Bayesian Variable Selection Under Collinearity

Joyee Ghosh* Andrew E. Ghattas†

Abstract

In this article we highlight some interesting facts about Bayesian variable selection methods for linear regression models in settings where the design matrix exhibits strong collinearity. We first demonstrate via real data analysis and simulation studies that summaries of the posterior distribution based on marginal and joint distributions may give conflicting results for assessing the importance of strongly correlated covariates. The natural question is which one should be used in practice. The simulation studies suggest that summary statistics that evaluate the importance of correlated covariates jointly are more appropriate, and some priors may be more adversely affected in such a setting. To obtain a better understanding behind the phenomenon we study some toy examples with Zellner's g-prior. The results show that strong collinearity may lead to a multimodal posterior distribution over models, in which joint summaries are more robust than marginal summaries. Thus we recommend a routine examination of the correlation matrix and calculation of the joint inclusion probabilities for correlated covariates, in addition to marginal inclusion probabilities, for assessing the importance of covariates in Bayesian variable selection.

Key Words: Bayesian model averaging; Linear regression; Marginal inclusion probability; Markov chain Monte Carlo; Median probability model; Multimodality; Zellner's g-prior.

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

*Joyee Ghosh is Assistant Professor, Department of Statistics and Actuarial Science, The University of Iowa, Iowa City, IA 52242 (Email joyee-ghosh@uiowa.edu)

†Andrew E. Ghattas is Ph.D. candidate, Department of Biostatistics, The University of Iowa, Iowa City, IA 52242 (Email andrew-ghattas@uiowa.edu)

covariates. We first present a brief overview of Bayesian variable selection. Let $\mathbf{Y} = (Y_1, \ldots Y_n)'$ denote the vector of response variables, and let $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$ denote the p covariates. Models corresponding to different subsets of covariates may be represented by the vector $\boldsymbol{\gamma} = (\gamma_1, \ldots \gamma_p)'$, such that $\gamma_j = 1$ when $\boldsymbol{x}_j$ is included in the model and $\gamma_j = 0$ otherwise. Let Γ denote the model space of 2^p possible models and $p_{\boldsymbol{\gamma}} = \sum_{j=1}^{p} \gamma_j$ denote the number of covariates in model $\boldsymbol{\gamma}$, excluding the intercept. The linear regression model is

$$\mathbf{Y} \mid \beta_0, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \phi, \boldsymbol{\gamma} \sim \mathsf{N}(\mathbf{1}\beta_0 + \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \mathbf{I}_n/\phi), \tag{1}$$

where $\mathbf{1}$ is an $n \times 1$ vector of ones, β_0 is the intercept, $\mathbf{X}_{\boldsymbol{\gamma}}$ is the $n \times p_{\boldsymbol{\gamma}}$ design matrix and $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ is the $p_{\boldsymbol{\gamma}} \times 1$ vector of regression coefficients under model $\boldsymbol{\gamma}$, ϕ is the reciprocal of the error variance, and $\mathbf{I}_n$ is an $n \times n$ identity matrix. The intercept is assumed to be included in every model. The models in (1) are assigned a prior distribution $p(\boldsymbol{\gamma})$ and the vector of parameters under each model $\boldsymbol{\gamma}$ is assigned a prior distribution $p(\boldsymbol{\theta}_{\boldsymbol{\gamma}} \mid \boldsymbol{\gamma})$, where $\boldsymbol{\theta}_{\boldsymbol{\gamma}} = (\beta_0, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \phi)$. The posterior probability of any model is obtained using Bayes' rule as:

$$p(\boldsymbol{\gamma} \mid \mathbf{Y}) = \frac{p(\mathbf{Y} \mid \boldsymbol{\gamma})p(\boldsymbol{\gamma})}{\sum_{\boldsymbol{\gamma} \in \Gamma} p(\mathbf{Y} \mid \boldsymbol{\gamma})p(\boldsymbol{\gamma})}, \tag{2}$$

where $p(\mathbf{Y} \mid \boldsymbol{\gamma}) = \int p(\mathbf{Y} \mid \boldsymbol{\theta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}) p(\boldsymbol{\theta}_{\boldsymbol{\gamma}} \mid \boldsymbol{\gamma}) d\boldsymbol{\theta}_{\boldsymbol{\gamma}}$ is the marginal likelihood of $\mathbf{Y}$ under model $\boldsymbol{\gamma}$. 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 jth covariate is:

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

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).

It is known that strong collinearity in the design matrix could make the variance of the ordinary least squares estimates unusually high. As a result the standard t-test statistics may all be insignificant in spite of the corresponding covariates being associated with the response variable. In this article we study a Bayesian analogue of this phenomenon. Note that our goal is not to raise concerns about Bayesian variable selection methods, rather we describe in what ways they are affected by collinearity and how to address such problems in a straightforward manner. In Section 2 we use real data analysis to demonstrate that marginal and joint summaries of the posterior distribution over models may provide conflicting conclusions about the importance of covariates in a high collinearity situation. In Section 3 we illustrate via simulation studies that joint summaries are more likely to be correct than marginal summaries under collinearity. Further, independent normal priors and their scale mixtures generally perform better than Zellner's g-prior (Zellner, 1986) and its mixtures in this context. In Section 4 we provide some theoretical insight into the

problem using the g-prior for the parameters under each model and a discrete uniform prior for the model space. Our results show that collinearity leads to a multimodal posterior distribution which could lead to incorrect assessment of the importance of variables when using marginal inclusion probabilities. A simple solution is to use the joint inclusion probabilities (and joint Bayes factors) that still provide accurate results. In Section 5 we conclude with some suggestions to cope with the problem of collinearity in Bayesian variable selection.

2 Biscuit Dough Data

To motivate the problem studied in this article, we begin with an analysis of the biscuit dough dataset, available as `cookie` in the `R` package `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 *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. An enumerable model space ensures that there is no ambiguity in the results due to approximation. We use the training sample of size 39 in `ppls` for estimation, and the test set of size 31 for prediction.

We use a discrete uniform prior for the model space which assigns equal probability to each of the 2^{15} models and non-informative priors for the intercept β_0 and precision parameter ϕ, given by $p(\beta_0, \phi) \propto 1/\phi$. For the model specific regression coefficients $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$, 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. 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). 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 posterior marginal inclusion probability for each covariate is less than 0.5. The marginal Bayes factor for $\gamma_j = 1$ vs. $\gamma_j = 0$ is the ratio of the posterior odds to the prior odds, $\frac{p(\gamma_j=1|\mathbf{Y})/p(\gamma_j=0|\mathbf{Y})}{p(\gamma_j=1)/p(\gamma_j=0)}$, for $j = 1, \ldots, p$. Because $p(\gamma_j = 1|\mathbf{Y}) < 0.5$ the posterior odds are less than 1, and the prior odds are equal to 1 under an uniform prior, hence the marginal Bayes factors are less than 1. Here the MPM is the null model with only an intercept, and its prediction mean squared error (MSE) is 3.945. As all 15 covariates are correlated we next calculate the Bayes factor $\text{BF}(\text{H}_\text{A} : \text{H}_0)$, 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) 11,073, 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 are different (in magnitude) under different priors, but they unanimously provide strong evidence against H_0. This suggests that it could be worthwhile to consider a model with at least one covariate. Because all the covariates are correlated with each other, the full model is worth an investigation. The MSEs 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. We also look at predictions using the highest probability model (HPM) under each prior. For all priors except the independent t, this is the model with covariate 13 only. For the independent t prior, the HPM includes covariates 1,4,9,12,14. For this prior, the marginal likelihoods are not known exactly making the HPM more difficult to estimate. The MSEs for the HPM under both variants of the g-priors are 2.03, and are 2.03 and 2.01 for the independent priors.

This example shows three main findings. First, marginal and joint summaries for covariates may differ widely when dealing with correlated covariates. Second, the predictive performance of independent priors is more robust to the choice of the model than g-priors under collinearity. For a given model $\boldsymbol{\gamma}$, the posterior mean of the vector of regression coefficients, $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$, under the g-prior (as specified in equation (3) in Section 4) is $\frac{g}{1+g}\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$, where $\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ is the ordinary least squares (OLS)

estimate of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ (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. Third, under collinearity, the HPM could provide better prediction than the MPM. 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 phenomenon can happen in practice. In the following two Sections we try to understand the problem better by using simulation studies and theoretical toy examples.

3 Simulation Studies

Our goal is to compare the performance of marginal and joint summaries of the posterior distribution for different priors under collinearity. We consider the four priors used for the real dataset in the previous Section. The marginal summaries are evaluated via the MPM and the joint summaries using Bayes factors. It is of interest to evaluate whether the covariates in the "true" model can be identified by the different priors and/or estimates. We agree with the Associate Editor that a model cannot be completely "true", however, we think like many authors that studying the performance of different priors and/or estimates under a "true" model may give us insight about their behavior. From now on by important covariates we would refer to covariates with nonzero regression coefficients in the "true" model. One could also define "importance" in terms of predictive ability of the model, and we comment on these issues in more details in the Discussion Section.

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 $\boldsymbol{z}$, 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 $\boldsymbol{z}$. This results in pairwise correlations of about 0.997 among the correlated covariates. The remaining $(p - q)$ covariates are generated independently as $\mathsf{N}(0, 1)$ 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 correlated covariates associated with the response variable, the predictive performance of the model will not be adversely affected. Thus we also consider the cases when the MPM drops the entire group of "true" correlated covariates, when $p(\gamma_j = 1 \mid \mathbf{Y}) < 0.5$ for all q correlated covariates. Results are summarized in Table 1 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 "true" correlated covariates simultaneously, and iv) BF: proportion of datasets in which the Bayes factor $\mathrm{BF}(\mathrm{H_A} : \mathrm{H_0}) \geq 10$, where $\mathrm{H_0}$ is the hypothesis that $\gamma_j = 0$ for all the q correlated covariates and $\mathrm{H_A}$ denotes its complement.

Table 1 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 (Barbieri and Berger, 2004; 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 the correlated covariates. The underperformance of the g-priors could be partly explained by their somewhat inaccurate representation of prior belief in the scenarios under consideration. This issue is discussed in more details in Section 4.

Prior	$q = 2$				$q = 3$				$q = 4$			
	FN	FP	Null	BF	FN	FP	Null	BF	FN	FP	Null	BF
g-prior	0.36	0.06	0.00	0.99	0.78	0.05	0.43	1.00	0.86	0.05	0.51	1.00
Zellner-Siow	0.21	0.08	0.00	0.99	0.77	0.06	0.38	1.00	0.87	0.04	0.54	1.00
Independent normal	0.01	0.06	0.00	0.99	0.14	0.05	0.00	1.00	0.15	0.05	0.00	1.00
Independent t	0.01	0.05	0.00	0.99	0.11	0.05	0.00	0.99	0.14	0.04	0.00	1.00

Table 1: 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 2. 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 correlated covariates that are not associated with the response variable. The Bayes factors based on joint inclusion indicators lead to a correct conclusion 99-100% of the time.

Prior	$q = 2$			$q = 3$			$q = 4$		
	FN	FP	BF	FN	FP	BF	FN	FP	BF
g-prior	0.15	0.03	0.00	0.16	0.02	0.01	0.15	0.03	0.00
Zellner-Siow	0.11	0.07	0.00	0.11	0.06	0.01	0.10	0.07	0.00
Independent normal	0.14	0.08	0.00	0.16	0.03	0.00	0.14	0.00	0.00
Independent t	0.16	0.07	0.00	0.17	0.02	0.00	0.15	0.01	0.00

Table 2: 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 Zellner's g-prior and Collinearity

In the previous simulation studies and real data analysis Zellner's g-prior (Zellner, 1986) has been shown to be most affected. In this Section we explore this prior further with empirical and theoretical toy examples to get a better understanding of its behavior under collinearity. Zellner's g-prior

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 focus on the unit information g-prior with $g = n$, in the presence of strong collinearity. Letting $\mathbf{X}$ denote the design matrix under the full model, we assume that the columns of $\mathbf{X}$ 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 $\mathbf{X}'\mathbf{X}$ is n times the observed correlation matrix of the predictor variables. Under model $\boldsymbol{\gamma}$ the g-prior is given by:

$$\begin{aligned} p(\beta_0, \phi \mid \boldsymbol{\gamma}) &\propto 1/\phi \\ \boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \boldsymbol{\gamma}, \phi &\sim \mathsf{N}\left(\mathbf{0}, \frac{g}{\phi}(\boldsymbol{X}_{\boldsymbol{\gamma}}{}'\boldsymbol{X}_{\boldsymbol{\gamma}})^{-1}\right). \end{aligned} \tag{3}$$

We first explain why the information contained in this prior is in strong disagreement with the data, for the scenarios considered in Section 3. For simplicity of exposition we take a small example with $p = 2$, and denote the sample correlation coefficient between the two covariates by r. For given g and ϕ, the prior variance of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ in the full model $\boldsymbol{\gamma} = (1, 1)'$ is given by

$$\frac{g}{\phi}\left(\mathbf{X}'_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}}\right)^{-1} = \frac{g}{\phi}\left[n\begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}\right]^{-1} = \frac{g}{n\phi(1-r^2)}\begin{pmatrix} 1 & -r \\ -r & 1 \end{pmatrix}.$$

When $r \approx 1$, the prior correlation coefficient between β_1 and β_2 is $-r \approx -1$. Thus the g-prior strongly encourages the coefficients to move in opposite directions when the covariates are strongly positively correlated. Krishna *et al.* (2009) have given similar arguments for not preferring the g-prior in high collinearity situations. An effect of a prior distribution may be better understood by examining the posterior distribution that arises under it, which is studied in the rest of this Section.

Let $\hat{\mathbf{Y}}_{\boldsymbol{\gamma}} = \mathbf{1}\hat{\beta}_0 + \boldsymbol{X}_{\boldsymbol{\gamma}}\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$, where $\hat{\beta}_0 = \bar{Y} = \sum_{i=1}^n Y_i/n$ and $\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}} = (\boldsymbol{X}_{\boldsymbol{\gamma}}{}'\boldsymbol{X}_{\boldsymbol{\gamma}})^{-1}\boldsymbol{X}_{\boldsymbol{\gamma}}{}'\mathbf{Y}$ are the ordinary least squares estimates of β_0 and $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$. Let the regression sum of squares for model $\boldsymbol{\gamma}$ be $\mathrm{SSR}_{\boldsymbol{\gamma}} = \sum_{i=1}^n \left(\hat{\mathbf{Y}}_{\boldsymbol{\gamma}i} - \bar{Y}\right)^2$ and the total sum of squares be $\mathrm{SST} = \sum_{i=1}^n \left(Y_i - \bar{Y}\right)^2$. Then the

coefficient of determination (see for example, Christensen (2002) Section 14.1.1) for model $\boldsymbol{\gamma}$ is $R^2_{\boldsymbol{\gamma}} = \text{SSR}_{\boldsymbol{\gamma}}/\text{SST}$. When $\boldsymbol{\gamma}$ is the null model with only the intercept term, $\hat{\mathbf{Y}}_{\boldsymbol{\gamma}} = \mathbf{1}\bar{Y}$, thus its $\text{SSR}_{\boldsymbol{\gamma}} = 0$ and $R^2_{\boldsymbol{\gamma}} = 0$. The marginal likelihood for the g-prior can be calculated analytically as:

$$p(\mathbf{Y} \mid \boldsymbol{\gamma}) \propto (1+g)^{\frac{n-p_{\boldsymbol{\gamma}}-1}{2}} \{1 + g(1 - R^2_{\boldsymbol{\gamma}})\}^{-\frac{(n-1)}{2}}, \tag{4}$$

where $p_{\boldsymbol{\gamma}} = \sum_{j=1}^{p} \gamma_j$ denotes the number of covariates in model $\boldsymbol{\gamma}$ (excluding the intercept), and the constant of proportionality does not depend on $\boldsymbol{\gamma}$ (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 $p(\boldsymbol{\gamma}) = 1/2^p$ for all models. For exploration of non-enumerable model spaces MCMC may be used such that $p(\boldsymbol{\gamma} \mid \mathbf{Y})$ is the target distribution of the Markov chain. George and McCulloch (1997) discuss fast updating schemes for MCMC sampling with the g-prior.

Next, we consider a small simulation study for $p = 3$ with strong collinearity among the covariates, 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. In later subsections 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, $p(\boldsymbol{\gamma}|\mathbf{Y})$ will converge to 1 in probability, if $\boldsymbol{\gamma} \in \Gamma$ 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 come into effect.

4.1 Simulated data for $p = 3$

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 $\boldsymbol{z}$, 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 $\boldsymbol{z}$. 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 3 and 4.

	$R^2_{\boldsymbol{\gamma}}$					$p(\boldsymbol{\gamma} \mid \mathbf{Y})$				
$\boldsymbol{\gamma}$	$n = 25$	$n = 10^2$	$n = 10^3$	$n = 10^4$	$n = 10^5$	$n = 25$	$n = 10^2$	$n = 10^3$	$n = 10^4$	$n = 10^5$
$(0,0,0)'$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$(0,0,1)'$	0.736	0.582	0.666	0.691	0.691	**0.274**	**0.144**	**0.548**	0.000	0.000
$(0,1,0)'$	0.734	0.589	0.665	0.691	0.691	**0.253**	**0.329**	**0.086**	0.018	0.000
$(0,1,1)'$	0.736	0.590	0.666	0.692	0.692	0.054	0.035	0.026	**0.909**	0.000
$(1,0,0)'$	0.738	0.591	0.665	0.690	0.691	**0.293**	**0.398**	**0.282**	0.000	0.000
$(1,0,1)'$	0.738	0.592	0.666	0.691	0.692	0.058	0.047	0.040	0.000	0.000
$(1,1,0)'$	0.738	0.591	0.666	0.692	0.692	0.057	0.041	0.017	**0.046**	0.000
$(1,1,1)'$	0.738	0.594	0.667	0.692	0.692	0.011	0.006	0.001	**0.026**	**1.000**

Table 3: 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.

Table 3 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. Note that for each sample size n, a single dataset was generated, and the same data generating model was used for all n. The differences in R^2 values for a given model across different sample sizes is due to sampling variability, and it stabilizes to a common value once n starts getting large. For $n \geq 1,000$, one model usually receives a large chunk of the posterior probability, but under repeated sampling there is considerable variability regarding which model gets the large mass.

Table 4 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 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 $(1 - p((0,0,0)' \mid \mathbf{Y})) = 1$. 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.

Even though the HPM is not the "true" model, it may still be effective for predictions in this

high collinearity situation. When the main goal is prediction, whether the "true" model has been selected or not may be irrelevant. However, sometimes it may be of practical interest to find the covariates associated with the response variable, as in a genetic association study. In this case it would be desirable to select the "true" model for a better understanding of the underlying biological process, and both the HPM and the MPM could fail to do so under high collinearity.

	$n = 25$	$n = 10^2$	$n = 10^3$	$n = 10^4$	$n = 10^5$
$p(\gamma_1 = 1 \mid \mathbf{Y})$	0.419	0.492	0.341	0.072	1.000
$p(\gamma_2 = 1 \mid \mathbf{Y})$	0.375	0.411	0.130	1.000	1.000
$p(\gamma_3 = 1 \mid \mathbf{Y})$	0.397	0.232	0.615	0.936	1.000

Table 4: Simulation study for $p = 3$, to demonstrate the effect of collinearity on posterior marginal inclusion probabilities of covariates.

In the following subsections 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.

4.2 Assumptions About R^2 and Collinearity

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

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

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

$$\mathrm{BF}(\boldsymbol{\gamma} = (0, 0, \ldots, 0)' : \boldsymbol{\gamma} = (1, 0, \ldots, 0)') \approx 0, \tag{5}$$

with probability 1, if the specified conditions hold.

For a discrete uniform prior for the model space, that is $p(\boldsymbol{\gamma}) = 1/2^p$ for all models $\boldsymbol{\gamma} \in \Gamma$, the

posterior probability of any model $\boldsymbol{\gamma}$ may be expressed entirely in terms of Bayes factors as:

$$p(\boldsymbol{\gamma} \mid \mathbf{Y}) = \frac{p(\mathbf{Y} \mid \boldsymbol{\gamma})(1/2^p)}{\sum_{\boldsymbol{\gamma} \in \Gamma} p(\mathbf{Y} \mid \boldsymbol{\gamma})(1/2^p)} = \frac{p(\mathbf{Y} \mid \boldsymbol{\gamma})}{\sum_{\boldsymbol{\gamma} \in \Gamma} p(\mathbf{Y} \mid \boldsymbol{\gamma})} = \frac{p(\mathbf{Y} \mid \boldsymbol{\gamma})/p(\mathbf{Y} \mid \boldsymbol{\gamma}^\star)}{\sum_{\boldsymbol{\gamma} \in \Gamma} p(\mathbf{Y} \mid \boldsymbol{\gamma})/p(\mathbf{Y} \mid \boldsymbol{\gamma}^\star)} = \frac{\mathrm{BF}(\boldsymbol{\gamma} : \boldsymbol{\gamma}^\star)}{\sum_{\boldsymbol{\gamma} \in \Gamma} \mathrm{BF}(\boldsymbol{\gamma} : \boldsymbol{\gamma}^\star)}, \tag{6}$$

where $\boldsymbol{\gamma}^\star \in \Gamma$ (Berger and Molina, 2005). Taking $\boldsymbol{\gamma} = (0, 0, \ldots, 0)'$ and $\boldsymbol{\gamma}^\star = (1, 0, \ldots, 0)'$ in (6), and using (5), for large enough n we have the following with probability 1,

$$p(\boldsymbol{\gamma} = (0, 0, \ldots, 0)' \mid \mathbf{Y}) \approx 0. \tag{7}$$

As the null model receives negligible posterior probability we may omit it when computing the normalizing constant of $p(\boldsymbol{\gamma} \mid \mathbf{Y})$, 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 Γ. We provide a formal justification of this approximation in Appendix B.

We now make an assumption about strong collinearity among the covariates, so that $R^2_{\boldsymbol{\gamma}}$ for all non-null models $\boldsymbol{\gamma} \in \Gamma - \{(0, 0, \ldots, 0)'\}$ 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 $\left\{ \frac{1+n(1-R^2_{\boldsymbol{\gamma}})}{1+n(1-R^2_{\boldsymbol{\gamma}'})} \right\}^{-\frac{(n-1)}{2}}$ can be taken to be approximately 1, for any pair of distinct non-null models $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}'$, 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 $n \to \infty$. One would usually expect a group or multiple groups of correlated covariates to occur, instead of all p of them being highly correlated. This simplified assumption is made for exploring the behavior theoretically, but the phenomenon holds under more general conditions. This has been already demonstrated in the simulation studies in Section 3, where a subset (of varying size) of the p covariates was assumed to be correlated rather than all of them. Our empirical results suggest that this assumption will usually not hold when the correlations are smaller than 0.9 or so. Thus it will probably not occur frequently, but cannot be ruled out either, as evident from the real data analysis in Section 2. 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 subsection.

4.3 Collinearity Example for $p = 2$

Under Assumptions 1 and 2 and the discrete uniform prior for the model space, $p(\boldsymbol{\gamma}) = \frac{1}{2^2},\ \boldsymbol{\gamma} \in \Gamma$, the posterior probabilities of the 2^2 models can be approximated as follows, with probability 1:

$$\begin{aligned} p(\boldsymbol{\gamma} = (0,0)' \mid \mathbf{Y}) &\approx 0, \qquad & p(\boldsymbol{\gamma} = (0,1)' \mid \mathbf{Y}) &\approx \frac{1}{2}, \\ p(\boldsymbol{\gamma} = (1,0)' \mid \mathbf{Y}) &\approx \frac{1}{2}, \qquad & p(\boldsymbol{\gamma} = (1,1)' \mid \mathbf{Y}) &\approx 0. \end{aligned} \tag{8}$$

The detailed calculations are given in Appendix C. The calculations in (8) have the following implications with probability 1. First, the posterior is approximately bimodal with modes at the one-variable models, $\boldsymbol{\gamma} = (0,1)'$ and $\boldsymbol{\gamma} = (1,0)'$. Second, as each mode receives approximately 0.5 posterior probability, this implies that the marginal posterior inclusion probabilities $p(\gamma_1 = 1 \mid \mathbf{Y})$ and $p(\gamma_2 = 1 \mid \mathbf{Y})$ 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 $p(\gamma_1 = 1 \text{ or } \gamma_2 = 1) = 1 - p(\boldsymbol{\gamma} = (0,0)') = 1 - (1/2)^2 = 3/4$, under the discrete uniform prior for the model space. The posterior probability that at least one of the important covariates is included is $P(\gamma_1 = 1 \text{ or } \gamma_2 = 1 \mid \mathbf{Y}) = 1 - p(\boldsymbol{\gamma} = (0,0)' \mid \mathbf{Y}) \approx 1$, by (8). Let $\mathrm{H_0}$ denote $\boldsymbol{\gamma} = (0,0)'$ and $\mathrm{H_A}$ denote its complement, that is $\mathrm{H_A}$ denotes $\gamma_1 = 1$ or $\gamma_2 = 1$. Then the prior odds $P(\mathrm{H_A})/\mathrm{P}(\mathrm{H_0}) = (3/4)/(1/4) = 3$ and the posterior odds $P(\mathrm{H_A} \mid \mathbf{Y})/\mathrm{P}(\mathrm{H_0} \mid \mathbf{Y})$ is expected to be very large, because $P(\mathrm{H_A} \mid \mathbf{Y}) \approx 1$ and $P(\mathrm{H_0} \mid \mathbf{Y}) \approx 0$, by (8). Then it is clear from the calculations of the prior and posterior odds that the Bayes factor $\mathrm{BF}(\mathrm{H_A} : \mathrm{H_0}) = \frac{\mathrm{P}(\mathrm{H_A}|\mathbf{Y})/\mathrm{P}(\mathrm{H_0}|\mathbf{Y})}{\mathrm{P}(\mathrm{H_A})/\mathrm{P}(\mathrm{H_0})}$ will be very large with probability 1, under the above assumptions.

4.4 Collinearity Example for General p

Consider a similar set up with p highly correlated covariates and $\boldsymbol{\gamma} = (1, 1, \ldots, 1)'$ as the "true" model. Under Assumptions 1 and 2 the following results hold with probability 1, which is implicitly assumed throughout this Section. For large n, under Assumption 1, the null model has nearly zero posterior probability by (7), so it is not considered in the calculation of the normalizing constant for posterior probabilities of models as before. Under Assumption 2, taking $g = n$ in (4), all $(2^p - 1)$ non-null models have the term $\{1 + n(1 - R^2_{\boldsymbol{\gamma}})\}^{-\frac{(n-1)}{2}}$ (approximately) in common. Ignoring common terms the marginal likelihood for any model of dimension $p_{\boldsymbol{\gamma}}$ is approximately proportional to $(1+n)^{\frac{n-p_{\boldsymbol{\gamma}}-1}{2}}$. Given n, this term decreases as $p_{\boldsymbol{\gamma}}$ increases, so the models with $p_{\boldsymbol{\gamma}} = 1$ will have the highest posterior probability, and the posterior will have p modes at each of the one-dimensional models. The posterior inclusion probability for the jth covariate is

$$p(\gamma_j = 1 \mid \mathbf{Y}) = \frac{\sum_{\boldsymbol{\gamma}\in\Gamma:\gamma_j=1} p(\mathbf{Y} \mid \boldsymbol{\gamma})}{\sum_{\boldsymbol{\gamma}\in\Gamma} p(\mathbf{Y} \mid \boldsymbol{\gamma})} \approx \frac{\sum_{\boldsymbol{\gamma}\in\Gamma:\gamma_j=1} p(\mathbf{Y} \mid \boldsymbol{\gamma})}{\sum_{\boldsymbol{\gamma}\in\Gamma-\{(0,0,\ldots,0)'\}} p(\mathbf{Y} \mid \boldsymbol{\gamma})} \approx \frac{\sum_{p_{\boldsymbol{\gamma}}=1}^{p} \binom{p-1}{p_{\boldsymbol{\gamma}}-1}(1+n)^{\frac{n-p_{\boldsymbol{\gamma}}-1}{2}}}{\sum_{p_{\boldsymbol{\gamma}}=1}^{p} \binom{p}{p_{\boldsymbol{\gamma}}}(1+n)^{\frac{n-p_{\boldsymbol{\gamma}}-1}{2}}}, \quad (9)$$

where the last approximation is due to Assumption 2 regarding collinearity. The expression in (9) follows the fact that i) all $p_{\boldsymbol{\gamma}}$-dimensional models have the marginal likelihood proportional to $(1+n)^{\frac{n-p_{\boldsymbol{\gamma}}-1}{2}}$ approximately (using Assumption 2 in (4)), ii) there are altogether $\binom{p}{p_{\boldsymbol{\gamma}}}$ such models, and iii) exactly $\binom{p-1}{p_{\boldsymbol{\gamma}}-1}$ of these have $\gamma_j = 1$. Dividing the numerator and denominator of (9) by $(1+n)^{\frac{n-2}{2}}$ we have

$$p(\gamma_j = 1 \mid \mathbf{Y}) \approx \frac{1 + \sum_{p_{\boldsymbol{\gamma}}=2}^{p} \binom{p-1}{p_{\boldsymbol{\gamma}}-1}(1+n)^{-\frac{(p_{\boldsymbol{\gamma}}-1)}{2}}}{p + \sum_{p_{\boldsymbol{\gamma}}=2}^{p} \binom{p}{p_{\boldsymbol{\gamma}}}(1+n)^{-\frac{(p_{\boldsymbol{\gamma}}-1)}{2}}} \approx \frac{1}{p},$$

where the last approximation follows for fixed p and sufficiently large n, as the terms in the sum over $p_{\boldsymbol{\gamma}}$ (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 $p = 2$.

Let $\mathrm{H}_0 : \boldsymbol{\gamma} = (0, 0, \ldots, 0)'$ and H_A: complement of H_0. Because the prior odds $P(\mathrm{H}_\mathrm{A})/P(\mathrm{H}_0) = (2^p - 1)$ is fixed (for fixed p), and the posterior odds is large for sufficiently large n, the Bayes factor $\mathrm{BF}(\mathrm{H}_\mathrm{A} : \mathrm{H}_0)$ will be large. This useful result suggests that while marginal inclusion probabilities

(marginal Bayes factors) may give misleading conclusions about the importance of covariates, the joint inclusion probabilities (joint Bayes factors) would correctly indicate that at least one of the covariates should be included in the model. These results are in agreement with the simulation studies in Section 3 and provide a theoretical justification for them.

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 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.

An interesting question was raised by the reviewer: should we label all the covariates appearing in the "true" model as important even in cases of high collinearity. The definition of important covariates largely depends on the goal of the study. For example, in genetic association studies there could be some highly correlated genetic markers, all associated with the response variable, and the goal of the study is often identifying such markers. In this case they would all be deemed important. In recent years statisticians have focused on this aspect of variable selection with correlated covariates, where it is desired that correlated covariates are to be simultaneously included in (or excluded from) a model as a group. The elastic net by Zou and Hastie (2005) is a regularization method with such grouping effect. Bayesians have formulated priors that will induce the grouping effect (Krishna *et al.*, 2009; Liu *et al.*, 2014). In some of these papers the authors have shown that

including correlated covariates in a group with appropriate regularization or shrinkage rules may improve predictions.

If the goal is to uncover the model with best predictive performance, then including highly correlated covariates simultaneously in the model may not necessarily lead to the best predictive model. The MPM is the optimal predictive model under squared error loss, and certain conditions. For the optimality conditions to be satisfied, the design matrix has to be orthogonal in the all submodels scenario, and certain types of priors must be used. In general, the MPM does quite well under non-orthogonality too, but may not do as well under high collinearity. One possibility would be to find the model with best predictive ability, measured by expected squared error loss, with expectation taken with respect to the posterior predictive distribution (see for example, Lemma 1 of Barbieri and Berger (2004)). This would be feasible for conjugate priors and small model spaces that can be enumerated. For large model spaces one could use the same principle to find the best model among the set of sampled models. However, for general priors, when the posterior means of the regression coefficients are not available in closed form, the problem would become computationally challenging. Thus for genuine applications, it would be good practice to report out of sample predictive performance of both the HPM and the MPM. When finding the best predictive model in the list of all/sampled models is computationally feasible it would be desirable to report it as well.

Acknowledgment

The authors thank the Editor, the Associate Editor, and one reviewer for many useful suggestions that led to a much improved paper. The research of Joyee Ghosh was partially supported by the NSA Young Investigator grant H98230-14-1-0126.

Appendix A: Proof of Proposition 1

Proof. To simplify the notation let $R^2_{\boldsymbol{\gamma}} = R^2$ for $\boldsymbol{\gamma} = (1, 0, \ldots, 0)'$. Then putting $g = n$ and using the expression for marginal likelihood of the g-prior given in (4) we have,

$$\begin{aligned}
\mathrm{BF}(\boldsymbol{\gamma} = (0,0,\ldots,0)' : \boldsymbol{\gamma} = (1,0,\ldots,0)') &= p(\mathbf{Y} \mid \boldsymbol{\gamma} = (0,0,\ldots,0)')/p(\mathbf{Y} \mid \boldsymbol{\gamma} = (1,0,\ldots,0)') \\
&= \frac{1}{(1+n)^{(n-2)/2}\{1+n(1-R^2)\}^{-(n-1)/2}} \\
&= \frac{1}{(1+n)^{(n-2)/2}\left[\frac{\{1+n(1-R^2)\}(1+n)}{(1+n)}\right]^{-(n-1)/2}} \\
&= \frac{1}{(1+n)^{(n-2)/2}\left(\frac{1+n-nR^2}{1+n}\right)^{-(n-1)/2}(1+n)^{-(n-1)/2}} \\
&= \frac{1}{(1+n)^{(n-2-n+1)/2}\left(1-\frac{n}{1+n}R^2\right)^{-(n-1)/2}} \\
&= \frac{1}{(1+n)^{-1/2}\left(1-\frac{n}{1+n}R^2\right)^{-(n-1)/2}} \\
&= (1+n)^{1/2}\left(1-\frac{n}{1+n}R^2\right)^{(n-1)/2}
\end{aligned} \tag{10}$$

Taking the logarithm of (10) the following result holds with probability 1, by Assumption 1:

$$\begin{aligned}
\log(\mathrm{BF}(\boldsymbol{\gamma} = (0,0,\ldots,0)' : \boldsymbol{\gamma} = (1,0,\ldots,0)')) &= \log\left((1+n)^{1/2}\left(1-\frac{n}{1+n}R^2\right)^{(n-1)/2}\right) \\
&= \frac{1}{2}\log(1+n) + \frac{(n-1)}{2}\log(1-\frac{n}{n+1}R^2) \\
&< \frac{1}{2}\log(1+n) + \frac{(n-1)}{2}\log(1-\frac{n}{n+1}\delta_1)
\end{aligned} \tag{11}$$

As n goes to infinity, the first term in (11) goes to ∞ 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 (11) goes to $-\infty$ at a polynomial

rate in n, of degree 1. Thus, as $n \to \infty$, with probability 1,

$$\begin{aligned} \log(\text{BF}(\boldsymbol{\gamma} = (0,0,\ldots,0)' : \boldsymbol{\gamma} = (1,0,\ldots,0)')) &\to -\infty, \text{ or} \\ \text{BF}(\boldsymbol{\gamma} = (0,0,\ldots,0)' : \boldsymbol{\gamma} = (1,0,\ldots,0)') &\to 0. \end{aligned} \tag{12}$$

From (12) it follows that for sufficiently large n, we can make $\text{BF}(\boldsymbol{\gamma} = (0,0,\ldots,0)' : \boldsymbol{\gamma} = (1,0,\ldots,0)') < \epsilon$, for given $\epsilon > 0$, with probability 1. This completes the proof. □

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 $\sum_{\boldsymbol{\gamma} \in \Gamma} p(\mathbf{Y} \mid \boldsymbol{\gamma})$

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 $\frac{\sum_{i=2}^{m} a_{in}}{\sum_{i=1}^{m} a_{in}} \to 1$.*

Proof.

$$\begin{aligned} \lim_{n\to\infty} \frac{\sum_{i=2}^{m} a_{in}}{\sum_{i=1}^{m} a_{in}} &= \lim_{n\to\infty} \frac{\sum_{i=2}^{m} a_{in}/a_{2n}}{\sum_{i=1}^{m} a_{in}/a_{2n}} \\ &= \lim_{n\to\infty} \frac{\sum_{i=2}^{m} a_{in}/a_{2n}}{(a_{1n}/a_{2n}) + \sum_{i=2}^{m} a_{in}/a_{2n}} \\ &= \frac{\lim_{n\to\infty} \sum_{i=2}^{m} a_{in}/a_{2n}}{\lim_{n\to\infty}(a_{1n}/a_{2n}) + \lim_{n\to\infty} \sum_{i=2}^{m} a_{in}/a_{2n}} \\ &= \frac{\lim_{n\to\infty} \sum_{i=2}^{m} a_{in}/a_{2n}}{0 + \lim_{n\to\infty} \sum_{i=2}^{m} a_{in}/a_{2n}} \\ &= 1 \end{aligned}$$

□

Corollary 1. *If Assumption 1 holds, then given $\eta > 0$, however small, for sufficiently large n, we can make $\left(1 - \frac{\sum_{\boldsymbol{\gamma} \in \Gamma - \{(0,0,\ldots,0)'\}} p(\mathbf{Y}|\boldsymbol{\gamma})}{\sum_{\boldsymbol{\gamma} \in \Gamma} p(\mathbf{Y}|\boldsymbol{\gamma})}\right) < \eta$, with probability 1.*

Proof. We have $p(\mathbf{Y} \mid \boldsymbol{\gamma}) > 0,\ \boldsymbol{\gamma} \in \Gamma$ and $\frac{p(\mathbf{Y}|\boldsymbol{\gamma}=(0,0,\ldots,0)')}{p(\mathbf{Y}|\boldsymbol{\gamma}=(1,0,\ldots,0)')} \to 0$ as $n \to \infty$, with probability 1, by (12). Then as $n \to \infty$ we have the following, with probability 1, by Lemma 1:

$$\frac{\sum_{\boldsymbol{\gamma} \in \Gamma - \{(0,0,\ldots,0)'\}} p(\mathbf{Y} \mid \boldsymbol{\gamma})}{\sum_{\boldsymbol{\gamma} \in \Gamma} p(\mathbf{Y} \mid \boldsymbol{\gamma})} \quad \to \quad 1.$$

The proof follows immediately. □

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 (7). We derive the posterior probabilities of the non-null models under Assumptions 1 and 2 here. For any non-null model $\boldsymbol{\gamma} \in \Gamma - \{(0,0)'\}$,

$$\begin{aligned}
p(\boldsymbol{\gamma} \mid \mathbf{Y}) &= \frac{p(\boldsymbol{\gamma}) p(\mathbf{Y} \mid \boldsymbol{\gamma})}{\sum_{\boldsymbol{\gamma} \in \Gamma} p(\boldsymbol{\gamma}) p(\mathbf{Y} \mid \boldsymbol{\gamma})} \\
&= \frac{(1/2^2) p(\mathbf{Y} \mid \boldsymbol{\gamma})}{\sum_{\boldsymbol{\gamma} \in \Gamma} (1/2^2) p(\mathbf{Y} \mid \boldsymbol{\gamma})} \text{ (because } p(\boldsymbol{\gamma}) = 1/2^2 \text{ for } \boldsymbol{\gamma} \in \Gamma) \\
&= \frac{p(\mathbf{Y} \mid \boldsymbol{\gamma})}{\sum_{\boldsymbol{\gamma} \in \Gamma} p(\mathbf{Y} \mid \boldsymbol{\gamma})} \\
&\approx \frac{p(\mathbf{Y} \mid \boldsymbol{\gamma})}{\sum_{\boldsymbol{\gamma} \in \Gamma - \{(0,0)'\}} p(\mathbf{Y} \mid \boldsymbol{\gamma})}, \qquad (13)
\end{aligned}$$

with probability 1. The last approximation in (13) follows from Corollary 1 in Appendix B.

We will use the expression in (13) to derive the posterior probabilities. First note that under Assumption 2 the term $\{1 + n(1 - R^2_{\boldsymbol{\gamma}})\}^{-\frac{(n-1)}{2}}$ in the expression of marginal likelihood $p(\mathbf{Y} \mid \boldsymbol{\gamma})$ in (4) 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 (13). Then by

(4), (13), and substituting $g = n$ we have with probability 1,

$$p(\boldsymbol{\gamma} = (0,1)' \mid \mathbf{Y}) \approx \frac{(1+n)^{(n-1-1)/2}}{(1+n)^{(n-1-1)/2} + (1+n)^{(n-1-1)/2} + (1+n)^{(n-2-1)/2}}. \quad (14)$$

Dividing the numerator and denominator of the right hand side of (14) by $(1+n)^{(n-2)/2}$,

$$\begin{aligned} p(\boldsymbol{\gamma} = (0,1)' \mid \mathbf{Y}) &\approx \frac{1}{2 + (1+n)^{-1/2}} \\ &\approx \frac{1}{2}, \end{aligned}$$

for large enough n, with probability 1.

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

$$p(\boldsymbol{\gamma} = (1,0)' \mid \mathbf{Y}) \approx \frac{1}{2},$$

for large enough n, with probability 1.

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

$$\begin{aligned} p(\boldsymbol{\gamma} = (1,1)' \mid \mathbf{Y}) &\approx \frac{(1+n)^{(n-2-1)/2}}{(1+n)^{(n-1-1)/2} + (1+n)^{(n-1-1)/2} + (1+n)^{(n-2-1)/2}} \\ &\approx \frac{1}{2(1+n)^{1/2} + 1} \\ &\approx 0, \end{aligned}$$

for large enough n, with probability 1.

References

Barbieri, M. and Berger, J. (2004). Optimal predictive model selection. *Annals of Statistics* **32**, 3, 870–897.

Berger, J. O. and Molina, G. (2005). Posterior model probabilities via path-based pairwise priors.

Statistica Neerlandica **59**, 3–15.

Brown, P. J., Fearn, F., and Vannucci, M. (2001). Bayesian wavelet regression on curves with application to a spectroscopic calibration problem. *Journal of the American Statistical Association* **96**, 454, 398–408.

Christensen, R. (2002). *Plane Answers to Complex Questions*. Springer, 3rd edn.

Clyde, M. A., Ghosh, J., and Littman, M. L. (2011). Bayesian adaptive sampling for variable selection and model averaging. *Journal of Computational and Graphical Statistics* **20**, 1, 80–101.

Fernández, C., Ley, E., and Steel, M. F. (2001). Benchmark priors for Bayesian model averaging. *Journal of Econometrics* **100**, 381–427.

George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica* **7**, 339–374.

Ghosh, J. and Clyde, M. A. (2011). Rao-Blackwellization for Bayesian variable selection and model averaging in linear and binary regression: A novel data augmentation approach. *Journal of the American Statistical Association* **106**, 495, 1041–1052.

Ghosh, J. and Reiter, J. P. (2013). Secure Bayesian model averaging for horizontally partitioned data. *Statistics and Computing* **23**, 311–322.

Kraemer, N. and Boulesteix, A.-L. (2012). *ppls: Penalized Partial Least Squares*. R package version 1.05.

Krishna, A., Bondell, H. D., and Ghosh, S. K. (2009). Bayesian variable selection using an adaptive powered correlation prior. *Journal of Statistical Planning and Inference* **139**, 8, 2665–2674.

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

Liu, F., Chakraborty, S., Li, F., Liu, Y., and Lozano, A. C. (2014). Bayesian regularization via graph laplacian. *Bayesian Analysis* **9**, 2, 449–474.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, 233–243. North-Holland/Elsevier.

Zellner, A. and Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. In *Bayesian Statistics: Proceedings of the First International Meeting held in Valencia (Spain)*, 585–603.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B* **67**, 2, 301–320.