Bayesian variable selection using I-priors

Haziq Jamil
Mathematical Sciences, Faculty of Science
Universiti Brunei Darussalam

Wicher Bergsma
Department of Statistics
London School of Economics
and Political Science

February 15, 2020

Abstract

In statistical modelling, there is often a genuine interest to learn the most reasonable, parsimonious, and interpretable model that fits the data. We turn our attention to the problem of variable selection in the context of ordinary linear regressions. Model selection is indeed a vastly covered topic, so our focus is on the Bayesian approach to model selection, emphasising the selection of variables through inferences on model probabilities. The appeal of Bayesian methods are that it reduces the selection problem to one of estimation, rather than a true search of the variable space for the model that optimises a certain criterion. The I-prior (Bergsma, 2019) further enhances this appeal, as we discover I will talk about several Bayesian variable selection methods out there in the literature, and how we can make use of I-priors to improve on results in the presence of multicollinearity.

Chapter 8

Bayesian variable selection using I-priors

Consider the normal linear regression model, where response variables $y_1, \dots, y_n$ relate to several independent variables linearly through the following equation:

$$\begin{aligned} y_i &= \sum_{k=1}^{p} x_{ik}\beta_k + \epsilon_i \\ \epsilon_i &\overset{\text{iid}}{\sim} \text{N}(0, \sigma^2). \end{aligned} \tag{8.1}$$

The topic of interest for this chapter is model selection for linear regression models. That is, from a set of p covariates or predictors $\{X_1, \dots, X_p\}$, the task is to determine the best choice of subset(s) of variables that should be included in a regression model used to explain the variation in the response variable. As such, the term *variable selection* is synonymous to model selection for linear regression models. Fundamental to this notion of variable selection is an inherent belief in sparseness of the true data generative process surrounding the response variable, i.e. not all of the variables need be used to predict the response. Model selection is indeed a huge topic to cover fully. We broadly classify variable selection into three categories: 1) (pairwise) model comparison using some criterion; 2) shrinkage to induce sparsity; and 3) Bayesian model selection. We understand that different categorisations and hence categories of model selection exist in the literature, but our focus is on the discussion of the three types as mentioned.

Model selection criteria, both from a frequentist and Bayesian standpoint, can either be of a predictive nature (e.g. R^2, mean squared error of prediction (MSEP), C_p (Mallows, 1973), k-fold cross-validation MSEP, etc.), or be based on likelihood (e.g. likelihood ratios, Bayes factors, Akaike information criterion (AIC, Akaike, 1973), Bayesian information criterion (BIC, Schwarz, 1978), etc.). Selecting a model based on either of these criteria requires comparison of all 2^p criteria, which is not feasible for large p. Typically, these criteria are used in conjunction with stepwise procedures such as forward-selection or backward-deletion to restrict attention to a smaller number of potential subsets (George and McCulloch, 1993; Miller, 2002).

On the other hand, regularised least squares regression (ridge regression (Hoerl and Kennard, 1970), Lasso (Tibshirani, 1996), or a convex combination of the two via elastic nets (Zou and Hastie, 2005), etc.) provide additional information to the regression model in order to provide a sparse solution to linear system of equations in $\boldsymbol{\beta}$. These methods are proven to be popular as they are fast and perform exceptionally well in many situations, even in cases where $p > n$. Additionally, the Lasso produces solutions for $\boldsymbol{\beta}$ which are exactly zero. However, the Lasso in general produces estimates which are biased towards zero, are inconsistent, and have no valid standard errors (Friedman et al., 2001; Kyung et al., 2010). Further criticisms of the Lasso include its inability to select more than n predictors in a $p > n$ situation, and poor performance when multicollinearity exists among the covariates.

From a Bayesian perspective, regularisation is akin to placing priors on the β_k's to shrink the effects of the β_k's: the ridge regression has a Bayesian interpretation of placing normal priors on the regression coefficients, while for the Lasso, a Laplace or double exponential prior (Park and Casella, 2008). The term adaptive shrinkage has been used for the method in which hyperpriors are placed on the scale parameter of the prior for the β_k's. The idea is to adaptively shape the prior depending on the importance of the variable in the regression model. Bayesian shrinkage includes the task of specifying tuning parameters. This could potentially affect chain mixing in a Markov chain Monte Carlo method (MCMC) procedure, the estimation method that is commonly used.

True Bayesian model selection is probabilistic in nature: a priori, one assigns probabilities over the set of models, and then after observing the data, posterior model probabilities (PMPs) are used to discern which of the models was likeliest to have been behind the data generative process of the observed responses. Of course, with large p, calculation of all 2^p posterior model probabilities to ascertain which is highest will be a challenge, if not impossible. But, as with most Bayesian applications, MCMC can be applied as a practical means of overcoming this intractability. This stochastic approach to variable selection was pioneered by George and McCulloch (1993), and studied by others such as Dellaportas et al. (2002), Kuo and Mallick (1998), and Ntzoufras (2011). Unlike shrinkage methods, Bayesian model selection is able to quantify the amount of times a variable "enters the model" (inclusion probabilities), and thereby measuring its worth as a predictor.

Note that, in addition to model probabilities and inclusion probabilities, estimates of regression coefficients are obtained simultaneously in Bayesian variable selection. When several competing models have high posterior probabilities, regression coefficients from each model, or indeed any quantity of interest, may be combined and weighted by their posterior model probabilities—a technique known as *Bayesian model averaging* (Hoeting et al., 1999; Madigan and Raftery, 1994). Averaging over a set of models takes into account the uncertainty surrounding model selection, which other standard statistical procedures ignore upon selection of a single model from which to do inference. It is known to be the case that predictive accuracy of the model-averaged quantity is improved, as measured by a logarithmic scoring rule (Raftery et al., 1997).

Bayesian model selection is not without criticism, however. For complex models with many predictors or samples, MCMC is slow and may mix poorly (O'Hara and Sillanpää, 2009). Often, there are a lot of tuning parameters that need to be set correctly for the problem at hand. Also, the choice of priors for model parameters affects consistency of Bayesian model selection procedures. Specifically, improper priors cannot be used to calculate posterior model probabilities (Casella et al., 2009)—otherwise, one risks running into Lindley's paradox[1] (Lindley, 1957).

[1] Briefly, in testing a point null hypothesis of the mean of a normally distributed parameter, the null hypothesis is increasingly accepted as the prior variance of the parameter approaches infinity, regardless of evidence for or against the null. The paradox is also termed Jeffreys-Lindley paradox (Robert, 2014).

The plan for this chapter is to describe a fully Bayesian model for variable selection using I-priors, a novel objective prior for the regression function proposed by Bergsma (2019). The approach that we take is a stochastic search of the model space due to Kuo and Mallick (1998), realised through a simple Gibbs sampling procedure. The main motivation behind using I-priors in Bayesian variable selection is its suitability in accommodating to datasets with strong multicollinearity and being able to run with little to no prior information about the parameters. A simulation study is conducted and several real-world examples presented to demonstrate this fact.

8.1 Brief introduction to I-priors

For n pairs of data $\{(y_1, x_1), \dots, (y_n, x_n)\}$, a general regression model is considered:

$$\begin{aligned} y_i &= f(x_i) + \epsilon_i \\ \epsilon_i &\overset{\text{iid}}{\sim} \text{N}(0, \sigma^2). \end{aligned} \tag{8.2}$$

where $y_i \in \mathbb{R}$, $x_i \in \mathbb{R}^p$ and f belongs to some reproducing kernel Hilbert space (RKHS) $\mathcal{F}$ with kernel $h : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$. We are interested in estimating the regression function f.

For this purpose, we define the I-prior as the distribution maximizing entropy (subject to suitable constraints) based on the Fisher information on the regression function. For the present model, it is Gaussian with covariance kernel proportional to the Fisher information, and mean chosen a priori (e.g., 0). Note that an RKHS $\mathcal{F}$ is spanned by the functions $h(\cdot, x)$ (Berlinet and Thomas-Agnan, 2004). Therefore, f belonging to a linear RKHS $\mathcal{F}$ with kernel $h(x, x') = x^\top x'$, f can be represented by

$$f(x_i) = \sum_{j=1}^n h(x_i, x_j) w_j = \sum_{j=1}^n x_i^\top x_j w_j \tag{8.3}$$

for some real-valued constants $w_1, \dots, w_n$. We see that $x^\top x_i = \sum_{k=1}^p x_{ik} x_{jk}$, and therefore

$$\begin{aligned} f(x_i) &= \sum_{j=1}^n x_i^\top x_j w_j \\ &= \sum_{j=1}^n \left(\sum_{k=1}^p x_{ik} x_{jk} \right) w_j \\ &= \sum_{k=1}^p \underbrace{\left(\sum_{j=1}^n x_{jk} w_j \right)}_{\beta_k} x_{ik} \end{aligned}$$

which is identical to the (8.1), the normal linear regression model described at the beginning of this chapter. See Chapter 4 of Jamil (2018b) for more details.

It is well known that the Fisher information for the regression coefficients of model (8.1) is

$$\mathcal{I}(\beta) \propto X^\top X$$

where X is an $n \times p$ matrix whose rows are x_i. Therefore, an I-prior for $\beta = (\beta_1, \dots, \beta_p)^\top$ is

$$\beta \sim \mathrm{N}_p(0, \kappa\sigma^2 X^\top X) \tag{8.4}$$

where κ is some scale parameter that needs to be estimated. Having the σ^2 appear in the covariance of β proves to be an advantage when estimating the variable selection model fully Bayesian under a Gibbs sampling approach (see section 8.4).

The I-prior has the intuitively appealing property that the more information is available about a linear functional of the regression function, the larger its prior variance, and, broadly speaking, the less influential the prior is on the posterior. This is in contrast to the g-prior, which has covariance proportional to the *inverse* Fisher information. Analysis of some real data sets and a small-scale simulation study show competitive performance of the I-prior methodology (Bergsma, 2019; Jamil, 2018b).

8.2 Preliminary: model probabilities, model evidence and Bayes factors

The paradigm of model selection is as follows. From a finite set of models $\mathcal{M} = \{M_1, \dots, M_K\}$, pairs of data $\{(y_1, \mathbf{x}_1), \dots, (y_n, \mathbf{x}_n)\}$, $y_i \in \mathbb{R}$ and $\mathbf{x}_i \in \mathcal{X} \equiv \mathbb{R}^p$, had been generated according to the generative process dictated by one of the models $M_k \in \mathcal{M}$ and its respective parameters Θ_k. Having observed only this data set, the goal is to infer which of the models had generated the data, and consequently obtain estimates for the parameters. It is perhaps most natural to ponder which of the models is most likely to be the "true" one given the data presented, and thus, this natural way of thinking leads one to the concept of *model probabilities.* From a Bayesian perspective in particular, *posterior model probabilities* allow us to quantify the certainty to which any model is behind the data generative process, after taking into account relevant evidence (observation of the data) and prior beliefs about model and parameter uncertainty.

Let $p(M_1), \dots, p(M_K)$ be prior probabilities assigned to the model space $\mathcal{M}$, and $p(\Theta_k|M_k)$ be the prior on the parameters of model M_k. For any model $M_k \in \mathcal{M}$, the posterior model probability for model m is

$$p(M_k|\mathbf{y}) = \frac{p(\mathbf{y}|M_k)p(M_k)}{\sum_{k=1}^K p(\mathbf{y}|M_k)p(M_k)} \tag{8.5}$$

where

$$p(\mathbf{y}|M_k) = \int p(\mathbf{y}|M_k, \Theta_k)p(\Theta_k|M_k)\,\mathrm{d}\Theta_k \tag{8.6}$$

is known as the marginal likelihood, or *evidence*, for model M_k. As a remark, the prior distributions for the parameters do not necessarily need to depend on the model, so we might have that $p(\Theta_k|M_k) = p(\Theta_k)$. A natural strategy for model selection is to select the model such that $p(M_k|\mathbf{y})$ is largest (the *highest probability model*, HPM), but several models rather than just a single one may be reported to convey model uncertainty (Chipman et al., 2001).

Note, that models may be pairwise compared based on these posterior model probabilities, for which the posterior odds

$$\frac{p(M_k|\mathbf{y})}{p(M_0|\mathbf{y})} = \overbrace{\frac{p(\mathbf{y}|M_k)}{p(\mathbf{y}|M_0)}}^{\text{Bayes factor}} \times \overbrace{\frac{p(M_k)}{p(M_0)}}^{\text{prior odds}} \tag{8.7}$$

provide a point summary for comparing model M_k against model M_0. The first term on the right-hand side is the Bayes factor for comparing any model $M_k \in \mathcal{M}$ to another model $M_0 \in \mathcal{M}$, and is denoted by $\mathrm{BF}(M_k, M_0)$. Thus, model selection based on posterior model probabilities can be formalised as the Bayesian alternative to classical hypothesis testing using Bayes factors (Kass and Raftery, 1995).

The issue that is faced with Bayesian model selection is that all posterior model probabilities must be calculated in order for a full comparison to be made. When the model space is very large, this can prove to be an insurmountable task. In the case of linear regression, where each of the p variables may be selected or not, the size of the model space is 2^p. Even for moderate sized p this can already be a challenge computationally. In the coming sections, we shall see that this problem is alleviated by the use of MCMC methods to evaluate posterior model probabilities.

8.3 The Bayesian variable selection model

We shall loosely refer to a model as a subset of variables selected from the full set of variables $\{X_1, \dots, X_p\}$. It would be useful to be able to index each of these 2^p possible models somehow, and we achieve this by the use of indicator variables $\gamma = (\gamma_1, \dots, \gamma_p) \in \{0,1\}^p$. Let $\gamma_j = 1$ if the variable X_j is selected, and $\gamma_j = 0$ otherwise, for $j = 1, \dots, p$. As an example, the full model, where all the variables are included in the model, is denoted by $\gamma = (1, \dots, 1)$, while the intercept only model is denoted by $\gamma = (0, \dots, 0)$. Note that we do not consider the intercept to be selectable.

Following Kuo and Mallick (1998), the linear model in eq. (8.1) is expanded to include the indicator variables to form

$$\begin{gathered} y_i = \alpha + \sum_{k=1}^{p} x_{ik}\gamma_k\beta_k + \epsilon_i \\ \epsilon_i \overset{\text{iid}}{\sim} \mathrm{N}_n(0, \sigma^2). \end{gathered} \tag{8.8}$$

Hence, in addition to the usual model parameters $(\boldsymbol{\beta}, \sigma, \alpha)$, we are also interested in conducting model inferences through the posterior distribution of the γ's. The priors for the parameters are described below:

- **Model indicators** γ_j. An independent Bernoulli prior is specified for the model indicators

  $$p(\gamma) = \prod_{j=1}^{p} \pi_j^{\gamma_j}(1-\pi_j)^{1-\gamma_j}. \tag{8.9}$$

  We may choose to set all $\pi_j = 0.5$ a priori to reflect equally likely probabilities that any model may be chosen. Alternatively, we might have some subjective beliefs about which predictor is more likely or unlikely to be included in the model. We may also choose to include π_j in the estimation procedure by assigning a hyperprior on π_j such as the Beta(1, 1) (uniform distribution), Beta(1/2, 1/2) (Jeffreys prior), or any other suitable hyperprior.

In any case, in this chapter we consider the simplest case of setting all $\pi_j = 0.5$.

- **Regression coefficients $\boldsymbol{\beta}$.** The Kuo and Mallick (1998) model is often known as the independent sampler due to the independence of model parameters and the indicator variables, i.e., $p(\boldsymbol{\beta}, \gamma) = p(\boldsymbol{\beta})p(\gamma)$. As such, prior choices for the regression coefficients can be any of the usual priors on $\boldsymbol{\beta}$, including (but not limited to)
  - the independent prior $\boldsymbol{\beta} \sim \mathrm{N}_p(\mathbf{0}, c^2\mathbf{I}_p)$ for some choice of c (e.g. $c = 10$);
  - the g-prior $\boldsymbol{\beta}|\sigma, g \sim \mathrm{N}_p(\mathbf{0}, g\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1})$ for some g either chosen a priori or estimated (Bayes or empirical Bayes); or
  - the I-prior $\boldsymbol{\beta}|\sigma, \kappa \sim \mathrm{N}_p(\mathbf{0}, \kappa\sigma^2\mathbf{X}^\top\mathbf{X})$, which is the focus of this chapter.
- **Intercept** α. A normal prior $\alpha \sim \mathrm{N}(0, \sigma^2 A)$.
- **Scale** σ. An inverse gamma prior $\sigma \sim \Gamma^{-1}(c, d)$.

Priors for the intercept and scale parameters are chosen so as to maintain conjugacy to the normal regression model. Choices for the prior hyperparameters depend on the user's prior beliefs, but it is reasonable to set vague and uninformative hyperparameters to let the data speak as much as it can, especially in the absence of prior information. With this in mind, we may choose large values of A (e.g. 100) and small values of the shape and scale parameters for the inverse gamma (e.g. 0.001). Note that as $c, d \to 0$ in the inverse gamma distribution we get the Jeffreys prior[2] for scale parameters.

Remark 8.1. The BVS model eq. (8.8) together with the choice of Bernoulli priors on γ and a normal prior $\mathrm{N}_p(\mathbf{0}, \mathbf{V}_\beta)$ for $\boldsymbol{\beta}$ can be seen a *spike-and-slab prior* prior for linear regression models, a mixture of a point mass at zero and a normal density (Geweke, 1996; Mitchell and Beauchamp, 1988). Write $\boldsymbol{\theta} = (\gamma_1\beta_1, \dots, \gamma_p\beta_p)^\top$, which is interpreted as the "model-specific" regression coefficients. Then, the prior on $\boldsymbol{\theta}$ is equivalently written

$$\boldsymbol{\theta}|\gamma \sim \begin{cases} \mathrm{N}_p(\mathbf{0}, \mathbf{V}_\beta) & \text{w.p. } p(\gamma) \\ 0 & \text{w.p. } 1 - p(\gamma). \end{cases}$$

A subtle fact of these spike-and-slab priors is that the posterior distribution for $\boldsymbol{\theta}$ will also be a combination of a point mass and a normal density (with appropriate posterior parameters). Looking at it from this perspective, regression coefficients are assigned zero values with positive probability, and it is this fact that allows covariates to be dropped from the model. As pointed out by Kuo and Mallick (1998), the form of the variable selection model allows the selection of important variables, while simultaneously shrinking the coefficients via prior information.

8.4 Gibbs sampling for the I-prior BVS model

The Bayesian variable selection model can be estimated using Gibbs sampling, as demonstrated originally by Kuo and Mallick (1998). In this section, we describe the Gibbs sampling procedure to obtain posterior samples of the parameters. For the I-prior specifically, the joint density of the responses and the priors is

$$p(\mathbf{y}, \gamma, \boldsymbol{\beta}, \alpha, \sigma^2, \kappa) = p(\mathbf{y}|\gamma, \boldsymbol{\beta}, \alpha, \sigma^2)p(\boldsymbol{\beta}|\sigma^2, \kappa)p(\alpha|\sigma^2)p(\gamma)p(\sigma^2)p(\kappa),$$

[2] The Jeffreys prior for a parameter θ is defined as $p(\theta) \propto |\mathcal{I}(\theta)|^{1/2}$ (Jeffreys, 1946).

where the distribution of the model $p(\mathbf{y}|\gamma,\boldsymbol{\beta},\alpha,\sigma^2)$ and of the priors have been described in the previous section (except for κ, which we now assign an inverse gamma distribution). Let us denote $\Theta = \{\alpha,\boldsymbol{\beta},\gamma,\sigma^2,\kappa\}$ to be the full set of parameters that we wish to obtain posterior samples for. Starting with suitable initial values $\Theta^{(0)}$, we then proceed to obtain samples $\Theta^{(1)},\dots,\Theta^{(T)}$ by sampling each parameter from the conditional posterior density of that parameter given the rest of the parameters. A suggested set of initial values are the maximum likelihood (ML) estimates of Θ or the posterior mean estimate of Θ under the full model $\gamma = (1,\dots,1)$ after an initial MCMC run.

The Gibbs conditional densities are straightforward to obtain on account of model conjugacy (details of the derivation are given in **??**). We start with $\boldsymbol{\beta}$. The conditional density of $\boldsymbol{\beta}$ given $\alpha,\gamma,\sigma^2,\kappa$ is multivariate normal with mean $\tilde{\mathbf{B}}(\mathbf{y}-\alpha\mathbf{1}_n)$ and covariance matrix $\sigma^2\tilde{\mathbf{B}}$, where $\tilde{\mathbf{B}} = \mathbf{X}_\gamma^\top\mathbf{X}_\gamma + (\kappa\mathbf{X}^\top\mathbf{X})^{-1}$, and $\mathbf{X}_\gamma = (\gamma_1 X_1 \cdots \gamma_p X_p)$. Interestingly, when X_j is dropped from the model ($\gamma_j = 0$), the posterior mean and variance for j'th component of $\boldsymbol{\beta}$ is entirely informed by the prior (Kuo and Mallick, 1998). The data-driven I-prior incorporates information from the covariates into the prior, which then informs the posterior. In a similar manner, the conditional density for the intercept α is found to be $\mathrm{N}\left(\sum_{i=1}^n (y_i - \mathbf{x}_i^\top\boldsymbol{\theta})/\tilde{A}, \sigma^2\tilde{A}\right)$, where $\tilde{A} = n + A^{-1}$ and A is the prior variance for α.

The (conditional) posterior samples of $\gamma = (\gamma_1,\dots,\gamma_p)$ are obtained componentwise, and each conditional probability mass function for γ_j is Bernoulli with success probability $\tilde{\pi}_j = u_j/(u_j+v_j)$, where

$$u_j = \pi_j \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{y}-\alpha\mathbf{1}_n - \mathbf{X}\boldsymbol{\theta}_j^{[1]}\|^2\right)$$

and

$$v_j = (1-\pi_j) \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{y}-\alpha\mathbf{1}_n - \mathbf{X}\boldsymbol{\theta}_j^{[0]}\|^2\right).$$

Here, we have used the notation $\boldsymbol{\theta}_j^{[1]}$ to indicate the vector $\boldsymbol{\theta}$ with the j'th component replaced by $\boldsymbol{\beta}_j$, and $\boldsymbol{\theta}_j^{[0]}$ to indicate the vector $\boldsymbol{\theta}$ with the j'th component replaced by 0. Values of 1 for γ are more likely to be sampled when the ratio u_j/v_j is greater than the prior odds $\pi_j/(1-\pi_j)$. Specifically when the prior probabilities π_j are all set to be 0.5, then γ_j will be more likely to be sampled as '1' if $u_j > v_j$, i.e. if the residual sum of squares (RSS) $\|\mathbf{y}-\alpha\mathbf{1}_n-\mathbf{X}\boldsymbol{\theta}\|^2$ is *smaller* when the j'th component of $\boldsymbol{\theta}$ is non-zero, compared to the RSS when the j'th component of $\boldsymbol{\theta}$ is zero.

We can in fact draw parallels to a Bayesian hypothesis test, with the null hypothesis being $\mathrm{H}_0 : \beta_j = 0$ and the alternative being $\mathrm{H}_1 : \beta_j \neq 0$, conditional on knowing all other values of the parameters. Under H_k, $\mathbf{y}|\Theta \sim \mathrm{N}_n(\alpha\mathbf{1}_n + \mathbf{X}\boldsymbol{\theta}_j^{[k]}, \sigma^2\mathbf{I}_n)$, $k = 0,1$. The conditional Bayes factor comparing the model in the alternative hypothesis M_1 to the model in the null hypothesis M_0 is therefore

$$\mathrm{BF}(M_1,M_0) = \frac{u_j/\pi_j}{v_j/(1-\pi_j)} = \frac{\tilde{\pi}_j}{1-\tilde{\pi}_j} \bigg/ \frac{\pi_j}{1-\pi_j}.$$

Thus, it can be seen that the decision to include or exclude the j'th variable from the model relates a hypothesis test using the Bayes factor rule, and this decision is embedded in the conditional posterior probabilities $\tilde{\pi}_j$. The Gibbs sampling procedure does something that can be described as “an automated stochastic F-test for subset selection” (Kuo and Mallick, 1998).

Both scale parameters σ^2 and κ follow the conditional inverse gamma distributions

$$\sigma^2|\alpha,\beta,\gamma,\kappa \sim \Gamma^{-1}\big(n/2 + c_\sigma + 1, \|\mathbf{y} - \alpha\mathbf{1}_n - \mathbf{X}\boldsymbol{\theta}\|^2/2 + d_\sigma\big)$$

and

$$\kappa|\alpha,\beta,\gamma,\sigma^2 \sim \Gamma^{-1}\big(p/2 + c_\kappa + 1, \boldsymbol{\beta}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\boldsymbol{\beta}/\sigma^2 + d_\kappa\big).$$

Note that the inverse gamma distribution that we specify here is defined by its shape (s) and scale parameter (σ), and has the density function $p(x) = \Gamma(s)^{-1}\sigma^s x^{-(s+1)}e^{-\sigma/x}$. Here, $\{c_\sigma, d_\sigma\}$ and $\{c_\kappa, d_\kappa\}$ are the shape and scale hyperparameters of the inverse gamma priors on σ^2 and κ respectively.

8.5 Posterior inferences

Having obtained posterior samples $\Theta^{(t)} = \{\alpha^{(t)}, \boldsymbol{\beta}^{(t)}, \gamma^{(t)}, \sigma^{2(t)}, \kappa^{(t)}\}$, there are two quantities of interest in relation to model inferences. The first is an estimate of posterior model probabilities, given by

$$\widehat{\mathrm{P}}(\gamma = \gamma'|\mathbf{y}) = \frac{1}{T}\sum_{i=1}^{T}[\gamma^{(t)} = \gamma'], \tag{8.10}$$

where $[\cdot]$ is the Iverson bracket. This gives an estimate of the probability of a model coded by γ' appearing in the posterior state space of models. The second is a quantification of the posterior inclusion for each of the p variables $X_1, \ldots, X_p$, known as *posterior inclusion probabilities* (PIPs) for a variable being selected in any model. This is given by

$$\widehat{\mathrm{P}}(\gamma_j = 1|\mathbf{y}) = \frac{1}{T}\sum_{i=1}^{T}[\gamma_j^{(t)} = 1], \qquad j = 1, \ldots, p. \tag{8.11}$$

Posterior inclusion probabilities are the marginals of the posterior model probabilities across each variable.

Table 8.1: Illustration of samples of γ from the Gibbs sampler for $p = 3$. As an example, to estimate the posterior model probability of $\{X_1,X_3\}$, we count the occurrences of the combination $\gamma^{(t)} = (1, 0, 1)$ in the sample and divide by T. To estimate posterior inclusion probabilities for any of the three variables, we take the sample mean of the binary variates column-wise.

t	$\gamma_1^{(t)}$	$\gamma_2^{(t)}$	$\gamma_3^{(t)}$
1	1	0	1
2	1	0	0
3	1	1	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$
T	1	0	1

Note, that the regression coefficient of interest is not $\boldsymbol{\beta}$, but rather the “model averaged” regression coefficients $\boldsymbol{\theta} = (\gamma_1\beta_1, \dots, \gamma_p\beta_p)^\top$ (Madigan and Raftery, 1994). Posterior variances for $\boldsymbol{\theta}$ will typically be larger than variances for $\boldsymbol{\beta}$, because posterior estimates surrounding $\boldsymbol{\theta}$ will have incorporated model uncertainty, but $\boldsymbol{\beta}$ on the other hand, will not. Thus, any inferential procedure surrounding the regression coefficients avoids the risk of over-confidence. Note that, since $\boldsymbol{\theta}$ will contain values of exactly zero when predictors are dropped out of the model, the posterior density for $\boldsymbol{\theta}$ is a mixture of a point mass at zero and a normal density. In any case, the likelihood only provides sufficient information to identify the product of $\boldsymbol{\beta}$ and γ, but not each of them separately (Kuo and Mallick, 1998).

Remark 8.2. The intention of computing model-averaged regression coefficients $\boldsymbol{\theta}$ is solely for the inclusion of model uncertainty. There is a strong agreement in the Bayesian variable selection literature that that such coefficients are practically meaningless when it comes to explanatory inferences. Banner and Higgs (2017) writes that "regression coefficients... may not hold equivalent interpretations across all of the models in which they appear", and one reason for this might be "interpretation of partial regression coefficients can depend on other variables that have been included in the model". The use of model-averaged effect sizes may thus result in misleading inferences (Cade, 2015).

Finally, any quantity of interest Δ can be incorporated as part of the Gibbs sampling procedure. That is, at each Gibbs iteration $t = 1, \dots, T$, calculate $\Delta^{(t)}$ as a function of the parameter values at iteration t. This can be done during the Gibbs sampling process, or even after the fact as part of a post-processing procedure. Any inference on the posterior of Δ will then have incorporated the model uncertainty from a model averaging standpoint, as discussed earlier. As an example, suppose we are interested in the predicted value at a new covariate value $\mathbf{x}_{\text{new}} \in \mathbb{R}^p$. For each Gibbs sample, calculate

$$y_{\text{new}}^{(t)} = \alpha^{(t)} + \mathbf{x}_{\text{new}}^\top(\gamma_1\beta_1, \dots, \gamma_p\beta_p),$$

and obtain a point estimate $\hat{y}_{\text{new}}$ using the posterior mean or mode. The uncertainty for this estimate is contained in the standard deviation calculated from the sample $y_{\text{new}}^{(1)}, \dots, y_{\text{new}}^{(T)}$, from which a 95% credibility interval for this estimate can be obtained from the empirical upper and lower 0.025 cut off points.

8.6 Two stage procedure

The variable selection procedure can be improved by a "preselection" of variables to trim off unimportant variables which reduces the size of the model space being explored. Without appealing to other external preselection methods, there is actually information that we could use from Bayesian variable selection models in the form of posterior inclusion probabilities. The procedure would work as follows:

1. Run the Bayesian variable selection model and obtain posterior inclusion probabilities for each variable.
2. Discard variables with inclusion probablities less than a certain treshold, τ.
3. Re-run the Bayesian variable selection model on the set of reduced variables.

A natural choice for τ would be 0.5, and therefore a two-stage approach to Bayesian variable selection can then be motivated as selecting the subset of variables which constitutes what is known as the *median probability model*. The median probability model is obtained by selecting all variables with a posterior inclusion probability of greater than or equal to a half. Barbieri and Berger (2004) show that the median probability model has the property of being optimally predictive (minimises squared error loss for predictions) under certain strict conditions.

The notion of a two-stage approach is not new, as many variable selection methods in the literature generally employ a preselection method of some kind before running their selection process proper. This can be based on subjective preconceptions about which variables to retain, substantive theory, or even an objective preselection criterion. Two-stage procedures for Bayesian variable selection models have been used in works by Fouskakis and Draper (2008) and Ntzoufras (2011).

In the simulation studies conducted and observations from real-data examples, this two-stage approach does seem to provide a benefit. The complexity of estimating all model probabilities grows exponentially with p, therefore reducing this benefits the model selection procedure because the search of the model space is less cluttered. Of course, this idea works if the "correct" variables are deleted when proceeding to the second stage. We posit that the p posterior inclusion probabilities are easier to estimate than the 2^p posterior model probabilities from the same amount of information coming from the MCMC samples. As a result, information summarised through the posterior inclusion probabilities are more precise than the posterior model probabilities.

8.7 Simulation study

In this section, we conduct a simulation study to compare the performance of different priors in the Bayesian variable selection framework described above. The priors on $\boldsymbol{\beta}$ that are compared are those mentioned in section 8.3, i.e. the I-prior, the independent prior with large prior variance (flat/uninformative prior), and the g-prior with $g = n$ (unit information prior, Ntzoufras, 2011). We also make a comparison the variable selection performance of the Lasso, which, from a Bayesian perspective, is similar to setting a double-exponential or Laplace priors on the regression coefficients (Park and Casella, 2008). For clarity, the Lasso model employed in the simulations is of a frequentist regularisation framework as per Tibshirani (1996), and is neither a Bayesian variable selection model as described earlier, nor a fully Bayes implementation as per Park and Casella (2008). We felt it interesting to compare the Lasso as it is widely used for variable selection of linear models.

The experiment is to select from a total of $p = 100$ variables of an artificial dataset of sample size $n = 150$, which has pairwise correlations induced between the variables. This was inspired by the studies done by George and McCulloch (1993) and Kuo and Mallick (1998) in their respective papers, albeit on a larger scale (in theirs, $p = 30$). Five different scenarios were looked at. For each scenario, only s out of 100 variables were selected to form the "true" model and generate the responses according to the linear model $\mathbf{y} \sim \mathrm{N}_{100}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_{150})$. The signal-to-noise ratio (SNR) as a percentage is defined as s%, and the five scenarios are made up of varying SNR from high to low: 90%, 75%, 50%, 25%, and 10%. Variables that were included in the model had true $\boldsymbol{\beta}$ coefficients equal to one. That is, $\boldsymbol{\beta}_{\text{true}} = (\mathbf{1}_s, \mathbf{0}_{100-s})^\top$, where $\mathbf{1}_s$ is a row-vector of s ones, and $\mathbf{0}_{100-s}$ is a row-vector of $100 - s$ zeroes.

The data generation process is summarised as follows:

- Draw $\mathbf{Z}_1, \dots, \mathbf{Z}_{100} \overset{\text{iid}}{\sim} \text{N}_{150}(\mathbf{0}, \mathbf{I}_{150})$.
- Draw $\mathbf{U} \sim \text{N}_{150}(\mathbf{0}, \mathbf{I}_{150})$.
- Set $\mathbf{X} = (\mathbf{Z}_1 + \mathbf{U}, \dots, \mathbf{Z}_{100} + \mathbf{U})$. This induces pairwise correlations of about $1/2$ between the columns of $\mathbf{X}$.[3]
- Draw $\mathbf{y} \sim \text{N}_{150}(\mathbf{X}\boldsymbol{\beta}_{\text{true}}, \sigma^2\mathbf{I}_{150})$, with $\sigma = 2$.

In each scenario, we are interested in obtaining the highest probability model and counting the number of false choices made in this model after a two-stage procedure of variable selection. False choices can either be selecting variables wrongly (false inclusion) or failing to select a variable (false exclusion). Each scenario was repeated a total of 100 times to account for variability in the data generation process, and the results averaged. A summary of the results is presented in table 8.2. The overall results are also plotted in the form a frequency polygon (see fig. 8.1).

Table 8.2: Simulation results (proportion of false choices) for the Bayesian variable selection experiment using the I-prior, an independent prior, the g-prior and the Lasso across varying SNR. Standard errors are given in parentheses.

	Signal-to-noise ratio				
False choices	90%	75%	50%	25%	10%
I-prior					
0-2	**0.93** (0.03)	**0.92** (0.03)	**0.90** (0.03)	**0.79** (0.04)	**0.55** (0.05)
3-5	0.07 (0.03)	0.07 (0.03)	0.10 (0.03)	0.20 (0.04)	0.27 (0.04)
>5	0.00 (0.00)	0.01 (0.01)	0.00 (0.00)	0.01 (0.01)	0.18 (0.04)
Ind. prior					
0-2	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	**0.44** (0.05)	**1.00** (0.00)
3-5	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.30 (0.05)	0.00 (0.00)
>5	**1.00** (0.00)	**1.00** (0.00)	**1.00** (0.00)	0.26 (0.04)	0.00 (0.00)
g-prior					
0-2	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	**0.78** (0.04)	**0.86** (0.03)
3-5	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.14 (0.03)	0.13 (0.03)
>5	**1.00** (0.00)	**1.00** (0.00)	**1.00** (0.00)	0.08 (0.03)	0.01 (0.01)
Lasso					
0-2	0.03 (0.02)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
3-5	0.19 (0.04)	0.02 (0.01)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
>5	**0.78** (0.04)	**0.98** (0.01)	**1.00** (0.00)	**1.00** (0.00)	**1.00** (0.00)

The simulation results seem to indicate that the I-prior performs consistently well across all five scenarios, making no more than five false choices out of 100 (i.e. a 95% correct selection rate) in at least 82% of the time in the worst scenario. We do not observe much difference between the g-prior and the independent prior, and while they behave poorly in high SNR scenarios, these two

[3] For any row of $\mathbf{X}$, $\text{Cov}[X_j, X_k] = \text{Cov}[Z_j + U, Z_k + U] = \text{Var}\ [U] = 1$, and $\text{Var}[X_j] = \text{Var}[Z_j + U] = 2$. Thus, $\text{Corr}[X_j, X_k] = \text{Cov}[X_j, X_k]/(\text{Var}[X_j]\text{Var}[X_k])^{1/2} = 1/2$.

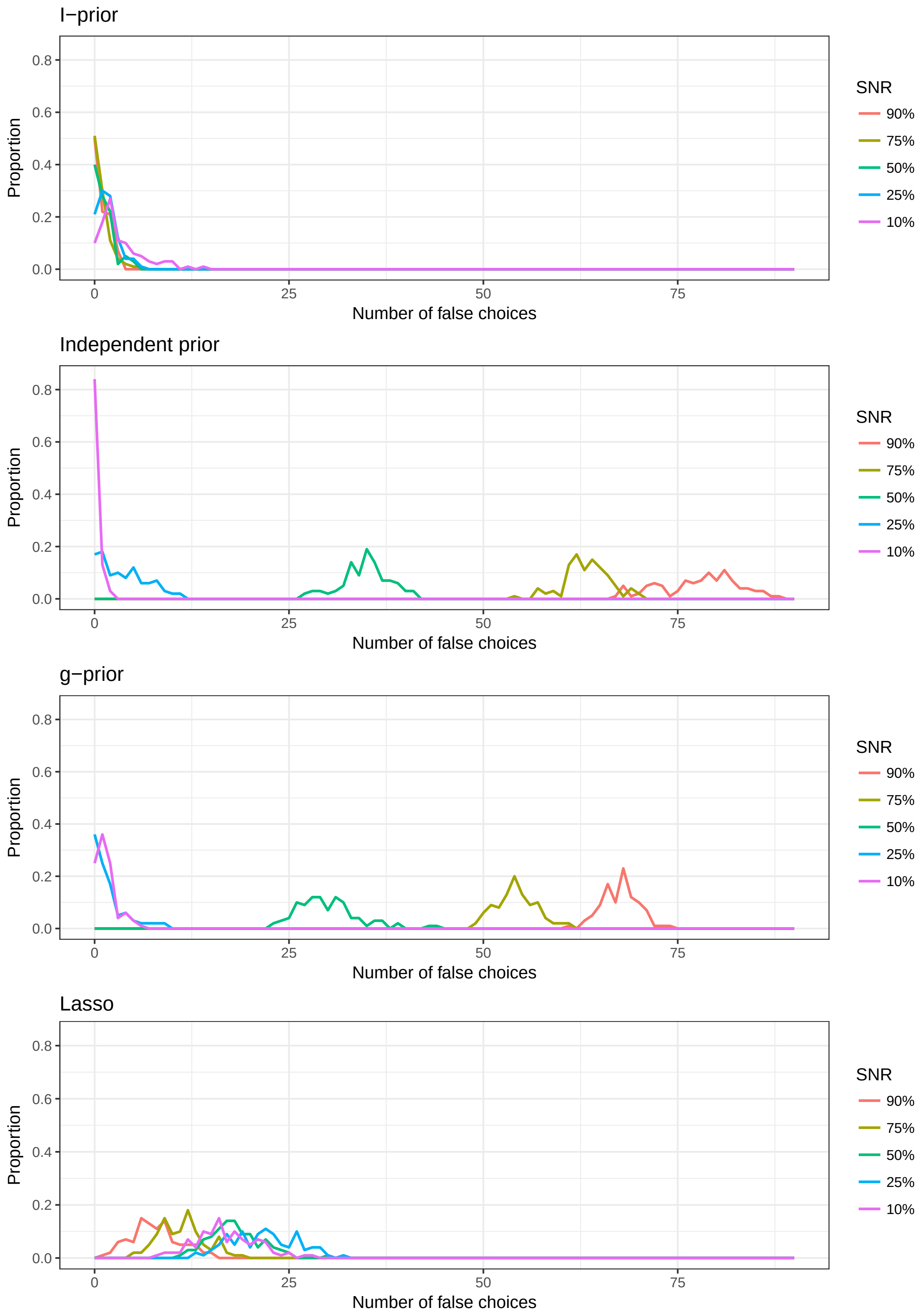

Figure 8.1: Frequency polygons for the number of false choices for each of the four priors. The I-prior performs robustly well across the five scenarios tested, mostly yielding five or fewer false inclusions or exclusions. Spurious exclusions led to the independent and g-prior simultaneously performing well in low SNR and badly in high SNR scenarios. The Lasso is known to be unreliable in the presence of collinearity.

priors seem to perform extremely well in low SNR scenarios. A high propensity to drop variables in these scenarios is a likely explanation, which does not necessarily indicate good performance—they perform well by contentiously omitting of a large number of unnecessary variables, especially in a two-stage procedure. Finally, the Lasso is well known to yield poor selection performance under multicollinearity, so the results are expected. The Lasso procedure was not subject to a two-stage approach because the Lasso does not provide information regarding posterior inclusion probabilities for individual variables.

We also inspect the sensitivity of the hyperprior choice of π_j for the indicator variables on the number of false choices made. fig. 8.2 plots the mean number of false choices made in each of the five SNR scenarios with varying hyperprior setting for π_j. From the plot, it is seen that for high SNR scenarios, setting π_j too low increases the number of false exclusions. Conversely, for low SNR scenarios, setting π_j too high increases the number of false inclusions. This makes sense: when the true model size is small, then setting π_j too high encourages variables to be retained in the model, and vice versa. While the optimal π_j corresponds directly to the true SNR (e.g. SNR $=$ 10% performs best under $\pi_j = 0.10$), fig. 8.2 makes a case for $\pi_j = 0.5$ to be a "safe choice" in the face of prior ignorance on model size.

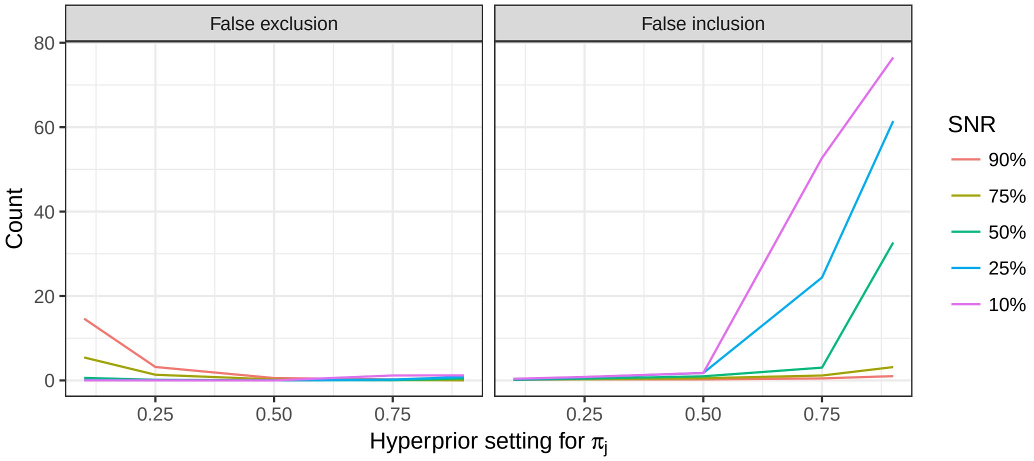

Figure 8.2: Average number of false choices (false inclusions or false exlusions) for the five different scenarios (SNR varied between 90%, 75%, 50%, 25% and 10%) with different hyperprior settings for $\gamma_j \sim \text{Bern}(\pi_j)$.

8.8 Examples

Now, we apply our I-prior Bayesian variable selection model to three real-world data sets that have all been previously analysed in the variable selection literature. All examples were analysed in R using our **ipriorBVS** package (Jamil, 2018a) which contains a wrapper to JAGS (Plummer, 2003). Reproducible code is available at `http://myphdcode.haziqj.ml`. In all analyses, a two-stage procedure was conducted for the I-prior model, where each stage consists of obtaining 10,000 MCMC samples.

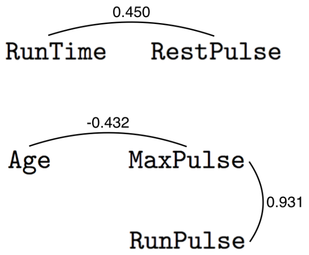

Figure 8.3: The sample correlations of interest in the aerobic fitness dataset. These show variables with correlations greater than 0.4 in magnitude.

8.8.1 Aerobic data set

This dataset appeared in the *SAS/STAT® User's Guide* (SAS Institute Inc., 2008) and was also analysed by Kuo and Mallick (1998). It involves understanding the factors which affect aerobic fitness, which is measured by the ability to consume oxygen. A sample of $n = 30$ male participants' had their physical fitness measured by means of simple exercise tests. The response variable contains measurement of oxygen uptake rate in mL/kg body weight per minute. The six covariates were the participants' age (X_1), weight (X_2), time taken to run one mile (X_3), resting heart rate (X_4), heart rate while running (X_5), and maximum heart rate during the exercise (X_6). This dataset, although small in size, is interesting to analyse because of the correlations between the variables, mainly due to the measurements being taken during the same exercise test. The sample correlations of interest are shown in fig. 8.3.

Table 8.3: Results for variable selection of the Aerobic data set. Note that the Bayes factors reported are the Bayes factors comparing any of the models to Model 1 (base model).

	PIP	Coef. (SD)	Model 1	Model 2	Model 3	Model 4
X_1	0.685	-0.169 (0.14)	✓		✓	
X_2	0.205	-0.017 (0.05)				
X_3	1.000	-0.745 (0.12)	✓	✓	✓	✓
X_4	0.168	-0.013 (0.05)				
X_5	0.663	-0.163 (0.15)	✓			✓
X_6	0.275	0.003 (0.10)				
		PMP	0.564	0.235	0.105	0.096
		BF	1.000	0.418	0.187	0.170

Notice that table 8.3 reports only on four out of $2^6 = 64$ possible models, but the sum of the posterior model probabilities add to one. Naturally, models which are deemed important by virtue of data evidence are sampled more often, and in fact, models which are unpromising may not even get sampled. So, MCMC methods does not need to list out all possible models because models which are never visited in the posterior state space are assigned a probability of zero. The highest posterior model was found to be the model with the variables X_1, X_3 and X_5 (PMP = 0.564). In fig. 8.4, we can see that the point mass at zero overwhelms the rest of the values in the density plots for X_2, X_4 and X_6, and hence these variables were dropped.

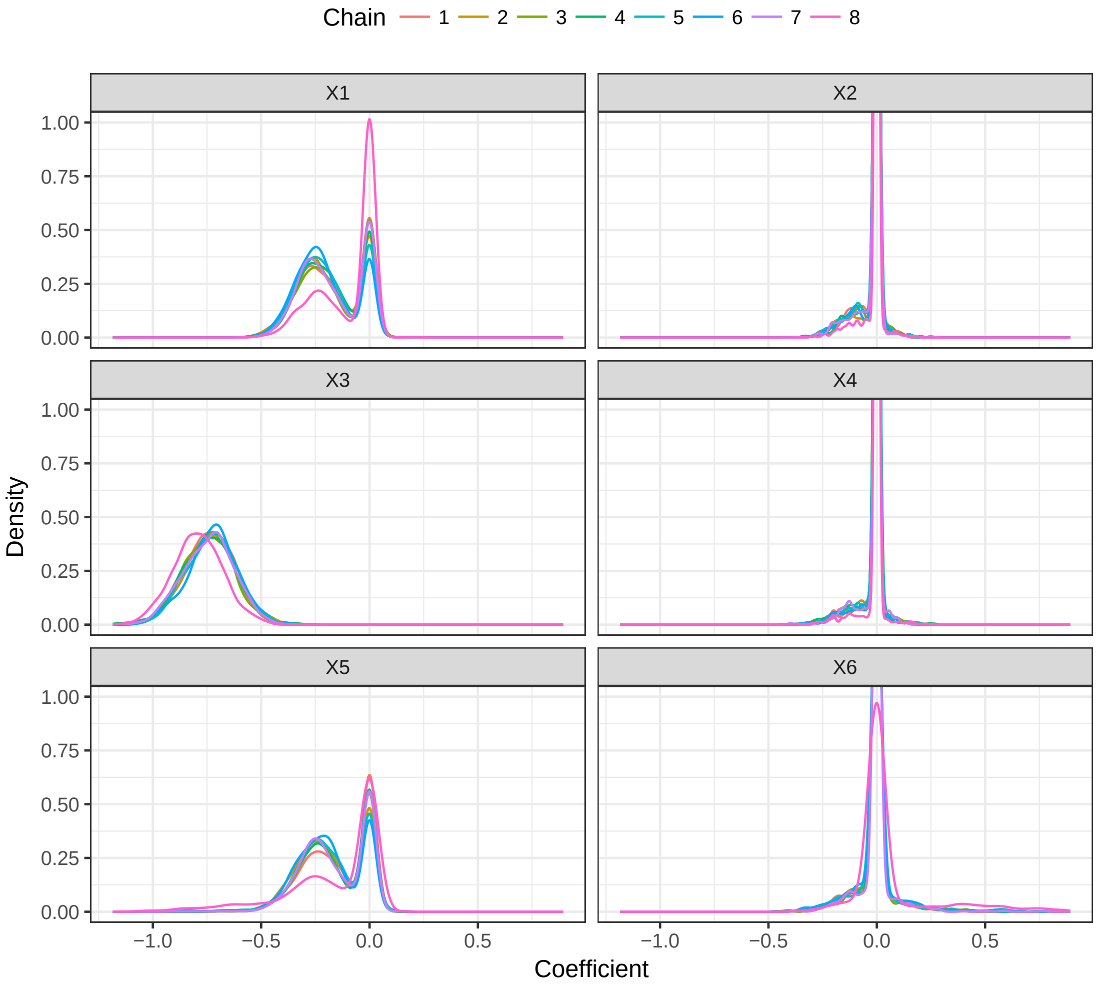

Figure 8.4: Posterior density plots of the regression coefficients $\boldsymbol{\theta}$ for the aerobic data set. The spike at zero observed in the density plots for X_2, X_4 and X_6 is indicative of these variable being dropped often in the posterior samples.

8.8.2 Mortality and air pollution data

The next real world application comes from a paper by McDonald and Schwing (1973). In it, the effects of air pollution on mortality in a US metropolitan area ($n = 60$ and $p = 15$) were studied. The response variable is the total age adjusted mortality rate, and the main pollution effects of interest were that of hydrocarbons (HC), oxides of nitrogen (NO_x) and sulphur dioxide (SO_2). Several other environmental and socioeconomic considerations were taken into account, otherwise the model may include unexplained variation caused by factors other than pollution. For example, a metropolitan area with a high proportion of the elderly should expect to have a higher mortality rate than one with a low proportion. All of the variables can be considered as continuous and real; table 8.4 provides a description of the variables.

Table 8.4: Description of the air pollution data set.

Variable	Description
Mortality	Total age adjusted mortality rate
Precipitation	Mean annual precipitation (in)
Relative humidity	Percent relative humidity, annual average at 1 p.m.
January temperature	Mean January temperature (°F)
July temperature	Mean July temperature (°F)
Population density	Population per square mile in urbanised area
Household size	Population per household
Education	Median school years completed for those over 25
Sound housing units	Percentage of sound housing units (no defects)
Age >65 years	Percent of population that is 65 years of age or over
Non-white	Percent of urbanised area population that is non-white
White collar	Percent employment in white-collar urbanised occupations
Income <\$3,000	Percent of families with income under \$3,000
HC	Relative population potential of hydrocarbons
NO_x	Relative population potential of oxides of nitrogen
SO_2	Relative population potential of sulphur dioxide

This dataset also contains several highly correlated variables which impedes a meaningful regression analysis. When the full model is fitted using ordinary least squares, none of the pollutant effects were found to be significant. Clearly, a variable selection method was required. McDonald and Schwing (1973) used a ridge regression technique to determine which variables to select and eliminate "unstable" coefficients found from a trace analysis. In addition, the authors also looked at a variable elimination method based on total squared error via Mallow's C_p. The results are summarised in table 8.5.

In this case, the I-prior BVS model concurred with the overall finding of McDonald and Schwing (1973), in that SO_2 was found to be a significant contributing factor towards mortality rates, while the rest of the pollutants were not. the I-prior BVS model also obtained a model with the largest R^2 and the smallest size. We note that the effect size for SO_2 is slightly larger under an I-prior, but generally, the rest of the I-prior coefficients are similar in magnitude and sign to the coefficients of the other two models.

8.8.3 Ozone data set

In this section, we replicate the Bayesian variable selection analysis of the Ozone dataset done by Casella and Moreno (2006, abbr. C&M) which appeared initially in Breiman and Friedman (1985, abbr. B&F), and also show how Bayesian variable selection can help select important interaction terms. The data consists of daily ozone readings and various meteorological quantities, and the aim was to see which of these quantities contributed to the ozone concentration. The variables considered are explained in table 8.6.

The data contains 366 points, one for each day of the leap year 1976. There are 163 data points containing missing data on some of the predictors, so we did a complete case analysis on the remaining 203 samples. Out of these 203, we randomly set aside 25 to use for validation, thus the n used to train the model was $n = 178$. The training and test set were repeated multiple times and results averaged in order to make a comparison to the unknown training and test

Table 8.5: A comparison of the coefficient values obtained using ordinary least squares (full model), McDonald and Schwing's minimum C_p and ridge analysis, and the I-prior. Standard errors/posterior standard deviations are given in parentheses. Values shaded grey indicate OLS regression coefficients not significant at the 10% level.

	Full model	Min. C_p	Ridge	I-prior
Environmental factors				
Precipitation	0.306 (0.14)	0.247 (0.07)	0.230 (0.07)	0.254 (0.12)
Relative humidity	0.009 (0.10)			
January temperature	-0.318 (0.18)	-0.164 (0.06)	-0.172 (0.06)	-0.195 (0.11)
July temperature	-0.237 (0.15)	-0.073 (0.07)		
Demographic factors				
Population density	0.084 (0.09)		0.091 (0.06)	
Household size	-0.232 (0.15)			
Education	-0.233 (0.16)	-0.190 (0.06)	-0.171 (0.07)	-0.151 (0.12)
Sound housing units	-0.052 (0.15)			
Age >65 years	-0.213 (0.20)			
Non-white	0.640 (0.19)	0.481 (0.07)	0.462 (0.07)	0.517 (0.10)
White collar	-0.014 (0.12)			
Income <\$3,000	-0.009 (0.22)			
Pollution potential				
HC	-0.979 (0.72)			
$\mathrm{NO_x}$	0.983 (0.75)			
$\mathrm{SO_2}$	0.090 (0.15)	0.255 (0.06)	0.232 (0.06)	0.302 (0.09)
Size	15	6	6	5
R^2	0.764	0.541	0.553	0.676

set used in the other studies. Out-of-sample prediction RMSE were obtained, as well as the coefficient of determination R^2.

C&M removed the variables X_{11} and X_{12} before running their selection model, citing multicollinearity causing ill-conditioned design matrices. Upon inspection, there are indeed correlations among the variables as high as 0.93 for some of them, but not enough to cause rank deficiency in the design matrix and a degenerate $\mathbf{X}^\top\mathbf{X}$ matrix. The sample correlations $\widehat{\mathrm{Corr}}(X_7, X_{11}) = 0.91$ and $\widehat{\mathrm{Corr}}(X_{11}, X_{12}) = 0.93$ seemed to drive the decision to drop the two variables, and while it is a valid concern, we shall still conduct variable selection on the full set of 12 variables to see the performance of I-priors in the presence of multicollinearity in this real-world data set. On another note, the variables X_1, X_2 and X_3 were presumably intended to be categorical as in modelling seasonality in a time series data, but these were treated as continuous, as did C&M. The results are compared in table 8.7.

What we found was that the model selected using the I-prior does better in terms of R^2 as well as RMSE compared to the methods used by C&M and B&F. The average posterior model probability for $\{X_1, X_6, X_{11}\}$ as found by the I-prior was 0.722[4]. One thing to note is that the I-prior model selected the variable X_{11} instead of its highly correlated proxy X_7, which is what C&M selected. These two variables are temperature measurements at different locations

Table 8.6: Description of the ozone data set used in this analysis. The data is available from the R package **mlbench** (Leisch and Dimitriadou, 2010).

Variable	Description
y	Daily maximum one-hour-average ozone reading (ppm) at Upland, CA
X_1	Month: $1 = \text{January}, \dots, 12 = \text{December}$
X_2	Day of month: $1, 2, \dots$
X_3	Day of week: $1 = \text{Monday}, \dots, 7 = \text{Sunday}$
X_4	500-millibar pressure height (m) measured at Vandenberg Air Force Base
X_5	Wind speed (mph) at Los Angeles International Airport (LAX)
X_6	Humidity (%) at LAX
X_7	Temperature (°F) measured at Sandberg, CA
X_8	Inversion base height (feet) at LAX
X_9	Pressure gradient (mmHg) from LAX to Daggett, CA
X_{10}	Visibility (mi) measured at LAX
X_{11}	Temperature (°F) measured at El Monte, CA
X_{12}	Inversion base temperature (degrees Fahrenheit) at LAX

Table 8.7: Results for variable selection of the Ozone data set using only linear predictors.

Method	Variables	Size	R^2	RMSE
I-prior	X_1, X_6, X_{11}	3	0.708	0.554
Casella and Moreno (C&M)	X_6, X_7, X_8	3	0.686	0.992
Breiman and Friedman (B&F)	X_7, X_8, X_9, X_{10}	4	0.669	1.056

in California. As C&M excluded X_{11} from the model search it was of course never considered in their model selection process, and because we included it in ours, the variable selection model was able to consider both variables together and decide on the more appropriate one.

Interestingly, the distance as the crow flies between Sandberg, CA (location of temperature measurements for X_7) and Upland, CA (location of ozone readings) is roughly 121 km, but El Monte, CA (location of temperature measurements for X_{11}) is just 35 km away from Upland, CA. It stands to reason that X_{11} provides more geographical reliability than X_7. Unless there is a strong insistence on deleting variables beforehand, we might not know for sure whether the variable was rightfully removed from consideration, as this example seems to prove. Out of curiosity, running the variable selection model on the reduced variable space as C&M did, we arrive at the same results as theirs.

We then used the I-prior method to select between the squared terms and all level two interactions, in addition to all the variables, in an effort to improve model prediction. For 12 such variables, the number of variables to select becomes $12 + 12 + 12(12 - 1)/2 = 90$. By doing so, we were able to improve the model to give a slightly better predictive ability. The results are shown below in table 8.8. The I-prior again selected a model which was superior in terms of R^2

[4] Since the total model space used was different between our method, C&M and B&F, it does not make sense to compare posterior model probabilities which we obtained. C&M reported a model probability of 0.491 for their model, but this model was not selected at all using the I-prior.

[5] Map tiles by Stamen Design, under CC BY 3.0. Data by OpenStreetMap, under CC BY-SA 3.0. Created using the **ggmap** package (Kahle and Wickham, 2013) in R.

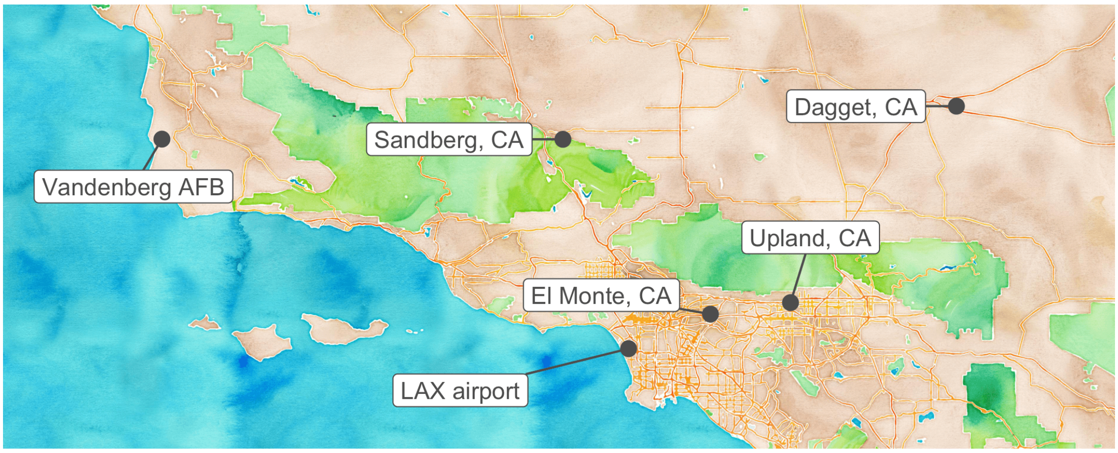

Figure 8.5: Locations[5] of the various points of interest in California, USA, related to the ozone measurements.

and RMSE compared to that obtained by C&M.

Table 8.8: Results for variable selection of the Ozone data set using linear, squared and two-way interaction terms.

Method	Variables	Size	R^2	RMSE
I-prior	$X_1, X_5, X_6, X_{11}, X_{12}, X_1^2, X_9^2, X_6X_{11}, X_6X_{12}, X_7X_9$	10	0.812	0.503
C&M	$X_2, X_1^2, X_7^2, X_9^2, X_1X_5, X_2X_6, X_3X_7, X_4X_6, X_6X_8, X_6X_{10}$	10	0.758	0.873

8.9 Conclusion

The model selection problem is an important one in statistics, but highly contentious. Miller (2002) writes that many statisticians view model selection as "unclean" or "distasteful", and that "terms such as 'fishing expeditions', 'torturing the data until they confess', 'data mining', and others are used as descriptions of these practices". The disagreement with the principle of model selection stems from the belief in the mantra that models should only be built by thoughtfully choosing variables which are expected to influence the response by appealing to substantive theory, and not by virtue of optimising some model selection criterion. However, variable selection as an exploratory study is certainly justified by many practical applications, especially when there is a genuine desire to know the most reasonable, parsimonious and interpretable model. Through variable selection exercises, we can learn which covariates are important, and which are negligible, in explaining the variation in the response.

The Bayesian variable selection method that we have seen has the appeal of reducing the problem of model search into one of estimation. At the outset, we aimed to seek a model which: 1) requires little tuning on the part of the user; 2) would work well in the presence of multicollinearity; and 3) is able to work well with little to no prior information. The I-prior on the regression coefficients in Kuo and Mallick's (1998) spike-and-slab stochastic search framework achieves this aim.

The attractive feature of a Bayesian approach to variable selection is the ability to simultaneously shrink and select predictors, thereby incorporating model uncertainty in the regressors. Sparsification is not "hard coded", in the sense that regression coefficients are assigned a value of zero with some positive probability in the posterior. This is unlike the regularisation or penalised log-likelihood approach to variable selection using the Lasso, elastic net, and so on, whereby sparsity is induced at the mode, but not in the posterior distribution (Scott and Varian, 2014). This translates to being provided with a single variable selection decision, rather than information that is coded through a probability distribution.

We discuss three areas to concentrate on for future research and improvement:

1. **$\boldsymbol{p > n}$ cases**. Typically, when there is insufficient information in the data to inform the estimation, then additional information is sought from the priors. In our case, the I-prior covariance involves the inverse of a low rank matrix which is not invertible. A p-variate normal distribution with a singular covariance matrix will only have a probability distribution defined on a low dimensional subspace. The issue may however be computational—it might be worth exploring the generalised inverse, or study ways in which to avoid the inverse computation in the Gibbs sampler. As a matter of fact, we note that the posterior precision for $\boldsymbol{\beta}$ can be written as

$$\begin{aligned}\tilde{\mathbf{B}}^{-1} &= \big(\mathbf{X}_\gamma^\top \mathbf{X}_\gamma + (\kappa \mathbf{X}^\top \mathbf{X})^{-1}\big)^{-1} \\ &= \mathbf{X}_\gamma^\top \mathbf{X}_\gamma \big((\mathbf{X}_\gamma^\top \mathbf{X}_\gamma)^2 + \kappa \mathbf{I}_p\big)^{-1}\end{aligned}$$

   which avoids the need for inverting the low-rank matrix $\mathbf{X}_\gamma^\top \mathbf{X}_\gamma$.

2. **Improvement in computational time**. Although the model itself is not computationally intensive to run (roughly $O(np^2)$ in time per Gibbs iteration), the main bottleneck is the reliance on a stochastic sampling algorithm. As in the previous chapter, variational inference is a promising area to look into, especially given that the Gibbs conditional distributions were straightforward to obtain, and these might be similar to a mean-field variational distribution. If this is successful, then it is expected to reduce computational time and avoid convergence issues that comes with traditional MCMCs. Variational inference with spike-and-slab priors on regression coefficients was studied by Ormerod et al. (2017).

3. **Extension to generalised linear models**. Kuo and Mallick (1998) in their paper already provided a sketch of how the variable selection model would work. In a similar fashion, we can extend the I-prior variable selection to categorical responses when the continuous latent propensities are modelled using linear functions. Such an approach has been implemented in gene selection studies, for which the variables are gene expressions and the responses are presence of a particular disease (Lee et al., 2003).

Finally, it should be mentioned that more complex variable selection models can be coded with the γ indicators. For instance, in selecting squared or interaction terms, we can insist on having the model select the main term if the squared or interaction term is selected, by specifying

$$y_i = \alpha + \max(\gamma_1, \gamma_3)\beta_1 x_{1i} + \max(\gamma_2, \gamma_3)\beta_2 x_{2i} + \gamma_3\beta_3 x_{1i} x_{2i}.$$

Or perhaps, we could use a single γ indicator for the dummy variables which make up a single categorical covariate, which we would then infer on the selection of the single covariate rather than each individual category of the covariate.