Mixtures of g-priors for Bayesian Variable Selection

January 28, 2007

Abstract

Zellner's g-prior remains a popular conventional prior for use in Bayesian variable selection, despite several undesirable consistency issues. In this paper, 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 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 $\mathbf{Y} = (y_1, \ldots, y_n)^T$ normally distributed with mean vector $\boldsymbol{\mu} = (\mu_1, \ldots \mu_n)$ and covariance $\mathbf{I}_n/\phi$, where ϕ is a precision parameter (the inverse of the usual variance) and $\mathbf{I}_n$ is a $n \times n$ identity matrix. Given a set of potential predictor variables $\mathbf{X}_1, \ldots, \mathbf{X}_p$, we assume that the mean vector $\boldsymbol{\mu}$ is in the span of $\mathbf{1}_n, \mathbf{X}_1, \ldots, \mathbf{X}_p$, where $\mathbf{1}_n$ is a vector of ones of length n. The model choice problem involves selecting a subset of predictor variables which places additional restrictions on the subspace that contains the mean. Under model $\mathcal{M}_{\boldsymbol{\gamma}}$, $\boldsymbol{\mu}$ may be expressed in vector form as

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

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

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

$$p(\mathfrak{M}_{\boldsymbol{\gamma}} \mid \mathbf{Y}) = \frac{p(\mathfrak{M}_{\boldsymbol{\gamma}}) p(\mathbf{Y} \mid \mathfrak{M}_{\boldsymbol{\gamma}})}{\sum_{\mathfrak{M}_{\boldsymbol{\gamma}} \in \Gamma} p(\mathfrak{M}_{\boldsymbol{\gamma}}) p(\mathbf{Y} \mid \mathfrak{M}_{\boldsymbol{\gamma}})}. \tag{2}$$

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

$$p(\mathbf{Y} \mid \mathfrak{M}_{\boldsymbol{\gamma}}) = \int_{\boldsymbol{\Theta}_{\boldsymbol{\gamma}}} p(\mathbf{Y} \mid \boldsymbol{\theta}_{\boldsymbol{\gamma}}, \mathfrak{M}_{\boldsymbol{\gamma}}) p(\boldsymbol{\theta}_{\boldsymbol{\gamma}} \mid \mathfrak{M}_{\boldsymbol{\gamma}}) d\boldsymbol{\theta}_{\boldsymbol{\gamma}} \tag{3}$$

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

Whereas Bayesian variable selection has a long history (Leamer, 1978a,b; Mitchell and Beauchamp, 1988; Zellner, 1971, Sec 10.4), the advent of Markov chain Monte Carlo methods catalyzed Bayesian model selection and averaging in regression models (George and McCulloch, 1993, 1997; Geweke, 1996; Raftery, Madigan and Hoeting, 1997; Smith and Kohn, 1996; Clyde and George, 2004; Hoeting, Madigan, Raftery and Volinsky, 1999). Prior density choice for Bayesian model selection and model averaging, however, remains an open area (Clyde and George, 2004; Berger and Pericchi, 2001). Subjective elicitation of priors for model-specific coefficients is often precluded, particularly in high-dimensional model spaces, such as in non-parametric regression using spline and wavelet bases. Thus, it is often necessary to resort to specification of priors using some formal method (Berger and Pericchi, 2001; Kass and Wasserman, 1996). 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 $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ 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 $\boldsymbol{\theta}_{\boldsymbol{\gamma}}$ 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 $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$,

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

has been widely adopted because of computational efficiency in evaluating marginal likelihoods and model search, and perhaps most importantly, because of its simple, understandable interpretation as arising from the analysis of a conceptual sample generated using the same design matrix X as employed in the current sample (George and McCulloch, 1997; Smith and Kohn, 1996; Fernández, Ley and Steel, 2001). George and Foster (2000) showed how g could be calibrated based on many popular model selection criteria, such as AIC and BIC. 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 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.

In this paper, 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 miss-specification of g. The Zellner and 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, that leads to closed form marginal likelihoods in terms of the Gaussian hyper-$g$eometric function. Both the Cauchy and hyper-g priors provide similar computational efficiency, adaptivity

and nonlinear shrinkage found in EB procedures.

The paper is organized as follows: in the next section, 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 behaviour 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 behavior of posteriors. To investigate small sample performance, we compare the Zellner-Siow Cauchy and hyper-g priors to other approaches in a simulation study (Section 5) and in examples from the literature (Section 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:

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

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 hypothesis, he did not explicitly consider nested models, where the null hypothesis restricts the values for a sub-vector 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 non-nested comparisons. In the Bayesian

solution, the posterior probabilities of models can be expressed through the Bayes factor for pairs of hypotheses, namely,

$$p(\mathcal{M}_{\boldsymbol{\gamma}} \mid \mathbf{Y}) = \frac{p(\mathcal{M}_{\boldsymbol{\gamma}})\ \mathrm{BF}[\mathcal{M}_{\boldsymbol{\gamma}} : \mathcal{M}_b]}{\sum_{\mathcal{M}_k \in \Gamma}\ p(\mathcal{M}_k)\ \mathrm{BF}[\mathcal{M}_k : \mathcal{M}_b]}, \tag{5}$$

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

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

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

$$\mathrm{BF}(\mathcal{M}_j : \mathcal{M}_k) = \frac{\mathrm{BF}(\mathcal{M}_j : \mathcal{M}_b)}{\mathrm{BF}(\mathcal{M}_k : \mathcal{M}_b)}.$$

In principle, the choice of 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 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, $\mathcal{M}_{\boldsymbol{\gamma}}$ and $\mathcal{M}_b$, a pair of nested models. We will refer to the choice of $\mathcal{M}_N$, the null model for the base 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 $\mathcal{M}_{\boldsymbol{\gamma}}$ with the null model $\mathcal{M}_N$ through the hypotheses: $H_0 : \boldsymbol{\beta}_{\boldsymbol{\gamma}} = 0$, and $H_1 : \boldsymbol{\beta}_{\boldsymbol{\gamma}} \in \mathbb{R}^{p_{\boldsymbol{\gamma}}}$. Without loss of generality, we may assume that the columns of $\mathbf{X}_{\boldsymbol{\gamma}}$ have been centered, such that $\mathbf{1}^T\mathbf{X}_{\boldsymbol{\gamma}} = \mathbf{0}$, in which case the intercept α may be regarded as a common parameter to both $\mathcal{M}_{\boldsymbol{\gamma}}$ and $\mathcal{M}_N$. This, and arguments based on orthogonal parameterizations and invariance to scale and location transformations (Jeffreys, 1961; Eaton, 1989; Berger, Pericchi and Varshavsky, 1998), have led to the adoption of

$$p(\alpha, \phi \mid \mathcal{M}_{\boldsymbol{\gamma}}) = 1/\phi \tag{6}$$

$$\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \phi, \mathcal{M}_{\boldsymbol{\gamma}} \sim N\left(\mathbf{0}, \frac{g}{\phi}(\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{X}_{\boldsymbol{\gamma}})^{-1}\right) \tag{7}$$

as a default prior specification for α, $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$, and ϕ under $\mathcal{M}_{\boldsymbol{\gamma}}$. Note that eq(7) is the same as eq(4) which is from Zellner (1986), while eq(6) is not in Zellner (1986). In fact, most references to g-priors in the variable selection literature refer to the above version (Berger and Pericchi, 2001; George and Foster, 2000; Clyde and George, 2000; Hansen and Yu, 2001; Fernández *et al.*, 2001). Continuing with this tradition, we will also refer to the priors in (6-7) 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 (6-7), the marginal likelihood is given by

$$p(\mathbf{Y} \mid \mathcal{M}_{\boldsymbol{\gamma}}, g) = \frac{\Gamma(\frac{n-1}{2})}{\sqrt{\pi}^{(n-1)}\sqrt{n}} \left(\|\mathbf{Y} - \bar{\mathbf{Y}}\|^2\right)^{-\frac{n-1}{2}} \frac{(1+g)^{(n-1-p_{\boldsymbol{\gamma}})/2}}{\left[1 + g(1 - R^2_{\boldsymbol{\gamma}})\right]^{(n-1)/2}} \tag{8}$$

where $R^2_{\boldsymbol{\gamma}}$ is the ordinary coefficient of determination for regression model $\mathcal{M}_{\boldsymbol{\gamma}}$. Though the marginal of the null model $p(\mathbf{Y} \mid \mathcal{M}_N)$ does not involve the hyper-parameter g, it can be obtained as a special case of expression (8) with $R^2_{\boldsymbol{\gamma}} = 0$ and $p_{\boldsymbol{\gamma}} = 0$. The resulting Bayes factor for comparing any model $\mathcal{M}_{\boldsymbol{\gamma}}$ to the null model is

$$\mathrm{BF}[\mathcal{M}_{\boldsymbol{\gamma}} : \mathcal{M}_N] = (1+g)^{(n-p_{\boldsymbol{\gamma}}-1)/2} \, [1 + g(1 - R^2_{\boldsymbol{\gamma}})]^{-(n-1)/2}. \tag{9}$$

2.2 Full-Based Bayes Factors

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

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

where $\mathbf{X}_{-\gamma}$ refers to the columns of $\mathbf{X}$ excluded in model $\mathcal{M}_{\boldsymbol{\gamma}}$. Model $\mathcal{M}_{\boldsymbol{\gamma}}$ corresponds to the hypothesis $H_0 : \boldsymbol{\beta}_{-\boldsymbol{\gamma}} = 0$, while the hypothesis $H_1 : \boldsymbol{\beta}_{-\boldsymbol{\gamma}} \in \mathbb{R}^{p-p_{\boldsymbol{\gamma}}}$ corresponds to the full model $\mathcal{M}_F$, where common parameters α and $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ 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, that is, $\mathbf{1}^T[\mathbf{X}_{\boldsymbol{\gamma}}, \mathbf{X}_{-\boldsymbol{\gamma}}] = \mathbf{0}$ and $\mathbf{X}^T_{\boldsymbol{\gamma}}\mathbf{X}_{-\boldsymbol{\gamma}} = \mathbf{0}$, in order to justify treating α and $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ as common parameters to both models (Zellner and Siow, 1980). This leads to the following g-priors for the full-based Bayes factors,

$$\mathcal{M}_{\boldsymbol{\gamma}} : \quad p(\alpha, \phi, \boldsymbol{\beta}_{\boldsymbol{\gamma}}) \propto 1/\phi \tag{10}$$

$$\mathcal{M}_F : \quad p(\alpha, \phi, \boldsymbol{\beta}_{\boldsymbol{\gamma}}) \propto 1/\phi, \quad \boldsymbol{\beta}_{-\boldsymbol{\gamma}} \mid \phi \sim N\left(0, \frac{g}{\phi}(\mathbf{X}^T_{-\boldsymbol{\gamma}}\mathbf{X}_{-\boldsymbol{\gamma}})^{-1}\right) \tag{11}$$

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

$$\mathrm{BF}[\mathcal{M}_{\boldsymbol{\gamma}} : \mathcal{M}_F] = (1+g)^{-(n-p-1)/2} \left[1 + g\frac{1-R^2_F}{1-R^2_{\boldsymbol{\gamma}}}\right]^{(n-p_{\boldsymbol{\gamma}}-1)/2} \tag{12}$$

where $R^2_{\boldsymbol{\gamma}}$ and R^2_F are the usual coefficients of the determination of models $\mathcal{M}_{\boldsymbol{\gamma}}$ 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 (11) for $\boldsymbol{\beta}$ in $\mathcal{M}_F$ depends on $\mathcal{M}_{\boldsymbol{\gamma}}$, which changes with each model comparison. Nonetheless, posterior probabilities (5) can still be formally defined using the Bayes factor with respect to the full model (12). 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 (2002) 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.

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 below.

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 $p_{\boldsymbol{\gamma}}$ are fixed, the Bayes factor (9) for comparing $\mathcal{M}_{\boldsymbol{\gamma}}$ to $\mathcal{M}_N$ will go to 0. That is, large spread of the prior induced by the non-informative 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", although the phenomenon was certainly well understood and discussed by Jeffreys (1961).

Information Paradox: Suppose in comparing the null model and a particular model $\mathcal{M}_{\boldsymbol{\gamma}}$, we have overwhelming information supporting $\mathcal{M}_{\boldsymbol{\gamma}}$. For example, suppose $\|\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}\|^2$ goes to infinity, so that $R^2_{\boldsymbol{\gamma}} \to 1$ or, equivalently, the usual F-statistic goes to ∞ with both n and $p_{\boldsymbol{\gamma}}$ fixed. In any conventional sense, one would expect that $\mathcal{M}_{\boldsymbol{\gamma}}$ should receive high posterior probability and that

the Bayes factor $\text{BF}(\mathcal{M}_{\boldsymbol{\gamma}} : \mathcal{M}_N)$ would go to infinity as the information against $\mathcal{M}_N$ accumulates. However, in this situation the Bayes factor (9) with a fixed choice of g, tends to a constant $(1+g)^{(n-p_{\boldsymbol{\gamma}}-1)/2}$, as $R^2_{\boldsymbol{\gamma}} \to 1$ (Zellner, 1986; Berger and Pericchi, 2001). Since 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 Choice 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 BIC.

- **Risk Inflation Criterion:** Foster and George (1994) calibrated priors for model selection based on the Risk Inflation Criterion (RIC) and recommend 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 conclude 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 (8), an EB estimate of g is the maximum (marginal) likelihood estimate constrained to be non-negative, which turns out to be

$$\hat{g}^{\text{EBL}}_{\boldsymbol{\gamma}} = \max\{F_{\boldsymbol{\gamma}} - 1, 0\} \tag{13}$$

where

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

is the usual F statistic for testing $\boldsymbol{\beta}_{\boldsymbol{\gamma}} = 0$. An asymptotic SE based on the observed information for the estimate of g is straightforward to derive.

- **Global Empirical Bayes :** The global EB procedure assumes one common g in all models, but borrows strength from all models by estimating g from the marginal likelihood of the data, averaged over all models,

$$\hat{g}^{\text{EBG}} = \text{argmax}_{g>0} \sum_{\boldsymbol{\gamma}} p(\mathcal{M}_{\boldsymbol{\gamma}}) \frac{(1+g)^{(n-p_{\boldsymbol{\gamma}}-1)/2}}{[1+g(1-R^2_{\boldsymbol{\gamma}})]^{-(n-1)/2}}. \tag{14}$$

In general, this marginal likelihood is not tractable and does not provide a closed form solution for the MLE of g, although numerical optimization may be used (George and Foster, 2000). Here we propose an EM algorithm based on treating both the model indicator and precision ϕ as latent data in order to find the marginal maximum likelihood estimator of g. The E-step consists of the following expectations

$$\begin{aligned} E[\phi^{(i)} \mid \mathcal{M}_{\boldsymbol{\gamma}}, \mathbf{Y}, \hat{g}^{(i)}] &= \frac{n-1}{\|\mathbf{Y}-\bar{\mathbf{Y}}\|^2 \left(1 - \frac{\hat{g}^{(i)}}{1+\hat{g}^{(i)}} R^2_{\boldsymbol{\gamma}}\right)} \\ E[\mathcal{M}_{\boldsymbol{\gamma}} \mid \mathbf{Y}, \hat{g}^{(i)}] &= \frac{p(\mathbf{Y} \mid \mathcal{M}_{\boldsymbol{\gamma}}, \hat{g}^{(i)})}{\sum_{\boldsymbol{\gamma}} p(\mathbf{Y} \mid \mathcal{M}_{\boldsymbol{\gamma}}, \hat{g}^{(i)})} \equiv \hat{p}^{(i)}(\mathcal{M}_{\boldsymbol{\gamma}}|\mathbf{Y}) \end{aligned} \tag{15}$$

evaluated at the current estimate of g, and where the marginal likelihood $p(\mathbf{Y} \mid \mathcal{M}_{\boldsymbol{\gamma}}, g)$ is based on equation (8). After simplification, the marginal maximum likelihood estimate of g from the M-step is

$$\hat{g}^{(i+1)} = \max \left\{ \sum_{\boldsymbol{\gamma}} \hat{p}^{(i)}(\mathcal{M}_{\boldsymbol{\gamma}}|\mathbf{Y}) \frac{R^2_{\boldsymbol{\gamma}} / \sum_{\boldsymbol{\gamma}'} \hat{p}^{(i)}(\mathcal{M}_{\boldsymbol{\gamma}'}|\mathbf{Y}) p_{\boldsymbol{\gamma}'}}{\left(1 - \frac{\hat{g}^{(i)}}{1+\hat{g}^{(i)}} R^2_{\boldsymbol{\gamma}}\right)/(n-1)} - 1, 0 \right\} \tag{16}$$

where the terms inside the summation can be viewed as a weighted Bayesian F-statistic. The global EB estimate of g, $\hat{g}^{\text{EBG}}$ is the estimate of g from (16) after convergence. A side benefit of the EM algorithm is that the EB-Global posterior model probabilities are obtained from (15) at convergence. When the dimension of the model space prohibits enumeration, the Global EB estimates may be based on a subset of models, for example, obtained using stochastic search and sampling from the EB-local posterior. One may obtain an asymptotic

SE using the method of Louis (1982) using output from the EM algorithm or derive the information directly.

The unit information prior, risk inflation criterion, and benchmark prior do not resolve the information paradox for fixed n and p since 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 below.

Theorem 1 *In the setting of the information paradox with fixed n, $p < n$, and $R^2_{\boldsymbol{\gamma}} \to 1$, the Bayes factor (9) for comparing $\mathcal{M}_{\boldsymbol{\gamma}}$ to $\mathcal{M}_N$ goes to ∞ under either the local or global EB estimate of g.*

Proof: It is easy check that the Bayes factor (9) with $g = \hat{g}^{\text{EBL}}$ goes to infinity when $R^2_{\boldsymbol{\gamma}}$ goes to 1. It implies that the maximum of the right side of equation (14) also goes to infinity, and so does the leading term $\text{BF}[\mathcal{M}_{\boldsymbol{\gamma}} : \mathcal{M}_N]$ with $g = \hat{g}^{\text{EBG}}$. ‡

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 $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$, which typically provide 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 $p(\mathbf{Y} \mid \mathcal{M}_{\boldsymbol{\gamma}})$ is proportional to

$$\text{BF}[\mathcal{M}_{\boldsymbol{\gamma}} : \mathcal{M}_N] = \int_0^\infty (1+g)^{(n-1-p_{\boldsymbol{\gamma}})/2} \left[1 + (1 - R^2_{\boldsymbol{\gamma}})g\right]^{-(n-1)/2} \pi(g)\, dg \tag{17}$$

in the null-based approach. Similar expressions for the full-based approach can be obtained using expression (12). Under selection of a model $\mathcal{M}_{\boldsymbol{\gamma}} \neq \mathcal{M}_N$, the posterior mean of $\boldsymbol{\mu}$, $\mathbb{E}[\boldsymbol{\mu} \mid \mathcal{M}_{\boldsymbol{\gamma}}, \mathbf{Y}]$, is

$$\mathbb{E}[\boldsymbol{\mu} \mid \mathcal{M}_{\boldsymbol{\gamma}}, \mathbf{Y}] = \mathbf{1}_n \hat{\alpha} + \mathbb{E}[\frac{g}{1+g} \mid \mathcal{M}_{\boldsymbol{\gamma}}, \mathbf{Y}]\ \mathbf{X}_{\boldsymbol{\gamma}} \hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}} \tag{18}$$

where $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ are the ordinary least squares estimate of α and $\boldsymbol{\beta}$, respectively, under model $\mathcal{M}_{\boldsymbol{\gamma}}$. Under the fixed g-prior, the posterior mean for $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ 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

$$\mathbb{E}[\boldsymbol{\mu} \mid \mathbf{Y}] = \mathbf{1}_n \hat{\alpha} + \sum_{\boldsymbol{\gamma}: \mathcal{M}_{\boldsymbol{\gamma}} \neq \mathcal{M}_N} p(\mathcal{M}_{\boldsymbol{\gamma}}|\mathbf{Y}) \mathbb{E}[\frac{g}{1+g} \mid \mathcal{M}_{\boldsymbol{\gamma}}, \mathbf{Y}] \; \mathbf{X}_{\boldsymbol{\gamma}} \hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}} \tag{19}$$

which provides multiple nonlinear adaptive shrinkage through the expectation of the linear shrinkage factor and through the posterior model probabilities. Because g not only appears in Bayes factors and model probabilities, but also appears 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 importantly, 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, essentially, Jeffreys (1961) rejected normal priors 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 Cauchy priors on the regression coefficients as suitable multivariate extensions to Jeffreys' work on the univariate normal mean problem. If the two models under comparison are nested, the Zellner-Siow priors 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 (6) and

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

a multivariate Cauchy centered at the null model: $\boldsymbol{\beta}_{\boldsymbol{\gamma}} = \mathbf{0}$ 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 Inverse Gamma prior, Inv-Gamma$(g \mid 1/2, n/2)$, on g, namely,

$$\pi(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \phi) \propto \int N(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \mathbf{0}, \tfrac{g}{\phi}(\mathbf{X}_{\boldsymbol{\gamma}}^T \mathbf{X}_{\boldsymbol{\gamma}})^{-1}) \; \pi(g) \; dg \tag{20}$$

where

$$\pi(g) = \frac{(n/2)^{1/2}}{\Gamma(1/2)} \; g^{-3/2} e^{-n/(2g)} \; . \tag{21}$$

One may take advantage of the mixture of g-prior representation (20) to first integrate out $\boldsymbol{\theta}_{\boldsymbol{\gamma}}$ given g, leaving a one dimensional integral over g as in (17) which is independent of the model dimension. This one dimensional integral can be carried out using standard numerical integration or using a Laplace approximation to the integral. 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

$$\pi(g) = \frac{a-2}{2}(1+g)^{-a/2} \qquad g > 0 \tag{22}$$

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 (2004) in the case of known 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 (24), 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 (22) with $a > 2$.

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

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

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 four, 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 0.80 equal to 0.45. 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,

$$p(g \mid \mathbf{Y}, \mathcal{M}_{\boldsymbol{\gamma}}) = \frac{p_{\boldsymbol{\gamma}} + a - 2}{2 \; {}_2F_1(\frac{n-1}{2}, 1; \frac{p_{\boldsymbol{\gamma}}+a}{2}; R^2_{\boldsymbol{\gamma}})} \; (1+g)^{(n-1-p_{\boldsymbol{\gamma}}-a)/2} \big[1 + (1 - R^2_{\boldsymbol{\gamma}})g\big]^{-(n-1)/2} \tag{23}$$

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

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

$$\begin{aligned} \text{BF}[\mathcal{M}_{\boldsymbol{\gamma}} : \mathcal{M}_N] &= \frac{a-2}{2} \int_0^\infty (1+g)^{(n-1-p_{\boldsymbol{\gamma}}-a)/2} \big[1 + (1 - R^2_{\boldsymbol{\gamma}})g\big]^{-(n-1)/2} dg \\ &= \frac{a-2}{p_{\boldsymbol{\gamma}} + a - 2} \; {}_2F_1(\frac{n-1}{2}, 1; \frac{p_{\boldsymbol{\gamma}}+a}{2}; R^2_{\boldsymbol{\gamma}}) \end{aligned} \tag{24}$$

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

$$\mathbb{E}[g \mid \mathcal{M}_{\boldsymbol{\gamma}}, \mathbf{Y}] = \frac{2}{p_{\boldsymbol{\gamma}} + a - 4} \frac{{}_2F_1(\frac{n-1}{2}, 2; \frac{p_\gamma + a}{2}; R^2_{\boldsymbol{\gamma}})}{{}_2F_1(\frac{n-1}{2}, 1; \frac{p_\gamma + a}{2}; R^2_{\boldsymbol{\gamma}})} \tag{25}$$

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,

$$\begin{aligned} \mathbb{E}[\frac{g}{1+g} \mid \mathbf{Y}, \mathcal{M}_{\boldsymbol{\gamma}}] &= \frac{\int g(1+g)^{\frac{n-1-p_{\boldsymbol{\gamma}}-a}{2}-1}\big[1 + (1 - R^2_{\boldsymbol{\gamma}})g\big]^{-(n-1)/2} dg}{\int (1+g)^{\frac{n-1-p_{\boldsymbol{\gamma}}-a}{2}}\big[1 + (1 - R^2_{\boldsymbol{\gamma}})g\big]^{-(n-1)/2} dg} \\ &= \frac{2}{p_{\boldsymbol{\gamma}} + a} \frac{{}_2F_1(\frac{n-1}{2}, 2; \frac{p_\gamma + a}{2} + 1; R^2_{\boldsymbol{\gamma}})}{{}_2F_1(\frac{n-1}{2}, 1; \frac{p_\gamma + a}{2}; R^2_{\boldsymbol{\gamma}})} \end{aligned} \tag{26}$$

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 $R^2_{\boldsymbol{\gamma}}$. Similar numerical difficulties with the ${}_2F_1$ have been encountered by Butler and Wood (2002), who developed a Laplace approximation to the integral representation

$${}_2F_1(a, b; c; z) = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)} \int_0^1 \frac{t^{b-1}(1-t)^{c-b-1}}{(1-tz)^a} dt. \tag{27}$$

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 to $(0, 1)$ or $(0, \infty)$. Details of the fully exponential Laplace approximations (Tierney and Kadane, 1986) of order $O(n^{-1})$ to the expression (24), and of order $O(n^{-2})$ for the ratios in (25) and (26) are given in Appendix A.

4 Consistency

So far in this paper, 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 $R^2_{\boldsymbol{\gamma}} \to 1$ as described in subsection 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.

4.1 Information paradox

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

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

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

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 (17) is a monotonic increasing function of $R^2_{\boldsymbol{\gamma}}$. Therefore when $R^2_{\boldsymbol{\gamma}}$ goes to 1, it goes to $\int (1+g)^{(n-1-p_{\boldsymbol{\gamma}})/2}\pi(g)dg$ by the Monotone Convergence Theorem. So the non-integrability of $(1+g)^{(n-1-p_{\boldsymbol{\gamma}})/2}\pi(g)$ 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 $a \leq n - p_{\boldsymbol{\gamma}} + 1$, which in the case of the minimal sample size, suggests that we take $2 < a \leq 3$. 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 $g \leq \infty$ will resolve the paradox.

4.2 Model Selection Consistency

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

$$\text{plim}_n \; p(\mathcal{M}_{\boldsymbol{\gamma}} \mid \mathbf{Y}) = 1 \quad \text{when } \mathcal{M}_{\boldsymbol{\gamma}} \text{ is the true model,} \tag{28}$$

where "plim" denotes convergence in probability and the probability distribution here is the sampling distribution under the true model $\mathcal{M}_{\boldsymbol{\gamma}}$. By the relationship between posterior probabilities and Bayes factors (5), the consistency property (28) is equivalent to

$$\text{plim}_n \; \text{BF}[\mathcal{M}_{\boldsymbol{\gamma}'} : \mathcal{M}_{\boldsymbol{\gamma}}] = 0, \quad \text{for all } \mathcal{M}_{\boldsymbol{\gamma}'} \neq \mathcal{M}_{\boldsymbol{\gamma}}. \tag{29}$$

Under the assumption that, under any model $\mathcal{M}_{\boldsymbol{\gamma}'}$ that does not contain $\mathcal{M}_{\boldsymbol{\gamma}}$, the true model,

$$\lim_{n\to\infty} \frac{\boldsymbol{\beta}_{\boldsymbol{\gamma}}{}^t \mathbf{X}_{\boldsymbol{\gamma}}^t (I - P_{\boldsymbol{\gamma}'}) \mathbf{X}_{\boldsymbol{\gamma}} \boldsymbol{\beta}_{\boldsymbol{\gamma}}}{n} = b_{\boldsymbol{\gamma}'} \in (0, \infty) \tag{30}$$

where $P_{\boldsymbol{\gamma}'} = \mathbf{X}_{\boldsymbol{\gamma}'}(\mathbf{X}_{\boldsymbol{\gamma}'}^t \mathbf{X}_{\boldsymbol{\gamma}'})^{-1} \mathbf{X}_{\boldsymbol{\gamma}'}$ is the projection matrix onto the span of $\mathbf{X}_{\boldsymbol{\gamma}'}$. Fernández *et al.* (2001) have shown that consistency holds for BRIC and BIC. We consider the case for mixtures of g-priors and the Empirical Bayes setting.

Theorem 3 *Assume (30) holds. When the true model is not the null model, i.e., $\mathcal{M}_{\boldsymbol{\gamma}} \neq \mathcal{M}_N$, posterior probabilities under Empirical Bayes, Zellner-Siow priors, and hyper-g priors are consistent for model selection; when $\mathcal{M}_{\boldsymbol{\gamma}} = \mathcal{M}_N$, 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 was already noted by George and Foster (2000). Looking at the proof of Theorem 3 one can actually see that while the EB and hyper-g priors are not consistent in the sense of (28) under the null model, the null model will 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, leading to the hyper-g/n prior:

$$\pi(g) = \frac{a-2}{2n}(1 + g/n)^{-a/2} \tag{31}$$

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 $(\mathbf{Y}, \mathbf{X}_1, \ldots, \mathbf{X}_p)$ 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 BMA prediction given by

$$\hat{Y}_n^* = \hat{\alpha} + \sum_{\boldsymbol{\gamma}} \mathbf{x}_{\boldsymbol{\gamma}}^{*\,T} \hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}} \; p(\mathfrak{M}_{\boldsymbol{\gamma}} \mid \mathbf{Y}) \int_0^\infty \frac{g}{1+g} \pi(g \mid \mathfrak{M}_{\boldsymbol{\gamma}}, \mathbf{Y}) dg. \tag{32}$$

The local and global EB estimators can be obtained by replacing $\pi(g \mid \mathfrak{M}_{\boldsymbol{\gamma}}, \mathbf{Y})$ by a degenerate distribution with point mass at $\hat{g}^{\text{EBL}}$ and $\hat{g}^{\text{EBG}}$, respectively. When the true sampling distribution is known, i.e., $(\mathfrak{M}_{\boldsymbol{\gamma}}, \alpha, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \phi)$ 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

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

where plim denotes convergence in probability and the probability distribution here is the true sampling distribution under model $\mathfrak{M}_{\boldsymbol{\gamma}}$. We next state a result concerning prediction consistency under the mixtures considered in this paper.

Theorem 4 *The BMA estimators $\hat{Y}_n^*$ (32) under Empirical Bayes, the hyper-g, hyper-g/n and Zellner-Siow priors are consistent under prediction.*

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

When $\mathfrak{M}_{\boldsymbol{\gamma}} \neq \mathfrak{M}_N$, by Theorem 3, $\pi(\mathfrak{M}_{\boldsymbol{\gamma}} \mid \mathbf{Y})$ goes to one in probability. Using the consistency of the least squares estimators, it suffices to show that

$$\text{plim}_n \int_0^\infty \frac{g}{1+g} \pi(g \mid \mathfrak{M}_{\boldsymbol{\gamma}}, \mathbf{Y}) dg = 1 \,. \tag{33}$$

The integral above can be rewritten as

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

where $L(g) = (1+g)^{-p_{\boldsymbol{\gamma}}/2} \big[1 - R^2_{\boldsymbol{\gamma}} \frac{g}{1+g}\big]^{-(n-1)/2}$ is maximized at $\hat{g}^{\text{EBL}}_{\boldsymbol{\gamma}}$ given by (13). Applying a Laplace approximation to the denominator and numerator of the ratio above along the lines of (46), we get that

$$\int_0^\infty \frac{g}{1+g} \pi(g \mid \mathfrak{M}_{\boldsymbol{\gamma}}, \mathbf{Y}) dg = \frac{\hat{g}^{\text{EBL}}_{\boldsymbol{\gamma}}}{1 + \hat{g}^{\text{EBL}}_{\boldsymbol{\gamma}}} \; (1 + O(1/n)) \; .$$

It is clear that $\hat{g}_{\boldsymbol{\gamma}}^{\mathrm{EBL}}$ goes to infinity in probability under $\mathfrak{M}_{\boldsymbol{\gamma}}$, and hence we can conclude that the limit in (33) 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 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 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 conduct 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 $\boldsymbol{\epsilon} \sim N(0, \mathbf{I}_n/\phi)$, $\phi = 1$, $\alpha = 2$, and sample size $n = 100$. Following Cui and George (2004) we set $\mathbf{X}'\mathbf{X} = \mathbf{I}_p$, 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 $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$, the first p_γ components of $\boldsymbol{\beta}$, as $N(0, g/\phi\mathbf{I}_{p_\gamma})$ and set the remaining components of $\boldsymbol{\beta}$ to zero. We used $g = 5, 25$ as in Cui and George (2004), 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 twelve procedures, listed in Table 1. Among the procedures, the theoretical MSE for **oracle** and **full** procedures are known to be $p_{\boldsymbol{\gamma}} + 1$ and $p + 1$, respectively. For each Bayesian procedure, 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 $p(\beta_j \neq 0 \mid \mathbf{Y}) > 1/2$ (Barbieri and Berger, 2003), and Bayesian Model Averaging (BMA). In both HPM and MPM, the point estimate is the posterior mean of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ under the selected model. For BIC, the log marginal for

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 $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ and uniform prior on $(\alpha, \log(\phi))$.
ZS-F	Base model in Bayes factor taken as the Full model.
	Cauchy prior for $\boldsymbol{\beta}_{(-\boldsymbol{\gamma})}$ and uniform prior for $(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \alpha, \log(\phi))$.
HG-3	Hyper-g prior with $a = 3$
HG-4	Hyper-g prior with $a = 4$
HG-n	Hyper-g/n prior with $a = 3$.

Table 1: Description of the twelve procedures used in the simulation study and examples.

model $\mathcal{M}_{\boldsymbol{\gamma}}$ is defined as

$$\log p(\mathbf{Y} \mid \mathcal{M}_{\boldsymbol{\gamma}}) \equiv -\frac{1}{2}\{n \log(\hat{\sigma}^2_{\boldsymbol{\gamma}}) + p_{\boldsymbol{\gamma}} \log(n)\} \tag{34}$$

where $\hat{\sigma}^2_{\boldsymbol{\gamma}} = \text{RSS}_{\boldsymbol{\gamma}}/n$ is the MLE of σ^2 under model $\mathcal{M}_{\boldsymbol{\gamma}}$. These marginals are used for calculating posterior model probabilities for determining the highest posterior probability model and the median probability model, and for calculating quantities under model averaging. For AIC, the penalty for model complexity in the log marginal is taken as $2p_{\boldsymbol{\gamma}}$ rather than $\log(n)p_{\boldsymbol{\gamma}}$ in the above expression (34) for BIC. For both AIC and BIC, the point estimate of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ is the OLS estimate under model $\mathcal{M}_{\boldsymbol{\gamma}}$. 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 $p_{\boldsymbol{\gamma}}$, this was replicated 1000 times and the average MSE was reported.

Average MSE results from the simulation study are shown in Figure 1. For $p_{\boldsymbol{\gamma}} > 0$, MSE results for the two EB procedures, the Zellner-Siow 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 generate the data, overall it is intermediate between AIC and BIC. Differences between the fully Bayes and other procedures is 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. Cui and George (2004) found that the global Empirical Bayes out-performed fully Bayes procedures (under the assumption of known ϕ). We have used an uniform prior on the model space (for both the EB and fully Bayes procedures), whereas Cui and George (2004) place independent Bernoulli(ω) priors on variable inclusion, and compare EB estimates of ω with fully Bayes procedures that place a uniform (Beta) prior on ω. While we have not addressed prior distributions over models, this is an important aspect. Additionally, the simulations in Cui and George (2004) are for the $p = n$ case. Although we show that fully Bayes procedures are consistent as $n \longrightarrow \infty$ for fixed p, additional study of their theoretical properties is necessary for the situation when p is close to the sample size.

6 Examples with Real Data

In this section, we explore the small sample properties of the two mixture g-priors on real data sets 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

Fernández *et al.* (2001) revisited the crime data used by Raftery *et al.* (1997) as an illustration of prior choice for Bayesian model averaging. The cross-sectional data comprise aggregate measures of the crime rate for 47 states, and include 15 explanatory variable, leading to $2^{15} = 36,768$ potential

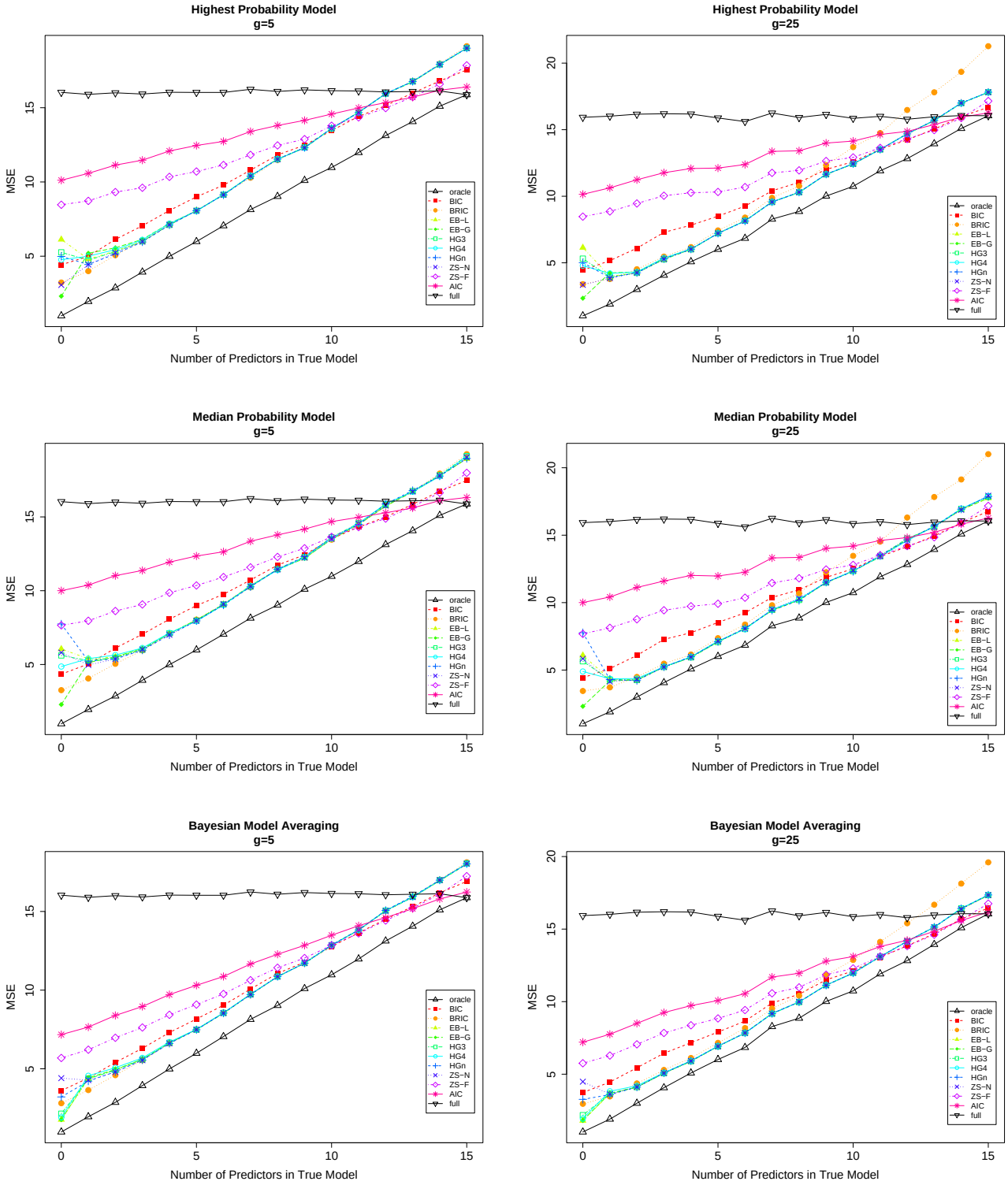

Figure 1: Average MSE from 100 simulations for each method with $p = 15$ and $n = 100$ using the oracle (—), AIC (▲), BIC (■), BRIC (●), EB-local (□), EB-global (○), hyper-g with a = 3 (+), Zellner-Siow null (×), Zellner-Siow full (◇), and full model (- - -)

regression models. As in the earlier analyses of this data, all variables, except the indicator variable S, have been log-transformed.

While the priors used by Raftery *et al.* (1997) are in the conjugate Normal-Inverse Gamma family, their use requires specification of several hyper-parameters. Using the benchmark prior, that compromises between BIC and RIC, ($g = \max\{n, p^2\} = 15^2$), Fernández *et al.* (2001) came to roughly the same conclusions as Raftery *et al.* (1997). Table 2 illustrates the effect of prior

	BRIC	HG-n	HG3	HG4	EB-L	EB-G	ZS-N	ZS-F	BIC	AIC
log(AGE)	0.75	0.85	0.84	0.84	0.85	0.86	0.85	0.88	0.91	0.98
S	0.15	0.27	0.29	0.31	0.29	0.29	0.27	0.36	0.23	0.36
log(Ed)	0.95	0.97	0.97	0.96	0.97	0.97	0.97	0.97	0.99	1.00
log(Ex0)	0.66	0.66	0.66	0.66	0.67	0.67	0.67	0.68	0.69	0.74
log(Ex1)	0.39	0.45	0.47	0.47	0.46	0.46	0.45	0.50	0.40	0.47
log(LF)	0.08	0.20	0.23	0.24	0.22	0.21	0.20	0.30	0.16	0.34
log(M)	0.09	0.20	0.23	0.24	0.22	0.22	0.20	0.30	0.17	0.39
log(N)	0.23	0.37	0.39	0.39	0.39	0.38	0.37	0.46	0.36	0.57
log(NW)	0.51	0.69	0.69	0.68	0.70	0.70	0.69	0.75	0.78	0.92
log(U1)	0.11	0.25	0.27	0.28	0.27	0.27	0.25	0.35	0.23	0.41
log(U2)	0.45	0.61	0.61	0.61	0.62	0.62	0.61	0.68	0.70	0.86
log(W)	0.18	0.35	0.38	0.39	0.38	0.38	0.36	0.47	0.36	0.64
log(X)	0.99	1.00	0.99	0.99	1.00	1.00	1.00	0.99	1.00	1.00
log(prison)	0.78	0.89	0.89	0.89	0.90	0.90	0.90	0.92	0.95	0.99
log(time)	0.19	0.37	0.38	0.39	0.39	0.38	0.37	0.47	0.41	0.65

Table 2: Marginal inclusion probabilities for each variable under 10 prior scenarios. The median probability model includes variables where the marginal inclusion probability is greater than or equal to 1/2.

choice on the marginal inclusion probabilities and median probability model. BRIC corresponds to the benchmark prior of Fernández *et al.* (2001), although here we enumerate the model space and calculate marginal likelihoods analytically, rather than using MCMC to sample high probability models. BRIC leads to the most parsimonious model, and is more conservative than BIC. In

contrast, 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 $g = 15^2$ under BRIC. The highest probability model under HG-3 ($a = 3$), HG-4 ($a = 4$), EB-local, EB-global, ZS-null, ZS-full, and AIC are all the same, although the probability on the highest posterior model ranges from 0.01 to 0.03. These low values do not imply that these are "bad" models, but rather indicate that there is significant model uncertainty about what is the best model (Hoeting *et al.*, 1999) and is a common phenomenon in model spaces with a large number of predictors where posterior mass tends to be "diluted" across many models (George, 1999). Of these, HG-3, HG-4, EB-local, EB-global, and ZS-null all lead to very similar marginal inclusion probabilities.

6.2 Ozone

Our last example uses the ground level ozone data analyzed in Breiman and Friedman (1985), and more recently by Miller (2001) and Casella and Moreno (2002). The dataset consists of daily measurements of the maximum ozone concentration near Los Angeles and 8 meteorological variables (the description for the variables are listed in Appendix C). Following Miller (2001) and Casella and Moreno (2002), we examine regression models using the 8 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. 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 (RMSE) under each method, where

$$\text{RMSE}(\mathcal{M}) = \sqrt{\frac{\sum_{i \in V} (Y_i - \hat{Y}_i)^2}{n_V}} \tag{35}$$

V is the validation set, and 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, but lead to slightly different RMSEs because of different shrinkage estimators. The ZS procedure with the full model based BFs leads to selection of the most complex model (11 variables.)

Prior	$\mathcal{M}^*$	R^2	$p_{\mathcal{M}^*}$	RMSE($\mathcal{M}^*$)
HG-3	hum, ibh, dpg, ibt, ibh.dpg, hum.ibt, ibh.ibt	0.762	7	4.436
HG-4	ibh, dpg, ibt, dpg^2, ibt^2, hum.ibh	0.757	6	4.528
HG-n	ibh, dpg, ibt, dpg^2, ibt^2, hum.ibh	0.757	6	4.523
ZSnull	ibh, dpg, ibt, dpg^2, ibt^2, hum.ibh	0.757	6	4.525
ZSfull	hum, ibh, dpg, ibt, hum^2, ibt^2, vh.temp, vh.ibh, temp.dpg, ibh.dpg, hum.ibt	0.780	11	4.434
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	0.798	18	4.570
BIC	ibh, dpg, ibt, dpg^2, ibt^2, hum.ibh	0.757	6	4.521
BRIC	dpg, ibt, hum.ibt	0.715	3	4.603
EB-L	ibh, dpg, ibt, dpg^2, ibt^2, hum.ibh	0.757	6	4.526
EB-G	ibh, dpg, ibt, dpg^2, ibt^2, hum.ibh	0.757	6	4.528

Table 3: Out of sample prediction errors for the ozone data with the highest probability model using linear, quadratic and interactions of the 8 meteorological variables under each of the priors. RMSE is the square root of the mean of the squared prediction errors on the validation set using predictions from the highest probability model.

7 Discussion

In this paper, 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-priors perform as well 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, robustness to miss-specification of g and 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. Most Bayesian variable selection implementations place independent Bernoulli(ω) priors on variable inclusion. Setting $\omega = 1/2$ corresponds to an uniform prior on the model space, which is what we have used throughout. Alternatively one can take fully Bayesian or EB approaches as in Cui and George (2004). Other types of priors include dilution priors (George, 1999) that "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 $\mathbf{X}_{\boldsymbol{\gamma}}$ is full rank, the g-prior formulation may be extended to the non-full rank setting such as in ANOVA models by replacing the inverse of $\mathbf{X}'_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}}$ in the g-prior with a generalized inverse and $p_{\boldsymbol{\gamma}}$ by the rank of the projection matrix. Because marginal likelihood calculations depend only on properties of the projection on the space spanned by $\mathbf{X}_{\boldsymbol{\gamma}}$ (which is unique), results will not depend on the choice of generalized inverse. For the hyper-g priors, the rank $p_{\boldsymbol{\gamma}}$ 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 $p_{\boldsymbol{\gamma}} < n-2$. 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 (with or without point masses at zero) have become popular for inducing sparsity without restricting *a priori* the number of variables in the model (Wolfe, Godsill and Ng, 2004; Johnstone and Silverman, to appear). Many of these priors may be represented as scale mixtures of normals, and with rescaling of the explanatory variables by their standard deviations, may be thought of as scale mixtures of independent g-priors. Even without point masses at zero, model estimates under certain independent mixtures of normals may effectively lead to variable selection as the modal estimate of a β_j may be zero. The modal estimates can be obtained by EM algorithms or by optimization procedures, but may still have many non-zero coefficients. In conjunction with point masses at zero, however, 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 paper, an unresolved issue is the recommendation of multivariate versus independent mixtures of g-priors.

References

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

Barbieri, M. M. and Berger, J. (2003) Optimal predictive model selection. *Ann. Statist.* To appear.

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*, vol. 38 of *IMS Lecture Notes – Monograph Series*, (ed. P. Lahiri), pp. 135–193. Institute of Mathematical Statistics.

Berger, J. O., Pericchi, L. R. and Varshavsky, J. A. (1998) Bayes factors and marginal distributions in invariant situations. *Sankhya, Ser. A*, **60**, 307–321.

Breiman, L. and Friedman, J. (1985) Estimating optimal transformations for multiple regression and correlation. *J. Amer. Statist. Assoc.*, pp. 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. (2002) Objective Bayes variable selection. Tech. Rep. 2002-023, Dept. of Statistics, Univ. of Florida.

Clyde, M. and George, E. I. (2000) Flexible empirical Bayes estimation for wavelets. *J. Roy.Statist. Soc. Ser. B*, **62**, 681–698.

Clyde, M. and George, E. I. (2004) Model uncertainty. *Statist. Sci.*, **19**, to appear.

Cui, W. and George, E. I. (2004) Empirical Bayes vs. fully Bayes variable selection. Tech. rep., The Wharton School, University of Pennsylvania.

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

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

Foster, D. P. and George, E. I. (1994) The risk inflation criterion for multiple regression. *Ann. Statist.*, **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.*

George, E. I. (2000) The variable selection problem. *J. Amer. Statist. Assoc.*, **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. *J. Amer. Statist. Assoc.*, **88**, 881–889.

George, E. I. and McCulloch, R. E. (1997) Approaches for Bayesian variable selection. *Statist. 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*, pp. 609–620.

Hansen, M. H. and Yu, B. (2001) Model selection and the principle of minimum description length. *J. Amer. Statist. Assoc.*, **96**, 746–774.

Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T. (1999) Bayesian model averaging: a tutorial (with discussion). *Statist. Sci.*, **14**, 382–401. Corrected version at http://www.stat.washington.edu/www/research/online/hoeting1999.pdf.

Jeffreys, H. (1961) *Theory of Probability.* Oxford Univ. Press.

Johnstone, I. and Silverman, B. (to appear) Empirical Bayes selection of wavelet thresholds. *Annals of Statistics.*

Kass, R. E. and Raftery, A. E. (1995) Bayes factors. *J. Amer. Statist. Assoc.*, **90**, 773–795.

Kass, R. E. and Wasserman, L. (1995) A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J. Amer. Statist. Assoc.*, **90**, 928–934.

Kass, R. E. and Wasserman, L. (1996) The selection of prior distributions by formal rules (Corr: 1998V93 p412). *J. Amer. Statist. Assoc.*, **91**, 1343–1370.

Leamer, E. E. (1978a) Regression selection strategies and revealed priors. *J. Amer. Statist. Assoc.*, **73**, 580–587.

Leamer, E. E. (1978b) *Specification searches: Ad hoc inference with nonexperimental data*. Wiley.

Louis, T. A. (1982) Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B, Methodological*, **44**, 226–233.

Miller, A. J. (2001) *Subset selection in regression*. Chapman & Hall.

Mitchell, T. J. and Beauchamp, J. J. (1988) Bayesian variable selection in linear regression (with discussion). *J. Amer. Statist. Assoc.*, **83**, 1023–1032.

Pauler, D. K., Wakefield, J. C. and Kass, R. E. (1999) Bayes factors and approximations for variance component models. *J. Amer. Statist. Assoc.*, **94**, 1242–1253.

Raftery, A. E., Madigan, D. and Hoeting, J. A. (1997) Bayesian model averaging for linear regression models. *J. Amer. Statist. Assoc.*, **92**, 179–191.

Smith, M. and Kohn, R. (1996) Nonparametric regression using Bayesian variable selection. *J. Econometrics*, **75**, 317–343.

Strawderman, W. E. (1971) Proper Bayes minimax estimators of the multivariate normal mean. *Ann. Math. Statist.*, **42**, 385–388.

Tierney, L. and Kadane, J. (1986) Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Assoc.*, **81**, 82–86.

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, Series B*, **66**, 575–589.

Zellner, A. (1971) *An Introduction to Bayesian Inference in Econometrics*. John Wiley & Sons, INC.

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

Zellner, A. and Min, C. (1997) Baysian analysis, model selection and prediction. In *Bayesian analysis in econometrics and statistics: the Zellner view and papers*, (ed. A. Zellner), pp. 389–399. Edward Elgar Publishing Limited.

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

Appendices

A Laplace Approximations to Bayes Factors

We provide details of Laplace approximations to the integral (17) and 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's series expansion about $\hat{\theta}$, the mode of h. The Laplace approximation leads to an $O(n^{-1})$ approximation to the integral,

$$\int_\Theta \exp(h(\theta))\ d\theta \approx \sqrt{2\pi}\ \hat{\sigma}_h\ h(\hat{\theta}) \tag{36}$$

where

$$\hat{\sigma}_h = \left[\frac{-d^2h(\theta)}{d\theta^2}|_{\theta=\hat{\theta}}\right]^{-1/2}. \tag{37}$$

Write $L(g) = (1+g)^{(n-p_\gamma-1)/2}\ [1+(1-R^2_{\boldsymbol{\gamma}})g]^{-(n-1)/2}$ for the (marginal) likelihood of g and let $h_d(g) = \log(L(g)) + \log(\pi(g))$. For $\mathbb{E}(f(g) \mid \mathcal{M}_{\boldsymbol{\gamma}}, \mathbf{Y})$ for positive functions $f(g)$, let $h_n(g) = \log(f(g)) + \log(L(g)) + \log(\pi(g))$. The expected value of $f(g)$ is obtained as the ratio of two Laplace approximations,

$$\mathbb{E}(f(g) \mid \mathcal{M}_{\boldsymbol{\gamma}}, \mathbf{Y}) = \frac{\int_0^\infty \exp(h_n(g))\ dg}{\int_0^\infty \exp(h_d(g))\ dg} \approx \frac{\hat{\sigma}_{h_n}}{\hat{\sigma}_{h_d}} \frac{\exp\left(h_n(\hat{g}_n)\right)}{\exp\left(h_d(\hat{g}_d)\right)}$$

where $\hat{g}_{h_n}$ and $\hat{g}_{h_f}$ are the modes of $h_n(g)$ and $h_d(g)$, respectively, and $\hat{\sigma}_{h_n}$ and $\hat{\sigma}_{h_d}$ are defined as in (37) using $h_n(g)$ and $h_d(g)$ evaluated at their respective modes. The Laplace approximation to the integral in the denominator is exactly the required expression for the Bayes Factor in (17). Using a Laplace approximation to approximate both the numerator and denominator integrals leads to an $O(n^{-2})$ approximation to $\mathbb{E}(f(g) \mid \mathcal{M}_{\boldsymbol{\gamma}}, \mathbf{Y})$ (Tierney and Kadane, 1986).

A.1 Laplace Approximations with Zellner-Siow Priors

For the Zellner-Siow prior, the univariate integrals for the marginal likelihood in (17) and more generally for $\mathbb{E}[g^a(1+g)^b] \mid \mathbf{Y}, \mathcal{M}_{\boldsymbol{\gamma}}]$ is given by

$$\int_0^\infty \exp\left(h(g)\right)\ dg = \int_0^\infty (1+g)^{(n-p_\gamma-1+2b)/2}\ [1+(1-R^2_{\boldsymbol{\gamma}})g]^{-(n-1)/2} g^{a-3/2} e^{-n/(2g)} dg \tag{38}$$

where the marginal likelihood corresponds to $a = b = 0$ and the numerator Laplace approximation for 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

$$-(1-R^2_{\boldsymbol{\gamma}})(p_{\boldsymbol{\gamma}}+3-2(a-b))g^3+[n-p_{\boldsymbol{\gamma}}+2b-4+2(a+b-(1-a)(1-R^2_{\boldsymbol{\gamma}}))]g^2+[n(2-R^2_{\boldsymbol{\gamma}})+2a-3]g+n = 0 \tag{39}$$

with second derivative

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

We 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 (39) (see Abramowitz and Stegun, 1970, page 17). For the marginal likelihood ($a = b = 0$) and expected value of the shrinkage factor ($a = 1, b = -1$), it is clear that

$$\lim_{g\to 0}\frac{d}{dg}h(g) > 0 \quad \text{and} \quad \lim_{g\to\infty}\frac{d}{dg}h(g) < 0,$$

and since $h(g)$ is continuous on $\Re^+$, 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 equation (39) has more than one real solution, then all three solutions are real. From equation (39), we know that the product of the three solutions is equal to $n/[(1-R^2_{\boldsymbol{\gamma}})(p_{\boldsymbol{\gamma}}+3)-2(a-b)]$ which is positive for the functions of interest. Since we already known that one of the solutions is positive, the other two have to be both negative or both positive. However, the latter cannot occur since equation (39) implies that the summation of all pair-products of the three solutions is negative.

A.2 Hyper-g Prior

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

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

For $a = 0$, this is equivalent to the local EB estimate of g. While this provides excellent agreement for large n and R_2 near one, when $\hat{g}_{\boldsymbol{\gamma}} = 0$, 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 $h_d(g)$ over the expanded support, $\hat{g}_{\boldsymbol{\gamma}} = \frac{R^2_{\boldsymbol{\gamma}}/(p_{\boldsymbol{\gamma}}+a}{(1-R^2_{\boldsymbol{\gamma}})/(n-1-p_{\boldsymbol{\gamma}}-a)} - 1$. However, this gives worse behavior than the original approximation when $R^2_{\boldsymbol{\gamma}}$ is small, i.e., when $\hat{g}_{\boldsymbol{\gamma}} = 0$.

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 (24), the posterior expectation of g (25) and the posterior shrinkage factor (19). These integrals can all be expressed as in the following general form:

$$\int_0^\infty g^{b-1}(1+g)^{(n-1-p_{\boldsymbol{\gamma}}-a)/2}\big[1+(1-R^2_{\boldsymbol{\gamma}})g\big]^{-(n-1)/2}dg$$

where b is a constant, for example, the Bayes factor (24) corresponds to $b = 1$. With the transformation $g = e^\tau$, the integral above is equal to

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

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

$$(2b-p_{\boldsymbol{\gamma}}-a)(1-R^2_{\boldsymbol{\gamma}})e^{2\tau} + \big[4b-p_{\boldsymbol{\gamma}}-a+R^2_{\boldsymbol{\gamma}}(n-1-2b)2b\big]\ e^\tau + 2b.$$

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

$$e^{\hat{\tau}} = \frac{\sqrt{\big[4b-p_{\boldsymbol{\gamma}}-a+R^2_{\boldsymbol{\gamma}}(n-1-ab)\big]^2 - 8b(2b-p_{\boldsymbol{\gamma}}-a)(1-R^2_{\boldsymbol{\gamma}})} - \big[4b-p_{\boldsymbol{\gamma}}-a+R^2_{\boldsymbol{\gamma}}(n-1-ab)\big]}{2(ab-p_{\boldsymbol{\gamma}}-a)(1-R^2_{\boldsymbol{\gamma}})}.$$

The corresponding variance $\hat{\sigma}^2_h$ in (37) is equal to

$$\hat{\sigma}^2_h = \frac{1}{-h''(\hat{\tau})}|_{\tau=\hat{\tau}} = \left[-\frac{n-1-p_{\boldsymbol{\gamma}}-a}{2}\frac{e^\tau}{(1+e^\tau)^2} + \frac{n}{2}\frac{(1-R^2_{\boldsymbol{\gamma}})e^\tau}{\big[1+(1-R^2_{\boldsymbol{\gamma}})e^\tau\big]^2}\right]^{-1}|_{\tau=\hat{\tau}}.$$

A.3 Hyper-g/n

The integrals can be expressed as in the following general form:

$$\int_0^\infty g^{b-1}(1+g)^{(n-1-p_{\boldsymbol{\gamma}})/2}\big[1+(1-R^2_{\boldsymbol{\gamma}})g\big]^{-(n-1)/2}(1+\frac{g}{n})^{-a/2}dg$$

where b is a constant, for example, the Bayes factor (24) corresponds to $b = 1$. With the transformation $g = e^\tau$, the integral above is equal to

$$\int_{-\infty}^{\infty} e^{b\tau}(1+e^\tau)^{(n-1-p_{\boldsymbol{\gamma}})/2}\big[1+(1-R^2_{\boldsymbol{\gamma}})e^\tau\big]^{-(n-1)/2}(1+\frac{e^\tau}{n})^{-a/2}d\tau.$$

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

$$2\frac{dh(\tau)}{d\tau} = 2b + (n-1-p_{\boldsymbol{\gamma}})\frac{e^\tau}{1+e^\tau} - (n-1)\frac{(1-R^2_{\boldsymbol{\gamma}})e^\tau}{1+(1-R^2_{\boldsymbol{\gamma}})e^\tau} - a\frac{e^\tau/n}{1+e^\tau/n}.$$

Setting $h'(\tau) = 0$ gives a cubic equation of e^τ as below

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

The second derivative $h^{''}(\tau)$ is given by

$$2\frac{d^2h(\tau)}{d\tau^2} = -a\frac{ne^\tau}{(1+ne^\tau)^2} - (n-1)\frac{(1-R^2_{\boldsymbol{\gamma}})e^\tau}{(1+(1-R^2_{\boldsymbol{\gamma}})e^\tau)^2} + (n-p_{\boldsymbol{\gamma}}-1)\frac{e^\tau}{(1+e^\tau)^2}.$$

B Proof for Theorem 3

We first cite some preliminary results from Fernández *et al.* (2001) without proof. Under the sampling model $\mathcal{M}_{\boldsymbol{\gamma}}$: (i) if $\mathcal{M}_{\boldsymbol{\gamma}}$ is nested within or equal to a model $\mathcal{M}_{\boldsymbol{\gamma}'}$, then

$$\text{plim}_{n\to\infty}\ \frac{\text{RSS}_{\boldsymbol{\gamma}'}}{n} = 1/\phi \tag{40}$$

and (ii) for any model $\mathcal{M}_{\boldsymbol{\gamma}'}$ that does not contain $\mathcal{M}_{\boldsymbol{\gamma}}$, under assumption (30),

$$\text{plim}_{n\to\infty}\frac{\text{RSS}_{\boldsymbol{\gamma}'}}{n} = 1/\phi + b_{\boldsymbol{\gamma}'} \tag{41}$$

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

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

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

$$\hat{g}^{\text{EBL}}_{\boldsymbol{\gamma}'} = \left[\frac{R^2_{\boldsymbol{\gamma}'}/p_{\boldsymbol{\gamma}'}}{(1-R^2_{\boldsymbol{\gamma}'})/(n-1-p_{\boldsymbol{\gamma}'})} - 1\right](1+o_P(1)), \tag{42}$$

and

$$\mathrm{BF}_{\mathrm{EBL}}[\mathcal{M}_{\boldsymbol{\gamma}'} : \mathcal{M}_N] \overset{P}{\sim} \frac{1}{(1-R^2_{\boldsymbol{\gamma}'})^{\frac{n-1-p_{\boldsymbol{\gamma}'}}{2}}} \frac{(n-1-p_{\boldsymbol{\gamma}'})^{\frac{n-1-p_{\boldsymbol{\gamma}'}}{2}}}{(n-1)^{\frac{n-1}{2}}}, \tag{43}$$

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

$$\mathrm{BF}_{\mathrm{EBL}}[\mathcal{M}_{\boldsymbol{\gamma}'} : \mathcal{M}_{\boldsymbol{\gamma}}] \overset{P}{\sim} \frac{1}{n^{(p_{\boldsymbol{\gamma}'}-p_{\boldsymbol{\gamma}})/2}} \left(\frac{\mathrm{RSS}_{\boldsymbol{\gamma}}/n}{\mathrm{RSS}_{\boldsymbol{\gamma}'}/n}\right)^{n/2}. \tag{44}$$

Consider the following three situations:

(a) $\mathcal{M}_{\boldsymbol{\gamma}} \cap \mathcal{M}_{\boldsymbol{\gamma}'} \neq \emptyset$ and $\mathcal{M}_{\boldsymbol{\gamma}} \nsubseteq \mathcal{M}_{\boldsymbol{\gamma}'}$: Applying (40) and (30), we have

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

which converges to zero (in probability) exponentially fast with respect to n since $b_{\boldsymbol{\gamma}'}$ is a positive constant. Therefore, no matter what value $p_{\boldsymbol{\gamma}'} - p_{\boldsymbol{\gamma}}$ takes, the Bayes factor (44) goes to zero (in probability);

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

(c) $\mathcal{M}_{\boldsymbol{\gamma}} \cap \mathcal{M}_{\boldsymbol{\gamma}'} = \emptyset$. In this case, we have $nR^2_{\boldsymbol{\gamma}'}$ converges in distribution to $\chi^2_{p_{\boldsymbol{\gamma}'}}/(1+\phi b_{\boldsymbol{\gamma}})$ where $b_{\boldsymbol{\gamma}}$ is defined in (41). Since

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

we have $\mathrm{BF}_{\mathrm{EBL}}[\mathcal{M}_{\boldsymbol{\gamma}'} : \mathcal{M}_N] = O_P(1)$. On the other hand, since $R^2_{\boldsymbol{\gamma}}$ goes to a constant strictly between zero and one, by (43) we have

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

where the second term goes to infinity exponentially fast. So the Bayes factor goes to zero in probability.

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

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

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

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

$$\text{BF}_\pi[\mathcal{M}_{\boldsymbol{\gamma}'} : \mathcal{M}_N] = \int L(g)\pi(g)dg = \int \left(1 - R^2_{\boldsymbol{\gamma}'}\frac{g}{1+g}\right)^{-\frac{n-1}{2}} \frac{\pi(g)}{(1+g)^{p_{\boldsymbol{\gamma}'}/2}}\, dg. \tag{45}$$

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 (36). The relative error is still $O(1/n)$ (Kass and Raftery, 1995). As such, we can write (45) as

$$\text{BF}_\pi[\mathcal{M}_{\boldsymbol{\gamma}'}, \mathcal{M}_{\boldsymbol{\gamma}}] = \frac{\pi(\hat{g}^{\text{EBL}}_{\boldsymbol{\gamma}'})}{\pi(\hat{g}^{\text{EBL}}_{\boldsymbol{\gamma}})}\frac{\tilde{\sigma}_{\boldsymbol{\gamma}'}}{\tilde{\sigma}_{\boldsymbol{\gamma}}}\, \text{BF}_{\text{EBL}}[\mathcal{M}_{\boldsymbol{\gamma}'}, \mathcal{M}_{\boldsymbol{\gamma}}](1 + O(1/n)), \tag{46}$$

where

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

is similar to (37). When $\mathcal{M}_{\boldsymbol{\gamma}'} \cap \mathcal{M}_{\boldsymbol{\gamma}} \neq \emptyset$, $R^2_{\boldsymbol{\gamma}'}$ converges in probability to a constant strictly between zero and one, so we have the first two terms are bounded in probability since $\tilde{\sigma}_{\boldsymbol{\gamma}'} = O_P(n)$, $\pi(\hat{g}^{\text{EBL}}_{\boldsymbol{\gamma}'}) = O_P(n^{-3/2})$ for the Zellner-Siow, $\pi(\hat{g}^{\text{EBL}}_{\boldsymbol{\gamma}'}) = O_P(n^{-a/2})$ for the hyper-g prior, and $\pi(\hat{g}^{\text{EBL}}_{\boldsymbol{\gamma}'}) = O_P(1)$ 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 $\mathcal{M}_{\boldsymbol{\gamma}'} \cap \mathcal{M}_{\boldsymbol{\gamma}} = \emptyset$, following the same reasoning used for local EB estimate in this case, we have that $nR^2_{\boldsymbol{\gamma}'}$ converges in distribution to $\chi^2_{p_{\boldsymbol{\gamma}'}}/(1 + \phi b_{\boldsymbol{\gamma}})$. Then, when n is large, we have

$$\text{BF}[\mathcal{M}_{\boldsymbol{\gamma}'} : \mathcal{M}_N] \leq \exp(C\chi^2_{p_{\boldsymbol{\gamma}'}})(1+\epsilon)\int \frac{\pi(g)}{(1+g)^{p_{\boldsymbol{\gamma}'}/2}}\, dg \leq 2\exp(C\chi^2_{p_{\boldsymbol{\gamma}'}}) \tag{47}$$

where C is some constant. Therefore, it does not go to infinity. On the other hand, we have that $\text{BF}_\pi[\mathcal{M}_{\boldsymbol{\gamma}} : \mathcal{M}_N]$ goes to infinity using an approximation as in (46). Therefore $\text{BF}[\mathcal{M}_{\boldsymbol{\gamma}'} : \mathcal{M}_{\boldsymbol{\gamma}}] \to 0$.

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

Both the local and global empirical Bayes approaches are not consistent in this situation. For any non-null model $\mathcal{M}_{\boldsymbol{\gamma}'}$, we have that $R^2_{\boldsymbol{\gamma}'}$ goes to zero and $\hat{g}_{\boldsymbol{\gamma}'} = \max(F_{p_{\boldsymbol{\gamma}'}, n-1-p_{\boldsymbol{\gamma}'}} - 1, 0)$ where F

denotes and a F-distribution with degrees of freedom $p_{\boldsymbol{\gamma}'}$ and $n-1-p_{\boldsymbol{\gamma}'}$ respectively. This F-distributed random variable converges in distribution to $\chi^2_{p_{\gamma'}}/p_{\boldsymbol{\gamma}'}$ and hence $\hat{g}_{\boldsymbol{\gamma}'} = O_P(1)$. Therefore, since $\text{BF}[\mathcal{M}_{\boldsymbol{\gamma}'} : \mathcal{M}_N] \geq (1+\hat{g}_{\boldsymbol{\gamma}'})^{-p_\gamma/2}$, $\text{BF}[\mathcal{M}_{\boldsymbol{\gamma}'} : \mathcal{M}_N]$ can not go to zero. The argument for the global Empirical Bayes approach is similar.

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

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

no proper prior which does not depend on n can lead to consistency under the null. However, the first inequality in (47) shows that if the integral above vanishes as n goes to infinity, 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 lead to consistency under the null.

C Ozone Data

Variables used in the ozone pollution example:

- ozone – Daily ozone concentration (maximum one 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 (percent) at LAX.
- temp – Sandburg Air Force Base temperature (F^o).
- ibh – Inversion base height at LAX.
- ibt – Inversion base temperature at LAX.
- dpg – Daggett Pressure gradient (mm Hg) from LAX to Daggett, CA.
- vis – Visibility (miles) at LAX.