Econometric Theory, **24**, 2008, 377–403. Printed in the United States of America.
DOI: 10.1017/S026646660808016X

AN IN-DEPTH LOOK AT HIGHEST POSTERIOR MODEL SELECTION

TANUJIT DEY
Case Western Reserve University

HEMANT ISHWARAN
Case Western Reserve University
and
Cleveland Clinic

J. SUNIL RAO
Case Western Reserve University

We consider the properties of the highest posterior probability model in a linear regression setting. Under a spike and slab hierarchy we find that although highest posterior model selection is total risk consistent, it possesses hidden undesirable properties. One such property is a marked underfitting in finite samples, a phenomenon well noted for Bayesian information criterion (BIC) related procedures but not often associated with highest posterior model selection. Another concern is the substantial effect the prior has on model selection. We employ a rescaling of the hierarchy and show that the resulting rescaled spike and slab models mitigate the effects of underfitting because of a perfect cancellation of a BIC-like penalty term. Furthermore, by drawing upon an equivalence between the highest posterior model and the median model, we find that the effect of the prior is less influential on model selection, as long as the underlying true model is sparse. Nonsparse settings are, however, problematic. Using the posterior mean for variable selection instead of posterior inclusion probabilities avoids these issues.

1. INTRODUCTION

In a Bayesian model averaging setting, at least from a predictive point of view, it is well acknowledged that averaging over models by their posterior model probabilities is an effective way to mitigate model uncertainty. The resulting predictor, called the Bayesian model averaged predictor, or BMA predictor, is often found to perform well in applied settings (Hoeting, Madigan, Raftery, and Volinsky, 1999). However, when the goal is model selection, the paradigm shifts dramatically from trying to predict the response Y to identifying which variables are influential in predicting Y. Because model selection ultimately

Address correspondence to Hemant Ishwaran, Department of Quantitative Health Sciences, Cleveland Clinic, Cleveland, OH 44195, USA; e-mail: hemant.ishwaran@gmail.com.

 0266-4666/08 $15.00

forces us to choose a single model from our class of models, it generally rules out the BMA predictor, which does not correspond to any one specific model. For that matter, model averaging as an approach is often perceived to be at odds with model selection.

Rather than employing model averaging, an often used approach to model selection is to choose the model with the highest posterior value, the so-called highest posterior probability model. This is the model whose integrated marginal density when multiplied by the prior probability is the highest among all such models. The highest posterior model is conceptually simple to understand, which makes it alluring as a model selection strategy. It is also attractive because it is tied to several well-accepted ideas. Under a uniform prior, for example, it is the model with a Bayes factor greater than or equal to one for all pairwise model comparisons. Furthermore, using a decision theoretic approach, one can show that the highest posterior model is the model that maximizes the expected utility under the BMA predictive distribution (Bernardo and Smith, 1994; Chipman, George, and McCulloch, 2001; Gelfand, Dey, and Chang, 1992).

The main contribution of the paper is to provide a rigorous analysis of the highest posterior model as a variable selection tool. Our discussion centers on linear regression models. We assume that $Y_1, \ldots, Y_n$ are independent responses such that

$$Y_i = \beta_1 x_{i,1} + \cdots + \beta_K x_{i,K} + \varepsilon_i = \mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{1}$$

where $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are nonrandom (fixed design) K-dimensional covariates. Here $\{\varepsilon_i\}$ are independent variables such that $\mathbb{E}(\varepsilon_i) = 0$ and $\mathbb{E}(\varepsilon_i^2) = \sigma^2 > 0$.

Equation (1) represents the true data generating mechanism. Our goal is to present a frequentist study of the asymptotic properties for the highest posterior model under this framework. For concreteness we focus specifically on highest posterior model selection using a class of Bayesian models refered to as *spike and slab models* (a detailed description of these models is given in Sect. 3). Such models have become popular as a method for variable selection in regression settings, making a detailed study of the highest posterior model under such a paradigm of great interest. Spike and slab models have been used in a wide variety of applications spanning problems from wavelets (Clyde, Parmigiani, and Vidakovic, 1998) to the identification of differentially expressing genes from microarray data (Ishwaran and Rao, 2003, 2005a). The prototype spike and slab model was introduced in Mitchell and Beauchamp (1988). Modern variants were given in George and McCulloch (1993). For background and further references see Ishwaran and Rao (2005b).

Throughout the paper we confine attention to the orthogonal regression setting. It is assumed throughout that $\mathbf{X}'\mathbf{X} = n\mathbf{I}$, where $\mathbf{X}$ is the $n \times K$ design matrix corresponding to (1). Focusing on the orthogonal model is advantageous because it leads to closed form expressions that help amplify certain theoretical results. We do provide examples, however, that suggest that the theory extends to nonorthogonal settings.

1.1. Main Results

Surprisingly, our analysis reveals several deficiencies in highest posterior model selection. This is interesting, because although empirical evidence suggesting caution in using the highest posterior model is documented (Atkinson, 1978), it is generally believed that the highest posterior model possesses good model selection properties. Our key findings are as follows:

1. In the setting considered, the highest posterior model coincides with the *median model* (Barbieri and Berger, 2004). This is a direct consequence of orthogonality of the regressors. The median model is defined as the model consisting of those variables with overall posterior inclusion probability greater than or equal to 50%. The posterior inclusion probability for a variable k is the posterior probability for the set consisting of all models that include k. We find that the highest posterior model, and hence the median model, is consistent.
2. Unfortunately, the promising asymptotics do not speak to finite-sample performance. The rate at which overfitting vanishes is only of order $\log(n)$, and even more worrisome is that among underfit models, the highest posterior model may favor smaller ones with large coefficients. As a result, finite-sample performance can be far from optimal in terms of total risk. This is brought out rigorously using a local asymptotic argument in which all coefficients have small values. This setup shows that the highest posterior model eventually concentrates on the null model. This turns out to be a general phenomenon shared by certain procedures (see Leeb and Pötscher, 2005, Rmk. 4.4).
3. We explore a rescaled spike and slab hierarchy where the responses $\{Y_i\}$ are rescaled by the square root of the sample size, and we show how this mitigates the underfitting tendencies of the highest posterior model. Specifically, we find a perfect cancellation of a $\log(n)$ term from the log posterior model probability that is the culprit leading to underfitting in finite samples. The trade-off, however, is a lack of consistency.
4. The choice of prior plays a heavy role in highest posterior model selection, although less so under rescaling. By focusing on posterior inclusion probabilities (which is motivated by the fact that the highest posterior model and the median model still coincide under a rescaled spike and slab framework), we show that the effect of the prior under rescaling is less pronounced in sparse models. However, if the true model is nonsparse, there is a delicate interplay between overall model complexity and the prior belief of the informativeness of a predictor that may lead to the highest posterior model including too many truly zero coefficients.

1.2. The Effect of the Prior on Model Selection

To amplify the preceding point of how the prior can impact model selection, and to help fix ideas, consider Table 1. The table is based on a highest posterior

Table 1. Analysis of body fat data from Hoeting et al. (1999). The first column records predictor name. The next two columns are models determined using BIC (Schwarz, 1978) and AIC (Akaike, 1973) penalization. The next three columns are highest posterior model, $\mathcal{M}_{\hat{\alpha},\gamma}$, selected under different choices for the hypervariance ($\gamma_k = \gamma$ for $\gamma = 1, 10, n$). The last three columns are the corresponding posterior inclusion probabilities, $p_{k,\gamma}$. Terminology and notation are explained later in the paper.

Predictors	BIC	AIC	$\mathcal{M}_{\hat{\alpha},1}$	$\mathcal{M}_{\hat{\alpha},10}$	$\mathcal{M}_{\hat{\alpha},n}$	$p_{k,1}$	$p_{k,10}$	$p_{k,n}$
Knee	✓	✓	✓	✓	✓	1.000	1.000	1.000
Abdomen	✓	✓	✓	✓	—	0.995	0.900	0.561
Wrist	—	✓	✓	—	—	0.673	0.449	0.149
Neck	✓	✓	—	✓	✓	0.453	0.889	0.976
Ankle	—	✓	✓	—	—	0.993	0.404	0.036
Weight	—	✓	—	—	—	0.388	0.379	0.096
Chest	✓	✓	✓	✓	—	0.735	0.642	0.286
Forearm	—	—	✓	—	—	0.910	0.195	0.031
Thigh	—	✓	✓	—	—	0.729	0.367	0.120
Age	—	—	—	—	—	0.364	0.287	0.138
Biceps	—	—	✓	—	—	0.614	0.195	0.045
Hip	—	—	—	—	—	0.332	0.151	0.037
Height	—	—	—	—	—	0.243	0.116	0.027

probability analysis of the body fat data of Hoeting et al. (1999). The data analyzed consist of $n = 251$ subjects, with each subject being measured for percentage of body fat. Also recorded are the age, weight, height, and 10 other body circumference measurements for the individual. As in Hoeting et al. (1999), we use body fat for the response in our regression model. The remaining 13 measurements are predictors for predicting body fat (see column 1 of Table 1 for a list of these variables). Note that with 13 predictors there is a total of $2^{13} = 8{,}192$ models, thus making an all-subsets highest posterior model search feasible.

Listed in columns 4 through 6 of Table 1 are the highest posterior models selected under various prior specifications. Models were computed under a nonscaled spike and slab hierarchy with a uniform model prior (terminology will be spelled out later in the paper). The three models correspond to different values for the prior variance (hypervariance) of the β_k coefficients. As can be seen, as the hypervariance γ increases, the highest posterior model quickly moves from a model larger than the Akaike information criterion (AIC) selected model to a model even sparser than the one selected using the Bayesian information criterion (BIC). This shows how sensitive highest posterior model selection can be to the choice of prior. Another interesting point, and one that came as somewhat of a surprise to us, is the relationship model size has with respect to γ.

Typically, large hypervariances in spike and slab models are used to elicit a prior belief that a coefficient is informative. In contrast, Table 1 suggests that large hypervariances induce sparse models under highest posterior model selection, thus going in the opposite direction in which priors are typically chosen. Later we provide a theoretical explanation for this curious property.

Last, consider the final three columns of Table 1, which list the posterior inclusion probabilities for the three sets of priors used. Close inspection shows that the highest posterior model nearly coincides with the median model, that is, the model formed by choosing predictors k with $p_{k,\gamma} \geq 0.5$. This is interesting because it suggests that our theory, which takes advantage of the equivalence of the two models under orthogonal designs, may apply reasonably well even in highly correlated settings such as the one considered here.

1.3. Organization of the Paper

The paper is organized as follows. Section 2 introduces notation and the concept of total risk for variable selection. Section 3 introduces spike and slab models and formally defines the highest posterior model and the median model. Section 4 studies the asymptotic selection performance of the highest posterior model, and Section 5 makes use of a local asymptotic argument. Section 6 introduces the idea of rescaled spike and slab hierarchal models and studies their asymptotic selection performance and impact rescaling has on appropriate choice of priors. Section 7 presents a detailed simulation for studying performance of the procedures. Section 8 finishes with a discussion. The Appendixes supply proofs of results and a discussion of the asymptotics for null and full models.

2. NOTATION

The problem of selecting variables in the linear regression setting can be recast formally as a model selection problem using the following notation. For each subset $\alpha \subseteq \{1, \ldots, K\}$, let $\boldsymbol{\beta}_\alpha$ be the components of $\boldsymbol{\beta}$ that are indexed by the elements of α. One can think of $\boldsymbol{\beta}_\alpha$ equivalently in terms of the underlying model associated with it, which we call $\mathcal{M}_\alpha$. Formally, this model is defined as

$$Y_i = \mathbf{x}'_{i,\alpha}\boldsymbol{\beta}_\alpha + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where $\mathbf{x}_{i,\alpha}$ are the components of $\mathbf{x}_i$ indexed by α (this notation of extracting coordinates indexed by α will be used repeatedly throughout the paper). When we think of $\boldsymbol{\beta}_\alpha$, we will think of this interchangeably as either α or $\mathcal{M}_\alpha$. We define $K_\alpha = \#\mathcal{M}_\alpha$ to be the size of $\mathcal{M}_\alpha$ or, equivalently, the number of variables in $\boldsymbol{\beta}_\alpha$.

For example, for (1), which is often referred to as the *full model*, $\alpha = \{1, \ldots, K\}$ and $K_\alpha = K$. The other extreme model, the *null model* denoted by $\mathcal{M}_\emptyset$, corresponds to $\alpha = \emptyset$ and the model $\mathcal{M}_\emptyset$ defined as

$$Y_i = \varepsilon_i, \qquad i = 1, \ldots, n. \tag{2}$$

The size of the null model is $K_\emptyset = 0$.

2.1. The True Model

Let $\boldsymbol{\beta}_0 = (\beta_{1,0}, \ldots, \beta_{K,0})'$ denote the true value for $\boldsymbol{\beta}$. We indicate the true model by $\mathcal{M}_{\alpha_0}$, or just α_0, where α_0 are the indices of the nonzero coefficients of $\boldsymbol{\beta}_0$. Throughout it is assumed that $\mathcal{M}_{\alpha_0}$ is neither the null nor the full model. In other words, it is assumed that

$$1 \leq K_{\alpha_0} \leq K - 1. \tag{3}$$

Such an assumption might seem at odds with our choice of prior (to be discussed shortly), which allows for the posterior to concentrate on both the null and full models. Our reason for assuming (3) is primarily for ease of presentation of asymptotic results. As a convenient way of summarizing the behavior of the highest posterior model our asymptotic results are cast in the form of comparisons of posterior model probabilities for overfit and underfit models relative to $\mathcal{M}_{\alpha_0}$ (this style of presentation is common in the model selection literature; see, e.g., Shao, 1993; Zhang, 1993a). Unfortunately, for the notion of an overfit and underfit model to make sense, we must exclude the case when $\mathcal{M}_{\alpha_0}$ is either the full or the null model. See, however, Appendix B for how our results are adjusted when (3) is violated.

2.2. Total Risk

To measure performance of a model selection procedure, we introduce the notion of total risk. The total risk is essentially a composite loss function under zero-one loss misclassification. To explain this, first note that a variable selection procedure can always be recast as a decision rule. If $\mathcal{M}_{\hat{\alpha}}$ is the model selected by the procedure, the corresponding decision rule is $\hat{\boldsymbol{\delta}} = (\hat{\delta}_1, \ldots, \hat{\delta}_K)'$, where

$$\hat{\delta}_k = \begin{cases} 1 & \text{if } k \text{ is included in } \hat{\alpha} \\ 0 & \text{if } k \text{ is excluded from } \hat{\alpha}. \end{cases}$$

Let $\delta_{k,0} = 1$ if and only if $\beta_{k,0} \neq 0$. We define the *total risk* for a model selection procedure $\hat{\boldsymbol{\delta}}$ as

$$\mathcal{R}(\hat{\boldsymbol{\delta}}) = \sum_{k=1}^{K} |\hat{\delta}_k - \delta_{k,0}| = \sum_{k \in \alpha_0} \mathbb{I}\{\hat{\delta}_k = 0\} + \sum_{k \in \alpha_0^c} \mathbb{I}\{\hat{\delta}_k = 1\}.$$

We say that a variable selection procedure $\hat{\boldsymbol{\delta}}$ is total risk consistent if $\mathcal{R}(\hat{\boldsymbol{\delta}}) \xrightarrow{\text{a.s.}} 0$. Notice that this is equivalent to the statement that the model estimator is consistent,

$$\mathbb{P}\{\mathcal{M}_{\hat{\alpha}} = \mathcal{M}_{\alpha_0}\} \to 1.$$

However, the measure of total risk is more meaningful for finite samples, as it provides a way of comparing procedures in terms of misclassification rates. See Ishwaran and Rao (2003, 2005a, 2005b) for further details. Throughout the paper we use total risk as a measure of performance.

3. BAYESIAN HIERARCHY

In this section we introduce a spike and slab framework that will be used to define our highest posterior model selection procedure, $\mathcal{M}_{\hat{\alpha}}$. The properties of this selection procedure under the data generating mechanism (1) will then be explored in Section 4.

Two ingredients are required in specifying the spike and slab framework. The first involves specifying the prior used for the model space, which we will denote by π. This is chosen so that $\pi(\mathcal{M}_\alpha) > 0$ for each model α and $\sum_\alpha \pi(\mathcal{M}_\alpha) = 1$ (the sum being over all possible 2^K models). The second ingredient is the likelihood structure used for modeling each regression problem $\mathcal{M}_\alpha$. This is specified by making use of a Bayesian hierarchy.

Section 3.1 discusses our choice for π in detail and how the resulting hierarchy is related to a popular class of models referred to as spike and slab models. Here we discuss the likelihood for $\mathcal{M}_\alpha$. This is based on a normal-normal conjugate hierarchical model. For each nonnull model $\mathcal{M}_\alpha$, we assume

$$\begin{aligned}(\mathbf{Y}|\,\mathcal{M}_\alpha, \mathbf{X}, \boldsymbol{\beta}) &\sim \text{Normal}(\mathbf{X}_\alpha \boldsymbol{\beta}_\alpha, \hat{\sigma}_n^2 \mathbf{I}),\\ (\boldsymbol{\beta}_\alpha|\boldsymbol{\Gamma}_\alpha) &\sim \text{Normal}(\mathbf{0}, \boldsymbol{\Gamma}_\alpha),\end{aligned} \tag{4}$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)'$, $\boldsymbol{\Gamma}_\alpha$ is the diagonal matrix extracted from $\boldsymbol{\Gamma}$ using only those coordinates in α, and $\boldsymbol{\Gamma} = \text{diag}\{\gamma_1, \ldots, \gamma_K\}$ where $0 < \gamma_k < \infty$ are prespecified hypervariances. Typically $\{\gamma_k\}$ are selected to reflect a prior belief in a coefficient. Large γ_k values induce large posterior values for β_k and reflect a belief that the coefficient is informative, whereas small values give small β_k coefficients and indicate a prior belief that the coefficient is noninformative.

Although a normal-normal hierarchical structure makes sense for nonnull models, special allowance has to be made for the null model $\mathcal{M}_\emptyset$. To properly accommodate model (2), we assume

$$(\mathbf{Y}|\,\mathcal{M}_\emptyset) \sim \text{Normal}(\mathbf{0}, \hat{\sigma}_n^2 \mathbf{I}). \tag{5}$$

In both (4) and (5), $\hat{\sigma}_n^2 \xrightarrow{\text{p}} \sigma^2$ is assumed to be some consistent estimator for σ^2. The rationale for introducing $\hat{\sigma}_n^2$ in the hierarchy (4) rather than taking a more traditional approach of introducing σ^2 as a Bayesian parameter with a prior is that it greatly simplifies our analysis. However, this is not to say that there is a loss in inferential capability in doing so. For example, Ishwaran and Rao (2005a) introduced both $\hat{\sigma}_n^2$ and a Bayesian parameter σ^2 in their setup

and found only small incremental improvements in their selection procedure due to the inclusion of σ^2.

3.1. Spike and Slab Models

For π we use an independence prior of the form

$$\pi(\mathcal{M}_\alpha) = \prod_{k=1}^{K} w_k^{\mathbb{I}\{k\in\alpha\}}(1-w_k)^{\mathbb{I}\{k\notin\alpha\}}, \tag{6}$$

where $0 < w_k < 1$ are prespecified probabilities reflecting the prior belief that a coefficient is nonzero. Independence priors are a popular choice in Bayesian variable selection procedures, in both orthogonal and nonorthogonal settings, because they are easy to specify and simplify computations (Clyde, DeSimone, and Parmigiani, 1996; George and McCulloch, 1993, 1997; Hoeting et al., 1999; Smith and Kohn, 1996). One common choice for (6) is to set $w_k = \frac{1}{2}$ for each k. This gives $\pi(\mathcal{M}_\alpha) = 2^{-K}$ for each α and is often referred to as an *indifference*, or uniform, prior. More elaborate setups that involve a prior for w_k are also used (Ishwaran and Rao, 2005a). However, we will not discuss these here.

When π is specified by (6), the hierarchy (4) and (5) can be written more compactly using a multivariate normal scale mixture distribution for $\boldsymbol{\beta}$. Let $b_k \sim \text{Bernoulli}(w_k)$ be independent Bernoulli random variables. That is,

$$\mathbb{P}\{b_k = 1\} = w_k = 1 - \mathbb{P}\{b_k = 0\}.$$

The vector $\mathbf{b} = (b_1, \ldots, b_K)'$ can be uniquely mapped to a specific α model. Thus, if we use the informal notation of denoting a degenerate multivariate normal by using zero values for the variance, then (4)–(6) can be more compactly expressed as

$$\begin{aligned}
(\mathbf{Y}|\mathbf{X},\boldsymbol{\beta}) &\sim \text{Normal}(\mathbf{X}\boldsymbol{\beta}, \hat{\sigma}_n^2\mathbf{I}),\\
(\boldsymbol{\beta}|\mathbf{b},\boldsymbol{\Gamma}) &\sim \text{Normal}(\mathbf{0}, \text{diag}\{b_1\gamma_1, \ldots, b_K\gamma_K\}),\\
(b_k|w_k) &\overset{\text{ind}}{\sim} \text{Bernoulli}(w_k), \qquad k = 1, \ldots, K.
\end{aligned} \tag{7}$$

When written this way, the hierarchy (7) is often referred to as a spike and slab model (Ishwaran and Rao, 2005a, 2005b).

3.2. The Highest Posterior Model

It is widely believed that the highest posterior probability model possesses optimal model selection properties. As we show shortly this may not be the case at all. To define this model, observe by Bayes theorem that the posterior probability for a model $\mathcal{M}_\alpha$ is

$$\pi(\mathcal{M}_\alpha|\mathbf{Y}) = \frac{\pi(\mathcal{M}_\alpha)f(\mathbf{Y}|\mathcal{M}_\alpha)}{\sum_{\alpha'}\pi(\mathcal{M}_{\alpha'})f(\mathbf{Y}|\mathcal{M}_{\alpha'})},$$

where

$$f(\mathbf{Y}|\mathcal{M}_\alpha) = \int f(\mathbf{Y}|\mathcal{M}_\alpha, \mathbf{X}, \boldsymbol{\beta})f(\boldsymbol{\beta}_\alpha)\, d\boldsymbol{\beta}_\alpha$$

is the marginal density for Y under $\mathcal{M}_\alpha$. The model $\mathcal{M}_{\hat{\alpha}}$ is said to be *the highest posterior model* if

$$\pi(\mathcal{M}_{\hat{\alpha}}|\mathbf{Y}) \geq \max_\alpha \pi(\mathcal{M}_\alpha|\mathbf{Y}).$$

Equivalently, $\mathcal{M}_{\hat{\alpha}}$ is the highest posterior model if

$$\frac{\pi(\mathcal{M}_{\hat{\alpha}})f(\mathbf{Y}|\mathcal{M}_{\hat{\alpha}})}{\pi(\mathcal{M}_\alpha)f(\mathbf{Y}|\mathcal{M}_\alpha)} \geq 1, \quad \text{for each } \alpha.$$

Notice that under a flat prior for π this indicates that the highest posterior model has a Bayes factor that is greater than or equal to one over all model comparisons.

3.3. The Median Model

The issue of whether the highest posterior model is appropriate for inference was partially considered in Barbieri and Berger (2004). There it was shown in orthogonal settings that the optimal model from a Bayesian predictive viewpoint was not the highest posterior model but the median model. The median model can be defined formally as follows. Let $\Delta_k = \{\alpha : k \in \alpha\}$ denote the set of all models containing variable k. The model $\mathcal{M}_{\hat{\alpha}}$ is said to be *the median model* if $\hat{\alpha} = \{k : p_k \geq \frac{1}{2}\}$ where p_k is the posterior inclusion probability defined by

$$p_k = \sum_{\alpha \in \Delta_k} \pi(\mathcal{M}_\alpha|\mathbf{Y}). \tag{8}$$

In other words, $\mathcal{M}_{\hat{\alpha}}$ is the model containing variables that appear with at least 50% posterior probability over all models.

The optimality result for the median model hinges on the definition of predictive optimality. This was defined as follows. For each α, let $\hat{\boldsymbol{\beta}}_\alpha = \mathbb{E}(\boldsymbol{\beta}|\mathbf{Y}, \mathcal{M}_\alpha)$ be the posterior mean for $\boldsymbol{\beta}$ from $\mathcal{M}_\alpha$. For convenience we write $\hat{\boldsymbol{\beta}}_\alpha$ as a K-dimensional vector by setting coordinates not in α to equal zero. Let $\hat{\boldsymbol{\beta}}$ be the BMA estimator for $\boldsymbol{\beta}$ from (7). Then,

$$\hat{\boldsymbol{\beta}} = \sum_\alpha \mathbb{E}(\boldsymbol{\beta}|\mathbf{Y}, \mathcal{M}_\alpha)\pi(\mathcal{M}_\alpha|\mathbf{Y}) = \sum_\alpha \hat{\boldsymbol{\beta}}_\alpha \pi(\mathcal{M}_\alpha|\mathbf{Y}).$$

Let $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ be the BMA predictor and $\hat{\mathbf{Y}}_\alpha = \mathbf{X}\hat{\boldsymbol{\beta}}_\alpha$ the predictor for $\mathcal{M}_\alpha$. The model $\mathcal{M}_{\hat{\alpha}}$ is said to be *predictively optimal* if

$$\mathbb{E}\|\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{\hat{\alpha}}\|^2 \leq \min_{\alpha} \mathbb{E}\|\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{\alpha}\|^2,$$

where $\|\cdot\|$ denotes the $\mathcal{L}_2$-norm and the expectation is over $\mathbf{Y}$ with respect to the BMA predictive distribution. Thus, $\mathcal{M}_{\hat{\alpha}}$ is the single model whose prediction is closest to the BMA in an $\mathcal{L}_2$-sense.

Although Barbieri and Berger (2004) illustrated several settings where the median model and highest posterior model differed, they also showed interesting examples where they coincided. Arguments given by Barbieri and Berger (2004) show in fact that spike and slab models of the form (7) are examples where the two approaches are equivalent.

THEOREM 1 (Barbieri and Berger, 2004). *The highest posterior model and the median model coincide for spike and slab models of the form (7).*

We come back to Theorem 1 later when we more closely examine the expression for the posterior inclusion probabilities (8) (in fact, we will provide an independent proof of a similar result; see Theorem 5). For now, it is enough to note the two procedures are identical and in particular that their total risks are equal. That is, $\mathcal{R}(\hat{\boldsymbol{\delta}}_H) = \mathcal{R}(\hat{\boldsymbol{\delta}}_M)$ where $\hat{\boldsymbol{\delta}}_H$ and $\hat{\boldsymbol{\delta}}_M$ denote the decision rules for the highest posterior and median models, respectively.

4. SOME PRELIMINARY ASYMPTOTIC RESULTS

To present our asymptotic findings we follow Shao (1993) by categorizing models into two types depending upon whether they contain all the nonzero values of $\boldsymbol{\beta}_0$ or not (see also Nishii, 1984). The two categories of models are

$$\text{Category I} = \{\alpha : \text{At least one nonzero component of } \boldsymbol{\beta}_0 \text{ is not in } \mathcal{M}_\alpha\},$$

$$\text{Category II} = \{\alpha : \mathcal{M}_\alpha \text{ contains all nonzero components of } \boldsymbol{\beta}_0\}.$$

Category I models are underfit incorrect models that exclude at least one nonzero coefficient. Note that a Category I model may include regressors whose coefficients are truly zero. On the other hand, Category II models consist of the true model $\mathcal{M}_{\alpha_0}$, the model with the smallest dimension, in addition to all models containing $\mathcal{M}_{\alpha_0}$ as a subset. These latter models are overfit in the sense that they include all nonzero coefficients plus at least one truly zero coefficient. In Shao (1993, 1996) it was shown that one could not consistently recover the true model by minimizing the estimated prediction error using leave-one-out cross-validation or bootstrapping. Although the method distinguishes Category I models from Category II models, and therefore does not underfit, it cannot distinguish between Category II models in the limit, and therefore it

overfits asymptotically. See also the papers by Stone (1977a, 1977b), who drew similar conclusions about cross-validation. Also see Geweke and Meese (1981); Hannan and Quinn (1979); Shibata (1976) for similar results in terms of Akaike's criteria.

Perhaps not surprisingly, the highest posterior model does not suffer from the inconsistency of the previously mentioned procedures. When Bayesians use normal hierarchical models for model selection they are often implicitly taking advantage of a BIC penalization, and this penalization makes it possible to discern the true model from other Category II models, thus avoiding overfitting asymptotically. In fact, we show in the following theorem that the highest posterior model is consistent (for more on BIC-like consistency in linear regression see Geweke and Meese, 1981; Nishii, 1984; Rao and Wu, 1989; Schwarz, 1978). However, although these results may appear promising, we find some undesirable properties.

THEOREM 2. *Assume that* $\{\varepsilon_i\}$ *are independent such that* $\mathbb{E}(\varepsilon_i) = 0$, $\mathbb{E}(\varepsilon_i^2) = \sigma^2$, *and* $\mathbb{E}(\varepsilon_i^4) \leq M$ *for some* $M < \infty$. *Assume that* $\max_{1\leq i\leq n}\|\mathbf{x}_i\|/\sqrt{n} \to 0$. *If (1) represents the data model and (7) the Bayesian model used for defining the posterior model selection procedure* $\mathcal{M}_{\hat{\alpha}}$, *then the following conditions hold.*

(i) $\mathcal{R}(\hat{\boldsymbol{\delta}}_M) = \mathcal{R}(\hat{\boldsymbol{\delta}}_H) \xrightarrow{\text{a.s.}} 0$.

(ii) If $\mathcal{M}_\alpha$ *is a Category II model and* $\mathcal{M}_{\alpha'}$ *a Category I model,*

$$\mathbb{P}\{\pi(\mathcal{M}_\alpha|\mathbf{Y}) \geq \pi(\mathcal{M}_{\alpha'}|\mathbf{Y})\} \to 1.$$

(iii) If $\mathcal{M}_\alpha$ *and* $\mathcal{M}_{\alpha'}$ *are Category I models,*

$$\mathbb{P}\{\pi(\mathcal{M}_\alpha|\mathbf{Y}) \geq \pi(\mathcal{M}_{\alpha'}|\mathbf{Y})\} \to \begin{cases} 1 & \text{if } \|\boldsymbol{\beta}_{0,\alpha}\|^2 > \|\boldsymbol{\beta}_{0,\alpha'}\|^2 \\ 0 & \text{if } \|\boldsymbol{\beta}_{0,\alpha}\|^2 < \|\boldsymbol{\beta}_{0,\alpha'}\|^2. \end{cases} \tag{9}$$

(iv) If $\mathcal{M}_\alpha$ *and* $\mathcal{M}_{\alpha'}$ *are two distinct Category II models such that* $K_\alpha = K_{\alpha'}$,

$$0 < \lim_{n\to\infty} \mathbb{P}\{\pi(\mathcal{M}_\alpha|\mathbf{Y}) \geq \pi(\mathcal{M}_{\alpha'}|\mathbf{Y})\} < 1.$$

Moreover, $\mathbb{P}\{\pi(\mathcal{M}_\alpha|\mathbf{Y}) \geq \pi(\mathcal{M}_{\alpha'}|\mathbf{Y})\} \to \frac{1}{2}$ *if* $\gamma_1 = \cdots = \gamma_K$ *and* π *is uniform.*

Part (i) of Theorem 2 shows that highest posterior model selection is asymptotically consistent, but results (iii) and (iv) suggest suboptimal finite-sample performance. Consider result (iv), which pertains to Category II models. A careful inspection of the proof of the theorem (see equation (A.1) in the proof) shows that if α is a Category II model,

$$\log(\pi(\mathcal{M}_\alpha|\mathbf{Y})) - \log(\pi(\mathcal{M}_{\alpha_0}|\mathbf{Y})) = \tfrac{1}{2}(K_{\alpha_0} - K_\alpha)\log(n) + C(\alpha, \alpha_0) + o_p(1),$$

where $C(\alpha, \alpha_0)$ is a finite constant depending upon the prior. The dominating term in the expression is a BIC-like penalty of order $\log(n)$ that ensures, at least asymptotically, that the highest posterior model will identify the true model $\mathcal{M}_{\alpha_0}$ over all other Category II models. However, for a fixed sample size n, there is a trade-off between the BIC penalty and the other terms in the expression, making it possible for $\mathcal{M}_{\hat{\alpha}}$ to equal some Category II model other than $\mathcal{M}_{\alpha_0}$.

Given this, it is of interest to consider what the distribution of $\pi(\alpha|\mathbf{Y})$ might look like over the space of Category II models. This would provide some insight into the behavior of $\mathcal{M}_{\hat{\alpha}}$ over this space. Result (iv) provides some answers. For example, the result shows that under a uniform model prior and equal hypervariances no model has a posterior probability dominating Category II models of the same dimension.

Under such a setting, this would suggest that if the highest posterior model selects a Category II model, such selection is approximately uniform from a class of Category II models of the same dimension (note that this is only an informal argument as result (iv) is not conditional on the behavior of $\mathcal{M}_{\hat{\alpha}}$). Each instance of selecting a Category II model results in excess variability due to estimating zero coefficients of the true model. One way to mitigate this effect would be to use an estimator that averages over models from the same contour. However, the highest posterior model does not take advantage of model averaging, and thus we expect it would perform suboptimally when compared to procedures that use model averaging for selection (for examples of such procedures, see Ishwaran and Rao, 2003, 2005a, 2005b).

Result (iii) is also of concern in finite samples. As is well known, BIC penalization often leads to underfitting in practice, and thus for a fixed sample size, it is more than likely that the highest posterior model will be a Category I model. Result (iii) shows that over Category I models, $\pi(\mathcal{M}_\alpha|\mathbf{Y})$ is asymptotically larger over those models with large $\mathcal{L}_2$-coefficient norms. Thus, over Category I models, the highest posterior model could easily favor small models with high signal coefficients (again, this is only an informal argument as result (iii) is not conditional on the behavior of $\mathcal{M}_{\hat{\alpha}}$). Although this might not impact prediction error performance, in terms of variable selection it is undesirable to select a small high-signal model over one with almost the same $\mathcal{L}_2$-coefficient norm containing many small nonzero coefficients. Thus, in a finite-sample setting the highest posterior model may have large total risk.

5. LOCAL ASYMPTOTICS

The tendency for the highest posterior model to favor smaller models can be studied more carefully by using a local asymptotic argument. We look at its behavior in the setting when coefficients shrink to zero at an $n^{-1/2}$ rate. This

will indicate how the highest posterior model might perform in a finite-sample setting in which all coefficients are small.

For our analysis we assume that we have a triangular array of observations. That is, we assume that the true data generating mechanism is

$$Y_{ni} = \mathbf{x}_i' \boldsymbol{\beta}_n + \varepsilon_{ni}, \qquad i = 1, \ldots, n, \tag{10}$$

where for each n, $\{\varepsilon_{ni} : i = 1, \ldots, n\}$ are independent random variables with mean 0 and variance σ^2. In (10), the true $\boldsymbol{\beta}_n$ parameter is $\boldsymbol{\beta}_{n,0} = \sigma n^{-1/2} \boldsymbol{\beta}_0$.

THEOREM 3. *Assume that for every n, $\{\varepsilon_{ni} : i = 1, \ldots, n\}$ satisfy the same conditions as $\{\varepsilon_i\}$ in Theorem 2. If (10) represents the data model, and (7) the Bayesian model used for defining the posterior model selection procedure $\mathcal{M}_{\hat{\alpha}}$, then under the same conditions for $\mathbf{x}_i$ as in Theorem 2 the following conditions hold.*

(i) For any two models $\mathcal{M}_\alpha$ and $\mathcal{M}_{\alpha'}$ such that $K_\alpha < K_{\alpha'}$,

$$\mathbb{P}\{\pi(\mathcal{M}_\alpha | \mathbf{Y}) \geq \pi(\mathcal{M}_{\alpha'} | \mathbf{Y})\} \to 1.$$

In particular, $\mathcal{R}(\hat{\boldsymbol{\delta}}_H) \xrightarrow{\text{a.s.}} K_{\alpha_0}$ because $\mathbb{P}\{\mathcal{M}_{\hat{\alpha}} = \mathcal{M}_\emptyset\} \to 1$.

(ii) If π is uniform and $\mathcal{M}_\alpha$ and $\mathcal{M}_{\alpha'}$ are any two models such that $K_\alpha = K_{\alpha'}$,

$$\lim_{n \to \infty} \mathbb{P}\{\pi(\mathcal{M}_\alpha | \mathbf{Y}) \geq \pi(\mathcal{M}_{\alpha'} | \mathbf{Y})\}$$

$$= \mathbb{P}\left\{ \sum_{k \in \alpha} (\beta_{k,0} + Z_k)^2 - \sum_{k \in \alpha} \log(\gamma_k) \right.$$

$$\left. \geq \sum_{k \in \alpha'} (\beta_{k,0} + Z_k)^2 - \sum_{k \in \alpha'} \log(\gamma_k) \right\},$$

where Z_k are independent and identically distributed (i.i.d.) $N(0,1)$ random variables.

Part (i) of Theorem 3 shows that the highest posterior model is inconsistent under a local asymptotic framework and therefore cannot be uniformly consistent. Moreover, part (i) shows that asymptotically, $\pi(\mathcal{M}_\alpha | \mathbf{Y})$ will be larger over small models, which in combination with part (iii) of Theorem 2 suggests that the highest posterior model will tend to favor small models if many of the true coefficients have small values. This favoring of smaller models coincides with what is often seen using BIC. In fact, the phenomenon (i) where the null model is preferentially selected under a sequence of shrinking alternatives is known to be true generally for any pointwise consistent model selection procedure (see Leeb and Pötscher, 2005, Rmk. 4.4). Also see Leeb and Pötscher (2007) and

Yang (2005) for more discussion on why consistent model selection procedures can perform poorly in finite samples.

Part (ii) of Theorem 3 shows over models of the same size that the posterior probability involves a delicate interplay between the true size of a coefficient and the prior hypervariance. In particular, notice that if γ_k is small for a given predictor k, then $\pi(\mathcal{M}_\alpha|\mathbf{Y})$ will become large if $k \in \alpha$, whereas if γ_k is large, then $\pi(\mathcal{M}_\alpha|\mathbf{Y})$ is much smaller. This is quite curious, and undesirable, because it is counterintuitive to the spike and slab framework. The hypervariances in spike and slab models are designed so that large γ_k values reflect a prior belief that a covariate is informative, whereas small values reflect a belief that the variable is noninformative. In fact, recall how this counterintuitive result was found earlier in Table 1. Theorem 3 provides a reasonable explanation for why this occurs.

6. RESCALED SPIKE AND SLAB MODELS

Recently Ishwaran and Rao (2005b) discussed the role shrinkage plays in spike and slab models for high-dimensional variable selection. They introduced the notion of a *rescaled spike and slab model*, arguing that such models are desirable because of their property of maintaining a shrinkage effect in the limit. The effect of rescaling was used by Ishwaran and Rao (2005b) in tandem with a spike and slab hierarchy involving continuous bimodal priors. Although Ishwaran and Rao (2005b) considered a more delicate setup than here, the impact of rescaling nevertheless still holds, and it is of interest to study this.

In the context considered here, the rescaled spike and slab version of (7) is defined as follows:

$$\begin{aligned}
(\mathbf{Y}^*|\mathbf{X},\boldsymbol{\beta}) &\sim \text{Normal}(\mathbf{X}\boldsymbol{\beta}, n\mathbf{I}),\\
(\boldsymbol{\beta}|\mathbf{b},\boldsymbol{\Gamma}) &\sim \text{Normal}(\mathbf{0}, \text{diag}\{b_1\gamma_1,\ldots,b_K\gamma_K\}),\\
(b_k|w_k) &\overset{\text{ind}}{\sim} \text{Bernoulli}(w_k), \qquad k = 1,\ldots,K,
\end{aligned} \tag{11}$$

where $\mathbf{Y}^* = \hat{\sigma}_n^{-1} n^{1/2}\mathbf{Y}$. Rescaling amounts to replacing $\mathbf{Y}$ with an $n^{1/2}$ scaled value and then replacing the variance in the hierarchical model with the value n.

6.1. Inclusion Probabilities

The following theorem provides an explicit characterization of the posterior inclusion probability for a variable. Define

$$(\xi_{1,n},\ldots,\xi_{K,n})' = \hat{\sigma}_n^{-1} n^{-1/2}\mathbf{X}'\mathbf{Y}. \tag{12}$$

This represents the signal contained in the data for the regression coefficients. Let $p_k^* = \sum_{\alpha\in\Delta_k}\pi(\mathcal{M}_\alpha|\mathbf{Y}^*)$ denote the posterior inclusion probability for k under (11).

THEOREM 4. *Under the rescaled spike and slab model (11) we have*

$$(p_k^*)^{-1} = 1 + (w_k^{-1} - 1)(\gamma_k + 1)^{1/2} \exp(-\tfrac{1}{2} d_k \xi_{k,n}^2), \qquad k = 1, \ldots, K,$$

where $d_k = \gamma_k/(\gamma_k + 1)$. *This should be compared to the posterior inclusion probability under the nonscaled model (7):*

$$p_k^{-1} = 1 + (w_k^{-1} - 1)(n\hat{\sigma}_n^{-2}\gamma_k + 1)^{1/2} \exp(-\tfrac{1}{2} d_{k,n} \xi_{k,n}^2), \qquad k = 1, \ldots, K,$$

where $d_{k,n} = \gamma_k/(\gamma_k + \hat{\sigma}_n^2/n)$.

Theorem 4 shows that p_k^* and p_k differ significantly over the zero coefficients. Under the conditions of Theorem 2, $p_k^* \xrightarrow{\text{p}} 1$ and $p_k \xrightarrow{\text{p}} 1$ if k is a nonzero coefficient, but if k corresponds to a zero coefficient, $p_k \xrightarrow{\text{p}} 0$, whereas, because the effect of γ_k does not vanish under rescaling, p_k^* has the following distributional limit:

$$p_k^* \overset{\text{d}}{\rightsquigarrow} (1 + (w_k^{-1} - 1)(\gamma_k + 1)^{1/2} \exp(-\tfrac{1}{2} d_k Z_k^2))^{-1}.$$

6.2. Asymptotic Results

From these results it is tempting to conclude that it is better to use a nonscaled model (7). However, in the following theorem we show that rescaling helps to correct some problems seen for the highest posterior model. Our proof takes advantage of a perfect cancellation of a BIC penalty appearing in the expansion of the log of the posterior model probability. Although the trade-off is that the highest posterior model selected is no longer risk consistent, the flip side is that the method no longer breaks down in the local asymptotic case.

THEOREM 5. *Under the Bayesian model (11) the following conditions hold.*

(i) The highest posterior model and the median model are equivalent.

(ii) Fixed-parameter asymptotics: results (ii) and (iii) for Theorem 2 hold, where in (iii) replace $\boldsymbol{\beta}_0$ *by* $\mathbf{D}^{1/2}\boldsymbol{\beta}_0$ *for* $\mathbf{D} = \operatorname{diag}\{d_1, \ldots, d_K\}$.

(iii) Local asymptotics: under the conditions of Theorem 3, if π *is uniform and* $\mathcal{M}_\alpha$ *and* $\mathcal{M}_{\alpha'}$ *are any two models,*

$$\lim_{n\to\infty} \mathbb{P}\{\pi(\mathcal{M}_\alpha|\mathbf{Y}^*) \geq \pi(\mathcal{M}_{\alpha'}|\mathbf{Y}^*)\}$$

$$= \mathbb{P}\left\{ \sum_{k\in\alpha} (d_k(\beta_{k,0} + Z_k)^2 - \log(\gamma_k + 1)) \right.$$

$$\left. \geq \sum_{k\in\alpha'} (d_k(\beta_{k,0} + Z_k)^2 - \log(\gamma_k + 1)) \right\}.$$

Parts (i) and (ii) are essentially similar to the nonscaled case. It is under the local asymptotic setting, however, where we start to see differences between the two methods. Part (iii) shows that the highest posterior model no longer concentrates on the null model; in fact, notice the presence of d_k in the limiting probability on the right. If the true value of a coefficient is zero, and γ_k is selected moderately small, then the value $Z_k^2 = (\beta_{k,0} + Z_k)^2$ is appropriately down weighted by the value d_k. The presence of d_k, therefore, discourages selection of models with zero coefficients. Notice also the presence of $\log(\gamma_k + 1)$ appearing in the limit on the right. This differs from the nonscaled case where a $\log(\gamma_k)$ term was found. Now a small hypervariance $0 < \gamma_k < 1$, which might be associated with a coefficient thought to be noninformative, no longer heavily favors selection of models with $k \in \alpha$.

6.3. The Impact of Rescaling

The preceding argument implicitly applies to a paradigm where coefficients are shrinking to zero. To further study the relationship between the prior and highest posterior model selection under rescaling, we exploit the equivalence between the highest posterior model and the median model. By Theorem 4, the posterior inclusion probability satisfies $p_k^* \geq \frac{1}{2}$ if and only if $\xi_{k,n}^2 \geq \psi^*(\gamma_k, w_k)$, where

$$\psi^*(\gamma, w) = \frac{2(\gamma+1)}{\gamma} \log\left(\frac{1}{w} - 1\right) + \frac{\gamma+1}{\gamma} \log(\gamma+1).$$

To understand how p_k^* varies in terms of these two parameters, we plotted the contour values of $\psi^*(\gamma, w)$ over a range of γ and w values. See Figure 1.

The thick line superimposed on the figure identifies the contour values for the function $\psi^*(\cdot, 0.5)$, which corresponds to the use of a uniform prior for π. For example, note how the line lies between the values of 2 and 3 when $\gamma = 10$, whereas its value is approximately 1 when γ is small. Compare this to the nonscaled setting. By Theorem 4, a predictor k is selected if and only if $\xi_{k,n}^2 \geq \psi(\gamma_k, w_k, n, \hat{\sigma}_n^2)$, where

$$\psi(\gamma, w, n, \sigma^2) = \frac{2(\gamma + \sigma^2/n)}{\gamma} \log\left(\frac{1}{w} - 1\right) + \frac{\gamma + \sigma^2/n}{\gamma} \log(n\sigma^{-2}\gamma + 1).$$

Figure 2 plots the contour values of this function over γ and w for $n = 50$ and $\sigma^2 = 1$. The two plots confirm the mitigation of underfitting effects due to rescaling. Even for the small sample size illustrated in Figure 2, one can see how much larger the contour values are compared to Figure 1. For example, under a flat prior, the contour value for small γ values is more than 3 as compared to near 1 under rescaling. It is clear that nonscaled highest posterior model selection will tend to favor smaller models.

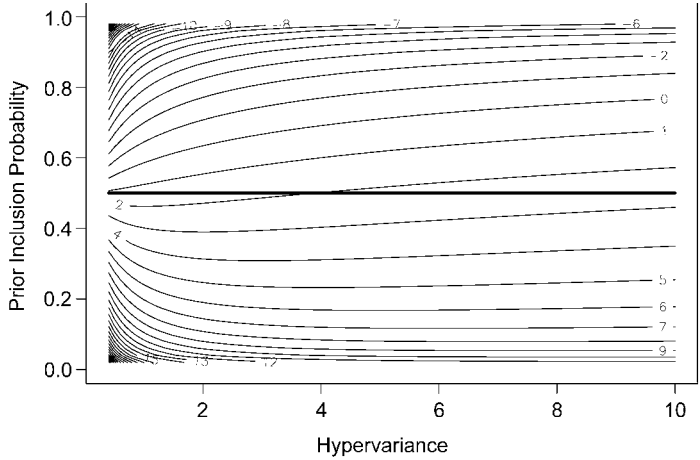

FIGURE 1. Contour plot of $\psi^*(\gamma, w)$ as a function of the hypervariance, γ, and prior inclusion probability, w. Note that $\psi^*(\gamma, w)$ is positive for small values of w and becomes negative as w increases to one. The thick line is horizontal value $w = \frac{1}{2}$ and identifies contour values for the function $\psi^*(\cdot, 0.5)$.

Rescaling also allows the γ_k to be set in a more intuitive way. Consider what happens if $w_k = \frac{1}{2}$. If $\beta_{k,0}$ is nonzero, then using a large value for γ_k shows that k will be selected with high probability because $\xi^2_{k,n}$ will eventually be larger than $\psi^*(\gamma_k, 0.5)$, which will be relatively small in comparison. This is exactly

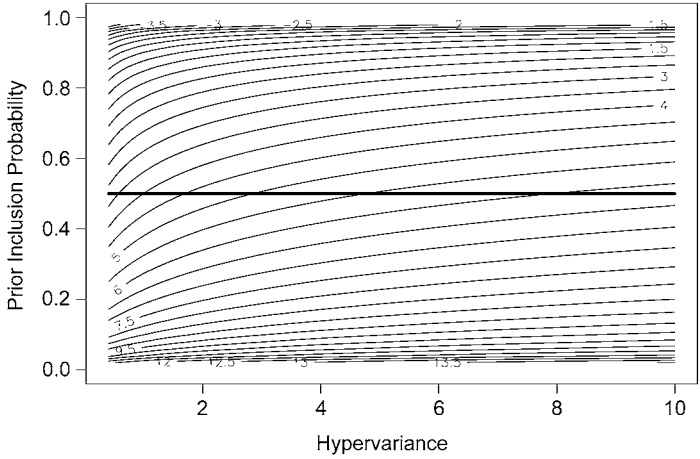

FIGURE 2. Contour plot of $\psi(\gamma, w, n, \sigma^2)$ for $n = 50$ and $\sigma^2 = 1$. The thick line is horizontal value $w = \frac{1}{2}$ and identifies contour values for the function $\psi^*(\cdot, 0.5, n, \sigma^2)$.

what we would like, because large γ_k values are used to elicit a belief that a covariate is informative.

On the other hand, consider what happens if $\beta_{k,0} = 0$. Now if we select γ_k to be small, $\psi^*(\gamma_k, 0.5) \approx 1$. Thus, if $\beta_{k,0} = 0$, the predictor k will be selected with asymptotic probability $\mathbb{P}\{\chi_1^2 \geq 1\} \approx 0.32$. Although this appears too large, and brings into doubt whether γ_k should be chosen this way, one should note if $\beta_{k,0}$ is nonzero, but very small, then k will be selected with roughly the same probability. Thus, although one pays a price in overfitting if the coefficient is truly zero, there is a built-in robustness to misspecification. In fact, this is exactly the same effect discussed in Section 6.2 that keeps the rescaled highest posterior model from breaking down in a local asymptotic setting.

Figure 1 also indicates how the highest posterior model might be affected by the underlying size of the model. Consider the case when π is chosen such that $w_k = w$ for some hyperparameter $w > 0$ reflecting overall size or model complexity. The preceding discussion shows how to choose γ_k in a meaningful way when $w = \frac{1}{2}$. Namely, we use small values for covariates we expect to be noninformative and large values for informative covariates. This type of calibration also applies when the value for $w < \frac{1}{2}$ and the model is anticipated to be sparse. In fact, now the use of small γ_k values to indicate noninformative predictors makes even more sense. On the other hand, if $w > \frac{1}{2}$ is large, then we run into calibration problems. Now $\psi^*(\gamma, w)$ can be negative for *all* values of γ. Clearly overfitting will occur, and there is no sensible way of selecting γ_k. This suggests that highest posterior model selection might operate best under rescaling only under sparse settings.

7. BREIMAN SIMULATIONS

In this section we use simulations to study the empirical performance of highest posterior model selection under both the scaled and nonscaled settings. Our simulations followed those used by Breiman (1992). Specifically, data were generated by taking ε_i to be i.i.d. $N(0,1)$ variables, and covariates $\mathbf{x}_i$ were simulated independently from a multivariate normal distribution such that $\mathbb{E}(x_{i,k}) = 0$ and $\mathbb{E}(x_{i,j}x_{i,k}) = \rho^{|j-k|}$, where $0 < \rho < 1$ represents a correlation parameter. All simulations involved $K = 14$ predictors with a sample size of $n = 50$.

We considered four sets of simulations, (A), (B), (A′), and (B′), reflecting different correlation values ρ and different values for β_k. For simulation (A), 12 of the 14 predictors were chosen to be nonzero. All coefficients were set to the same small value, this being chosen such that the theoretical R^2 value for the model was 0.5 (for a discussion of this point, see Breiman, 1992). Thus, simulation (A) reflects a setting involving many nonzero weak predictors. In contrast, simulation (B) was designed to reflect a setting with few nonzero strong predictors. Of the 14 predictors, only six were chosen to be nonzero. Of these, two had the same large value; the other four had the same medium-sized value. The size of coefficients was selected so that the large coefficients

were 2.4 times larger than the medium-sized ones and such that the theoretical $R^2 = 0.75$. Both simulations (A) and (B) were based on a correlation of $\rho = 0$, thus reflecting an orthogonal design setting. On the other hand, simulations (A′) and (B′) used $\rho = 0.9$, reflecting a highly correlated design. Excepting this, simulation (A′) was the same as (A), and simulation (B′) was the same as (B).

Each of the four simulations was repeated 1,000 times independently. For each experiment we kept track of the false discovery rate (FDR), false nondiscovery rate (FNR) and total risk performance (TotalRisk) of a procedure. The FDR and FNR are the false discovery and false nondiscovery rates defined as the false positive and false negative rates for those coefficients identified as nonzero and zero, respectively. The TotalRisk equals the total number of falsely identified nonzero coefficients and falsely identified zero coefficients. Also computed was the prediction error performance (PE) of a procedure. This was defined as the mean square error of the estimated predictor computed over the original data when compared to a freshly drawn set of Y_i responses using the original $\mathbf{x}_i$ values.

Highest posterior model selection (both scaled and nonscaled) was investigated under a uniform model prior for π. Hypervariances were set to $\gamma_k = \gamma$, where three different values for γ were considered: $\gamma = 1, 10, n$ (this is similar to what was done in Table 1). Additionally, we introduced a hybrid procedure based on preselected hypervariances. Prior to estimating the highest posterior model we ran the spike and slab Gibbs sampler outlined in Ishwaran and Rao (2005b). This was done using both scaled and nonscaled responses. We then used the posterior means of γ_k estimated from the Gibbs sampler as the hypervariances used for highest posterior model selection. Note that a different set of hypervariances and a different set of highest posterior models were estimated for the scaled and nonscaled settings. Finally, in addition to examining highest posterior model selection, we also kept track of models selected using BIC and AIC penalization. These are the models with lowest BIC and AIC penalties over all models. The results for all these procedures are recorded in Table 2. Values reported are averaged values from the 1,000 simulations.

Our conclusions are summarized as follows:

1. In simulations (A) and (A′), total risk performance for highest posterior model selection, and also for other measures of performance, degrades as γ increases. This is true for both the scaled and nonscaled cases. This is because the simulations reflect a setting where almost all coefficients are relatively small and so the optimal way to select γ would be to choose a small value. If γ is large, estimated models are too small, and performance suffers. Notice also the mitigation effects of rescaling. This is seen by the better total risk performance due to a lower FNR brought on by a less aggressive tendency to underfit—particularly evident here because of the weak underlying signal.

TABLE 2. Breiman simulations

	Simulation (A)				Simulation (A′)			
	FDR	FNR	TotalRisk	PE	FDR	FNR	TotalRisk	PE
				Many nonzero weak predictors				
BIC	0.031	0.782	6.962	1.526	0.085	0.850	10.290	1.201
AIC	0.051	0.724	4.931	1.409	0.111	0.851	9.482	1.218
$\mathcal{M}^*_{\hat{\alpha},1}$	0.057	0.665	3.669	1.359	0.085	0.736	4.254	1.167
$\mathcal{M}^*_{\hat{\alpha},10}$	0.040	0.742	5.559	1.444	0.091	0.845	9.373	1.164
$\mathcal{M}^*_{\hat{\alpha},n}$	0.031	0.781	6.972	1.527	0.083	0.847	10.051	1.197
$\mathcal{M}^*_{\hat{\alpha},\bullet}$	0.045	0.721	4.873	1.408	0.097	0.841	8.785	1.150
$\mathcal{M}_{\hat{\alpha},1}$	0.030	0.782	7.008	1.525	0.082	0.847	10.070	1.196
$\mathcal{M}_{\hat{\alpha},10}$	0.019	0.812	8.555	1.627	0.055	0.847	10.507	1.206
$\mathcal{M}_{\hat{\alpha},n}$	0.014	0.823	9.314	1.684	0.041	0.847	10.666	1.223
$\mathcal{M}_{\hat{\alpha},\bullet}$	0.045	0.743	5.446	1.432	0.101	0.848	9.603	1.182
	Simulation (B)				Simulation (B′)			
	FDR	FNR	TotalRisk	PE	FDR	FNR	TotalRisk	PE
				Few nonzero strong predictors				
BIC	0.091	0.135	1.800	1.287	0.167	0.351	4.570	1.204
AIC	0.195	0.088	2.135	1.274	0.342	0.363	5.185	1.239
$\mathcal{M}^*_{\hat{\alpha},1}$	0.353	0.102	3.754	1.286	0.494	0.022	5.993	1.228
$\mathcal{M}^*_{\hat{\alpha},10}$	0.151	0.101	1.897	1.271	0.245	0.266	3.735	1.164
$\mathcal{M}^*_{\hat{\alpha},n}$	0.090	0.134	1.792	1.287	0.208	0.338	4.468	1.199
$\mathcal{M}^*_{\hat{\alpha},\bullet}$	0.177	0.088	1.963	1.265	0.263	0.258	3.755	1.164
$\mathcal{M}_{\hat{\alpha},1}$	0.086	0.133	1.758	1.284	0.209	0.340	4.494	1.200
$\mathcal{M}_{\hat{\alpha},10}$	0.041	0.182	2.057	1.334	0.105	0.345	4.353	1.192
$\mathcal{M}_{\hat{\alpha},n}$	0.026	0.214	2.365	1.382	0.076	0.344	4.280	1.187
$\mathcal{M}_{\hat{\alpha},\bullet}$	0.121	0.114	1.789	1.273	0.273	0.342	4.693	1.209

Note: BIC and AIC are models selected using BIC and AIC penalization; $\mathcal{M}^*_{\hat{\alpha},\gamma}$ and $\mathcal{M}_{\hat{\alpha},\gamma}$ are highest posterior scaled and nonscaled models, respectively, where $\gamma_k = \gamma$; $\mathcal{M}^*_{\hat{\alpha},\bullet}$ and $\mathcal{M}_{\hat{\alpha},\bullet}$ are highest posterior scaled and nonscaled models using estimated hypervariances.

2. The opposite effect is seen for simulations (B) and (B′) under rescaling. Now highest posterior model performance improves as γ increases. This is because in these sets of simulations only six coefficients are nonzero, two being large, the other four nonzero coefficients having moderate size. As γ increases estimated models become small and approach the true model. This backfires for nonscaled models though, because sparse models are approached very rapidly as γ becomes larger. Performance degrades in this setting because models are just too small and false negative rates

too high. This behavior could have been predicted from Figure 2, where contour values under a uniform prior (i.e., the thick line) can be seen to increase quickly with increasing γ. As a result, only a very strong signal can be detected.

3. Total risk performance and PE for rescaled highest posterior model selection is better than in nonscaled selection in almost all examples. This is true when comparing models for the same hypervariance or when comparing models based on preselected hypervariances.
4. Total risk performance and PE for rescaled highest posterior model under preselected hypervariances, $\mathcal{M}^*_{\hat{\alpha},\bullet}$, is better than AIC and BIC in almost all examples. The exception occurs in (B), where BIC is best with respect to total risk. In this particular scenario, BIC appears to discourage overfitting enough without inducing too much underfitting. Interestingly, $\mathcal{M}^*_{\hat{\alpha},\bullet}$ is highly competitive in PE with BIC even in this situation. It is also interesting to note that BIC is no longer better than $\mathcal{M}^*_{\hat{\alpha},\bullet}$ in the correlated case, (B′).
5. The results for simulations (A′) and (B′) are in tune with those found for (A) and (B). This suggests that our theory may also apply to correlated settings, at least approximately.

8. DISCUSSION

In this paper we have shown rigorously that underfitting using the highest posterior model occurs because of a BIC-like penalization term appearing in the log-posterior model probability. Various limiting probabilities were given in simple closed form expressions to help quantify what this means in practice. These kinds of specific details are not often found in the literature (one nice example is Zhang, 1993b, where rates at which over- and underfitting occur were established within the context of fixed-penalty information criteria). Our results show that underfitting is much worse than first thought—that the highest posterior model will tend to favor the smallest models with high signal coefficients among the group of underfit models for finite samples. This was further amplified by a local asymptotic argument that showed that the highest posterior model will eventually concentrate on the null model in cases where true coefficients are all weak.

We proposed a remedy for this by rescaling the responses and suitably altering the Bayesian hierarchy. Highest posterior model selection under the resulting rescaled spike and slab models, although no longer risk consistent, will be more evenly balanced, and the problems of underfitting will be substantially mitigated. In addition, calibration of the γ_k can be done in a way that is intuitively consistent with the spike and slab hierarchy. Interestingly though, under nonsparse prior settings, things break down because there is no sensible way to calibrate the γ_k. Here, regardless of how the γ_k are set, overfitting will be the result. This suggests that there is a disconnection between the general use of

the highest posterior model for model selection and the manner in which hyperparameters in the prior are chosen.

One promising alternative that does not suffer from these problems is model selection based on the posterior mean from a rescaled spike and slab hierarchical model. Specifically, let $\hat{\boldsymbol{\beta}}_n^* = (\hat{\beta}_{1,n}^*, \ldots, \hat{\beta}_{K,n}^*)'$ denote the posterior mean for $\boldsymbol{\beta}$ from (11). Taking advantage of conjugacy and orthogonality, one can show that

$$\mathbb{E}(\beta_k|\mathbf{Y}^*, \mathcal{M}_\alpha) = \begin{cases} \dfrac{\gamma_k}{1+\gamma_k}\,\xi_{k,n} & \text{if } k \in \alpha \\ 0 & \text{if } k \notin \alpha. \end{cases}$$

Therefore,

$$\begin{aligned} \hat{\beta}_{k,n}^* &= \sum_{\alpha \in \Delta_k} \mathbb{E}(\beta_k|\mathbf{Y}^*, \mathcal{M}_\alpha)\,\pi(\mathcal{M}_\alpha|\mathbf{Y}^*) \\ &= \frac{\gamma_k}{1+\gamma_k}\, p_k^*\, \xi_{k,n}, \qquad k = 1, \ldots, K. \end{aligned} \tag{13}$$

Proper calibration of priors is now no longer difficult, even for the nonsparse setting. For example, using the notation of Section 6.3, suppose that the model complexity value is $w > \frac{1}{2}$. If γ_k is chosen to be small for zero coefficients and large for nonzero coefficients, then p_k^* will be nearly one for all coefficients. However, the posterior mean, because it includes the shrinkage effect $\gamma_k/(\gamma_k + 1)$, will shrink the value of p_k^* and consequently will be *small* for zero coefficients while still being large for nonzero coefficients.

Thus, selection based on using the posterior mean will be free from calibration problems in all settings. In fact, effective strategies for model selection using the posterior mean from rescaled spike and slab models have already been developed (Ishwaran and Rao, 2005b). These operate by extending the hierarchy to include a continuous bimodal prior for the γ_k and provide automatic adaptive estimation of the γ_k (recall how the effectiveness of such values was demonstrated in the Breiman simulations). The resulting model averaged posterior mean estimates coupled with hard thresholding are then used for selection. A detailed theoretical treatment and empirical validation supporting the use of the posterior mean can be found in Ishwaran and Rao (2003, 2005a, 2005b) for the interested reader.

REFERENCES

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In B.N. Petrov & F. Csaki (eds.), *Proceedings of the Second International Symposium on Information Theory*, pp. 267–281. Akademia Kiado.

Atkinson, A. (1978) Posterior probabilities for choosing a regression model. *Biometrika* 65, 39–48.

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

Bernardo, J. & A. Smith (1994) *Bayesian Theory*. Wiley.
Breiman, L. (1992) The little bootstrap and other methods for dimensionality selection in regression: *X*-fixed prediction error. *Journal of the American Statistical Association* 87, 738–754.
Chipman, H., E. George, & R. McCulloch (2001) The practical implementation of Bayesian model selection. In P. Lahiri (ed.), *Model Selection*, IMS Monograph 38, pp. 67–116. IMS.
Clyde, M., H. DeSimone, & G. Parmigiani (1996) Prediction via orthogonalized model mixing. *Journal of the American Statistical Association* 91, 1197–1208.
Clyde, M., G. Parmigiani, & B. Vidakovic (1998) Multiple shrinkageand subset selection in wavelets. *Biometrika* 85, 391–402.
Gelfand, A.E., D. Dey, & H. Chang (1992) Model determination using predictive distributions with implementations via sampling-based methods. In J.M. Bernardo, J.O. Berger, A.P. Dawid, & A.F.M. Smith (eds.), *Bayesian Statistics*, vol. 4, pp. 147–167. Oxford University Press.
George, E. & R. McCulloch (1993) Variable selection via Gibbs sampling. *Journal of the American Statistical Association* 88, 881–889.
George, E. & R. McCulloch (1997) Approaches for Bayesian variable selection. *Statistica Sinica* 7, 339–373.
Geweke, J. & R. Meese (1981) Estimating regression models of finite but unknown order. *International Economic Review* 22, 55–70.
Hannan, E. & B. Quinn (1979) The determination of the order of an autoregression. *Journal of the Royal Statistical Society, Series B* 41, 190–195.
Hoeting, J.A., D. Madigan, A.E. Raftery, & C.T. Volinsky (1999) Bayesian model averaging: A tutorial. *Statistical Science* 14, 382–417.
Ishwaran, H. & J. Rao (2003) Detecting differentially expressed genes in microarrays using Bayesian model selection. *Journal of the American Statistical Association* 98, 438–455.
Ishwaran, H. & J. Rao (2005a) Spike and slab gene selection for multigroup microarray data. *Journal of the American Statistical Association* 100, 764–780.
Ishwaran, H. & J. Rao (2005b) Spike and slab variable selection: Frequentist and Bayesian strategies. *Annals of Statistics* 33, 730–773.
Leeb, H. & B. Pötscher (2007) Sparse Estimators and the Oracle Property, or the Return of Hodges' Estimator. April. Cowles Foundation Discussion Paper no. 1500.
Leeb, H. & B. Pötscher (2005) Model selection and inference: Facts and fiction. *Econometric Theory* 21, 21–59.
Mitchell, T. & J. Beauchamp (1988) Bayesian variable selection in linear regression. *Journal of the American Statistical Association* 83, 1023–1036.
Nishii, R. (1984) Asymptotic properties of criteria for selection of variables in multiple regression. *Annals of Statistics* 12, 758–765.
Rao, C. & Y. Wu (1989) A strongly consistent procedure for model selection in a regression problem. *Biometrika* 76, 369–374.
Schwarz, G. (1978) Estimating the dimension of a model. *Annals of Statistics* 6, 461–464.
Serfling, R. (2002) *Approximation Theorems of Mathematical Statistics*. Wiley.
Shao, J. (1993) Linear model selection by cross-validation. *Journal of the American Statistical Association* 88, 486–494.
Shao, J. (1996) Bootstrap model selection. *Journal of the American Statistical Association* 91, 655–665.
Shibata, R. (1976) Selection of the order of an autoregressive model by Akaike's information. *Biometrika* 63, 117–126.
Smith, M. & R. Kohn (1996) Nonparametric regression using Bayesian variable selection. *Journal of Econometrics* 75, 317–344.
Stone, M. (1977a) An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society, Series B* 39, 44–47.
Stone, M. (1977b) Asymptotics for and against cross-validation. *Biometrika* 64, 29–35.
Yang, Y. (2005) Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika* 92, 937–950.

Zhang, P. (1993a) Model selection via multifold cross validation. *Annals of Statistics* 21, 299–313.
Zhang, P. (1993b) On the convergence rate of model selection criteria. *Communications in Statistical Theory and Methods* 22, 2765–2775.

APPENDIX A: Proofs

Proof of Theorem 2. With some algebra, and using $\mathbf{X}'_\alpha \mathbf{X}_\alpha = n\mathbf{I}_\alpha$, one can show for each nonnull model $\mathcal{M}_\alpha$,

$$f(\mathbf{Y}|\mathcal{M}_\alpha) = C_\alpha \exp\left(\frac{1}{2}\sum_{k\in\alpha} d_{k,n}\xi_{k,n}^2\right),$$

where $\xi_{k,n}$ is defined as in (12), $d_{k,n}$ is defined as in Theorem 4, and

$$C_\alpha = (2\pi)^{-n/2} n^{-K_\alpha/2} \hat{\sigma}_n^{-n+K_\alpha} \exp\left(-\frac{1}{2\hat{\sigma}_n^2}\mathbf{Y}'\mathbf{Y}\right) \prod_{k\in\alpha} (\gamma_k + \hat{\sigma}_n^2/n)^{-1/2}.$$

Therefore, if $\mathcal{M}_\alpha$ and $\mathcal{M}_{\alpha'}$ are two distinct nonnull models,

$$\frac{f(\mathbf{Y}|\mathcal{M}_\alpha)}{f(\mathbf{Y}|\mathcal{M}_{\alpha'})} = \left(\frac{\hat{\sigma}_n^2}{n}\right)^{(K_\alpha - K_{\alpha'})/2} \frac{\prod_{k\in\alpha} (\gamma_k + \hat{\sigma}_n^2/n)^{-1/2}}{\prod_{k\in\alpha'} (\gamma_k + \hat{\sigma}_n^2/n)^{-1/2}}$$

$$\times \exp\left(\frac{1}{2}\sum_{k\in\alpha} d_{k,n}\xi_{k,n}^2 - \frac{1}{2}\sum_{k\in\alpha'} d_{k,n}\xi_{k,n}^2\right).$$

Taking logs, and keeping in mind $\hat{\sigma}_n^2 \xrightarrow{\text{p}} \sigma^2 > 0$, it follows that

$$\begin{aligned}
&\log(\pi(\mathcal{M}_\alpha|\mathbf{Y})) - \log(\pi(\mathcal{M}_{\alpha'}|\mathbf{Y})) \\
&\quad = \frac{1}{2}(K_{\alpha'} - K_\alpha)\log(n) + \frac{1}{2}\sum_{k\in\alpha} d_{k,n}\xi_{k,n}^2 - \frac{1}{2}\sum_{k\in\alpha'} d_{k,n}\xi_{k,n}^2 \\
&\qquad + \frac{1}{2}(K_\alpha - K_{\alpha'})\log(\hat{\sigma}^2) + \frac{1}{2}\sum_{k\in\alpha'} \log(\gamma_k) - \frac{1}{2}\sum_{k\in\alpha} \log(\gamma_k) \\
&\qquad + \log(\pi(\mathcal{M}_\alpha)) - \log(\pi(\mathcal{M}_{\alpha'})) + O_p(n^{-1}). \qquad \textbf{(A.1)}
\end{aligned}$$

Notice that the first term on the right-hand side represents a BIC penalty that penalizes larger models by a factor of $\log(n)$. There is a trade-off in size, though, involving the sums containing $\xi_{k,n}^2$. These terms can be evaluated using a central limit theorem. In particular, it follows that if $\beta_{k,0} = 0$, then $\xi_{k,n}^2 \xrightarrow{\text{d}} \chi_1^2$, whereas if $\beta_{k,0} \neq 0$, we have $\xi_{k,n} = \hat{\sigma}_n^{-1}(n^{1/2}\beta_{k,0} + O_p(1))$. Thus over the nonzero coefficients, $\xi_{k,n}^2$ is order $O_p(n)$, which becomes the dominant term in (A.1). Because of this, it is clear that each Category II model $\mathcal{M}_\alpha$ will have exponentially larger posterior probability than any Category I model $\mathcal{M}_{\alpha'}$, and thus (ii) follows (for the case of the null model $\alpha' = \emptyset$, simply

substitute 0 for $K_{\alpha'}$ and remove all sums involving α' in (A.1)). Over Category II models,

$$\frac{1}{2}\sum_{k\in\alpha} d_{k,n}\xi_{k,n}^2 - \frac{1}{2}\sum_{k\in\alpha'} d_{k,n}\xi_{k,n}^2$$

$$= \frac{1}{2}\sum_{k\in\alpha\cap\alpha_0^c} \xi_{k,n}^2 - \frac{1}{2}\sum_{k\in\alpha'\cap\alpha_0^c} \xi_{k,n}^2 + O_p(n^{-1}). \tag{A.2}$$

Because (A.2) only involves zero coefficients, the term is order $O_p(1)$, and thus the BIC penalty becomes the dominant term in (A.1). Because this penalty penalizes larger models, deduce (i). Furthermore, if $K_\alpha = K_{\alpha'}$, and if π is uniform and $\gamma_1 = \cdots = \gamma_K$, then (A.1) reduces to (A.2) over Category II models. If $\mathcal{M}_\alpha$ and $\mathcal{M}_{\alpha'}$ are distinct, the sum on the right-hand side of (A.2) converges in distribution to $\frac{1}{2}$ of the difference between two i.i.d. χ_k^2 variables (with $2k$ equaling the total number of distinct zero coefficients in both models). From this deduce the second part of (iv). The first part of (iv) follows because (A.1) reduces to (A.2) plus an incidental finite constant if π and γ_k are not necessarily uniform.

Finally to prove (iii), observe that $d_{k,n}\xi_{k,n}^2 = \hat{\sigma}_n^{-2} n\beta_{k,0}^2 + O_p(n^{1/2})$ if $\beta_{k,0} \neq 0$. These variables represent the dominating terms in working out the desired probability and are of order n. Dividing (A.1) throughout by $\hat{\sigma}_n^{-2} n$, and using $\hat{\sigma}_n^2 \xrightarrow{\text{p}} \sigma^2 > 0$, it follows that the event on the left-hand side of (9) occurs if and only if

$$\sum_{k\in\alpha\cap\alpha_0} \beta_{k,0}^2 \geq \sum_{k\in\alpha'\cap\alpha_0} \beta_{k,0}^2 + O_p(n^{-1/2}) + \text{smaller order terms.}$$ ■

Proof of Theorem 3. A triangular central limit theorem (Serfling, 2002) shows under the local asymptotic setting (10) that $\xi_{k,n} = \sigma\beta_{k,0}/\hat{\sigma}_n + Z_{k,n}$ where $Z_{k,n} \xrightarrow{\text{d}} Z_k$. Also, under (10), identity (A.1) continues to hold. Thus, whenever $K_\alpha \neq K_{\alpha'}$, the BIC penalty is the dominant term in (A.1), and part (i) follows. If $K_\alpha = K_{\alpha'}$, the BIC term disappears. Part (ii) follows upon using $\xi_{k,n} \xrightarrow{\text{d}} \beta_{k,0} + Z_k$. ■

Proof of Theorem 4. By definition,

$$p_k = \frac{\sum_{\alpha\in\Delta_k} \pi(\mathcal{M}_\alpha|\mathbf{Y})}{\sum_{\alpha\in\Delta_k} \pi(\mathcal{M}_\alpha|\mathbf{Y}) + \sum_{\alpha\notin\Delta_k} \pi(\mathcal{M}_\alpha|\mathbf{Y})}$$

$$= \left(1 + \frac{\sum_{\alpha\notin\Delta_k} \pi(\mathcal{M}_\alpha|\mathbf{Y})}{\sum_{\alpha\in\Delta_k} \pi(\mathcal{M}_\alpha|\mathbf{Y})}\right)^{-1}. \tag{A.3}$$

In the proof of Theorem 2 we have shown for each nonnull model $\mathcal{M}_\alpha$,

$$f(\mathcal{M}_\alpha|\mathbf{Y}) = Cn^{-K_\alpha/2}\hat{\sigma}_n^{K_\alpha} \exp\left(\frac{1}{2}\sum_{k'\in\alpha} d_{k',n}\xi_{k',n}^2\right) \prod_{k'\in\alpha} (\gamma_{k'} + \hat{\sigma}_n^2/n)^{-1/2},$$

where C is a constant independent of α (in fact $C = \pi(\mathcal{M}_\emptyset|\mathbf{Y})$). Assume that $\alpha \in \Delta_k$. Let $\alpha_{-k} = \alpha - \{k\}$. Removing the contribution from the variable k, we have

$$f(\mathcal{M}_\alpha|\mathbf{Y}) = (n\hat{\sigma}_n^{-2}\gamma_k + 1)^{-1/2} \exp(\tfrac{1}{2}d_{k,n}\xi_{k,n}^2) f(\mathcal{M}_{\alpha_{-k}}|\mathbf{Y}).$$

Furthermore, because of the special nature of the prior, if we extract the contribution of k, we have $\pi(\mathcal{M}_\alpha) = w_k(1 - w_k)^{-1}\pi(\mathcal{M}_{\alpha_{-k}})$, and therefore

$$\pi(\mathcal{M}_\alpha|\mathbf{Y}) = w_k(1 - w_k)^{-1}(n\hat{\sigma}_n^{-2}\gamma_k + 1)^{-1/2} \exp(\tfrac{1}{2}d_{k,n}\xi_{k,n}^2)\pi(\mathcal{M}_{\alpha_{-k}}|\mathbf{Y}). \quad \textbf{(A.4)}$$

There is a 1:1 mapping between Δ_k and Δ_k^c. In particular, observe that

$$\sum_{\alpha\in\Delta_k} \pi(\mathcal{M}_{\alpha-k}|\mathbf{Y}) = \sum_{\alpha\notin\Delta_k} \pi(\mathcal{M}_\alpha|\mathbf{Y}).$$

Hence, if we extract the term in (A.4) depending upon k, the ratio of the sums in (A.3) simplifies to

$$R_k = (w_k^{-1} - 1)(n\hat{\sigma}_n^{-2}\gamma_k + 1)^{1/2} \exp(-\tfrac{1}{2}d_{k,n}\xi_{k,n}^2).$$

Therefore, $p_k^{-1} = 1 + R_k$. By carefully tracking back through the arguments, substituting n for $\hat{\sigma}_n^2$ and $\mathbf{Y}^*$ for $\mathbf{Y}$, one can deduce a similar result for p_k^*. ■

Proof of Theorem 5. We first prove (i) by showing that the posterior model probability factorizes into a product of the underlying posterior inclusion probabilities. To do so we will make use of the product rule of probability. For convenience we will use products hereafter to indicate intersection of sets. Let Δ be any set of models such that $\Delta_k \cap \Delta \neq \emptyset$ and $\Delta_k^c \cap \Delta \neq \emptyset$. We have

$$\pi(\alpha \in \Delta_k|\alpha \in \Delta, \mathbf{Y}^*) = \frac{\sum_{\alpha\in\Delta_k\cap\Delta} \pi(\mathcal{M}_\alpha|\mathbf{Y}^*)}{\sum_{\alpha\in\Delta_k\cap\Delta} \pi(\mathcal{M}_\alpha|\mathbf{Y}^*) + \sum_{\alpha\in\Delta_k^c\cap\Delta} \pi(\mathcal{M}_\alpha|\mathbf{Y}^*)}$$

$$= \left(1 + \frac{\sum_{\alpha\in\Delta_k^c\cap\Delta} \pi(\mathcal{M}_\alpha|\mathbf{Y}^*)}{\sum_{\alpha\in\Delta_k\cap\Delta} \pi(\mathcal{M}_\alpha|\mathbf{Y}^*)}\right)^{-1}.$$

By arguing as in the proof of Theorem 4 (cf. eqn. (A.3)), one can show this equals p_k^*. The other relevant case for Δ occurs when $\Delta_k \cap \Delta = \{k\}$ and $\Delta_k^c \cap \Delta = \emptyset$. In other words, $\Delta = \{\emptyset, \{k\}\}$. Let R_k^* be the value of R_k when n is substituted for $\hat{\sigma}_n^2$ and d_k for $d_{k,n}$. Then, arguing as before, it is not hard to see that

$$\pi(\alpha \in \Delta_k|\alpha \in \Delta, \mathbf{Y}^*) = \left(1 + \frac{\pi(\mathcal{M}_\emptyset|\mathbf{Y}^*)}{\pi(\mathcal{M}_\emptyset|\mathbf{Y}^*)/R_k^*}\right)^{-1} = p_k^*.$$

From this, and the product rule of probability, deduce that for any model

$$\pi(\mathcal{M}_\alpha|\mathbf{Y}^*) = \pi\left(\prod_{k\in\alpha}\Delta_k\prod_{k\notin\alpha}\Delta_k^c|Y^*\right) = \prod_{k\in\alpha}p_k^*\prod_{k\notin\alpha}(1-p_k^*).$$

It is clear that the median model is the model that maximizes the posterior model probability.

Now to prove (ii) and (iii). Using arguments similar to Theorem 2 deduce

$$\begin{aligned}&\log(\pi(\mathcal{M}_\alpha|\mathbf{Y}^*)) - \log(\pi(\mathcal{M}_{\alpha'}|\mathbf{Y}^*))\\ &\quad= \frac{1}{2}\sum_{k\in\alpha}d_k\xi_{k,n}^2 - \frac{1}{2}\sum_{k\in\alpha'}d_k\xi_{k,n}^2\\ &\qquad+\frac{1}{2}\sum_{k\in\alpha'}\log(\gamma_k+1) - \frac{1}{2}\sum_{k\in\alpha}\log(\gamma_k+1) + \log(\pi(\mathcal{M}_\alpha)) - \log(\pi(\mathcal{M}_{\alpha'})).\end{aligned}$$

Argue as in the proofs of Theorems 2 and 3 to deduce (ii) and (iii). ■

APPENDIX B: Asymptotics for Null and Full Models

We now indicate how our main asymptotic results are affected when assumption (3) is violated. That is, if $\mathcal{M}_{\alpha_0}$ is allowed to be either the null or full model.

1. Consider first the case when $\mathcal{M}_{\alpha_0}$ is the full model. Then the space of Category II models is simply $\mathcal{M}_{\alpha_0}$. Because this space contains only one model, part (iv) of Theorem 2 no longer applies. However, the remainder of the theorem continues to hold. Theorems 3 and 5 also continue to hold. This can be seen by noting that the key expression (A.1) applies in the full model case.
2. Now suppose $\mathcal{M}_{\alpha_0}$ is the null model. Then the space of Category I models is defined to be empty, and all models are of a Category II type. Hence, in Theorem 2, parts (ii) and (iii) no longer apply because the Category I model space is empty. The remainder of the theorem continues to hold, however. This follows by noting that (A.1) applies to the null model by setting the relevant dimension to zero and removing all sums involving the null model. Theorem 3 continues to apply, although the highest posterior model is obviously no longer inconsistent. Theorem 3 in fact is noninformative in the null model case because the point was to consider the setting in which coefficients may be small and nonzero. Theorem 5 also continues to apply except for part (ii), which applies to Category I models. Note that in Theorems 3 and 5 if α or α' is set to the null model, simply replace the relevant sum with the value 0. Also note that $\beta_{k,0} = 0$ for each k.