Bayesian Adjustment for Multiplicity

by

James G. Scott

Department of Statistical Science
Duke University

Date: ______________________________
Approved:

____
Dr. James O. Berger, Supervisor

____
Dr. M.J. Bayarri

____
Dr. Merlise Clyde

____
Dr. Mike West

Dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Department of Statistical Science
in the Graduate School of Duke University

2009

Abstract

(Statistical Science)

Bayesian Adjustment for Multiplicity

by

James G. Scott

Department of Statistical Science
Duke University

Date: ______________________________

Approved:

____
Dr. James O. Berger, Supervisor

____
Dr. M.J. Bayarri

____
Dr. Merlise Clyde

____
Dr. Mike West

An abstract of a dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Department of Statistical Science
in the Graduate School of Duke University

2009

Abstract

This thesis explores Bayesian approaches for handling multiplicity. It considers three common multiple-testing scenarios: tests of exchangeable experimental units, tests for variable inclusion in linear regresson models, and tests for conditional independence in jointly normal vectors. Though the modeling approach throughout is Bayesian, frequentist reasoning regarding error rates will often be employed.

Chapter 1 frames the issues in the context of historical debates about Bayesian multiplicity adjustment. Chapter 2 confronts the problem of large-scale screening of functional data, where control over Type-I error rates is a crucial issue. Chapter 3 develops new theory for comparing Bayes and empirical-Bayes approaches for multiplicity correction in regression variable selection. Chapters 4 and 5 describe new theoretical and computational tools for Gaussian graphical-model selection, where multiplicity arises in performing many simultaneous tests of pairwise conditional independence. Chapter 6 introduces a new approach to sparse-signal modeling based upon local shrinkage rules. Here the focus is not on multiplicity *per se*, but rather on using ideas from Bayesian multiple-testing models to motivate a new class of multivariate scale-mixture priors. Finally, Chapter 7 describes some directions for future study, many of which are the subjects of my current research agenda.

Acknowledgements

I'm grateful to all of the people who have helped me and supported me during my time here at Duke. Jim Berger has been everything a student could ask for in a mentor; I will always be thankful for his patience and generosity of spirit, and for all that he has taught me. I must also thank Mike West for his lessons on graphical models, and for his counsel on all matters great and small. Further thanks go to Merlise Clyde, Susie Bayarri, Bill Jefferys, and Pepe Quintana for their help along the way.

Nick Polson has been like a second advisor to me over the last year, and I am glad to count him as a teacher and friend. One of these days, I might understand one of his "little notes" on the first try!

Further thanks go to Mumtaz Ahmed, Michael Raynor, Lige Shao, Jim Guszcza, and Jim Wappler of Deloitte Consulting, along with Andy Henderson of the University of Texas, for their insight into the application discussed in Chapter 2, and for access to their data. I am much obliged to the National Science Foundation, the International Society for Bayesian Analysis, and the James B. Duke Scholarship Fund for their help in financing my graduate education and conference travel. I am also grateful to James Aiken for all his editing help.

It's been my privilege to meet many smart, funny, and kindhearted friends while living in Durham. To Dan, Jarad, Camille, Eric, Scotland, Leanna, Richard, Joe, Kristian, Jeff, Melanie, Ioanna, Scott, Matt, Gavino, Jill, Chip, and others I'm sure I have left out: thank you.

I'm especially indebted to Carlos Carvalho, who has been an insightful collabo-

rator, generous mentor, and loyal friend. May there be many papers, squash games, and spectacularly pointless debates yet to come.

To my wonderful Abbie: thank you for your love and kindness, and for letting me walk at your side the rest of the way. I love you.

Finally, to my parents Anne and George: thank you for your sacrifices, your guidance, and your love. Without you, none of this—Davidson, NASA, Texas, Cambridge, Duke, and Texas once again—would have happened. I could not have hoped for more supportive parents, and I will carry your example with me wherever I go.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The multiplicity problem in statistics can be stated very simply: how should we adjust our standard of impressiveness if we give a data set more than one crack at impressing us? Flip a coin ten times, and a run of ten heads is impressive. Flip that same coin 100,000 times, and a run of ten heads is nearly guaranteed. Somewhere between ten and 100,000 flips, there is room for doubt. How can that doubt be quantified?

This fundamental problem of "multiplicity adjustment"—in the sense of adjusting one's tolerance for surprise as the set of potentially surprising events grows large—arises in all manner of modern high-throughput experiments. These include microarrays, functional magnetic-resonance imagery, environmental sensor networks, combinatorial chemistry, proteomics, and many more besides. These experiments share a common inferential goal: to filter lower-dimensional signals from higher-dimensional noise. Discoveries, moreover, typically require subsequent validation, and too many

Type-I errors will mean too many expensive wild-goose chases. Hence the case for a testing procedure that displays good frequentist properties is very compelling.

But so too is the case for a model-based Bayesian procedure. These experiments may involve thousands of separate tests, and such a large volume of data often allows the distributional properties of “signals” and “noise” to be characterized quite precisely. Ignoring this information means forfeiting discriminatory power.

The central argument of this thesis is that there exists a conceptually simple, general-purpose Bayesian solution to many kinds of multiplicity problems; that this solution involves nothing more than the appropriate choice of prior model probabilities; and that this solution need not come at the expense of inflated Type-I-error rates.

In certain types of simultaneous-testing scenarios, the Bayesian solution to multiplicity is almost as simple to state as the problem. Model the signals with one distribution F_1. Model the noise with another distribution F_0. Let w be an unknown probability, and assume that the observations y_i come from the mixture

$$y_i \sim w \cdot F_1 + (1 - w) \cdot F_0 \,. \tag{1.1}$$

Apply Bayes’ rule to compute the posterior probabilities that each y_i arises from F_1. The multiplicity problem will be handled automatically through data-dependent adaptation of the posterior probabilities upon the common mixing weight w.

The sense in which this procedure is automatic, and the range of hypothesis-testing problems to which it applies, are two of the main topics of this thesis. In these problems, the only difficulty—though it may still be a substantial one, as we will see in some examples—lies in making intelligent choices about F_0 and F_1. But I will also consider problems in which (1.1) is not immediately, or at all, applicable. These include variable selection in linear regression, and tests of structural association, as in graphical models and factor models.

All of these problems generate many questions regarding multiplicity. Though I will sometimes make use of frequentist reasoning about error rates, I will focus on the questions that are most relevant to Bayesians, and to a specifically Bayesian understanding of multiplicity adjustment.

Some of these questions are foundational in nature. How, for example, does the Bayesian multiplicity penalty differ from, and how does it relate to, the Bayesian "Occam's Razor" effect?

Other questions are theoretical or methodological. Is there a sense in which the empirical-Bayes approach for assigning prior model probabilities is a good approximation to the fully Bayesian approach? What about Bayesian versus empirical-Bayes approaches for characterizing F_0 and F_1 in the simple exchangeable model? How robust are decisions with respect to misspecification of these two distributions? And are there ways of approximating the two-groups model of (1.1) with a simpler one-group model that still, in some sense, adjusts for multiplicities?

Finally, still other questions are computational. How can such large model spaces be searched efficiently enough to ensure that the multiplicity penalty applies in practice, and not just in theory? Are practitioners best served by Markov-chain Monte Carlo, or by stochastic-search methods?

These are some of the questions this thesis will attempt to answer.

1.1 Types of multiplicities

Gopalan and Berry (1998) identify at least ten common kinds of multiplicity. Some of these—for example, the use of several different test statistics on the same data—have no analogue in Bayesian inference. Others, such as interim analyses of sequentially collected data, are not a source of concern to Bayesians, for whom stopping rules are irrelevant. Still others are equally vexing to Bayesians as they are to non-Bayesians, and seem to admit no general solution. For example, publication bias and analysis bias, which is the tendency of practitioners to analyze only those data sets that have already been flagged as interesting ahead of time, pose difficulties in all schools of statistical thought.

This thesis will focus on three kinds of multiplicity for which Bayesian solutions are both readily available and quite different from classical solutions. These are:

Multiple tests in exchangeable settings, as in the simple model of (1.1). The primary goal is to flag which y_i are signals and which are noise, with the second goal often being to estimate the size of the signals. Motivating contexts include microarrays, fMRI scans, and quantitative-trait-loci mapping in genetics.

Multiple tests in linear models, where structural relationships between a response y and a basket of predictors $\{x_j\}_{j=1}^{p}$ are of primary interest.

Multiple associations of the kind that arise in fitting structured low-dimensional models to describe high-dimensional joint distributions—for example, Gaussian graphical models.

These three kinds of multiplicity arise in a wide variety of applied contexts, but are united by at least three features. The object of inferential interest is usually high-dimensional. Under an appropriate parametrization, this object is sparse, in the sense that some of its components are zero or essentially zero. Finally, the extent of this sparsity is unknown. A recognition of this third fact, along with a willingness to let the data itself characterize the prevailing rate of sparsity, is the core of the Bayesian approach to multiplicity adjustment.

1.2 An example: exchangeable normal means

As a simple illustration of the above ideas, suppose we observe $\mathbf{y} = (y_1, \ldots, y_N)$, where the y_i arise independently from normal densities $y_i \sim \mathrm{N}(\theta_i, \sigma^2)$. The y_i's, for example, may be the observed log-fold-change values from a microarray, with the θ_i's representing the mean differential expression levels for each of many thousands of genes. The multiple-testing problem is to assess whether each θ_i is zero or nonzero.

The model from (1.1) provides a natural hierarchical Bayesian approach: assume that each θ_i is nonzero with some common prior probability w, and that the nonzero θ_i come from a common $\mathrm{N}(0, \tau^2)$ density. As has been suggested, the crucial insight that makes this model work for multiple testing is to let the data choose w, and then to test on the basis of the posterior inclusion probabilities $p_i = \Pr(\theta_i \neq 0 \mid \mathbf{y})$.

This results in what is now increasingly being referred to as the Bayesian multiplicity penalty. The effect can most easily be seen if one imagines repeatedly testing a fixed number of signals (nonzero θ_i's) in the presence of an increasing number of true nulls. As noise comes to dominate the cohort of tested units, the posterior mass of w will concentrate near 0, making it increasingly difficult for all units—even the signals—to overcome the prior belief in their irrelevance. This yields much the same effect as choosing a small value for w after the fact, but Bayesian learning about w obviates the need to make an arbitrary choice for the hyperparameter.

This effect can be seen in Table 1.1, from Scott and Berger (2006). The table shows

Table 1.1: Posterior inclusion probabilities for nine signal means of varying strength in the presence of increasing N(0, 1) noise. The top row is the size of the signal; the left-most column is the number N of noise observations in addition to the signals.

	Signal Size								
N	-5.56	-2.62	-1.20	-1.01	-0.90	-0.15	1.65	1.94	3.57
25	97	71	31	28	26	20	43	51	88
100	99	47	21	20	19	16	26	31	75
500	100	34	7	6	6	4	11	15	79
5000	100	11	2	2	2	1	3	4	42

the posterior inclusion probabilities w_i of nine fixed signals that remain constant in the presence of an increasingly large number N of standard-normal noise observations: first 25, then 50, then 500, and finally 5000. The multiplicity penalty can clearly be seen, for example, in the decay of the inclusion probability for the signal of size -2.62. With only 25 noise observations, this signal is nonzero with 71% probability. But with 5,000 noise observations, it is nonzero with only 11% probability. The precipitous drop can be attributed partially to a change in the implied Bayes factor for $\theta_i \neq 0$ versus $\theta_i = 0$ due to posterior learning about the variance components τ^2 and σ^2. Mostly, however, it happens because of a major change in the implied prior odds ratio, $w/(1-w)$.

This example has several interesting features. First, in order for multiplicity adjustment to happen, w must be random rather than fixed. Consequently, if w is estimated to be large due to the presence of many signals, then the "multiplicity penalty" is actually a "multiplicity advantage."

Second, this approach sidesteps the classical debate, outlined by Hochberg and Tamhane (1987), as to whether the *comparison-wise* or *experiment-wise* error rate should be controlled. Instead, the Bayesian approach involves a comparison-wise measure of evidence (the posterior inclusion probability) that arises from assuming a specific experiment-wise relationship among the means (exchangeability). The relative tolerance for false positives and false negatives is encoded indirectly through the choice of loss function, rather than specified directly in terms of desired error rates.

Finally, the change in the prior odds as N grew did not require the choice of an arbitrary penalty term. Rather, it happened merely by introducing more noise observations into the data set and testing all of the means indiscriminately—precisely the kind of "data-dredging" that classical multiplicity correction is meant to guard against. The fact that a (careful) Bayesian approach guards against this scenario automatically is a powerful frequentist argument in its favor.

1.3 A review of Bayesian multiplicity adjustment

This thesis is primarily about multiple testing: on or off, in or out, signal or noise. Yet issues of multiplicity also arise in comparison problems, where it is relationships among elements of a vector that are of primary inferential interest. From a classical standpoint, testing and comparison appear to be very similar problems, especially in light of the similar ways in which they are handled using adjustment methods—e.g., Bonferroni correction—for raw p-values. Yet historically, Bayesian approaches to these two types of multiplicity have been quite different, and so it is worth tracing the lineage of each of these sets of ideas independently.

Multiple Testing

Any understanding of multiple testing must begin simply with testing—and the classic statement on modern Bayesian testing is due to Harold Jeffreys in 1939 (see Section 5.2 of Jeffreys, 1961, a later edition). Jeffreys imagined testing whether the mean θ in a normal sampling model was zero or nonzero, and wondered what sort of default prior $\pi(\theta)$ one should adopt for this mean under the assumption that it was nonzero. Through a detailed series of ad-hoc arguments about how the resulting Bayes factor must behave, he arrived at the conclusion that $\pi(\theta)$ should be symmetric about zero, that it should be scaled by the variance of the sampling model, and that it should have no finite (integer) moments.

The reasoning behind this third criterion forms the basis of modern debates about information consistency, an issue that will come up again in the context of graphical models. More immediately relevant for the multiple-testing problem, however, is Jeffreys' recognition that testing often requires the use of priors that are conventional in some sense, and that a well-behaved test boils down to the choice of a suitable conventional prior. This has come to be recognized as an important consideration in multiple-testing problems, where the sheer number of tests being done usually precludes, from a practical standpoint, both a full elicitation and a full study of posterior robustness to subjectively chosen priors.

Of course, as in the normal-means example, well-behaved marginal likelihoods are only part of the story in situations where multiplicity is present, because the choice of prior over model space matters a great deal. Jeffreys clearly recognized this. Although none of the solutions he offered could be described as systematic, he had much to say about priors in model-choice problems.

In fact, Jeffreys even seemed to understand multiplicity issues in model selection, though this is not completely obvious from his writings. At stake here is what Jeffreys meant by the phrase "correcting for selection," which was something he asserted must take place through the choice of prior model probabilities. This phrase is somewhat murky owing to its lack of clear provenance in the literature, but the most plausible reading is that Jeffreys meant to imply something like what modern statisticians refer to as "multiplicity adjustment."

Complicating the issue is that Jeffreys invoked an Occam's-Razor-like notion in at least two different senses. This can be a source of confusion as to what he meant when he talked about "simplicity" and "selection."

First, Jeffreys clearly anticipated the modern understanding of Bayes' Theorem as an automatic Occam's Razor in model-selection problems. A more complex law has a more diffuse predictive distribution, and will therefore suffer by comparison with a simpler law in the face of data that are consistent with both:

> For if we have a set of hypotheses $q_1, \ldots, q_m$, all asserting that a quantity x will like in a range of $\pm\epsilon$, we may denote their disjunction by q, which will assert the same. Suppose that $\sim q$ would permit the quantity to lie in a range of $\pm E$, where E is much greater than ϵ. Suppose further that x is measured and found to be in the range indicated by q. Then if p denotes this proposition, $P(p \mid qh) = 1$, and $P(p \mid \sim qh)$ is of order ϵ/E. [Here h is used to denote prior information.] Hence
>
> $$\frac{P(q \mid ph)}{P(\sim q \mid ph)} = O\left(\frac{E}{\epsilon}\right) \frac{P(q \mid h)}{P(\sim q \mid h)}$$
>
> Thus if E/ϵ is large and q is a serious possibility, a single verification

> may send its probability nearly up to 1. . . . With this rule, therefore, we can with a few verifications exclude from serious consideration any vaguely stated hypotheses that would require the observed results to be remarkable coincidences (Jeffreys, 1961, page 45).

This is precisely the sense in which modern Bayesians construe Occam's Razor: as a phenomenon arising from the marginal likelihoods of the data under the competing hypotheses (see, for example, Jefferys and Berger, 1992). The result that the more complex model will be penalized is simply a consequence of deeper ideas—that models with extra parameters have additional sources of uncertainty; that uncertainty about parameters translates into uncertainty about predictions; and that Bayesian inference rewards sharp predictions when they turn out to be correct. This has little to do with multiple testing, and is something that arises in any pairwise comparison between models, regardless of how many other models are in play.

Just two pages later, however, Jeffreys discusses an example where the goal is to choose how many terms should appear in a polynomial model for gravitation. Here he invokes a second notion of simplicity that arises not from the likelihood, but rather from the prior probabilities of models themselves:

> Precise statement of the prior probabilities of the laws . . . requires that they should actually be put in an order of decreasing prior probability. But this corresponds to actual scientific procedure. A physicist would first test whether the whole variation is random as against the existence of a linear trend; then a linear law against a quadratic one, then proceeding in order of increasing complexity. All we have to say is that the simpler laws have the greater prior probabilities. This is what Wrinch and I called the

> *simplicity postulate* (page 47 of Jeffreys, 1961, emphasis in original).

Since Jeffreys clearly identified the Occam's-Razor-like properties of Bayesian marginal likelihoods, he must have been describing a different principle here. While he was not very helpful about how to embody his "postulate" in a general context, it is plausible, in light of the few specific examples he did give, that he meant something like multiplicity correction.

This is where the phrase "correcting for selection" comes in. For instance, Jeffreys imagined testing a data set for the presence of many possible periodicities of differing lengths, denoted by hypotheses $q_1, \ldots, q_m$ whose disjunction is q':

> So far we have considered the comparison of the null hypothesis with a simple alternative . . . . Sometimes, however . . . some previous consideration suggests that some one of a group of alternative hypotheses may be right without giving any clear indication of which. For instance, the chief periods in the tides and the motion of the moon were detected by first noticing that the observed quantity varied systematically and then examining the departures in detail. In such a case . . . the presence of one period by itself would give little or no reason to expect another . . . . Suppose then that the alternatives are m in number, all with probability k initially, and that
>
> $$P(q \mid H) = P(q' \mid H) = 1/2\,.$$
>
> Since we are taking the various alternatives as irrelevant [independent] the probability that they are all false is $(1-k)^m$. But the proposition

> that they are all false is q; hence
>
> $$\begin{aligned} (1-k)^m &= 1/2 \\ k = 1 - 2^{-1/m} &\approx 0.7m \end{aligned}$$
>
> if m is large. Thus if we test the hypothesis q_1 separately we shall have
>
> $$\frac{P(q \mid H)}{P(q_1 \mid H)} = \frac{1}{2k} \approx \frac{m}{2 \log 2} = 0.7m$$
>
> nearly. If K is found by taking $P(q \mid H) = P(q_1 \mid H)$, we can *correct for selection* by multiplying K by $0.7m$ (Jeffreys, 1961, page 253, emphasis added).

A similar phrase about the effect of "selection" makes another appearance two sections later:

> [Suppose that] two sets of observations made by the same method are used to detect a new parameter by their difference. According to the rule that we are adopting, any series of observations is suspected of being subject to disturbance until there is reason to the contrary. When we are comparing two series, therefore, we are really considering four hypotheses, not two as in the test of agreement of a location parameter with zero; for neither may be disturbed, or either or both may. We continue to denote the hypothesis that both location parameters are λ by q, but q' is broken up into three, which we shall denote q_1, q_2, q_{12} (Jeffreys, 1961, page 278).

Then in a small footnote, Jeffreys adds that the prior probabilities of these four hypotheses, "in view of the convergence criterion and *allowance for selection* … might be multiplied by 1, $\frac{1}{4}$, $\frac{1}{4}$, $\frac{1}{8}$" (emphasis added).

This manner of “adjusting” or “allowing” for selection may involve a different vernacular, but it is strikingly similar to the modern notion of adjusting for multiple tests in model-choice problems. Indeed, this second example offers an obvious parallel to variable selection in regression: two variables being tested for inclusion require the comparison of four different linear models, each of which must be assigned a prior probability.

Despite their intuitive appeal, however, Jeffreys’ recommendations for handling the issue were ad hoc. His preferred assignments of prior probabilities to different models typically did not arise from a deeper generative model, and were instead designed to induce a specific kind of behavior in the problem at hand. It is fair to say that Jeffreys succeeded in showing that something like Occam’s Razor arises from deeper principles, but that he could not show the same thing for multiplicity correction. This, of course, he acknowledged himself: “I do not know whether the simplicity postulate will ever be stated in a sufficiently precise form to give exact prior probabilities to all laws; I do know that it has not been so stated yet” (Jeffreys, 1961, page 49).

Nowadays, of course, models such as (1.1) are very popular, but it is worth understanding their historical significance. While the statistical community still has not managed to state Jeffreys’ simplicity postulate in total generality, the two ideas behind (1.1)—exchangeability, along with uncertainty about common parameters—go a long way in staking the common practice of “correcting for selection” to a deeper, principled foundation.

Though Jeffreys provided the basic framework for multiplicity correction using prior model probabilities, much of the credit for articulating these deeper principles goes to Berry (1988), who framed the discrete mixture model as a natural extension of empirical-Bayes methodology. He seems to be the first author to discuss this procedure as a possible Bayesian solution to the simultaneous-testing problem, and many subsequent authors have fleshed out these ideas. Indeed, what Jeffreys did for William of Occam, the modern Bayesian community seems to have done for Jeffreys.

Many of the technical details characterizing the behavior of (1.1) can be found in Scott and Berger (2006) and Bogdan et al. (2008b), who outline various properties of the resulting Bayes rules, and of the joint posterior distribution for nuisance parameters. These proofs are given under the assumption that the nonzero means follow a normal distribution, as in the example of the previous section. Do et al. (2005) also provide an interesting variation wherein the nonzero means are modeled nonparametrically using Dirichlet processes.

Of course, while the focus here is on fully Bayesian versions of multiplicity adjustment, many of the same issues also come up in empirical-Bayes analysis. See, for example, Johnstone and Silverman (2004), Abramovich et al. (2006), and Dahl and Newton (2007).

Muller et al. (2006) and Bogdan et al. (2008a) both describe the relationship between Bayesian multiple testing and recent classical approaches that control the false-discovery rate, or FDR (Benjamini and Hochberg, 1995). This thesis does not consider the FDR-based approach, but over the last ten years it has become the dominant classical paradigm. A recent review is in Efron (2008). Many authors have commented on possible Bayesian interpretations of FDR; see, for example, Efron et al. (2001) and Storey (2003).

Finally, two other useful references give Bayesian versions of traditional multiplicity penalties involving p-values, which differ from Bayesian approaches that make use of the now-dominant discrete-mixture approach. Westfall et al. (1997) give conditions under which Bonferroni-adjusted p-values can approximate a Bayesian analysis in one-sided multiple-testing problems. Meng and Dempster (1987), on the other hand, note that *post-hoc* adjustment of p-values may not even be necessary under the assumption of exchangeability among the treatment means, with adjustment provided automatically by the resulting "Bayesian p-values."

Multiple Comparisons

Multiple comparisons are a second common form of multiplicity, and a quite different set of methodologies has evolved to handle them. The first explicitly model-based Bayesian approach to multiple comparisons seems to be that of Duncan, who wrote a series of papers (Duncan, 1961, 1965; Waller and Duncan, 1969) that introduced and refined a Bayes rule for comparing individual responses in the usual one-way ANOVA setting. Indeed, these seem to be the first systematic Bayesian treatments of any multiplicity problem.

The basic setup can be understood from a simple example in Duncan (1965)—namely, that of observing $N(N-1)/2$ t-statistics t_{ij} of the form

$$t_{ij} = \frac{|\bar{x}_i - \bar{x}_j|}{\sqrt{2}s_{ij}} \, ,$$

with each $\bar{x}_k \sim \mathrm{N}(\theta_k, \sigma^2_{\bar{x}})$, and with s_{ij} denoting an unbiased estimator of $\sigma_{\bar{x}}$, the common standard deviation of the sample means. The goal is to assess whether $\theta_i - \theta_j > \Delta$, $\theta_i - \theta_j < -\Delta$, or $|\theta_i - \theta_j| \leq \Delta$ for all distinct $i, j \in (1, \ldots, N)$, and for some "insignificant difference" Δ.

The upshot of these results is that, under an additive linear loss function, the Bayes rule for the whole decision problem reduces to the successive application of the Bayes rule for each subcomponent problem. (An additive linear loss function is one in which the losses of the component problems depend linearly upon the size of the error, and add together to give the loss for the joint problem.) Only a single constant, one that describes the relative seriousness of Type-I and Type-II errors, needs to be specified, and a Bayesian model is used for inference.

This can be viewed as a loose Bayesian justification for controlling the comparison-wise error rate rather than the experiment-wise error rate, since the number of tests performed enters the decision rule only insofar as it affects estimates of parameters at the top level of the hierarchical model for each θ_i. The model itself differs little from traditional Bayesian ANOVA models (see, for example, Tiao and Tan, 1965), though the questions posed are quite different.

The key innovations of these early papers were threefold: using hierarchical models for multiple comparisons; phrasing the issue in terms of a formal decision-theoretic framework as a way of adjudicating the debate over controlling "experiment-wise" versus "comparison-wise" error rates; and adapting to apparent heterogeneity (or lack thereof) in the data, since the procedure depended upon the F statistic. These broad themes, elucidated in the 1960's, were soon to be echoed in subsequent Bayesian developments on the multiple-testing problem.

This literature sparked the development of a wide range of "quasi-Bayesian" multiple-comparison procedures. These relied upon Bayesian models like that of Waller and Duncan, but then calibrated the choice of hyperparameters so that the resulting inferences met a pre-specified frequentist criterion—for example, control over the family-wise error rate at the 0.05 level. These procedures made use of Bayesian machinery but not Bayesian reasoning, and so are not discussed here; see Berry and Hochberg (1999) for a review and list of references.

Other Bayesian treatments of ANOVA-type models for multiple comparisons can be found in Berger and Deely (1988) and DuMouchel (1988).

A second, entirely different class of Bayesian multiple-comparison procedures abandons the use of hierarchical ANOVA models, opting instead for the explicit assignment of prior probabilities to hypotheses that certain means are identical. These are essentially models for clustering, with the object of inferential interest being the implied partition of experimental units. The use of Dirichlet-process mixtures for this purpose was proposed by Berry (1988) and developed in Gopalan and Berry (1998), who give guidelines for choosing the total mass parameter α of the Dirichlet process prior. The product-partition models of Hartigan (1990) and Crowley (1997) provide a second possibility here, with the amount of clustering controlled by the choice of an appropriate "cohesion" function.

1.4 Outline

Chapter 2 introduces a set of flexible tools for multiple testing of autoregressive time series, and functional data more generally. The initial models follow closely in the footsteps of the simple two-groups model (1.1) for testing normal means, with subsequent variations involving more complicated nonparametric models. This chapter is intended primarily as a case study involving a difficult applied multiple-testing question that arises in management theory. Here the general framework of Bayesian multiplicity adjustment interacts with a complex class of null and alternative models to generate a novel large-scale screening methodology that is both powerful and computationally tractable.

Chapter 3 then considers the problem of variable selection in regression, which can be thought of as a form of multiple testing for predictors. The first goal of the chapter is to clarify when, and how, multiplicity correction is automatic in Bayesian regression analysis, and to show how this phenomenon interacts with the Bayesian Occams-Razor effect. The second goal is to contrast Bayes and empirical-Bayes approaches to the problem using a mix of theory, simulation, and examples. The chapter's main result along these lines is a theorem that characterizes a surprising discrepancy between fully Bayes and empirical-Bayes versions of standard variable-selection priors. This discrepancy arises from a different source than the failure to account for uncertainty in the empirical-Bayes estimate, which is the usual issue in such problems. Indeed, it will be shown that even at the extreme, when the empirical-Bayes estimate converges asymptotically to the true parameter value, the potential for a serious difference remains.

Chapter 4 extends these ideas to the problem of Gaussian graphical-model selection. These models pose a special kind of variable-selection problem for an ensemble of related linear regressions, and multiplicity issues are therefore just as salient in this context as they are in linear models. The chapter will describe a novel default prior for graphically constrained covariance matrices, called the hyper-inverse Wishart g-prior, and will show how this corresponds to an implied fractional prior that arises from using fractional Bayes factors to perform model selection. This approach turns out to yield significant improvement in handling the multiplicity issue. In addition, theoretical results regarding a novel form of information consistency for graphs are derived.

Chapter 5 turns to computational questions that arise in large model spaces, with specific attention to graphical-model spaces. The chapter focuses on a serial algorithm called feature-inclusion stochastic search, or FINCS, that uses online estimates of edge-inclusion probabilities to guide Bayesian model determination. FINCS is compared to MCMC, to Metropolis-based search methods, and to the popular lasso. It is found to be superior along a variety of dimensions, leading to better sets of discovered models, greater speed and stability, and reasonable estimates of edge-inclusion probabilities. The new search procedure is then used to study the out-of-sample predictive properties of the new prior proposed in Chapter 4. In these experiments, real data involving mutual-fund returns are used, and the fractional prior is shown to outperform a variety of competing methods, both Bayesian and non-Bayesian.

Chapter 6 introduces a new approach to sparse-signal modeling called the horseshoe estimator. The horseshoe is a close cousin of other widely used Bayes rules arising from, for example, double-exponential and Cauchy priors, in that it is a member of the same family of multivariate scale mixtures of normals. Its advantage, however, is its robustness at handling unknown sparsity and large outlying signals. This chapter gives analytical results showing why the horseshoe is a good default for robust estimation of sparse normal means, and proves a new representation theorem for the posterior mean under normal scale mixtures. This theorem is related to classic results of Stein and Masreliez, and gives qualitative insight into some aspects of robust Bayesian analysis in sparse settings. Most importantly, the horseshoe estimator is shown to correspond quite closely to the answers one would get by pursuing a full Bayesian model-averaging approach using a traditional discrete-mixture prior to model signals and noise. This correspondence holds both for the posterior mean itself and for the classification rule induced by a simple thresholding scheme, meaning that the resulting ‘thresholded horseshoe’ can also be viewed as a novel Bayes multiple-testing procedure.

Finally, Chapter 7 contains some concluding remarks, and describes some directions for future research.

Chapter 2

Flexible Multiple Testing of Functional Data

This chapter introduces a framework for large-scale simultaneous testing of functional data, which is an issue that arises in areas as diverse as epidemiology, physics, genomics, and business. The goal of such analyses is to decide whether an unknown function is zero or nonzero on the basis of noisy data, and to do so for many thousands of functions at once. I will present a motivating example where the goal is to screen a large database of corporate performance statistics in an attempt to flag publicly traded companies that have consistently (and nonrandomly) outperformed their peer groups over time.

This chapter will focus on longitudinal observations, which are among the most common type of functional data encountered. Nonetheless, generalizations to other kinds of functional data—for example, spatial intensity surfaces or drug-response curves—are straightforward. Possible areas of application for these tools include time-course microarray modeling, *ex post* fraud detection in financial data, and drug discovery.

2.1 Motivating example and background

Suppose a single time series of length T is observed for each of N different units. Two possible models for each time series are entertained: a simple autoregressive null model H_0 and a more complex alternative model H_1. The goal is to determine which units come from the alternative model.

This is a frequent problem in the analysis of multiple time series. Although many details will be context-dependent, certain common themes emerge. Of key interest is how, in repeatedly applying a procedure used for testing a single time series, the rate of Type-I errors can be controlled. Model-based approaches are an attractive option, but model errors can become overwhelming in the face of massive multiplicity. One of this chapter's main results is that great care must be taken in characterizing H_0 and H_1 in order to keep false positives at bay, with the suggested robustification step involving the use of nonparametric Bayesian methods.

A running example from management theory will be used to motivate and study the proposed methodology on a real data set covering up to 50 years of annual performance statistics for over 24,000 publicly traded American companies. Performance is operationalized by return on assets, a common accounting measure of how efficiently a company uses its assets to generate revenue. The goal of the analysis is to flag firms whose historical performance trajectories are highly unlikely to have occurred by random chance, since these firms may have good (or bad) management practices that are discernible through follow-up case studies.

Longitudinal performance stratification is a classic topic in management theory. Indeed, one of the primary aims of strategic-management research, and the ambition of many best-selling books, is to explain why some firms fail and others succeed.

Much academic work in this direction focuses on decomposing observed variation into market-level, industry-level, and firm-level components (Bowman and Helfat, 2001; Hawawini et al., 2003). Of the work that attempts to identify specific non-random performers, much of it relies upon model-free clustering algorithms (for example, Harrigan, 1985), which allow no probabilistic assessment of whether the clusters found are significantly different from one another. Other approaches employ simple classical tests (Ruefli and Wiggins, 2000), often based upon ordinal time series. These have the advantage of being model-free, but are typically not based upon sufficient statistics, and suffer from the fact that available multiplicity-correction approaches (e.g. Bonferroni correction) tend toward the overly aggressive.

Due to the number of firms for which public financial data is available, this problem makes an excellent testbed for the study of general-purpose multiple-testing methodology in time-series analysis. There is, however, no theoretical ideal of what an "average-performing" company should look like, beyond the notion that it should revert to the population-level mean even if it has some randomly good or bad years. The Bayesian approach requires that suitable notions of randomness and nonrandomness be embodied in a statistical model. This model must confront an obvious multiplicity problem: many thousands of companies will be tested, and false positives will make expensive wild-goose chases out of any follow-up studies seeking to explain possible sources of competitive advantage.

Robustness and trustworthy Type-I error characteristics are therefore crucial practical considerations, and so even though this chapter's modeling approach is Bayesian, it also contains much frequentist reasoning regarding Type-I error rates. Indeed, the procedure developed here will be shown to echo the basic properties outlined in the introduction in the context of testing normal means—namely, that well-constructed Bayesian testing procedures yield excellent control over false positives.

2.2 Testing a zero-mean AR(1) model

2.2.1 The model

This section outlines a basic framework for multiple testing that, for the sake of illustration, will be purposely simplistic. Nonetheless, it will provide a useful jumping-off point for the methodological developments of subsequent sections, and will show why more flexible models are typically needed in order to achieve reasonable Type-I error performance.

Let y_{it} be the observation for unit i at time t, and let $\mathbf{y}_i$ be the vector of observations for unit i. In the management-theory example, y is standardized performance metric called Return on Assets (ROA), which measures how efficiently a company's assets generate earnings. Each company's ROA values were regressed upon a set of covariates judged to be relevant by three experts in management theory collaborating on the project. These include the company's size, debt-to-equity ratio, and market share, along with categorical variables for year and for industry membership. (See Ruefli and Wiggins, 2002, for a summary of the literature regarding covariate effects on observed firm performance.) The actual values used in the following analyses were the residuals from this regression. Also, since the question at issue is one of relative performance, not absolute performance, these residuals were standardized by CDF transform to follow a N$(0, 1)$ distribution.

Since we do not expect random gains or losses in one year to be completely erased by the following year, a model accounting for serial autocorrelation seems mandatory. Management-theoretic support for this assumption in the present context can be found in Denrell (2003) and Denrell (2005); analogous situations in engineering, finance, and biology are very common.

The null hypothesis is then a stationary AR(1) model depending upon parameter $\boldsymbol{\theta} = (\phi, v)$:

$$
\begin{aligned}
y_{it} &= \phi y_{i,(t-1)} + \nu_{it} \\
\nu_{it} &\overset{iid}{\sim} \mathrm{N}(0, v)\,.
\end{aligned}
$$

This assumption allows for long runs of good or bad performance due simply to chance: a large shock (ν_t) may take quite long to decay depending upon the value of ϕ, which is assumed to lie on $(-1, 1)$.

Non-null companies can then be modeled as AR(1) processes that revert to a nonzero mean. Placing a mixture prior on this unknown mean will then encode the relevant hypothesis test:

$$
\begin{aligned}
\mathbf{y}_i &\sim \mathrm{N}(\mathbf{y}_i \mid m_i \mathbf{1}, \Sigma_\theta) && (2.1) \\
\boldsymbol{\theta} &\sim \mathrm{N}(\phi \mid d, D) \times \mathrm{IG}(v \mid a, b) && (2.2) \\
m_i &\sim w \cdot \mathrm{N}(0, \sigma^2) + (1 - w)\delta_0\,, && (2.3)
\end{aligned}
$$

with $\mathbf{1}$ is the vector of all ones, Σ_θ is the familiar AR(1) variance matrix, δ_0 is a point mass at 0, and $w \in [0, 1]$ is the prior probability of arising from the alternative hypothesis. The exchangeable normal prior on the nonzero means m_i reflects the prior belief that, among firms that systematically deviate from zero, most of the deviations will be relatively small. The posterior probabilities $w_i = P(m_i \neq 0 \mid Y)$ then can be used to flag non-null units.

The model must be completed by specifying priors for ϕ, v, and σ^2. In the example at hand, the data are actually residuals from a regression model that adjusts for industry and other covariates, so it makes sense to choose σ^2 to match the variance of these residuals. In more general settings, the prior for σ^2 must be appropriately scaled by ϕ and v in the absence of strong prior information, since the marginal variance of the residual autoregressive process is the only quantity that provides a default scale for the problem.

The conditional likelihoods of each data vector under the two hypotheses ($m_i \neq 0$ versus $m_i = 0$) are available in closed form:

$$P(\mathbf{y}_i \mid m_i = 0, \boldsymbol{\theta}) = \mathrm{N}\left(\mathbf{y}_i \mid \mathbf{0}, \Sigma_\theta\right) \tag{2.4}$$

$$P(\mathbf{y}_i \mid m_i \neq 0, \boldsymbol{\theta}, \sigma^2) = \mathrm{N}\left(\mathbf{y}_i \mid \mathbf{0}, \Sigma_\theta + \sigma^2(\mathbf{11}^t)\right) , \tag{2.5}$$

where $(\mathbf{11}^t)$ is the matrix of all ones.

The ratio of (2.5) to (2.4) gives the Bayes factor, conditional upon ϕ, v, and σ^2, for testing an individual time series against the null model. Just as in the example from the introduction, multiplicity adjustment occurs through the posterior inclusion probabilities, which will depend upon the unknown mixing weight w.

2.2.2 Results on ROA data

Historical ROA time series for 3,459 publicly traded American companies between 1965 and 2004 were used to fit the above model. This encompasses almost every public company over that period for which at least 20 years of data were available.

Standard independent conjugate priors for ϕ and v were used:

$$\phi \sim \mathrm{N}_U(0.5, 0.25^2) \tag{2.6}$$

$$v \sim \mathrm{IG}(2, 1)\,, \tag{2.7}$$

where N_U indicates that the normal prior for ϕ is truncated to lie on the appropriate interval. These priors were chosen to reflect the expectations of the collaborating management theorists regarding the persistence and scale of random ROA fluctuations from year to year. See Berger et al. (1998) for a general guidelines on choosing common hyperparameters in model-selection problems.

In many cases the results of the fit seemed reasonable. Most firms were assigned to H_0 with high probability, while companies with obvious patterns of sustained excellence or inferiority were flagged as being from H_1 with very high probability. Figure 2.1 contains instructive examples: two excellent companies (WD-40 and Coca-Cola), along with one obviously poor company (Oglethorpe Power) were assigned greater than 95% probability of being non-null. A fourth example, Texas Intruments, had several intermittent years of good performance but no pattern of sustained excellence, and the model gave it greater than 90% probability of being from the null model.

On the other hand, the model displayed two serious shortcomings:

- Many firms diverged in obvious ways (for example, via the appearance of long-term trends) from the expectations of a single AR(1) model. Discussion of this important issue is postponed until Section 2.4.

- More subtly, the model imposed a homogeneous error structure on data that seemed rather heterogeneous. Some fairly basic exploratory data analysis indicated that firms displayed differing degrees of “stickiness” in their trajectories.

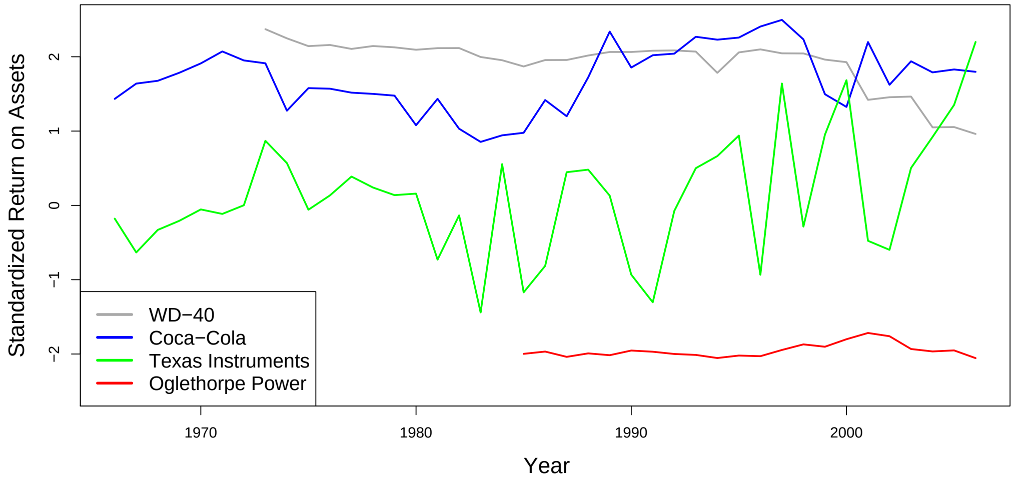


Figure 2.1: ROA histories for four example firms.

This suggested that a single value of ϕ for the entire data set might be unsatisfactory. Likewise, some firms appeared systematically more volatile than others, making a single-variance model equally questionable.

2.2.3 Robustness simulations

The possibility of model errors in (2.1)–(2.3) bring the issue of robustness to the forefront. This section describes the results of a simulation study that shows just how poorly this model can perform when a particular type of model error is encountered: deviation from the "single ϕ, single v" approach to describing the AR(1) residuals of all companies in the sample.

Several data sets displaying different levels of heterogeneity were simulated. The homogeneous (i.e. single ϕ, single v) model was subsequently fitted to each simulated data set in order to assess the robustness of the procedure's Type-I error performance.

Each simulated data set had 3500 times series of length $T = 40$, with each time series drawn from a mixture distribution of AR(1) models. These distributions ranged from trivial one-component mixtures (for which the assumed model was true) to complex nine-component mixtures (for which the assumed model was quite a bad approximation). These conditions are summarized in Table 2.1. In the four- and nine-component models, all components were equiprobable. Since all simulated companies had $m_i = 0$, ideally there should be no positive flags.

For the purposes of classification, thresholding is reported at the $w_i \geq 0.5$ and the $w_i \geq 0.9$ levels, where w_i is the posterior inclusion probability for company i. The first ($w_i \geq 0.5$) threshold reflects a 0–1 loss function that symmetrically penalizes false positives and false negatives. The second threshold is arbitrary, but is meant to reflect a more conservative approach to identifying signals. A full decision-theoretic analysis incorporating more realistic loss functions would yield a different data-adaptive threshold.

Table 2.1 supports two conclusions:

- The proposed model exhibits very strong control over false positives when its assumptions are met: 3 false positives and 0 false positives in the two cases investigated, out of 3,500 units tested.

- This excellent Type-I error profile is not at all robust to a violation of the autoregressive model's assumptions. In the most extreme case, nearly a third of units (1045 out of 3500) tested had inclusion probabilities $w_i \geq 50\%$, when in reality none were from the alternative model. In other less extreme cases, the false positives still numbered in the hundreds, which is clearly unsatisfactory.

These results dramatically illustrate the effect of heterogeneity in the autoregressive profiles of each tested unit. If such heterogeneity exists but is ignored, the Type-I

Table 2.1: Robustness of the multiple-testing procedure's Type-I error performance to heterogeneity in the autoregressive profiles of tested units. Here $\hat{w}$ refers to the posterior mode of the mixing ratio w, and the w_i's are the posterior probabilities that each $m_i \neq 0$.

Number of Components: Model	$\hat{w}$	# $w_i \geq 0.5$	# $w_i \geq 0.9$
1: $(\phi, v) = (0.5, 0.25)$	0.01	3	0
1: $(\phi, v) = (0.9, 0.5)$	0.02	0	0
4: $(\phi, v) \in \{0.5, 0.7\} \times \{0.25, 0.5\}$	0.02	30	4
4: $(\phi, v) \in \{0.4, 0.6, 0.8\} \times \{0.05, 0.25, 0.5\}$	0.08	152	66
9: $(\phi, v) \in \{0.2, 0.95\} \times \{0.05, 0.5\}$	0.37	1045	560
9: $(\phi, v) \in \{0.2, 0.6, 0.95\} \times \{0.05, 0.25, 0.5\}$	0.29	797	493

error performance of the procedure may be severely compromised.

2.3 A nonparametric null model

In the previous section, a specific form of model error—different groups of companies following different AR models—was shown to be a source of overwhelming Type-I errors. Hence a natural extension is to consider a more complicated autoregressive model for the residuals that accounts for the possibility of stratification.

The Dirichlet process (Ferguson, 1973) offers a straightforward nonparametric technique for accommodating uncertainty about this random distribution. Let $\mathbf{z}_i$ represent the response vector for unit i, for now ignoring any contribution due to a nonzero mean. Recall that for parameter $\boldsymbol{\theta} = (\phi, v)$, Σ_θ denotes the AR(1) variance matrix. The DP mixture model can then be written as a hierarchical model:

$$\mathbf{z}_i \sim \mathrm{N}(\mathbf{z}_i \mid \mathbf{0}, \Sigma_{\theta_i}) \tag{2.8}$$

$$\boldsymbol{\theta_i} \sim G, \; G \sim \mathrm{DP}(\alpha, G_0) \tag{2.9}$$

$$G_0 = \mathrm{N}(\phi \mid d, D) \times \mathrm{IG}(v \mid a, b) \,, \tag{2.10}$$

where the hyperparameters (d, D) and (a, b) must be chosen to reflect the expected properties of the base measure G_0 (which is a product of two independent distributions, a normal and an inverse-gamma), and where α controls the degree of expected departure from the base measure.

Dirichlet-process priors for nonparametric Bayesian density estimation were popularized by Escobar and West (1995), and their use in nonlinear autoregressive time series dates to Müller et al. (1997). For another reference on the use of DP priors in the context of Bayesian multiple comparisons, see Gopalan and Berry (1998).

Realizations of the Dirichlet process are discrete with probability 1, and so we expect some of the $\boldsymbol{\theta}_i$'s to be the same across companies. This is the DP framework's primary strength here, since it will facilitate borrowing of information across time series. Simply allowing each time series to have its own ϕ and own v would make for a simpler model (albeit with more parameters), but the DP prior reflects the subject-specific knowledge that significant clustering of autoregressive parameters should be expected.

This will lead to behavior similar to that predicted by a finite mixture of AR(1) models, such as the kind considered by Frühwirth-Schnatter and Kaufmann (2008). The Dirichlet-process prior, however, avoids the complicated task of directly computing marginal likelihoods for mixture models of different sizes, and so makes computation much simpler. Note that since the marginal distribution of one draw from a Dirichlet-process mixture depends only on the base measure, the DP acts like a mixture model that is predictively matched to a single observation.

It is important to consider, of course, how choices for α and G_0 affect the implied prior distributions both for the number of mixture components and for the parameters associated with each component. The marginal prior for the parameters of each mixture component is simply given by the base measure, while the prior for α can be described in terms of n and k, the desired number of mixture components, using results from Antoniak (1974).

2.4 A nonparametric alternative model

Section 2.2 considered a simple constant-mean AR(1) model for non-null units, and Section 2.3 modified the AR(1) assumption to account for a richer autoregressive structure. This section now modifies the constant-mean assumption to allow for time-varying nonzero trajectories upon which the autoregressive residuals are superimposed. Most management teams, after all, do not stay the same for 40 or 50 years, and we should not expect their performance to stay the same, either.

Firm performance trajectories can be viewed as continuous random functions that are observed at discrete (in this case, annual) intervals. This is essentially a nonparametric version of a mixed-effects model for longitudinal data (Kleinman and Ibrahim, 1998). Recent examples of such models include Bigelow and Dunson (2005), Dunson and Herring (2006), and, in a spatial context, Gelfand et al. (2005).

Let $\mathbf{f}_i = \{y_i(t), t \in \mathbb{R}^+\}$ be a continuous-time stochastic process for each observed unit, and let $\mathbf{t}_i$ denote the vector of times at which each unit was observed. Then the model is

$$\mathbf{y}_i = \mathbf{f}_i(\mathbf{t}_i) + \mathbf{z}_i \tag{2.11}$$

$$\mathbf{z}_i \sim \mathrm{N}(\mathbf{z}_i \mid \mathbf{0}, \Sigma_{\theta_i}), \quad \boldsymbol{\theta_i} \sim G \tag{2.12}$$

$$\mathbf{f}_i \sim w \cdot F + (1 - w) \cdot \delta_{F_0} \, , \tag{2.13}$$

where G is the nonparametric residual model defined in Section 2.3, δ_{F_0} represents a point mass at the zero function $F_0(t) = 0$, and F is a random distribution over a function space Ω. As in Section 2.2, w is the unknown prior probability of coming from the alternative model H_1, represented in this case by the distribution F. It is convenient to represent each hypothesis test using a model index parameter γ_i: $\gamma_i = 0$ if $\mathbf{f}_i = F_0$ (i.e. the null model H_0 is true for unit i), and $\gamma_i = 1$ otherwise.

The crucial consideration in using the above model for hypothesis testing is that the space Ω from which each $\mathbf{f}_i$ is drawn must be restricted to a sufficiently small class of functions. This would be necessary even if F were only being estimated, and not tested against a simpler model: if Ω is too broad, then the alternative model itself will not be likelihood-identified, since any pattern of residuals could equally well be absorbed by the mean function.

Following this guideline is, if anything, more important for model selection. An over-broad class of functions will mean that the random distribution F is vague, in the sense that the predictive distribution of observables will be diffuse. It is widely known that using vague priors for model selection can produce very misleading results, and will typically have the unintended consequence of sending the Bayes factor in favor of the simpler model to infinity; see, for example, Berger and Pericchi (2001). This is often known as Bartlett's paradox in the simple context of testing normal means (Bartlett, 1957), but the same principle applies here.

It may also be the case that elements of Ω depend upon some parameter $\boldsymbol{\eta}$. Since this parameter appears only in the alternative model, $\boldsymbol{\eta}$ needs a proper prior, or else the marginal likelihoods will be defined only up to an arbitrary multiplicative constant.

Similar challenges occur in all model-selection problems. Some examples of general approaches and guidelines for choosing priors on nonshared parameters can be found in Laud and Ibrahim (1995), O'Hagan (1995), Berger and Pericchi (1996), and Berger et al. (1998). Unfortunately, few tools of analogous generality have been developed for nonparametric problems, with most work concentrating on how to compute Bayes factors for pre-specified models (Basu and Chib, 2003), or how to test a parametric null against a nonparametric alternative of a suitably restricted form (Berger and Guglielmi, 2001).

This leaves just two obvious criteria for choosing Ω and F in the face of weak prior information:

1. Elements of Ω should be smooth, i.e. slowly varying on the unit-time scale of the residual model. This will allow deconvolution of the mean process from the residual, and reflects the prior belief that the mean function will describe long-term departures from 0 in the face of short-term autoregressive jitters. (Indeed, these departures are precisely what the methodology is meant to detect.)

2. F should be centered at the null model, and should concentrate most of its mass on elements of Ω that predict $\mathbf{y}$ values on a scale similar to those predicted by the null model. This will avoid Bartlett's paradox, and generalizes the argument made by Jeffreys (1961) in recommending an appropriately scaled Cauchy prior for testing normal means.

These criteria allow much wiggle room, but at least provide a starting point. Unfortunately there is no objective solution, in this or in any model-selection problem, though the closest thing to a default approach is to simply choose the marginal variance of the alternative process to exactly match the marginal variance of the null process. Best, of course, is to conduct a robustness study, where the features of the nonparametric alternative not shared by the null are varied in order to assess changes in the conclusions. This will usually be quite difficult in large multiple-testing problems, since computations for just a single version of the alternative model may be intensive.

The choice of α, the precision parameter for the residual Dirichlet-process prior, is relatively free by comparison, since this parameter appears in both the null and alternative models. Strictly speaking, in order to use a noninformative prior for α, verification of the conditions in Berger et al. (1998) regarding group invariance is necessary, which is difficult in this case. (The issue is that a parameter does not necessarily mean the same thing in both H_0 and H_1 just because it is assigned the same symbol in each.) In the absence of a formal justification for using a noninformative prior, the conservative approach is to elicit priors for α in terms of the expected number of AR(1) mixture components in each of H_0 and H_1. Often there will be extrinsic justification for choosing α to be the same under both models.

2.5 A model for the corporate-performance data

As an example of how tests involving (2.11)–(2.13) can be constructed, this section outlines a nonparametric model for a larger subset of the corporate-performance data containing 5,498 firms. This larger data set contains every publicly traded American company between 1965 and 2005 for which at least 15-year histories were available.

The class of Gaussian processes with some known covariance function is ideally suited for modeling nonzero trajectories, since the covariance function can be chosen to yield smooth functions with probability 1, and since the prior marginal variance of the process can be controlled exactly (so that Bartlett's paradox may be easily avoided). Gaussian processes have the added advantage of analytical tractability, which is very important in hypothesis testing because of the need to evaluate the marginal likelihood of the data under the alternative model. More general classes of functions are certainly possible, though perhaps computationally challenging in the face of massive multiplicity.

One additional feature to account for is clustering, since management theorists are interested in identifying a small collection of archetypal trajectories that may correspond to different sources of competitive advantage. Partitioning of firms into shared trajectories is especially relevant for advocates of the so-called "resource-based view" of the firm (Wernerfelt, 1984). Additionally, clustering on treatment effects is known to increase power in multiple-testing problems (Dahl and Newton, 2007).

The approach considered here is similar to that of Dunson and Herring (2006), with nonzero random functions modeled using a functional Dirichlet process:

$$F \sim \mathrm{FDP}(\nu, \mathrm{GP}(C_{\boldsymbol{\kappa}})) \tag{2.14}$$

$$C_{\boldsymbol{\kappa}}(t_1, t_2) = \kappa_1 \cdot \exp\left(-0.5 \cdot \frac{|t_1 - t_2|}{\kappa_2}\right)^2 . \tag{2.15}$$

The functional Dirichlet process in (2.14) has precision parameter ν and is centered at a Gaussian process with covariance function $C_{\boldsymbol{\kappa}}$. At time t, the value of the function $\mathbf{f}_i$ has a Dirichlet-process marginal distribution: $\mathbf{f}_i(t) \sim F(t)$, where $F(t) \sim \mathrm{DP}(\nu, \mathrm{N}(0, \kappa_1))$. In choosing the hyperparameter $\boldsymbol{\kappa}$, close attention must be paid to the marginal variance of the residual model, so that variance inflation in (2.15) does not overwhelm the likelihood in the Bayes-factor computations. For greater detail on Gaussian processes for nonparametric regression and function estimation, see Rasmussen and Williams (2006).

This model is significantly richer than the simplistic framework developed in Section 2.2, but is similar in two crucial ways:

Centering at the null model, since the Gaussian process in (2.15) leads to $\mathrm{E}(\mathbf{f_i} \mid \gamma_i = 1) = \mathbf{0}$. As before, it is equally likely *a priori* that a firm's trajectory will be predominantly negative or predominantly positive.

Variance inflation under the alternative model is controlled through the choice of a single hyperparameter, with κ_1 in (2.15) playing the role of σ^2 in (2.3). Hence despite the complicated nonparametric wrapper, the Occam's-razor effect upon the implied marginal likelihoods still happens in the familiar way.

The model also solves both of the major problems encountered in the ROA data: time-varying nonzero trajectories, and clustering both of trajectories and of company-specific parameters for the autoregressive residual.

An extensive prior-elicitation process was undertaken with three experts in management theory who had originally compiled the data. For the base measure of the Dirichlet-process mixture of AR(1) covariance models, the same hyperparameters from the parametric model in (2.6) and (2.7) were used. Hence the starting point for this elicitation was: $\kappa_1 \approx 1.94$, which is the prior marginal variance of the residual AR process (assessed by simulation), and $\kappa_i = 15$ (on a 41-year time scale), which reflected the experts' judgments about the long-term effects of strategic choices made by firms. The elicitees were repeatedly shown trajectories drawn from this prior and other similar priors, and soon settled upon $\kappa_1 = 1.25$ and $\kappa_2 = 13$ as values that better reflected their expectations. Additionally, they chose $\alpha = 10/\log N$, and $\nu = 15/\log N$ on the basis of how many clusters they expected.

For the actual data, a third trajectory model was also introduced: a DP mixture of constant trajectories rather than of Gaussian-process trajectories. This entails

Table 2.2: Some selected firms that were flagged as non-null with greater than 95% probability in the original analysis. GV key is the unique Compustat identifier for each firm; GICS refers to the company's code under the global industry classification system.

GV Key	Company Name	GICS	Industry
10326	Tambrands, Inc	30302010	Personal Products
1920	Avon Products	30302010	Personal Products
2269	H&R Block	25302020	Specialized Consumer Services
9860	Southern Natural Gas	55102010	Gas Utilities
1478	Wyeth	35202010	Pharmaceuticals
8208	Overnite Transportation	20304020	Trucking
8570	Macfrugals Bargains	25503020	General Merchandise Stores
9526	Scripps Howard Broadcasting	25401020	Broadcasting & Cable TV
10225	Swiss Chalet, Inc	25301020	Hotels, Resorts, Cruise Lines
10974	UST Inc	30203010	Tobacco
10920	United Parcel Service	20301010	Air Freight, Logistics
7435	3M	20105010	Industrial Conglomerates
4062	Dow Jones	25401040	Publishing
6830	Lubrizol	15101050	Specialty Chemicals
1878	Autodesk	45103010	Application Software
4094	Dun & Bradstreet	20201030	Diversified Commercial Services
8633	Plantronics	45201020	Telecommunications Equipment
11535	Winn-Dixie Stores	30101030	Food Retail
12540	Adobe Systems	45103010	Application Software

only a slight complication of the analysis, in that now (2.13) is a three-component mixture rather than a two-component mixture. This is equivalent to including the limiting-linear-model framework of Gramacy (2005)—whereby a flat trajectory is given nonzero probability as an explicit limiting case of the Gaussian process—inside the base measure of the functional Dirichlet process itself.

The above model was implemented using the blocked Gibbs-sampling algorithm of Ishwaran and James (2001) to draw from the nonparametric distributions F and G. Convergence was assessed through multiple restarts from different starting points, and was judged to be satisfactory.

Overall, 981 of 5,498 firms were flagged as being from the alternative model with greater than 50% probability, representing an overall discovery rate of about 18%. Of these, only 196 firms were from the alternative model with greater than 90% probability. A small sample of these firms can be found in Table 2.2.

To assess robustness to the hyperparameter choices κ_1 and κ_2, which control the marginal variance and temporal range of the Gaussian process base measure in (2.15), the results were recomputed for a coarse grid of 12 pairs of values spanning $0.75 \leq \kappa_1 \leq 2.25$ and $5 \leq \kappa_2 \leq 20$. This reflected the lower and upper ends of what the collaborating management theorists considered reasonable on the basis of observing draws from these priors.

As expected, larger values of κ_1 tended to yield fewer non-null classifications (due to variance inflation in the marginal likelihoods), while larger values of κ_2 tended to punish firms whose peaks and valleys in performance were short-lived over the 41-year time horizon. Many firms that were borderline in the original analysis (that is, having w_i just barely larger than 50%) were reclassified as "noise" for certain other values

Table 2.3: Posterior probabilities for six firms that were flagged as being from trajectory GP 8 with greater than 50% probability in the MLE clustering analysis; see Figure 2.2 for labels. (GV refers to a unique corporate identifier.)

GV	Company Name	Flat 1	Flat 2	GP 7	GP 8	Other	Null
11535	Winn-Dixie Stores	4	1	10	78	5	2
4828	Delhaize America	7	2	11	75	4	1
6830	Lubrizol Corp	20	9	2	64	4	1
7139	Maytag Corp	25	2	8	57	2	6
4323	Emery Air Freight	5	2	29	57	4	3
7734	National Gas & Oil	18	13	11	53	3	2

of $\boldsymbol{\kappa}$. Yet a stable cohort of 246 firms were flagged as non-null in all 12 analyses, suggesting a reasonable degree of robustness with respect to hyperparameter choice.

The behavior of individual firms can be further characterized using the MCMC history to get a maximum-likelihood estimate of the nonparametric alternative model. The almost-sure discreteness of the Dirichlet process means that this estimate is a mixture of a small number of flat and Gaussian-process trajectories. The 17 highest-weight trajectories in the MLE are in Figure 2.2; they are split into four loose categories reflecting different archetypes of company performance.

This allows an MLE clustering analysis: if F in (2.14) is frozen at the mixture model in Figure 2.2 and the MCMC rerun, it is possible to flag companies that come from specific clusters. (Strictly speaking, only the first 17 atoms in the stick-breaking approximation of F were frozen; others atoms were still considered, but they were allowed to vary.) Examples of the kinds of summaries available are in Table 2.3 and Figure 2.3.

Some general features of the methodology are apparent from these results:

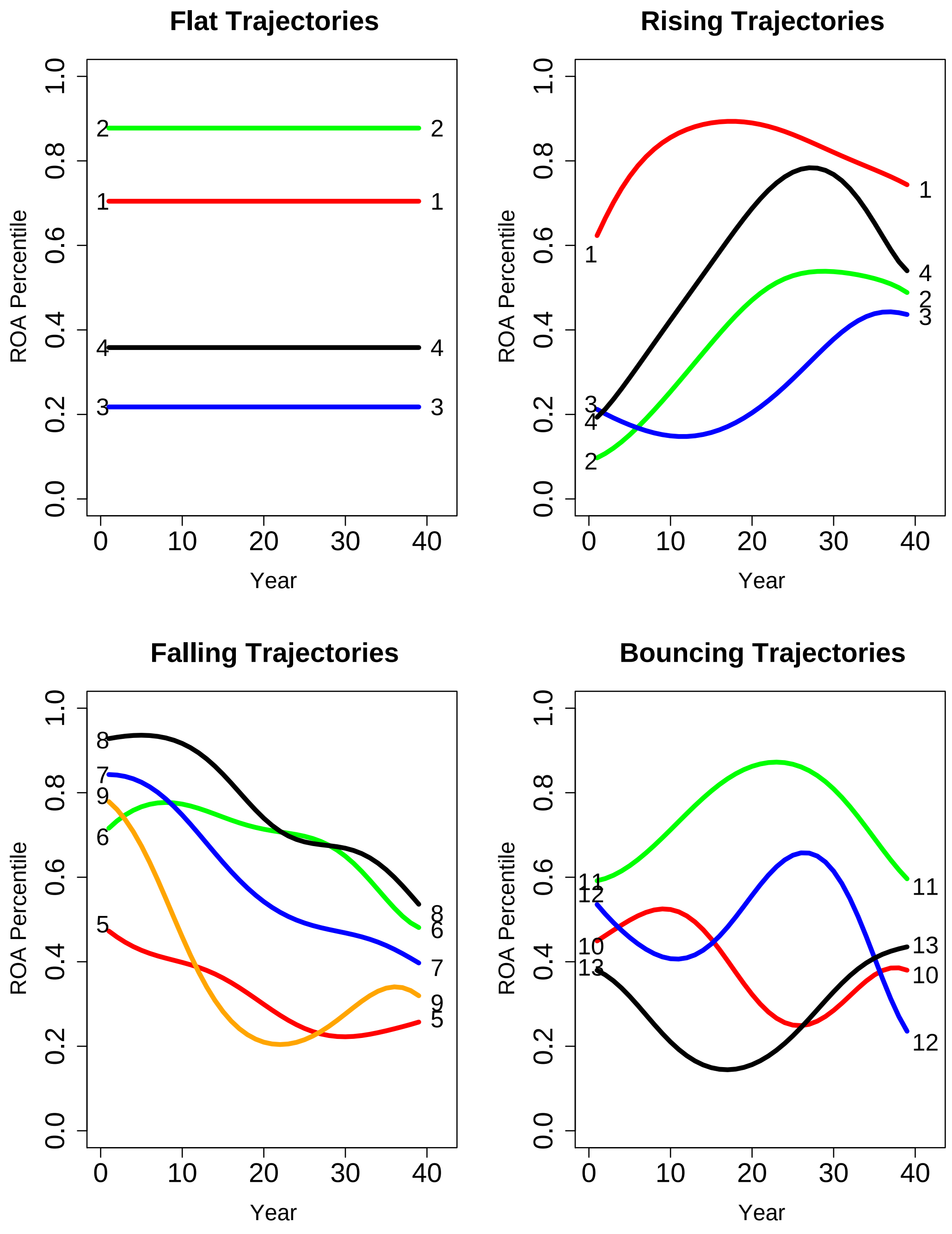


Figure 2.2: The 17 highest-weight trajectories in the MLE estimate of the alternative model, split into four loose categories. The y-axis is shown on the normal inverse-cdf scale to reflect the quantile of each firm's performance.

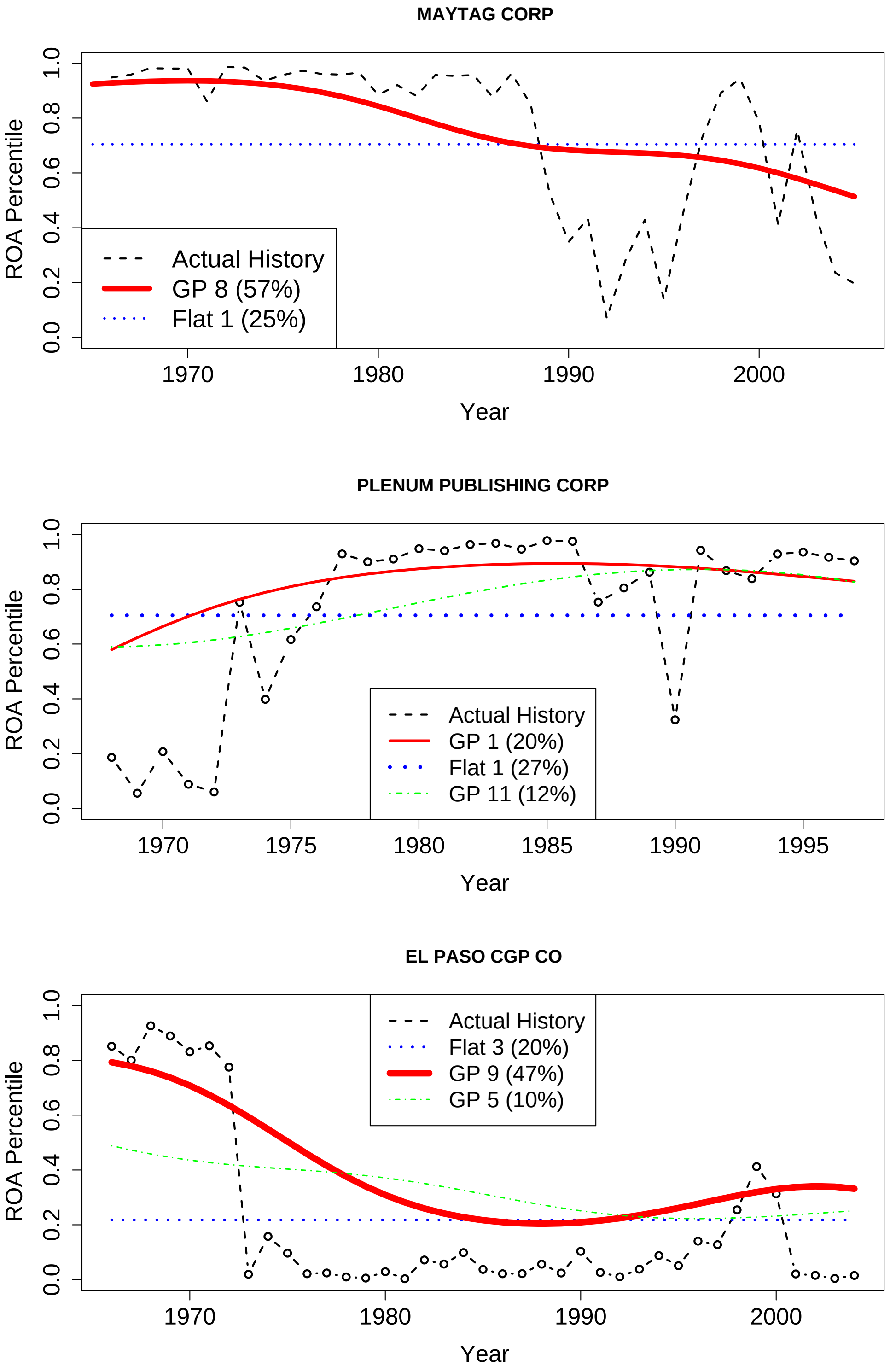


Figure 2.3: Maytag, Plenum Publishing, and El Paso CGP: actual ROA histories along with trajectory-membership probabilities from the MLE clustering analysis.

- There is substantial shrinkage of estimated mean trajectories back toward the global average (i.e. the 50th percentile). This is itself a form of multiplicity correction, in that extreme outcomes are quite likely to be attributed to chance even among those firms flagged as being from the alternative model.

- Often there is no dominant trajectory in the MLE cluster set that characterizes a specific firm's history. For the three firms in Figure 2.3, the probability is split among two or even three trajectories.

- Model-averaged predictions of future performance are available with very little extra work, since full MCMC histories of each firm's trajectory and residual model are available.

- The MLE clustering analysis can provide evidence for historical evolution within specific firms, which is of great interest as the subject of follow-up case studies. Maytag, for example, displays markedly different performance patterns before and after 1987, which is reflected in its high probability of being from a falling trajectory (GP 8).

It must be emphasized that any such clustering analysis is at best an approximation, aside from the fact that the models themselves are also approximations. It relies, after all, upon a single point estimate of the trajectories composing the random distribution F, and as such ignores uncertainty about the trajectories themselves. In the example at hand, several independent runs were conducted; each yielded a different MLE cluster set, but the same broad patterns (e.g. something like GP 8, something like Flat 3, and so on) emerged each time, suggesting at least some degree of robustness of the qualitative conclusions.

Still, the most reliable quantities are the inclusion probabilities $\{w_i\}$ computed from the full nonparametric analysis, which should form the basis of claims regarding which units are from the null and which are not.

2.6 Type-I error performance: a simulation study

This section recapitulates the simulation study of Section 2.2.3 using the more complicated models. The goal is to assess Type-I error performance by applying the methodology outlined here to a simulated data set where the number of nonzero trajectories is known.

The true residual model G was constructed using a single MCMC draw from the stick-breaking representation of the nonparametric residual model for the corporate-performance data. This corresponded roughly to a 21-component mixture of AR(1) models (since 39 of the 60 atoms in the stick-breaking representation of G had trivial weights), with many (ϕ, v) pairs that differed starkly in character.

To simulate the true mean trajectories, two independent sets of $5,500$ draws were taken from the prior in (2.11)–(2.13). In the first simulated set of trajectories, the true value of w was fixed at 1/5 in order to roughly approximate the fraction of discoveries on the real ROA data. In the second set, w was fixed at 1/55, to reflect a much sparser collection of signals. Upon sampling, these yielded 1126 and 98 nonzero trajectories, respectively.

In both studies, each of these trajectories was convolved with a single independent vector of autoregressive noise drawn from the true G—in other words, with a highly complex pattern of residual variation of the type that was shown to flummox the "single AR(1)" model of Section 2.2. For the sake of comparison, the same set of

5500 noise vectors was used for each experiment.

The results of these simulations were very encouraging for the Type-I error performance of the more complicated model. In the first simulated data set, 297 of the 1126 nonzero trajectories were flagged as non-nulls with greater than 50% probability. Only 24 of the 4374 null companies were falsely flagged, and of these, only 3 had larger than a 70% inclusion probability.

In the second simulated data set, of the 98 known nonzero trajectories, 24 had posterior inclusion probabilities greater than 50%, and 6 had inclusion probabilities greater than 90%. Of the 5402 null trajectories, only 2 had inclusion probabilities greater than 50% (these two were only 63% and 67%).

If $w_i \geq 0.5$ is taken as the decision rule for *a posteriori* classification of a trajectory as non-null (which again reflects a symmetric 0–1 loss function), then the realized false-discovery rates were only 7.7% on 26 discoveries in the sparser case, and only 7.5% on 321 discoveries in the denser case. (The closeness of these two realized false-discovery rates may simply be a coincidence due to the particular values of w chosen).

This suggests that a sizeable fraction of the 981 firms flagged as non-null trajectories in Section 2.5 represent non-average performers, and are not false positives.

2.7 Summary

This chapter has described a framework for Bayesian multiple hypothesis testing in time-series analysis that arises naturally out the simple "exchangeable normal-means" model considered in the introduction. The proposed methodology requires specifying only a few key hyperparameters for the nonparametric null and alternative models, and general guidelines for choosing these quantities have been given.

Naïve characterizations of the null hypothesis are shown to have poor error performance, suggesting that the Bayesian procedure is highly sensitive to the accuracy of the null model used to describe an "average" time series. Yet once a sufficiently complex model for residual variation is specified, the procedure exhibits very strong control over the number of false-positive declarations, even in the face of firms with different autoregressive profiles. The difference between the results in Section 2.2.3 and Section 2.6 highlights the effectiveness and practicality of using nonparametric methods as a general error-robustification tactic in multiple-testing problems.

Posterior inference for a specific time series can be summarized in at least three ways: by quoting w_i (the probability of that unit's being from the alternative model), by performing an MLE clustering analysis as in Section 2.5, or by plotting the posterior draws of the unit's nonzero mean trajectory. Plots such as Figure 2.3 can be quite useful for communicating inferences to nonexperts, as in the management-theory example considered throughout.

The companies flagged as impressive performers, of course, can only be judged so with respect to a particular notion of randomness: the DP mixture of autoregressive models described in Section 2.3. Inference on the nonzero trajectories can still reflect model misspecification, and cannot unambiguously identify companies that have found a source of sustained competitive advantage. This Dirichlet-process model, however, is a much more general statement of the null models postulated by Denrell (2003) and Denrell (2005), suggesting that the procedure described here can identify firms that, with high posterior probability, depart from randomness in a specific way that may be interesting to researchers in strategic management.

Chapter 3

Multiplicity Adjustment in Variable Selection

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

$$Y_i = \alpha + X_{ij_1}\beta_{j_1} + \ldots + X_{ij_k}\beta_{j_k} + \epsilon_i \tag{3.1}$$

for some $\{j_1, \ldots, j_k\} \subset \{1, \ldots, p\}$, where $\epsilon_i \overset{iid}{\sim} \mathrm{N}(0, \phi^{-1})$, the error precision ϕ being unknown.

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

The chapter has two main objectives. The first is to clarify how multiplicity correction enters Bayesian variable selection: through choice of the prior model probabilities, just as Jeffreys argued 70 years ago (albeit not for this particular problem). The discussion highlights the fact that not all Bayesian analyses automatically adjust for multiplicity.

The second objective is to study marked dissimilarity between fully Bayesian answers and empirical-Bayes approaches to the problem—a dissimilarity arising from a different source than the failure to account for uncertainty in the empirical-Bayes estimate (which is the usual issue in such problems). Indeed, even at the extreme, when the empirical-Bayes estimate converges asymptotically to the true parameter value, the potential for a serious discrepancy remains.

The existence of a such a discrepancy between fully Bayesian answers and empirical-Bayes answers is of immediate interest to Bayesians, who often use empirical Bayes as a computational simplification. The discrepancy is also of interest to non-Bayesians in light of some disturbing properties of the standard empirical-Bayes analysis:

- It seems to have a bias toward extreme answers that can produce too many false positives (or false negatives).

- It frequently collapses to a degenerate solution, resulting in an inappropriate statement of certainty in the selected regression model.

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

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

3.1 Notation

All models are assumed to include an intercept term α. Let H_0 denote the null model with only this intercept term, and let H_F denote the full model with all covariates under consideration. The full model thus has parameter vector $\boldsymbol{\theta}' = (\alpha, \boldsymbol{\beta}')$, $\boldsymbol{\beta}' = (\beta_1, \ldots, \beta_p)'$. Submodels $H_{\boldsymbol{\gamma}}$ are indexed by a binary vector $\boldsymbol{\gamma}$ of length m indicating a set of $k_{\boldsymbol{\gamma}} \leq p$ 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 $\{0, 1\}^p$, which has 2^p members. Inference relies upon the prior probability of each model, $p(H_{\boldsymbol{\gamma}})$, along with the marginal likelihood of the data under each model:

$$f(\mathbf{Y} \mid H_{\boldsymbol{\gamma}}) = \int f(\mathbf{Y} \mid \boldsymbol{\theta}_{\boldsymbol{\gamma}}, \phi) \; \pi(\boldsymbol{\theta}_{\boldsymbol{\gamma}}, \phi) \; \mathrm{d}\boldsymbol{\theta}_{\boldsymbol{\gamma}} \; \mathrm{d}\phi \, , \tag{3.2}$$

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:

$$p(H_{\boldsymbol{\gamma}} \mid \mathbf{Y}) \propto p(H_{\boldsymbol{\gamma}}) f(\mathbf{Y} \mid H_{\boldsymbol{\gamma}}) \, . \tag{3.3}$$

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, assume that all covariates have been centered so that $\mathbf{1}$ and $\mathbf{X}_{\boldsymbol{\gamma}}$ are orthogonal. Also assume that the common choice $\pi(\alpha) = 1$ is made for the parameter α in each model (see Berger et al., 1998, for a justification of this choice of prior).

Often all models will have small posterior probability, in which case more useful summaries of the posterior distribution are quantities such as the posterior inclusion probabilities of the individual variables. These are defined much as they are in the exchangeable-means case:

$$w_i = \Pr(\gamma_i \neq 0 \mid \mathbf{Y}) = \sum_{\boldsymbol{\gamma}} 1_{\gamma_i = 1} \cdot p(H_{\boldsymbol{\gamma}} \mid \mathbf{Y}) \, . \tag{3.4}$$

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

The choice of priors for model-specific parameters poses a different set of concerns. 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 (3.2) upon the choice of priors for model-specific parameters.

This chapter uses null-based Zellner-Siow priors (Zellner and Siow, 1980) for computing the marginal likelihoods in (3.2), which are heavy-tailed versions of Zellner's canonical g-prior (Zellner, 1986); explicit expressions can be found in Appendix A. The chief rationale for using these priors has to do with the notion of information consistency. See Berger and Pericchi (2001) and Liang et al. (2008) for overviews of information consistency; Jeffreys (1961) for a discussion of the issue in the context of testing normal means; and Appendix A for an example involving g-priors.

3.2 Multiple testing and Occam's Razor

While the Occam's-Razor effect mentioned in the introduction does function as a penalty against more complex models, it is not a multiple-testing penalty *per se*; the Bayes factor between two fixed models will not change as more possible variables are thrown into the mix, and hence will not exert control over the number of false positives as p grows large.

Instead, as earlier examples have shown, multiplicity must be handled through the choice of prior probabilities of models. 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 p 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 p grows. Indeed, note that this "pseudo-objective" prior reflects an *a-priori* expected model size of $p/2$ with a standard deviation of $\sqrt{p}/2$, meaning that the prior for the fraction of included covariates becomes very tightly concentrated around 1/2 as p grows. (One might see 7 heads in 10 flips of a fair coin, but not 700 heads in 1000 flips.)

The standard modern practice in Bayesian variable-selection problems is to proceed much as in the straightforward multiple-testing case—that is, to treat variable inclusions as exchangeable Bernoulli trials with common success probability w. This implies that the prior probability of a model is given by

$$p(H_{\boldsymbol{\gamma}} \mid p) = w^{k_{\boldsymbol{\gamma}}} \; (1-w)^{p-k_{\boldsymbol{\gamma}}} \,, \tag{3.5}$$

with $k_{\boldsymbol{\gamma}}$ representing the number of included variables in the model.

As before, w must be random in order to yield an automatic multiple-testing penalty.

Two different strategies for estimating w have evolved. The empirical-Bayes approach was popularized by George and Foster (2000), and is a common method of treating the prior inclusion probability w in (3.5) in a data-dependent way. The most straightforward version of empirical-Bayes is to estimate the prior inclusion probability by maximum likelihood, maximizing the marginal likelihood of w summed over model space (often called Type-II maximum likelihood):

$$\hat{w} = \arg\max_{w \in [0,1]} \sum_{\boldsymbol{\gamma}} p(H_{\boldsymbol{\gamma}} \mid w) \cdot f(\mathbf{Y} \mid H_{\boldsymbol{\gamma}}) \,. \tag{3.6}$$

One uses this in (3.5) to define the *ex-post* prior probabilities $p(H_{\boldsymbol{\gamma}} \mid \hat{w}) = \hat{w}^{k_{\boldsymbol{\gamma}}}(1 - \hat{w})^{p-k_{\boldsymbol{\gamma}}}$, resulting in final model posterior probabilities

$$p(H_{\boldsymbol{\gamma}} \mid \mathbf{Y}) \propto \hat{w}^{k_{\boldsymbol{\gamma}}} \cdot (1-\hat{w})^{p-k_{\boldsymbol{\gamma}}} f(\mathbf{Y} \mid H_{\boldsymbol{\gamma}})\,. \tag{3.7}$$

The EB solution $\hat{w}$ can 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: if there are only k true variables and p grows large, then $\hat{w} \to 0$.

Fully Bayesian variable-selection priors are a second possible strategy for handling w. This approach has been discussed by Ley and Steel (2007), Cui and George (2008), and Carvalho and Scott (2009), among others. Typically w is assumed to have a Beta distribution, $w \sim \mathrm{Be}(a, b)$, giving:

$$p(H_{\boldsymbol{\gamma}}) = \int_0^1 p(H_{\boldsymbol{\gamma}} \mid w)\pi(w)\ \mathrm{d}w \propto \frac{\beta(a+k_{\boldsymbol{\gamma}}, b+p-k_{\boldsymbol{\gamma}})}{\beta(a,b)}\,, \tag{3.8}$$

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

$$p(H_{\boldsymbol{\gamma}}) = \frac{(k_{\boldsymbol{\gamma}})!(p-k_{\boldsymbol{\gamma}})!}{(p+1)(p!)} = \frac{1}{p+1}\binom{p}{k_{\boldsymbol{\gamma}}}^{-1}. \tag{3.9}$$

Utilizing this in (3.3) would yield posterior model probabilities of

$$p(H_{\boldsymbol{\gamma}} \mid \mathbf{Y}) \propto \frac{1}{p+1}\binom{p}{k_{\boldsymbol{\gamma}}}^{-1} f(\mathbf{Y} \mid H_{\boldsymbol{\gamma}})\,. \tag{3.10}$$

This has the air of paradox: in contrast to (3.7), where the multiplicity adjustment is apparent, in these expressions w has been marginalized away. How can w then be "adjusted" by the data so as to induce a multiplicity-correction effect?

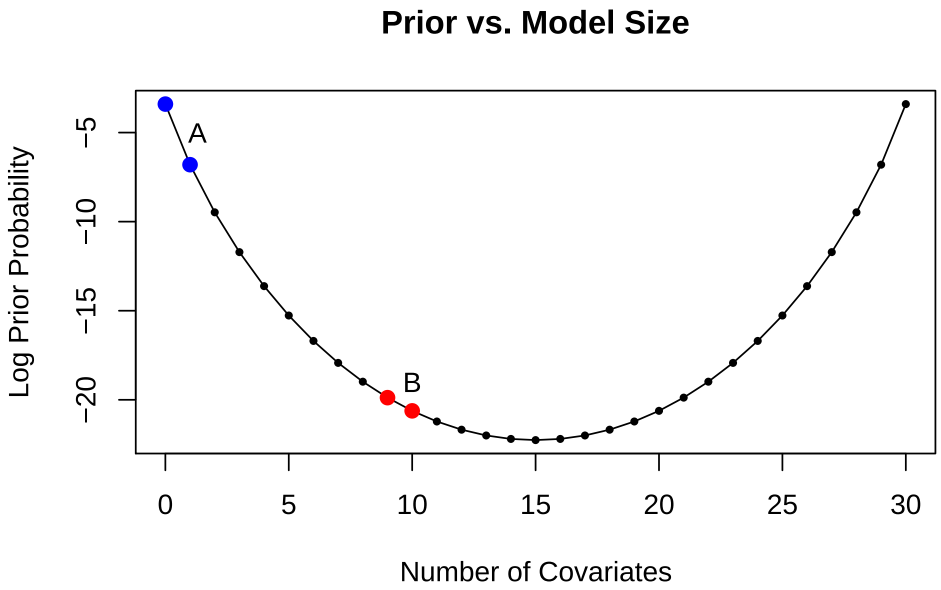


Figure 3.1: Prior probability versus model size.

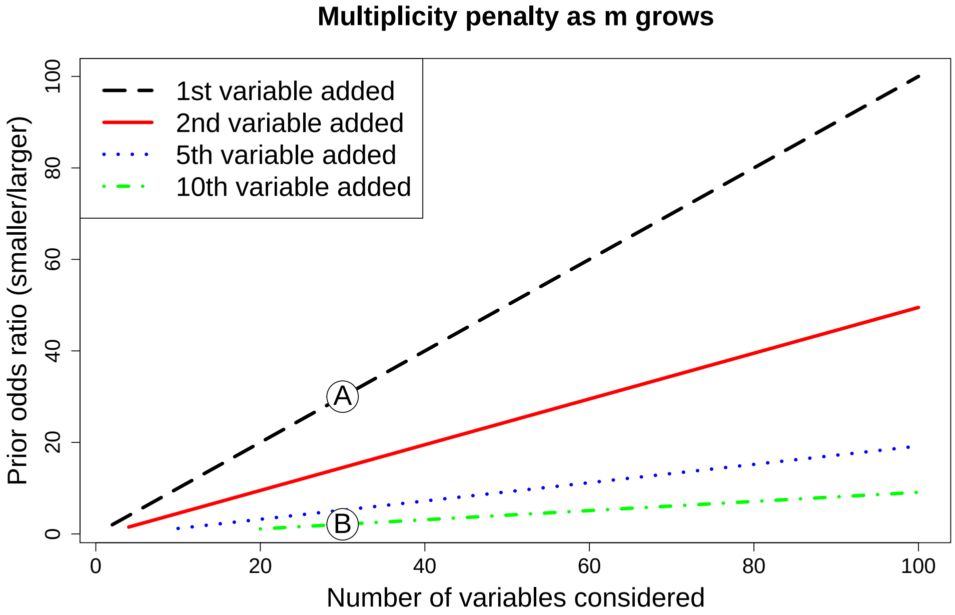


Figure 3.2: Multiplicity penalties as p grows.

Figures 3.1 and 3.2 hint at the answer, which is that the multiplicity penalty was always in the prior probabilities in (3.9) to begin with; it was just hidden. In Figure 3.1 the prior log-probability is plotted as a function of model size for a particular value of p (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 over those of size 10 by a factor of only 2.1 (label B).

Figure 3.2 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.

This effect is markedly different from the usual Occam's-Razor penalty coming through the marginal likelihoods. But marginal likelihoods are clearly relevant. They determine where models will sit along the curve in Figure 3.1, 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 3.1. This is also consistent with the empirical-Bayes answer: if $\hat{w} > 0.5$, then the analysis will increase the chance of variables entering the model.

Interestingly, the uniform prior on w also gives every variable a marginal prior inclusion probability of 1/2; these marginal probabilities are the same as those induced

Table 3.1: Posterior inclusion probabilities for the 10 real variables in the simulated data set, along with the number of false positives (posterior inclusion probability greater than 1/2) among the "pure noise" columns in the design matrix. Marginal likelihoods were calculated using null-based Zellner-Siow priors by enumerating the model space in the $p = 11$ and $p = 20$ cases, and by 5 million iterations of the feature-inclusion stochastic-search algorithm (Berger and Molina, 2005; Scott and Carvalho, 2008) in the $p = 50$ and $p = 100$ cases.

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

by the "pseudo-objective" choice of $w = 1/2$. Yet because probability is apportioned among models in a very different way, profoundly different behaviors emerge.

Table 3.1 compares these two regimes on a simulated data set for which the true value of k was fixed at 10. This study used simulated $n = \max\{75, p + 2\}$ by p design matrices of $\mathrm{N}(0, 1)$ covariates and 10 regression coefficients that differed from zero, along with $p - 10$ coefficients that were identically zero, for various values of p. The table summarizes the inclusion probabilities of the 10 real variables as they are tested along with an increasing number of noise variables (first 1, then 10, 40, and 90). It also indicates how many false positives (defined as having inclusion probability ≥ 0.5) are found among the noise variables. Here, "uncorrected" refers to giving all models equal prior probability by setting $w = 1/2$. "Oracle Bayes" is the result from choosing w to reflect the known fraction of nonzero covariates.

The following points can be observed:

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

- On the simulated data, proper multiplicity adjustment yields reasonably strong control over false positives, in the sense that the number of false positives appears bounded (and small) as m increases. In contrast, the number of false positives appears to be increasing linearly for the uncorrected Bayesian analysis, as would be expected.

- The full Bayes and oracle Bayes answers are qualitatively very similar; indeed, if one adopted the (median probability model) prescription of selecting those variables with posterior inclusion probability greater than 1/2, they would both always select the same variables, except in two instances.

Table 3.2 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 ($w = 1/2$) and fully Bayesian analyses under a variety of different marginal likelihood computations.

All variables appear uniformly less impressive when adjusted for multiplicity. This happens regardless of how one computes marginal likelihoods, indicating that, indeed,

Table 3.2: Posterior inclusion probabilities for the important main effects, quadratic effects, and cross-product effects for ozone-concentration data under various marginal likelihoods, with and without full Bayesian multiplicity correction. Key: GN = null-based g-priors, GF = full-based g-priors, ZSN = null-based Zellner-Siow priors.

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

the multiplicity penalty is logically distinct from the prior on regression coefficients and instead results from the prior distribution across model space.

3.3 Theoretical properties of empirical-Bayes

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

Lemma 3.3.1. *In the variable-selection problem, if the null model H_0 has the (strictly)* *largest marginal likelihood, then the Type-II MLE estimate of w is $\hat{w} = 0$. Similarly, if the full model H_F has the (strictly) largest marginal likelihood, then $\hat{w} = 1$.*

Proof. Since $p(H_{\boldsymbol{\gamma}})$ sums to 1 over $\boldsymbol{\gamma}$, the marginal likelihood of the data for a given value of w satisfies

$$f(\mathbf{Y}) = \sum_{\gamma} f(\mathbf{Y} \mid H_{\boldsymbol{\gamma}})\, p(H_{\boldsymbol{\gamma}}) \leq \max_{\boldsymbol{\gamma} \in \Gamma} f(\mathbf{Y} \mid H_{\boldsymbol{\gamma}})\,. \tag{3.11}$$

Furthermore, the inequality is strict under the conditions of the lemma (because the designated marginals are strictly largest), unless the prior assigns $p(H_{\boldsymbol{\gamma}}) = 1$ to the maximizing marginal likelihood. The only way that $p(H_{\boldsymbol{\gamma}}) = w^{k_{\boldsymbol{\gamma}}} \cdot (1-w)^{p-k_{\boldsymbol{\gamma}}}$ can equal 1 is for w to be 0 or 1 and for the model to be H_0 or H_F, respectively. At these values of w, equality is indeed achieved in (3.11) under the stated conditions, and the lemma follows. □

As a consequence, the empirical Bayes approach here would assign final probability 1 to H_0 whenever it has the largest marginal likelihood, and final probability 1 to H_F whenever it has the largest marginal likelihood. These are clearly very unsatisfactory answers.

3.3.1 Comparison of Empirical Bayes and Fully Bayesian Analysis

Note that the motivating lemma above referred to a clearly undesirable property of empirical-Bayes analysis in variable selection. A number of the following results have the same character. Mostly, however, the focus will be on comparing empirical-Bayes and fully Bayesian analysis.

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

To contrast this with a full Bayesian analysis, suppose one has a prior density $\pi(\boldsymbol{\lambda})$ for $\boldsymbol{\lambda}$ and a target function $\psi(\boldsymbol{\theta}, \mathbf{Y} \mid \boldsymbol{\lambda})$. 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

$$\int_\Lambda \psi(\boldsymbol{\theta}, \mathbf{Y} \mid \boldsymbol{\lambda})\pi(\boldsymbol{\lambda} \mid \mathbf{Y})\ \mathrm{d}\boldsymbol{\lambda} \approx \psi(\boldsymbol{\theta}, \mathbf{Y} \mid \hat{\boldsymbol{\lambda}})\,, \tag{3.12}$$

i.e. that the full Bayesian answer on the left can be well approximated by the empirical-Bayes answer on the right. The justification for (3.12) would be based on the fact that, typically, $\pi(\boldsymbol{\lambda} \mid \mathbf{Y})$ will be collapsing to a point mass near the true $\boldsymbol{\lambda}$ as the sample size increases, so that (3.12) will hold for appropriately smooth functions $\psi(\boldsymbol{\theta}, \mathbf{Y} \mid \boldsymbol{\lambda})$ when the sample size is large.

Note that there are typically better approximations to the left-hand side of (3.12), 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*. Higher-order empirical-Bayes analysis will typically yield better results, but the issue lies 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 (3.12) could fail for non-smooth $\psi(\boldsymbol{\theta}, \mathbf{Y} \mid \boldsymbol{\lambda})$ is no surprise. But this failure can also occur for very common functions, such as the conditional posterior density itself. Indeed, in choosing $\psi(\boldsymbol{\theta}, \mathbf{Y} \mid \boldsymbol{\lambda}) = \pi(\boldsymbol{\theta} \mid \boldsymbol{\lambda}, \mathbf{Y})$, the left-hand side of (3.12) is just the posterior density of $\boldsymbol{\theta}$ given $\mathbf{Y}$, which (by definition) can be written as

$$\pi(\boldsymbol{\theta} \mid \mathbf{Y}) \propto f(\mathbf{Y} \mid \boldsymbol{\theta}) \int_{\Lambda} \pi(\boldsymbol{\theta} \mid \boldsymbol{\lambda}) \pi(\boldsymbol{\lambda}) \, \mathrm{d}\boldsymbol{\lambda} \,. \tag{3.13}$$

On the other hand, for this choice of ψ, (3.12) becomes

$$\pi(\boldsymbol{\theta} \mid \mathbf{Y}) \approx \pi(\boldsymbol{\theta} \mid \mathbf{Y}, \hat{\boldsymbol{\lambda}}) \propto f(\mathbf{Y} \mid \boldsymbol{\theta}) \cdot \pi(\boldsymbol{\theta} \mid \hat{\boldsymbol{\lambda}}) \,, \tag{3.14}$$

and the two expressions on the right-hand sides of (3.13) and (3.14) can be very different.

As an indication of what goes wrong in (3.12) for this choice of ψ, note that

$$\begin{aligned}
\pi(\boldsymbol{\theta} \mid \mathbf{Y}) &= \int_{\Lambda} \pi(\boldsymbol{\theta} \mid \boldsymbol{\lambda}, \mathbf{Y}) \cdot \pi(\boldsymbol{\lambda} \mid \mathbf{Y}) \, \mathrm{d}\boldsymbol{\lambda} \\
&= \int_{\Lambda} \frac{\pi(\boldsymbol{\theta}, \boldsymbol{\lambda} \mid \mathbf{Y})}{\pi(\boldsymbol{\lambda} \mid \mathbf{Y})} \cdot \pi(\boldsymbol{\lambda} \mid \mathbf{Y}) \, \mathrm{d}\boldsymbol{\lambda} & (3.15) \\
&= \int_{\Lambda} \frac{f(\mathbf{Y} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta} \mid \boldsymbol{\lambda}) \pi(\boldsymbol{\lambda})}{f(\mathbf{Y}) \pi(\boldsymbol{\lambda} \mid \mathbf{Y})} \cdot \pi(\boldsymbol{\lambda} \mid \mathbf{Y}) \, \mathrm{d}\boldsymbol{\lambda} \,, & (3.16)
\end{aligned}$$

which leads to (3.13) upon canceling $\pi(\boldsymbol{\lambda} \mid \mathbf{Y})$ in the numerator and denominator. It simply does not matter that $\pi(\boldsymbol{\lambda} \mid \mathbf{Y})$ is collapsing to a point about the true $\boldsymbol{\lambda}$, because it occurs in both the numerator and the denominator of the integrand.

The usual appeal to the asymptotic consistency of the empirical-Bayes estimator $\hat{\boldsymbol{\lambda}}$ is therefore irrelevant in judging the appropriateness of the approximation in (3.12).

In the remainder of this section, the focus will be on comparing (3.13) with (3.14) since, for model selection, the full posteriors are most relevant. The "closeness" of the two distributions will be measured by Kullback-Leibler divergence, a standard measure for comparing a pair of distributions P and Q over parameter space Θ:

$$\mathrm{KL}(P \parallel Q) = \int_{\Theta} P(\boldsymbol{\theta}) \log \left(\frac{P(\boldsymbol{\theta})}{Q(\boldsymbol{\theta})} \right) \mathrm{d}\boldsymbol{\theta} \, . \tag{3.17}$$

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

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

Note that the KL-convergence criterion is defined with respect to a particular function of the parameter, along with a particular prior distribution on the hyperparameter. If, for a specific target function $\psi(\boldsymbol{\theta})$, it is not possible to find a reasonable prior $\pi(\boldsymbol{\lambda})$ that leads to KL convergence, then estimating $\psi(\boldsymbol{\theta})$ by empirical Bayes is clearly suspect on Bayesian grounds. This is because a Bayesian could not replicate such a procedure even asymptotically; there would be an unbridgeable information gap between the two procedures that would not disappear even in the limit. 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:

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

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

3.3.2 Posteriors for Normal Means

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

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

$$\pi_E(\boldsymbol{\theta} \mid \mathbf{y}, \hat{\mu}_E) = \mathrm{N}\left(\frac{1}{2}\left(\mathbf{y} + \bar{y}\mathbf{1}\right),\ \frac{1}{2}\mathbf{I}\right) \tag{3.18}$$

$$\pi_F(\boldsymbol{\theta} \mid \mathbf{y}) = \mathrm{N}\left(\frac{1}{2}\left(\mathbf{y} + \bar{y}\mathbf{1}\right) - \left(\frac{1}{nA+2}\right)\bar{y}\mathbf{1},\ \frac{1}{2}\mathbf{I} + \frac{A}{2(nA+2)}(\mathbf{1}\mathbf{1}^t)\right)\ , \tag{3.19}$$

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

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

$$\pi_E(\theta_1 \mid \mathbf{y}, \hat{\mu}_E) = \mathrm{N}\left([y_1 + \bar{y}]/2,\ 1/2\right) \tag{3.20}$$

$$\pi_F(\theta_1 \mid \mathbf{y}) = \mathrm{N}\left([y_1 + \bar{y}]/2,\ 1/2 + [2n]^{-1}\right)\ . \tag{3.21}$$

It is easy to check that $\mathrm{KL}(\pi_F \parallel \pi_E) \to 0$ as $n \to \infty$. 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 (3.18) and (3.19).

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

$$\mathrm{KL} = \frac{1}{2}\left[\log\left(\frac{\det \Sigma_E}{\det \Sigma_F}\right) + \mathrm{tr}(\Sigma_E^{-1}\Sigma_F) + (\hat{\boldsymbol{\theta}}_E - \hat{\boldsymbol{\theta}}_F)^t\ \Sigma_E^{-1}\ (\hat{\boldsymbol{\theta}}_E - \hat{\boldsymbol{\theta}}_F) - n\right] \tag{3.22}$$

$$= \frac{1}{2}\left[-\log\left(1 + \frac{nA}{nA+2}\right) + \frac{nA}{nA+2} + 2n\left(\frac{1}{nA+2}\right)^2 \bar{y}^2\right]\ . \tag{3.23}$$

For any nonzero choice of A and for any finite value of the hyperparameter μ, it is clear that under $p(\mathbf{y}, \boldsymbol{\theta} \mid \mu)$ the quantity $[2n/(nA+2)^2] \cdot \bar{y}^2 \to 0$ in probability as $n \to \infty$. Hence for any value of A (including $A = \infty$), the KL divergence in (3.23) converges to $(1 - \log 2)/2 > 0$ as n grows.

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

3.3.3 Results for Variable Selection

For the variable-selection problem, explicit expressions for the KL divergence between empirical-Bayes and fully Bayes posterior distributions are not available. It is therefore not possible to provide a general characterization of when the empirical-Bayes variable-selection procedure is KL-convergent, in the sense defined above, to a fully Bayesian procedure.

Two interesting sets of results, however, are still available: one regarding the KL divergence between the prior probability distributions of the fully Bayesian and empirical-Bayesian procedures, and the other regarding the *expected* posterior KL divergence.

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

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 p goes to infinity. I refer to these lower bounds as "information gaps," and give them in both Kullback-Leibler (Theorem 3.3.2) and Hellinger (Theorem 3.3.3) versions.

Theorem 3.3.2. *Let* $\underline{G}(p) = \min_{\hat{w}} KL(p_F(\boldsymbol{\gamma}) \parallel p_E(\boldsymbol{\gamma}))$*. Then* $\underline{G}(p) \to \infty$ *as* $p \to \infty$*.*

Proof. The KL divergence is

$$\mathrm{KL} = \sum_{k=0}^{p} \frac{1}{p+1} \left[\log \left(\frac{1}{p+1} \binom{p}{k}^{-1} \right) - \log \left(\hat{w}^k \cdot (1-\hat{w})^{p-k} \right) \right] \tag{3.24}$$

$$= -\log(p+1) - \frac{1}{p+1} \sum_{k=0}^{p} \left[\log \binom{p}{k} + k \log \hat{w} + (p-k) \log(1-\hat{w}) \right] .$$

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

$$\underline{\mathrm{G}}(p) = -\log(p+1) - \frac{1}{p+1} \sum_{k=0}^{p} \left[\log \binom{p}{k} + p \log(1/2) \right]$$

$$= p \log 2 - \log(p+1) - \frac{1}{p+1} \sum_{k=0}^{p} \log \binom{p}{k} . \tag{3.25}$$

The first (linear) term in (3.25) dominates the second (logarithmic) term, whereas results in Gould (1964) show the third term to be asymptotically linear in p with slope 1/2. Hence $\underline{\mathrm{G}}(p)$ grows linearly with p, with asymptotic positive slope of $\log 2 - 1/2$. □

Theorem 3.3.3. *Let* $\underline{H}^2(p) = \min_{\hat{w}} H^2(p_F(\boldsymbol{\gamma}) \parallel p_E(\boldsymbol{\gamma}))$*. Then* $\underline{H}^2(p) \to 1$ *as* $p \to \infty$*.*

Proof. Clearly

$$\mathrm{H}^2(p_F(\boldsymbol{\gamma}) \parallel p_E(\boldsymbol{\gamma})) = 1 - \frac{1}{\sqrt{p+1}} \sum_{k=0}^{p} \sqrt{\binom{p}{k} \hat{w}^k (1-\hat{w})^{p-k}} . \tag{3.26}$$

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

$$\underline{\mathrm{H}}^2(p) = 1 - (p+1)^{-1/2} \cdot 2^{-p/2} \cdot \sum_{k=0}^{p} \sqrt{\binom{p}{k}} \, . \tag{3.27}$$

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

$$\lim_{p \to \infty} \left[(p+1)^{-1/2} \cdot 2^{-p/2} \cdot \sum_{k=0}^{p} \sqrt{\binom{p}{k}} \right] = 0 \, , \tag{3.28}$$

from which the result follows immediately. □

In summary, the *ex-post* prior distribution associated with the EB procedure is particularly troubling when the number of tests p grows without bound. On the one hand, when the true value of k remains fixed or grows at a rate slower than p—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 $\hat{w} \to 0$ and the EB prior $p_E(\boldsymbol{\gamma})$ becomes an ever-poorer approximation to $p_F(\boldsymbol{\gamma})$. On the other hand, if the true k is growing at the same rate as p, then the best one can hope for is that $\hat{w} = 1/2$—and even then, the information gap between $p_F(\boldsymbol{\gamma})$ and $p_E(\boldsymbol{\gamma})$ grows linearly without bound (for KL divergence), or converges to 1 (for Hellinger distance).

Posterior KL Divergence

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

Theorem 3.3.4. *In the variable-selection problem, let p, $n > p$, and $\phi > 0$ be fixed. Suppose $\mathbf{X}_{\boldsymbol{\gamma}}$ is of full rank for all models and that the family of priors for model-specific parameters, $\{\pi(\boldsymbol{\beta}_{\boldsymbol{\gamma}})\}$, are such that $p(\boldsymbol{\beta}_{\boldsymbol{\gamma}} = \mathbf{0}) < 1$ for all $H_{\boldsymbol{\gamma}}$. Then, for any true model $H^T_{\boldsymbol{\gamma}}$, the expected posterior KL divergence $\mathrm{E}[KL(p_F(\boldsymbol{\gamma} \mid \mathbf{Y}) \parallel p_E(\boldsymbol{\gamma} \mid \mathbf{Y}))]$ under this true model is infinite.*

Proof. The posterior KL divergence is

$$\mathrm{KL}(p_F(\boldsymbol{\gamma} \mid \mathbf{Y}) \parallel p_E(\boldsymbol{\gamma} \mid \mathbf{Y})) = \sum_{\gamma} p_F(H_{\boldsymbol{\gamma}} \mid \mathbf{Y}) \cdot \log\left(\frac{p_F(H_{\boldsymbol{\gamma}} \mid \mathbf{Y})}{p_E(H_{\boldsymbol{\gamma}} \mid \mathbf{Y})}\right) . \tag{3.29}$$

This is clearly infinite if there exists a model $H_{\boldsymbol{\gamma}}$ for which $p_E(H_{\boldsymbol{\gamma}} \mid \mathbf{Y}) = 0$ but $p_F(H_{\boldsymbol{\gamma}} \mid \mathbf{Y}) > 0$. Since the fully Bayesian posterior assigns nonzero probability to all models under the conditions of the theorem, this will be the case whenever the empirical-Bayesian solution is $\hat{w} = 0$ or $\hat{w} = 1$.

Thus it suffices to show that $\hat{w}$ will be 0 with positive probability under any true model.

Assume without loss of generality that $\phi = 1$. Recall that $\pi(\alpha) = 1$ for all models, and that the intercept is orthogonal to all other covariates. Letting $\boldsymbol{\beta}^{\boldsymbol{*}}_{\boldsymbol{\gamma}} = (\alpha, \boldsymbol{\beta}_{\boldsymbol{\gamma}})^t$ for model $H_{\boldsymbol{\gamma}}$, and letting $L(\cdot)$ stand for the likelihood, the marginal likelihood for any model can then be written

$$f(\mathbf{Y} \mid H_{\boldsymbol{\gamma}}) = L(\hat{\boldsymbol{\beta}}^*_{\gamma}) \cdot \sqrt{2\pi/n} \int_{\mathbb{R}^{k_\gamma}} g(\boldsymbol{\beta}_{\boldsymbol{\gamma}}) \pi(\boldsymbol{\beta}_{\boldsymbol{\gamma}}) \, \mathrm{d}\boldsymbol{\beta}_{\boldsymbol{\gamma}} \, , \tag{3.30}$$

where

$$g(\boldsymbol{\beta}_{\boldsymbol{\gamma}}) = \exp\left\{ -\frac{1}{2} (\boldsymbol{\beta}_{\boldsymbol{\gamma}} - \hat{\boldsymbol{\beta}}_{\gamma})^t \; \mathbf{X}^t_{\boldsymbol{\gamma}} \mathbf{X}_{\boldsymbol{\gamma}} \; (\boldsymbol{\beta}_{\boldsymbol{\gamma}} - \hat{\boldsymbol{\beta}}_{\gamma} \, . \right\}$$

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

$$B_\gamma(\mathbf{Y}) = \frac{f(\mathbf{Y} \mid H_0)}{f(\mathbf{Y} \mid H_{\boldsymbol{\gamma}})} \ ,$$

which from (3.30) is clearly continuous as a function of $\mathbf{Y}$ for every $\boldsymbol{\gamma}$. Evaluated at $\mathbf{Y} = \mathbf{0}$, this Bayes factor satisfies

$$B_\gamma(\mathbf{0}) = \left(\int_{\mathbb{R}^{k_\gamma}} \exp\left\{ -\frac{1}{2} (\hat{\boldsymbol{\beta}}_\gamma - \boldsymbol{\beta}_{\boldsymbol{\gamma}})^t \ \mathbf{X}_{\boldsymbol{\gamma}}^t \mathbf{X}_{\boldsymbol{\gamma}} \ (\hat{\boldsymbol{\beta}}_\gamma - \boldsymbol{\beta}_{\boldsymbol{\gamma}}) \right\} \pi(\boldsymbol{\beta}_{\boldsymbol{\gamma}}) \ \mathrm{d}\boldsymbol{\beta}_{\boldsymbol{\gamma}} \right)^{-1} > 1 \quad (3.31)$$

for each $H_{\boldsymbol{\gamma}}$ under the assumptions of the theorem.

By continuity, for every model $H_{\boldsymbol{\gamma}}$ there exists an ϵ_γ such that $B_\gamma(\mathbf{Y}) > 1$ for any $\|\mathbf{Y}\| < \epsilon_\gamma$. Let $\epsilon^* = \min_{\boldsymbol{\gamma}} \epsilon_\gamma$. Then for $\mathbf{Y}$ satisyfing $\|\mathbf{Y}\| < \epsilon^*$, $B_\gamma(\mathbf{Y}) > 1$ for all non-null models, meaning that H_0 will have the largest marginal likelihood. By Lemma 3.3.1, $\hat{p} = 0$ when such a $\mathbf{Y}$ is observed.

But there is positive probability of observing $|\mathbf{Y}| < \epsilon^*$ under any model, 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 $\mathrm{KL}(p_F(\boldsymbol{\gamma} \mid \mathbf{Y}) \parallel p_E(\boldsymbol{\gamma} \mid \mathbf{Y}))$ is infinite under the sampling distribution $p(\mathbf{Y} \mid H_{\boldsymbol{\gamma}})$, and so its expectation is clearly infinite. □

Since the expected KL divergence is infinite for any number m of variables being tested, and for any true model, it is clear that $\mathrm{E}(KL)$ does not converge to 0 as $p \to \infty$.

In Theorem 3.3.4, the expectation can be taken with respect to either of two possible sampling distributions:

- the sampling distribution under a specific model $H_{\boldsymbol{\gamma}}$, with $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ fixed

- the sampling distribution under a specific model H_{γ}, with$\boldsymbol{\beta}_{\gamma}$ drawn from a prior satisfying the regularity conditions of the theorem.

The result also holds, with only slight modifications to the proof, under other obvious choices of the sampling distribution—for example, under the Bernoulli model for $\boldsymbol{\gamma}$ in (3.5), with w either fixed or random.

3.4 Numerical Investigation of Empirical-Bayes Variable Selection

The theoretical results in Section 3.3 indicate that empirical-Bayes analysis cannot always be trusted in the variable selection problem. This section presents numerical results that indicate that these concerns are of practical significance, and not mere theoretical curiosities. As in the previous section, most of the investigation involves comparing empirical-Bayes and fully Bayesian analysis, but at least some of the findings point out obviously inappropriate properties of the empirical-Bayes procedure itself.

3.4.1 Results under Properly Specified Priors

The following simulation was performed 75,000 times for each of four different sample sizes:

1. Draw a random $n \times p$ design matrix $\mathbf{X}$ of independent $\mathrm{N}(0,1)$ covariates.

2. Draw a random $w \sim \mathrm{U}(0,1)$, and draw a sequence of m independent Bernoulli trials with success probability w to yield a binary vector $\boldsymbol{\gamma}$ encoding the true set of regressors.

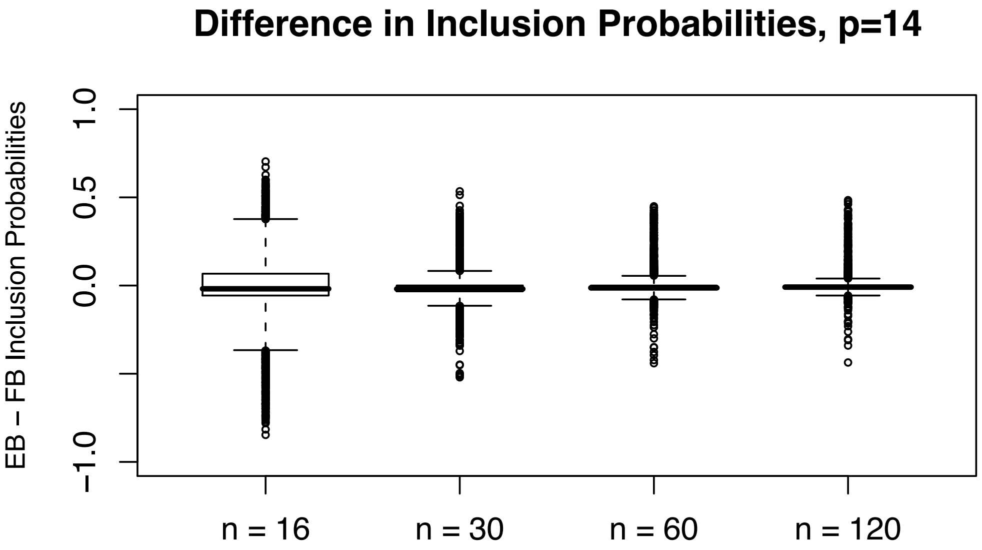


Figure 3.3: Differences in inclusion probabilities between EB and FB analyses in the simulation study.

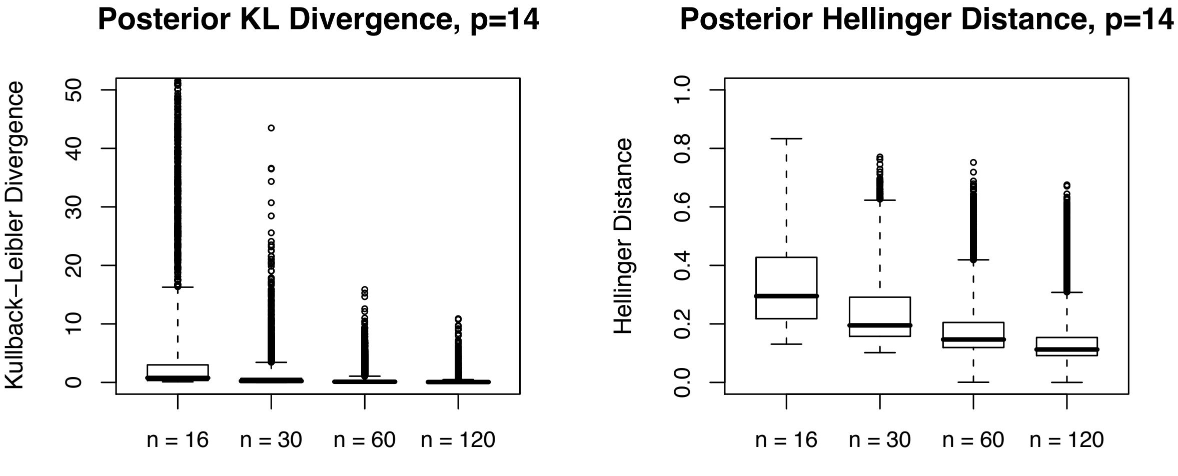


Figure 3.4: Realized KL divergence and Hellinger distance between FB and EB posterior distributions in the simulation study.

3. Draw $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$, the vector of regression coefficients corresponding to the nonzero elements of $\boldsymbol{\gamma}$, from a Zellner-Siow prior. Set the other coefficients $\boldsymbol{\beta}_{-\boldsymbol{\gamma}}$ to 0.

4. Draw a random vector of responses $\mathbf{Y} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{I})$.

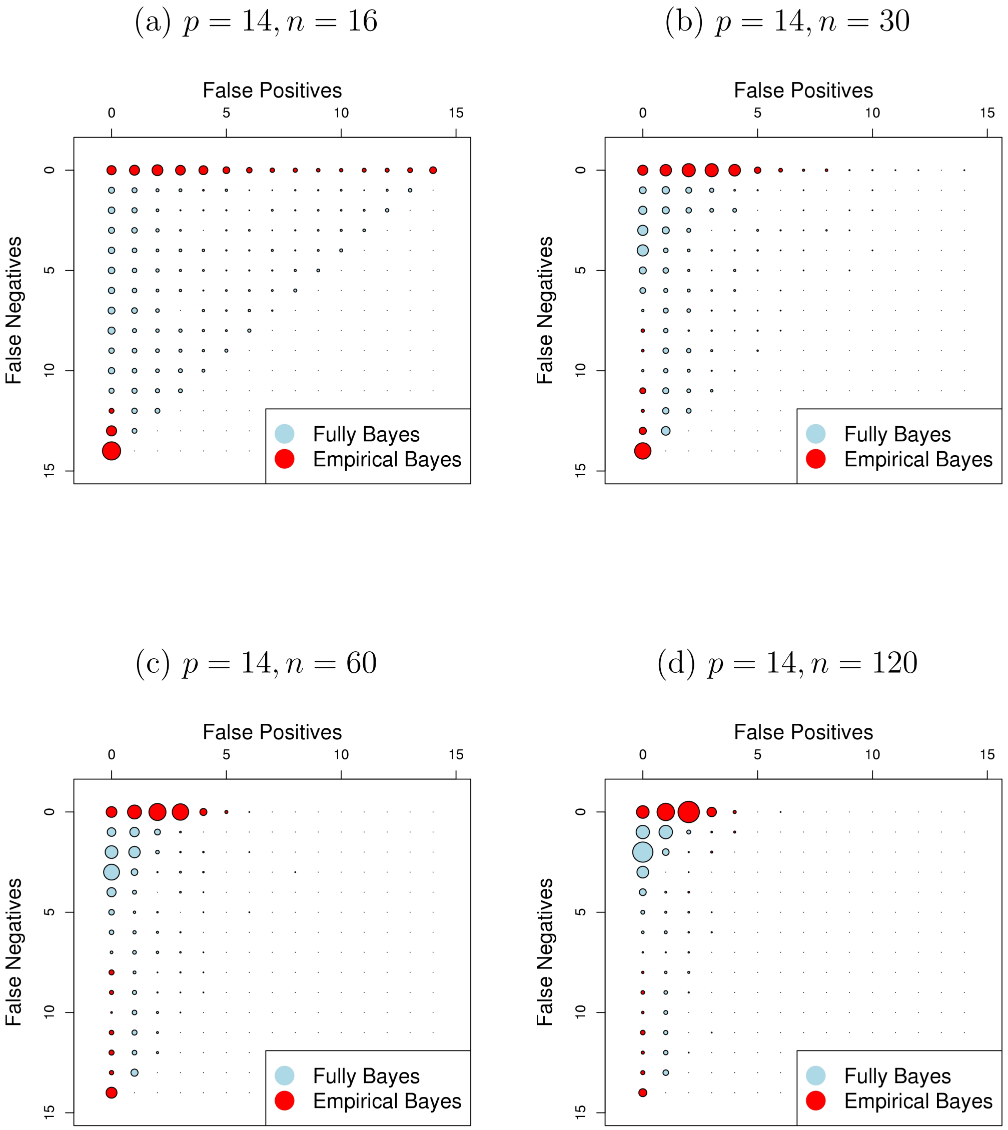
5. Compute marginal likelihoods (assuming Zellner-Siow priors) for all 2^p possible models; use these quantities to compute $\hat{w}$ along with the EB and FB posterior distributions across model space.

In all cases p 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. I repeated the experiment for four different sample sizes ($n = 16$, $n = 30$, $n = 60$, and $n = 120$) to simulate a variety of different p/n ratios.

Three broad patterns emerged from these experiments. First, the two procedures often reached very different conclusions about which covariates were important. Figure 3.3 shows frequent large discrepancies between the posterior inclusion probabilities given by the EB and FB procedures. This happened even when n was relatively large compared to the number of parameters being tested, suggesting that even large sample sizes do not render a data set immune to this difference.

Second, Figure 3.4 shows that, even when $n = 120$, the EB procedure can differ significantly from the FB posterior as measured by KL divergence.

Finally, both procedures produce plenty of mistakes when classifying variables as being in or out of the model, but these mistakes differ substantially in their overall character. For each simulated data set, the number of false positives and false negatives identified by the EB and FB median-probability models were recorded. These two numbers give an (x, y) pair that can then be plotted (along with the pairs from



(a) $p = 14, n = 16$

(b) $p = 14, n = 30$

(c) $p = 14, n = 60$

(d) $p = 14, n = 120$

Figure 3.5: Orthogonal design, $p = 14$. The differential pattern of errors under the 75,000 simlated data sets. The area of the circle represents how overrepresented the given procedure (red for EB, blue for FB) is in that cell. The circle at (3 right, 2 down), for example, represents an error involving 3 false positives and 2 false negatives.

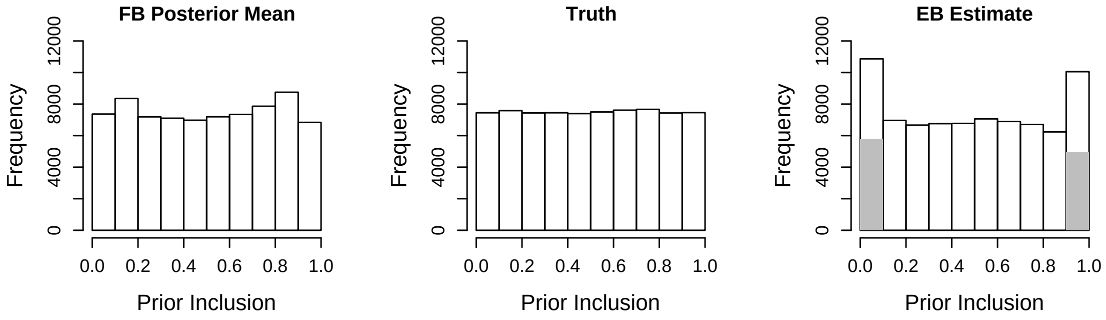


Figure 3.6: Distribution of $\hat{w}$ in the simulation study with a correctly specified (uniform) prior for w. The grey bars indicated the number of times, among values of $\hat{w}$ in the extremal bins, that the empirical-Bayes solution collapsed to the degenerate $\hat{w} = 0$ or $\hat{w} = 1$.

all other simulated data sets) to give a graphical representation of the kind of mistakes each procedure produces under repetition. The four panes of Figure 3.5 show these plots for all four sample sizes. Each integer (x, y) location contains a circle whose color—red for EB, blue for FB—shows which procedure produced that kind of mistake more often, and whose area shows how much more often it produced that mistake.

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

Much of this overshooting can be explained by Figure 3.6. The EB procedure gives the degenerate $\hat{w} = 0$ or $\hat{w} = 1$ solution much too often—over 15% of the time even when n is fairly large—suggesting that the issues raised by Theorem 3.3.4 can be quite serious in practice.

3.4.2 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. As a criticism of empirical-Bayes analysis, however, there was an obvious bias: the fully Bayesian procedure was being evaluated under its true prior distribution, with respect to which it is necessarily optimal.

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

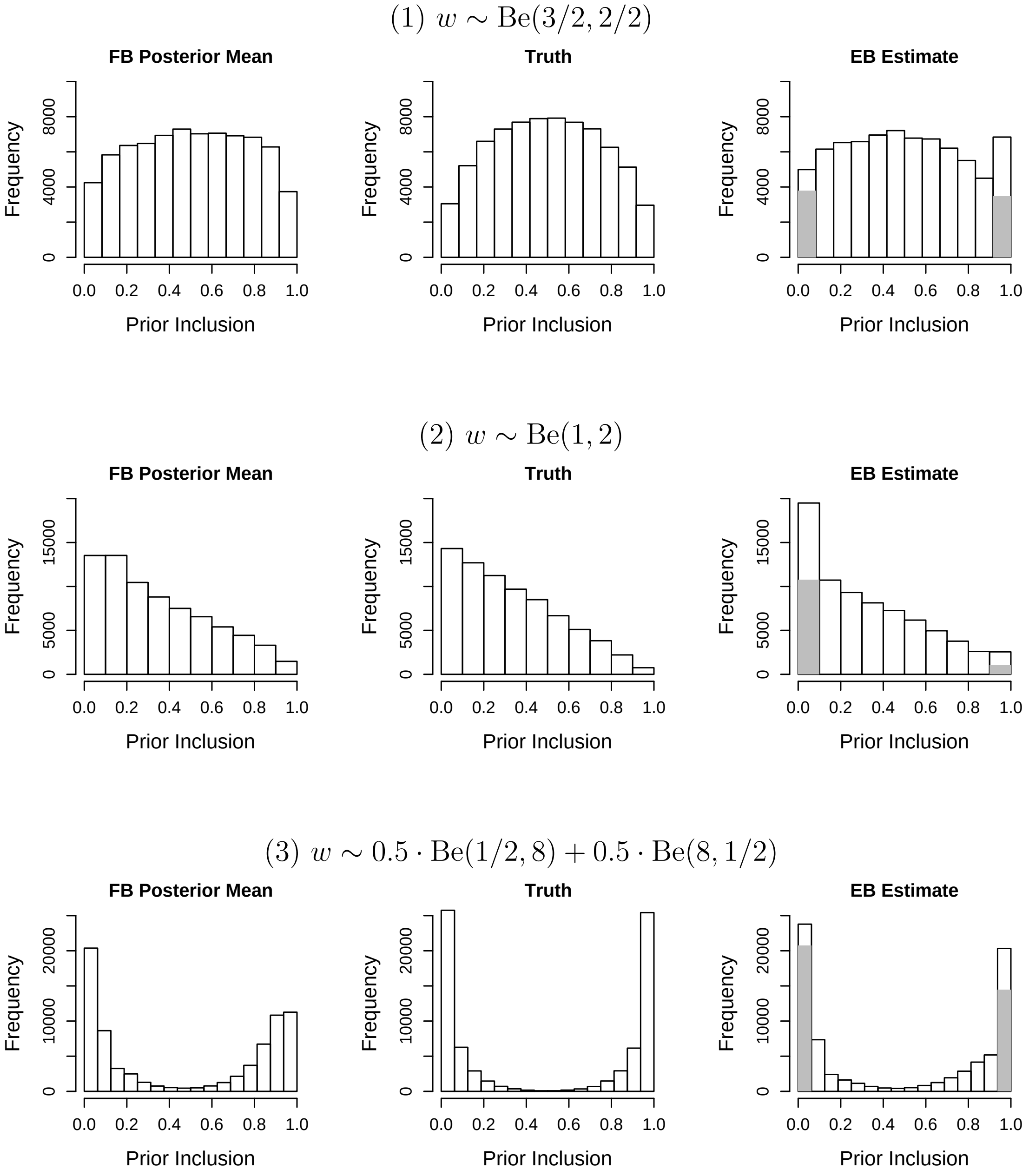
1. $w \sim \text{Be}(3/2, 3/2)$, yielding mainly moderate (but not uniform) values of w.

2. $w \sim \text{Be}(1, 2)$, yielding mainly smaller values of w.

3. $w \sim 0.5 \cdot \text{Be}(1/2, 8) + 0.5 \cdot \text{Be}(8, 1/2)$, yielding primarily values of w close to 0 or 1.

The results are summarized in Figure 3.7. In each case the central pane shows the true distribution of w, with the left pane showing the Bayesian posterior means under the uniform prior and the right pane showing the empirical-Bayes estimates $\hat{w}$.

As expected, the incorrectly specified Bayesian model tends to shrink the estimated values of w 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 w. This gives the illusion that empirical-Bayes tends to do better here.



(1) $w \sim \mathrm{Be}(3/2, 2/2)$

(2) $w \sim \mathrm{Be}(1, 2)$

(3) $w \sim 0.5 \cdot \mathrm{Be}(1/2, 8) + 0.5 \cdot \mathrm{Be}(8, 1/2)$

Figure 3.7: Distribution of $\hat{w}$ in different versions of the simulation study, where the fully Bayesian model had a misspecified (uniform) prior on p. The grey bars indicated the number of times, among values of $\hat{w}$ in the left- and right-most bins, that the empirical-Bayes solution collapsed to the degenerate $\hat{w} = 0$ or $\hat{w} = 1$.

Notice, however, the grey bars in the right-most panes. These bars indicate the percentage of time, among values of $\hat{w}$ 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 $\hat{w} \in [0, 0.1)$ 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 3.7 shows that, paradoxically, where the fully Bayesian model is most incorrect, its advantages over the empirical-Bayes procedure are the strongest. In the mixture model where w often took values very close to 0 or 1 (as would not be expected under a uniform prior for w), the empirical-Bayes procedure collapses to a degenerate solution nearly half the time. Even if the null (or full) model is true in most of the 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. Moreover, 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.

3.4.3 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 p/n ratio.

Table 3.3: Exact (to 3 decimal places) 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 1960–90	0.164	0.095
Defense Spending Share	0.156	0.085
Civil Liberties	0.154	0.075
Average Inflation 1960–90	0.150	0.064
Real Exchange Rate Distortions	0.146	0.071
Interior Density	0.139	0.067

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 $w = 1/2$ prior) on these regressions.

I consider 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 marginal likelihoods for different values of w 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 3.3. 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 were used for model selection, quite different results would emerge. For instance, if median-probability models were selected (i.e., one includes only those variables with inclusion probability greater than 1/2), the FB analysis would include the first four variables (and would almost choose the fifth variable), while the EB analysis would select only the first two variables (and almost the third). While there are better options that simply choosing a model outright, note that doing so would result in fundamentally different economic pictures for the FB and EB analysis.

Finally, the EB procedure gives different answers 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. Clearly the issue with empirical Bayes—indeed, with any plug-in choice of w—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.

3.5 Summary

The investigations described in this chapter started out as an attempt to more fully understand when, and how, multiplicity correction automatically occurs in Bayesian variable selection, 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 w.

The resolution to this conflict—that the multiplicity correction is indeed pre-fixed in the prior probabilities, but the amount of correction employed will depend on the data—led to another conflict: how can the empirical-Bayes approach to variable selection be an accurate approximation to the full Bayesian analysis?

Indeed, this chapter has shown that empirical-Bayes variable selection can lead to results that are quite different from those under the full Bayesian analysis. This difference has been demonstrated 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 $\hat{w}$, all supported the general conclusions about empirical-Bayes variable selection that were mentioned in the beginning of the chapter.

Of course, there are many fine empirical-Bayes analyses that have been done in model selection and variable selection, and it would be unfair and simplistic to say that any such analysis is wrong. These results do suggest, however, that empirical-Bayes variable selection does not carry the automatic guarantee of performance that accompanies empirical-Bayes methodology in many other contexts, and so additional care should be taken in its use.

Chapter 4

Multiplicity Adjustment in Gaussian Graphical Models

Gaussian graphical models are tools for modeling conditional independence relationships, and they offer many practical advantages in high-dimensional problems. They can make computing more efficient by alleviating the need to handle large matrices; they can yield better predictions by fitting sparser models; and they can aid scientific understanding by breaking down a global model into a collection of local models that are easier to parse.

Yet often the graph itself must be inferred from the data, meaning that two quantities must be specified: the prior distribution for Σ under each graph, and a prior distribution over different possible graphs.

The first specification is difficult because there is no common covariance matrix shared by all graphs, but rather an entire collection of covariance matrices $\{\Sigma_G\}$ indexed by all possible graphs. Different graphs imply different numbers of free elements in Σ, and so it is not possible to use an improper prior for each Σ_G as one might do for covariance estimation under a fixed graph, since this would leave the resulting model probabilities defined only up to an arbitrary constant. Instead, one must either elicit a subjective prior for each Σ_G (clearly intractable in high dimensions), or else choose some default proper prior that is neither too vague nor too precise. Regardless, it is clear that any answer will depend on the priors chosen for the various Σ_G's. This chapter will introduce an objective-Bayesian approach using fractional Bayes factors to handle this difficulty.

The second task, specifying a prior across different graphs, involves many of the same issues that are encountered in specifying priors for variable selection. As in linear models, there is a substantial multiplicity issue: graphical model selection is a problem of simultaneously testing many pairwise conditional independence relationships, each of which is its own null hypothesis. In fact, since graphical model selection is nothing but an ensemble variable-selection problem for a set of related linear models under an assumption of joint Gaussianity, the parallels with the previous chapter are more than metaphorical. As with linear models, the seemingly objective choice of assigning all graphs equal prior probability will be shown to flag many false-positive edges, and this chapter shows how a familiar class of fully Bayesian edge-selection priors can avoid this problem.

4.1 Notation

An undirected graph is a pair $G = (V, E)$ with vertex set V and edge set $E = \{(i, j)\}$ for some pairs $(i, j) \in V$. Nodes i and j are adjacent, or neighbors, if $(i, j) \in E$. Complete graphs are those having $(i, j) \in E$ for every $i, j \in V$. Maximal complete subgraphs $C \subset V$ are called cliques; two cliques that overlap in a set S are said to have S as a separator. A decomposition is a recursive partitioning of a graph G into subgraphs (A, S, B) such that $V = A \cup B$, $S = A \cap B$ is complete, and any path from a node in A to a node in B goes through the separator S. All graphs in this thesis are assumed to be decomposable (unless otherwise noted), thus admitting a decomposition into a sequence of cliques. Denote the set of cliques and separators of a graph by $\mathcal{C}$ and $\mathcal{S}$ respectively. A decomposable graph G can be represented by a perfect ordering of cliques and separators. An ordering of cliques $C_i \in \mathcal{C}$ and separators $S_i \in \mathcal{S}$ is said to be perfect if for every $i = 2, \dots, |V|$ there exists a $j < i$ such that

$$S_i = C_i \cap H_{i-1} \subset C_j,$$

where

$$H_{i-1} = \bigcup_{j=1}^{i-1} C_j.$$

Finally, a perfect numbering of vertices in G is obtained by taking first the vertices in C_1, then those in $C_2 \setminus H_1$, $C_3 \setminus H_2$, and so on (Lauritzen, 1996, Lemma 2.12). Taking this order in reverse creates what is called a perfect vertex-elimination scheme of G.

A Gaussian graphical model uses a graphical structure to define a set of pairwise conditional-independence relationships on a p-dimensional zero-mean, normally distributed random vector $x \sim \mathrm{N}(0, \Sigma)$. The unknown covariance matrix Σ is restricted by its Markov properties; given $\Omega = \Sigma^{-1}$, elements x_i and x_j of the vector x are conditionally independent, given their neighbors, if and only if $\Omega_{ij} = 0$. If $G = (V, E)$ is an undirected graph describing the joint distribution of x, $\Omega_{ij} = 0$ for all pairs $(i, j) \notin E$. The covariance matrix Σ is in $M^+(G)$, the set of all symmetric positive-definite matrices having elements in Σ^{-1} set to zero for all $(i,j) \notin E$.

The hyper-inverse Wishart distribution is a general class of hyper-Markov laws introduced by Dawid and Lauritzen (1993) for a covariance matrix $\Sigma \in M^+(G)$, where $G = (V, E)$ is a decomposable graph. The notation is $(\Sigma \mid G) \sim \mathrm{HIW}_G(b, D)$, where $b \in \mathbb{R}^+$ is a degrees-of-freedom parameter, and where $D \in M^+(G)$ is a symmetric positive-definite scale matrix. The density of this distribution is defined with respect to the product of Lebesgue measures for the (i,j) elements of Σ for which $(i,j) \in E$, subject to the conditions that Σ_C is symmetric and positive definite for all $C \in \mathcal{C}$.

The density of Σ can be obtained from the clique-specific marginal densities as a ratio of products over cliques and separators:

$$p(\Sigma \mid G) = \frac{\prod_{C \in \mathcal{C}} p(\Sigma_C \mid b, D_C)}{\prod_{S \in \mathcal{S}} p(\Sigma_S \mid b, D_S)}, \tag{4.1}$$

where, for each clique $C \in \mathcal{C}$ (and separators $S \in \mathcal{S}$), $\Sigma_C \sim \mathrm{IW}(b, D_C)$ with density

$$p(\Sigma_C \mid b, D_C) \propto |\Sigma_C|^{-(b/2+|C|)} \exp\left\{-\frac{1}{2}\mathrm{tr}\left(\Sigma_C^{-1} D_C\right)\right\}. \tag{4.2}$$

The factorization in (4.1) holds assuming that, if $S = C_1 \cap C_2$, then the elements of Σ_S are common in Σ_{C_1} and Σ_{C_2}. For further details and explanations, refer to Dawid and Lauritzen (1993), Giudici and Green (1999) and Letac and Massam (2007).

Next, suppose one observes n samples $(x_1, \ldots, x_n)$ of p-dimensional vectors, where each $x_i \sim \mathrm{N}(0, \Sigma)$, with unknown covariance matrix Σ. Let X be the $n \times p$ matrix of samples and let X_j refer to column j of X and x_i to row i; also, let X_C refer to the columns of X corresponding to the nodes in clique C; and assume that $\Sigma \in M^+(G)$ for some unknown decomposable graph G on p nodes. The posterior probability of a graph G is

$$p(G \mid X) \propto p(G) \int_{\Sigma \in M^{+}(G)} p(X \mid \Sigma, G) \cdot p(\Sigma \mid G) \, \mathrm{d}\Sigma \,, \tag{4.3}$$

where $p(G)$ is the prior probability of the graph, and where the integral is the marginal likelihood of the data under G. If $\Sigma \sim \mathrm{HIW}_G(b, D)$, the integral in (4.3) is available in closed form using the ratio of the prior and posterior normalizing constants:

$$p(X \mid G) = (2\pi)^{-np/2} \frac{h(G, b, D)}{h(G, b^{*}, D^{*})} \,, \tag{4.4}$$

where $b^{*} = b + n$ and $D^{*} = D + X'X$. The normalizing constant $h(\cdot)$ is

$$h(G, b, D) = \frac{\prod_{C \in \mathcal{C}} \left| \frac{1}{2} D_C \right|^{\frac{(b+|C|-1)}{2}} \Gamma_{|C|} \left(\frac{b+|C|-1}{2} \right)^{-1}}{\prod_{S \in \mathcal{S}} \left| \frac{1}{2} D_S \right|^{\frac{(b+|S|-1)}{2}} \Gamma_{|S|} \left(\frac{b+|S|-1}{2} \right)^{-1}} \,, \tag{4.5}$$

where $\Gamma_p(x) = \pi^{p(p-1)/4} \cdot \prod_{j=1}^{p} \Gamma(x + (1-j)/2)$ is the multivariate gamma function.

4.2 Model-selection priors for restricted covariance matrices

4.2.1 Criteria for model-selection priors

The expression for the marginal likelihood in (4.3) involves an integral over the prior for Σ under the graph G. This integral will typically be very sensitive to different choices of the prior (as discussed in Jones et al., 2005), which, as has been highlighted in previous chapters, is a general phenomenon in model selection.

In all but the smallest of problems, $p(\Sigma \mid G)$ must be a conjugate hyper-inverse Wishart prior; otherwise it will not be possible to make use of (4.4) for computing marginal likelihoods. Other priors will require approximating the integrals in (4.3), and in such a large model space, the need to do so repeatedly will usually pose an insurmountable obstacle.

"Well-behaved" is somewhat difficult to judge, since distributions over the space of constrained covariance matrices are not very intuitive. Priors for regression parameters, however, are—and it is typically easier to assess priors for graphically constrained covariance matrices by studying the properties of the priors they induce on all implied conditional regression models. This point will be developed in Section 4.4.

4.2.2 A conventional proper prior

The most popular choice of prior for use in Gaussian graphical models is $\Sigma \sim \mathrm{HIW}_G(\delta, \tau I)$, where the scale matrix $D = \tau I$ is proportional to the identity matrix. This might be called the "conventional proper prior" for graphical model selection, by analogy with Berger and Pericchi (2001). Examples of these or similar priors being used to compute marginal likelihoods can be found in Giudici (1996), Giudici and Green (1999), Dobra et al. (2004), Jones et al. (2005), Atay-Kayis and Massam (2005), and Carvalho and West (2007), among others. The scale parameter τ must be chosen to match the expected scale of the data; if τ is too large, the prior for Σ under each graph will wash out the likelihood, and there will be no basis for discriminating among competing graphs.

The now-standard notation of the conventional proper prior can be confusing: a single scale matrix D may used to specify $p(\Sigma \mid G)$ for all graphs, but this scale matrix means different things under each graph, since different graphs imply different configurations of free elements in Σ. Under G, only the (i, j) elements of D for which the edge $(i, j) \in G$ are relevant in determining the distribution of Σ; other, nonfree, elements of Σ must be filled in using the (deterministic) completion operation described in Massam and Neher (1998); see also Dawid and Lauritzen (1993) and Carvalho et al. (2007). Hence the notation $\Sigma \sim \mathrm{HIW}_G(b, D)$ must be taken as convenient shorthand for the statement that Σ depends upon the free elements of D implied by the graph G.

4.2.3 The hyper-inverse Wishart g-prior

Given the practical need for conjugacy, consider another possible form of the hyper-inverse Wishart distribution, one where D involves the cross-product matrix:

$$(\Sigma \mid G) \sim \mathrm{HIW}_G(\delta, gX'X)\,,$$

where g is some suitably small fraction such as $1/n$.

This hyper-inverse Wishart g-prior, by analogy with Zellner's g-prior in linear regression (Zellner, 1986), provides an alternative to the conventional $\mathrm{HIW}(\delta, \tau I)$ for use in graphical model-selection. The similarity to the g-prior in regression is more than superficial, since this prior will be shown to induce a g-like prior for the univariate conditional regression models implied by G. This along with other theoretical and methodological properties will be examined in Section 4.4 and Section 4.5.

4.3 Fractional Bayes factors for Gaussian graphical models

Direct use of the hyper-inverse Wishart g-prior for selecting graphs is incoherent, since it involves a double use of the data. Luckily, the prior arises very naturally through the use of fractional Bayes factors, meaning that double use of the data can be avoided.

O'Hagan (1995) proposed Fractional Bayes factors as a default Bayesian model-selection technique for use when prior information is weak. The idea is to train a noninformative prior for each model using a small fractional power g of the likelihood function. This is done simultaneously for all models being considered, converting all noninformative priors into proper priors that are then used to select a model with the remainder of the likelihood.

Choose $g \in (0, 1)$ and let $p_N(\Sigma \mid G)$ be a noninformative (typically improper) prior for Σ under a decomposable graph G. The fractional Bayes factor for graphs G_1 and G_2 is then $\mathrm{FBF}_g(G_1, G_2) = Q_g(X \mid G_1)/Q_g(X \mid G_2)$, where

$$Q_g(X \mid G) = \frac{\int p_N(\Sigma \mid G) p(X \mid \Sigma, G) \, \mathrm{d}\Sigma}{\int p_N(\Sigma \mid G) p(X \mid \Sigma, G)^g \, \mathrm{d}\Sigma} \, . \tag{4.6}$$

Then $p^*(\Sigma \mid G) \propto p_N(\Sigma \mid G) \cdot p(X \mid \Sigma, G)^g$ is called the *implied fractional prior*, where the constant of proportionality is the integral of the given expression, and $p(X \mid \Sigma, G)^{1-g}$ is called the *implied fractional likelihood.*

Equation (4.6) clearly depends upon the choice of a noninformative prior for Σ. The obvious choice, given the need for conjugacy, is to simply take

$$p_N(\Sigma \mid G) \propto \frac{\prod_{C \in \mathcal{C}} |\Sigma_C|^{-|C|}}{\prod_{S \in \mathcal{S}} |\Sigma_S|^{-|S|}} \, , \tag{4.7}$$

for $\Sigma \in M^+(G)$, an improper prior that makes use of the same factorization over cliques and separators, and is defined with respect the same measure, as the hyper-inverse Wishart distribution. Interestingly, this prior has clique marginals that correspond to one of many versions of Fisher's fiducial prior (Sun and Berger, 2006), so called because it yields Fisher's fiducial distribution for marginal variances. These clique marginals, $p(\Sigma_C) \propto |\Sigma_C|^{-|C|}$, also have the same form as the priors that yield exact frequentist matching for means and variances (Geisser and Cornfield, 1963) when used for covariance estimation.

My goal here is not to make any analogous claims about frequentist matching for the prior in (4.7). Instead, I will use it because it is a usefully conjugate generalization of a familiar prior to the space of covariance matrices with graphical structure, and because it allows the following result:

Theorem 4.3.1. *The hyper-inverse Wishart g-prior* $(\Sigma \mid G) \sim \mathrm{HIW}_G(gn, gX'X)$ *is the implied fractional prior for* Σ *corresponding to the prior in (4.7), where* $0 < g < 1$ *is the fraction of the likelihood used for training. The fractional marginal likelihood is*

$$Q_g(X \mid G) = (2\pi)^{-np/2} \frac{h(G, gn, gX'X)}{h(G, n, X'X)}\,, \tag{4.8}$$

with $h(G, b, D)$ *defined as in (4.5).*

Proof. Follows immediately from the conjugacy of the hyper-inverse Wishart prior with the normal likelihood, and from Equation (4.4). □

This procedure does not merely specify a single prior distribution, but rather a whole cohort of objective prior distributions for all $\Sigma_G \in M^+(G)$.

An obvious issue with the use of this methodology is the choice of g. If the hyper-inverse Wishart g-prior were used as a real prior, it would be possible to place a hyper-prior on g, and not anchor the procedure to a specific value; indeed, Liang et al. (2008) recommend exactly this approach toward g-priors in linear-model selection.

Yet when interpreting the hyper-inverse Wishart g-prior in terms of fractional Bayes factors, it is no longer possible to put a prior on g. This is because g is not a model parameter about which there is information in the likelihood, but rather the fractional power of the likelihood itself used for training the noninformative prior in (4.7). This fraction must be chosen outright in order for the fractional marginal likelihoods in (4.8) to be well-defined.

Several criteria help to guide this choice. First, there is an established tradition of using "minimal training sample" sizes to calibrate default Bayes factors; see, for example, O'Hagan (1995), Berger and Pericchi (1996), and Berger and Pericchi (2001). A minimal training sample is the smallest sample size needed to convert all improper priors such as (4.7) into proper priors. The intuition is that as much of the data as possible should be held back to choose between models. It is easy to see from (4.1) and (4.2) that the minimal training sample size is 1, suggesting that g be $1/n$.

Second, it is clear that g must be $O(1/n)$ in order for the implied fractional prior to correspond asymptotically to the gold standard of a carefully elicited subjective prior distribution. If g decreases too slowly as a function of n, the implied fractional prior will asymptotically overwhelm the likelihood; if it decreases too fast, the prior will become arbitrarily diffuse. Neither behavior could possibly result from the choices of a careful elicitee making intelligent decisions about each $p(\Sigma \mid G)$. (This hypothetical "careful elicitation" a useful ideal to keep in mind, even if dimensionality makes this ideal impossible to attain.)

Finally, the implied fractional prior for Σ should have heavy tails. Choosing $gn = 1$ implies that the vector x is marginally Cauchy, and that each $p(\Sigma \mid G)$ is heavy-tailed without being too vague. This choice dovetails with the advice given by Liang et al. (2008), who themselves generalize the recommendations of Jeffreys (1961) and Zellner and Siow (1980).

As a default choice, I recommend setting $g = 1/n$, though this is not a hard rule, and other choices that decay like $1/n$ may be reasonable (and can be judged by the reasonableness of their effect on the implied fractional prior). Robustness to these choices should be considered, just as it should be in subjective analyses.

4.4 Properties of the hyper-inverse Wishart g-prior

4.4.1 Information consistency

The consistency of fractional Bayes factors as $n \to \infty$ is a well-known result from O'Hagan (1995). This section considers a second notion of consistency, often called *information consistency* or *finite-sample consistency*, that describes how a Bayes factor behaves for fixed n with respect to a test statistic that would be used to perform a classical test of significance on the same problem.

The canonical example of an information-inconsistent procedure is model selection in linear regression using fixed-g versions of Zellner's g-prior (Zellner and Siow, 1980; Liang et al., 2008). Imagine testing a specific regression model M_A having k possible covariates against the null model M_0 having only an intercept term. If the usual F statistic for testing M_A against M_0 goes to infinity for fixed n and $k < n - 1$, the evidence against M_0 is overwhelming, and one would expect the Bayes factor $\mathrm{BF}(M_A : M_0)$ to diverge.

But under the standard g-prior, this Bayes factor instead converges to the fixed constant $(1 + g)^{(n-k-1)/2}$. This gives an intrinsic limitation (one not shared by the F statistic) to how strongly the Bayes factor may support the bigger model. Such behavior is intuitively unappealing, since $\mathrm{pr}(F > C \mid M_0) \to 0$ as $C \to \infty$.

A natural question is whether the fractional Bayes factors defined above exhibit a similar information paradox. They do not, in two related senses.

4.4.2 Tests against the null graph

Let G_0 denote the null graph having no edges, and let G_A denote the graph to be compared with the null. The Bayes factor for comparing these two models is

$$\begin{aligned} \mathrm{BF}(G_0 : G_A) = K \quad &\cdot \quad \frac{\prod_{j=1}^{p} \left| \frac{g}{2} X_j' X_j \right|^{\frac{gn}{2}}}{\prod_{j=1}^{p} \left| \frac{1}{2} X_j' X_j \right|^{\frac{n}{2}}} \cdot \frac{\prod_{S \in \mathcal{S}} \left| \frac{g}{2} X_S' X_S \right|^{\frac{gn+|S|-1}{2}}}{\prod_{C \in \mathcal{C}} \left| \frac{g}{2} X_C' X_C \right|^{\frac{gn+|C|-1}{2}}} \\ &\cdot \quad \frac{\prod_{C \in \mathcal{C}} \left| \frac{1}{2} X_C' X_C \right|^{\frac{n+|C|-1}{2}}}{\prod_{S \in \mathcal{S}} \left| \frac{1}{2} X_S' X_S \right|^{\frac{n+|S|-1}{2}}} \, , \end{aligned} \tag{4.9}$$

where g is fixed, $\mathcal{C}$ and $\mathcal{S}$ are the cliques and separators of G_A, and the leading term K is

$$K = \left[\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{gn}{2}\right)} \right]^p \cdot \frac{\prod_{S \in \mathcal{S}} \Gamma_{|S|} \left(\frac{n+|S|-1}{2} \right)}{\prod_{C \in \mathcal{C}} \Gamma_{|C|} \left(\frac{n+|C|-1}{2} \right)} \cdot \frac{\prod_{C \in \mathcal{C}} \Gamma_{|C|} \left(\frac{gn+|C|-1}{2} \right)}{\prod_{S \in \mathcal{S}} \Gamma_{|S|} \left(\frac{gn+|S|-1}{2} \right)} \, .$$

It remains to define a suitable test statistic b as a basis for assessing information consistency. Following Lauritzen (1996), let $\widehat{\Omega}_0$ be the maximum-likelihood estimate for the precision matrix under G_0, and let $\widehat{\Omega}_A$ be the maximum-likelihood estimate under G_A. Then there is a nested sequence $G_0 \subset \ldots \subset G_d = G_A$ of decomposable graphs that differ only by a single edge. Let e_i denote the edge in G_i but not in G_{i-1}, and let C_i be the (unique) clique of G_i containing e_i. This sets the groundwork for the following proposition (see Proposition 5.14 of Lauritzen, 1996, for a proof):

Proposition 4.4.1. *The test of significance for G_A against the null graph G_0 can be performed by rejecting G_0 for sufficiently small values of $b = |\widehat{\Omega}_0|/|\widehat{\Omega}_A|$. Under G_0, b is distributed as the product of independent beta random variables $B_1 \cdots B_d$, with $B_i \sim Be\{(n - |C_i|)/2, 1/2\}$.*

This defines the relevant test statistic b, which allows a precise statement of information consistency for the fractional Bayes factors in Theorem 4.3.1.

Theorem 4.4.2. *Let G_A be a decomposable graph having cliques $\mathcal{C}$, let G_0 be the null graph, and let $FBF_g(G_A : G_0)$ be the fractional Bayes factor, given data X, corresponding to the noninformative prior in (4.7). For any finite $n > \max_{C \in \mathcal{C}} |C|$ and for any $0 < g < 1$, $FBF_g(G_0 : G_A) \to 0$ as $b \to 0$.*

Proof. The Bayes factor in (4.9) simplifies to

$$K \cdot \left(\frac{1}{g}\right)^{(S_1 - gnp)/2} \left(\frac{1}{2}\right)^{(n-gn)(S_2-p)/2} \cdot \prod_{j=1}^{p} \left|X_j' X_j\right|^{-(n-gn)/2} \cdot \frac{\prod_{C \in \mathcal{C}} |X_C' X_C|^{(n-gn)/2}}{\prod_{S \in \mathcal{S}} |X_S' X_S|^{(n-gn)/2}} \,,$$

where the exponent terms S_1 and S_2 are

$$\begin{aligned} S_1 &= \sum_{C \in \mathcal{C}} |C| \cdot (gn + |C| - 1) - \sum_{S \in \mathcal{S}} |S| \cdot (gn + |S| - 1) \\ S_2 &= \sum_{C \in \mathcal{C}} |C| - \sum_{S \in \mathcal{S}} |S| \,. \end{aligned}$$

Now apply the formula of Lauritzen (1996) for the determinant of $\widehat{\Omega}_G$, which will exist due to the restriction that $n > \max_{C \in \mathcal{C}} |C|$:

$$|\widehat{\Omega}_G| = n^p \frac{\prod_{S \in \mathcal{S}} |X_S' X_S|}{\prod_{C \in \mathcal{C}} |X_C' X_C|} \,.$$

This gives

$$\mathrm{BF}(G_0 : G_A) = C \cdot \left(\frac{|\widehat{\Omega}_0|}{|\widehat{\Omega}_A|} \right)^{(n-gn)/2} ,$$

where C is a fixed, finite term involving g, p, n, and the structure of the graph G_A. The proof of information consistency now follows immediately by plugging the test statistic b into the above equation, and noticing that for fixed $0 < g < 1$, $\mathrm{BF}(G_0 : G_A) \to 0$ as $b \to 0$. □

4.4.3 Tests for an implied conditional regression model

Information consistency is important in a second sense: nonzero entries in a precision matrix imply a set of nonzero conditional regression coefficients for each element of x upon the other elements, and Bayes factors for model selection in Gaussian graphical models perform variable selection on all of these implied regressions simultaneously. Observing the behavior of these implied conditional regression models provides a useful window on the behavior of $p(\Sigma \mid G)$, which is far harder to understand intuitively.

The following lemma provides a characterization of the implied conditionals.

Lemma 4.4.3. *Suppose $(x \mid \Sigma) \sim N(0, \Sigma)$ and $\Sigma \sim \mathrm{HIW}_G(b, D)$ for some decomposable graph G. Suppose $x = (z, y)'$ where z is a scalar and the vertices are numbered following the perfect vertex elimination scheme of G, with z as the first vertex. Let Σ and D be partitioned as*

$$\Sigma = \begin{pmatrix} \Sigma_{zz} & \Sigma_{zy} \\ \Sigma_{yz} & \Sigma_{yy} \end{pmatrix}, \quad D = \begin{pmatrix} D_{zz} & D_{zy} \\ D_{yz} & D_{yy} \end{pmatrix} .$$

Then:

(i) $\Sigma_{z|y}^{-1} = \left(\Sigma_{zz} - \Sigma_{zy}\Sigma_{yy}^{-1}\Sigma_{yz}\right)^{-1} \sim Ga\left(\frac{b+k}{2}, \frac{D_{z|y}}{2}\right)$

(ii) $\left(\Sigma_{zy}\Sigma_{yy}^{-1} \mid \Sigma_{z|y}\right) \sim N\left(D_{zy}D_{yy}^{-1}, \Sigma_{z|y}D_{yy}^{-1}\right)$

with k representing the number of neighbors of z under G.

Proof. Let $\Omega = \Sigma^{-1} = \Phi'\Phi$ be partitioned as:

$$\begin{pmatrix} \Omega_{zz} & \Omega_{zy} \\ \Omega_{yz} & \Omega_{yy} \end{pmatrix} = \begin{pmatrix} \Phi'_{zz} & 0 \\ \Phi_{yz} & \Phi'_{yy} \end{pmatrix} \begin{pmatrix} \Phi_{zz} & \Phi_{zy} \\ 0 & \Phi_{yy} \end{pmatrix}.$$

Recall that $(\Sigma_{zz} - \Sigma_{zy}\Sigma_{yy}^{-1}\Sigma_{yz})^{-1} = \Sigma_{z|y}^{-1} = \Omega_{zz} = \Phi'_{zz}\Phi_{zz}$ and $\Sigma_{zy}\Sigma_{yy}^{-1} = -\Phi_{zz}^{-1}\Phi_{zy} = \Gamma_{z|y}$. Theorem 1 of Roverato (2000) (see also Paulsen et al., 1989; Wermuth, 1980) shows that if G is decomposable and the vertices are listed in a perfect vertex elimination scheme, the pattern of zeros of Ω are preserved in Φ. Now, if $(\Sigma|G) \sim \text{HIW}_G(b, D)$, properties of the Cholesky decomposition of the hyper-inverse Wishart as defined in Theorem 3 of Roverato (2000) (see also equation 27 of Atay-Kayis and Massam, 2005) allow us to write $\Psi = \Phi T^{-1}$ where $D^{-1} = T'T$ with

$$\Psi_{zz}^2 \sim \text{Ga}\left(\frac{b+k}{2}, \frac{1}{2}\right) \quad \text{and} \quad \Psi_{zy_i} \sim \text{N}(0,1)\,, \tag{4.10}$$

for all y_i in the neighborhood of z. For the non neighbors of z in y, $\Phi_{zy_i} = \Psi_{zy_i} = 0$.

It is now straightforward to see from (4.10) that

$$\Omega_{zz} = \Phi'_{zz}\Phi_{zz} = \Phi_{zz}^2 = (\Psi_{zz}T_{zz})^2 \sim \text{Ga}\left(\frac{b+k}{2}, \frac{T_{zz}^{-2}}{2}\right)\,,$$

so that

$$\Sigma_{z|y}^{-1} \sim \text{Ga}\left(\frac{b+k}{2}, \frac{D_{z|y}}{2}\right)\,,$$

which proves part (i) of the Lemma. Turning the focus to the form of $\Gamma_{z|y} = \Sigma_{zy}\Sigma_{yy}^{-1} =$

$-\Phi_{zz}^{-1}\Phi_{zy}$ and writing it as a function of Ψ and T, one gets

$$\gamma_i = -\left(\frac{T_{y_i z}}{T_{zz}} + \frac{1}{\Psi_{zz}T_{zz}}\sum_{j=1}^{i}\Psi_{zy_j}T_{y_j y_i}\right) . \tag{4.11}$$

Given Φ_{zz}, this is just a linear combination of independent standard normals, so that

$$(\Gamma|\Phi_{zz}) \quad \sim \quad \mathrm{N}\left(-T_{zz}^{-1}T_{zy}, \frac{1}{\Phi_{zz}^2}T_{yy}'T_{yy}\right) \tag{4.12}$$

$$(\Sigma_{zy}\Sigma_{yy}^{-1}|\Sigma_{z|y}^{-1}) \quad \sim \quad \mathrm{N}\left(D_{zy}D_{yy}^{-1}, \Sigma_{z|y}D_{yy}^{-1}\right) , \tag{4.13}$$

proving part (ii) of the Lemma. □

This lemma leads to the following theorem, which is intended to be understood as if the implied fractional prior were a true prior, and the implied fractional likelihood were a true likelihood. Since the fractional prior has the exact functional form of a (proper) hyper-inverse Wishart prior, and the fractional likelihood has the functional form of a Gaussian likelihood, there is no mathematical ambiguity in operationally defining "prior" and "likelihood" this way, even if the interpretation is difficult from a pure subjectivist viewpoint.

Theorem 4.4.4. *Let $X = (z\ Y)$ be the $n \times p$ matrix of observed data having rows $x_i = (z_i, y_i)$. Let G_A be a decomposable graph having cliques $\mathcal{C}$, and let M_A be the conditional regression model for z in terms of Y implied by the neighbors of z in G_A, and let M_0 be the null regression model for z. Assume further, that the vertices are listed in the perfect vertex elimination scheme of G_A, with z as the first vertex. Let F be the usual F-statistic for testing M_A against M_0, and let $BF_g(M_A : M_0)$ be the likelihood ratio $p(z|Y, M_A)/p(z|Y, M_0)$, where these marginals are defined under the fractional prior and the fractional likelihood in (4.6) for a fixed g. For any finite $n > \max_{C\in\mathcal{C}} |C|$ and for any $0 < g < 1$, $F \to \infty$ implies $BF_g(M_A : M_0) \to \infty$.*

Proof. Let Y_z denote the columns of the matrix X corresponding to the neighbors of z under M_A. Applying Lemma 4.4.3 under the assumption that $n > \max_{C \in \mathcal{C}} |C|$, the hyper-inverse Wishart g-prior gives the following regression relationship:

$$\begin{aligned} z &= Y_z f + \epsilon, \quad \epsilon \sim \mathrm{N}(0, \phi^{-1} I) & (4.14) \\ (f \mid \phi, M_A) &\sim \mathrm{N}\left(\hat{f}, (g\phi)^{-1}(Y_z' Y_z)^{-1}\right) & (4.15) \\ (\phi \mid M_A) &\sim \mathrm{Ga}\left(\frac{gn+k}{2}, \frac{gr}{2}\right), & (4.16) \end{aligned}$$

where ϕ is the conditional precision or $\Sigma_{z|Y_z}^{-1}$; I is the $n \times n$ identity matrix, $\hat{f}' = (Y_z' Y_z)^{-1} Y_z' z$ is the traditional least-squares estimate for f; and $r = z'(I - P_{Y_z})z$ with P_{Y_z} denoting the perpendicular projection matrix onto the column space of Y_z. Hence r is the residual sum of squares after regressing z upon Y_z.

Marginalizing over f and ϕ, taking care to use the fractional likelihood rather than the full likelihood, gives

$$P(z \mid Y, M_A) = (\pi)^{-n/2} g^{\frac{gn+2k}{2}} (1-g)^{n/2} \cdot \frac{\Gamma\left(\frac{n+gn+k}{2}\right)}{\Gamma\left(\frac{gn+k}{2}\right)} \cdot r^{-n/2} . \qquad (4.17)$$

Assuming $0 < g < 1$, the relevant Bayes factor can then be computed by recognizing the null model M_0 as a special case of (4.17) with $k = 0$ and $r = z'z$:

$$\mathrm{BF}(M_A : M_0) = C\left(1 - R^2_{M_A}\right)^{-n/2} , \qquad (4.18)$$

where C is a fixed term involving g, n, and k, and where $R^2_{M_A}$ is the usual coefficient of determination for model M_A. As the F-statistic grows without bound, $R^2_{M_A} \to 1$, and the Bayes factor in (4.18) clearly diverges. □

The hierarchical model in (4.14), (4.15), and (4.16) is immediately recognizable as a modified form of Zellner's g-prior for the vector of conditional regression coefficients. Unlike the g-prior, however, this procedure avoids the information paradox.

4.4.4 Remarks

One possible source of concern is that the priors in Equations (4.15)–(4.16) are centered at their maximum-likelihood estimates. These priors, however, are fractional priors, not real priors. They are only used in conjunction with a diminished fractional power of the likelihood function, and so the fact that they are centered should not be viewed as an improper double-use of the data, at least in the way that "double-use" is normally understood. This notion can be formalized by solving a particular set of equations to yield the intrinsic prior, which is the real (non likelihood-centered) prior to which a default Bayes factor corresponds asymptotically; see Berger and Pericchi (2001) for further discussion.

Theorem 4.4.2 applies to any true decomposable graph, and so offers a universal guarantee of information consistency with respect to the b-statistic. Theorem 4.4.4, however, does not apply to all possible univariate conditional regression models, only to those in which the variable z to be predicted comes first in a perfect vertex elimination scheme. There are multiple such schemes, meaning that the theorem will apply to multiple models for any given true graph. Yet without a more general distributional result analogous to Lemma 4.4.3, it is not possible to establish information consistency for all univariate regression models. Nonetheless, the theorem does demonstrate information consistency with respect to the F-statistic for a useful, albeit restricted, subclass of these conditional sparse regression models. This "proof-of-concept" result, while not completely general, is still enough to demonstrate that the HIW g-prior has certain desirable properties that the conventional proper prior lacks, and that these properties relate to precisely the concept of information consistency familiar from linear models.

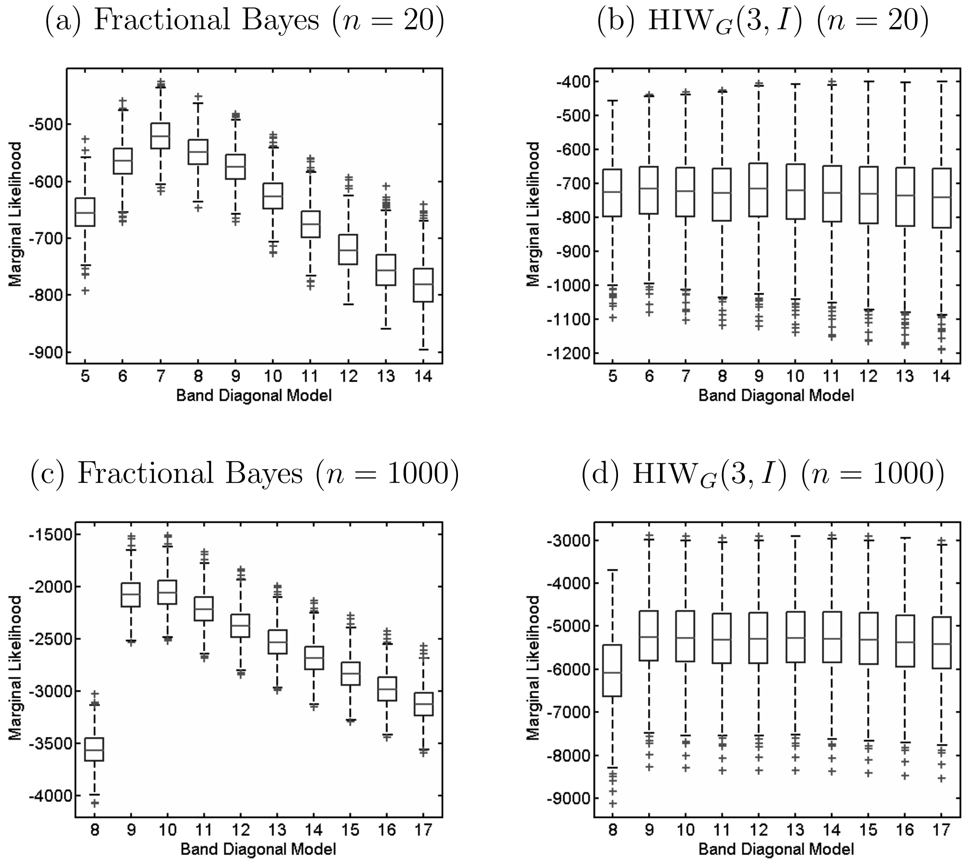
4.5 Fractional marginal likelihoods: a simulation study

This section studies the behavior of fractional marginal likelihoods through simulations that compare models of differing complexity. The baseline for comparison will be the conventional alternative, the $\mathrm{HIW}_G(\delta, \tau I)$ prior. These results illustrate that the casual use of the conventional prior undermines the ability to choose between competing models, and that the objective procedure presented so far is well-behaved.

Data sets of various sizes ($n = 20, 50, 100, 500, 1000$) were simulated from the true model: each $x_i \sim \mathrm{N}(0, \Sigma)$, where Σ was the 50-dimensional correlation matrix of a stationary Gaussian AR(10) process. This represents a Gaussian graphical model due to the band-diagonal form of the precision matrix. Note that $p = 50$ is the number of nodes (i.e. the length of each observation x_i), but that n is the number of such independent draws observed from the true model.

An appropriate choice for the conventional proper prior is $\Sigma \sim \mathrm{HIW}_G(3, I)$, where the choice of $\delta = 3$ reflects the standard advice to give $p(\Sigma \mid G)$ a finite first moment (Jones et al., 2005). I have chosen $\tau = 1$, since Σ is known to be a correlation matrix. Typically τ is very hard to specify without looking at the data, making this choice, if anything, overly favorable to the conventional prior.

For each sample size, I simulated 1000 data sets from the true model and computed marginal likelihoods for 21 different candidate graphs corresponding to band-diagonal



(a) Fractional Bayes ($n = 20$)

(b) $\mathrm{HIW}_G(3, I)$ ($n = 20$)

(c) Fractional Bayes ($n = 1000$)

(d) $\mathrm{HIW}_G(3, I)$ ($n = 1000$)

Figure 4.1: Boxplots of realized marginal likelihoods in the simulation study. The x-axis is the bandwidth of the precision matrix; the y-axis is marginal log-likelihood.

precision matrices of bandwidth 0 through 20 (with the true model having bandwidth 10). Figure 4.1 gives the frequency distributions of marginal log-likelihoods for each candidate model, which show substantially better separation under the fractional Bayes approach. Results are given for both the smallest ($n = 20$) and largest ($n = 1000$) sample sizes, although the same pattern emerged for all sample sizes.

For each data set, the model with the highest marginal likelihood was noted. This gives a Monte-Carlo estimate for the frequency distribution of the preferred band-diagonal size under the two priors (Fig. 4.2). As n grows, the fractional Bayes factors favor the true model (bandwidth 10) with increasing accuracy, whereas the results

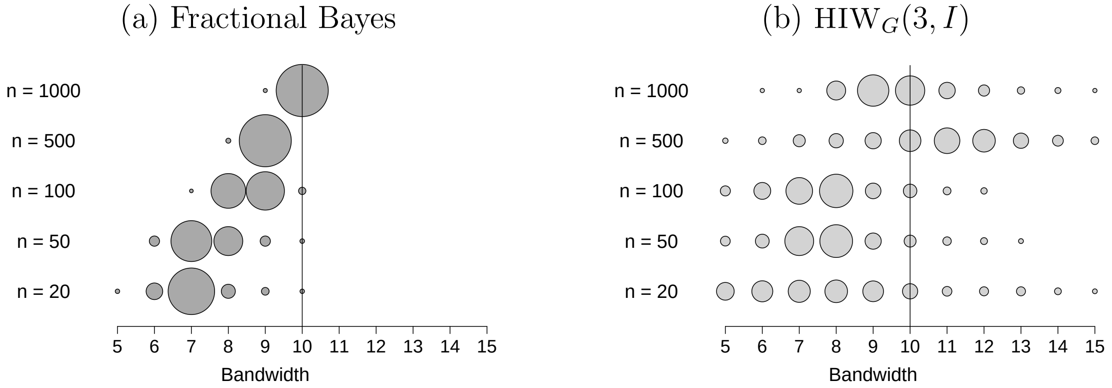


Figure 4.2: The empirical distribution of the chosen bandwidth under repeated sampling from the true model in the simulation study. The area of each circle represents the fraction out of 1000 independent data sets in which each model had the highest marginal likelihood under the given prior.

from the $\text{HIW}_G(3, I)$ are highly erratic.

It is clear that, unlike the conventional prior, the fractional-Bayes procedure prefers sparse models in the absence of enough data to justify extra edges. Yet as the sample size increases, the fractional approach favors more complex models, and eventually chooses the true one almost every time. The conventional prior does not exhibit this tendency nearly as strongly, displaying an unintuitively high level of variation in the choice of model.

One might imagine that the conventional prior, by shrinking the covariance structure toward the identity matrix with its strong pattern of off-diagonal zeros, would yield systematically smaller models. This expectation is not confirmed by the simulation study, indicating that intuitions about shrinkage gleaned from covariance estimation do not necessarily apply to model selection. This can be understood in terms of the Occam's-razor property of Bayesian marginal likelihoods (Jefferys and Berger, 1992). The conventional prior spreads its mass out quite broadly, which is an advantage in estimation problems but crippling in model selection due to the lack of predictions sharp enough to yield any posterior separation of models.

The results from this simulation strengthen the theoretical results developed so far. Note that the conventional prior induces a set of ridge-regression priors on the complete conditional models considered in Section 4.4.3 (as can be shown through a straightforward application of Lemma 4.4.3). The problems with ridge-regression priors for variable selection are well understood (Zellner and Siow, 1980; Liang et al., 2008), and give some intuition as to why the conventional $\mathrm{HIW}_G(\delta, \tau I)$ prior is sub-optimal for model selection.

4.6 Priors over graphs

There are two simple approaches to assigning prior probabilities to graphs themselves. The first is to give every graph the same prior probability κ^{-1}, where κ is the number of decomposable graphs on p nodes (which is not trivial to compute). As Giudici and Green (1999) note, this prior is quite heavily concentrated on graphs of middling size due to combinatorial explosion. This is the same problem that affects the uniform prior over linear models, as described in the previous chapter.

The second alternative, rapidly becoming the standard for graphical models much as it has for linear models, is to imagine a sequence of edge inclusions as having a binomial distribution with success probability w (Dobra et al., 2004; Jones et al., 2005). This yields priors of the form

$$p(G) \propto w^k (1-w)^{m-k} \tag{4.19}$$

for decomposable graphs, with non-decomposable graphs given prior probability 0 by construction. (Recall that k is the number of edges, and $m = p(p-1)/2$ is the maximum number of possible edges.) The constant of proportionality is hard to compute since it is typically unknown how many decomposable graphs there are of a given size, but this constant is the same for all models and can thus be ignored.

If the expected fraction of included edges is known quite precisely, this framework may be attractive. Yet often this fraction is not known, making an arbitrary choice of w seem heavy-handed. Instead, placing a prior on w is as easy and attractive an option as it proved to be in variable selection. Assuming the conjugate beta prior $w \sim \text{Be}(a, b)$ allows the same explicit marginalization:

$$p(G) \propto \int_0^1 p(G \mid w) p(w) \, \mathrm{d}w \propto \frac{\beta(a + k, b + m - k)}{\beta(a, b)} \, ,$$

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

$$p(G) \propto \frac{(k)!(m-k)!}{(m+1)(m!)} = \frac{1}{m+1} \binom{m}{k}^{-1} \, , \tag{4.20}$$

which is quite familiar. In the previous chapter, these priors yielded an automatic penalty for testing irrelevant covariates in a regression model. A quick set of simulations are enough to confirm that, as expected, the same effect holds in graphical models.

Beginning with a correlation matrix corresponding to the ten-node graph from Figure 4.3, progressively more "noise" nodes were added to the data set. This refers to nodes unconnected both from the true graph and from each other, but that lead to combinatorial explosion in the number of edges that must be tested for inclusion. The numbers of noise nodes chosen were 5, 15, and 40, which in addition to the 10 connected nodes in the true graph imply 60, 300, and 1225 separate hypothesis tests,

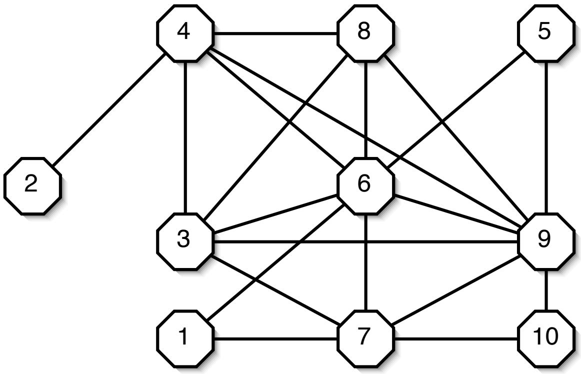


Figure 4.3: The true 10-node, decomposable graph used in the multiplicity-correction study. All noise nodes were completely unconnected both from this graph and from each other, meaning that any edges involving them are false positives.

respectively. In all cases the number of true hypotheses remained fixed at 22, one for each edge in the 10-node graph.

Three sets of tests were performed under each of three different priors on models. These results are summarized in Table 4.1. Here, "Fully Bayes" uses the model probabilities from (4.20), while "Oracle Bayes" involves plugging the true value of w into (4.19) to compute prior graph probabilities. All marginal likelihoods were computed using fractional Bayes factors, with $g = 1/n$.

Notice that the fully Bayesian multiplicity-correction prior in (4.20) suppresses false postives very effectively. The difference between corrected and uncorrected versions is substantial; in the 50-node example, giving all models the same prior probability yields 40 false positives (inclusion probability greater than 50%), whereas imposing the multiplicity-correction prior yields none. Importantly, this approach does not depend upon the choice of an arbitrary hyper-parameter w, although if subjective inputs are required, they can be accommodated through a different beta prior while still retaining closed-form answers.

Other differences from standard priors also emerge. Consider, for example, edges

Table 4.1: Estimated inclusion probabilities for specific edges (left-most column) as the number of unconnected noise nodes grows from 5 to 15 to 40. The last line of the table shows the number of falsely positive flags, which are other, non-enumerated edges that have edge-inclusion probability greater than 0·5. All probabilities were calculated using five million iterations of the FINCS algorithm (Scott and Carvalho, 2008).

	Number of Noise Nodes								
	No Correction			Oracle Bayes			Fully Bayes		
Edge	5	15	40	5	15	40	5	15	40
(1,6)	99	99	99	99	99	99	99	99	99
(3,4)	16	2	0	1	0	3	1	1	1
(3,6)	99	99	99	99	99	97	99	99	99
(3,8)	18	0	0	10	2	0	6	1	0
(3,9)	31	3	0	0	0	0	0	0	0
(4,6)	99	99	99	99	99	99	99	99	99
(4,9)	99	99	99	99	99	99	99	99	99
(5,6)	16	5	0	22	21	18	24	23	22
(5,9)	36	14	31	31	30	34	31	31	33
(6,7)	58	76	74	14	2	0	8	3	0
(6,9)	22	3	0	1	0	0	1	0	0
(8,9)	43	4	7	14	2	0	7	1	0
(9,10)	89	95	99	71	69	76	60	60	69
FPs:	6	11	40	1	1	1	1	1	0

$(5, 9)$ and $(6, 7)$. If all models are given equal prior probabilities, adding more noise edges makes $(6, 7)$ appear stronger and $(5, 9)$ appear slightly weaker. Yet the opposite happens using the multiplicity-correction priors: the addition of more noise nodes makes $(6, 7)$ disappear entirely and yet retains $(5, 9)$ at close to its original strength.

This behavior suggests that (4.20) does not merely shrink edge-inclusion probabilities to 0 uniformly as k remains fixed and m grows. Rather, it differentially rewards edges that participate in more parsimonious models, suggesting a fundamental difference from (4.19) in the way mass is apportioned across model space.

4.7 Discussion

This chapter has introduced a new method of model selection for Gaussian graphical models based upon objective Bayesian ideas. The strengths of the approach are the theoretical guarantees of Section 4.4, the intuitive behavior of fractional marginal likelihoods demonstrated in Section 4.5, and the strong control over false positives shown in Section 4.6.

The missing ingredient from this chapter is a comparison of methodologies on a real data set. This has been postponed until the next chapter, where new computational tools are introduced that allow these questions to be investigated more fully.

Chapter 5

Computation in Graphical Model Spaces

The previous chapter developed new methodological tools for Gaussian graphical modeling. Yet inferring the conditional independence structure of a random vector also presents a substantial problem in stochastic computation. These model spaces are usually enormous; p nodes in a graph mean $m = p(p-1)/2$ possible edges, and hence 2^m possible graphs corresponding to all combinations of individual edges being in or out of the model. Beyond $p = 7$, enumeration becomes a practical impossibility, and yet there are substantial gains to be had by fitting graphical models to far larger data sets—from portfolio selection problems with dozens or hundreds of assets, to biological applications involving thousands of genes. This motivates the development of accurate, scalable search methodologies that are capable of finding good models, or at least distinguishing the important edges from the irrelevant ones.

One obvious candidate is the reversible-jump MCMC approach of Giudici and

Green (1999), which works well on very small problems—for example, four to six nodes. Yet many authors express skepticism that MCMC is well-suited for problems that are even slightly larger; see, for example, Dobra et al. (2004), Jones et al. (2005), and Hans et al. (2007), who find little comfort in infinite-runtime guarantees when MCMC is deployed in discrete model spaces of such size and complexity. As these and many other authors note, assessing whether a Markov chain over a multimodal space has converged to a stationary distribution is devilishly tough, and theoretical results exist only for the smallest of problems (Woodard, 2007). Even when state-of-the-art "mixing" tactics are used—simulated tempering, parallel chains, adaptive proposal distributions—apparent finite-time convergence can prove to be a mirage.

Two alternative classes of graphical model-selection procedures have been developed to overcome these problems: compositional methods and direct search.

Compositional methods exploit the fact that graphs are models for conditional independence. They proceed by selecting a sparse regression model for each variable in terms of all the others, and then assembling the resulting set of conditional relationships into a graph to yield a valid joint distribution. Several methods for choosing each regression model are available, some based upon $L1$-regularization (Meinshausen and Buhlmann, 2006; Yuan and Lin, 2007) and others based upon stepwise selection (Dobra et al., 2004).

Like MCMC, direct-search methods operate in the space of graphs rather than the space of conditional regressions. Unlike MCMC, however, they abandon the goal of converging to a stationary distribution—which may be unattainable, and is usually unassessable—in favor of simply listing and scoring a collection of good models. (Typically the scores are either Bayesian marginal likelihoods or posterior probabilities.) The two most prominent search procedures are the serial Metropolis-based algorithm of Jones et al. (2005) and the parallel Shotgun Stochastic Search of Hans et al. (2007).

The goal of this chapter is to decribe an alternative search procedure called FINCS, for feature-inclusion stochastic search. FINCS is a serial procedure that relies upon constantly updated estimates of edge-inclusion probabilities in order to propose new regions of model space for exploration, and incorporates a novel, efficient global move based upon recent graph-triangulation algorithms. It is strongly related to the method recommended by Berger and Molina (2005) in the context of linear-model selection.

The FINCS algorithm will first be described in detail. Then the remainder of the chapter compares FINCS to other search procedures, to MCMC, and to compositional methods. These comparisons show FINCS to be superior on three objective criteria: stability, model scores, and predictive accuracy of models discovered. Sections 5.2 and 5.3 describe these results on simulated problems at the lower (25-node) and upper (100-node) ends of what might be considered moderate-dimensional. Then Section 5.4 assesses the predictive performance of models discovered using FINCS on a real 59-dimensional example involving monthly mutual-fund returns. FINCS will also be used to show the superiority of fractional marginal likelihoods over those arising from conventional priors, as measured by out-of-sample predictive performance.

5.1 Graphical model determination

5.1.1 Existing methods

Markov-chain Monte Carlo

Giudici and Green (1999) popularized the use of MCMC for graphical models, implementing a reversible-jump sampler (Green, 1995) over all model parameters including graphical constraints.

Jones et al. (2005) then considered a version of MCMC that eliminated the complexities of reversible-jump by explicitly marginalizing over most parameters and working directly with graph marginal likelihoods. This version of the algorithm makes one-edge moves through graph space, accepting proposed moves with probability

$$\alpha = \min\left(1, \frac{p(G')h(G \mid G')}{p(G)h(G' \mid G)}\right) , \tag{5.1}$$

where $p(\cdot)$ is the posterior probability of the graph (available in closed form due to conjugacy assumptions), and $h(\cdot)$ is the proposal probability.

Another obvious MCMC algorithm to apply to graphical models is Gibbs sampling, whereby each edge indicator variable in turn is sampled conditional upon all the others. This is the graphical-model equivalent of the SSVS procedure described by George and McCulloch (1993) for linear models. Yet this algorithm, despite being a true MCMC with demonstrable theoretical convergence to a stationary distribution, does not receive any attention in the literature. This is perhaps explained by its tendency to miss very large modes in model space, as the results below will demonstrate.

Metropolis-based stochastic search

In practice, the Metropolis criterion is less useful as an MCMC transition kernel, and far more useful as a search heuristic for finding and cataloguing good models. This is true for two reasons.

First, guaranteeing ergodicity of the Markov chain requires that each step involve a costly enumeration of all possible one-edge moves that maintain decomposability (which in general will not be symmetric). If this fact is not accounted for, then the proposal densities $h(\cdot)$ in (5.1) will be wrong, and the chain will not converge to a stationary distribution. This enumeration is possible, but costly (Deshpande et al., 2001).

The more important reason, however, is the lack of MCMC convergence diagnostics on such complex, multimodal problems. The model spaces are simply too large to trust the usual rules of thumb, and it makes little sense to evaluate models by their frequency of occurrence in a Monte Carlo sample when one can simply list the best ones instead. Hence it is best to view the Metropolis algorithm as a tool for stochastic search, and not as true MCMC.

Jones et al. (2005) and Hans et al. (2007) describe one recent advancement called Shotgun Stochastic Search (SSS). This algorithm powerfully exploits a distributed computing environment to consider all possible local moves in parallel at each step, moving to a new model in proportion to how much each possible move improves upon the current model. Yet for those who only have access to serial computing environments, evaluating all possible neighbors of a given graph may be prohibitively time-consuming.

Compositional methods

In compositional search, one first defines the neighborhood $\mathrm{ne}(i)$ of each node by fitting a single sparse regression model of x_i upon a subset of $\{x_j : j \neq i\}$. Dobra et al. (2004) perform a Bayesian selection procedure to define each neighborhood, while Meinshausen and Buhlmann (2006) and Yuan and Lin (2007) use variants of the lasso (Tibshirani, 1996). Regardless of the variable-selection method used, the resulting set of regressions implicitly defines a graph.

Such procedures do not, in general, yield a valid joint distribution: often $i \in \mathrm{ne}(j)$ but $j \notin \mathrm{ne}(i)$, which is impossible in an undirected graph. There are two ways of proceeding to a valid edge set E, called (for obvious reasons) the AND graph and the OR graph:

AND graph: $(i, j) \in E$ if $i \in \mathrm{ne}(j) \wedge j \in \mathrm{ne}(i)$

OR graph: $(i, j) \in E$ if $i \in \mathrm{ne}(j) \vee j \in \mathrm{ne}(i)$

Each edge set must then be triangulated to yield decomposable graphs. It is easy to see that $E_\wedge \subset E_\vee$. There is no principled way to decide which edge set to use in practice, though Meinshausen and Buhlmann (2006) give conditions under which the two will converge to the same answer asymptotically.

Note that defining a graph by composition also involves a (possibly difficult) search procedure, since each sparse regression model must be chosen from 2^{p-1} possible candidates. Lasso does not involve an explicit search over models, but the $L1$ penalty term that induces sparsity is typically chosen by cross-validation, which often takes just as long when done carefully.

5.1.2 FINCS: feature-inclusion stochastic search

As an alternative, this section presents a serial algorithm called FINCS, or feature-inclusion stochastic search, that combines three types of moves through graph space: local moves, resampling moves, and global moves. Related to simpler algorithm introduced by Berger and Molina (2005) in the context of regression variable selection, FINCS is motivated by a simple observation: edge moves that have improved some models are more likely to improve other models as well, or at least more likely to do so than a randomly chosen move.

FINCS also recognizes the tension between two conflicting goals: local efficiency and global mode-finding. Results from Giudici and Green (1999) and Wong et al. (2003) suggest that pairwise comparisons for edge-at-a-time moves in graph space are much faster than those for multiple-edge moves due to the local structure implied by the hyper-inverse Wishart distribution. Both a graph's junction tree representation and its marginal log-likelihood can be updated quite efficiently under such local moves, which will affect at most two cliques and one separator.

Yet one must also confront the familiar problem of multimodality, severely exacerbated by the restriction to decomposable graphs. Each local move changes not only the graph itself but also the local topology of reachable space, opening some doors to immediate exploration and closing others. As the number of nodes increases, moreover, the stepwise-decomposable paths in model space between two far-flung graphs become vanishingly small as a proportion of all possible paths between them. The theory guarantees that such a path always exists (Frydenberg and Lauritzen, 1989), but this path may be very difficult to find.

These principles suggest that a sound computational strategy must include a blend of local and global moves—local moves to explore concentrated regions of good graphs with minimal numerical overhead, and global moves to avoid missing important regions that aren't easily reachable from one another by a series of local moves that maintain stepwise decomposability.

Motivated by these concerns, FINCS interweaves three different kinds of moves:

Local move: Starting with graph G_{t-1}, generate a new graph G_t by randomly choosing to add or delete an edge that will maintain decomposability. If adding, do so in proportion to $\hat{q}_{ij}$, the current estimates of edge-inclusion probabilities. If deleting, do so in inverse proportion to these probabilities.

Resampling move: Revisit one of $\{G_1, G_2, \ldots G_{t-1}\}$ in proportion to their posterior model probabilities, and begin making local moves from the resampled graph.

Global move: Jump to a new region of graph space by generating a *randomized median triangulation pair*, or RMTP. This can be done in three steps:

1. Begin with an empty graph and iterate through all possible edges once, independently adding each one in proportion to its estimated inclusion probability $\hat{q}_{ij}$. In general this will yield a nondecomposable graph G_N. A simpler variation involves deterministically choosing the median graph.

2. Compute a minimally sandwiching triangulation pair for G_N. This pair comprises a minimal decomposable supergraph $G^+ \supset G_N$ along with a maximal decomposable subgraph $G^- \subset G_N$, wherein no candidate edge can be added to G^- or removed from G^+ while still maintaining the decomposability of each.

3. Evaluate each member of the pair, and choose G_t in proportion to their posterior probabilities.

An RMTP move immediately transcends the limitations of stepwise-decomposable moves, bridging nondecomposable valleys in model space with a minimal set of fill edges and allowing the search to escape local modes. Several algorithms are available for computing these inclusion-minimal triangulations and inclusion-maximal subtriangulations in $O(nk)$ time, where n is the number of nodes and k is the number of edges in G_N. One especially useful algorithm due to Berry et al. (2006) allows simultaneous computing of both G^+ and G^-; another can be found in Berry et al. (2004). It should be noted that these triangulations are neither unique nor globally optimal, since computing a minimum triangulation is a different (NP-complete) problem.

After each step, simply compute the posterior probability of G_t, and use it (assuming it hasn't been visited already) to update the estimated inclusion probabilities of all the edges:

$$\hat{q}_{ij}(t) = \frac{\sum_{k=1}^{k=t} 1_{(i,j) \in G_k} \cdot P(X \mid G_k) \cdot \pi(G_k)}{\sum_{k=1}^{k=t} P(X \mid G_k) \cdot \pi(G_k)} \,. \tag{5.2}$$

It is important to emphasize that these inclusion probabilities $\hat{q}_{ij}$ are simply a search heuristic. There is no sense in which they "converge" to the true inclusion probabilities, except in the trivial sense that FINCS will eventually enumerate all the models. Present theory and computing technology simply do not allow any definitive statements about the true inclusion probabilities; hence the only unambiguous measure of a search procedure is: which models does it find, and how good are they?

The estimated inclusion probabilities nonetheless provide a useful summary of the search, since one can tell just by glancing at them how important each edge seems to be among the cohort of models discovered. They also help in assessing stability, since it would be unwise to trust a procedure that yields highly volatile estimates under repetition.

5.1.3 Details of data structures and implementation

Substantial gains in efficiency for the resampling step are possible by implementating storage of previously visited models in a binary search tree (a map in the C++ STL) indexed by model score. This requires normalizing model probabilities to $(0, 1]$ and maintaining a lower-dimensional representation of the empirical distribution of model scores on that interval. This purpose is well served by a beta distribution whose parameters are updated each time a substantial pocket of probability is found in model space. By drawing a score from this distribution and then resampling the model corresponding to that score, resampling can be done in $\log T$ time, where T is the current number of saved models. This allows only approximate resampling, but experience suggests that it is much better that the costly linear-time step required by exact resampling (which is not, after all, a theoretical requirement of the algorithm).

Subsequent local moves can also be greatly streamlined by saving a copy of the junction tree for each graph so that it doesn't need to be recomputed upon resampling.

In my experience, a blend of 80–90% local moves with 10-15% resampling moves seems to work well, with the remaining fraction devoted to global moves; these are most effective when used sparingly due to the computational expense of triangulating the graph and rebuilding the junction tree, which grows quite rapidly with dimension. It also seems advisable to allow for an unusually long run of local moves following a global move. This accounts for the fact that the global move can be expected to find other hills in the model space, but is very unlikely to jump directly to the tops of those hills. A longer-than-normal run of local moves following such a jump allows the search procedure to climb to the top, thereby helping to solve the multimodality problem.

In larger problems, it is important to initialize the search with a cohort of promising graphs for resampling, which can be done automatically and quite rapidly before beginning the search. I have found that on small-to-moderate-dimensional problems such as the 25-node example in Section 5.2, this step can safely be skipped, since FINCS converges quite rapidly to the same answer regardless of the initial graph.

Finally, it is necessary to bound the inclusion probabilities away from 0 and 1 in order to allow new edges to enter the picture. This is done by renormalizing the online estimates for probabilities to $(\delta, 1-\delta)$ for some suitably chosen small value. Different choices of δ will either flatten or sharpen the choice of edges; a reasonable default choice in moderate-dimensional problems is between 0.01 and 0.05.

5.2 A simulated 25-node example

This section compaes FINCS with both Metropolis and Gibbs on a 25-node graph obtained by starting with the 10-node graph in Figure 4.3 and adding 15 unconnected nodes. This problem foregrounds two challenges: to find the smaller 10-node graph embedded in the larger 25-node space, and to avoid flagging false edges that follow from getting stuck in local modes.

The results of these comparisons are summarized in Figures 5.1, 5.2, and 5.3. Here the global-move version of FINCS resamples an old model every 10 iterations and makes an RMTP move every 50 iterations. “Gibbs” refers to stochastic-search

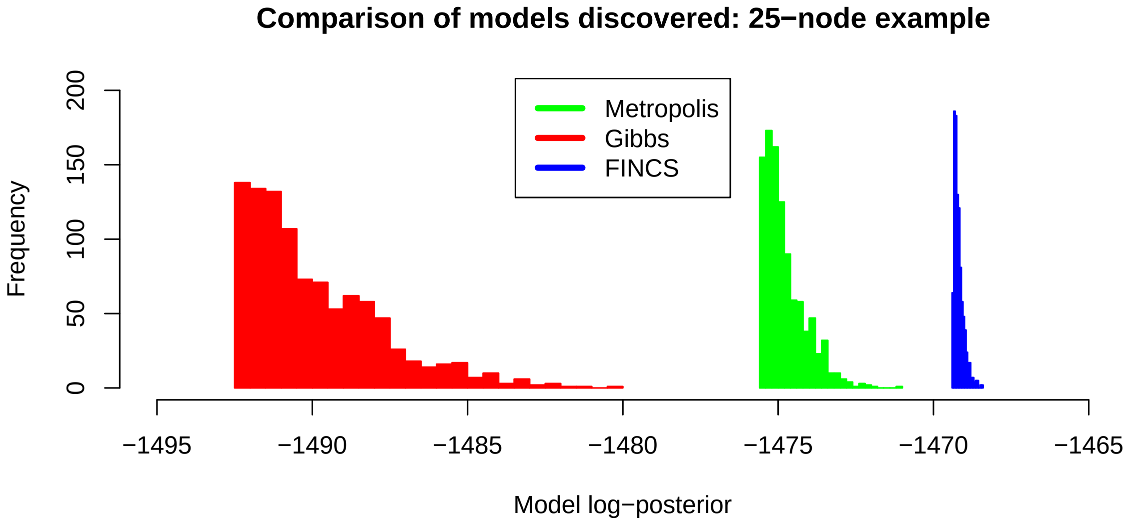


Figure 5.1: Comparison of model-search algorithms on the 25-node example. These are the posterior probabilities of the top 1000 models found using Gibbs (10000 iterations), FINCS (3 million iterations), and Metropolis (10 million iterations). The Gibbs and FINCS runs each had the same number of marginal likelihood evaluations.

edge-selection, modeled after George and McCulloch (1993); “Metropolis” refers to the random-walk stochastic-search algorithm of Jones et al. (2005).

Three questions are of interest:

1. Which search method finds the best collection of models, as measured by posterior probability?

2. Which search procedures are stable under repetition?

3. Are the estimated inclusion probabilities from FINCS and Metropolis intuitively reasonable?

Figure 5.1 gives histograms of the best 1000 models discovered during single long runs of FINCS, Gibbs, and Metropolis. This decisively answers the first question: all of the top 1000 models discovered by FINCS are more probable than the single

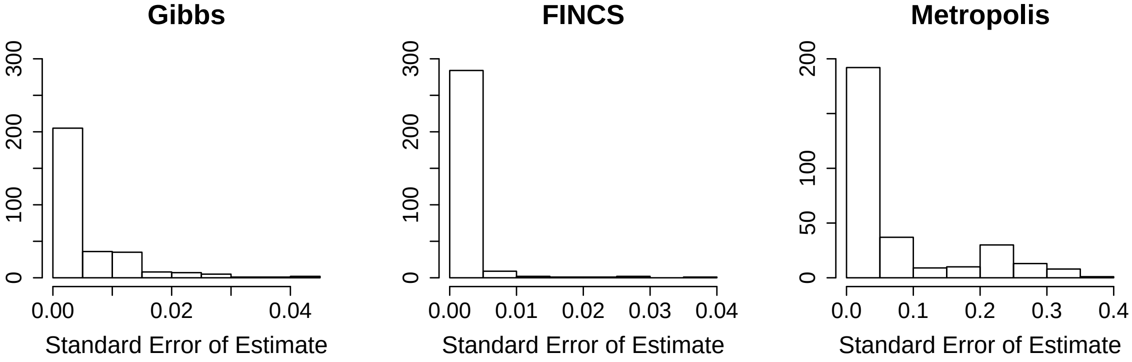


Figure 5.2: Standard errors (under 20 repetitions) of estimates for the 300 pairwise inclusion probabilities in the 25-node example.

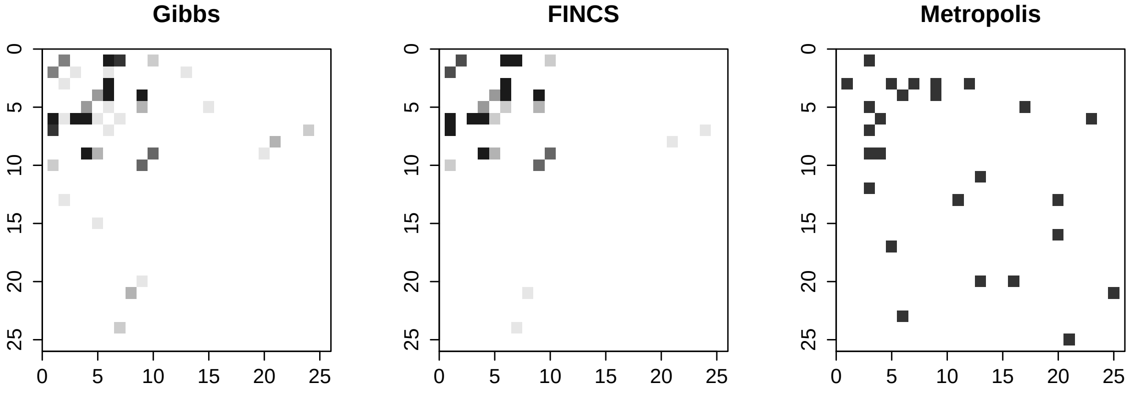


Figure 5.3: Grayscale images of the estimated inclusion probabilities from each algorithm on the 25-node example.

best model discovered by either Metropolis or Gibbs. Gibbs does particularly poorly here; its best model was 11 orders of magnitude worse than the thousandth-best model discovered by FINCS. Such large systematic differences were unexpected, and yet were very stable under repeated restarts, both from the same initial model and very different initial models. On this example, FINCS always found better models than both Metropolis and Gibbs, and there were always substantial gaps between the poorest of the top 1000 models discovered by FINCS and the best discovered by either competing method.

To assess stability, I conducted 20 short runs of each algorithm, starting from the null graph each time. (The run lengths were 2000, 300000, and 1 million for Gibbs, FINCS, and Metropolis, respectively.) Each run yielded an estimate for all 300 pairwise inclusion probabilities. The standard errors of these estimates provide an excellent proxy for stability, since highly variable inclusion probabilities indicate that the results of a single run are unreliable.

Figure 5.2 shows the results of this exercise. The estimated inclusion probabilities for the Metropolis algorithm are extremely erratic, displaying run-to-run standard deviations as high as 40% for some edges. Indeed, these edges were estimated at 0% inclusion probability in some runs and 100% in others. Gibbs and FINCS, on the other hand, were fairly stable, with estimated inclusion probabilities rarely differing by more than 5% from run to run.

Finally, Figure 5.3 gives the estimated edge-inclusion probabilities for long runs of all three algorithms. As the above results foreshadow, Metropolis did quite poorly: it flagged several false positives outside the 10-node subgraph, and missed many edges corresponding to strong partial correlations.

While it is easy to say that Metropolis gets the inclusion probabilities wrong, it is more difficult to say whether FINCS gets them right. For both Gibbs and FINCS, there is very strong rank-correlation ($> 90\%$) between pairwise inclusion probabilities and pairwise partial correlations, which suggests that both procedures are flagging important edges. And as Figure 5.3 shows, FINCS does yield inclusion probabilities that are fairly close to those given by Gibbs, with slight-but-systematic biases in favor of strong edges and against weak ones (as might be expected, given the "expand about the modes" nature of the procedure). These biases seem rather minor given the enormous size of the model space—indeed, both procedures give the same median-probability graph, and yield the same qualitative conclusions.

The Gibbs inclusion probabilities do, of course, represent a true MCMC estimate, and they meet the usual informal convergence criterion: after a very brief burn-in, they are substantially the same run after run, regardless of the starting point. Readers willing to trust the Gibbs estimates, therefore, are likely to conclude that FINCS gets things slightly wrong here (though not nearly so wrong as Metropolis).

The Gibbs estimates are not very trustworthy, however, in light of the evidence in Figure 5.1. The Gibbs inclusion probabilities, despite being highly stable under repetition, are estimated without ever visiting a single model within 11 orders of magnitude of the best model found by FINCS. This suggests that the usual convergence diagnostics may be misleading. This is uncertain; it is certain, however, that the Gibbs sampler routinely misses enormous pockets of probability in model space, and produces a demonstrably inferior list of models.

In the absence of better theory or more trustworthy convergence diagnostics, it seems imprudent to use the Gibbs answer when FINCS finds thousands of models that are each tens of thousands of times more probable than any found by Gibbs. At worst, FINCS is identifying the most likely models while giving a useful, albeit slightly biased, picture of the inclusion probabilities.

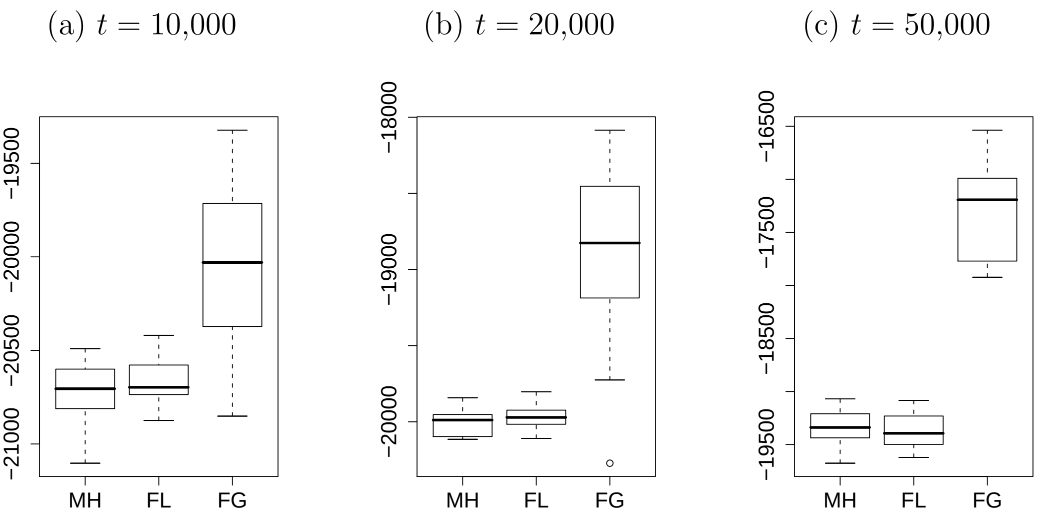
5.3 A simulated 100-node example

To assess the performance of FINCS on a bigger problem, I simulated a data set of size 250 from the correlation matrix of a stationary AR(10) process of length 100. This represents a graphical model due to the block-diagonal form of the 100-dimensional precision matrix. I then ran FINCS-global, FINCS-local, and Metropolis 20 different times, always starting from the null graph, and recorded the top marginal log-likelihood discovered in the course of the search. Here, FINCS-local resamples a previously visited model every 5 iterations. FINCS-global resamples every 5 iterations and performs a global move every 1000 iterations; each global move is followed by 100 local moves as described in Section 5.1.3. Results are presented in the Figure 5.4, while runtime information is in Table 5.1.

To give a better sense of time efficiency, the study incorporated two pairs of approximately equal-time, moderate-length runs of Metropolis and FINCS-global on the same data set. One pair of runs started from the null graph with no edges, while another pair started from an initial graph corresponding to a set of conditional regressions. Table 5.2 gives the top marginal log-likelihoods discovered, along with runtime information for each search.

There are four lessons to take from these experiments.

First, the proposed global move is very helpful for escaping local modes in model space. The overhead required to triangulate and compute a new junction tree means



(a) $t = 10{,}000$ (b) $t = 20{,}000$ (c) $t = 50{,}000$

Figure 5.4: 100-node example: Boxplots of the top marginal log-likelihoods discovered (y-axis) on 20 restarts of 3 different run lengths for the Metropolis (MH), FINCS-local (FL), and FINCS-global (FG) algorithms.

Table 5.1: 100-node example: Mean (standard deviation) runtime for each algorithm.

	Runtime in seconds		
Iterations	Metropolis	FINCS-local	FINCS-global
10,000	20.98 (0.58)	27.22 (1.00)	36.36 (2.96)
20,000	38.25 (1.08)	62.86 (3.53)	68.03 (5.29)
50,000	99.28 (2.63)	155.38 (5.12)	176.21 (14.42)

that FINCS-global takes about twice as long as Metropolis for an equivalent number of steps, implying that a single global move takes roughly the same amount of time as 1000 local moves for this 100-node problem. (This ratio gets steeper as the number of nodes increases). Yet the advantage conferred by this global move is clear; 10,000 iterations of FINCS-global, for example, dramatically outperformed 20,000 iterations of Metropolis despite taking about 25% less raw time on average.

The necessity of the global move becomes even clearer with more iterations ($t =$

Table 5.2: Marginal log-likelihoods of best models discovered from two different starting points. The guessed graph corresponds to a set of conditional regressions (marginal log-likelihood of -5170.29 after triangulation), where an edge (i, j) was flagged if x_j yielded a z-statistic ≥ 3.0 in absolute value as a predictor of x_i (or vice versa).

Algorithm	Start	Iterations	Time (s)	Best model
FINCS-global	Null graph	80,000	458.90	-16897.32
Metropolis	Null graph	220,000	456.70	-18689.34
FINCS-global	Guess ($\lvert z\rvert > 3.0$)	50,000	318.74	-3211.09
Metropolis	Guess ($\lvert z\rvert > 3.0$)	110,000	334.72	-3330.84

50,000 in the right panel of Figure 5.4). By 50,000 iterations, both Metropolis and FINCS-local have leveled off as a result of getting stuck in local modes, whereas FINCS-global has continued to climb at a rapid pace.

Second, there are stark differences in character between the posterior summaries yielded by Metropolis and by FINCS. Metropolis acts essentially as a stochastic optimizer; in an average run of 10,000 iterations, it visited fewer than 300 distinct models. FINCS, on the other, both finds modes and expands about them, cataloguing many thousands of distinct models (including hundreds that are nearly as good as best model found) in each run of 10,000 iterations. This yields a much richer summary of model space, more than compensating for the marginal time penalty paid to store and resample models.

Third, no search algorithm, no matter how well tuned, can compensate for a poor starting point on problems of this size. Even FINCS-global, despite strong performance compared to other competing algorithms, takes a very long time to bridge the many thousands of orders of magnitude between the unreasonable null graph and a reasonable guess. This is very different from the 25-node problem, where the choice of initialization did not matter at all.

Finally, even when a smart initial guess is supplied, FINCS still substantially outperforms Metropolis. Given the same amount of computing time, FINCS found models 119 orders of magnitude better than those found by Metropolis. While this looks small compared to the 2000 orders of magnitude separating the best models that each procedure found in an uninitialized search, it still corresponds to a Bayes factor of 4.8×10^{51}—an enormous improvement.

5.4 A real 59-node example: mutual funds

This section's goal is to compare FINCS and Metropolis (both to each other and to the lasso, a popular compositional method) on a real 59-node example involving mutual-fund data. Comparisons between fractional and conventional marginal likelihoods are also given, shedding light on the practical differences of the methodologies described in the previous chapter.

The example involves graphical-model selection for a set of 59 mutual funds in several different sectors: 13 U.S. bond funds, 30 U.S. stock funds, 7 balanced funds investing in both U.S. stocks and bonds, and 9 international stock funds. The example is motivated by the need for accurate, stable estimates of variances and pairwise correlations of assets in dynamic portfolio-selection problems. Graphical models, as Carvalho and West (2007) show, offer a potent tool for regularization and stabilization of these estimates, leading to portfolios with the potential to uniformly dominate their traditional counterparts in terms of risk, transaction costs, and overall profitability.

A fair barometer of performance here is prediction, since the study includes a non-Bayesian technique (the lasso) for which there is no notion of the marginal likelihoods and posterior probabilities used as measures in previous sections.

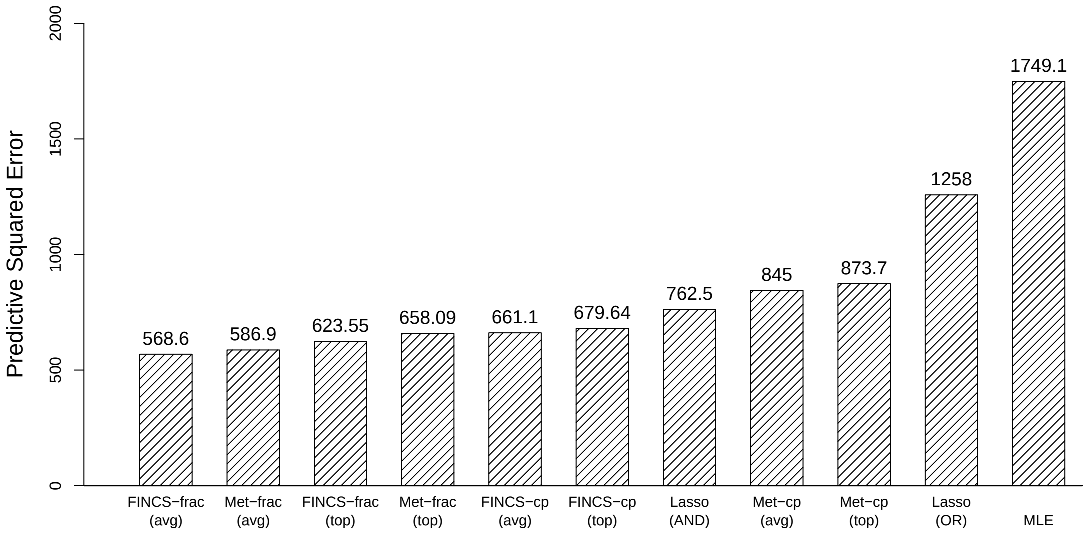


Figure 5.5: Sum of squared errors in predicting missing values, 59-node mutual-fund example. Frac: fractional marginal likelihoods. CP: marginal likelihoods under the conventional proper prior. Top: predictions under top model. Avg: predictions under model averaging. I allowed 8 million iterations for Metropolis and 3 million for FINCS. Runtimes were: FINCS-frac (12m, 21s); FINCS-cp (13m, 1s); Met-frac (27m, 46s); Met-cp (28m, 31s).

I split the 86-month sample into a 60-month training set and a 26-month prediction set. I then used the training set to search for good models (using both FINCS-global and Metropolis) and compute posterior means $\{\hat{\Sigma}\}$ under each of the 500 best models found during the course of the search. I then used these estimates to predict observations in the 26-month validation set. In each month, the 56 "observed" returns, along with the estimated $\hat{\Sigma}$'s, were used to compute the conditional expectations of the remaining 3 "missing" returns. (The numbers 56 and 3 were chosen since it is possible to enumerate all $\binom{59}{3} = 32509$ combinations of 3 unobserved assets, thereby eliminating the possibility of error due to sampling.)

These imputations were performed using both search procedures, along with two different methods of computing graph marginal likelihoods: the fractional approach used on the previous examples, and the conventional HIW$(\delta, \tau I)$ prior. The lasso solutions were computed using leave-one-out cross-validation to choose the ℓ^1 penalty term. The predictions of both the lasso-AND graph and the lasso-OR graph are included, along with those of the MLE for the sake of comparison. Results are given both for the predictions of the top model discovered and for the model-averaged predictions of the top 500 models.

The total squared-errors of these imputations are given in Figure 5.5. These results show that FINCS leads to better predictions in less computing time than Metropolis, confirming the trends seen on simulated data.

Using fractional priors, Metropolis took over twice as long as FINCS to achieve a comparable level of error. An even larger discrepancy between algorithms appears under the HIW$(\delta, \tau I)$ prior, where Metropolis does 28% worse than FINCS in mean-squared error despite running over twice as long. The known adverse "mode-flattening" effect of the conventional prior (Carvalho and Scott, 2009) appears to hamstring the Metropolis algorithm far more than it does FINCS—a gap which narrows, but does not close entirely, under the well-behaved fractional prior.

It is also interesting that three of the four Bayesian estimates beat the lasso-AND graph. This is true regardless of whether one uses the top model or averages over the top 500 models, with the model-averaged version providing a 25% reduction in mean-squared error compared to the best lasso solution. All four Bayesian models (with and without model-averaging) beat the lasso-OR graph quite considerably—even the Bayesian procedure using the inferior prior and search algorithm. The two lasso graphs are also very different from each other, meaning that the asymptotic guarantee that the two will converge to the same answer is not especially helpful here. Without useful guidelines for choosing between them, one could easily choose the vastly inferior lasso-OR graph *ex ante*.

5.5 Discussion

This chapter has introduced a stochastic-search algorithm based upon a novel approach to computation in Gaussian graphical models: using online estimates of inclusion probabilities both to drive the choice of models to visit, and to guide a new form of global move in graph space that allows escape from the local modes that tend to thwart other procedures. Simulations suggest that FINCS outperforms Metropolis regardless of the problem size, and regardless of the assessment metric.

On balance, FINCS seems to be more reliable in this context than a pure "off-the-shelf" MCMC method. Despite apparent convergence, Gibbs sampling has an alarming tendency to miss large, important parts of model space, calling its utility here into question.

FINCS is a serial algorithm, yet gives reasonable answers on moderate-dimensional problems that up to now have required parallel methods such as Shotgun Stochastic Search. It therefore provides a crucial bridge between small problems for which Metropolis is clearly adequate, and large problems for which no serial algorithm will be competitive. It seems particularly well-suited to problems like the 59-node mutual-fund example of Section 5.4, where the predictive context naturally calls for Bayesian model averaging. In this context, FINCS gives a demonstrably better cohort of models in substantially less time than Metropolis, and is not nearly as susceptible to the mode-flattening effect of a suboptimal model-selection prior on the covariance matrix.

FINCS also finds Bayesian models yielding much better predictions than the lasso, a popular classical method often specifically lauded for its predictive optimality.

Chapter 6

Local Shrinkage Rules for Modeling Sparse Signals

The introduction of this thesis outlined a simple procedure for testing normal means based upon a two-groups model, $\theta_i \sim w \cdot g(\theta_i) + (1-w) \cdot \delta_0$ for some prior g. This procedure facilitates classification of the means as either signal or noise on the basis of the posterior inclusion probabilities w_i. As other chapters have demonstrated, this classification scheme exhibits strong control over Type-I errors, not just in the normal means example but in a wide variety of more complicated models.

Less appreciated, but just as useful, is that the two-groups model gives rise to a Bayes rule for estimating each θ_i, namely the posterior mean under the discrete mixture: $\mathrm{E}(\theta_i \mid \mathbf{y}) = w_i \cdot \mathrm{E}_g(\theta_i \mid \theta_i \neq 0, \mathbf{y})$. Under many kinds of priors g, the second term $\mathrm{E}_g(\theta_i \mid \theta_i \neq 0, \mathbf{y})$ will simply be a shrinkage rule that is linear in y_i (often depending upon, for example, the posterior expectation of a variance term, or some function thereof). Hence it is often convenient to think of this extra amount of shrinkage as being combined with the inclusion probability w_i to yield a single idiosyncratic shrinkage term $w_i^\star$, such that $\mathrm{E}(\theta_i \mid \mathbf{y}) = w_i^\star y_i$. Because of the unique structural features of the discrete-mixture model, Bayes rules of this type perform very well in the presence of sparsity, even when the goal is pure estimation (Johnstone and Silverman, 2004).

This chapter runs the same argument in the opposite direction. Instead of arriving at an estimator $w_i^\star y_i$ by beginning with a classification scheme, the $w_i^\star$ terms will be modeled directly, without making use of the discrete mixture. (Recall that even under the discrete-mixture model, $w_i^\star$ will not be identically zero, since under the usual regularity conditions there will be positive probability that $\theta_i \neq 0$.)

Why not just use the two-groups model directly? Certainly when dealing with exchangeable normal means, there is no clear rationale for abandoning the discrete mixture if one believes sparsity to be present. But in many more complicated scenarios, from generalized linear models to covariance regularization to function estimation, the marginal likelihoods necessary to compute posterior inclusion probabilities may require a difficult, often intractable integral. In situations like this, there may be considerable advantages in working with a one-group model that, in some sense, behaves like a two-groups model.

A second reason for studying this question is the popularity of the lasso (Tibshirani, 1996) for use in jointly classifying and estimating elements of, for example, a sparse vector of regression coefficients. The lasso solution has a Bayesian interpretation as the posterior mode under double-exponential priors; at the same time, the posterior mean under these priors has exactly the form of these "local shrinkage rules" described above. Hence it is important to develop theoretical tools for understanding how the double-exponential prior measures up when considered as just one member of a much larger family of Bayesian models. To put the matter simply: if the double-exponential prior turns out to be deficient as a model for describing sparsity, then the deficiency will not be repaired by using the posterior mode rather than the mean.

Finally, there are often purely sociological reasons for not using discrete mixtures. Some practitioners simply have an aversion to putting zeros in models (see, e.g. Gelman et al., 2004, page 180), and may find it more appealing to use absolutely continuous shrinkage priors.

The goal, then, is to understand what kinds of models for the local shrinkage factors $w_i^\star$ will yield performance similar to that of the two-groups model. I will study this question using exchangeable normal means as a testbed example, and will end up recommending a particular default local shrinkage rule called the horseshoe prior.

6.1 Overview of proposed methodology

Suppose one observes a p-dimensional vector $(\mathbf{y}|\boldsymbol{\theta}) \sim \mathrm{N}(\boldsymbol{\theta}, \sigma^2 I)$. Suppose further that $\boldsymbol{\theta}$ is believed to be sparse, in the sense that many of its entries are zero, or nearly so.

This chapter's proposed estimator for $\boldsymbol{\theta}$ arises as the posterior mean under an exchangeable model $\pi_H(\theta_i)$ called the horseshoe prior. This model has an interpretation as a scale mixture of normals:

$$\begin{aligned} (\theta_i \mid \lambda_i, \tau) &\sim \mathrm{N}(0, \lambda_i^2 \tau^2) \\ \lambda_i &\sim \mathrm{C}^+(0,1)\,, \end{aligned}$$

where $\mathrm{C}^+(0,1)$ is a standard half-Cauchy distribution on the positive reals.

Observe the difference from a typical model involving a common variance component τ: each θ_i is mixed over its own λ_i, and each λ_i has an independent half-Cauchy prior. This places the horseshoe in the widely studied class of multivariate scale mixtures of normals (West, 1987; Angers and Berger, 1991; Denison and George, 2000; Griffin and Brown, 2005). These might be called "local shrinkage rules," to distinguish them from global shrinkage rules with only a shared scale parameter τ.

The name "horseshoe prior" arises from the observation that

$$\mathrm{E}(\theta_i \mid \mathbf{y}) = \int_0^1 (1-\kappa_i)\, y_i \; \pi(\kappa_i \mid \mathbf{y}) \; \mathrm{d}\kappa_i = \{1 - \mathrm{E}\,(\kappa_i \mid \mathbf{y})\}\; y_i \,,$$

where $\kappa_i = 1/(1+\lambda_i^2)$, assuming $\sigma^2 = \tau^2 = 1$. After the half-Cauchy prior on λ_i is transformed, κ_i is seen to have a $\mathrm{Be}(1/2, 1/2)$ prior. By virtue of being symmetric and unbounded at both 0 and 1, this density function resembles a horseshoe. Its utility for discriminating signal from noise is readily apparent. The left side of the horseshoe, $\kappa_i \approx 0$, yields virtually no shrinkage; the right side of the horseshoe, $\kappa_i \approx 1$, yields near-total shrinkage.

Unlike most models for sparsity, the approach considered here does not involve a mixture of a point mass for noise and a density for signals. It nonetheless proves impressively accurate at handing a wide variety of sparse situations. Indeed, many of the desirable properties of heavy-tailed discrete-mixture models (Johnstone and Silverman, 2004) will be shown to obtain for the horseshoe prior as well. These "two-group" models, following the language of Efron (2008), are the recognized gold standard for handling sparsity. I will show that the horseshoe, despite being a one-group model, essentially matches the conclusions of the two-group model, with significantly less computational effort.

The horseshoe estimator has three main strengths. First, it is highly adaptive, both to unknown sparsity and to unknown signal-to-noise ratio. Second, it is robust to large, outlying signals, as I will demonstrate analytically after proving a new representation theorem similar to that of Pericchi and Smith (1992). Finally, it exhibits a strong form of multiplicity control by limiting the number of spurious signals. In this respect, the horseshoe shares one of the most attractive features of Bayesian and empirical-Bayes model-selection techniques: after a simple thresholding rule is applied, the horseshoe exhibits an automatic penalty for multiple hypothesis testing. The nature of this multiple-testing penalty is well understood in discrete mixture models, and this chapter will clarify how a similar effect occurs when the thresholded horseshoe is used instead.

6.2 The horseshoe density function

The density $\pi_H(\theta_i)$ is not expressible in closed form, but very tight upper and lower bounds in terms of elementary functions are available.

Theorem 6.2.1. *The horseshoe prior satisfies the following:*

(a) $\lim_{\theta \to 0} \pi_H(\theta) = \infty$

(b) *For* $\theta \neq 0$,

$$\frac{K}{2} \log\left(1 + \frac{4}{\theta^2}\right) < \pi_H(\theta) < K \log\left(1 + \frac{2}{\theta^2}\right) , \tag{6.1}$$

where $K = 1/\sqrt{2\pi^3}$.

Proof. Clearly,

$$\pi_H(\theta) = \int_0^\infty \frac{1}{\sqrt{2\pi\lambda^2}} \exp\left\{\frac{-\theta^2}{2\lambda^2}\right\} \frac{2}{\pi(1+\lambda^2)} \, \mathrm{d}\lambda \, .$$

Let $u = 1/\lambda^2$. Then

$$\pi_H(\theta) = K \int_0^\infty \frac{1}{1+u} \exp\left\{-\frac{\theta^2 u}{2}\right\} \, \mathrm{d}u \, ,$$

or equivalently, for $z = 1 + u$:

$$\begin{aligned} \pi_H(\theta) &= K e^{\theta^2/2} \int_1^\infty \frac{1}{z} e^{-z\theta^2/2} \, \mathrm{d}z & (6.2) \\ &= K e^{\theta^2/2} \, E_1(\theta^2/2) \, , & (6.3) \end{aligned}$$

where $E_1(\cdot)$ is the exponential integral function (closely related to the upper incomplete gamma function). This function satisfies very tight upper and lower bounds:

$$\frac{e^{-t}}{2} \log\left(1 + \frac{2}{t}\right) < E_1(t) < e^{-t} \log\left(1 + \frac{1}{t}\right)$$

for all $t > 0$, which proves Part (b). Part (a) then follows from the lower bound in Equation (6.1), which approaches ∞ as $\theta \to 0$. □

Plots of π_H, the standard double-exponential, and the standard Cauchy densities are in Figure 6.1, in which several interesting features are apparent:

- It is symmetric about zero.
- It has heavy, Cauchy-like tails that decay like θ_i^{-2}.
- It has an infinitely tall spike at 0, in the sense that the density approaches ∞ logarithmically fast as $\theta_i \to 0$ from either side.

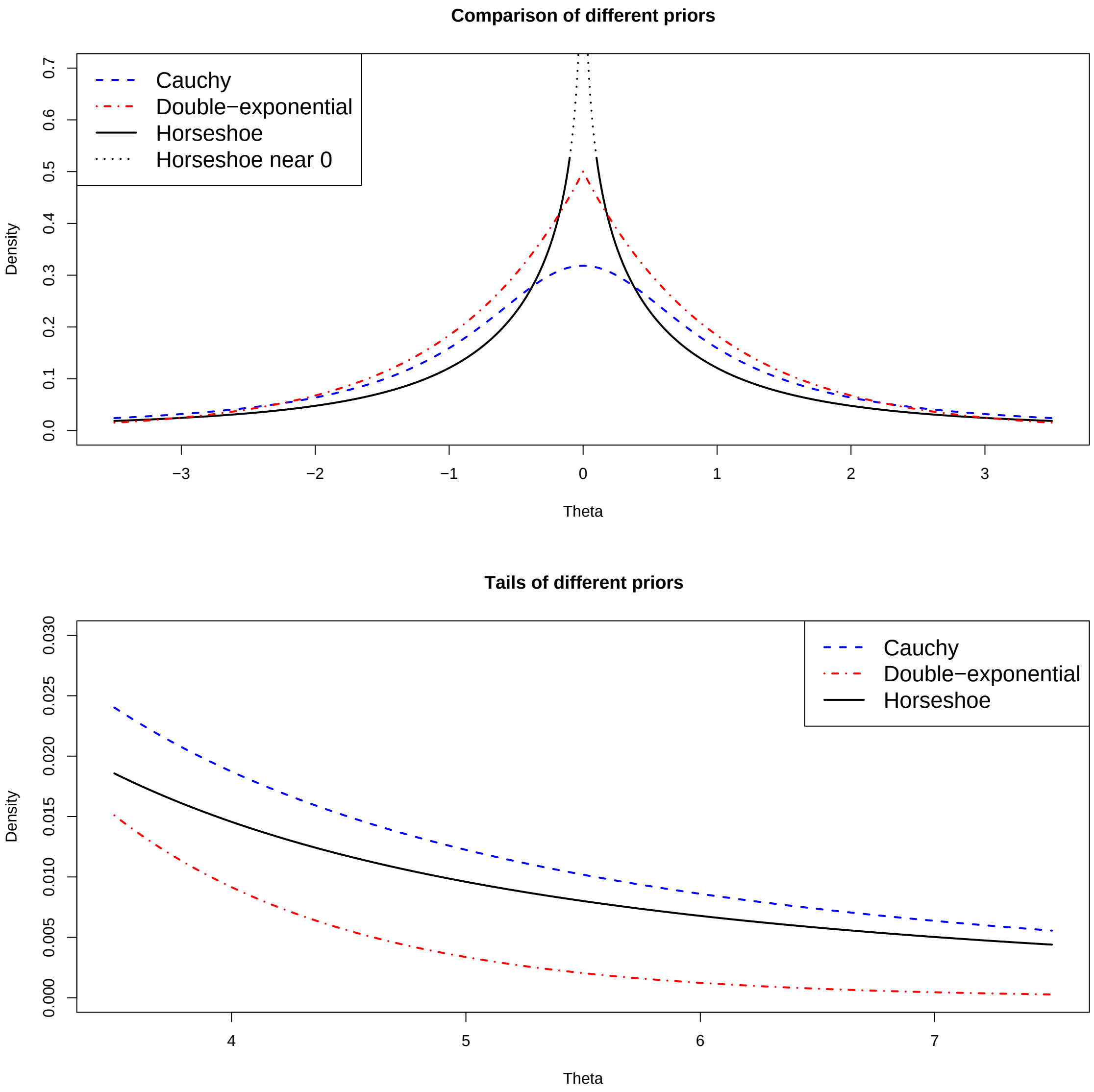


Figure 6.1: A comparison of π_H versus standard Cauchy and double-exponential densities; the dotted lines indicate that π_H approaches ∞ near 0.

The prior's flat tails allow each θ_i to be large if the data warrant such a conclusion, and yet its infinitely tall spike at zero means that the estimate can also be quite severely shrunk. These properties—heavy tails, but also a heavy spike—parallel the $\mathrm{Be}(1/2, 1/2)$ density for the shrinkage coefficient κ_i, which is unbounded both at 0 and 1.

6.3 Robust shrinkage of sparse signals

6.3.1 A representation of the posterior mean

Recall the following well-known result. If $p(y - \theta)$ is a normal likelihood of known variance, $\pi(\theta)$ is the prior for the mean θ, and $m(y) = \int p(y - \theta)\pi(\theta) \, \mathrm{d}\theta$ is the marginal density for y, then, for one sample of y:

$$\mathrm{E}(\theta \mid y) = y + \frac{\mathrm{d}}{\mathrm{d}y} \ln m(y) \,, \tag{6.4}$$

where $\sigma^2 = 1$ without loss of generality, unless otherwise noted. Versions of this result appear in Masreliez (1975), Polson (1991) and Pericchi and Smith (1992); analogous representation theorems are discussed by Brown (1971) and Stein (1981).

The theorem is useful for the insight it gives about an estimator's behavior in situations where y is very different from the prior mean. In particular, it shows that "Bayesian robustness" may be achieved by choosing a prior for θ such that the derivative of the log predictive density is bounded as a function of y. Ideally this bound should converge to 0, which from (6.4) will lead to $\mathrm{E}(\theta \mid y) \approx y$ for large $|y|$. This means that the data far out in the tails can almost completely overrule the prior.

The representation in (6.4) does not directly apply to the horseshoe prior, as $\pi(\theta)$ fails to satisfy the condition that $\pi(\theta)$ be bounded. The following theorem, however, provides an alternative representation for all $\pi(\theta)$ in the scale mixture of normals family. This family's key feature is that $\pi(\theta \mid \lambda)$ is bounded, thereby allowing the required interchange of integration and differentiation. The theorem applies to likelihoods that may be non-normal, including the normal case with σ^2 unknown.

Theorem 6.3.1. *Given likelihood $p(y-\theta)$, suppose that $\pi(\theta)$ is a mean-zero scale mixture of normals: $(\theta \mid \lambda) \sim N(0, \lambda^2)$, with λ having proper prior $\pi(\lambda)$ such that $m(y)$ is finite. Define*

$$\begin{aligned} m^\star(y) &= \int p(y-\theta)\ \pi^\star(\theta)\ d\theta \\ \pi^\star(\theta) &= \int_0^\infty \pi(\theta \mid \lambda)\ \pi^\star(\lambda)\ d\lambda \\ \pi^\star(\lambda) &= \lambda^2 \pi(\lambda)\,. \end{aligned}$$

Then

$$\begin{aligned} E(\theta \mid y) &= \frac{m^\star(y)}{m(y)} \frac{d}{dy} \ln m^\star(y) \\ &= \frac{1}{m(y)} \frac{d}{dy} m^\star(y)\,. \end{aligned} \tag{6.5}$$

Proof. Notice first that $m^\star(y)$ exists by the case $\pi(\lambda^2) \equiv 1$, which leads to the harmonic estimator in the case of a normal likelihood. This is sufficient to allow the interchange of integration and differentiation. Also note the following identities:

$$\frac{\mathrm{d}}{\mathrm{d}y} p(y-\theta) = -\frac{\mathrm{d}}{\mathrm{d}\theta} p(y-\theta) \ \text{ and } \ \lambda^2 \frac{\mathrm{d}}{\mathrm{d}\theta}\left\{\mathrm{N}(\theta \mid 0, \lambda^2)\right\} = \theta\ \mathrm{N}(\theta \mid 0, \lambda^2)\,.$$

Clearly,

$$\mathrm{E}(\theta | y) = \frac{1}{m(y)} \int \theta\ p(y-\theta)\ \mathrm{N}(\theta \mid 0, \lambda^2) \pi(\lambda)\ \mathrm{d}\theta\ \mathrm{d}\lambda\,.$$

Using integration by parts and the above identities gives

$$\mathrm{E}(\theta | y) = \frac{1}{m(y)} \int \frac{\mathrm{d}}{\mathrm{d}y} p(y-\theta) \mathrm{N}(\theta \mid 0, \lambda^2) \pi^\star(\lambda)\ \mathrm{d}\theta\ \mathrm{d}\lambda\,,$$

from which the result follows directly. □

After some algebra, (6.5) can be shown to reduce to (6.4) where $p(\mathbf{y} - \boldsymbol{\theta})$ is a normal likelihood. This extends the results from Masreliez (1975), Polson (1991), and Pericchi and Smith (1992) to a more general family of scale mixtures of normals. Hence the more familiar (6.4) will be used in the rest of the chapter.

This result also provides a complementary insight regarding an estimator's behavior for y near zero—precisely the case for sparse data. The key element is $\pi^\star(\theta)$, the unnormalized prior density for θ under $\pi^\star(\lambda) \equiv \lambda^2 \pi(\lambda)$. If $\pi^\star(\theta)$ ensures that $\mathrm{d}/\mathrm{d}y \; m^\star(y)$ is small, then $\mathrm{E}(\theta \mid y)$ will be strongly shrunk to 0. The region where $m^\star(y)$ is flat thus corresponds to the region where the estimator suggests that y is noise.

To understand at an intuitive level why the horseshoe prior yields both kinds of robustness, note two facts. First, its Cauchy-like tails ensure a redescending score function, in the venerable tradition of robust Bayes estimators involving heavy-tailed priors. Second, since $\pi(\lambda) \propto 1/(1+\lambda^2)$, then $\pi^\star(\lambda) \propto \lambda^2/(1+\lambda^2)$. This is essentially uniform, leading to $m^\star(y)$ having small derivative in a larger neighborhood near the origin than other priors. By Theorem 6.3.1, this will yield strong shrinkage of noise.

6.3.2 The horseshoe score function

The following results illustrate the horseshoe's robustness to large, outlying signals.

Theorem 6.3.2. *Suppose $y \sim \mathrm{N}(\theta, 1)$. Let $m_H(y)$ denote the predictive density under the horseshoe prior for known scale parameter τ, i.e. where $(\theta \mid \lambda) \sim \mathrm{N}(0, \tau^2\lambda^2)$ and $\lambda \sim C^+(0,1)$. Then*

$$\lim_{|y| \to \infty} \frac{d}{dy} \ln m_H(y) = 0 \,.$$

Proof. Clearly,

$$m_H(y) = \frac{1}{\sqrt{2\pi^3}} \int_0^\infty \exp\left(-\frac{y^2/2}{1+\tau^2\lambda^2}\right) \frac{1}{\sqrt{1+\tau^2\lambda^2}} \frac{1}{1+\lambda^2} \, \mathrm{d}\lambda \, .$$

Make a change of variables to $z = 1/(1+\tau^2\lambda^2)$. Then

$$\begin{aligned} m_H(y) &= \frac{1}{\sqrt{2\pi^3}} \int_0^1 \exp(-zy^2/2) \, (1-z)^{-1/2} \, \left\{\frac{1}{\tau^2} + \left(1 - \frac{1}{\tau^2}\right) z\right\}^{-1} \mathrm{d}z \\ &= \frac{2}{\tau\sqrt{2\pi^3}} \exp\left(-\frac{y^2}{2}\right) \Phi_1\left(\frac{1}{2}, 1, \frac{3}{2}, \frac{y^2}{2}, 1 - \frac{1}{\tau^2}\right) , \qquad (6.6) \end{aligned}$$

where $\Phi_1(\alpha, \beta, \gamma, x, y)$ is the degenerate hypergeometric function of two variables (Gradshteyn and Ryzhik, 1965, 9.261).

By a similar transformation, it is apparent that

$$\frac{\mathrm{d}}{\mathrm{d}y} m_H(y) = -\frac{4y}{3\tau\sqrt{2\pi^3}} \Phi_1\left(\frac{1}{2}, 1, \frac{5}{2}, \frac{y^2}{2}, 1 - \frac{1}{\tau^2}\right) .$$

Hence

$$\frac{\mathrm{d}}{\mathrm{d}y} \ln m_H(y) = -\frac{2y \; \Phi_1\left(\frac{1}{2}, 1, \frac{3}{2}, \frac{y^2}{2}, 1 - \frac{1}{\tau^2}\right)}{3 \; \Phi_1\left(\frac{1}{2}, 1, \frac{5}{2}, \frac{y^2}{2}, 1 - \frac{1}{\tau^2}\right)} \qquad (6.7)$$

Next, note the following identity from Gordy (1998):

$$\Phi_1(\alpha, \beta, \gamma, x, y) = \exp(x) \sum_{n=0}^{\infty} \frac{(\alpha)_n (\beta)_n}{(\gamma)_n} \frac{y^n}{n!} \; {}_1F_1(\gamma - \alpha, \gamma + n, -x)$$

for $0 \leq y < 1$, $0 < \alpha < \gamma$, where ${}_1F_1(a, b, x)$ is Kummer's function of the first kind, and $(a)_n$ is the rising factorial. Also note that for $y < 0$, $0 < \alpha < \gamma$,

$$\Phi_1(\alpha, \beta, \gamma, x, y) = \exp(x)(1-y)^{-\beta} \; \Phi_1\left(\gamma - \alpha, \beta, \gamma, -x, \frac{y}{y-1}\right) .$$

The final identities necessary are from Chapter 4 of Slater (1960):

$$
\begin{aligned}
{}_1F_1(a,b,x) &= \frac{\Gamma(a)}{\Gamma(b)}\, e^x\, x^{a-b} \left\{1 + O(x^{-1})\right\}, \quad x > 0 \\
{}_1F_1(a,b,x) &= \frac{\Gamma(a)}{\Gamma(b-a)}\, (-x)^{-a} \left\{1 + O(x^{-1})\right\}, \quad x < 0
\end{aligned}
$$

for real-valued x.

Hence regardless of the sign of $1 - 1/\tau^2$, expansion of (6.7) by combining these identities yields a polynomial of order y^2 or greater remaining in the denominator, from which the result follows. □

Using the previously quoted identities, it is apparent that, for fixed τ, the difference between y and $\mathrm{E}(\theta \mid y)$ is bounded for all y. The horseshoe prior is therefore of bounded influence, with the bound decaying to 0 independently of τ for large $|y|$. Denote this bound by b_τ and the horseshoe posterior mean by $\hat{\boldsymbol{\theta}}^H$ for a vector $\boldsymbol{\theta}$ of fixed dimension p. This gives the following corollary.

Corollary 6.3.3. *$E_{\mathbf{y}|\boldsymbol{\theta}}(\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}^H\|^2)$ is bounded for all $\boldsymbol{\theta}$.*

Proof.

$$
\begin{aligned}
\mathrm{E}\left\{\sum_{i=1}^{p}(\theta_i - \hat{\theta}_i^H)^2\right\} &< \mathrm{E}\left\{\sum_{i=1}^{p}(|\theta_i - y| + b_\tau)^2\right\} \\
&= p + pb_\tau^2
\end{aligned}
$$

□

Finally, by applying the result from (6.4) to (6.7), one gets an explicit form for the posterior mean:

$$\mathrm{E}_H(\theta_i \mid y_i) = y_i \left\{ 1 - \frac{2\Phi_1\left(\frac{1}{2}, 1, \frac{3}{2}, \frac{y_i^2}{2}, 1 - \frac{1}{\tau^2}\right)}{3\Phi_1\left(\frac{1}{2}, 1, \frac{5}{2}, \frac{y_i^2}{2}, 1 - \frac{1}{\tau^2}\right)} \right\} . \tag{6.8}$$

Combining (6.8) with the closed-form marginal in (6.6) allows an empirical-Bayes estimate $\mathrm{E}(\theta \mid y, \hat{\tau})$ to be computed extremely rapidly, even in very high dimensions.

6.3.3 Joint distribution for τ and the λ_i's

With the exception of Corollary 6.3.3, the above results describe the behavior of the horseshoe estimator for a single θ_i when τ is known. Usually, however, one must confront p different θ_i's along with unknown τ, leading to a joint distribution for $(\tau, \lambda_1, \ldots, \lambda_p)$. Inspection of this joint distribution yields an understanding of how sparsity is handled under the horseshoe model.

Let $\mathbf{y} = (y_1, \ldots, y_p)$. Recall that $\kappa_i = 1/(1 + \tau^2 \lambda_i^2)$, and let $\kappa = (\kappa_1, \ldots, \kappa_p)$. For the horseshoe prior, $\pi(\lambda_i) \propto 1/(1 + \lambda_i^2)$, and so

$$\pi(\kappa_i \mid \tau) \propto \kappa_i^{-1/2} (1 - \kappa_i)^{-1/2} \, \frac{1}{1 + (\tau^2 - 1)\kappa_i} .$$

Some straightforward algebra leads to

$$p(\mathbf{y}, \kappa, \tau^2) \propto \pi(\tau^2) \, \tau^p \, \prod_{i=1}^{p} \frac{e^{-\kappa_i y_i^2/2}}{\sqrt{1 - \kappa_i}} \, \prod_{i=1}^{p} \frac{1}{\tau^2 \kappa_i + 1 - \kappa_i} . \tag{6.9}$$

Form (6.9) yields several insights. As in other common multivariate scale mixtures, the global shrinkage parameter τ is conditionally independent of $\mathbf{y}$, given κ. Similarly, the κ_i's are conditionally independent of each other, given τ.

More interestingly, the marginal posterior density for κ_i is always unbounded as $\kappa_i \longrightarrow 1$, regardless of τ. The density, of course, remains integrable on the unit interval, with τ strongly influencing the amount of probability given to any set of positive Lebesgue measure containing $\kappa_i = 1$. Indeed, despite its unboundedness near 1, the posterior density of κ_i can quite heavily favor $\kappa_i \approx 0$, implying very little shrinkage in the posterior mean.

Finally, (6.9) clarifies that the global shrinkage parameter τ is estimated by the average "signal density." To see this, observe that if p is large, the conditional posterior distribution for τ^2, given κ, is well approximated by substituting $\bar{\kappa} = p^{-1} \sum_{i=1}^{p} \kappa_i$ for each κ_i. Ignoring the contribution of the prior for τ^2, this gives, up to a constant,

$$\begin{aligned} p(\tau^2 \mid \kappa) &\approx (\tau^2)^{-p/2} \left(1 + \frac{1-\bar{\kappa}}{\tau^2 \bar{\kappa}}\right)^{-p} \\ &\approx (\tau^2)^{-p/2} \exp\left\{-\frac{1}{\tau^2} \, \frac{p(1-\bar{\kappa})}{\bar{\kappa}}\right\} , \end{aligned}$$

or approximately a $\mathrm{Ga}(\frac{p+2}{2}, \frac{p-p\bar{\kappa}}{\bar{\kappa}})$ distribution for $1/\tau^2$. If $\bar{\kappa}$ is close to 1, implying that most observations are shrunk to nearly 0, then τ^2 will be very small with high probability, with an approximate mean $\mu = 2(1-\bar{\kappa})/\bar{\kappa}$ and standard deviation of $\mu/\sqrt{p-2}$.

6.4 Comparison with other Bayes estimators

6.4.1 Comparison with multivariate scale mixtures

The advantage of the horseshoe estimator can now be stated directly. In sparse situations, posterior learning about τ allows most noise observations to be shrunk to nearly zero. Yet this small value of τ will not inhibit the estimation of large, obvious

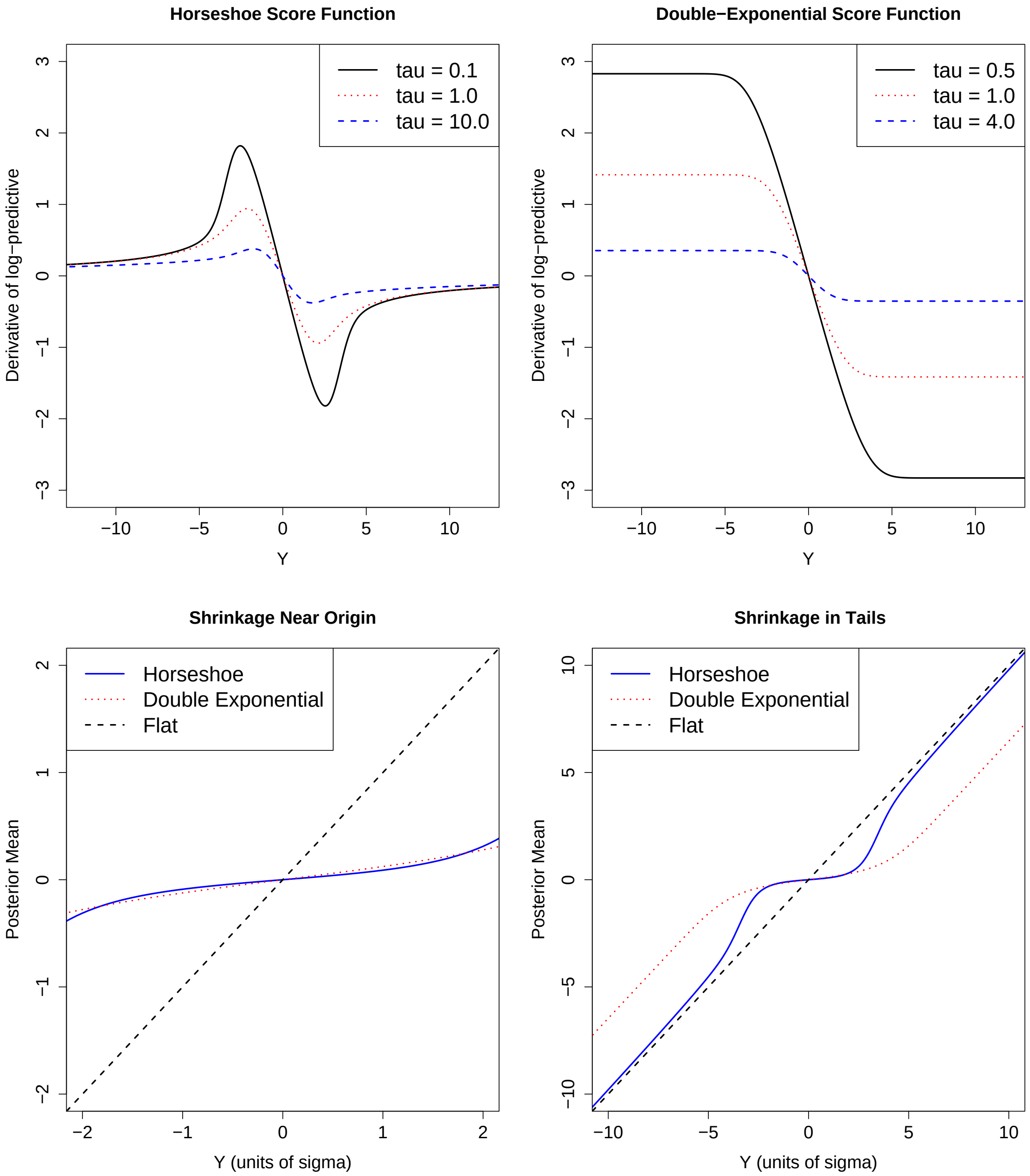


Figure 6.2: Top: A comparison of the score function for horseshoe and double–exponential priors for different values of τ. Bottom: a comparison of the posterior mean versus y for horseshoe and double-exponential priors for different values of τ, with the left pane zoomed in near the origin, and the right pane giving a broad view to encompass large signals.

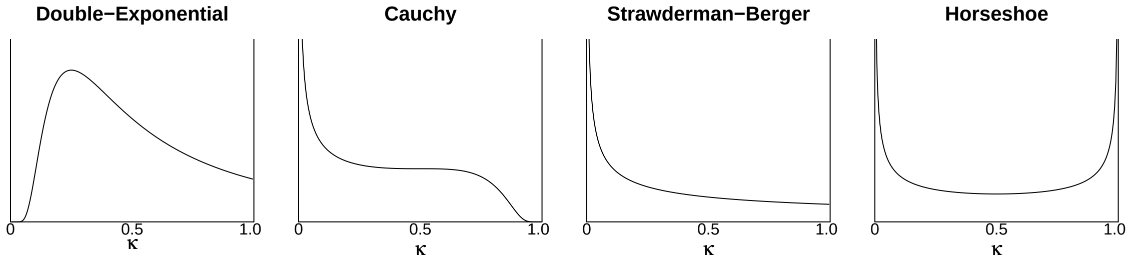


Figure 6.3: A comparison of the implied density for the shrinkage weights $\kappa_i \in [0, 1]$ for four different priors, where $\kappa_i = 0$ means no shrinkage and $\kappa_i = 1$ means total shrinkage to zero.

Table 6.1: The implied priors for κ_i and λ_i associated with some common priors for shrinkage and sparsity.

Prior for θ_i	Prior for λ_i	Prior for κ_i
Double-exponential	$\lambda_i^2 \sim \text{Ex}(2)$	$\pi_{DE}(\kappa_i) \propto \kappa_i^{-2} e^{-\frac{1}{2\kappa_i}}$
Cauchy	$\lambda_i \sim \text{IG}(1/2, 1/2)$	$\pi_C(\kappa_i) \propto \kappa_i^{-\frac{1}{2}} (1-\kappa_i)^{-\frac{3}{2}} e^{-\frac{\kappa_i}{2(1-\kappa_i)}}$
Strawderman–Berger	$\pi(\lambda_i) \propto \lambda_i (1+\lambda_i^2)^{-3/2}$	$\pi_{SB}(\kappa_i) \propto \kappa_i^{-\frac{1}{2}}$
Horseshoe	$\lambda_i \sim \text{C}^+(0, 1)$	$\pi_H(\kappa_i) \propto \kappa_i^{-1/2} (1-\kappa_i)^{-1/2}$

signals due to a redescending score function.

This set of features is not shared by common Bayes estimators based upon other multivariate scale mixtures of normals. Take, for example, the commonly used double-exponential prior, where each λ_i has an independent Ex(2) distribution. Results from Pericchi and Smith (1992) and Mitchell (1994) show that

$$
\begin{aligned}
\mathrm{E}_{DE}(\theta_i \mid y_i) &= w_i(y_i + b) + (1 - w_i)(y_i - b) \\
w_i &= F(y_i)/\{F(y_i) + G(y_i\} \\
F(y_i) &= e^{c_i}\ \Phi(-y - b) \\
G(y_i) &= e^{-c_i}\ \Phi(-y + b) \\
b &= \frac{\sqrt{2}}{\tau}\,,\ c_i = \frac{\sqrt{2}(y - \mu)}{\tau}\,,
\end{aligned}
$$

where Φ is the normal cumulative distribution function. The double-exponential posterior mean thus has an interpretation as a data-based average of $y - b$ and $y + b$, with w_i becoming arbitrarily close to 0 for large positive y, or to 1 for large negative y. This can be seen in the score function, plotted in Figure 6.2.

Under the double-exponential prior, posterior learning about τ can also cause noise observations to be squelched. This phenomenon is well understood (Park and Casella, 2008; Hans, 2008); small values of τ lead to strong shrinkage near the origin, just as under the horseshoe.

Less appreciated, however, is the effect this has upon estimation in the tails. Notice that small values of τ yield large values of b. If the overall level of sparsity in $\boldsymbol{\theta}$ is increased, τ must shrink to reduce risk at the origin. Yet the risk must be simultaneously increased in the tails, since $|\mathrm{E}_{DE}(\theta_i \mid y_i) - y_i| \approx b$ for large $|y_i|$. When $\boldsymbol{\theta}$ is sparse, estimation of τ under the double-exponential model must strike a balance between reducing risk in estimating noise, and reducing risk in estimating large signals. This is demonstrated by the bottom two panels of Figure 6.2, which also show that no such tradeoff exists under the horseshoe prior.

These properties, and the properties of other common Bayes rules, can be understood intuitively in terms of the shrinkage coefficient κ_i. Table 6.1 gives the priors for κ_i implied by four different multivariate-scale-mixture priors: the double-exponential, the Strawderman–Berger density (Strawderman, 1971; Berger, 1980), the Cauchy, and the horseshoe. Figure 6.3 also plots these four densities.

Notice that at $\kappa_i = 0$, both $\pi_C(\kappa_i)$ and $\pi_{SB}(\kappa_i)$ are both unbounded, while $\pi_{DE}(\kappa_i)$ vanishes. This suggests that Cauchy and Strawderman–Berger estimators will be good, and double-exponential estimators will be mediocre, at leaving large signals unshrunk. Meanwhile, at $\kappa_i = 1$, π_C tends to zero, while both π_{DE} and π_{SB} tend to fixed constants. This suggests that Cauchy estimators will be poor, while double-exponential and Strawderman–Berger estimators will be mediocre, at correctly shrinking noise all the way to 0.

The horseshoe prior, on the other hand, implies a Beta$(1/2, 1/2)$ distribution for κ_i. This is clearly unbounded both at 0 and 1, allowing both signals and noise to accommodated by a single prior.

6.4.2 An illustrative example

An example will help illustrate these ideas. Two standard normal observations were simulated for each of 1000 means: 10 signals of mean 10, 90 signals of mean 2, and 900 noise of mean 0. Two models were then fit to this data: one that used independent horseshoe priors for each θ_i, and one that used independent double-exponential priors. For both models, a half-Cauchy prior was used for the global scale parameter τ following the advice of Gelman (2006) and Scott and Berger (2006), and Jeffreys' prior $\pi(\sigma) \propto 1/\sigma$ was used for the error variance.

The shrinkage characteristics of these two fits are summarized in Figure 6.4. These

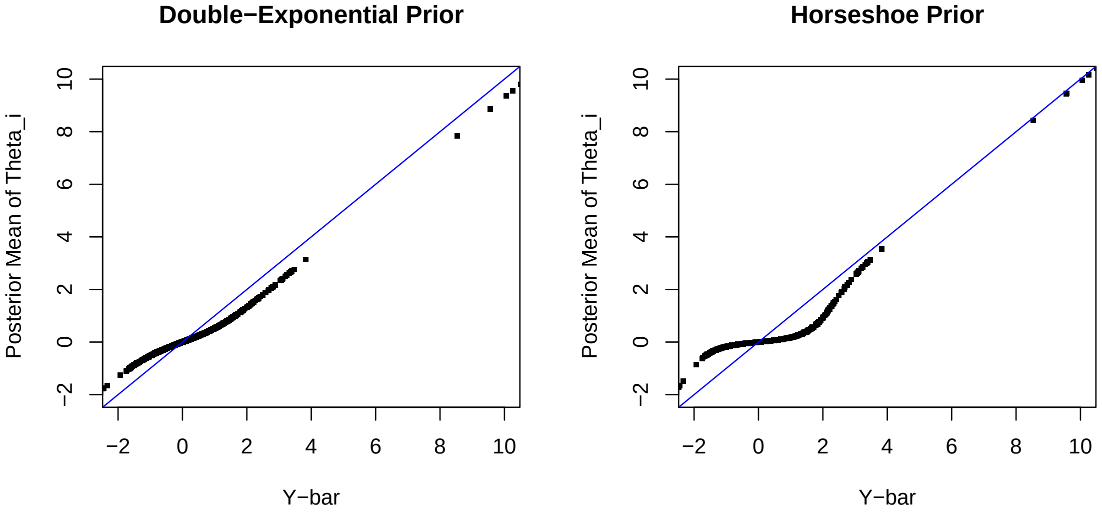


Figure 6.4: Plots of $\bar{y}_i$ versus $\hat{\theta}_i$ for double-exponential (left) and horseshoe (right) priors on data where most of the means are zero. The diagonal lines are where $\hat{\theta}_i = \bar{y}_i$.

plots show the posterior mean $\hat{\theta}_i^{DE}$ and $\hat{\theta}_i^{H}$ as a function of the observed data $\bar{y}_i$, with the diagonal lines showing where $\hat{\theta}_i = \bar{y}_i$. Key differences occur near $\bar{y}_i \approx 0$ and $\bar{y}_i \approx 10$. Compared with the horseshoe prior, the double-exponential prior tends to under-shrink small observations and over-shrink large ones.

These differences can also be seen in Figure 6.5. The left panel shows that the global shrinkage parameter τ is estimated to be much smaller under the horseshoe model than under the double-exponential model, roughly 0.2 versus 0.7. But under the horseshoe model, the local shrinkage parameters can take on quite large values and hence overrule this global shrinkage; this handful of large λ_i's under the horseshoe prior, corresponding to the observations near 10, can be seen in the right panel.

The horseshoe prior easily handles both aspects of the problem: leaving large signals unshrunk while squelching most of the noise. The double-exponential prior, in contrast, requires a delicate balancing act in estimating τ, which affects error near

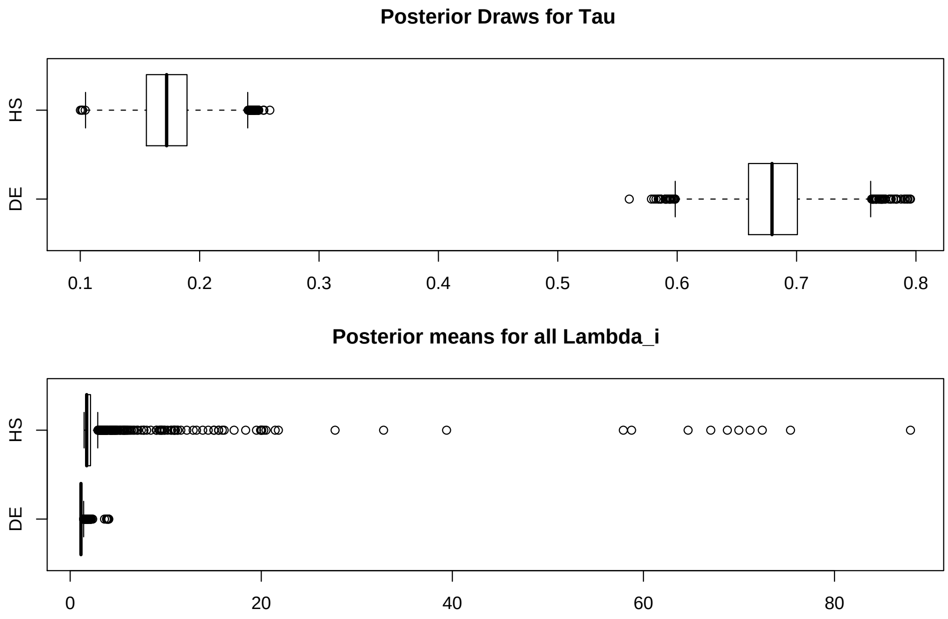


Figure 6.5: Top: posterior draws for the global shrinkage parameter τ under both double-exponential and horseshoe for the toy example. Bottom: boxplot of $\hat{\lambda}_i$'s, the posterior means of the local shrinkage parameters λ_i.

0 and error in the tails. That this balance is often hard to strike is reflected in the mean-squared error, which was about 25% lower under the horseshoe model for this example.

6.4.3 Comparison with discrete-mixture rules

Recall the discrete mixture model mentioned in the introduction of this chapter:

$$\theta_i \sim (1-w)\delta_0 + w \; g(\theta_i) \,, \tag{6.10}$$

with the mixing probability w being unknown.

The crucial choice here is that of g, which must allow convolution with a normal likelihood in order to evaluate the predictive density under the alternative model g. One common choice is a normal prior, the properties of which are well understood in a multiple-testing context (Scott and Berger, 2006; Bogdan et al., 2008a). Also see Johnstone and Silverman (2004) for an empirical-Bayes treatment of a heavy-tailed version of this model.

For these and other conjugate choices choices of g, it is straightforward to compute the posterior inclusion probabilities, $w_i = \Pr(\theta_i \neq 0 \mid \mathbf{y})$. As noted before, these quantities will adapt to the level of sparsity in the data through shared dependence upon the unknown mixing probability w. The discrete mixture can be thought of as adding a point mass at $\kappa_i = 1$, allowing total shrinkage. In broad terms, this is much like the unbounded density under the horseshoe prior as $\kappa_i \to 1$, suggesting that these models may be similar in the degree to which they shrink noise variables to 0.

Hence it is interesting to consider the differences between shrinkage profiles of the horseshoe (π_H) and the discrete mixture using Strawderman–Berger priors (π_{DM}):

$$\pi_{DM}: \ (\theta_i \mid \kappa_i) \sim (1-w)\ \delta_0 + w\ \mathrm{N}\left(0, \frac{\tau^2+\sigma^2}{2\kappa_i} - \sigma^2\right),$$

with $\kappa_i \sim \mathrm{Be}(1/2, 1)$ and $\tau > \sigma$. The Strawderman–Berger prior has a number of desirable properties for describing the nonzero θ_i's, since it is both heavy-tailed and yet still allows closed-form convolution with the normal likelihood.

Both the horseshoe and the discrete mixture have global scale parameters τ and local shrinkage weights κ_i. Yet τ plays a very different role in each model. Under the horseshoe prior,

$$\mathrm{E}_H(\theta_i | \kappa_i, \tau, y_i) = \left(1 - \frac{\kappa_i}{\kappa_i + \tau^2(1-\kappa_i)}\right) y_i\,, \tag{6.11}$$

recalling that σ^2 is assumed to be 1. And in the mixture model,

$$\mathrm{E}_{DM}(\theta_i|\kappa_i, w_i, y_i) = w_i \left(1 - \frac{2\kappa_i}{1+\tau^2}\right) y_i \, ,$$

where w_i is the posterior inclusion probability. Let $G^*(y_i \mid \tau)$ denote the predictive density, evaluated at y_i, under the Strawderman–Berger prior. Then these probabilities are

$$\frac{w_i}{1-w_i} = \frac{w \; G^*(y_i \mid \tau)}{(1-w) \; \mathrm{N}(y_i \mid 0, 1)} \, . \tag{6.12}$$

Several differences between the approaches are apparent:

- In the discrete model, local shrinkage is controlled by the Bayes factor in (6.12), which is a function of the signal-to-noise ratio τ/σ. In the scale-mixture model, local shrinkage is determined entirely by the κ_i's, or equivalently the λ_i's.

- In the discrete model, global shrinkage is primarily controlled through the prior inclusion probability w. In the scale-mixture model, global shrinkage is controlled by τ, since if τ is small in (6.11), then κ_i must be very close to 0 for extreme shrinkage to be avoided. Hence in the former model, w adapts to the overall sparsity of $\boldsymbol{\theta}$, while in the latter model, τ performs this role.

- There will be a strong interaction between w and τ in the discrete model which is not present in the scale-mixture model. Intuitively, as w changes, τ must adapt to the scale of the observations that are reclassified as signals or noise.

Nonetheless, these structural differences between the procedures are small compared to their operational similarities, as Sections 6.5 and 6.6 will show. Both models imply priors for κ_i that are unbounded at 0 and at 1. They have similarly favorable risk properties for estimation under squared-error or absolute-error loss. Perhaps most remarkably, they yield thresholding rules that are nearly indistinguishable in practice despite originating from very different goals.

6.5 Estimation risk

6.5.1 Overview

This section describes the results of a large bank of simulation studies intended to assess the risk properties of the horseshoe prior under both squared-error and absolute-error loss. I will benchmark its performance against three alternatives: the double-exponential model, along with fully Bayesian and empirical-Bayes versions of the discrete-mixture model with Strawderman–Berger priors. The study involves simulating repeatedly from models that correspond to archetypes of "strong" sparsity (in Experiment 1) and "weak" sparsity (in Experiment 2), but that match none of the priors under evaluation.

For both the double-exponential and the horseshoe, the following default hyperpriors were used in all studies:

$$
\begin{aligned}
\pi(\sigma) &\propto 1/\sigma \\
(\tau \mid \sigma) &\sim \mathrm{C}^{+}(0, \sigma)\,.
\end{aligned}
$$

A slight modification is necessary in the fully Bayesian discrete-mixture model, since under the Strawderman–Berger prior of (6.4.3), τ must be larger than σ:

$$
\begin{aligned}
w &\sim \mathrm{U}(0,1) \\
\pi(\sigma) &\propto 1/\sigma \\
(\tau \mid \sigma) &\sim \mathrm{C}(\sigma, \sigma)\; 1_{\tau \geq \sigma}\,.
\end{aligned}
$$

These are similar to the recommendations of Scott and Berger (2006), with the half-Cauchy prior on τ being appropriately scaled by σ and yet having only a polynomial rate of decay. In the empirical-Bayes approach, w, σ, and τ were estimated by marginal maximum likelihood, subject to the constraint that $\tau \geq \sigma$.

Experiment 1: Strongly sparse signals

Strongly sparse signals are vectors in which some of the components are identically zero; this notion is operationalized in the following model:

$$\begin{aligned} y_i &\sim \mathrm{N}(\theta_i, \sigma^2) \\ \theta_i &\sim w \; t_3(0, \tau) + (1 - w) \; \delta_0 \\ w &\sim \mathrm{Be}(1, 4) \,, \end{aligned}$$

where the nonzero θ_i's follow a t distribution with 3 degrees of freedom, and where $\boldsymbol{\theta}$ has 20% nonzero entries on average.

I simulated from this model under many different configurations of the signal-to-noise ratio. In all cases τ was fixed at 3, and results are given for 1000 simulated data sets for each of $\sigma^2 = 1$ and $\sigma^2 = 9$. Each simulated vector was of length 250.

Experiment 2: Weakly sparse signals

A vector $\boldsymbol{\theta}$ is considered weakly sparse if none of its components are identically zero, but its component nontheless follow some kind of power-law or l^α decay; see Johnstone and Silverman (2004) for a more formal description. Weakly sparse vectors have most of their total "energy" concentrated on relatively few elements.

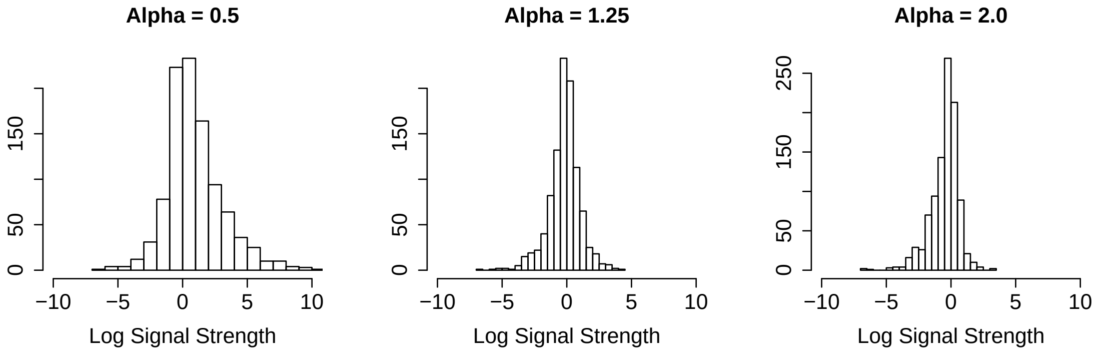


Figure 6.6: Log signal strength for three weak-l^α vectors of 1000 means, with α at 0.5, 1.25, and 2.0.

I simulated 1000 weakly sparse data sets of 250 means each, where

$$\begin{aligned} y_i &\sim \mathrm{N}(\theta_i, \sigma^2) \\ (\theta_i \mid \eta, \alpha) &\sim \mathrm{U}(-\eta c_i, \eta c_i) \\ \eta &\sim \mathrm{Ex}(2) \\ \alpha &\sim \mathrm{U}(a, b)\,, \end{aligned}$$

for $c_i = n^{1/\alpha}\ i^{-1/\alpha}$ for $i = 1, \ldots, n$. These $\boldsymbol{\theta}$'s correspond to a weak-l^α bound on the coefficients, as described by Johnstone and Silverman (2004): the ordered θ_i's follow a power-law decay, with the size of the largest coefficients controlled by the exponentially distributed random bound η, and the speed of the decay controlled by the random norm α.

Two experiments were conducted: one where α was drawn uniformly on $(0.5, 1)$, and another where α was uniform on $(1, 2)$. Small values of α give vectors where the cumulative signal strength is concentrated on a few very large elements. Larger values of α, on the other hand, yield vectors where the signal strength is more uniformly distributed among the components. For illustration, see Figure 6.6, which shows the

Table 6.2: Risk under squared-error loss and absolute-error loss in experiment 1. Bold diagonal entries in the top and bottom halves are median sum of squared-errors and absolute errors, respectively, in 1000 simulated data sets. Off-diagonal entries are average risk ratios, risk of row divided by risk of column, in units of σ. DE: double exponential. HS: horseshoe. DMF: discrete mixture, fully Bayes. DME: discrete mixture, empirical Bayes.

		$\sigma^2 = 1$				$\sigma^2 = 9$			
		DE	HS	DMF	DME	DE	HS	DMF	DME
SE Loss	DE	**209**	1.62	1.62	1.71	**850**	1.47	1.51	1.50
	HS		**77**	0.95	1.04		**416**	0.99	0.99
	DMF			**93**	1.18			**440**	1.00
	DME				**74**				**437**
AE Loss	DE	**178**	1.50	1.60	1.73	**341**	1.56	1.75	1.76
	HS		**80**	1.02	1.13		**142**	1.10	1.10
	DMF			**83**	1.20			**123**	1.00
	DME				**60**				**122**

log signal-strength of three weak-l^α vectors of 1000 means with α fixed at each of 0.5, 1.25, and 2.0.

6.5.2 Results

The results of Experiments 1 and 2 are summarized in Tables 6.2 and 6.3

Two conclusions are readily apparent from the tables. First, the double-exponential is systematically inferior to the horseshoe, and to both versions of the discrete mixture rule, under both squared-error and absolute-error loss. The difference in performance is substantial. In experiment 1, the double-exponential averaged between 50% and 75% more risk regardless of the specific value of σ^2 and regardless of which loss function is used. In experiment 2, the double-exponential typically had $\approx$ 25–40% more risk.

Close inspection of some of these simulated data sets led us to conclude that the double-exponential prior suffers in two different ways here, much as it did in the toy example summarized in Figures 6.4 and 6.5. It lacks tails that are heavy enough

Table 6.3: Risk under squared-error loss and absolute-error loss in experiment 2. Bold diagonal entries in the top and bottom halves are median sum of squared-errors and absolute errors, respectively, in 1000 simulated data sets. Off-diagonal entries are average risk ratios, risk of row divided by risk of column, in units of σ. DE: double exponential. HS: horseshoe. DMF: discrete mixture, fully Bayes. DME: discrete mixture, empirical Bayes.

		$\alpha \in (0.5, 1.0)$				$\alpha \in (1.0, 2.0)$			
		DE	HS	DMF	DME	DE	HS	DMF	DME
SE Loss	DE	**231**	1.36	1.42	1.38	**139**	1.34	1.34	1.32
	HS		**170**	1.05	1.01		**69**	0.97	0.96
	DMF			**194**	0.95			**73**	0.99
	DME				**227**				**73**
AE Loss	DE	**189**	1.23	1.31	1.30	**144**	1.23	1.24	1.23
	HS		**148**	1.07	1.05		**91**	1.00	0.99
	DMF			**142**	0.98			**92**	1.00
	DME				**150**				**92**

to robustly estimate the large t_3 signals, and it also lacks sufficient mass near 0 to adequately squelch the substantial noise in $\boldsymbol{\theta}$.

Like any shrinkage estimator, the double-exponential model is fine when its prior describes reality, but suffers in problems that correspond to very reasonable, commonly held notions of sparsity. The horseshoe, on the other hand, allows each shrinkage coefficient κ_i to be arbitrarily close to 0 or 1, and hence can accurately estimate both signal and noise.

Second, it is equally interesting that no meaningful systematic edges could be found for any of the other three approaches: the horseshoe, the Bayes discrete mixture, or the empirical-Bayes discrete mixture. All three models have heavy tails, and all three models can shrink y_i arbitrarily close to 0. Though only a limited set of results are summarized here, this trend held for a wide variety of other data sets.

In general, the horseshoe estimator, despite providing a one-group answer, acts in a similar fashion to the model-averaged Bayes estimator arising from a two-group mixture.

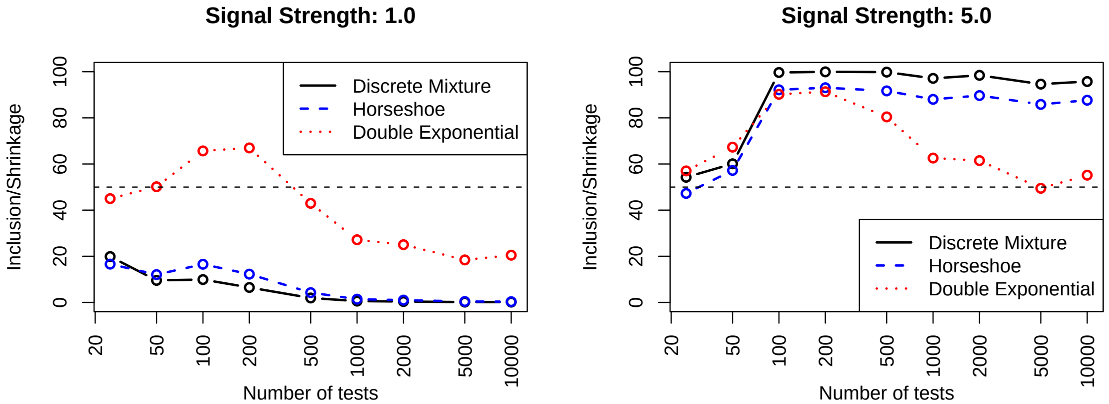


Figure 6.7: Inclusion probabilities/significance weights w_i versus number of tests n in Experiment 3.

6.6 Classification risk

6.6.1 Overview

This section now describes a simple thresholding rule for the horseshoe estimator that can yield accurate decisions about whether each θ_i is signal or noise. These classifications turn out to be nearly instinguishable from those of the Bayesian discrete-mixture model under a simple 0–1 loss function, suggesting an even deeper correspondence between the two procedures than was shown in the previous section.

Recall that under the under the discrete mixture model of (6.10), the Bayes estimator for each θ_i is $\hat{\theta}_i^{DM} = w_i E_g(\theta_i \mid y_i)$, where w_i is the posterior inclusion probability for θ_i. For appropriately heavy-tailed g such as the Strawderman–Berger prior of Section 6.4.3, this expression is approximately $w_i y_i$, meaning that w_i can be construed in two different ways:

- as a posterior probability, which forms the basis for a classification rule that is optimal in both Bayesian and frequentist senses.

- as an indicator of how much shrinkage should be performed on y_i, thereby giving rise to an estimator $\hat{\theta}_i^{DM} \approx w_i y_i$ with excellent risk properties under squared-error and absolute-error loss.

The horseshoe estimator also yields "significance weights" $w_i = 1 - \kappa_i$, with $\hat{\theta}_i^H = w_i y_i$. Though these lack an interpretation as posterior probabilities, the previous section showed that these weights behave similarly to those arising from the discrete mixture. Hence by analogy with the decision rule one would apply to the discrete-mixture w_i's under a symmetric 0–1 loss function, one possible threshold is to call θ_i a signal if the horseshoe yields $w_i \geq 0.5$, and to call it noise otherwise.

The following simulations will demonstrate the surprising fact that, even though the horseshoe w_i's are not posterior probabilities, and even though the horseshoe model itself makes no allowance for two different groups, this simple thresholding rule nonetheless displays very strong control over the number of false-positive classifications. Indeed, in problem after problem, it is hard to tell the difference between the w_i's from the two-group model and those from the horseshoe.

I will study two different asympotic scenarios: "fixed-k" asymptotics and "ideal signal-recovery" asymptotics.

Experiment 3: Fixed-k asymptotics

Under fixed-k asymptotics, the number of true signals remains fixed, while the number of noise observations grows without bound. I studied this asymptotic regime by fixing 10 true signals that are repeatedly tested in the face of an increasingly large number of noise observation. The error variance σ^2 remains fixed at 1 throughout. The 10 signals were the half-integers between 0.5 and 5.0 with random signs.

Experiment 4: Ideal signal-recovery asymptotics

Unfortunately, fixed-k asymptotics are in some sense hopeless for signal recovery: as $n \to \infty$, every signal must eventually be classified as noise under the discrete mixture model, since each Bayes factor remains bounded while the prior odds ratio comes to favor the null hypothesis arbitrarily strongly.

A set of asymptotic conditions does exist, however, that makes near-perfect signal recovery possible. Suppose that the nonzero θ_i's follow a $\mathrm{N}(0, \tau^2)$ distribution. Define $s = \tau^2/\sigma^2$, and recall that p is the fraction of signals among all observations. Bogdan et al. (2008b) show that if

$$s \to \infty \tag{6.13}$$

$$w \to 0 \tag{6.14}$$

$$s^{-1} \log \left(\frac{1-w}{w} \right) \to C \quad \text{for some} \quad 0 < C < \infty \,, \tag{6.15}$$

then as the number of tests n grows without bound, the probability of a false positive ($w_i \geq 0.5, \theta_i = 0$) converges to 0, while the probability of a false negative ($w_i < 0.5, \theta_i \neq 0$) converges to a fixed constant less than one, the exact value of which will depend upon C.

Essentially, the situation is one in which the fraction of signal observations can approach 0 as long as the signal-to-noise ratio s is growing larger at a sufficiently rapid rate. The Bayesian mixture model can then recover all but a fixed fraction of the true signals without committing any Type-I errors.

To study ideal signal-recovery asymptotics, I used the same 10 signal observations as under the fixed-k asymptotics. The variance of the noise observations, however, decayed to 0 as n grew, instead of remaining fixed at 1:

$$\sigma_n^2 = \frac{D}{\log\left(\frac{n-k}{k}\right)} \, ,$$

where D is any constant, n is the total number of tests being performed, and k is the fixed number of true signals. This experiment set $D = 0.4$ and $k = 10$. It is straightforward to show that as $n \to \infty$, the conditions of (6.13), (6.14), and (6.15) will hold, and the results of Bogdan et al. (2008b) will obtain, if for each sample size the noise variance is σ_n^2.

6.6.2 Results

The general character of the results can be understood from Figure 6.7. These two plots shows the inclusion probabilities/significance weights for two signals—a weak signal of 1σ, and a strong signal of 5σ—as the number of noise observations increases gradually from 10 to $10,000$ under fixed-k asymptotics.

Intuitively, the weak signal should rapidly be overwhelmed by noise, while the strong signal should remain significant. This is precisely what happens under both the discrete mixture and the horseshoe prior, whose significance weights coincide almost perfectly as n grows. This is not, however, what happens under the double-exponential prior, which shrinks the strong signal almost as much as it shrinks the weak signal at all levels.

Comprehensive results for Experiments 3 and 4 are given in Tables 6.4–6.9. A close inspection of these numbers confirms the findings in Figure 6.7: regardless of the asymptotic regime and regardless of the signal strength, the horseshoe significance weights are a good stand-in for the posterior inclusion probabilities under the

Table 6.4: Posterior probabilities as n grows for 10 fixed signals; discrete mixture model, fixed-k asymptotics.

	Signal Strength										
# Tests	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	FP
25	18	20	23	27	31	35	39	44	49	54	0
50	8	10	12	15	19	25	32	41	50	60	0
100	8	10	15	27	46	69	86	95	99	100	0
200	5	6	11	21	42	70	90	98	100	100	2
500	1	2	3	6	14	35	67	91	98	100	1
1000	0	1	1	2	3	9	24	55	85	97	0
2000	0	0	1	1	3	8	24	59	89	98	0
5000	0	0	0	0	1	3	9	32	72	95	0
10000	0	0	0	0	1	3	9	32	74	96	3

Table 6.5: Significance weights as n grows for 10 fixed signals; horseshoe prior, fixed-k asymptotics.

	Signal Strength										
# Tests	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	FP
25	15	17	18	21	24	28	32	37	42	47	0
50	11	12	14	17	20	26	33	40	49	57	0
100	14	17	22	31	46	62	75	85	89	92	0
200	11	12	17	26	43	63	79	87	91	93	1
500	4	4	6	10	18	36	61	80	89	92	1
1000	1	1	2	3	5	10	25	52	76	88	0
2000	1	1	1	2	4	9	24	54	80	90	0
5000	0	0	0	1	1	3	10	33	67	86	0
10000	0	0	0	1	1	3	10	30	68	88	2

Table 6.6: Significance weights as n grows for 10 fixed signals; double-exponential prior, fixed-k asymptotics.

	Signal Strength										
# Tests	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	FP
25	44	45	46	48	50	51	53	55	56	57	0
50	47	50	52	55	58	60	62	64	66	67	5
100	60	66	72	77	81	84	86	88	89	90	8
200	59	67	73	79	83	86	88	89	90	91	29
500	39	43	49	56	62	68	72	76	78	80	15
1000	26	27	30	34	39	44	50	54	59	63	0
2000	23	25	27	31	36	41	47	53	58	61	0
5000	18	18	20	22	25	30	35	40	45	49	0
10000	19	20	22	25	29	34	39	45	50	55	2

Table 6.7: Posterior probabilities as n grows for 10 fixed signals; discrete mixture model, ideal signal-recovery asymptotics.

	Signal Strength										
# Tests	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	FP
25	12	13	15	18	20	24	27	31	36	42	0
50	45	63	82	94	98	100	100	100	100	100	11
100	21	37	68	92	99	100	100	100	100	100	3
200	8	23	69	97	100	100	100	100	100	100	3
500	3	13	67	99	100	100	100	100	100	100	2
1000	2	11	76	100	100	100	100	100	100	100	1
2000	1	9	79	100	100	100	100	100	100	100	2
5000	1	5	76	100	100	100	100	100	100	100	0
10000	0	4	82	100	100	100	100	100	100	100	1

Table 6.8: Significance weights as n grows for 10 fixed signals; horseshoe prior, ideal signal-recovery asymptotics.

	Signal Strength										
# Tests	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	FP
25	12	13	14	16	18	21	24	28	32	38	0
50	38	54	71	84	90	94	95	97	97	98	0
100	25	40	64	82	91	94	96	97	97	98	1
200	16	30	64	86	92	95	96	97	98	98	2
500	7	19	62	89	94	96	97	98	98	99	1
1000	5	15	69	91	95	96	97	98	98	99	0
2000	3	11	70	92	95	97	98	98	99	99	0
5000	1	6	70	93	96	97	98	98	99	99	0
10000	1	5	74	94	96	97	98	99	99	99	0

Table 6.9: Significance weights as n grows for 10 fixed signals; double-exponential prior, ideal signal-recovery asymptotics.

	Signal Strength										
# Tests	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	FP
25	37	39	40	41	43	44	45	47	48	50	0
50	89	93	95	96	97	97	98	98	98	98	0
100	95	97	98	98	99	99	99	99	99	99	4
200	81	90	93	95	96	97	97	97	98	98	3
500	71	84	89	92	94	95	95	96	96	97	6
1000	69	83	89	92	93	94	95	96	96	97	7
2000	57	74	83	87	90	91	93	94	94	95	12
5000	44	63	75	81	85	87	89	91	92	92	11
10000	36	54	69	77	81	85	87	88	90	91	9

discrete mixture. They lead to nearly identical numerical summaries of the strength of evidence in the data, and nearly identifical classifications of signal versus noise. One anomaly worth mentioning occurs in Table 6.6.2, which shows 11 false positives on 50 tests for the discrete mixture, against none for the horseshoe. This difference seems large, but further inspection showed that of these 11 discrepencies, all but one had their w_i's within 5% for the horseshoe and the discrete mixture.

The horseshoe thresholding rule is, of course, quite accurate in its own right, aside from its correspondence with the discrete mixture. As the tables show, it exhibits very strong control over the number of false positive declarations while retaining a reasonable amount of power, even under fixed-k asymptotics.

On the other hand, there is no sense in which the significance weights from the double-exponential can be relied upon to sort signal from noise. These weights are inappropriately uniform as a function of signal strength, suggesting that the underlying joint model for τ and κ_i cannot adapt sufficiently to the degree of sparsity in the data.

6.7 Estimation of hyperparameters

6.7.1 Bayes, empirical Bayes, and cross-validation

This section discusses the estimation of key model hyperparameters.

One possibility is to proceed with a fully Bayesian solution by placing priors upon model hyperparameters. An excellent reference on hyperpriors for variance components can be found in Gelman (2006). A second possibility is to estimate σ and τ, along with w if the discrete mixture model is being used, by empirical Bayes. Marginal maximum-likelihood solutions along these lines are explored in, for example, George and Foster (2000) and Johnstone and Silverman (2004). A third possibility is cross-validation, a common approach when fitting double-exponential priors to regression models (Tibshirani, 1996).

Empirical Bayes solutions under the horseshoe prior are extremely fast to compute, since once a two-dimensional optimization over τ and σ is undertaken, all posterior expectations are known. Caution, however, is in order. When few signals are present, it is quite common for the posterior mass of τ to concentrate near 0 and for the signals to be flagged via large values of the local shrinkage parameters λ_i. The marginal maximum-likelihood solution is therefore always in danger of collapsing to the degenerate $\hat{\tau} = 0$. See, for example, Tiao and Tan (1965).

Cross-validation will not pose this problem, but it still involves plugging in a point estimate for the signal-to-noise ratio. This can be misleading, given that σ and τ will typically have an unknown correlation structure. Indeed, as the following example serves to illustrate, careful handling of uncertainty in the joint distribution for τ and σ can be crucial.

Example: Suppose the true model is $\theta = 20$ and $\sigma^2 = 1$. Two observations are available: $y_1 = 19.6$ and $y_2 = 20.4$. Two different Bayesian versions of the horseshoe model in (6.1) are considered. In both cases, σ is unknown and assigned the noninformative prior $1/\sigma$. But in the first fit, τ is assigned a $\mathrm{C}^+(0, 1)$ prior, while in the second fit, τ is assigned a $\mathrm{C}^+(0, \sigma)$ distribution, allowing it to scale with the uncertain error variance.

The two posterior distributions for θ under these fits are shown in Figure 6.8. In the first fit using absolute scaling for τ, the posterior is bimodal, with one mode

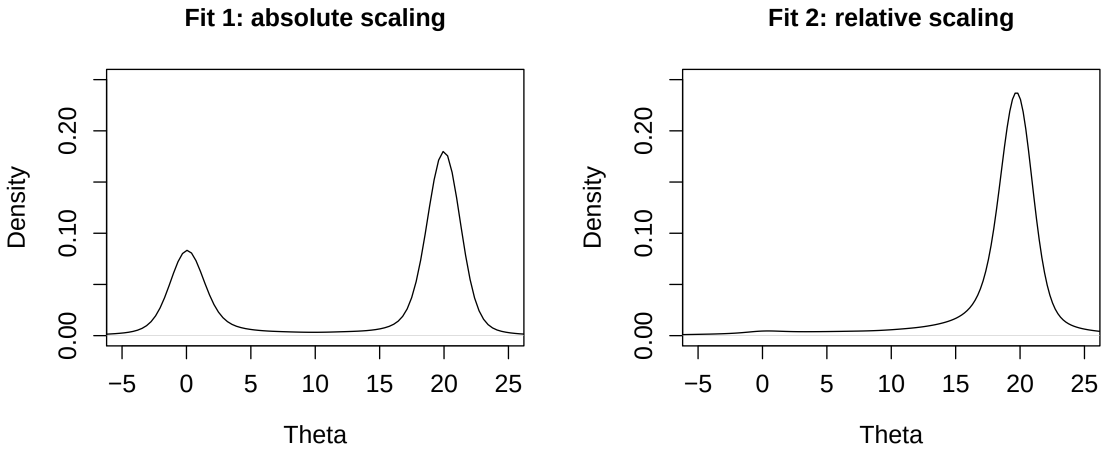


Figure 6.8: Example 1. Left: the posterior for θ when $\tau \sim C^{+}(0,1)$. Right: the posterior when $\tau \sim \mathrm{C}^{+}(0,\sigma)$.

around 20 and the other around 0. This bimodality is absent in the second fit, where τ was allowed to scale relative to σ.

A situation with only two observations is highly stylized, yet the differences between the two fits are still striking. Note that the issue is not one of failing to condition on σ in the prior for τ; indeed, the first fit involved plugging the true value of σ into the prior for τ, which is exactly what an empirical-Bayes analysis aims to accomplish asymptotically. Rather, the issue is one of averaging over uncertainty about σ in estimating the signal-to-noise ratio. Similar phenomena can be observed with other scale mixtures; see Fan and Berger (1992) for a general discussion of the issue.

6.7.2 A notable discrepency under the double-exponential prior

Chapter 3 gave a detailed comparison of Bayes and empirical-Bayes approaches for handling w, the prior inclusion probability, in the context of discrete mixture models for variable selection. Since in the horseshoe model, τ plays the role of w in exercising multiplicity control, an analogous set of issues may arise here.

A full theoretical discussion of these issues is beyond the scope of this thesis. Nonetheless, the following example serves as a warning that marginalizing over uncertainty in hyperparameters can drastically change the implied regularization penalty. In light of the results from Chapter 3, it is entirely possible that this difference between Bayesian and plug-in analyses will not disappear even in the limit. While I do not have results that mirror the specificity of Chapter 3, my conjecture is that something similar happens here.

Suppose that in the basic normal-means problem, $\theta_i = \mu + \tau \eta_i$, where $\eta_i \sim \mathrm{DE}(2)$ has a double-exponential distribution. Hence

$$\pi(\boldsymbol{\theta} \mid \mu, \tau) \propto \tau^{-p} \exp\left(-\frac{1}{\tau} \sum_{i=1}^{p} |\theta_i - \mu| \right) ,$$

leading to the joint distribution

$$p(\boldsymbol{\theta}, \mathbf{y} \mid \mu, \nu) \propto \nu^{-p} \exp\left\{ -\frac{1}{2\sigma^2} \left(\sum_{i=1}^{p} (y_i - \theta_i)^2 + \nu^{-1} \sum_{i=1}^{p} |\theta_i - \mu| \right) \right\} ,$$

where ν is the regularization penalty (for known σ).

The plug-in solution is to estimate μ and ν by cross-validation or marginal maximum likelihood. Meanwhile, a reasonable fully Bayesian solution is to use the noninformative prior $\pi(\mu, \tau) \propto 1/\tau$. This yields a marginal prior distribution for $\boldsymbol{\theta}$ of

$$\begin{aligned} \pi(\boldsymbol{\theta}) &= \int \pi(\boldsymbol{\theta}|\mu, \tau)\ \pi(\mu, \tau)\ \mathrm{d}\mu\ \mathrm{d}\tau \\ &\propto \exp\left\{-\frac{1}{2}Q(\boldsymbol{\theta})\right\}\ , \end{aligned}$$

where $Q(\boldsymbol{\theta})$ is piecewise linear and depends upon the order statistics $\theta_{(j)}$ (Uthoff, 1973). Specifically, define $v_j(\boldsymbol{\theta}) \equiv v_j = \sum_{i=1}^{p} |\theta_{(i)} - \theta_{(j)}|$. Then

$$\pi(\boldsymbol{\theta}) = (p-2)!\ 2^{-p+1} \sum_{j=1}^{p} w_j^{-1}\,, \tag{6.16}$$

where

$$w_j = \begin{cases} 4v_j^{p-1}\left(j - \frac{p}{2}\right)\left(\frac{p}{2} + 1 - j\right), & j \neq \frac{p}{2}, \frac{p}{2}+1 \\ 4v_j^{p-1}\left[1 + (p-1)\left(\theta_{(p/2+1)} - \theta_{(p/2)}\right) v_j^{-1}\right], & j = \frac{p}{2}, \frac{p}{2}+1 \end{cases}.$$

Hence the non-Bayesian estimates $\boldsymbol{\theta}$ using

$$\pi_{EB}(\boldsymbol{\theta} \mid \mathbf{y}, \hat{\nu}, \hat{\mu}) \propto \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{p}(y_i - \theta_i)^2 + \hat{\nu}^{-1}\sum_{i=1}^{p}|\theta_i - \hat{\mu}|\right)\right\}\ , \tag{6.17}$$

while the Bayesian estimates $\boldsymbol{\theta}$ using

$$\pi_{FB}(\boldsymbol{\theta} \mid \mathbf{y}) \propto \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{p}(y_i - \theta_i)^2\right) + \frac{(p-2)!}{2^{p-1}}\log\left(\sum_{i=1}^{p}[w_i(\boldsymbol{\theta})]^{-1}\right)\right\}\ . \tag{6.18}$$

The former is the traditional double-exponential prior, while the latter prior exhibits a rather complicated dependence upon the order statistics of the θ_i's (which do not appear in the plug-in expression). It is by no means certain that the two procedures will reach similar answers asymptotically, since this difference in functional form persists for all p.

The double-exponential prior coupled with the noninformative prior on μ and τ is just one example where the marginalization in (6.16) is analytically tractable. But it serves to convey the essence of the problem, which is quite general. The Bayes and plug-in approaches for estimating τ imply fundamentally different regularization penalties for $\boldsymbol{\theta}$, regardless of whether $\boldsymbol{\theta}$ is estimated by the mean or the mode, and regardless of whether marginal maximum likelihood or cross-validation is used.

Neither prior is wrong *per se*, but the stark difference between (6.17) and (6.18) is interesting in its own right, and also calls into question the extent to which the plug-in analysis can approximate the fully Bayesian one. In light of these issues, plug-in analysis should be undertaken with caution.

6.8 Discussion

The horseshoe, of course, is not a panacea for sparse problems. But it does seem to be a good default option. It is both surprising and interesting that its answers coincide so closely with the answers from the two-group gold standard of a Bayesian mixture model. Indeed, these results show an interesting duality between the two procedures. While the discrete mixture arrives at a good shrinkage rule by way of a multiple-testing procedure, the horseshoe estimator goes in the opposite direction, arriving at a good multiple-testing procedure by way of a shrinkage rule. Its combination of strong global shrinkage through τ, along with robust local adaptation to signals through the λ_i's, is unmatched by other common Bayes rules using scale mixtures.

This chapter studies sparsity in the simplified context where $\boldsymbol{\theta}$ is a vector of normal means: $(\mathbf{y}|\boldsymbol{\theta}) \sim \mathrm{N}(\boldsymbol{\theta}, \sigma^2 I)$, where σ^2 may be unknown. It is here that the lessons drawn from a comparison of different approaches for modeling sparsity are most readily understood, but these lessons apply to more difficult problems—regression, covariance regularization, function estimation—where many of the challenges of modern statistics lie. Multivariate scale mixtures of normals are often used as shrinkage priors in regression and basis expansion; see, for example, Park and Casella (2008), Hans (2008), and Dunson (2008). In these situations, the behavioral similarities of the horseshoe and the Bayesian two-group model may be exploited even more profitably.

Chapter 7

Concluding Remarks

This thesis has considered a variety of interesting data sets and methodological problems, with multiplicity adjustment as the common theme. These topics included functional data, linear models, and Gaussian graphical models.

In Chapter 2, I introduced a framework for large-scale simultaneous testing of functional data, which is an issue that arises in areas as diverse as epidemiology, physics, genomics, and business. The goal of such analyses is to decide whether an unknown function is zero or nonzero on the basis of noisy data, and to do so for many thousands of functions at once. I presented a motivating example that involved screening the entire Compustat database in an attempt to flag publicly traded companies that have consistently (and nonrandomly) outperformed their peer group over time.

In the area of variable selection, I have considered the nature of Bayesian multiplicity adjustment as it relates to the Occam's-Razor penalty, and developed a theory of Kullback-Leibler convergence to describe the asymptotic correspondence between Bayes and empirical-Bayes procedures. In particular, Chapter 3 proved a theorem that characterized a surprising discrepancy along these lines between fully Bayes and empirical-Bayes approaches to multiplicity adjustment.

In the area of Gaussian graphical modeling, I have introduced a default version of the hyper-inverse Wishart prior, the HIW g-prior, and shown how it corresponds to the implied fractional prior for covariance selection using fractional Bayes factors. That such a default procedure exists, and provides easy-to-compute answers in closed form, is itself of considerable interest. I have also developed the a novel theory of information consistency for covariance selection, which involved formally defining two new notions of consistency that describe the limiting behavior of Bayes factors as information in favor of a particular graph becomes arbitrarily strong. As Chapter 4 showed, fractional marginal likelihoods, in contrast to those based on conventional priors, show both types of finite-sample consistency. This provides a theoretical guarantee of reasonable performance that has heretofore been absent from the graph-selection literature.

I also connected these developments with the more general problem of multiple-testing for edges in a graph. Extensive numerical experiments show that the combined use of fractional priors for Σ, along with edge-selection priors over graph space, will strongly control the rate of false edges admitted into the graph, and that ignoring the multiplicity issue can produce unacceptably high numbers of false positives. Moreover, these priors can either be left alone, allowing the data itself to characterize the prevailing rate of edge inclusion, or they can be used to encode prior information about different edge inclusion probabilities for different parts of the precision matrix.

This offers substantial flexibility for complicated structural information to be built directly into the model.

Computational tools for graphical modeling were considered in Chapter 5. The upshot of this chapter is that adaptive learning of edge-inclusion probabilities can offer substantial improvement over existing graph-search algorithms.

Finally, Chapter 6 introduced the horseshoe estimator for sparse signals. The horseshoe prior is a member of the family of multivariate scale mixtures of normals, and is therefore closely related to widely used approaches for sparse Bayesian learning, including, among others, double-exponential (LASSO) and Student-t priors (relevance vector machines). The advantages of the horseshoe are its robustness at handling unknown sparsity and large outlying signals. These properties were justified theoretically via a representation theorem and accompanied by comprehensive empirical experiments that compared its performance to benchmark alternatives.

I will now briefly describe some future directions for research relating to these areas.

7.1 Future work on graphical models

One of my current research goals is to extend the stochastic-search algorithm of Chapter 5 to a full adaptive MCMC. The primary challenge here is to account for the set of possible moves arising in the global "jumping" step. As the examples in Chapter 5 showed, this type of move is important for efficiency. Yet the triangulation algorithms used here involve a complicated blend of deterministic and stochastic elements, and this makes characterizing the proposal distribution for the global move—which is necessary to ensure ergodicity of the chain—somewhat difficult.

A second goal of mine is to define novel forms of priors over graph space that involve more than simply the number of edges included in the graph. Recall that throughout this thesis, I have used priors of the form $p(G) \propto w^k(1-w)^{m-k}$, where k is the number of edges, and $m = p(p-1)/2$ is the maximum number of possible edges. The inclusion probability w controlled the underlying rate of edge inclusion, thereby handling the implied multiple-testing problem.

The rationale for moving beyond these simple priors comes from an analogy with linear models, where the interplay between the marginal likelihood and the prior over model space is well understood (see Chapter 3). For example, under g-priors, the Bayes factor for testing $M_{\boldsymbol{\gamma}}$ against the null model is

$$\mathrm{BF}(M_{\boldsymbol{\gamma}} : M_0) = (1+g)^{(n-k_{\boldsymbol{\gamma}}-1)/2}[1+(1-R^2_{\boldsymbol{\gamma}})g]^{-(n-1)/2} \,,$$

where $R^2_{\boldsymbol{\gamma}} \in (0,1]$ is the usual coefficient of determination for model $M_{\boldsymbol{\gamma}}$. The term $(1+g)^{n-k_{\boldsymbol{\gamma}}}$ encodes an explicit dependence upon $k_{\boldsymbol{\gamma}}$, the number of variables in the model. Hence it is quite reasonable to use variable-selection priors that also depend upon $k_{\boldsymbol{\gamma}}$, since the marginal likelihoods are directly informative about the prior variable-inclusion probability w.

The relationship between priors over graph space and graph marginal likelihoods, however, is much less direct. Recall Equation 4.9, which gives the fractional Bayes factor for testing a graph G_A against the null graph G_0:

$$\begin{aligned}\mathrm{BF}(G_0 : G_A) = K \quad &\cdot \quad \frac{\prod_{j=1}^{p} \left|\frac{g}{2}X_j'X_j\right|^{\frac{gn}{2}}}{\prod_{j=1}^{p} \left|\frac{1}{2}X_j'X_j\right|^{\frac{n}{2}}} \cdot \frac{\prod_{S\in\mathcal{S}} \left|\frac{g}{2}X_S'X_S\right|^{\frac{gn+|S|-1}{2}}}{\prod_{C\in\mathcal{C}} \left|\frac{g}{2}X_C'X_C\right|^{\frac{gn+|C|-1}{2}}} \\ &\cdot \quad \frac{\prod_{C\in\mathcal{C}} \left|\frac{1}{2}X_C'X_C\right|^{\frac{n+|C|-1}{2}}}{\prod_{S\in\mathcal{S}} \left|\frac{1}{2}X_S'X_S\right|^{\frac{n+|S|-1}{2}}} \,,\end{aligned}$$

where g is fixed, $\mathcal{C}$ and $\mathcal{S}$ are the cliques and separators of G_A, and the leading term K is

$$K = \left[\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{gn}{2}\right)} \right]^p \cdot \frac{\prod_{S \in \mathcal{S}} \Gamma_{|S|}\left(\frac{n+|S|-1}{2}\right)}{\prod_{C \in \mathcal{C}} \Gamma_{|C|}\left(\frac{n+|C|-1}{2}\right)} \cdot \frac{\prod_{C \in \mathcal{C}} \Gamma_{|C|}\left(\frac{gn+|C|-1}{2}\right)}{\prod_{S \in \mathcal{S}} \Gamma_{|S|}\left(\frac{gn+|S|-1}{2}\right)} \, .$$

This is a complicated expression involving the graph topology, one whose dependence upon the number of edges k is at best indirect. The foundational case for using vanilla edge-selection priors is therefore attenuated, since the graph marginal likelihoods are not directly informative about w in the same way that linear-model marginal likelihoods are for variable-selection priors.

In light of this fact, a second research goal of mine is to define a prior over graph space—and, I hope, a corresponding generative probability model—whose hyperparameters are more directly informed by graph marginal likelihoods. This will involve a potentially novel form of multiplicity correction, where the units being tested are no longer individual edges, but rather more interesting structural features involving cliques and separators of the graph.

7.2 Future work on local shrinkage rules

As Chapter 6 intimated, the horseshoe prior can be applied to a much wider class of problems than the estimation of many normal means. Initial investigations show it to be a useful procedure in all kinds of problems where the goal is to estimate a set of coefficients $\boldsymbol{\beta} = \{\beta_i\}_{i=1}^p$ that determine some functional relationship between a set of inputs $\{x_i\}_{i=1}^p$ and a target variable y. This framework encompasses problems of regression, classification, function estimation, covariance regularization, and many others.

Discrete mixtures offer the correct representation of sparse problems by placing positive prior probability on $\beta_i = 0$, but pose several difficulties. These include foundational issues related to the specification of priors for trans-dimensional model comparison, and computational issues related both to the calculation of marginal likelihoods and to the rapid combinatorial growth of the solution set. Shrinkage priors, on the other hand, can be very attractive computationally, but they create their own set of challenges, since the posterior probability mass on $\{\beta_i = 0\}$ (a set of Lebesgue measure zero) is never positive. Truly sparse solutions can therefore be achieved only through artifice.

I hope to study the horseshoe estimator as a default shrinkage procedure in a wide variety of sparse situations. Indeed, preliminary results have shown a happy, and remarkably consistent, fact about the horseshoe's performance: that it quite closely mimics the answers one would get by performing Bayesian model-averaging under a heavy-tailed discrete-mixture model. Bayesian model averaging is clearly the predictive gold standard for such problems (see, e.g. Raftery et al., 1997), and a large part of the horseshoe prior's appeal stems from its ability to provide "BMA-like" performance without the attendant computational fuss.

As a quick demonstration, consider the following two examples. First, I chose two fixed vectors of ten nonzero coefficients: $\boldsymbol{\beta}_{1:10} = (2, 2, 2, 2, 2, 2, 2, 2, 2, 5, 20)$ and $\boldsymbol{\beta}_{1:10} = (1, 2, 3, 4, 5, 6, 7, 8, 9, 10)$. I then "padded" these with $(p-10)$ zeros for several different choices of p, simulated random design matrices with moderately correlated entries, and simulated $\mathbf{y}$ by adding standard normal errors to the true linear predictor $\mathbf{X}\boldsymbol{\beta}$. In all cases, n scaled linearly with p.

For this example, the horseshoe was benchmarked against the LASSO (i.e. the

Table 7.1: Mean-squared error in estimating $\boldsymbol{\beta}$ in first example.

Case 1: $\boldsymbol{\beta}_{1:10} = (2, 2, 2, 2, 2, 2, 2, 2, 5, 20)$

p	20	50	100	200	400
n	24	60	120	240	480
Lasso	1.86	0.78	0.34	0.13	0.12
HS	1.28	0.33	0.11	0.06	0.07

Case 2: $\boldsymbol{\beta}_{1:10} = (1, 2, 3, 4, 5, 6, 7, 8, 9, 10)$

p	20	50	100	200	400
n	25	55	105	205	405
Lasso	0.61	0.40	0.48	0.21	0.23
HS	0.31	0.23	0.12	0.09	0.08

posterior mode under double-exponential priors), with the regularization scale τ chosen through cross-validation. Results are presented in Table 7.1.

Second, I fixed $p = 50$, but rather than fixing the non-zero values of $\boldsymbol{\beta}$, I simulated 1000 data sets with varying levels of sparsity, where non-zero β_i's were generated from a standard Student-t with 2 degrees of freedom. (The coefficients were 80% sparse on average, with nonzero status decided by a weighted coin flip.) The horseshoe was compared both with the LASSO and with Bayesian model-averaging using Zellner-Siow priors. Results for both estimation error and out-of-sample prediction error are displayed in Figure 7.1.

As these results show, both BMA and the horseshoe prior systematically outperform the LASSO in sparse regression problems, without either one enjoying a noticeable advantage over the other. This is an interesting (and as yet under-explored) phenomenon that may prove very useful in ultra-high-dimensional situations, where the computational challenges associated with Bayesian model averaging may be very cumbersome indeed.

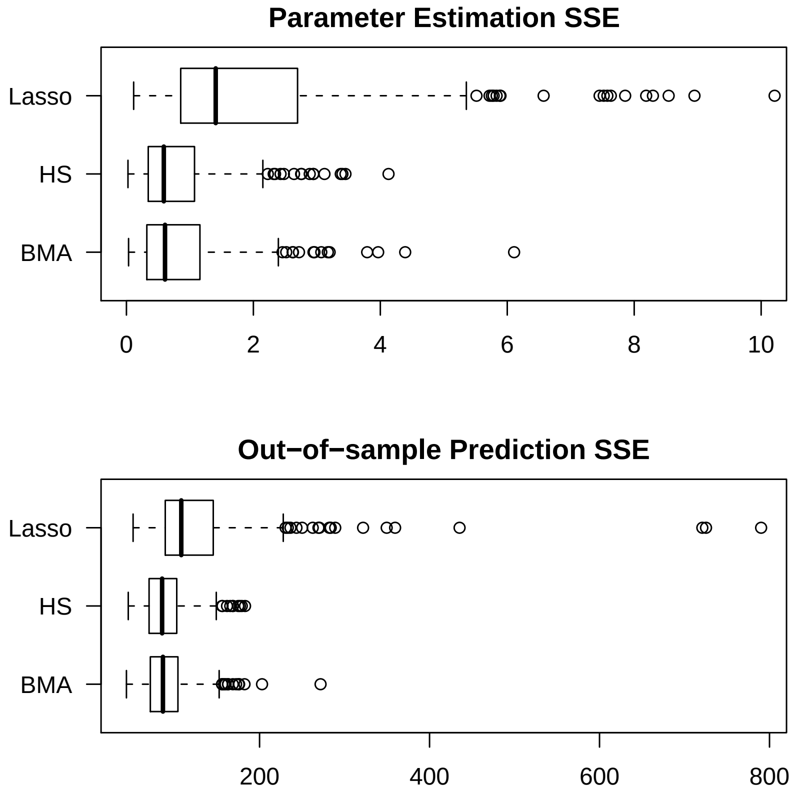


Figure 7.1: Results for second example. "BMA" refers to the model-averaged results under Zellner-Siow priors. "Lasso" refers to the posterior mode under Laplacian priors.

The horseshoe prior also gives rise to a very natural generalization. Suppose that $\theta \sim \mathrm{N}(0, \frac{1-\kappa}{\kappa})$. Then there exists a four-parameter family of hypergeometric–beta priors for κ that has the following form:

$$\pi(\kappa) = C^{-1} \cdot \kappa^{\alpha-1} \, (1-\kappa)^{\beta-1} \, \left\{ \frac{1}{\tau^2} + \left(1 - \frac{1}{\tau^2} \right) \kappa \right\}^{-1} \exp(-s\kappa) \, , \tag{7.1}$$

where $\alpha, \beta, \tau > 0$ and $s \in \mathbb{R}$.

The normalizing constant,

$$C = \int_0^1 \kappa^{\alpha-1} \, (1-\kappa)^{\beta-1} \, \left\{ \frac{1}{\tau^2} + \left(1 - \frac{1}{\tau^2} \right) \kappa \right\}^{-1} \exp(-s\kappa) \, \mathrm{d}\kappa \, , \tag{7.2}$$

can be computed using hypergeometric series, following the results of Chapter 6:

$$C = e^{-s} \ \mathrm{Be}(\alpha, \beta) \ \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} \frac{(\beta)_{m+n}}{(\alpha+\beta)_{m+n} \ m! \ n!} \ s^m \ (1 - 1/\tau^2)^m \,, \tag{7.3}$$

where $\mathrm{Be}(\cdot, \cdot)$ is the beta function and $(a)_n$ is the rising factorial. Equivalently,

$$C = e^{-s} \ \mathrm{Be}(\alpha, \beta) \ \Phi_1(\beta, 1, \alpha+\beta, s, 1 - 1/\tau^2) \,, \tag{7.4}$$

where Φ_1 is the degenerate hypergeometric function of two variables (Gradshteyn and Ryzhik, 1965, 9.261).

The reason for working with a transformed variable is that κ has an interpretation as a shrinkage coefficient in a two-stage normal model. Suppose that $y \sim \mathrm{N}(\theta, 1)$, and that $\theta \sim \mathrm{N}(0, \lambda^2)$. Then $\kappa = 1/(1+\lambda^2)$ is the amount of weight that the Bayes estimator places on 0, the prior mean, once the data y have been observed:

$$\mathrm{E}(\theta \mid y, \lambda^2) = \left(\frac{\lambda^2}{1+\lambda^2} \right) y + \left(\frac{1}{1+\lambda^2} \right) 0 = (1-\kappa) y \,.$$

Since $\kappa_i \in [0,1]$, this is clearly finite, and so by Fubini's theorem,

$$\mathrm{E}(\theta \mid y) = \int_0^1 (1-\kappa) y \ \pi(\kappa \mid y) \ \mathrm{d}\kappa = \{1 - \mathrm{E}\left(\kappa \mid y\right)\} \ y \,. \tag{7.5}$$

Hence the estimator for θ is determined by the posterior expectation of κ. But if κ has a hypergeometric–beta prior, then it will also have a hypergeometric–beta posterior, since

$$p(y, \kappa) \propto \kappa^{\alpha' - 1} \ (1-\kappa)^{\beta-1} \ \left\{ \frac{1}{\tau^2} + \left(1 - \frac{1}{\tau^2} \right) \kappa \right\}^{-1} \ e^{-\kappa s'}$$

where $s' = s + y^2/2$ and $\alpha' = \alpha + 1/2$. This expression is in the same family as (7.1). The moment-generating function given by a special case of the expressions in Gordy

(1998):

$$M(t) = e^t \, \frac{\Phi_1(\beta, 1, \alpha' + \beta, s' - t, 1 - 1/\tau^2)}{\Phi_1(\beta, 1, \alpha' + \beta, s', 1 - 1/\tau^2)} \, .$$

Therefore,

$$\mathrm{E}(\kappa^n \mid y) = \frac{(\alpha')_n}{(\alpha' + \beta)_n} \frac{\Phi_1(\beta, 1, \alpha' + \beta + n, s', 1 - 1/\tau^2)}{\Phi_1(\beta, 1, \alpha' + \beta, s', 1 - 1/\tau^2)} \, . \tag{7.6}$$

Combining (7.5) with (7.6) yields

$$\mathrm{E}(\theta \mid y) = \left\{ 1 - \frac{\alpha'}{\alpha' + \beta} \frac{\Phi_1(\beta, 1, \alpha' + \beta + 1, s', 1 - 1/\tau^2)}{\Phi_1(\beta, 1, \alpha' + \beta, s', 1 - 1/\tau^2)} \right\} y \, , \tag{7.7}$$

will all other posterior moments for θ following in turn.

One important special case of the family is the Strawderman prior (Strawderman, 1971), which corresponds to $\alpha = 1/2$, $\beta = 1$, $s = 0$, and $\tau = 1$. Another special case is the half-Cauchy prior on the scale factor λ, studied by Gelman (2006) and in Chapter 6. This corresponds to $\alpha = \beta = 1/2$, $s = 0$, and $\tau = 1$. Yet a third special case is the uniform-shrinkage prior, where $\alpha = \beta = 1$, $s = 0$, and $\tau = 1$. All of these can be seen in the upper-left pane of Figure 7.2.

As these special cases show, (7.1) can lead to many standard-looking shapes that are similar to other normal scale mixtures. Yet it can also produce a wide variety of other implied densities for λ^2 that are inaccessible through other families. Examples can be seen in the other three panes of Figure 7.2.

The parameters (α, β) operate much in the same way as the parameters of a beta distribution for κ, to which (7.1) reduces when $\tau = 1$ and $s = 0$. Smaller values of α encourage heavier tails in $\pi(\theta)$, with $\alpha = 1/2$, for example, yielding Cauchy-like tails. Smaller values of β encourage $\pi(\theta)$ to have more mass near the origin, and eventually to become unbounded; $\beta = 1/2$ yields, for example, $\pi(\theta) \approx \log(1 + 1/\theta^2)$ near 0.

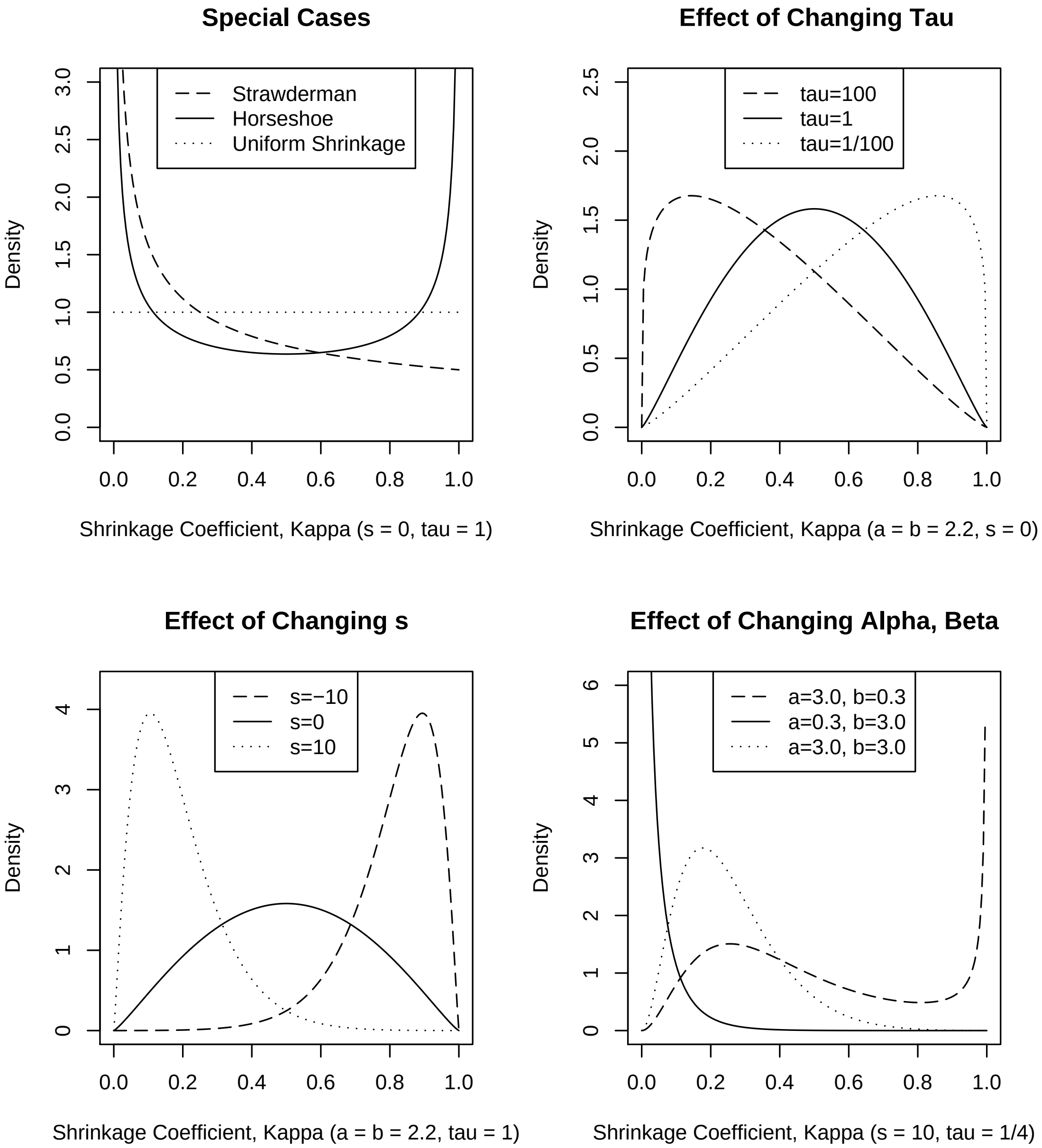


Figure 7.2: Effect of changing the four parameters (α, β, s, τ) on the density for the shrinkage coefficient κ.

The parameters s and τ are global scaling factors, though with different effects on the shape of the density. The exponential term involving s in (7.1) varies from e^{-s} to 1 as κ ranges from 1 to 0 (or equivalently, as λ^2 ranges from 0 to ∞). Similarly, large values of τ encourage small values of κ, and vice versa. It is useful to think of τ as a global scale parameter, and of s as a convenient vehicle for regressing shrinkage coefficients upon external covariates.

To appreciate the wide applicability of these hypergeometric–beta priors, consider the following two examples of their possible use:

- These priors yield closed-form expressions for marginal likelihoods under normal convolution, a fact that is very important for rapidly computing posterior model probabilities in high-dimensional Bayesian model selection and hypothesis testing.
- These priors generate a wide variety of Bayes estimators that are provably minimax under quadratic loss (making use of results in Fourdrinier et al., 2008). Moreover, their risk properties can be assessed merely by simulating chi-squared random variables (that is, without a second level of approximation) due to the existence of closed-form moments.

These are but two examples, and one of my future research goals is to study this class of priors more fully. I anticipate that they can be applied in a wide variety of problems where scale mixtures of normals are useful. This includes classical shrinkage estimation (Strawderman, 1971; Stein, 1981; Fourdrinier et al., 1998), the analysis of variance (Box and Tiao, 1964; Tiao and Tan, 1965), and Bayesian hierarchical modeling (Berger and Bernardo, 1992; Daniels, 1999; Gelman, 2006). It also includes robust Bayesian estimation (Pericchi and Smith, 1992) and various approaches for sparsity described in Chapter 6.

Appendix A

Variations on Zellner's g-prior

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

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

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

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

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

$$\mathrm{BF}(M_1 : M_2) = \frac{\mathrm{BF}(M_1 : M_B)}{\mathrm{BF}(M_2 : M_B)} \, . \tag{A.1}$$

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. While this choice of M_B is free in principle, there are only two such choices which yield a pair of nested models in all comparisons: the null model and the full model.

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

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

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

$$\mathrm{BF}(M_{\boldsymbol{\gamma}} : M_0) = (1+g)^{(n-k_{\boldsymbol{\gamma}}-1)/2}[1 + (1 - R^2_{\boldsymbol{\gamma}})g]^{-(n-1)/2} \, , \tag{A.2}$$

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

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 made by writing the full model as:

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

with the design matrix partitioned in the obvious way. Then a g-prior is specified for the parameters in the full model not shared by the smaller model, which again has k regression parameters excluding the intercept:

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

This does not lead to a coherent "within-model" prior specification for the parameters of the full model, since their prior distribution depends upon which submodel is considered. Nevertheless, marginal likelihoods can still be consistently defined in the manner of Equation A.1. Conditional upon g, this yields a Bayes factor in favor of the full model of

$$\mathrm{BF}(M_F : M_{\boldsymbol{\gamma}}) = (1+g)^{(n-p-1)/2} \; (1+gW)^{-(n-k-1)/2} \tag{A.3}$$

where $W = (1 - R^2_F)/(1 - R^2_{\boldsymbol{\gamma}})$.

The existence of these simple expressions has made the use of g-priors very popular. Yet g-priors yield display a strange type of behavior often called the "information paradox." This can be seen in (A.2): the Bayes factor in favor of $M_{\boldsymbol{\gamma}}$ goes to the finite constant $(1+g)^{n-p-1}$ as $R^2_{\boldsymbol{\gamma}} \to 1$ (which can only happen if $M_{\boldsymbol{\gamma}}$ 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 information about model choice that can be gleaned from 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 $\int_0^\infty (1+g)^{n-k_{\boldsymbol{\gamma}}-1}\pi(g)\ \mathrm{d}g = \infty$ for all $k_{\boldsymbol{\gamma}} \leq p$, which is a generalization of the condition given in Jeffreys (1961) (see Chapter 5.2, Equations 10 and 14).

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

$$(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \phi) \sim C\left(0, \frac{n}{\phi}(\mathbf{X}'_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}})^{-1}\right) \tag{A.4}$$

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, $g \sim \mathrm{IG}(1/2, 2/n)$. The Zellner-Siow null-based Bayes factor under model $M_{\boldsymbol{\gamma}}$ then takes the form:

$$\mathrm{BF}(M_{\boldsymbol{\gamma}} : M_0) = \int_0^\infty (1+g)^{(n-k_{\boldsymbol{\gamma}}-1)/2}[1+(1-R^2_{\boldsymbol{\gamma}})g]^{-(n-1)/2} g^{-3/2} \exp(-n/(2g)\ \mathrm{d}g \tag{A.5}$$

A similar formula exists for the full-based version:

$$\mathrm{BF}(M_F : M_{\boldsymbol{\gamma}}) = \int_0^\infty (1+g)^{(n-p-1)/2}[1+Wg]^{-(n-k-1)/2} g^{-3/2} \exp(-n/(2g)\ \mathrm{d}g \tag{A.6}$$

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. Fortunately, 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 the widespread practical use of Zellner-Siow priors.

Bibliography

Abramovich, F., Benjamini, Y., Donoho, D., and Johnstone, I. (2006), “Adapating to Unknown Sparsity by Controlling the False-Discovery Rate,” *The Annals of Statistics*, 34, 584–653.

Angers, J. and Berger, J. (1991), “Robust hierarchical Bayes estimation of exchangeable means,” *The Canadian Journal of Statistics*, 19, 39–56.

Antoniak, C. (1974), “Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems,” *Annals of Statistics*, 2, 1152–74.

Atay-Kayis, A. and Massam, H. (2005), “A Monte Carlo method for computing the marginal likelihood in nondecomposable Gaussian graphical models,” *Biometrika*, 92.

Barbieri, M. and Berger, J. O. (2004), “Optimal predictive model selection,” *The Annals of Statistics*, 32, 870–897.

Bartlett, M. (1957), “A comment on D.V. Lindley’s statistical paradox,” *Biometrika*, 44, 533–4.

Basu, S. and Chib, S. (2003), “Marginal likelihood and Bayes factors for Dirichlet process mixture models,” *Journal of the American Statistical Association*, 98, 224–35.

Benjamini, Y. and Hochberg, Y. (1995), “Controlling the false-discovery rate: a practical and powerful approach to multiple testing,” *Journal of the Royal Statistical Society, Series B*, 57, 289–300.

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

Berger, J. O. (1980), “A robust generalized Bayes estimator and confidence region for a multivariate normal mean,” *The Annals of Statistics*, 8, 716–761.

Berger, J. O. (1985), *Statistical Decision Theory and Bayesian Analysis*, Springer-Verlag, 2nd edn.

Berger, J. O. and Bernardo, J. M. (1992), “On the development of reference priors (with discussion),” in *Bayesian Statistics 4*, eds. J. Bernardo, J. Berger, A. Dawid, and A. Smith, pp. 35–60, Oxford University Press.

Berger, J. O. and Deely, J. (1988), “A Bayesian Approach to Ranking and Selection of Related Means With Alternatives to Analysis-of-Variance Methodology,” *Journal of the American Statistical Association*, 83, 364–73.

Berger, J. O. and Guglielmi, A. (2001), "Bayesian and conditional frequentist testing of parametric model versus nonparametric alternatives," *Journal of the American Statistical Association*, 96, 174–84.

Berger, J. O. and Molina, G. (2005), "Posterior model probabilities via path-based pairwise priors," *Statistica Neerlandica*, 59, 3–15.

Berger, J. O. and Pericchi, L. (1996), "The intrinsic Bayes factor for model selection and prediction," *Journal of the American Statistical Association*, 91, 109–122.

Berger, J. O. and Pericchi, L. (2001), "Objective Bayesian methods for model selection: introduction and comparison," in *Model Selection*, vol. 38 of *Institute of Mathematical Statistics Lecture Notes – Monograph Series*, pp. 135–207, Beachwood.

Berry, A., Blair, J., Heggernes, P., and Peyton, B. (2004), "Maximum Cardinality Search for Computing Minimal Triangulations of Graphs," *Algorithmica*, 39, 287–298.

Berry, A., Heggernes, P., and Villander, Y. (2006), "A vertex incremental approach for maintaining chordality," *Discrete Mathematics*, 306, 318–336.

Berry, D. (1988), "Multiple Comparisons, Multiple Tests, and Data Dredging: A Bayesian Perspective," in *Bayesian Statistics 3*, eds. J. Bernardo, M. DeGroot, D. Lindley, and A. Smith, pp. 79–94, Oxford University Press.

Berry, D. and Hochberg, Y. (1999), "Bayesian perspectives on multiple comparisons," *Journal of Statistical Planning and Inference*, 82, 215–277.

Bigelow, J. and Dunson, D. (2005), "Semiparametric Classification in Hierarchical Functional Data Analysis," Tech. rep., Duke University Department of Statistical Science.

Bogdan, M., Ghosh, J. K., and Tokdar, S. T. (2008a), "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*, vol. 1, pp. 211–30, Institute of Mathematical Statistics.

Bogdan, M., Chakrabarti, A., and Ghosh, J. K. (2008b), "Optimal rules for multiple testing and sparse multiple regression," Tech. rep., Purdue University.

Bowman, E. H. and Helfat, C. E. (2001), "Does corporate strategy matter?" *Strategic Management Journal*, 22, 1–23.

Box, G. and Tiao, G. C. (1964), "A Bayesian approach to the importance of assumptions applied to the comparison of variances," *Biometrika*, 51, 153–67.

Brown, L. (1971), "Admissible estimators, recurrent diffusions and insoluble boundary problems," *The Annals of Mathematical Statistics*, 42, 855–903.

Carlin, B. and Louis, T. (2000), "Empirical Bayes: Past, Present and Future," *Journal of the American Statistical Association*, 95, 1286–89.

Carvalho, C. and West, M. (2007), "Dynamic Matrix-Variate Graphical Models," *Bayesian Analysis*, 2, 69–96.

Carvalho, C., Massam, H., and West, M. (2007), "Simulation of hyper-inverse Wishart distributions in graphical models," *Biometrika*, 94, 647–59.

Carvalho, C. M. and Scott, J. G. (2009), "Objective Bayesian model selection in Gaussian graphical models," *Biometrika*, to appear.

Casella, G. and Moreno, E. (2002), "Objective Bayes variable selection," Tech. Rep. 023, University of Florida.

Crowley, E. (1997), "Product Partition Models for Normal Means," *Journal of the American Statistical Association*, 92, 192–8.

Cui, W. and George, E. I. (2008), "Empirical Bayes vs. fully Bayes variable selection," *Journal of Statistical Planning and Inference*, 138, 888–900.

Dahl, D. B. and Newton, M. A. (2007), "Multiple Hypothesis Testing by Clustering Treatment Effects," *Journal of the American Statistical Association*, 102, 517–26.

Daniels, M. J. (1999), "A prior for the variance in hierarchical models," *The Canadian Journal of Statistics*, 27, 567–78.

Dawid, A. P. and Lauritzen, S. L. (1993), "Hyper-Markov laws in the statistical analysis of decomposable graphical models," *The Annals of Statistics*, 3, 1272–1317.

Denison, D. and George, E. (2000), "Bayesian prediction using adaptive ridge estimators," Tech. rep., Imperial College, London.

Denrell, J. (2003), "Vicarious Learning, Undersampling of Failure, and the Myths of Management." *Organizational Science*.

Denrell, J. (2005), "Selection Bias and the Perils of Benchmarking," *Harvard Business Review*.

Deshpande, A., Garofalakis, M. N., and Jordan, M. I. (2001), "Efficient stepwise selection in decomposable models," in *Uncertainty in Artificial Intelligence (UAI), Proceedings of the Seventeenth Conference*, eds. J. Breese and D. Koller.

Do, K.-A., Muller, P., and Tang, F. (2005), "A Bayesian mixture model for differential gene expression," *Journal of the Royal Statistical Society, Series C*, 54, 627–44.

Dobra, A., Jones, B., Hans, C., Nevins, J., and West, M. (2004), "Sparse graphical models for exploring gene expression data," *Journal of Multivariate Analysis*, 90, 196–212.

DuMouchel, W. (1988), "A Bayesian model and prior elicitation procedure for multiple comparisons," in *Bayesian Statistics 3*, eds. J. Bernardo, M. DeGroot, D. Lindley, and A. Smith, pp. 79–94, Oxford University Press.

Duncan, D. B. (1961), "Bayes Rules for a Common Multiple Comparisons Problem and Related Student-t Problem," *The Annals of Mathematical Statistics*, 32, 1013–33.

Duncan, D. B. (1965), "A Bayesian Approach to Multiple Comparisons," *Technometrics*, 7, 171–222.

Dunson, D. and Herring, A. (2006), "Semiparametric Bayesian latent trajectory models," Tech. rep., Duke University Department of Statistical Science.

Dunson, D. B. (2008), "Kernel local partition processes for functional data," Tech. rep., Duke University Department of Statistical Science.

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

Efron, B. (2008), "Microarrays, Empirical Bayes and the two-groups model (with discussion)," *Statistical Science*, 1, 1–22.

Efron, B., R., T., Storey, J., and Tusher, V. (2001), "Empirical Bayes analysis of a microarray experiment," *Journal of American Statistical Association*, 96, 1151–1160.

Escobar, M. and West, M. (1995), "Bayesian density estimation and inference using mixtures," *Journal of the American Statistical Association*, 90, 577–88.

Fan, T. and Berger, J. O. (1992), "Behaviour of the posterior distribution and inferences for a normal mean with t prior distributions," *Stat. Decisions*, 10, 99–120.

Ferguson, T. (1973), "A Bayesian analysis of some nonparametric problems," *The Annals of Statistics*, 1, 209–30.

Fernandez, C., Ley, E., and Steel, M. (2001), "Model Uncertainty in Cross-Country Growth Regressions," *Journal of Applied Econometrics*, 16, 563–76.

Fourdrinier, D., Strawderman, W., and Wells, M. T. (1998), "On the construction of Bayes minimax estimators," *The Annals of Statistics*, 26, 660–71.

Fourdrinier, D., Kortbi, O., and Strawderman, W. (2008), "Bayes minimax estimators of the mean of a scale mixture of multivariate normal distributions," *Journal of Multivariate Analysis*, 99, 74–93.

Frühwirth-Schnatter, S. and Kaufmann, S. (2008), "Model-Based Clustering of Multiple Time Series," *Journal of Business and Economic Statistics*, 26, 78–89.

Frydenberg, M. and Lauritzen, S. L. (1989), "Decomposition of maximum likelihood in mixed models," *Biometrika*, 76, 539–555.

Geisser, S. and Cornfield, J. (1963), "Posterior distributions for multivariate normal parameters," *Journal of the Royal Statistical Society, Series B*, 25, 368–376.

Gelfand, A., Kottas, A., and MacEachern, S. (2005), "Bayesian nonparametric spatial modeling with Dirichlet process mixing," *Journal of the American Statistical Association*, 100, 1021–35.

Gelman, A. (2006), "Prior distributions for variance parameters in hierarchical models," *Bayesian Anal.*, 1, 515–33.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2004), *Bayesian Data Analysis*, Chapman and Hall/CRC, 2nd edn.

George, E. I. and Foster, D. P. (2000), "Calibration and empirical Bayes variable selection," *Biometrika*, 87, 731–747.

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

Giudici, P. (1996), "Learning in graphical Gaussian models," in *Bayesian Statistics 5*, eds. J. Berger, J. Bernardo, A. Dawid, and A. Smith, pp. 621–8, Oxford University Press.

Giudici, P. and Green, P. J. (1999), "Decomposable graphical Gaussian model determination," *Biometrika*, 86, 785–801.

Gopalan, R. and Berry, D. (1998), "Bayesian multiple comparisons using Dirichlet process priors," *Journal of the American Statistical Association*, 93, 1130–1139.

Gordy, M. B. (1998), "A generalization of generalized beta distributions," Tech. rep., Board of Governors of the Federal Reserve System, Finance and Economics Discussion Series.

Gould, H. (1964), "Sums of logarithms of binomial coefficients," *The American Mathematical Monthly*, 71, 55–58.

Gradshteyn, I. and Ryzhik, I. (1965), *Table of Integrals, Series, and Products*, Academic Press.

Gramacy, R. (2005), "Bayesian treed Gaussian process models," Ph.D. thesis, University of California–Santa Cruz.

Green, P. J. (1995), "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika*, 82, 711–732.

Griffin, J. and Brown, P. (2005), "Alternative prior distributions for variable selection with very many more variables than observations," Tech. rep., University of Warwick.

Hans, C., Dobra, A., and West, M. (2007), "Shotgun stochastic search in regression with many predictors," *Journal of the American Statistical Association*, 102, 507–516.

Hans, C. M. (2008), "Bayesian Lasso Regression," Tech. rep., Ohio State University.

Harrigan, K. (1985), "An Application of Clustering for Strategic Group Analysis," *Strategic Management Journal*, 6, 55–73.

Hartigan, J. (1990), "Partition models," *Communications in Statistics: Theory and Methods*.

Hawawini, G., Subramanian, V., and Verdin, P. (2003), "Is performance driven by industry- or firm-specific factors? A new look at the evidence," *Strategic Management Journal*, 24, 1–16.

Hochberg, Y. and Tamhane, C. (1987), *Multiple Comparison Procedures*, Wiley.

Ishwaran, H. and James, L. (2001), "Gibbs sampling methods for stick-breaking priors," *Journal of the American Statistical Association*, 96, 161–173.

Jefferys, W. and Berger, J. (1992), "Ockham's Razor and Bayesian Analysis," *American Scientist*, 80, 64–72.

Jeffreys, H. (1961), *Theory of Probability*, Oxford University Press, 3rd edn.

Johnstone, I. M. and Silverman, B. W. (2004), "Needles and Straw in Haystacks: Empirical-Bayes estimates of possibly sparse sequences," *The Annals of Statistics*, 32, 1594–1649.

Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C., and West, M. (2005), "Experiments in Stochastic Computation for High-dimensional Graphical Models," *Statistical Science*, 20, 388–400.

Kleinman, K. and Ibrahim, J. (1998), "A semiparametric Bayesian approach to the random effects model," *Biometrics*, 54, 921–38.

Laud, P. and Ibrahim, J. (1995), "Predictive Model Selection," *Journal of the Royal Statistical Society, Series B*, 57, 247–62.

Lauritzen, S. L. (1996), *Graphical Models*, Clarendon Press, Oxford.

Letac, G. and Massam, H. (2007), "Wishart distributions on decomposable graphs," *Ann. Statist.*, 35, 1278–1323.

Ley, E. and Steel, M. F. (2007), "On the Effect of Prior Assumptions in Bayesian Model Averaging with Applications to Growth Regression," No. 4238 in Policy Research Working Paper Series, World Bank.

Liang, F., Paulo, R., Molina, G., Clyde, M., and Berger, J. (2008), "Mixtures of g-priors for Bayesian variable selection," *Journal of the American Statistical Association*, 103, 410–23.

Masreliez, C. (1975), "Approximate non-Gaussian filtering with linear state and observation relations," *IEEE. Trans. Autom. Control.*

Massam, H. and Neher, E. (1998), "Estimation and Testing for Lattice Conditional Independence Models on Euclidean Jordan Algebras," *The Annals of Statistics*, 26, 1051–1082.

Meinshausen, N. and Buhlmann, P. (2006), "High dimensional graphs and variable selection with the Lasso," *Annals of Statistics*, 34, 1436–1462.

Meng, C. and Dempster, A. (1987), "A Bayesian approach to the multiplicity problem for significance testing with binomial data," *Biometrics*, 43, 301–11.

Mitchell, A. F. (1994), "A Note on Posterior Moments for a Normal Mean with Double-Exponential Prior," *Journal of the Royal Statistical Society, Series B*, 56, 605–10.

Müller, P., West, M., and MacEachern, S. (1997), "Bayesian models for non-linear auto-regressions," *Journal of Time Series Analysis*, 18, 593–614.

Muller, P., Parmigiani, G., and Rice, K. (2006), "FDR and Bayesian Multiple Comparisons Rules," in *Proceedings of the 8th Valencia World Meeting on Bayesian Statistics*, Oxford University Press.

O'Hagan, A. (1995), "Fractional Bayes factors for model comparison," *Journal of the Royal Statistical Society, Series B*, 57, 99–138.

Park, T. and Casella, G. (2008), "The Bayesian Lasso," *Journal of the American Statistical Association*, 103, 681–6.

Paulsen, V., Power, S., and Smith, R. (1989), "Schur products and matrix completions," *J. Funct. Anal.*, 85, 151–78.

Pericchi, L. R. and Smith, A. (1992), "Exact and Approximate Posterior Moments for a Normal Location Parameter," *Journal of the Royal Statistical Society (Series B)*, 54, 793–804.

Polson, N. G. (1991), "A representation of the posterior mean for a location model," *Biometrika*, 78, 426–30.

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997), "Bayesian Model Averaging for Linear Regression Models," *Journal of the American Statistical Association*, 92, 1197–1208.

Rasmussen, C. E. and Williams, C. (2006), *Gaussian Processes for Machine Learning*, MIT Press.

Roverato, A. (2000), "Cholesky decomposition of a hyper-inverse Wishart matrix," *Biometrika*, 87, 99–112.

Ruefli, T. W. and Wiggins, R. R. (2000), "Longitudinal Performance Stratification: An Iterative Kolmogorov-Smirnov Approach," *Management Science*, 46, 685–92.

Ruefli, T. W. and Wiggins, R. R. (2002), "Sustained Competitive Advantage: Temporal Dynamics and the Incidence and Persistence of Superior Economic Performance," *Organization Science*, 13, 81–105.

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. and Berger, J. O. (2006), "An exploration of aspects of Bayesian multiple testing," *Journal of Statistical Planning and Inference*, 136, 2144–2162.

Scott, J. G. and Carvalho, C. M. (2008), "Feature-inclusion stochastic search for Gaussian graphical models," *Journal of Computational and Graphical Statistics*, 17.

Slater, L. J. (1960), *Confluent Hypergeometric Functions*, Cambridge University Press.

Stein, C. (1981), "Estimation of the mean of a multivariate normal distribution," *The Annals of Statistics*, 9, 1135–51.

Storey, J. (2003), "The positive false-discovery rate: a Bayesian interpretation and the q-value," *The Annals of Statistics*, 31, 2013–35.

Strawderman, W. (1971), "Proper Bayes minimax estimators of the multivariate normal mean," *The Annals of Statistics*, 42, 385–8.

Sun, D. and Berger, J. O. (2006), "Objective priors for the multivariate normal model," in *Proceedings of the 8th Valencia World Meeting on Bayesian Statistics*, ISBA.

Tiao, G. C. and Tan, W. (1965), "Bayesian analysis of random-effect models in the analysis of variance. I. Posterior distribution of variance components," *Biometrika*, 51, 37–53.

Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso," *J. Royal. Statist. Soc B.*, 58, 267–88.

Uthoff, V. (1973), "The Most Powerful Scale and Location Invariant Test of the Normal Versus the Double Exponential," *The Annals of Statistics*, 1, 170–4.

Waller, R. and Duncan, D. (1969), "A Bayes rule for the symmetric multiple comparison problem," *Journal of the American Statistical Association*, 64, 1484–1503.

Wermuth, N. (1980), "Linear recursive equations, covariance selection adn path analysis," *J. Am. Statist. Assoc.*, 75, 963–72.

Wernerfelt, B. (1984), "The Resource-Based View of the Firm," *Strategic Management Journal*, 5, 171–180.

West, M. (1987), "On scale mixtures of normal distributions," *Biometrika*, 74, 646–8.

Westfall, P. H., Johnson, W. O., and Utts, J. M. (1997), "A Bayesian perspective on the Bonferroni adjustment," *Biometrika*, 84, 419–27.

Wong, F., Carter, C., and Kohn, R. (2003), "Efficient estimation of covariance selection models," *Biometrika*, 90, 809–830.

Woodard, D. (2007), "Conditions for Rapid and Torpid Mixing of Parallel and Simulated Tempering on Multimodal Distributions," Ph.D. thesis, Duke University.

Yuan, M. and Lin, Y. (2007), "Model selection and estimation in the Gaussian graphical model," *Biometrika*, 94, 19–35.

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

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

Biography

James Scott was born on May 15, 1982, and was raised in Katy, Texas. His parents are Anne and George Scott, who still live in Katy. James attended the Plan II Honors program at the University of Texas at Austin, receiving his bachelor's degree in mathematics in May of 2004. He was awarded a Marshall Scholarship for study in Great Britain, and spent two years at Trinity College, Cambridge, reading mathematics and statistics. Upon returning to the United States in 2006, he began studying statistics at Duke University, completing his Ph.D in April of 2009.