Resrarch Journal of Applied Sciences 11 (7): 428-438, 2016
ISSN: 1815-932X

Bayesian Variable Selection under Collinearity of Parameters

Samaneh Mohammad Ijarchelo, Khosro Afereydoon and Leyla Zamanzadeh
Department of Mathematical Statistics, Payame Noor University, Tabriz, Iran

Abstract: In this study, 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 osterior inclusion probabilities and Bayes factors 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 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 appropriate 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, median probability model, multimodality, Zellner's g-prior

INTRODUCTION

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

$$Y|\beta_{0,}\ \beta_{\gamma,}\varphi,\ \gamma?N(1\beta_0+X_\gamma\beta_{\gamma,}In/\varphi) \quad (1)$$

Where:

- I = An n×1 vector of ones
- β_0 = The intercept
- X_γ = The $n\times p_\gamma$ design matrix
- β_γ = The $p_\gamma\times1$ vector of regression coefficients under model γ
- φ = The reciprocal of the error variance
- I_n = An n×n identity matrix

The intercept is assumed to be included in every model. The models in Eq. 1 areassigned a prior distribution $p(\gamma)$ and the vector of parameter sunder each model γ is assigned a prior distribution $p(\theta_\gamma|\gamma)$ where $\theta_\gamma = (\beta_0, \beta_\gamma, \varphi)$. The posterior probability of any modelis obtained using Bayes' rule as:

$$P(\gamma|Y)=p(Y|\gamma)p(\gamma)\div 1_{(\gamma\in\Gamma)}|(p(Y|\gamma)p(\gamma)) \quad (2)$$

where, $p(Y|\gamma)=\int p(Y|\theta y,\gamma)p(\theta y|y)d\theta y$ is the marginal distribution of Y under model γ. This is also referred to as the marginal likelihood of the model γ. We will consider scenarios when the marginal likelihood may or may not exist 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 theoretical arguments. When p is larger than 25-30, the posterior probabilities in Eq. 2 are not available for general design matrices due to computation allimitations, 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 samplesize 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 HPMa challenging task.

Moreover, for large model spaces, the HPM may have a very small posterior probability, so it is not clear if

Corresponding Author: Samaneh Mohammad Ijarchelo, Department of Mathematical Statistics, Payame Noor University, Tabriz, Iran

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 $p(\gamma j = 1 \mid Y) = ?$. The use of these can be further motivated by the Median Probability Model (MPM) by Barbieri and Berger (2004). The MPMincludes all variables whose posterior marginal inclusion probabilities are ≥ 0.5. Instead of selectinga single best model, another option is to consider a weight edaverage 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). For the optimality conditions to be satisfied, the columns of the design matrixneed to be orthogonal in the all submodels scenario and the priors must also satisfy some conditions. Independent conjugate normal priors belong to the class of priors that satisfies these conditions. Barbieri and Berger (2004) suggested that in practicethe MPM often outperforms the HPM even if the condition of orthogonality is not satisfied. It is known that strong collinearity in the design matrix could make the variance of the ordinary least-square estimates unusually high. As a result the standard t-test statistics may all be in significant in spite of the corresponding covariates being associated with the response variable. In this study, we study a Bayesian analog of this phenomenon. Note that our goal is not toraise 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 straight forward manner. Further, independent normal priors generally perform better than Zellner's g-prior (Zellner, 1986) and its mixtures in this context. We provide some theoretical insight into the problemusing 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) which still provide accurate results. We conclude with some suggestions to cope with the problem of collinearity in Bayesian variable selection.

Biscuit dough data: To motivate the problem studied in this study, we begin withan analysis of the biscuit dough dataset, available as cookie in the R package. The dataset was obtained from an experiment that used Near-In Frared (NIR) spectroscopy to analyze the composition of biscuit dough pieces. The experiment of Osborne *et al.* (1984) investigated whether NIR spectroscopy could be used for automatic quality control in the biscuit baking industry. Compared to classical chemical methods these are nondestructive and fast. Hence, the method could potentially be used for automatic online control. An NIR reflectance spectrum for each dough is a continuous curve measured at many equally spaced wavelengths. The goal of the experiment was to extract the information contained inthis curve to predict the chemical composition of the dough.

The package contains the training and test samples used in the original experiment by Osborne *et al.* (1984) where 39 samples were used for calibration and 31 samples made with a similar recipe were used for prediction. We use the same training and test data. Osborne *et al.* (1984) concluded that the method was capable of predicting the fat content in the biscuit doughs sufficiently accurately. Brown *et al.* (2001) omitted the first 140 and last 49 of the available 700 wavelengths to reduce the computational burden because these were thought to contain little information. For our analysis, we choose the wavelengths 191-205 to have $p = 15$ covariates with high pairwise correlations (around 0.999) among all of them and the percentage of fat as the response variable. Considering all possible subsets of the full model, this results in a model space of 2^{15} models. The model space is small enough that all posterior probabilities (and thus, posterior inclusion probabilities) can be calculated exactly or approximately by a Laplace approximation. This ensuresthat there is no ambiguity in the results due to Monte Carlo approximation.

We use a discrete uniform prior for the model space, which assigns equal probability to each of the 2^{15} models and diffuse priors for the intercept β_0 and precision parameter φ, given by $p(\beta_0, \varphi) \propto 1/\varphi$. For the model specific regression coefficients β_γ, we consider the multivariate normal g-prior (Zellner, 1986) with $g = n$, the multivariate Zellner-Siow (Zellner and Siow, 1980) Cauchy prior and independent normal priors.

The marginal likelihood for the g-prior is given in Eq. 5. For the Zellner-Siow prior, marginal likelihoods are approximated by a Laplace approximation for a one-dimensional integral over g, for example, Appendix 1 by Liang *et al.* (2012) for more details. The posterior computation for the different versions of g-priors is done by enumerating all 2^{15} models with the BAS algorithm (Clyde *et al.*, 2012). For the independent normal priors, we use the same hyper parameters as Ghosh and Clyde (2012). We enumerate all models and calculate the marginal likelihoods using Eq. 4 by Ghosh and Clyde (2012) for posterior computation. Suppose, it is of interest to predict a set of future response variables Y_f, at a set of covariates X_f, from the same process that generated the observed data Y. We use the mean of the Bayesian predictive distribution $p(Y_f|\gamma,Y)$, for a given model γ where:

$$p(Y_f|\gamma,Y) \tag{3}$$

The mean of the above distribution is of the form $1\tilde{\beta}_0+X_{f\gamma}\tilde{\beta}_\gamma$ where $\tilde{\beta}_0$ and $\tilde{\beta}_\gamma$ are the posterior means of β_0 and β_γ under the model γ. Liang *et al.* (2012), Ghosh and Clyde (2012). For every prior considered, the posterior marginal inclusion probability for each covariate is <0.5. The marginal Bayes factor for $Y_j = 1$ versus $Y_j = 0$ is the ratio of the posterior odds to the prior odds:

$$\frac{p(\gamma j=1|Y)/p(\gamma j=0|Y)}{p(\gamma j=1)/p(\gamma j=0)}$$

for j = 1, ..., p. Because $p(\gamma_j = 1|Y) < 0.5$ the posterior odds are <1 and the prior odds are equal to 1 under a uniform prior, hence the marginal Bayes factors are <1. Here, the MPM is the null model with only an intercept. The predicted values are calculated using Eq. 3 and the Prediction Mean Squared Error (PMSE) is 3.95.

As all 15 covariates are correlated, we next calculate the Bayes factor $BF(H_A: H_0)$, where H_0 is the model with only the intercept and HA denotes its complement. For the existence of the marginal likelihood under the null model with only an intercept, we need the sample size to be at least two and the sample variance of the response variable to be strictly positive that is the values of the response variable cannot be all equal. These conditions are satisfied in this example and would usually hold for continuous response variables. The Bayes factors are 114, 69 and 11.073, for the g-prior, Zellner-Siow prior and independent normal 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 worth while 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 PMSEs for the full model under the three priors are 4.55, 3.98 and 2.00, respectively. We also consider prediction using the Highest Probability Model (HPM) under each prior; for all priors the HPM only selects covariate 13. The PMSEs for the HPM under all priors are 2.03.

This example illustrates three main points. First, for a group of highly correlated covariates, the marginal posterior inclusion probabilities for all of them may be low even when the joint posterior inclusion probability that at least one of them is includedis very high. Second, the predictive performance using g-priors could be more adversely affected than using independent priors when highly correlated covariates are included in the model. Third, if one has to select a single model for prediction, the HPM could provide better prediction than the MPM under collinearity because unlike the MPM, the HPM does not discard the entire set of correlated covariates associated with the response variable. Note that a different choice of wavelengths as covariates may not lead to selection of the null model as was the case here using the MPM but our goal is to illustrate that this phenomenon can happen in practice. For a given model γ, the posterior mean of the vector of regression coefficients, β_γ, under the g-prior (as specified in Eq. 4 is $g/1+g\tilde{\beta}\gamma$ where $\tilde{\beta}_\gamma$ is the Ordinary Least-Square (OLS) estimate of β_γ (Liang *et al.*, 2012; 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 estimate under the independent normal priors is a ridge regression estimate (Ghosh and Clyde, 2012) which is known to be more stable under collinearity. In the following two sections, we try to underst and the problem better by using simulation studies and theoretical examples.

MATERIALS AND METHODS

Simulation studies: Our goal is to compare the performance of marginal and joint summaries of the posterior distribution for different priors under collinearity. It is of interest to evaluate whether the covariates in the "true" model can be identified by using different priors and/or estimates. We agree with the associate editor that amodel cannot be completely "true," however, we think like many researchers that studying the performance of procedures based ondifferent priors and/or estimates under a "true" model may giveus 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.

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 Eq. 1 with $\varphi = 1/4$ and the procedure is repeated to generate 100 datasets. For all priors, the model space of $2^{10} = 1024$ models is 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 leastone 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 $p(\gamma j = 1 \mid Y) < 0.5$ 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 marginal inclusion probabilities that are used to determine the MPM.They are defined as follows:

FNR: False negative rate defined as $\sum_{i=1}^{m} \text{FNRi}/m$ where mis the number of simulated datasets and FNRi is the numberof true covariates that are not selected by the MPM for theith dataset, divided by the total number of true covariates (q).

FPR: False positive rate defined as:

$$\sum_{i=1}^{m} \frac{\text{FNRi}}{m}$$

where, FPRi is the number of noise covariates that are selected by the MPM for the ith dataset, divided by the total number ofnoise covariates (p-q = 10-q).

Null: Proportion of datasets in which the MPM discards all "true" correlated covariates simultaneously.

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 1 shows that the false negative rate is much higher for q>2 than q = 2. With q = 4, this rate is higher than 80% for the g-priors and >10% for the independent priors. The false positive rate is generally lowand 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 (George and McCulloch, 1997; Barbieri and Berger, 2004; Berger and Molina, 2005) 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 under performance of the estimates based on the g-priors could bepartly explained by their some what in accurate representation of prior belief in the scenarios under consideration.

Unimportant correlated covariates: In this simulation study, we consider the same values of n, p and q as before. We now set the regression coefficients for the q correlated covariates at zero and the remaining (p-q) coefficients at one. We generate the covariates and the response variable. The results based on repeating the procedure 100 times are presented in Table 2. The false negative rates are similar across (Table 2).

Table 1: Simulation study with p = 10 covariates of which q correlated covariates are included in the "true" model as signals and (pq) uncorrelated covariates denote noise

	q = 2				q = 3				q = 4			
Prior	FNR	FPR	Null	BF	FNR	FPR	Null	BF	FNR	FPR	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

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 assignals

	q = 2			q = 3			q = 4		
Prior	FNR	FPR	BF	FNR	FPR	BF	FNR	FPR	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

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 tohave any problems in discarding correlated covariates that arenot associated with the response variable. The Bayes factors based on joint inclusion indicators lead to a correct conclusion 99-100% of the time.

Zellner's g-prior and collinearity: In the previous simulation studies and real data analysis, Zellner'sg-prior (Zellner, 1986) has been shown to be most affected. In this study, we explore this prior further with empirical and the oretical 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 ofchoices have been proposed in the literature. In this study, we 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 columnis $\sqrt{n}$, as in Ghosh and Clyde (2012). For the standardizeddesign matrix, $X^{t}X$ is n times the observed correlation matrix ofthe predictor variables. Under model γ, the g-prior is given by:

$$p(\beta_0,\varphi|\gamma)\,\alpha 1/\varphi \quad (4)$$

We first explain why the information contained in this prioris in strong disagreement with the data, for the scenarios considered previously. 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 $\beta\gamma$ in the full model $\gamma = (1, 1)$ is given by:

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

When r = 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) gave similar arguments for not preferring the g-prior inhigh collinearity situations. An effect of a prior distribution may be better understood by examining the posterior distribution that arises under it. Now let $Y\hat{\gamma}=1\beta_{\hat{0}}+X\gamma\beta_{\gamma}$ where:

$$\hat{\beta}_0=\bar{Y}=\sum_{i=1}^{n}Y_i/n$$

and;

$$\hat{\beta}_{\gamma}=(X^{1}{}_{\gamma}X_{\gamma})^{-1}X^{1}{}_{\gamma}Y$$

are the ordinary least-square estimates of β_0 and β_γ. Let the regression sum of squares for model γ be $SSR_{\gamma}=\sum_{i=1}^{n}(\hat{Y}_{\gamma i}\bar{Y})^2$ and the total sum of squares be $SSR_{\gamma}=\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2$. Then, the coefficient of determination (Christensen, 2002). For model γ is R^2_{γ} = SSR_γ/SST. When, γ is the null model with only the intercept term, $Y\gamma$ = 1Y, thus its SSR_γ = 0 and R^2_γ = 0, in this special case. The marginal likelihood forthe g-prior can be calculated analytically as Eq. 5, where:

$$p_{\gamma}=\sum_{j=1}^{p}\gamma j \quad (5)$$

denotes the number of covariates in model γ (excluding the intercept) and the constant of proportionality does not depend on γ Eq. 5 (Liang *et al.* 2012). 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 none numerable model spaces, MCMC may beused such that $p(\gamma|Y)$ is the target distribution of the Markov chain. George and McCulloch (1997) discussed 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. We consider some toy examples to explain this problem the oretically and hence, obtain a better understanding of the properties ofthe MPM. For our theoretical examples, we will deal with finite and large n under conditions of severe collinearity. Our results complement the results of Fernandez *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 $\gamma\epsilon I$ 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 take effect.

RESULTS

Simulated data for p = 3: We generate highly correlated covariates by sampling 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 Eq. 1 with $\varphi = 1/4$. We look at a range of moderate to extremely large sample sizes in Table 3. For each sample size n, a single dataset is generated and the same data generating model is 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 when n is large.

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, inspite of the full model being the "true" model. The full model does not have a high posterior probability even for $n = 10^4$, finally posterior consistency takes effect when n is as large as 10^5. For $n \geq 1000$, one model usually has a high posterior probability but under repeated sampling there is considerable variability regarding which model gets the large mass. We run the experiment a second time to illustrate the sampling variability and report the results in Table 4.

Finally, Table 5 studies the posterior inclusion probabilities of covariates corresponding to the data sets generated in Table 3. We find that for n = 25 and n = 100, the marginalin clusion 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 covariatesis included in the model is 1-p((0, 0, 0)_|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 will very likely be effective for predictions in this high collinearity situation because it never discards all the important covariates. 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.

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.

Assumptions about R^2 and collinearity: First note that for the null model Y = (0, 0, ..., 0)', we have

Table 3: Simulation study for p 3, to demonstrate the effect of collinearity on posterior probabilities of models; the posterior probabilities of the top three models appear in bold

	$R^2\gamma$					$p(\gamma\|Y)$				
Y	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 4: Replicate of simulation study for p = 3 with collinearity in the design matrix, to demonstrate the effect of sampling variability; the posterior probabilities of the top three models appear in bold

	$R^2\gamma$					$p(\gamma\|Y)$				
Y	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)'	0.612	0.759	0.691	0.685	0.691	0.204	0.254	0.017	0.000	0.000
(0, 1, 0)'	0.629	0.761	0.693	0.685	0.691	0.332	0.344	0.915	0.000	0.000
(0, 1, 1)'	0.643	0.761	0.693	0.686	0.692	0.097	0.036	0.032	0.061	0.000
(1, 0, 0)'	0.617	0.760	0.690	0.685	0.691	0.231	0.293	0.004	0.000	0.000
(1, 0,1)'	0.617	0.760	0.691	0.686	0.692	0.045	0.032	0.001	0.175	0.000
(1, 1, 0)'	0.633	0.761	0.693	0.686	0.692	0.072	0.037	0.029	0.609	0.000
(1, 1, 1)	0.643	0.761	0.693	0.686	0.692	0.019	0.004	0.001	0.155	1.000

Table 5: Effect of collinearity on posterior marginal inclusion probabilities of covariates corresponding to the simulation study reported in Table 3

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

$R^2_\gamma = 0$ by definition. To deal with random R^2_γ for non-null models, we make the following assumption:

Assumption 1: Assume that the "true" model is the full model and that for all sample size n and for all non-null models γ with probability 1.

Proposition 1: If Assumption 1 holds then for given ≥0 and for g = n sufficiently large, the Bayes factor for comparing γ = (0, 0, ..., 0) and γ = (1, 0, ..., 0)[1] can be made smaller than with probability 1. The proof is given in Appendix 1. This result implies that the Bayes factor:

$$BF\left(\gamma = (0, 0, \ldots, 0)' : \gamma = (1, 0, \ldots, 0)'\right) \approx 0 \qquad (6)$$

with probability 1 if the specified conditions hold. For a discrete uniform prior for the model space that is p (γ) = $1/2^p$ for all models γ, the posterior probability of any model γ may be expressed entirely in terms of Bayes factors as:

$$p(\gamma|Y) = \frac{\left(p(Y|\gamma)(1/2^p)\right)}{\left(\sum_{\gamma\epsilon\Gamma} p(Y|\gamma)(1/2^p)\right)}$$

$$\frac{\frac{\left(p(Y|\gamma)\right)}{\left(\sum_{\gamma\epsilon\Gamma} p(Y|\gamma)\right)} = \left(\frac{p(Y|\gamma)}{p(Y|y^*)}\right)}{\frac{\left(\sum_{\gamma\epsilon\Gamma} p(Y|\gamma)\right)}{p(Y|y^*)} = \frac{BF(y:y*)}{\sum_{\gamma\epsilon\Gamma} BF(y:y*)}} \qquad (7)$$

where, γ*¡ϵΓ (Berger and Molina, 2005). Taking γ = (0, 0, ..., 0) and γ* = (1, 0, ..., 0) in Eq. 6 and 7, for large enough n, we have the following with probability 1:

$$p\left(\gamma = (0, 0, \ldots, 0)' \middle| Y\right) \approx 0 \qquad (8)$$

As the null model receives negligible posterior probability, we may omit it when computing the normalizing constant of p(γ|Y) that is, we may compute the posterior probabilities of non-null models by renormalizing over the set Γ-{(0, 0, ..., 0)} instead of Γ. We provide a formal justification of this approximation in Appendix 2.

We now make an assumption about strong collinearity among the covariates so that R^2_γ for all non-null models γϵΓ-{(0, 0, ..., 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:

$$\left\{\frac{1 + n\left(1 - R_{Y^2}\right)}{\left(1 + n\left(1 - R_{Y^2}\right)\left(\left(-(n-1)\right)/2\right)\right.}\right\}$$

can be taken to be approximately 1 for any pair of distinct non-null models γ and y' 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→∞. 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 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. We next study the posterior distribution of 2^2 models for an example with p = 2 highly correlated covariates and extend theresults to the general p scenario in the following.

Collinearity example for p = 2: Under assumptions 1 and 2 and the discrete uniform prior for the model space, p(γ) = $1/2^2$, the posterior probabilities of the 2^2 models can be approximated as follows with probability 1:

$$(0,0')|Y)p\left(y == OP\left(y = \left((0,1)'\middle|Y\right) \approx \frac{1}{2}\right.\right.$$
$$P\left(y = (1,0)'\middle|Y\right) \approx \frac{1}{2}, P\left(Y = (1,1)'\middle|Y\right) \approx 0 \qquad (9)$$

The detailed calculations are given in Appendix 3 and the results in Eq. 9 have the following implications with probability 1.

The marginal posterior inclusion probabilities for both covariates would be close to 0.5, so the MPM will most likely include at least one of the two important covariates which happened in all our simulations. The prior probability that at least one of the important covariates is included is 1-p(γ = (0, 0) = 1-(1/2)^2 = 3/4. The posterior probability of the same event is 1-p(γ = |Y)≈1 by Eq. 9. Let H_0 denote γ = (0,) and H_A denote its complement. Then the prior odds $P(H_A)/P(H_0)$ = (3/4)/(1/4) = 3 and the posterior odds $P(H_A|Y)/P(H_0|Y)$ is expected to be very large because $P(H_A|Y)\approx 1$ and $P(H_0|Y)\approx 0$ by Eq. 9. Thus, the Bayes factor BF(H_A: H_0) = ($P(H_A/Y)$ /$P(H_O/Y)$))/($P(H_{(A)}$/($P(H_0)$))) will be very large with probability 1, under the above assumptions.

Collinearity example for general p: Consider a similar setup with p highly correlated covariates and γ = (1, 1, ..., 1)' as the "true" model. Under Assumptions1 and 2, the following results hold with probability 1 which is implicitly assumed throughout this section. For large n, under Assumption 1, the nullmodel has nearly zero posterior probability by Eq. 8, 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 Eq. 5, all (2^p-1) non-nullmodels have the term {1+n (1-R_y^2)}-(n-1)/2 (approximately) incommon. Ignoring common terms, the marginal likelihood forany model of dimension p_γ is approximately proportional to (1+n) n-py-1/2. Given n, this term decreases as p_γ increases, so the models with p_γ = 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 thejth covariate is:

$$p\left(y_j-1\middle|Y=\frac{\sum Y\in : Y_{j=1}P_{(Y|y)}}{\sum_{y\in}p(Y|y)}\right)\approx\frac{\sum_{y\in:y_{j=1}}p^{(Y|y)}}{\sum y\in-\{(0,0,...,0)'\}p^{(Y|y)}}\approx\frac{\sum_{p_y=1}^{p}\binom{p-1}{p_y-1}\left((1+n)\frac{n-py-1}{2}\right)}{\sum_{p_y=1}^{p}\binom{p}{p_y}\left((1+n)\frac{n-py-1}{2}\right)} \quad (10)$$

where, the last approximation is due to assumption 2 regarding collinearity. The expression in Eq. 10 follows the fact that all p_γ-dimensional models have the marginal likelihood proportional to(1+n) n-py-1/2 2 approximately (using assumption 2 in Eq. 5), there are altogether (p/p_y) such models and exactly (p-1/p_y-1) of these have γ_j = 1. Dividing the numeratorand denominator of Eq. 10 by $(1+n)^{n-2/2}$, we have:

$$\left(y_{(j=1)}|Y\right)\approx\left(1+S_{(PY=2)}\right)^{P}=\frac{\left(P-1_{@P}(Y-1)\right)}{\left(P+\sum_{(PY=2)}{}^{P}\left(P_{@P}Y\right)\right)}$$

where, the last approximation follows for fixed p and sufficiently large n as the terms in the sumover p_γ (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≥3 compared to p = 2.

Let H_0: γ = (0, 0, ..., 0)' and H_A: complement of H_0. Because, the prior odds $P(H_A)/P(H_0)$ = (2^p-1) is fixed (for fixed p) and the posterior odds is large for sufficiently largen, the Bayes factor BF(H_A: 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 and provide a theoretical justification for them.

DISCUSSION

Based on the empirical results, it seems preferable to use independent priors for model matrices with high collinearity instead of scale mixtures of g-priors. The MPM is the model that includes all covariates with posterior marginal inclusion probabilities ≥0.5 so it is easy to understand, straight for ward to estimate and it generally has good performance except in cases of severe collinearity. As the threshold of 0.5 may not be appropriate for highly correlated covariates were commend 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 priors used in this study. Thus, development of priors along the lines by 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 itis 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 a grouping effect. Bayesians have formulated priors that will induce the grouping effect (Krishna *et al*., 2009; Liu *et al*., 2014). In some of these studies, the researchers have shown that including correlated covariates in a groupwith appropriate regularization or shrinkage rules may improvepredictions.

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. Ingeneral, the MPM does quite well under nonorthogonality too but may not do as well under high collinearity. One possibility would be to find the model with best predictive ability froma Bayesian point of view, measured by expected squared error loss with expectation taken with respect to the predictive distribution (e.g., Lemma 1) (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 thelist of all/sampled models is computationally feasible one could report it as well.

Appendix 1 (proof of proposition 1):

Proof. To simplify the notation let $R^2_\gamma = R^2$ for γ = Then putting g = n and using the expression for marginal likelihoodof the g-prior given in Eq. 5 we have:

$$BF\ (y = (0, 0, \ldots, 0)' : y = (1, 0, \ldots, 0)'\ (A'1)$$

$$\left(Y \middle| y = \frac{(0,0,\ldots,0)'}{p\left(Y \middle| y = (1,0,\ldots 0)'\right)}\right)$$

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

$$= \frac{1}{\dfrac{n-2}{2_{(1+n)}\left[\dfrac{1+n\left(1-R^2\right)}{1+N}\right]}}$$

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

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

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

Taking the logarithm of (A.1), the following result holds with probability 1, by assumption 1:

$$\log\left(BF\left(y = (0, 0, \ldots, 0)' : y = (1, 0, \ldots, 0)\right)\right)$$
$$\log\left((1+n)\left(1/2\left(1-n/\left(1+nR^2\right)\right)\right)\right)$$
$$= 1/2\log(1+n) + ((n-1))/2\log\left(1-n\ (n+1)\right)\left(R^2\right)$$
$$< 1/2\log(1+n) + ((n-1))/2\log\left(1-n/(n+1)\right)(\delta_1)$$

As n goes to infinity, the first term in (A.2) goes to∞at a logarithmicrate in n. Logarithm is a continuous function so goes to log (1-δ1) as n goes to infinity. Because 0<δ1<1, we have log (1-δ1) <0. This implies that the second term in (A.2) goes to -∞at apolynomial rate in n, of degree 1. Thus, as n→∞ with probability 1:

$$\log\ (BF(y = (0, 0, \ldots, 0)' : y = (1, 0, \ldots, 0)'))\to -\infty \qquad (A.3)$$

$$BF(y = (0, 0, \ldots, 0)' : (y = (1, 0, \ldots, 0)')^0 \qquad (A.4)$$

From (A.3) it follows that for sufficiently large n, we can make: BF(y = (0, 0, …, 0)', y = (1, 0, …, 0)'<, for given >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→∞. 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 2 (justification for omission of the null model for computing the normalizing constant):

$\sum p\ (Y/y)$ we first establish the following lemma. This shows that for computing a finite sum of positive quantities, if one of the quantities isnegligible compared to another, then the sum can be computed accurately even if the quantity with negligible contribution is omitted from the sum.

Lemma A.1: Consider $a_{in}>0$, i = 1, 2, ...m. If $a_{in}/a2_{in}$ as n →∞ then $\dfrac{\sum_{i=2}^{m} a_{in}}{\sum_{i=1}^{m} a_{in}} \to 1$.

Proof.

$$= \lim_{n\to\infty} \frac{\sum_{i=2}^{m} a_{in}}{\sum_{i=1}^{m} a_{in}} \lim_{n\to\infty} \frac{\sum_{i=2}^{m} a_{\frac{in}{a_{2n}}}}{\sum_{i=1}^{m} a_{\frac{in}{a_{2n}}}}$$

$$\lim_{n\to\infty} \frac{\sum_{i=2}^{m} a_{\frac{in}{a_{2n}}}}{\left(a_{\frac{in}{a_{2n}}} + \sum_{i=2}^{m} a_{\frac{in}{a_{2n}}} \right)}$$

$$= \lim_{n\to\infty} \sum_{i=2}^{m} \frac{a_{\frac{in}{a_{2n}}}}{\lim_{n\to\infty} (a_{in} / a_{2n}) + \lim_{n\to\infty} \sum_{i=2}^{m} a_{\frac{in}{a_{2n}}}}$$

$$= \lim_{n\to\infty} \sum_{i=0}^{m} \frac{a_{\frac{in}{a_{2n}}}}{0 + \lim_{n\to\infty} \sum_{i=2}^{m} a_{\frac{in}{a_{2n}}}} = 1$$

Corollary A.1. If assumption 1 holds, then given $\eta>0$, how ever small, for sufficiently large n, we can make $(1 - (y \in -\{(0, 0, ..., 0)'\})p(Y|y) / \Sigma(y \in p)(Y|y))$ with probability 1. Proof we have $p(Y|\gamma)>0$, $\gamma\epsilon\Gamma$ $p(Y|y = (0, 0, ..., 0)')/(Y|y = (1, 0, ..., 0)') \to\infty$ as $n\to\infty$ with probability 1, by (A.3). Then as $n\to\infty$ we have the following with probability 1, by Lemma A.1:

$$\frac{\sum_{y\epsilon-\{(0,0,...,0)'\}} p(Y|y)}{\sum_{y\epsilon} p(Y|y)} \to 1$$

The proof follows immediately:

Appendix 3 (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 Eq. 8. We derive the posterior probabilities of thenon-null models under assumptions 1 and 2 here. For any non-null model $\gamma\epsilon\tau$- {(0, 0)_}:

$$p(Y|y) = \frac{p(y)P(Y|y)}{\sum_{y\epsilon\tau\, P(y)} p(Y|y)} =$$

$$\left(\frac{\frac{1}{2^{(2)} p\left(\left(Y/y\right)\right)}}{\left(\sum_{y\epsilon\tau} \left(\left(\frac{1}{2^{(2)}}\right) p(Y|y)\right)\right)} \right) = \sum_{y\epsilon(\{(0,0)\}-p(Y|y))}$$

with probability 1. The last approximation in (C.1) follows from Corollary A.1 in Appendix 2. We will use the expression in (C.1) to derive the posterior probabilities. First note that under assumption 2 the term $\{1+n\ (1-R_Y^2)/-(N-1/2)\}$in the expression of marginal likelihood $p(Y|\gamma)$ in Eq. 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 (C.1). Then by Eq. 5 (C.1) and substituting g = n, we have with probability 1:

$$p(\gamma = (0,1)'|Y) \approx$$

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

Dividing the numerator and denominator of the right-hand side of (C.2) by $(1+n)^{(n?2)/2}$, (0, 1) $Y\approx 1/(2+(1+n)(-1/2))$ $P(y = 1/2)$ for large enough n with probability 1. Under assumption 2, we note that $P(y = (0.1)Y)\approx 1/2$ would have an identical expression as $p(\gamma = (0.1)Y)$. Hence, $p(\gamma = (1.0)Y)\approx 1$ for large enough n with probability 1. We finally derive $p(\gamma = (1.1)\_|Y)$ in a similar manner as follows:

$$\approx \frac{(1+n)^{\frac{n-2-1}{2}}}{(1+n)^{\frac{n-1-1}{2}} + (1+n)^{\frac{n-1-1}{2}} + (1+n)^{\frac{n-2-1}{2}}}$$

for large enough n with probability 1. [Received May 2013. Revised March 2015].

REFERENCES

Barbieri, M.M. and J.O. Berger, 2004. Optimal predictive model selection. Ann. Stat., 32: 870-897.

Berger, J.O. and G. Molina, 2005. Posterior model probabilities via path-based pairwise priors. Statistica Neerlandica, 59: 3-15.

Brown, P.J., T. Fearn and M. Vannucci, 2001. Bayesian wavelet regression on curves with application to a spectroscopic calibration problem. J. Am. Stat. Assoc., 96: 398-408.

Christensen, R., 2002. Plane Answers to Complex Questions. 3rd Edn., Springer, New York.

Clyde, M.A., J. Ghosh and M.L. Littman, 2012. Bayesian adaptive sampling for variable selection and model averaging. J. Comput. Graphical Stat., 20: 80-101.

Fernandez, C., E. Ley and M.F. Steel, 2001. Benchmark priors for Bayesian model averaging. J. Econ., 100: 381-427.

George, E.I. and R.E. McCulloch, 1997. Approaches for Bayesian variable selection. Stat. Sin., 7: 339-373.

Ghosh, J. and J.P. Reiter, 2013. Secure Bayesian model averaging for horizontally partitioned data. Stat. Comput., 23: 311-322.

Ghosh, J. and M.A. Clyde, 2012. Rao-blackwellization for bayesian variable selection and model averaging in linear and binary regression: A novel data augmentation approach. J. Am. Stat. Assoc., 106: 1041-1052.

Krishna, A., H.D. Bondell and S.K. Ghosh, 2009. Bayesian variable selection using an adaptive powered correlation prior. J. Stat. Plann. Inference, 139: 2665-2674.

Liang, F., R. Paulo, G. Molina, M.A. Clyde and J.O. Berger, 2012. Mixtures of g priors for Bayesian variable selection. J. Am. Stat. Assoc., 103: 410-423.

Liu, F., S. Chakraborty, F. Li, Y. Liu and A.C. Lozano, 2014. Bayesian regularization via graph Laplacian. Bayesian Anal., 9: 449-474.

Osborne, B.G., T. Fearn, A.R. Miller and S. Douglas, 1984. Application of near infrared reflectance spectroscopy to the compositional analysis of biscuits and biscuit doughs. J. Sci. Food Agric., 35: 99-105.

Zellner, A. and A. Siow, 1980. Posterior odds ratios for selected regression hypotheses. Proceedings of the 1st International Meeting on Bayesian Statistics, May 28-June 2, 1980, Valencia, pp: 585-603.

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, Goel, P.K. and A. Zellner (Eds.). North-Holland/ Elsevier, Amsterdam, The Netherlands, pp: 233-243.

Zou, H. and T. Hastie, 2005. Regularization and variable selection via the elastic net. J. Royal Stat. Soc., 67: 301-320.