Bayesian Variable Selection Under Collinearity

Joyee Ghosh*  Andrew E. Ghattas†

June 3, 2014

Abstract

In this article we provide some guidelines to practitioners who use Bayesian variable selection for linear regression models in settings where the design matrix exhibits strong collinearity. We first use toy examples with Zellner’s g-prior to gain some theoretical insight about the behavior of the posterior distribution over models in such a setting. The results suggest that strong collinearity leads to a multimodal posterior distribution and if there are three or more important covariates that are strongly correlated with each other, the median probability model of Barbieri and Berger could potentially discard all of them. Using simulation studies and real data, we illustrate that several other popular priors may also be adversely affected by collinearity. Both theoretical and empirical results suggest that a routine examination of the correlation matrix and calculation of the joint inclusion probabilities for correlated covariates can help practitioners to cope with the problem.

Key Words: Bayesian variable selection; Linear regression; Markov chain Monte Carlo; Median probability model; Multimodality.

1 Introduction

In the Bayesian approach to variable selection in linear regression, all models are embedded in a hierarchical mixture model, with mixture components being models with different subsets of covariates. We first present a brief overview of Bayesian variable selection. Let $Y = (Y_1, \ldots, Y_n)'$
denote the vector of response variables, and let $x_1, x_2, \ldots, x_p$ denote the $p$ covariates. Models corresponding to different subsets of covariates may be represented by the vector $\gamma = (\gamma_1, \ldots, \gamma_p)'$, such that $\gamma_j = 1$ when $x_j$ is included in the model and $\gamma_j = 0$ otherwise. Let $\Gamma$ denote the model space of $2^p$ possible models and $p_\gamma = \sum_{j=1}^p \gamma_j$ denote the number of covariates in model $\gamma$, excluding the intercept. The linear regression model is

$$Y \mid \beta_0, \beta_\gamma, \phi, \gamma \sim \mathcal{N}(1\beta_0 + X_\gamma \beta_\gamma, I_n/\phi),$$

where $1$ is an $n \times 1$ vector of ones, $\beta_0$ is the intercept, $X_\gamma$ is the $n \times p_\gamma$ design matrix and $\beta_\gamma$ is the $p_\gamma \times 1$ vector of regression coefficients under model $\gamma$, $\phi$ is the reciprocal of the error variance, and $I_n$ is an $n \times n$ identity matrix. The intercept is assumed to be included in every model. The models are assigned a prior distribution $p(\gamma)$ and the vector of parameters under each model $\gamma$ is assigned a prior distribution $p(\theta_\gamma \mid \gamma)$, where $\theta_\gamma = (\beta_0, \beta_\gamma, \phi)$. The posterior probability of any model is obtained using Bayes’ rule as:

$$p(\gamma \mid Y) = \frac{p(Y \mid \gamma)p(\gamma)}{\sum_{\gamma \in \Gamma} p(Y \mid \gamma)p(\gamma)},$$

where $p(Y \mid \gamma) = \int p(Y \mid \theta_\gamma, \gamma)p(\theta_\gamma \mid \gamma)d\theta_\gamma$ is the marginal likelihood of $Y$ under model $\gamma$. We will consider scenarios when the marginal likelihood may or may not be available in closed form. For model selection, a natural choice would be the highest probability model (HPM). This model is theoretically optimal for selecting the true model under a 0–1 loss function using decision-theoretic arguments.

When $p$ is larger than $25 – 30$ the posterior probabilities in (2) are not available in closed form for general design matrices due to computational limitations, irrespective of whether the marginal likelihoods can be calculated in closed form or not. Generally one resorts to Markov chain Monte Carlo (MCMC) or other stochastic sampling based methods to sample models. The MCMC sample size is typically far smaller than the dimension $(2^p)$ of the model space, when $p$ is large. As a result Monte Carlo estimates of posterior probabilities of individual models can be unreliable, which makes accurate estimation of the HPM a challenging task. Moreover, for large model spaces the
HPM may have a very small posterior probability, so it is not clear if variable selection should be based on the HPM alone as opposed to combining the information across models. Thus variable selection is often performed with the marginal posterior inclusion probabilities, for which more reliable estimates are available from the MCMC output. The marginal inclusion probability for the $j$th covariate is:

$$p(\gamma_j = 1 \mid Y) = \sum_{\gamma \in \Gamma: \gamma_j = 1} p(\gamma \mid Y).$$

(3)

The use of these can be further motivated by the median probability model (MPM) of Barbieri and Berger (2004). The MPM includes all variables whose posterior marginal inclusion probabilities are greater than or equal to 0.5. Instead of selecting a single best model another option is to consider a weighted average of quantities of interest over all models with weights being the posterior probabilities of models. This is known as Bayesian model averaging (BMA) and it is optimal for predictions under a squared error loss function. However, sometimes from a practical perspective a single model may need to be chosen for future use. In such a situation the MPM is the optimal predictive model under a squared error loss function under certain conditions (Barbieri and Berger, 2004).

In this article we show that the use of marginal inclusion probabilities or the MPM for model selection has the potential to completely miss a set of important correlated covariates when the correlation is fairly strong. In Section 2 we give some theoretical insight into the problem using Zellner’s $g$-prior (Zellner, 1986) for the parameters under each model and a discrete uniform prior for the model space. In Section 3 we illustrate via simulation studies that independent normal priors and their scale mixtures generally perform better than the $g$-prior and its mixtures in the context of collinearity. However, a real data analysis in Section 4 reveals that for a large number of correlated covariates all priors may incorrectly choose the null model. In Section 5 we conclude with some suggestions to cope with the problem of collinearity.
2 Zellner’s $g$-prior and Collinearity

Zellner’s $g$-prior (Zellner, 1986) and its variants are widely used for the model specific parameters in Bayesian variable selection. A key reason for the popularity is perhaps its computational tractability in high-dimensional model spaces. The choice of $g$ is critical in model selection and a variety of choices have been proposed in the literature. In this Section, we shall focus on the unit information $g$-prior with $g=n$, in the presence of strong collinearity. Letting $X$ denote the design matrix under the full model, we assume that the columns of $X$ have been centered to have mean 0 and scaled so that the norm of each column is $\sqrt{n}$, as in Ghosh and Clyde (2011). For the standardized design matrix $X'X$ is $n$ times the observed correlation matrix. Under model $\gamma$ the $g$-prior is given by:

$$p(\beta_0, \phi \mid \gamma) \propto 1/\phi \begin{pmatrix} 0 & g(X_\gamma'X_\gamma)^{-1} \end{pmatrix} \begin{pmatrix} \beta_\gamma \mid \gamma, \phi \end{pmatrix} \sim N.$$ (4)

Let $\hat{Y}_\gamma = 1\hat{\beta}_0 + X_\gamma\hat{\beta}_\gamma$, where $\hat{\beta}_0 = \bar{Y} = \sum_{i=1}^{n} Y_i/n$ and $\hat{\beta}_\gamma = (X_\gamma'X_\gamma)^{-1}X_\gamma'Y$ are the ordinary least squares estimates of $\beta_0$ and $\hat{\beta}_\gamma$. Let the regression sum of squares for model $\gamma$ be $SSR_\gamma = \sum_{i=1}^{n} (\hat{Y}_\gamma - \bar{Y})^2$ and the total sum of squares be $SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$. Then the coefficient of determination for model $\gamma$ is $R^2_\gamma = SSR_\gamma/SST$. When $\gamma$ is the null model with only the intercept term, $\hat{Y}_\gamma = 1\bar{Y}$, thus its SSR$_\gamma = 0$ and $R^2_\gamma = 0$. The marginal likelihood for the $g$-prior can be calculated analytically as follows:

$$p(Y \mid \gamma) \propto (1 + g)^{-\frac{n-p}{2}} \left\{1 + g(1 - R^2_\gamma)\right\}^{-\frac{n-p}{2}},$$ (5)

where the constant of proportionality does not depend on $\gamma$ (see Section 2.1 equation (5) of Liang et al. (2008)). We assume throughout that we have a discrete uniform prior for the model space so that $p(\gamma) = 1/2^p$ for all models. For exploration of non-enumerable model spaces MCMC may be used such that $p(\gamma \mid Y)$ is the target distribution of the Markov chain. George and McCulloch (1997) discuss fast updating schemes for MCMC sampling with the $g$-prior.

First, we consider a small simulation study for $p = 3$ with strong collinearity among the covari-
ates, as a motivating example for this article. In later sections we consider some toy examples to explain this problem theoretically and hence obtain a better understanding of the properties of the MPM. For our theoretical examples, we will deal with finite and large \( n \) under conditions of severe collinearity. Our results complement the results of Fernández et al. (2001) who showed that model selection consistency holds for the \( g \)-prior with \( g = n \). Their result implies that under appropriate assumptions, \( p(\gamma|Y) \) will converge to 1 in probability, if \( \gamma \in \Gamma \) is the true model. Our simulations and theoretical calculations demonstrate that under severe collinearity the posterior distribution over models may become multimodal and very large values of \( n \) may be needed for consistency to kick in.

2.1 Motivating Simulation Study for \( p = 3 \)

First we conduct a small simulation study with \( p = 3 \) so that we can explicitly list each of the \( 2^3 \) models along with their \( R^2 \) values and posterior probabilities, to demonstrate the problem associated with severe collinearity empirically. We generate the covariates such that they are all highly correlated with each other, as follows.

We sample a vector of \( n \) standard normal variables, say \( z \), and then generate each of the covariates by adding another vector of \( n \) independent normal variables with mean 0 and standard deviation 0.05 to \( z \). This results in pairwise correlations of about 0.997 among all the covariates. We set the intercept and the regression coefficients for all the covariates equal to one, and generate the response variable as in model (1) with \( \phi = 1/4 \). We look at a range of moderate to extremely large sample sizes in Tables 1 and 2.

Table 1 shows that high positive correlations among the important covariates lead to similar \( R^2 \) values across all non-null models. For the \( g \)-prior this translates into high posterior probabilities for the single variable models, in spite of the full model being the true model. The full model does not have a high posterior probability even for \( n = 10^4 \), finally the posterior consistency comes into play when \( n \) is as large as \( 10^5 \).

Table 2 studies the corresponding posterior inclusion probabilities of covariates. We find that for \( n = 25 \) and \( n = 100 \), the marginal inclusion probabilities are all smaller than 0.5, so the MPM
Table 1: Simulation study for $p = 3$, to demonstrate the effect of collinearity on posterior probabilities of models; the posterior probabilities of the top 3 models have been highlighted.

will be the null model. However, for all values of $n$, the joint inclusion probability that at least one of the correlated covariates is included in the model is 1. This suggests that the joint inclusion probabilities are still effective measures of importance of covariates even when the MPM or the HPM are adversely affected by collinearity.

Table 2: Simulation study for $p = 3$, to demonstrate the effect of collinearity on posterior marginal and joint inclusion probabilities of covariates.

In the following sections we first introduce a few assumptions and propositions and then conduct a theoretical study of the $p = 2$ case followed by that for the general $p$ case.

### 2.2 Assumptions About $R^2$ and Collinearity

First note that for the null model $\gamma = (0,0,\ldots,0)'$, we have $R^2_\gamma = 0$, by definition. To deal with random $R^2_\gamma$ for non-null models, we make the following assumption:

**Assumption 1.** Assume that the true model is the full model and that $0 < \delta_1 < R^2_\gamma < \delta_2 < 1$, for all sample size $n$ and for all non-null models $\gamma \in \Gamma - \{(0,0,\ldots,0)\}'$, with probability 1.

**Proposition 1.** If Assumption 1 holds, then for given $\epsilon > 0$, and for $g = n$ sufficiently large, the
Bayes factor for comparing $\gamma = (0, 0, \ldots, 0)'$ and $\gamma = (1, 0, \ldots, 0)'$ can be made smaller than $\epsilon$, with probability 1.

The proof is given in Appendix A. This result implies that the Bayes factor,

$$BF(\gamma = (0, 0, \ldots, 0)' : \gamma = (1, 0, \ldots, 0)') \approx 0,$$

with probability 1, if the specified conditions hold.

For a discrete uniform prior for the model space, that is $p(\gamma) = 1/2^p$ for all models $\gamma \in \Gamma$, the posterior probability of any model $\gamma$ may be expressed entirely in terms of Bayes factors as:

$$p(\gamma \mid Y) = \frac{p(Y \mid \gamma)(1/2^p)}{\sum_{\gamma \in \Gamma} p(Y \mid \gamma)(1/2^p)} = \frac{p(Y \mid \gamma)}{\sum_{\gamma \in \Gamma} p(Y \mid \gamma)} = \frac{p(Y \mid \gamma)/p(Y \mid \gamma^*)}{\sum_{\gamma \in \Gamma} p(Y \mid \gamma)/p(Y \mid \gamma^*)} = \frac{BF(\gamma : \gamma^*)}{\sum_{\gamma \in \Gamma} BF(\gamma : \gamma^*)},$$

where $\gamma^* \in \Gamma$ (Berger and Molina, 2005). Taking $\gamma = (0, 0, \ldots, 0)'$ and $\gamma^* = (1, 0, \ldots, 0)'$ in (7), and using (6), for large enough $n$ we have the following with probability 1,

$$p(\gamma = (0, 0, \ldots, 0)' \mid Y) \approx 0.$$

As the null model receives negligible posterior probability we may omit it when computing the normalizing constant of $p(\gamma \mid Y)$, that is we may compute the posterior probabilities of non-null models by re-normalizing over the set $\Gamma - \{(0, 0, \ldots, 0)\}'$ instead of $\Gamma$. We provide a formal justification of this approximation in Appendix B.

We now make an assumption about strong collinearity among the covariates, so that $R_{\gamma}^2$ for all non-null models $\gamma \in \Gamma - \{(0, 0, \ldots, 0)\}'$ are sufficiently close to each other, with probability 1.

**Assumption 2.** Assume that the $p$ covariates are highly correlated with each other such that the ratio

$$\left\{\frac{1+n(1-R_{\gamma}^2)}{1+n(1-R_{\gamma'}^2)}\right\}^{\frac{n-1}{2}}$$

can be taken to be approximately 1, for any pair of distinct non-null models $\gamma$ and $\gamma'$, with probability 1.

The above assumption is not made in an asymptotic sense, instead it assumes that the collinearity is strong enough for the condition to hold over a range of large $n$, but not necessarily as $n \to \infty$. 
We next study the posterior distribution of $2^2$ models for an example with $p = 2$ highly correlated covariates and extend the results to the general $p$ scenario in the following section.

### 2.3 Collinearity Example for $p = 2$

Under the discrete uniform prior for the model space we have,

$$p(\gamma) = \frac{1}{2^2}, \gamma \in \Gamma. \quad (9)$$

Under Assumptions 1 and 2, the posterior probabilities of the $2^2$ models can be approximated as follows, with probability 1:

$$p(\gamma = (0, 0)' \mid Y) \approx 0, \quad (10a)$$
$$p(\gamma = (0, 1)' \mid Y) \approx \frac{1}{2}, \quad (10b)$$
$$p(\gamma = (1, 0)' \mid Y) \approx \frac{1}{2}, \quad (10c)$$
$$p(\gamma = (1, 1)' \mid Y) \approx 0. \quad (10d)$$

The detailed calculations are given in Appendix C. The calculations in (10) have the following implications with probability 1. First, the posterior is approximately bimodal with modes at the one-variable models, $\gamma = (0, 1)'$ and $\gamma = (1, 0)'$. Second, as each mode receives approximately 0.5 posterior probability, this implies that the marginal posterior inclusion probabilities $p(\gamma_1 = 1 \mid Y)$ and $p(\gamma_2 = 1 \mid Y)$ would also be close to 0.5. Third, the MPM picks all variables whose marginal inclusion probabilities are greater than or equal to 0.5, so in this case it is likely that the MPM will include at least one of the two important covariates, which happened in all our simulations in Section 3.

Finally note that the prior probability that at least one of the important covariates is included is $p(\gamma_1 = 1 \text{ or } \gamma_2 = 1) = 1 - p(\gamma = (0, 0)') = 1 - (1/2)^2 = 3/4$, by (9), under the discrete uniform prior for the model space. The posterior probability that at least one of the important covariates is included is $P(\gamma_1 = 1 \text{ or } \gamma_2 = 1 \mid Y) = 1 - p(\gamma = (0, 0)' \mid Y) \approx 1$, by (10a). Let $H_0$ denote
\( \gamma = (0,0)' \) and \( H_A \) denote its complement, that is \( H_A \) denotes \( \gamma_1 = 1 \) or \( \gamma_2 = 1 \). Then the prior odds \( P(H_A)/P(H_0) = (3/4)/(1/4) = 3 \) and the posterior odds \( P(H_A \mid Y)/P(H_0 \mid Y) \) is expected to be very large, because \( P(H_A \mid Y) \approx 1 \) and \( P(H_0 \mid Y) \approx 0 \), by (10). Then it is clear from the calculations of the prior and posterior odds that the Bayes factor \( BF(H_A : H_0) = \frac{P(H_A \mid Y)/P(H_0 \mid Y)}{P(H_A)/P(H_0)} \) will be very large with probability 1, under the above assumptions.

### 2.4 Collinearity Example for General \( p \)

Consider a similar set up as above with \( p \) highly correlated covariates that are all important, that is the full model \( \gamma = (1,1,\ldots,1)' \) is the true model. Under Assumptions 1 and 2 the following results hold with probability 1. We implicitly assume throughout this section that the results hold with probability 1, but refrain from mentioning it explicitly, for ease of exposition.

Under Assumption 1, provided that \( n \) is large enough, the null model has nearly zero posterior probability by (8), so it is not considered in the calculation of the normalizing constant for posterior probabilities of models as before. Under Assumption 2 and using \( g = n \) in (5), all \((2^p - 1)\) non-null models have the term \( \{1 + n(1 - R_\gamma^2)\}^{-(n-1)/2} \) approximately in common. Ignoring terms that are common across all non-null models (exactly or approximately), the marginal likelihood for any model of dimension \( p_\gamma \) is approximately proportional to \( (1 + n)^{-\frac{n - p_\gamma - 1}{2}} \). For a given sample size \( n \), this term decreases as \( p_\gamma \) increases, so the lowest dimensional non-null models with \( p_\gamma = 1 \) will have the highest posterior probability. This implies that the posterior will be multimodal with \( p \) modes at each of the \( p \) one-dimensional models.

Using the standard definition of posterior marginal inclusion probability for the \( j \)th covariate:

\[
p(\gamma_j = 1 \mid Y) = \frac{\sum_{\gamma \in \Gamma: \gamma_j = 1} p(Y \mid \gamma)}{\sum_{\gamma \in \Gamma} p(Y \mid \gamma)} \approx \frac{\sum_{\gamma \in \Gamma: \gamma_j = 1} p(Y \mid \gamma)}{\sum_{\gamma \in (0,0,\ldots,0)'} p(Y \mid \gamma)}, \text{ (by (18) in the proof of Corollary 1 in Appendix B)}
\]

\[
\approx \sum_{p_\gamma = 1}^p \left( \frac{p-1}{p_\gamma} \right) \left( 1 + n \right)^{-\frac{n - p_\gamma - 1}{2}},
\]

where the last approximation sign is due to Assumption 2 (\( R_\gamma^2 \) are nearly the same for all non-null}
models due to high multicollinearity). The expression in (11) is derived using the fact that i) all \( p_{\gamma} \)-dimensional models have the marginal likelihood proportional to \((1 + n)^{n-p_{\gamma}-1}/2\) approximately (using Assumption 2 in (5)), ii) there are altogether \( (p_{\gamma}) \) such models, and iii) exactly \((p_{\gamma}-1)\) of these have \( \gamma_j = 1 \). Dividing the numerator and denominator of (11) by \((1 + n)^{n-2}/2\) and writing separately the leading term in the sum corresponding to \( p_{\gamma} = 1 \), the expression in (11) simplifies to

\[
p(\gamma_j = 1 \mid Y) \approx \frac{1 + \sum_{p_{\gamma} = 2}^{p} \left( \frac{p_{\gamma}-1}{p_{\gamma}} \right) \left( 1 + n \right)^{(p_{\gamma}-1)/2}}{p + \sum_{p_{\gamma} = 2}^{p} \left( \frac{p_{\gamma}}{p} \right) \left( 1 + n \right)^{(p_{\gamma}-1)/2}} \approx \frac{1}{p},
\]

(12)

where the last approximation follows for fixed \( p \) and sufficiently large \( n \), as the terms in the sum over \( p_{\gamma} \) (from 2 to \( p \)) involve negative powers of \((1 + n)\). This result suggests that the MPM will have greater problems due to collinearity for \( p \geq 3 \) compared to \( p = 2 \). When \( n \) is not so large in practice, the inclusion probabilities may not be shrunk to a value as small as \( 1/p \). However, we shall see in Section 4 that all of them may be shrunk below 0.5, which in turn would affect the MPM.

The prior probability that at least one of the \( p \) (important) covariates is included in the model is \( p(\gamma_1 = 1 \text{ or } \gamma_2 = 1 \ldots \text{ or } \gamma_p = 1) = 1 - p(\gamma = (0,0,\ldots,0)^T) = 1 - (1/2^p) \) under the discrete uniform prior for the model space, and the corresponding posterior probability \( P(\gamma_1 = 1 \text{ or } \gamma_2 = 1 \ldots \text{ or } \gamma_p = 1 \mid Y) \approx 1 - p(\gamma = (0,0,\ldots,0)^T \mid Y) \approx 1 \), by (8). Let \( H_0 \) denote \( \gamma = (0,0,\ldots,0)^T \) and \( H_A \) denote its complement. It is clear that the posterior odds \( P(H_A \mid Y)/P(H_0 \mid Y) \) will be very large for sufficiently large \( n \), because \( P(H_A \mid Y) \approx 1 \) and \( P(H_0 \mid Y) \approx 0 \). Since \( p \) is fixed the prior odds \( P(H_A)/P(H_0) = (2^p - 1) \) is fixed, so the Bayes factor \( BF(H_A : H_0) = \frac{P(H_A \mid Y)/P(H_0 \mid Y)}{P(H_A)/P(H_0)} \) is expected to be very large as in the \( p = 2 \) case, provided \( n \) is large enough. This is a useful result and suggests that while marginal inclusion probabilities (and hence marginal Bayes factors) may give misleading conclusions about the importance of covariates, the joint inclusion probabilities and joint Bayes factors would still indicate that at least one of the covariates should be included in the model. We next conduct simulation studies to demonstrate the usefulness of these theoretical results.
3 Simulation Studies

Our goal is to compare the performance of the MPM for different priors under collinearity. In particular, we consider i) the multivariate normal $g$-prior (Zellner, 1986) with $g = n$, ii) the multivariate Zellner-Siow (Zellner and Siow, 1980) Cauchy prior, iii) independent normal priors, and iv) independent $t$ priors with 4 degrees of freedom. For the independent priors we use the same formulation as Ghosh and Clyde (2011), and for the independent normal priors we choose the same hyperparameters. We set $p = 10$ so that the model space can be enumerated, and posterior marginal inclusion probabilities can be calculated analytically for the normal priors. This ensures that there is no ambiguity in the results due to approximation. For the Zellner-Siow prior, marginal likelihoods are approximated by a Laplace approximation for a one-dimensional integral over $g$, see for example Liang et al. (2008) for more details. Marginal likelihoods for independent $t$ priors involve higher dimensional integrals that are not available in closed form, so we use the orthogonal data augmentation (ODA) MCMC sampler of Ghosh and Clyde (2011). The posterior computation for the $g$-priors was done by enumerating the model space with the BAS algorithm (Clyde et al., 2011).

3.1 Important Correlated Covariates

We take $n = 50$, $p = 10$, and $q = 2, 3, 4$, where $q$ is the number of correlated covariates. We sample a vector of $n$ standard normal variables, say $z$, and then generate each of the $q$ correlated covariates by adding another vector of $n$ independent normal variables with mean 0 and standard deviation 0.05 to $z$. This results in pairwise correlations of about 0.997 among the correlated covariates. The remaining $(p - q)$ covariates are generated independently as $N(0, 1)$ variables. We set the intercept and the regression coefficients for the correlated covariates equal to one, and all other regression coefficients equal to zero. The response variable is generated according to model (1) with $\phi = 1/4$ and the procedure is repeated to generate 100 datasets. For the independent $t$ priors the MCMC sampler is run for 10,000 iterations and the first 1,000 are discarded as burn-in. For other priors all $2^{10} = 1,024$ models are enumerated.

If a covariate included in the true model is not selected by the MPM, it is considered a false
negative. If a noise variable is selected by the MPM, that leads to a false positive. It could be argued that as long as the MPM includes at least one of the important correlated covariates, predictive performance of the model will not be adversely affected. Thus we also consider the cases when the MPM drops the entire group of important correlated covariates, when \( p(\gamma_j = 1 \mid Y) < 0.5 \) for all \( q \) correlated covariates. Results are summarized in Table 3 in terms of four quantities, of which the first three measure the performance of the MPM. They are i) FN: proportion of false negatives, ii) FP: proportion of false positives, iii) Null: proportion of datasets in which the MPM discards all important correlated covariates simultaneously, and iv) BF: proportion of datasets in which the Bayes factor \( BF(H_A : H_0) \geq 10 \), where \( H_0 \) is the hypothesis that \( \gamma_j = 0 \) for all the \( q \) correlated covariates and \( H_A \) denotes its complement.

Table 3 shows that the proportion of false negatives is much higher for \( q > 2 \) than \( q = 2 \). With \( q = 4 \) this rate is higher than 80% for the \( g \)-priors and higher than 10% for the independent priors. The false positive rate is generally low and the performance is similar across all priors. For \( q = 2 \) none of the priors drop all correlated covariates together. However, for \( q = 3, 4 \), the \( g \)-priors show this behavior in about 40-50% cases. This problem may be tackled by considering joint inclusion probabilities for correlated covariates (Berger and Molina, 2005; George and McCulloch, 1997), and the corresponding Bayes factors lead to a correct conclusion 99-100% of the time. The independent priors seem more robust to collinearity and they never discard all important correlated covariates.

<table>
<thead>
<tr>
<th>Prior</th>
<th>( q = 2 )</th>
<th>( q = 3 )</th>
<th>( q = 4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FN</td>
<td>FP</td>
<td>Null</td>
</tr>
<tr>
<td>( g )-prior</td>
<td>0.36</td>
<td>0.06</td>
<td>0.00</td>
</tr>
<tr>
<td>Zellner-Siow</td>
<td>0.21</td>
<td>0.08</td>
<td>0.00</td>
</tr>
<tr>
<td>Independent normal</td>
<td>0.01</td>
<td>0.06</td>
<td>0.00</td>
</tr>
<tr>
<td>Independent ( t )</td>
<td>0.01</td>
<td>0.05</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Table 3: Simulation study with \( p = 10 \) covariates, of which \( q \) correlated covariates are included in the true model as signals, and \((p - q)\) uncorrelated covariates denote noise.

### 3.2 Unimportant Correlated Covariates

In this simulation study we consider the same values of \( n, p, \) and \( q, \) and generate the design matrix as before. We now set the regression coefficients for the \( q \) correlated covariates at zero, and the
remaining \((p - q)\) coefficients at one.

The results based on repeating the procedure 100 times are presented in Table 4. The false negative rates are similar across priors. This is expected because these are affected by the uncorrelated covariates only. The false positive rates are generally small and similar across priors, so the MPM does not seem to have any problems in discarding unimportant correlated covariates. The Bayes factors based on joint inclusion indicators lead to a correct conclusion 99-100% of the time.

<table>
<thead>
<tr>
<th>Prior</th>
<th>(q = 2)</th>
<th>(q = 3)</th>
<th>(q = 4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(g)-prior</td>
<td>0.15 0.03 0.00</td>
<td>0.16 0.02 0.01</td>
<td>0.15 0.03 0.00</td>
</tr>
<tr>
<td>Zellner-Siow</td>
<td>0.11 0.07 0.00</td>
<td>0.11 0.06 0.01</td>
<td>0.10 0.07 0.00</td>
</tr>
<tr>
<td>Independent normal</td>
<td>0.14 0.08 0.00</td>
<td>0.16 0.03 0.00</td>
<td>0.14 0.00 0.00</td>
</tr>
<tr>
<td>Independent (t)</td>
<td>0.16 0.07 0.00</td>
<td>0.17 0.02 0.00</td>
<td>0.15 0.01 0.00</td>
</tr>
</tbody>
</table>

Table 4: Simulation study with \(p = 10\) covariates, of which \(q\) correlated noise variables are not included in the true model, and \((p - q)\) uncorrelated covariates are included in the true model as signals.

4 Application to Biscuit Dough Data

We analyze the biscuit dough data, available as \texttt{cookie} in the \texttt{R} package \texttt{ppls} (Kraemer and Boulesteix, 2012). The dataset was obtained from an experiment that used near-infrared (NIR) spectroscopy to analyze the composition of biscuit dough pieces. An NIR reflectance spectrum for each dough is a continuous curve measured at many equally spaced wavelengths. Brown \textit{et al.} (2001) omitted the first 140 and last 49 of the available 700 wavelengths because these contained small information. For our analysis we choose the wavelengths 191 – 205 to have an enumerable model space with \(p = 15\) covariates with high pairwise correlations (around 0.999) among all of them, and the percentage of fat as the response variable. We use the training sample of size 39 in \texttt{ppls} for estimation, and the test set of size 31 for prediction. For the independent \(t\) priors, we run the MCMC sampler for 100,000 iterations after a burn-in of 25,000 samples.

For every prior the MPM is the model with only an intercept, and its prediction mean squared error (MSE) is 3.945. We next calculate the Bayes factor \(\text{BF}(H_A : H_0)\), where \(H_0\) is the model with only the intercept and \(H_A\) denotes its complement. The Bayes factors are i) 114, ii) 69, iii)
11,750, and iv) 6,425,129, for the i) $g$-prior, ii) Zellner-Siow prior, iii) independent normal priors, and iv) independent $t$ priors. The Bayes factors differ among the priors in magnitude, but they unanimously provide strong evidence against $H_0$. This suggests that the full model is worth an investigation. The MSE for the full model under the $g$-priors are 4.55 and 3.98, and are 2.00 and 1.98 for the independent priors.

This example illustrates that all priors could potentially drop a group of important correlated covariates, and the predictive performance of $g$-priors can be more adversely affected by collinearity than that of independent priors. For a given model $\gamma$, the posterior mean of $\beta,\gamma$ under the $g$-prior with a fixed $g$ is $\frac{g}{1+g}\hat{\beta},\gamma$, where $\hat{\beta},\gamma$ is the ordinary least squares (OLS) estimate of $\beta,\gamma$ (Liang et al., 2008; Ghosh and Reiter, 2013). It is well-known that OLS estimates can be unstable due to high variance under collinearity, so it is not surprising that the $g$-prior inherits this property. The corresponding estimate under the independent normal priors is a ridge regression estimate (Ghosh and Clyde, 2011), which is known to be more stable under collinearity. Note that a different choice of wavelengths as covariates may not lead to selection of the null model as the MPM, but our goal is to illustrate that this can happen in practice and that data analysts need to be cautious about these potential problems.

5 Discussion

Based on the empirical results it seems preferable to use scale mixtures of independent priors for design matrices with high collinearity instead of scale mixtures of $g$-priors. The MPM is easy to understand, straightforward to estimate, and it generally has good performance except in cases of severe collinearity. Thus we recommend a two-step procedure: using the MPM for variable selection as a first step, followed by an inspection of joint inclusion probabilities and Bayes factors for groups of correlated covariates, as a second step. For complex correlation structures it may be desirable to incorporate that information in the prior. Krishna et al. (2009) proposed a new powered correlation prior for the regression coefficients and a new model space prior with this objective. The posterior computation for their prior will be computationally very demanding for high dimensions compared to some of the other standard priors like independent normal and $t$ priors used in this paper. Thus
development of priors along the lines of Krishna et al. (2009) that scale well with the dimension of the model space is a promising direction for future research.

Appendix A: Proof of Proposition 1

Proof. To simplify the notation let $R_\gamma^2 = R^2$ for $\gamma = (1, 0, \ldots, 0)'$. Then putting $g = n$ and using the expression for marginal likelihood of the $g$-prior given in (5) we have,

$$BF(\gamma = (0, 0, \ldots, 0)': \gamma = (1, 0, \ldots, 0)') = \frac{p(Y | \gamma = (0, 0, \ldots, 0)')}{p(Y | \gamma = (1, 0, \ldots, 0)')}$$

$$= \frac{1}{(1 + n)^{(n-2)/2} \left( 1 + n(1 - R^2) \right)^{- (n-1)/2}}$$

$$= \frac{1}{(1 + n)^{(n-2)/2} \left( \frac{1 + n(1 - R^2)}{1 + n} \right)^{- (n-1)/2}}$$

$$= \frac{1}{(1 + n)^{(n-2)/2} \left( \frac{1 + n - n R^2}{1 + n} \right)^{- (n-1)/2} (1 + n)^{- (n-1)/2}}$$

$$= \frac{1}{(1 + n)^{(n-2)/2} \left( 1 - \frac{n}{1 + n} R^2 \right)^{- (n-1)/2}}$$

$$= \frac{1}{(1 + n)^{-1/2} \left( 1 - \frac{n}{1 + n} R^2 \right)^{- (n-1)/2}}$$

$$= (1 + n)^{1/2} \left( 1 - \frac{n}{1 + n} R^2 \right)^{(n-1)/2}$$

(13)

Taking the logarithm of (13) the following result holds with probability 1, by Assumption 1:

$$\log(BF(\gamma = (0, 0, \ldots, 0)': \gamma = (1, 0, \ldots, 0)')) = \log \left( (1 + n)^{1/2} \left( 1 - \frac{n}{1 + n} R^2 \right)^{(n-1)/2} \right)$$

$$= \frac{1}{2} \log(1 + n) + \frac{(n - 1)}{2} \log(1 - \frac{n}{n + 1} R^2)$$

(14)

$$< \frac{1}{2} \log(1 + n) + \frac{(n - 1)}{2} \log(1 - \frac{n}{n + 1} \delta_1)$$

(15)

As $n$ goes to infinity, the first term in (15) goes to $\infty$ at a logarithmic rate in $n$. Logarithm is a continuous function so $\log(1 - \frac{n}{n + 1} \delta_1)$ goes to $\log(1 - \delta_1)$ as $n$ goes to infinity. Because $0 < \delta_1 < 1$,
we have \( \log(1 - \delta_1) < 0 \). This implies that the second term in (15) goes to \(-\infty\) at a polynomial rate in \( n \), of degree 1. Thus, as \( n \to \infty \), with probability 1,

\[
\log(BF(\gamma = (0,0,\ldots,0)' : \gamma = (1,0,\ldots,0)')) \to -\infty, \quad \text{or} \quad (16)
\]

\[
BF(\gamma = (0,0,\ldots,0)' : \gamma = (1,0,\ldots,0)') \to 0. \quad (17)
\]

From (17) it follows that for sufficiently large \( n \), we can make \( BF(\gamma = (0,0,\ldots,0)' : \gamma = (1,0,\ldots,0)') < \epsilon \), for given \( \epsilon > 0 \), with probability 1. This completes the proof. \( \square \)

Note that the above proof is based on an argument where we consider the limit as \( n \to \infty \). However, for other results concerning collinearity, we assume that \( n \) is large but finite. Thus we avoid the use of limiting operations in the main body of the article to avoid giving the reader an impression that we are doing asymptotics.

**Appendix B: Justification for Omission of the Null Model for Computing the Normalizing Constant \( \sum_{\gamma \in \Gamma} p(Y | \gamma) \)**

We first establish the following lemma. This shows that for computing a finite sum of positive quantities, if one of the quantities is negligible compared to another, then the sum can be computed accurately even if the quantity with negligible contribution is omitted from the sum.

**Lemma 1.** Consider \( a_{in} > 0, \ i = 1, 2, \ldots m \). If \( \frac{a_{1n}}{a_{2n}} \to 0 \) as \( n \to \infty \) then \( \sum_{i=1}^{m} \frac{a_{in}}{a_{2n}} \to 1 \).

**Proof.**

\[
\lim_{n \to \infty} \frac{\sum_{i=2}^{m} a_{in}}{\sum_{i=1}^{m} a_{in}} = \lim_{n \to \infty} \frac{\sum_{i=2}^{m} a_{in}/a_{2n}}{\sum_{i=1}^{m} a_{in}/a_{2n}} = \lim_{n \to \infty} \frac{\sum_{i=2}^{m} a_{in}/a_{2n}}{(a_{1n}/a_{2n}) + \sum_{i=2}^{m} a_{in}/a_{2n}} = \lim_{n \to \infty} \frac{\sum_{i=2}^{m} a_{in}/a_{2n}}{\lim_{n \to \infty} (a_{1n}/a_{2n}) + \lim_{n \to \infty} \sum_{i=2}^{m} a_{in}/a_{2n}} = \lim_{n \to \infty} \frac{\sum_{i=2}^{m} a_{in}/a_{2n}}{0 + \lim_{n \to \infty} \sum_{i=2}^{m} a_{in}/a_{2n}} = 1
\]
Corollary 1. If Assumption 1 holds, then given \( \eta > 0 \), however small, for sufficiently large \( n \), we can make \( \left( 1 - \frac{\sum_{\gamma \in \Gamma - \{(0,0,0,\ldots,0)'} p(Y|\gamma)}{\sum_{\gamma \in \Gamma} p(Y|\gamma)} \right) < \eta \), with probability 1.

Proof. We have \( p(Y|\gamma) > 0, \gamma \in \Gamma \) and \( \frac{p(Y|\gamma=(0,0,\ldots,0)')}{p(Y|\gamma=(1,0,\ldots,0)')} \to 0 \) as \( n \to \infty \), with probability 1, by (17). Then as \( n \to \infty \) we have the following, with probability 1, by Lemma 1:

\[
\frac{\sum_{\gamma \in \Gamma - \{(0,0,\ldots,0)'} p(Y|\gamma)}{\sum_{\gamma \in \Gamma} p(Y|\gamma)} \to 1. \tag{18}
\]

The proof follows immediately. \( \square \)

Appendix C: Calculation of Posterior Probabilities of all \( 2^2 \) Models for \( p = 2 \)

The posterior probability of the null model was shown to be approximately zero in (8). We derive the posterior probabilities of the non-null models under Assumptions 1 and 2 here. For any non-null model \( \gamma \in \Gamma - \{(0,0)'} \),

\[
p(\gamma | Y) = \frac{p(\gamma)p(Y|\gamma)}{\sum_{\gamma \in \Gamma} p(\gamma)p(Y|\gamma)} = \frac{(1/2^2)p(Y|\gamma)}{\sum_{\gamma \in \Gamma}(1/2^2)p(Y|\gamma)} \quad \text{(because } p(\gamma) = 1/2^2 \text{ for } \gamma \in \Gamma) \\
= \frac{p(Y|\gamma)}{\sum_{\gamma \in \Gamma} p(Y|\gamma)} \approx \frac{p(Y|\gamma)}{\sum_{\gamma \in \Gamma - \{(0,0)'} p(Y|\gamma)}, \tag{19}
\]

with probability 1. The last approximation in (19) follows from Corollary 1 in Appendix B.

We will use the expression in (19) to derive the posterior probabilities. First note that under Assumption 2 the term \( \{1 + n(1 - R^2_{\gamma})\}^{-\frac{(n-1)}{2}} \) in the expression of marginal likelihood \( p(Y|\gamma) \) in (5) is approximately the same across all non-null models with probability 1. Thus this term does not have to be taken into account when computing the posterior probabilities by (19). Then by
(5), (19), and substituting \( g = n \) we have with probability 1,

\[
p(\gamma = (0, 1)' | \mathbf{Y}) \approx \frac{(1 + n)^{(n-1-1)/2}}{(1 + n)^{(n-1-1)/2} + (1 + n)^{(n-1-1)/2} + (1 + n)^{(n-2-1)/2}}.
\]  

(20)

Dividing the numerator and denominator of the right hand side of (20) by \((1 + n)^{(n-2)/2}\),

\[
p(\gamma = (0, 1)' | \mathbf{Y}) \approx \frac{1}{2 + (1 + n)^{-1/2}} \\
\approx \frac{1}{2},
\]

(21)

for large enough \( n \), with probability 1.

Under Assumption 2, we note that \( p(\gamma = (1, 0)' | \mathbf{Y}) \) would have an identical expression as \( p(\gamma = (0, 1)' | \mathbf{Y}) \). Hence

\[
p(\gamma = (1, 0)' | \mathbf{Y}) \approx \frac{1}{2},
\]

(22)

for large enough \( n \), with probability 1.

We finally derive \( p(\gamma = (1, 1)' | \mathbf{Y}) \) in a similar manner as follows,

\[
p(\gamma = (1, 1)' | \mathbf{Y}) \approx \frac{(1 + n)^{(n-2-1)/2}}{(1 + n)^{(n-1-1)/2} + (1 + n)^{(n-1-1)/2} + (1 + n)^{(n-2-1)/2}} \\
\approx \frac{1}{2(1 + n)^{1/2} + 1} \\
\approx 0,
\]

(23)

for large enough \( n \), with probability 1.

**References**


