Multiple Testing and the Variable-Selection Problem

BY JAMES G. SCOTT

Department of Statistical Science, Duke University,
Durham, North Carolina 27708-0251, U.S.A.
james@stat.duke.edu

AND JAMES O. BERGER

Department of Statistical Science, Duke University,
Durham, North Carolina 27708-0251, U.S.A.
berger@stat.duke.edu

SUMMARY
We study the multiplicity-correction effect of standard Bayesian variable-selection priors in linear regression. Specifically, we compare empirical-Bayes (EB) and fully Bayesian (FB) approaches for handling the prior inclusion probability $p$ required by these priors. Several new information-theoretic results, along with extensive computer experiments, lead us to conclude that the empirical-Bayes approach to variable selection has serious flaws as a multiple-testing procedure. We also compare EB and FB results on a real data set from the literature on GDP-growth regressions, showing that these flaws can have significant practical impact.

Some key words: linear regression; empirical Bayes; multiple testing; Bayesian model selection
Consider the usual variable-selection problem: upon observing responses \( Y \) along with an \( n \times m \) matrix of covariates \( X \) upon whose columns \( Y \) may depend, we wish to select a linear model \( Y = X_\gamma \beta_\gamma + \epsilon \), where \( \gamma \) is a binary vector indicating a subset of variables included in the model, and where \( \epsilon_i \sim N(0, \sigma^2) \).

In Bayesian model selection, the model is itself a random variable that takes values in the discrete space \( \Gamma = \{0, 1\}^m \). Inference relies upon the prior probability assigned to each model, \( \pi(\gamma) \), along with the marginal likelihood of the data under the model, \( f(Y | \gamma) \), which together specify the posterior probability \( \pi(\gamma | Y) \).

We are concerned with the multiple-testing problem implied by variable selection, where a single null hypothesis claims that a specific variable ought to be excluded from the model. The advantages of a Bayesian approach to multiplicity correction include its lack of ad-hoc penalty terms, its strong control over false positives, and its flexible accommodation of nonparametric hypotheses. These benefits are discussed in, for example, Waller and Duncan (1969), Gopalan and Berry (1998), Berry and Hochberg (1999), Do et al. (2005), Scott and Berger (2006), and Carvalho and Scott (2007).

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. Many scientists express concerns over the lack of reproducibility of such studies; these concerns are doubtlessly fueled by widespread failure to account for the increasing multiplicity of parameters in large regression problems.

In Section 2 we give a brief historical and methodological overview of multiplicity adjustment in variable selection. We focus on two common procedures, empirical Bayes (EB) and fully Bayes (FB) variable-selection priors, and describe why empirical Bayes can yield misleading answers. We also use two data sets—one simulated, one real—to illustrate the need for multiplicity adjustment.

The remainder of the paper argues that empirical-Bayes methods can be seriously flawed along a number of dimensions. In Section 3 we introduce the notion of an information gap between fully Bayesian and empirical-Bayes procedures, proving a handful of short theorems that characterize how their behavior differs. Section 4 then discusses some experiments that show three systematic pathologies of EB estimates: they have a bias toward extreme answers that produce many more false positives; they frequently collapse to a degenerate solution indicating prior certainty in the regression model; and they can yield drastically different answers about which variables are important.

In Section 5 we give a real example from economics where the EB answer and FB answer differ substantially, providing motivation and context to the issues we
consider. We conclude in Section 6 with a discussion of our results.

2 Approaches to multiple testing

2.1 Bayes factors and Ockham’s Razor

It is often erroneously assumed that the Bayesian “Ockham’s-razor effect” renders multiplicity issues irrelevant when doing variable selection. It is important to understand why this is not so.

Bayesian marginal likelihoods, to be sure, do contain a built-in penalty for model complexity that is logically independent of the priors assigned to models (Jefferys and Berger, 1992). This is a general phenomenon resulting from the need to integrate the likelihood across a higher-dimensional parameter space under the more complex model. For example, when using $g$-priors (Zellner, 1986) to compare model $\gamma$ to an intercept-only model $\gamma_0$, the Bayes factor is:

$$\text{BF}(\gamma : \gamma_0) = (1 + g)^{(n-k_\gamma-1)/2} \cdot [1 + g(1 - R^2_\gamma)]^{-(n-1)/2}$$

where $g$ is typically $O(n)$ and $R^2_\gamma$ is the coefficient of determination.

One can immediately see the Ockham’s-razor penalty in the term involving $k_\gamma$, the number of parameters. Yet this does not look anything like a multiple-testing penalty: the Bayes factor clearly exhibits no dependence upon the number of pairwise comparisons, and hence will not control the number of false positives at all as $m$ grows large.

In fact, Bayes factors cannot have this property without violating the likelihood principle, since sufficient statistics for a given pair of models do not depend upon which other models are considered. Prior probabilities, however, do—which makes it natural to view the choice of prior distribution over models as an opportunity to address the multiple-testing problem.

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, we show that the fully Bayesian solution to multiplicity correction has something of this flavor.

We note an obvious corollary to this reasoning: that giving all models equal prior probability can be very bad from a multiple-testing standpoint, since this identifies posterior odds ratios with Bayes factors. In fact, for $m$ variables considered, this
“pseudo-objective” prior yields an expected model size of \( m/2 \) with a standard deviation of \( \sqrt{m}/2 \), meaning that the prior distribution for the fraction of included covariates becomes very tight around \( 1/2 \) as \( m \) grows. This is a surefire recipe for false positives; see Ley and Steel (2007) and Carvalho and Scott (2007) for further discussion of this issue.

2.2 Variable-selection priors and empirical Bayes

The standard modern practice in subset-selection problems is to treat variable inclusions as exchangeable Bernoulli trials with common success probability \( p \):

\[
\pi(\gamma \mid p) = p^{k_\gamma} (1 - p)^{m - k_\gamma},
\]

with \( k_\gamma \) representing the number of included variables in the model.

The key intuition in using these priors for multiplicity correction is to treat \( p \) as a model parameter to be estimated from the data. This yields an automatic multiple-testing penalty, which is often explained to be a result of shrinkage: as \( m \) grows with the true \( k \) remaining fixed, the posterior mass of \( p \) will concentrate near 0, making it harder for all variables to overcome the increasingly strong prior belief in their irrelevance (Scott and Berger, 2006).

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 (2) in a data-dependent way. Empirical Bayes might be described as a Bayesian analogue of profile likelihood. Noting that:

\[
\pi(\gamma \mid Y) = \int_0^1 \pi(\gamma \mid Y, p) \cdot \pi(p \mid Y) \, dp,
\]

the idea is choose the prior inclusion probability \( \hat{p} \) to maximize the marginal probability of the data:

\[
\hat{p} = \arg \max_{p \in [0, 1]} \sum_\gamma \pi(\gamma \mid p) \cdot f(Y \mid \gamma)
\]

This gives ex-post prior model probabilities \( \pi(\gamma \mid \hat{p}) = \hat{p}^{k_\gamma} \cdot (1 - \hat{p})^{m - k_\gamma} \). The EB solution \( \hat{p} \) can found either by direct numerical optimization or by the EM algorithm detailed in Liang et al. (2007).

To understand the theoretical rationale for this procedure, note that if \( p^* \) is the true value of \( p \), then \( \pi(p \mid Y) \overset{D}{\longrightarrow} \delta_{p^*} \) and \( \hat{p} \to p^* \) under the usual regularity conditions. It is then asserted that the integral in (3) can be approximated by:

\[
\int_0^1 \pi(\gamma \mid Y, p) \cdot \pi(p \mid Y) \, dp \approx \pi(\gamma \mid Y, \hat{p}) \propto f(Y \mid \gamma) \cdot \pi(\gamma \mid \hat{p})
\]

where the proportionality is up to a constant that is the same for all models.
The fallacy of this final step is made plain after expanding (3) a bit further:

\[
\pi(\gamma \mid Y) = \int_0^1 \pi(\gamma \mid Y, p) \cdot \pi(p \mid Y) \, dp \\
= \int_0^1 \frac{\pi(\gamma \mid p)f(Y \mid \gamma, p)}{f(Y \mid p)} \cdot \frac{\pi(p)f(Y \mid p)}{f(Y)} \, dp \quad (6)
\]

After cancelling the common \( f(Y \mid p) \) term and noting that \( Y \) is conditionally independent of \( p \) given \( \gamma \), we have (again, up to a common constant):

\[
\pi(\gamma \mid Y) \propto f(Y \mid \gamma) \int_0^1 \pi(\gamma \mid p) \cdot \pi(p) \, dp \quad (7)
\]

Note the sharp disagreement with (5), which is tempting but wrong. The cancellation of the \( f(Y \mid p) \) term means that the correct integral is not over the posterior but rather over the prior, which is clearly not well approximated by \( \pi(\gamma \mid \hat{p}) \).

A useful contrast here is with Equation 9 in Scott and Berger (2006), in which such cancellation is not possible and where, as a result, empirical Bayes makes for a good large-\( n \) approximation.

### 2.3 A fully Bayesian version

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

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

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

\[
\pi(\gamma) = \frac{(k_\gamma)!}{(m + 1)(m!)} = \frac{1}{m + 1} \left( \frac{m}{k_\gamma} \right)^{-1} \quad (9)
\]

This has the air of paradox: once we have marginalized \( p \) away, how can the data tell us about it so that shrinkage may induce a multiplicity-correction effect?

Figures 1 and 2 hint at the answer, which is that the multiplicity penalty was always in the prior to begin with. In Figure 1 we see the prior log-probability 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: 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.

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. For this study, we generated a fake $n = 75$ by $m = 100$ design matrix of $N(0, 1)$ covariates and 10 regression coefficients that differ from zero, along with 90 coefficients that are 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 (inclusion probability $\geq 50\%$) are found among the noise variables. Here, “uncorrected” refers to giving all models equal prior probability by setting $p = 1/2$. For the sake of comparison, we also give the “oracle Bayes” result in which $p$ reflects the known fraction of nonzero covariates.

Second, 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.

On the simulated data, proper multiplicity adjustment yields strong control over false positives, qualitatively reproducing the behavior of an oracle that knows how many variables are in the model. And on both the real and simulated data, adjustment tends to make all variables appear uniformly less impressive for large $m$, suggesting that false positives are squelched for the real data, as well. This happens regardless of how one computes marginal likelihoods, showing that the multiplicity penalty is logically distinct from the prior on regression coefficients and instead results from the prior distribution across model space.
Despite the ease with which fully Bayesian methods can be implemented for multiple testing and variable selection, empirical-Bayes alternatives remain popular—see George and Foster (2000), Efron et al. (2001), and Cui and George (2007) for examples and other recent references. In light of this popularity, it is worth asking how well these empirical-Bayes methods approximate fully Bayesian ones.

We find that in many cases, this approximation is quite bad. This may not be surprising in light of the discrepancy between Equations 5 and 7, but we now characterize, in information-theoretic terms, how sharply the approaches can disagree.

Denote the empirical-Bayes prior distribution over model space by $P^E_{\gamma}$, and the fully-Bayesian distribution (with uniform prior on $p$) by $P^F_{\gamma}$; these expressions are given in (2) and (9), respectively. After observing data $D$, we then have posterior distributions over model space, denoted $P^E_{\gamma|Y}$ and $P^F_{\gamma|Y}$.

Two standard measuring sticks for comparing the information content in these two pairs of distributions will be useful:

- The Kullback-Leibler (or KL) divergence of distribution $Q$ from distribution $P$ over parameter space $\Theta$ is:
  \[
  \text{KL}(P \parallel Q) = \int_{\Theta} P(\theta) \log \left( \frac{P(\theta)}{Q(\theta)} \right) d\theta
  \]  
(10)

- The squared Hellinger distance between two distributions $P$ and $Q$ over parameter space $\Theta$ is:
  \[
  H^2(P \parallel Q) = \frac{1}{2} \int_{\Theta} \left( \sqrt{P(\theta)} - \sqrt{Q(\theta)} \right)^2 d\theta
  \]  
(11)

The KL divergence lies on $[0, \infty)$, while the squared Hellinger distances lies on $[0, 1]$; each is 0 if and only if its two arguments are equal, and both satisfy the intuitive criterion that larger values signify greater disparity in information content.

We now prove four short theorems about the discrepancy between the EB and FB procedures as measured by KL divergence and Hellinger distance. The first shows that it is possible for the EB prior to be arbitrarily bad as an approximation to the FB prior; the second shows the same to be true of the EB posterior vis-à-vis the FB posterior.

**Theorem 3.1.** For fixed $m$, the prior divergence $\text{KL}(P^F_{\gamma} \parallel P^E_{\gamma})$ can be arbitrarily large.
Proof. The prior divergence can be written as:

\[
\text{KL}(P^F \parallel P^E) = \sum_{k=0}^{m} \frac{1}{m+1} \left[ \log \left( \frac{1}{m+1} \binom{m}{k} \right) - \log \left( \hat{p}^k \cdot (1 - \hat{p})^{m-k} \right) \right]
\]  

(12)

Since \( \lim_{\hat{p} \to 0} k \log \hat{p} = -\infty \) and \( \lim_{\hat{p} \to 0} k \log(1 - \hat{p}) = 1 \) for \( k \geq 1 \), the summand in (12) clearly diverges as \( \hat{p} \to 0 \); a symmetric result holds if \( \hat{p} \to 1 \).

**Theorem 3.2.** For fixed \( m \), the posterior divergence \( \text{KL}(P^F_{\gamma|Y} \parallel P^E_{\gamma|Y}) \) can be arbitrarily large.

**Proof.** The theorem is a corollary of Theorem 3.1 if there exists a configuration of marginal likelihoods over models for which \( \hat{p} = 0 \) or \( \hat{p} = 1 \). Such a configuration indeed exists: assume that all regressions models have a marginal likelihood equal to a constant \( W \) except for the full model, which has a slightly larger marginal likelihood equal to \( W + \epsilon \). Then assuming without loss of generality that \( W = 1 \):

\[
f(Y \mid p) = \sum_{\gamma \in \Gamma} \pi(\gamma \mid p) \cdot f(Y \mid \gamma)
\]

\[
= \left[ \sum_{k=0}^{m} \binom{m}{k} p^k \cdot (1 - p)^{m-k} \right] - p^m + (1 + \epsilon)p^m
\]

which for any positive \( \epsilon \) is clearly maximized at \( p = 1 \).

A symmetric configuration wherein the marginal likelihood of the null model is \( W + \epsilon \) yields \( \hat{p} = 0 \).

The upshot of Theorem 3.2 is that EB and FB answers can, at least in principle, disagree ferociously. Conditional upon the same data, the empirical-Bayes procedure can be absolutely certain that all variables are in (or all out), even while the fully Bayesian procedure yields inclusion probabilities arbitrarily close to 50%. This alone should give one pause before using empirical Bayes—and while the theorem only states that such degenerate behavior is theoretically possible, in Section 4 we show that something similar does indeed happen with surprising regularity even in realistic problems.

The next 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 high 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 3.3) and Hellinger (Theorem 3.4) flavors.
**Theorem 3.3.** Let $G(m) = \min_{\hat{p}} KL(P_{\gamma}^F \parallel P_{\gamma}^E)$. Then $G(m) \to \infty$ as $m \to \infty$.

**Proof.** The KL divergence is clearly minimized for $\hat{p} = 1/2$ regardless of $m$ or $k$, meaning that:

$$G(m) = \min_{\hat{p}} KL(P_{\gamma}^F \parallel P_{\gamma}^E) = -\log(m + 1) - \frac{1}{m + 1} \sum_{k=0}^{m} \log\left(\frac{m}{k}\right) + m \log(1/2)$$

$$= m \log 2 - \log(m + 1) - \frac{1}{m + 1} \sum_{k=0}^{m} \log\left(\frac{m}{k}\right)$$

(13)

The first (linear) term in (13) 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 $G(m)$ grows linearly with $m$, with asymptotic positive slope of $\log 2 - 1/2$.

**Theorem 3.4.** Let $H^2(m) = \min_{\hat{p}} H^2(P_{\gamma}^F \parallel P_{\gamma}^E)$. Then $\lim_{m \to \infty} H^2(m) = 1$

**Proof.**

$$H^2(P_{\gamma}^F \parallel P_{\gamma}^E) = 1 - \frac{1}{\sqrt{m + 1}} \sum_{k=0}^{m} \sqrt{\binom{m}{k} \hat{p}^k (1 - \hat{p})^{m-k}}$$

(14)

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

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

(15)

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

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

(16)

from which the result follows immediately.

In summary, we find the behavior of the EB procedure to be 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 $\hat{p} \to 0$ and the EB prior $P_{\gamma}^E$ becomes arbitrarily bad as an approximation to $P_{\gamma}^F$. 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 $P_{\gamma}^F$ and $P_{\gamma}^E$ grows linearly without bound (for KL divergence), or converges to 1 (for Hellinger distance).
The theoretical results in §3 suggest that empirical Bayes may be untrustworthy as a multiple-testing procedure—or, at the very least, that one should not automatically assume it inherits the desirable properties of a fully Bayesian approach. In this section, we present the results of extensive computer experiments showing that these concerns are of practical significance, and not mere theoretical curiosities.

We performed the following simulation 30,000 times for each of four different sample sizes:

1. Draw a random $m \times n$ design matrix $X$ 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 $\gamma$ encoding the true set of regressors.

3. Draw $\beta_{\gamma}$, the vector of regression coefficients corresponding to the nonzero elements of $\gamma$, from a Zellner-Siow prior (Zellner and Siow, 1980; Liang et al., 2007). Set the other coefficients $\beta_{-\gamma}$ to 0.

4. Draw a random vector of responses $Y \sim N(X\beta, I)$.

5. Using only $X$ and $Y$, 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 30,000 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 startlingly different conclusions about which covariates were important. As Figure 5 shows, we frequently saw 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 pathology.

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 as fast as we anticipated (Figure 4). These boxplots show long upper tails even when $n = 120$, indicating that, with nontrivial frequency, the EB procedure yields quite a bad approximation to 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, we tabulated how many false positives and
false negatives were declared by the EB and FB median-probability models, which are the models containing covariates with greater than 50% inclusion probabilities (Barbieri and Berger, 2004). 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 3 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, 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, often wildly missing the mark. It is also more susceptible to large numbers of Type-I errors—particularly worrying for a putative multiple-testing procedure.

Much of this overshooting can be explained by Table 3 and Figure 6. The EB procedure gives the degenerate \(\hat{p} = 0\) or \(\hat{p} = 1\) solution distressingly often—over 15% of the time even when \(n\) is fairly large—suggesting that the issues raised by Theorem 3.2 can be quite serious in practice.

To get some intuition as to why this is so, it helps to write the global marginal likelihood in (4) as:

\[
f(Y) = \sum_{k=0}^{m} C_k \cdot p^k \cdot (1-p)^{m-k}
\]

where \(C_k\) is just the sum of the marginal likelihoods for all models of size \(k\). Under repeated sampling from the prior, this objective function is simply a random polynomial of degree \(m\). It is unsurprising that for certain configurations of the \(\{C_k\}\), this polynomial will fail to have a local maximum in the interior of \([0, 1]\), leaving only the boundary points as possible solutions.

5 Example: Determinants of Economic Growth

We now show the differences between EB and FB answers on a real data set 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 (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 4. Two results are worth noting. First, the EB inclusion probabilities are nontrivially different from their FB counterparts, often disagreeing by 10% or more. Indeed, when judged by their median-probability models, the two analyses paint fundamentally different economic pictures. The FB analysis, for example, gives roughly two-to-one odds that a country’s life expectancy in 1960 can explain its GDP growth, whereas the EB analysis claims that this factor is more likely to be irrelevant. The two procedures also disagree on whether coastal population density from 1960 is in the median-probability model.

Second, the EB procedure disagrees not just about the numerical strength of specific variables, but also about the global ordering of variables. A simple glance at the table (ordered by decreasing FB posterior inclusion probability) shows the nonmonotonicity of the EB column. We conclude that the problem with empirical Bayes—indeed, with any plug-in choice of $p$—is not merely one of uniform over- or under-shrinkage to 0 compared to the FB procedure, but concerns something much more fundamental about its apportioning of mass across model space.

6 Summary

We have given an information-theoretic characterization of EB and FB approaches to multiple testing in variable selection. In particular, we find that the empirical-Bayes procedure exhibits suspect behavior as the number of tests for variable inclusion $m$ grows without bound, since it becomes progressively worse as an approximation to the well-studied fully Bayesian procedure. Additionally, our computer experiments show that these theoretical concerns have practical significance in a wide variety of circumstances. The EB procedure yields a fundamentally different—and, in our view, qualitatively worse—pattern of errors under long-run repetition, and it can also give very different posterior inclusion probabilities on specific problems, such as the GDP-growth regressions considered in §5. This can be a source of unnecessary confusion for practitioners; the EB/FB disagreement, for example, would likely frustrate any economist who hoped to use the median-probability GDP-growth model as a jumping-off point for theoretical development.

For all of these reasons, we recommend the fully Bayesian procedure.

References


<table>
<thead>
<tr>
<th>Signal</th>
<th>Uncorrected</th>
<th>Fully Bayes</th>
<th>Oracle Bayes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>10</td>
<td>40</td>
</tr>
<tr>
<td>$\beta_1$: -1.08</td>
<td>.999</td>
<td>.999</td>
<td>.999</td>
</tr>
<tr>
<td>$\beta_2$: -0.84</td>
<td>.999</td>
<td>.999</td>
<td>.999</td>
</tr>
<tr>
<td>$\beta_3$: -0.74</td>
<td>.999</td>
<td>.999</td>
<td>.999</td>
</tr>
<tr>
<td>$\beta_4$: -0.51</td>
<td>.977</td>
<td>.977</td>
<td>.999</td>
</tr>
<tr>
<td>$\beta_5$: -0.30</td>
<td>.292</td>
<td>.289</td>
<td>.288</td>
</tr>
<tr>
<td>$\beta_6$: +0.07</td>
<td>.259</td>
<td>.286</td>
<td>.055</td>
</tr>
<tr>
<td>$\beta_7$: +0.18</td>
<td>.219</td>
<td>.248</td>
<td>.244</td>
</tr>
<tr>
<td>$\beta_8$: +0.35</td>
<td>.773</td>
<td>.771</td>
<td>.994</td>
</tr>
<tr>
<td>$\beta_9$: +0.41</td>
<td>.927</td>
<td>.912</td>
<td>.999</td>
</tr>
<tr>
<td>$\beta_{10}$: +0.63</td>
<td>.995</td>
<td>.995</td>
<td>.999</td>
</tr>
<tr>
<td>FPs</td>
<td>0</td>
<td>2</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 1: Posterior inclusion probabilities for the 10 real variables in the simulated data set, along with the number of false positives among the “pure noise” columns in the design matrix, under different multiplicity correction regimes. 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, 2007) in the $p = 50$ and $p = 100$ cases. False positives are noise variables with posterior inclusion probability higher than 50%. 
Table 2: Posterior inclusion probabilities for ozone-concentration data under various marginal likelihoods, with and without full Bayesian multiplicity correction. Probabilities are computed with stochastic search by visiting 5 million models, starting from the null model. Key: GN = null-based $g$-priors, GF = full-based $g$-priors, ZSN = null-based Zellner-Siow priors, EBIC = extended BIC.

<table>
<thead>
<tr>
<th>Case</th>
<th>All models equal</th>
<th>Fully Bayesian, $p \sim U(0, 1)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GN</td>
<td>GF</td>
</tr>
<tr>
<td>x1</td>
<td>.860</td>
<td>.892</td>
</tr>
<tr>
<td>x2</td>
<td>.052</td>
<td>.051</td>
</tr>
<tr>
<td>x3</td>
<td>.030</td>
<td>.029</td>
</tr>
<tr>
<td>x4</td>
<td>.985</td>
<td>.987</td>
</tr>
<tr>
<td>x5</td>
<td>.195</td>
<td>.219</td>
</tr>
<tr>
<td>x6</td>
<td>.186</td>
<td>.226</td>
</tr>
<tr>
<td>x8</td>
<td>.960</td>
<td>.962</td>
</tr>
<tr>
<td>x9</td>
<td>.029</td>
<td>.035</td>
</tr>
<tr>
<td>x10</td>
<td>.999</td>
<td>.999</td>
</tr>
<tr>
<td>x1–x1</td>
<td>.999</td>
<td>.999</td>
</tr>
<tr>
<td>x9–x9</td>
<td>.999</td>
<td>.999</td>
</tr>
<tr>
<td>x1–x2</td>
<td>.577</td>
<td>.607</td>
</tr>
<tr>
<td>x1–x7</td>
<td>.076</td>
<td>.088</td>
</tr>
<tr>
<td>x3–x7</td>
<td>.021</td>
<td>.019</td>
</tr>
<tr>
<td>x4–7</td>
<td>.330</td>
<td>.353</td>
</tr>
<tr>
<td>x6–x8</td>
<td>.776</td>
<td>.785</td>
</tr>
<tr>
<td>x7–x10</td>
<td>.975</td>
<td>.952</td>
</tr>
</tbody>
</table>

Table 3: Simulated data; number of times in 30000 simulated data sets (with everything drawn from the prior) that the empirical-Bayes estimate of the prior inclusion probability collapsed to the degenerate solution $\hat{p} = 0$ or $\hat{p} = 1$, indicating prior certainty in the correct model.

<table>
<thead>
<tr>
<th>Case</th>
<th>$\hat{p} = 0$</th>
<th>$\hat{p} = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>n = 16</td>
<td>2460</td>
<td>2343</td>
</tr>
<tr>
<td>n = 30</td>
<td>2403</td>
<td>2292</td>
</tr>
<tr>
<td>n = 60</td>
<td>2349</td>
<td>2298</td>
</tr>
<tr>
<td>n = 120</td>
<td>2169</td>
<td>2241</td>
</tr>
</tbody>
</table>
Figure 3: Orthogonal design, \( m = 14 \). This is the differential pattern of errors under 30,000 fake data sets, with everything drawn from the prior. In all cases the inclusion probabilities were computed by enumeration of the model space and direct calculation of all marginal likelihoods. 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.
Figure 4: Realized KL divergence and Hellinger distance between FB and EB posterior distributions on model space in 30,000 fake data sets, \( m = 14 \), everything drawn from the prior, for a variety of different data-set sizes: \( n = \{16, 30, 60, 120\} \).
Figure 5: Empirical distribution of difference in inclusion probabilities between EB and FB, 30000 fake data sets with 14 possible covariates in each one, everything drawn from the prior.
Figure 6: Frequency distribution of $\hat{p}$ in the 14-variable simulation exercise for the $n = 60$ (left) and the $n = 120$ (right) cases. Since the distribution of true values for $p$ is uniform, these spikes at 0 and 1 suggest that empirical-Bayes has systematic bias toward extreme answers.
<table>
<thead>
<tr>
<th>Covariate</th>
<th>Fully Bayes</th>
<th>Emp. Bayes</th>
</tr>
</thead>
<tbody>
<tr>
<td>East Asian Dummy</td>
<td>0.983</td>
<td>0.983</td>
</tr>
<tr>
<td>Fraction of Tropical Area</td>
<td>0.727</td>
<td>0.653</td>
</tr>
<tr>
<td>Life Expectancy in 1960</td>
<td>0.624</td>
<td>0.499</td>
</tr>
<tr>
<td>Population Density Coastal in 1960s</td>
<td>0.518</td>
<td>0.379</td>
</tr>
<tr>
<td>GDP in 1960 (log)</td>
<td>0.497</td>
<td>0.313</td>
</tr>
<tr>
<td>Outward Orientation</td>
<td>0.417</td>
<td>0.318</td>
</tr>
<tr>
<td>Fraction GDP in Mining</td>
<td>0.389</td>
<td>0.235</td>
</tr>
<tr>
<td>Land Area</td>
<td>0.317</td>
<td>0.121</td>
</tr>
<tr>
<td>Higher Education 1960</td>
<td>0.297</td>
<td>0.148</td>
</tr>
<tr>
<td>Investment Price</td>
<td>0.226</td>
<td>0.130</td>
</tr>
<tr>
<td>Fraction Confucian</td>
<td>0.216</td>
<td>0.145</td>
</tr>
<tr>
<td>Latin American Dummy</td>
<td>0.189</td>
<td>0.108</td>
</tr>
<tr>
<td>Ethnolinguistic Fractionalization</td>
<td>0.188</td>
<td>0.117</td>
</tr>
<tr>
<td>Political Rights</td>
<td>0.188</td>
<td>0.081</td>
</tr>
<tr>
<td>Primary Schooling in 1960</td>
<td>0.167</td>
<td>0.093</td>
</tr>
<tr>
<td>Hydrocarbon Deposits in 1993</td>
<td>0.165</td>
<td>0.093</td>
</tr>
<tr>
<td>Fraction Spent in War 1960–90</td>
<td>0.164</td>
<td>0.095</td>
</tr>
<tr>
<td>Defense Spending Share</td>
<td>0.156</td>
<td>0.085</td>
</tr>
<tr>
<td>Civil Liberties</td>
<td>0.154</td>
<td>0.075</td>
</tr>
<tr>
<td>Average Inflation 1960–90</td>
<td>0.150</td>
<td>0.064</td>
</tr>
<tr>
<td>Real Exchange Rate Distortions</td>
<td>0.146</td>
<td>0.071</td>
</tr>
<tr>
<td>Interior Density</td>
<td>0.139</td>
<td>0.067</td>
</tr>
</tbody>
</table>

Table 4: Exact inclusion probabilities for 22 variables in a linear model for GDP growth among a group of 30 countries.