Article type: Focus Article

Bayesian Model Selection Using the Median Probability Model

Article ID

Joyee Ghosh

Department of Statistics and Actuarial Science, The University of Iowa, Iowa City, IA 52242

Abstract

In the Bayesian approach to model selection, models and model specific parameters are treated as unknown quantities and uncertainty about them are expressed through prior distributions. Given the observed data, updating of the prior distribution to the posterior distribution occurs via Bayes' theorem. The posterior probability of a given model may be interpreted as the support it gets based on the observed data. The highest probability model (HPM) which receives the maximum support from the data is a possible choice for model selection. For large model spaces Markov chain Monte Carlo (MCMC) algorithms are commonly used to estimate the posterior distribution over models. However, estimates of posterior probabilities of individual models based on MCMC output are not reliable because the number of MCMC samples is typically far smaller than the size of the model space. Thus the HPM is difficult to estimate and for large model spaces it often has a very small posterior probability. An alternative to the HPM is the median probability model (MPM) of Barbieri and Berger [1], which has been shown to be the optimal model for prediction using a squared error loss function, under certain conditions. In this article we review some of the conditions for which the MPM is optimal, and provide real data examples to evaluate the performance of the MPM under small and large model spaces. We also discuss the behavior of the MPM under collinearity.

Keywords

Bayesian model averaging, Bayesian variable selection, Highest probability model, Marginal inclusion probability, Markov chain Monte Carlo

INTRODUCTION

We begin with a review of the Bayesian approach to variable selection. The full linear regression model with all p covariates is

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

where $\mathbf{Y} = (Y_1, \dots Y_n)'$ denotes the vector of response variables, $\boldsymbol{X}$ denotes the $n \times p$ design matrix with full rank, $\boldsymbol{\beta}$ is the $p \times 1$ vector of regression coefficients, ϕ is the reciprocal of the error variance, and $\mathbf{I}_n$ is an $n \times n$ identity matrix. Models corresponding to different subsets of covariates may be represented by the vector $\boldsymbol{\gamma} = (\gamma_1, \dots \gamma_p)'$, such that $\gamma_j = 1$ when the jth covariate $\boldsymbol{x}_j$ is included in the model and $\gamma_j = 0$ otherwise. Let $p_{\boldsymbol{\gamma}} = \sum_{j=1}^{p} \gamma_j$ denote the number of covariates in model $\boldsymbol{\gamma}$. The linear regression submodel under γ is

$$\mathbf{Y} \mid \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \phi, \boldsymbol{\gamma} \sim \mathsf{N}(\mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \mathbf{I}_n/\phi), \tag{2}$$

where $\mathbf{X}_{\boldsymbol{\gamma}}$ is the $n \times p_{\boldsymbol{\gamma}}$ design matrix containing the columns of $\boldsymbol{X}$ corresponding to the nonzero components of $\boldsymbol{\gamma}$, and $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ is the corresponding $p_{\boldsymbol{\gamma}} \times 1$ vector of regression coefficients under model $\boldsymbol{\gamma}$. The first column of $\boldsymbol{X}$ is usually taken to be an $n \times 1$ vector of ones corresponding to the intercept. In the Bayesian framework, models are treated as additional unknown parameters and they are assigned a prior distribution $p(\boldsymbol{\gamma})$. Then parameters within each model $\boldsymbol{\gamma}$ are assigned a prior distribution $p(\boldsymbol{\theta}_{\boldsymbol{\gamma}} \mid \boldsymbol{\gamma})$, where $\boldsymbol{\theta}_{\boldsymbol{\gamma}} = (\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \phi)$. The posterior probability of any model may now be obtained using Bayes' rule as:

$$p(\boldsymbol{\gamma} \mid \mathbf{Y}) = \frac{p(\mathbf{Y} \mid \boldsymbol{\gamma})p(\boldsymbol{\gamma})}{\sum_{\boldsymbol{\gamma} \in \Gamma} p(\mathbf{Y} \mid \boldsymbol{\gamma})p(\boldsymbol{\gamma})}, \tag{3}$$

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 the model $\boldsymbol{\gamma}$, and Γ is the space of models under consideration. For model selection, the highest probability model (HPM) is a possible candidate. If the goal is choosing the true model, then the HPM can be shown to be the optimum Bayesian model under a $0-1$ loss function, provided the true model is in the list of models under consideration [3, 13].

In the absence of any additional information, the most common scenario in the model selection framework is to consider all possible subsets of the p covariates leading to a model space Γ containing 2^p candidate models. When p is moderately large (bigger than $25-30$) enumerating all 2^p models and their posterior probabilities in (3) becomes computationally prohibitive due to the enormous size of the model space. Moreover, even for small model spaces the integral needed to calculate the marginal likelihood may not be available in closed form. Markov chain Monte Carlo (MCMC) [11, 12, 21, 20, 22, 4] or other stochastic search algorithms [2, 18, 8] are generally used to sample models for large model spaces. Typically the sample size for these algorithms is much smaller than the size of the model space, so the estimates of posterior model probabilities based on such samples are not very reliable. Thus accurate estimation of the HPM is regarded as a difficult task. Moreover, for large model spaces the HPM often has a negligible posterior probability, so inference based solely on the HPM ignores many other competing models with similar posterior probability.

Because of the difficulties mentioned above, the HPM is not always the preferred choice for model selection. It would be appealing to have a model selection method that incorporates the information across models and lends itself to more stable estimation. This motivates the use of posterior inclusion probabilities in Bayesian variable selection, which are more accurately estimated from the MCMC output compared to posterior probabilities of individual models, especially when the marginal likelihoods do not have closed form expressions. The posterior marginal inclusion probability for the jth covariate is defined as:

$$p(\gamma_j = 1 \mid \mathbf{Y}) = \sum_{\boldsymbol{\gamma} \in \Gamma : \gamma_j = 1} p(\boldsymbol{\gamma} \mid \mathbf{Y}). \tag{4}$$

The inclusion probabilities measure the importance of a covariate based on all models in which the covariate is included. This incorporates model uncertainty and the median probability model (MPM) of Barbieri and Berger [1] provides further justification for the use of these. The MPM is defined as the model which includes all covariates with posterior marginal inclusion probabilities greater than or equal to 0.5. Marginal inclusion probabilities can be estimated in a straightforward manner based on MCMC output, so the MPM can be estimated easily. In the next section we briefly discuss the optimal properties of the MPM and some conditions that are required for the optimality results to hold.

MEDIAN PROBABILITY MODEL

In this section we provide background for the MPM by highlighting some of the main ideas in the paper by Barbieri and Berger [1]. For a review with more technical details we refer interested readers to Section 9.8 of Ghosh et al. [13] which contains a clear and concise review of several main results concerning the MPM in a theorem proof style.

Motivation

In practical applications it is often of interest to find a model that yields good predictions rather than finding the true model. When the goal is predicting a future observation $y^* = \boldsymbol{x}^* \boldsymbol{\beta} + \epsilon$, at a given value of the covariates $\boldsymbol{x}^* = (x_1{}^*, \ldots, x_p{}^*)$, a common strategy is to use all the models in the list, and the resulting method is called Bayesian model averaging (BMA). For a quadratic loss function, that is

$$\mathcal{L}(y^*, \widehat{y}^*) = (y^* - \widehat{y}^*)^2, \tag{5}$$

where $\widehat{y}^*$ is a generic notation for the predicted value of y^*, the optimal predictor is the BMA estimate [3, 1, 6, 13]. Here optimality is defined in terms of minimizing the expected loss in (5), with expectation being taken over the posterior predictive distribution of y^* given the observed data $\mathbf{Y}$. The BMA estimate is a weighted average which weighs the predictions from each model with weights equal to their posterior probabilities as follows:

$$\widehat{y}^*_{BMA} = \boldsymbol{x}^* \widehat{\boldsymbol{\beta}}_{BMA} = \boldsymbol{x}^* \sum_{\boldsymbol{\gamma} \in \Gamma} p(\boldsymbol{\gamma} \mid \mathbf{Y}) S_{\boldsymbol{\gamma}} \tilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}, \tag{6}$$

where $S_{\boldsymbol{\gamma}}$ is a $p \times p_{\boldsymbol{\gamma}}$ matrix such that $\boldsymbol{x}^* S_{\boldsymbol{\gamma}}$ is the $p_{\boldsymbol{\gamma}}$-dimensional subvector of $\boldsymbol{x}^*$ with components corresponding to the nonzero coordinates of the model $\boldsymbol{\gamma}$, and $\tilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ is the posterior mean of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ under model $\boldsymbol{\gamma}$. As the dimension of $\tilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ varies across different models in Γ but $\boldsymbol{x}^*$ is always p-dimensional, the introduction of the matrix $S_{\boldsymbol{\gamma}}$ is required so that the matrices are conformable for multiplication in (6).

Because of practical considerations, sometimes one may need to select a single model to be used for repeated future predictions. Thus this framework rules out the possibility of using BMA for prediction. For example, there may be limited resources (in terms of time and/or money) which does not permit collecting information on all covariates in the future. If a single model is used for prediction, one will need information on the covariates in that model only. If the model is sparse, which is often the case for high-dimensional problems, this could be a more practical strategy. A somewhat problematic aspect of BMA is its lack of interpretability. Because BMA includes many models, it is difficult to assess the importance of covariates in the overall prediction. If a single best model is used for prediction, one may easily identify the important predictors as the ones that appeared in the best model. Barbieri and Berger [1] point out that a common misperception was that the HPM is the best model for prediction, but this is true only in special cases. For example, the optimality of the HPM holds if the model space contains only two models [3], and it sometimes holds for orthogonal design matrices in linear regression models. Barbieri and Berger [1] show that the MPM is the optimal predictive model under some general conditions. Of course in situations when the HPM and MPM are the same model, the HPM will also be optimal for prediction. Formally the MPM, say $\boldsymbol{\gamma}^*$, is defined as follows,

$$\gamma_j^* = \begin{cases} 1 & \text{if } p(\gamma_j = 1 \mid \mathbf{Y}) \geq 1/2, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

Existence

The MPM may not exist for an arbitrary model space. However, Barbieri and Berger consider a particular class of models which has a "graphical model structure", for which the MPM is guaranteed to exist. We provide their definition of this special class of models below.

Definition 1 *(Barbieri and Berger, 2004 [1]) Assume that for each covariate index j, there is an index set $I(j)$ of other covariates. A subclass of linear models is said to have "graphical model structure" if it consists of all models satisfying the condition that "for each j, if $\boldsymbol{x}_j$ is included in the model, then covariates $\boldsymbol{x}_k$ with $k \in I(j)$ are also included in the model."*

The above class includes the class of models with all possible subsets of the p covariates, which is one of the most commonly considered frameworks in variable selection. To see that, define $I(j)$ as the null set. Another class of models having the "graphical model structure" is a sequence of nested models defined as:

$$\boldsymbol{\gamma}(j),\ j=0,\ldots,p,\ \text{such that}\ \boldsymbol{\gamma}(j)=(\underbrace{1,\ldots\ldots,1}_{j\text{ ones}},\underbrace{0,\ldots\ldots,0}_{(p-j)\text{ zeroes}})^{'}. \tag{8}$$

The above notation basically says that in this sequence the first model does not have any covariate, the second one has only $\boldsymbol{x}_1$, the third model includes $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, and so on. To see that this has "graphical model structure" take $I(j)=\{k:1\leq k<j\}$ for $j\geq 2$ and otherwise take $I(j)$ as the null set. A sequence of nested models may arise in a polynomial regression scenario where j denotes the degree of the polynomial. For the special class of nested models (8) the MPM may be represented in a simpler way. One needs to calculate the cumulative sum of posterior probabilities of the models, starting from the smallest model and in ascending order, until the sum is greater than or equal to $1/2$. The MPM will then be the first model where this sum is $\geq 1/2$. This representation perhaps makes the name "median probability model" more intuitive. Barbieri and Berger derive several results under the nested models scenario. As nested models are not that common in the variable selection set up, at least in our experience, we do not discuss them any further in this section. We conclude this section with their optimality results relevant for the all possible subsets scenario.

Optimality

Let $\boldsymbol{x}^*$ denote the covariates where predictions will be made in the future, assume that these will follow some distribution such that

$$R=\mathsf{E}[(\boldsymbol{x}^*)^{'}\boldsymbol{x}^*]$$

exists and is positive definite. One possible choice is $R=\boldsymbol{X}^{'}\boldsymbol{X}$, implying that covariates for which future predictions will be made are similar to the ones in the observed data. Let the posterior mean of $\boldsymbol{\beta}$ in the full model be denoted as $\tilde{\boldsymbol{\beta}}$ and suppose the posterior means $\tilde{\boldsymbol{\beta}_{\boldsymbol{\gamma}}}$ in submodels $\boldsymbol{\gamma}$ satisfy

$$\tilde{\boldsymbol{\beta}_{\boldsymbol{\gamma}}}=S_{\boldsymbol{\gamma}}^{'}\tilde{\boldsymbol{\beta}}. \tag{9}$$

The above condition implies that posterior means of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ in submodels may be obtained by simply picking the corresponding coordinates of the full model posterior mean $\tilde{\boldsymbol{\beta}}$. The authors mention the following scenario with independent conjugate normal priors, where condition (9) will hold. Suppose $\boldsymbol{X}^{'}\boldsymbol{X}$ is diagonal and the prior for the full model is a p-dimensional normal distribution

$$p(\boldsymbol{\beta}\mid\phi)=\mathsf{N}_p(\boldsymbol{m},\boldsymbol{L}/\phi), \tag{10}$$

where $\boldsymbol{L}$ is a known diagonal matrix. Assume that the submodel priors are of the form

$$p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}\mid\phi)=N_{p_{\boldsymbol{\gamma}}}(S_{\boldsymbol{\gamma}}^{'}\boldsymbol{m},S_{\boldsymbol{\gamma}}^{'}\boldsymbol{L}S_{\boldsymbol{\gamma}}/\phi). \tag{11}$$

Then condition (9) will hold for a fixed ϕ or for any prior for ϕ.

Corollary 1 *(Barbieri and Berger, 2004 [1]) If R is diagonal with diagonal elements $r_j>0$, condition (9) holds, and any submodel of the full model is allowed, then the best predictive model is the median probability model given by (7). Further if ϕ is known for the priors in (10,11) and the prior model probabilities are of the form*

$$p(\boldsymbol{\gamma})=\prod_{j=1}^{p}{\pi_j}^{\gamma_j}(1-\pi_j)^{(1-\gamma_j)}, \tag{12}$$

where π_j is the prior probability for covariate $\boldsymbol{x}_j$ to be in the model, then the median probability model is the model with highest posterior probability.

In Corollary 1 the best predictive model refers to the model that minimizes the expected predictive loss, where the expectation is first taken with respect to the predictive distribution of y^* given $\mathbf{Y}$ and $\boldsymbol{x}^*$, and then a further expectation is taken over $\boldsymbol{x}^*$. Barbieri and Berger note that the above corollary holds when all models have common parameters, and one can define $\pi_j = 1$ for them. For example, an intercept is a parameter that is often included in all models. Corollary 1 has two important implications. First, the MPM is the optimal predictive model under an orthogonal design matrix, $R = \boldsymbol{X}^{'}\boldsymbol{X}$, and independent normal conjugate priors. Second, the additional conditions needed for the HPM and MPM to be the same are fairly restrictive. Thus even for orthogonal design matrices and independent conjugate normal priors on the regression coefficients, the HPM may not be the best predictive model (MPM). This suggests that it would be good practice to routinely report the predictions from the MPM, in addition to the HPM, when one is interested in model selection for prediction. The authors also consider some generalizations when the assumptions of a diagonal R can be relaxed, in the nested models case. In the next section we illustrate how to estimate the MPM in practice and evaluate its predictive performance.

EXAMPLES

We analyze two classic real datasets in this section. Hald's dataset has a very small model space so all models can be enumerated. The ozone dataset has a much larger model space so we use MCMC to explore the model space.

Hald's data: nested models

Hald's dataset consists of 4 covariates, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_4$, that are ingredients of a cement mix and a response variable, $\mathbf{Y}$, which is the heat evolved during the process of hardening. An interesting aspect of this example is high collinearity in the covariates, there are two pairs of highly correlated variables $\{\boldsymbol{x}_1, \boldsymbol{x}_3\}$ and $\{\boldsymbol{x}_2, \boldsymbol{x}_4\}$ with pairwise correlations being -0.82 and -0.97 respectively. There are 13 observations in this dataset.

We consider a sequence of 5 nested models all of which include an intercept: $\boldsymbol{\gamma}(0) = (1,0,0,0,0)'$, $\boldsymbol{\gamma}(1) = (1,1,0,0,0)'$, $\boldsymbol{\gamma}(2) = (1,1,1,0,0)'$, $\boldsymbol{\gamma}(3) = (1,1,1,1,0)'$, and $\boldsymbol{\gamma}(4) = (1,1,1,1,1)'$. Here $\boldsymbol{\gamma}(0)$ includes the intercept term only, $\boldsymbol{\gamma}(1)$ includes the intercept term and $\{\boldsymbol{x}_1\}$, $\boldsymbol{\gamma}(2)$ includes the intercept term and $\{\boldsymbol{x}_1, \boldsymbol{x}_2\}$, etc. This dataset has been used by Barbieri and Berger [1] but in this article we analyze it with a different prior under a full Bayesian framework. We use equal prior probabilities for the models, that is $p(\boldsymbol{\gamma}) = 1/5$ for all models, and Zellner's g-prior [23] for the model specific parameters. As the intercept is included in all models we first make a slight change of notation for a simpler representation. Let β_1 denote the intercept and $\boldsymbol{\beta}_{(1)\boldsymbol{\gamma}}$ denote the regression coefficients corresponding to the covariates in model $\boldsymbol{\gamma}$, excluding the first one, that is the intercept. For example, for $\boldsymbol{\gamma}(2)$, $\boldsymbol{\beta}_{\boldsymbol{\gamma}} = (\beta_1, \boldsymbol{\beta}_{(1)\boldsymbol{\gamma}}{}')^{'} = (\beta_1, \beta_2, \beta_3)^{'}$. Let $(\mathbf{1}, \mathbf{X}^c)$ denote the design matrix under the full model where $\mathbf{1}$ denotes an $n \times 1$ vector of ones corresponding to the intercept, and $\mathbf{X}^c$ contains the column vectors corresponding to the 4 covariates, assuming that they have been centered and scaled so that the mean of each column of $\mathbf{X}^c$ is 0 and the norm of each column is $\sqrt{n}$, as in Ghosh and Clyde [14]. Let $\mathbf{X}^c_{\boldsymbol{\gamma}}$ denote the submatrix of $\mathbf{X}^c$ under model $\boldsymbol{\gamma}$. Zellner's g-prior is given as follows:

$$\begin{aligned} p(\beta_1, \phi \mid \boldsymbol{\gamma}) &\propto 1/\phi, \\ \boldsymbol{\beta}_{(1)\boldsymbol{\gamma}} \mid \boldsymbol{\gamma}, \phi &\sim \mathsf{N}\left(\mathbf{0}, \frac{g}{\phi}(\mathbf{X}^{c}_{\boldsymbol{\gamma}}{}'\mathbf{X}^{c}_{\boldsymbol{\gamma}})^{-1}\right). \end{aligned} \tag{13}$$

An explicit analytic expression is available for the marginal likelihood for this prior (see Section 2.1 equation (5) of Liang et al. [19]),

$$p(\mathbf{Y} \mid \boldsymbol{\gamma}) \propto (1+g)^{\frac{n-p_{\boldsymbol{\gamma}}-1}{2}} \{1 + g(1-R^2_{\boldsymbol{\gamma}})\}^{-\frac{(n-1)}{2}}, \tag{14}$$

where $R^2_{\boldsymbol{\gamma}}$ is the standard coefficient of determination for model $\boldsymbol{\gamma}$ (see Section 14.1 of Christensen [5] for definition), $p_{\boldsymbol{\gamma}}$ is the number of nonzero components of $\boldsymbol{\gamma}$ excluding the first, and the constant of proportionality does not depend

Table 1 Posterior summaries for the nested models scenario for Hald's dataset.

	$\boldsymbol{\gamma}(0)$	$\boldsymbol{\gamma}(1)$	$\boldsymbol{\gamma}(2)$	$\boldsymbol{\gamma}(3)$	$\boldsymbol{\gamma}(4)$
$p(\boldsymbol{\gamma} \mid \mathbf{Y})$	0.0000	0.0001	0.7020	0.2348	0.0631
Cumulative sum of $p(\boldsymbol{\gamma} \mid \mathbf{Y})$	0.0000	0.0001	0.7021	0.9369	1.0000

on the model $\boldsymbol{\gamma}$. We set the hyperparameter $g = n$ [8]. The posterior probabilities, $p(\boldsymbol{\gamma} \mid \mathbf{Y})$ are obtained by putting $p(\boldsymbol{\gamma}) = 1/5$ and $p(\mathbf{Y} \mid \boldsymbol{\gamma})$ in (3), and the marginal inclusion probabilities, $p(\gamma_j = 1 \mid \mathbf{Y})$, are calculated by (4).

From Table 1 the MPM is $\boldsymbol{\gamma}(2)$, using the representation of the MPM for a sequence of nested models. According to this definition, $\boldsymbol{\gamma}(2)$ is the MPM because it is the first model in the sequence where the cumulative sum is $\geq 1/2$. This representation provides an intuition for the name "median probability model". Alternatively we can also use the more general definition of the MPM based on $p(\gamma_j = 1 \mid Y)$, which for this dataset are $1.0000, 0.9999, 0.2979, 0.0631$ for $\boldsymbol{x}_1 - \boldsymbol{x}_4$ respectively. The MPM includes all covariates with $p(\gamma_j = 1 \mid Y) \geq 1/2$, which would lead to selecting $\{\boldsymbol{x}_1, \boldsymbol{x}_2\}$ in addition to the intercept, which is indeed the same model $\boldsymbol{\gamma}(2)$. Here the MPM ended up being the middlemost model in the sequence, however this may not happen in general. For example if the last model in the sequence had a posterior probability greater than 0.5, it would have been the MPM. Here $\boldsymbol{\gamma}(2)$ has the maximum posterior probability so it is also the HPM. Finally, for a choice of $\mathsf{E}[(\boldsymbol{x}^*)^{'}\boldsymbol{x}^*] = a\mathbf{X}^{'}\mathbf{X}$, for some $a > 0$, this example satisfies the three conditions listed in Section 4 of Barbieri and Berger's paper, so by their Corollary 4 for nested models, the MPM $\boldsymbol{\gamma}(2)$ (which is also the HPM here) is the optimal predictive model.

Ozone data: all submodels

We use the popular ozone dataset used by several authors in earlier papers [9, 19, 14]. The response variable is ground level ozone in Los Angeles and there are 8 meteorological covariates. The description of the covariates is given in the Appendix. We consider a model with second order interactions and square terms to capture the nonlinearity [19, 14]. This leads to a total of 44 covariates and we consider all possible combinations of the covariates, while always including the intercept, like some of the previous papers. Unlike the previous example, all models cannot be enumerated due to the sheer size (2^{44}) of the model space. There are some considerably high correlations in the design matrix which is not uncommon when including square and interaction terms. The g-prior may have greater problems in the presence of collinearity than independent priors [15], so in addition to the g-prior, we also use conjugate independent normal priors [14] for this analysis, as described below. Let $p_{\boldsymbol{\gamma}}$ denote the number of covariates included in model $\boldsymbol{\gamma}$, excluding the intercept. Then consider the prior

$$\begin{aligned} p(\beta_1, \phi) &\propto 1/\phi, \\ \boldsymbol{\beta}_{(1)\boldsymbol{\gamma}} \mid \boldsymbol{\gamma}, \phi &\sim \mathsf{N}\left(\mathbf{0}, \Lambda_{\boldsymbol{\gamma}(1)}^{-1}/\phi\right). \end{aligned} \tag{15}$$

where $\Lambda_{\boldsymbol{\gamma}}$ is a $(p_{\boldsymbol{\gamma}} + 1) \times (p_{\boldsymbol{\gamma}} + 1)$ diagonal matrix with the first diagonal $\lambda_1 = 0$ and the rest of the $p_{\boldsymbol{\gamma}}$ diagonal elements being the subset of $\{\lambda_2, \ldots, \lambda_p\}$ for which the corresponding $\gamma_j = 1$, and $\Lambda_{\boldsymbol{\gamma}(1)}$ is the submatrix obtained by excluding the first row and first column of $\Lambda_{\boldsymbol{\gamma}}$. Under this prior, the posterior mean for $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ is given as

$$\tilde{\boldsymbol{\beta}_{\boldsymbol{\gamma}}} = (\mathbf{X}_{\boldsymbol{\gamma}}{}'\mathbf{X}_{\boldsymbol{\gamma}} + \Lambda_{\boldsymbol{\gamma}})^{-1}\mathbf{X}_{\boldsymbol{\gamma}}^{'}\mathbf{Y}, \tag{16}$$

where $\mathbf{X}_{\boldsymbol{\gamma}}$ denotes the design matrix under model $\boldsymbol{\gamma}$, whose first column corresponds to the intercept. We assume as before all the columns of $\mathbf{X}$ have been standardized to have mean 0 and norm $\sqrt{n}$, except the first column. The marginal likelihood under this prior can be obtained in closed form as

$$p(\mathbf{Y} \mid \boldsymbol{\gamma}) \propto |\Lambda_{\boldsymbol{\gamma}(1)}|^{1/2}|\mathbf{X}_{\boldsymbol{\gamma}}{}^{'}\mathbf{X}_{\boldsymbol{\gamma}} + \Lambda_{\boldsymbol{\gamma}}|^{-1/2}\left(\|\mathbf{Y} - \mathbf{X}_{\boldsymbol{\gamma}}\tilde{\boldsymbol{\beta}_{\boldsymbol{\gamma}}}\|^2 + \tilde{\boldsymbol{\beta}_{\boldsymbol{\gamma}}}^{'}\Lambda_{\boldsymbol{\gamma}}\tilde{\boldsymbol{\beta}_{\boldsymbol{\gamma}}}\right)^{-\frac{n-1}{2}}, \tag{17}$$

where $\{\lambda_2, \ldots, \lambda_p\}$ are positive hyperparameters. Following Ghosh and Clyde [14] we choose $\lambda_j = 1$ for $j = 2, \ldots, p$. The model space prior is taken to be a discrete uniform prior, $p(\boldsymbol{\gamma}) = 1/(2^{44})$ for $\boldsymbol{\gamma} \in \Gamma$, for both the g-prior as well as independent normal priors. For exploring the model space we use a Metropolis Hastings algorithm with a mixture kernel that randomly chooses between an add/delete step and a swap step so that, the add/delete step adds or removes a covariate from the current model, and the swap step exchanges a covariate in the current model with one that is not included. The swap step in this algorithm specially helps when there is collinearity [8, 14]. Alternatively sandwich algorithms [17] may be used to improve mixing.

Our goal is to compare the out of sample predictive performance of BMA, the HPM, and the MPM, under the two priors described above. We split the data randomly into two halves, and use one half for training and the other half for predicting the response variables. Each half consists of 165 observations. For a given training dataset, we run the MCMC algorithm for a million iterations under each prior and discard the first 200,000 samples as burn in. We repeat the whole procedure 100 times. For each replicate, we compute the mean squared error (MSE) for predicting the response variables in the corresponding test dataset. The MSE is computed for BMA, the HPM, and the MPM under both priors. The estimator under BMA was defined in equation (6). We replace the sum over all models by the sum over sampled models, as commonly done for large model spaces. Here $p(\boldsymbol{\gamma} \mid \mathbf{Y})$ is not known because the model space is too large to enumerate, so we use their standard Monte Carlo estimates. For a model $\boldsymbol{\gamma}$ this estimate is simply the proportion of times it was sampled by the MCMC algorithm, after burn in. Alternatively one could also use the marginal likelihoods to form another type of renormalized estimator [7, 10]. To calculate the BMA estimate, the remaining ingredients, $\tilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$'s, need to be calculated for each sampled model. These are calculated exactly by (16) under independent priors, and by the expression $\tilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}} = (\tilde{\beta}_1, \tilde{\boldsymbol{\beta}}'_{(1)\boldsymbol{\gamma}})' = (\bar{\mathbf{Y}}, g\hat{\boldsymbol{\beta}}'_{(1)\boldsymbol{\gamma}}/(1+g))'$ for the g-prior, where $\hat{\boldsymbol{\beta}}_{(1)\boldsymbol{\gamma}}$ is the ordinary least squares estimate of $\boldsymbol{\beta}_{(1)\boldsymbol{\gamma}}$ [19, 16]. Next, the Monte Carlo estimate of $p(\gamma_j = 1 \mid \mathbf{Y})$ is used (which is simply the fraction of times the index j was sampled) to determine the MPM. Once the covariates to be included in the MPM are known, the exact posterior mean of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ under it is obtained as in BMA. The HPM is estimated as the sampled model with maximum marginal likelihood and then posterior mean of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ is calculated under it.

The top panel in Figure 1 shows boxplots of square-root of MSE (RMSE) for the different estimators and priors. This plot gives an overall idea about the RMSE for each method (combination of prior and estimator) but it does not show explicitly the relative performance of different methods for a given test sample. The plot in the second panel displays the relative RMSE, defined as the RMSE of a method divided by the smallest RMSE obtained for a given test sample. For the best method this will ratio will be 1 and for other methods this will be larger than 1. It is clear from Figure 1 that the MPM under the g-prior has weaker predictive performance compared to other methods. This result is in agreement with that of Ghosh and Ghattas [15] who show that the MPM under the g-prior may have a tendency to drop a group of strongly correlated variables associated with the response variable, if there are more than two correlated variables. The MPM produces two unusually large RMSE values which makes it difficult to visualize the other boxplots. Thus we create similar plots in Figure 2 after removing these two values. The second panel of Figure 2 shows that for each prior, BMA has boxplots closest to 1, so it is frequently the best method, followed by the HPM and the MPM respectively. This is an example where the conditions of optimality of the MPM are not met. In principle one can determine the model whose predictive performance is closest to that under BMA, among the sampled models based on MCMC [8]. Clyde et al. [8] report that they found the HPM to be closer to this optimal model than the MPM, when there is high collinearity. Thus our results seem to align with this observation. The difference among the estimators is more pronounced for the g-prior and the independent priors have better predictive performance overall. However, the plot also reveals that predictions with BMA under the two priors have negligible difference.

There is considerable variability in the list of covariates included in the estimated HPM/MPM across different replicates because the replicates correspond to different random splits of the dataset. To obtain a better understanding of the phenomenon when the MPM is outperformed by the HPM in predictive accuracy we focus on 3 particular replicates. We consider some characteristics of the MPM and the HPM under the g-prior in Tables 2 and 3 below. The results in Table 2 show that the HPM may include covariates with relatively low marginal posterior inclusion probability and that it tends to have slightly larger model sizes compared to the MPM. Table 3 shows that for replicate 40, a very high RMSE of the MPM is accompanied by a very low marginal likelihood compared to the HPM. The HPM seems to include covariates with somewhat larger correlation than those included in the MPM. For example for replicate 1, the highest correlation among the covariates in the HPM is 0.96 compared to 0.71 in the MPM.

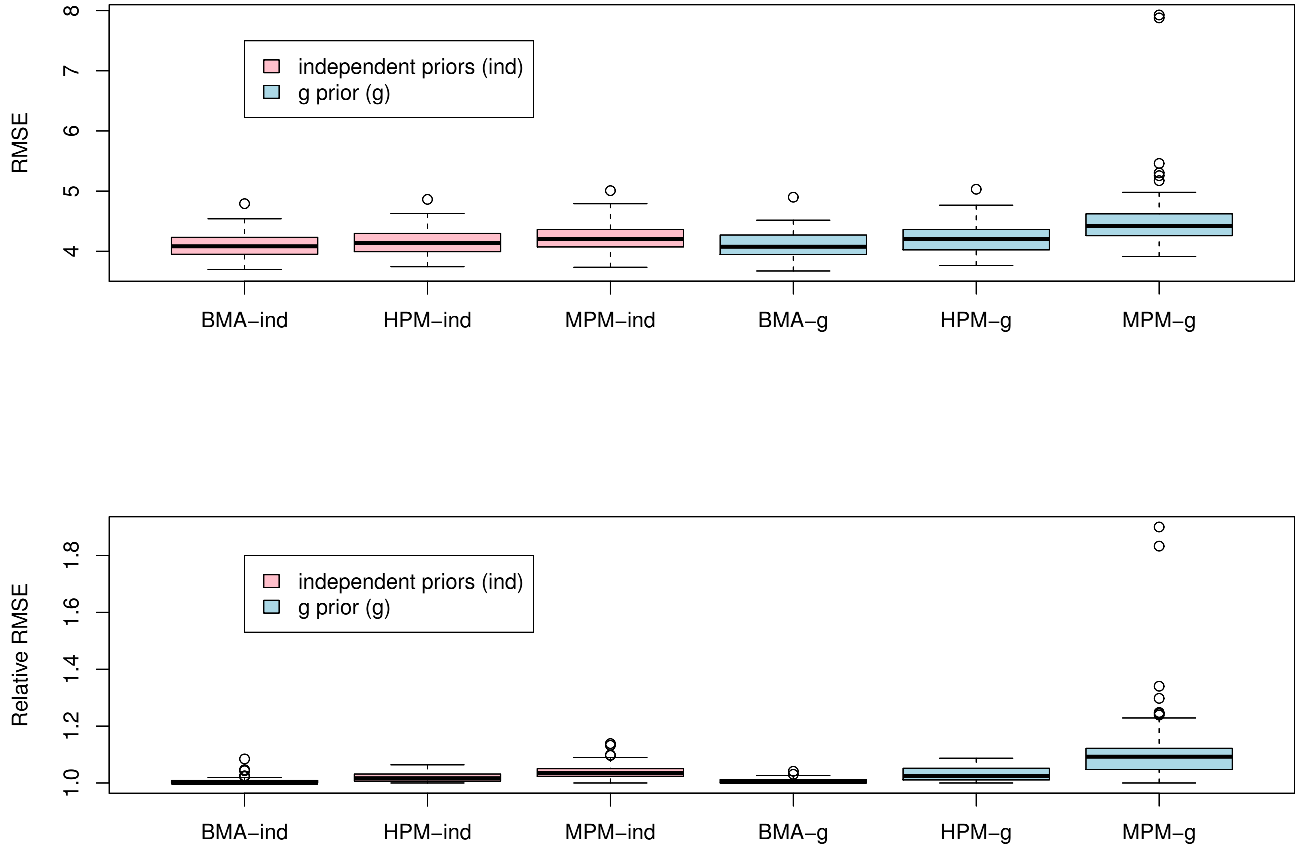

Figure 1: The root mean squared error (RMSE) for out of sample predictions for the ozone dataset based on Bayesian model averaging (BMA), the highest probability model (HPM) and the median probability model (MPM) under independent priors and the g-prior. Relative RMSE of any method is its RMSE relative to that of the best method (smallest RMSE). The boxplots are based on 100 replicates.

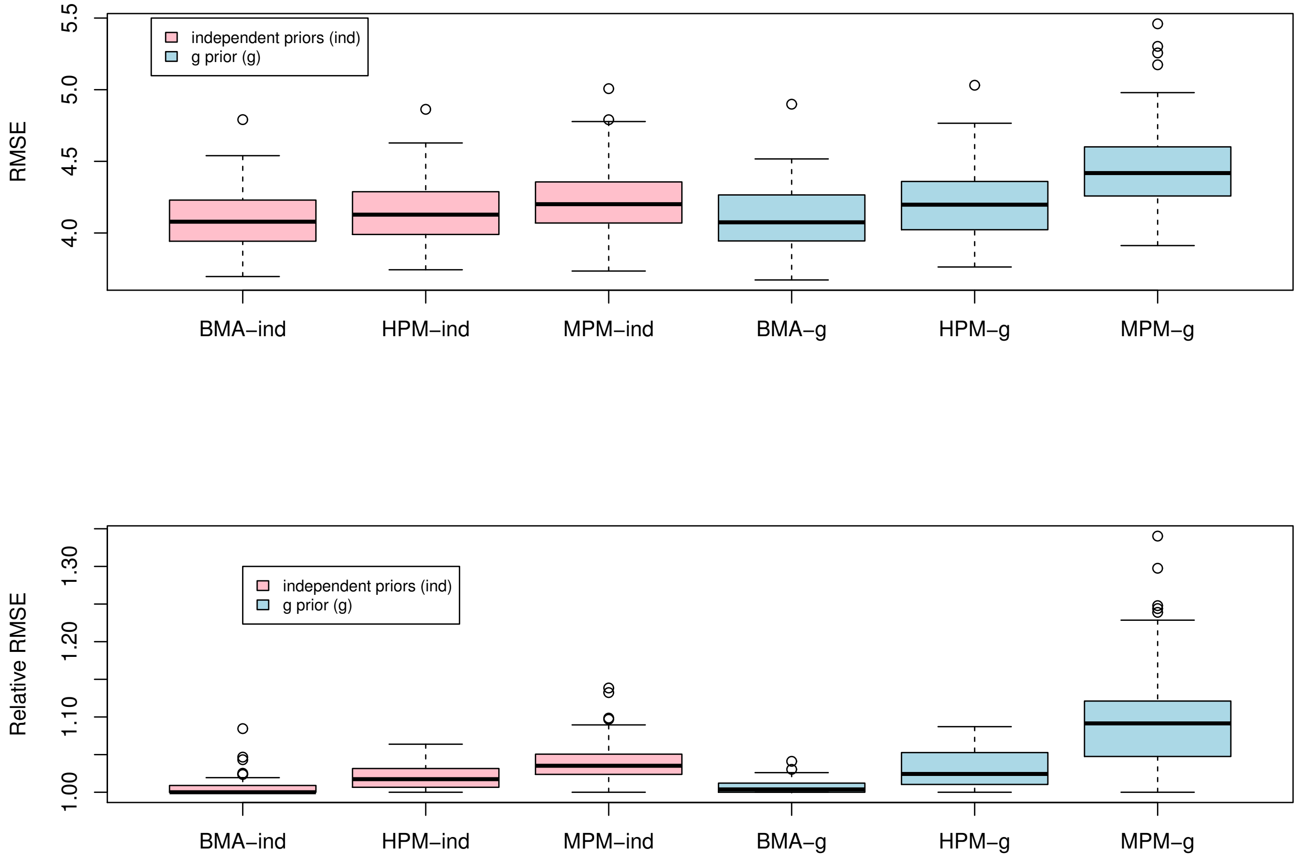

Figure 2: The root mean squared error (RMSE) for out of sample predictions for the ozone dataset based on Bayesian model averaging (BMA), the highest probability model (HPM) and the median probability model (MPM) under independent priors and the g-prior. Relative RMSE of any method is its RMSE relative to that of the best method (smallest RMSE). The boxplots are based on 98 replicates after removing two unusually large RMSE values for the MPM under the g-prior.

Table 2 List of all variables included in estimated HPM/MPM for replicates 1, 2 and 40 where +/* denotes if the variable is included in the estimated HPM/MPM. Estimated marginal posterior inclusion probabilities are provided in parentheses.

Replicate						
1	hum.2	dpg.2	hum.ibt	ibt	hum.dpg	temp.ibt
	(0.61)+*	(0.57)+*	(0.67)+*	(0.26)+	(0.47)+	(0.44)+
2	ibt	hum.ibt	hum.ibh	temp.ibt		
	(0.52)+*	(0.73)+*	(0.38)+	(0.28)+		
40	dpg.2	ibt.vis	vh.dpg	temp.ibt	ibh.dpg	
	(0.73)+*	(0.59)+*	(0.34)+	(0.47)+	(0.45)+	

Table 3 Bayes factor of MPM versus HPM and RMSE of BMA, HPM, and MPM for replicates 1, 2 and 40.

Replicate	BF(MPM:HPM)	BMA	MPM	HPM
		RMSE	RMSE	RMSE
1	1.60e-05	3.79	4.07	3.80
2	9.47e-07	4.14	4.47	4.08
40	1.22e-48	4.15	7.88	4.33

DISCUSSION

Based on theoretical and empirical results, using Bayesian model averaging for prediction instead of any single model seems to be a more robust strategy. However, when a single model needs to be chosen for constraints or interpretation, Barbieri and Berger recommend considering the median probability model in addition to the highest probability model. This is reasonable as they point out that the MPM is the only model for which some optimality results are known and it is quite straightforward to estimate based on MCMC output.

In this article we have considered two real data examples with high collinearity, the MPM (same as the HPM there) is theoretically optimal in one of the examples, while the HPM does better empirically in the other. Barbieri and Berger provide examples where optimality conditions of the MPM are not formally satisfied but the MPM still emerges as superior than the HPM by a large margin, so this motivates the use of the MPM even when theoretical results are not directly applicable. Moreover, we considered priors for which the marginal likelihoods are available in closed form, but for more general priors, estimating the HPM can be much more challenging. This suggests that both the posterior median and the modal model are worth reporting in an application. Finally, Barbieri and Berger mention the importance of joint inclusion probabilities for correlated covariates. Ghosh and Ghattas [15] demonstrate the usefulness of joint inclusion probabilities over marginal inclusion probabilities under severe collinearity in the design matrix so these can serve as useful additional summaries for correlated data.

ACKNOWLEDGMENT

The author thanks the Editor and two reviewers for several helpful suggestions that led to a much improved paper. The author was partially supported by the NSA Young Investigator grant H98230-14-1-0126.

APPENDIX: Ozone Data

Covariates	Description
vh	the altitude at which the pressure is 500 millibars
wind	the wind speed (mph)
hum	the humidity (in percent)
temp	the temperature (F)
ibh	the temperature inversion base height (feet)
dpg	the pressure gradient (mm Hg)
ibt	the inversion base temperature (degrees F)
vis	the visibility (miles)

The full model contains the above 8 main effects, 8 square terms, and 28 interaction terms. Square terms are denoted by adding ".2" to the main effect terms and interaction terms are indicated by joining the main effect terms by a dot.

References

[1] Maria Maddalena Barbieri and James O. Berger. Optimal predictive model selection. *Annals of Statistics*, 32(3):870–897, 2004. 1, 2, 3, 4, 5

[2] James O. Berger and German Molina. Posterior model probabilities via path-based pairwise priors. *Statistica Neerlandica*, 59:3–15, 2005. 2

[3] José M. Bernardo and Adrian F.M. Smith. *Bayesian Theory*. Wiley New York, 1994. 2, 3

[4] Leonardo Bottolo and Sylvia Richardson. Evolutionary stochastic search for Bayesian model exploration. *Bayesian Analysis*, 5(3):583–618, 2010. 2

[5] Ronald Christensen. *Plane Answers to Complex Questions*. Springer, 3 edition, 2002. 5

[6] Merlise A. Clyde and Edward I. George. Model uncertainty. *Statistical Science*, 19(1):81–94, 2004. 3

[7] Merlise A. Clyde and Joyee Ghosh. Finite population estimators in stochastic search variable selection. *Biometrika*, 99(4):981–988, 2012. 7

[8] Merlise A. Clyde, Joyee Ghosh, and Micahel L. Littman. Bayesian adaptive sampling for variable selection and model averaging. *Journal of Computational and Graphical Statistics*, 20(1):80–101, 2011. 2, 6, 7

[9] Jerome H. Friedman and Bernard W. Silverman. Flexible parsimonious smoothing and additive modeling. *Technometrics*, 31:3–39, 1989. 6

[10] Gonzalo García-Donato and Miguel Angel Martínez-Beneito. On sampling strategies in Bayesian variable selection problems with large model spaces. *Journal of the American Statistical Association*, 108(501):340–352, 2013. 7

[11] Edward I. George and Robert E. McCulloch. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88:881–889, 1993. 2

[12] Edward I. George and Robert E. McCulloch. Approaches for Bayesian variable selection. *Statistica Sinica*, 7:339–374, 1997. 2

[13] Jayanta K. Ghosh, Mohan Delampady, and Tapas Samanta. *An introduction to Bayesian analysis: theory and methods*. Springer, New York, 2006. 2, 3

[14] Joyee Ghosh and Merlise A. Clyde. 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, 2011. 5, 6, 7

[15] Joyee Ghosh and Andrew E. Ghattas. Bayesian variable selection under collinearity. Technical Report #417, The University of Iowa, 2014. 6, 7, 10

[16] Joyee Ghosh and Jerome P. Reiter. Secure Bayesian model averaging for horizontally partitioned data. *Statistics and Computing*, 23:311–322, 2013. 7

[17] Joyee Ghosh and Aixin Tan. Sandwich algorithms for Bayesian variable selection. *Computational Statistics and Data Analysis*, 81:76–88, 2015. 7

[18] Chris Hans, Adrian Dobra, and Mike West. Shotgun stochastic search for ”large p” regression. *Journal of the American Statistical Association*, 102:507–516, 2007. 2

[19] Feng Liang, Rui Paulo, German Molina, Merlise A. Clyde, and James O. Berger. Mixtures of g-priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103:410–423, 2008. 5, 6, 7

[20] David J. Nott and Peter J. Green. Bayesian variable selection and the Swendsen-Wang algorithm. *Journal of Computational and Graphical Statistics*, 13(1):141–157, 2004. 2

[21] Adrian E. Raftery, David Madigan, and Jennifer A. Hoeting. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92:179–191, 1997. 2

[22] Melanie A. Wilson, Edwin S. Iversen, Merlise A. Clyde, Scott C. Schmidler, and Joellen M. Schildkraut. Bayesian model search and multilevel inference for SNP association studies. *Annals of Applied Statistics*, 4(3):1342–1364, 2010. 2

[23] Arnold Zellner. 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*, pages 233–243. North-Holland/Elsevier, 1986. 5