

BAYES AND EMPIRICAL-BAYES MULTIPLICITY ADJUSTMENT IN THE VARIABLE-SELECTION PROBLEM

BY JAMES G. SCOTT * AND JAMES O. BERGER

Department of Statistical Science, Duke University

This paper studies the multiplicity-correction effect of standard Bayesian variable-selection priors in linear regression. The first goal of the paper is to clarify when, and how, multiplicity correction is automatic in Bayesian analysis, and contrast this multiplicity correction with the Bayesian Ockham's-razor effect. Secondly, we contrast empirical-Bayes and fully Bayesian approaches to variable selection, through examples, theoretical results, and simulations. Considerable differences between the results of the two approaches are found, which suggest that considerable care be taken with the empirical-Bayes approach in variable selection.

1. Introduction. Consider the usual problem of variable selection in linear regression. Given a vector \mathbf{Y} of n responses and an $n \times m$ design matrix \mathbf{X} , the goal is to select k predictors out of m possible ones for fitting a model of the form:

$$(1) \quad Y_i = \alpha + X_{ij_1}\beta_{j_1} + \dots + X_{ij_k}\beta_{j_k} + \epsilon_i$$

for some $\{j_1, \dots, j_k\} \subset \{1, \dots, m\}$, where $\epsilon_i \stackrel{iid}{\sim} N(0, \phi^{-1})$, the precision ϕ being unknown.

Since the number of possible predictors in many scientific problems today is huge, there has been a marked increase in the attention being paid to the multiple-testing problem that arises in deciding whether or not each variable should be in the model. Multiplicity issues are particularly relevant when researchers have little reason to suspect one model over another, and simply want the data to flag interesting covariates from a large pool. In such cases, variable selection is treated less as a formal inferential framework and more as an exploratory tool used to generate insights about complex, high-dimensional systems. Still, the results of such studies are often used

*This research was supported by the U.S. National Science Foundation under a Graduate Research Fellowship.

AMS 2000 subject classifications: Primary 62J05, 62J15

Keywords and phrases: Multiple testing, Bayesian model selection, empirical Bayes

to buttress scientific conclusions or guide policy decisions—conclusions or decisions that may be quite wrong if the multiple-testing problem is ignored.

The focus in this paper is on Bayesian and empirical-Bayesian approaches to multiplicity correction in variable selection. These approaches have the attractive feature that they can automatically adjust for multiple testing without needing to introduce ad-hoc penalties. This is done by assuming that regression coefficients come from some common, unknown mixture distribution. Some useful references on this idea include Waller and Duncan (1969), Meng and Dempster (1987), Berry (1988), Westfall et al. (1997), Berry and Hochberg (1999), and Scott and Berger (2006). Similar ideas in a classical context are discussed by Johnstone and Silverman (2004) and Abramovich et al. (2006).

This paper has two main objectives. The first is to clarify how multiplicity correction enters the Bayesian analysis: through choice of the prior model probabilities. The discussion highlights the fact that not all Bayesian analyses automatically adjust for multiplicity; it also clarifies the difference between multiplicity correction and the Bayesian Ockham's-razor effect (see Jefferys and Berger, 1992), which induces a quite different type of penalty on more complex models. The striking differences between Bayesian analyses with and without multiplicity correction will be illustrated on a variety of examples.

While investigating this original objective, a series of surprises involving empirical-Bayes variable selection was encountered, and this became a second focus of the paper. The main surprise was a marked dissimilarity between fully Bayesian answers and empirical-Bayes answers—a dissimilarity arising from a different source than the failure to account for uncertainty in the empirical-Bayes estimate (which is the usual issue in such problems). Indeed, even at the extreme, when the empirical-Bayes estimate converges asymptotically to the true parameter value, the potential for a serious discrepancy remains.

The existence of a such a discrepancy between fully Bayesian answers and empirical-Bayes answers is of immediate interest to Bayesians, who often use empirical Bayes as a computational simplification. The discrepancy is also of interest to non-Bayesians for several reasons. First, frequentist complete-class theorems suggest that, if an empirical-Bayes analysis does not approximate some fully Bayesian analysis, then it may be suboptimal and needs alternative justification. Such justifications can be found for a variety of situations in George and Foster (2000), Efron et al. (2001), Johnstone and Silverman (2004), Cui and George (2008), and Bogdan et al. (2008).

Second, theoretical and numerical investigations of the discrepancy re-

vealed some practically disturbing properties of standard empirical-Bayes analysis in variable selection, including the following:

- It seems to have a bias toward extreme answers that can produce too many false positives (or false negatives).
- It frequently collapses to a degenerate solution, resulting in an inappropriate statement of certainty in the selected regression model.

As a simple example of the second point, suppose the usual variable-selection prior is used, where each variable is presumed to be in the model, independent of the others, with an unknown common probability p . A common empirical-Bayes method is to estimate p by marginal maximum likelihood (or Type-II maximum likelihood, as it is commonly called; see Section 3.2), and use this estimated \hat{p} to determine the prior probabilities of models. This procedure will be shown to have the startlingly inappropriate property of assigning final probability 1 to either the full model or the intercept-only (null) model whenever the full (or null) model has the largest marginal likelihood, even if this marginal likelihood is only slightly larger than that of the next-best model.

This is certainly not the first situation in which the Type-II MLE approach to empirical Bayes has been shown to have problems. But the extent of the problem in variable selection, and its unusual character, seem not to have been recognized. Of course, there are alternatives to Type-II MLE in estimation of p (as shown in some of the above papers), and the results in this paper suggest that such alternatives be seriously considered.

Section 2 introduces notation. Section 3 gives a brief historical and methodological overview of multiplicity correction for Bayesian variable selection, and focuses on the issue of clarifying the source and nature of the correction. Section 4 introduces a theoretical framework for characterizing the differences between fully Bayesian and empirical-Bayes analyses, and gives several examples and theoretical results concerning the differences. Section 5 presents numerical results indicating the practical nature of the differences, through a simulation experiment and a practical example. Section 6 gives further discussion of the results.

2. Preliminaries.

2.1. *Notation.* All models are assumed to include an intercept term α . Let M_0 denote the null model with only this intercept term, and let M_F denote the full model with all covariates under consideration. The full model thus has parameter vector $\boldsymbol{\theta}' = (\alpha, \boldsymbol{\beta}')$, $\boldsymbol{\beta}' = (\beta_1, \dots, \beta_m)'$. Submodels M_γ

are indexed by a binary vector $\boldsymbol{\gamma}$ of length m indicating a set of $k_\gamma \leq m$ nonzero regression coefficients $\boldsymbol{\beta}_\gamma$:

$$\gamma_i = \begin{cases} 0 & \text{if } \beta_i = 0 \\ 1 & \text{if } \beta_i \neq 0. \end{cases}$$

It is most convenient to represent model uncertainty as uncertainty in $\boldsymbol{\gamma}$, a random variable that takes values in the discrete space $\{0, 1\}^m$, which has 2^m members. Inference relies upon the prior probability of each model, $p(M_\gamma)$, along with the marginal likelihood of the data under each model:

$$(2) \quad f(\mathbf{Y} \mid M_\gamma) = \int f(\mathbf{Y} \mid \boldsymbol{\theta}_\gamma, \phi) \pi(\boldsymbol{\theta}_\gamma, \phi) d\boldsymbol{\theta}_\gamma d\phi,$$

where $\pi(\boldsymbol{\theta}_\gamma, \phi)$ is the prior for model-specific parameters. These together define, up to a constant, the posterior probability of a model:

$$(3) \quad p(M_\gamma \mid \mathbf{Y}) \propto p(M_\gamma) f(\mathbf{Y} \mid M_\gamma).$$

Let \mathbf{X}_γ denote the columns of the full design matrix \mathbf{X} given by the nonzero elements of $\boldsymbol{\gamma}$, and let \mathbf{X}_γ^* denote the concatenation $(\mathbf{1} \ \mathbf{X}_\gamma)$, where $\mathbf{1}$ is a column of ones corresponding to the intercept α . For simplicity, we will assume that all covariates have been centered so that $\mathbf{1}$ and \mathbf{X}_γ are orthogonal. We will also assume that the common choice $\pi(\alpha) = 1$ is made for the parameter α in each model (see Berger et al., 1998, for a justification of this choice of prior).

Often all models will have small posterior probability, in which case more useful summaries of the posterior distribution are quantities such as the posterior inclusion probabilities of the individual variables:

$$(4) \quad p_i = \Pr(\gamma_i \neq 0 \mid \mathbf{Y}) = \sum_{\boldsymbol{\gamma}} 1_{\gamma_i=1} \cdot p(M_\gamma \mid \mathbf{Y}).$$

These quantities also define the median-probability model, which is the model that includes those covariates having posterior inclusion probability at least $1/2$. Under many circumstances, this model has greater predictive power than the most probable model (Barbieri and Berger, 2004).

2.2. Priors for Model-Specific Parameters. There is an extensive body of literature confronting the difficulties of Bayesian model choice in the face of weak prior information. These difficulties arise due to the obvious dependence of the marginal likelihoods in (2) upon the choice of priors for

model-specific parameters. In general one cannot use improper priors on these parameters, since this leaves the resulting Bayes factors defined only up to an arbitrary multiplicative constant.

This paper chiefly uses null-based Zellner-Siow priors (Zellner and Siow, 1980) for computing the marginal likelihoods in (2), which are heavy-tailed versions of Zellner’s canonical g -prior (Zellner, 1986); explicit expressions can be found in Appendix A. The chief rationale for using these priors has to do with the notion of information consistency. See Berger and Pericchi (2001) and Liang et al. (2008) for overviews of information consistency, along with the appendix for an example involving g -priors and Jeffreys (1961) for a discussion of the issue in the general normal-means testing context.

3. Approaches to multiple testing.

3.1. *Bayes Factors, Ockham’s Razor, and Multiplicity.* In both Bayes and empirical-Bayes variable selection, the marginal likelihood contains a built-in penalty for model complexity that is often called the Bayesian “Ockham’s-razor effect” (Jefferys and Berger, 1992). This penalty arises in integrating the likelihood across a higher-dimensional parameter space under the more complex model. For example, when inspecting the null-based g -prior Bayes factors in (33), one can immediately see the Ockham’s-razor penalty in the term involving k_γ , the number of parameters. A complex model will always yield a larger value of R^2 than a smaller model, but will have a smaller marginal likelihood unless this increase outweighs the penalty paid for larger values of k_γ .

While this is a penalty against more complex models, it is not a multiple-testing penalty *per se*; the Bayes factor between two fixed models will not change as more possible variables are thrown into the mix, and hence will not exert control over the number of false positives as m grows large.

Instead, multiplicity must be handled through the choice of prior probabilities of models. The earliest recognition of this idea seems to be that of Jeffreys in 1939, who gave a variety of suggestions for apportioning probability across different kinds of model spaces (see Sections 1.6, 5.0, and 6.0 of Jeffreys (1961), a later edition). Jeffreys paid close attention to multiplicity adjustment, which he called “correcting for selection.” In scenarios involving an infinite sequence of nested models, for example, he recommended using model probabilities that formed a convergent geometric series, so that the prior odds ratio for each pair of neighboring models (that is, those differing by a single parameter) was a fixed constant. Another suggestion, appropriate for more general contexts, was to give all models of size k a single lump of probability to be apportioned equally among models of that size. Below,

in fact, the fully Bayesian solution to multiplicity correction will be shown to have exactly this flavor.

It is interesting that, in the variable-selection problem, assigning all models equal prior probability (which is equivalent to assigning each variable prior probability of $1/2$ of being in the model) provides no multiplicity control. This is most obvious in the orthogonal situation, which can be viewed as m independent tests of $H_i : \beta_i = 0$. If each of these tests has prior probability of $1/2$, there will be no multiplicity control as m grows. Indeed, note that this “pseudo-objective” prior reflects an *a-priori* expected model size of $m/2$ with a standard deviation of $\sqrt{m}/2$, meaning that the prior for the fraction of included covariates becomes very tight around $1/2$ as m grows.

3.2. Variable-Selection Priors and Empirical Bayes. The standard modern practice in Bayesian variable-selection problems is to treat variable inclusions as exchangeable Bernoulli trials with common success probability p , which implies that the prior probability of a model is given by

$$(5) \quad p(M_\gamma | p) = p^{k_\gamma} (1 - p)^{m - k_\gamma},$$

with k_γ representing the number of included variables in the model.

We saw above that selecting $p = 1/2$ does not provide multiplicity correction. Treating p as an unknown parameter to be estimated from the data will, however, yield an automatic multiple-testing penalty. The intuition is that, as m grows with the true k remaining fixed, the posterior distribution of p will concentrate near 0, so that the situation is the same as if one had started with a very low prior probability that a variable should be in the model (Scott and Berger, 2006). Note that one could adjust for multiplicity subjectively, by specifying p to reflect subjective belief in the proportion of variables that should be included. No fixed choice of p that is independent of m , however, can adjust for multiplicity.

The empirical-Bayes approach to variable selection was popularized by George and Foster (2000), and is a common strategy for treating the prior inclusion probability p in (5) in a data-dependent way. The most common approach is to estimate the prior inclusion probability by maximum likelihood, maximizing the marginal likelihood of p summed over model space (often called Type-II maximum likelihood):

$$(6) \quad \hat{p} = \arg \max_{p \in [0,1]} \sum_{\gamma} p(M_\gamma | p) \cdot f(\mathbf{Y} | M_\gamma).$$

One uses this in (5) to define the *ex-post* prior probabilities $p(M_\gamma | \hat{p}) =$

$\hat{p}^{k_\gamma}(1 - \hat{p})^{m - k_\gamma}$, resulting in final model posterior probabilities

$$(7) \quad p(M_\gamma | \mathbf{Y}) \propto \hat{p}^{k_\gamma} \cdot (1 - \hat{p})^{m - k_\gamma} f(\mathbf{Y} | M_\gamma).$$

The EB solution \hat{p} can be found either by direct numerical optimization or by the EM algorithm detailed in Liang et al. (2008). For an overview of empirical-Bayes methodology, see Carlin and Louis (2000).

It is clear that the empirical-Bayes approach will control for multiplicity in the same way as the fully Bayesian approach: if there are only k true variables and m grows large, then $\hat{p} \rightarrow 0$.

3.3. A Fully Bayesian Version. Fully Bayesian variable-selection priors have been discussed by Ley and Steel (2007), Cui and George (2008), and Carvalho and Scott (2007), among others. These priors assume that p has a Beta distribution, $p \sim \text{Be}(a, b)$, giving:

$$(8) \quad p(M_\gamma) = \int_0^1 p(M_\gamma | p) \pi(p) dp \propto \frac{\beta(a + k_\gamma, b + m - k_\gamma)}{\beta(a, b)},$$

where $\beta(\cdot, \cdot)$ is the beta function. For the default choice of $a = b = 1$, implying a uniform prior on p , this reduces to:

$$(9) \quad p(M_\gamma) = \frac{(k_\gamma)!(m - k_\gamma)!}{(m + 1)(m!)} = \frac{1}{m + 1} \binom{m}{k_\gamma}^{-1}.$$

Utilizing this in (3) would yield posterior model probabilities

$$(10) \quad p(M_\gamma | \mathbf{Y}) \propto \frac{1}{m + 1} \binom{m}{k_\gamma}^{-1} f(\mathbf{Y} | M_\gamma).$$

This has the air of paradox: in contrast to (7), where the multiplicity adjustment is apparent, here p has been marginalized away. How can p then be adjusted by the data so as to induce a multiplicity-correction effect?

Figures 1 and 2 hint at the answer, which is that the multiplicity penalty was always in the prior probabilities in (9) to begin with; it was just hidden. In Figure 1 the prior log-probability is plotted as a function of model size for a particular value of m (in this case 30). This highlights the marginal penalty that one must pay for adding an extra variable: in moving from the null model to a model with one variable, the fully Bayesian prior favors the simpler model by a factor of 30 (label A). This penalty is not uniform:

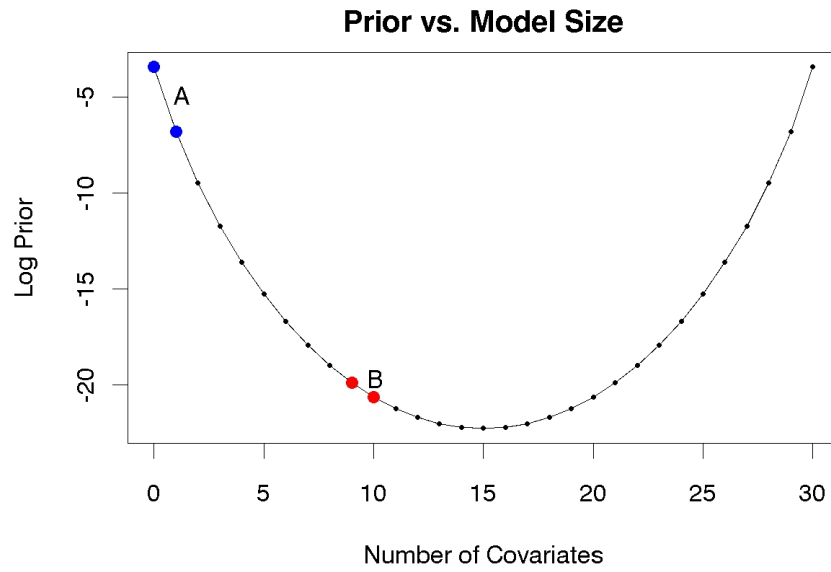


FIG 1. *Prior probability versus model size.*

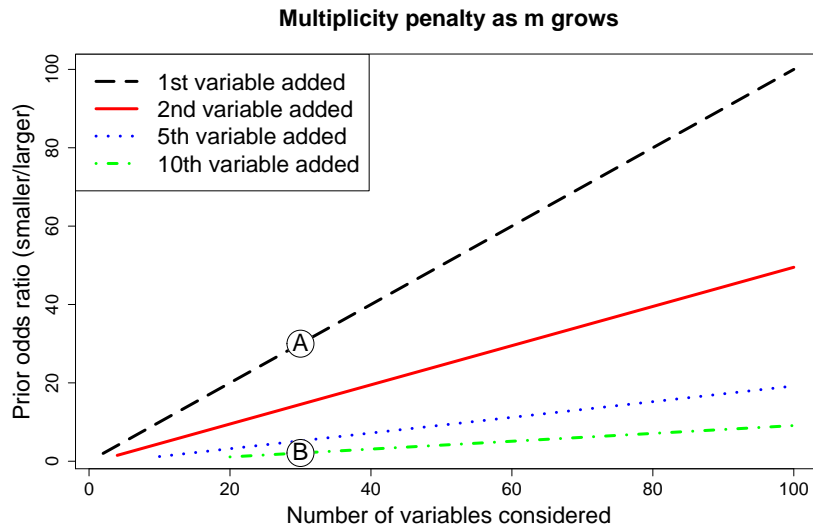


FIG 2. *Multiplicity penalties as m grows.*

models of size 9, for example, are favored to those of size 10 by a factor of only 2.1 (label B).

Figure 2 then shows these penalties getting steeper as one considers more models. Adding the first variable incurs a 30-to-1 prior-odds penalty if one tests 30 variables (label A as before), but a 60-to-1 penalty if one tests 60 variables. Similarly, the 10th-variable marginal penalty is about two-to-one for 30 variables considered (label B), but would be about four-to-one for 60 variables.

We were careful above to distinguish this effect from the Ockham’s-razor penalty coming from the marginal likelihoods. But marginal likelihoods are clearly relevant. They determine where models will sit along the curve in Figure 1, and thus will determine whether the prior-odds multiplicity penalty for adding another variable to a good model will be more like 2, more like 30, or something else entirely. Indeed, note that, if only large models have significant marginal likelihoods, then the ‘multiplicity penalty’ will now become a ‘multiplicity advantage’ as one is on the increasing part of the curve in Figure 1. (This is also consistent with the empirical-Bayes answer: if $\hat{p} > 0.5$, then the analysis will increase the chance of variables entering the model.)

Interestingly, the uniform prior on p also gives every variable a marginal prior inclusion probability of $1/2$; these marginal probabilities are the same as those induced by the “psuedo-objective” choice of $p = 1/2$. Yet because probability is apportioned among models in a very different way, profoundly different behaviors emerge.

Table 1 compares these two regimes on a simulated data set for which the true value of k was fixed at 10. This study used a simulated $n = 75$ by $m = 100$ design matrix of $N(0, 1)$ covariates and 10 regression coefficients that differed from zero, along with 90 coefficients that were identically zero. The table summarizes the inclusion probabilities of the 10 real variables as we test them along with an increasing number of noise variables (first 1, then 10, 40, and 90). It also indicates how many false positives (defined as having inclusion probability ≥ 0.5) are found among the noise variables. Here, “uncorrected” refers to giving all models equal prior probability by setting $p = 1/2$. “Oracle Bayes” is the result from choosing p to reflect the known fraction of nonzero covariates.

The following points can be observed:

- The fully Bayes procedure exhibits a clear multiplicity adjustment; as the number of noise variables increases, the posterior inclusion probabilities of variables decrease. The uncorrected Bayesian analysis shows no such adjustment and can, rather bizarrely, sometimes have the inclusion probabilities increase as noise variables are added.

Signal	Number of noise variables											
	Uncorrected				Fully Bayes				Oracle Bayes			
	1	10	40	90	1	10	40	90	1	10	40	90
$\beta_1 : -1.08$.99	.99	.99	.99	.99	.99	.99	.99	.99	.99	.99	.99
$\beta_2 : -0.84$.99	.99	.99	.99	.99	.99	.99	.98	.99	.99	.99	.99
$\beta_3 : -0.74$.99	.99	.99	.99	.99	.99	.99	.99	.99	.99	.99	.99
$\beta_4 : -0.51$.97	.97	.99	.99	.91	.94	.71	.34	.99	.97	.85	.52
$\beta_5 : -0.30$.29	.28	.28	.12	.55	.24	.04	.00	.79	.28	.06	.01
$\beta_6 : +0.07$.26	.28	.05	.01	.51	.25	.03	.01	.78	.28	.05	.01
$\beta_7 : +0.18$.21	.24	.24	.27	.45	.21	.03	.01	.70	.24	.04	.01
$\beta_8 : +0.35$.77	.77	.99	.99	.89	.68	.30	.05	.97	.77	.45	.11
$\beta_9 : +0.41$.92	.91	.99	.99	.96	.86	.56	.22	.99	.91	.72	.35
$\beta_{10} : +0.63$.99	.99	.99	.99	.99	.99	.92	.73	.99	.99	.97	.87
FPs	0	2	5	10	0	1	0	0	0	2	1	0

TABLE 1

Posterior inclusion probabilities for the 10 real variables in the simulated data set, along with the number of false positives (posterior inclusion probability greater than $1/2$) among the “pure noise” columns in the design matrix. Marginal likelihoods were calculated using null-based Zellner-Siow priors by enumerating the model space in the $p = 11$ and $p = 20$ cases, and by 5 million iterations of the feature-inclusion stochastic-search algorithm (Berger and Molina, 2005; Scott and Carvalho, 2008) in the $p = 50$ and $p = 100$ cases.

- On the simulated data, proper multiplicity adjustment yields reasonably strong control over false positives, in the sense that the number of false positives appears bounded (and small) as m increases. In contrast, the number of false positives appears to be increasing linearly for the uncorrected Bayesian analysis, as would be expected.
- The full Bayes and oracle Bayes answers are qualitatively very similar; indeed, if one adopted the (median probability model) prescription of selecting those variables with posterior inclusion probability greater than $1/2$, they would both always select the same variables, except in two instances.

Table 2 shows the inclusion probabilities for a model of ozone concentration levels outside Los Angeles that includes 10 atmospheric variables along with all squared terms and second-order interactions ($m = 65$). Probabilities are given for uncorrected ($p = 1/2$) and fully Bayesian analyses under a variety of different marginal likelihood computations. All variables appear uniformly less impressive when adjusted for multiplicity. This happens regardless of how one computes marginal likelihoods, indicating that, indeed, the multiplicity penalty is logically distinct from the prior on regression coefficients and instead results from the prior distribution across model space.

Other examples of such multiplicity correction put into practice can be found throughout the literature. For nonparametric problems, see Gopalan

	All models equal				Fully Bayesian, $p \sim U(0, 1)$			
	GN	GF	ZSN	PBIC	GN	GF	ZSN	PBIC
x1	.860	.892	.943	.750	.419	.450	.478	.297
x2	.052	.051	.060	.045	.018	.021	.023	.011
x3	.030	.029	.033	.022	.011	.013	.014	.008
x4	.985	.987	.995	.954	.767	.791	.817	.667
x5	.195	.219	.306	.163	.040	.049	.051	.026
x6	.186	.226	.353	.152	.033	.038	.030	.036
x7	.200	.202	.215	.248	.301	.288	.273	.381
x8	.960	.962	.977	.929	.739	.747	.758	.755
x9	.029	.035	.054	.029	.014	.018	.016	.010
x10	.999	.999	.999	.998	.986	.986	.998	.974
x1-x1	.999	.999	.999	.999	.986	.991	.995	.977
x9-x9	.999	.999	.999	.998	.872	.894	.918	.782
x1-x2	.577	.607	.732	.498	.153	.176	.196	.119
x4-7	.330	.353	.459	.236	.086	.101	.108	.057
x6-x8	.776	.785	.859	.671	.258	.285	.314	.205
x7-x8	.266	.288	.296	.274	.103	.119	.113	.082
x7-x10	.975	.952	.952	.929	.935	.927	.957	.933

TABLE 2

Posterior inclusion probabilities for the important main effects, quadratic effects, and cross-product effects for ozone-concentration data under various marginal likelihoods, with and without full Bayesian multiplicity correction. Key: GN = null-based g -priors, GF = full-based g -priors, ZSN = null-based Zellner-Siow priors.

and Berry (1998); for gene-expression studies, see Do et al. (2005); for econometrics, see Ley and Steel (2007); for Gaussian graphical models, see Carvalho and Scott (2007); and for time-series data, see Scott (2008).

4. Theoretical Investigation of Empirical-Bayes Variable Selection.

4.1. *Motivation.* Here is a surprising lemma that indicates the need for caution with empirical Bayes methods in variable selection. The lemma refers to the variable-selection problem, with the prior variable inclusion probability p being estimated by marginal (or Type-II) maximum likelihood in the empirical-Bayes approach.

LEMMA 4.1. *In the variable-selection problem, if M_0 has the (strictly) largest marginal likelihood, then the Type-II MLE estimate of p is $\hat{p} = 0$. Similarly, if M_F has the (strictly) largest marginal likelihood, then $\hat{p} = 1$.*

PROOF. Since $p(M_\gamma)$ sums to 1 over γ , the marginal likelihood of p sat-

satisfies

$$(11) \quad f(\mathbf{Y}) = \sum_{\Gamma} f(\mathbf{Y} | M_{\gamma}) p(M_{\gamma}) \leq \max_{\gamma \in \Gamma} f(\mathbf{Y} | M_{\gamma}).$$

Furthermore, the inequality is strict under the conditions of the lemma (because the designated marginals are strictly largest), unless the prior assigns $p(M_{\gamma}) = 1$ to the maximizing marginal likelihood. The only way that $p(M_{\gamma}) = p^{k_{\gamma}} \cdot (1-p)^{m-k_{\gamma}}$ can equal 1 is for p to be 0 or 1 and for the model to be M_0 or M_F , respectively. At these values of p , equality is indeed achieved in (11) under the stated conditions, and the results follow. \square

As a consequence, the empirical Bayes approach here would assign final probability 1 to M_0 whenever it has the largest marginal likelihood, and final probability 1 to M_F whenever it has the largest marginal likelihood. These are clearly very unsatisfactory answers.

4.2. *Comparison of Empirical Bayes and Fully Bayesian Analysis.* Note that the motivating lemma above referred to a clearly undesirable property of empirical-Bayes analysis in variable selection. A number of the following results have the same character, but we will, for the most part, focus on a comparison between empirical-Bayes and fully Bayesian analysis. Comparison of the two is certainly of interest, both to Bayesians who might consider empirical Bayes as a computational approximation, and to frequentists for the reasons mentioned in the introduction.

To explore the difference between empirical-Bayes and fully Bayesian analysis, it is useful to abstract the problem somewhat and suppose simply that the data \mathbf{Y} have sampling density $f(\mathbf{Y} | \boldsymbol{\theta})$, and let $\boldsymbol{\theta} \in \Theta$ have prior density $\pi(\boldsymbol{\theta} | \boldsymbol{\lambda})$ for some unknown hyperparameter $\boldsymbol{\lambda} \in \Lambda$. Empirical-Bayes methodology typically proceeds by estimating $\boldsymbol{\lambda}$ from the data using a consistent estimator. (The Type-II MLE approach would estimate λ by the maximizer of the marginal likelihood $m(\mathbf{Y} | \boldsymbol{\lambda}) = \int_{\Lambda} f(\mathbf{Y} | \boldsymbol{\theta})\pi(\boldsymbol{\theta} | \boldsymbol{\lambda}) d\boldsymbol{\theta}$, and this will typically be consistent in empirical Bayes settings.) It is then argued that (at least asymptotically) the Bayesian analysis with $\hat{\boldsymbol{\lambda}}$ will be equivalent to the Bayesian analysis if one knew $\boldsymbol{\lambda}$.

To contrast this with a full Bayesian analysis, suppose we have a prior density $\pi(\boldsymbol{\lambda})$ for $\boldsymbol{\lambda}$ and a target function $\psi(\boldsymbol{\theta}, \mathbf{Y} | \boldsymbol{\lambda})$. For instance, ψ could be the posterior mean of $\boldsymbol{\theta}$ given $\boldsymbol{\lambda}$ and \mathbf{Y} , or it could be the conditional posterior distribution of $\boldsymbol{\theta}$ given $\boldsymbol{\lambda}$ and \mathbf{Y} . The empirical-Bayesian claim, in this context, would be that

$$(12) \quad \int_{\Lambda} \psi(\boldsymbol{\theta}, \mathbf{Y} | \boldsymbol{\lambda})\pi(\boldsymbol{\lambda} | \mathbf{Y}) d\boldsymbol{\lambda} \approx \psi(\boldsymbol{\theta}, \mathbf{Y} | \hat{\boldsymbol{\lambda}}),$$

i.e. that the full Bayesian answer on the left can be well approximated by the empirical-Bayes answer on the right. The justification for (12) would be based on the fact that, typically, $\pi(\boldsymbol{\lambda} \mid \mathbf{Y})$ will be collapsing to a point mass near the true $\boldsymbol{\lambda}$ as the sample size increases, so that (12) will hold for appropriately smooth functions $\psi(\boldsymbol{\theta}, \mathbf{Y} \mid \boldsymbol{\lambda})$ when the sample size is large.

Note that there are typically better approximations to the left hand side of (12), such as the Laplace approximation. These, however, are focused on reproducing the full-Bayes analysis through an analytic approximation, and are not ‘empirical-Bayes’ per se. Likewise, higher order empirical-Bayes analysis will likely yield better results here, but the issue is in realizing when one needs to resort to such higher-order analysis in the first place, and in understanding why this is so for problems such as variable selection.

That (12) could fail for non-smooth $\psi(\boldsymbol{\theta}, \mathbf{Y} \mid \boldsymbol{\lambda})$ is no surprise. But what may come as a surprise is that this failure can occur for very common functions, such as the conditional posterior density itself. Indeed, in choosing $\psi(\boldsymbol{\theta}, \mathbf{Y} \mid \boldsymbol{\lambda}) = \pi(\boldsymbol{\theta} \mid \boldsymbol{\lambda}, \mathbf{Y})$, the left-hand side of (12) is just the posterior density of $\boldsymbol{\theta}$ given \mathbf{Y} , which (by definition) can be written as

$$(13) \quad \pi(\boldsymbol{\theta} \mid \mathbf{Y}) \propto f(\mathbf{Y} \mid \boldsymbol{\theta}) \int_{\Lambda} \pi(\boldsymbol{\theta} \mid \boldsymbol{\lambda}) \pi(\boldsymbol{\lambda}) \, d\boldsymbol{\lambda}.$$

On the other hand, for this choice of ψ , (12) becomes

$$(14) \quad \pi(\boldsymbol{\theta} \mid \mathbf{Y}) \approx \pi(\boldsymbol{\theta} \mid \mathbf{Y}, \hat{\boldsymbol{\lambda}}) \propto f(\mathbf{Y} \mid \boldsymbol{\theta}) \cdot \pi(\boldsymbol{\theta} \mid \hat{\boldsymbol{\lambda}}),$$

and the two expressions on the right-hand sides of (13) and (14) can be very different.

As an indication as to what goes wrong in (12) for this choice of ψ , note that

$$(15) \quad \begin{aligned} \pi(\boldsymbol{\theta} \mid \mathbf{Y}) &= \int_{\Lambda} \pi(\boldsymbol{\theta} \mid \boldsymbol{\lambda}, \mathbf{Y}) \cdot \pi(\boldsymbol{\lambda} \mid \mathbf{Y}) \, d\boldsymbol{\lambda} \\ &= \int_{\Lambda} \frac{\pi(\boldsymbol{\theta}, \boldsymbol{\lambda} \mid \mathbf{Y})}{\pi(\boldsymbol{\lambda} \mid \mathbf{Y})} \cdot \pi(\boldsymbol{\lambda} \mid \mathbf{Y}) \, d\boldsymbol{\lambda} \end{aligned}$$

$$(16) \quad = \int_{\Lambda} \frac{f(\mathbf{Y} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta} \mid \boldsymbol{\lambda}) \pi(\boldsymbol{\lambda})}{f(\mathbf{Y}) \pi(\boldsymbol{\lambda} \mid \mathbf{Y})} \cdot \pi(\boldsymbol{\lambda} \mid \mathbf{Y}) \, d\boldsymbol{\lambda},$$

which leads to (13) upon canceling $\pi(\boldsymbol{\lambda} \mid \mathbf{Y})$ in the numerator and denominator; it does not help that $\pi(\boldsymbol{\lambda} \mid \mathbf{Y})$ is collapsing to a point about the true $\boldsymbol{\lambda}$, because it occurs in both the numerator and the denominator of the integrand.

In the remainder of this section, the focus will be on comparison of (13) and (14) since, for model selection, the full posteriors are most relevant. The

“closeness” of the two distributions will be measured by Kullback-Leibler divergence, a standard measure for comparing a pair of distributions P and Q over parameter space Θ :

$$(17) \quad \text{KL}(P \parallel Q) = \int_{\Theta} P(\boldsymbol{\theta}) \log \left(\frac{P(\boldsymbol{\theta})}{Q(\boldsymbol{\theta})} \right) d\boldsymbol{\theta}.$$

The KL divergence lies on $[0, \infty)$, equals 0 if and only if its two arguments are equal, and satisfies the intuitive criterion that larger values signify greater disparity in information content. KL divergence can be used to formalize the notion of empirical-Bayes convergence to fully Bayesian analysis as follows:

KL Empirical-Bayes Convergence: Suppose the data \mathbf{Y} and parameter $\boldsymbol{\theta}$ have joint distribution $p(\mathbf{Y}, \boldsymbol{\theta} \mid \boldsymbol{\lambda})$, where $\boldsymbol{\theta} \in \Theta$ is of dimension m , and where $\boldsymbol{\lambda} \in \Lambda$ is of fixed dimension that does not grow with m . Let $\pi_E = \pi(\psi(\boldsymbol{\theta}) \mid \mathbf{Y}, \hat{\boldsymbol{\lambda}})$ be the empirical-Bayes posterior distribution for some function of the parameter $\psi(\boldsymbol{\theta})$, and let $\pi_F = \pi(\psi(\boldsymbol{\theta}) \mid \mathbf{Y}) = \int_{\Lambda} \pi(\psi(\boldsymbol{\theta}) \mid \mathbf{Y}, \boldsymbol{\lambda}) \cdot \pi(\boldsymbol{\lambda}) d\boldsymbol{\lambda}$ be the corresponding fully Bayesian posterior under the prior $\pi(\boldsymbol{\lambda})$. If, for every $\boldsymbol{\lambda} \in \Lambda$, $\text{KL}(\pi_F \parallel \pi_E) \rightarrow 0$ in probability under $p(\mathbf{Y}, \boldsymbol{\theta} \mid \boldsymbol{\lambda})$ as $m \rightarrow \infty$, then π_E will be said to be KL-convergent to the fully Bayesian posterior π_F .

Note that KL convergence is defined with respect to a particular function of the parameter, along with a particular prior distribution on the hyperparameter. If, for a given function $\psi(\boldsymbol{\theta})$, it is not possible to find a reasonable prior $\pi(\boldsymbol{\lambda})$ that leads to KL convergence, then estimating $\psi(\boldsymbol{\theta})$ by empirical Bayes is clearly suspect: a Bayesian could not replicate such a procedure even asymptotically, while a frequentist would be concerned by complete-class theorems. (A “reasonable” prior is a necessarily vague notion, but obviously excludes things such as placing a point mass at $\hat{\boldsymbol{\lambda}}$.)

Instead of KL divergence, of course, one might instead use another distance or divergence measure. The squared Hellinger distance is one such possibility:

$$H^2(P \parallel Q) = \frac{1}{2} \int_{\Theta} \left(\sqrt{P(\boldsymbol{\theta})} - \sqrt{Q(\boldsymbol{\theta})} \right)^2 d\boldsymbol{\theta}.$$

Most of the subsequent results, however, use KL divergence because of its familiarity and analytical tractability.

4.3. Posteriors for Normal Means. As a simple illustration of the above ideas, consider the following two examples of empirical-Bayes analysis, one that satisfies the convergence criterion and one that does not.

Imagine observing a series of independent random variables $y_i \sim N(\theta_i, 1)$, where each $\theta_i \sim N(\mu, 1)$. (Thus the hyperparameter $\lambda = \mu$ here.) The natural empirical-Bayes estimate of μ is the sample mean $\hat{\mu}_E = \bar{y}$. A standard hyperprior for a fully Bayesian analysis would be $\mu \sim N(0, A)$, for some specified A . (The objective hyperprior $\pi(\mu) = 1$ is essentially the limit of this as $A \rightarrow \infty$.) Let $\boldsymbol{\theta} = (\theta_1 \dots \theta_n)$ and $\mathbf{y} = (y_1 \dots y_n)$. Using the expressions given in, for example, Berger (1985), the empirical-Bayes and full Bayes posteriors are

$$(18) \quad \pi_E(\boldsymbol{\theta} \mid \mathbf{y}, \hat{\mu}_E) = N\left(\frac{1}{2}(\mathbf{y} + \bar{y}\mathbf{1}), \frac{1}{2}\mathbf{I}\right)$$

$$(19) \quad \pi_F(\boldsymbol{\theta} \mid \mathbf{y}) = N\left(\frac{1}{2}(\mathbf{y} + \bar{y}\mathbf{1}) - \left(\frac{1}{nA + 2}\right)\bar{y}\mathbf{1}, \frac{1}{2}\mathbf{I} + \frac{A}{2(nA + 2)}(\mathbf{1}\mathbf{1}^t)\right),$$

where \mathbf{I} is the identity matrix and $\mathbf{1}$ is a column vector of all ones.

Example 1: Suppose only the first normal mean, θ_1 , is of interest, meaning that $\psi(\boldsymbol{\theta}) = \theta_1$. Then sending $A \rightarrow \infty$ yields

$$(20) \quad \pi_E(\theta_1 \mid \mathbf{y}, \hat{\mu}_E) = N([y_1 + \bar{y}]/2, 1/2)$$

$$(21) \quad \pi_F(\theta_1 \mid \mathbf{y}) = N([y_1 + \bar{y}]/2, 1/2 + [2n]^{-1}).$$

It is easy to check that $\text{KL}(\pi_F \parallel \pi_E) \rightarrow 0$ as $n \rightarrow \infty$. Hence $\pi_E(\theta_1)$ arises from a KL-convergent EB procedure under a reasonable prior, since it corresponds asymptotically to the posterior given by the objective prior on the hyperparameter μ .

Example 2: Suppose now that $\boldsymbol{\theta}$, the entire vector of means, is of interest (hence $\psi(\boldsymbol{\theta}) = \boldsymbol{\theta}$). The relevant distributions are then the full π_E and π_F given in (18) and (19).

A straightforward computation shows that $\text{KL}(\pi_F \parallel \pi_E)$ is given by:

$$(22) \quad \text{KL} = \frac{1}{2} \left[\log\left(\frac{\det \Sigma_E}{\det \Sigma_F}\right) + \text{tr}(\Sigma_E^{-1}\Sigma_F) + (\hat{\boldsymbol{\theta}}_E - \hat{\boldsymbol{\theta}}_F)^t \Sigma_E^{-1} (\hat{\boldsymbol{\theta}}_E - \hat{\boldsymbol{\theta}}_F) - n \right]$$

$$(23) \quad = \frac{1}{2} \left[-\log\left(1 + \frac{nA}{nA + 2}\right) + \frac{nA}{nA + 2} + 2n \left(\frac{1}{nA + 2}\right)^2 \bar{y}^2 \right].$$

For any nonzero choice of A and for any finite value of the hyperparameter μ , it is clear that under $p(\mathbf{y}, \boldsymbol{\theta} \mid \mu)$ the quantity $[2n/(nA + 2)^2] \cdot \bar{y}^2 \rightarrow 0$ in probability as $n \rightarrow \infty$. Hence for any value of A (including $A = \infty$), the KL divergence in (23) converges to $(1 - \log 2)/2 > 0$ as n grows.

Of course, this only considers priors of the form $\mu \sim N(0, A)$, but the asymptotic normality of the posterior can be used to prove the result for essentially any prior that satisfies the usual regularity conditions, suggesting that there is no reasonable prior for which $\pi_E(\boldsymbol{\theta})$ is KL-convergent.

The crucial difference here is that, in the second example, the parameter of interest increases in dimension as information about the hyperparameter μ accumulates. This is not the usual situation in asymptotic analysis. Hence even as $\hat{\boldsymbol{\theta}}_F$ and $\tilde{\boldsymbol{\theta}}_F$ are getting closer to each other elementwise, the KL divergence does not shrink to 0 as expected. This is distressingly similar to EB inference in linear models, where one learns about the prior inclusion probability p only as the dimension of $\boldsymbol{\gamma}$, the parameter of interest, grows.

4.4. Results for Variable Selection. For the variable-selection problem, explicit expressions for the KL divergence between empirical-Bayes and fully Bayes procedures are not available. It is therefore not possible to give a general characterization of whether the empirical-Bayes variable-selection procedure is KL-convergent, in the sense defined above, to a fully Bayesian procedure. Two interesting sets of results are available, however: one regarding the KL divergence between the prior probability distributions of the fully Bayesian and empirical-Bayesian procedures, and the other regarding the *expected* posterior KL divergence.

Denote the empirical-Bayes prior distribution over model indicators by $p_E(\boldsymbol{\gamma})$ and the fully-Bayesian distribution (with uniform prior on p) by $p_F(\boldsymbol{\gamma})$. Similarly, after observing data D , write $p_E(\boldsymbol{\gamma} \mid \mathbf{Y})$ and $p_F(\boldsymbol{\gamma} \mid \mathbf{Y})$ for the posterior distributions.

4.4.1. Prior KL Divergence. The first two theorems prove the existence of lower bounds on how close the EB and FB priors can be, and show that these lower bounds become arbitrarily large as the number of tests m goes to infinity. We refer to these lower bounds as “information gaps,” and give them in both Kullback-Leibler (Theorem 4.2) and Hellinger (Theorem 4.3) flavors.

THEOREM 4.2. *Let $\underline{G}(m) = \min_{\hat{p}} KL(p_F(\boldsymbol{\gamma}) \parallel p_E(\boldsymbol{\gamma}))$. Then $\underline{G}(m) \rightarrow \infty$ as $m \rightarrow \infty$.*

PROOF. The KL divergence is

$$\begin{aligned}
 (24) \quad \text{KL} &= \sum_{k=0}^m \frac{1}{m+1} \left[\log \left(\frac{1}{m+1} \binom{m}{k}^{-1} \right) - \log \left(\hat{p}^k \cdot (1-\hat{p})^{m-k} \right) \right] \\
 &= -\log(m+1) - \frac{1}{m+1} \sum_{k=0}^m \left[\log \binom{m}{k} + k \log \hat{p} + (m-k) \log(1-\hat{p}) \right].
 \end{aligned}$$

This is minimized for $\hat{p} = 1/2$ regardless of m or k , meaning that:

$$\begin{aligned}
 (25) \quad \underline{\mathbf{G}}(m) &= -\log(m+1) - \frac{1}{m+1} \sum_{k=0}^m \left[\log \binom{m}{k} + m \log(1/2) \right] \\
 &= m \log 2 - \log(m+1) - \frac{1}{m+1} \sum_{k=0}^m \log \binom{m}{k}.
 \end{aligned}$$

The first (linear) term in (25) dominates the second (logarithmic) term, whereas results in Gould (1964) show the third term to be asymptotically linear in m with slope $1/2$. Hence $\underline{\mathbf{G}}(m)$ grows linearly with m , with asymptotic positive slope of $\log 2 - 1/2$. \square

THEOREM 4.3. *Let $\underline{\mathbf{H}}^2(m) = \min_{\hat{p}} H^2(p_F(\gamma) \parallel p_E(\gamma))$. Then $\underline{\mathbf{H}}^2(m) \rightarrow 1$ as $m \rightarrow \infty$.*

PROOF.

$$(26) \quad H^2(p_F(\gamma) \parallel p_E(\gamma)) = 1 - \frac{1}{\sqrt{m+1}} \sum_{k=0}^m \sqrt{\binom{m}{k} \hat{p}^k (1-\hat{p})^{m-k}}.$$

This distance is also minimized for $\hat{p} = 1/2$, meaning that:

$$(27) \quad \underline{\mathbf{H}}^2(m) = 1 - (m+1)^{-1/2} \cdot 2^{-m/2} \cdot \sum_{k=0}^m \sqrt{\binom{m}{k}}.$$

A straightforward application of Stirling's approximation to the factorial function shows that:

$$(28) \quad \lim_{m \rightarrow \infty} \left[(m+1)^{-1/2} \cdot 2^{-m/2} \cdot \sum_{k=0}^m \sqrt{\binom{m}{k}} \right] = 0,$$

from which the result follows immediately. \square

In summary, the *ex-post* prior distribution associated with the EB procedure is particularly troubling when the number of tests m grows without bound. On the one hand, when the true value of k remains fixed or grows at a rate slower than m —that is, when concerns over false positives become the most trenchant, and the case for a Bayesian procedure exhibiting strong multiplicity control becomes the most convincing—then $\hat{p} \rightarrow 0$ and the EB prior $p_E(\gamma)$ becomes arbitrarily bad as an approximation to $p_F(\gamma)$. On the other hand, if the true k is growing at the same rate as m , then the best one can hope for is that $\hat{p} = 1/2$. And even then, the information gap between $p_F(\gamma)$ and $p_E(\gamma)$ grows linearly without bound (for KL divergence), or converges to 1 (for Hellinger distance).

4.4.2. *Posterior KL Divergence.* The next theorem shows that, under very mild conditions, the expected KL divergence between FB and EB posteriors is infinite. This version assumes that the error precision ϕ is fixed, but the generalization to an unknown ϕ is straightforward.

THEOREM 4.4. *In the variable-selection problem, let $m, n > m$, and $\phi > 0$ be fixed. Suppose \mathbf{X}_γ is of full rank for all models and that the family of priors for model-specific parameters, $\{\pi(\beta_\gamma)\}$, are such that $p(\beta_\gamma = \mathbf{0}) < 1$ for all M_γ . Then, for any true model M_γ^T , the expected posterior KL divergence $E[KL(p_F(\gamma | \mathbf{Y}) \| p_E(\gamma | \mathbf{Y}))]$ under this true model is infinite.*

PROOF. The posterior KL divergence is

$$(29) \quad KL(p_F(\gamma | \mathbf{Y}) \| p_E(\gamma | \mathbf{Y})) = \sum_{\Gamma} p_F(M_\gamma | \mathbf{Y}) \cdot \log \left(\frac{p_F(M_\gamma | \mathbf{Y})}{p_E(M_\gamma | \mathbf{Y})} \right).$$

This is clearly infinite if there exists a model M_γ for which $p_E(M_\gamma | \mathbf{Y}) = 0$ but $p_F(M_\gamma | \mathbf{Y}) > 0$. Since the fully Bayesian posterior assigns nonzero probability to all models, this condition is met whenever the empirical-Bayesian solution is $\hat{p} = 0$ or $\hat{p} = 1$. Thus it suffices to show that \hat{p} will be 0 with positive probability under any true model.

Assume without loss of generality that $\phi = 1$. Recall that we are also assuming that $\pi(\alpha) = 1$ for all models, and that the intercept is orthogonal to all other covariates. Letting $\beta_\gamma^* = (\alpha, \beta_\gamma)^t$ for model M_γ , and letting $L(\cdot)$ stand for the likelihood, the marginal likelihood for any model can then be written

$$(30) \quad f(\mathbf{Y} | M_\gamma) = L(\hat{\beta}_\gamma^*) \cdot \sqrt{2\pi/n} \int_{\mathbb{R}^{k_\gamma}} g(\beta_\gamma) \pi(\beta_\gamma) d\beta_\gamma,$$

where

$$g(\beta_\gamma) = \exp \left\{ -\frac{1}{2}(\beta_\gamma - \hat{\beta}_\gamma)^t \mathbf{X}_\gamma^t \mathbf{X}_\gamma (\beta_\gamma - \hat{\beta}_\gamma) \right\}$$

The Bayes factor for comparing the null model to any model is:

$$B_\gamma(\mathbf{Y}) = \frac{f(\mathbf{Y} | M_0)}{f(\mathbf{Y} | M_\gamma)},$$

which from (30) is clearly continuous as a function of \mathbf{Y} for every γ . Evaluated at $\mathbf{Y} = \mathbf{0}$, this Bayes factor satisfies

$$(31) \quad B_\gamma(\mathbf{0}) = \left(\int_{\mathbb{R}^{k_\gamma}} \exp \left\{ -\frac{1}{2}(\hat{\beta}_\gamma - \beta_\gamma)^t \mathbf{X}_\gamma^t \mathbf{X}_\gamma (\hat{\beta}_\gamma - \beta_\gamma) \right\} \pi(\beta_\gamma) d\beta_\gamma \right)^{-1} > 1$$

for each M_γ under the assumptions of the theorem.

By continuity, for every model M_γ there exists an ϵ_γ such that $B_\gamma(\mathbf{Y}) > 1$ for any $|\mathbf{Y}| < \epsilon_\gamma$. Let $\epsilon^* = \min_\gamma \epsilon_\gamma$. Then for \mathbf{Y} satisfying $|\mathbf{Y}| < \epsilon^*$, $B_\gamma(\mathbf{Y}) > 1$ for all non-null models, meaning that M_0 will have the largest marginal likelihood. By Lemma 4.1, $\hat{p} = 0$ when such a \mathbf{Y} is observed.

But under any model, there is positive probability of observing $|\mathbf{Y}| < \epsilon^*$ for any positive ϵ^* , since this set has positive Lebesgue measure. Hence regardless of the true model, there is positive probability that the KL divergence $\text{KL}(p_F(\gamma | \mathbf{Y}) || p_E(\gamma | \mathbf{Y}))$ is infinite under the sampling distribution $p(\mathbf{Y} | M_\gamma)$, and so its expectation is clearly infinite. \square

Since the expected KL divergence is infinite for any number m of variables being tested, and for any true model, it is clear that $E(KL)$ does not converge to 0 as $m \rightarrow \infty$. This, of course, is a weaker conclusion than would be a lack of KL divergence in probability, but it is nevertheless interesting.

In Theorem 4.4, the expectation is with respect to the sampling distribution under a specific model M_γ , with β_γ either fixed or marginalized away with respect to a prior distribution. But this result implies an infinite expectation with respect to other reasonable choices of the expectation distribution—for example, under the Bernoulli sampling model for γ in (5) with fixed prior inclusion probability p .

5. Numerical Investigation of Empirical-Bayes Variable Selection.

5.1. *Overview.* The theoretical results in Section 4 indicate that empirical-Bayes analysis cannot always be trusted in the variable selection problem. This section presents numerical results that indicate that these concerns

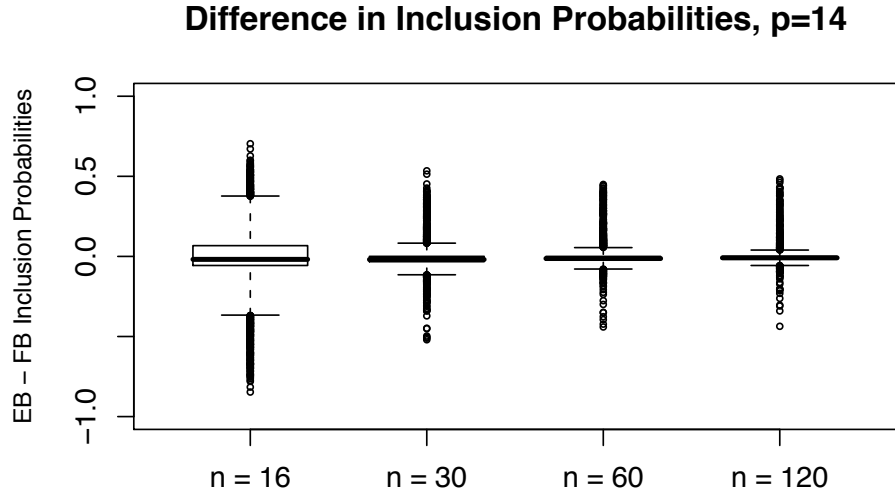


FIG 3. Differences in inclusion probabilities between EB and FB analyses in the simulation study.

are of practical significance, and not mere theoretical curiosities. As in the previous section, most of the investigation is phrased as a comparison of empirical-Bayes and fully Bayesian analysis, but at least some of the findings point out obviously inappropriate properties of the empirical-Bayes procedure itself.

5.2. *Results under Properly Specified Priors.* The following simulation was performed 75,000 times for each of four different sample sizes:

1. Draw a random $m \times n$ design matrix \mathbf{X} of independent $N(0, 1)$ covariates.
2. Draw a random $p \sim U(0, 1)$, and draw a sequence of m independent Bernoulli trials with success probability p to yield a binary vector γ encoding the true set of regressors.
3. Draw β_γ , the vector of regression coefficients corresponding to the nonzero elements of γ , from a Zellner-Siow prior. Set the other coefficients $\beta_{-\gamma}$ to 0.
4. Draw a random vector of responses $\mathbf{Y} \sim N(\mathbf{X}\beta, \mathbf{I})$.
5. Using only \mathbf{X} and \mathbf{Y} , compute marginal likelihoods (assuming Zellner-Siow priors) for all 2^m possible models; use these quantities to compute \hat{p} along with the EB and FB posterior distributions across model space.

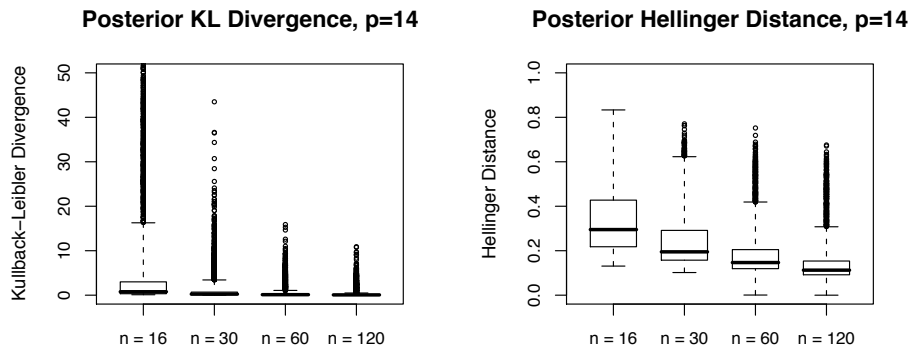


FIG 4. Realized KL divergence and Hellinger distance between FB and EB posterior distributions in the simulation study, $n = \{16, 30, 60, 120\}$.

In all cases m was fixed at 14, yielding a model space of size 16,384—large enough to be interesting, yet small enough to be enumerated 75,000 times in a row. We repeated the experiment for four different sample sizes ($n = 16$, $n = 30$, $n = 60$, and $n = 120$) to simulate a variety of different m/n ratios.

Three broad patterns emerged from these experiments. First, the two procedures often reached very different conclusions about which covariates were important. Figure 3 shows frequent large discrepancies between the posterior inclusion probabilities given by the EB and FB procedures. This happened even when n was relatively large compared to the number of parameters being tested, suggesting that even large sample sizes do not render a data set immune to this difference.

Second, while the realized KL divergence and Hellinger distance between EB and FB posterior distributions did tend to get smaller with more data, they did not approach 0 particularly fast (Figure 4). These boxplots show long upper tails even when $n = 120$, indicating that, with nontrivial frequency, the EB procedure can differ significantly from the FB posterior.

Finally, both procedures make plenty of mistakes when classifying variables as being in or out of the model, but these mistakes differ substantially in their overall character. For each simulated data set, the number of false positives and false negatives declared by the EB and FB median-probability models were recorded. These two numbers give an (x, y) pair that can then be plotted (along with the pairs from all other simulated data sets) to give a graphical representation of the kind of mistakes each procedure makes under repetition. The four panes of Figure 5 show these plots for all four sample sizes. Each integer (x, y) location contains a circle whose color—red for EB, blue for FB—shows which procedure made that kind of mistake more often,

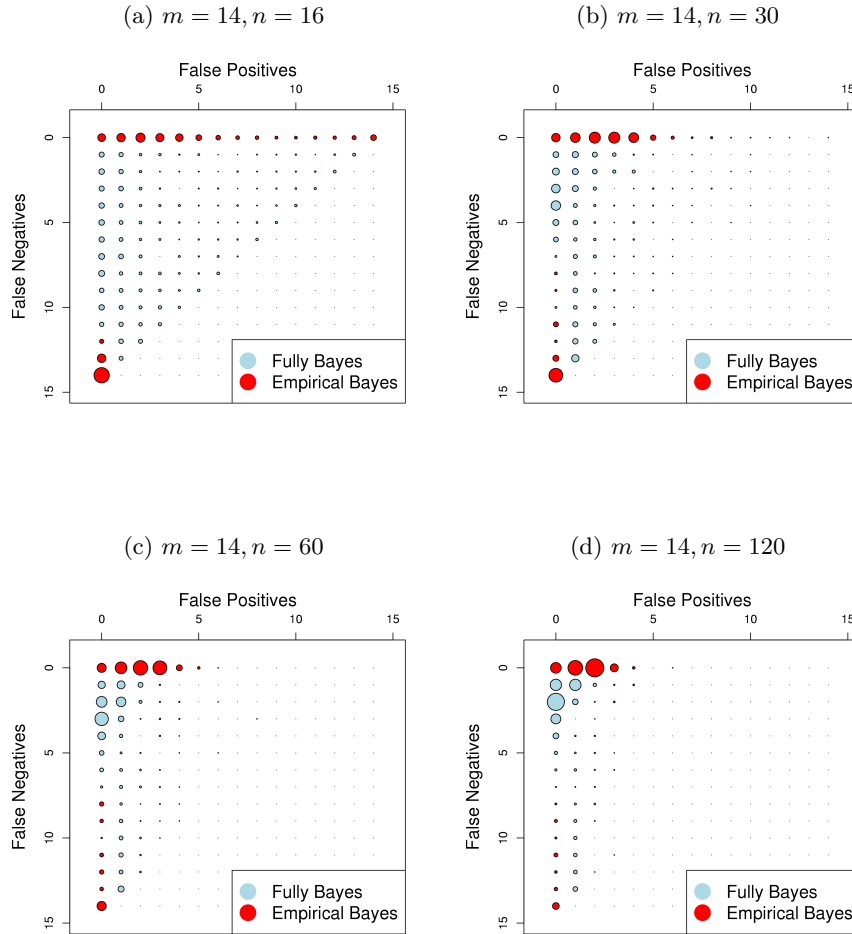


FIG 5. *Orthogonal design, $m = 14$. The differential pattern of errors under the 75,000 simulated data sets. The area of the circle represents how overrepresented the given procedure (red for EB, blue for FB) is in that cell. The circle at (3 right, 2 down), for example, represents an error involving 3 false positives and 2 false negatives.*

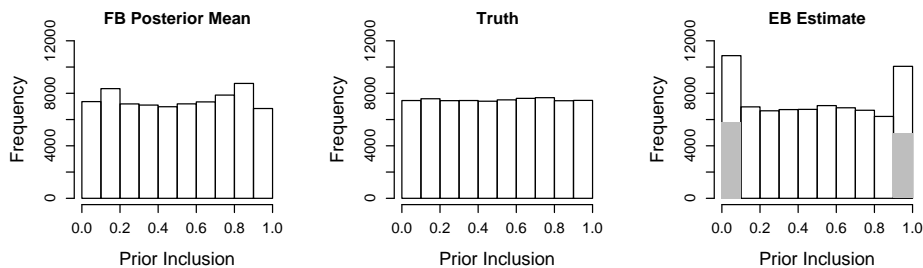


FIG 6. Distribution of \hat{p} in the simulation study with a correctly specified (uniform) prior for p . The grey bars indicated the number of times, among values of \hat{p} in the extremal bins, that the empirical-Bayes solution collapsed to the degenerate $\hat{p} = 0$ or $\hat{p} = 1$.

and whose area shows how much more often it made that mistake.

Notice that, regardless of sample size, the EB procedure tends to give systematically more extreme answers. In particular, it seems more susceptible to making Type-I errors—worrying for a multiple-testing procedure.

Much of this overshooting can be explained by Figure 6. The EB procedure gives the degenerate $\hat{p} = 0$ or $\hat{p} = 1$ solution much too often—over 15% of the time even when n is fairly large—suggesting that the issues raised by Theorem 4.4 can be quite serious in practice.

5.3. Results Under Improperly Specified Priors. The previous section demonstrated that significant differences can exist between fully Bayesian and empirical-Bayes variable selection in finite-sample settings. As a criticism of empirical-Bayes analysis, however, there was an obvious bias: the fully Bayesian procedure was being evaluated under its true prior distribution, with respect to which it is necessarily optimal.

It is thus of interest to do a similar comparison for situations in which the prior distribution is specified incorrectly: the fully Bayesian answers will assume a uniform prior p , but p will actually be drawn from a non-uniform distribution. We limit ourselves to discussion of the analogue of Figure 6 for various situations, all with $m = 14$ and $n = 60$. Three different choices of the true distribution for p were investigated, again with 75,000 simulated data sets each:

1. $p \sim \text{Be}(3/2, 3/2)$, yielding mainly moderate (but not uniform) values of p .
2. $p \sim \text{Be}(1, 2)$, yielding mainly smaller values of p .
3. $p \sim 0.5 \cdot \text{Be}(1/2, 8) + 0.5 \cdot \text{Be}(8, 1/2)$, yielding primarily values of p close to 0 or 1.

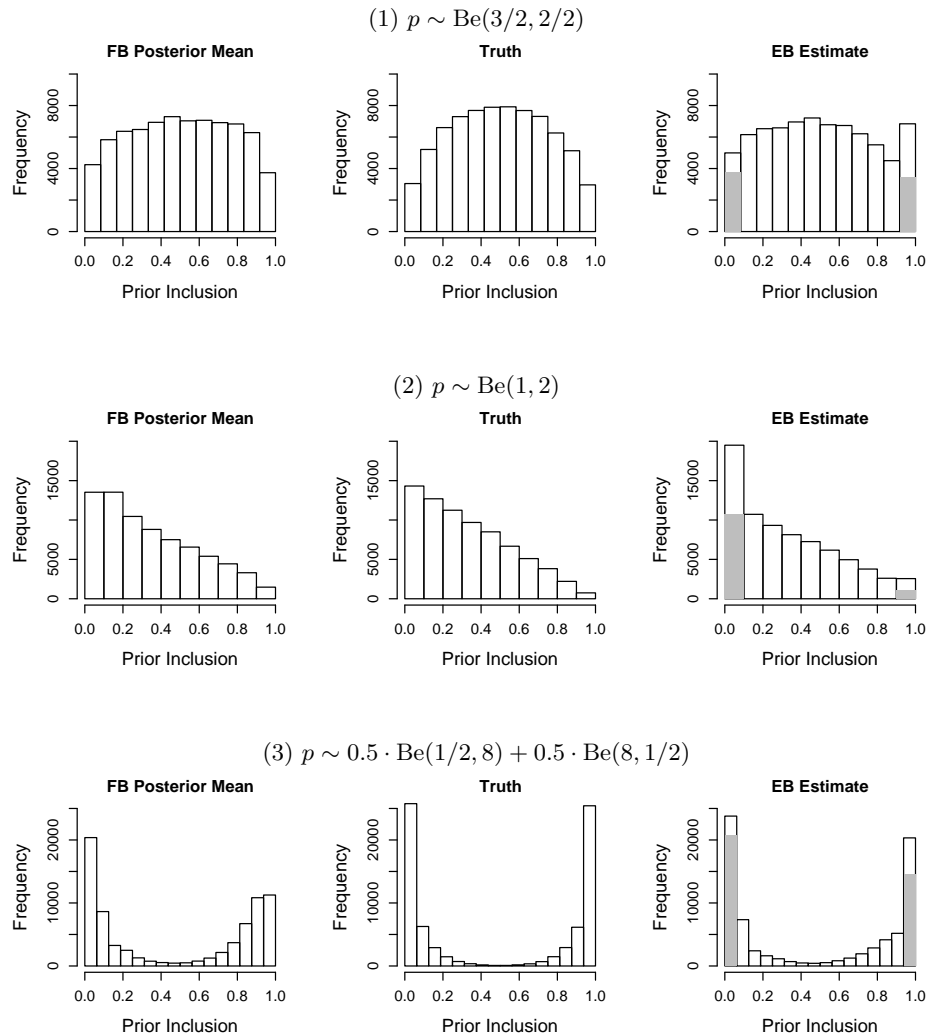


FIG 7. Distribution of \hat{p} in different versions of the simulation study, where the fully Bayesian model had a misspecified (uniform) prior on p . The grey bars indicated the number of times, among values of \hat{p} in the extremal bins, that the empirical-Bayes solution collapsed to the degenerate $\hat{p} = 0$ or $\hat{p} = 1$.

The results are summarized in Figure 7. In each case the central pane shows the true distribution of p , with the left pane showing the Bayesian posterior means under the uniform prior and the right pane showing the empirical-Bayes estimates \hat{p} .

As expected, the incorrectly specified Bayesian model tends to shrink the estimated values of p back to the prior mean of 0.5. This tendency is especially noticeable in Case 3, where the true distribution contains many extreme values of p . This gives the illusion that empirical-Bayes tends to do better here.

Notice, however, the grey bars in the right-most panes. These bars indicate the percentage of time, among values of \hat{p} that fall in the left- or right-most bins of the histogram, that the empirical-Bayes solution is exactly 0 or 1 respectively. For example, of the roughly 20,000 times that $\hat{p} \in [0, 0.1)$ in Case 2, it was identically 0 more than 10,000 of those times. (The fully Bayesian posterior mean, of course, is never exactly 0 or 1.)

The bottom panel of Figure 7 shows that, paradoxically, where the fully Bayesian model is most incorrect, its advantages over the empirical-Bayes procedure are the strongest. In the mixture model giving many values of p very close to 0 or 1, empirical-Bayes collapses to a degenerate solution nearly half the time. Even if the extremal model is true in most of these cases, recall that the empirical Bayes procedure would result in an inappropriate statement of certainty in the model. Of course, this would presumably be noticed and some correction would be entertained, but the frequency of having to make the correction is itself worrisome.

In these cases, while the fully Bayesian posterior mean is necessarily shrunk back to the prior mean, this shrinkage is not very severe, and the uniform prior giving rise to such shrinkage can easily be modified if it is believed to be wrong. And in cases where the uniform prior is used incorrectly, a slight amount of unwanted shrinkage seems a small price to pay for the preservation of real prior uncertainty.

5.4. *Example: Determinants of Economic Growth.* The following data set serves to illustrate the differences between EB and FB answers in a scenario of typical size, complexity, and m/n ratio.

Many econometricians have applied Bayesian methods to the problem of GDP-growth regressions, where long-term economic growth is explained in terms of various political, social, and geographical predictors. Fernandez et al. (2001) popularized the use of Bayesian model averaging in the field; Sala-i Martin et al. (2004) used a Bayes-like procedure called BACE, similar to BIC-weighted OLS estimates, for selecting a model; and Ley and Steel

Covariate	Fully Bayes	Emp. Bayes
East Asian Dummy	0.983	0.983
Fraction of Tropical Area	0.727	0.653
Life Expectancy in 1960	0.624	0.499
Population Density Coastal in 1960s	0.518	0.379
GDP in 1960 (log)	0.497	0.313
Outward Orientation	0.417	0.318
Fraction GDP in Mining	0.389	0.235
Land Area	0.317	0.121
Higher Education 1960	0.297	0.148
Investment Price	0.226	0.130
Fraction Confucian	0.216	0.145
Latin American Dummy	0.189	0.108
Ethnolinguistic Fractionalization	0.188	0.117
Political Rights	0.188	0.081
Primary Schooling in 1960	0.167	0.093
Hydrocarbon Deposits in 1993	0.165	0.093
Fraction Spent in War 1960–90	0.164	0.095
Defense Spending Share	0.156	0.085
Civil Liberties	0.154	0.075
Average Inflation 1960–90	0.150	0.064
Real Exchange Rate Distortions	0.146	0.071
Interior Density	0.139	0.067

TABLE 3

Exact inclusion probabilities for 22 variables in a linear model for GDP growth among a group of 30 countries.

(2007) considered the effect of prior assumptions (particularly the pseudo-objective $p = 1/2$ prior) on these regressions.

We study a subset of the data from Sala-i Martin et al. (2004) containing 22 covariates on 30 different countries. A data set of this size allows the model space to be enumerated and the EB estimate \hat{p} to be calculated explicitly, which would be impossible on the full data set. The 22 covariates correspond to the top 10 covariates flagged in the BACE study, along with 12 others chosen uniformly at random from the remaining candidates.

Summaries of exact EB and FB analyses (with Zellner-Siow priors) can be found in Table 3. Two results are worth noting. First, the EB inclusion probabilities are nontrivially different from their FB counterparts, often disagreeing by 10% or more.

Second, if these are used for model selection, quite different results would emerge. For instance, if median-probability models were selected (i.e., one includes only those variables with inclusion probability greater than $1/2$), the FB analysis would include the first four variables (and would almost choose the fifth variable), while the EB analysis would select only the first

two variables (and almost the third). While we would not endorse simply choosing a model here, note that doing so would result in fundamentally different economic pictures for the FB and EB analysis.

6. Summary. This paper started out as an attempt to more fully understand when, and how, multiplicity correction automatically occurs in Bayesian analysis, and to examine the importance of ensuring that such multiplicity correction is included. That the correction can only happen through the choice of appropriate prior probabilities of models seemed to conflict with the intuition that multiplicity correction occurs through data-based adaptation of the prior-inclusion probability p .

The resolution to this conflict—that the multiplicity correction is indeed pre-fixed in the prior probabilities, but the amount of correction employed will depend on the data—led to another conflict: how can the empirical-Bayes approach to variable selection be an accurate approximation to the full Bayesian analysis? Indeed, we have seen in the paper that empirical-Bayes variable selection can lead to results quite different than those from the full Bayesian analysis. This difference was evidenced through examples (both simple pedagogical examples and a more realistic practical example), through simulation studies, and through information-based theoretical results. These studies, as well as the results about the tendency of empirical-Bayes variable selection to choose extreme \hat{p} , all supported the general conclusions about empirical-Bayes variable selection that were mentioned in the introduction.

Of course, there are many fine empirical-Bayes analyses that have been done in model selection and variable selection, and we have not said that any such analysis is wrong. We are suggesting, however, that empirical-Bayes variable selection does not carry the automatic guarantee of performance that accompanies empirical-Bayes methodology in many other contexts, and so additional care should be taken in its use.

References.

- F. Abramovich, Y. Benjamini, D. Donoho, and I. Johnstone. Adapating to unknown sparsity by controlling the false-discovery rate. *The Annals of Statistics*, 34(2):584–653, 2006.
- M. Barbieri and J. O. Berger. Optimal predictive model selection. *The Annals of Statistics*, 32(3):870–897, 2004.
- J. Berger, L. Pericchi, and J. Varshavsky. Bayes factors and marginal distributions in invariant situations. *Sankhya, Ser. A*, 60:307–321, 1998.
- J. O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, 2nd edition, 1985.
- J. O. Berger and G. Molina. Posterior model probabilities via path-based pairwise priors. *Statistica Neerlandica*, 59(1):3–15, 2005.

- J. O. Berger and L. Pericchi. Objective Bayesian methods for model selection: introduction and comparison. In *Model Selection*, volume 38 of *Institute of Mathematical Statistics Lecture Notes – Monograph Series*, pages 135–207. Beachwood, 2001.
- D. Berry. Multiple comparisons, multiple tests, and data dredging: A Bayesian perspective. In J. Bernardo, M. DeGroot, D. Lindley, and A. Smith, editors, *Bayesian Statistics 3*, pages 79–94. Oxford University Press, 1988.
- D. Berry and Y. Hochberg. Bayesian perspectives on multiple comparisons. *Journal of Statistical Planning and Inference*, 82:215–277, 1999.
- M. Bogdan, J. K. Ghosh, and S. T. Tokdar. A comparison of the Benjamini-Hochberg procedure with some Bayesian rules for multiple testing. In *Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen*, volume 1, pages 211–30. Institute of Mathematical Statistics, 2008.
- B. Carlin and T. Louis. Empirical Bayes: Past, present and future. *Journal of the American Statistical Association*, 95(452):1286–89, 2000.
- C. M. Carvalho and J. G. Scott. Objective Bayesian model selection in Gaussian graphical models. Technical report, Institute of Statistics and Decision Sciences, 2007.
- G. Casella and E. Moreno. Objective Bayes variable selection. Technical Report 023, University of Florida, 2002.
- W. Cui and E. I. George. Empirical Bayes vs. fully Bayes variable selection. *Journal of Statistical Planning and Inference*, 138:888–900, 2008.
- K.-A. Do, P. Muller, and F. Tang. A Bayesian mixture model for differential gene expression. *Journal of the Royal Statistical Society, Series C*, 54(3):627–44, 2005.
- M. Eaton. *Group Invariance Applications in Statistics*. Institute of Mathematical Statistics, 1989.
- B. Efron, T. R., J. Storey, and V. Tusher. Empirical Bayes analysis of a microarray experiment. *Journal of American Statistical Association*, 96:1151–1160, 2001.
- C. Fernandez, E. Ley, and M. Steel. Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics*, 16:563–76, 2001.
- E. I. George and D. P. Foster. Calibration and empirical Bayes variable selection. *Biometrika*, 87(4):731–747, 2000.
- R. Gopalan and D. Berry. Bayesian multiple comparisons using Dirichlet process priors. *Journal of the American Statistical Association*, 93:1130–1139, 1998.
- H. Gould. Sums of logarithms of binomial coefficients. *The American Mathematical Monthly*, 71(1):55–58, 1964.
- W. Jefferys and J. Berger. Ockham’s razor and Bayesian analysis. *American Scientist*, 80:64–72, 1992.
- H. Jeffreys. *Theory of Probability*. Oxford University Press, 3rd edition, 1961.
- I. M. Johnstone and B. W. Silverman. Needles and straw in haystacks: Empirical-Bayes estimates of possibly sparse sequences. *The Annals of Statistics*, 32(4):1594–1649, 2004.
- E. Ley and M. F. Steel. On the effect of prior assumptions in Bayesian model averaging with applications to growth regression. Number 4238 in Policy Research Working Paper Series. World Bank, 2007.
- F. Liang, R. Paulo, G. Molina, M. Clyde, and J. Berger. Mixtures of g -priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103:410–23, 2008.
- C. Meng and A. Dempster. A Bayesian approach to the multiplicity problem for significance testing with binomial data. *Biometrics*, 43:301–11, 1987.
- X. Sala-i Martin, G. Doppelhofer, and R. I. Miller. Determinants of long-term growth: A Bayesian averaging of classical estimates (bace) approach. *American Economic Review*, 94:813–835, 2004.
- J. G. Scott. Nonparametric Bayesian multiple-hypothesis testing of autoregressive time

- series. Discussion Paper 2008-09, Duke University Department of Statistical Science, 2008.
- J. G. Scott and J. O. Berger. An exploration of aspects of Bayesian multiple testing. *Journal of Statistical Planning and Inference*, 136(7):2144–2162, 2006.
- J. G. Scott and C. M. Carvalho. Feature-inclusion stochastic search for gaussian graphical models. *Journal of Computational and Graphical Statistics*, 2008. to appear.
- R. Waller and D. Duncan. A Bayes rule for the symmetric multiple comparison problem. *Journal of the American Statistical Association*, 64:1484–1503, 1969.
- P. H. Westfall, W. O. Johnson, and J. M. Utts. A Bayesian perspective on the Bonferroni adjustment. *Biometrika*, 84:419–27, 1997.
- A. Zellner. On assessing prior distributions and Bayesian regression analysis with g -prior distributions. In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, pages 233–243. Elsevier, 1986.
- A. Zellner and A. Siow. Posterior odds ratios for selected regression hypotheses. In *Bayesian Statistics: Proceedings of the First International Meeting held in Valencia*, pages 585–603, 1980.

APPENDIX A: VARIATIONS ON ZELLNER’S G -PRIOR

Conventional variable-selection priors rely upon the conjugate normal-gamma family of distributions, which yields closed-form expression for the marginal likelihoods. To give an appropriate scale for the normal prior describing the regression coefficients, Zellner (1986) suggested a particular form of this family:

$$\begin{aligned}(\boldsymbol{\beta} \mid \phi) &\sim N\left(\boldsymbol{\beta}_0, \frac{g}{\phi}(\mathbf{X}'\mathbf{X})^{-1}\right) \\ \phi &\sim \text{Ga}\left(\frac{\nu}{2}, \frac{\nu s}{2}\right),\end{aligned}$$

with prior mean $\boldsymbol{\beta}_0$, often chosen to be 0. The conventional choice $g = n$ gives a prior covariance matrix for the regression parameters equal to the unit Fisher information matrix for the observed data \mathbf{X} . This prior can be interpreted as encapsulating the information arising from a single observation under a hypothetical experiment with the same design as the one to be analyzed.

Zellner’s g -prior was originally formulated for testing a precise null hypothesis, $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$, versus the alternative $H_A : \boldsymbol{\beta} \in \mathbb{R}^p$. But others have adapted Zellner’s methodology to the more general problem of testing nested regression models by placing a flat prior on the parameters shared by the two models and using a g -prior only on the parameters not shared by the smaller model. This seems to run afoul of the general injunction against improper priors in model selection problems, but can nonetheless be formally justified by arguments appealing to orthogonality and group invariance; see, for example, Berger et al. (1998) and Eaton (1989). These arguments apply

to cases where all covariates have been centered to have a mean of zero, which is assumed without loss of generality to be true.

A full variable-selection problem, of course, involves many non-nested comparisons. Yet Bayes factors can still be formally defined using the “encompassing model” approach of Zellner and Siow (1980), who operationally define all marginal likelihoods in terms of Bayes factors with respect to a base model M_B :

$$(32) \quad \text{BF}(M_1 : M_2) = \frac{\text{BF}(M_1 : M_B)}{\text{BF}(M_2 : M_B)}.$$

Since the set of common parameters which are to receive improper priors depends upon the choice of base model, different choices yield a different ensemble of Bayes factors and imply different “operational” marginal likelihoods. And while this choice of M_B is free in principle, there are only two such choices which yield a pair of nested models in all comparisons: the null model and the full model.

In the null-based approach, each model is compared to the null model consisting only of the intercept α . This parameter, along with the precision ϕ , is common to all models, leading to a prior specification that has become the most familiar version of Zellner’s g -prior:

$$\begin{aligned} (\alpha, \phi \mid \gamma) &\propto 1/\phi \\ (\beta_\gamma \mid \phi, \gamma) &\sim \text{N}\left(0, \frac{g}{\phi}(\mathbf{X}'_\gamma \mathbf{X}_\gamma)^{-1}\right). \end{aligned}$$

This gives a simple expression for the Bayes factor for evaluating a model γ with k regression parameters (excluding the intercept):

$$(33) \quad \text{BF}(M_\gamma : M_0) = (1 + g)^{(n-k_\gamma-1)/2} [1 + (1 - R_\gamma^2)g]^{-(n-1)/2},$$

where $R_\gamma^2 \in (0, 1]$ is the usual coefficient of determination for model M_γ .

Adherents of the full-based approach, on the other hand, compare all models to the full model, on the grounds that the full model is usually much more scientifically reasonable than the null model and provides a more sensible yardstick (Casella and Moreno, 2002). This comparison can be done by writing the full model as:

$$M_F : \mathbf{Y} = \mathbf{X}_\gamma^* \theta_\gamma + \mathbf{X}_{-\gamma} \beta_{-\gamma}$$

with the design matrix partitioned in the obvious way. Then a g -prior is specified for the parameters in the full model not shared by the smaller

model, which again has k regression parameters excluding the intercept:

$$\begin{aligned} (\alpha, \beta_\gamma, \phi \mid \gamma) &\propto 1/\phi \\ (\beta_{-\gamma} \mid \phi, \gamma) &\sim N\left(0, \frac{g}{\phi}(\mathbf{X}'_{-\gamma}\mathbf{X}_{-\gamma})^{-1}\right) \end{aligned}$$

This does not lead to a coherent “within-model” prior specification for the parameters of the full model, since their prior distribution depends upon which submodel is considered. Nevertheless, marginal likelihoods can still be consistently defined in the manner of Equation 32. Conditional upon g , this yields a Bayes factor in favor of the full model of

$$(34) \quad \text{BF}(M_F : M_\gamma) = (1 + g)^{(n-m-1)/2} (1 + gW)^{-(n-k-1)/2}$$

where $W = (1 - R_F^2)/(1 - R_\gamma^2)$.

The existence of these simple expressions has made the use of g -priors very popular. Yet g -priors yield display a disturbing type of behavior often called the “information paradox.” This can be seen in (33): the Bayes factor in favor of M_γ goes to the finite constant $(1 + g)^{n-m-1}$ as $R_\gamma^2 \rightarrow 1$ (which can only happen if M_γ is true and the residual variance goes to 0). For typical problems this will be an enormous number, but still quite a bit smaller than infinity. Hence the paradox: the Bayesian procedure under a g -prior places an intrinsic limit upon the possible degree of convincingness to be found in the data, a limit which is confirmed neither by intuition nor by the behavior of the classical test statistic.

Liang et al. (2008) detail several versions of information-consistent g -like priors. One way is to estimate g by empirical-Bayes methods (George and Foster, 2000). A second, fully Bayesian, approach involves placing a prior upon g that satisfies the condition $\int_0^\infty (1 + g)^{n-k_\gamma-1} \pi(g) \, dg = \infty$ for all $k_\gamma \leq p$, which is a generalization of the condition given in Jeffreys (1961) (see Chapter 5.2, Equations 10 and 14).

This second approach generalizes the recommendations of Zellner and Siow (1980), who compare models by placing a flat prior upon common parameters and a g -like Cauchy prior on non-shared parameters:

$$(35) \quad (\beta_\gamma \mid \phi) \sim C\left(0, \frac{n}{\phi}(\mathbf{X}'_\gamma\mathbf{X}_\gamma)^{-1}\right)$$

These have come to be known as Zellner-Siow priors, and their use can be shown to resolve the information paradox. Although they do not yield closed-form expressions for marginal likelihoods, one can exploit the scale-mixture-of-normals representation of the Cauchy distribution to leave one-dimensional integrals over standard g -prior marginal likelihoods with respect

to an inverse-gamma prior, $g \sim \text{IG}(1/2, 2/n)$. The Zellner-Siow null-based Bayes factor under model M_γ then takes the form:

$$(36) \quad \text{BF}(M_\gamma : M_0) = \int_0^\infty (1+g)^{(n-k_\gamma-1)/2} [1+(1-R_\gamma^2)g]^{-(n-1)/2} g^{-3/2} \exp(-n/(2g)) dg$$

A similar formula exists for the full-based version:

$$(37) \quad \text{BF}(M_F : M_\gamma) = \int_0^\infty (1+g)^{(n-m-1)/2} [1+Wg]^{-(n-k-1)/2} g^{-3/2} \exp(-n/(2g)) dg$$

with W given above.

These quantities can be computed by one-dimensional numerical integration, but in high-dimensional model searches this will be a bottleneck. Luckily there exists a closed-form approximation to these integrals first noted in Liang et al. (2008). It entails computing the roots of a cubic equation, and extensive numerical experiments show the approximation to be quite accurate. These Bayes factors seem to offer an excellent compromise between good theoretical behavior and computational tractability, thereby overcoming the single biggest hurdle to the widespread practical use of Zellner-Siow priors.

DUKE UNIVERSITY
Box 90251
DURHAM, NC 27708 USA
E-MAIL: james@stat.duke.edu