Mixtures of *g* Priors for Bayesian Variable Selection

Feng LIANG, Rui PAULO, German MOLINA, Merlise A. CLYDE, and Jim O. BERGER

Zellner's *g* prior remains a popular conventional prior for use in Bayesian variable selection, despite several undesirable consistency issues. In this article we study mixtures of *g* priors as an alternative to default *g* priors that resolve many of the problems with the original formulation while maintaining the computational tractability that has made the *g* prior so popular. We present theoretical properties of the mixture *g* priors and provide real and simulated examples to compare the mixture formulation with fixed *g* priors, empirical Bayes approaches, and other default procedures.

KEY WORDS: AIC; Bayesian model averaging; BIC; Cauchy; Empirical Bayes; Gaussian hypergeometric functions; Model selection; Zellner–Siow priors.

1. INTRODUCTION

The problem of variable selection or subset selection in linear models is pervasive in statistical practice; see George (2000) and Miller (2001). We consider model choice in the canonical regression problem with response vector $\mathbf{Y} = (y_1, \dots, y_n)^T$ normally distributed with mean vector $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)^T$ and covariance $\mathbf{I}_n/\phi$, where ϕ is a precision parameter (the inverse of the usual variance) and $\mathbf{I}_n$ is an $n \times n$ identity matrix. Given a set of potential predictor variables $\mathbf{X}_1, \dots, \mathbf{X}_p$, we assume that the mean vector $\boldsymbol{\mu}$ is in the span of $\mathbf{1}_n, \mathbf{X}_1, \dots, \mathbf{X}_p$, where $\mathbf{1}_n$ is a vector of 1's of length n. The model choice problem involves selecting a subset of predictor variables that places additional restrictions on the subspace that contains the mean. We index the model space by $\boldsymbol{\gamma}$, a p-dimensional vector of indicators with $\gamma_j = 1$, meaning that $\mathbf{X}_j$ is included in the set of predictor variables, and with $\gamma_j = 0$, meaning that $\mathbf{X}_j$ is excluded. Under model $\mathcal{M}_\gamma$, $\boldsymbol{\mu}$ may be expressed in vector form as

$$\mathcal{M}_\gamma: \quad \boldsymbol{\mu} = \mathbf{1}_n\alpha + \mathbf{X}_\gamma\boldsymbol{\beta}_\gamma,$$

where α is an intercept that is common to all models, $\mathbf{X}_\gamma$ represents the $n \times p_\gamma$ design matrix under model $\mathcal{M}_\gamma$, and $\boldsymbol{\beta}_\gamma$ is the p_γ-dimensional vector of nonzero regression coefficients.

The Bayesian approach to model selection and model uncertainty involves specifying priors on the unknowns $\boldsymbol{\theta}_\gamma = (\alpha, \boldsymbol{\beta}_\gamma, \phi) \in \boldsymbol{\Theta}_\gamma$ in each model and, in turn, updating prior probabilities of models $p(\mathcal{M}_\gamma)$ to obtain posterior probabilities of each model:

$$p(\mathcal{M}_\gamma|\mathbf{Y}) = \frac{p(\mathcal{M}_\gamma)p(\mathbf{Y}|\mathcal{M}_\gamma)}{\sum_\gamma p(\mathcal{M}_\gamma)p(\mathbf{Y}|\mathcal{M}_\gamma)}.$$

A key component in the posterior model probabilities is the marginal likelihood of the data under model $\mathcal{M}_\gamma$:

$$p(\mathbf{Y}|\mathcal{M}_\gamma) = \int_{\boldsymbol{\Theta}_\gamma} p(\mathbf{Y}|\boldsymbol{\theta}_\gamma, \mathcal{M}_\gamma)p(\boldsymbol{\theta}_\gamma|\mathcal{M}_\gamma)\,d\boldsymbol{\theta}_\gamma,$$

obtained by integrating the likelihood with respect to the prior distribution for model-specific parameters $\boldsymbol{\theta}_\gamma$.

Whereas Bayesian variable selection has a long history (Zellner 1971, sec 10.4; Leamer 1978a,b; Mitchell and Beauchamp 1988), the advent of Markov chain Monte Carlo methods catalyzed Bayesian model selection and averaging in regression models (George and McCulloch 1993, 1997; Geweke 1996; Smith and Kohn 1996; Raftery, Madigan, and Hoeting 1997; Hoeting, Madigan, Raftery, and Volinsky 1999; Clyde and George 2004). Prior density choice for Bayesian model selection and model averaging, however, remains an open area (Berger and Pericchi 2001; Clyde and George 2004). Subjective elicitation of priors for model-specific coefficients is often precluded, particularly in high-dimensional model spaces, such as in nonparametric regression using spline and wavelet bases. Thus, it is often necessary to resort to specification of priors using some formal method (Kass and Wasserman 1996; Berger and Pericchi 2001). In general, the use of improper priors for model-specific parameters is not permitted in the context of model selection, as improper priors are determined only up to an arbitrary multiplicative constant. In inference for a given model, these arbitrary multiplicative constants cancel in the posterior distribution of the model-specific parameters. However, these constants remain in marginal likelihoods leading to indeterminate model probabilities and Bayes factors (Jeffreys 1961; Berger and Pericchi 2001). To avoid indeterminacies in posterior model probabilities, proper priors for $\boldsymbol{\beta}_\gamma$ under each model are usually required.

Conventional proper priors for variable selection in the normal linear model have been based on the conjugate Normal–Gamma family for $\boldsymbol{\theta}_\gamma$ or limiting versions, allowing closed-form calculations of all marginal likelihoods (George and McCulloch 1997; Raftery et al. 1997; Berger and Pericchi 2001). Zellner's (1986) g prior for $\boldsymbol{\beta}_\gamma$,

$$\boldsymbol{\beta}_\gamma|\phi, \mathcal{M}_\gamma \sim \mathrm{N}\left(0, \frac{g}{\phi}(\mathbf{X}_\gamma^T\mathbf{X}_\gamma)^{-1}\right), \tag{1}$$

has been widely adopted because of its computational efficiency in evaluating marginal likelihoods and model search and, perhaps most important, because of its simple, understandable interpretation as arising from the analysis of a conceptual sample generated using the same design matrix $\mathbf{X}$ as employed in the current sample (Zellner 1986; Smith and Kohn 1996; George and McCulloch 1997; Fernández, Ley, and Steel 2001).

Feng Liang is Assistant Professor, Department of Statistics, University of Illinois at Urbana–Champaign, Champaign, IL 61820 (E-mail: *liangf@uiuc.edu*). Rui Paulo is Assistant Professor, Department of Mathematics, ISEG, Technical University of Lisbon, Lisbon, Portugal (E-mail: *rui@iseg.utl.pt*). German Molina is Quantitative Trading Researcher at Tudor Investment Corporation, London, England (E-mail: *german@alumni.duke.edu*). Merlise A. Clyde is Associate Professor, Department of Statistical Science, Duke University, Durham, NC 27708 (E-mail: *clyde@stat.duke.edu*). Jim O. Berger is the Arts and Sciences Professor of Statistics, Department of Statistical Science, Duke University, Durham, NC 27708 (E-mail: *berger@stat.duke.edu*). The authors thank Ed George and two anonymous referees for many helpful comments that have improved this manuscript. This work was supported by National Science Foundation grants DMS-0112069, DMS-0422400, DMS-0406115 and U.S. Environmental Protection Agency grant R828686-01-0. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or the U.S. Environmental Protection Agency.

Journal of the American Statistical Association
March 2008, Vol. 103, No. 481, Review Article
DOI 10.1198/016214507000001337

George and Foster (2000) showed how g could be calibrated based on many popular model selection criteria, such as the Akaike information criterion (AIC), the Bayesian information criterion (BIC), and the risk information criterion (RIC). To avoid the difficulty of preselecting g, while providing adaptive estimates, George and Foster (2000) and Clyde and George (2000) proposed and developed empirical Bayes (EB) methods using a common (global) estimate of g from the marginal likelihood of g. Motivated by information theory, Hansen and Yu (2001) developed related approaches that use model-specific (local EB) estimates of g. These EB approaches provide automatic prior specifications that lead to model selection criteria that bridge the AIC and BIC and provide nonlinear, rather than linear, shrinkage of model coefficients while still maintaining the computational convenience of the g-prior formulation. As many Bayesians are critical of empirical Bayes methods on the grounds that they do not correspond to solutions based on Bayesian or formal Bayesian procedures, a natural alternative to data-based EB priors are fully Bayes specifications that place a prior on g. While Zellner (1986) suggested that a prior on g should be introduced and g could be integrated out, this approach has not taken hold in practice, primarily for perceived computational difficulties.

In this article we explore fully Bayes approaches using mixtures of g priors. As calculation of marginal likelihoods using a mixture of g priors involves only a one-dimensional integral, this approach provides the attractive computational solutions that made the original g priors popular while providing robustness to misspecification of g. The Zellner–Siow (1980) Cauchy priors can be viewed as a special case of mixtures of g priors. Perhaps because Cauchy priors do not permit closed-form expressions of marginal likelihoods, they have not been adopted widely in the model choice community. Representing the Zellner–Siow Cauchy prior as a scale mixture of g priors, we develop a new approximation to Bayes factors that allows simple, tractable expressions for posterior model probabilities. We also present a new family of priors for g, the hyper-g prior family, which leads to closed-form marginal likelihoods in terms of the Gaussian hypergeometric function. Both the Cauchy and the hyper-g priors provide similar computational efficiency, adaptivity, and nonlinear shrinkage found in EB procedures. The same family of priors has also been independently proposed and studied by Cui and George (2007) for Bayesian variable selection where they focus on the case of known error variance.

This article is organized as follows. In Section 2 we review Zellner's g prior, with suggested specifications for g from the literature, and discuss some of the paradoxes associated with fixed g priors. In Section 3 we present mixtures of g priors. Motivated by Jeffrey's; desiderata for the properties of Bayes factors, we specify conditions on the prior distribution for g that resolve the Bayes factor paradoxes associated with fixed g priors. We discuss theoretical properties of the Zellner–Siow Cauchy and hyper-g priors and other asymptotic properties of posteriors in Section 4. To investigate small-sample performance, we compare the Zellner–Siow Cauchy and hyper-g priors to other approaches in a simulation study (Sec. 5) and in examples from the literature (Sec. 6). Finally, in Section 7 we conclude with recommendations for priors for the variable selection problem and unresolved issues.

2. ZELLNER'S *g* PRIORS

In constructing a family of priors for a Gaussian regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, Zellner (1986) suggested a particular form of the conjugate Normal–Gamma family, namely, a g prior:

$$p(\phi) \propto \frac{1}{\phi}, \qquad \boldsymbol{\beta}|\phi \sim \mathrm{N}\left(\boldsymbol{\beta}_a, \frac{g}{\phi}(\mathbf{X}^T\mathbf{X})^{-1}\right),$$

where the prior mean $\boldsymbol{\beta}_a$ is taken as the anticipated value of $\boldsymbol{\beta}$ based on imaginary data and the prior covariance matrix of $\boldsymbol{\beta}$ is a scalar multiple g of the Fisher information matrix, which depends on the observed data through the design matrix $\mathbf{X}$. In the context of hypothesis testing with $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ versus $H_1 : \boldsymbol{\beta} \in \mathbb{R}^k$, Zellner suggested setting $\boldsymbol{\beta}_a = \boldsymbol{\beta}_0$ in the g prior for $\boldsymbol{\beta}$ under H_1 and derived expressions for the Bayes factor for testing H_1 versus H_0.

While Zellner (1986) derived Bayes factors using g priors for testing precise hypotheses, he did not explicitly consider nested models, where the null hypothesis restricts the values for a subvector of $\boldsymbol{\beta}$. We (as have others) adapt Zellner's g prior for testing nested hypotheses by placing a flat prior on the regression coefficients that are common to both models and using the g prior for the regression parameters that are only in the more complex model. This is the strategy used by Zellner and Siow (1980) in the context of other priors. While such an approach leads to coherent prior specifications for a pair of hypotheses, variable selection in regression models is essentially a multiple hypothesis-testing problem, leading to many nonnested comparisons. In the Bayesian solution the posterior probabilities of models can be expressed through the Bayes factor for pairs of hypotheses, namely,

$$p(\mathcal{M}_\gamma|\mathbf{Y}) = \frac{p(\mathcal{M}_\gamma)\,\mathrm{BF}[\mathcal{M}_\gamma : \mathcal{M}_b]}{\sum_{\gamma'} p(\mathcal{M}_{\gamma'})\,\mathrm{BF}[\mathcal{M}_{\gamma'} : \mathcal{M}_b]}, \qquad (2)$$

where the Bayes factor, $\mathrm{BF}[\mathcal{M}_\gamma : \mathcal{M}_b]$, for comparing each of $\mathcal{M}_\gamma$ to a base model $\mathcal{M}_b$ is given by

$$\mathrm{BF}[\mathcal{M}_\gamma : \mathcal{M}_b] = \frac{p(\mathbf{Y}|\mathcal{M}_\gamma)}{p(\mathbf{Y}|\mathcal{M}_b)}.$$

To define the Bayes factor of any two models $\mathcal{M}_\gamma$ and $\mathcal{M}_{\gamma'}$, we utilize the "encompassing" approach of Zellner and Siow (1980) and define the Bayes factor for comparing any two models $\mathcal{M}_\gamma$ and $\mathcal{M}_{\gamma'}$ to be

$$\mathrm{BF}(\mathcal{M}_\gamma : \mathcal{M}_{\gamma'}) = \frac{\mathrm{BF}(\mathcal{M}_\gamma : \mathcal{M}_b)}{\mathrm{BF}(\mathcal{M}_{\gamma'} : \mathcal{M}_b)}.$$

In principle, the choice of the base model $\mathcal{M}_b$ is completely arbitrary as long as the priors for the parameters of each model are specified separately and do not depend on the comparison being made. However, because the definition of common parameters changes with the choice of the base model, improper priors for common parameters in conjunction with g priors on the remaining parameters lead to expressions for Bayes factors that do depend on the choice of the base model. The null model and the full model are the only two choices for $\mathcal{M}_b$ which make each pair, $\mathcal{M}_\gamma$ and $\mathcal{M}_b$, a pair of nested models. We will refer to the choice of the base model being $\mathcal{M}_N$ (the null model) as the *null-based* approach. Similarly the *full-based* approach utilizes $\mathcal{M}_F$ (the full model) as the base model.

2.1 Null-Based Bayes Factors

In the null-based approach to calculating Bayes factors and model probabilities, we compare each model $\mathcal{M}_\gamma$ with the null model $\mathcal{M}_N$ through the hypotheses $H_0 : \boldsymbol{\beta}_\gamma = 0$ and $H_1 : \boldsymbol{\beta}_\gamma \in \mathbb{R}^{p_\gamma}$. Without loss of generality, we may assume that the columns of $\mathbf{X}_\gamma$ have been centered, so that $\mathbf{1}^T \mathbf{X}_\gamma = \mathbf{0}$, in which case the intercept α may be regarded as a common parameter to both $\mathcal{M}_\gamma$ and $\mathcal{M}_N$. This, along with arguments based on orthogonal parameterizations and invariance to scale and location transformations (Jeffreys 1961; Eaton 1989; Berger, Pericchi, and Varshavsky 1998), has led to the adoption of

$$p(\alpha, \phi | \mathcal{M}_\gamma) = \frac{1}{\phi}, \tag{3}$$

$$\boldsymbol{\beta}_\gamma | \phi, \mathcal{M}_\gamma \sim \mathrm{N}\left(\mathbf{0}, \frac{g}{\phi}(\mathbf{X}_\gamma^T \mathbf{X}_\gamma)^{-1}\right), \tag{4}$$

as a default prior specification for α, $\boldsymbol{\beta}_\gamma$, and ϕ under $\mathcal{M}_\gamma$. Most references to g priors in the variable selection literature refer to the previous version (Clyde and George 2000; George and Foster 2000; Berger and Pericchi 2001; Fernández et al. 2001; Hansen and Yu 2001). Continuing with this tradition, we will also refer to the priors in (3)–(4) simply as Zellner's g prior.

A major advantage of Zellner's g prior is the computational efficiency due to the closed-form expression of all marginal likelihoods. Under (3)–(4), the marginal likelihood is given by

$$p(\mathbf{Y} | \mathcal{M}_\gamma, g) = \frac{\Gamma((n-1)/2)}{\sqrt{\pi}^{(n-1)} \sqrt{n}} \|\mathbf{Y} - \bar{\mathbf{Y}}\|^{-(n-1)} \times \frac{(1+g)^{(n-1-p_\gamma)/2}}{[1 + g(1 - R_\gamma^2)]^{(n-1)/2}}, \tag{5}$$

where R_γ^2 is the ordinary coefficient of determination of regression model $\mathcal{M}_\gamma$. Though the marginal of the null model $p(\mathbf{Y} | \mathcal{M}_N)$ does not involve the hyperparameter g, it can be obtained as a special case of (5) with $R_\gamma^2 = 0$ and $p_\gamma = 0$. The resulting Bayes factor for comparing any model $\mathcal{M}_\gamma$ to the null model is

$$\mathrm{BF}[\mathcal{M}_\gamma : \mathcal{M}_N] = (1+g)^{(n-p_\gamma-1)/2} [1 + g(1 - R_\gamma^2)]^{-(n-1)/2}. \tag{6}$$

2.2 Full-Based Bayes Factors

For comparing model $\mathcal{M}_\gamma$ with covariates $\mathbf{X}_\gamma$ to the full model, we will partition the design matrix associated with the full model as $\mathbf{X} = [\mathbf{1}, \mathbf{X}_\gamma, \mathbf{X}_{-\gamma}]$, so that the full model $\mathcal{M}_F$, written in partitioned form, is represented as

$$\mathcal{M}_F: \quad \mathbf{Y} = \mathbf{1}\alpha + \mathbf{X}_\gamma \boldsymbol{\beta}_\gamma + \mathbf{X}_{-\gamma} \boldsymbol{\beta}_{-\gamma} + \boldsymbol{\epsilon},$$

where $\mathbf{X}_{-\gamma}$ refers to the columns of $\mathbf{X}$ excluded in model $\mathcal{M}_\gamma$. Model $\mathcal{M}_\gamma$ corresponds to the hypothesis $H_0 : \boldsymbol{\beta}_{-\gamma} = 0$, while the hypothesis $H_1 : \boldsymbol{\beta}_{-\gamma} \in \mathbb{R}^{p - p_\gamma}$ corresponds to the full model $\mathcal{M}_F$, where common parameters α and $\boldsymbol{\beta}_\gamma$ are unrestricted under both models. For comparing these two models, we assume (without loss of generality) that the full model has been parameterized in a block-orthogonal fashion such that $\mathbf{1}^T [\mathbf{X}_\gamma, \mathbf{X}_{-\gamma}] = \mathbf{0}$ and $\mathbf{X}_\gamma^T \mathbf{X}_{-\gamma} = \mathbf{0}$, in order to justify treating α and $\boldsymbol{\beta}_\gamma$ as common parameters to both models (Zellner and Siow 1980). This leads to the following g priors for the full-based Bayes factors:

$$\begin{aligned}
\mathcal{M}_\gamma: &\quad p(\alpha, \phi, \boldsymbol{\beta}_\gamma) \propto \frac{1}{\phi}, \\
\mathcal{M}_F: &\quad p(\alpha, \phi, \boldsymbol{\beta}_\gamma) \propto \frac{1}{\phi}, \\
&\quad \boldsymbol{\beta}_{-\gamma} | \phi \sim \mathrm{N}\left(0, \frac{g}{\phi}(\mathbf{X}_{-\gamma}^T \mathbf{X}_{-\gamma})^{-1}\right),
\end{aligned} \tag{7}$$

with the resulting Bayes factor for comparing any model $\mathcal{M}_\gamma$ to the full model given by

$$\mathrm{BF}[\mathcal{M}_\gamma : \mathcal{M}_F] = (1+g)^{-(n-p-1)/2} \left[1 + g \frac{1 - R_F^2}{1 - R_\gamma^2}\right]^{(n-p_\gamma-1)/2}, \tag{8}$$

where R_γ^2 and R_F^2 are the usual coefficients of the determination of models $\mathcal{M}_\gamma$ and $\mathcal{M}_F$, respectively.

It should be noted that, unlike the null-based approach, the full-based approach does not lead to a coherent prior specification for the full model, because the prior distribution (7) for $\boldsymbol{\beta}$ in $\mathcal{M}_F$ depends on $\mathcal{M}_\gamma$, which changes with each model comparison. Nonetheless, posterior probabilities (2) can still be formally defined using the Bayes factor with respect to the full model (8). A similar formulation, where the prior on the full model depends on which hypothesis is being tested, has also been adapted by Casella and Moreno (2006) in the context of intrinsic Bayes factors. Their rationale is that the full model is the scientific "null" and that all models should be judged against it. Because in most of the literature g priors refer to the null-based approach, in the rest of this article, we will mainly focus on the null-based g prior and its alternatives unless specified otherwise.

2.3 Paradoxes of g Priors

The simplicity of the g-prior formulation is that just one hyperparameter g needs to be specified. Because g acts as a dimensionality penalty, the choice of g is critical. As a result, Bayes factors for model selection with fixed choices of g may exhibit some undesirable features, as discussed next.

Bartlett's Paradox. For inference under a given model, the posterior can be reasonable even if g is chosen very large in an effort to be noninformative. In model selection, however, this is generally a bad idea. In fact, in the limiting case when $g \to \infty$ while n and p_γ are fixed, the Bayes factor (6) for comparing $\mathcal{M}_\gamma$ to $\mathcal{M}_N$ will go to 0. That is, large spread of the prior induced by the noninformative choice of g has the unintended consequence of forcing the Bayes factor to favor the null model, the smallest model, regardless of the information in the data. Such a phenomenon has been noted in Bartlett (1957) and is often referred to as "Bartlett's paradox," which was well understood and discussed by Jeffreys (1961).

Information Paradox. Suppose, in comparing the null model and a particular model $\mathcal{M}_\gamma$, we have overwhelming information supporting $\mathcal{M}_\gamma$. For example, suppose $\|\hat{\boldsymbol{\beta}}_\gamma\|^2$ goes to ∞, so that $R^2_\gamma \to 1$, or, equivalently, the usual F statistic goes to ∞ with both n and p_γ fixed. In any conventional sense, one would expect that $\mathcal{M}_\gamma$ should receive high posterior probability and that the Bayes factor $\mathrm{BF}(\mathcal{M}_\gamma : \mathcal{M}_N)$ would go to ∞ as the information against $\mathcal{M}_N$ accumulates. However, in this situation, the Bayes factor (6) with a fixed choice of g tends to a constant $(1+g)^{(n-p_\gamma-1)/2}$ as $R^2_\gamma \to 1$ (Zellner 1986; Berger and Pericchi 2001). Because this paradox is related to the limiting behavior of the Bayes factor as information accumulates, we will refer to it as the "information paradox."

2.4 Choices of *g*

Under uniform prior model probabilities, the choice of g effectively controls model selection, with large g typically concentrating the prior on parsimonious models with a few large coefficients, whereas small g tends to concentrate the prior on saturated models with small coefficients (George and Foster 2000). Recommendations for g have included the following:

- *Unit information prior.* Kass and Wasserman (1995) recommended choosing priors with the amount of information about the parameter equal to the amount of information contained in one observation. For regular parametric families, the "amount of information" is defined through Fisher information. In the normal regression case, the unit information prior corresponds to taking $g = n$, leading to Bayes factors that behave like the BIC.
- *Risk inflation criterion.* Foster and George (1994) calibrated priors for model selection based on the RIC and recommended the use of $g = p^2$ from a minimax perspective.
- *Benchmark prior.* Fernández et al. (2001) did a thorough study on various choices of g with dependence on the sample size n or the model dimension p and concluded with the recommendation to take $g = \max(n, p^2)$. We refer to their "benchmark prior" specification as "BRIC" as it bridges BIC and RIC.
- *Local empirical Bayes.* The local EB approach can be viewed as estimating a separate g for each model. Using the marginal likelihood after integrating out all parameters given in (5), an EB estimate of g is the maximum (marginal) likelihood estimate constrained to be nonnegative, which turns out to be

  $$\hat{g}^{\mathrm{EBL}}_\gamma = \max\{F_\gamma - 1, 0\}, \tag{9}$$

  where

  $$F_\gamma = \frac{R^2_\gamma / p_\gamma}{(1 - R^2_\gamma)/(n - 1 - p_\gamma)}$$

  is the usual F statistic for testing $\boldsymbol{\beta}_\gamma = 0$. An asymptotic SE (standard error) based on the observed information for $\hat{g}^{\mathrm{EBL}}_\gamma$ is straightforward to derive.
- *Global empirical Bayes.* The global EB procedure assumes one common g for all models and borrows strength from all models by estimating g from the marginal likelihood of the data, averaged over all models,

  $$\hat{g}^{\mathrm{EBG}} = \underset{g>0}{\arg\max} \sum_\gamma p(\mathcal{M}_\gamma) \frac{(1+g)^{(n-p_\gamma-1)/2}}{[1 + g(1 - R^2_\gamma)]^{(n-1)/2}}. \tag{10}$$

  In general, this marginal likelihood is not tractable and does not provide a closed-form solution for $\hat{g}^{\mathrm{EBG}}$, although numerical optimization may be used (George and Foster 2000). Here we propose an EM algorithm based on treating both the model indicator and the precision ϕ as latent data. The E step consists of the following expectations:

  $$\begin{aligned} E\big[\phi^{(i)} \mid \mathcal{M}_\gamma, \mathbf{Y}, \hat{g}^{(i)}\big] &= \frac{n-1}{\|\mathbf{Y} - \bar{\mathbf{Y}}\|^2 (1 - (\hat{g}^{(i)}/(1+\hat{g}^{(i)}))R^2_\gamma)}, \\ E\big[\mathcal{M}_\gamma | \mathbf{Y}, \hat{g}^{(i)}\big] &= \frac{p(\mathbf{Y}|\mathcal{M}_\gamma, \hat{g}^{(i)})}{\sum_{\gamma'} p(\mathbf{Y}|\mathcal{M}_{\gamma'}, \hat{g}^{(i)})} \\ &\equiv \hat{p}^{(i)}(\mathcal{M}_\gamma|\mathbf{Y}), \end{aligned} \tag{11}$$

  evaluated at the current estimate of g and where the marginal likelihood $p(\mathbf{Y}|\mathcal{M}_\gamma, g)$ is based on (5). After simplification, the marginal maximum likelihood estimate of g from the M step is

  $$\begin{aligned} \hat{g}^{(i+1)} = \max\Bigg\{ & \sum_\gamma \hat{p}^{(i)}(\mathcal{M}_\gamma|\mathbf{Y}) \\ & \times \frac{R^2_\gamma / \sum_{\gamma'} \hat{p}^{(i)}(\mathcal{M}_{\gamma'}|\mathbf{Y}) p_{\gamma'}}{(1 - (\hat{g}^{(i)}/(1+\hat{g}^{(i)}))R^2_\gamma)/(n-1)} - 1, 0 \Bigg\}, \end{aligned} \tag{12}$$

  where the terms inside the summation can be viewed as a weighted Bayesian F statistic. The global EB estimate of g, $\hat{g}^{\mathrm{EBG}}$, is the estimate of g from (12) after convergence. A side benefit of the EM algorithm is that the global EB posterior model probabilities are obtained from (11) at convergence. When the dimension of the model space prohibits enumeration, the global EB estimates may be based on a subset of models obtained, for example, using stochastic search and sampling from the local EB posterior. One may obtain an asymptotic SE using the method of Louis (1982) with output from the EM algorithm or deriving the information directly.

The unit information prior, risk inflation criterion, and benchmark prior do not resolve the information paradox for fixed n and p because the choices of g are fixed values not depending on the information in the data. However, the two EB approaches do have the desirable behavior as stated later.

Theorem 1. In the setting of the information paradox with fixed n, $p < n$, and $R^2_\gamma \to 1$, the Bayes factor (6) for comparing $\mathcal{M}_\gamma$ to $\mathcal{M}_N$ goes to ∞ under either the local or the global EB estimate of g.

Proof. It is easy to check that the Bayes factor (6) with $g = \hat{g}^{\mathrm{EBL}}$ goes to ∞ when R^2_γ goes to 1. It implies that the maximum of the right side of (10) also goes to ∞, and so does the leading term $\mathrm{BF}[\mathcal{M}_\gamma : \mathcal{M}_N]$ with $g = \hat{g}^{\mathrm{EBG}}$.

The EB priors provide a resolution of the information paradox that arises when using a fixed g in the g priors. One may view the marginal maximum likelihood estimate of g as a posterior mode under a uniform (improper) prior distribution for g.

Rather than using a plug-in estimate to eliminate g, a natural alternative is the integrated marginal likelihood under a proper prior on g. Consequently, a prior on g leads to a mixture of g priors for the coefficients $\boldsymbol{\beta}_\gamma$, which typically provides more robust inference. In the next section we explore various mixing distributions that maintain the computational convenience of the original g prior and have attractive theoretical properties as in the EB approaches.

3. MIXTURES OF g PRIORS

Letting $\pi(g)$ (which may depend on n) denote the prior on g, the marginal likelihood of the data $p(\mathbf{Y}|\mathcal{M}_\gamma)$ is proportional to

$$\mathrm{BF}[\mathcal{M}_\gamma : \mathcal{M}_N] = \int_0^\infty (1+g)^{(n-1-p_\gamma)/2} \times [1 + (1 - R_\gamma^2)g]^{-(n-1)/2} \pi(g)\, dg \quad (13)$$

in the null-based approach. Similar expressions for the full-based approach can be obtained using (8). Under selection of a model $\mathcal{M}_\gamma \neq \mathcal{M}_N$, the posterior mean of $\boldsymbol{\mu}$, $\mathbb{E}[\boldsymbol{\mu}|\mathcal{M}_\gamma, \mathbf{Y}]$, is

$$\mathbb{E}[\boldsymbol{\mu}|\mathcal{M}_\gamma, \mathbf{Y}] = \mathbf{1}_n \hat{\alpha} + \mathbb{E}\left[\frac{g}{1+g} \,\middle|\, \mathcal{M}_\gamma, \mathbf{Y}\right] \mathbf{X}_\gamma \hat{\boldsymbol{\beta}}_\gamma,$$

where $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}_\gamma$ are the ordinary least squares estimates of α and $\boldsymbol{\beta}$, respectively, under model $\mathcal{M}_\gamma$. Under the fixed g prior, the posterior mean for $\boldsymbol{\beta}_\gamma$ under a selected model is a linear shrinkage estimator with a fixed shrinkage factor $g/(1+g)$; thus mixtures of g priors allow adaptive data-dependent shrinkage. The optimal (Bayes) estimate of $\boldsymbol{\mu}$ under squared error loss is the posterior mean under model averaging given by

$$\mathbb{E}[\boldsymbol{\mu}|\mathbf{Y}] = \mathbf{1}_n \hat{\alpha} + \sum_{\gamma:\mathcal{M}_\gamma \neq \mathcal{M}_N} p(\mathcal{M}_\gamma|\mathbf{Y}) \times \mathbb{E}\left[\frac{g}{1+g} \,\middle|\, \mathcal{M}_\gamma, \mathbf{Y}\right] \mathbf{X}_\gamma \hat{\boldsymbol{\beta}}_\gamma, \quad (14)$$

which provides multiple nonlinear adaptive shrinkage through the expectation of the linear shrinkage factor and through the posterior model probabilities. Because g appears not only in Bayes factors and model probabilities but also in posterior means and predictions, the choice of prior on g should ideally allow for tractable computations for all these quantities.

While tractable calculation of marginal likelihoods and predictions is desirable, more important, we would like priors that lead to consistent model selection and have desirable risk properties. We explore in detail two fully Bayesian approaches: Zellner–Siow's Cauchy prior (Zellner and Siow 1980), which is obtained using an Inverse-Gamma prior on g, and the hyper-g prior, which is an extension of the Strawderman (1971) prior to the regression context.

3.1 Zellner–Siow Priors

In the context of hypothesis testing regarding a univariate normal mean, Jeffreys (1961) rejected normal priors essentially for reasons related to the Bayes factor paradoxes described earlier and found that the Cauchy prior was the simplest prior to satisfy basic consistency requirements for hypothesis testing. Zellner and Siow (1980) introduced multivariate Cauchy priors on the regression coefficients as suitable multivariate extensions to Jeffreys's work on the univariate normal mean problem. If the two models under comparison are nested, the Zellner–Siow strategy is to place a flat prior on common coefficients and a Cauchy prior on the remaining parameters. For example, in the null-based approach, the prior on (α, ϕ) is given by (3) and

$$\pi(\boldsymbol{\beta}_\gamma|\phi) \propto \frac{\Gamma(p_\gamma/2)}{\pi^{p_\gamma/2}} \left|\frac{\mathbf{X}_\gamma^T \mathbf{X}_\gamma}{n/\phi}\right|^{1/2} \left(1 + \boldsymbol{\beta}_\gamma^T \frac{\mathbf{X}_\gamma^T \mathbf{X}_\gamma}{n/\phi} \boldsymbol{\beta}_\gamma\right)^{-p_\gamma/2},$$

a multivariate Cauchy centered at the null model, $\boldsymbol{\beta}_\gamma = \mathbf{0}$, with precision suggested by the form of the unit Fisher information matrix.

Arguably, one of the reasons why the Zellner–Siow prior has never become quite as popular as the g prior in Bayesian variable selection is the fact that closed-form expressions for marginal likelihoods are not available. Zellner and Siow (1980) derived approximations to the marginal likelihoods by directly approximating the integral over $\mathbb{R}^{p_\gamma}$ with respect to the multivariate Cauchy prior. However, as the model dimensionality increases, the accuracy of the approximation degrades.

It is well known that a Cauchy distribution can be expressed as a scale mixture of normals. The Zellner–Siow priors can be represented as a mixture of g priors with an Inv-Gamma$(1/2, n/2)$ prior on g, namely,

$$\pi(\boldsymbol{\beta}_\gamma|\phi) \propto \int N\left(\boldsymbol{\beta}_\gamma \,\middle|\, \mathbf{0}, \frac{g}{\phi}(\mathbf{X}_\gamma^T \mathbf{X}_\gamma)^{-1}\right) \pi(g)\, dg, \quad (15)$$

with

$$\pi(g) = \frac{(n/2)^{1/2}}{\Gamma(1/2)} g^{-3/2} e^{-n/(2g)}.$$

One may take advantage of the mixture of g-prior representation (15) to first integrate out $\boldsymbol{\theta}_\gamma$ given g, leaving a one-dimensional integral over g as in (13), which is independent of the model dimension. This one-dimensional integral can be carried out using standard numerical integration techniques or using a Laplace approximation. The Laplace approximation involves expanding the unnormalized marginal posterior density of g about its mode and leads to tractable calculations for approximating marginal likelihoods as the marginal posterior mode of g is a solution to a cubic equation. Furthermore, the posterior expectation of $g/(1+g)$, necessary for prediction, can also be approximated using the same form of Laplace approximation, and again the associated mode is the solution to a cubic equation. Details can be found in Appendix A and are implemented in an R package available from the authors.

3.2 Hyper-g Priors

As an alternative to the Zellner–Siow prior for the model choice problem, we introduce a family of priors on g:

$$\pi(g) = \frac{a-2}{2}(1+g)^{-a/2}, \qquad g > 0, \quad (16)$$

which is a proper distribution for $a > 2$. This family of priors includes priors used by Strawderman (1971) to provide improved mean square risk over ordinary maximum likelihood estimates in the normal means problem. These priors have also been studied by Cui and George (2007) for the problem of variable selection in the case of known error variance.

When $a \leq 2$ the prior $\pi(g) \propto (1+g)^{-a/2}$ is improper; both the reference prior and the Jeffreys prior correspond to $a = 2$. When $1 < a \leq 2$ we will see that the marginal density, given below in (17), is finite, so that the corresponding posterior distribution is proper. Even though the choice of $1 < a \leq 2$ leads to proper posterior distributions, because g is not included in the null model, the issue of arbitrary constants of proportionality leads to indeterminate Bayes factors. For this reason we will limit attention to the prior in (16) with $a > 2$.

More insight on hyperparameter specification can be obtained by instead considering the corresponding prior on the shrinkage factor $g/(1+g)$, where

$$\frac{g}{1+g} \sim \mathrm{Beta}\left(1, \frac{a}{2} - 1\right),$$

which is a Beta distribution with mean $2/a$. For $a = 4$ the prior on the shrinkage factor is uniform. Values of a greater than 4 tend to put more mass on shrinkage values near 0, which is undesirable a priori. Taking $a = 3$ places most of the mass near 1, with the prior probability that the shrinkage factor is greater than .80 equal to .45. We will work with $a = 3$ and $a = 4$ for future examples, although any choice $2 < a \leq 4$ may be reasonable.

An advantage of the hyper-g prior is that the posterior distribution of g given a model is available in closed form:

$$p(g|\mathbf{Y}, \mathcal{M}_\gamma) = \frac{p_\gamma + a - 2}{2\,{}_2F_1((n-1)/2, 1; (p_\gamma + a)/2; R_\gamma^2)} \times (1+g)^{(n-1-p_\gamma-a)/2}[1 + (1 - R_\gamma^2)g]^{-(n-1)/2},$$

where ${}_2F_1(a, b; c; z)$ in the normalizing constant is the Gaussian hypergeometric function (Abramowitz and Stegun 1970, sec. 15). The integral representing ${}_2F_1(a, b; c; z)$ is convergent for real $|z| < 1$ with $c > b > 0$ and for $z = \pm 1$ only if $c > a + b$ and $b > 0$. As the normalizing constant in the prior on g is also a special case of the ${}_2F_1$ function with $z = 0$, we refer to this family of prior distributions as the hyper-g priors.

The Gaussian hypergeometric function appears in many quantities of interest. The normalizing constant in the posterior for g leads to the null-based Bayes factor

$$\begin{aligned}\mathrm{BF}[\mathcal{M}_\gamma : \mathcal{M}_N] &= \frac{a-2}{2} \int_0^\infty (1+g)^{(n-1-p_\gamma-a)/2} \\ &\quad \times [1 + (1 - R_\gamma^2)g]^{-(n-1)/2}\,dg \\ &= \frac{a-2}{p_\gamma + a - 2} \\ &\quad \times {}_2F_1\left(\frac{n-1}{2}, 1; \frac{p_\gamma + a}{2}; R_\gamma^2\right), \qquad (17)\end{aligned}$$

which can be easily evaluated. The posterior mean of g under $\mathcal{M}_\gamma$ is given by

$$\begin{aligned}&\mathbb{E}[g|\mathcal{M}_\gamma, \mathbf{Y}] \\ &\quad = \frac{2}{p_\gamma + a - 4} \frac{{}_2F_1((n-1)/2, 2; (p_\gamma + a)/2; R_\gamma^2)}{{}_2F_1((n-1)/2, 1; (p_\gamma + a)/2; R_\gamma^2)}, \qquad (18)\end{aligned}$$

which is finite if $a > 3$. Likewise, the expected value of the shrinkage factor under each model can also be expressed using the ${}_2F_1$ function:

$$\begin{aligned}&\mathbb{E}\left[\frac{g}{1+g} \,\middle|\, \mathbf{Y}, \mathcal{M}_\gamma\right] \\ &\quad = \frac{\int g(1+g)^{(n-1-p_\gamma-a)/2-1}[1 + (1 - R_\gamma^2)/g]^{-(n-1)/2}\,dg}{\int (1+g)^{(n-1-p_\gamma-a)/2}[1 + (1 - R_\gamma^2)g]^{-(n-1)/2}\,dg} \\ &\quad = \frac{2}{p_\gamma + a} \frac{{}_2F_1((n-1)/2, 2; (p_\gamma + a)/2 + 1; R_\gamma^2)}{{}_2F_1((n-1)/2, 1; (p_\gamma + a)/2; R_\gamma^2)}, \qquad (19)\end{aligned}$$

which unlike the ordinary g prior leads to nonlinear data-dependent shrinkage.

While subroutines in the Cephes library (*http://www.netlib.org/cephes*) are available for evaluating Gaussian hypergeometric functions, numerical overflow is problematic for moderate to large n and large R_γ^2. Similar numerical difficulties with the ${}_2F_1$ have been encountered by Butler and Wood (2002), who developed a Laplace approximation to the integral representation

$${}_2F_1(a, b; c; z) = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)} \int_0^1 \frac{t^{b-1}(1-t)^{c-b-1}}{(1-tz)^a}\,dt. \qquad (20)$$

Because the Laplace approximation involves an integral with respect to a normal kernel, we prefer to develop the expansion after a change of variables to $\tau = \log(g)$, carrying out the integration over the entire real line. This avoids issues with modes on the boundary (as in the local empirical Bayes solution) and leads to an improved normal approximation to the integral as the variable of integration is no longer restricted. Details of the fully exponential Laplace approximations (Tierney and Kadane 1986) of order $O(n^{-1})$ to the expression (17) and of order $O(n^{-2})$ for the ratios in (18) and (19) are given in Appendix A.

4. CONSISTENCY

So far in this article we have considered several alternatives to fixed g priors: local and global empirical Bayes, Zellner–Siow priors, and hyper-g priors. In this section we investigate the theoretical properties of mixtures of g priors. In particular, three aspects of consistency are considered here: (1) the "information paradox" where $R_\gamma^2 \to 1$ as described in Section 2.3, (2) the asymptotic consistency of model posterior probabilities where $n \to \infty$ as considered in Fernández et al. (2001), and (3) the asymptotic consistency for prediction. While agreeing that no model is ever completely true, many (ourselves included) do feel it is useful to study the behavior of procedures under the assumption of a true model.

4.1 Information Paradox

A general result providing conditions under which mixtures of g priors resolve the information paradox is given next.

Theorem 2. To resolve the information paradox for all n and $p < n$, it suffices to have

$$\int_0^\infty (1+g)^{(n-1-p_\gamma)/2}\pi(g)\,dg = \infty \quad \forall p_\gamma \leq p.$$

In the case of minimal sample size (i.e., $n = p + 2$), it suffices to have $\int_0^\infty (1+g)^{1/2}\pi(g)\,dg = \infty$.

Proof. The integrand function in the Bayes factor (13) is a monotonic increasing function of R_γ^2. Therefore, when R_γ^2 goes to 1, it goes to $\int (1+g)^{(n-1-p_\gamma)/2}\pi(g)\,dg$ by the monotone convergence theorem. So the nonintegrability of $(1+g)^{(n-1-p_\gamma)/2}\pi(g)$ is a sufficient and necessary condition for resolving the "information paradox." The result for the case with minimal sample size is straightforward.

It is easy to check that the Zellner–Siow prior satisfies this condition. For the hyper-g prior, there is an additional constraint that $a \le n - p_\gamma + 1$, which, in the case of the minimal sample size, suggests that we take $2 < a \le 3$. As a fixed g prior corresponds to the special case of a degenerate prior that is a point mass at a selected value of g, it is clear that no fixed choice of $g \le \infty$ will resolve the paradox.

4.2 Model Selection Consistency

The following definition of posterior model consistency for model choice is considered in Fernández et al. (2001), namely,

$$\operatorname{plim}_n p(\mathcal{M}_\gamma|\mathbf{Y}) = 1 \quad \text{when } \mathcal{M}_\gamma \text{ is the true model,} \tag{21}$$

where "plim" denotes convergence in probability and the probability measure here is the sampling distribution under the assumed true model $\mathcal{M}_\gamma$. By the relationship between posterior probabilities and Bayes factors (2), the consistency property (21) is equivalent to

$$\operatorname{plim}_n \mathrm{BF}[\mathcal{M}_{\gamma'} : \mathcal{M}_\gamma] = 0 \quad \text{for all } \mathcal{M}_{\gamma'} \ne \mathcal{M}_\gamma.$$

For any model $\mathcal{M}_{\gamma'}$ that does not contain the true model $\mathcal{M}_\gamma$, we will assume that

$$\lim_{n\to\infty} \frac{\boldsymbol{\beta}_\gamma^T \mathbf{X}_\gamma^T (I - P_{\gamma'})\mathbf{X}_\gamma \boldsymbol{\beta}_\gamma}{n} = b_{\gamma'} \in (0,\infty), \tag{22}$$

where $P_{\gamma'}$ is the projection matrix onto the span of $\mathbf{X}_{\gamma'}$. Under this assumption, Fernández et al. (2001) have shown that consistency holds for BRIC and BIC. Here we consider the case for mixtures of g priors and the empirical Bayes approaches.

Theorem 3. Assume (22) holds. When the true model is not the null model (i.e., $\mathcal{M}_\gamma \ne \mathcal{M}_N$), posterior probabilities under empirical Bayes, Zellner–Siow priors, and hyper-g priors are consistent for model selection; when $\mathcal{M}_\gamma = \mathcal{M}_N$, consistency still holds true for the Zellner–Siow prior, but does not hold for the hyper-g or local and global empirical Bayes.

The proof is given in Appendix B. A key feature in the consistency of posterior model probabilities under the null model with the Zellner–Siow prior is that the prior on g depends on the sample size n; this is not the case in the EB or hyper-g priors. The inconsistency under the null model of the EB prior has been noted by George and Foster (2000). Looking at the proof of Theorem 3, one can see that while the EB and hyper-g priors are not consistent in the sense of (21) under the null model, the null model will still be the highest probability model, even though its posterior probability is bounded away from 1. Thus, the priors will be consistent in a weaker sense for the problem of model selection under a 0–1 loss.

The lack of consistency under the null model motivates a modification of the hyper-g prior, which we refer to as the hyper-g/n prior:

$$\pi(g) = \frac{a-2}{2n}\left(1 + \frac{g}{n}\right)^{-a/2},$$

where the normalizing constant for the prior is another special case of the Gaussian hypergeometric family. While no analytic expressions are available for the distribution or various expectations (this form of the prior is not closed under sampling), it is straightforward to approximate quantities of interest using Laplace approximations as detailed in Appendix A.

4.3 Prediction Consistency

In practice, prediction sometimes is of more interest than uncovering the true model. Given the observed data $(\mathbf{Y}, \mathbf{X}_1, \dots, \mathbf{X}_p)$ and a new vector of predictors $\mathbf{x}^* \in \mathbb{R}^p$, we would like to predict the corresponding response Y^*. In the Bayesian framework, the optimal point estimator (under squared error loss) for Y^* is the Bayesian model averaging (BMA) prediction given by

$$\hat{Y}_n^* = \hat{\alpha} + \sum_\gamma \mathbf{x}_\gamma^{*T} \hat{\boldsymbol{\beta}}_\gamma p(\mathcal{M}_\gamma|\mathbf{Y}) \times \int_0^\infty \frac{g}{1+g}\pi(g|\mathcal{M}_\gamma, \mathbf{Y})\,dg. \tag{23}$$

The local and global EB estimators can be obtained by replacing $\pi(g|\mathcal{M}_\gamma, \mathbf{Y})$ by a degenerate distribution with point mass at $\hat{g}^{\mathrm{EBL}}$ and $\hat{g}^{\mathrm{EBG}}$, respectively. When the true sampling distribution is known, that is, $(\mathcal{M}_\gamma, \alpha, \boldsymbol{\beta}_\gamma, \phi)$ are known, it is optimal (under squared error loss) to predict Y^* by its mean. Therefore, we call $\hat{Y}_n^*$ consistent under prediction if

$$\operatorname{plim}_n \hat{Y}_n^* = \mathbb{E}Y^* = \alpha + \mathbf{x}_\gamma^{*T}\boldsymbol{\beta}_\gamma,$$

where plim denotes convergence in probability and the probability measure here is the sampling distribution under model $\mathcal{M}_\gamma$.

Theorem 4. The BMA estimators $\hat{Y}_n^*$ in (23) under empirical Bayes, the hyper-g, hyper-g/n, and Zellner–Siow priors are consistent under prediction.

Proof. When $\mathcal{M}_\gamma = \mathcal{M}_N$, by the consistency of least squares estimators, we have $\|\hat{\boldsymbol{\beta}}_\gamma\| \to 0$, so the consistency of the BMA estimators follows.

When $\mathcal{M}_\gamma \ne \mathcal{M}_N$, by Theorem 3, $\pi(\mathcal{M}_\gamma|\mathbf{Y})$ goes to 1 in probability. Using the consistency of the least squares estimators, it suffices to show that

$$\operatorname{plim}_n \int_0^\infty \frac{g}{1+g}\pi(g|\mathcal{M}_\gamma, \mathbf{Y})\,dg = 1. \tag{24}$$

The preceding integral can be rewritten as

$$\frac{\int_0^\infty (g/(1+g))L(g)\pi(g)\,dg}{\int_0^\infty L(g)\pi(g)\,dg},$$

where $L(g) = (1+g)^{-p_\gamma/2}[1 - R_\gamma^2 \frac{g}{1+g}]^{-(n-1)/2}$ is maximized at $\hat{g}_\gamma^{\mathrm{EBL}}$ given by (9). Applying a Laplace approximation to the denominator and numerator of the preceding ratio along the lines of (B.7), we have

$$\int_0^\infty \frac{g}{1+g}\pi(g|\mathcal{M}_\gamma, \mathbf{Y})\,dg = \frac{\hat{g}_\gamma^{\mathrm{EBL}}}{1+\hat{g}_\gamma^{\mathrm{EBL}}}\left(1 + O\left(\frac{1}{n}\right)\right).$$

It is clear that $\hat{g}_\gamma^{\mathrm{EBL}}$ goes to ∞ in probability under $\mathcal{M}_\gamma$, and, hence, we can conclude that the limit in (24) is equal to 1, as we desired. The consistency of the local empirical Bayes procedures is a direct consequence. Because $\hat{g}^{\mathrm{EBG}}$ is of the same order as $\hat{g}^{\mathrm{EBL}}$, the consistency of the global empirical Bayes follows.

We have shown that Zellner–Siow and hyper-g/n priors are consistent for model selection under a 0–1 loss for any assumed true model. Additionally, the hyper-g priors and EB procedures are also consistent for model selection for all models except the null model. However, all of the mixture of g priors and EB procedures are consistent for prediction under squared error loss. Because the asymptotic results do not provide any discrimination among the different methods, we conducted a simulation study to compare mixture of g priors with empirical Bayes and other default model selection procedures.

5. SIMULATION STUDY

We generated data for the simulation study as $\mathbf{Y} = \mathbf{1}_n\alpha + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, with $\boldsymbol{\epsilon} \sim \mathrm{N}(0, \mathbf{I}_n/\phi)$, $\phi = 1$, $\alpha = 2$, and sample size $n = 100$. Following Cui and George (2007), we set $\mathbf{X}^T\mathbf{X} = \mathbf{I}_p$ but took $p = 15$ so that all models may be enumerated, thus avoiding extra Monte Carlo variation due to stochastic search of the model space. For a model of size p_γ, we generated $\boldsymbol{\beta}_\gamma$ as $\mathrm{N}(0, g/\phi\mathbf{I}_{p_\gamma})$ and set the remaining components of $\boldsymbol{\beta}$ to 0. We used $g = 5, 25$ as in Cui and George (2007), representing weak and strong signal-to-noise ratios.

We used squared error loss

$$\mathrm{MSE}(m) = \|\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\hat{\boldsymbol{\beta}}^{(m)}\|^2,$$

where $\hat{\boldsymbol{\beta}}^{(m)}$ is the estimator of $\boldsymbol{\beta}$ using method m, which may entail both selection and shrinkage. We compared the 10 methods listed in Table 1. Among them, the theoretical mean squared errors (MSEs) for the oracle and full methods are known to be $p_\gamma + 1$ and $p + 1$, respectively. For each Bayesian method, we considered the following criteria for model choice: selection of the highest posterior probability model (HPM), selection of the median probability model (MPM), which is defined as the model where a variable is included if the marginal inclusion probability $p(\beta_j \neq 0|\mathbf{Y}) > 1/2$ (Barbieri and Berger 2004), and Bayesian model averaging (BMA). In both HPM and MPM, the point estimate is the posterior mean of $\boldsymbol{\beta}_\gamma$ under the selected model. For BIC, the log marginal for model $\mathcal{M}_\gamma$ is defined as

$$\log p(\mathbf{Y}|\mathcal{M}_\gamma) \equiv -\frac{1}{2}\{n\log(\hat{\sigma}_\gamma^2) + p_\gamma\log(n)\}, \qquad (25)$$

where $\hat{\sigma}_\gamma^2 = \mathrm{RSS}_\gamma/n$ is the maximum likelihood estimator (MLE) of σ^2 under model $\mathcal{M}_\gamma$. These marginals are used for calculating posterior model probabilities for determining HPM and MPM and for calculating quantities under model averaging. For AIC, the penalty for model complexity in the log marginal is taken as $2p_\gamma$ rather than $p_\gamma\log(n)$ as in (25) for BIC. For both AIC and BIC, the point estimate of $\boldsymbol{\beta}_\gamma$ is the ordinary least squares (OLS) estimate under model $\mathcal{M}_\gamma$. Uniform prior probabilities on models were used throughout.

For each value of g and $p_\gamma = 0, 1, \ldots, 15$, we generated $\mathbf{Y}$ and calculated the MSE under each method. For each combination of method, g and true model size p_γ, this was replicated 1,000 times, and the average MSE was reported.

Average MSE results from the simulation study are shown in Figure 1. For $p_\gamma > 0$, MSE results for the two EB procedures, the Zellner–Siow (ZS) null-based approach, and the hyper-g priors are virtually identical, outperforming other default specifications for a wide range of model sizes (to simplify the figure, only the hyper-g with $a = 3$ is pictured as the other hyper-g results are indistinguishable from it). While the ZS full-based procedure performs better than the other fully Bayes procedures when the full model generates the data, overall it is intermediate between AIC and BIC. Differences between the fully Bayes and other procedures are most striking under the null model. Despite the theoretical asymptotic inconsistency of the global EB procedure for model selection, it is the best overall under the null model. This may be partly explained by the fact that the estimate of g "borrows" strength from all models and is more likely to estimate g as 0 when the null is true. However, with model averaging, we see that the local EB and the hyper-g prior do almost as well as the global EB procedure.

Interestingly, we found that all of the fully Bayes mixture g priors do as well as the global EB with model selection, except under the null model, whereas Cui and George (2007) found that the global EB outperformed fully Bayes procedures (under the assumption of known ϕ). We have used a uniform prior on the model space (for both the EB and the fully Bayes procedures), whereas Cui and George (2007) placed independent Bernoulli(ω) priors on variable inclusion and compared EB estimates of ω with fully Bayes procedures that place a uniform prior on ω. While we have not addressed prior distributions over models, this is an important aspect and may explain some of the difference in findings. Additionally, the simulations in Cui and George (2007) are for the $p = n$ case. Although we show that fully Bayes procedures are consistent as $n \to \infty$ for fixed p, additional study of their theoretical properties is necessary for the situation when p is close to the sample size.

Table 1. Description of the 10 methods used in the simulation study and examples

Oracle	Ordinary least squares using the true model
Full	Ordinary least squares under the full model
BIC	Bayesian information criterion
AIC	Akaike information criterion
BRIC	g prior with $g = \max(n, p^2)$
EB–L	Local EB estimate of g in g prior
EB–G	Global EB estimate of g in g prior
ZS–N	Base model in Bayes factor taken as the null model; Cauchy prior for $\boldsymbol{\beta}_\gamma$ and uniform prior on $(\alpha, \log(\phi))$
ZS–F	Base model in Bayes factor taken as the full model; Cauchy prior for $\boldsymbol{\beta}_{(-\gamma)}$ and uniform prior for $(\boldsymbol{\beta}_\gamma, \alpha, \log(\phi))$
HG3	Hyper-g prior with $a = 3$

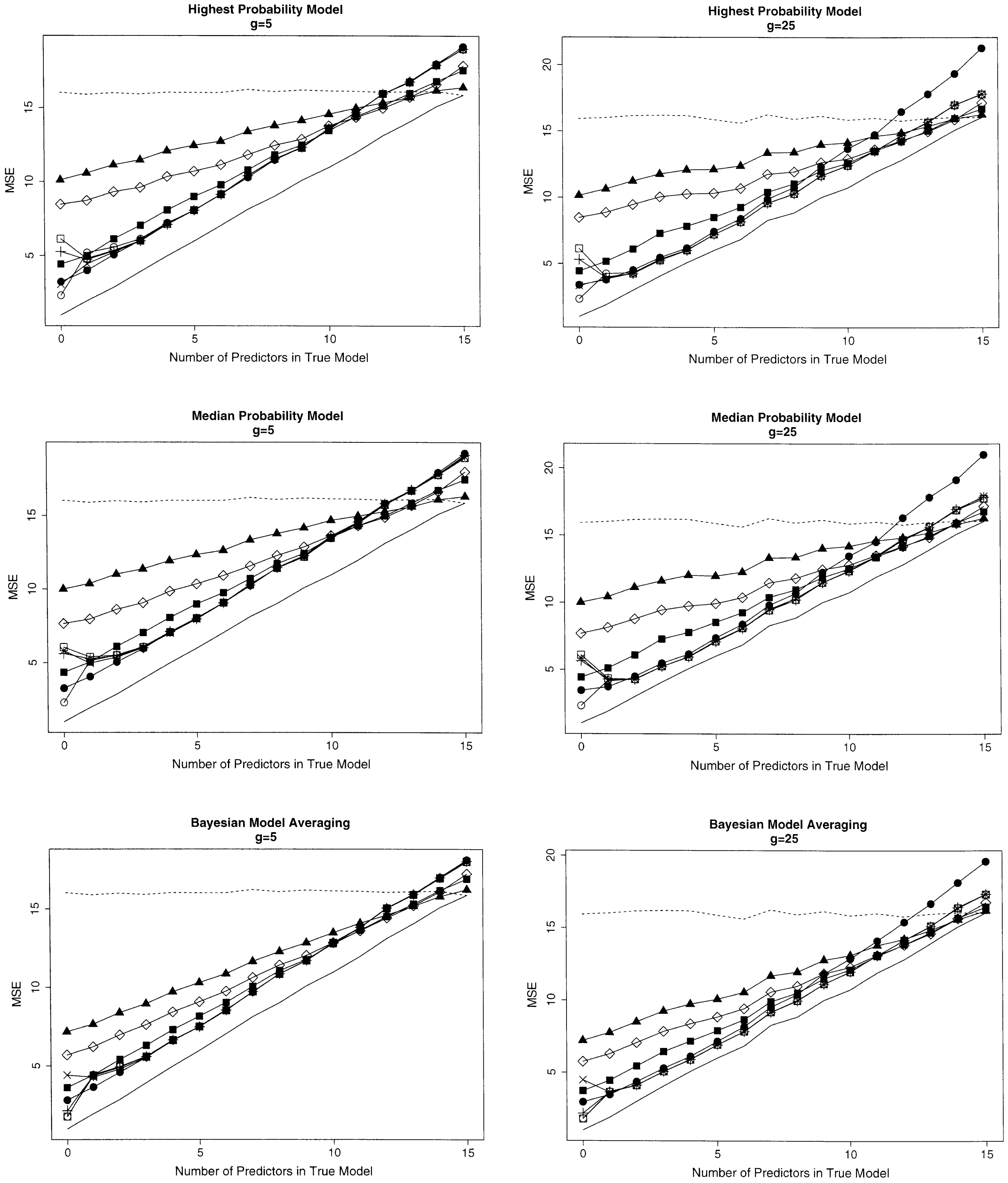

Figure 1. Average MSE from 1,000 simulations for each method with $p = 15$ and $n = 100$ using the oracle (–), AIC (▲), BIC (■), BRIC (●), EB-local (□), EB-global (○), hyper-g with $a = 3$ (+), Zellner–Siow null (×), Zellner–Siow full (◇), and full model (- -).

6. EXAMPLES WITH REAL DATA

In this section we explore the small-sample properties of the two mixture g priors on real datasets and contrast our results using other model selection procedures such as AIC, BIC, the benchmark prior (BRIC), EB-local, and EB-global.

6.1 Crime Data

Raftery et al. (1997) used the crime data of Vandaele (1978) as an illustration of Bayesian model averaging in linear regression. The cross-sectional data comprise aggregate measures of the crime rate for 47 states and include 15 explanatory variables, leading to $2^{15} = 36{,}768$ potential regression models. The prior distributions used by Raftery et al. (1997) are in the conjugate Normal–Inv-Gamma family, but their use requires specification of several hyperparameters. Fernández et al. (2001) revisited the crime data using g priors and explored the choice of g on model choice. Using the benchmark prior, which is a compromise between BIC and RIC ($g = \max\{n, p^2\} = 15^2$), Fernández et al. (2001) came to roughly the same conclusions regarding model choice and variable importance as Raftery et al. (1997). We continue along these lines by investigating how the mixture g priors affect variable importance. As in the earlier analyses of these data, all variables, except the indicator variable, have been log-transformed. The data are available in the R library MASS under UScrime, and all calculations have been done with the R package BAS available from *http://www.stat.duke.edu/~clyde/BAS*.

Table 2 illustrates the effect of 10 prior choices on the marginal inclusion probabilities and the median probability model. BRIC is equivalent to the benchmark prior of Fernández et al. (2001) and in this case corresponds to RIC ($g = 15^2$). [Our results may differ slightly from those of Fernández et al. 2001, who used Markov chain Monte Carlo (MCMC) to sample high-probability models and estimate quantities based on ergodic averages, while we enumerate the model space and calculate all quantities analytically.] BRIC leads to the most parsimonious model and is more conservative than BIC ($g \approx 47$) or any of the other approaches in terms of variable inclusion with marginal inclusion probabilities shrunk toward 0. In contrast to the predetermined values of g in BRIC, the global EB estimate of g is 19.5 (SE = 11.2), while the local EB for g under the highest probability model is 24.3 (SE = 13.6), dramatically lower than g under BRIC. The fully Bayesian mixture g priors HG3, HG4, and ZS-null all lead to very similar marginal inclusion probabilities as the data-based adaptive empirical Bayes approaches EB-global and EB-local. These all lead to the same median probability model. As is often the case, here the marginal inclusion probabilities under AIC are larger, leading to the inclusion of two additional variables in the AIC median probability model compared to the median probability model under the mixture g priors and the EB priors.

6.2 Ozone

Our last example uses the ground-level ozone data analyzed by Breiman and Friedman (1985) and, more recently, by Miller (2001) and Casella and Moreno (2006). The dataset consists of daily measurements of the maximum ozone concentration near Los Angeles and eight meteorological variables (a description of the variables is given in App. C). Following Miller (2001) and Casella and Moreno (2006), we examine regression models using the eight meteorological variables, plus interactions and squares, leading to 44 possible predictors. Enumeration of all possible models is not feasible, so instead we use stochastic search to identify the highest probability models using the R package BAS. We compared the different procedures on the basis of out-of-sample predictive accuracy by taking a random split (50/50) of the data and reserving half for calculating the average prediction error (root mean squared error; RMSE) under each method, where

$$\mathrm{RMSE}(\mathcal{M}) = \sqrt{\frac{\sum_{i\in V}(Y_i - \hat{Y}_i)^2}{n_V}},$$

V is the validation set, n_V is the number of observations in the validation set ($n_V = 165$), and $\hat{Y}_i$ is the predicted mean for Y_i under the highest probability model. From Table 3, the two EB procedures, BIC, and HG ($\alpha = 4$) all identify the same model. The ZS procedure with the full-based Bayes factors and AIC lead to selection of the most complex models with 11 and 18 variables, respectively. While the hyper-g prior with $\alpha = 3$ has the smallest RMSE, overall the differences in RMSE are not enough to suggest that any method dominates the others in terms of prediction. Based on their theoretical properties, we continue to recommend the mixtures of g priors, such as the hyper-g prior with $\alpha = 3$.

7. DISCUSSION

In this article we have shown how mixtures of g priors may resolve several consistency issues that arise with fixed g priors, while still providing computational tractability. Both real and simulated examples have demonstrated that the mixture g

Table 2. Marginal inclusion probabilities for each variable under 10 prior scenarios

	BRIC	HG-n	HG3	HG4	EB-L	EB-G	ZS-N	ZS-F	BIC	AIC
log(AGE)	.75	.85	.84	.84	.85	.86	.85	.88	.91	.98
S	.15	.27	.29	.31	.29	.29	.27	.36	.23	.36
log(Ed)	.95	.97	.97	.96	.97	.97	.97	.97	.99	1.00
log(Ex0)	.66	.66	.66	.66	.67	.67	.67	.68	.69	.74
log(Ex1)	.39	.45	.47	.47	.46	.46	.45	.50	.40	.47
log(LF)	.08	.20	.23	.24	.22	.21	.20	.30	.16	.34
log(M)	.09	.20	.23	.24	.22	.22	.20	.30	.17	.39
log(N)	.23	.37	.39	.39	.39	.38	.37	.46	.36	.57
log(NW)	.51	.69	.69	.68	.70	.70	.69	.75	.78	.92
log(U1)	.11	.25	.27	.28	.27	.27	.25	.35	.23	.41
log(U2)	.45	.61	.61	.61	.62	.62	.61	.68	.70	.86
log(W)	.18	.35	.38	.39	.38	.38	.36	.47	.36	.64
log(X)	.99	1.00	.99	.99	1.00	1.00	1.00	.99	1.00	1.00
log(prison)	.78	.89	.89	.89	.90	.90	.90	.92	.95	.99
log(time)	.19	.37	.38	.39	.39	.38	.37	.47	.41	.65

NOTE: The median probability model includes variables where the marginal inclusion probability is greater than or equal to 1/2.

Table 3. Out-of-sample prediction errors for the ozone data with the highest probability model using linear, quadratic, and interactions of the eight meteorological variables under each of the priors

Prior	$\mathcal{M}^*$	R^2	$p_{\mathcal{M}^*}$	RMSE($\mathcal{M}^*$)
HG3	hum, ibh, dpg, ibt, ibh.dpg, hum.ibt, ibh.ibt	.762	7	4.4
HG4	ibh, dpg, ibt, dpg^2, ibt^2, hum.ibh	.757	6	4.5
HG-n	ibh, dpg, ibt, dpg^2, ibt^2, hum.ibh	.757	6	4.5
ZS–N	ibh, dpg, ibt, dpg^2, ibt^2, hum.ibh	.757	6	4.5
ZS–F	hum, ibh, dpg, ibt, hum^2, ibt^2, vh.temp, vh.ibh, temp.dpg, ibh.dpg, hum.ibt	.780	11	4.4
AIC	hum, ibh, dpg, ibt, hum^2, ibt^2, vh.wind, vh.ibh, wind.ibh, vh.dgp, ibh.dpg, vh.ibt, wind.ibt, humid.ibt, wind.vis, dpg.vis	.798	18	4.6
BIC	ibh, dpg, ibt, dpg^2, ibt^2, hum.ibh	.757	6	4.5
BRIC	dpg, ibt, hum.ibt	.715	3	4.6
EB–L	ibh, dpg, ibt, dpg^2, ibt^2, hum.ibh	.757	6	4.5
EB–G	ibh, dpg, ibt, dpg^2, ibt^2, hum.ibh	.757	6	4.5

NOTE: RMSE is the square root of the mean of the squared prediction errors on the validation set using predictions from the highest probability model.

priors perform as well as or better than other default choices. Because the global EB procedure must be approximated when the model space is too large to enumerate, the mixture g priors such as the Zellner–Siow Cauchy prior or the hyper-g priors provide excellent alternatives in terms of adaptivity and shrinkage properties and robustness to misspecification of g, and still permit fast marginal likelihood calculations, a necessary feature for exploring high-dimensional model spaces.

Priors on the model space are also critical in model selection and deserve more attention. Many Bayesian variable selection implementations place independent Bernoulli(ω) priors on variable inclusion. Setting $\omega = 1/2$ corresponds to a uniform prior on the model space, which we have used throughout. Alternatively, one may specify a hierarchical model over the model space by placing a prior on ω and take a fully Bayesian or EB approach. For example, Cui and George (2007) use a uniform prior on ω, which induces a uniform prior over the model size and, therefore, favors models with small or large sizes, and contrast this to EB estimates of ω. Other types of priors include dilution priors (George 1999), which "dilute" probabilities across neighborhood of similar models, and priors that correct the so-called "selection effect" in choice among many models (Jeffreys 1961; Zellner and Min 1997).

While we have assumed that $\mathbf{X}_\gamma$ is full rank, the g-prior formulation may be extended to the non-full-rank setting such as in analysis of variance (ANOVA) models by replacing the inverse of $\mathbf{X}_\gamma^T\mathbf{X}_\gamma$ in the g prior with a generalized inverse and p_γ by the rank of the projection matrix. Because marginal likelihood calculations depend only on properties of the projection on the space spanned by $\mathbf{X}_\gamma$ (which is unique), results will not depend on the choice of generalized inverse. For the hyper-g priors, the rank p_γ must be less than $(n - 3 - a)$ in order for the Gaussian hypergeometric function to be finite and posterior distributions to be proper. For the Zellner–Siow priors, we require $p_\gamma < n - 2$. For the $p > n$ setting, proper posterior distributions can be ensured by placing zero prior probability on models of rank greater than or equal to $(n - 3 - a)$ or $(n - 2)$ for the hyper-g and Zellner–Siow priors, respectively. In the small-n setting, this is not an unreasonable restriction.

For the large-p, small-n setting, independent priors on regression coefficients have become popular for inducing shrinkage (Wolfe, Godsill, and Ng 2004; Johnstone and Silverman 2005). Many of these priors are represented as scale mixtures of normals and, therefore, may be thought of as scale mixtures of independent g priors. In conjunction with point masses at 0, these independent g priors also induce sparsity without restricting a priori the number of variables in the model; however, the induced prior distribution of the mean is no longer invariant to the choice of basis for a given model. Furthermore, closed-form solutions for marginal likelihoods are no longer available and Monte Carlo methods must be used to explore both model spaces and parameter spaces. While beyond the scope of this article, an unresolved and interesting issue is the recommendation of multivariate versus independent mixtures of g priors.

APPENDIX A: LAPLACE APPROXIMATIONS TO BAYES FACTORS

Here, we provide details of Laplace approximations to the integral (13) and to posterior expectations of $g/(1+g)$ under the Zellner–Siow and hyper-g priors.

For integrals of the form $\int_\Theta \exp(h(\theta))\,d\theta$, we make repeated use of the fully exponential Laplace approximation (Tierney and Kadane 1986), based on expanding a smooth unimodal function $h(\theta)$ in a Taylor series expansion about $\hat\theta$, the mode of h. The Laplace approximation leads to an $O(n^{-1})$ approximation to the integral,

$$\int_\Theta \exp(h(\theta))\,d\theta \approx \sqrt{2\pi}\,\hat\sigma_h h(\hat\theta), \tag{A.1}$$

where

$$\hat\sigma_h = \left[\frac{-d^2 h(\theta)}{d\theta^2}\bigg|_{\theta=\hat\theta}\right]^{-1/2}. \tag{A.2}$$

Write $L(g) = (1+g)^{(n-p_\gamma-1)/2}[1 + (1 - R_\gamma^2)g]^{-(n-1)/2}$ for the (marginal) likelihood of g. For positive functions $f(g)$, let $h_1(g) = \log(f(g)) + \log(L(g)) + \log(\pi(g))$. Define also $h_2(g) = \log(L(g)) + \log(\pi(g))$. The expected value of $f(g)$ is obtained as the ratio of two Laplace approximations:

$$\mathbb{E}(f(g)|\mathcal{M}_\gamma, \mathbf{Y}) = \frac{\int_0^\infty \exp(h_1(g))\,dg}{\int_0^\infty \exp(h_2(g))\,dg} \approx \frac{\hat\sigma_{h_1}}{\hat\sigma_{h_2}}\frac{\exp(h_1(\hat g_1))}{\exp(h_2(\hat g_2))},$$

where $\hat g_1$ and $\hat g_2$ are the modes of $h_1(g)$ and $h_2(g)$, respectively, and $\hat\sigma_{h_1}$ and $\hat\sigma_{h_2}$ are defined as in (A.2) using $h_1(g)$ and $h_2(g)$—instead of $h(\theta)$—evaluated at their respective modes. The Laplace approximation to the integral in the denominator is exactly the required expression for the Bayes factor in (13). Using a Laplace approximation to estimate both the numerator and the denominator integrals leads to an $O(n^{-2})$ approximation to $\mathbb{E}(f(g)|\mathcal{M}_\gamma, \mathbf{Y})$ (Tierney and Kadane 1986).

A.1 Laplace Approximations With Zellner–Siow Priors

For the Zellner–Siow prior, the univariate integral for the marginal likelihood in (13) and, more generally, for $\mathbb{E}[g^a(1+g)^b]|\mathbf{Y}, \mathcal{M}_\gamma]$ is given by

$$\int_0^\infty \exp(h(g))\,dg \equiv \int_0^\infty (1+g)^{(n-p_\gamma-1+2b)/2} \times [1 + (1 - R_\gamma^2)g]^{-(n-1)/2} g^{a-3/2} e^{-n/(2g)}\,dg,$$

where the marginal likelihood corresponds to $a = b = 0$ and the numerator of the expected value of the shrinkage factor corresponds to setting $a = 1, b = -1$. The mode, $\hat{g}_h$, is provided by the solution to the cubic equation

$$\begin{aligned}
&-(1 - R_\gamma^2)\big(p_\gamma + 3 - 2(a - b)\big)g^3 \\
&\quad + \big[n - p_\gamma + 2b - 4 + 2\big(a + b - (1 - a)(1 - R_\gamma^2)\big)\big]g^2 \\
&\quad + [n(2 - R_\gamma^2) + 2a - 3]g + n = 0, \qquad \text{(A.3)}
\end{aligned}$$

with second derivative

$$\frac{d^2h(g)}{dg^2} = \frac{1}{2}\left[\frac{(n-1)(1 - R_\gamma^2)}{(1 + g(1 - R_\gamma^2))^2} - \frac{n - p_\gamma - 1}{(1+g)^2} + \frac{3 - 2a}{g^2} - \frac{2n}{g^3}\right].$$

We next show that there is a unique, positive mode for $h(g)$ in the interior of the parameter space. In general, there are three (possibly complex) roots, available in closed form, for the solution to (A.3) (see Abramowitz and Stegun 1970, p. 17). For the marginal likelihood ($a = b = 0$) and numerator of the expected value of the shrinkage factor ($a = 1, b = -1$), it is clear that

$$\lim_{g\to 0} \frac{dh(g)}{dg} > 0 \qquad \text{and} \qquad \lim_{g\to\infty} \frac{dh(g)}{dg} < 0,$$

and because $h(g)$ is continuous in $\Re^+$, there exists at least one positive (real) solution in the interior of the parameter space. The following argument shows that there exists only one positive solution: If (A.3) has more than one real solution, then all three solutions are real. From (A.3), we know that the product of the three solutions is equal to $n/[(1 - R_\gamma^2)(p_\gamma + 3) - 2(a - b)]$, which is positive for the functions of interest. Because we already know that one of the solutions is positive, the other two have to be both negative or both positive. However, the latter cannot occur because (A.3) implies that the summation of all pair products of the three solutions is negative.

A.2 Laplace Approximations With Hyper-*g* Priors

For the hyper-g prior, the integrand function $\exp(h_1(g)) = L(g) \times \pi(g)$ is maximized at g equal to

$$\hat{g}_\gamma = \max\left(\frac{R_\gamma^2/(p_\gamma + a)}{(1 - R_\gamma^2)/(n - 1 - p_\gamma - a)} - 1, 0\right).$$

For $a = 0$, this is equivalent to the local EB estimate of g. While this provides excellent agreement for large n and R_γ^2 near 1, when $\hat{g}_\gamma = 0$ the usual large-sample Laplace approximation to the integral is not valid because the maximizer is a boundary point.

There are several alternatives to the standard Laplace approximation. One approach when the mode is on the boundary is to use a Laplace approximation over the expanded parameter space as in Pauler, Wakefield, and Kass (1999). The likelihood function, $L(g)$, is well defined over an extended parameter space $g \geq -1$ with maximizer of the function $h_2(g)$ over the expanded support given by $\hat{g}_\gamma = R_\gamma^2/(p_\gamma + a)/(1 - R_\gamma^2)/(n - 1 - p_\gamma - a) - 1$. However, this gives worse behavior than the original approximation when R_γ^2 is small, that is, when $\hat{g}_\gamma = 0$.

To avoid problems with the boundary, we instead apply the Laplace approximation after a change of variables to $\tau = \log g$ in approximating all the integrals related with hyper-g priors, including the Bayes factor (17), the posterior expectation of g in (18), and the posterior shrinkage factor (14). These integrals can all be expressed as in the following general form:

$$\int_0^\infty g^{b-1}(1 + g)^{(n-1-p_\gamma-a)/2}[1 + (1 - R_\gamma^2)g]^{-(n-1)/2}\,dg,$$

where b is some constant; for example, the Bayes factor (17) corresponds to $b = 1$. With the transformation $g = e^\tau$, the preceding integral is equal to

$$\int_{-\infty}^\infty e^{(b-1)\tau}(1 + e^\tau)^{(n-1-p_\gamma-a)/2}[1 + (1 - R_\gamma^2)e^\tau]^{-(n-1)/2}e^\tau\,d\tau,$$

where the extra e^τ comes from the Jacobian of the transformation of variables. Denote the logarithm of the integrand function by $h(\tau)$. Setting $h'(\tau) = 0$ gives a quadratic equation in e^τ:

$$\begin{aligned}
&(2b - p_\gamma - a)(1 - R_\gamma^2)e^{2\tau} \\
&\qquad + [4b - p_\gamma - a + R_\gamma^2(n - 1 - 2b)2b]e^\tau + 2b = 0.
\end{aligned}$$

It is easy to check that only one of the roots is positive, which is given by

$$\begin{aligned}
e^{\hat\tau} = &\Big(\big([4b - p_\gamma - a + R_\gamma^2(n - 1 - ab)]^2 - 8b(2b - p_\gamma - a) \\
&\times (1 - R_\gamma^2)\big)^{1/2} - [4b - p_\gamma - a + R_\gamma^2(n - 1 - ab)]\Big) \\
&\Big/ \big(2(ab - p_\gamma - a)(1 - R_\gamma^2)\big).
\end{aligned}$$

The corresponding variance $\hat\sigma_h^2$ in (A.2) is equal to

$$\begin{aligned}
\hat\sigma_h^2 &= \frac{1}{-h''(\hat\tau)}\bigg|_{\tau=\hat\tau} \\
&= \left[-\frac{n - 1 - p_\gamma - a}{2}\frac{e^\tau}{(1 + e^\tau)^2}\right. \\
&\qquad \left. + \frac{n}{2}\frac{(1 - R_\gamma^2)e^\tau}{[1 + (1 - R_\gamma^2)e^\tau]^2}\right]^{-1}\bigg|_{\tau=\hat\tau}.
\end{aligned}$$

A.3 Laplace Approximations With Hyper-*g*/*n* Priors

The integrals can be expressed in the following general form:

$$\begin{aligned}
&\int_0^\infty g^{b-1}(1 + g)^{(n-1-p_\gamma)/2} \\
&\qquad \times [1 + (1 - R_\gamma^2)g]^{-(n-1)/2}\left(1 + \frac{g}{n}\right)^{-a/2} dg,
\end{aligned}$$

where b is some constant; for example, the Bayes factor (17) corresponds to $b = 1$. With the transformation $g = e^\tau$, the preceding integral is equal to

$$\begin{aligned}
&\int_{-\infty}^\infty e^{b\tau}(1 + e^\tau)^{(n-1-p_\gamma)/2} \\
&\qquad \times [1 + (1 - R_\gamma^2)e^\tau]^{-(n-1)/2}\left(1 + \frac{e^\tau}{n}\right)^{-a/2} d\tau.
\end{aligned}$$

Denote the logarithm of the integrand function by $h(\tau)$. Its derivative is equal to

$$\begin{aligned}
2\frac{dh(\tau)}{d\tau} &= 2b + (n - 1 - p_\gamma)\frac{e^\tau}{1 + e^\tau} \\
&\qquad - (n - 1)\frac{(1 - R_\gamma^2)e^\tau}{1 + (1 - R_\gamma^2)e^\tau} - a\frac{e^\tau/n}{1 + e^\tau/n}.
\end{aligned}$$

Setting $h'(\tau) = 0$ gives a cubic equation in e^τ:

$$\begin{aligned}
&2bn + (2b - p_\gamma - a)(1 - R_\gamma^2)e^{3\tau} \\
&\quad + \big\{[(1 - R_\gamma^2)(p_\gamma - ab) - R_\gamma^2]n + R_\gamma^2 + p_\gamma \\
&\quad + (2 - R_\gamma^2)(a - 2b)\big\}e^{2\tau} \\
&\quad + n[R_\gamma^2(n - 1) - p_\gamma - a/n + 2b(1/n + 2 - R_\gamma^2)]e^\tau.
\end{aligned}$$

The second derivative of h is given by

$$\begin{aligned}
2\frac{d^2h(\tau)}{d\tau^2} &= -a\frac{ne^\tau}{(1 + ne^\tau)^2} - (n - 1)\frac{(1 - R_\gamma^2)e^\tau}{(1 + (1 - R_\gamma^2)e^\tau)^2} \\
&\qquad + (n - p_\gamma - 1)\frac{e^\tau}{(1 + e^\tau)^2}.
\end{aligned}$$

APPENDIX B: PROOF FOR THEOREM 3

We first cite some preliminary results from Fernández et al. (2001) without proof. Under the assumed true model $\mathcal{M}_{\gamma}$,

1. If $\mathcal{M}_{\gamma}$ is nested within or equal to a model $\mathcal{M}_{\gamma'}$, then

$$\operatorname*{plim}_{n\to\infty} \frac{\mathrm{RSS}_{\gamma'}}{n} = \frac{1}{\phi}. \tag{B.1}$$

2. For any model $\mathcal{M}_{\gamma'}$ that does not contain $\mathcal{M}_{\gamma}$, under assumption (22),

$$\operatorname*{plim}_{n\to\infty} \frac{\mathrm{RSS}_{\gamma'}}{n} = \frac{1}{\phi + b_{\gamma'}}, \tag{B.2}$$

where $\mathrm{RSS}_{\gamma} = (1 - R^2_{\gamma})\|\mathbf{Y} - \bar{\mathbf{Y}}\|^2$ is the residual sum of squares.

Case 1: $\mathcal{M}_{\gamma} \neq \mathcal{M}_N$

We first show that the consistency result holds true for the local EB estimate. For any model $\mathcal{M}_{\gamma'}$ such that $\mathcal{M}_{\gamma'} \cap \mathcal{M}_{\gamma} \neq \emptyset$, because $R^2_{\gamma'}$ goes to some constant strictly between 0 and 1 in probability, we have

$$\hat{g}^{\mathrm{EBL}}_{\gamma'} = \left[\frac{R^2_{\gamma'}/p_{\gamma'}}{(1 - R^2_{\gamma'})/(n - 1 - p_{\gamma'})} - 1\right](1 + o_P(1)) \tag{B.3}$$

and

$$\mathrm{BF}_{\mathrm{EBL}}[\mathcal{M}_{\gamma'} : \mathcal{M}_N]$$
$$\overset{P}{\sim} \frac{1}{(1 - R^2_{\gamma'})^{(n-1-p_{\gamma'})/2}} \frac{(n - 1 - p_{\gamma'})^{(n-1-p_{\gamma'})/2}}{(n - 1)^{(n-1)/2}}, \tag{B.4}$$

where the notation $X_n \overset{P}{\sim} Y_n$ means that X_n/Y_n goes to some nonzero constant in probability. Therefore,

$$\mathrm{BF}_{\mathrm{EBL}}[\mathcal{M}_{\gamma'} : \mathcal{M}_{\gamma}] \overset{P}{\sim} \frac{1}{n^{(p_{\gamma'} - p_{\gamma})/2}} \left(\frac{\mathrm{RSS}_{\gamma}/n}{\mathrm{RSS}_{\gamma'}/n}\right)^{n/2}. \tag{B.5}$$

Consider the following three situations:

a. $\mathcal{M}_{\gamma} \cap \mathcal{M}_{\gamma'} \neq \emptyset$ and $\mathcal{M}_{\gamma} \not\subseteq \mathcal{M}_{\gamma'}$. Applying (B.1) and (22), we have

$$\operatorname*{plim}_{n\to\infty} \left(\frac{\mathrm{RSS}_{\gamma}/n}{\mathrm{RSS}_{\gamma'}/n}\right)^{n/2} = \lim_{n\to\infty} \left(\frac{1/\phi}{1/\phi + b_{\gamma'}}\right)^{n/2},$$

which converges to 0 (in probability) exponentially fast with respect to n since $b_{\gamma'}$ is a positive constant. Therefore, no matter what value $p_{\gamma'} - p_{\gamma}$ takes, the Bayes factor (B.5) goes to 0 (in probability).

b. $\mathcal{M}_{\gamma} \subseteq \mathcal{M}_{\gamma'}$. By the result in Fernández et al. (2001), $(\mathrm{RSS}_{\gamma}/\mathrm{RSS}_{\gamma'})^{n/2}$ converges in distribution to $\exp(\chi^2_{p_{\gamma'} - p_{\gamma}}/2)$. Combining this result with the fact that the first term goes to 0 (because $p_{\gamma'} > p_{\gamma}$), we have that the Bayes factor converges to 0.

c. $\mathcal{M}_{\gamma} \cap \mathcal{M}_{\gamma'} = \emptyset$. In this case we have $nR^2_{\gamma'}$ converging in distribution to $\chi^2_{p_{\gamma'}}/(1 + \phi b_{\gamma})$, where b_{γ} is defined in (B.2). Because

$$\mathrm{BF}_{\mathrm{EBL}}[\mathcal{M}_{\gamma'} : \mathcal{M}_N] = \frac{(1 + g)^{(n-1-p_{\gamma'})/2}}{[1 + (1 - R^2_{\gamma'})g]^{(n-1)/2}}$$
$$\le (1 - R^2_{\gamma'})^{-(n-1)/2},$$

we have $\mathrm{BF}_{\mathrm{EBL}}[\mathcal{M}_{\gamma'} : \mathcal{M}_N] = O_P(1)$. On the other hand, because R^2_{γ} goes to a constant strictly between 0 and 1, by (B.4) we have

$$\mathrm{BF}_{\mathrm{EBL}}[\mathcal{M}_{\gamma} : \mathcal{M}_N] \overset{P}{\sim} (n - 1)^{-p_{\gamma}/2}(1 - R^2_{\gamma})^{-n/2},$$

where the second term goes to ∞ exponentially fast. So the Bayes factor goes to 0 in probability.

Next, we show the consistency result for the global EB approach. Recall that

$$\hat{g}^{\mathrm{EBG}} = \arg\max_{g>0} \sum_{\gamma'} p(\mathcal{M}_{\gamma'})\,\mathrm{BF}[\mathcal{M}_{\gamma'} : \mathcal{M}_N].$$

Our result for the local EB estimate implies that maximizing the right side is equivalent to maximizing $\mathrm{BF}[\mathcal{M}_{\gamma} : \mathcal{M}_N]$. So $\hat{g}^{\mathrm{EBG}}$ will be of the same order as $\hat{g}^{\mathrm{EBL}}_{\gamma} = O_P(n)$. Consequently, the global empirical Bayes approach is asymptotically equivalent to the unit information prior (BIC), and the consistency result follows.

Finally, we prove the consistency result for three mixtures of g priors: Zellner–Siow priors, hyper-g priors, and hyper-g/n priors. Recall that for a $\pi(g)$-mixture g prior,

$$\mathrm{BF}_{\pi}[\mathcal{M}_{\gamma'} : \mathcal{M}_N]$$
$$= \int L(g)\pi(g)\,dg$$
$$= \int \left(1 - R^2_{\gamma'}\frac{g}{1 + g}\right)^{-(n-1)/2} \frac{\pi(g)}{(1 + g)^{p_{\gamma'}/2}}\,dg. \tag{B.6}$$

A variation on the Laplace approximation uses the MLE and the square root of the reciprocal of the observed Fisher information as opposed to the posterior mode and to (A.1). The relative error is still $O(1/n)$ (Kass and Raftery 1995). As such, we can write (B.6) as

$$\mathrm{BF}_{\pi}[\mathcal{M}_{\gamma'}, \mathcal{M}_{\gamma}]$$
$$= \frac{\pi(\hat{g}^{\mathrm{EBL}}_{\gamma'})}{\pi(\hat{g}^{\mathrm{EBL}}_{\gamma})} \frac{\tilde{\sigma}_{\gamma'}}{\tilde{\sigma}_{\gamma}} \mathrm{BF}_{\mathrm{EBL}}[\mathcal{M}_{\gamma'}, \mathcal{M}_{\gamma}]\left(1 + O\left(\frac{1}{n}\right)\right), \tag{B.7}$$

where

$$\tilde{\sigma}_{\gamma} = \left[\frac{-d^2L}{dg^2}\bigg|_{g=\hat{g}^{\mathrm{EBL}}_{\gamma}}\right]^{-1/2}$$

is similar to (A.2). When $\mathcal{M}_{\gamma'} \cap \mathcal{M}_{\gamma} \neq \emptyset$, $R^2_{\gamma'}$ converges in probability to a constant strictly between 0 and 1, so we have that the first two terms are bounded in probability [because $\tilde{\sigma}_{\gamma'} = O_P(n)$, $\pi(\hat{g}^{\mathrm{EBL}}_{\gamma'}) = O_P(n^{-3/2})$ for the Zellner–Siow prior; $\pi(\hat{g}^{\mathrm{EBL}}_{\gamma'}) = O_P(n^{-a/2})$ for the hyper-g prior; and $\pi(\hat{g}^{\mathrm{EBL}}_{\gamma'}) = O_P(1)$ for the hyper-g/n prior]. In light of the consistency for the local empirical Bayes approach, we have established consistency in these circumstances in the mixture case.

When $\mathcal{M}_{\gamma'} \cap \mathcal{M}_{\gamma} = \emptyset$, following the same reasoning used for the local EB estimate in this case, we have that $nR^2_{\gamma'}$ converges in distribution to $\chi^2_{p_{\gamma'}}/(1 + \phi b_{\gamma})$. Then, when n is large, we have

$$\mathrm{BF}_{\pi}[\mathcal{M}_{\gamma'} : \mathcal{M}_N] \le \exp(C\chi^2_{p_{\gamma'}})(1 + \epsilon)\int \frac{\pi(g)}{(1 + g)^{p_{\gamma'}/2}}\,dg$$
$$\le 2\exp(C\chi^2_{p_{\gamma'}}), \tag{B.8}$$

where C is some constant. Therefore, it does not go to ∞. On the other hand, we have that $\mathrm{BF}_{\pi}[\mathcal{M}_{\gamma} : \mathcal{M}_N]$ goes to ∞ using an approximation as in (B.7). Therefore, $\mathrm{BF}_{\pi}[\mathcal{M}_{\gamma'} : \mathcal{M}_{\gamma}] \to 0$.

Case 2: $\mathcal{M}_{\gamma} = \mathcal{M}_N$

Both the local and the global empirical Bayes approaches are not consistent in this situation. For any nonnull model $\mathcal{M}_{\gamma'}$, we have that $R^2_{\gamma'}$ goes to 0 and $\hat{g}_{\gamma'} = \max(F_{p_{\gamma'}, n-1-p_{\gamma'}} - 1, 0)$, where $F_{p_{\gamma'}, n-1-p_{\gamma'}}$ denotes a random variable with an F distribution with degrees of freedom $p_{\gamma'}$ and $n - 1 - p_{\gamma'}$. This F-distributed random variable converges in distribution to $\chi^2_{p_{\gamma'}}/p_{\gamma'}$, and, hence, $\hat{g}_{\gamma'} =$

$O_P(1)$. Therefore, because $\mathrm{BF}[\mathcal{M}_{\gamma'} : \mathcal{M}_N] \geq (1 + \hat{g}_{\gamma'})^{-p_\gamma/2}$, $\mathrm{BF}[\mathcal{M}_{\gamma'} : \mathcal{M}_N]$ cannot go to 0. The argument for the global empirical Bayes approach is similar.

Hyper-g priors are not consistent in this situation either. Indeed, because

$$\mathrm{BF}_\pi[\mathcal{M}_{\gamma'} : \mathcal{M}_N] \geq \int (1+g)^{-p'_\gamma/2} \pi(g)\, dg,$$

no proper prior that does not depend on n can lead to consistency under the null. However, the first inequality in (B.8) shows that if the preceding integral vanishes as n goes to ∞, then we achieve consistency. The prior on g must, therefore, depend on n. It is now easy to show that both the Zellner–Siow and hyper-g/n priors lead to consistency under the null.

APPENDIX C: OZONE DATA

Variables used in the ozone pollution example:

ozone	Daily ozone concentration (maximum 1-hour average, parts per million) at Upland, CA
vh	Vandenburg 500-millibar-pressure height (m)
wind	Wind speed (mph) at Los Angeles International Airport (LAX)
hum	Humidity (%) at LAX
temp	Sandburg Air Force Base temperature (°F)
ibh	Inversion base height at LAX
ibt	Inversion base temperature at LAX
dpg	Daggett pressure gradient (mmHg) from LAX to Daggett, CA
vis	Visibility (miles) at LAX

[Received December 2006. Revised August 2007.]

REFERENCES

Abramowitz, M., and Stegun, I. (1970), *Handbook of Mathematical Functions*, New York: Dover.

Barbieri, M. M., and Berger, J. (2004), "Optimal Predictive Model Selection," *The Annals of Statistics*, 32, 870–897.

Bartlett, M. (1957), "A Comment on D. V. Lindley's Statistical Paradox," *Biometrika*, 44, 533–534.

Berger, J. O., and Pericchi, L. (2001), "Objective Bayesian Methods for Model Selection: Introduction and Comparison," in *Model Selection*, ed. P. Lahiri, Hayward, CA: Institute of Mathematical Statistics, pp. 135–193.

Berger, J. O., Pericchi, L. R., and Varshavsky, J. A. (1998), "Bayes Factors and Marginal Distributions in Invariant Situations," *Sankhyā*, Ser. A, 60, 307–321.

Breiman, L., and Friedman, J. (1985), "Estimating Optimal Transformations for Multiple Regression and Correlation," *Journal of the American Statistical Association*, 80, 580–598.

Butler, R. W., and Wood, A. T. A. (2002), "Laplace Approximations for Hypergeometric Functions With Matrix Argument," *The Annals of Statistics*, 30, 1155–1177.

Casella, G., and Moreno, E. (2006), "Objective Bayes Variable Selection," *Journal of the American Statistical Association*, 101, 157–167.

Clyde, M., and George, E. I. (2000), "Flexible Empirical Bayes Estimation for Wavelets," *Journal of the Royal Statistics Society*, Ser. B, 62, 681–698.

——— (2004), "Model Uncertainty," *Statistical Science*, 19, 81–94.

Cui, W., and George, E. I. (2007), "Empirical Bayes vs. Fully Bayes Variable Selection," *Journal of the Statistical Planning and Inference*, in press.

Eaton, M. L. (1989), *Group Invariance Applications in Statistics*, Hayward, CA: Institute of Mathematical Statistics.

Fernández, C., Ley, E., and Steel, M. F. (2001), "Benchmark Priors for Bayesian Model Averaging," *Journal of Econometrics*, 100, 381–427.

Foster, D. P., and George, E. I. (1994), "The Risk Inflation Criterion for Multiple Regression," *The Annals of Statistics*, 22, 1947–1975.

George, E. (1999), Discussion of "Model Averaging and Model Search Strategies," by M. Clyde, in *Bayesian Statistics 6: Proceedings of the Sixth Valencia International Meeting*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford, U.K.: Oxford University Press, pp. 157–185.

——— (2000), "The Variable Selection Problem," *Journal of the American Statistical Association*, 95, 1304–1308.

George, E. I., and Foster, D. P. (2000), "Calibration and Empirical Bayes Variable Selection," *Biometrika*, 87, 731–747.

George, E. I., and McCulloch, R. E. (1993), "Variable Selection via Gibbs Sampling," *Journal of the American Statistical Association*, 88, 881–889.

——— (1997), "Approaches for Bayesian Variable Selection," *Statistica Sinica*, 7, 339–374.

Geweke, J. (1996), "Variable Selection and Model Comparison in Regression," in *Bayesian Statistics 5: Proceedings of the Fifth Valencia International Meeting*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford, U.K.: Oxford University Press, pp. 609–620.

Hansen, M. H., and Yu, B. (2001), "Model Selection and the Principle of Minimum Description Length," *Journal of the American Statistical Association*, 96, 746–774.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999), "Bayesian Model Averaging: A Tutorial" (with discussion), *Statistical Science*, 14, 382–401. Corrected version at *http://www.stat.washington.edu/www/research/online/hoeting1999.pdf.*

Jeffreys, H. (1961), *Theory of Probability*, New York: Oxford University Press.

Johnstone, I., and Silverman, B. (2005), "Empirical Bayes Selection of Wavelet Thresholds," *The Annals of Statistics*, 33, 1700–1752.

Kass, R. E., and Raftery, A. E. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773–795.

Kass, R. E., and Wasserman, L. (1995), "A Reference Bayesian Test for Nested Hypotheses and Its Relationship to the Schwarz Criterion," *Journal of the American Statistical Association*, 90, 928–934.

——— (1996), "The Selection of Prior Distributions by Formal Rules," *Journal of the American Statistical Association*, 91, 1343–1370; Corrigenda (1998), 93, 412.

Leamer, E. E. (1978a), "Regression Selection Strategies and Revealed Priors," *Journal of the American Statistical Association*, 73, 580–587.

——— (1978b), *Specification Searches: Ad hoc Inference With Nonexperimental Data*, New York: Wiley.

Louis, T. A. (1982), "Finding the Observed Information Matrix When Using the EM Algorithm," *Journal of the Royal Statistical Society*, Ser. B, 44, 226–233.

Miller, A. J. (2001), *Subset Selection in Regression*, New York: Chapman & Hall.

Mitchell, T. J., and Beauchamp, J. J. (1988), "Bayesian Variable Selection in Linear Regression" (with discussion), *Journal of the American Statistical Association*, 83, 1023–1032.

Pauler, D. K., Wakefield, J. C., and Kass, R. E. (1999), "Bayes Factors and Approximations for Variance Component Models," *Journal of the American Statistical Association*, 94, 1242–1253.

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997), "Bayesian Model Averaging for Linear Regression Models," *Journal of the American Statistical Association*, 92, 179–191.

Smith, M., and Kohn, R. (1996), "Nonparametric Regression Using Bayesian Variable Selection," *Journal of Econometrics*, 75, 317–343.

Strawderman, W. E. (1971), "Proper Bayes Minimax Estimators of the Multivariate Normal Mean," *The Annals of Mathematical Statistics*, 42, 385–388.

Tierney, L., and Kadane, J. (1986), "Accurate Approximations for Posterior Moments and Marginal Densities," *Journal of the American Statistical Association*, 81, 82–86.

Vandaele, W. (1978), "Participation in Illegitimate Activities: Ehrlich Revisited," in *Deterrence and Incapacitation*, Washington, DC: U.S. National Academy of Sciences, pp. 270–335.

Wolfe, P. J., Godsill, S. J., and Ng, W.-J. (2004), "Bayesian Variable Selection and Regularisation for Time-Frequency Surface Estimation," *Journal of the Royal Statistical Society*, Ser. B, 66, 575–589.

Zellner, A. (1971), *An Introduction to Bayesian Inference in Econometrics*, New York: Wiley.

——— (1986), "On Assessing Prior Distributions and Bayesian Regression Analysis With g-Prior Distributions," in *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, eds. P. K. Goel and A. Zellner, Amsterdam: North-Holland/Elsevier, pp. 233–243.

Zellner, A., and Min, C. (1997), "Bayesian Analysis, Model Selection and Prediction," in *Bayesian Analysis in Econometrics and Statistics: The Zellner View and Papers*, ed. A. Zellner, London: Edward Elgar, pp. 389–399.

Zellner, A., and Siow, A. (1980), "Posterior Odds Ratios for Selected Regression Hypotheses," in *Bayesian Statistics: Proceedings of the First International Meeting Held in Valencia*, eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, Valencia, Spain: University of Valencia Press, pp. 585–603.