathbf{0}, \mathbf{D}_{\boldsymbol{\gamma}}\mathbf{R}_{\boldsymbol{\gamma}}\mathbf{D}_{\boldsymbol{\gamma}})$	$\prod_j \mathsf{Bern}(w_j)$
George and McCulloch (1997)	$\mathsf{N}_p(\mathbf{0}, \sigma^2\mathbf{D}_{\boldsymbol{\gamma}}\mathbf{R}_{\boldsymbol{\gamma}}\mathbf{D}_{\boldsymbol{\gamma}})$	$\prod_j \mathsf{Bern}(w_j)$
George and Foster (2000)	$\mathsf{N}_{q_\gamma}\left(\mathbf{0}, g\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\right)$	$w^{q_\gamma}(1-w)^{p-q_\gamma}$
Cripps, Kohn, and Nott (2006)	$\mathsf{N}_{q_\gamma}\left(\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}, \mathbf{V}_{\boldsymbol{\gamma}}\right)$	$w^{q_\gamma-1}(1-w)^{p-q_\gamma}$
Krishna et al. (2008)	$\mathsf{N}_{q_\gamma}\left(\mathbf{0}, g\sigma^2 k(\mathbf{X}^T\mathbf{X})^{\lambda}\right)$	$q_{\boldsymbol{\gamma}}^{q_\gamma}(1-q_{\boldsymbol{\gamma}})^{p-q_\gamma}\times$ $\det(\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{X}_{\boldsymbol{\gamma}})^{\lambda/2}$

used in model selection situations (Robert, 2007, Section 5.2.5). One approach to this problem is to use the data to construct priors. Although this approach violates the strict subjectivist approach to Bayesian statistics, it has the advantage of being automatic and can be more useful in practical situations where scant prior information is available.

Cripps et al. (2006) derive priors from standard least squares estimates of regression coefficients. They base their model on the hierarchical model structure $p(\mathbf{y}|\boldsymbol{\beta}_{\boldsymbol{\gamma}})p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma})p(\boldsymbol{\gamma})$. Their model can be summarized by

$$\begin{aligned}
(Y_i|\beta_0, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}, \sigma^2) &\sim \mathsf{N}\left(\beta_0/\sqrt{n} + \mathbf{x}_{\boldsymbol{\gamma},i}^T\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2\right) \\
(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma}) &\sim \mathsf{N}_{q_\gamma}\left(\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}^{LS}, c_1\sigma^2(\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{X}_{\boldsymbol{\gamma}})^{-1}\right) \\
\beta_0 &\sim \mathsf{N}\left(\sqrt{n}\hat{\beta}_0^{LS}, c_0\right) \\
(\boldsymbol{\gamma}|\omega) &\sim \omega^{q_\gamma-1}(1-\omega)^{p-q_\gamma} \\
\omega &\sim \mathsf{Beta}(a, b) \\
\sigma^{-2} &\sim \mathsf{Gam}\left(\frac{m}{2} - 1, \frac{m\,\mathrm{SSE}\,(\boldsymbol{\gamma})}{2(n-q_{\boldsymbol{\gamma}})}\right),
\end{aligned}$$

where the columns of $\mathbf{X}_{\boldsymbol{\gamma}}$ have been centered about their sample means, $\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}^{LS} = (\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{X}_{\boldsymbol{\gamma}})^{-1}\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{y}$, $\hat{\beta}_0^{LS}$ is the least squares estimate of β_0, and $\mathrm{SSE}\,(\boldsymbol{\gamma}) = \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}_{\boldsymbol{\gamma}}(\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{X}_{\boldsymbol{\gamma}})^{-1}\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{y}$. The prior for $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ is similar to Zellner's g-prior. The $(\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{X}_{\boldsymbol{\gamma}})^{-1}$ term in the prior variance sets the covariance of the prior equal to that found in the

sample. However, instead of being centered at zero like Zellner's prior, Cripps et al. (2006) center their prior at the least squares estimates for model $\boldsymbol{\gamma}$. The parameter values for σ^2 are chosen such that the mode of the prior distribution is an unbiased estimate of σ^2 for model $\boldsymbol{\gamma}$. Cripps et al. (2006) recommend setting $m = 7$, which allows the prior for σ^2 to be less informative than the likelihood.

The intrinsic priors of Berger and Pericchi (1996b) are a data-based method for constructing prior distributions that have been used in a a variable selection context (see Berger and Pericchi, 1996a; Casella and Moreno, 2006). As noted previously, improper noninformative priors cannot be used for model-selection procedures. However, Berger and Pericchi suggest using a noninformative prior on the smallest subset of the data such that the posterior distribution of the parameters is proper when conditioned on this "minimal sample." The posterior distribution of the model parameters based on the minimal sample can then be used as a proper prior to construct Bayes factors for model comparison. In order to avoid having the Bayes factor depend on any particular minimal sample, Berger and Pericchi suggest averaging Bayes factors over all possible minimal samples, and they call Bayes factors constructed in this manner "intrinsic" Bayes factors. Intrinsic Bayes factors are not Bayes factors in a technical sense, however, it was shown by Berger and Pericchi that there exists prior distributions such that, asymptotically, intrinsic Bayes factors correspond to real Bayes factors with such priors.

1.2 Variable selection in Additive Models and Additive Models with Interactions

As described previously, additive models are a nice balance between fully nonparametric approaches and the restrictive linear models. In this section we review both Bayesian and non-Bayesian approaches to variable selection in additive models.

Smith and Kohn (1996) use a Bayesian variable selection approach to fit additive models. They use a free-knot-spline basis expansion to model each functional component. Since the basis expansion can be written in a linear model form, Smith and

Kohn (1996) then use a variable selection technique to fit the additive model. They use Zellner's g-prior on the spline coefficients. The degree of smoothing for each functional component of the regression function is then controlled by the variable selection on the spline coefficients.

Shively, Kohn, and Wood (1999) also present a Bayesian approach to variable selection in the additive model. They model the j^{th} additive regression component $f_j(\cdot)$ as

$$f_j(x) = \beta_j x + (\tau_j^2)^{1/2} \int_0^\infty W_j(u)\mathrm{d}u, \tag{1.17}$$

where $W_j(\cdot)$ is a Wiener process with $\mathsf{Var}(W(x)) = x$ and $W_j(0) = 0$. This structure for $f_j(\cdot)$ balances a linear and a nonlinear fit. If τ^2 is zero then $f_j(\cdot)$ is a simple linear function. If τ^2 is positive, then the integrated Wiener process allows for flexibility in fitting nonlinear functions. If both β_j and τ_j are zero, then the j^{th} covariate is removed from the regression. Shively et al. (1999) define binary variables $\gamma_j^{(\beta)}$ and $\gamma_j^{(\tau)}$ for each predictor that indicate which component of (1.17) should enter the model. If $\gamma_j^{(\beta)} = 0$, then $\beta_j = 0$, otherwise β_j is nonzero. Binary variables $\gamma_j^{(\tau)}$ are defined similarly for the τ_j.

Shively et al. (1999) construct a data-based prior for model parameters by running a preliminary Gibbs sampler on the full model using non-informative uniform priors for all model parameters. They then use a multivariate normal distribution with mean vector equal to the sample mean vector of the preliminary posterior draws and a covariance matrix equal to the covariance matrix of the preliminary posterior draws multiplied by the sample size as the prior distribution for model parameters. They justify this prior by showing its similarity to Schwarz's model selection criterion (Schwarz, 1978).

Avalos, Grandvalet, and Ambroise (2003) present a generalization of the LASSO to the case of additive models. They base their generalization on a connection between adaptive ridge regression (ARR) and the LASSO for classical linear regression. Grandvalet (1999) showed that the ARR estimator,

$$\underset{\boldsymbol{\beta},\xi_1,\ldots,\xi_p}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^{p} \xi_j \beta_j^2 \tag{1.18}$$

subject to $\sum_{j=1}^{p} 1/\xi_j = p/\lambda$ and $\xi_j > 0$, is equivalent to the LASSO estimator. Based on the relation between the LASSO and ARR estimators, Avalos et al. (2003) propose a generalization of ARR for additive models with cubic splines

$$\underset{\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_j}{\operatorname{argmin}} \left\| \mathbf{y} - \sum_{j=1}^{p} \mathbf{W}_j \boldsymbol{\beta}_j \right\|^2 + \lambda \sum_{j=1}^{p} \xi_j \boldsymbol{\beta}_j^T \Omega_j \boldsymbol{\beta}_j \tag{1.19}$$

subject to $\sum_{j=1}^{p} 1/\xi_j = p/\lambda$, where $\mathbf{W}_j$ is a B-spline basis expansion for the j^{th} predictor and Ω_j is the penalization matrix corresponding to the penalty on the second derivative of $f_j(\cdot)$. (See also the COSSO, which was developed independently by Lin and Zhang, 2006.)

Avalos et al. (2003) note that the solution to (1.19) has the ability to restrict certain predictors to a linear effect in the regression equation, but the solution does not have the ability to remove predictors from the regression equation. To adjust for this, Avalos et al. (2003) note that the smoother matrix of a cubic smoothing spline $\mathbf{S}$ can be partitioned such that $\mathbf{S} = \mathbf{L} + \tilde{\mathbf{W}}$ into a component associated with a linear fit $\mathbf{L}$ and a nonlinear fit $\tilde{\mathbf{W}}$.

Equation (1.19) can then be altered by separating out the linear and nonlinear components of each $f_j(\cdot)$ as in

$$\underset{\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_p}{\operatorname{argmin}} \left\| \mathbf{y} - \sum_{j=1}^{p} \mathbf{x}_j \alpha_j + \sum_{j=1}^{p} \tilde{\mathbf{W}}_j \tilde{\boldsymbol{\beta}}_j \right\|^2 + \sum_{j=1}^{p} \mu_j \alpha_j^2 + \sum_{j=1}^{p} \xi_j \tilde{\boldsymbol{\beta}}_j^T \Omega_j \tilde{\boldsymbol{\beta}}_j \tag{1.20}$$

subject to $\sum_{j=1}^{p} 1/\mu_j = p/\mu$ and $\sum_{j=1}^{p} 1/\xi_j = p/\lambda$.

Avalos et al. (2003) present a backfitting algorithm (Hastie and Tibshirani, 1990) for finding the solution to (1.20) where the parameters ξ_j and μ_j are set to

$$\xi_j = \lambda \frac{\sum_{k=1}^{p} \sqrt{\tilde{\boldsymbol{\beta}}_k^T \Omega_k \tilde{\boldsymbol{\beta}}_k}}{p \sqrt{\tilde{\boldsymbol{\beta}}_j^T \Omega_j \tilde{\boldsymbol{\beta}}_j}}$$

$$\mu_j = \mu \frac{\sum_{j=1}^{p} |\alpha_j|}{p|\alpha_j|}$$

and the values of μ and λ are chosen by generalized cross-validation (Craven and Wahba, 1979).

Cantoni, Flemming, and Ronchetti (2006) present a generalization of the nonnegative garrote to the case of additive models. Their estimators of $f_j(\cdot)$ are $\hat{c}_j \hat{f}_j(\cdot; \lambda_j)$, where the $\hat{c}_j$ are the solutions to

$$\underset{c_1,\ldots,c_p}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \alpha - \sum_{j=1}^{p} c_j \hat{f}_j(x_{ij}; \lambda_j))^2$$

subject to $c_j \geq 0$ and $\sum_{j=1}^{p} c_j \leq s$, $\hat{f}_j(\cdot; \lambda_j)$ are initial estimates of $f_j(\cdot)$, and λ_j are tuning parameters used to fit the $\hat{f}_j(\cdot; \lambda_j)$. Cantoni et al. (2006). note that the original nonnegative garrote is a special case of their formulation where $\hat{f}_j(x; \lambda_j) = \hat{\beta}_j x$ with $\hat{\beta}_j$ the typical least-squares estimate. Thus, their solution involves two steps—an initial fit of the additive model to obtain $\hat{f}_j(\cdot; \lambda_j)$ followed by a nonnegative garrote to perform variable selection.

Cantoni et al. (2006) give several different methods for choosing the values of the tuning parameters λ_j and s. They run several simulations to show that their approach is comparable with the COSSO (to be described shortly).

Yuan and Lin (2006) give several different methods for performing variable selection on grouped variables. We discuss their group LASSO method here. This method can be used for factors in an ANOVA setting (where the grouped variables are sets of dummy variables for a particular predictor) or for basis expansions in additive models. The group LASSO is the solution to

$$\underset{\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_p}{\operatorname{argmin}} \left\| \mathbf{y} - \sum_{j=1}^{p} \mathbf{W}_j \boldsymbol{\beta}_j \right\|^2 + \lambda \sum_{j=1}^{p} \|\boldsymbol{\beta}_j\|_{\mathbf{K}_j} \tag{1.21}$$

where $\mathbf{W}_j$ is a matrix of dummy variables or a basis expansion for the j^{th} covariate and $\|\boldsymbol{\beta}\|_{\mathbf{K}} = \sqrt{\boldsymbol{\beta}^T \mathbf{K} \boldsymbol{\beta}}$. The matrices $\mathbf{K}_j$ can be chosen to be the identity matrix, a diagonal matrix with the number of columns of $\mathbf{W}_j$ along the diagonal, a penalization matrix as in Avalos et al. (2003), or some other matrix more appropriate for a given application.

Yuan and Lin (2006) provide an iterative algorithm for fitting the group LASSO that is derived from the Karush-Kuhn-Tucker conditions (Kuhn and Tucker, 1951).

Lin and Zhang (2006) present a method for variable selection in nonparametric smoothing spline ANOVA models. They call their method the Component Selec-

tion and Smoothing Operator (COSSO). This method is another penalized regression approach and uses the sum of function norms as a penalty. In other words, we let $\mathcal{F} = \{1\} \oplus \bigoplus_{\eta=1}^{m} \mathcal{F}^{\eta}$, where the function spaces $\mathcal{F}^{\eta}$ correspond to spaces of main effect functions, two-way interaction functions, and so on, according to the functional ANOVA decomposition. The COSSO is the solution to

$$\underset{f \in \mathcal{F}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} [y_i - f(x_i)]^2 + \tau_n^2 \sum_{\eta=1}^{m} \|P^{\eta} f\|, \tag{1.22}$$

where $P^{\ell} f$ is the projection of f onto the ℓ^{th} component of $\bigoplus_{\eta=1}^{m} \mathcal{F}^{\eta}$ and τ_n^2 is a smoothing parameter. Lin and Zhang (2006) show that when the component functions are restricted to be linear the COSSO reduces to the LASSO. Lin and Zhang (2006) derive an equivalent formulation of the COSSO that is easier for computation and choose the value of the tuning parameter by cross validation or generalized cross validation.

1.3 Variable Selection in Nonparametric Regression

In the previous sections, we reviewed variable selection procedures which assumed some sort of structure on the regression function $f(\cdot)$—linear, additive, SS-ANOVA, etc. In this section, we review a few variable selection methods that do not make the assumptions of the previous methods.

Friedman (1991) develops the Multivariate Adaptive Regression Splines (MARS) procedure that uses a forward stepwise regression procedure to construct a regression function from "reflected pairs" of basis functions (see also, Hastie et al., 2001). A reflected pair is two functions of the form $[x - \xi]_+$ and $[\xi - x]_+$, where ξ is a knot and $[z] = \max(z, 0)$. In the MARS procedure, a collection $\mathcal{C}$ of n reflected pairs is constructed for each of the p predictors, where each of the n reflected pairs use observed values of the predictors as the knots. Thus, there are $2np$ basis functions in this collection. The MARS procedure proceeds via a forward stepwise procedure that, at the j^{th} step, selects a $g_j(\cdot)$ from among the $2np$ functions in $\mathcal{C}$ or a product

of any number of the functions in $\mathcal{C}$ to include in

$$f(\mathbf{x}) = \beta_0 + \sum_{j=1}^{M} \beta_j B_j(\mathbf{x}).$$

MARS typically results in overfitting, so a backward deletion procedure is often used to remove terms from the regression equation.

Denison, Mallick, and Smith (1998b) develop a Bayesian version of the MARS procedure. They parametrize the MARS regression function as

$$f(\mathbf{x}) = \beta_0 + \sum_{j=1}^{\omega} \beta_j B_j(\mathbf{x})$$

where

$$B_j(\mathbf{x}) = \sum_{k=1}^{\kappa_j} [\gamma_{jk}(x_{\nu_{jk}} - \xi_{jk})]_+,$$

ω is the total number of basis functions in the regression equation, γ_{jk} is a parameter that takes on values in $\{-1, 1\}$, ν_{jk} is a parameter that takes on values in $\{1, \ldots, p\}$, ξ_{jk} is a knot that can take on values in $x_{1,\nu_{jk}}, \ldots, x_{n,\nu_{jk}}$, and κ_j controls the order of interaction for the $B_j(\cdot)$ and can take values in $\{1, \ldots, M\}$ (where M is a maximum allowable order, usually 2). Priors are assigned to all parameters in the regression function, including the parameter ω that gets a truncated Poisson distribution. Because ω has a distribution the parameter space is of variable dimension and Denison, Mallick, and Smith (1998a) develop a reversible-jump MCMC algorithm (Green, 1995) to sample from the posterior distribution of the Bayesian MARS model.

Linkletter, Bingham, Hengartner, and Higdon (2006) use a Gaussian process to develop a nonparametric variable-selection procedure. The regression function $f(\cdot)$ is given a Gaussian process prior (Rasmussen and Williams, 2005), where, by definition of a Gaussian process, $f(\mathbf{x})$ and $f(\mathbf{x}^\star)$ have a bivariate normal distribution with mean vector $(m(\mathbf{x}), m(\mathbf{x}^\star))$ and covariance $k(\mathbf{x}, \mathbf{x}^\star)$. For their variable selection procedure, Linkletter et al. (2006) let the mean function be identically zero ($m(\mathbf{x}) \equiv 0$) and the covariance function be a function of covariates

$$k(\mathbf{x}, \mathbf{x}^\star) = \frac{1}{\lambda} \prod_{j=1}^{p} \rho_j^{2^{\alpha_j} |x_j - x_j^\star|^{\alpha_j}}. \tag{1.23}$$

Under this covariance function, if the parameter ρ_j is close to one (or equal to one), then the j^{th} covariate has very little effect on the Gaussian-process regression function. Thus, Linkletter et al. (2006) construct a prior for each ρ_j that is a mixture of a point mass at 1 and a uniform distribution on $(0, 1)$. To determine which predictors have real effects and which predictors are spurious, Linkletter et al. include a false predictor in their model (i.e. a predictor "randomly sampled from the design space of $\mathbf{X}$"(Linkletter et al., 2006, page 481) and, therefore, is known to have no effect on the response). The posterior distribution of the coefficient ρ_F on the false predictor can be used to gauge which of the real predictors have a real effect on the response. Linkletter et al. recommend running their MCMC analysis several times and recording the median of the posterior draws of ρ_F. A distribution for the posterior median of the false predictor can be approximated with these values, and the posterior medians of the real predictors can be compared to this distribution.

Reich, Storlie, and Bondell (2008) combine Bayesian variable selection with functional-ANOVA models to perform nonparametric variable selection. As with other functional-ANOVA approaches, they decompose the function space of the regression function $f(\cdot)$ into orthogonal subspaces that correspond to main effects of the p predictors, two-interactions, and a "left-over" space that corresponds to all other higher order interactions. Using results from Wahba (1990), they put Gaussian process priors on all components of the regression function. Each regression component is also given an indicator variable that determines if the component function is included in the model. They sample from the posterior distribution of the parameters using an MCMC method.

1.4 Stick-Breaking Priors

In this section we review some of the literature on Bayesian clustering. In particular, we focus on stick-breaking priors, as these priors will be used later in this dissertation.

A stick-breaking prior is a probability measure over a space of discrete probability measures (Ishwaran and James, 2001, contains a good review of stick-breaking priors).

A random measure $Q(\cdot)$ from a stick-breaking prior can be constructed via

$$Q(\cdot) = \sum_{j=1}^{N} w_j \delta_{Z_j}(\cdot), \tag{1.24}$$

where $N \in \{1, 2, \ldots\} \cup \{\infty\}$, the w_j are constructed via a stick breaking process

$$\begin{aligned} w_1 &= V_1 \\ w_j &= (1 - V_1)(1 - V_2) \cdots (1 - V_j) V_j \qquad \text{for } j \geq 2 \\ V_j &\sim \mathsf{Beta}(a_j, b_j)\,, \end{aligned}$$

the Z_j are random draws from a base measure $Q_0(\cdot)$, $\mathbf{a} = (a_1, \ldots, a_N)$ and $\mathbf{b} = (b_1, \ldots, b_N)$ are sequences of parameters for a Beta distribution, and $\delta_Z(A)$ is a probability measure that equals one when the set A contains the point Z and equals zero otherwise.

The most well known stick-breaking prior is the Dirichlet process. The Dirichlet process is defined as a random probability measure $Q(\cdot)$ on $(\Omega, \mathcal{B})$ such that for any partition $B_1, \ldots, B_k$ of Ω with $B_j \in \mathcal{B}$ the random quantity $Q(B_1), \ldots, Q(B_k)$ follows a Dirichlet distribution with parameters $\alpha Q_0(B_1), \ldots, \alpha Q_0(B_k)$, where α and $Q_0(\cdot)$ are the parameters of the Dirichlet process.

The parameter $Q_0(\cdot)$ is the base distribution or the mean of the Dirichlet process because, by the definition of the Dirichlet process, $Q(B) \sim \mathsf{Bern}\left(\frac{\alpha Q_0(B)}{\alpha(Q_0(B)+\alpha Q_0(\Omega - B)}\right)$ and therefore $\mathsf{E}[Q(B)] = \frac{\alpha Q_0(B)}{\alpha Q_0(B)+\alpha Q_0(\Omega - B)} = \frac{Q_0(B)}{(Q_0(B)+Q_0(\Omega - B)} = Q_0(B)$.

The parameter α is known as the precision parameter of the Dirichlet process because this parameter controls how "close" the random draws from the Dirichlet process are to the base distribution. This can be seen by examination of the Pólya urn scheme for generating draws from a random distribution from a Dirichlet process. Blackwell and MacQueen (1973) showed that random draws $Y_1, \ldots, Y_n$ may be obtained from a random distribution with a Dirichlet process with parameters α and $Q_0(\cdot)$ by the following scheme:

1. For $j = 1$, generate Y_j from $Q_0(\cdot)$.

2. For $j > 1$, generate Y_j from $\frac{\alpha}{\alpha+j+1}Q_0(\cdot) + \frac{1}{\alpha+j+1}\sum_{k=1}^{j-1}\delta_{Y_j}(\cdot)$

According the above scheme, for j greater than 1, a new draw is generated from the base distribution with probability $\frac{\alpha}{\alpha+j+1}$ or is set equal to one of the previous draws Y_k (for $1 \leq k \leq j-1$) with probability $\frac{1}{\alpha+j+1}$. Therefore, as $\alpha \to \infty$ (i.e. as the precision increases) the probability that a new draw Y_j will be taken from the base distribution goes to one.

Sethuraman (1994) shows that the Dirichlet process can be represented as a stick breaking prior. When $N = \infty$ and $a_j \equiv 1$ and $b_j \equiv \alpha$, the stick-breaking prior in (1.24) is a Dirichlet process.

Another special case of the stick-breaking prior is the Pitman-Yor process (Pitman and Yor, 1997). The Pitman-Yor process can be constructed using $N = \infty$, $a_j \equiv 1-a$, and $b_j = b + ja$ (Pitman, 1995, 1996). The Dirichlet process is immediately seen as a special case of the Pitman-Yor process with $a = 0$ and $b = \alpha$.

Stick breaking priors are suitable in clustering applications because draws $Y_1, \ldots, Y_n$ from a distribution which has been generated via a stick-breaking prior have positive probability of $Y_j = Y_{j'}$ for $j \neq j'$. Thus, the draws $Y_1, \ldots, Y_n$ will form clusters with all Y_j in a cluster equal.

1.5 Shape-Restricted Inference

In this section we review the literature on shape-restricted regression. The literature begins with Hildreth (1954) who considers the typical univariate regression situation (or multivariate situation with all independent variables except one held constant), where observations Y_{ij} are a function of an unknown mean function covariates x_i and an error term

$$Y_{ij} = m(x_i) + \epsilon_{ij},$$

where $i = 1, \ldots, p$ (with p the number of distinct values of the x's), and $j = 1, \ldots, n_i$. Hildreth derives maximum likelihood estimates for the ordinates $m(x_i)$ under restrictions of concavity ($\frac{m(x_{i+1})-m(x_i)}{x_{i+1}-x_i} \geq \frac{m(x_{i+2})-m(x_{i+1})}{x_{i+2}-x_{i+1}}$) by converting the problem into one of quadratic programming and using the famous conditions of Kuhn and Tucker

(1951). This procedure does not, however, give estimates of the function $m(\cdot)$ that satisfy concavity for values of the predictor variable not included in the original data set.

Brunk (1955) derives maximum likelihood estimates for monotone parameters in the exponential family of distributions. Observations Y_{ij} are assumed to come from a distribution $f(y|\theta_i)$, where $i = 1, \ldots, p$ indexes the p populations from which the Y_{ij} are drawn, $j = 1, \ldots, n_i$ indexes the independent draws from the i^{th} population, $\mathbf{x}_i = (x_{i1}, \ldots, x_{im})$ is an m-dimensional vector of values of the independent variables for the i^{th} population, and θ_i is a scalar parameter associated with the values of the independent variable for the i^{th} population. Brunk (1955) gives the example where θ_i is the probability that an observer fails to see an object that is at a certain distance and angle from the observer. In the example, $\mathbf{x}_i = (x_{i1}, x_{i2})$, x_{i1} is the distance from the observer and x_{i2} is the angle ($0 \leq x_{i2} \leq \pi/2$). A reasonable assumption is that the probability an observer will fail to detect an object is nondecreasing in both the distance from the observer and the angle of the object to the observer. As with the method of Hildreth (1954), this method provides estimates of θ_i only for values of $\mathbf{x}$ actually observed in the data set.

Barlow, Bartholomew, Bremner, and Brunk (1972) derive the pool-adjacent-violators (PAV) algorithm as a solution to the problem of isotonic regression (see also Brunk, 1958; Ayer, Brunk, Ewing, Reid, and Silverman, 1955). The isotonic regression problem is a special case of the problem considered by Brunk (1955). If we let Y_i (for $i = 1, \ldots, n$) be values of a response variable, then the problem of isotonic regression is to find values $\hat{m}_i$ that minimize $\sum_{i=1}^{n}(y_i - \hat{m}_i)^2$ subject to a monotonicity constraint $\hat{m}_{i+1} \geq \hat{m}_i$. Friedman and Tibshirani (1984) provide an intuitive description of this algorithm. Start by considering a scatterplot of the index i versus the values of the response variable y_i and by setting $\hat{m}_i = y_i$ for all i. Now compare $\hat{m}_1$ and $\hat{m}_2$. If $\hat{m}_2 \geq \hat{m}_1$, then move on and compare $\hat{m}_3$ to $\hat{m}_2$. Continue in this manner until $\hat{m}_n$ is reached or until $\hat{m}_i < \hat{m}_{i-1}$. If $\hat{m}_i < \hat{m}_{i-1}$, then set $\hat{m}_i = \hat{m}_{i-1} = \frac{y_i + y_{i-1}}{2}$. Now check that $\hat{m}_{i-1} \geq \hat{m}_{i-2}$. If not, then set $\hat{m}_i = \hat{m}_{i-1} = \hat{m}_{i-2} = \frac{y_i + y_{i-1} + y_{i-2}}{3}$. Continue in this manner until monotonicity has been restored in all of the previous $\hat{m}$ values.

Although the PAV algorithm produces the maximum likelihood estimates under monotonicity constraints, there are several unsatisfying aspects of these estimates. The first is that if the response variables already satisfy the monotonicity constraint, then the PAV estimates $\hat{m}_i$ are just the y_i—that is the isotonic regression fit exactly interpolates the data. Also, if each y_i has an associated covariate x_i and the PAV fit is used to supply an estimate of the unknown regression function for the data, the resulting regression function estimate is a step function. In many applications, it is safe to assume that an unknown regression function is smooth. Thus, the "solution to the isotonic-regression problem is not the solution to the problem of monotone smoothing because the solution sequence $\hat{m}_1, \ldots, \hat{m}_n$ is not necessarily smooth" (Friedman and Tibshirani, 1984, p. 244).

Friedman and Tibshirani (1984) develop a monotonic scatterplot smoother by combining a "running-mean" smoother with the PAV algorithm. Their method proceeds by deriving fits to the data (Y_i, x_i) with a running mean smoother

$$\hat{s}_k(x_i) = \frac{\text{ave}(y_i)_{i \in A}}{|A|},$$

where $A = \{j \in \{1, \ldots, n\} : j \leq \min(j + k, n) \text{ and } j \geq \max(1, j - k)\}$ and $|A|$ is the cardinality of the set A. Loosely, this smoother is the average of the current y value and the k neighbors to the left and right of the current y value. To obtain a monotone smoother, Friedman and Tibshirani (1984) use the PAV algorithm on the $\hat{s}_k(x_i)$ values.

The approach of Friedman and Tibshirani (1984) starts by smoothing the data and finishes by applying the PAV algorithm to "monotonize" the fit. Mukerjee (1988) reverses the order of these procedures by first computing the isotonic regression fits with the PAV algorithm and then applying a kernel smoother to the fits. Mukerjee (1988) shows that this new estimator has nice statistical properties such as consistency of the estimator and its derivative at a point in the domain of the function.

Ramsay (1988) restricts the M-spline basis to derive a regression-spline basis—called I-splines—for monotonic regression functions. For observed data (y_i, x_i) (for $i = 1, \ldots, n$), regression splines are piecewise polynomials joined at "knots" (designated positions along the x-axis) $\xi_1 \leq \ldots \leq \xi_n$. The polynomial pieces are defined

by their order k or their degree $k-1$ and the number of smoothness constraints they must attain at each of the knots. For example, a spline of degree four (or order 3) with 3 smoothness restrictions has cubic polynomial pieces between each knot where the values of adjacent polynomials and the values of their first and second derivatives must be equal at their common knot.

Regression splines are fitted using a linear combination of basis functions that spans the space of splines of the desired order and number of smoothness restrictions. The set of basis functions used by Ramsay (1988) are known as M-splines. Each function in the M-spline basis is nonnegative, which suggests that each M-spline basis function could be integrated from $\min(x_1, \ldots, x_n)$ to t to form a new set of basis functions for monotone regression splines. When the coefficients on each of these I-spline functions are nonnegative, the set of I-splines forms a basis for monotonic regression splines. Ramsay notes that the isotonic regression estimator of Barlow et al. (1972) is a special case of the I-spline approach where the order is $k = 1$. Ramsay and Abrahamowicz (1989) generalize the I-spline approach to binomial regression.

The approach of Brunk (1955) includes the fitting of isotonic regression functions of more than one predictor. However, as is the case with fully nonparametric regression, interpreting the individual effects of each of the predictors is difficult. Bacchetti (1989) presents an additive isotonic regression model where the regression function is of the form

$$f(\mathbf{x}_i) = h\left(\sum_{j=1}^{p} f_j(x_{ij})\right)$$

where $h(\cdot)$ is a link function that allows this approach to be used in generalized linear models. Bacchetti defines the cyclic, pool-adjacent-violators algorithm for fitting additive isotonic regression models. The algorithm borrows from the alternating-conditional-expectations algorithm of Breiman and Friedman (1985), the local-scoring algorithm of Hastie and Tibshirani (1986), and the pool-adjacent-violators algorithm of Ayer et al. (1955) and Barlow et al. (1972). The algorithm proceeds by cycling through each predictor variable $\mathbf{x}_j$ to find the best fits $\hat{f}(\mathbf{x}_j) = (\hat{f}(x_{1j}), \ldots, \hat{f}(x_{nj}))$ given the current fits of the other variables and subject to the monotonicity constraints. Because this method relies on the PAV algorithm, this method cannot

escape some of the drawbacks of the PAV algorithm like the step-function behavior of the final fit. However, in some situations, the step-function behavior of the fits can be seen as a strength because it can provide useful partitions of the observations into groups.

Geyer (1991) notes that estimating the regression-function ordinates is simply a constrained optimization problem—a problem which has been studied extensively in the field of optimization theory and has reasonable solutions (see, for example, Nocedal and Wright, 2006). Geyer uses constrained optimization methods to compute parameter estimates under shape-restrictions and the parametric bootstrap (Efron, 1979) to obtain likelihood ratio tests of various hypotheses.

Mammen (1991) motivates the use of regression splines in shape-restricted regression by proving that for any function in a large class of functions with qualitative smoothness restrictions (e.g., convexity or monotonicity), there exists a regression spline such that the values of the regression spline at the design points of the data are equal to the value of the function at the design points of the data. Mammen (1991) presents an algorithm for estimation of the regression spline and derives a convergence rate to the true regression function.

Lavine and Mockus (1995) develop a nonparametric Bayesian approach to the isotonic regression problem using the Dirichlet process. The Dirichlet process is a prior distribution on a space of distribution functions (or, equivalently, probability measures) (Ferguson, 1973), and distribution functions are nondecreasing. Thus, in order to put a prior distribution on the space of all nondecreasing regression functions, Lavine and Mockus (1995) assume that only x-values in a particular range (say (a, b)) are of interest. They then transform the nondecreasing regression function to the interval $(0, 1)$ by

$$m_z(x) = \frac{m(x) - m(a)}{m(b) - m(a)}.$$

Of course, in most applied problems, the true values of the regression function at endpoints a and b are not known, so Lavine and Mockus (1995) put a prior distribution on the values $m(a)$ and $m(b)$ and then condition the Dirichlet process prior on $m(\cdot)$ on $m(a)$ and $m(b)$. As noted by Lavine and Mockus (1995), the use of a Dirichlet

process has some drawbacks. The most notable is that probability measures from a Dirichlet process are almost surely discrete. Thus, if the it is believed that the true regression function is smooth, the Dirichlet process may not be an appropriate prior distribution.

Schell and Singh (1997) take advantage of the fact that the isotonic regression estimate of Barlow et al. (1972) is a step function. In some applications, it may be useful to think of observations as coming from several groups where the values of a response of interest are the same for all values of the independent variable in a certain range. Additionally, it may be appropriate to assume that value of the regression function with respect to the independent variable is monotonic. Thus, an isotonic regression estimate that results in a step function seems a natural choice to fit data for which these assumptions hold. Schell and Singh (1997) take this approach one step further by developing a method for combining groups defined by the isotonic regression fit. They propose a backward-elimination-type scheme to combine groups from the isotonic regression fit that do not differ significantly from each other. That is, Schell and Singh (1997) first compute the isotonic regression estimate and then compare adjacent groups defined by this fit. If the groups do not differ significantly according to a test statistic that they derive, then Schell and Singh (1997) combine the group into one and use a pooled estimate of the isotonic regression values for the new pooled groups.

Similar to Ramsay (1988), He and Shi (1998) use splines to compute a monotonic regression estimate. He and Shi (1998), however, use B-splines and find linear restrictions on the B-spline coefficients that guarantee a monotonic estimate of the regression function. More precisely, if $B_1(\cdot), \ldots, B_N(\cdot)$ are normalized B-splines of degree two, then the linear restrictions

$$B'(\xi_j)^T \boldsymbol{\alpha} \geq 0 \qquad \text{for } j = 1, \ldots, n_k,$$

where n_k is the number of knots and $\boldsymbol{\alpha}$ is the vector of spline coefficients, ensure that the resulting spline fit is monotonic. He and Shi (1998) use an L_1 objective function that, when combined with the linear restrictions above, forms a constrained optimization problem that can be solved with standard linear programming methods.

To select the appropriate number of knots, He and Shi select an initial number of knots (placed at quantiles of the x values in the data) by choosing the number of knots that gives the smallest local minimum of an AIC-type criterion and then systematically drop knots from the model to see if the fit improves (as measured by their AIC-type criterion). They stop when removing any knot does not result in improved fit.

Ramsay (1998) shows that the functions which satisfy the differential equation $\frac{\partial^2 f(x)}{\partial x^2} = w(x)\frac{\partial f(x)}{\partial x}$ are strictly monotone. He solves this equation and provides an explicit formula for such functions

$$f(x) = \beta_0 + \beta_1 D^{-1}[\exp(D^{-1}w(x))],$$

where $D^{-1}g(x) = \int_0^x g(t)\mathrm{d}t$ is Ramsay's notation for the partial integration operator and the function $w(x)$ is a Lebesgue square integrable function. The advantage of the above representation is that the problem of estimating a monotone function is now a problem of estimating an unconstrained function $w(x)$. Ramsay estimates this function with a suitable basis expansion and by minimizing the criterion

$$n^{-1}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 D^{-1}[\exp(D^{-1}\mathbf{b}^T\psi(x))])^2 + \lambda\int_0^1 (\mathbf{b}^T\psi(t))^2 dt,$$

where λ is a tuning parameter that is chosen by cross-validation and $\mathbf{b}^T\psi(x)$ is a basis expansion. The introduction of the basis expansion involves a second tuning parameter—the number of basis functions. Selection of this parameter was not addressed and remains a drawback of this approach.

Hall and Huang (2001) produce a monotone regression fit by starting with a kernel regression fit and constraining it to be monotone. A kernel regression estimator is a special case of a linear smoother, which can be written in the following form

$$\hat{m}(x) = \sum_{i=1}^{n} q_i(x)y_i,$$

where $q_i(\cdot)$ are weights that depend on a kernel function. The approach of Hall and Huang (2001) is to augment the terms in the sum above with values $\mathbf{p} = p_1, \ldots, p_n$ where $p_i \geq 0$ for all $i = 1, \ldots, n$ and $\sum_{i=1}^{n} p_i = 1$ such that

$$\hat{m}(x) = \sum_{i=1}^{n} p_i q_i(x)y_i.$$

A monotone estimate can be obtained by finding values of $\mathbf{p}$ that minimize a suitably defined distance between $\mathbf{p}$ and the discrete uniform distribution $p_i = 1/n$ subject to the constraint $\hat{m}'(x) \geq 0$. The optimization can be performed using standard quadratic programming techniques.

Mammen, Marron, Turlach, and Wand (2001) put the constrained regression problem (and more generally, the unconstrained regression smoothing problem) in the context of finding appropriate projections. They begin by defining a vector space $\mathcal{V}$ of vectors of functions that includes the space of all possible data vectors $\mathcal{V}_Y$ (redefined as a vector of constant functions $(f_1(\cdot) \equiv Y_i, \ldots, f_n(\cdot) \equiv Y_n)$ and denoted as $\underset{\rightarrow}{Y}$), the space of all functions which are candidate smooth functions $\mathcal{V}_s$ (redefined as a vector of the same function $(f_1(\cdot) \equiv m(\cdot), \ldots, f_n(\cdot) \equiv m(\cdot))$ and denoted as $\underset{\rightarrow}{m}$), and the space of all candidate smooth functions with the appropriate constraint $\mathcal{V}_{s,c}$, e.g. monotonicity. Note that $\mathcal{V}_{s,c} \subset \mathcal{V}_s$. With an appropriate norm, the solution to the constrained regression problem is the projection of the data vector into the space of constrained functions

$$\hat{m}_c = \underset{\underset{\rightarrow}{m} \in \mathcal{V}_{s,c}}{\operatorname{argmin}} \left\| \underset{\rightarrow}{Y} - \underset{\rightarrow}{m} \right\|.$$

The term $\left\| \underset{\rightarrow}{Y} - \underset{\rightarrow}{m} \right\|$ can be decomposed into

$$\left\| \underset{\rightarrow}{Y} - \underset{\rightarrow}{m} \right\| = \left\| \underset{\rightarrow}{Y} - \hat{m} \right\| + \left\| \hat{m} - \underset{\rightarrow}{m} \right\|,$$

where $\hat{m}$ is the unconstrained smoother (i.e., $\hat{m} = \operatorname{argmin}_{\underset{\rightarrow}{m} \in \mathcal{V}_s} \left\| \underset{\rightarrow}{Y} - \underset{\rightarrow}{m} \right\|$). The decomposition above suggests fitting a constrained regression estimate by first fitting a smooth function and then finding the projection of that smooth function in the constrained space. This approach has advantages in that the unconstrained smoother need only be altered in regions where the constraint is not met. Thus, the constrained smoother inherits the good theoretical properties of the unconstrained smoother in those areas where the constraint is met.

Holmes and Heard (2003) present a Bayesian, piecewise-constant model for the unknown regression function. The number of constant pieces and the location of these pieces (the knots) are considered random and given a prior distribution. Holmes and Heard (2003) then use a reversible-jump Markov chain Monte Carlo (RJMCMC)

procedure to sample from the posterior distribution of the regression function. The procedure itself does not use any formal monotonicity constraints. To obtain a monotonic regression fit, Holmes and Heard (2003) discard all posterior simulations that do not satisfy the monotonicity constraint. This Bayesian procedure a procedure which could be very inefficient. Also, although the simple form of the piecewise regression functions allow for some computational efficiency, in most practical problems, prior beliefs about the true regression function make a piecewise constant prior inappropriate.

Neelon and Dunson (2004) use a piecewise-linear model for the unknown regression function. For a given set of knots $\xi_0, \ldots, \xi_k$, they model the unknown regression function as

$$f(x) = \beta_0 + \sum_{j=1}^{k} \beta_j w_j(x),$$

where $w_j(x) = I(x \geq \xi_{j-1})(\min(x, \xi_j) - \xi_{j-1})$. Prior distributions on $\beta_1, \ldots, \beta_k$ are induced by the formula $\beta_j = I(\alpha_j \geq \delta)\alpha_j$, where δ is a small, positive number and the α_j are latent variables. The latent variables are given an autoregressive prior

$$\begin{aligned} \beta_1 &\sim \mathsf{N}(m, v) \\ \alpha_j &\sim \mathsf{N}\left(\alpha_{j-1}, \lambda^{-1}\right) \qquad \text{for } j = 2, \ldots, k, \end{aligned}$$

where m and v are chosen to be the researchers' best guess as to the average slope of the regression function and the confidence in this guess, respectively. The model specification is completed by assigning gamma priors to δ and λ. Neelon and Dunson (2004) provide an MCMC algorithm to sample from the posterior distribution of their model. The method of Neelon and Dunson (2004) is a simple generalization of monotonic regression procedures which fit piecewise-constant functions to the data (e.g., Holmes and Heard, 2003; Brunk, 1958). However, the reliance of piecewise linear functions to model the unknown regression function may not be appropriate in many practical situations where the researcher believes the unknown regression function should satisfy certain differentiability conditions.

Dunson (2005) uses a reparametrization of the regression function to impose isotonicity of the regression function at the x-values in the data. Dunson (2005) considers

the model

$$Y_i \sim \mathsf{Poi}(\xi_i \exp\left\{m(x_i) + \mathbf{z}_i^T \boldsymbol{\beta}\right\})$$

where the ξ_i are given a Dirichlet process prior to lessen the dependence on the assumption of a Poisson distribution for the data, $\mathbf{z}_i$ and $\boldsymbol{\beta}$ are other covariates and their regression coefficients, and $m(\cdot)$ is a monotonic regression function. To impose isotonicity on $m(\cdot)$, Dunson (2005) applies the following transformation. Assume that $x_1, \ldots, x_n$ are arranged such that $x_1 \leq \ldots \leq x_n$, then the mean of the Poisson distribution above can be written

$$\begin{aligned}\xi_i \exp\{m(x_i)\} \exp\{\mathbf{z}_i^T \boldsymbol{\beta}\} &= \xi_i \exp\{\textstyle\prod_{j=1}^{i} \frac{g(x_j)}{g(x_{j-1})}\} \exp\{\mathbf{z}_i^T \boldsymbol{\beta}\} \\ &= \xi_i \exp\{\textstyle\prod_{j=1}^{i} \gamma_j\} \exp\{\mathbf{z}_i^T \boldsymbol{\beta}\},\end{aligned}$$

where $g(x) = \exp\{f(x)\}$ and $g(x_0) = 1$. Monotonicity is imposed on $m(\cdot)$ at values $x_1, \ldots, x_n$ by assigning priors to the γ_j with support $[1, \infty)$. Dunson (2005) uses a mixture prior with a point mass and one and gamma distribution truncated to the interval $(1, \infty)$. Flat portions of $m(\cdot)$ are given positive probability because of the point mass at one in the prior distribution. The main draw back to this approach is that values of $m(\cdot)$ are only estimated at the data values $x_1, \ldots, x_n$ and so there is no straightforward way to extrapolate to values of $m(\cdot)$ at points other than those in the data.

Dette, Neumeyer, and Pilz (2006) combine an unrestricted nonparametric regression function estimate and a density function estimate to obtain a strictly monotone regression function estimate. They motivate their method by considering a random sample of n uniform variates $U_1, \ldots, U_n$. The density function of a monotone transformation of these variables $V_i = m(U_i)$ can be written as $(m^{-1})'(v) \mathbb{1}_{(m(0), m(1))} (m^{-1}(v))$ and can be estimated by

$$\frac{1}{N h_d} \sum_{i=1}^{n} K_d \left(\frac{m(U_i) - v}{h_d} \right),$$

where $K_d(\cdot)$ and h_d are an appropriate smoothing kernel and bandwidth, respectively. Integrating the function $(m^{-1})'(v) \mathbb{1}_{(m(0), m(1))} (m^{-1}(v))$ from $-\infty$ to (say) t

gives $m^{-1}(v)$, so an estimate of this quantity can be had by integrating the kernel density estimate

$$\hat{m}^{-1}(t) = \frac{1}{Nh_d} \sum_{i=1}^{n} \int_{-\infty}^{t} K_d \left(\frac{m(U_i) - v}{h_d} \right) \, dt.$$

Finally, taking the inverse of this estimate gives the estimate of the function $m(x)$.

Based on the above motivation, the monotone fitting procedure recommended by Dette et al. (2006) is to fit a kernel smoother to the data to obtain estimates of the unconstrained regression function. Use these estimates to fit density estimate as explained above. Then integrate the density estimate and take the inverse to obtain the estimate of the constrained estimate. Dette et al. (2006) prove that, under certain conditions, the estimate is asymptotically normal. However, this procedure requires the selection of two different tuning parameters denoted h_d and h_r for *d*ensity and *r*egression respectively, which can be a deterrent to its practical implementation.

Chang, Chien, Hsiung, Wen, and Wu (2007) use a Bernstein polynomial expansion for shape-restricted regression (see also Chak, Madras, and Smith, 2005). The Bernstein polynomial of a function $m(\cdot)$ can be written

$$\sum_{j=0}^{M} m(j/M) b_M(x, j),$$

where $b_M(x, j) = \binom{M}{j} x^j (1-x)^{M-j}$ is the j^{th} Bernstein polynomial of order M. The famous theorem of Bernstein (1912) shows that as M increases, the Bernstein polynomial of a function becomes arbitrarily close to the function $m(\cdot)$ uniformly. Thus, if we would like to estimate an unknown regression function, we can set $\beta_j = m(j/M)$ and estimate the coefficients β_j for a suitably large value of M.

The form of the Bernstein polynomial makes it especially amenable to shape restricted regression. Because the coefficients β_j are equal to the unknown regression function at equally spaced points in the domain of the function, it is relatively simple to derive the necessary constraints on the coefficients that will give different shape restrictions. Chang et al. (2007) give shape restrictions for monotonic, unimodal, and unimodal-concave functions. A Bayesian approach to fitting a shape-restricted regression with Bernstein polynomials is to place a prior on the coefficients that satisfies

the appropriate shape-constraints. Chang et al. (2007) do not actually give explicit priors for the coefficients. Instead they give a RJMCMC algorithm for sampling from the posterior distribution of the model parameters and the order of the Bernstein polynomials M. It is not clear from their exposition, however, that the algorithm they present is a valid RJMCMC sampling algorithm.

1.6 Plan of Dissertation

In this document we summarize some of our own research on variable selection methods and applications. We present this research in the form of four papers that have been submitted or are in preparation for submission to academic journals. In Chapter 2, we present a method for clustering and selection of regression coefficients in linear regression. In Chapter 3, we present a method for computing approximate posterior model probabilities in additive models. In Chapter 4, we discuss an application of variable selection to the fitting of monotonic regression functions. In Chapter 5, we present methods to fit shape-restricted regression functions using quadratic programming. We conclude with Chapter 6, where we discuss further research we have conducted in response to feedback on the methods described in the previous chapters and provide recommendations for future research.

Chapter 2

A Bayesian Approach to Multicollinearity and the Simultaneous Selection and Clustering of Predictors in Regression

2.1 Introduction

Multicollinearity is one of the most dreaded problems in regression analysis. A substantial amount of effort has been devoted to researching multicollinearity's deleterious effects, it's detection, and possible remedies (see Hill and Adkins, 2003, for a review). The literature contains book-length presentations of the topic (Belsley, Kuh, and Welsch, 1980; Belsley, 1991), lively discussions (Belsley, 1984), and accounts of the perils of improper corrective measures (Kennedy, 1982, 1983; Buse, 1994). The discussion of multicollinearity has even, at times, taken a religious tone, with Blanchard (1987) declaring "multicollinearity is God's will, not a problem with OLS or statistical techniques in general."

The crux of the multicollinearity problem is that in the typical linear model

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i \tag{2.1}$$

the coefficients β_j (for $j = 1, \ldots, p$) represent the effect of the j^{th} covariate x_{ij} on the response y_i, holding all other covariates constant. When the covariates exhibit multicollinearity, the data do not contain much information on the effects of one covariate holding the others constant, because, at least in the data at hand, the covariates "move" together. In other words, the data are deficient for determining the independent effects of a covariate on the response because the covariates themselves are not independent.

The effects of multicollinearity are many. Standard errors for the estimated coefficients will be extremely large. Some coefficients may have signs opposite of what is to be expected given the subject matter of the data. The overall F-test for significance of the regression may have a small p-value, even though none of the individual coefficients may be statistically significant. Small changes in the model specification (e.g., removing one covariate) can have dramatic effects on the resulting parameter estimates.

To underscore the point that multicollinearity is a problem with the data and not necessarily a problem with statistical methodology, Goldberger (1991) invents the term "micronumerosity" to describe the condition of having a data set with very few observations. When data exhibit micronumerosity, the coefficient estimates have large standard errors—but rightfully so! Large standard errors imply high uncertainty about the model parameters, and a small data set does not contain much information about the parameters. Similarly, when the data exhibit multicollinearity, the standard errors of the parameter estimates will be large, because the data do not contain much information about the regression coefficients.

Since weak data is to blame for the adverse effects of multicollinearity, one obvious solution is to obtain better data—that is, obtain data for values of the covariates not found (or underrepresented) in the original data set. In most cases, particularly in the social sciences, this solution is not feasible as the data must come from observational studies where the researcher has no option to impose different levels of covariates.

In some cases, the pattern of multicollinearity is itself an object of interest. In microarray data, where the expression level of many genes is used to predict a response (e.g., phenotype or survival time), correlated gene expressions could indicate groups of genes that are working together in a genetic pathway (Segal, Dahlquist, and Conklin, 2003). A model which tries to estimate independent effects does not seem appropriate for this type of data.

In situations such as those above, it seems appropriate to use a model that can specifically account for covariates whose effect on the response is not independent of other covariates in the model. One approach is to impose linear restrictions on the coefficients in (2.1). Common restrictions are $\beta_j = 0$ and $\beta_j = \beta_{j'}$. The latter restriction essentially forces the covariates x_{ij} and $x_{ij'}$ into a group.

There are problems with this approach in that the restrictions must be specified *a priori*. In some fields, like economics, with a rich theoretical history, restrictions like $\beta_j = \beta_{j'}$ might be justified on theoretical grounds alone. However, such justifications may not exist, or the restrictions may not hold exactly. Further, restrictions such as $\beta_j = 0$ seem silly *a priori*.

Theil and Goldberger (1961) and Theil (1963) provide a way to relax the exactness of the linear restrictions by adding random noise. If $\boldsymbol{\beta}$ is the vector of regression coefficients, and $\mathbf{R}$ is a matrix of restrictions, then a partial prior distribution on the regression coefficients can be imposed using

$$\mathbf{c} = \mathbf{R}\boldsymbol{\beta} + \boldsymbol{\eta}$$

where $\boldsymbol{\eta} \sim \mathcal{N}_p(0, \Sigma)$ and $\mathbf{c}$ is often a vector of zeros. However, even this clever approach requires the matrix of linear restrictions $\mathbf{R}$ to be specified *a priori*.

In this chapter, we propose a prior distribution on the space of all linear restrictions of the form $\beta_j = \beta_{j'}$ $(j \neq j')$ and $\beta_j = 0$. Under this prior, the linear restrictions on the coefficient parameters are determined by the data and do not need to be specified *a priori*. Our Bayesian approach results in a procedure that simultaneously selects covariates for inclusion in the model and clusters highly correlated covariates by setting their coefficients equal. In what follows, we discuss some related penalized regression methods (Section 2.2), present our Bayesian prior for linear restrictions

(Section 2.3), discuss simliar methods in the literature (Section 2.4), compare the performance of our method with competing methods through a simulation study (Section 2.5), and conclude with an analysis of some real data sets (Section 2.6) and some final thoughts (Section 2.7).

2.2 Penalized Least Squares

Penalized least squares is a procedure that consists of minimizing the sum of the squared residuals subject to a size constraint on the regression coefficients. More precisely, penalized least squares produces estimates of regression coefficients by finding values of β_0 and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ that minimize

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + P(\boldsymbol{\lambda}, \boldsymbol{\beta}) \tag{2.2}$$

where $P(\cdot, \cdot)$ is a positive function and $\boldsymbol{\lambda}$ is a (possibly) vector-valued tuning parameter that balances goodness-of-fit with the penalty criterion. The choice of penalty function $P(\cdot, \cdot)$ can have surprising effects on the estimates of the regression coefficients. We briefly describe some of the most popular penalty functions below.

One well known penalized least squares method is ridge regression (Hoerl and Kennard, 1970). Ridge regression was introduced as a way to correct for multicollinearity and uses the sum of the squared coefficients ($P(\lambda, \boldsymbol{\beta}) = \lambda \sum_{j=1}^{p} \beta_j^2$) as a penalty. The effect of the penalty on the resulting parameter estimates is that the estimates are "shrunk" closer to zero than the OLS estimates.

The Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani, 1996) uses the sum of the absolute values of the coefficients ($P(\lambda, \boldsymbol{\beta}) = \lambda \sum_{j=1}^{p} |\beta_j|$) as a penalty. Because of the geometry of the LASSO penalty, some of coefficient estimates can be set exactly equal to zero. Thus, the LASSO penalty simultaneously shrinks the coefficient estimates toward zero and removes some covariates from the model by setting their coefficients equal to zero.

For highly correlated covariates, the LASSO can have the undesirable property of arbitrarily choosing just one of two highly correlated covariates to remain in the

model. As a solution to this problem Zou and Hastie (2005) propose the "elastic net" procedure. The elastic net uses a penalty that is a convex combination of the ridge regression penalty ($P(\boldsymbol{\lambda}, \boldsymbol{\beta}) = \lambda_1 \sum_{j=1}^{p} \beta_j^2 + \lambda_2 \sum_{j=1}^{p} |\beta_j|$). The geometry of the elastic net penalty still has the ability to set some coefficient estimates exactly equal to zero. However, the ridge regression portion of the penalty allows for highly collinear covariates to be included in the model together.

The Octagonal Shrinkage and Clustering Algorithm for Regression (OSCAR) (Bondell and Reich, 2008) uses a convex combination of the lasso penalty and the pairwise maximum of coefficients ($P(\boldsymbol{\lambda}, \boldsymbol{\beta}) = \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j<k} \max\{\beta_j, \beta_k\}$) as a penalty. This penalty also has the LASSO-type ability to set certain coefficients equal to zero. Additionally, this penalty has the ability to set the coefficients of highly correlated covariates equal to each other. Thus, the OSCAR can identify clusters of covariates by setting their coefficients equal. Depending on the specific application, the ability to identify clusters can be a desirable improvement over the elastic net.

2.3 A Bayesian Approach to Simultaneous Shrinkage, Selection and Clustering

Some of the penalized least squares methods in the previous section have a Bayesian analogue. The connection to Bayesian inference comes from the combination of a normal likelihood and a prior that gives a posterior distribution whose (log) kernel is of the form (2.2). For example, a normal sampling density with known variance σ^2 and a p-dimensional normal prior for $\boldsymbol{\beta}$ with a zero mean vector and a covariance matrix $2\sigma^2/\lambda\mathbf{I}$ gives a posterior proportional to

$$\exp\left\{-\frac{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p}\beta_j^2}{2\sigma^2}\right\}. \tag{2.3}$$

Minimizing (2.2) with the ridge-regression penalty function is equivalent to maximizing (2.3). Thus, the ridge-regression estimates are equivalent to the maximum a posteriori (MAP) estimates.

As the above example illustrates, the penalty function in the penalized least squares method corresponds to a prior distribution in the Bayesian approach. In general, for a normal sampling density, the prior $p(\boldsymbol{\beta}|\sigma^2) \propto \exp\{-P(\lambda, \boldsymbol{\beta})/(2\sigma^2)\}$ leads to a posterior distribution whose MAP estimator is equivalent to the estimator obtained by minimizing (2.2).

We propose a prior that, like the OSCAR, simultaneously selects and clusters covariates by setting coefficients equal to each other or setting coefficients equal to zero. Thus, in effect, our approach is to put a prior on all possible linear restrictions of the form $\beta_j = \beta_{j'}$ and $\beta_j = 0$. We describe the full Bayesian model below.

Throughout the rest of this chapter, we assume that the response variable has been centered and that each predictor has been centered and scaled (i.e., for each predictor $\mathbf{x}_j$, $\sum_{i=1}^n x_{ij} = 0$ and $\sum_{i=1}^n x_{ij}^2 = n-1$), so that we may ignore the intercept term β_0.

We begin with the sampling density for the data which is the typical normal density with mean $\beta_1 x_{i1} + \cdots + \beta_p x_{ip}$ and variance σ^2. We induce a prior on the regression coefficients β_j (for $j > 0$) by combining random draws $\theta_1, \ldots, \theta_p$ from a distribution $\mathcal{P}(\cdot)$, where $\mathcal{P}(\cdot)$ is a random distribution from a Dirichlet process, and $\gamma_1, \ldots, \gamma_p$ from a Bernoulli distribution.

The Dirichlet process (Ferguson, 1973) is indexed by two parameters—a "base" distribution $H(\cdot)$ and a precision parameter α. The base distribution can be thought of informally as a the center of the random distributions from the Dirichlet process, and the precision parameter α controls how "close" the random distributions from the Dirichlet process are to the base distribution $H(\cdot)$.

We use the Dirichlet process because of its well known clustering properties (Blackwell and MacQueen, 1973; Neal, 2003; Ghosh and Ramamoorthi, 2003, Chapter 3). A sequence of random draws $\theta_1, \ldots, \theta_n$ from a realization of the Dirichlet process $\mathcal{P}(\cdot)$ gives positive probability to events $\theta_i = \theta_j$, where $i \neq j$.

We use a normal distribution with mean 0 and variance η^2 as a base distribution for $\mathcal{P}$. The normal base distribution essentially serves as a prior distribution for the regression coefficients, except that using the normal distribution in the Dirichlet process allows for clustering of the predictors by allowing the possibility of $\theta_j = \theta_{j'}$.

We note that our approach here is similar to Gopalan and Berry (1998) who use the Dirichlet process prior in a similar ANOVA setting to control for multiple comparisons among population means.

To allow for the removal of covariates from the model by setting the corresponding regression coefficient equal to zero, we introduce latent variables $\gamma_1, \ldots, \gamma_p$. These latent variables have a Bernoulli distribution with parameter π, where π is given a uniform prior on $(0, 1)$. Each θ_j and γ_j are multiplied to induce the desired prior on the regression coefficient $\beta_j = \gamma_j \theta_j$.

The full hierarchical Bayesian model can be summarized as

$$
\begin{aligned}
Y_i | \boldsymbol{\beta}, \sigma^2 &\sim \mathsf{N}\left(\mathbf{x}_i^T \boldsymbol{\beta}, \sigma^2\right) && \text{for } i = 1, \ldots, n \\
\beta_j &= \gamma_j \theta_j && \text{for } j = 1, \ldots, p \\
\gamma_j &\sim \mathsf{Bern}(\pi) && \text{for } j = 1, \ldots, p \\
\theta_j | \mathcal{P} &\sim \mathcal{P}(\cdot) && \text{for } j = 1, \ldots, p \\
\mathcal{P} &\sim \mathsf{DiriProc}(\alpha, \Phi_{0,\eta^2}(\cdot)) \\
\pi &\sim \mathsf{Unif}(0, 1) \\
\sigma^2 &\sim \mathsf{InvGam}(a_\sigma, b_\sigma) \\
\eta^2 &\sim \mathsf{InvGam}(a_\eta, b_\eta)
\end{aligned}
$$

where $\Phi_{m,s^2}(\cdot)$ denotes the normal distribution with mean m and variance s^2, $\mathsf{InvGam}(a, b)$ denotes an inverse gamma distribution with density function $p(x) = b^a / \Gamma(a) x^{-(a+1)} \exp\{-b/x\}$, and a_σ, b_σ, a_η, and b_η are constants (more on this later).

In practice, we approximate the Dirichlet process using the finite-dimensional Dirichlet process as outlined in Ishwaran and Zarepour (2002, section 4). In our case, the finite Dirichlet process can be constructed for the regression coefficients by letting $p_1, \ldots, p_M$ be draws from a Dirichlet distribution with parameters $\alpha/M, \ldots, \alpha/M$ for a suitably large M (see the appendix for more details) and $\xi_1, \ldots, \xi_M$ be draws from $\mathcal{N}(0, \eta^2)$. Then, we let $S_1, \ldots, S_p$ be draws from a discrete distribution on the integers $1, \ldots, M$ with probabilities $p_1, \ldots, p_M$. The prior on regression coefficients is induced by setting $\beta_j = \gamma_j \xi_{S_j}$.

The full hierarchical Bayesian model with the finite Dirichlet process can be summarized as

$$
\begin{aligned}
Y_i|\boldsymbol{\beta}, \sigma^2 &\sim \mathsf{N}\left(\mathbf{x}_i^T\boldsymbol{\beta}, \sigma^2\right) && \text{for } i = 1, \ldots, n \\
\beta_j &= \gamma_j \xi_{S_j} && \text{for } j = 1, \ldots, p \\
\gamma_j &\sim \mathsf{Bern}(\pi) && \text{for } j = 1, \ldots, p \\
S_j &\sim \mathsf{Multinom}(p_1, \ldots, p_M) && \text{for } j = 1, \ldots, p \\
\xi_j|\eta^2 &\sim \mathsf{N}\left(0, \eta^2\right) && \text{for } j = 1, \ldots, M \\
(p_1, \ldots, p_M) &\sim \mathsf{Dirichlet}(\alpha/M, \ldots, \alpha/M) \\
\pi &\sim \mathsf{Unif}(0, 1) \\
\sigma^2 &\sim \mathsf{InvGam}(a_\sigma, b_\sigma) \\
\eta^2 &\sim \mathsf{InvGam}(a_\eta, b_\eta)\,.
\end{aligned}
$$

The details on sampling from the posterior distribution of the model parameters are given in the appendix.

2.4 Similar Methods in the Literature

Since the original draft of this chapter, we have found two approaches to correlated predictors in the literature that are similar to the method we have proposed here. First, Nott (2008) uses a Dirichlet process prior with a normal base distribution for the regression coefficients. This allows for clustering of regression coefficients but does not allow for regression coefficients to be zero. Nott (2008) notes the similarity of his approach with ridge regression. As the Dirichlet precision parameter α approaches ∞, the Dirichlet process prior becomes arbitrarily "close" to the normal base distribution, and the method in Nott (2008) essentially becomes ridge regression.

Second, MacLehose, Dunson, Herring, and Hoppin (2007) place a Dirichlet process prior with a mixture distribution as the base distribution on the regression coefficients. The mixture they use is a combination of a point mass at zero and a normal distribution. Thus, the method of MacLehose et al. (2007) allows clustering of the regression

coefficients through the Dirichlet process and setting of the coefficients equal to zero through the point mass at zero in the base distribution. This method is essentially the same as the method we have proposed here.

2.5 Simulation Study

We conducted five simulations to compare the Bayesian approach to other methods—OLS and the penalized least-squares methods mentioned previously. Our simulation study is similar to the one conducted by Bondell and Reich (2008) (see also Tibshirani, 1996). We simulated 50 data sets from each of five different data-generating schemes. Each data set was generated from the standard linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon \qquad \text{for } i = 1, \ldots, n$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I}_n)$, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T \sim \mathcal{N}(\mathbf{0}, \mathbf{V})$, σ^2, and $\boldsymbol{\beta}$ were varied for each of the 5 simulations as described below

Sim	n	p	$\boldsymbol{\beta}$	σ	$\mathbf{V}_{ij}$
1	20	8	$(3, 2, 1.5, 0, 0, 0, 0, 0)^T$	3	$0.7^{\lvert i-j\rvert}$
2	20	8	$(3, 0, 0, 2, 0, 0, 0, 1.5)^T$	3	$0.7^{\lvert i-j\rvert}$
3	20	8	$(0.85, \ldots, 0.85)^T$	3	$0.7^{\lvert i-j\rvert}$
4	50	20	$(\underbrace{0, \ldots, 0}_{5}, \underbrace{2, \ldots, 2}_{5}, \underbrace{0, \ldots, 0}_{5}, \underbrace{2, \ldots, 2}_{5})$	15	$\begin{cases} 0.5 & \text{for } i \neq j \\ 1.0 & \text{otherwise} \end{cases}$
5	50	20	$(\underbrace{3, \ldots, 3}_{9}, \underbrace{0, \ldots, 0}_{11})$	15	$\text{diag}(\underbrace{\mathbf{V}^*, \ldots, \mathbf{V}^*}_{3}, \mathbf{I}_{11})$

where $\text{diag}(\mathbf{A}_1, \ldots, \mathbf{A}_r)$ denotes a block diagonal matrix with blocks $\mathbf{A}_1, \ldots, \mathbf{A}_r$ and

$$\underset{3\times3}{\mathbf{V}^*} = \begin{pmatrix} 1.16 & 1.00 & 1.00 \\ 1.00 & 1.16 & 1.00 \\ 1.00 & 1.00 & 1.16 \end{pmatrix}.$$

The first simulation models a situation where the three active predictors are highly correlated with each other and less correlated with the inactive predictors. The second simulation models a situation where the three active predictors are highly correlated

with other inactive predictors and less correlated with each other. The third simulation models a situation where all predictors are active and are correlated with each other. The fourth simulation models a situation where two groups of predictors are active but every predictor is equally correlated with every other predictor. Finally, the fifth simulation models a situation where there are three groups of active predictors, each predictor in an active group is correlated with the other predictors in its group, and the predictors in the inactive group are not correlated with any other predictors.

The elastic net, LASSO, OSCAR, and ridge regression require the selection of tuning parameters. We used 10-fold cross-validation to select tuning parameters for the elastic net, LASSO, and ridge regression. To select the tuning parameters for the OSCAR we used a variation of AIC (Akaike, 1973; McQuarrie and Tsai, 1998)

$$\mathrm{AIC}_c = \log \frac{\mathrm{RSS}}{n} + \frac{n + \mathrm{df}}{n - \mathrm{df} - 2},$$

where RSS is the residual sum of squares and df is the effective number of parameters as defined by Bondell and Reich (2008).

For our Bayesian model we let $\alpha = 1$, $a_\sigma = b_\sigma = a_\eta = b_\eta = 0.1$, and $M = p + 25$ (these values are also used for the examples in Section 2.6). We fit the model using a Gibbs sampler constructed with the `OpenBUGS` software (Thomas, O'Hara, Ligges, and Sturtz, 2006). We used 3 chains per parameter and used 6,000 iterations per chain with 1,000 per chain discarded as burn-in. We used the means of the posterior draws of the parameters as point estimates for calculation of MSE. (See the appendix for computational details and code.)

The fit of each method was evaluated by computing

$$\mathrm{MSE}(\hat{\boldsymbol{\beta}}_m) = (\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta})^T \mathbf{V} (\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}) \tag{2.4}$$

where $\hat{\boldsymbol{\beta}}_m$ is the estimate of the true beta vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ for the m^{th} method and and $\mathbf{V}$ is the population covariance matrix of the predictors (see Tibshirani, 1996).

The results of each simulation are presented in Figure 2.1. The first row of plots in Figure 2.1 contains boxplots of the MSE for each of the 50 data sets and each

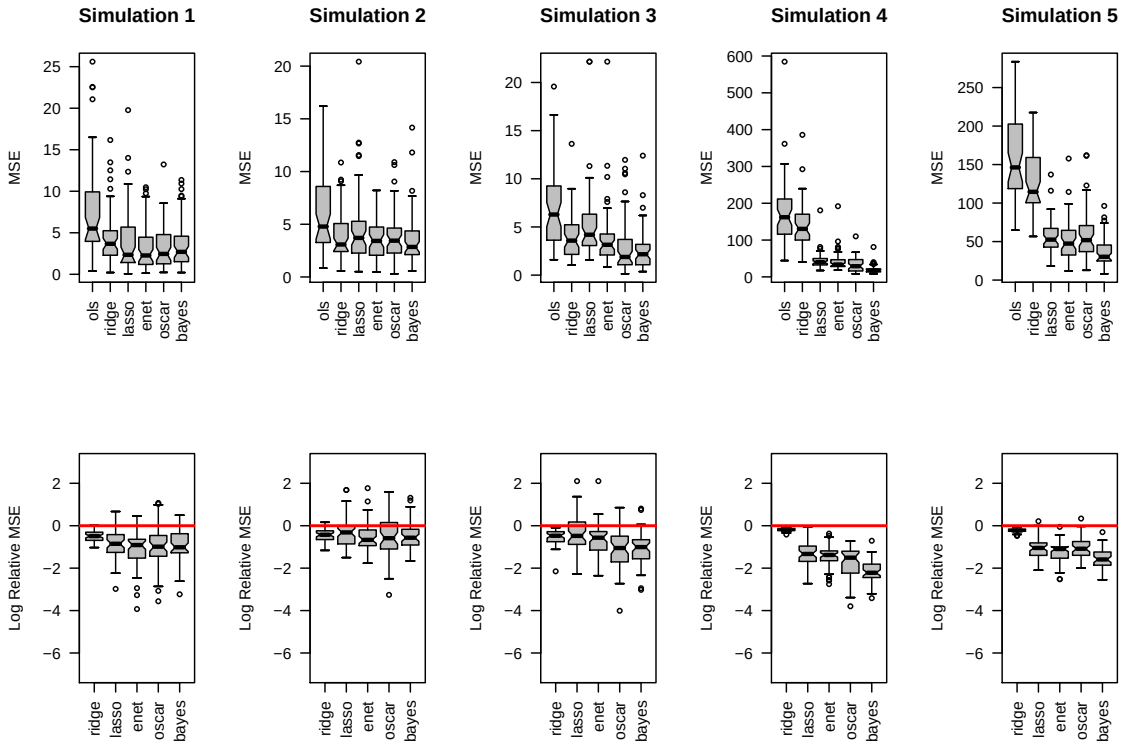

Figure 2.1: Boxplots of simulation results. Boxplots in the first row of plots represent the MSE's from 50 data sets. Boxplots in the second row of plots represent LRMSE's from 50 data sets.

of the methods. It is immediately apparent from the plots that all of the methods perform better than the OLS estimates (on average). The plot also indicates that the Bayesian method we have proposed is highly competitive with the other methods. In three of the five simulations (simulations 2, 4, and 5), the Bayesian method had the lowest median MSE of the competing methods.

The second row of plots in Figure 2.1 contains boxplots of the log of the ratio of the MSE of each method to the MSE of the OLS estimates. That is, for each data set, we calculated

$$\text{LRMSE}_m = \log(\text{MSE}_m/\text{MSE}_{OLS}) \tag{2.5}$$

where MSE_m is the MSE of the m^{th} method and LRMSE_m stands for the Log Relative MSE. This metric essentially measures the improvement (or not) of the m^{th} method

over the standard OLS estimates. Values of LRMSE_m greater than zero indicate that the m^{th} method had a larger (i.e., worse) MSE than the OLS estimates. Values of LRMSE_m less than zero indicate that the m^{th} method had a smaller MSE (i.e., better) than the MSE of the OLS estimates.

From the second row of plots in Figure 2.1, it appears that most of the methods are comparable in terms of their improvement over standard OLS estimates (with ridge regression, in general, having a worse improvement over OLS than the other methods). We see that, in simulations 4 and 5, the median value of LRMSE_m for the Bayesian method is lower than all other methods and is close to the lowest in the remaining three simulations.

2.6 Illustrations with Real Data

2.6.1 Hald Data

We begin with the famous "Hald" data set (Woods, Steinour, and Starke, 1932; Hald, 1952). The data contain 13 observations and five variables. The response variable (y) is the heat evolved for particular mixture of cement, and the covariates are tricalcium aluminate (x_1), tricalcium silicate (x_2), tetracalcium alumino ferrite (x_3), and dicalcium silicate (x_4).

Table 2.1 contains a standard OLS analysis of the Hald data. The variance inflation factors (VIF) for each of these predictors are all extremely high, indicating troublesome multicollinearity. Also, the overall regression is significant (with a p-value less than 0.0001) even though none of the individual predictors is significant.

Figure 2.2 contains a plot of the estimates of the Hald coefficients for each of the methods we used in the simulation study (also, see Table 2.2). Tuning parameters for the OSCAR, elastic net, LASSO, and ridge regression were selected as in Section 2.5, and estimates from each method are represented on the plot by the first letter of the name of the method (e.g., OSCAR estimates are denoted by an "o"). Bayesian estimates (denoted on the plot by a circle) for each regression coefficient were obtained from 15,000 MCMC simulations (3 chains, 1,000 burn in, 5,000 iterations per chain af-

Table 2.1: Standard regression analysis of the Hald data.

Coef	Estimate	SE	p-val	VIF
β_1	9.12	4.13	0.055	38
β_2	7.94	10.62	0.474	254
β_3	0.65	4.56	0.890	47
β_4	−2.41	11.19	0.834	283

Table 2.2: Estimates of the Hald coefficients under different estimation methods.

Method	β_1	β_2	β_3	β_4
Bayes	8.90	9.10	0.00	0.00
OLS	9.12	7.94	0.65	-2.41
Ridge	6.63	4.50	-1.64	-5.81
LASSO	8.71	6.91	0.20	-3.49
E. Net	7.06	4.90	-1.55	-5.97
OSCAR	4.57	4.57	-3.51	-4.57

ter thinning by 10) by taking the median of all nonzero MCMC draws. OLS estimates are denoted on the plot by an "X." Ninety-five percent credible (confidence) intervals were obtained for Bayesian estimates (black lines) and for the OLS estimates (gray lines). Estimates from each method are listed in Table 2.2. The Bayesian estimates in the table are posterior medians of all (including zeroes) MCMC simulations.

As expected, the Bayesian intervals are substantially shorter than the OLS confidence bands. The medians of the posterior distributions of two coefficients (β_3 and β_4) are zero. In comparison, the LASSO set β_3 to zero and the elastic net and OSCAR set none of the coefficients to zero. The OSCAR has also set three of the four coefficients equal to each other in magnitude with the last coefficients differing from the first two coefficients in sign only.

Unlike other other methods, the Bayesian analysis automatically gives multiple cluster-configurations with an estimated posterior probability for each configuration. Different clusters of predictors can be obtained simply by counting their frequencies

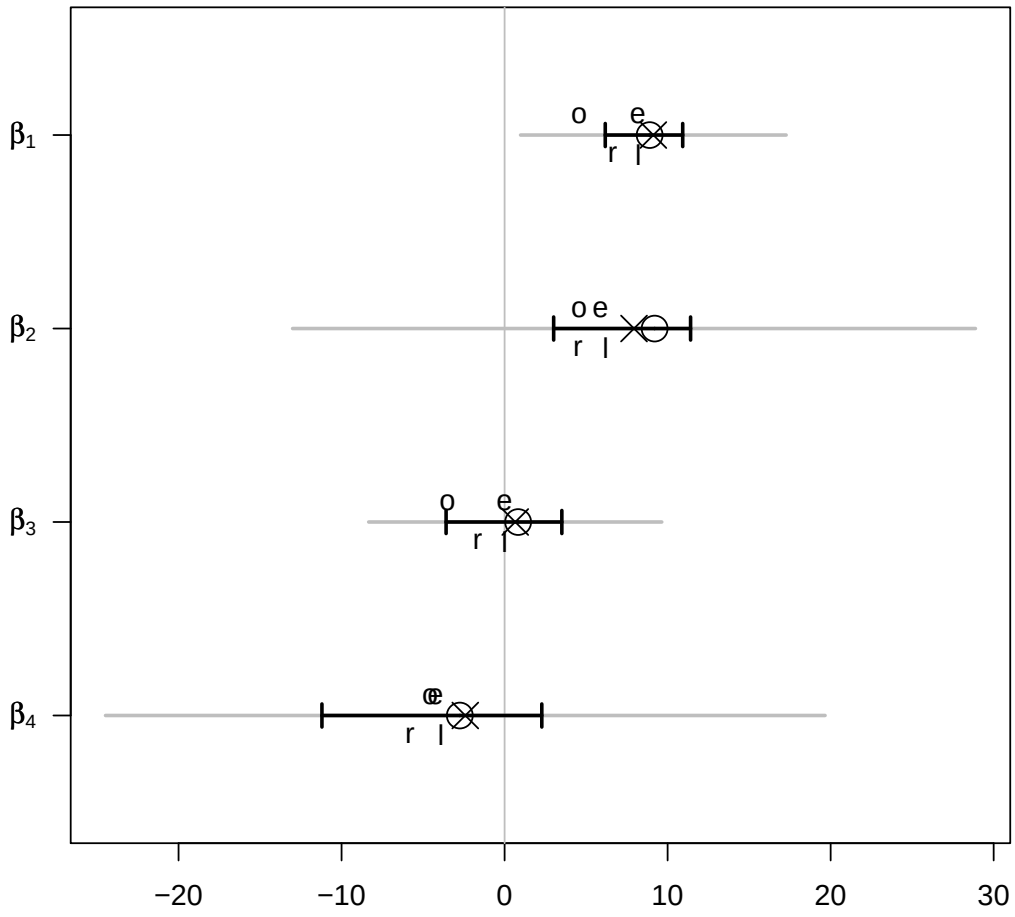

Figure 2.2: Estimates for the coefficients in the Hald data example. The Bayes estimates are the posterior medians of all non-zero MCMC draws and are denoted on the graph by a circle. The OLS estimates are denoted by an "X". Each of the remaining estimates are denoted by the first letter of the name of the method—LASSO with an "l", ridge regression with an "r", OSCAR with an "o", and elastic net with an "e". The dark lines denote Bayesian 90% credible (confidence) intervals taken from the non-zero MCMC draws and the light gray lines denote 90% confidence intervals for the OLS estimates.

Table 2.3: Frequency of clusters in the Hald data example. Point estimates (posterior medians) of the regression coefficients are given for 13 of the most common models.

x_1	x_2	x_3	x_4	Percentage	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
□	□	∅	∅	17.2	9.4	9.4	0.00	0.0
□	□	∅	△	15.7	8.4	8.4	0.00	-2.1
□	△	∅	◊	11.8	8.5	6.2	0.00	-4.2
□	□	△	∅	11.3	10.1	10.1	1.65	0.0
□	△	∅	∅	10.8	8.6	10.2	0.00	0.0
□	□	△	◊	10.2	8.8	8.8	0.42	-1.7
□	△	◊	⋈	8.3	7.8	5.2	-0.73	-5.3
□	□	△	△	3.0	10.6	10.6	1.29	1.3
□	∅	△	◊	2.9	6.2	0.0	-2.60	-10.7
□	△	◊	∅	2.6	9.8	10.1	1.41	0.0
□	∅	∅	△	2.6	8.4	0.0	0.00	-10.2
□	△	◊	◊	2.1	9.7	11.5	1.33	1.3
□	△	△	◊	1.0	5.4	-3.5	-3.50	-14.3

in the MCMC simulations. Table 2.3 contains the frequencies of various clusters of coefficients in the posterior draws. In the table, clusters are denoted by two or more predictors having the same symbol in their respective columns. The empty set symbol denotes a predictor that was set to zero.

The most common configuration of coefficients for the Hald data was $\beta_1 = \beta_2$ with β_3 and β_4 set to zero. This configuration was found in 17.2% of the posterior draws. The LASSO configuration, where β_3 is set equal zero, appears third most often with 11.8% of the posterior draws. The OSCAR configuration, where $\beta_1 = \beta_2 = -\beta_4$, does not appear in the posterior draws as the Bayesian procedure does not allow coefficients with equal magnitudes but opposite signs. Finally, the configuration with all coefficients nonzero and none of them clustered appears in the posterior distribution 8.3% of the time.

Table 2.4: Description of variables in the Ehrlich (1973) crime data set (see `R` documentation for `UScrime` data set in the `MASS` package, Venables and Ripley, 2002).

Variable	Description
`M`	Percentage of males aged 14-24
`So`	Indicator variable for a southern state
`Ed`	Mean years of schooling
`Po1`	Police expenditure in 1960
`Po2`	Police expenditure in 1959
`LF`	Labor force participation rate
`M.F`	Number of males per 1000 females
`Pop`	State population
`NW`	Number of nonwhites per 1000 people
`U1`	Unemployment rate of urban males 14-24
`U2`	Unemployment rate of urban males 35-39
`GDP`	Gross domestic product per head
`Ineq`	Income inequality
`Prob`	Probability of imprisonment
`Time`	Average time served in state prisons
`y`	Rate of crimes in a particular category per head of population

2.6.2 Crime Data

For our next example, we use the crime data of Ehrlich (1973) and Ehrlich (1975) (and subsequently corrected by Vandaele, 1978). A description of each variable is in Table 2.4 (Venables and Ripley, 2002).

As has been done by previous authors, we use the log of all variables (response and predictors) before fitting the regression. We begin with a standard OLS analysis found in Table 2.5. The variance inflation factors for two of the variables—`Po1` and `Po2`—are large enough to cause concern.

Estimates for the Bayesian model were based on a MCMC sample of 5,000 it-

Table 2.5: Standard OLS analysis of the crime data.

Coef	Est	SE	p-val	VIF
M	0.143	0.042	0.002	2.5
Ed	0.235	0.059	4e-04	5.3
Po1	0.269	0.251	0.292	93.5
Po2	-0.012	0.254	0.962	96.3
LF	0.033	0.044	0.449	2.8
M.F	-0.072	0.053	0.186	4.2
Pop	-0.082	0.048	0.097	3.4
NW	0.143	0.048	0.005	3.4
U1	-0.032	0.054	0.561	4.3
U2	0.115	0.054	0.042	4.4
GDP	0.137	0.076	0.082	8.7
Ineq	0.332	0.071	5e-05	7.5
Prob	-0.153	0.048	0.003	3.4
Time	-0.072	0.046	0.129	3.2

erations from 3 chains after thinning each chain by 5 and burning the first 1,000 iterations of each chain.

Figure 2.3 contains point estimates of the coefficients for the crime data. The Bayes estimates are posterior medians of all nonzero MCMC draws and are denoted by a circle. OLS estimates are denoted by an "X." Estimates of the remaining methods are denoted by the first letter in the name of the method (as was done in the Hald example). Ninety-percent credible intervals for the Bayesian estimates are denoted by the dark lines and 90% confidence intervals for the OLS estimates are denoted by the gray lines.

Figure 2.3 shows that the Bayesian intervals for the coefficients on `Po1` and `Po2` are substantially smaller than the confidence intervals for the OLS estimates. Also, both Bayesian estimates of the coefficients for `Po1` and `Po2` are of the same sign, whereas the OLS estimates have opposite signs.

Figure 2.4 contains the estimated posterior probability of a zero coefficient for each predictor in the crime data. We see a slight difference between the Bayesian analysis and the OLS analysis. With p-values of 0.299 and 0.962, the police expenditure variables are not significant according to the standard OLS analysis. However, the estimated posterior probabilities of zero police-expenditure coefficients are 0.165 and 0.225.

Unlike the Hald data, no interpretable clustering pattern emerges in the crime data example. However, we can compute posterior probabilities of certain cluster patterns of interest. For example, the posterior probability that the two crime expenditure variables are in the same cluster is 0.249.

2.7 Discussion

The Bayesian method that we have presented here has several distinct advantages over the competing methods that we have examined in this chapter. The first is that our method does not require the specification of one or more tuning parameters. Hyperparameters must be specified, but the "non-informative" values for these hyperparameters that we have used in this chapter have proven to be effective in the

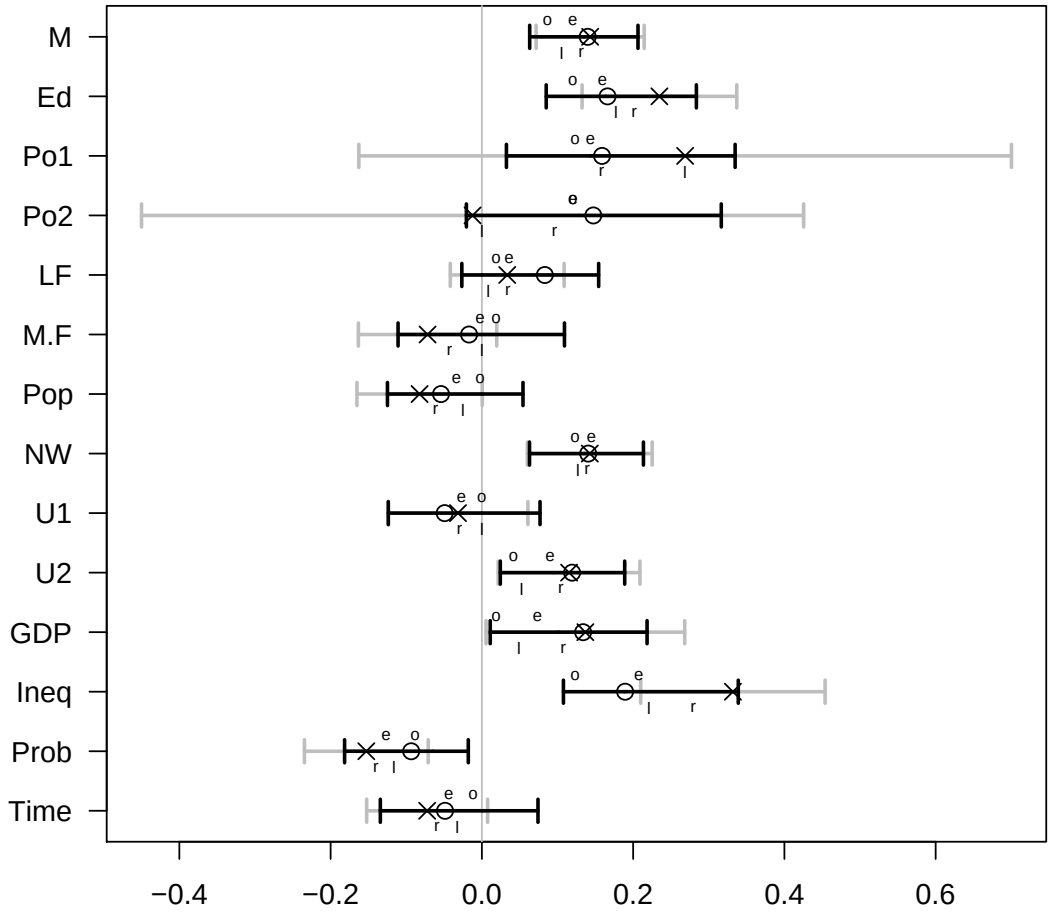

Figure 2.3: Estimates for the coefficients in the crime data example. The Bayes estimates are the posterior medians of all nonzero MCMC draws and are denoted on the graph by a circle. The OLS estimates are denoted by an "X". Each of the remaining estimates are denoted by the first letter of the name of the method—LASSO with an "l", ridge regression with an "r", OSCAR with an "o", and elastic net with an "e". The dark lines denote Bayesian credible (confidence) intervals and the light gray lines denote confidence intervals for the OLS estimates.

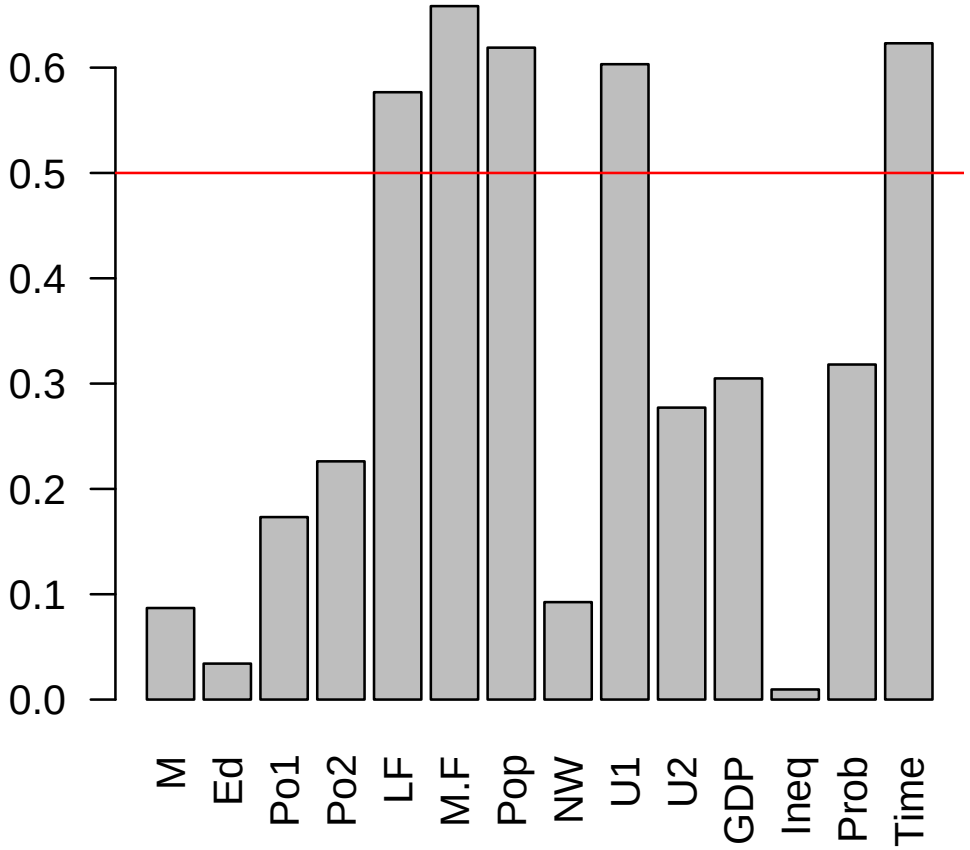

Figure 2.4: Posterior probability that each coefficient equals zero in the crime data example.

simulation study and in the examples. Thus, we can consider the hyperparameter values that we have chosen here as good default values for the Bayesian procedure.

Second, because we use Gibbs sampling to search the posterior model space we do not have to enumerate all possible models to obtain posterior probability estimates (George and McCulloch, 1993).

Third our Bayesian procedure comes with an automatic way to produce uncertainty calculations for quantities of interest. Calculation of Bayesian confidence intervals is straightforward (as seen in Figures 2.2 and 2.3). Calculation of cluster uncertainty can easily be calculated by examining the clustering patterns in the MCMC draws (as in Table 2.3).

Finally, the Bayesian method does not require that $p < n$. That is, the Bayesian procedure that we have used here may be used in the "large p, small n" problem (West, 2003) without any further adaptation.

Chapter 3

Approximate Posterior Model Probabilities in Additive Models via the Group LASSO

3.1 Introduction

The literature abounds in variable selection methods for the linear model (see, for example, Miller, 2002; George, 2000). One particular method that has generated a substantial amount of research is the Least Absolute Shrinkage and Selection Operator or LASSO (Tibshirani, 1996). This method involves minimizing penalized sums of squares where the penalty is the sum of the absolute values of the coefficients. For certain values of a tuning parameter, the minimizer of this penalized sum of squares can set one or more coefficients to zero. Many other variable selection approaches are variations on this penalized regression theme and typically differ from the LASSO by varying the form of the penalty. See, for example, Breiman (1995), Fan and Li (2001), Zou (2006), Zou and Hastie (2005) and Bondell and Reich (2008).

In many practical applications, the linear model assumptions are too restrictive and nonparametric regression models are preferred. In recent years several authors have proposed variable selection techniques for fully nonparametric regression. Friedman (1991) uses a forward stepwise regression procedure to construct a regression

function from "reflected pairs" of basis functions. Linkletter et al. (2006) define the covariance function of a Gaussian process to be a function of individual predictors. Variables are selected by inclusion or exclusion from the covariance function. Lafferty and Wasserman (2008) use derivatives of the nonparametric function estimates with respect to smoothing parameters to find sparse solutions to the nonparametric variable selection problem.

Although, fully nonparametric regression models are attractive in that they make relatively few assumptions about the regression function, they also lack the interpretability of the classical linear model. Additive models (Stone, 1985; Buja, Hastie, and Tibshirani, 1989; Hastie and Tibshirani, 1990) are a nice compromise between the restrictive linear model and the fully nonparametric formulation. The additive model assumes that each predictor's contribution to the mean of the response can be modeled by an unspecified smooth function, thereby retaining some of the benefits of fully nonparametric regression. These models are typically fitted with a backfitting algorithm (Hastie and Tibshirani, 1990). Additive models retain some of the benefits of interpretability found in classical linear models because each predictor has its own functional effect on the response. In addition, the simplifying assumptions of additive functional effects allow additive models to avoid the curse of dimensionality. Additive models can also be extended to functional-ANOVA models that allow for higher order interactions among the predictors (Barry, 1986; Wahba, 1990; Gu, 2002).

A handful of variable selection techniques exist for additive models. Chen (1993) develops a bootstrap procedure for model selection in functional-ANOVA models. Shively et al. (1999) develop a Bayesian model where the functional effect of each predictor is given a prior with a linear component and a nonlinear Wiener process component. Shi and Tsai (1999) give a modified version of Akaike's Information Criterion (AIC) (Akaike, 1974) suitable for selection of regression models with linear and additive components. Gustafson (2000) presents a Bayesian variable selection technique for regression models that allow predictors to have linear or functional effects and two-way interactions. Wood, Kohn, Shively, and Jiang (2002) develop a Bayesian method, based on the Bayesian Information Criterion (BIC) (Schwarz, 1978), for selecting between a linear regression model, a model with additive functional effects,

or a fully nonparametric regression model. Lin and Zhang (2006) present the Component Selection and Smoothing Operator (COSSO) that is a generalization of the LASSO based on fitting a penalized functional-ANOVA model where the penalty is the norm of the projection of each functional component into a partition of the model space. Reich et al. (2008) develop a Bayesian variable selection technique for functional-ANOVA models with Gaussian process priors.

Yuan and Lin (2006, hereafter YL06) present several variations of model selection for predictors that form natural groupings (e.g., sets of dummy variables for factors or basis expansions for additive models). Their approach generalizes several penalized linear regression methods, such as Least Angle Regression (LARS) (Efron et al., 2004) and the nonnegative garrote (Breiman, 1995).

Of particular interest to the topic of this chapter is the generalization of the LASSO that YL06 call the group LASSO. Avalos et al. (2003) also develop a similar procedure for the special case of additive models using a B-spline basis. The group LASSO is a penalized least-squares method that uses a special form of penalty to eliminate prespecified groups of variables from the model. More specifically, if $\mathbf{y}$ is an $n \times 1$ vector of responses, $\mathbf{X}_j$ is an $n \times m_j$ matrix of variables associated with the j^{th} predictor and $\boldsymbol{\beta}_j$ is an $m_j \times 1$ vector of coefficients, then group LASSO solutions minimize

$$\text{argmin}_{\boldsymbol{\beta}} \left\| \mathbf{y} - \sum_{j=1}^{g} \mathbf{X}_j \boldsymbol{\beta}_j \right\|^2 + \lambda \sum_{j=1}^{g} \|\boldsymbol{\beta}_j\|_{\mathbf{K}_j} \tag{3.1}$$

where $\|\mathbf{z}\|_{\mathbf{A}} = (\mathbf{z}^T \mathbf{A} \mathbf{z})^{1/2}$ (for $\underset{d\times 1}{\mathbf{z}} \in \mathbb{R}^d$ and $\underset{d\times d}{\mathbf{A}}$ a positive definite matrix), $\boldsymbol{\beta} = (\underset{m_1\times 1}{\boldsymbol{\beta}_1}^T, \ldots, \underset{m_g\times 1}{\boldsymbol{\beta}_g}^T)^T$ and g is the number of groups. YL06 show that for some values of the tuning parameter λ, the solution to (3.1) includes $\boldsymbol{\beta}_j = \mathbf{0}$ for some subset of $j = 1, \ldots, g$.

One drawback to most variable selection methods is that they do not provide a measure of model uncertainty. Variable selection methods typically give one model as best, without giving some measurement of uncertainty for this estimated model.

The exceptions to this are methods that use a Bayesian paradigm, which typically provide a measure of model uncertainty by calculating the number of times a particular model is visited in the posterior draws from a Markov chain Monte Carlo

(MCMC) simulation (George and McCulloch, 1993). However, MCMC methods are computationally expensive and it can be hard to assess convergence when MCMC methods must traverse a space of differing dimensions.

In this chapter, we present a method for calculating approximate posterior model probabilities after fitting the group LASSO. We generalize the method of Yuan and Lin (2005, hereafter YL05), who develop a similar method in the case of the classical linear regression model. YL05 present an empirical Bayes method to obtain posterior model probabilities in the linear model using Laplace's approximation and the LASSO. They use a normal model for errors and a Laplacian prior for regression coefficients and show that, for a particular choice of model prior probabilities, choosing the model with highest posterior probability is equivalent to choosing the model selected by the LASSO. YL05 then derive empirical Bayes estimates of the tuning parameter (the hyperparameter from the Laplacian prior) and the variance parameter from the sampling density. Although YL05 provide formulas to compute posterior model probabilities, in practice (i.e., in their simulations and examples), they use their empirical Bayes formulation only to select the value of the tuning parameter in the LASSO.

In what follows, we present our Bayesian method for posterior model probabilities in additive model (Section 3.2), discuss issues surrounding what YL05 termed "nonregular" models (Section 3.3), examine the performance of our method using simulated examples (Section 3.5) and close with an illustration of our method on real data (Section 3.6) and some concluding remarks (Section 3.7).

3.2 Model Formulation

Following YL05, we use a hierarchical model structure to carry out the variable selection. We model the joint distribution of data $\mathbf{y}$, basis function coefficients $\boldsymbol{\beta}$ and variable-selection parameters $\boldsymbol{\gamma}$ as

$$p(\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = p(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\gamma})p(\boldsymbol{\beta}|\boldsymbol{\gamma})p(\boldsymbol{\gamma})$$

The model begins with n independent observations $\mathbf{y} = (y_1, \ldots, y_n)$. The response

for the i^{th} observation y_i depends on p predictor vectors $x_{i1}, \ldots, x_{ip}$ in the following fashion

$$y_i = \sum_{j=1}^{p} f_j(x_{ij}) + \epsilon_i.$$

where $\epsilon_i \sim \mathsf{N}(0, \sigma^2)$. Each $f_j(\cdot)$ is modeled with a basis expansion of m basis functions $\psi_{j1}(\cdot), \ldots, \psi_{jm}(\cdot)$ with coefficients $\beta_{j1}, \ldots, \beta_{jm}$, where m is chosen to be $\lfloor n^{1/5} \rfloor$ (where $\lfloor z \rfloor$ is the largest integer not greater than z). Our choice for m is motivated by the fact that the bias of a basis-expansion estimate decays like m^{-2} when the true regression function is twice continuously differentiable, while the variance decays like m/n. Thus the choice $m = Cn^{1/5}$ leads to bias-variance tradeoff giving the smallest order of mean squared error. However, we do not address the issue of choosing the optimal value for C and simply set this value equal to one.

We let $\boldsymbol{\gamma}$ be the vector containing the p variable selection parameters γ_j, where $\gamma_j = 1$ if predictor j is in the model and $\gamma_j = 0$ otherwise. Writing the basis functions as a matrix

$$\underset{n \times mp}{\boldsymbol{\Psi}} = \begin{pmatrix} \psi_{11}(x_{11}) & \cdots & \psi_{1m}(x_{11}) & \cdots & \psi_{p1}(x_{1p}) & \cdots & \psi_{pm}(x_{1p}) \\ \psi_{11}(x_{21}) & \cdots & \psi_{1m}(x_{21}) & \cdots & \psi_{p1}(x_{2p}) & \cdots & \psi_{pm}(x_{2p}) \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \psi_{11}(x_{n1}) & \cdots & \psi_{1m}(x_{n1}) & \cdots & \psi_{p1}(x_{np}) & \cdots & \psi_{pm}(x_{np}) \end{pmatrix}$$

and the coefficients as a vector

$$\underset{pm \times 1}{\boldsymbol{\beta}} = (\underset{m \times 1}{\boldsymbol{\beta}_1^T}, \ldots, \underset{m \times 1}{\boldsymbol{\beta}_p}^T)^T = (\beta_{11}, \ldots, \beta_{1m}, \ldots, \beta_{p1}, \ldots, \beta_{pm})^T,$$

we can write our model for $\mathbf{y}$ conditioned on $\boldsymbol{\gamma}$ as

$$\underset{n \times 1}{\mathbf{y}} \sim \mathsf{N}_n\left(\underset{n \times mk}{\boldsymbol{\Psi}_{\boldsymbol{\gamma}}} \underset{mk \times 1}{\boldsymbol{\beta}_{\boldsymbol{\gamma}}}, \sigma^2 \underset{n \times n}{\mathbf{I}_n}\right)$$

where k is the number of nonzero elements of $\boldsymbol{\gamma}$, $\boldsymbol{\Psi}_{\boldsymbol{\gamma}}$ and $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ contain the columns of $\boldsymbol{\Psi}$ and the elements of $\boldsymbol{\beta}$ that correspond to nonzero elements of $\boldsymbol{\gamma}$ and $\mathbf{I}_n$ is an $n \times n$ identity matrix.

Conditional on $\boldsymbol{\gamma}$, the prior density (with respect to the sum of counting and Lebesgue measure) for $\boldsymbol{\beta}_j$ is a mixture of a point-mass at zero and a multivariate

Laplace distribution (Ernst, 1998), that is,

$$p(\boldsymbol{\beta}_j|\boldsymbol{\gamma}) = (1-\gamma_j)I_{\{\mathbf{0}\}}(\boldsymbol{\beta}_j) + \gamma_j \frac{\Gamma(m/2)}{2\pi^{m/2}\Gamma(m)}\tau^m \exp\left\{-\tau \left\|\boldsymbol{\beta}_j\right\|\right\}$$

where $\|\mathbf{x}\| = \sqrt{\mathbf{x}^T\mathbf{x}}$ and $I_A(\cdot)$ is one if its argument is the set A and zero otherwise. Thus, for the full coefficient vector $\boldsymbol{\beta}$, the prior density (with respect to the product of sums of counting and Lebesgue measure) is

$$p(\boldsymbol{\beta}|\boldsymbol{\gamma}) = \left(\prod_{j\notin J_\gamma} I_{\{\mathbf{0}\}}(\boldsymbol{\beta}_j)\right)\left(\frac{\Gamma(m/2)\tau^m}{2\pi^{m/2}\Gamma(m)}\right)^{|\boldsymbol{\gamma}|}\exp\left\{-\tau\sum_{j\in J_\gamma}\left\|\boldsymbol{\beta}_j\right\|\right\}$$

where $J_{\boldsymbol{\gamma}} = \{k : \gamma_k = 1\}$.

The final piece of our hierarchical specification is a prior distribution on all models $\boldsymbol{\gamma}$. We let

$$p(\boldsymbol{\gamma}) \propto d_{\boldsymbol{\gamma}} q^{|\boldsymbol{\gamma}|}(1-q)^{p-|\boldsymbol{\gamma}|}$$

where $q \in (0,1)$ and $d_{\boldsymbol{\gamma}}$ is a measure of dependence among the $|\boldsymbol{\gamma}|$ variables in the model. The quantity $d_{\boldsymbol{\gamma}}$ in our specification is similar in spirit to the term $\det(\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{X}_{\boldsymbol{\gamma}})$ in the model formulation of YL05. In the formulation of YL05, the term $\det(\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{X}_{\boldsymbol{\gamma}})$ (where $\mathbf{X}_{\boldsymbol{\gamma}}$ is a matrix of predictors $\mathbf{x}_j = (x_{1j},\ldots,x_{nj})$ included in model $\boldsymbol{\gamma}$) is small for models with highly correlated and, therefore, redundant predictors.

In order to account for nonlinear relationships between predictors in our model, we let $d_{\boldsymbol{\gamma}}$ be the determinant of the matrix of Kendall's tau for all pairings of predictors in model $\boldsymbol{\gamma}$. More formally, let κ_{ij} be Kendall's tau for predictors $\mathbf{x}_i$ and $\mathbf{x}_j$ and let $\mathbf{K}$ be the matrix (κ_{ij}), then $d_{\boldsymbol{\gamma}} = \det(\mathbf{K}_{\boldsymbol{\gamma}})$, where $\mathbf{K}_{\boldsymbol{\gamma}}$ is a submatrix of $\mathbf{K}$ corresponding to nonzero elements of $\boldsymbol{\gamma}$. The term $d_{\boldsymbol{\gamma}}$ serves to penalize redundant models that have a high degree of dependence among the predictors.

We note that there have arisen two philosophies with regard to redundant predictors. The first is that if two predictors are highly related, then one or the other should be included in the model but not both. This philosophy is exemplified by the approach of YL05. The other philosophy is that if two predictors are highly related, then they should both be included in the model as a group (or excluded from the model as a group). This approach is exemplified by Zou and Hastie (2005) and Bondell and Reich (2008).

Thus, we explored a few other variations to the prior on $\boldsymbol{\gamma}$. For example, one method assumed an ordering to the predictors—for example, least costly to measure to most costly to measure—and penalized models that included "higher-cost" predictors that were highly correlated with excluded predictors of "lower cost". In practice (i.e., in simulation studies), however, these different priors had little effect on the outcome of the model selection routine. Therefore, we chose the simpler prior based on the determinant of the Kendall's τ matrix as described previously.

Also, measures of nonlinear association other than Kendall's τ may be used in constructing $d_{\boldsymbol{\gamma}}$. We considered measures of association based on summary measures of the empirical copula between pairs of predictors. Ultimately, however, we used the Kendall's τ matrix because it is computationally more efficient.

With the above model formulation, we can now write the joint posterior for $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ and $\boldsymbol{\gamma}$ as

$$p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}|\mathbf{y}) \propto (1-q)^p (2\pi\sigma^2)^{-n/2} d_{\boldsymbol{\gamma}} \left(\frac{q}{1-q} \frac{\Gamma(m/2)\tau^m}{2\pi^{m/2}\Gamma(m)} \right)^{|\boldsymbol{\gamma}|} \times \exp\left\{ -\frac{\|\mathbf{y} - \boldsymbol{\Psi}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}\|^2 + \lambda \sum_{j \in J_{\boldsymbol{\gamma}}}}{2\sigma^2} \right\}.$$

The marginal posterior probability for model $\boldsymbol{\gamma}$ can be obtained by integrating out $\boldsymbol{\beta}$

$$p(\boldsymbol{\gamma}|\mathbf{y}) \propto C(\mathbf{y}) B(\boldsymbol{\gamma}) \int_{\mathbb{R}^{mp}} \exp\left\{ -\frac{\|\mathbf{y} - \boldsymbol{\Psi}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}\|^2 + \lambda \sum_{j \in J_{\boldsymbol{\gamma}}} \|\boldsymbol{\beta}_j\|}{2\sigma^2} \right\} d\boldsymbol{\beta} \tag{3.2}$$

with

$$C(\mathbf{y}) = (1-q)^p (2\pi\sigma^2)^{-n/2},$$
$$B(\boldsymbol{\gamma}) = \left(\frac{q}{1-q} \frac{\Gamma(m/2)\tau^m}{2\pi^{m/2}\Gamma(m)} \right)^{|\boldsymbol{\gamma}|}.$$

The integral in (3.2) can be approximated using the Laplace's approximation. Following YL05, we let $\boldsymbol{\beta}^*_{\boldsymbol{\gamma}} = \text{argmin}_{\boldsymbol{\beta}_{\boldsymbol{\gamma}}} \|\mathbf{y} - \boldsymbol{\Psi}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}\|^2 + \lambda \sum_{j \in J_{\boldsymbol{\gamma}}} \|\boldsymbol{\beta}_j\|$. That is, $\boldsymbol{\beta}^*_{\boldsymbol{\gamma}}$ is

the group LASSO solution. Then, $\boldsymbol{\beta}_{\boldsymbol{\gamma}} = \boldsymbol{\beta}^*_{\boldsymbol{\gamma}} + \mathbf{u}$. Substituting this quantity into (3.2) gives the expression

$$C(\mathbf{y})B(\boldsymbol{\gamma})\exp\left\{-\frac{\min_{\boldsymbol{\beta}_{\boldsymbol{\gamma}}}\left(\|\mathbf{y}-\boldsymbol{\Psi}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}\|^2+\lambda\sum_{j\in J_{\boldsymbol{\gamma}}}\|\boldsymbol{\beta}_j\|\right)}{2\sigma^2}\right\}$$
$$\times\int_{\mathbb{R}^{mp}}\exp\left\{-\frac{1}{2\sigma^2}\left[\|\boldsymbol{\Psi}_{\boldsymbol{\gamma}}\mathbf{u}\|^2-2\mathbf{u}^T\boldsymbol{\Psi}^T_{\boldsymbol{\gamma}}\mathbf{y}^*+\lambda\sum_{j\in J_{\boldsymbol{\gamma}}}\left(\|\boldsymbol{\beta}^*_j+\mathbf{u}_j\|-\|\boldsymbol{\beta}^*_j\|\right)\right]\right\}d\boldsymbol{\beta} \quad (3.3)$$

where $\mathbf{y}^* = \mathbf{y}-\boldsymbol{\Psi}_{\boldsymbol{\gamma}}\boldsymbol{\beta}^*_{\boldsymbol{\gamma}}$ and $\underset{m\times 1}{\boldsymbol{\beta}^*_j}$ and $\underset{m\times 1}{\mathbf{u}_j}$ are the elements of $\boldsymbol{\beta}^*_{\boldsymbol{\gamma}}$ and $\mathbf{u}$ that correspond to the basis functions of the j^{th} predictor in model $\boldsymbol{\gamma}$.

Now let

$$f(\mathbf{u}) = \frac{1}{\sigma^2}\left[\|\boldsymbol{\Psi}_{\boldsymbol{\gamma}}\mathbf{u}\|^2-2\mathbf{u}^T\boldsymbol{\Psi}^T_{\boldsymbol{\gamma}}\mathbf{y}^*+\lambda\sum_{j\in J_{\boldsymbol{\gamma}}}\left(\|\boldsymbol{\beta}^*_j+\mathbf{u}_j\|-\|\boldsymbol{\beta}^*_j\|\right)\right].$$

Note that $f(\mathbf{u})$ is minimized at $\mathbf{u}=\mathbf{0}$ by construction. Further,

$$\left.\frac{\partial f(\mathbf{u})}{\partial\mathbf{u}\partial\mathbf{u}^T}\right|_{\mathbf{u}=\mathbf{0}} = \frac{1}{\sigma^2}(2\boldsymbol{\Psi}^T_{\boldsymbol{\gamma}}\boldsymbol{\Psi}_{\boldsymbol{\gamma}}+\lambda\mathbf{A}), \quad (3.4)$$

where

$$\underset{mk\times mk}{\mathbf{A}} = \begin{bmatrix} -\frac{\boldsymbol{\beta}^*_1\boldsymbol{\beta}^{*T}_1}{\|\boldsymbol{\beta}^*_1\|^3}+\frac{\mathbf{I}_m}{\|\boldsymbol{\beta}^*_1\|} & \underset{m\times m}{\mathbf{O}} & \cdots & \underset{m\times m}{\mathbf{O}} \\ \underset{m\times m}{\mathbf{O}} & -\frac{\boldsymbol{\beta}^*_2\boldsymbol{\beta}^{*T}_2}{\|\boldsymbol{\beta}^*_2\|^3}+\frac{\mathbf{I}_m}{\|\boldsymbol{\beta}^*_2\|} & \cdots & \underset{m\times m}{\mathbf{O}} \\ \vdots & \vdots & \ddots & \vdots \\ \underset{m\times m}{\mathbf{O}} & \underset{m\times m}{\mathbf{O}} & \cdots & -\frac{\boldsymbol{\beta}^*_k\boldsymbol{\beta}^{*T}_k}{\|\boldsymbol{\beta}^*_k\|^3}+\frac{\mathbf{I}_m}{\|\boldsymbol{\beta}^*_k\|} \end{bmatrix},$$

and $\mathbf{O}$ is a matrix of zeroes of the appropriate size.

The above equations can be used to apply Laplace's approximation to the quantity

in (3.3), which gives

$$
\begin{aligned}
p(\boldsymbol{\gamma}|\mathbf{y}) &\propto C(\mathbf{y})B(\boldsymbol{\gamma})\exp\left\{-\frac{\min_{\boldsymbol{\beta}_\gamma}\left(\|\mathbf{y}-\boldsymbol{\Psi}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}\|^2+\lambda\sum_{j\in J_\gamma}\|\boldsymbol{\beta}_j\|\right)}{2\sigma^2}\right\} \\
&\quad\times\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\exp\left\{-\frac{1}{2\sigma^2}f(\mathbf{u})\right\}d\mathbf{u} \\
&\approx C(\mathbf{y})B(\boldsymbol{\gamma})\exp\left\{-\frac{\min_{\boldsymbol{\beta}_\gamma}\left(\|\mathbf{y}-\boldsymbol{\Psi}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}\|^2+\lambda\sum_{j\in J_\gamma}\|\boldsymbol{\beta}_j\|\right)}{2\sigma^2}\right\} \\
&\quad\times\exp\left\{-\frac{1}{2}f(\mathbf{0})\right\}(2\pi)^{m|\boldsymbol{\gamma}|/2}\left|\frac{1}{2}\frac{\partial f(\mathbf{0})}{\partial\mathbf{u}\partial\mathbf{u}^T}\right|^{-1/2} \\
&= C(\mathbf{y})B(\boldsymbol{\gamma})\exp\left\{-\frac{\min_{\boldsymbol{\beta}_\gamma}\left(\|\mathbf{y}-\boldsymbol{\Psi}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}\|^2+\lambda\sum_{j\in J_\gamma}\|\boldsymbol{\beta}_j\|\right)}{2\sigma^2}\right\} \\
&\quad\times(2\pi)^{m|\boldsymbol{\gamma}|/2}\left|\tfrac{1}{\sigma^2}(\boldsymbol{\Psi}_{\boldsymbol{\gamma}}^T\boldsymbol{\Psi}_{\boldsymbol{\gamma}}+\tfrac{\lambda}{2}\mathbf{A})\right|^{-1/2} \\
&= (1-q)^p(2\pi\sigma^2)^{-n/2}d_{\boldsymbol{\gamma}}\left(\frac{q}{1-q}\frac{\Gamma(m/2)\tau^m}{2\pi^{m/2}\Gamma(m)}\right)^{|\boldsymbol{\gamma}|} \\
&\quad\times(2\pi)^{m|\boldsymbol{\gamma}|/2}\left|\tfrac{1}{\sigma^2}(\boldsymbol{\Psi}_{\boldsymbol{\gamma}}^T\boldsymbol{\Psi}_{\boldsymbol{\gamma}}+\tfrac{\lambda}{2}\mathbf{A})\right|^{-1/2} \\
&\quad\times\exp\left\{-\frac{\min_{\boldsymbol{\beta}_\gamma}\left(\|\mathbf{y}-\boldsymbol{\Psi}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}\|^2+\lambda\sum_{j\in J_\gamma}\|\boldsymbol{\beta}_j\|\right)}{2\sigma^2}\right\}.
\end{aligned}
\tag{3.5}
$$

3.3 Nonregular models

The approximation in (3.5) holds only if all components of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}^* = (\boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_p^*)$ are non-zero—else the derivative in (3.4) does not exist. This happens when the group LASSO sets one or more of the elements of $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_p^T)^T$ to $\underset{mp\times 1}{\mathbf{0}}$. YL05, in the context of linear models, define a "nonregular" model as any model where at least one coefficient is set to zero by the LASSO. YL05 show that in the special case of an orthogonal design matrix, for every nonregular model $\boldsymbol{\gamma}$ there exists a submodel of $\boldsymbol{\gamma}$, $\boldsymbol{\gamma}^*$, with only those predictors in $\boldsymbol{\gamma}$ whose coefficients were not set to zero by the LASSO, with higher asymptotic posterior probability. YL05 conclude, therefore, that nonregular models are not of interest.

Similarly, we define a nonregular additive model as any model $\boldsymbol{\gamma}$ for which $\boldsymbol{\beta}_j^* =$

$\underset{m\times 1}{\mathbf{0}}$ for at least one $j \in J_{\boldsymbol{\gamma}}$. For a given λ, any nonregular model is essentially equivalent to the submodel that has removed predictors whose coefficients were set to zero by the group LASSO. Therefore, we do not need calculate posterior probabilities for nonregular models.

3.4 Estimation of λ and σ^2

To select a value for λ we begin with an approach similar to the approach taken by YL05 for linear models. We start with the joint distribution of $\mathbf{y}$ and $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ conditional on all other model parameters

$$p(\mathbf{y}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma}, \lambda, \sigma^2) = (2\pi\sigma^2)^{-n/2} \left[\frac{\Gamma(m/2)}{2\Gamma(m)} \left(\frac{\lambda}{2\sigma^2\pi^{1/2}}\right)^m\right]^{|\boldsymbol{\gamma}|} \times \exp\left\{-\frac{\|\mathbf{y}-\boldsymbol{\Psi}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}\|^2+\lambda\sum_{j\in J_{\boldsymbol{\gamma}}}\|\boldsymbol{\beta}_j\|}{2\sigma^2}\right\}.$$

then integrate out $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ using Laplace's approximation as in (3.5)

$$\begin{aligned} p(\mathbf{y}|\boldsymbol{\gamma}, \lambda, \sigma^2) &= (2\pi\sigma^2)^{-n/2} \left[\frac{\Gamma(m/2)}{2\Gamma(m)} \left(\frac{\lambda}{2\sigma^2\pi^{1/2}}\right)^m\right]^{|\boldsymbol{\gamma}|} \\ &\quad \times \int_{\mathbb{R}^{m|\boldsymbol{\gamma}|}} \exp\left\{-\frac{\|\mathbf{y}-\boldsymbol{\Psi}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}\|^2+\lambda\sum_{j\in J_{\boldsymbol{\gamma}}}\|\boldsymbol{\beta}_j\|}{2\sigma^2}\right\} d\boldsymbol{\beta}_{\boldsymbol{\gamma}} \\ &\approx (2\pi)^{-n/2}(\sigma^2)^{-\left(\frac{n+m|\boldsymbol{\gamma}|}{2}\right)} \left(\frac{\Gamma(m/2)}{\Gamma(m)} 2^{(m+2)/2}\lambda^m\right)^{|\boldsymbol{\gamma}|} \\ &\quad \times \exp\left\{-\frac{\min_{\boldsymbol{\beta}_{\boldsymbol{\gamma}}}\left(\|\mathbf{y}-\boldsymbol{\Psi}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}\|^2+\lambda\sum_{j\in J_{\boldsymbol{\gamma}}}\|\boldsymbol{\beta}_j\|\right)}{2\sigma^2}\right\} \\ &\quad \times \left|\boldsymbol{\Psi}_{\boldsymbol{\gamma}}^T\boldsymbol{\Psi}_{\boldsymbol{\gamma}} + \tfrac{\lambda}{2}\mathbf{A}\right|^{-1/2} \end{aligned} \tag{3.6}$$

If we set $\boldsymbol{\gamma}$ in (3.6) equal to $\hat{\boldsymbol{\gamma}}_\lambda$, the model chosen by the group LASSO for a given λ, then maximizing (3.6) with respect to σ^2 gives

$$\hat{\sigma^2} = \frac{\left\|\mathbf{y}-\boldsymbol{\Psi}_{\hat{\boldsymbol{\gamma}}_\lambda}\boldsymbol{\beta}^*_{\hat{\boldsymbol{\gamma}}_\lambda}\right\|^2+\lambda\sum_{j\in J_{\hat{\boldsymbol{\gamma}}_\lambda}}\left\|\boldsymbol{\beta}^*_j\right\|}{n+|\hat{\boldsymbol{\gamma}}_\lambda|m} \tag{3.7}$$

Substituting (3.7) back into (3.6) and taking -2 times the natural logarithm of (3.6) gives

$$\begin{aligned} h(\lambda) = & -2|\hat{\boldsymbol{\gamma}}_\lambda|[\ln(m/2) - \ln(m)] + |\hat{\boldsymbol{\gamma}}_\lambda|(m+2)\ln 2 - 2m|\hat{\boldsymbol{\gamma}}_\lambda|\ln\lambda \\ & + (n + m|\hat{\boldsymbol{\gamma}}_\lambda|)\left[\ln\left(\frac{\left\|\mathbf{y} - \boldsymbol{\Psi}_{\hat{\boldsymbol{\gamma}}_\lambda}\boldsymbol{\beta}^*_{\hat{\boldsymbol{\gamma}}_\lambda}\right\|^2 + \lambda\sum_{j\in J_{\hat{\boldsymbol{\gamma}}_\lambda}}\left\|\boldsymbol{\beta}^*_{\hat{\boldsymbol{\gamma}}_\lambda}\right\|}{n + m|\hat{\boldsymbol{\gamma}}_\lambda|}\right)\right] \\ & + \ln|\boldsymbol{\Psi}^T_{\hat{\boldsymbol{\gamma}}_\lambda}\boldsymbol{\Psi}_{\hat{\boldsymbol{\gamma}}_\lambda} + \tfrac{\lambda}{2}\mathbf{A}|. \end{aligned} \tag{3.8}$$

An estimate of λ can then be found by minimizing (3.8) by a grid search, for instance.

Simulations have shown that choosing λ based on (3.8) results in overparametrized models. Therefore, we present an alternative BIC-type criterion for selecting λ.

YL06 use an estimate of the risk to select a value of the tuning parameter in their group LASSO algorithm. They obtain a value of λ by minimizing

$$C_p(\hat{\boldsymbol{\mu}}_\lambda) = \frac{\|\mathbf{y} - \hat{\boldsymbol{\mu}}\|^2}{\hat{\sigma^2}} - n + 2\widehat{\mathrm{df}}, \tag{3.9}$$

where $\hat{\boldsymbol{\mu}}_\lambda = \mathbf{X}\hat{\boldsymbol{\beta}}$ is an estimate of $E(\mathbf{Y}|\mathbf{X})$ that depends on λ through $\hat{\boldsymbol{\beta}}$, the group LASSO estimates and

$$\widehat{\mathrm{df}} = \sum_j^p I\left(\left\|\hat{\boldsymbol{\beta}}_j\right\| > 0\right) + \sum_j^p \frac{\left\|\hat{\boldsymbol{\beta}}_j\right\|}{\left\|\hat{\boldsymbol{\beta}}^{LS}_j\right\|}(p_j - 1), \tag{3.10}$$

where p_j is the degrees of freedom associated with the j^{th} predictor (in our case, $p_j = m$—the number of columns in the basis expansion of the j^{th} predictor) and $\hat{\boldsymbol{\beta}}^{LS}$ are the least squares estimates.

Using the criterion suggested by YL06 also results in overparametrized models. Thus, we use a slightly modified version of (3.9) that is similar in spirit to BIC to select a value for λ. The criterion is

$$C_p(\hat{\boldsymbol{\mu}}_\lambda) = \frac{\|\mathbf{y} - \hat{\boldsymbol{\mu}}\|^2}{\hat{\sigma^2}} - n + \log(n)\widehat{\mathrm{df}}, \tag{3.11}$$

where

$$\widehat{\mathrm{df}} = m\sum_j^p I\left(\left\|\hat{\boldsymbol{\beta}}_j\right\| > 0\right). \tag{3.12}$$

This selection criterion uses the more severe term $\log(n)$ in the penalty of C_p. Also, the second term on the right-hand side of (3.10) attenuates the penalty for inclusion of the j^{th} predictor in the model by adjusting the penalty through $\left\|\hat{\boldsymbol{\beta}}_j\right\| / \left\|\hat{\boldsymbol{\beta}}_j^{LS}\right\|$ which is always less than one. Our degrees-of-freedom term does not attenuate the penalty in this manner, because we are less concerned with the actual values of the coefficients and more concerned with the inclusion (or not) of each predictor in the model. Therefore, our chosen penalty penalizes each predictor the same amount for entering the model.

3.5 Simulation Study

To examine the performance of our method of computing approximate posterior probabilities, we conducted a simulation study, where all computation was executed in the `R` environment for statistical computing (R Development Core Team, 2007). We simulated data sets from a model with 5 "active" predictors and 5 "inactive" predictors

$$y_i = \sum_{j=1}^{10} f_j(x_{ij}) + e_i \tag{3.13}$$

$$\begin{aligned}
f_1(x) &= \exp(1.1x^3) - 2 \\
f_2(x) &= 2x - 1 \\
f_3(x) &= \sin(4\pi x) \\
f_4(x) &= \log\{(\exp(2) - 1)x + 1\} - 1 \\
f_5(x) &= -32(x - 0.5)^2/4 + 1 \\
f_j(x) &= 0 \qquad (\text{for } j = 6, \ldots, 10),
\end{aligned}$$

where $e_i \sim \mathsf{N}(0, 1)$, each f_j is scaled to have a range of 2 and lie in $[-1, 1]$ when $x \in [0, 1]$ (see Shively et al., 1999).

The x_{ij} variables were generated from three different sampling schemes. The first scheme—the independent x scheme—generates each x variable independently from

a uniform distribution, i.e $X_{ij} \sim U(0,1)$. The second scheme—the AR(1) scheme—generates the i^{th} row of the x matrix from a multivariate normal distribution with an AR(1) covariance structure, i.e. $(X_{i1}, \ldots, X_{i10}) \sim \mathsf{N}_{10}(\mathbf{0}, \mathbf{\Sigma})$ where $\mathbf{\Sigma}_{ij} = 0.7^{|i-j|}$. The third simulation scheme—the split-plot scheme—generates the i^{th} row of the x matrix from a multivariate normal with a zero mean vector and a "split-plot" style covariance structure, i.e. $(X_{i1}, \ldots, X_{i10}) \sim \mathsf{N}_{10}(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is a block diagonal matrix with the first block equal to

$$\mathbf{\Sigma}_1 = \begin{bmatrix} 1.16 & 1.00 & 1.00 \\ 1.00 & 1.16 & 1.00 \\ 1.00 & 1.00 & 1.16 \end{bmatrix}$$

the second block is a 2×2 submatrix of the first block and the third block is a 5×5 identity matrix. After each x matrix is generated, the columns of the x matrix are scaled to lie in $[0, 1]$.

For each of the x-matrix-generating schemes, we simulated 1,000 data sets and calculated approximate posterior model probabilities. We recorded the proportion of times that the model with the highest posterior probability was the true model. We also recorded the number of times the model with the highest posterior probability failed to include an "active" predictor (a "false negative") and each time the model incorrectly included and "inactive" predictor (a "false positive"). We recorded this same information for the models with the second and third highest posterior probability and for the model selected by the group LASSO alone. The results are presented in Tables 3.1, 3.2 and 3.3, where the column "true.mod" contains the proportion of times a method selected the true model, the column "false.neg" contains the average number active variables that were not in the selected model of a given method and the column "false.pos" contains the average number of inactive variables were included in the model of a given method.

Overall, the group LASSO and the model with the highest posterior probability were similar (almost exactly the same) in the number of "active" predictors that were selected. The model with the highest posterior probability tended to select less "inactive" predictors than the group LASSO. The difference between the two methods

Table 3.1: Simulation results for uncorrelated predictors. The column "true.mod" contains the proportion of times (out of 1000 simulated data sets) a method selected the true model. The column "false.neg" contains the average number of active variables that were not in the selected model of a given method. The column "false.pos" contains the average number of inactive variables that were included in the model of a given method. The rows marked by "Bayes" contain results for the model with the highest posterior probability. The rows marked "Bayes2" contain results for the model with the second highest posterior probability. Finally, the rows marked "Bayes 3" contain results for the model with the third highest posterior probability.

	n	true.mod	false.neg	false.pos
Bayes	100	0.269 (0.014)	0.187 (0.020)	1.261 (0.039)
Bayes2	100	0.228 (0.013)	0.526 (0.024)	0.729 (0.034)
Bayes3	100	0.029 (0.005)	0.761 (0.023)	0.946 (0.031)
G. LASSO	100	0.268 (0.014)	0.184 (0.020)	1.398 (0.047)
Bayes	200	0.426 (0.016)	0.000 (0.000)	0.993 (0.035)
Bayes2	200	0.297 (0.014)	0.421 (0.016)	0.458 (0.027)
Bayes3	200	0.000 (0.000)	0.715 (0.014)	0.772 (0.026)
G. LASSO	200	0.421 (0.016)	0.000 (0.000)	1.096 (0.040)
Bayes	500	0.509 (0.016)	0.000 (0.000)	0.763 (0.030)
Bayes2	500	0.292 (0.014)	0.062 (0.008)	0.287 (0.021)
Bayes3	500	0.001 (0.001)	0.024 (0.005)	0.300 (0.020)
G. LASSO	500	0.508 (0.016)	0.000 (0.000)	0.793 (0.032)

Table 3.2: Simulation results for AR(1) predictors. The column "true.mod" contains the proportion of times (out of 1000 simulated data sets) a method selected the true model. The column "false.neg" contains the average number of active variables that were not in the selected model of a given method. The column "false.pos" contains the average number of inactive variables that were included in the model of a given method. The rows marked by "Bayes" contain results for the model with the highest posterior probability. The rows marked "Bayes2" contain results for the model with the second highest posterior probability. Finally, the rows marked "Bayes 3" contain results for the model with the third highest posterior probability.

	n	true.mod	false.neg	false.pos
Bayes	100	0.191 (0.012)	1.094 (0.034)	0.442 (0.024)
Bayes2	100	0.026 (0.005)	1.729 (0.034)	0.296 (0.020)
Bayes3	100	0.030 (0.005)	1.750 (0.034)	0.313 (0.020)
G. LASSO	100	0.191 (0.012)	1.088 (0.034)	0.460 (0.026)
Bayes	200	0.363 (0.015)	0.491 (0.023)	0.460 (0.024)
Bayes2	200	0.067 (0.008)	1.298 (0.027)	0.288 (0.018)
Bayes3	200	0.067 (0.008)	1.287 (0.027)	0.300 (0.019)
G. LASSO	200	0.363 (0.015)	0.491 (0.023)	0.470 (0.025)
Bayes	500	0.573 (0.016)	0.160 (0.013)	0.376 (0.021)
Bayes2	500	0.169 (0.012)	0.916 (0.020)	0.134 (0.012)
Bayes3	500	0.041 (0.006)	1.048 (0.017)	0.274 (0.016)
G. LASSO	500	0.573 (0.016)	0.160 (0.013)	0.381 (0.021)

in this regard was not large, but the trend is persistent across all 3 simulations and across all sample sizes.

As we have pointed out before, the group LASSO (and other model selection procedures) select only one model. In contrast, the Bayesian procedure that we have derived gives multiple models with a degree of confidence in each as measured by the posterior probability. The benefit of this feature is best demonstrated in the case of independent predictors. For a sample size of 100, the model with the highest posterior probability was the true model in 26.9% of the simulations. However, the model with the second highest posterior probability was the true model in 22.8% of the posterior

Table 3.3: Simulation results for split-plot predictors. The column "true.mod" contains the proportion of times (out of 1000 simulated data sets) a method selected the true model. The column "false.neg" contains the average number of active variables that were not in the selected model of a given method. The column "false.pos" contains the average number of inactive variables that were included in the model of a given method. The rows marked by "Bayes" contain results for the model with the highest posterior probability. The rows marked "Bayes2" contain results for the model with the second highest posterior probability. Finally, the rows marked "Bayes 3" contain results for the model with the third highest posterior probability.

	n	true.mod	false.neg	false.pos
Bayes	100	0.109 (0.010)	1.576 (0.035)	0.289 (0.020)
Bayes2	100	0.000 (0.000)	2.203 (0.034)	0.254 (0.019)
Bayes3	100	0.006 (0.002)	2.125 (0.035)	0.201 (0.017)
G. LASSO	100	0.109 (0.010)	1.573 (0.035)	0.291 (0.020)
Bayes	200	0.212 (0.013)	1.017 (0.029)	0.286 (0.019)
Bayes2	200	0.000 (0.000)	1.939 (0.029)	0.240 (0.018)
Bayes3	200	0.012 (0.003)	1.772 (0.029)	0.222 (0.017)
G. LASSO	200	0.212 (0.013)	1.013 (0.029)	0.296 (0.021)
Bayes	500	0.488 (0.016)	0.433 (0.020)	0.230 (0.017)
Bayes2	500	0.030 (0.005)	1.367 (0.022)	0.164 (0.014)
Bayes3	500	0.045 (0.007)	1.254 (0.021)	0.146 (0.013)
G. LASSO	500	0.488 (0.016)	0.433 (0.020)	0.230 (0.017)

simulations. Thus, the model with the highest or second highest posterior probability was the true model in close to 50% of the simulations. This trend also holds across all simulations, although it is more dramatic in the case of independent predictors.

3.6 Illustration of Method on Real Data

We demonstrate our method on the NCAA data set from Mangold, Bean, and Adams (2003). We use a reduced version of the full data set where three observations with missing data have been removed from the 97 total observations (Boos and Ste-

fanski, 2008). The data contain six-year graduation rates at Division I universities and 19 covariates which could affect graduation rates. Table 3.4 contains a description of each predictor in the data set.

Table 3.4: Selected models for the NCAA data set using the method presented in this chapter (in the Bayes column), the LASSO and the COSSO. Descriptions of the predictors in the NCAA data set are from Boos and Stefanski (2008).

Variables	Bayes	LASSO	COSSO	Description
top10		✓		% Students in top 10% HS
act25	✓	✓	✓	ACT composite 25^{th}
oncampus	✓	✓	✓	% Students living on campus
ft.grad	✓	✓		% First-time undergraduates
size		✓		Total enrollment/1000
tateach		✓	✓	% Courses taught by TAs
bbindex				Composite of basketball ranking
tuition				In-state tuition/1000
board		✓		Room and board/1000
attend			✓	Avg basketball home attendance
full.sal	✓	✓	✓	Full professor salary
sf.ratio			✓	Student to faculty ratio
white		✓	✓	% White
ast.sal	✓		✓	Assistant professor salary
pop				Population of city where located
phd		✓	✓	% Faculty with PhD
accept	✓		✓	Acceptance rate
l.pct			✓	% Receiving loans
outstate			✓	% Out of state

We fit our Bayesian model to the data set and, for comparison, also fit the LASSO and the COSSO (where tuning parameters were selected using 10-fold cross validation). The model selected by each method is listed in Table 3.4.

The LASSO and the COSSO were both more "liberal" in this example i.e. they

both selected models with a larger number of predictors than the Bayesian method. The LASSO selected 10 of 19 and the COSSO selected 12 of 19 predictors compared to the Bayesian method that selected 6 of 19 predictors. Also, the Bayesian fit included one variable not included in the COSSO model (`ft.grad`) and two variables not included in the LASSO.

3.7 Conclusion

We have presented a method to obtain posterior model probabilities in additive models using Laplace's approximation and the group LASSO. Our method gives a measure of model variability, a feature which most other variable selection methods do not have. Also, our method avoids MCMC methods which can be computationally expensive. Additionally, our method takes advantage of the generality of the additive model formulation and does not assume the restrictions in the classical linear model.

Chapter 4

A Variable Selection Approach to Bayesian Monotonic Regression with Bernstein Polynomials

4.1 Introduction

Much of applied statistics consists of determining the relationship between several variables. This relationship is often represented by assuming a response variable is a stochastic function of one or more predictor variables. Substantive prior information often exists that dictates the functional relationship between the response and predictors should exhibit certain shape restrictions. Examples of this can be found in nearly all areas of applied statistics. In microeconomics, the assumption of diminishing marginal returns of factor inputs restricts the production possibilities frontier to be concave. Similarly, in the theory of the consumer, the assumption of diminishing marginal rate of substitution restricts indifference curves to be convex (Nicholson, 1992). In biomedical applications, dose response models are assumed to be non-decreasing with the possibility that the dose-response relationship is flat over certain regions. In distance sampling (Buckland, Anderson, Burnham, Laake, Borchers, and Thomas, 2001), the probability that an observer collecting data will detect a hidden object is assumed to be monotonically decreasing in the distance

between the object and the observer. In survival analysis, the survivor function is non-increasing. In reliability and survival analysis, the hazard rate and the failure rate are often assumed to have a "U" shape (Reboul, 2005). In actuarial studies, the mean residual life function $m(x)$ must satisfy the shape restrictions $m'(x) + 1 \geq 0$ and $m(x) \geq 0$.

Shape restricted regression has a long history in the literature. Studies on shape-restricted regression begin with Hildreth (1954) who proposes a method for estimating points along a concave function and Brunk (1955) who presents maximum likelihood estimators for estimating monotone restrictions on parameters. Barlow et al. (1972) propose the pool-adjacent-violators (PAV) algorithm for fitting an isotonic regression function, which averages over adjacent observations that do not satisfy the isotonicity restraint. Other more recent studies include Gallant and Golub (1984), Friedman and Tibshirani (1984), Ramsay (1988), Mukerjee (1988), Lavine and Mockus (1995), Mammen, Marron, Turlach, and Wand (2001), and Hall and Huang (2001) to name a few.

The approach we take in this chapter uses Bernstein polynomials to approximate the unknown regression function. Stadtmüller (1986) first used Bernstein polynomials to approximate an unknown regression function, and Tenbusch (1997) generalized this method to the case of more than one predictor. Brown and Chen (1999) adapt the weights of a Bernstein polynomial smoother using kernels of a beta density to obtain a regression function estimate.

The Bernstein polynomial is a natural fit for shape-restricted regression. Consider a Bernstein polynomial of degree M for the continuous function $f(\cdot)$

$$g(x) = \sum_{k=0}^{M} f(k/M) \binom{M}{k} x^k (1-x)^{M-k},$$

for $x \in [0,1]$. If we let $\beta_k = f(k/M)$, then it is easy to see that restricting the coefficients β_k is equivalent to restricting the function values at the $(k/M)^{th}$ quantile of the predictor variable. We note that by a suitable change of variables, we can restrict the range of x to $[0,1]$ without loss of generality. For the rest of this chapter, we assume that $x \in [0,1]$.

Also, the form of the derivatives of $g(\cdot)$ can be useful in determining appropriate shape restrictions for a particular application. The ℓ^{th} derivative of $g(\cdot)$ can be written

$$g^{(\ell)}(x) = M(M-1)\cdots(M-\ell+1)\sum_{k=0}^{M-\ell}(\nabla^{(\ell)}\beta_k)\binom{M-\ell}{k}x^k(1-x)^{M-\ell-k}$$

where $\nabla^{(1)}\beta_k = \beta_{k+1} - \beta_k$, $\nabla^{(2)}\beta_k = \nabla^{(1)}\beta_{k+1} - \nabla^{(1)}\beta_k = \beta_{k+2} - 2\beta_{k+1} + \beta_k$, and $\nabla^{(\ell)}\beta_k = \nabla^{(\ell-1)}\beta_{k+1} - \nabla^{(\ell-1)}\beta_k$. The first and second derivatives simplify to

$$g'(x) = M\sum_{k=0}^{M-1}(\beta_{k+1} - \beta_k)\binom{M-1}{k}x^k(1-x)^{M-1-k}$$
$$g''(x) = M(M-1)\sum_{k=0}^{M-2}(\beta_{k+2} - 2\beta_{k+1} + \beta_k)\binom{M-2}{k}x^k(1-x)^{M-2-k},$$

Because the quantities $\binom{M}{k}x^k(1-x)^{M-k}$ are non-negative for all $x \in [0,1]$, it is easy to see how to construct restrictions on the coefficients that will impose restrictions on the first and second derivatives.

Chak, Madras, and Smith (2005) introduce the idea of using Bernstein polynomials for shape-restricted regression. They give restrictions on the coefficients β_k that enforce the following shapes:

- Nonnegativity: $\beta_k \geq 0$ for all k
- Isotonicity: $\beta_k \leq \beta_{k+1}$ for $k = 0, \ldots, M-1$
- Concavity: $\beta_{k+1} - 2\beta_k + \beta_{k-1} < 0$ for $k = 1, \ldots, M-1$

In addition to monotonicity, Chang, Chien, Hsiung, Wen, and Wu (2007) give restrictions on the coefficients that enforce the following shapes:

- Unimodality: $\beta_0 = \cdots = \beta_{r_1} < \beta_{r_1+1} \leq \cdots \leq \beta_{r_2}$ and $\beta_{r_2} \geq \beta_{r_2+1} \geq \cdots \geq \beta_{r_3} > \beta_{r_3+1} = \cdots = \beta_M$ for $0 \leq r_1 < r_2 < r_3 \leq M$
- Unimodal concavity: $\beta_1 - \beta_0 > 0$, $\beta_M - \beta_{M-1} < 0$, and $\beta_{k+1} - 2\beta_k + \beta_{k-1} \leq 0$ for $k = 1, \ldots, M-1$

We take a Bayesian approach to estimating a shape-restricted regression function. Other authors have developed Bayesian techniques. Neelon and Dunson (2004) use a piecewise linear function with constraints on the slope parameters to ensure the regression function is isotonic. Holmes and Heard (2003) use an unconstrained, piecewise-constant function with random knots to model the regression function. To enforce monotonicity, they develop a reversible-jump MCMC algorithm (Green, 1995) to sample from the posterior and throw away posterior draws that do not satisfy the constraint. Such an approach may require a huge amount of sampling as the constraint can be violated making the algorithm inefficient. Also, the convergence of a reversible-jump MCMC algorithm may not be easy to diagnose.

The approach we take in this chapter is to construct a prior distribution for the coefficients $\beta_0, \ldots, \beta_M$ in g that satisfies the requisite shape restrictions. In this chapter we present a prior on the Bernstein coefficients for the case of monotonic regression, although the idea can be extended to other shape restrictions and to generalized linear models (e.g. logistic or probit regression, and Poisson regression). We show that through a simple reparametrization of the β_k coefficients in the Bernstein polynomial expansion, the problem of monotonic regression is equivalent to the problem of variable selection (Section 4.2). By adapting the variable selection model of Geweke (1996), we obtain a monotonic regression fit through Gibbs sampling (Sections 4.3 and 4.4). We demonstrate the usefulness of this prior in several simulated examples (Section 4.5). We conclude the chapter with an analysis of two real data sets and some final comments (Sections 4.6 and 4.7).

4.2 Prior on Bernstein Coefficients

4.2.1 A simple prior

Consider, again, the Bernstein polynomial approximation to the unknown regression function

$$g(x, \boldsymbol{\beta}) = \sum_{k=0}^{M} \beta_k \binom{M}{k} x^k (1-x)^{M-k} = \sum_{k=0}^{M} \beta_k b_M(x, k). \tag{4.1}$$

Our task is to define a prior on $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_M) \in \mathbb{R}^{M+1}$ such that $\beta_0 \leq \cdots \leq \beta_M$. A first attempt (see Chang et al., 2007) would be to define a continuous prior (e.g., a normal distribution) on auxiliary variables $U_0, \ldots, U_M$, then set $\beta_0 = U_{(0)}, \ldots, \beta_M = U_{(M)}$, where $U_{(0)}, \ldots, U_{(M)}$ are order statistics for the U_i's. This approach, however, does not give positive probability to events such as $\beta_k = \beta_{k+1}$—events which can be crucial when modeling flat portions of a regression function.

This difficulty is easily demonstrated using a simulated example. Consider the model

$$Y_i = f(X_i) + \varepsilon_i,$$

where $i = 1, \ldots, 50$, $f(x) = 1$ for all $x \in [0, 1]$, $X_i \sim \mathcal{U}(0, 1)$, and $\varepsilon_i \sim \mathsf{N}(0, 0.1^2)$. We simulated one data set from this model and fitted the model

$$\begin{aligned} Y_i|\boldsymbol{\beta}, \sigma^2 &\sim \mathsf{N}\big(g(x_i, \boldsymbol{\beta}), \sigma^2\big) \\ \sigma^{-2} &\sim \mathsf{Gam}(0.01, 0.01) \\ \beta_0 = U_{(0)} \cdots \beta_M &= U_{(M)} \\ U_0, \ldots, U_M &\sim \mathsf{N}\big(\alpha, \nu^2\big) \\ \alpha &\sim \mathsf{N}(0, 1) \\ \nu^{-2} &\sim \mathsf{Gam}(1.0, 1.0) \end{aligned}$$

to the data, where $\mathsf{Gam}(a, b)$ is a gamma distribution with mean $\frac{a}{b}$ and $\mathsf{N}(m, s^2)$ is the normal distribution with mean m and variance s^2, and $M = 25$. The model was fitted using an MCMC algorithm constructed by the `WinBUGS` software (Spiegelhalter, Thomas, and Best, 1999) with 3 chains, 5,000 draws per chain, after 1,000 draws of burn-in, and a plot of the model fit is displayed in Figure 4.1. The thick, black, solid line is the posterior median from the above model and the thin, black, solid lines are pointwise 95% credible bands. For reference, the solid white line is the true regression function. Also, the dashed, thick, black line is the posterior median and the gray shaded area represent pointwise confidence bands from the model we describe in this chapter. It is clearly apparent that the order-statistic prior is inadequate to capture the flatness of the regression function. The prior has essentially forced

the estimates of the regression function to have positive slope, whereas our method captures the true regression function in the credible bands.

4.2.2 A Reparametrization

In order to define a prior for shape-restricted regression that allows for events $\beta_{k+1} = \beta_k$ or, more generally, events $\beta_{k_1} = \beta_{k_1+1} = \cdots = \beta_{k_2}$ for $0 \leq k_1 \leq k_2 \leq M$, we reparametrize $g(x, \boldsymbol{\beta})$ by taking the differences between adjacent coefficients. In other words, we obtain new parameters $\mathbf{u} = (u_0, u_1, \ldots, u_M)$ where $u_0 = \beta_0$ and $u_k = \beta_k - \beta_{k-1}$ for $k = 1, \ldots, M$. Using this reparametrization we obtain

$$\begin{aligned}
g(x, \boldsymbol{\beta}) &= \sum_{j=0}^{M} \beta_j b_M(x, j) \\
&= \sum_{j=0}^{M} (\sum_{k=0}^{j} u_k) b_M(x, j) \\
&= \sum_{k=0}^{M} u_k \sum_{j=k}^{M} b_M(x_i, j) \\
&= \sum_{k=0}^{M} u_k w_M(x, k) \\
&= u_0 + u_1 w_M(x, 1) + \cdots + u_M w_M(x, M) \\
&\qquad \text{(because } \textstyle\sum_{j=0}^{M} b_M(x_i, j) = 1\text{)} \\
&= g(x, \mathbf{u})
\end{aligned}$$

Note that, under the reparametrization, the Bernstein polynomial regression function has simplified to a linear combination of the parameters $u_0, \ldots, u_M$ and "covariates" $w_M(x, 1), \ldots, w_M(x, M)$, where $w_M(x, 1), \ldots, w_M(x, M)$ can be computed with the

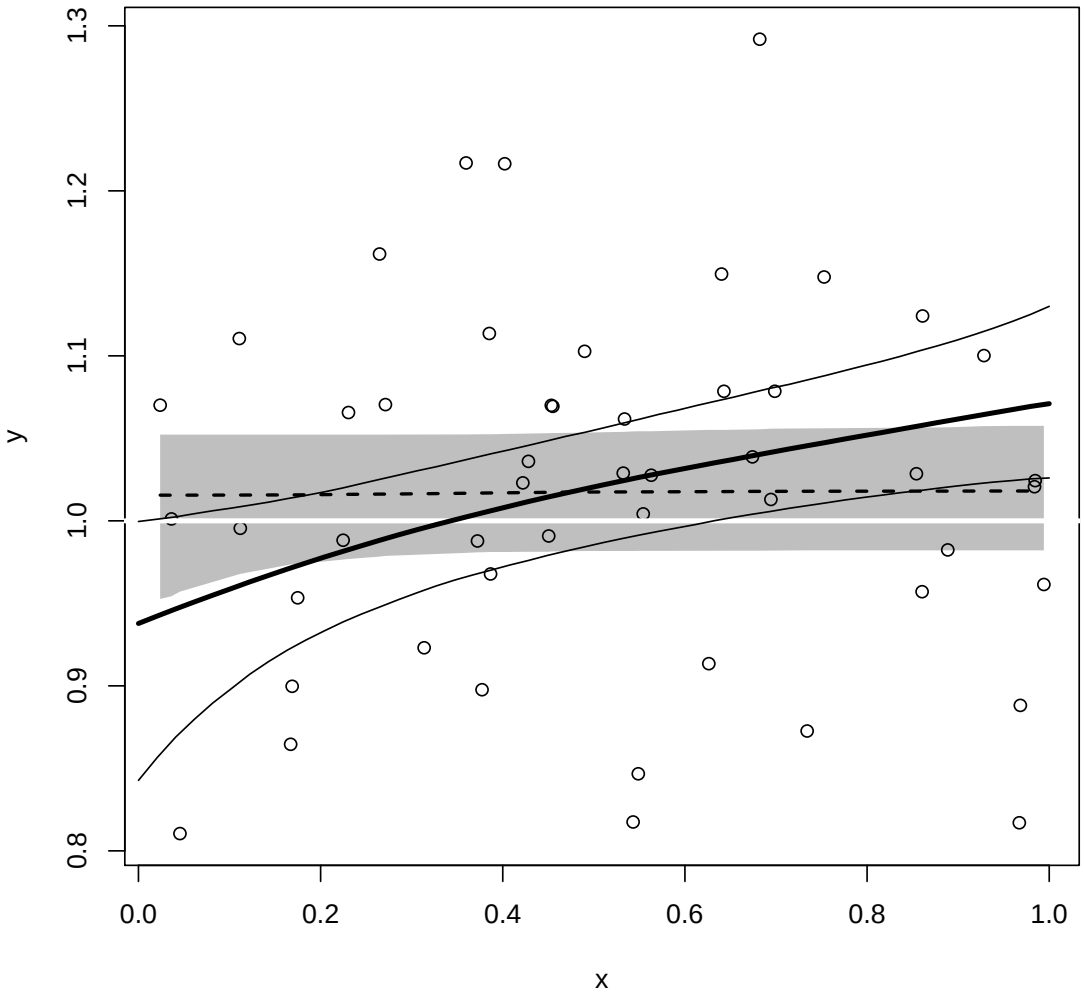

Figure 4.1: Plot of simulated data from a uniformly flat regression function with a Bayesian model that does not allow for flat portions of the regression curve.

incomplete beta function

$$\begin{aligned} w_M(x,k) &= \sum_{j=k}^{M} b_M(x,j) \\ &= F_{B(k,M-k+1)}(x) \\ &= \int_0^x \frac{v^{K-1}(1-v)^{(M-k)}}{B(k,M-k+1)} dv, \end{aligned}$$

where $F_{B(k,M-k+1)}(\cdot)$ is the cdf of a Beta random variable with parameters k and $M-k+1$ and $B(a,b)$ is the beta function. Under this parametrization, if $u_k = 0$ then $\beta_k = \beta_{k-1}$, and if $u_k > 0$ then $\beta_k > \beta_{k-1}$.

4.2.3 Variable Selection and Monotonic Regression

The reparametrization above suggests some form of variable selection may be used to determine the fit of the monotonic regression function. A variable selection prior on $U_1, \ldots, U_p$ would give a prior on the $\beta_1, \ldots, \beta_p$ with positive probability of adjacent β's being exactly equal, which would satisfy the monotonicity constraint while allowing for flat portions of the regression function.

The literature contains several studies on Bayesian variable selection (e.g., Mitchell and Beauchamp (1988), George and McCulloch (1993), George and McCulloch (1997), George and Foster (2000), Cripps, Kohn, and Nott (2006), Casella and Moreno (2006)). In a situation similar to our own, Dunson (2005) reparametrizes the regression function for count data by using ratios of adjacent function values. Dunson then uses a mixture of a point-mass at one and a gamma distribution truncated to be greater than one to impose a monotonic constraint to the regression function.

The approach of Geweke (1996) is particularly germane to the problem of monotonic regression as outlined in the previous section. To perform variable selection in the linear model, Geweke uses a prior for each regression coefficient that is a mixture of a point mass at zero and a truncated normal distribution. This mixture prior is a natural fit for our situation in which the prior for U_i must satisfy two requirements—events $U_k = 0$ must have positive probability and $U_k \geq 0$ with probability one.

4.3 Model Specification

By adapting the variable selection approach of Geweke (1996), we can specify the complete Bayesian model. The sampling density for the vector $\mathbf{y}$ of n observations is

$$p(\mathbf{y}|\mathbf{u},\sigma^2) = (2\pi\sigma^2)^{-n/2}\exp\left\{-\frac{\sum_{i=1}^{n}(y_i - \mathbf{w}_i^T\mathbf{u})^2}{2\sigma^2}\right\}.$$

where $\mathbf{w}$ is as defined previously. We give the parameter u_0, which is not restricted to be positive or zero, a normal prior with mean m_0 and variance s_0^2

$$p(u_0) = (2\pi s_0^2)^{-1/2}\exp\left\{-\frac{(u_0 - m_0)^2}{2s_0^2}\right\}.$$

As has been done by previous authors, we condition the prior for $u_1, \ldots, u_M$ on binary latent indicator variables $\gamma_1, \ldots, \gamma_M$. We define the k^{th} indicator variable to be one if $u_k = 0$ and zero otherwise. The prior for u_k conditioned on the latent indicator γ_k can then be written

$$p(u_k|\tau^2,\gamma_k) = \gamma_k \mathbb{1}_{\{0\}}(u_k) + (1-\gamma_k)2(2\pi\tau^2)^{-1/2}\exp\left\{-\frac{u_k^2}{2\tau^2}\right\}\mathbf{1}_{(0,\infty)}(u_k)$$

where the function $\mathbb{1}_A(\cdot)$ is one if its argument is in the set A and zero otherwise. Each γ_k is given a Bernoulli prior with parameter p_γ

$$p(\gamma_k) = p_\gamma^{\gamma_k}(1-p_\gamma)^{1-\gamma_k},$$

and p_γ is given a uniform prior on $(0,1)$.

To complete the prior specification, we give the variance parameters σ^2 and τ^2 conditionally conjugate inverse gamma priors

$$p(\sigma^2) \propto (\sigma^2)^{-(a_\sigma+1)}\exp\left\{-b_\sigma/\sigma^2\right\},$$
$$p(\tau^2) \propto (\tau^2)^{-(a_\tau+1)}\exp\left\{-b_\tau/\tau^2\right\}.$$

The final step in the model specification requires choosing values for a_σ, b_σ, a_τ, b_τ, and M. For values of a_σ and b_σ,we use small, "noninformative" values (e.g., 1 or 0.1). Similarly, for a_τ and b_τ we recommend setting both hyperparameters equal to

1. In practice, the values of a_τ and b_τ have very little effect on the quality of the fit of the monotonic regression.

To choose a value for M, we take an approach similar to the approach taken by Neelon and Dunson (2004) for choosing the number of knots. We choose a "large" value for M—possibly as large as the number of unique values of the predictor variable—and let the variable selection procedure remove redundant columns of the Bernstein expansion. This approach makes the procedure adaptive (in a loose sense) in that the procedure "chooses" a value of M by removing redundant columns of the Bernstein basis expansion. This approach is also similar to Smith and Kohn (1996) who used variable selection on the coefficients in a basis expansion to fit an unconstrained, nonparametric regression curve.

4.4 Posterior Sampling

A Markov Chain Monte Carlo approach can be used to sample from the joint posterior distribution of the parameters in the regression model (Tierney, 1994). A Gibbs sampling scheme (Geman and Geman, 1984; Gelfand and Smith, 1990) can proceed by following the algorithm specified in Geweke (1996), except that our algorithm includes two additional Gibbs updates for parameters τ^2 and p_γ. Computational details, including an outline of the Gibbs sampler, are included in the appendices.

4.5 A Simulation Study

To test our monotonic regression method, we ran several simulations and compared the performance of our approach with competing monotonic regression methods. We generated 50 data sets from each of four models $y_i = h_s(x_i) + \epsilon_i$ (for $s = 1, \ldots, 4$ and $i = 1, \ldots, 100$), where $\epsilon_i \sim \mathsf{N}(0, 0.1^2)$, $x_i \sim UnifDist01$. The regression function $h_s(\cdot)$ took on the following values

Flat:	$h_1(x) = 1$
Linear:	$h_2(x) = x$
Flat and Nonlinear:	$h_3(x) = I(x > 0.5)\sqrt{2x-1}$
Wavy:	$h_4(x) = \frac{\sin(3\pi x)+3\pi x}{3\pi}$.

All functions were chosen to have a range of one, and the standard deviation of ϵ_i was chosen to be one-tenth of the range.

We compared our method with three other methods that are accessible as functions in the `R` statistical software (R Development Core Team, 2007; Ihaka and Gentleman, 1996). The first is the local regression method as implemented by the `loess` function (Cleveland, Grosse, and Shyu, 1992). The second is the isotonic regression estimator of Barlow et al. (1972) as implemented in the `isoreg` function. The third is the monotonic regression estimator of Dette, Neumeyer, and Pilz (2006) as implemented in the `monreg` function of the `monreg` package (Pilz and Titoff, 2005).

The `loess` method requires the user to specify one tuning parameter. We used five-fold cross-validation to choose the value for this parameter. The `monreg` function requires the specification of two tuning parameters—λ_d and λ_r. As per the recommendation of Dette et al. (2006, section 4.1), we set $\lambda_d = \lambda_r^2$. We then used five-fold cross-validation to choose the value of λ_r.

To fit our Bayesian model to each data set, we used a Bernstein polynomial expansion of order 40 (i.e., $M = 40$). We ran the MCMC algorithm for 110,000 iterations, discarded the first 10,000 iterations as burn-in, and kept every 10th iteration of the resulting chain (for a final chain length of 10,000).

After fitting each regression method to a simulated data set, we evaluated the quality of the fit by computing fitted values for each method over a grid of 100 equally spaced "x-values" on the unit interval. The absolute difference between the true function and the fitted value was standardized (by dividing by the true standard deviation of 0.1), and the mean over the grid of 100 x-values was recorded for each method.

The results of the simulation are contained in Table 4.1 and summarized graphically in Figure 4.2. In general, the Bayesian method performed very well relative to

Table 4.1: Simulation results for comparison of the Bayesian procedure and competing monotonic regression methods. Values in the table are means of mean absolute standardized deviation of fitted values from the true function values at an equally-spaced grid of 100 points. Standard errors are in parentheses.

	bayes	loess	isoreg	monreg
Flat	0.25	0.41	0.36	0.47
	(0.014)	(0.013)	(0.014)	(0.007)
Linear	0.48	0.42	0.57	0.43
	(0.008)	(0.012)	(0.006)	(0.010)
Flat/Sqrt	0.48	0.51	0.52	0.50
	(0.008)	(0.008)	(0.009)	(0.007)
Sine	0.46	0.48	0.56	0.50
	(0.008)	(0.010)	(0.007)	(0.009)

the other methods. In three of the four simulations, the Bayesian method had the lowest mean standardized absolute deviation from the true function. As expected, the Bayesian method performs substantially better than the other three methods when the true regression function is flat. The Bayesian method also outperforms the other methods when the underlying function is the flat/square root function or the sine function, although the `monreg` function is a close competitor. Predictably, the Bayesian method did not perform as well when the true regression function had no flat portions (i.e., was linear), however, it was not the worst method in this case.

4.6 Illustrations Using Real Data Sets

We try our method on two real data sets. The first example uses data with a continuous response. The second example uses data with a binomial response.

4.6.1 Monotone Regression Model for Continuous Response

The first example uses data from Ramsay (1998) on the growth of a 10-year-old boy. The data contain 83 height measurements over a period of 312 days. Growth

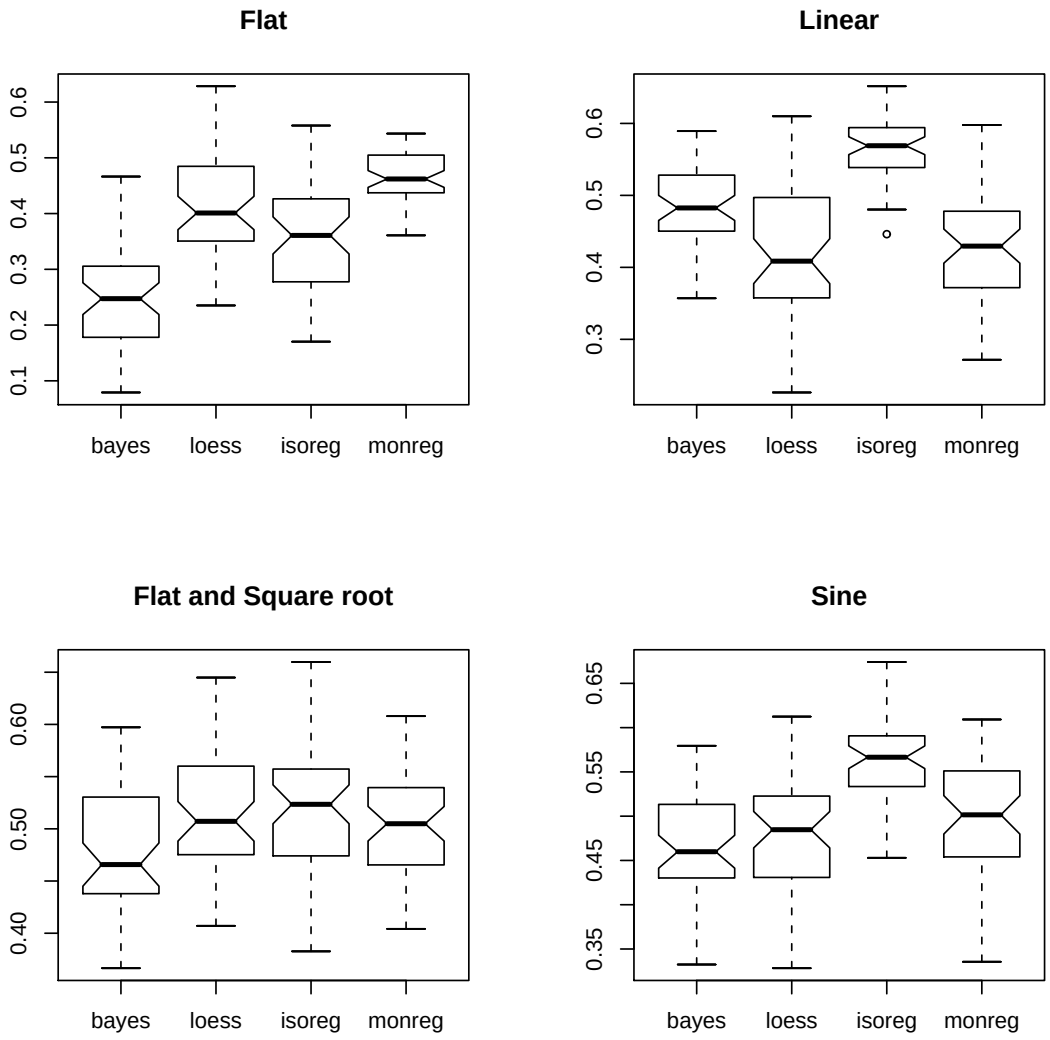

Figure 4.2: Boxplots of the mean of the standardized, absolute deviations of fitted values from the true function values at an equally-spaced grid of 100 points.

curves should exhibit monotonicity and allow for relatively flat regions consistent with the belief that growth occurs in spurts rather than continuously over time (Thalange, Foster, Gill, Price, and Clayton, 1996).

To fit the model with our Bayesian procedure, we rescaled the time (day) variable to the interval $[0, 1]$. We fit the Bayesian model using Bernstein polynomials of order 80. We ran the MCMC algorithm of Section 4.4 three times in order to assess convergence using the Gelman-Rubin (GR) diagnostics (Gelman and Rubin, 1992). Each MCMC run consisted of 160,000 iterations with 10,000 iterations discarded as burn-in and only every tenth iteration kept for plotting and analysis (i.e., the final MCMC size was 15,000 per MCMC run). The GR diagnostics for each parameter were all equal to one (up to two decimal places), except for a handful of parameters which had GR diagnostics very nearly one (the maximum GR diagnostic from this group was 1.02). Figure 4.3 contains the fit from the Bayesian model along with fits from `isoreg`, `loess`, and `monreg`. All fits (with the exception of `isoreg`) appear relatively smooth. A small portion of the `monreg` fit is not covered by the pointwise 95% credible bands from the Bayes method.

4.6.2 Monotone Regression Model for Discrete Response

The Bayesian variable selection method for monotonic regression can easily be extended to generalized linear models. For example, the sampling probability of a binomial observation is

$$p(y_i|n_i, p_i) = \binom{n_i}{y_i} p_i^{y_i}(1-p_i)^{n_i-y_i}$$

where n_i is the number of trials and p_i is the probability of success on given trial. Using the logit transformation to model p_i we have

$$\log\left(\frac{p_i}{1-p_i}\right) = \mathbf{w}_i^T\mathbf{u}$$

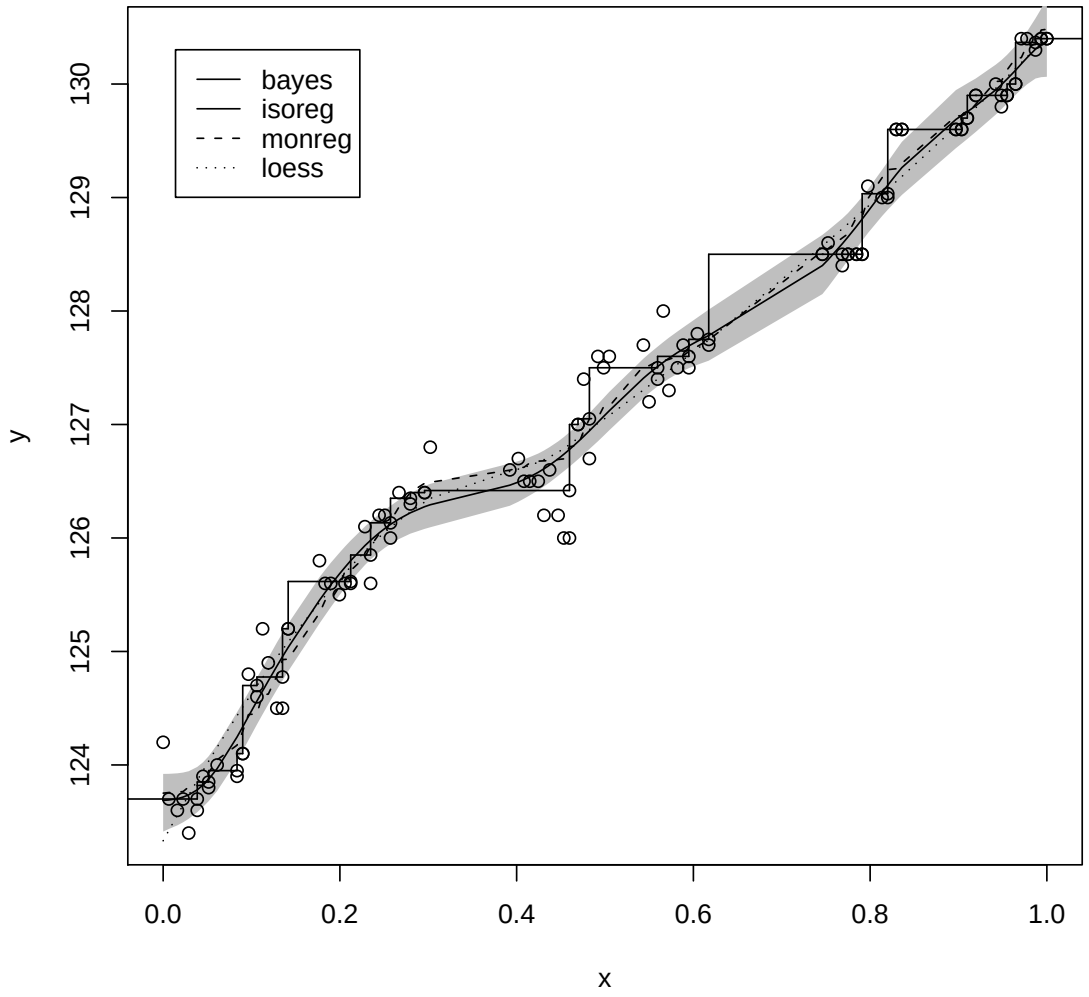

Figure 4.3: Plot of child height versus day (standardized to fall in the range $[0, 1]$). The smooth, solid line is the Bayesian fit, the step function solid line is the `isoreg` fit, the dashed line is the `monreg` fit and the dotted line is the `loess` fit. The shaded region indicates 95% pointwise credible (confidence) region for the Bayesian fit.

where u_0, $\mathbf{w}$, and $\mathbf{u}$ are as previously defined in Section 4.2.2.

A Gibbs sampler for this model can be constructed using the `WinBUGS` software. We fit this model to the Downs syndrome data of Geyer (1991). The data consist of the number of Downs syndrome births out of all births recorded by the British Columbia Health Surveillance Registry according to the mother's age category. For these data, it is not unreasonable to assume that the probability of a Downs syndrome birth increases monotonically with the age of the birth mother. The fit of the model to this data is plotted in Figure 4.4 with 95% confidence bands.

4.7 Conclusion

In this chapter, we have presented a connection between monotonic regression and variable selection. This connection opens up new ways to fit monotonic regression models. We presented one Bayesian approach in this chapter and were able to exploit the already existing MCMC techniques in variable selection to derive a Gibbs sampler for our monotonic regression method. We have demonstrated the effectiveness of our method in a simulation comparison of other competing methods. Finally, we have demonstrated the flexibility of our method by fitting our model to two very different data sets consisting of continuous and discrete-valued response variables. One immediate advantage or our proposed Bayesian method is that posterior interval bands can be obtained in a straightforward manner without appealing to any sort of asymptotic-based inference for both continuous and discrete-valued responses.

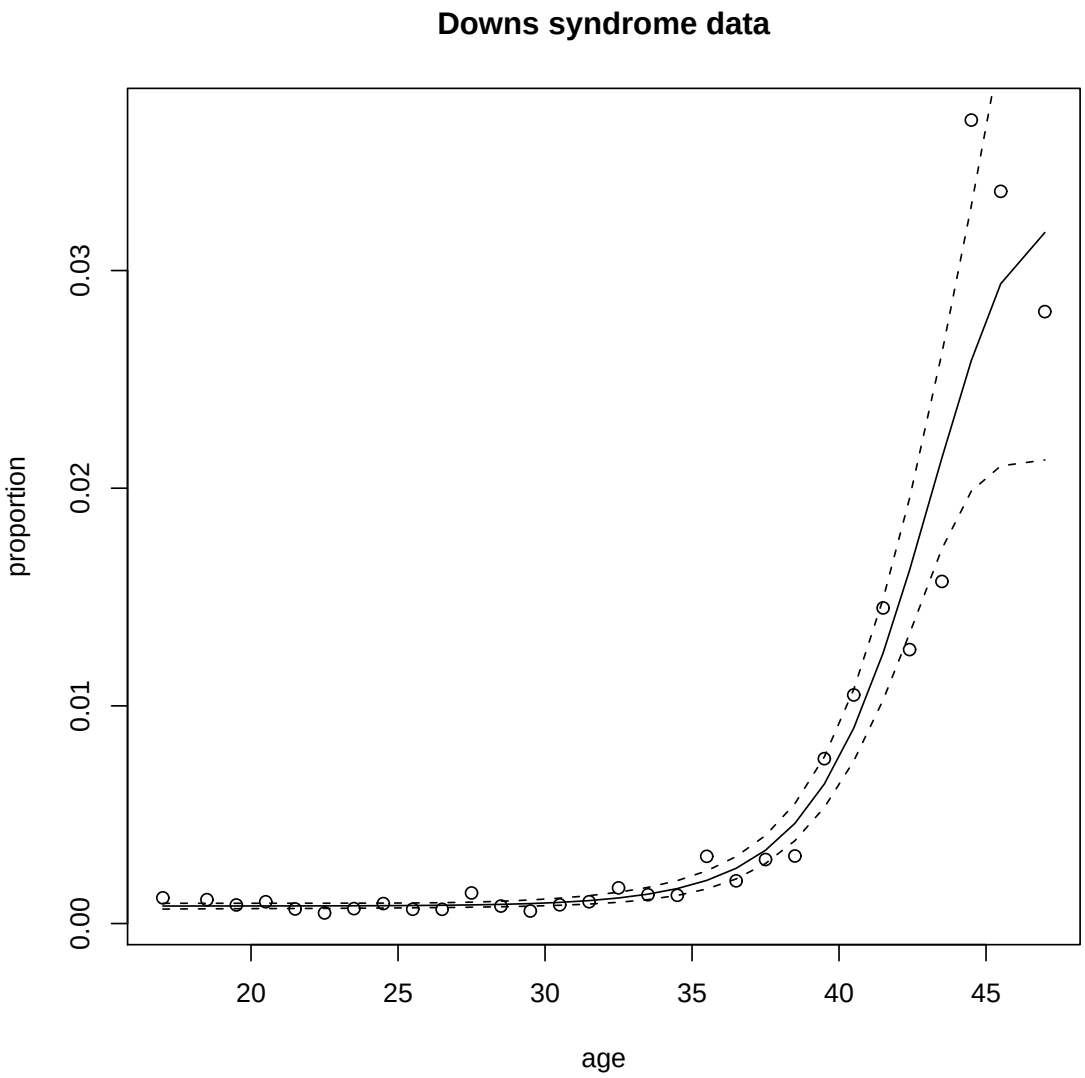

Figure 4.4: Plot of proportion of Downs syndrome births versus age of birth mother. The solid line is the posterior median and the dashed lines are 95% confidence bands.

Chapter 5

Bernstein Polynomials, Quadratic Programming, and Shape-Restricted Regression

5.1 Introduction

The "workhorse" in empirical research is the classical linear model with normal errors. The linear model provides simplicity in its interpretation and elegance in its theory (see, for example, Monahan, 2008; Rencher and Schaalje, 2008). When the data are not normal, the linear model can be extended to analyze other types of data (see McCullagh and Nelder, 1989).

The assumptions embodied in the linear model are often too restrictive for many applications, however. In such cases, nonparametric regression techniques offer an alternative that requires only minimal assumptions on the smoothness of the regression function (Efromovich, 1999).

But in many applied problems, there exists substantive subject-matter information on the unknown regression function beyond the usual smoothness constraints nonparametric regression techniques. In biomedical applications, dose-response functions are assumed to be monotonically increasing. In distance sampling (Buckland et al., 2001), the probability that an observer detects an animal (or other object)

is assumed to be monotonically decreasing with increasing distance. In reliability studies, the failure rate is assumed to have a "bathtub" shape (Reboul, 2005). In actuarial studies, the mean residual life function $m(x)$ is required to be nonnegative and to have a derivative $m'(x)$ which satisfies $m'(x) + 1 \geq 0$.

Several researchers have devised methods for fitting nonparametric regression functions with shape restrictions. Hildreth (1954) used quadratic programming to compute estimates of the ordinates of a concave regression function. Barlow et al. (1972) provide an algorithm for computing the maximum likelihood estimate of an isotonic regression function called the Pool Adjacent Violators (PAV) algorithm. Friedman and Tibshirani (1984) fit an isotonic regression curve by first fitting a running mean smoother to the data and then applying the PAV algorithm to the smoothed fits. Mukerjee (1988) reverses the order the Friedman and Tibshirani (1984) approach by first fitting the data with the PAV algorithm and then applying a kernel smoother to the fits. Ramsay (1988) and He and Shi (1998) alter spline basis expansions to ensure monotonicity in spline regression. Ramsay (1998) derives monotone regression functions from a solution to differential equations. Lavine and Mockus (1995) use the Dirichlet process to give a Bayesian solution to the monotone regression function. Other Bayesian approaches include Holmes and Heard (2003), who model the regression curve as piecewise constant and discard any draws from the posterior distribution that violate monotonicity, Neelon and Dunson (2004), who model the regression curve as piecewise linear and use a prior on coefficients that restricts the curve to be monotonic, Chang et al. (2007) who model the regression curve with a restricted Bernstein polynomial and compute posterior draws using a reversible jump Markov chain Monte Carlo (RJMCMC) algorithm for inference (see also Chak et al., 2005).

In this chapter, we describe a method to fit shape-restricted regression curves using Bernstein polynomials. Our model relies on the restrictions derived in Chang et al. (2007) and Chak et al. (2005). We use methods from quadratic programming to fit the restricted regression curve, and thereby avoid costly computations associated with the RJMCMC algorithm used in Chang et al. (2007). In the sections that follow, we present the Bernstein polynomial model for shape-restricted regression (Section 5.2), describe our method for fitting this model using quadratic programming (Section

5.3), use our method on real data sets (Section 5.4), and finish with some concluding remarks (Section 5.5).

5.2 Shape Restricted Inference with Bernstein Polynomials

Consider a sample $Y_1, \ldots, Y_n$ of size n with associated predictors $x_1, \ldots, x_n \in [0, 1]$ from a regression model

$$Y_i = f(x_i) + \epsilon_i$$

where $\epsilon_i \sim \mathsf{N}(0, \sigma^2)$ the regression function f is defined on the unit interval. Note that values of the predictor variable can be easily transformed to the unit interval, therefore, we lose no generality by assuming that $x_1, \ldots, x_n \in [0, 1]$. The Bernstein polynomial expansion of $f(x)$ can be written

$$g(x) = \sum_{k=0}^{M} f(k/M) \binom{M}{k} x^k (1-x)^{M-k} = \sum_{k=0}^{M} f(k/M) b_M(x, k) \tag{5.1}$$

If we let $f(k/M) = \beta_k$, then an estimate of the unknown regression function f can be obtained by estimating the coefficients in

$$g_M(x, \boldsymbol{\beta}) = \sum_{k=0}^{M} \beta_k b(x, k) \tag{5.2}$$

(see Stadtmüller, 1986; Tenbusch, 1997, for more on using Bernstein polynomials in nonparametric regression).

Because of the relationship $f(k/M) = \beta_k$, it is straightforward to derive restrictions on the Bernstein polynomial coefficients that will impose shape restrictions on the underlying regression function. Chang et al. (2007) and Chak et al. (2005) derive coefficient constraints for various shape restrictions. In this chapter, we will focus on four particular shape constraints—isotonicity (nondecreasing), antitonicity (nonincreasing), concavity, and convexity. These restrictions can be imposed on the Bernstein regression function with the following constraints on the parameters:

Isotonicity	$\beta_0 \leq \cdots \leq \beta_M$,
Antitonicity	$\beta_0 \geq \cdots \geq \beta_M$,
Convexity	$\beta_k - 2\beta_{k-1} + \beta_{k-2} \geq 0$ for $k = 2, \ldots, M$,
Concavity	$2\beta_{k-1} - \beta_k - \beta_{k-2} \geq 0$ for $k = 2, \ldots, M$.

5.3 Fitting Shape Restricted Regression Curves using Quadratic Programming

As mentioned previously, Chang et al. (2007) fit the Bernstein regression function by assigning a prior on regression coefficients β_k, which incorporates the appropriate shape restrictions. Chang et al. (2007) use a RJMCMC algorithm to sample from the posterior distribution of the coefficients and the order of the Bernstein polynomial M. Curtis and Ghosh (2008) note that the prior used in Chang et al. (2007) for monotonic regression does not allow for events $\beta_{k-1} = \beta_k$ with positive probability, which can be a detriment when modeling regression functions with flat regions. To correct for this, Curtis and Ghosh (2008) reparametrize the Bernstein regression function in terms of differences between adjacent coefficients and assign a variable-selection prior to the transformed parameters. Curtis and Ghosh (2008) use the variable-selection Gibbs sampler of Geweke (1996) to sample from the posterior distribution of the model parameters.

Both of the approaches mentioned above require extensive computations to compute the fit of the shape-restricted regression curve. The approach of Chang et al. (2007) requires a sampler which moves between parameter spaces of different dimensions, which can make it hard to assess convergence to the posterior distribution and can require a large number of simulations. The approach of Curtis and Ghosh (2008) also requires a large number of MCMC simulations to obtain adequate effective sample sizes of the posterior draws because of the high correlation between values $b_M(x, k)$ for different k.

We, therefore, propose a simple method for fitting shape-restricted, Bernstein-polynomial regression functions that uses quadratic programming. We first define

some notation. Let $\mathbf{y} = (y_1, \ldots, y_n)$ be the vector of observed values of the response variables $Y_1, \ldots, Y_n$. Let $\mathbf{b}_i^{(M)} = (b_M(x_i, 0), \ldots, b_M(x_i, M))$ be the Bernstein basis expansion for the i^{th} value of the predictor, and let the matrix $\mathbf{B}_M$ be the $n \times (M+1)$ matrix, where the i^{th} row of $\mathbf{B}_M$ is $\mathbf{b}_i^{(M)}$. Let $\mathbf{D}_{(1)}$ be an $M \times (M+1)$ matrix equal to

$$\begin{bmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{bmatrix},$$

and let $\mathbf{D}_{(2)}$ be an $(M-1) \times (M+1)$ matrix equal to

$$\begin{bmatrix} 1 & -2 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & & & \\ 0 & 0 & 0 & 0 & \cdots & 1 & -2 & 1 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1 & -2 & 1 \end{bmatrix}.$$

The matrix $\mathbf{D}_{(1)}$ when post-multiplied by a vector $\boldsymbol{\beta}$ of length $(M+1) \times 1$ gives a vector length M with first-order differences of the elements of $\boldsymbol{\beta}$. Similarly, $\mathbf{D}_{(2)}$ when post multiplied by $\boldsymbol{\beta}$ gives a vector of length $M-1$ containing second-order differences of the elements of $\boldsymbol{\beta}$.

Our approach minimizes

$$\|\mathbf{y} - \mathbf{B}_M\boldsymbol{\beta}\|^2 + \lambda\boldsymbol{\beta}'\mathbf{D}'_{(2)}\mathbf{D}_{(2)}\boldsymbol{\beta} \tag{5.3}$$

subject to constraints relevant to the shape restrictions described in Section 5.2. These constraints are listed in Table 5.1 in the notation just defined.

The first term in (5.3) contains the typical sum of squares used to measure goodness of fit. The second term in (5.3) is a penalty term that is a sum of squared, second-order differences of the regression coefficients $\boldsymbol{\beta}$. This penalty is motivated by the second order differences of B-spline coefficients used in Eilers and Marx (1996).

Table 5.1: Table of constraints for different shape restrictions.

Type	Constraint		
Isotonicity	$\mathbf{D}_{(1)}\boldsymbol{\beta}$	$\geq$	$\mathbf{0}$
Antitonicity	$-\mathbf{D}_{(1)}\boldsymbol{\beta}$	$\geq$	$\mathbf{0}$
Convexity	$\mathbf{D}_{(2)}\boldsymbol{\beta}$	$\geq$	$\mathbf{0}$
Concavity	$-\mathbf{D}_{(2)}\boldsymbol{\beta}$	$\geq$	$\mathbf{0}$

Eilers and Marx (1996) show that the second order differences of the B-spline coefficients approximate the integral of the second derivative of the regression function that is used in smoothing splines to penalize overfitting the data.

Quadratic programming problems are optimization problems where the solution is a minimizer of a quadratic objective function subject to linear constraints (see, for example, Nocedal and Wright, 2006, Chapter 16). More precisely, the solution to a quadratic programming problem is the minimizer of

$$-\mathbf{d}'\mathbf{b} + \frac{1}{2}\mathbf{b}'\mathbf{D}\mathbf{b} \tag{5.4}$$

with respect to $\mathbf{b}$ subject to constraints

$$\mathbf{A}'\mathbf{b} \geq \mathbf{b}_0, \tag{5.5}$$

where $\mathbf{b}$ and $\mathbf{d}$ are $p \times 1$ vectors, $\mathbf{D}$ is a $p \times p$ matrix, $\mathbf{A}$ is a $p \times r$ matrix specifying the r linear inequality constraints in $\mathbf{b}_0$.

It is easy to see that (5.3) can be written in the form of (5.4). After algebraic manipulation, (5.3) becomes

$$-\mathbf{y}'\mathbf{B}_M\boldsymbol{\beta} + \frac{1}{2}\boldsymbol{\beta}'(\mathbf{B}_M'\mathbf{B}_M + \lambda\mathbf{D}_{(2)}'\mathbf{D}_{(2)})\boldsymbol{\beta}$$

with constraints as defined in Table 5.1. Thus, in our notation for the quadratic

programming problem, we have

$$\begin{aligned}\mathbf{D} &= \mathbf{B}'_M\mathbf{B}_M + \lambda\mathbf{D}'_{(2)}\mathbf{D}_{(2)}\\ \mathbf{d} &= \mathbf{B}'_M\mathbf{y}\\ \mathbf{A}' &= \mathbf{D}_{(1)}\end{aligned}$$

for isotonicity, for example.

5.4 Illustrations Using Real Data

We demonstrate our method on two real data sets. The first data set (from Ramsay, 1998) contains 83 height measurements of a 10-year-old boy over a period of 312 days. In growth studies, it is a reasonable assumption that growth is nondecreasing as a function of time with periods of rapid growth interspersed with periods of relatively little growth (Thalange et al., 1996).

In order to fit our method to the child-growth data, we must specify the value of two tuning parameters—the order M of the Bernstein polynomial expansion and the parameter λ that balances goodness of fit with the penalty term. Hastie et al. (2001, Chapter 7) provide a summary of many of the different methods of selecting tuning parameters, such as C_p (Mallows, 1973), AIC (Akaike, 1974), cross validation (Stone, 1974), and generalized cross validation (Craven and Wahba, 1979).

We propose to use the Schwarz's criterion (Schwarz, 1978)—commonly known as BIC—to select the best values for the tuning parameters. Schwarz's criterion is derived as an approximation to the marginal likelihood of a Bayesian model (for further details, see Raftery, 1995). The value of BIC for a given λ and M can be calculated as

$$\text{BIC}_{M,\lambda} = -2\ell(\hat{\boldsymbol{\beta}}_{M,\lambda}, \hat{\sigma}^2_{M,\lambda}) + d\log n, \tag{5.6}$$

where d is the number of unique values in $\hat{\boldsymbol{\beta}}$ and is an estimate of the degrees of freedom of the model (Meyer and Woodroofe, 2000), $\hat{\boldsymbol{\beta}}_{M,\lambda}$ is an estimate of $\boldsymbol{\beta}$ based on the quadratic programming method for given values of M and λ, $\hat{\sigma}^2 = \|\mathbf{y} -$

$\mathbf{B}_M\hat{\boldsymbol{\beta}}_{M,\lambda}\|^2/(n-d)$ is an estimate of the variance σ^2, $\ell(\hat{\boldsymbol{\beta}}_{M,\lambda}, \hat{\sigma}^2_{M,\lambda})$ is the value of the log likelihood evaluated at $\hat{\boldsymbol{\beta}}_{M,\lambda}$, $\hat{\sigma}^2_{M,\lambda}$. Lower values of BIC indicate better model fit. To find the optimal values of the tuning parameters, we compute BIC for different values of λ and M and choose the values which minimize BIC.

Figure 5.1 contains plots of BIC values versus different values of M for six different values of λ. The lowest values of BIC are obtained for values of λ less than 0.001. Also, the value of M that gives the minimum BIC does not change for values of $\lambda \leq 0.001$.

Figure 5.2 contains a plot of the optimal fit for the child-growth data using the quadratic programming approach and using a local-regression smoother (loess) (see Cleveland et al., 1992, for example). The tuning parameter of the loess method was selected based on 5-fold cross validation (see, for example, Hastie et al., 2001, Section 7.10). The fits of the two methods are similar over a large portion of the range of the x values. However, for values of x between $\sim$100 and $\sim$125, the loess smoother estimates a decreasing mean for height, which violates assumptions appropriate for growth data.

The second example that we consider uses data from Dudzinski and Mykytowycz (1961), which has subsequently been analyzed by Ratkowsky (1983), Wei (1998), and Birke and Dette (2007). The data contain the dry weight of the eye lenses and the age of 71 rabbits.

As in the previous example, we fit our quadratic-programming concave regression function for several different values of λ and M and calculate the BIC for each fit. Figure 5.3 contains plots of the BIC versus M for different values of λ. The lowest values of BIC are attained when $\lambda \leq 0.01$. For values of λ less than or equal to 0.01, the plots of BIC versus M are virtually the same, and the optimal M is 6.

Ratkowsky (1983) proposes a parametric, concave nonlinear function

$$\mathsf{E}(Y|x) = \alpha \exp\left\{-\frac{\xi}{x+\gamma}\right\}$$

for these data, where x is the rabbit age and Y is the eye-lens weight at age x. Ratkowsky (1983) reports parameter estimates of $\hat{\alpha} = 5.63991$, $\hat{\xi} = 130.584$, and $\hat{\gamma} = 37.6029$.

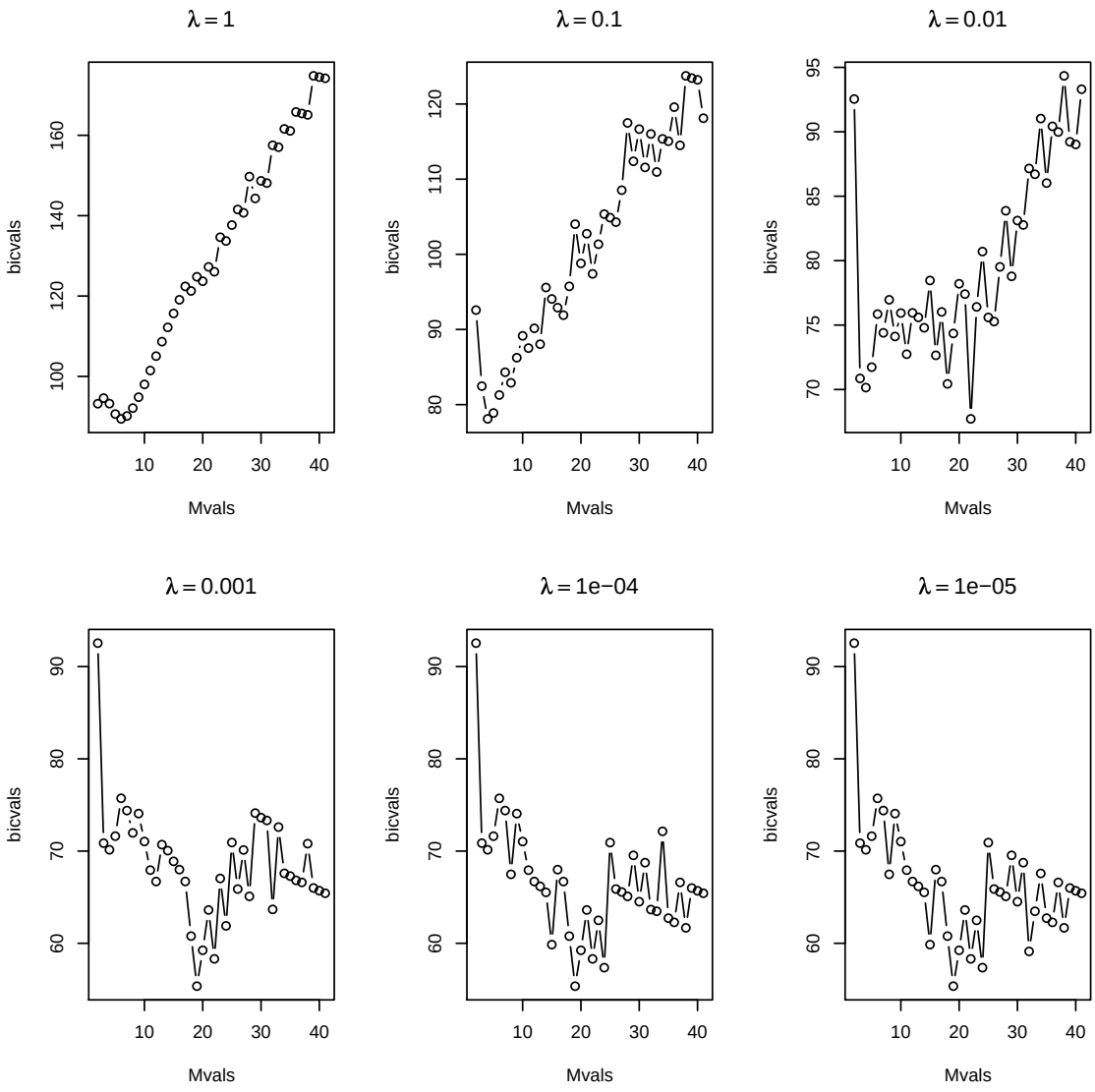

Figure 5.1: Plots of BIC for different values of M and λ.

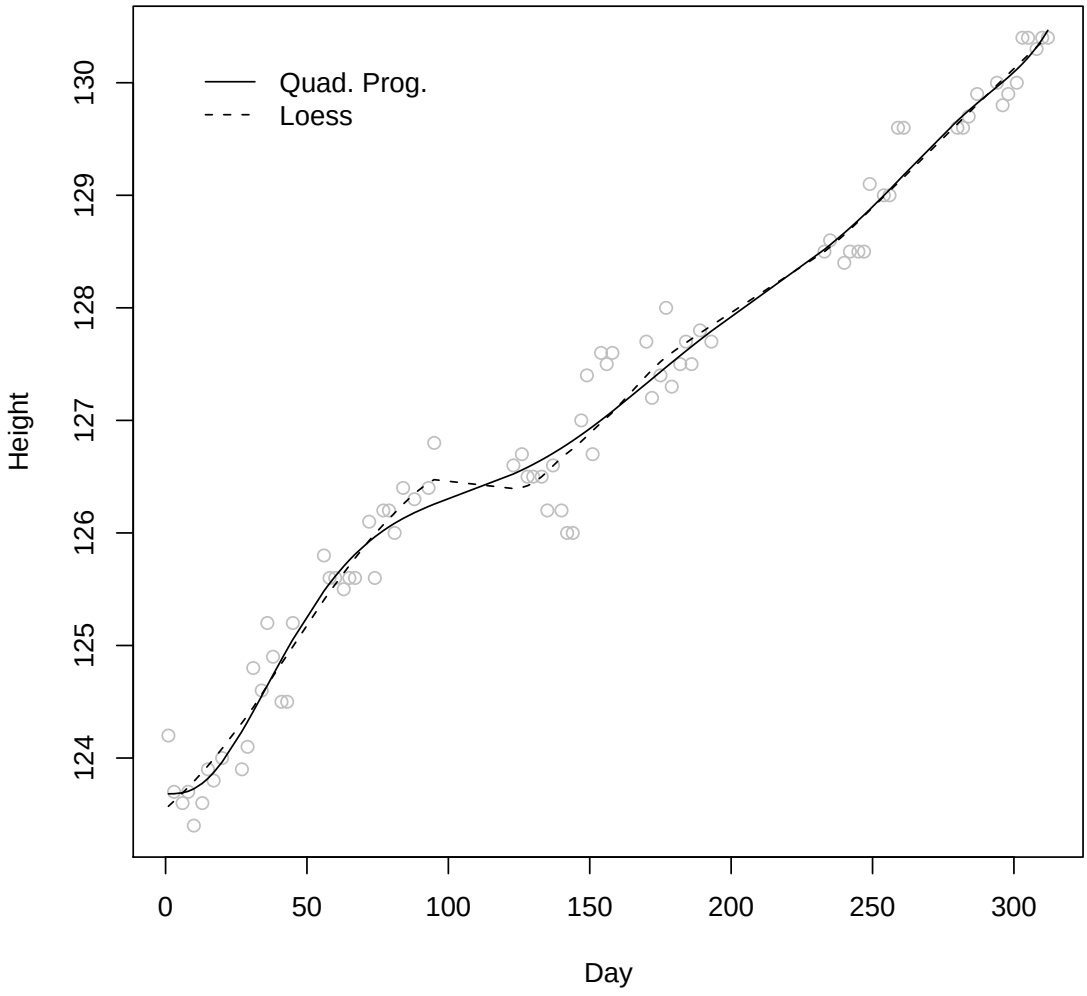

Figure 5.2: Plot of the quadratic programming fit for the child-growth data.

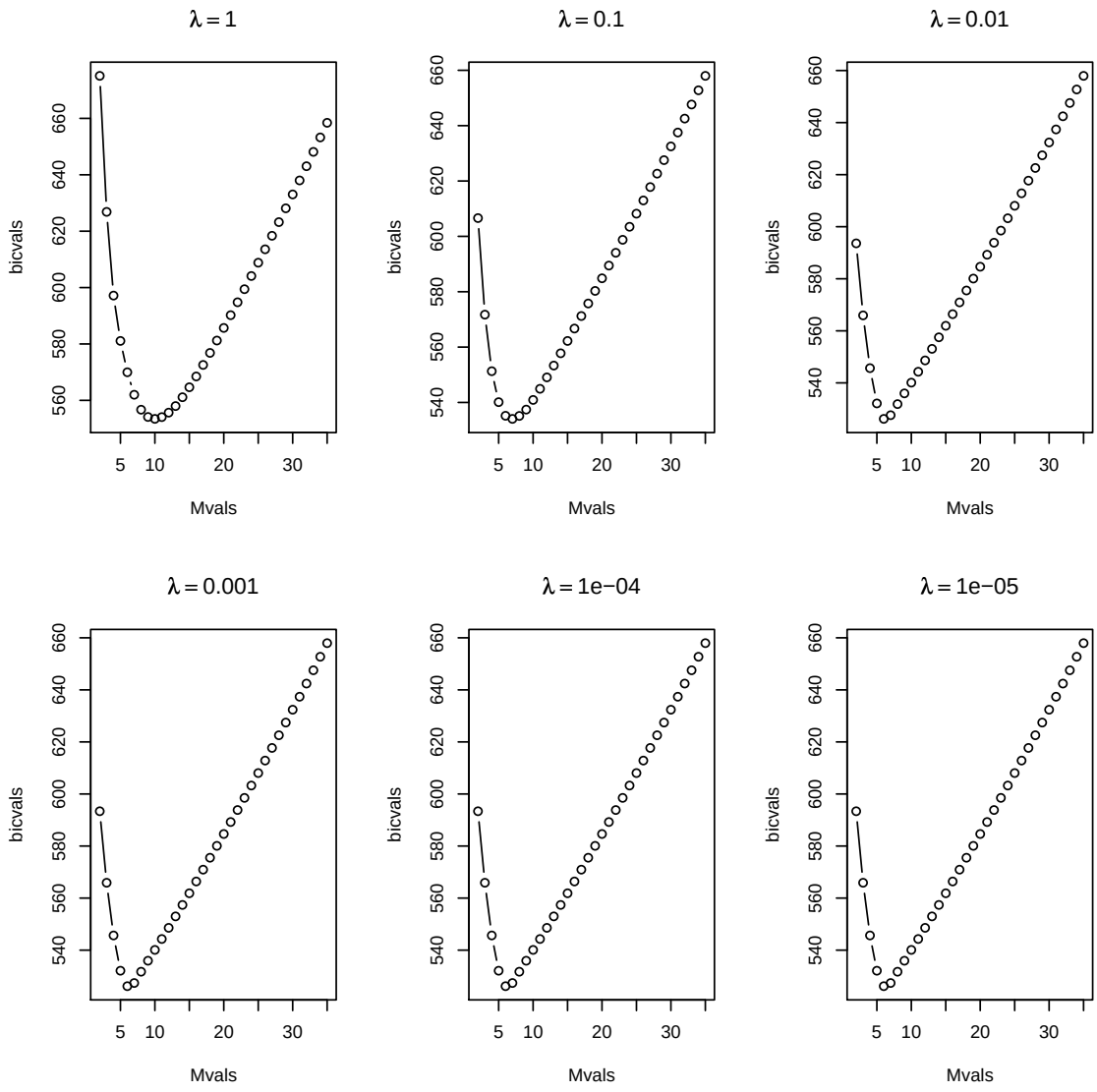

Figure 5.3: BIC versus M for different values of λ in the rabbit-data example.

Figure 5.4 contains a plot of the rabbit data with the fits from the quadratic programming method, the loess method, and the parametric model from Ratkowsky (1983). In this example, there is very little difference between the model fits of the different methods, although the loess method is less smooth than the other methods.

5.5 Conclusions

We have presented a simple way to fit shape-restricted regression curves. The approach presented in this chapter makes the derivation of shape restrictions and the fitting of the model to the data relatively easy. The use of Bernstein polynomials provides an intuitive method for deriving the necessary shape restrictions on the regression curve, and the use of quadratic programming provides a straightforward method for fitting these curves.

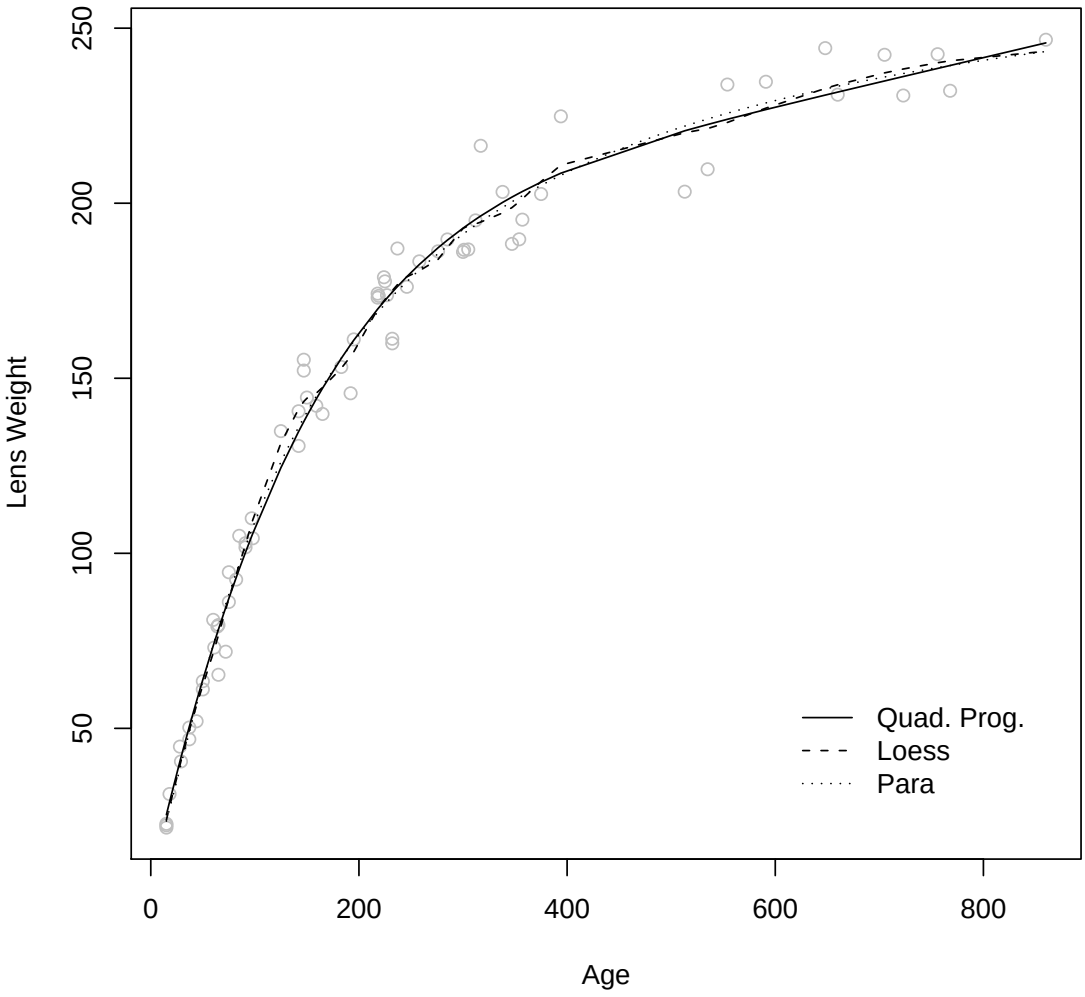

Figure 5.4: Plot of lens weight versus age with model fits for the quadratic programming, local-regression, and parametric methods.

Chapter 6

Further Research

In this chapter, we review further research that has been conducted since the writing of the papers contained in the previous chapters.

6.1 "Flat" Portions of Bernstein Regression Curves

We claim in Chapter 4 that by allowing for positive probability of events $\beta_k = \beta_{k-1}$ the Bernstein polynomial regression function can model flat portions of a regression curve. However, for the monotonic regression function (and for Bernstein regression functions in general), the Bernstein polynomial regression curve can never be exactly flat along any interval in $[0, 1]$ (Ghosal, 2008), unless all $\beta_k = 0$. This can easily be seen from the form of the derivative:

$$g'(x) = M \sum_{k=0}^{M-1} (\beta_{k+1} - \beta_k) \binom{M-1}{k} x^k (1-x)^{M-1-k}$$

Because of the restrictions on the coefficients β_k that enforce monotonicity, the terms $(\beta_{k+1} - \beta_k)$ in the sum are all positive or zero. Hence, the derivative can never be zero unless all β_k are equal.

In this section, we attempt to show more rigorously how the Bernstein polynomial regression curve can model flat portions of the regression function. We note that,

as per the discussion in the previous paragraph, if a particular application requires a model that allows for regions of regression curve that are exactly flat, then the Bernstein polynomial approach is not appropriate. We will show, however, how closely the Bernstein polynomial model can approximate a flat regression curve on a given interval in $[0, 1]$.

We first define a measure of flatness for the Bernstein polynomial regression curve. For an arbitrary interval $[a, b]$ in $[0, 1]$ we let the flatness of the regression curve (call it Ψ) be the integral of the first derivative of the regression curve over the interval $[a, b]$. More precisely,

$$\Psi = \frac{\int_a^b g'_M(x, \boldsymbol{\beta})\mathrm{d}x}{b - a}. \tag{6.1}$$

This is simply the average value of the slope of the Bernstein regression curve over the interval $[a, b]$. The flatness Ψ can be simplified to

$$\Psi = \frac{g_M(b, \boldsymbol{\beta}) - g_M(a, \boldsymbol{\beta})}{b - a},$$

the slope of the line through the points a and b.

We note a few properties of the flatness measure Ψ. First, under the restrictions on $\boldsymbol{\beta}$ for monotonicity, Ψ is always greater than or equal to zero. The flatness Ψ is zero if and only if all β_k are equal. Thus, Ψ may not be an appropriate measure of flatness for regression curves other than the monotonic regression curve because, under other restrictions, Ψ can be zero even though the curve is not flat. In such situations, a different measure of flatness could be used such as the integral of the square of the first derivative—$\int_a^b [g'_M(x, \boldsymbol{\beta})]^2 \mathrm{d}x$—or the integral of the absolute value of the first derivative—$\int_a^b |g'_M(x, \boldsymbol{\beta})| \mathrm{d}x$.

Second, the prior on $\boldsymbol{\beta}$ induces a prior distribution on Ψ. Thus we can find the expected value and variance of Ψ conditional on the values of the other model parameters (τ and p). The conditional expectation of Ψ simplifies to a nice expression (see Appendix B.2 for the derivation)

$$\mathsf{E}(\Psi | \tau^2, p) = \sqrt{2\pi\tau^2}(1 - p)M. \tag{6.2}$$

The variance of Ψ can be written (see Appendix B.3 for derivation)

$$\mathsf{Var}(\Psi|\tau^2, p) = \frac{q\tau^2\left[1 - \frac{2}{\pi}q\right]}{(b-a)^2}\sum_{k=1}^{M}[w_M(b,k) - w_M(a,k)]^2, \tag{6.3}$$

where $1 - p = q$.

It is easy to see from (6.2) that if q and τ do not depend on M, the expectation of Ψ increases with M. The behavior of $\mathsf{Var}(\Psi|\tau^2, p)$ for increasing M is not apparent from (6.3). Thus, we plotted $\mathsf{Var}(\Psi|\tau^2, p)$ versus M when $a = 0.25$, $b = 0.75$, and $(1-p)\tau^2[1 - \frac{2}{\pi}q] = 1$. The plot is contained in Figure 6.1 and shows an increasing trend in $\mathsf{Var}(\Psi|\tau^2, p)$ for increasing M.

The increasing trend in both $\mathsf{E}(\Psi|\tau^2, p)$ and $\mathsf{Var}(\Psi|\tau^2, p)$ for increasing M suggests letting the prior variance of the U_k and the probability of $U_k = 0$ depend on M as a way to control the prior flatness of the Bernstein polynomial regression curve. One possibility is to let $1 - p = 1/M$ and $\tau^2 = \eta^2/M$. Then, we have

$$\begin{aligned}\mathsf{E}(\Psi|\eta^2) &= \sqrt{2\pi}\frac{\eta}{\sqrt{M}} \\ \mathsf{Var}(\Psi|\eta^2) &= \frac{\eta^2\left[M - \frac{2}{\pi}\right]}{M^3(b-a)^2}\sum_{k=1}^{M}[w_M(b,k) - w_M(a,k)]^2.\end{aligned}$$

We plot the value of $\mathsf{Var}(\Psi)$ versus M for $[a, b] = [0.25, 0.75]$ in Figure 6.2. As the plot indicates, $\mathsf{Var}(\Psi|\eta^2)$ decreases with M.

Thus, under certain prior settings we can restrict the flatness of the Bernstein regression function to an arbitrarily small degree.

6.2 Soil Data Example for Bayesian Variable Selection and Clustering

In this section, we add another real-data example to those of Chapter 2. We use the soil data from Bondell and Reich (2008). The data include 20 observations of 500-meter-squared plots of land in the North Carolina Appalachian Mountains. The response is the number of different plant species observed on each individual plot.

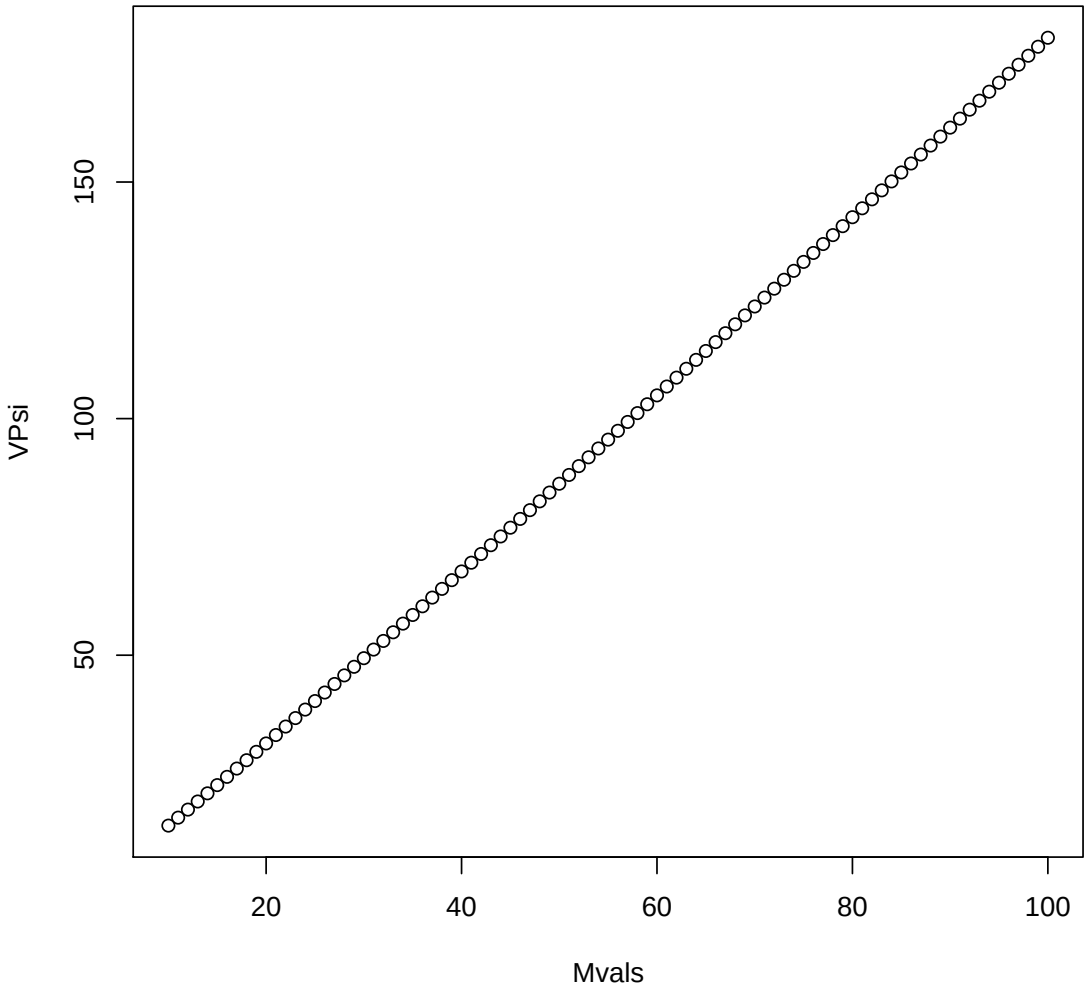

Figure 6.1: Plot of the variance of Ψ versus values of M for the interval $[0.25, 0.75]$ when $q\tau^2[1 - \frac{2}{\pi}q] = 1$.

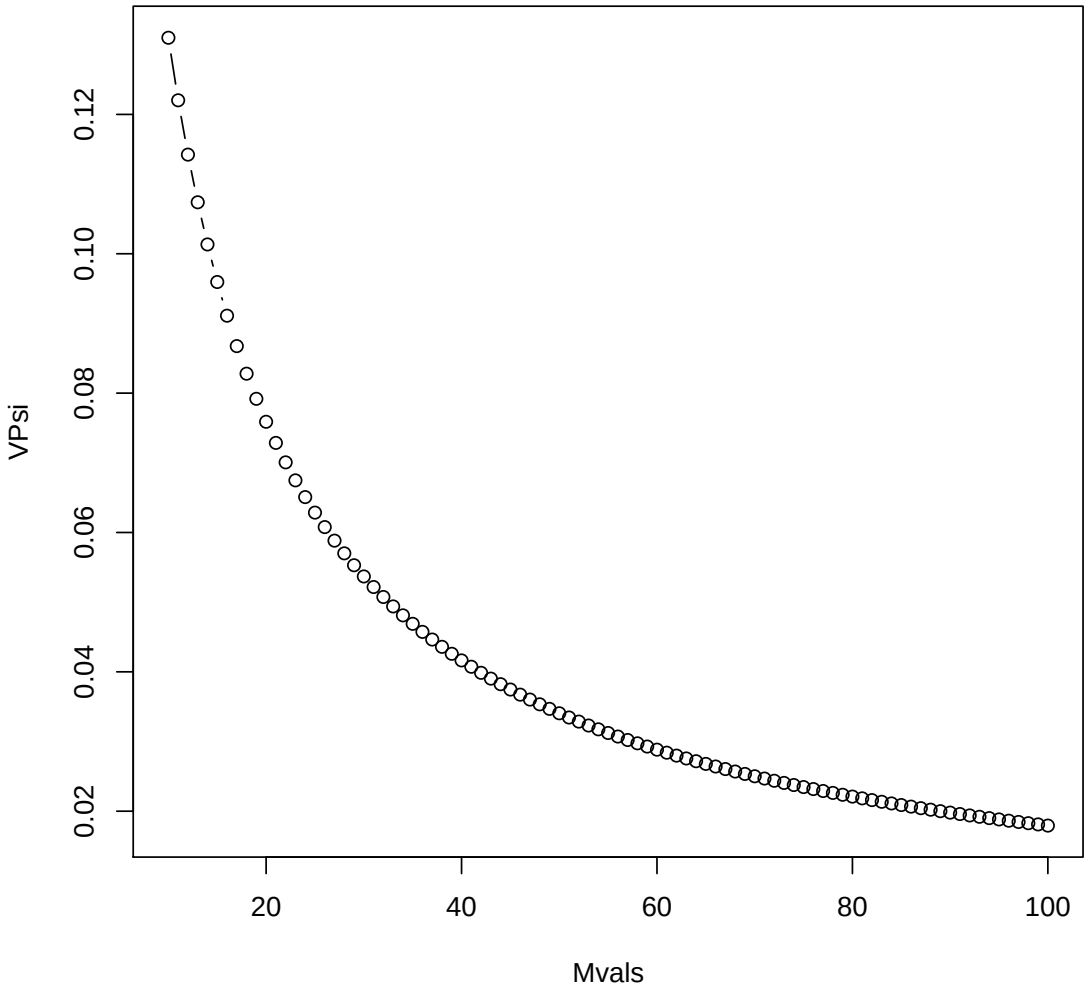

Figure 6.2: Plot of the variance of Ψ versus values of M for the interval $[0.25, 0.75]$ when $q = 1/M$ and $\tau^2 = \eta^2/M$ and $\eta^2 = 1$.

Fifteen different soil characteristics were collected for each plot and included in the data as predictors.

We include the soil data example for three reasons. First, we would like to compare our method more carefully with the OSCAR. Second, we would like to examine the effect of different prior distributions on the variance parameters in our model. Finally, we would also like to consider the effect of postprocessing of posterior draws (as recommended in Medvedovic and Sivaganesan, 2002) to assess clustering structure of the predictors.

6.2.1 Noninformative Priors for Variance Parameters

In Chapter 2, we used inverse-gamma priors with small values of the hyperparameters (0.1) for the variance parameters τ^2 and σ^2. Other authors have recommended different priors for variance parameters in hierarchical models, and we explore these in this section.

Scott and Berger (2006)

Scott and Berger (2006) consider the multiple testing problem of determining active genes from microarray data. The i^{th} observation Y_i is assumed to come from a normal distribution with mean $\gamma_i \mu_i$ and variance σ^2. The γ_i variables are binary variables indicating whether the i^{th} observation comes from an active ($\gamma_i = 1$) or an inactive ($\gamma_i = 0$) gene. The μ_i are assumed to come from a normal distribution with mean zero and variance τ^2.

Scott and Berger (2006) use the following prior for τ^2 and σ^2

$$\begin{aligned} p(\tau^2, \sigma^2) &= p(\tau^2|\sigma^2)p(\sigma^2) \\ &= \left[\frac{1}{\sigma^2}\left(1 + \frac{\tau^2}{\sigma^2}\right)^{-2}\right]\frac{1}{\sigma^2} \\ &= (\tau^2 + \sigma^2)^{-2}. \end{aligned}$$

Scott and Berger (2006) note that the prior $p(\tau^2|\sigma^2)$ is proper and *must* be a proper prior since the term τ^2 does not appear in all models (the term does not

appear in the model with all $\mu_j = 0$). They also note (Scott and Berger, 2006, page 2146) that the prior is

> ... scaled by σ^2 (which is commonly done for a hypervariance), and its form is an obvious modification of the usual objective prior $(\tau^2 + \sigma^2)^{-1}$ (which cannot be used here because it is improper).

Scott and Berger (2006, page 2146) also point out that the

> ... mild decrease in $p(\tau^2|\sigma^2)$ ensures that the prior is not overly influential in the analysis.

Gelman (2006)

Gelman (2006) compares several different noninformative priors for the hierarchical variance τ^2 or standard deviation τ in the model

$$\begin{aligned} Y_{ij} &\sim \mathsf{N}\left(\mu + \alpha_j, \sigma^2\right), \qquad i = 1, \ldots, n_j, \qquad j = 1, \ldots, J \\ \alpha_j &\sim \mathsf{N}\left(\mathbf{0}, \tau^2\right) \\ \sigma^2 &= p(\sigma^2) = 1/\sigma^2 \end{aligned}$$

Table 6.1 contains a list of all the priors discussed in Gelman (2006) and some of the recommendations and comments made by Gelman (2006) on each of these priors. In Chapter 2, we use the prior $\mathsf{InvGam}(\epsilon, \epsilon)$.

The prior recommended by Gelman (2006) is the half Cauchy prior distribution. If X is a random variable with a half Cauchy distribution, then the density of X is $p_X(x) = (s^2 + x^2)^{-1}$. The half Cauchy is a special case of the folded, noncentral t distribution. The folded, noncentral-t distribution is defined as the distribution of a random variable $T = |Z|S$, where $Z \sim \mathsf{N}(\xi, 1)$ and $S^2 \sim \mathsf{InvGam}\left(\frac{\nu}{2}, \frac{\nu}{2}s^2\right)$ (a scaled inverse χ^2 distribution). The half Cauchy distribution is obtained when $\nu = 1$ and $\xi = 0$.

Analysis of the Soil Data with Different Priors on τ^2 and σ^2

We investigate the effects of four different prior distributions on τ^2 and σ^2 on the analysis of the soil data described previously. The four prior distributions are summarized in Table 6.2. The first prior (IG) places a noninformative prior $p(\sigma^2) \propto \sigma^{-2}$

Prior	Comments
$\log \tau \sim c$	Results in an improper posterior.
$\log \tau \sim \mathsf{Unif}(-K, K)$	Depends strongly on $-K$.
$\tau \sim \mathsf{Unif}(0, \infty)$	Results in an improper posterior when $J \leq 2$.
	Miscalibrated toward larger values.
$\tau \sim \mathsf{HalfCauchy}(s^2)$	Recommended by Gelman (2006)
$\tau^2 \sim \mathsf{Unif}(0, \infty)$	Results in an improper posterior when $J \leq 3$.
	Miscalibrated toward larger values.
$\tau^2 \sim \mathsf{InvGam}(\epsilon, \epsilon)$	Results in an improper posterior as $\epsilon \to 0$.
	For data which support low values of τ, inferences are sensitive to ϵ.

Table 6.1: List of priors for the hypervariance parameter τ^2 considered by Gelman (2006). Comments are based on remarks in Gelman (2006)

on σ^2 and an $\mathsf{InvGam}(\epsilon, \epsilon)$ prior on τ^2 (where $\epsilon = 0.1$). The second prior (IGIG) puts $\mathsf{InvGam}(\epsilon, \epsilon)$ on both σ^2 and τ^2 (where, again, $\epsilon = 0.1$). The third prior (SB) is the prior from Scott and Berger (2006). The final prior (HC) combines the noninformative prior $p(\sigma^2) \propto \sigma^{-2}$ on σ^2 and the half Cauchy prior on τ^2.

As in Bondell and Reich (2008), we standardize all variables in the data set before running any analyses (i.e. we standardize each predictor such that for the j^{th} $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})'$, $\sum_{i=1}^n x_{ij} = 0$, and $\sum_{i=1}^n x_{ij}^2 = n - 1$, and we standardize the response similarly).

Table 6.3 contains the results of a classical regression analysis of the data as given in R. None of the predictor variables is significant (the lowest p-value is 0.249). The overall regression F-test is significant at the 0.10 level (with a p-value of 0.097).

Apparently, the correlation between the predictors is high enough that regression algorithm detects a singularity in the X-matrix. The correlation between the predictors is depicted in Figure 6.3. As indicated in the plot, several of the predictors are highly correlated. Several pairs of predictors have absolute correlations higher than 0.96.

Table 6.2: Four prior distributions used for the parameters σ^2 and τ^2 on the soil data in the Bayesian selection and clustering model.

Prior	Abbr.	Form
Inverse Gamma / Noninformative	IG	$p(\sigma^2, \tau^2) \propto \sigma^{-2}(\tau^2)^{-(\epsilon+1)} \exp\left\{-\frac{\epsilon}{\tau^2}\right\}$
Inverse Gamma / Inverse Gamma	IGIG	$p(\sigma^2, \tau^2) \propto (\sigma^2\tau^2)^{-(\epsilon+1)} \exp\left\{-\left(\frac{\epsilon}{\sigma^2} + \frac{\epsilon}{\tau^2}\right)\right\}$
Scott and Berger (2006)	SB	$p(\sigma^2, \tau^2) \propto (\tau^2 + \sigma^2)^{-2}$
Half Cauchy	HC	$p(\sigma^2, \tau) \propto \sigma^{-2}(s^2 + \tau^2)^{-1}$

Table 6.3: Results from a classical analysis of the soil data.

	Estimate	Std. Error	t value	Pr(>|t|)
BaseSat	0.31	2.04	0.15	0.88
SumCation	−9.50	29.73	−0.32	0.76
CECbuffer	−4.12	43.15	−0.10	0.93
Ca	10.14	51.24	0.20	0.85
Mg	2.80	9.31	0.30	0.77
K	−0.19	1.82	−0.10	0.92
Na	NA	NA	NA	NA
P	0.23	0.38	0.60	0.57
Cu	0.51	0.50	1.02	0.35
Zn	−0.45	0.46	−0.99	0.36
Mn	0.35	0.27	1.28	0.25
HumicMatter	−0.89	0.86	−1.03	0.34
Density	−0.09	0.75	−0.12	0.91
pH	0.49	0.76	0.65	0.54
ExchAc	1.97	14.88	0.13	0.90

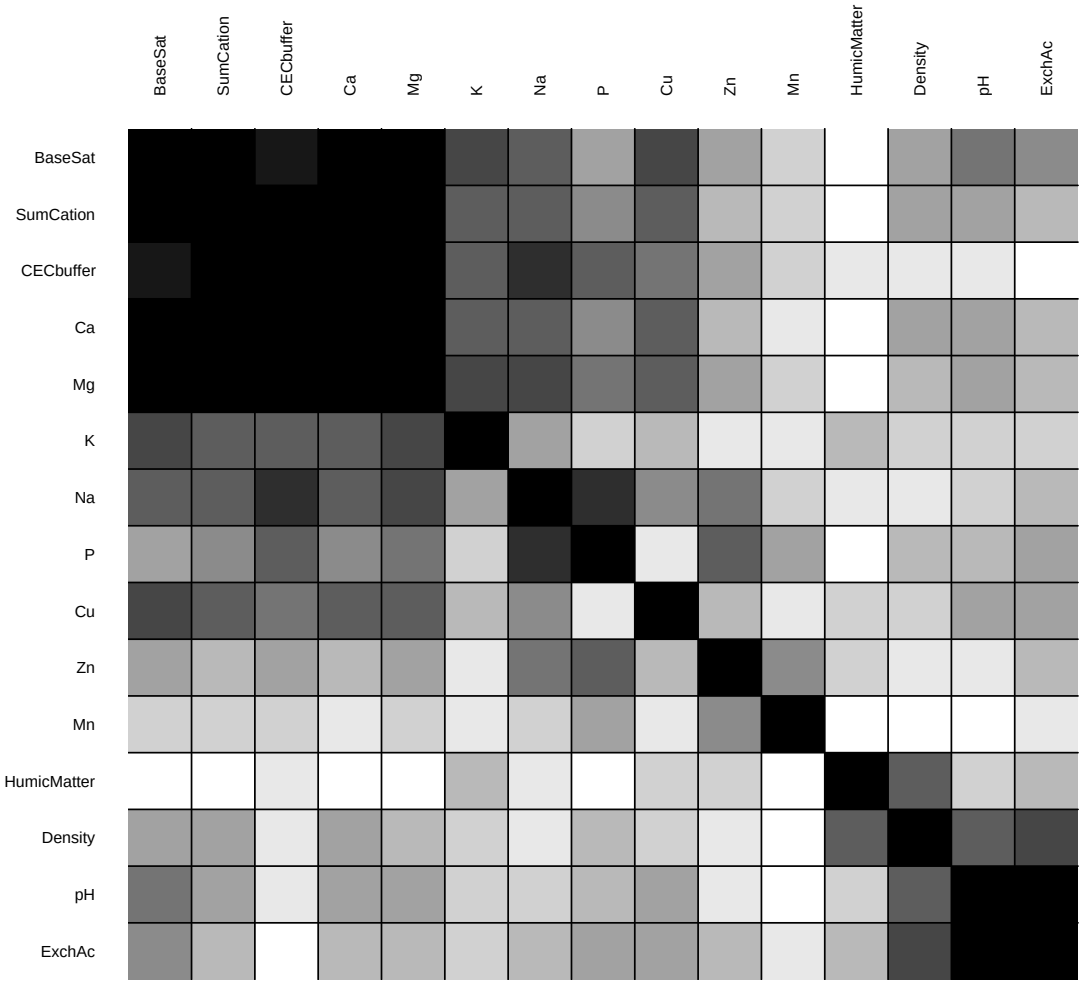

Figure 6.3: Image plot of the absolute value of the correlation matrix for the predictors in the soil data. Levels of gray correspond to the levels of absolute correlation, where black corresponds to an absolute correlation of 1 and white corresponds to an absolute correlation of 0.

Table 6.4: Effective sample sizes for each prior in the soil-data example.

Parameter	IG	SB	HC	IGIG
β_1	6300	3200	4800	910
β_2	840	290	390	1000
β_3	1400	420	510	2900
β_4	730	1500	2600	15000
β_5	15000	10000	550	7800
β_6	870	1000	2400	7300
β_7	2600	5300	7200	7200
β_8	1200	1100	330	9500
β_9	1700	660	210	690
β_{10}	7900	1500	2000	15000
β_{11}	1200	350	120	960
β_{12}	210	180	97	670
β_{13}	2700	740	1000	12000
β_{14}	1600	1000	680	920
β_{15}	15000	700	950	15000
σ^2	250	880	76	690
π	88	15000	30	410
τ^2	24	11	15	210

We fit the Bayesian model of Chapter 2 using the different prior distributions for the variance parameters as summarized in Table 6.2 using a modification of the `openbugs` function in the `R2WinBUGS` package in R. For each model, we ran 3 chains of 30,000 iterations. We burned the first 5,000 of these iterations and thinned the remiaining observations by 5. Thus, our final posterior sample size was 5,000 per chain for a total of 15,000 iterations. Trace plots of the model fits are contained in Figure 6.4 and the Gelman-Rubin (GR) statistics (Gelman and Rubin, 1992, listed as $\hat{R}$ in the plots) are contained in Figure 6.5.

The SB prior had the worst convergence of the four models. The Gelman-Rubin

Figure 6.4: Trace plots of MCMC iterations for the four models with priors outlined in Table 6.2.

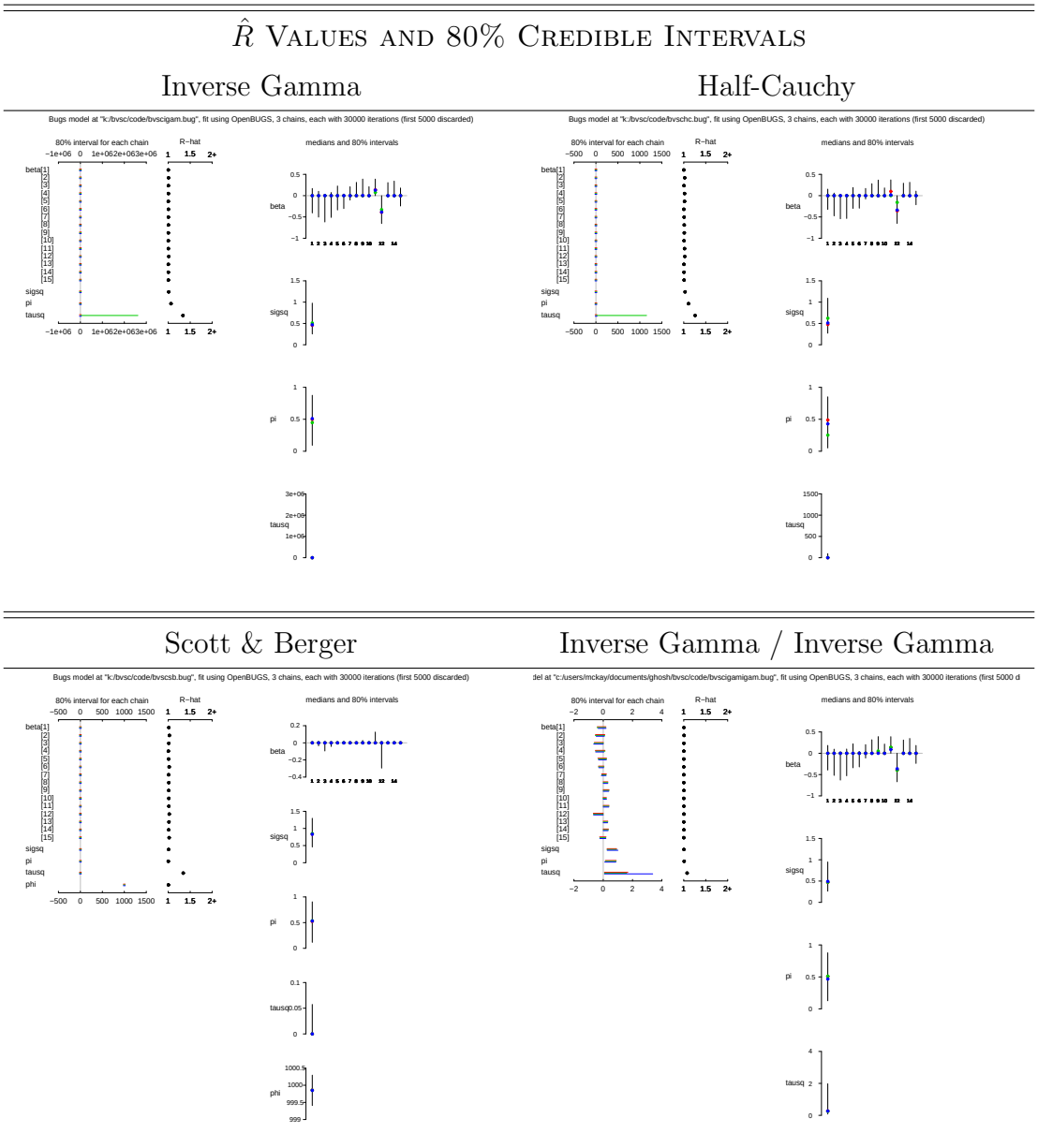

Figure 6.5: Convergence diagnostics for MCMC algorithm and 80% credible intervals for model parameters in the soil data analysis.

(GR) diagnostic for the hypervariance τ^2 was not near one and the trace plots (particularly for the regression coefficients) indicate that something is wrong with the mixing of the chains. The IG prior and the HC prior give trace plots that look much better than the traceplots for the SB priors, but the GR diagnostics for τ^2 are about as bad as for the SB prior. The IGIG prior got better mixing than the others and the GR diagnostic for the problematic τ^2 is much closer to one than for the other priors. The trace plots show, however, that there is still one chain that takes a long excursion out into the extreme right tail of the τ^2 distribution.

Except for the IGIG prior, the effective sample sizes for π and for τ^2 are disappointingly small (< 30) for all analyses. The effective sample sizes of τ^2 under the IGIG prior are ten to twenty times larger than for the other priors, although the effective sample size is still only 210.

The diagnostics that we have analyzed indicate that the only prior that produces MCMC output that can be reliably used for inference is the IGIG prior. Better results might be obtained by altering the MCMC algorithm. The algorithm used here is the default algorithm as used in the `openbugs` function. A better, customized algorithm might be used to improve MCMC outcomes.

6.2.2 Clustering Configurations in the Soil Data Set

Because of the relatively poor performance of the MCMC algorithm for three of the four priors, we proceed with our analysis of the soil data using all four models, but focus mainly on the IGIG model. A plot of the posterior probability of nonzero coefficients for each of the models is contained in Figure 6.6. All models give similar patterns of posterior probabilities of nonzero coefficients. That is, different priors on the variance terms did not give substantial differences in the posterior probability of nonzero coefficients relative to the other coefficients. Table 6.5 contains the rankings of each coefficient from each model in descending order.

If we use a cutoff probability (say 0.5, as indicated by the horizontal line in the plots of Figure 6.6) for inclusion of a predictor into the model, then we can compare the results of the Bayesian clustering algoirthm with the OSCAR and LASSO

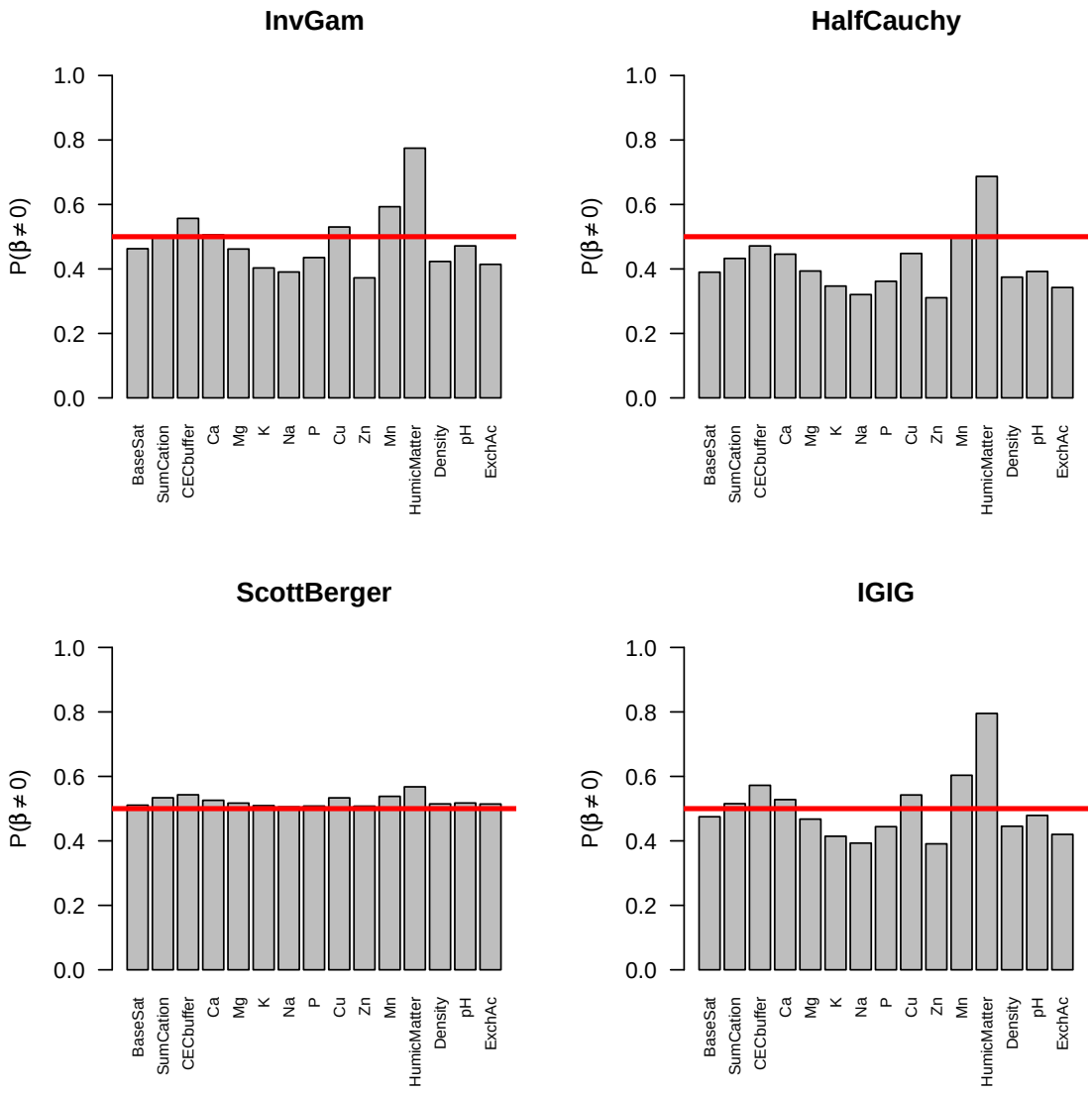

Figure 6.6: Posterior probability of nonzero regression coefficients in the soil-data example for models with different priors on the variance terms.

Table 6.5: Rankings of regression coefficients under models with the four different priors in Table 6.2.

	InvGam	ScottBerger	HalfCauchy	IGIG
1	HumicMatter	HumicMatter	HumicMatter	HumicMatter
2	Mn	CECbuffer	Mn	Mn
3	CECbuffer	Mn	CECbuffer	CECbuffer
4	Cu	SumCation	Cu	Cu
5	Ca	Cu	Ca	Ca
6	SumCation	Ca	SumCation	SumCation
7	pH	pH	Mg	pH
8	BaseSat	Mg	pH	BaseSat
9	Mg	Density	BaseSat	Mg
10	P	ExchAc	Density	Density
11	Density	BaseSat	P	P
12	ExchAc	K	K	ExchAc
13	K	P	ExchAc	K
14	Na	Zn	Na	Na
15	Zn	Na	Zn	Zn

results in Bondell and Reich (2008). Table 6.6 is a reproduction of the table in Bondell and Reich (2008) and contains estimates of the regression coefficients using the OSCAR and LASSO under 5-fold cross validation (5CV) and generalized cross validation (GCV). Clusters of regression coefficients can be identified by coefficients with equal (in absolute value) estimates.

The OSCAR with 5CV [OSCAR(5CV)] creates 7 clusters—one cluster of four predictors (sum cations, CEC, calcium, potassium), five clusters of one predictor each (phosphorus, copper, manganese, humic matter, pH), and one cluster of predictors with zero coefficients. Simlilarly, the OSCAR with GCV [OSCAR(GCV)] also creates 7 clusters—one cluster of five predictors (sum cations, CEC, calcium, potassium, pH), five clusters of one predictor each (% base saturation, phosphorus, copper, manganese, humic matter), and one cluster of predictors with zero coefficients. The LASSO under 5CV [LASSO(5CV)] selects two of the four predictors from the largest cluster of the OSCAR(5CV)—CEC and Potassium—and includes all other variables in the OSCAR(5CV). The LASSO with GCV [LASSO(GCV)] selects three of the five variables in the largest cluster of the OSCAR(GCV) estimates—calcium, potassium, pH—and all other predictors selected by OSCAR(GCV) plus one more—zinc.

If we use a cutoff probability of 0.5, the the IGIG model selects three of the four predictors in the largest cluster of the OSCAR(5CV)—sum cation, CEC, calcium—and selects three of the five remaining predictors of the OSCAR(5CV) model—copper, manganese, humic matter. If we lower the cutoff proability to 0.45, then the Bayesian model includes all of the predictors in the model selected by OSCAR(GCV), except phosphorus. Thus, the Bayesian model appears to be selecting models that are more similar to OSCAR models than to LASSO models.

However, choosing a model by selecting predictors with posterior probabilities that are higher than a specified cutoff does not give information on different clusterings of regression coefficients. One approach to examining clustering of the regression coefficients is to examine which clusterings were observed most often in the posterior draws. Unfortunately, this approach is not fruitful because the number of possible clusterings is large relative to the number of MCMC iterations, and the posterior distribution can assign a very small probability to a large number of cluster configurations.

Table 6.6: Estimates of regression coefficients using the OSCAR and the LASSO under 5-fold cross validation and generalized cross validiation. This table is reproduction of the table in Bondell and Reich (2008)

Variable	OSCAR (5-fold CV)	OSCAR (GCV)	LASSO (5-fold CV)	LASSO (GCV)
% Base saturation	0.000	−0.073	0.000	0.000
Sum cations	−0.178	−0.174	0.000	0.000
CEC	−0.178	−0.174	−0.486	0.000
Calcium	−0.178	−0.174	0.000	−0.670
Magnesium	0.000	0.000	0.000	0.000
Potassium	−0.178	−0.174	−0.189	−0.250
Sodium	0.000	0.000	0.000	0.000
Phosphorus	0.091	0.119	0.067	0.223
Copper	0.237	0.274	0.240	0.400
Zinc	0.000	0.000	0.000	−0.129
Manganese	0.267	0.274	0.293	0.321
Humic matter	−0.541	−0.558	−0.563	−0.660
Density	0.000	0.000	0.000	0.000
pH	0.145	0.174	0.013	0.225
Exchangeable acidity	0.000	0.000	0.000	0.000

For example, the most common cluster configuration in the IGIG model is the cluster with all regression coefficients set equal to zero. This cluster occurs in %5.2 of the posterior draws. The next most common cluster configuration is the model with only humic matter, which occurs in %1.6 of the posterior draws. The next four most common models each have two predictors—humic matter and some other predictor—and occur in 0.8%, 0.6%, 0.5%, and 0.5% percent of the total draws.

In a slightly different scenario, Medvedovic and Sivaganesan (2002) recommend "post-processing" of the posterior draws to search for viable clusters. In our regression setting, the approach of Medvedovic and Sivaganesan (2002) requires computing the pairwise posterior probability ϖ_{jk} that $\beta_j = \beta_k$ for all j and k from the posterior draws. The pairwise posterior proababilities can then be used in a hierarchical clustering procedure where $1 - \varpi_{jk}$ can be used as "distances" between regression coefficients (see Rencher, 2002, Chapter 14, for example).

Figure 6.7 contains a dendrogram based on a cluster analysis of the regression coefficients using posterior pairwise probabilities as the distance measure. The clustering configurations have some simlilarities and some differences to the cluster configurations induced by the OSCAR. The OSCAR(5CV) induces 7 clusters of the 15 regression coefficients (where one of the clusters contains all zero coefficients). If we "cut" the "tree" in Figure 6.7 so as to allow 7 clusters, we get the cluster configuration as delineated by the rectangles in Figure 6.7. The largest cluster contains magnesium, posphorus, density, pH, sodium, and zinc. With the exception of pH, all of these predictors are estimated to have zero regression coefficients by the OSCAR(5CV). The next largest cluster contains phosphorus, % base saturation, and exchangeable acidity. Again, only one of these predictors—phosphorus—has a nonzero coefficient, according to the OSCAR(5CV). The next largest cluster contains copper and Manganese, both of which have nonzero coefficients, but are not clustered, in the OSCAR(5CV) analysis. The predictors sum cations, calcium, and magnesium were clustered (along with potassium) in the OSCAR(5CV) but are not clustered in Figure 6.7. Finally, the predictor with the largest estimated effect in the OSCAR(5CV) analysis is assigned its own cluster in Figure 6.7 and in the OSCAR(5CV) analysis. It appears that the cluster configurations in Figure 6.7 match fairly closely with the OSCAR(5CV) con-

figurations when the coefficients are estimated to have a zero coefficient. However, when the predictors are estimated to have nonzero effects, the cluster configurations from Figure 6.7 and from the OSCAR(5CV) are markedly different.

6.3 Future Work

We conclude with some thoughts about future research, particularly regarding the clustering of predictors in regression. The work described previously leaves many questions open to further inquiry. We list some of these questions below:

- Under what theoretcial conditions is it optimal to set two coefficients equal to each other as opposed to setting one of the coefficients to zero?
- Because the priors propsed in the Bayesian variable selection and clustering procedure are "automatic" and do not necessarily correspond to any prior subjective beliefs, other theoretical justification should be given for using such a procedure. For example, what are asymptotic properties of the estimators from the Bayesian variable selection and clustering procedure? Do they satisfy oracle-type properties. That is, can we be assured that as the sample size increases, the probability that our procedure will select the "true" cluster configuration goes to one? Do the estimators of the nonzero coefficients converge to their true values as the sample size increases?
- Barbieri and Berger (2004) show that, in a variable selection setting, the median model has several good theoretical properties. When predictors are being clustered or removed from the model, is there a generalizaton of the median model? If so, what are its theoretical properties?

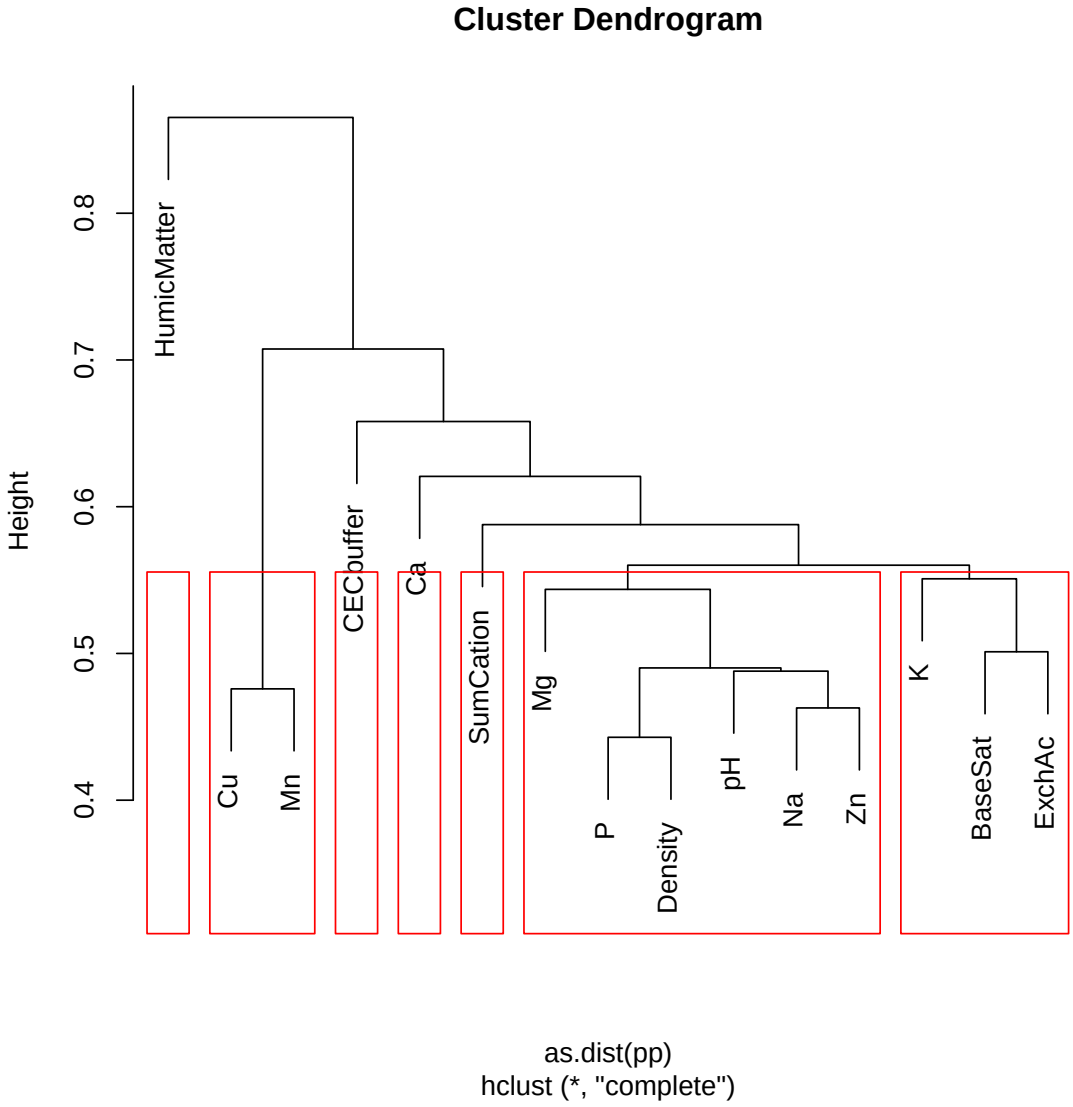

Figure 6.7: Dendrogram for the soil data when pairwise, posterior proabilities are used as a measure of distance. The rectangular boxes indicate the cluster configuration when the tree is cut to yield seven total clusters.

Bibliography

Akaike, H. (1973), "Information theory and an extension of the maximum likelihood principle," in *Second international symposium on information theory*, eds. Petrov, B. N. and Csaki, F.

— (1974), "A new look at statistical model identification." *IEEE Transactions on Automatic Control*, 19, 716–23.

Andrews, D. F. and Mallows, C. L. (1974), "Scale mixtures of normal distributions," *Journal of the Royal Statistical Society, Series B*, 36, 99–102.

Avalos, M., Grandvalet, Y., and Ambroise, C. (2003), "Regularization methods for additive models," in *Lecture Notes in Computer Science*, vol. 2779/2003, pp. 509–520.

Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T., and Silverman, E. (1955), "An empirical distribution function for sampling with incomplete information," *Annals of Mathematical Statistics*, 26, 641–647.

Bacchetti, P. (1989), "Additive isotonic models," *Journal of the American Statistical Association*, 84, 289–294.

Barbieri, M. M. and Berger, J. O. (2004), "Optimal predictive model selection," *Annals of Statistics*, 32, 870–897.

Barlow, R., Bartholomew, D., Bremner, J., and Brunk, H. (1972), *Statistical Inference under Order Restrictions*, Wiley.

Barry, D. (1986), "Nonparmetric Bayesian regression," *Annals of Statistics*, 14, 934–953.

Bartlett, M. (1957), "Comment on D. V. Lindley's statistical paradox," *Biometrika*, 44, 533–534.

Belsley, D. A. (1984), "Demeaning conditioning diagnostics through centering (with discussion)," *The American Statistician*, 38, 73–93.

— (1991), *Conditioning diagnostics: collinearity and weak data in regression*, Wiley.

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980), *Regression diagnostics: Identifying influential data and sources of collinearity*, Wiley.

Berger, J. O. (1985), *Statistical Decision Theory and Bayesian Analysis*, Springer.

Berger, J. O. and Pericchi, L. R. (1996a), "The intrinsic Bayes factor for linear models," in *Bayesian Statistics 6*, Oxford University Press, pp. 23–42.

— (1996b), "The intrinsic Bayes factor for model selection and prediction," *Journal of the American Statistical Association*, 91, 109–122.

Bernstein, S. (1912), "Démonstration du thèorém de Weierstrass, fondeé sur le calcul des probabilités," *Communications of Kharkov Mathematical Society*, 13, 1–2.

Birke, M. and Dette, H. (2007), "Esitmating a convex function in nonparametric regression," *Scandinavian Journal of Statistics*, 34, 384–404.

Blackwell, D. and MacQueen, J. B. (1973), "Ferguson distributions via Polya urn schemes," *Annals of Statistics*, 1, 353–355.

Blanchard, O. J. (1987), "Comment," *Journal of Business and Economic Statisitics*, 5, 449–451.

Bondell, H. D. and Reich, B. J. (2008), "Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR," *Biometrics*, 64, 115–123.

Boos, D. and Stefanski, L. (2008), "Boos-Stefanski variable selection home page," `http://www4.stat.ncsu.edu/~boos/var.select/ncaa.html`.

Breiman, L. (1995), "Better subset selection using the nonnegative garrotte," *Technometrics*, 37, 373–384.

Breiman, L. and Friedman, J. H. (1985), "Estimating optimal transformations for multiple regression and correlation," *Journal of the American Statistical Association*, 80, 580–597.

Brown, B. M. and Chen, S. X. (1999), "Beta-Bernstein smoothing for regression curves with compact support," *Scandinavian Journal of Statistics*, 26, 47–59.

Brunk, H. D. (1955), "Maximum likelihood estimates of monotone parameters," *Annals of Mathematical Statistics*, 26, 607–616.

— (1958), "On the estimation of parameters restriced by inequalities," *Annals of Mathematical Statistics*, 29, 437–454.

Buckland, S. T., Anderson, D. R., Burnham, K. P., Laake, J. L., Borchers, D. L., and Thomas, L. (2001), *Introduction to Distance Sampling: Estimating Abundance of Biological Populations*, Oxford University Press.

Buja, A., Hastie, T., and Tibshirani, R. (1989), "Linear smoothers and additive models," *Annals of Statistics*, 17, 453–555.

Buse, A. (1994), "Brickmaking and the collinear arts: A cautionary tale," *Canadian Journal of Economics*, 27, 408–414.

Cantoni, E., Flemming, J. M., and Ronchetti, E. (2006), "Variable selection in additive models by nonnegative garrote," `http://www.unige.ch/ses/metri/cahiers/2006_02.pdf`, Unversity of Geneva, Department of Economics Working Paper.

Casella, G. and Moreno, E. (2006), "Objective Bayesian variable selection," *Journal of the American Statistical Association*, 101, 157–167.

Chak, P. M., Madras, N., and Smith, B. (2005), "Semi-nonparametric estimation with Bernstein polynomials," *Economics Letters*, 89, 153–156.

Chang, I., Chien, L., Hsiung, C. A., Wen, C., and Wu, Y. (2007), "Shape restricted regression with random Bernstein polynomials," *IMS Lecture Notes—Monograph Series*, 54, 187–202.

Chen, Z. (1993), "Fitting multivariate regression functions by interaction spline models," *Journal of the Royal Statistical Society, Series B*, 55, 473–491.

Cleveland, W. S., Grosse, E., and Shyu, W. M. (1992), "Local regression models," in *Statistical Models in S*, eds. Chambers, J. M. and Hastie, T. J., Chapman & Hall / CRC, pp. 309–376.

Cox, D. R. and Snell, E. J. (1974), "The choice of varibles in obeservational studies," *Applied Statistics*, 23, 51–59.

Craven, P. and Wahba, G. (1979), "Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation," *Numerische Mathematik*, 31, 377–403.

Cripps, E., Kohn, R., and Nott, D. (2006), "Bayesian subset selection and model averaging using a centred and dispersed prior for the error variance," *Australian & New Zealand Journal of Statistics*, 48, 237–252.

Curtis, S. M. and Ghosh, S. K. (2008), "A Bayesian approach to multicollinearity and the simultaneous selection and clustering of predictors in linear regression," unpublished manuscript.

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.

— (1998b), "Bayesian MARS," *Statistics and Computing*, 8, 337–346.

Dette, H., Neumeyer, N., and Pilz, K. F. (2006), "A simple nonparametric estimator of a monotone regression function," *Bernoulli*, 12, 469–490.

Dudzinski, M. L. and Mykytowycz, R. (1961), "The eye lens as an indicator of age in the wild rabbit in Australia," *CSIRO Wildlife Research*, 6, 156–159.

Dunson, D. (2005), "Bayesian semiparametric isotonic regression for count data," *Journal of the American Statistical Association*, 100, 618–627.

Efromovich, S. (1999), *Noparmetric curve estimation: Methods, theory, and estimation*, Springer.

Efron, B. (1979), "Bootstrap methods: Another look at the jackknife," *Annals of Statistics*, 7, 1–26.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least angle regression," *Annals of Statistics*, 32, 407–499.

Ehrlich, I. (1973), "Participation in illegitimate activities: a theoretical and empirical investigation," *Journal of Political Economy*, 81, 521–567.

— (1975), "The deterrent effect of capital punishment: a question of life or death," *American Economic Review*, 65, 397–417.

Eilers, P. H. C. and Marx, B. D. (1996), "Flexible smoothing with B-splines and penalties (with discussion)," *Statistical Science*, 11, 89–121.

Ernst, M. D. (1998), "A multivariate generalized Laplace distribution," *Computational Statistics*, 13, 227–232.

Fan, J. and Li, R. (2001), "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, 96, 1348–1360.

Ferguson, T. (1973), "A Bayesian analysis of some nonparmetric problems," *Annals of Statistics*, 1, 209–230.

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

Friedman, J. and Tibshirani, R. (1984), "The monotone smoothing of scatterplots," *Technometrics*, 26, 243–250.

Friedman, J. H. (1991), "Multivariate adaptive regression splines," *Annals of Statistics*, 19, 1–141.

Gallant, A. R. and Golub, G. H. (1984), "Imposing curvature restrictions on flexible functional forms," *Journal of Econometrics*, 26, 295–321.

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

Gelman, A. (2006), "Prior distributions for variance parameters in hierarchical models," *Bayesian Analysis*, 1, 515–533.

Gelman, A. and Rubin, D. (1992), "Inference from iterative simulation using multiple sequences," *Statistical Science*, 7, 457–511.

Geman, S. and Geman, D. (1984), "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine*, 6, 721–741.

George, E. I. (2000), "The variable selection problem," *Journal of the American Statistical Association*, 95, 1304–1308.

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.

— (1997), "Approaches for Bayesian variable selection," *Statistica Sinica*, 7, 339–373.

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

Geyer, C. J. (1991), "Constrained maximum likelihood exemplified by isotonic convex logistic regression," *Journal of the American Statistical Association*, 86, 717–724.

Ghosal, S. (2008), personal communication.

Ghosh, J. K. and Ramamoorthi, R. V. (2003), *Bayesian Nonparametrics*, Springer.

Goldberger, A. S. (1991), *A Course in Econometrics*, Harvard University Press.

Gopalan, R. and Berry, D. A. (1998), "Bayesian multiple comparisons using Dirichlet process priors," *Journal of the American Statistical Association*, 93, 1130–1139.

Grandvalet, Y. (1999), "Least absolute shrinkage is equivalent to quadratic penalization," in *ICANN'98 (Perspectives in neural computing)*, Springer-Verlag Telos, pp. 201–206.

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

Greene, W. H. (2000), *Econometric analysis*, Prentice Hall, 4th ed.

Gu, C. (2002), *Smoothing spline ANOVA models*, Springer.

Gustafson, P. (2000), "Bayesian regression modeling with interactions and smooth effects," *Journal of the American Statistical Association*, 95, 795–806.

Hald, A. (1952), *Statistical Theory with Engineering Applications*, New York: Wiley.

Hall, P. and Huang, L. (2001), "Nonparametric kernel regression subject to monotonicity constraints," *Annals of Statistics*, 29, 624–647.

Hastie, T. and Tibshirani, R. (1986), "Generalized additive models," *Statistical Science*, 1, 297–318.

— (1990), *Generalized Additive Models*, Chapman and Hall.

Hastie, T., Tibshirani, R., and Friedman, J. (2001), *Elements of Statistical Learning*, Springer-Verlag.

He, X. and Shi, P. (1998), "Monotone B-spline smoothing," *Journal of the American Statistical Association*, 442, 643–649.

Hildreth, C. (1954), "Point estimate of ordinates of concave functions," *Journal of the American Statistical Association*, 49, 598–619.

Hill, R. C. and Adkins, L. C. (2003), "Collinearity," in *A Companion to Theoretical Econometrics*, ed. Baltagi, B. H., Blackwell Publishing, chap. 12, pp. 256–278.

Hoerl, A. E. and Kennard, R. W. (1970), "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, 12, 55–67.

Holmes, C. C. and Heard, N. A. (2003), "Generalized monotonic regression using random change points," *Statistics in Medicine*, 22, 623–638.

Ihaka, R. and Gentleman, R. (1996), "R: A language for data anlaysis and graphics," *Journal of Computational and Graphical Statistics*, 5, 299–314.

Ishwaran, H. and James, L. F. (2001), "Gibbs sampling methods for stick-breaking priors," *Journal of the American Statistical Association*, 96, 161–173.

Ishwaran, H. and Rao, J. S. (2005), "Spike and slab variable selection: Frequentist and Bayesian strategies," *Annals of Statistics*, 33, 730–773.

Ishwaran, H. and Zarepour, M. (2002), "Exact and approximate sum representations for the Dirichlet process," *The Canadian Journal of Statistics*, 30, 269–283.

Johnstone, I. M. and Silverman, B. W. (2005), "Empirical Bayes selection of wavelet thresholds," *Annals of Statistics*, 33, 1700–1752.

Kennedy, P. (1982), "Eliminating problems caused by multicollinearity: A warning," *Journal of Economic Education*, 13, 62–64.

Kennedy, P. E. (1983), "On an inappropriate means of reducing multicollinearity," *Regional Science and Urban Economics*, 13, 579–581.

Krishna, A., Bondell, H. D., and Ghosh, S. K. (2008), "Bayesian variable selection using an adaptive powered correlation prior," (under review).

Kuhn, H. W. and Tucker, A. W. (1951), "Nonlinear programming," in *Proceedings of second Berkeley symposium on mathematical statistics and probability*, ed. Neyman, J., Berkeley: University of California Press.

Kullback, S. and Leibler, R. A. (1951), "On information and sufficiency," *Annals of Mathematical Statistics*, 22, 72–86.

Lafferty, J. and Wasserman, L. (2008), "Rodeo: sparse, greedy nonparmetric regression," *Annals of Statistics*, 36, 28–63.

Lavine, M. and Mockus, A. (1995), "A nonparametric Bayes method for isotonic regression," *Journal of Statistical Planning and Inference*, 46, 235–248.

Liang, F., Paulo, R., Molina, G., Clyde, M., and Berger, J. O. (2008), "Mixtures of g-priors for Bayesian variable selection," *Journal of the American Statistical Association*, 103, 410–423.

Lin, Y. and Zhang, H. H. (2006), "Component selection and smoothing in multivariate nonparametric regression," *Annals of Statistics*, 34, 2272–2297.

Linkletter, C., Bingham, D., Hengartner, N., and Higdon, D. (2006), "Variable selection for Gaussian process models in computer experiments," *Technometrics*, 48, 478–490.

MacLehose, R. F., Dunson, D. B., Herring, A. H., and Hoppin, J. A. (2007), "Bayesian methods for highly correlated exposure data," *Epidemiology*, 18, 2.

Mallows, C. L. (1973), "Some comments on C_p," *Technometrics*, 15, 661–676.

Mammen, E. (1991), "Nonparametric regression under qualitative smoothness assumptions," *Annals of Statistics*, 19, 741–759.

Mammen, E., Marron, J. S., Turlach, B. A., and Wand, M. P. (2001), "A general projection framework for constrained smoothing," *Statistical Science*, 16, 232–248.

Mangold, W. D., Bean, L., and Adams, D. (2003), "The impact of intercollegiate athletics on graduation rates among major ncaa division I universities: Implications for college persistence theory and practice," *The Journal of Higher Education*, 74, 540–562.

Marin, J. and Robert, C. P. (2007), *Bayesian core: a practical approach to Bayesian statistics*, Springer.

McCullagh, P. and Nelder, J. A. (1989), *Generalized Linear Models*, Chapman & Hall/CRC, 2nd ed.

McQuarrie, A. D. R. and Tsai, C. (1998), *Regression and Time Series Model Selection*, World Scientific.

Medvedovic, M. and Sivaganesan, S. (2002), "Bayesian infinite mixture model based clustering of gene expression profiles," *Bioinformatics*, 18, 1194–1206.

Meyer, M. and Woodroofe, M. (2000), "On the degrees of freedom in shape-restricted regression," *Annals of Statistics*, 28, 1083–1104.

Miller, A. (2002), *Subset selection in regression*, Chapman & Hall / CRC, 2nd ed.

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

Monahan, J. F. (2008), *A Primer on Linear Models*, Chapman & Hall/CRC.

Mukerjee, H. (1988), "Monotone nonparametric regression," *Annals of Statistics*, 16, 741–750.

Neal, R. M. (2003), "Density modeling and clustering using Dirichlet diffusion trees," in *Bayesian Statistics 7*, Oxford University Press, pp. 619–629.

Neelon, B. and Dunson, D. B. (2004), "Bayesian isotonic regression and trend analysis," *Biometrics*, 60, 398–406.

Neter, J., Kutner, M. H., Wasserman, W., and Nachtsheim, C. J. (1996), *Applied Linear Statistical Models*, McGraw-Hill / Irwin.

Nicholson, W. (1992), *Microeconomic Theory*, The Dryden Press.

Nocedal, J. and Wright, S. J. (2006), *Numerical optimization*, Springer.

Nott, D. J. (2008), "Predictive performance of Dirichlet process shrinkage methods in linear regression," *Computational Statistics & Data Analysis*, 52, 3658–3669.

O'Hagan, A. (2004), "Dicing with the Unknown," *Significance*, 1, 132–133.

Pilz, K. and Titoff, S. (2005), `monreg`*: Nonparametric monotone regression*, R package version 0.1. Earlier developements by Holger Dette and Kay Pilz.

Pitman, J. (1995), "Exchangeable and partially exchangeable random partitions," *Probability Theory and Related Fields*, 102, 145–158.

— (1996), "Some developments of the Blackwell-MacQueen Pólya urn scheme," in *Statistics, Probability and Game Theory*, Hayward, CA: Institute of Mathematical Statistics, vol. 30 of *IMS Lecture Notes—Monograph Series*, pp. 245–267.

Pitman, J. and Yor, M. (1997), "The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator," *Annals of Probability*, 25, 855–900.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2007), `coda`*: Output analysis and diagnostics for MCMC*, r package version 0.12-1.

R Development Core Team (2007), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Raftery, A. E. (1995), "Bayesian model selection in social research," *Sociological Methodology*, 25, 111–163.

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997), "Bayesian model averaging for linear regression models," *Journal of the American Statistical Association*, 92, 179–191.

Ramsay, J. O. (1988), "Monotone regression splines in action," *Statistical Science*, 3, 425–461.

— (1998), "Estimating smooth monotone functions," *Journal of the Royal Statistical Society, Series B*, 60, 365–375.

Ramsay, J. O. and Abrahamowicz, M. (1989), "Binomial regression with monotone splines - a psychometric application," *Journal of the American Statistical Association*, 84, 906–915.

Rasmussen, C. E. and Williams, C. K. I. (2005), *Gaussian processes for machine learning*, MIT Press.

Ratkowsky, D. A. (1983), *Nonlinear Regression Modelling*, New York: Marcel Dekker.

Reboul, L. (2005), "Estimation of a function under shape restrictions. Applications to reliability," *Annals of Statistics*, 33, 1330–1356.

Reich, B. J., Storlie, C. B., and Bondell, H. B. (2008), "Bayesian variable selection for nonparametric regression," `http://www4.stat.ncsu.edu/~bondell/Bayes_nonp.pdf`, unpublished manuscript.

Rencher, A. C. (2002), *Methods of Multivariate Analysis*, Wiley-Interscience.

Rencher, A. C. and Schaalje, G. B. (2008), *Linear Models in Statistics*, Wiley-Interscience.

Robert, C. P. (1995), "Simulation of truncated normal variables," *Statistics and Computing*, 5, 121–125.

— (2007), *The Bayesian Choice*, Springer Texts in Statistics, paperback ed.

Schell, M. J. and Singh, B. (1997), "The reduced monotonic regression method," *Journal of the American Statistical Association*, 92, 128–135.

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

Scott, J. G. and Berger, J. O. (2006), "An exploration of aspects of Bayesian multiple testing," *Journal of Statistical Planning and Inference*, 136, 2144–2162.

Segal, M. R., Dahlquist, K. D., and Conklin, B. R. (2003), "Regression approaches for microarray data analysis," *Journal of Computational Biology*, 10, 961–980.

Sethuraman, J. (1994), "A Constructive Definition of Dirichlet Priors," *Statistica Sinica*, 4, 639–650.

Shi, P. and Tsai, C. (1999), "Semiparametric regression model selections," *Journal of Statistical Planning and Inference*, 77, 119–139.

Shively, T. S., Kohn, R., and Wood, S. (1999), "Variable selection and function estimation in additive nonparametric regression using a data-based prior," *Journal of the American Statistical Association*, 94, 777–794.

Smith, M. and Kohn, R. (1996), "Nonparametric regression using Bayesian variable selection," *Journal of Econometrics*, 75, 317–343.

Spiegelhalter, D. J., Thomas, A., and Best, N. G. (1999), *WinBUGS Version 1.2 User Manual*, MRC Biostatistics Unit.

Stadtmüller, U. (1986), "Asymptotic properties of nonparametric curve estimates," *Periodica Mathematica Hungarica*, 17, 83–108.

Stone, C. (1985), "Additive regression and other nonparametric models," *Annals of Statistics*, 13, 689–705.

Stone, M. (1974), "Cross-validatory choice and assessment of statistical predictions (with discussion)," *Journal of the Royal Statistical Society, Series B*, 36, 111–147.

Strawderman, W. E. (1971), "Proper Bayes minimax estimators of the multivariate normal mean," *Annals of Mathematical Statistics*, 42, 385–388.

Sturtz, S., Ligges, U., and Gelman, A. (2005), "R2WinBUGS: A Package for Running WinBUGS from R," *Journal of Statistical Software*, 12, 1–16.

Tenbusch, A. (1997), "Nonparametric curve estimation with Bernstein estimates," *Metrika*, 45, 1–30.

Thalange, N. K. S., Foster, P. J., Gill, M. S., Price, D. A., and Clayton, P. E. (1996), "Model of normal prepubertal growth," *Archives of Disease in Childhood*, 75, 427–431.

Theil, H. (1963), "On the use of incomplete prior information in regression analysis," *Journal of the American Statistical Association*, 58, 401–414.

Theil, H. and Goldberger, A. S. (1961), "On pure and mixed statistical estimation in economics," *International Economic Review*, 2, 65–78.

Thomas, A., O'Hara, B., Ligges, U., and Sturtz, S. (2006), "Making BUGS open," *R News*, 6, 12–17.

Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B*, 58, 267–288.

Tierney, L. (1994), "Markov chains for exploring posterior distributions," *Annals of Statistics*, 22, 1701–1728.

Vandaele, W. (1978), "Participation in illegitimate activities: Ehrlich revisited," in *Deterrence and Incapacitation*, eds. Blumstein, A., Cohen, J., and Nagin, D., National Academy of Sciences Press, pp. 270–335.

Venables, W. N. and Ripley, B. D. (2002), *Modern Applied Statistics with S*, Springer.

Wahba, G. (1990), *Spline models for observational data*, vol. 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics 59*, SIAM.

Wasserman, L. (2005), *All of nonparametric statistics*, Springer.

Wei, B.-C. (1998), *Exponential Family Nonlinear Models*, Springer.

West, M. (2003), "Bayesian factor regression in the 'large p small n' problem," in *Bayesian Statistics 7*, Oxford University Pressl, pp. 733–743.

Wood, S., Kohn, R., Shively, T., and Jiang, W. (2002), "Model selection in spline nonparametric regression," *Journal of the Royal Statistical Society, Series B*, 64, 119–139.

Woods, H., Steinour, H. H., and Starke, H. R. (1932), "Effect of composition of Portland cement on heat evolved during hardening," *Industrial Engineering and Chemistry*, 24, 1207–1214.

Wu, Y., Boos, D. D., and Stefanski, L. A. (2007), "Controlling variable selection by the addition of pseudovariables," *Journal of the American Statistical Association*, 102, 235–243.

Yuan, M. and Lin, Y. (2005), "Efficient empirical Bayes variable selection and estimation in linear models," *Journal of the American Statistical Association*, 100, 1215–1225.

— (2006), "Model selection and estimation in regression with grouped varibales," *Journal of the Royal Statistical Society, Series B*, 68, 49–67.

Zellner, A. (1986), "On assessing prior distributions and Bayesian regression analysis with g-prior distributions," in *Bayesian inference and decision techniques: essays in honor of Bruno de Finetti*, Elseivier Science Publishers, B. V., pp. 233–243.

Zellner, A. and Siow, A. (1980), "Posterior odds ratios for selected regression hypotheses," in *Bayesian statistics: Proceedings of the first international meeting held in Valencia*, Oxford University Press, pp. 585–603.

Zou, H. (2006), "The adaptive LASSO and its oracle properties," *Journal of the American Statistical Association*, 101, 1418–1429.

Zou, H. and Hastie, T. (2005), "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society, Series B*, 67, 301–320.

Appendices

Appendix A

Bayesian Selection and Clustering of Predictors in Linear Regression: Computational Details

The Gibbs sampler for the Bayesian variable selection and clustering model was constructed using the `OpenBUGS` software (Thomas et al., 2006) with the following code:

```
model{
  for(i in 1:n) {
    y[i] ~ dnorm(mu[i],isigsq)
    mu[i] <- inprod(X[i,],beta[])
    }

  alpha <- 1

  for(j in 1:M){
    xi[j] ~ dnorm(0, itausq)
    a.p[j] <- alpha/K
    }

  for(j in 1:K){
    S[j] ~ dcat(p[1:M])
    theta[j] <- xi[S[j]]
    gamma[j] ~ dbern(pi)
    beta[j] <- theta[j]*gamma[j]
    }

  p[1:M] ~ ddirch(a.p[])
```

```
pi ~ dunif(0,1)

# Prior for sigsq
isigsq ~ dgamma(0.1,0.1)
sigsq <- 1/isigsq

# Prior for tausq
itausq ~ dgamma(0.1,0.1)
tausq <- 1/itausq
}
```

We used functions from the `R` package `R2WinBUGS` (Sturtz, Ligges, and Gelman, 2005) to call `OpenBUGS` from `R`. Functions in the `R` package `coda` (Plummer, Best, Cowles, and Vines, 2007) were used for post-MCMC analysis and diagnostics checks. Specifically, the methods of Gelman and Rubin (1992) were used to check for convergence of the Gibbs sampler.

Appendix B

Monotonic Regression with Bernstein Polynomials

B.1 Computational Details

B.1.1 Programming Strategy

For speed, we coded our MCMC algorithm in a `C` function that we called from `R`. We compiled our `C` function using the GNU `C` compiler as found in `Rtools` "package" created by Brian Ripley and maintained by Duncan Murdoch (at `http://www.murdoch-sutherland.com/Rtools/`). Pseudo-random truncated normals were computed using the algorithm of Robert (1995). MCMC diagnostics and plots were prepared using the `coda` package in `R` (Plummer, Best, Cowles, and Vines, 2007). For the Downs syndrome example, we used the `R2WinBUGS` package (Sturtz, Ligges, and Gelman, 2005) to conveniently call `WinBUGS` from `R`.

B.1.2 Markov Chain Monte Carlo Algorithm

The following describes the Gibbs sampler for computing posterior draws from the Bayesian monotonic regression procedure we have outlined in this paper. The complete conditionals are as in Geweke (1996), except that Geweke's formulation implicitly relies on γ_k indicator variables for whether the corresponding u_k is zero.

In our formulation, we make this dependence explicit. We also add updates for the additional parameters τ^2 and p_γ.

Begin the MCMC algorithm by initializing all parameters—$u_0^{(0)}$, $\tau^{2(0)}$, $p_\gamma^{(0)}$, $\sigma^{2(0)}$, $\boldsymbol{\gamma}^{(0)} = (\gamma_1^{(0)}, \ldots, \gamma_M^{(0)})$, and $\mathbf{u}^{(0)} = (u_1^{(0)}, \ldots, u_M^{(0)})$ (conditional on the $\boldsymbol{\gamma}^{(0)}$). For $t = 1, \ldots, n_{iter}$ simulate in turn from each of the following distributions:

1. $\sigma^{2(t)} \sim \mathsf{InvGam}\Big(a_\sigma + \frac{n}{2}, b_\sigma + \sum_{i=1}^n (y_i - u_0^{(t-1)} - \mathbf{w}_i^T \mathbf{u}^{(t-1)})^2/2\Big)$

2. $u_0^{(t)} \sim \mathsf{N}\Big(\frac{m_0/s_0^2 + n\bar{z}/\sigma^2}{1/s_0^2 + n/\sigma^2}, (1/s_0^2 + n/\sigma^2)^{-1}\Big)$, where $z_i = y_i - \mathbf{w}_i^T \mathbf{u}^{(t-1)}$ and $\bar{z} = \sum_{i=1}^n z_i/n$.

3. For $k = 1, \ldots, M$, simulate jointly $\gamma_k^{(t)}$ and $u_k^{(t)}$ by first simulating

$$\gamma_k^{(t)} \sim \mathsf{Bern}\left(\frac{\varpi_1}{\varpi_1 + \varpi_0}\right),$$

and then setting $u_k^{(t)} = 0$ if $\gamma_k^{(t)} = 1$ or otherwise simulating

$$u_k^{(t)} \sim \mathsf{N}_{(0,\infty)}\big(m_*, \sigma_*^2\big)$$

where

$$\begin{aligned}
\varpi_1 &= \exp\left\{-\sum_{i=1}^n \frac{z_i^2}{2\sigma^{2(t)}}\right\} p_\gamma^{(t-1)}, \\
\varpi_0 &= 2(\sigma_*^2/\tau^{2(t-1)})^{1/2} \exp\left\{-\frac{\sum_{i=1}^n (z_i - b_k w_{ik})^2}{2\sigma^{2(t)}} - \frac{b_k^2}{2\sigma_k^2} + \frac{m_*^2}{2\sigma_*^2}\right\} \times \\
&\qquad \left[1 - \Phi\left(-\frac{m_*}{\sigma_*}\right)\right](1 - p_\gamma^{(t-1)}), \\
z_i &= y_i - u_0^{(t)} - w_{i,1}u_1^{(t)} - \cdots - w_{i,k-1}u_{k-1}^{(t)} - w_{i,k+1}u_{k+1}^{(t-1)} - \cdots - w_{i,M}u_M^{(t-1)}, \\
m_* &= \frac{\tau^{2(t-1)} b_k}{\tau^{2(t-1)} + \sigma_k^2}, \\
\sigma_*^2 &= \frac{\sigma_k^2 \tau^{2(t-1)}}{\tau^{2(t-1)} + \sigma_k^2}, \\
\sigma_k^2 &= \frac{\sigma^{2(t)}}{\sum_{i=1}^n w_{ik}^2}, \\
b_k &= \frac{\sum_{i=1}^n z_i w_{ik}}{\sum_{i=1}^n w_{ik}^2},
\end{aligned}$$

$\mathsf{N}_{(a,b)}(\xi, \nu^2)$ is a truncated normal distribution on the interval (a, b), and $\Phi(\cdot)$ is the cdf of the standard normal distribution.

4. $\tau^{2(t)} \sim \mathsf{InvGam}\left(a_\tau + \frac{M - \left\|\boldsymbol{\gamma}^{(t)}\right\|}{2}, b_\tau + \frac{\sum_{j=1}^{M}(u_j^{(t)})^2}{2}\right)$

5. $p_\gamma^{(t)} \sim \mathsf{Beta}\left(a_p + \left\|\boldsymbol{\gamma}^{(t)}\right\|, b_p + M - \left\|\boldsymbol{\gamma}^{(t)}\right\|\right)$, where values $a_p = 1$ and $b_p = 1$ are used for all examples and simulations in Chapter 4.

B.2 Derivation of the Expectation of Ψ

Define random variables $U_1, \ldots, U_M$ with a pdf that is a mixture of a point mass at zero (with weight p_k) and a truncated normal random variable truncated to be greater than zero (with weight $1 - p_k$). Thus, the expected value of U_k is

$$\mathsf{E}(U_k) = (1 - p_k)\mathsf{E}(Y)$$

where $Y \sim \mathsf{N}(\mu, \sigma^2)\,\mathbb{1}\,(\ell, \infty)$ with $\mu = 0$, $\sigma^2 = \tau^2$ (according to our notation in our paper), and $\ell = 0$. The expected value of a truncated random normal variate is (Greene, 2000, page 899)

$$\mathsf{E}(Y) = \mu + \sigma\lambda(\alpha)$$

where $\lambda(\alpha) = \frac{\phi(\alpha)}{1-\Phi(\alpha)}$, $\alpha = \frac{(\ell-\mu)}{\sigma}$, $\phi(\cdot)$ is the density function of a standard normal variate, and $\Phi(\cdot)$ is the distribution function of a standard normal variate. Thus, in our case with $\alpha = 0$, we have

$$\begin{aligned}\mathsf{E}(Y) &= \tau\frac{\phi(0)}{1 - \Phi(0)} \\ &= \tau(2\pi)^{-1/2}/0.5 \\ &= \tau\sqrt{2/\pi}\end{aligned}$$

Then

$$\mathsf{E}(U_k) = (1 - p_k)\tau\sqrt{2/\pi}.$$

Using this result, we can obtain the expected value of the average flatness Ψ over an interval (a, b) as a function of the user-specified parameter M. For simplicity, we will

assume that $p_k = p$ for all k and we let $q = 1 - p$ and $\mu = q\tau\sqrt{2/\pi}$. (Note, that this

expected value is conditional on τ, which, in our specification, is given gamma prior.)

$$
\begin{aligned}
\mathsf{E}(\Psi) &= \frac{\mathsf{E}\left(\sum_{k=1}^{M} U_k[w_M(b,k) - w_M(a,k)]\right)}{b-a} \\
&= \frac{\sum_{k=1}^{M} \mathsf{E}(U_k)[w_M(b,k) - w_M(a,k)]}{b-a} \\
&= \frac{\sum_{k=1}^{M} q\tau\sqrt{2\pi}[w_M(b,k) - w_M(a,k)]}{b-a} \\
&= \frac{q\tau\sqrt{2\pi}}{b-a} \sum_{k=1}^{M} [w_M(b,k) - w_M(a,k)] \\
&= \frac{\mu}{b-a} \sum_{k=1}^{M} \left[\sum_{j=k}^{M} b_M(b,j) - \sum_{j=k} b_M(a,j) \right] \\
&= \mu^* \sum_{k=1}^{M} \left[\sum_{j=k}^{M} \binom{M}{j} b^j (1-b)^{M-j} - \sum_{j=k} \binom{M}{j} a^j (1-a)^{M-j} \right] \\
&= \mu^* \sum_{k=1}^{M} \left[\sum_{j=k}^{M} P(X_M^{(b)} = j) - \sum_{j=k}^{M} P(X_M^{(a)} = j) \right] \\
&= \mu^* \sum_{k=1}^{M} \left[P(X_M^{(b)} \geq k) - P(X_M^{(a)} \geq k) \right] \\
&= \mu^* \sum_{k=1}^{M} \left[P(X_M^{(b)} > k) + P(X_M^{(b)} = k) - P(X_M^{(a)} > k) - P(X_M^{(a)} = k) \right] \\
&= \mu^* \sum_{k=1}^{M} \left\{ [1 - P(X_M^{(b)} \leq k)] + P(X_M^{(b)} = k) - [1 - P(X_M^{(a)} \leq k)] - P(X_M^{(a)} = k) \right\} \\
&= \mu^* \sum_{k=1}^{M} \left\{ [1 - F_{X_M^{(b)}}(k)] + P(X_M^{(b)} = k) - [1 - F_{X_M^{(a)}}(k)] - P(X_M^{(a)} = k) \right\} \\
&= \mu^* \sum_{k=1}^{M} \left\{ [1 - F_{X_M^{(b)}}(k)] - [1 - F_{X_M^{(a)}}(k)] \right\} + P(X_M^{(b)} \geq 1) - P(X_M^{(a)} \geq 1) \\
&= \mu^* \sum_{k=1}^{M} \left\{ [1 - F_{X_M^{(b)}}(k)] - [1 - F_{X_M^{(a)}}(k)] \right\} + [1 - P(X_M^{(b)} = 0)] - [1 - P(X_M^{(a)} = 0)] \\
&= \mu^* \sum_{k=1}^{M} \left\{ [1 - F_{X_M^{(b)}}(k)] - [1 - F_{X_M^{(a)}}(k)] \right\} + [1 - P(X_M^{(b)} \leq 0)] - [1 - P(X_M^{(a)} \leq 0)] \\
&= \mu^* \sum_{k=1}^{M} \left\{ [1 - F_{X_M^{(b)}}(k)] - [1 - F_{X_M^{(a)}}(k)] \right\} + [1 - F_{X_M^{(b)}}(0)] - [1 - F_{X_M^{(a)}}(0)] \\
&= \mu^* \sum_{k=0}^{M} \left\{ [1 - F_{X_M^{(b)}}(k)] - [1 - F_{X_M^{(a)}}(k)] \right\} \\
&= \mu^* \left\{ \sum_{k=0}^{M} [1 - F_{X_M^{(b)}}(k)] - \sum_{k=0}^{M} [1 - F_{X_M^{(a)}}(k)] \right\} \\
&= \mu^* \left[\mathsf{E}(X_M^{(b)}) - \mathsf{E}(X_M^{(a)}) \right] \\
&= \mu^*(Mb - Ma) \\
&= \mu^* M(b-a) \\
&= \frac{\mu}{b-a} M(b-a) \\
&= q\tau\sqrt{2\pi} M
\end{aligned}
$$

where $X_N^{(\xi)} \sim \mathsf{Bin}(N, \xi)$.

B.3 Derivation of the Variance of Ψ

To derive the variance of Ψ, we need the variance of U_k. (For convenience, we will drop the dependence of U on the subscript k.) The variance of a random truncated normal variable Y is (see Greene, 2000, page 899)

$$\sigma^2[1 - \delta(\alpha)],$$

where $\delta(\alpha) = \lambda(\alpha)[\lambda(\alpha) - \alpha]$ and $\lambda(\alpha)$ is as previously defined. Thus, if, as before, we have $\mu = 0$, $\sigma^2 = \tau^2$, $\ell = 0$, then $\alpha = 0$ and

$$\begin{aligned}
\mathsf{Var}(Y) &= \tau^2 \left[1 - \lambda(0)^2\right] \\
&= \tau^2 \left\{1 - \frac{\phi(0)^2}{[1 - \Phi(0)]^2}\right\} \\
&= \tau^2 \left\{1 - \frac{(2\pi)^{-1}}{1/2^2}\right\} \\
&= \tau^2 \left\{1 - \frac{2}{\pi}\right\}
\end{aligned}$$

To compute the value of $\mathsf{Var}(U)$, we define the latent variable $Z \sim \mathsf{Bern}(p)$. We can calculate the variance of U by conditioning on Z as in

$$\mathsf{Var}(U) = \mathsf{E}[\mathsf{Var}(U|Z)] + \mathsf{Var}[\mathsf{E}(U|Z)]$$

Thus, we have

$$\begin{aligned}
\mathsf{Var}(U|Z = 1) &= 0 \\
\mathsf{Var}(U|Z = 0) &= \tau^2 \left\{1 - \frac{2}{\pi}\right\}
\end{aligned}$$

and

$$\begin{aligned}\mathsf{E}(U|Z=1) &= 0\\ \mathsf{E}(U|Z=0) &= \tau\sqrt{2/\pi}\end{aligned}$$

Thus,

$$\begin{aligned}\mathsf{E}[\mathsf{Var}(U|Z)] &= p\cdot 0 + (1-p)\tau^2\left\{1-\frac{2}{\pi}\right\}\\ &= (1-p)\tau^2\left\{1-\frac{2}{\pi}\right\}\\ \mathsf{Var}[\mathsf{E}(U|Z)] &= \mathsf{E}[\mathsf{E}(U|Z)^2] - \mathsf{E}[\mathsf{E}(U|Z)]^2\\ &= [p\cdot 0 + (1-p)2\tau^2/\pi] - [p\cdot 0 + (1-p)\tau\sqrt{2/\pi}]^2\\ &= (1-p)2\tau^2/\pi - (1-p)^2 2\tau^2/\pi\\ &= p(1-p)2\tau^2/\pi,\end{aligned}$$

which gives

$$\begin{aligned}\mathsf{Var}(U) &= (1-p)\tau^2\left\{1-\frac{2}{\pi}\right\} + p(1-p)2\tau^2/\pi\\ &= (1-p)\tau^2\left[1-\frac{2}{\pi}+\frac{2p}{\pi}\right]\\ &= (1-p)\tau^2\left[1-\frac{2}{\pi}(1-p)\right]\\ &= q\tau^2\left[1-\frac{2}{\pi}q\right]\end{aligned}$$

Then the variance of Ψ is

$$\begin{aligned}\mathsf{Var}(\Psi) &= \mathsf{Var}\left\{\frac{\sum_{k=1}^{M} U_k[w_M(b,k)-w_M(a,k)]}{b-a}\right\}\\ &= \frac{\mathsf{Var}(U_1)\sum_{k=1}^{M}[w_M(b,k)-w_M(a,k)]^2}{(b-a)^2}\\ &= \frac{q\tau^2\left[1-\frac{2}{\pi}q\right]}{(b-a)^2}\sum_{k=1}^{M}[w_M(b,k)-w_M(a,k)]^2\end{aligned}$$