To-do list

Contents

Haziq Jamil
Department of Statistics
London School of Economics and Political Science
PhD thesis: 'Regression modelling using Fisher information covariance kernels (I-priors)'

Chapter 6

Bayesian variable selection using I-priors

chapter6

Earlier in Chapter 4, Section 4.1, we saw that model (1.1) subject to normal assumptions (1.2), model assumptions A1–A3, and f belonging to the canonical RKHS of functions over $\mathcal{X} \equiv \mathbb{R}^p$ yields the standard multilevel regression model

$$\begin{aligned} y_i &= \alpha + \sum_{k=1}^{p} x_{ik}\beta_k + \epsilon_i \\ \epsilon_i &\overset{\text{iid}}{\sim} \mathrm{N}_n(0, \sigma^2). \end{aligned} \tag{6.1}$$

{eq:linmod}

In this chapter, we use the notation $\sigma^2 = \psi^{-1}$ to denote the error variance. Furthermore, an I-prior on the regression coefficient entails prescribing the following normal prior the β_k's:

$$\boldsymbol{\beta} := (\beta_1, \ldots, \beta_p)^\top \sim \mathrm{N}(\mathbf{0}, \sigma^2 \kappa \mathbf{X}^\top \mathbf{X}).$$

This follows from (4.1) after a slight reparameterisation of the RKHS scale parameter $\kappa \mapsto \lambda^2/\sigma^4$. Throughout this chapter, we assume that the columns of the design matrix $\mathbf{X} = (X_1, \ldots, X_p)$ have been standardised, so that a single RKHS scale parameter is sufficient for the p covariates.

The topic of interest for this chapter is variable selection, or more generally, model selection. Model selection entails searching the entire model space to find the "best" model according to some criterion. There are many such criteria, from both frequentist and Bayesian perspectives, making model selection a huge topic to cover fully. These include

the (adjusted) R^2, Akaike's information criterion (AIC), Bayesian information criterion (BIC) and other similar information criteria, Mallow's C_p, (k-fold) cross-validation error, and so on. Methods such as forward or backward selection provide heuristic approaches to model selection, but never truly explore the entire model space.

On the other hand, regularised least squares regression (ridge regression, Lasso, elastic nets, etc.) provides additional information to the linear model in order to provide a sparse solution to linear system of equations in $\boldsymbol{\beta}$. These methods are proven to be popular and work well even in cases where $p > n$. From a Bayesian perspective, this is akin to placing priors on the β_k's to shrink the effects of the β_k's. Regularisation is fine when the ultimate goal is to 'cut down the number of variables' or to obtain a particular quantity of interest, such as prediction. However, often times there is a genuine desire to know the most reasonable, parsimonious and interpretable model—a goal which is achieved through model selection. Through variable selection, we learn which covariates are important, and which are negligible.

There is a massive literature concerning Bayesian variable selection (Chipman et al., 2001; O'Hara and Sillanpää, 2009). The approach that we take is a stochastic search of the model space due to George and McCulloch (1993) and Kuo and Mallick (1998) realised through a simple Gibbs sampling procedure. The plan for this chapter is to describe a fully Bayesian stochastic approach for variable selection using I-priors. 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.

new plan: variable selection via three methods: 1) model comparison via some criterion (frequentist and Bayes); 2) shrinkage to induce sparsity; 3) Bayesian model selection (probabilities on models).

problem: on linear regression models, choice of which subset of variables should be included in the model. explaining response variable with large number of explanatory variables. select a small subset, as close as possible to the true generative data process, which explains most of the variation in the response. belief of sparseness, that not all variables need to be included.

1) model selection criteria. methods described in (Miller, 2002). broadly three types: prediction criteria (R2, MSEP, Mallow's Cp, CV), likelihood/information criteria (AIC,BIC). requires comparison of all 2^p, so usually use stepwise procedures: forward-selection, backward-selection, etc. these are 'heuristic methods to restrict attention to a smaller number of potential subsets' (George and McCulloch, 1993). as a rough rule, possible if predictors are less than 25 (order 10^7).

2) shrinkage. induce sparseness by adding a penalty term to the likelihood in ML estimation (Ridge regression, Lasso, elastic net, etc.). idea is to add additional information to the system to regularise it. pros and cons?

Bayesian point of view is just to add priors. term adaptive shrinkage has been thrown around, to give suitable shape to the prior: if important, then flat, if useful then sharp peak. requires tuning, affects mixing in MCMC chain (O'Hara and Sillanpää, 2009). jeffreys' prior (no tuning), exponential prior (Bayesian Lasso (Park and Casella, 2008)).

3) bayesian model selection. probabilistic approach to model selection: put priors on model space, then estimate posterior model probabilities from data. slightly related to likelihood based criterion of model selection (post. model prob = BF calculations = marginal likelihoods), so if 2^p is large then difficult. however, the advantage of bayesian framework is the ability to apply MCMC as a practical means of overcoming the intractability. stochastic approach to model selection were pioneered by (George and McCulloch, 1993), and revisited by others (Kuo and Mallick, 1998; Dellaportas et al., 2002; Ntzoufras, 2011)

advantages: bayesian model averaging. useful when several competing models have high posterior probabilities.

criticisms: The topic of subset selection in regression is one that is viewed by many statisticians as 'unclean' or 'distasteful'. Terms such as 'fishing expeditions', 'torturing

the data until they confess', 'data mining', and others are used as descriptions of these practices. (Miller, 2002) in many studies variables are chosen because there is an expectation that they influence the response, and therefore regression model fitted is used to infer the strength of the influence. variable selection as an exploratory study. justified by many practical applications. arguably variable selection from a prescreening as part of data cleanup to remove problematic variables, such as those inducing a high degree of collinearity. how do we know we are deleting the correct variable?

In addition to model selection, any quantity of interest, say regression coefficients, predicted values, and any others, may be obtained by way of Bayesian model averaging. Instead of conditioning on a single selected model, a quantity of interest Δ is averaged over a set of models to avoid underestimation of the uncertainty surrounding it (**madigan1994model**). In essence, the posterior distribution of Δ may be found by averaging the model specific posterior distributions weighted by its posterior model probabilities

$$p(\Delta|y) = \sum_{k=1}^{K} p(\Delta|y, M_k)p(M_k|y). \tag{6.2}$$

{eq:bma}

Technically, the sum should be taken over the set of all possible models, but in practice this might not be achievable. Regardless, even if a smaller subset of models is used, it is well known that predictive accuracy of Δ is improved, as measured by a logarithmic scoring rule (**raftery1997bayesian**).

sec:bvs-prior

6.1 Motivation: model selection and a stochastic approach

The paradigm of model selection is as follows. From a finite set of models $\mathcal{M} = \{M_1, \ldots, M_K\}$, pairs of data $\{(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)\}$, $y_i \in \mathbb{R}$ 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).

Let P be a probability distribution over the model space $\mathcal{M}$, and denote the probability distribution function by $p(\cdot)$. Posterior model probabilities may then be evaluated as follows. For any model $M_k \in \mathcal{M}$, the posterior model probability for model m is

$$\begin{aligned} \mathrm{P}(M = M_k|\mathbf{y}) &\propto p(\mathbf{y}|M_k)p(M_k) \\ &\propto \int p(\mathbf{y}|M_k, \Theta_k)p(\Theta_k|M_k)\,\mathrm{d}\Theta_k\, p(M_k). \end{aligned} \tag{6.3}$$

eq:pmp

As a remark, the prior distributions for the parameters do not necessarily need to depend on the model, so we might have that $p(\Theta_k|M_k) = p(\Theta_k)$. Also, the normalising constant in (6.3) is of course the marginal distribution for $\mathbf{y}$, and is found by summing (6.3) over all possible models, i.e.

$$p(\mathbf{y}) = \sum_{k=1}^{K} \left\{ p(M_k) \int p(\mathbf{y}|M_k, \Theta_k)p(\Theta_k|M_k)\,\mathrm{d}\Theta_k \right\}. \tag{6.4}$$

eq:normconst

The integral found in (6.3) and (6.4) is known as *model evidence*. As the name suggests, it quantifies the amount of support for a particular model being the true model. It can also be viewed as the marginal likelihood under model M_k, so we denote it by $p(\mathbf{y}|M_k)$.

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 Bayes factor for comparing any model $M_k \in \mathcal{M}$ to a base model M_0 is given

by

$$\text{BF}(M_k, M_0) = \frac{p(\mathbf{y}|M_k)}{p(\mathbf{y}|M_0)}.$$

With a little manipulation, the posterior model probabilities can be calculated in terms of Bayes factors with respect to a base model. Also, the Bayes factor for comparing any two models $M_k, M_{k'} \in \mathcal{M}$ can also be calculated in terms of Bayes factors with respect to a base model:

$$\text{BF}(M_k, M_{k'}) = \frac{\text{BF}(M_k, M_0)}{\text{BF}(M_{k'}, M_0)}.$$

There are two potential issues with model comparison using posterior model probabilities or Bayes factors. The first is that the calculation of the model evidence may involve an integral which might be difficult to deal with. In the case of the linear model, this can be overcome by the use of conjugate priors for the regression model. As an aside, one of the main reasons why g-priors (Zellner, 1986) are very convenient to use for linear models is the fact that it completely simplifies the algebra in the posterior.

Even if all the relevant expressions can be obtained easily, there is the remaining issue that all of posterior probabilities must be calculated in order for a 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.

Markov chain Monte Carlo (MCMC) methods may be used to evaluate the posterior model probabilities. Naturally, models which are deemed important by virtue of data evidence are sampled more often in the posterior. In fact, models which are unpromising might never get sampled. MCMC methods might not list out all possible model probabilities, but it does not need to, because models which are never visited in the posterior state space are simply assigned probability zero. In the upcoming sections, we explain the Bayesian variable selection model that we use, and also the Gibbs sampling procedure employed to obtain the posterior model probabilities.

6.2 The Bayesian variable selection model

Considering variable selection as a special case of model selection, we shall loosely refer to a model as a subset of variables selected from the full set of variables $\{X_1, \ldots, X_p\}$.

Note that we do not consider the intercept to be selectable. If this were the case, this would imply a model as having intercept equal to zero as being possible. For most practical modelling purposes, the intercept is almost always non-zero.

It would be useful to be able to index each of these 2^p possible models somehow. Consider the one-to-one mapping $G : \mathcal{M} \to \{0, 1\}^p$ as defined by $M \mapsto \gamma = (\gamma_1, \ldots, \gamma_p)$, where $\gamma_j = 1$ if the variable X_j is selected, and 0 otherwise, for $j = 1, \ldots, p$. As an example, the full model, where all the variables are included in the model, is denoted by $\gamma = (1, \ldots, 1)$, while the intercept only model is denoted by $\gamma = (0, \ldots, 0)$.

Priors on the model can then be specified more concretely, for example

$$p(\gamma) = \prod_{j=1}^{p} \text{Bern}(\pi_j). \tag{6.5}$$

{eq:priorgamma}

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) (Jeffrey's prior), or any other suitable hyperprior. In any case, in this thesis we consider the simplest case of setting all $\pi_j = 0.5$.

Having considered the model prior, we now think about extending the linear model to include the indicator variables γ_j. Following Kuo and Mallick (1998), the linear model in (6.1) is expanded to

$$y_i | \alpha, \boldsymbol{\beta}, \gamma, \sigma^2 \overset{\text{iid}}{\sim} \text{N}\left(\alpha + \sum_{k=1}^{p} x_{ik} \gamma_k \beta_k \, , \, \sigma^2 \right). \tag{6.6}$$

{eq:km}

Hence, in addition to the usual model parameters, we are interested in conducting model inferences through the posterior distribution of the γ's. The Kuo and Mallick model is often known as the independent sampler due to the independence of model parameters and the indicator variables. Prior choices for the regression coefficients include an independent prior $\boldsymbol{\beta} \sim \text{N}_p(\mathbf{0}, c^2 \mathbf{I}_p)$, the g-prior $\boldsymbol{\beta} \sim \text{N}_p(\mathbf{0}, g\mathbf{X}^\top \mathbf{X}^{-1})$, or even the I-prior.

We complete the Bayesian variable selection model with conjugate prior choices on the remaining parameters—normal for α and inverse gamma for the scale parameters.

Choosing the I-prior for $\boldsymbol{\beta}$ in particular, the prior density on the model parameters is

$$\begin{aligned} p(\boldsymbol{\beta}, \alpha, \sigma^2, \kappa) &= p(\boldsymbol{\beta}|\sigma^2, \kappa)p(\alpha|\sigma^2)p(\sigma^2)p(\kappa) \\ &\equiv \mathrm{N}_p(\boldsymbol{\beta}|\mathbf{0}, \sigma^2\kappa\mathbf{X}^\top\mathbf{X}) \cdot \mathrm{N}(0, \sigma^2 A) \cdot \Gamma^{-1}(\sigma^2|c_\sigma, d_\sigma) \cdot \Gamma^{-1}(\kappa|c_\kappa, d_\kappa). \end{aligned} \tag{6.7}$$

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 $c_\sigma, d_\sigma, c_\kappa, d_\kappa \to 0$ then we get the Jeffrey's prior for scale parameters.

The BVS model (6.6) 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 for linear regression models; it is a combination of a point mass at zero and a normal prior (Mitchell and Beauchamp, 1988; Geweke, 1996). Write $\boldsymbol{\theta} = (\gamma_1\beta_1, \ldots, \gamma_p\beta_p)^\top$. Then, the prior on the model-specific regression coefficients $\boldsymbol{\theta}$ is

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

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 subtle 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.

As a remark, the regression coefficient of interest is not $\boldsymbol{\beta}$, but rather $\boldsymbol{\theta} = \boldsymbol{\gamma} \cdot \boldsymbol{\beta}$, which represents the "model averaged" regression coefficients. Posterior estimates surrounding $\boldsymbol{\theta}$ will have incorporated model uncertainty discussed earlier, but $\boldsymbol{\beta}$ on the other hand, will not. Posterior variances for $\boldsymbol{\theta}$ will typically be larger, but more "correct", than variances for $\boldsymbol{\beta}$.

6.3 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). The Gibbs sampling procedure about to be described is adapted from their work to include an I-prior on the regression coefficients, and estimation of an intercept and the scale hyperparameter κ for the I-prior.

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. Starting with suitable initial values $\Theta^{(0)}$, we then proceed to obtain further samples $\Theta^{(1)}, \ldots, \Theta^{(T)}$ 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 estimates of Θ, corresponding to the empirical Bayes estimate, under the full model $\gamma = (1, \ldots, 1)$.

Since the priors were chosen to be conjugate to the normal regression model, the Gibbs conditional densities are straightforward to obtain and familiar to those who work with Bayesian regression models. 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 $\tilde{B} = \mathbf{X}_\gamma^\top\mathbf{X}_\gamma + (\kappa\mathbf{X}^\top\mathbf{X})^{-1}$, 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 data into the prior which then informs the posterior, but this will not be the case if subjective priors or uninformative priors for β were used instead. In a similar manner, the conditional density for the intercept α is found to be $\mathrm{N}(\sum_{i=1}^n (y_i - \mathbf{x}_i^\top\boldsymbol{\theta})/\tilde{A}, \sigma^2\tilde{A})$, where $\tilde{A} = n + A^{-1}$ and A is the prior variance for α.

The (conditional) posterior samples of $\gamma = (\gamma_1, \ldots, \gamma_p)$ are obtained component-wise, 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 jth 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 jth component 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

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

Therefore 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 and scale parameter, and has the density function described in **??**. Here, c_σ, d_σ, c_κ and d_κ are the shape and scale hyperparameters of the inverse gamma priors on σ^2 and κ.

6.4 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

$$\widehat{\mathrm{P}}(\gamma = \gamma' | \mathbf{y}) = \frac{1}{T} \sum_{i=1}^{T} [\gamma^{(t)} = \gamma'], \tag{6.8}$$

{eq:pmp-prop}

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* for a variable being selected in any model. This is given by

$$\widehat{\mathrm{P}}(\gamma_j = 1 | \mathbf{y}) = \frac{1}{T} \sum_{i=1}^{T} [\gamma_j^{(t)} = 1], \qquad j = 1, \ldots, p. \tag{6.9}$$

{eq:pip-prop}

Posterior inclusion probabilities may also be thought of as the marginals of the posterior model probabilities across each variable.

As mentioned, posterior inferences for the regression coefficient should be done on samples for $\boldsymbol{\theta} = \gamma \cdot \boldsymbol{\beta}$ rather than $\boldsymbol{\beta}$ itself. This assures that model uncertainty is accounted for in any inferential procedure surrounding the regression coefficients. Note that, since $\boldsymbol{\theta}$ will contain values of exactly zero when predictors are dropped out of the model, the posterior density for $\boldsymbol{\theta}$ will consist of a point mass at zero combined with a normal density.

Finally, any quantity of interest Δ can be incorporated as part of the Gibbs sampling procedure. That is, at each Gibbs iteration $t = 1, \ldots, T$, 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

$$y_{\text{new}}^{(t)} = \alpha^{(t)} + \mathbf{x}_{\text{new}}^{\top} (\gamma^{(t)} \cdot \boldsymbol{\beta}^{(t)}),$$

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

6.5 Two stage procedure

The variable selection procedure can be made better by a "pre-selection" of variables to trim off unimportant variables which reduces the size of the model space being explored. Without appealing to other external pre-selection 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 under certain strict conditions.

The notion of a two-stage approaches are not new, as many variable selection methods in the literature generally employ a pre-selection 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 pre-selection criterion. Two-stage procedures for Bayesian variable selection models have been used in works by Fouskakis and Draper (2008) and Ntzoufras (2011).

6.6 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 beta that are compared are the I-prior, an 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 $p = 100$ variables of a $n = 150$ sample size artificial dataset which has pairwise correlations 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 $y \sim \mathrm{N}(X\beta, \sigma^2 I_{100})$. 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 β coefficients equal to one. That is, $\beta_{\text{true}} = (\mathbf{1}_s, \mathbf{0}_{100-s})^\top$, 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 $Z_1, \ldots, Z_{100} \overset{\text{iid}}{\sim} \mathrm{N}(0, 1)$.
- Draw $U \sim \mathrm{N}(0, 1)$.
- Set $X_j = Z_j + U$. This induces pairwise correlations of about 0.5.[1]
- Draw $\mathbf{y} \sim \mathrm{N}_{150}(\mathbf{X}\boldsymbol{\beta}_{\text{true}}, \sigma^2 \mathbf{I}_{150})$, 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 results are summarised in Table 6.1. The overall results are also plotted in the form a frequency polygon (see Figure 6.1).

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 priors seem to perform extremely well

[1] $\mathrm{Cov}(X_j, X_k) = \mathrm{Cov}(Z_j + U, Z_k + U) = \mathrm{Var}\, U = 1$, and $\mathrm{Var}(X_j) = \mathrm{Var}(Z_j + U) = 2$. Thus, $\mathrm{Corr}(X_j, X_k) = \mathrm{Cov}(X_j, X_k)/(\mathrm{Var}(X_j)\mathrm{Var}(X_k))^{1/2} = 1/2$.

Table 6.1: 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.

tab:simres

	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)
Independent 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)

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.

6.7 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. In all analyses, a two-stage procedure was conducted for the I-prior model, where each stage consists of obtaining 15,000 MCMC samples (including 5,000 for burn-in).

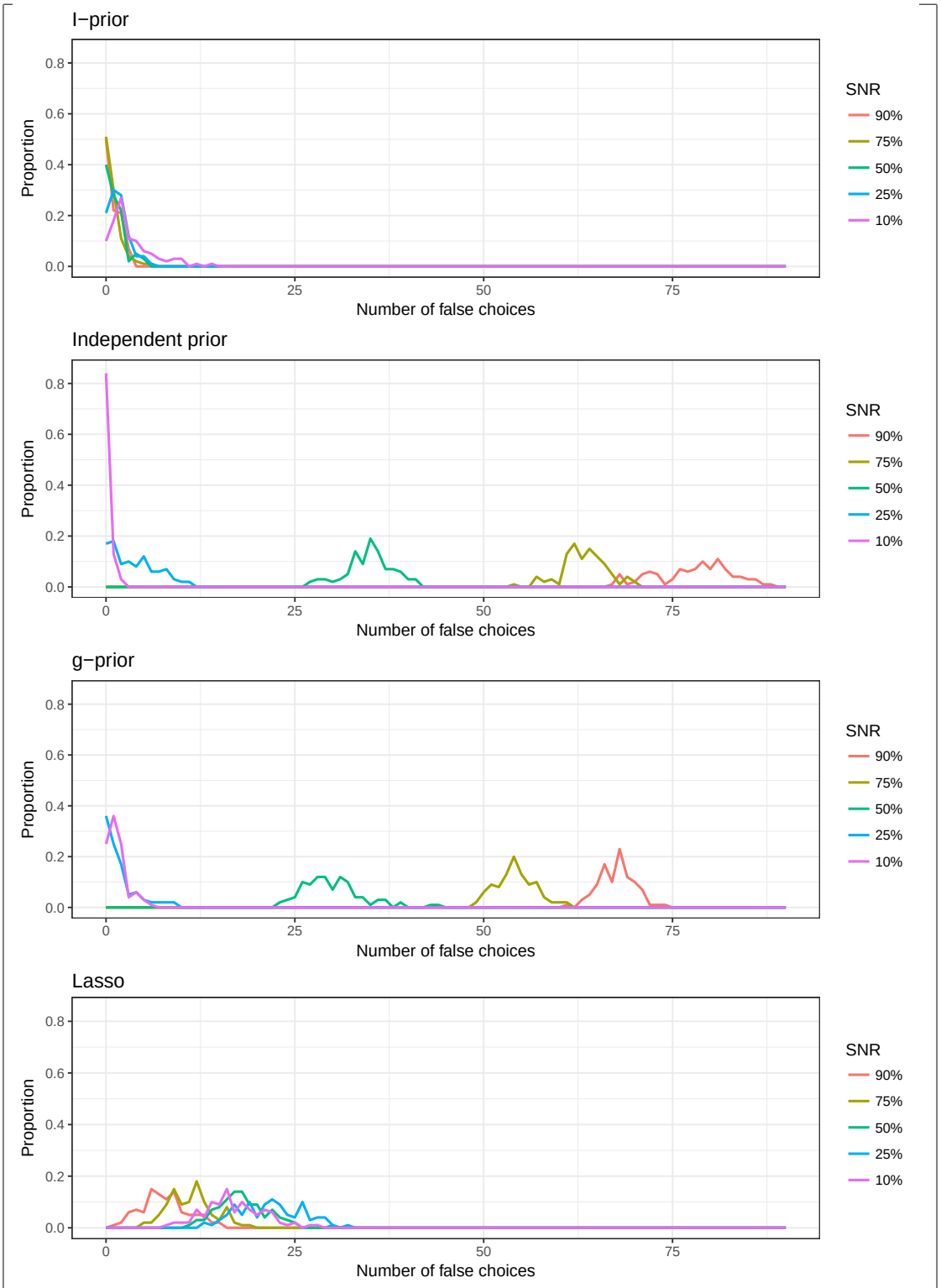

Figure 6.1: Frequency polygons for the number of false choices for each of the four priors.

fig:simres

6.7.1 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. Results are summarised in Table 6.2. The sample correlations of interest are shown in Figure 6.2 below:

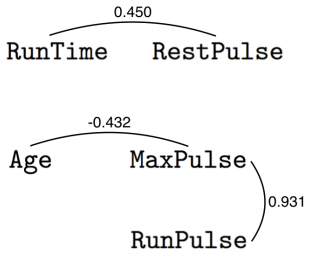

Figure 6.2: The sample correlations of interest in the aerobic fitness dataset. These show variables with correlations greater than 0.4 in magnitude.

fig:aerobic-cor

Table 6.2: Results for variable selection of the Aerobic data set.

tab:aerobic

	PIP	Model 1	Model 2	Model 3	Model 4
X_1	0.669	✓		✓	
X_2					
X_3	1.000	✓	✓	✓	✓
X_4					
X_5	0.659	✓			✓
X_6					
	PMP	0.564	0.235	0.105	0.096
	BF	1.000	0.418	0.187	0.170

The highest posterior model selected was the model with the variables X_1, X_3 and X_5. In Figure 6.3, 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.

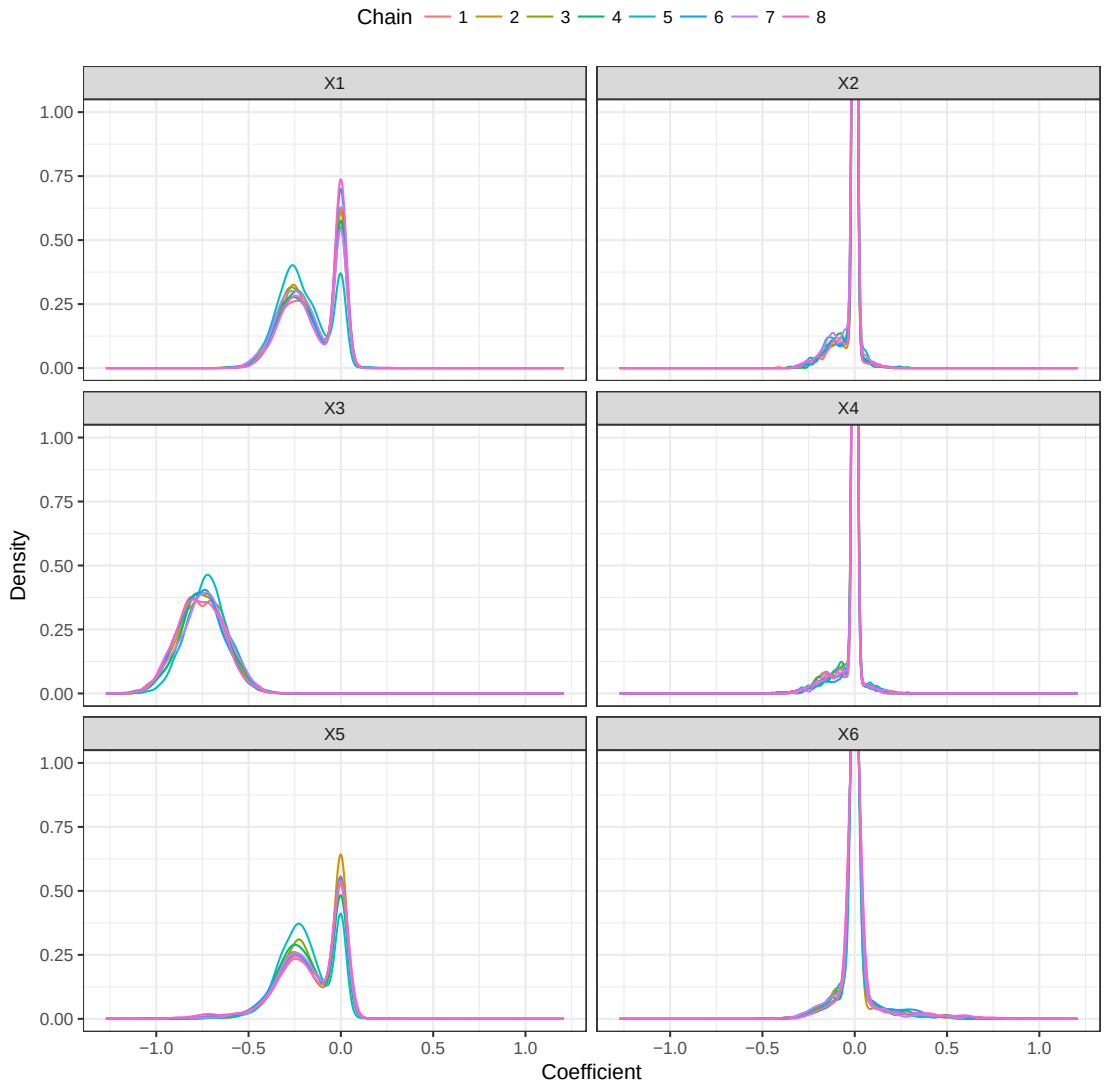

Figure 6.3: Posterior density plots of the regression coefficients $\boldsymbol{\theta}$ for the aerobic data set. The 'spike' at zero observed in the density plots for X_2, X_4 and X_6 is indicative of these variable being dropped often in the posterior samples.

fig:aerobic-densplot

sec:airpollution

6.7.2 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. A full description of the data can be found in **??**.

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 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 6.3.

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.

Table 6.3: 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 in the column of the full model indicate regression coefficients not significant at the 10% level.

tab:poll

	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)			
NO_x	0.983 (0.75)			
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

6.7.3 Ozone data set

sec:ozone

In this section, we replicate the Bayesian variable selection analysis of the Ozone dataset done by Casella and Moreno (2006) which appeared initially in Breiman and Friedman (1985), 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 are explained in **??**.

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 set used in the other studies. Out-of-sample prediction root mean squared errors (RMSE) were obtained, as well as the coefficient of determination R^2.

Casella and Moreno 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 correlations $\text{Corr}(X_7, X_{11}) = 0.91$ and $\text{Corr}(X_{11}, X_{12}) = 0.93$ seemed to drive the decision to drop the two variables, and while it is a valid concern, we will conduct variable selection on the full set of 12 variables. We can then 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 Casella and Moreno. The results are compared in Table 6.4.

Table 6.4: Results for variable selection of the Ozone data set using only linear predictors.

tab:ozoneres

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

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[2]. 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 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 method was able to consider both variables together and decide on the more appropriate one. 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 interest, 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 6.5. The I-prior again selected a model which was superior in terms of R^2 and RMSE compared to that obtained by C&M.

Table 6.5: Results for variable selection of the Ozone data set using linear, squared and two-way interaction terms.

tab:resozone2

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

6.8 Conclusion

The Bayesian variable selection methods that we have seen have 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

[2]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.

information. The I-prior on the regression coefficients in Kuo and Mallick's spike-and-slab stochastic search framework achieves this aim.

The attractive feature about the 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 variables 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.

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.

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 our case, it would also require careful consideration of the appropriate covariance matrix for the regressors, to keep in line with the definition of I-priors.

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:

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

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.

Bibliography

Barbieri2004

Barbieri, Maria Maddalena and James O Berger (2004). "Optimal predictive model selection". In: *Annals of Statistics* 32.3, pp. 870–897. ISSN: 00905364. DOI: 10.1214/009053604000000238. arXiv: 0406464v1 [arXiv:math].

Breiman1985

Breiman, L and J H Friedman (1985). "Estimating optimal transformations for multiple-regression and correlation". In: *Journal of the American Statistical Association* 80.391, pp. 614–619.

Casella2006

Casella, George and Elías Moreno (2006). "Objective Bayesian Variable Selection". In: *Journal of the American Statistical Association* 101.473, pp. 157–167. ISSN: 0162-1459. DOI: 10.1198/016214505000000646.

Chipman2001

Chipman, Hugh, Edward I George, and Robert E McCulloch (2001). "The Practical Implementation of Bayesian Model Selection". In: *IMS Lecture Notes - Monograph Series* 38, pp. 65–134.

dellaportas2002bayesian

Dellaportas, Petros, Jonathan J Forster, and Ioannis Ntzoufras (2002). "On Bayesian model and variable selection using MCMC". In: *Statistics and Computing* 12.1, pp. 27–36.

Fouskakis2008

Fouskakis, Dimitris and David Draper (2008). "Comparing Stochastic Optimization Methods for Variable Selection in Binary Outcome Prediction, With Application to Health Policy". In: *Journal of the American Statistical Association* 103.484, pp. 1367–1381. ISSN: 0162-1459. DOI: 10.1198/016214508000001048.

George1993

George, Edward I and Robert E McCulloch (1993). "Variable Selection Via Gibbs Sampling". In: *Journal of the American Statistical Association* 88.423, pp. 881–889. URL: http://www.jstor.org/stable/2290777?seq=1#page_scan_tab_contents.

geweke1996variable

Geweke, John (1996). "Variable selection and model comparison in regression". In: *In Bayesian Statistics 5*.

kass1995bayes
Kass, Robert E and Adrian E Raftery (1995). "Bayes factors". In: *Journal of the american statistical association* 90.430, pp. 773–795.

Kuo1998
Kuo, L and B Mallick (1998). "Variable selection for regression models". In: *Sankhya: The Indian Journal of Statistics, Series B* 60.1, pp. 65–81.

McDonald1973
McDonald, Gary C and Richard C Schwing (1973). "Instabilities of regression estimates relating air pollution to mortality". In: *Technometrics* 15.3, pp. 463–481.

miller2002subset
Miller, Alan (2002). *Subset selection in regression*. CRC Press.

mitchell1988bayesian
Mitchell, Toby J and John J Beauchamp (1988). "Bayesian variable selection in linear regression". In: *Journal of the American Statistical Association* 83.404, pp. 1023–1032.

Ntzoufras2008
Ntzoufras, Ioannis (2011). *Bayesian Modeling Using WinBUGS*. Wiley, pp. 389–433. DOI: 10.1002/9780470434567.ch11.

OHara2009
O'Hara, R B and M J Sillanpää (2009). "A review of Bayesian variable selection methods: what, how and which". In: *Bayesian Analysis* 4.1, pp. 85–117. ISSN: 1936-0975. DOI: 10.1214/09-BA403. URL: http://projecteuclid.org/euclid.ba/1340370391.

park2008bayesian
Park, Trevor and George Casella (2008). "The bayesian lasso". In: *Journal of the American Statistical Association* 103.482, pp. 681–686.

SAS2008
SAS Institute Inc. (2008). *SAS/STAT(R) 9.2 User's Guide*. 2nd. Cary, NC: SAS Institute Inc. ISBN: 978-1-60764-566-5. URL: https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_reg_sect055.htm.

scott2014predicting
Scott, Steven L and Hal R Varian (2014). "Predicting the present with bayesian structural time series". In: *International Journal of Mathematical Modelling and Numerical Optimisation* 5.1-2, pp. 4–23.

tibshirani1996regression
Tibshirani, Robert (1996). "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288.

zellner1986assessing
Zellner, Arnold (1986). "On assessing prior distributions and Bayesian regression analysis with g-prior distributions". In: *Bayesian inference and decision techniques*.