

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

BY JAMES G. SCOTT¹ AND JAMES O. BERGER

University of Texas at Austin and Duke University

This paper studies the multiplicity-correction effect of standard Bayesian variable-selection priors in linear regression. Our first goal is to clarify when, and how, multiplicity correction happens automatically in Bayesian analysis, and to distinguish this correction from the Bayesian Ockham's-razor effect. Our second goal is to contrast empirical-Bayes and fully Bayesian approaches to variable selection through examples, theoretical results and simulations. Considerable differences between the two approaches are found. In particular, we prove a theorem that characterizes a surprising asymptotic discrepancy between fully Bayes and empirical Bayes. This discrepancy arises from a different source than the failure to account for hyperparameter uncertainty in the empirical-Bayes estimate. Indeed, even at the extreme, when the empirical-Bayes estimate converges asymptotically to the true variable-inclusion probability, the potential for a serious difference remains.

1. Introduction. This paper addresses concerns about multiplicity in the traditional variable-selection problem for linear models. We focus on Bayesian and empirical-Bayesian approaches to the problem. These methods both have the attractive feature that they can, if set up correctly, account for multiplicity automatically, without the need for ad-hoc penalties.

Given the huge number of possible predictors in many of today's scientific problems, these concerns about multiplicity are becoming ever more relevant. They are especially critical 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 to buttress scientific conclusions or guide policy decisions—conclusions or decisions that may be quite wrong if the implicit multiple-testing problem is ignored.

Our first objective is to clarify how multiplicity correction enters Bayesian variable selection: by allowing the choice of prior model probabilities to depend upon

Received June 2008; revised November 2009.

¹Supported in part by the U.S. National Science Foundation under a Graduate Research Fellowship and Grants AST-0507481 and DMS-01-03265.

AMS 2000 subject classifications. 62J05, 62J15.

Key words and phrases. Bayesian model selection, empirical Bayes, multiple testing, variable selection.

the data in an appropriate way. Some useful references on this idea include Waller and Duncan (1969), Meng and Dempster (1987), Berry (1988), Westfall, Johnson and Utts (1997), Berry and Hochberg (1999) and Scott and Berger (2006). We also clarify the difference between multiplicity correction and the Bayesian Ockham's-razor effect [see Jefferys and Berger (1992)], which induces a very different type of penalty on model complexity. This discussion will highlight the fact that not all Bayesian analyses automatically adjust for multiplicity.

Our second objective is to describe and investigate a peculiar discrepancy between fully Bayes and empirical-Bayes variable selection. This discrepancy seems to arise from a different source than the failure to account for uncertainty in the empirical-Bayes estimate—the usual issue in such problems. Indeed, even when the empirical-Bayes estimate converges asymptotically to the true hyperparameter value, the potential for a serious difference remains.

The existence of such a discrepancy between fully Bayesian answers and empirical-Bayes answers—especially one that persists even in the limit—is of immediate interest to Bayesians, who often use empirical Bayes as a computational simplification. But the discrepancy is also of interest to non-Bayesians for at least two 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), Bogdan, Ghosh and Zak-Szatkowska (2008), Cui and George (2008), Bogdan, Chakrabarti and Ghosh (2008) and Bogdan, Ghosh and Tokdar (2008).

Second, theoretical and numerical investigations of the discrepancy revealed some unsettling properties of the standard empirical-Bayes analysis in variable selection. Of most concern is that empirical Bayes has the potential to collapse to a degenerate solution, resulting in an inappropriate statement of certainty in the selected regression model. As a simple example, suppose the usual variable-selection prior is used, where each variable is presumed to be in the model independently 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). This estimated \hat{p} is then used to determine the posterior 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 unusual character of the problem in variable selection seems not to have been recognized.

In bringing this issue to light, our goal is not to criticize empirical-Bayes analysis per se. Indeed, this paper will highlight many virtues of the empirical-Bayes approach to variable selection, especially compared to the nonadaptive model prior probabilities that are often used for variable selection. Our primary goal is comparative, rather than evaluative, in nature. In particular, we wish to explore the implications of the above discrepancy for Bayesians, who are likely to view empirical Bayes as an approximation to full Bayes analysis, and who wish to understand when the approximation is a good one. We recognize that others have alternative goals for empirical Bayes, and that these goals do not involve approximating full Bayes analysis. Also, there are non-Bayesian alternatives to marginal maximum likelihood in estimating p , as shown in some of the above papers. The results in this paper suggest that such alternatives be seriously considered by those wishing to adopt the empirical-Bayes approach, especially in potentially degenerate situations.

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. Sections 4 and 5 introduce 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 6 presents numerical results indicating the practical nature of the differences, through a simulation experiment and a practical example. Section 7 gives further discussion of the results.

2. Preliminaries.

2.1. *Notation.* 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

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

for some $\{j_1, \dots, j_k\} \subset \{1, \dots, m\}$, where $\varepsilon_i \stackrel{\text{i.i.d.}}{\sim} N(0, \phi^{-1})$ for an unknown variance ϕ^{-1} .

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

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

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, $p(M_{\boldsymbol{\gamma}})$, along with

the marginal likelihood of the data under each model:

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

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

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

Let $\mathbf{X}_{\boldsymbol{\gamma}}$ denote the columns of the full design matrix \mathbf{X} given by the nonzero elements of $\boldsymbol{\gamma}$, and let $\mathbf{X}_{\boldsymbol{\gamma}}^*$ denote the concatenation $(\mathbf{1} \ \mathbf{X}_{\boldsymbol{\gamma}})$, 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 $\mathbf{X}_{\boldsymbol{\gamma}}$ are orthogonal. We will also assume that the common choice $\pi(\alpha) = 1$ is made for the parameter α in each model [see Berger, Pericchi and Varshavsky (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:

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

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 g -priors [Zellner (1986)] for computing the marginal likelihoods in (2); explicit expressions can be found in the Appendix. See Liang et al. (2008) for a recent discussion of g -priors, and mixtures thereof, for variable selection.

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, resulting in a more diffuse predictive distribution for the data.

While this is a penalty against more complex models, it is not a multiple-testing penalty. Observe that 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 (i.e., 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. See Bogdan, Ghosh and Tokdar (2008) for extensive discussion of this issue.

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

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

with $k_{\mathcal{Y}}$ 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):

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

One uses this in (5) to define the ex-post prior probabilities $p(M_{\mathcal{Y}} | \hat{p}) = \hat{p}^{k_{\mathcal{Y}}} (1 - \hat{p})^{m - k_{\mathcal{Y}}}$, resulting in final model posterior probabilities

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

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 a straightforward way: if there are only k true variables and m grows large, then $\hat{p} \rightarrow 0$. This will make it increasingly more difficult for all variables to overcome the ever-stronger prior bias against their relevance.

3.3. *A fully Bayesian version.* Fully Bayesian variable-selection priors have been discussed by Ley and Steel (2009), Cui and George (2008) and Carvalho and Scott (2009), among others. These priors assume that p has a Beta distribution, $p \sim \text{Be}(a, b)$, giving

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

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

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

We call these expressions deriving from the uniform prior on p the “fully Bayes” version of variable selection priors, though of course many other priors could be used (including those incorporating subject-area information). Utilizing these prior probabilities in (3) yields the following posterior probabilities:

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

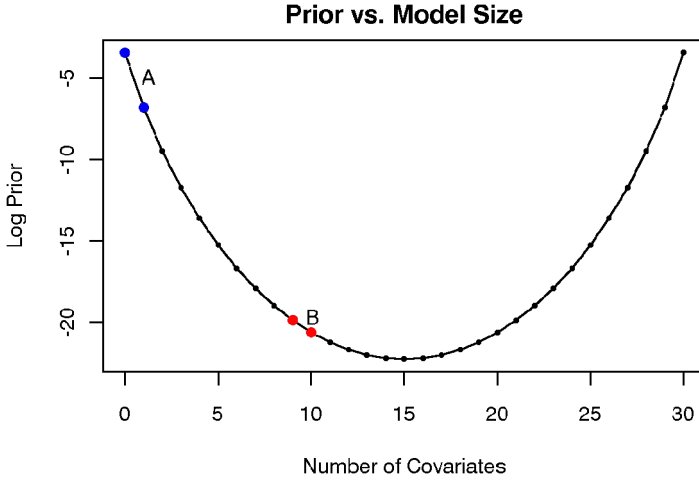


FIG. 1. *Prior model probability versus model size.*

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

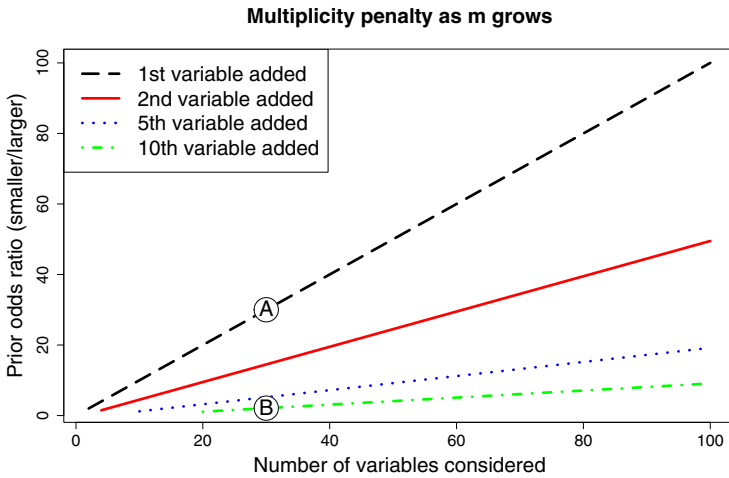


FIG. 2. *Multiplicity penalties as m grows.*

(label A). This penalty is not uniform: 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 "pseudo-objective" choice of $p = 1/2$. Yet because probability is apportioned among models in a very different way, profoundly different behaviors emerge.

For example, Table 1 compares these two regimes on a simulated data set for which the true value of k was fixed at 10. The goal of the study is, in essence, to understand how posterior probabilities adapt to situations of increasingly egregious "data dredging," where a set of true covariates is tested in the presence of an ever-larger group of spurious covariates. We used a simulated $m = 100$ 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 posterior 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 posterior 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 and empirical Bayes procedures both exhibit 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 posterior inclusion probabilities increase as noise variables are added.

TABLE 1

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

Signal	Method and number of noise variables															
	Uncorrected				Fully Bayes				Oracle Bayes				Empirical Bayes			
	1	10	40	90	1	10	40	90	1	10	40	90	1	10	40	90
-1.08	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99
-0.84	99	99	99	99	99	99	99	98	99	99	99	99	99	99	99	99
-0.74	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99
+0.63	99	99	99	99	99	99	92	73	99	99	97	87	99	99	93	80
-0.51	97	97	99	99	91	94	71	34	99	97	85	52	93	95	74	44
+0.41	92	91	99	99	96	86	56	22	99	91	72	35	97	88	60	25
+0.35	77	77	99	99	89	68	30	05	97	77	45	11	91	72	35	07
-0.30	29	28	28	12	55	24	04	00	79	28	06	01	64	25	04	01
+0.18	26	28	24	27	51	25	03	01	79	28	04	01	62	24	04	01
+0.07	21	24	05	01	45	21	03	01	70	24	05	01	56	22	03	01
FPs	0	2	5	10	0	1	0	0	0	2	1	0	0	1	1	0

- 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, empirical Bayes and oracle Bayes answers are all qualitatively (though not quantitatively) 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.

The differences between corrected and uncorrected analyses are quite stark, and calls into question the use of nonadaptive priors in situations with large numbers of potentially spurious covariates. For example, Table 2 shows the posterior 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 ($p = 1/2$), empirical Bayes and fully Bayesian analyses. All variables appear uniformly less impressive when adjusted for multiplicity.

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

TABLE 2
Posterior inclusion probabilities for the important main effects, quadratic effects and cross-product effects for ozone-concentration data under g-priors. Key: $p = 1/2$ implies that all models have equal prior probability; FB is fully Bayes; EB is empirical Bayes

	$p = 1/2$	FB	EB
x1	0.83	0.42	0.54
x2	0.13	0.03	0.05
x3	0.09	0.02	0.03
x4	0.94	0.73	0.84
x5	0.33	0.06	0.10
x6	0.38	0.07	0.10
x7	0.34	0.36	0.29
x8	0.78	0.74	0.77
x9	0.20	0.03	0.05
x10	0.96	0.96	0.97
x1-x1	1.00	0.97	0.99
x9-x9	0.95	0.82	0.91
x1-x2	0.48	0.16	0.24
x4-x7	0.33	0.10	0.15
x6-x8	0.43	0.25	0.34
x7-x8	0.31	0.13	0.18
x7-x10	0.71	0.86	0.85

(1998); for gene-expression studies, see Do, Muller and Tang (2005); for econometrics, see Ley and Steel (2009); for Gaussian graphical models, see Carvalho and Scott (2009); and for time-series data, see Scott (2009).

4. Theoretical comparison of Bayes and empirical Bayes.

4.1. *Motivation.* The previous section showed some examples where fully Bayes and empirical-Bayes methods gave qualitatively similar results. While this rough correspondence between the two approaches does seem to hold in a wide variety of applied problems, we now turn attention to the question of when, and how, it fails.

We begin with 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 $p(M_{\boldsymbol{\gamma}})$ sums to 1 over $\boldsymbol{\gamma}$, the marginal likelihood of p satisfies

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

Furthermore, the inequality is strict under the conditions of the lemma (because the designated marginals are strictly largest), unless the prior assigns $p(M_{\boldsymbol{\gamma}}) = 1$ to the maximizing marginal likelihood. The only way that $p(M_{\boldsymbol{\gamma}}) = p^{k_{\boldsymbol{\gamma}}} \cdot (1 - p)^{m - k_{\boldsymbol{\gamma}}}$ 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.

The above lemma does highlight a specific, undesirable property of the empirical-Bayes approach to variable selection—one whose practical significance we investigate by simulation in Section 6. For the most part, however, the rest of our results are of a fundamentally different character. We will not be evaluating either the fully Bayes or the empirical-Bayes approach according to an objective yardstick, such as how well each one does at recovering true relationships or suppressing false ones. Instead, we focus on comparing the two approaches to each other in a more formal way. As mentioned above, our fundamental goal is to understand when, and how, empirical Bayes corresponds asymptotically to full Bayes analysis. Such a comparison 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 these two approaches, 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 $\boldsymbol{\lambda}$ 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 $\boldsymbol{\lambda}$. (This claim is most interesting when the prior for $\hat{\boldsymbol{\lambda}}$ is unknown; if it is known, then there are also strong frequentist reasons to use this prior in lieu of empirical Bayes.)

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

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

that is, 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, $\pi(\lambda | \mathbf{Y})$ will collapse to a point mass near the true λ as the sample size increases, so that (12) will hold for appropriately smooth functions $\psi(\theta, \mathbf{Y} | \lambda)$ when the sample size is large.

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 nonsmooth $\psi(\theta, \mathbf{Y} | \lambda)$ is no surprise. But what may come as a surprise is that this failure can also occur for very common functions.

Most notably, it fails for the conditional posterior density itself. Indeed, in choosing $\psi(\theta, \mathbf{Y} | \lambda) = \pi(\theta | \lambda, \mathbf{Y})$, the left-hand side of (12) is just the posterior density of θ given \mathbf{Y} , which (by definition) can be written as

$$(13) \quad \pi_F(\theta | \mathbf{Y}) \propto f(\mathbf{Y} | \theta) \int_{\Lambda} \pi(\theta | \lambda) \pi(\lambda) d\lambda.$$

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

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

and the two expressions on the right-hand sides of (13) and (14) can be very different. [This difference may not matter, of course; for instance, if $f(\mathbf{Y} | \theta)$ is extremely concentrated as a likelihood, the prior being used may not matter.]

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

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

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

Of course, these elementary calculations simply lead to (13) after further algebra. But they illuminate the fact that, while $\pi(\lambda | \mathbf{Y})$ may indeed be collapsing to a point mass at the true λ , this term occurs in both the numerator and the denominator of the integrand and therefore cancels. The accuracy with which a point mass at $\hat{\lambda}$ approximates $\pi(\lambda | \mathbf{Y})$ is thus essentially irrelevant from the standpoint of full Bayes analysis.

4.2. *Comparison using Kullback–Leibler convergence.* Our goal, then, is to understand when (13) and (14) will yield the same answers, in an asymptotic sense. The closeness of these 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 :

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

Kullback–Leibler 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 [expectation] under $p(\mathbf{Y}, \boldsymbol{\theta} \mid \boldsymbol{\lambda})$ as $m \rightarrow \infty$, then π_E will be said to be KL-convergent in probability [expectation] 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. The intuition is the following. Suppose that in trying to estimate a given function $\psi(\boldsymbol{\theta})$, it is possible to construct a reasonable prior $\pi(\boldsymbol{\lambda})$ such that the KL-convergence criterion is met. Then the empirical Bayes and full Bayes analysis will disagree for every finite sample size, but are at least tending toward agreement asymptotically. If, on the other hand, 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 dubious from the fully Bayesian perspective. A Bayesian could not replicate such a procedure even asymptotically, while a frequentist may 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:

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

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

4.3. *An orthogonal example.* As a simple illustration of the above ideas, consider the following two examples of empirical-Bayes analysis. The first example satisfies the convergence criterion; the second does not. Both examples concern the same sampling model, in which we observe a series of conditionally independent

random variables $y_i \sim N(\theta_i, 1)$, and where we know that $\theta_i \sim N(\mu, 1)$. Thus, the hyperparameter $\lambda = \mu$ here. Let $\boldsymbol{\theta} = (\theta_1, \dots, \theta_m)$ and $\mathbf{y} = (y_1, \dots, y_m)$.

Alternatively, this can be thought of as an orthogonal regression problem where both the dimension and number of samples are growing at the same rate: $\mathbf{y} = X\boldsymbol{\theta} + \varepsilon$, with X being the $m \times m$ identity matrix. This framing makes the connection to variable selection much more plain.

The natural empirical-Bayes estimate of μ is the sample mean $\hat{\mu}_E = \bar{y}$, which is clearly consistent for μ as $m \rightarrow \infty$ and converges at the usual $1/\sqrt{m}$ rate. A standard hyperprior in a fully Bayesian analysis, on the other hand, would be $\mu \sim N(0, A)$ for some specified A ; the objective hyperprior $\pi(\mu) = 1$ is essentially the limit of this as $A \rightarrow \infty$. Using the expressions given in, for example, Berger (1985), the empirical-Bayes and full Bayes posteriors are

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

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

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 the target function $\psi(\boldsymbol{\theta}) = \theta_1$. Then sending $A \rightarrow \infty$ yields

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

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

It is easy to check that $\text{KL}(\pi_F \parallel \pi_E) \rightarrow 0$ as $m \rightarrow \infty$. 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), with parameters $(\hat{\boldsymbol{\theta}}_E, \Sigma_E)$ and $(\hat{\boldsymbol{\theta}}_F, \Sigma_F)$, respectively.

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

$$(22) \quad \begin{aligned} \text{KL} &= \frac{1}{2} \left[\log \left(\frac{\det \Sigma_E}{\det \Sigma_F} \right) + \text{tr}(\Sigma_E^{-1} \Sigma_F) + (\hat{\boldsymbol{\theta}}_E - \hat{\boldsymbol{\theta}}_F)^t \Sigma_E^{-1} (\hat{\boldsymbol{\theta}}_E - \hat{\boldsymbol{\theta}}_F) - m \right] \\ &= \frac{1}{2} \left[-\log \left(1 + \frac{mA}{mA + 2} \right) + \frac{mA}{mA + 2} + 2m \left(\frac{1}{mA + 2} \right)^2 \bar{y}^2 \right]. \end{aligned}$$

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 $[2m/(mA + 2)^2] \cdot \bar{y}^2 \rightarrow 0$ in probability as $m \rightarrow \infty$. Hence, for any value of A (including $A = \infty$), the KL divergence in (22) converges to $(1 - \log 2)/2 > 0$.

Of course, this only considers priors of the form $\mu \sim N(0, A)$, but the asymptotic normality of the posterior for μ 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 $\hat{\boldsymbol{\theta}}_E$ and $\hat{\boldsymbol{\theta}}_F$ are getting closer to each other elementwise, the KL divergence does not shrink to 0 as expected.

Two further comments are in order. First, a similar argument shows that the fully Bayes posterior is not KL-convergent to the so-called “oracle posterior” $\pi(\boldsymbol{\theta} \mid \mathbf{y}, \mu_T)$ —that is, the conditional posterior distribution for $\boldsymbol{\theta}$, given the true value of μ . This is not a source of worry for Bayesians, but it makes clear that the disagreement between EB and FB procedures cuts both ways, and is not merely a “failure” of empirical-Bayes; if a non-Bayesian’s goal is to reconstruct the oracle posterior, this could be achieved by empirical-Bayes analysis but not by full Bayes.

Second, the situation described above has the sample size n equal to the number of unknown parameters m . If n grows relative to m , the full Bayes and empirical-Bayes/oracle posteriors can indeed be KL-convergent. For instance, suppose there are r independent replicate observations for each μ_i . Then a similar calculation shows that $\text{KL}(\pi_F \parallel \pi_E) = O(1/r)$ as $r \rightarrow \infty$, so that KL convergence between the two approaches would obtain.

5. 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 also quite difficult to characterize the sampling distribution of \hat{p} , the empirical-Bayes estimate for the prior inclusion probability p . It is therefore not yet possible to give a general characterization of whether, and when, the empirical-Bayes variable-selection procedure is KL-convergent, in the sense defined above, to a fully Bayesian procedure.

Three interesting sets of results are available, however. First and most simply, we can characterize the KL divergence between the prior probability distributions of the fully Bayesian and empirical-Bayesian procedures. Second, we can characterize the limiting *expected* Kullback–Leibler divergence between EB and FB posteriors, even if we cannot characterize the limiting KL divergence itself. Third, we can compare the asymptotic behavior of the full Bayes and empirical-Bayes prior model probabilities for models in a size neighborhood of the true model.

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

5.1. *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 5.1) and Hellinger (Theorem 5.2) flavors.

THEOREM 5.1. *Let $\underline{G}(m) = \min_{\hat{p}} \text{KL}(p_F(M_{\mathcal{Y}}) \parallel p_E(M_{\mathcal{Y}}))$. Then $\underline{G}(m) \rightarrow \infty$ as $m \rightarrow \infty$.*

PROOF. The KL divergence is

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

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

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

The first (linear) term in (24) 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, $\underline{G}(m)$ grows linearly with m , with asymptotic positive slope of $\log 2 - 1/2$. \square

THEOREM 5.2. *Let $\underline{H}^2(m) = \min_{\hat{p}} H^2(p_F(M_{\mathcal{Y}}) \parallel p_E(M_{\mathcal{Y}}))$. Then $\underline{H}^2(m) \rightarrow 1$ as $m \rightarrow \infty$.*

PROOF.

$$(25) \quad H^2(p_F(M_{\mathcal{Y}}) \parallel p_E(M_{\mathcal{Y}})) = 1 - \frac{1}{\sqrt{m+1}} \sum_{k=0}^m \sqrt{\binom{m}{k} \hat{p}^k (1-\hat{p})^{m-k}}.$$

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

$$(26) \quad \underline{H}^2(m) = 1 - (m+1)^{-1/2} \cdot 2^{-m/2} \cdot \sum_{k=0}^m \sqrt{\binom{m}{k}}.$$

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

$$(27) \quad \lim_{m \rightarrow \infty} \left[(m + 1)^{-1/2} \cdot 2^{-m/2} \cdot \sum_{k=0}^m \sqrt{\binom{m}{k}} \right] = 0,$$

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 $p_E(M_{\mathcal{Y}})$ becomes arbitrarily bad as an approximation to $p_F(M_{\mathcal{Y}})$. (Here, the correction under the empirical-Bayes approach will be more aggressive compared with the Bayesian approach, and some may consider this additional aggressiveness to be a source of strength.) 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 $p_F(M_{\mathcal{Y}})$ and $p_E(M_{\mathcal{Y}})$ grows linearly without bound (for KL divergence), or converges to 1 (for Hellinger distance).

5.2. *Posterior KL divergence.* We now prove a theorem showing that, under very mild conditions, the expected KL divergence between FB and EB posteriors for the variable-selection problem is infinite. This version assumes that the error precision ϕ is fixed, but the generalization to an unknown ϕ is straightforward.

THEOREM 5.3. *In the variable-selection problem, let $m, n > m$, and $\phi > 0$ be fixed. Suppose $\mathbf{X}_{\mathcal{Y}}$ is of full rank for all models and that the family of priors for model-specific parameters, $\{\pi(\boldsymbol{\beta}_{\mathcal{Y}})\}$, is such that $p(\boldsymbol{\beta}_{\mathcal{Y}} = \mathbf{0}) < 1$ for all $M_{\mathcal{Y}}$. Then, for any true model $M_{\mathcal{Y}}^T$, the expected posterior KL divergence $E[\text{KL}\{p_F(M_{\mathcal{Y}} | \mathbf{Y}) \parallel p_E(M_{\mathcal{Y}} | \mathbf{Y})\}]$ under this true model is infinite.*

PROOF. The posterior KL divergence is

$$(28) \quad \text{KL}(p_F(M_{\mathcal{Y}} | \mathbf{Y}) \parallel p_E(M_{\mathcal{Y}} | \mathbf{Y})) = \sum_{\Gamma} p_F(M_{\mathcal{Y}} | \mathbf{Y}) \cdot \log \left(\frac{p_F(M_{\mathcal{Y}} | \mathbf{Y})}{p_E(M_{\mathcal{Y}} | \mathbf{Y})} \right).$$

This is clearly infinite if there exists a model $M_{\mathcal{Y}}$ for which $p_E(M_{\mathcal{Y}} | \mathbf{Y}) = 0$ but $p_F(M_{\mathcal{Y}} | \mathbf{Y}) > 0$. 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

$$(29) \quad f(\mathbf{Y} \mid M_{\gamma}) = L(\hat{\beta}_{\gamma}^*) \cdot \sqrt{2\pi/n} \int_{\mathbb{R}^{k_{\gamma}}} g(\beta_{\gamma})\pi(\beta_{\gamma}) d\beta_{\gamma},$$

where

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

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

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

which from (29) is clearly continuous as a function of \mathbf{Y} for every γ . Evaluated at $\mathbf{Y} = \mathbf{0}$ (so that $\hat{\beta}_{\gamma}$ then equals 0), this Bayes factor satisfies

$$(30) \quad B_{\gamma}(\mathbf{0}) = \left(\int_{\mathbb{R}^{k_{\gamma}}} \exp\left\{-\frac{1}{2}\beta_{\gamma}^t \mathbf{X}_{\gamma}^t \mathbf{X}_{\gamma} \beta_{\gamma}\right\} \pi(\beta_{\gamma}) d\beta_{\gamma} \right)^{-1} > 1$$

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 $|\mathbf{Y}| < \varepsilon_{\gamma}$. Let $\varepsilon^* = \min_{\gamma} \varepsilon_{\gamma}$. Then for \mathbf{Y} satisfying $|\mathbf{Y}| < \varepsilon^*$, $B_{\gamma}(\mathbf{Y}) > 1$ for all nonnull 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 $|\mathbf{Y}| < \varepsilon^*$ 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 $\text{KL}(p_F(M_{\gamma} \mid \mathbf{Y}) \parallel p_E(M_{\gamma} \mid \mathbf{Y}))$ is infinite under the sampling distribution $p(\mathbf{Y} \mid M_{\gamma})$, 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 $E(\text{KL})$ does not converge to 0 as $m \rightarrow \infty$. This, of course, is a weaker conclusion than would be a lack of KL convergence in probability.

In Theorem 5.3 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 .

5.3. *Asymptotic behavior of prior model probabilities under EB and FB procedures.* While interesting, the results in the previous two sections do not consider the usual type of asymptotic comparison, namely, how do the full Bayes and empirical Bayes posterior distributions converge as $m \rightarrow \infty$? It is not clear that such asymptotic comparisons are possible in general, although very interesting results can be obtained in particular contexts [cf. Bogdan, Chakrabarti and Ghosh (2008), Bogdan, Ghosh and Tokdar (2008)].

A rather general insight related to such comparison can be obtained, however, by focusing on the prior probabilities of “high posterior” models, as $m \rightarrow \infty$. To do so, we first need an approximation to the full Bayes prior probability of $M_{\mathcal{Y}}$, given in the following lemma. The proof is straightforward Laplace approximation, and is omitted.

LEMMA 5.4. *As $m \rightarrow \infty$, consider models of size $k_{\mathcal{Y}}$ such that $k_{\mathcal{Y}}/m$ is bounded away from 0 and 1. Then the Bayesian prior probability of $M_{\mathcal{Y}}$ with prior $\pi(p)$ is*

$$\begin{aligned}
 p_F(M_{\mathcal{Y}}) &= \int_0^1 p(M_{\mathcal{Y}} | p)\pi(p) dp \\
 &= \left(\frac{k_{\mathcal{Y}}}{m}\right)^{k_{\mathcal{Y}}} \left(1 - \frac{k_{\mathcal{Y}}}{m}\right)^{m-k_{\mathcal{Y}}} \left[\frac{(2\pi)(k_{\mathcal{Y}}/m)(1 - k_{\mathcal{Y}}/m)\pi(k_{\mathcal{Y}}/m)}{m}\right]^{1/2} \\
 &\quad \times \{1 + o(1)\},
 \end{aligned}$$

providing $\pi(\cdot)$ is continuous and nonzero.

Now suppose p_T is the true prior variable inclusion probability and consider the most favorable situation for empirical Bayes analysis, in which the empirical Bayes estimate for p_T satisfies

$$(31) \quad \hat{p} = p_T(1 + \varepsilon_E) \quad \text{where } \varepsilon_E \text{ is } O\left(\frac{1}{\sqrt{m}}\right) \text{ as } m \rightarrow \infty.$$

It is not known in general when this holds, but it does hold in exchangeable contexts where each variable is in or out of the model with unknown probability p , since such problems are equivalent to mixture model problems.

For models far from the true model, the prior model probabilities given by the Bayesian and empirical Bayesian approaches can be extremely different. Hence, it is most interesting to focus on models that are close to the true model for the comparison. In particular, we restrict attention to models whose size differs from the true model by $O(\sqrt{m})$.

THEOREM 5.5. *Suppose the true model size k_T satisfies $k_T/m = p_T + O(1/\sqrt{m})$ as $m \rightarrow \infty$, where $0 < p_T < 1$. Consider all models $M_{\mathcal{Y}}$ such that $k_T - k_{\mathcal{Y}} = O(\sqrt{m})$, and consider the optimal situation for EB in which (31) holds.*

Then the ratio of the prior probabilities assigned to such models by the Bayes approach and the empirical Bayes approach satisfies

$$\begin{aligned} \frac{p_F(M_{\mathbf{y}})}{p_E(M_{\mathbf{y}})} &= \left(\frac{k_{\mathbf{y}}}{m}\right)^{k_{\mathbf{y}}} \left(1 - \frac{k_{\mathbf{y}}}{m}\right)^{m-k_{\mathbf{y}}} \left[(2\pi) \left(\frac{k_{\mathbf{y}}}{m}\right) \left(1 - \frac{k_{\mathbf{y}}}{m}\right) \pi\left(\frac{k_{\mathbf{y}}}{m}\right) \right]^{1/2} \\ &\quad \times m^{-1/2} \{1 + o(1)\} \\ &\quad \times ((\hat{p})^{k_{\mathbf{y}}} (1 - \hat{p})^{m-k_{\mathbf{y}}})^{-1} \\ &= O\left(\frac{1}{\sqrt{m}}\right), \end{aligned}$$

providing $\pi(\cdot)$ is continuous and nonzero.

PROOF. Note that

$$(32) \quad p_E(M_{\mathbf{y}}) = (\hat{p})^{k_{\mathbf{y}}} (1 - \hat{p})^{m-k_{\mathbf{y}}} = \{p_T(1 + \varepsilon_E)\}^{k_{\mathbf{y}}} \{1 - p_T(1 + \varepsilon_E)\}^{m-k_{\mathbf{y}}}.$$

Taking the log and performing a Taylor expansion yields

$$\begin{aligned} \log p_E(M_{\mathbf{y}}) &= \log\{p_T^{k_{\mathbf{y}}} (1 - p_T)^{m-k_{\mathbf{y}}}\} + k_{\mathbf{y}} \log(1 + \varepsilon_E) \\ &\quad + (m - k_{\mathbf{y}}) \log\left\{1 - \frac{p_T}{(1 - p_T)} \varepsilon_E\right\} \\ &= \log\{p_T^{k_{\mathbf{y}}} (1 - p_T)^{m-k_{\mathbf{y}}}\} + k_{\mathbf{y}} \{\varepsilon_E + O(\varepsilon_E^2)\} \\ &\quad + (m - k_{\mathbf{y}}) \left\{-\frac{p_T}{(1 - p_T)} \varepsilon_E + O(\varepsilon_E^2)\right\} \\ &= \log\{p_T^{k_{\mathbf{y}}} (1 - p_T)^{m-k_{\mathbf{y}}}\} + \{k_T + O(\sqrt{m})\} \{\varepsilon_E + O(\varepsilon_E^2)\} \\ &\quad + \{m - k_T - O(\sqrt{m})\} \left\{-\frac{p_T}{(1 - p_T)} \varepsilon_E + O(\varepsilon_E^2)\right\} \\ &= \log\{p_T^{k_{\mathbf{y}}} (1 - p_T)^{m-k_{\mathbf{y}}}\} + O(\sqrt{m}\varepsilon_E) + O(m\varepsilon_E^2) \\ &= \log\{p_T^{k_{\mathbf{y}}} (1 - p_T)^{m-k_{\mathbf{y}}}\} + O(1). \end{aligned}$$

A nearly identical argument using Lemma 5.4 shows that the log Bayesian prior probability for these models is

$$(33) \quad \log\{p_F(M_{\mathbf{y}})\} = \log\{p_T^{k_{\mathbf{y}}} (1 - p_T)^{m-k_{\mathbf{y}}}\} - \log\sqrt{m} + O(1),$$

from which the result is immediate. \square

So we see that, even under the most favorable situation for the empirical-Bayes analysis, and even when only considering models that are close to the true model in terms of model size, the prior probabilities assigned by the Bayes approach are

smaller by a factor of order $1/\sqrt{m}$ than those assigned by the empirical-Bayes approach. The effect of this very significant difference in prior probabilities will be context dependent, but the result does provide a clear warning that the full Bayes and empirical-Bayes answers can differ—even when $m \rightarrow \infty$ and even when there is sufficient information in the data to guarantee the existence of a consistent estimator for p_T .

The theorem also shows that the empirical-Bayes procedure provides a better asymptotic approximation to the “oracle” prior probabilities, which may be argued by some to be the main goal of empirical-Bayes analysis. At least for this ideal scenario, the EB approach assigns larger prior probabilities to models which are closer to the true model. Of course, this fact is not especially relevant from the fully Bayesian perspective, and does not necessarily counterbalance the problems associated with ignoring uncertainty in the estimator for p .

Finally, this difference in prior probabilities will not always have a large effect. For instance, if $n \rightarrow \infty$ at a fast enough rate compared with m , then the Bayes and empirical-Bayes approach will typically agree simply because all of the posterior mass will concentrate on a single model [i.e., one of the marginal likelihoods $f(\mathbf{Y} | M_\gamma)$ will become dominant], and so the assigned prior probabilities will be irrelevant.

6. Numerical investigation of empirical-Bayes variable selection. This section presents numerical results that demonstrate practical, finite-sample significance of some of the qualitative differences mentioned above. As in the previous section, most of the investigation is phrased as a comparison of empirical-Bayes and full Bayes, taken from the fully Bayesian perspective.

Note that m here is taken to be moderate (14 for the simulation study and 22 for the real data set); the intent is to focus on the magnitude of the difference that one can expect in variable selection problems of such typical magnitude. Of course, such m are not large enough that one would automatically expect the empirical-Bayes approach to provide an accurate estimate of p , and so differences are to be expected, but it is still useful to see the magnitude of the differences. For a larger m situation, see Table 1; for the largest m in that table, the full Bayes and empirical-Bayes answers are much closer. The rationale for taking these values of m is that they allow the model space to be enumerated, avoiding potential confounding effects due to computational difficulties.

6.1. *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 $\boldsymbol{\gamma}$ encoding the true set of regressors.

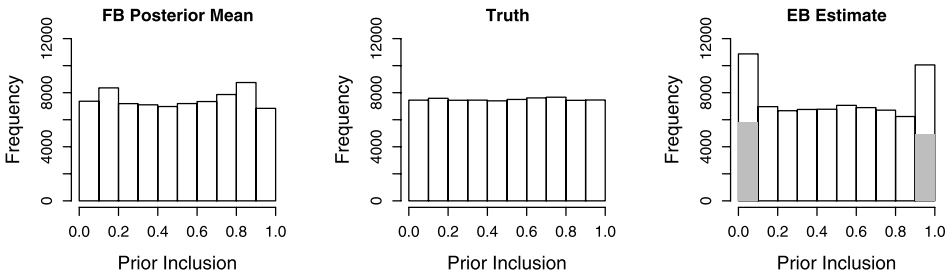


FIG. 3. Distribution of \hat{p} in the simulation study ($n = 60$) with a correctly specified (uniform) prior for p . The gray bars indicated the number of times, among values of \hat{p} in the extremal bins, that the empirical-Bayes solution collapsed to the degenerate $\hat{p} = 0$ or $\hat{p} = 1$.

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.

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.

Two broad patterns emerged from these experiments.

First, as Figure 3 shows, the EB procedure gives the degenerate $\hat{p} = 0$ or $\hat{p} = 1$ solution much too often. When $n = 60$, for example, almost 15% of cases collapsed to $\hat{p} = 0$ or $\hat{p} = 1$. This is essentially the same fraction of degenerate cases as when $n = 16$, which was 16%. This suggests that the issues raised by Theorem 5.3 can be quite serious in practice, even when n is large compared to m .

Second, even in nondegenerate situations, the two procedures often reached very different conclusions about which covariates were important. Figure 4 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. (Note that Figure 4 only depicts the differences that arise when the empirical-Bayes solution does not collapse to either 0 or 1.)

6.2. *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. There was an obvious bias, however, in that the fully Bayesian procedure was being evaluated under its true prior distribution, with respect to which it is necessarily optimal.

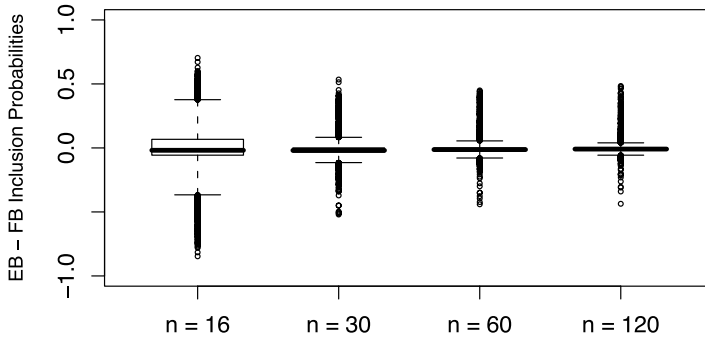


FIG. 4. Differences in all m inclusion probabilities between EB and FB analyses across all non-degenerate cases (i.e., where the EB solution does not collapse to the boundary). The percentage of points lying outside the boxplot whiskers (1.5 times the inter-quartile range) are as follows: 14% for $n = 16$, 12% for $n = 30$, 8% for $n = 60$ and 7% for $n = 120$.

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 nonuniform distribution. We limit ourselves to discussion of the analogue of Figure 3 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.

The results are summarized in Figure 5. 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 gray 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 5 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

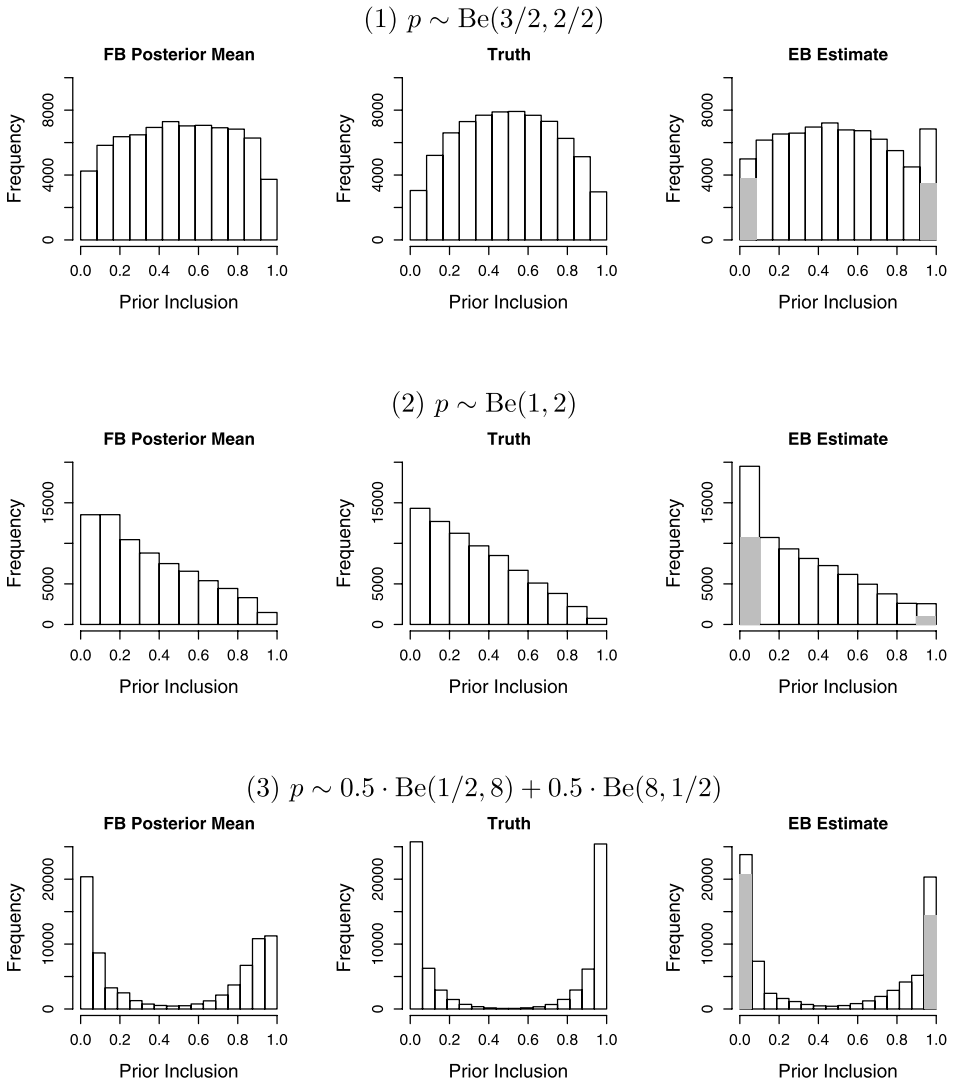


FIG. 5. Distribution of \hat{p} ($n = 60$) in different versions of the simulation study, where the fully Bayesian model had a misspecified (uniform) prior on p . The gray bars indicated the number of times, among values of \hat{p} in the extremal bins, that the empirical-Bayes solution collapsed to the degenerate $\hat{p} = 0$ or $\hat{p} = 1$.

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.

6.3. *Results when p is fixed.* We conducted a final version of the simulation with p fixed at 3 different values: $p = 0.10$, $p = 0.25$, and $p = 0.5$. Figure 6 plots the estimated values of p under the fully Bayes and empirical-Bayes procedures. (For the sake of visual clarity only the results from 2000 data sets are shown.)

It is clear that for the smallest value of $p = 0.1$, the degenerate solution $\hat{p} = 0$ occurs quite frequently. When p is moderate (as in the 0.25 or 0.5 cases), degeneracy occurs much less often.

It is also interesting to see the differences in how well the EB and FB analysis approximate the “oracle” inclusion probabilities, which are the posterior inclusion probabilities one would compute if one knew the true Bernoulli probability p . This can be measured by looking at the ℓ_1 distance from the oracle estimate:

$$\ell_1(\hat{\mathbf{p}}, \hat{\mathbf{p}}^{\text{or}}) = \sum_{j=1}^m |\hat{p}_j - \hat{p}_j^{\text{or}}|,$$

where \hat{p}_j^{or} is the oracle posterior inclusion probability for the j th variable.

The two procedures do quite similarly here, but with subtle differences. For example, on the “sparse” ($p = 0.1$) case, the mean ℓ_1 distance to the oracle answer across all Monte Carlo draws was 0.36 for the EB posterior, and 0.40 for the FB posterior. Yet the *median* ℓ_1 distance to the oracle answer was 0.27 for the FB posterior, and 0.30 for the EB posterior.

These differences were largely consistent across other values of p . This suggests that, while the FB procedure seems to reconstruct the oracle posterior inclusion probabilities better for a larger number of data sets (such as when the empirical-Bayes answer is degenerate), it tends to miss by a larger amount than the EB procedure does. This results in a worse level of average performance for the FB procedure in reconstructing the oracle posterior inclusion probabilities.

6.4. *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, Ley and Steel (2001) popularized the use of Bayesian model averaging in the field; Sala-i Martin, Doppelhofer and Miller (2004) used a Bayes-like procedure called BACE, similar to

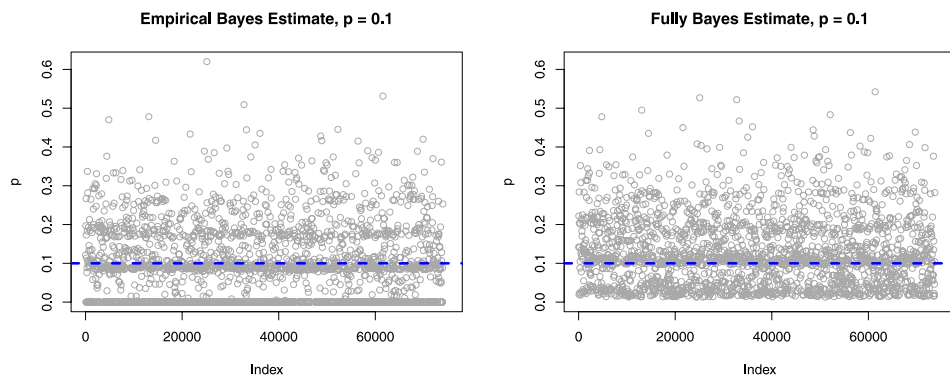
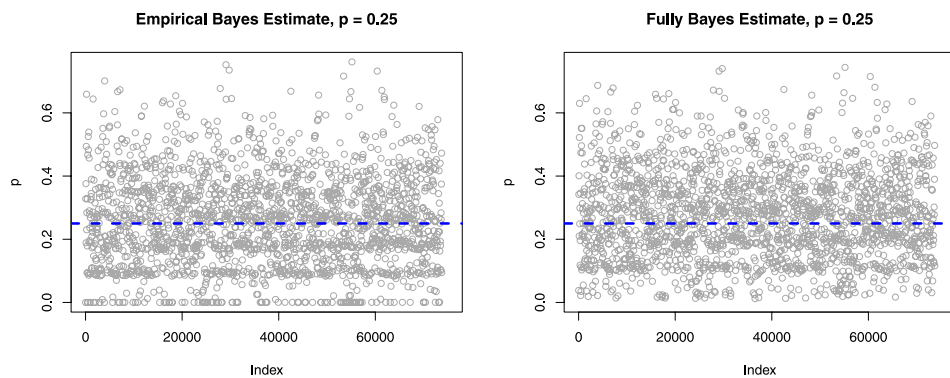
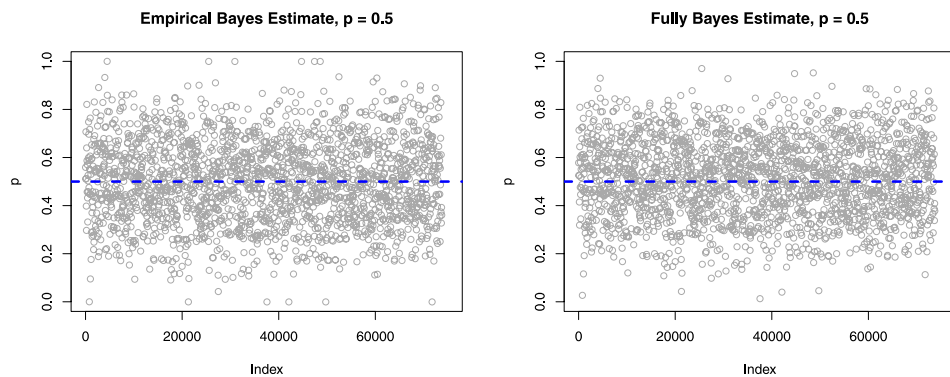
(1) $p = 0.1$ (2) $p = 0.25$ (3) $p = 0.5$ 

FIG. 6. Distribution of \hat{p} ($n = 60$) in the fixed- p versions of the simulation study (2000 subsamples of the fake data sets). The dashed line indicates the true value of p .

BIC-weighted OLS estimates, for selecting a model; and [Ley and Steel \(2009\)](#) 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, Doppelhofer and Miller \(2004\)](#) 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

TABLE 3
Exact inclusion probabilities for 22 variables in a linear model for GDP growth among a group of 30 countries

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 1960–1990	0.164	0.095
Defense spending share	0.156	0.085
Civil liberties	0.154	0.075
Average inflation 1960–1990	0.150	0.064
Real exchange rate distortions	0.146	0.071
Interior density	0.139	0.067

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.

7. 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 prefixed 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](#).

APPENDIX: 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:

$$(\boldsymbol{\beta} \mid \phi) \sim N\left(\boldsymbol{\beta}_0, \frac{g}{\phi}(\mathbf{X}'\mathbf{X})^{-1}\right),$$

$$\phi \sim \text{Ga}\left(\frac{\nu}{2}, \frac{\nu S}{2}\right)$$

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, Pericchi and Varshavsky (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 nonnested 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 :

$$(34) \quad \text{BF}(M_1 : M_2) = \frac{\text{BF}(M_1 : M_B)}{\text{BF}(M_2 : M_B)}.$$

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:

$$(\alpha, \phi \mid \boldsymbol{y}) \propto 1/\phi,$$

$$(\boldsymbol{\beta}_{\boldsymbol{y}} \mid \phi, \boldsymbol{y}) \sim N\left(0, \frac{g}{\phi} (\mathbf{X}'_{\boldsymbol{y}} \mathbf{X}_{\boldsymbol{y}})^{-1}\right).$$

This gives a simple expression for the Bayes factor for evaluating a model \boldsymbol{y} with k regression parameters (excluding the intercept):

$$(35) \quad \text{BF}(M_{\boldsymbol{y}} : M_0) = (1 + g)^{(n-k_{\boldsymbol{y}}-1)/2} [1 + (1 - R_{\boldsymbol{y}}^2)g]^{-(n-1)/2},$$

where $R_{\boldsymbol{y}}^2 \in (0, 1]$ is the usual coefficient of determination for model $M_{\boldsymbol{y}}$.

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

$$M_F : \mathbf{Y} = \mathbf{X}_{\boldsymbol{y}}^* \boldsymbol{\theta}_{\boldsymbol{y}} + \mathbf{X}_{-\boldsymbol{y}} \boldsymbol{\beta}_{-\boldsymbol{y}},$$

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:

$$(\alpha, \beta_{\mathcal{Y}}, \phi \mid \mathcal{Y}) \propto 1/\phi,$$

$$(\beta_{-\mathcal{Y}} \mid \phi, \mathcal{Y}) \sim N\left(0, \frac{g}{\phi}(\mathbf{X}'_{-\mathcal{Y}}\mathbf{X}_{-\mathcal{Y}})^{-1}\right).$$

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 sub-model is considered. Nevertheless, marginal likelihoods can still be consistently defined in the manner of (34). Conditional upon g , this yields a Bayes factor in favor of the full model of

$$(36) \quad \text{BF}(M_F : M_{\mathcal{Y}}) = (1 + g)^{(n-m-1)/2} (1 + gW)^{-(n-k-1)/2},$$

where $W = (1 - R_F^2)/(1 - R_{\mathcal{Y}}^2)$.

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 (35): the Bayes factor in favor of $M_{\mathcal{Y}}$ goes to the finite constant $(1 + g)^{n-m-1}$ as $R_{\mathcal{Y}}^2 \rightarrow 1$ (which can only happen if $M_{\mathcal{Y}}$ 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 $\int_0^\infty (1 + g)^{n-k_{\mathcal{Y}}-1} \pi(g) dg = \infty$ for all $k_{\mathcal{Y}} \leq p$, 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 nonshared parameters:

$$(37) \quad (\beta_{\mathcal{Y}} \mid \phi) \sim C\left(0, \frac{n}{\phi}(\mathbf{X}'_{\mathcal{Y}}\mathbf{X}_{\mathcal{Y}})^{-1}\right).$$

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 $M_{\boldsymbol{y}}$ then takes the form

$$(38) \quad \text{BF}(M_{\boldsymbol{y}} : M_0) = \int_0^\infty (1+g)^{(n-k_{\boldsymbol{y}}-1)/2} [1+(1-R_{\boldsymbol{y}}^2)g]^{-(n-1)/2} \\ \times g^{-3/2} \exp(-n/(2g)) dg.$$

A similar formula exists for the full-based version:

$$(39) \quad \text{BF}(M_F : M_{\boldsymbol{y}}) = \int_0^\infty (1+g)^{(n-m-1)/2} [1+Wg]^{-(n-k-1)/2} \\ \times g^{-3/2} \exp(-n/(2g)) dg$$

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.

REFERENCES

- BARBIERI, M. and BERGER, J. O. (2004). Optimal predictive model selection. *Ann. Statist.* **32** 870–897. [MR2065192](#)
- BERGER, J., PERICCHI, L. and VARSHAVSKY, J. (1998). Bayes factors and marginal distributions in invariant situations. *Sankhyā Ser. A* **60** 307–321. [MR1718789](#)
- BERGER, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. Springer, New York. [MR0804611](#)
- BERGER, J. O. and MOLINA, G. (2005). Posterior model probabilities via path-based pairwise priors. *Statist. Neerlandica* **59** 3–15. [MR2137378](#)
- BERRY, D. (1988). Multiple comparisons, multiple tests, and data dredging: A Bayesian perspective. In *Bayesian Statistics 3* (J. Bernardo, M. DeGroot, D. Lindley and A. Smith, eds.) 79–94. Oxford Univ. Press, New York. [MR1008045](#)
- BERRY, D. and HOCHBERG, Y. (1999). Bayesian perspectives on multiple comparisons. *J. Statist. Plann. Inference* **82** 215–277. [MR1736444](#)
- BOGDAN, M., GHOSH, J. K. and ZAK-SZATKOWSKA, M. (2008). Selecting explanatory variables with the modified version of the Bayesian information criterion. *Quality and Reliability Engineering International* **24** 627–641.
- BOGDAN, M., CHAKRABARTI, A. and GHOSH, J. K. (2008). Optimal rules for multiple testing and sparse multiple regression. Technical Report I-18/08/P-003, Wrocław Univ. Technology.
- BOGDAN, M., GHOSH, J. K. and TOKDAR, S. T. (2008). 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* 211–230. IMS, Beachwood, OH. [MR2462208](#)

- CARLIN, B. and LOUIS, T. (2000). Empirical Bayes: Past, present and future. *J. Amer. Statist. Assoc.* **95** 1286–1289. [MR1825277](#)
- CARVALHO, C. M. and SCOTT, J. G. (2009). Objective Bayesian model selection in Gaussian graphical models. *Biometrika* **96** 497–512.
- CASELLA, G. and MORENO, E. (2002). Objective Bayes variable selection. Technical Report 023, Univ. Florida.
- CUI, W. and GEORGE, E. I. (2008). Empirical Bayes vs. fully Bayes variable selection. *J. Statist. Plann. Inference* **138** 888–900. [MR2416869](#)
- DO, K.-A., MULLER, P. and TANG, F. (2005). A Bayesian mixture model for differential gene expression. *J. Roy. Statist. Soc. Ser. C* **54** 627–644. [MR2137258](#)
- EATON, M. (1989). *Group Invariance Applications in Statistics*. IMS, Hayward, CA.
- EFRON, B., TIBSHIRANI, R., STOREY, J. and TUSHER, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.* **96** 1151–1160. [MR1946571](#)
- FERNANDEZ, C., LEY, E. and STEEL, M. (2001). Model uncertainty in cross-country growth regressions. *J. Appl. Econometrics* **16** 563–576.
- GEORGE, E. I. and FOSTER, D. P. (2000). Calibration and empirical Bayes variable selection. *Biometrika* **87** 731–747. [MR1813972](#)
- GOPALAN, R. and BERRY, D. (1998). Bayesian multiple comparisons using Dirichlet process priors. *J. Amer. Statist. Assoc.* **93** 1130–1139. [MR1649207](#)
- GOULD, H. (1964). Sums of logarithms of binomial coefficients. *Amer. Math. Monthly* **71** 55–58. [MR1532480](#)
- JEFFERYS, W. and BERGER, J. (1992). Ockham's razor and Bayesian analysis. *American Scientist* **80** 64–72.
- JEFFREYS, H. (1961). *Theory of Probability*, 3rd ed. Clarendon Press, Oxford. [MR0187257](#)
- JOHNSTONE, I. and SILVERMAN, B. W. (2004). Needles and straw in haystacks: Empirical–Bayes estimates of possibly sparse sequences. *Ann. Statist.* **32** 1594–1649. [MR2089135](#)
- LEY, E. and STEEL, M. F. (2009). On the effect of prior assumptions in Bayesian model averaging with applications to growth regression. *J. Appl. Econometrics* **24** 651–674.
- LIANG, F., PAULO, R., MOLINA, G., CLYDE, M. and BERGER, J. (2008). Mixtures of g -priors for Bayesian variable selection. *J. Amer. Statist. Assoc.* **103** 410–423. [MR2420243](#)
- MENG, C. and DEMPSTER, A. (1987). A Bayesian approach to the multiplicity problem for significance testing with binomial data. *Biometrics* **43** 301–311. [MR0897406](#)
- SALA-I MARTIN, X., DOPPELHOFER, G. and MILLER, R. I. (2004). Determinants of long-term growth: A Bayesian averaging of classical estimates (bace) approach. *American Economic Review* **94** 813–835.
- SCOTT, J. G. (2009). Nonparametric Bayesian multiple testing for longitudinal performance stratification. *Ann. Appl. Statist.* **3** 1655–1674.
- SCOTT, J. G. and BERGER, J. O. (2006). An exploration of aspects of Bayesian multiple testing. *J. Statist. Plann. Inference* **136** 2144–2162. [MR2235051](#)
- SCOTT, J. G. and CARVALHO, C. M. (2008). Feature-inclusion stochastic search for Gaussian graphical models. *J. Comput. Graph. Statist.* **17** 790–808.
- WALLER, R. and DUNCAN, D. (1969). A Bayes rule for the symmetric multiple comparison problem. *J. Amer. Statist. Assoc.* **64** 1484–1503. [MR0362749](#)
- WESTFALL, P. H., JOHNSON, W. O. and UTTS, J. M. (1997). A Bayesian perspective on the Bonferroni adjustment. *Biometrika* **84** 419–427. [MR1467057](#)
- 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* (P. Goel and A. Zellner, eds.) 233–243. North-Holland, Amsterdam. [MR0881437](#)

ZELLNER, A. and STOW, A. (1980). Posterior odds ratios for selected regression hypotheses. In *Bayesian Statistics: Proceedings of the First International Meeting held in Valencia (Spain)* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 585–603. Univ. Press, Valencia.

DEPARTMENT OF STATISTICS
UNIVERSITY OF TEXAS AT AUSTIN
1 UNIVERSITY STATION, B6500
AUSTIN, TEXAS 78712
USA
E-MAIL: james.scott@mcombs.utexas.edu

DEPARTMENT OF STATISTICS
DUKE UNIVERSITY
BOX 90251
DURHAM, NORTH CAROLINA 27708
USA
E-MAIL: berger@stat.duke.edu