The Intrinsic Bayes Factor for Model Selection and Prediction

Article *in* Journal of the American Statistical Association · March 1996

DOI: 10.1080/01621459.1996.10476668

2 authors, including:

Luis Pericchi
University of Puerto Rico at Rio Piedras
65 PUBLICATIONS **2,323** CITATIONS

THE INTRINSIC BAYES FACTOR FOR MODEL SELECTION AND PREDICTION

by

James O. Berger and Luis R. Pericchi
Purdue University Universidad Simón Bolívar

Technical Report #93-43C

Department of Statistics
Purdue University

revised, March 1994

◇

THE INTRINSIC BAYES FACTOR FOR MODEL SELECTION AND PREDICTION*

by

James O. Berger and Luis R. Pericchi
Purdue University Universidad Simón Bolívar

Abstract

In the Bayesian approach to model selection or hypothesis testing with models or hypotheses of differing dimensions, it is typically not possible to utilize standard noninformative (or default) prior distributions. This has led Bayesians to use conventional proper prior distributions or crude approximations to Bayes factors. In this paper we introduce a new criterion called the *intrinsic Bayes factor*, which is fully automatic in the sense of requiring only standard noninformative priors for its computation, and yet seems to correspond to very reasonable actual Bayes factors. The criterion can be used for nested or nonnested models, and for multiple model comparison and prediction. From another perspective, the development suggests a general definition of a "reference prior" for model comparison.

1. INTRODUCTION

1.1 Is Another Model Selection Criterion Needed?

We obviously think so, but why? First, we feel that model selection should have a Bayesian basis. This is not so much based on generic Bayesian arguments, as on a belief that Bayesian methods of model selection and hypothesis testing are particularly needed for the following reasons:

(i) Measures based on frequentist computations, such as P-values (in, say, chi-squared testing of fit), are at best extremely difficult to interpret and at worst highly misleading (cf, Edwards, Lindman and Savage, 1963; Berger and Sellke, 1987; Berger and Delampady, 1987; and Delampady and Berger, 1990).

* This work was supported by the National Science Foundation, Grants DMS–8923071 and DMS–9303556, and by BID-CONICIT, Venezuela.

(ii) Analysis of non-nested and/or multiple models or hypotheses is very difficult in a frequentist framework.

(iii) Non-Bayesian methods have difficulty incorporating "Ockham's Razor," the notion that if two models explain data equally well, the simpler is to be preferred; Bayes factors do this automatically (cf, Spiegelhalter and Smith, 1982, and Jeffreys and Berger, 1992, and the references therein), while other methods require introduction of adhoc penalties for model complexity.

(iv) Prediction is often the real goal and, in accounting for model uncertainty in prediction, it is virtually necessary to use Bayesian methods, which can keep all models under consideration, weighted by their posterior probabilities (cf, Draper, 1994, for a review and earlier references, and also Section 5).

Discussions of these issues here would take us too far afield, but it is important to stress that we feel it to be *necessary* to do hypothesis testing and model selection in a Bayesian fashion (whereas in, say, estimation problems Bayesian analysis may be convenient, but is not always necessary). Among the many fine discussions of these issues are Jeffreys (1961), Edwards, Lindman, and Savage (1963), and Kass and Raftery (1993).

A second basic premise of our motivation is that one needs automatic methods of model selection. Within the Bayesian community there has been continual debate over whether subjective or objective Bayesian methods should be used; most Bayesians today accept that both can be useful. (The objective Bayesian methods are typically based on noninformative priors, and are sometimes called "default" or "automatic" Bayesian methods to avoid the loaded connotation of the label "objective.") The argument in favor of automatic methods of model selection is particularly compelling because one often initially entertains a wide variety of models, and careful subjective specification of prior distributions for all the parameters of all the models is typically not feasible. Another sense in which we seek to be automatic is to avoid specification of the "loss function" lying behind the model selection process and simply use Bayes factors for model selection. Bayes factors can be optimal in a Bayesian decision-theoretic framework (e.g., with $0-1$ loss), but are not necessarily so for other losses. While we would encourage consideration of the loss and use of the subjective decision-theoretic approach to model selection, it is likely

that default methods will dominate in practice.

Unfortunately, operation in strict accordance with the above two basic premises is not possible. The reason is that Bayes factors in hypothesis testing and model selection typically depend rather strongly on the prior distributions, much more so than in, say, estimation. (For instance, as the sample size grows, the influence of the prior distribution disappears in estimation, but does not in hypothesis testing or model selection.) And, for most model selection problems, one cannot use standard improper noninformative priors; such priors are defined only up to a constant multiple, and the Bayes factor is itself a multiple of this arbitrary constant. The conclusion is that one cannot proceed in any clearly optimal fashion: only subjective Bayesian analysis is truly defensible for model selection, but it is not practically feasible unless only a few simple models are being considered.

The best one can hope for is thus a method that is automatic and yet produces actual Bayes factors corresponding to reasonable (proper) prior distributions. An obvious way to achieve this is simply to choose "conventional" proper prior distributions for testing or model selection, priors that seem likely to be reasonable for typical problems. This was the approach espoused by Jeffreys (1961), who recommended specific proper priors for certain standard testing problems. (The conventional proper priors Jeffreys used for testing and model selection should not be confused with his more famous "Jeffreys priors" which are typically used as noninformative priors for estimation problems. Indeed, whenever we use the phrase "the Jeffreys prior" in this paper, we will be referring to the latter type of noninformative prior.)

This "conventional proper prior" approach has met with considerable resistance, from Bayesians as well as non-Bayesians. We suspect that the negative reaction is, to a large extent, an example of what I. J. Good has called the SUTC (sweep-under-the-carpet) attitude, by which methods that have unappealing features (such as conventional proper priors) that are highly visible are resisted, whereas methods with features that are much worse will be accepted if the undesirable features are not visible (i.e., are SUTC). An example of the latter occurs with BIC, the Bayesian information criterion developed by Schwarz (1978). This criterion starts with an asymptotic approximation to the Bayes factor, and

then simply ignores the term involving the prior (because it typically has a bounded effect asymptotically) even though this term affects the Bayes factor multiplicatively and can be very large or small; at first sight, BIC is thus as bad as use of a noninformative prior with an arbitrarily chosen constant multiple. Yet many Bayesians who criticize use of conventional proper priors will routinely use BIC. (We do not mean to be unduly critical of BIC — in fact, until now it has been our favorite general purpose model selection criterion — but it SUTCs something considerably worse than what is recommended by Jeffreys. Note, also, that Kass and Wasserman, 1992, give an additional justification of BIC; see Sections 1.3.2 and 2.4 for further discussion.)

So our target as a good automatic method of hypothesis testing and model selection is the conventional prior approach of Jeffreys. The chief difficulty with this method (besides its failure to SUTC its disadvantages) is that it requires development of a reasonable conventional proper prior. In Jeffreys (1961), considerable effort is expended to develop such priors, even though only simple situations are studied. And it is far from clear that Jeffreys's arguments for developing such priors can be formalized, so as to become a general method.

What we propose here is a completely general method of testing and model selection that will be argued to be essentially equivalent to a conventional proper prior approach, but without the need to determine a reasonable proper prior (which we, in effect, SUTC). Indeed, the answers we obtain seem to closely approximate Jeffreys answers for the problems he considered. From a different perspective, our approach can be thought of as automatically "correcting" BIC by inserting a reasonable value for the term that BIC ignores. Further desirable properties of the method will be discussed as we proceed.

1.2 Preliminaries

Models $M_1, M_2, \ldots, M_q$ are under consideration, with the data $\boldsymbol{X}$ having density $f_i(\boldsymbol{x}|\boldsymbol{\theta}_i)$ under model M_i. (The densities are assumed to be taken with respect to a common measure, which is otherwise irrelevant to our analysis.) The parameter vectors $\boldsymbol{\theta}_i$ are unknown, and are of dimension k_i.

Bayesian model selection proceeds by selecting prior distributions $\pi_i(\boldsymbol{\theta}_i)$ for the pa-

rameters of each model, together with prior probabilities p_i of each model being true. The posterior probability that M_i is true is then

$$P(M_i|\boldsymbol{x}) = (\sum_{j=1}^{q} \frac{p_j}{p_i} \cdot B_{ji})^{-1}, \tag{1.1}$$

where B_{ji}, the *Bayes factor of M_j to M_i*, is defined by

$$B_{ji} = \frac{m_j(\boldsymbol{x})}{m_i(\boldsymbol{x})} = \frac{\int f_j(\boldsymbol{x}|\boldsymbol{\theta}_j)\pi_j(\boldsymbol{\theta}_j)d\boldsymbol{\theta}_j}{\int f_i(\boldsymbol{x}|\boldsymbol{\theta}_i)\pi_i(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i}; \tag{1.2}$$

here $m_j(\boldsymbol{x})$ is the marginal or predictive density of $\boldsymbol{X}$ under M_j.

Although we use this standard Bayesian language, note that one does not strictly have to assume that one of the models is true; in particular, B_{ji} can be viewed as the "weighted" likelihood ratio of M_j to M_i, and hence can be interpreted solely in terms of comparative support of the data for the two models. (See Kass and Raftery (1993) for discussion and references concerning this viewpoint.) Also, although we formally discuss only model selection, our development will also apply to hypothesis testing. Thus, if it is desired to test H_i: $\boldsymbol{\theta} \in \Theta_i$, for $i = 1, \ldots, q$, where $\boldsymbol{X} \sim f(\boldsymbol{x}|\boldsymbol{\theta})$, then (1.1) and (1.2) are still valid with $f_i = f$, $\boldsymbol{\theta}_i = \boldsymbol{\theta}$, p_i being the prior probability of H_i, and $\pi_i(\boldsymbol{\theta})$ being the conditional prior density of $\boldsymbol{\theta}$ on Θ_i. (See, also, Bertolino, Piccinato, and Racugno, 1992.)

At this point, we focus on determination of the B_{ji}. We will return to issues surrounding determination of the $P(M_i|\boldsymbol{x})$ in Section 5, but the central issue is to compute the B_{ji}. Note that if all models were assigned equal prior probability $p_i = 1/q$, then the B_{ji} strictly determine the $P(M_i|\boldsymbol{x})$. Also, the B_{ji} have the separate comparative likelihood interpretation for model comparison, and hence can be motivated outside of strict Bayesian reasoning.

Computing B_{ji} requires specification of $\pi_i(\boldsymbol{\theta}_i)$ and $\pi_j(\boldsymbol{\theta}_j)$. Often in Bayesian analysis, one can effectively use noninformative (or default) priors $\pi_i^N(\boldsymbol{\theta}_i)$. Three common choices are the "uniform" prior, $\pi_i^U(\boldsymbol{\theta}_i) = 1$; the Jeffreys prior, $\pi_i^J(\boldsymbol{\theta}_i) = (\det(\boldsymbol{I}_i(\boldsymbol{\theta}_i)))^{1/2}$, where $I_i(\boldsymbol{\theta}_i)$ is the expected Fisher information matrix corresponding to M_i; and the reference prior, $\pi_i^R(\boldsymbol{\theta}_i)$, definitions of which can be found in Bernardo (1979) and Berger and Bernardo (1992).

Using any of the π_i^N in (1.2) would yield

$$B_{ji}^N = \frac{m_j^N(\boldsymbol{x})}{m_i^N(\boldsymbol{x})} = \frac{\int f_j(\boldsymbol{x}|\boldsymbol{\theta}_j)\pi_j^N(\boldsymbol{\theta}_j)d\boldsymbol{\theta}_j}{\int f_i(\boldsymbol{x}|\boldsymbol{\theta}_i)\pi_i^N(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i}. \tag{1.3}$$

The difficulty with this solution is that the π_i^N are typically improper, and hence defined only up to arbitrary constants c_i. Hence B_{ji}^N is defined only up to (c_j/c_i), which is itself arbitrary.

A common solution to this problem is to use part of the data as a *training sample.* Let $\boldsymbol{x}(\ell)$ denote the part of the data to be so used, and $\boldsymbol{x}(-\ell)$ represent the remainder of the data. The idea is that $\boldsymbol{x}(\ell)$ will be used to convert the $\pi_i^N(\boldsymbol{\theta}_i)$ to proper posterior distributions

$$\pi_i^N(\boldsymbol{\theta}_i|\boldsymbol{x}(\ell)) = f_i(\boldsymbol{x}(\ell)|\boldsymbol{\theta}_i)\pi_i^N(\boldsymbol{\theta}_i)/m_i^N(\boldsymbol{x}(\ell)), \tag{1.4}$$

where (slightly abusing notation) $f_i(\boldsymbol{x}(\ell)|\boldsymbol{\theta}_i)$ is the marginal density of $\boldsymbol{X}(\ell)$ under M_i and

$$m_i^N(\boldsymbol{x}(\ell)) = \int f_i(\boldsymbol{x}(\ell)|\boldsymbol{\theta}_i)\pi_i^N(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i. \tag{1.5}$$

The idea is to then compute the Bayes factors with the remainder of the data, $\boldsymbol{x}(-\ell)$, using the $\pi_i^N(\boldsymbol{\theta}_i|\boldsymbol{x}(\ell))$ as priors. The result is easily shown to be

$$\begin{aligned} B_{ji}(\ell) &= \frac{\int f_j(\boldsymbol{x}(-\ell)|\boldsymbol{\theta}_j, \boldsymbol{x}(\ell))\pi_j^N(\boldsymbol{\theta}_j|\boldsymbol{x}(\ell))d\boldsymbol{\theta}_j}{\int f_i(\boldsymbol{x}(-\ell)|\boldsymbol{\theta}_i, \boldsymbol{x}(\ell))\pi_i^N(\boldsymbol{\theta}_i|\boldsymbol{x}(\ell))d\boldsymbol{\theta}_i} \\ &= B_{ji}^N \cdot B_{ij}^N(\boldsymbol{x}(\ell)), \end{aligned} \tag{1.6}$$

where

$$B_{ij}^N(\boldsymbol{x}(\ell)) = m_i^N(\boldsymbol{x}(\ell))/m_j^N(\boldsymbol{x}(\ell)). \tag{1.7}$$

Clearly (1.6) removes the arbitrariness in the choice of constant multiples of the π_i^N: the arbitrary ratio c_j/c_i that multiplies B_{ji}^N would be cancelled by the ratio c_i/c_j that would then multiply $B_{ij}^N(\boldsymbol{x}(\ell))$. Note, also, that, while the first motivating expression in (1.6) seems to require the conditional distribution of $\boldsymbol{x}(-\ell)$ given $\boldsymbol{x}(\ell)$, the second expression only utilizes the typically much simpler marginal densities of $\boldsymbol{x}(\ell)$.

The above use of a training sample makes sense only if the $m_i^N(\boldsymbol{x}(\ell))$ in (1.5) are finite. This is formalized in the following definition.

Definition 1. A training sample, $\boldsymbol{x}(\ell)$, will be called *proper* if $0 < m_i^N(\boldsymbol{x}(\ell)) < \infty$ for all M_i, and *minimal* if it is proper and no subset is proper.

Example 1. Suppose $\boldsymbol{X} = (X_1, \ldots, X_n)$, where the X_i are i.i.d. $\mathcal{N}(\mu, \sigma_2^2)$ under M_2. Under M_1, the X_i are i.i.d. $\mathcal{N}(0, \sigma_1^2)$. (The common formulation here would write $\sigma_1^2 = \sigma_2^2 = \sigma^2$; we indicate in Section 2.1 why this should be avoided.) Consider the noninformative priors $\pi_1^N(\sigma_1) = 1/\sigma_1$ and $\pi_2^N(\mu, \sigma_2) = 1/\sigma_2^2$. (Although π_2^N is the formal Jeffreys prior, it is standard to instead use $1/\sigma_2$. We use the Jeffreys prior for illustration because certain computations are somewhat simpler.) It is straightforward to show that $m_2^N(x_i) = \infty$ for a single observation, so a training sample consisting of one observation is not proper. Training samples of two or more distinct observations are proper, however. Thus a training sample such as $\boldsymbol{x}(\ell) = (x_i, x_j)$ is minimal, and, indeed, then

$$m_1^N(\boldsymbol{x}(\ell)) = \frac{1}{2\pi(x_i^2 + x_j^2)}, \quad m_2^N(\boldsymbol{x}(\ell)) = \frac{1}{\sqrt{\pi}(x_i - x_j)^2}. \tag{1.8}$$

Typically, a minimal training sample is one for which all parameters in all models are identifiable. Often it will simply be a sample of size $\max\{k_i\}$; recall that k_i is the dimension of θ_i. It can be a smaller sample, however, especially if the π_i^N are proper in some variables. Indeed, if the π_i^N are actually proper densities, then the minimal training sample is the empty set, and $B_{ji}(\boldsymbol{x}(\ell)) = B_{ji}^N$. Although minimal training samples are well defined even for dependent data situations, such as time series, there may well be advantages in accomodating the dependence structure in the choice of a minimal training sample. This will be explored elsewhere.

1.3 Relationship to Other Bayesian Methods

We briefly list other "automatic" Bayesian approaches to the problem, emphasizing those of greatest relevance to our approach.

1.3.1 Use of Conventional Priors

Jeffreys (1961) introduced use of conventional proper priors for model selection and

hypothesis testing. In the situation of Example 1, for instance, he argued for use of

$$\pi_1(\sigma_1) = \frac{1}{\sigma_1}, \quad \pi_2(\mu, \sigma_2) = \frac{1}{\sigma_2} \cdot \frac{1}{\pi\sigma_2(1 + \mu^2/\sigma_2^2)}, \tag{1.9}$$

which utilize the standard noninformative priors for the scale parameters, but a (proper) $Cauchy(0, \sigma_2)$ density for the conditional prior of μ given σ_2. By choosing $\pi(\mu|\sigma_2)$ to be proper, the indeterminancy up to multiplicative constants of the Bayes factor is avoided, at least in terms of μ. Jeffreys would identify $\sigma_1^2 = \sigma_2^2 = \sigma^2$ in this situation, and hence would not worry about indeterminancy of $\pi(\sigma) = 1/\sigma$; if this prior occurs in both models, a multiplicative constant would cancel. We return to this issue in Sections 2.1 and 5.4.

Jeffreys argument for (1.9) is rather lengthy, and may or may not be viewed as convincing. His solution is, however, eminently reasonable; choosing the "scale" of the prior for μ to be σ_2 is natural, and Cauchy priors are known to be robust in various ways. Although it is easy to object to having such choices "imposed" upon the analysis, it is crucial to keep in mind that there is no alternative (except subjective elicitation). Alternative default methods either themselves correspond to imposition of some (proper) default prior or, worse, end up not corresponding to *any* actual Bayesian analysis. This issue is important enough to deserve emphasis:

Principle 1. *Methods that correspond to use of plausible default (proper) priors are preferable to those that do not correspond to any possible actual Bayesian analysis.*

As indicated earlier, it is not our purpose here to defend this principle; we highlight the issue primarily to clearly state our goal. In this regard, we will try to mention which default Bayesian methods are, and are not, consistent with this principle. Proposals that are consistent include those of Zellner and Siow (1980), Zellner (1984), Poirier (1985), Stewart (1987), Mitchell and Beauchamp (1988), Albert (1990), Madigan and Raftery (1991), George and McCulloch (1993), McCulloch and Rossi (1993), Raftery (1993), and Verdinelli and Wasserman (1993). The limitation of these approaches is that they tend to be problem specific, with careful thought going into construction of the default prior for the specific scenario. While we are not at all opposed to careful thought, our goal is still a completely general and automatic method, but one consistent with the above principle.

1.3.2 BIC and Asymptotic Methods

The famous BIC criterion of Schwarz (1978) is based on approximating B_{ji} by

$$B_{ji}^S = \frac{f_j(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_j)(\det \hat{\boldsymbol{I}}_j)^{-1/2}}{f_i(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_i)(\det \hat{\boldsymbol{I}}_i)^{-1/2}}, \tag{1.10}$$

where $\hat{\boldsymbol{I}}_i$ is the observed information matrix under model M_i and $\hat{\boldsymbol{\theta}}_i$ is the MLE. The motivation for this arises from approximating B_{ji} using Laplace's method, resulting in the asymptotic approximation

$$B_{ji}^L = \frac{f_j(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_j)(\det \hat{\boldsymbol{I}}_j)^{-1/2}}{f_i(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_i)(\det \hat{\boldsymbol{I}}_i)^{-1/2}} \cdot \frac{(2\pi)^{k_j/2}\pi_j(\hat{\boldsymbol{\theta}}_j)}{(2\pi)^{k_i/2}\pi_i(\hat{\boldsymbol{\theta}}_i)}. \tag{1.11}$$

As the sample size goes to infinity, the first factor of B_{ji}^L typically goes to 0 or ∞, while the second factor goes to a constant. This is the primary motivation for dropping the second factor in the definition of B_{ji}^S. (The BIC criterion for model selection is the log of B_{ji}^S; sometimes it is stated as $2\log B_{ij}^S$, and often the ratio of determinants in (1.10) is replaced by $n^{-(k_2-k_1)/2}$ in the i.i.d. case. Discussion and other references for such asymptotic expressions can be found in Haughton, 1988; Gelfand and Dey, 1992; Kass and Raftery, 1993; and Raftery, 1993.)

While the second factor in (1.11) is asymptotically constant, it can be arbitrarily large or small, and can in fact be the dominant factor for small n. It is hence non-ignorable in practice. Furthermore, for nested models this second factor typically favors the simpler model, often quite substantially, so that ignoring the second factor has the effect of systematically biasing the result in favor of the more complex model. This is the type of systematic violation of Principle 1 that we seek to avoid. Finally, for many problems, the above asymptotics are not even valid; the later Example 3 will provide an illustration. (That said, we should note that B_{ji}^S is one of the better Bayesian methods in terms of systematic bias; it is, at least, correct asymptotically up to a multiplicative constant. And Kass and Wasserman, 1992, show that the approximation does correspond, asymptotically, to an actual Bayes factor in nested model situations when the simpler model is true; see Section 2.4 for further discussion.)

The above asymptotic expression also has several theoretical uses. In Section 4, it will be used to help develop conventional priors. And, as suggested above, (1.11) is quite

useful as a criterion for evaluating proposed model selection criteria. If a proposed criterion (expressed in Bayes factor form) is not asymptotically equivalent to (1.11), then it does not behave like a true Bayes factor. In a sense, our goal in this paper is to even be "correct" to second order, so that methods which fail the first order criterion are far from what we are trying to achieve.

1.3.3 Conventional Noninformative Priors

In Section 1.2, it was observed that the problem with typical noninformative priors, π_i^N, is that they are defined only up to arbitrary multiplicative constants, c_i, and that such constants would be multiplicative factors of Bayes actors. Efforts have been made to conventionally specify the constants, c_i. For instance, Smith and Spiegelhalter (1980, 1981) and Spiegelhalter and Smith (1982) propose choosing the c_i so that (in our language) $B_{ji}(\boldsymbol{x}(\ell))$ equals 1, when $\boldsymbol{x}(\ell)$ is chosen to be the (imaginary) minimal training sample that would most favor the simpler model.

The Smith and Spiegelhalter method is a sensible method that comes close to satisfying Principle 1. It fails to completely do so, however, because it also has a systematic bias in favor of the more complex model. This bias arises because of the specification that $B_{ji}(\boldsymbol{x}(\ell))$ is to be 1, even though the (imaginary) training sample is chosen most favorable to the simpler model. This will be discussed further in Section 2.4.

1.3.4 Training Sample and Partial Likelihood Methods

The training sample idea, as discussed in Section 1.2, has been informally used many times. More formal developments of the idea can be found in Lempers (1971), Atkinson (1978), Geisser and Eddy (1979), Spiegelhalter and Smith (1982), San Martini and Spezzaferri (1984), and Gelfand, Dey, and Chang (1992), although not all these works utilize the idea with ordinary Bayes factors. Other references and the general asymptotic behavior of training sample methods can be found in Gelfand and Dey (1992).

In terms of the asymptotic criterion that was discussed at the end of Section 1.3.2, it can be shown that, if the size of the training sample increases with the sample size n, then the ensuing Bayes factor is *not* asymptotically equivalent to (1.11), up to a multiplicative

constant. Hence we will be concerned only with methods that have fixed training sample size, regardless of n. None of the above developments (except Smith and Spiegelhalter, with their imaginary training sample) have operated with Bayes factors as in (1.6), with fixed training sample size.

Aitkin (1991) can also be considered to be a training sample method; it takes the entire sample $\boldsymbol{x}$ as a training sample to obtain $\pi_i^N(\boldsymbol{\theta}_i|\boldsymbol{x})$, and then uses this as the prior in (1.6) to compute the Bayes factor. This double use of the data is, of course, not consistent with usual Bayesian logic, and the method violates the asymptotic criterion rather severely.

O'Hagan (1994) proposes using a fractional part of the entire likelihood, $[f(\boldsymbol{x}|\boldsymbol{\theta})]^\alpha$, instead of a training sample. This tends to produce a more stable answer than use of a particular training sample but will also fail the asymptotic criterion, unless $\alpha \propto 1/n$ as the sample size n grows. The behavior of fractional Bayes factors is well worth study, particularly choices such as $\alpha = m_0/n$, where m_0 is the minimal training sample size. This choice may result in Bayes factors that correspond to use of sensible default priors, at least for linear models and certain choices of the background noninformative priors; such justifications have yet to be formally established, however.

Independently of our work, de Vos (1993) has proposed a training sample method for linear models that is similar to our proposal. See Appendix 2 for discussion.

1.3.5 Bounds on Bayes Factors

In comparing M_2 with a nested model M_1, one can typically find an upper bound,

$$\bar{B}_{21} = \sup_{\{(\pi_1,\pi_2)\in\Gamma\}} B_{21}, \tag{1.12}$$

over a class, Γ, of appropriate prior distributions. Such a bound says "the comparative support in the data for M_2 versus M_1" is *at most* $\bar{B}_{21}$. Such bounds can be very useful in establishing a Bayesian Ockham's razor that is independent of prior opinion (cf., Jefferys and Berger, 1992), and in demonstrating severe evidential inadequacy of non-Bayesian measures such as P-values (cf., Edwards, Lindman, and Savage, 1963, Berger and Sellke, 1987, and Berger and Delampady, 1987). Indeed, in Delampady and Berger (1990) it is shown, in this way, that model selection via classical chi-squared testing of fit gives answers which (as commonly interpreted) are very biased in favor of the complex model.

Bounds, such as $\bar{B}_{21}$, cannot, however, be used as general model selection tools, because they operate only in one direction; corresponding lower bounds, $\underline{B}_{21}$, are typically zero. Also, $\bar{B}_{21}$ can be an unrealistically large upper bound and typically does not behave in accordance with (1.11) asymptotically.

1.4 Preview

In Section 2 we introduce the idea of the *intrinsic Bayes factor* (IBF) for nested models. Several variants of the IBF are presented, and illustrated on simple examples. Section 3 discusses the IBF for nonnested models or hypotheses, and illustrates the method on several data sets. (We consider only fairly basic examples in this paper, so that insight into the behavior of IBFs, and comparison with other methods, can be more readily seen.)

Section 4 introduces the notion of the *intrinsic prior*; this is the (often proper) conventional prior that would give answers similar to the IBF. The demonstration that IBFs correspond to Bayes factors for sensible conventional priors is a cornerstone of the justification for the approach. When computable, the intrinsic prior can itself be used to conduct the Bayesian analysis, in place of using the IBF. Thus another view of the developments here is that they provide a method for deriving conventional priors for model selection and hypothesis testing.

Section 5 considers problems of prediction and model uncertainty, and shows how IBFs can be used to overcome the difficult question of model weighting. Section 6 summarizes the results and provides practical guidelines.

2. INTRINSIC BAYES FACTORS FOR TWO NESTED MODELS OR HYPOTHESES

2.1 Nested Models

Assume that M_1 is nested in M_2, in the sense that we can write $\boldsymbol{\theta}_2 = (\boldsymbol{\xi}, \boldsymbol{\eta})$ and f_1 and f_2 satisfy

$$f_1(\boldsymbol{x}|\boldsymbol{\theta}_1) = f_2(\boldsymbol{x}|\boldsymbol{\xi} = \boldsymbol{\theta}_1,\ \boldsymbol{\eta} = \boldsymbol{\eta}_0), \tag{2.1}$$

where $\boldsymbol{\eta}_0$ is a specified value of $\boldsymbol{\eta}$. It will sometimes be convenient to identify $\boldsymbol{\theta}_1$ with $(\boldsymbol{\theta}_1, \boldsymbol{\eta}_0)$, so that $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ lie in the same space. Also, we will sometimes simply write

$\boldsymbol{\theta}_2 = (\boldsymbol{\theta}_1, \boldsymbol{\eta})$, although this is a dangerous (but common) practice. The danger is that $\boldsymbol{\theta}_1$ in $f_1(\boldsymbol{x}|\boldsymbol{\theta}_1)$ and $\boldsymbol{\theta}_1$ in $f_2(\boldsymbol{x}|(\boldsymbol{\theta}_1, \boldsymbol{\eta}))$ can have very different interpretations, yet because the symbols are the same it is all-too-easy to assign them the same prior (especially when default priors are being used).

Example 2. We wish to predict automotive fuel consumption, Y, from the weight, X_1, and engine size, X_2, of a vehicle. Two models are entertained:

$$\begin{aligned} M_1\text{: } & Y = X_1\beta_1 + \varepsilon_1, \quad \varepsilon_1 \sim \mathcal{N}(0, \sigma_1^2) \\ M_2\text{: } & Y = X_1\beta_1 + X_2\beta_2 + \varepsilon_2, \quad \varepsilon_2 \sim \mathcal{N}(0, \sigma_2^2). \end{aligned} \tag{2.2}$$

Thinking, first, about M_2, suppose the elicited prior is of the form $\pi_2(\beta_1, \beta_2, \sigma_2) = \pi_{21}(\beta_1) \cdot \pi_{22}(\beta_2) \cdot \pi_{23}(\sigma_2)$. It is then quite common to choose, as the M_1 prior, $\pi_1(\beta_1, \sigma_1) = \pi_{21}(\beta_1) \cdot \pi_{12}(\sigma_1)$, i.e., to use the same prior for β_1 as in Model 1. (Even worse, conceptually, is to equate σ_1 and σ_2 and give them the same prior.) The point, of course, is that β_1 has a different meaning (and value) under M_1 than under M_2. For instance, regressing fuel consumption on weight alone will yield a larger coefficient than regressing on both weight and engine size, because of the considerable positive correlation between weight and engine size in the data. (To clearly see this, consider the case where weight and engine size are exactly linearly related.)

This is an important issue because many of the schemes for developing conventional priors are based on a formalization of such parameter identifications, and are hence suspect. Intrinsic Bayes factors will naturally avoid the problem.

In Section 2.3 and Section 4, the following assumption, which is virtually always true for nested models, will be needed.

Assumption N. If M_1 is nested in M_2, assume that, as the sample size $n \to \infty$,

$$\hat{\boldsymbol{\theta}}_2 \overset{\text{under } M_1}{\longrightarrow} \boldsymbol{\theta}_2^* = (\boldsymbol{\theta}_1, \boldsymbol{\eta}_0), \tag{2.3}$$

where $\hat{\boldsymbol{\theta}}_2$ is the MLE under M_2

2.2 The Intrinsic Bayes Factor

For a given data set $\boldsymbol{x}$, there will typically be many minimal training samples as defined in Section 1.2. Let

$$\mathcal{X}_T = \{\boldsymbol{x}(1), \boldsymbol{x}(2), \ldots, \boldsymbol{x}(L)\} \tag{2.4}$$

denote the set of all minimal training samples, $\boldsymbol{x}(\ell)$. Clearly the Bayes factor $B_{21}(\ell)$, defined in (1.6), will depend on choice of the minimal training sample. To eliminate this dependence and increase stability, a natural idea is to average the $B_{21}(\ell)$ over all $\boldsymbol{x}(\ell) \in \mathcal{X}_T$. This average can be done either arithmetically or geometrically, leading to the *arithmetic intrinsic Bayes factor* (AIBF) and *geometric intrinsic Bayes factor* (GIBF) defined, respectively, by

$$B_{21}^{AI} = \frac{1}{L}\sum_{\ell=1}^{L} B_{21}(\ell) = B_{21}^{N} \cdot \frac{1}{L}\sum_{\ell=1}^{L} B_{12}^{N}(\boldsymbol{x}(\ell)), \tag{2.5}$$

$$B_{21}^{GI} = \left(\prod_{\ell=1}^{L} B_{21}(\ell)\right)^{1/L} = B_{21}^{N} \cdot \left(\prod_{\ell=1}^{L} B_{12}^{N}(\boldsymbol{x}(\ell))\right)^{1/L}, \tag{2.6}$$

where the $B_{12}^{N}(\boldsymbol{x}(\ell))$ are defined in (1.7). Note that $B_{21}^{GI} \leq B_{21}^{AI}$, since the geometric mean is less than or equal to the arithmetic mean. Thus B_{21}^{GI} will favor the nested (simpler) model to a greater extent than will B_{21}^{AI}.

Important Point 1: We *define* B_{12}^{AI} to be $1/B_{21}^{AI}$, and not by (2.5) with the indices reversed. The asymmetry arises because of M_1 being nested within M_2, and will be explained in Section 2.3. For B_{21}^{GI} there is no problem; reversing the indices in (2.6) clearly results in $1/B_{21}^{GI}$.

Important Point 2: If the sample size is very small, there will clearly be problems with using a part of the data as a training sample. For very small samples we recommend alternative versions of intrinsic Bayes factors, defined in Sections 2.3, 3.3, and Section 4.

Example 1 (continued). $\boldsymbol{X} = (X_1, \ldots, X_n)$ was an i.i.d. sample from M_1: $\mathcal{N}(0, \sigma_1^2)$ or M_2: $\mathcal{N}(\mu, \sigma_2^2)$. Using $\pi_1^N(\sigma_1) = 1/\sigma_1$ and $\pi_2^N(\mu, \sigma_2) = 1/\sigma_2^2$, computation yields

$$B_{21}^{N} = \sqrt{\frac{2\pi}{n}} \cdot (1 + \frac{n\bar{x}^2}{s^2})^{n/2}, \tag{2.7}$$

where $s^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2$. Together with (1.7) and (1.8), and noting that $\mathcal{X}_T$ consists of all $L = n(n-1)/2$ pairs of (different) observations, it follows that

$$B_{21}^{AI} = B_{21}^{N} \cdot \frac{1}{L} \sum_{\ell=1}^{L} \frac{(x_1(\ell) - x_2(\ell))^2}{2\sqrt{\pi}[x_1^2(\ell) + x_2^2(\ell)]}$$

$$= B_{21}^{N} \cdot \frac{1}{n(n-1)} \sum_{i<j} \frac{(x_i - x_j)^2}{\sqrt{\pi}[x_i^2 + x_j^2]}, \tag{2.8}$$

$$B_{21}^{GI} = B_{21}^{N} \cdot \left(\prod_{\ell=1}^{L} \frac{(x_1(\ell) - x_2(\ell))^2}{2\sqrt{\pi}[x_1^2(\ell) + x_2^2(\ell)]} \right)^{1/L}. \tag{2.9}$$

Note that B_{21}^{AI} and B_{21}^{GI} are defined for essentially any nested models, even those which are non-standard, such as the following.

Example 3. Assume that $\boldsymbol{X} = (X_1, \ldots, X_n)$ is an i.i.d. sample from either M_1: $X_i \sim \mathcal{N}(\theta_1, 1)$ with $\theta_1 < 0$; or M_2: $X_i \sim \mathcal{N}(\theta_2, 1)$ with $\theta_2 \in R^1$. Once again it is important to keep in mind that θ_1 and θ_2 might well be distinct quantities, apriori, even when $\theta_2 < 0$. It could be dangerous to formulate this problem by saying $X_i \sim \mathcal{N}(\theta, 1)$, with M_1: $\theta < 0$ and M_2: $\theta \in \mathrm{R}^1$. The danger is in being misled by the same symbol, θ, appearing in M_1 and M_2 which might cause one to assume that, say, $\pi_1(\theta)$ (under M_1) equals $\pi_2(\theta|\theta < 0)$ (under M_2). Such assumptions are simply not typically warranted.

The usual noninformative priors here are $\pi_1^N(\theta_1) = 1_{(-\infty,0)}(\theta_1)$ and $\pi_2^N(\theta_2) = 1$. Easy calculations then yield

$$B_{21}^{N} = 1/\Phi(-\sqrt{n}\bar{x}), \tag{2.10}$$

where Φ is the standard normal c.d.f. A minimal training sample is a single observation, since

$$m_1^N(x_i) = \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_i - \theta_1)^2} d\theta_1 = \Phi(-x_i)$$

and $m_2^N(x_i) = 1$ are both finite for a single observation. It follows from (2.5) and (2.6), that

$$B_{21}^{AI} = \frac{1}{\Phi(-\sqrt{n}\bar{x})} \cdot \frac{1}{n} \sum_{i=1}^{n} \Phi(-x_i), \tag{2.11}$$

$$B_{21}^{GI} = \frac{1}{\Phi(-\sqrt{n}\bar{x})} \cdot \left[\prod_{i=1}^{n} \Phi(-x_i) \right]^{1/n}. \tag{2.12}$$

That this is a non-standard example, which is difficult to handle by ordinary (default) methods, is partly indicated by the fact that the asymptotic expression (1.11) (and hence BIC) are not valid here. The correct asymptotics can be found in Haughton and Dudley (1992), which studies very general problems of this type. For later reference we record the asymptotic analogue of (1.11):

$$B_{21} \cong \frac{\pi_2(\bar{x})}{\Phi(-\sqrt{n}\bar{x})\pi_1(\min\{\bar{x}, 0\})}; \tag{2.13}$$

this is valid if π_2 is continuous and π_1 is continuous and has a finite limit, $\pi_1(0)$, at zero.

2.3 The Expected Intrinsic Bayes Factors

For small sample sizes, the training sample averages in (2.5) and (2.6) can have large variances (as statistics in a frequentist sense), which indicates an instability of IBFs when the sample size is small. Also, computation can be a problem if L is large (see, also, Section 2.5). One attractive solution to both these problems is to replace the averages in (2.5) and (2.6) by their expectations, evaluated at the MLE. Formally, we define the *expected arithmetic intrinsic Bayes factor* and *expected geometric intrinsic Bayes factor* by, respectively,

$$B_{21}^{EAI} = B_{21}^{N} \cdot \frac{1}{L} \sum_{i=1}^{L} E_{\hat{\theta}_2}^{M_2}[B_{12}^{N}(\boldsymbol{X}(\ell))], \tag{2.14}$$

$$B_{21}^{EGI} = B_{21}^{N} \cdot \exp\{\frac{1}{L} \sum_{i=1}^{L} E_{\hat{\theta}_2}^{M_2}[\log B_{12}^{N}(\boldsymbol{X}(\ell))]\}, \tag{2.15}$$

where the expectations are under M_2, with $\boldsymbol{\theta}_2$ set equal to the MLE $\hat{\boldsymbol{\theta}}_2$. If the $\boldsymbol{X}(\ell)$ are exchangeable, as is common, then the averages over L are clearly superfluous.

That (2.14) and (2.15) are justified as approximations to (2.5) and (2.6) for large L and under M_2 is obvious. However, they also are valid approximations under M_1 if Assumption N in Section 2.1 is satisfied, for then (under M_1) $\hat{\boldsymbol{\theta}}_2 \cong (\boldsymbol{\theta}_1, \boldsymbol{n}_0)$ which, together with (2.1), shows that the expectations in (2.14) and (2.15) are equivalent to those under M_1. This very helpful property is unique to nested models and will, unfortunately, prevent us from deriving analogues of (2.14) and (2.15) for nonnested problems.

Example 1 (continued). Here the $\boldsymbol{X}(\ell)$ are exchangeable, so from (2.14) and (2.15) we see that

$$B_{21}^{EAI} = B_{21}^{N} \cdot E_{\hat{\theta}_2}^{M_2}\left[\frac{(X_i - X_j)^2}{2\sqrt{\pi}(X_i^2 + X_j^2)}\right] = B_{21}^{N} \cdot \left(\frac{1 - \exp\{-n\bar{x}^2/s^2\}}{2\sqrt{\pi}[n\bar{x}^2/s^2]}\right); \tag{2.16}$$

see Berger and Pericchi (1993) for computation of the expectation. Also,

$$B_{21}^{EGI} = B_{21}^{N} \cdot \exp\left\{E_{\hat{\theta}_2}^{M_2}[\log \frac{(X_i - X_j)^2}{2\sqrt{\pi}(X_i^2 + X_j^2)}]\right\}, \tag{2.17}$$

where $\hat{\boldsymbol{\theta}}_2 = (\bar{x}, s^2/n)$. Here the expectation can be evaluated only as an infinite series (see Berger and Pericchi, 1993); but numerical computation is straightforward, as discussed in Section 2.5.

Example 3 (continued). Using (2.11) and (2.12), (2.14) and (2.15) become, respectively (again using exchangeability of the $\boldsymbol{X}(\ell)$),

$$\begin{aligned} B_{21}^{EAI} &= \frac{1}{\Phi(-\sqrt{n}\bar{x})} \cdot E_{\hat{\theta}_2}^{M_2}[\Phi(-X_i)] \\ &= \frac{1}{\Phi(-\sqrt{n}\bar{x})} \cdot \Phi(-\bar{x}/\sqrt{2}), \end{aligned} \tag{2.18}$$

$$B_{21}^{EGI} = \frac{1}{\Phi(-\sqrt{n}\bar{x})} \cdot \exp\{E_{\hat{\theta}_2}^{M_2}[\log \Phi(-X_i)]\}; \tag{2.19}$$

here $X_i \sim \mathcal{N}(\theta, 1)$ under M_2 and $\hat{\theta}_2 = \bar{x}$. Again, the expectation in (2.19) cannot be done in closed form.

As with B_{12}^{AI}, we *define* $B_{12}^{EAI} = 1/B_{21}^{EAI}$. In this there is no option, since in most problems (such as Examples 1 and 3)

$$E_{\hat{\theta}_2}^{M_2}[B_{21}^{N}(\boldsymbol{X}(\ell))] = \infty. \tag{2.20}$$

This also explains the definition of $B_{12}^{AI} = 1/B_{21}^{AI}$; although B_{12}^{AI} could be defined as in (2.5) with the indices reversed, the average of the $B_{12}^{N}(\boldsymbol{x}(\ell))$ would typically diverge as $L \to \infty$, resulting in a Bayes factor that would violate Principle 1.

2.4 Comparisons

We pause to compare the various intrinsic Bayes factors with each other and with certain other methods, so as to obtain insight in these simple cases as to whether our goals are being achieved. The comparisons we make are with the asymptotic expression (1.11), the Schwarz approximation (1.10), with Jeffreys (1961), and with Smith and Spiegelhalter (1980); as indicated in Section 1.3, we view these as among the best of the previously published approaches, when they apply. None of these approaches apply to Example 3; hence we delay discussion of that example until Section 4.

Example 1 (continued). Since $\hat{\mu} = \bar{x}$, $\hat{\sigma}_1 = (\sum_{i=1}^{n} x_i^2/n)^{1/2} = (\bar{x}^2 + s^2/n)^{1/2}$, $\hat{\sigma}_2 = (s^2/n)^{1/2}$, it is straightforward to compute the following:

Asymptotic Approximation: (1.11) becomes

$$B_{21}^L = B_{21}^N \cdot \frac{\hat{\sigma}_2^2}{\hat{\sigma}_1} \cdot \frac{\pi_2(\hat{\mu}, \hat{\sigma}_2)}{\pi_1(\hat{\sigma}_1)}. \tag{2.21}$$

Schwarz Approximation: (1.10) becomes

$$B_{21}^S = B_{21}^N \cdot \frac{1}{\sqrt{2\pi}}.$$

Jeffreys Approach: Jeffreys (1961) used the priors in (1.9), and approximated the resulting Bayes factor by

$$B_{21}^J = B_{21}^N \cdot \hat{\sigma}_2 \cdot \frac{1}{\hat{\sigma}_2 \pi [1 + \hat{\mu}^2/\hat{\sigma}_2^2]}. \tag{2.22}$$

Smith and Spiegelhalter (1980): Their Bayes factor is

$$B_{21}^{SS} = B_{21}^N \cdot \frac{1}{\sqrt{\pi}}.$$

It is useful to rewrite $\pi_2(\mu, \sigma_2)$ as

$$\pi_2(\mu, \sigma_2) = \pi_2(\mu|\sigma_2)\pi_2(\sigma_2). \tag{2.23}$$

If, now, the common noninformative choices $\pi_1(\sigma_1) = 1/\sigma_1$ and $\pi_2(\sigma_2) = 1/\sigma_2$ are made, then (2.21) becomes

$$B_{21}^{L} = B_{21}^{N} \cdot \hat{\sigma}_2 \cdot \pi_2(\hat{\mu}|\hat{\sigma}_2). \tag{2.24}$$

Note that B_{21}^{J} is of this form with $\pi_2(\mu|\sigma_2)$ being $Cauchy(0, \sigma_2)$, consistent with (1.9). Likewise, rewrite (2.16) and (2.17) as

$$B_{21}^{EAI} = B_{21}^{N} \cdot \hat{\sigma}_2 \cdot \left(\frac{1 - \exp\{-\hat{\mu}^2/\hat{\sigma}_2^2\}}{2\sqrt{\pi}[\hat{\mu}^2/\hat{\sigma}_2]} \right), \tag{2.25}$$

$$B_{21}^{GAI} = B_{21}^{N} \cdot \hat{\sigma}_2 \cdot \left(\frac{1}{\hat{\sigma}_2} E^{M_2}_{(\hat{\mu},\hat{\sigma}_2)}[\log \frac{(X_i - X_j)^2}{2\sqrt{\pi}(X_i^2 + X_j^2)}] \right). \tag{2.26}$$

Recall that one of our goals was to develop an automatic method that "reproduces" authentic sensible Bayes factors. B_{21}^{EAI} succeeds astonishingly well. Indeed, it can be shown that

$$\pi_2^I(\mu|\sigma_2) = \frac{1 - \exp\{-\mu^2/\sigma_2^2\}}{2\sqrt{\pi}[\mu^2/\sigma_2]} \tag{2.27}$$

is a proper prior (integrating to one over μ) and, furthermore, is virtually equivalent to Jeffreys $C(0, \sigma_2)$ choice of $\pi_2(\mu|\sigma_2)$; indeed, the two prior densities never differ by more than 15%, as can be seen in Figure 1, when $\sigma_2 = 1$.

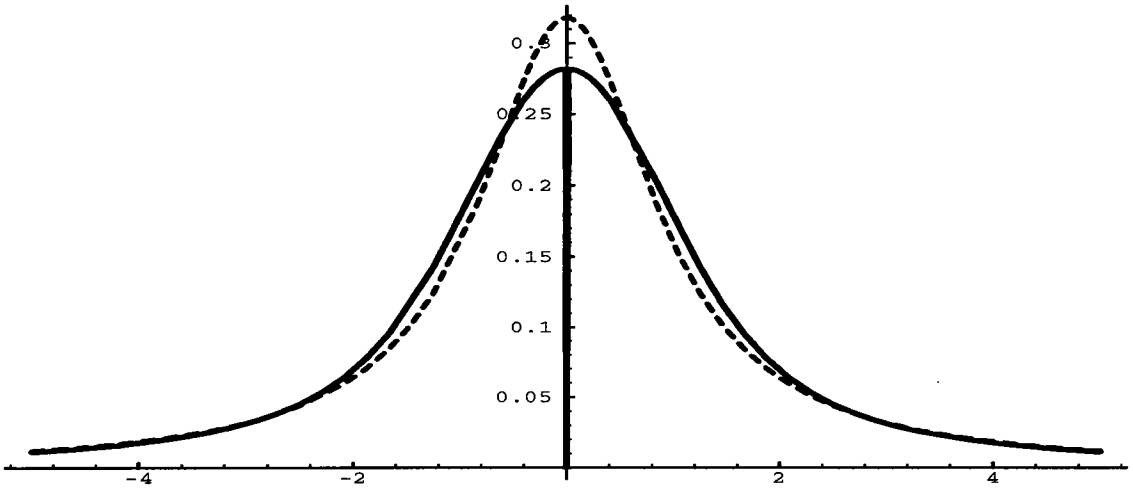

Figure 1. Comparison of the Cauchy (dashed) and intrinsic (solid) priors for Example 1.

This remarkable property of B_{21}^{EAI} is, unfortunately, not shared by B_{21}^{EGI}. The bracketed term in (2.26) does not correspond to a proper prior, although it does behave qualitatively similarly to $\pi_2^I(\mu|\sigma_2)$.

Comparing B_{21}^J or B_{21}^{EAI} with B_{21}^{SS}, we see that the latter is *larger* by a factor of about $\sqrt{\pi}(1+\hat{\mu}^2/\hat{\sigma}_2^2)$, which is always greater than one (and can be arbitrarily large). This phenomenon is quite generally true and provides support for the assertion in Section 1.3 that B_{21}^{SS} is biased towards the more complex model.

Similarly, B_{21}^S is *larger* by a factor of $\sqrt{\pi/2}(1+\hat{\mu}^2/\hat{\sigma}_2^2)$, and is itself thus biased towards the more complex model. If one had chosen $\pi_2(\mu|\sigma_2)$ to be $\mathcal{N}(0,\sigma_2^2)$, then B_{21}^L would equal $B_{21}^S \cdot \exp\{-\overline{x}^2/(2\hat{\sigma}_2^2)\}$. If M_1 were the true model and n were large, then $\overline{x} \cong 0$ and $B_{21}^L \cong B_{21}^S$. This is the basis for the argument in Kass and Wasserman (1992) that B_{21}^S is approximately a Bayes factor when the simpler model is true in a nested situation. Of course, ignoring $\exp\{-\overline{x}^2/2\hat{\sigma}_2^2\}$ is not always appropriate. (And this relationship between B_{21}^S and Bayes factors has only been established in somewhat special circumstances.)

The story forB_{21}^{AI} and B_{21}^{GI} is similar. Remarkably, the average in (2.5) again corresponds to a proper prior, even though it is a data-dependent proper prior (see Section 4.2). The corresponding term of (2.6) is qualitatively similar, but again does not correspond exactly to a proper prior.

2.5 Computational Issues

2.5.1 Computing B_{21}^{AI} and B_{21}^{GI}

There are three aspects to the computation of (2.5) or (2.6): computation of B_{21}^N, computation of the $B_{12}^N(\boldsymbol{x}(\ell))$, and the summation over ℓ.

(i) *Computing* B_{21}^N: This is a standard (though not necessarily easy) problem; see Rosenkrantz (1992), Kass and Raftery (1993) and Raftery (1993) for discussion.

(ii) *Computing the* $B_{12}^N(\boldsymbol{x}(\ell))$: Interestingly, these are often available in closed form for minimal training samples. This is true in the general linear model; and see Examples 4 and 6 in Section 3 for quite surprising illustrations. If the computation must be done numerically, it is obviously advantageous to use a numerical scheme which simultaneously does all the integrations. As an example, if one uses importance sampling to evaluate the integrals, it would be nice if one could generate a single importance sample that could be used for all the $B_{12}^N(\boldsymbol{x}(\ell))$ simultaneously. This is, in fact, quite feasible; because the $\boldsymbol{x}(\ell)$

are minimal training samples, it can be seen that the integrands in (1.5) tend to be quite diffuse, so that a "global" importance function can often be effective. Specific algorithms are currently being studied and will be reported elsewhere. See Gelfand, Dey, and Chang (1992) and Geyer and Thompson (1992) for related ideas.

(iii) *Summation over ℓ*: Often L, the number of minimal training samples, is $\binom{n}{m}$, where n is the sample size and m is the size of the minimal training sample. This can be enormous if n is moderate or large and $m \geq 2$. The natural solution to a too-large L is to sum only over a subset of $\mathcal{X}_T$, the set of minimal training samples. This could be a "random" selection of samples from $\mathcal{X}_T$, or could be systematic. As an example of the latter, in the situation of Example 1 it would be reasonable to choose disjoint pairs of observations, for instance $\{(x_1, x_2),\ (x_3, x_4), \ldots, (x_{n-1}, x_n)\}$, as the minimal training samples to be used in computing B_{21}^{AI} or B_{21}^{GI}. Unless n is quite small, this will yield essentially the same answer as use of all minimal training samples. Systematic choice of the training sample is a special case of using "weighted" averages to form IBFs; the "weights" in a systematic choice are simply zero or one.

2.5.2 Computing B_{21}^{EAI} and B_{21}^{EGI}

In the nonexchangeable case, (2.14) and (2.15) can be quite difficult to compute, unless the expectation over M_2 can be evaluated in closed form (which, however, is possible for the general linear model and B_{21}^{EAI}). In the exchangeable case, however, (2.14) and (2.15) are typically no harder to compute than B_{21}^{AI} and B_{21}^{GI}; one simply simulates the relevant expectations, using r i.i.d. samples (of sizes equal to the minimal training sample size), $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_r$, generated from M_2 with parameter $\boldsymbol{\theta}_2 = \hat{\boldsymbol{\theta}}_2$ (the MLE for the original data $\boldsymbol{x}$). For instance, the expectation in (2.14) becomes

$$E_{\hat{\boldsymbol{\theta}}_2}^{M_2}[B_{12}^r(\boldsymbol{X}(\ell))] \cong \frac{1}{r}\sum_{i=1}^{r} B_{12}^N(\boldsymbol{x}_i). \tag{2.28}$$

Note that use of (2.28) can be interpreted as a use of "imaginary training samples," as opposed to training samples from the actual data.

3. THE INTRINSIC BAYES FACTOR FOR NONNESTED MODELS OR HYPOTHESES

3.1 Standard IBFs and Trimmed IBFs

The expressions for B_{21}^{AI} and B_{21}^{GI} are computable for any models, and can potentially be used as IBFs even in nonnested cases.

Example 4 (Location-Scale). Suppose the M_i are location-scale densities

$$f_i(\boldsymbol{x}|\mu_i,\sigma_i) = \prod_{j=1}^{n} \sigma_i^{-1} g_i((x_j-\mu_i)/\sigma_i), \tag{3.1}$$

where $\boldsymbol{x} = (x_1,\ldots,x_n)$, and $\mu_i \in \mathbf{R}^1$ and $\sigma_i > 0$ are unknown. The standard noninformative priors here are $\pi_i^N(\mu_i,\sigma_i) = 1/\sigma_i$. Lemma A1 in Appendix 1 then shows that

$$m_i^N(x_j,x_k) = 1/(2|x_j - x_k|). \tag{3.2}$$

It is easy to see that $m_i^N(x_j) = \infty$ for a single observation, so that minimal training samples are any pair of (different) observations. Since (3.2), rather remarkably, does not depend on M_i, it is clear that $B_{21}^N(\boldsymbol{x}(\ell)) = B_{12}^N(\boldsymbol{x}(\ell)) = 1$ for any minimal training sample. It follows that

$$B_{21}^{AI} = B_{21}^{GI} = B_{21}^N, \quad B_{12}^{AI} = B_{12}^{GI} = B_{12}^N = 1/B_{21}^N, \tag{3.3}$$

which is basically the procedure considered by Spiegelhalter (1980). This is delightfully simple: all IBFs correspond to simply computing the Bayes factor for the ordinary noninformative priors. We will return to discussion of this situation in Section 4.3. See Pericchi and Pérez (1992) for an example.

In general, B_{21}^{AI} will not equal $1/B_{12}^{AI}$, and several difficulties can then arise. First, it is often not clear in nonnested situations which model is "more complex," and hence which model should be called M_2. This is irrelevant for B_{21}^{GI}, but can strongly affect B_{21}^{AI}. A second difficulty is that, in nonnested situations,

$$\bar{B}_{ij}^N \equiv \frac{1}{L}\sum_{\ell=1}^{L} B_{ij}^N(\boldsymbol{x}(\ell)) \tag{3.4}$$

can be extremely unstable. Indeed, its expectation can be infinite under one or both models, indicating serious potential problems.

Example 5. As an artificial — but illuminating — example, suppose that $X_1, \ldots, X_n$ are an i.i.d. sample from M_1: $\mathcal{N}(\theta_1, 1)$ or M_2:$\mathcal{N}(0, \theta_2^2)$. The usual noninformative priors are $\pi_1^N(\theta_1) = 1$ and $\pi_2^N(\theta_2) = 1/\theta_2$. For a single observation, x_i, $m_1^N(x_i) = 1$ and $m_2^N(x_i) = 1/(2|x_i|)$. Hence the minimal training samples are just the individual x_i.

Here there is a clear difficulty in stating which model is more complex. Note, however, that (3.4) becomes $\bar{B}_{12}^N = \frac{1}{n}\sum_{\ell=1}^{n} 2|x_\ell|$, which is well-behaved under either M_1 or M_2. In contrast, the corresponding $\bar{B}_{21}^N = \frac{1}{n}\sum_{\ell=1}^{n} 1/(2|x_\ell|)$ would not be well behaved (converging to ∞ as $n \to \infty$), and would result in a very unstable IBF. Hence, it is natural here to call M_2 "more complex" and use

$$B_{21}^{AI} = B_{21}^N \cdot \frac{2}{n}\sum_{\ell=1}^{n} |x_\ell|. \tag{3.5}$$

Of course, B_{21}^{GI} could also be used.

In this simple example the difficulties were resolvable, but in more complicated situations this might not be so. Hence, it is useful to consider possible solutions to the difficulties. One potential adhoc solution is to use trimmed averages instead of $\bar{B}_{12}^N$.

Definition 2. The *α-trimmed IBF*, $B_{21}^{\alpha AI}$ or $B_{21}^{\alpha GI}$, are defined by (2.5) or (2.6), but with the $(\alpha/2)L$ smallest and $(\alpha/2)L$ largest values of $B_{12}^N(\boldsymbol{x}(\ell))$ removed, and L replaced by $(1-\alpha)L$.

A moderate amount of trimming, say 10% or 20%, can dramatically improve the stability of B_{21}^{AI}, and is recommended if B_{21}^{AI} is used for nonnested models. Trimming also overcomes the purely numerical problem that, because of data rounding, it may happen that, what was thought to be a minimal training sample really is not, with the typical result that $m_2^N(\boldsymbol{x}(\ell)) = \infty$. For instance, if M_2 is a continuous location-scale density, then $\{x_i, x_j\}$ is theoretically a minimal training sample, since $x_i \neq x_j$ with probability one. But, in practice, data rounding may cause two observations to be equal, in which case use of (3.2) would clearly cause problems. Automatic trimming can eliminate this numerical

problem. (Note that trimmed IBFs are all different entities, in the sense that they do not converge to the expected IBFs as the sample size grows.)

As a final observation, note that one could trim *all* $B_{12}^N(\boldsymbol{x}(\ell))$ except the median value, B_{12}^{med}. This would result in

$$B_{21}^{\text{med}AI} = B_{21}^{\text{med}GI} = B_{21}^N \cdot B_{12}^{\text{med}} = \frac{1}{B_{12}^{\text{med}AI}} = \frac{1}{B_{12}^{\text{med}GI}}. \tag{3.6}$$

This complete trimming actually has considerable appeal, because the arithmetic and geometric Bayes factors then become the same, and because there is then no need to ascertain which model is more complex. This definition is not without its own difficulties, however. For instance, there is a consistency problem in definition if there are more than two models, since the medians of the $B_{ij}^N(\boldsymbol{x}(\ell))$ will typically occur at different $\boldsymbol{x}(\ell)$ for different (i, j) pairs. Nevertheless, this "median IBF" deserves further study.

We finish this section with a real-data example, indicating the severity of the types of problems discussed above.

Example 6. Proschan (1963) considers failure data arising from air conditioners on several different airplanes. For each individual airplane, he suggests that an exponential model fits the data well. To illustrate this, consider the following 30 failure times from a particular airplane: 23, 261, 87, 7, 120, 14, 62, 47, 225, 71, 246, 21, 42, 20, 5, 12, 120, 11, 3, 14, 71, 11, 14, 11, 16, 90, 1, 16, 52, 95. Three models are entertained for the (assumed independent) failure times, X_i:

$$M_1\colon f_1(x_i|\theta_1) = \theta_1^{-1}\exp\{-x_i/\theta_1\} \quad (\text{Exponential}(\theta_1)),$$

$$M_2\colon f_2(x_i|\mu,\sigma) = \frac{\exp\{-(\log x_i - \mu)^2/(2\sigma^2)\}}{\sqrt{2\pi}\sigma x_i} \quad (\text{Lognormal}(\mu,\sigma)),$$

$$M_3\colon f_3(x_i|\gamma,\beta) = \beta x_i^{(\beta-1)}\gamma^{-\beta}\exp\{-(x_i/\gamma)^\beta\} \quad (\text{Weibull}(\gamma,\beta)).$$

Note that these three models can represent very different behavior in terms of failure rates. The models here are not nested (except for M_1 within M_3), and the sample size is not large enough to trust asymptotics; hence, there are no established default Bayesian methods of model selection here.

For M_1 and M_2, the standard noninformative priors are $\pi_1^N(\theta_1) = 1/\theta_1$ and $\pi_2^N(\mu, \sigma) = 1/\sigma$. For M_3, both the Jeffreys prior, $\pi_3^J(\gamma, \beta) = 1/\gamma$, and the reference prior, $\pi_3^R(\gamma, \beta) = 1/(\gamma\beta)$, have been used; we will consider both. Calculation yields, for $\boldsymbol{x} = (x_1 \ldots, x_n)$,

$$m_1^N(\boldsymbol{x}) = \frac{\Gamma(n)}{(\Sigma x_i)^n}, \quad m_2^N(\boldsymbol{x}) = \frac{\Gamma((n-1)/2)}{(\prod_{i=1}^{n} x_i)\pi^{(n-1)/2} 2\sqrt{n} S_y^{(n-1)}}, \tag{3.7}$$

where $S_y^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2$, $y_i = \log x_i$. It is easy to see that minimal training samples are of the form $\boldsymbol{x}(\ell) = (x_i, x_j)$, $x_i \neq x_j$, so that

$$m_1^N(\boldsymbol{x}(\ell)) = \frac{1}{(x_i + x_j)^2}, \quad m_2^N(\boldsymbol{x}(\ell)) = \frac{1}{2x_i x_j |\log(x_i/x_j)|}. \tag{3.8}$$

For the reference prior and M_3, it can be shown, quite surprisingly, that $m_3^N(\boldsymbol{x}(\ell)) = m_2^N(\boldsymbol{x}(\ell))$. However, computation of $m_3^N(\boldsymbol{x})$ for the full sample, as well as all computations for M_3 and the Jeffreys prior, required one-dimensional numerical integration (over β; the integral over γ can be done in closed form). These were done using IMSL routines.

There are 435 pairs of observations, but nine of them have $x_i = x_j$. Hence those nine are not minimal training samples, and were ignored. Since M_1 is nested in M_3, we do not consider B_{13}^{AI}. And although M_1 is not strictly nested in M_2, it is intuitively clear that M_2 is a more complex model, so we also do not consider B_{12}^{AI}.

Table 1. IBFs for the failure data.

	B_{21}^{AI}	B_{21}^{GI}	B_{31}^{AI}	B_{31}^{GI}	B_{32}^{AI}	B_{32}^{GI}	B_{23}^{AI}	B_{23}^{GI}
Jeffreys prior	0.37	0.33	0.25	0.15	0.66	0.46	3.93	2.15
Reference prior	same	same	0.26	0.23	0.70	0.70	1.42	1.42

Table 1 presents the values of the various IBFs we have considered for this data. Note that the lognormal model is preferred over the Weibull model by about 1.4:1, but the exponential model is the clear favorite; this is Ockham's razor in operation. (For those who prefer thinking in terms of posterior probabilities, note that, when the prior probabilities of the models are equal and the arithmetic IBF with reference priors is used, the posterior probabilities given by (1.1) are $P(M_1|\boldsymbol{x}) = 0.613$, $P(M_2|\boldsymbol{x}) = 0.227$, and $P(M_3|\boldsymbol{x}) = 0.160$.)

The only particularly odd IBF in Table 1 is B_{23}^{AI}, with the Jeffreys prior, which is far from $1/B_{32}^{AI}$. This illustrates the possible inconsistency of B_{ji}^{AI} in nonnested models. Similarly, desirable relationships such as $B_{21}^{AI}/B_{31}^{AI} = B_{23}^{AI}$ do not typically hold (but see Section 5).

The IBFs in Table 1 that are based on the reference priors seem to be sensible and quite consistent across model comparisons. This appears to be a common phenomenon, and leads us to recommend using reference π_i^N rather than Jeffreys noninformative priors.

3.2 The Encompassing Model Approach

Except in nice situations, such as Example 4, B_{21}^{AI} or its partially trimmed versions still suffer from the potential ambiguity of requiring M_2 to be the more complex model. Often, it is simply not clear that there is a more complex model, and then different answers can result from arbitrary choice of one as more complex. Note that this problem never arises with B_{21}^{GI}, so that the following applies mainly to B_{21}^{AI}.

One solution to this problem is to embed M_1 and M_2 in a larger model. More generally, if $\{M_i\}$ is a collection of models being considered, then we call M_0 an *encompassing model* if all the M_i are nested within M_0. Ideally, M_0 would be chosen in a minimal way, but this is not absolutely necessary.

If M_0 is an encompassing model, then one can compute the B_{0i}^{AI} using the nested model definition. The intrinsic Bayes factor of M_j to M_i can then be defined as

$$B_{ji}^{0AI} = B_{0i}^{AI}/B_{0j}^{AI} = B_{ji}^{N} \cdot (\bar{B}_{i0}^{N}/\bar{B}_{j0}^{N}), \tag{3.9}$$

where $\bar{B}_{i0}^{N}$ and $\bar{B}_{j0}^{N}$ are defined as in (3.4). (Note, from the second expression in (3.9), that $m_0^N(\boldsymbol{x})$ for the full data need not be computed.)

Use of B_{ji}^{0AI} eliminates the issue of defining the "more complex" model, since $B_{ji}^{0AI} = 1/B_{ij}^{0AI}$; indeed (3.9) will provide multiple model consistency, as will be discussed in Section 5. And, since only nested model computations are involved, stability is not an issue, and adjustments such as trimming are typically unnecessary.

The disadvantages of using (3.9) are that one must be able to determine a (minimal) encompassing model M_0 (which itself will not necessarily be unique and which may intro-

duce identifiability problems; see the rejoinder of Gelfand, Dey, and Chang, 1992), and that the training sample size will now generally be larger (so as to assure that $m_0(\boldsymbol{x}(\ell)) < \infty$), which will typically make the $B_{i0}^N(\boldsymbol{x}(\ell))$ more expensive to compute. (In the multiple model scenario, there are, however, substantial computational advantages to (3.9); see Section 5.)

Example 5 (continued). Here the natural encompassing model is M_0: $\mathcal{N}(\mu, \sigma^2)$, with corresponding noninformative (formal Jeffreys) prior $\pi_1^N(\mu, \sigma) = 1/\sigma^2$. (Again, we use this prior, rather than the more common $1/\sigma$, for computational convenience.) A minimal training sample, $\boldsymbol{x}(\ell)$, consists of two observations $\{x_i, x_j\}$, and computation yields

$$m_0^N(\boldsymbol{x}(\ell)) = \frac{1}{\sqrt{\pi}(x_i - x_j)^2},\ m_1^N(\boldsymbol{x}(\ell)) = \frac{\exp\{-(x_i - x_j)^2/4\}}{2\sqrt{\pi}},\ m_2^N(\boldsymbol{x}(\ell)) = \frac{1}{2\pi(x_i^2 + x_j^2)}. \tag{3.10}$$

It follows that (see also Example 1)

$$\begin{aligned} \bar{B}_{10}^N &= \frac{1}{n(n-1)} \sum_{i<j} (x_i - x_j)^2 \exp\{-(x_i - x_j)^2/4\}, \\ \bar{B}_{20}^N &= \frac{1}{n(n-1)} \sum_{i<j} \frac{1}{\sqrt{\pi}} (x_i - x_j)^2/(x_i^2 + x_j^2). \end{aligned} \tag{3.11}$$

Then (3.9) defines B_{21}^{0AI}.

The encompassing model approach would appear to leave B_{21}^{GI} unchanged because, in the analogous expression, $B_{ji}^{0GI} = B_{0i}^{GI}/B_{0j}^{GI}$, all the factors involving M_0 would cancel. This is not quite true, however, because now the minimal training samples would be the typically larger samples needed to make $m_0(\boldsymbol{x}(\ell)) < \infty$. Thus B_{ji}^{0GI} will denote the geometric intrinsic Bayes factor, but with the training samples chosen to be minimal under the encompassing model.

It is a delicate question whether the larger training sample induced by the encompassing model is beneficial or harmful. A larger training sample provides more stability, but seems to correspond to somewhat less attractive proper Bayes factors (an issue that will be explored elsewhere). For multiple models, however, the pragmatic advantages of the encompassing model approach will be seen to be very considerable. We conclude this section with such a multiple model example.

Example 7 (Hald's Regression data). This classic data set (cf, Zellner, 1984) is typically analyzed using normal regression models. There are four potential regressors, which we denote by X_1, X_2, X_3, and X_4, and a possible constant term, which as a regressor we denote by $X_0 = 1$.

Suppose it is desired to compare the models M_1: $\{X_0, X_1, X_2\}$, M_2: $\{X_0, X_3, X_4\}$, and M_3: $\{X_0, X_1, X_4\}$, which are the standard normal regression models with the indicated regressors. This data set is an extreme "test" because of the very small sample size ($n = 13$), and the fact that the design matrix is nearly singular.

The obvious encompassing model here is M_0: $\{X_0, X_1, X_2, X_3, X_4\}$. Computation of the B_{0i}^{AI} and B_{0i}^{GI} is then relatively straightforward. Indeed, Appendix 2 presents the needed formulas for *any* normal linear model. As indicated there, the minimal training samples would be of size 6 in this example; and there are a total of 1715 such training samples. (This was small enough that we did not need to choose a subsample of training samples to do the computations.) Table 2 gives B_{0i}^{AI} and B_{0i}^{GI} for reference priors and the Jeffreys priors (see Appendix 2). For comparison purposes, Table 2 also gives the asymptotic (Schwarz) Bayes factor, and the P-value for testing H_0: M_i versus H_1: M_0.

Table 2. Hald's data; comparison of M_0 to M_1, M_2, and M_3

	M_0 versus M_1	M_0 versus M_2	M_0 versus M_3
P-value	0.47	0.0055	0.168
Schwarz	0.080	130.6	0.450
B_{0i}^{AI}, reference priors	0.18	13.1	0.458
B_{0i}^{AI}, Jeffreys prior	0.16	34.2	0.411
B_{0i}^{GI}, reference priors	0.082	4.60	0.201
B_{0i}^{GI}, Jeffreys priors	0.004	0.265	0.001

Note, first, the strange values of B_{0i}^{GI} with the Jeffreys priors. We have observed this in other linear model examples, and hence do not recommend the combination of the two.

The values of B_{0i}^{AI} are reasonable, and rather stable with respect to the choice of noninformative prior. The B_{0i}^{GI} for reference priors seem somewhat small, especially for

M_0 versus M_2, perhaps suggesting a somewhat excessive favoritism of B_{0i}^{GI} towards simpler models.

The asymptotic (Schwarz) Bayes factors are not unreasonable, except perhaps for the M_0 versus M_2 comparison. (Recall, we asserted that, by ignoring the term from the prior, the asymptotic answers tend to overly favor the complex model; here, M_0.) Similarly, the P-value for testing $H_0 : M_2$ versus $H_1 : M_0$ is 0.0055, which would seem to be very strong evidence for M_0; the intrinsic Bayes factors suggest, however, that the evidence is only moderate, an example of the well-known conflict between P-values and Bayes factors.

Finally, recall that the original goal was comparison among M_1, M_2, and M_3. From (3.9) and Table 2 for, say, reference priors, one obtains $B_{12}^{0AI} = 72.8$, $B_{13}^{0AI} = 2.54$, and $B_{23}^{0AI} = 0.035$. (Interestingly, these are reasonably close to the reference prior $B_{12}^{0GI} = 56.1$, $B_{13}^{0GI} = 2.45$, and $B_{23}^{0GI} = 0.044$.) Thus M_1 is moderately preferred to M_3 and quite strongly preferred to M_2.

3.3 Expected Intrinsic Bayes Factors

For nonnested models, there is no obvious analogue of B_{21}^{EAI} in (2.14) or B_{21}^{EGI} in (2.15). This is because, in replacing the averages by expectations, one does not know whether to take the expectations under M_1 or M_2.

Another benefit of the encompassing model formulation is that the expectations can be taken under M_0; as before, expectations under M_0 will approximately equal those under M_1 or M_2. We thus define (switching to the multiple model notation for later use)

$$B_{ji}^{E0AI} = \frac{B_{0i}^{EAI}}{B_{0j}^{EAI}}, \quad B_{ji}^{E0GI} = \frac{B_{0i}^{EGI}}{B_{0j}^{EGI}}. \tag{3.12}$$

Example 5 (continued). Computation yields

$$\begin{aligned} B_{01}^{EAI} &= B_{01}^{N} \cdot E_{(\hat{\mu},\hat{\sigma})}^{M_0}\left[\frac{1}{2}(X_i - X_j)^2 e^{-(X_i - X_j)^2/4}\right] = B_{01}^{N} \cdot \frac{\hat{\sigma}^2}{(\hat{\sigma}^2 + 1)}, \\ B_{02}^{EAI} &= B_{02}^{N} \cdot E_{(\hat{\mu},\hat{\sigma})}^{M_0}\left[\frac{(X_i - X_j)^2}{2\sqrt{\pi}(X_i^2 + X_j^2)}\right] = B_{02}^{N} \cdot \frac{(1 - \exp\{-\hat{\mu}^2/\hat{\sigma}^2\})}{2\sqrt{\pi}(\hat{\mu}^2/\hat{\sigma}^2)}, \end{aligned} \tag{3.13}$$

where $\hat{\mu} = \bar{x}$ and $\hat{\sigma}^2 = \Sigma(x_i - \bar{x})^2/n$. Thus

$$B_{21}^{E0AI} = \frac{B_{01}^{EAI}}{B_{02}^{EAI}} = B_{21}^{N} \cdot \frac{2\sqrt{\pi}\hat{\mu}^2}{(\hat{\sigma}^2 + 1)(1 - \exp\{-\hat{\mu}^2/\hat{\sigma}^2\})}. \tag{3.14}$$

Expressions for B_{21}^{E0GI} involve infinite series, and so are omitted here.

4. INTRINSIC PRIORS

4.1 Definition and Motivation.

In Example 1 (continued) in Section 2.4, we saw that B_{21}^{AI} and B_{21}^{EAI} were approximately equal to Bayes factors for the proper (conditional) prior (2.27). Such a prior, if it exists, is called an *intrinsic prior.* We view the fact that IBFs tend to correspond to actual Bayes factors w.r.t. (sensible) intrinsic priors to be their strongest justification. Hence, determination of the intrinsic priors is of inherent theoretical interest, as well as providing the best insight into the behavior of IBFs.

There are also potential practical benefits in determining intrinsic priors. One obvious benefit is that the intrinsic priors could themselves be used, in place of the π_i^N, to compute actual Bayes factors. This would eliminate the need for training sample computations and eliminate concerns about stability of the IBFs. Indeed, one could alternatively view the IBF procedure as a method to apply to "imaginary training samples," so as to determine actual conventional priors to be used for model selection and hypothesis testing. This could be viewed as the complement to, say, the reference prior theory (Bernardo, 1979; Berger and Bernardo, 1992), which also uses imaginary samples to develop conventional priors for estimation and related problems.

While this latter view of the IBF methodology has considerable philosophical appeal, there are pragmatic arguments against actually operating in this fashion. Foremost among these arguments is that it is often very difficult to determine intrinsic priors. In contrast, IBFs are typically extremely easy to determine.

The argument could be made that, for important and frequently used models, it is worthwhile to determine default priors that can become "packaged" with the models. This argument is quite strong for default priors in estimation and other model-based inference, since each model can have its associated default prior. For model selection, however, the intrinsic prior will typically depend on the *pair* of models being considered, and so intrinsic priors cannot be "packaged" with individual models. This will inhibit their routine use, except for situations in which the intrinsic prior can be derived in advance for an entire

class of model comparisons. One important situation in which this can be done is linear models; see Berger and Pericchi (1994).

As with reference priors, our definition of intrinsic priors will center around use of an (asymptotic) imaginary training sample. Frequently, an asymptotic argument can be used to verify the following approximation, which we state as a condition.

Condition A. Suppose that, as the sample size goes to infinity and for priors in an appropriate class, Γ, (1.2) can be approximated by

$$B_{ji} = B_{ji}^N \cdot \frac{\pi_j(\hat{\boldsymbol{\theta}}_j)\pi_i^N(\hat{\boldsymbol{\theta}}_i)}{\pi_j^N(\hat{\boldsymbol{\theta}}_j)\pi_i(\hat{\boldsymbol{\theta}}_i)} \cdot (1 + o(1)), \tag{4.1}$$

where $\hat{\boldsymbol{\theta}}_j$ and $\hat{\boldsymbol{\theta}}_i$ are the MLEs under M_j and M_i, respectively. (Here, $o(1) \to 0$ in, say, probability under M_j and M_i. Since we are only using this as a heuristic to motivate the definition of intrinsic priors, a precise statement of the condition is not needed.)

This condition can easily be seen to hold in the standard asymptotic situation of (1.11) but also holds in nonstandard situations such as that of Example 3 (see (2.10) and (2.13)).

To define intrinsic priors we begin by equating (4.1) with (2.5) or (2.6), yielding the equation

$$\frac{\pi_j(\hat{\boldsymbol{\theta}}_j)\pi_i^N(\hat{\boldsymbol{\theta}}_i)}{\pi_j^N(\hat{\boldsymbol{\theta}}_j)\pi_i(\hat{\boldsymbol{\theta}}_i)}(1 + o(1)) = \tilde{B}_{ij}^N, \tag{4.2}$$

where we define $\tilde{B}_{ij}^N$ to be either the arithmetic or geometric average of the $B_{ij}^N(\boldsymbol{x}(\ell))$. We next need to make some assumptions about the limiting behavior of the quantities in (4.2). The following is typically satisfied.

Condition B. As the sample size grows to infinity, the following hold:

(i) Under M_j, $\hat{\boldsymbol{\theta}}_j \to \boldsymbol{\theta}_j$, $\hat{\boldsymbol{\theta}}_i \to \psi_i(\boldsymbol{\theta}_j)$, and $\tilde{B}_{ij}^N \to B_j^*(\boldsymbol{\theta}_j)$.

(ii) Under M_i, $\hat{\boldsymbol{\theta}}_i \to \boldsymbol{\theta}_i$, $\hat{\boldsymbol{\theta}}_j \to \psi_j(\boldsymbol{\theta}_i)$, and $\tilde{B}_{ij}^N \to B_i^*(\boldsymbol{\theta}_i)$.

Typically, for $k = i$ or $k = j$,

$$B_k^*(\boldsymbol{\theta}_k) = \begin{cases} \lim\limits_{L\to\infty} \; E_{\boldsymbol{\theta}_k}^{M_k}\left[\frac{1}{L}\sum\limits_{\ell=1}^{L} B_{ij}^N(\boldsymbol{X}(\ell))\right] & \text{arithmetic case} \\ \lim\limits_{L\to\infty} \; \exp\left\{E_{\boldsymbol{\theta}_k}^{M_k}\left[\frac{1}{L}\sum\limits_{\ell=1}^{L} \log B_{ij}^N(\boldsymbol{X}(\ell))\right]\right\} & \text{geometric case;} \end{cases} \tag{4.3}$$

if the $\boldsymbol{X}(\ell)$ are exchangeable, then the limits and averages over L above can be removed.

Using Condition B, and passing to the limit in (4.2), first under M_j and then under M_i, results in the following two equations which define the *intrinsic prior* (π_j^I, π_i^I)

$$\frac{\pi_j^I(\boldsymbol{\theta}_j)\pi_i^N(\psi_i(\boldsymbol{\theta}_j))}{\pi_j^N(\boldsymbol{\theta}_j)\pi_i^I(\psi_i(\boldsymbol{\theta}_j))} = B_j^*(\boldsymbol{\theta}_j), \tag{4.4}$$

$$\frac{\pi_j^I(\psi_j(\boldsymbol{\theta}_i))\pi_i^N(\boldsymbol{\theta}_i)}{\pi_j^N(\psi_j(\boldsymbol{\theta}_i))\pi_i^I(\boldsymbol{\theta}_i)} = B_i^*(\boldsymbol{\theta}_i). \tag{4.5}$$

The motivation, again, is that priors which satisfy (4.4) and (4.5) would yield answers which are asymptotically equivalent to use of the intrinsic Bayes factors. We note that solutions are not necessarily unique, nor necessarily proper.

As a simple example, we have encountered situations in which $B_j^*(\boldsymbol{\theta}_j) = B_i^*(\boldsymbol{\theta}_i) = 1$ (e.g., Example 4 and Example 6 for lognormal versus Weibull with reference priors). It follows trivially that solutions to (4.4) and (4.5) are then

$$\pi_k^I(\boldsymbol{\theta}_k) = \pi_k^N(\boldsymbol{\theta}_k), \quad k = i, j. \tag{4.6}$$

Thus the intrinsic priors are merely the original noninformative priors. (While this may seem uninteresting, we argue in Section 4.4 that there is a very important "calibration" of π_i^N and π_j^N that is occurring here.)

4.2 Intrinsic Priors for Nested Models

In the nested model scenario of Section 2.1 and under Assumption N, solutions to (4.4) and (4.5) are trivially given by

$$\pi_1^I(\boldsymbol{\theta}_1) = \pi_1^N(\boldsymbol{\theta}_1), \qquad \pi_2^I(\boldsymbol{\theta}_2) = \pi_2^N(\boldsymbol{\theta}_2)B_2^*(\boldsymbol{\theta}_2). \tag{4.7}$$

Typically there are also many other solutions, perhaps even solutions that are proper distributions, but the solutions in (4.7) are the simplest.

Example 1 (continued). For B_{21}^{AI}, it follows from (2.16) and (4.3) that $B_2^*(\boldsymbol{\theta}_2) = \sigma_2 \cdot \pi_2^I(\mu|\sigma_2)$, where $\pi_2^I(\mu|\sigma_2)$ was defined in (2.27). Hence the intrinsic prior is

$$\begin{aligned} \pi_1^I(\sigma_1) &= \pi_1^N(\sigma_1) = 1/\sigma_1, \\ \pi_2^I(\mu, \sigma_2) &= \pi_2^N(\sigma_2)B_2^*(\mu, \sigma_2) = \frac{1}{\sigma_2} \cdot \pi_2^I(\mu|\sigma_2). \end{aligned} \tag{4.8}$$

Thus B_{21}^{AI} behaves (asymptotically) like the actual Bayes factor which uses reference non-informative priors for σ_1 and σ_2, and the proper $\pi_2^I(\mu|\sigma_2)$ for the conditional prior of μ given σ_2. Besides the propriety of $\pi_2^I(\mu|\sigma_2)$, it is also notable that the intrinsic prior for σ_2 is the reference prior $1/\sigma_2$, and not the (formal) Jeffreys prior $1/\sigma_2^2$ that was used to derive B_{21}^{AI}. We have observed this latter behavior in other examples also; the IBFs seem to try to convert the original π_i^N into reference priors for common, or similar, model parameters.

Example 3 (continued). For B_{21}^{AI}, we see from (2.11), (2.18), and (4.3) that $B_2^*(\theta_2) = \Phi(-\theta_2/\sqrt{2})$. Hence, (4.7) becomes

$$\pi_1^I(\theta_1) = \pi_1^N(\theta_1) = 1, \quad \pi_2^I(\theta_2) = 1 \cdot \Phi(-\theta_2/\sqrt{2}). \tag{4.9}$$

Two features of this intrinsic prior are of particular interest. First, on $(-\infty, 0)$, π_1^I and π_2^I are not even proportional. Recall we wrote the models as M_1: $\theta_1 < 0$, M_2: $\theta_2 \in \mathbf{R}^1$, as opposed to M_1: $\theta < 0$, M_2: $\theta \in \mathbf{R}^1$, to emphasize that the θ_i could have differing interpretations under each model and different priors, even on their common domain. This possibility appears to have been realized. Note, however, that, on $(-\infty, 0)$, π_1^I and π_2^I differ substantially only near zero.

The second interesting feature of the intrinsic prior is that

$$\int_0^\infty \pi_2^I(\theta_2)d\theta_2 = \int_0^\infty \Phi(-\theta_2/\sqrt{2})d\theta_2 = \frac{1}{\sqrt{\pi}}. \tag{4.10}$$

Hence, $\pi_2^I(\theta_2|\{\theta_2 > 0\})$ is proper.

The behavior of intrinsic priors that was observed in the above examples seems typical for nested models. "Common" parameters (or, at least, parameters that can be identified in the sense of (2.1)) typically have intrinsic priors that are standard noninformative priors or slight variants, while parameters that occur only in the more complex model (or that have extended domains in the more complex model) have (conditional) proper intrinsic priors. This corresponds with intuition and standard practice.

For nested problems in which (4.3) holds (which is typically the case), the above observations can be formalized in a quite interesting fashion. First, we give a key theorem.

Theorem 1. *For the arithmetic IBF, suppose that (4.3) holds and that $\pi_1^N(\boldsymbol{\theta}_1)$ is proper. Then $\pi_2^I(\boldsymbol{\theta}_2)$, defined in (4.7), is also proper.*

Proof. Because the limit is assumed to exist in (4.3), it is true that

$$\begin{aligned}
\int \pi_2^I(\boldsymbol{\theta}_2)d\boldsymbol{\theta}_2 &= \int \pi_2^N(\boldsymbol{\theta}_2)\left(\lim_{L\to\infty} E_{\boldsymbol{\theta}_2}^{M_2}\left[\frac{1}{L}\sum_{\ell=1}^{L} B_{12}^N(\boldsymbol{X}(\ell))\right]\right)d\boldsymbol{\theta}_2 \\
&= \lim_{L\to\infty}\frac{1}{L}\sum_{\ell=1}^{L}\int\int \pi_2^N(\boldsymbol{\theta}_2)f_2(\boldsymbol{x}(\ell)|\boldsymbol{\theta}_2)B_{12}^N(\boldsymbol{x}(\ell))d\boldsymbol{x}(\ell)d\boldsymbol{\theta}_2 \\
&= \lim_{L\to\infty}\frac{1}{L}\sum_{\ell=1}^{L}\int m_2^N(\boldsymbol{x}(\ell))\cdot[m_1^N(\boldsymbol{x}(\ell))/m_2^N(\boldsymbol{x}(\ell))]d\boldsymbol{x}(\ell) \\
&= \lim_{L\to\infty}\frac{1}{L}\sum_{\ell=1}^{L}(1) = 1,
\end{aligned}$$

the last line following from the fact that, if π_1^N is proper, then so is m_1^N. □

When $\pi_1^N(\boldsymbol{\theta}_1)$ is improper, one can consider a sequence of compact subsets of Θ_1 that converge to Θ_1, and obtain the intrinsic Bayes factors and intrinsic priors corresponding to each subset. With normalization, these priors will typically be proper (using Theorem 1), and the subset IBFs will typically converge to B_{21}^{AI}. We will formalize these ideas elsewhere, noting here simply that intrinsic priors are typically proper or unique limits of a sequence of proper priors. Note, also, that if a weighted average of the $B_{12}^N(\boldsymbol{x}(\ell))$ was used to define B_{21}^{AI}, then Theorem 1 would still remain valid; see de Vos (1993) for a use of weighted averaging.

As a final observation, note that there is no analogue of Theorem 1 for B_{21}^{GI}. Indeed, we saw in Section 2.4 that B_{21}^{GI} (equivalent to B_{21}^{EGI} in terms of intrinsic priors, because of (4.3)) does not seem to correspond to a Bayes factor with proper intrinsic priors.

4.3 Intrinsic Priors in Nonnested Models

For nonnested models, finding a solution to (4.4) and (4.5) is often more difficult.

Example 6 (continued.) Consider comparison of the nonnested models M_1: Exponential (θ_1) and M_2: Lognormal (μ, σ). It is easy to verify Condition A for "nice" priors. Condition

B is also true, where:

$$\text{under } M_1, \quad \hat{\boldsymbol{\theta}}_2 = (\hat{\mu}, \hat{\sigma}) = (\overline{y}, (S_y^2/n)^{1/2})$$
$$\stackrel{(n\to\infty)}{\longrightarrow} (E_{\theta_1}^{M_1}[\overline{Y}], (\frac{1}{n} E_{\theta_1}^{M_1}[S_y^2])^{1/2})$$
$$\stackrel{(n\to\infty)}{\longrightarrow} \psi_2(\theta_1) \equiv (\log\theta_1 - 0.5772, 1.2825); \tag{4.11}$$

$$\text{under } M_2, \quad \hat{\theta}_1 = \overline{x} \to \psi_1(\mu,\sigma) = E_{(\mu,\sigma)}^{M_2}[\overline{X}] = \exp\{\mu + \tfrac{1}{2}\sigma^2\}. \tag{4.12}$$

It follows that (4.3) becomes

$$B_1^* = \begin{cases} E_{\theta_1}^{M_1}\left[\frac{2X_iX_j|\log(X_i/X_j)|}{(X_i+X_j)^2}\right] & \text{arithmetic case} \\ \exp\left\{E_{\theta_1}^{M_1}\left[\log\left(\frac{2X_iX_j|\log(X_i/X_j)|}{(X_i+X_j)^2}\right)\right]\right\} & \text{geometric case} \end{cases}$$
$$= \begin{cases} 0.2954 & \text{arithmetic case} \\ 0.2383 & \text{geometric case;} \end{cases} \tag{4.13}$$

$$B_2^* = \begin{cases} E_{(\mu,\sigma)}^{M_2}\left[\frac{2X_iX_j|\log(X_i/X_j)|}{(X_i+X_j)^2}\right] & \text{arithmetic case} \\ \exp\left\{E_{(\mu,\sigma)}^{M_2}\left[\log\left(\frac{2X_iX_j|\log(X_i/X_j)|}{(X_i+X_j)^2}\right)\right]\right\} & \text{geometric case} \end{cases}$$
$$= \begin{cases} H^A(\sigma) \equiv E^Z\left[\frac{\sqrt{2}\sigma|Z|}{1+cosh(\sqrt{2}\sigma Z)}\right] & \text{arithmetic case} \\ H^G(\sigma) \equiv \frac{3\sigma}{2}\cdot\exp\left\{-2E^Z\left[\log\left(1+e^{\sqrt{2}\sigma Z}\right)\right]\right\} & \text{geometric case,} \end{cases} \tag{4.14}$$

where $Z \sim \mathcal{N}(0,1)$. (The derivations above are straightforward.)

For the arithmetic case, equations (4.4) and (4.5) thus become

$$\frac{\pi_2^I(\mu,\sigma)(1/\exp\{\mu+\frac{1}{2}\sigma^2\})}{(1/\sigma)\pi_1^I(\exp\{\mu+\frac{1}{2}\sigma^2\})} = H^A(\sigma), \tag{4.15}$$

$$\frac{\pi_2^I(\log\theta_1 - 0.5772, 1.2825)(1/\theta_1)}{(1/1.2825)\pi_1^I(\theta_1)} = (0.2954). \tag{4.16}$$

We have not attempted to characterize the solutions to (4.15) and (4.16) in general. The equations are fairly easy to solve, however, if one assumes that

$$\pi_2^I(\mu,\sigma) = \pi_{21}^I(\mu)\pi_{22}^I(\sigma). \tag{4.17}$$

Indeed, the solutions are then given (up to multiplication of π_1^I and division of π_2^I by an arbitrary positive constant) by

$$\pi_1^I(\theta_1) = 2/\theta_1^c$$
$$\pi_2^I(\mu,\sigma) = \frac{1}{2\sigma}H^A(\sigma)\exp\{(1-c)(\mu+\frac{1}{2}\sigma^2)\}, \tag{4.18}$$

where $c = 1.1291$. A similar analysis for the geometric IBF yields, as the intrinsic priors, the expressions in (4.18) with H^A replaced by H^G and $c = 1.2602$.

To obtain some insight into the behavior of these priors, it is useful to reparameterize M_2 by (ν, σ), where $\nu = \exp\{\mu + \sigma^2/2\}$ is the lognormal mean. Then

$$\pi_2^I(\mu, \sigma) \longrightarrow \frac{2}{\nu^c} \cdot \frac{H(\sigma)}{2\sigma},$$

where H is either H^A or H^G. The point of this transformation is that θ_1 and ν are then both the mean parameters of their respective distributions, and it is of considerable interest that they have the same intrinsic prior. Curiously, this prior is improper but is not the usual inverse noninformative prior. For some speculations as to why this is so, see the next section.

The "nuisance" parameter, σ, receives the prior $\pi_{22}^I(\sigma) = H(\sigma)/(2\sigma)$. It is easy to show that $\pi_{22}^I(\sigma)$ is monotonically decreasing, with the following limiting behavior:

$$\text{as } \sigma \to 0, \quad \pi_{22}^I(\sigma) \cong \begin{cases} 1/(2\sqrt{\pi}) & \text{arithmetic case,} \\ 3/16 & \text{geometric case,} \end{cases}$$

$$\text{as } \sigma \to \infty, \quad \pi_{22}^I(\sigma) \cong \begin{cases} 1/(\sqrt{\pi}\sigma^2) & \text{arithmetic case,} \\ \frac{3}{4}\exp(-2\sigma/\sqrt{\pi}) & \text{geometric case.} \end{cases}$$

It is thus clear that $\pi_{22}^I(\sigma)$ is integrable; indeed, we have normalized (4.18) so that, in the arithmetic case, $\pi_{22}^I(\sigma)$ is a proper density.

The pattern we have observed thus seems to be holding: for parameters that are in some sense "common," the intrinsic priors are the same and are of a noninformative type, while parameters that exist only in one of the models receive proper intrinsic priors.

4.4 Improper Intrinsic Priors and Matching Predictives

Our original goal was to develop an automatic Bayes factor that behaves similarly to sensible proper Bayes factors. For nested models, the examples and arguments in Section 4.2 suggest that this goal has been achieved by arithmetic IBFs. For nonnested models, however, IBFs seem to correspond to improper priors, and the extent to which the original goal has been met is unclear.

The "difficulty" in evaluating the situation for nonnested models is similar to the difficulty in dealing with "common" parameters in nested models. The discussion following

Theorem 1 indicated that this could be resolved, in the nested case, by taking limits of proper priors on the "common" parameters. A similar device would probably work in the nonnested case, but even then a certain degree of ambiguity will remain concerning whether the proper priors are appropriately "matched" across models.

An illuminating direct approach to this issue is to attempt to choose priors to match predictives. The underlying motivation is the foundational Bayesian view that one should concentrate on predictive distributions of observables; models and priors are, at best, convenient abstractions. According to this perspective, it is $m(\boldsymbol{y})$ that describes reality, where $\boldsymbol{y}$ is a variable of predictive interest. We can choose to represent $m(\boldsymbol{y})$ as $m_i(\boldsymbol{y}) = \int f_i(\boldsymbol{y}|\boldsymbol{\theta}_i)\pi_i(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i$, but (f_i, π_i) is thought of as merely a convenient abstraction.

From this perspective, if one is comparing models M_1: f_1 versus M_2: f_2, then the priors π_1 and π_2 should be chosen so that $m_1(\boldsymbol{y})$ and $m_2(\boldsymbol{y})$ are as close as possible. Thus we think of π_1 and π_2 as being properly calibrated if, when filtered through the models M_1 and M_2, they yield similar predictives. This could be assessed by defining some distance measure, $d(m_1, m_2)$, and calling π_1 and π_2 calibrated if $d(m_1, m_2)$ is small. A key issue in operationalizing this idea is that of choosing the variable $\boldsymbol{y}$ at which a predictive match is desired. It seems natural, in the exchangeable case, to choose $\boldsymbol{y}$ to be an "imaginary" minimal training sample; this is typically the smallest set of observations for which the various model parameters are identifiable. Keeping $\boldsymbol{y}$ minimal seems natural because, for the full data, we of course want m_1 and m_2 to discriminate between the models. We explore this formal approach elsewhere, here being content simply with showing that intrinsic priors seem to be well-calibrated in our examples.

The clearest examples of this predictive matching notion are Example 4 and Example 6 (lognormal versus Weibull case), where $m_i^N(\boldsymbol{y}) = m_j^N(\boldsymbol{y})$ for *any* minimal training sample, $\boldsymbol{y}$, and the reference priors $\pi_i^N = \pi_i^R$. Hence the reference priors seem to be completely calibrated in these situations. It is interesting that the formal Jeffreys prior is not completely calibrated here.

In the other examples considered in the paper, it is harder to establish the extent of the calibration of intrinsic priors. There are, however, intriguing suggestions of calibration: in Example 1, the intrinsic prior for σ_2 was the reference prior, not the original Jeffreys prior;

and, in Example 6 (continued) of Section 4.3, the intrinsic priors for the mean parameters seemed to be matched. (We suspect that this matched prior for the mean parameters is not the usual noninformative prior because $2/\nu^c$ may provide a better predictive match; we explore this elsewhere.)

The ideas here are related to ideas of elicitation through predictives (cf, Kadane, et. al, 1980). Also, similar uses of predictive matching to define priors for model selection can be found in Laud and Ibrahim (1992) and Garthwaite and Dickey (1992).

5. COMPARISON OF MULTIPLE MODELS AND PREDICTION

5.1 Multiple Model Coherency

If M_i, M_j, and M_k are three models under consideration, actual Bayes factors satisfy the conditions

$$B_{ij} = \frac{1}{B_{ji}}; \qquad \frac{B_{ij}}{B_{kj}} = B_{ik}. \tag{5.1}$$

For IBFs to satisfy (5.1), it is first important to ensure that minimal training samples are defined relative to all the models $\{M_1, M_2, \ldots, M_q\}$ simultaneously. It is then trivial to see that the geometric IBF always satisfies (5.1) (cf, (5.3)). This is a very appealing feature of B_{ij}^{GI}.

Arithmetic IBFs will typically not satisfy (5.1). An exception is when the encompassing model approach of Section 3.2 is used, in which case it is, again, trivial to verify (5.1). If it is desired to use the arithmetic IBF, but the encompassing model approach cannot be implemented, the following scheme can adjust the B_{ij}^{AI} so as to satisfy (5.1).

Step 1. Relabel, so that $\{M_1, M_2, \ldots, M_q\}$ are listed in order of increasing complexity.

Step 2. Compute B_{ij}^{AI} for all $i > j$. Note that the $m_k^N(\boldsymbol{x})$ and $m_k^N(\boldsymbol{x}(\ell))$ need be computed only once; determination of all B_{ij}^{AI} is then just algebra.

Step 3. Define $B_{ij}^{MAI} = m_i^*/m_j^*$, where

$$m_k^* = \left[\prod_{j=1}^{k-1} B_{kj}^{AI} / \prod_{j=k+1}^{q} B_{jk}^{AI} \right]^{1/p}. \tag{5.2}$$

It is obvious that the *multiple arithmetic intrinsic Bayes factors*, B_{ij}^{MAI}, then satisfy (5.1).

Example 6 (continued). M_1 (exponential), M_2 (lognormal), and M_3 (Weibull) are already ordered reasonably in terms of complexity. From Table 1 for the Jeffreys priors, we have that $B_{21}^{AI} = 0.37$, $B_{31}^{AI} = 0.25$, and $B_{32}^{AI} = 0.66$. Thus (5.2) yields $m_1^* = (B_{21}^{AI} \cdot B_{31}^{AI})^{-1/3} = ((0.37)(0.25))^{-1/3} = 2.21$. Similarly, $m_2^* = 0.82$ and $m_3^* = 0.55$. Then $B_{21}^{MAI} = m_2^*/m_1^* = 0.37$, $B_{31}^{MAI} = 0.25$, and $B_{32}^{MAI} = 0.67$. Thus the "coherency adjustment" is here very minor. The numbers in Table 1 for the reference priors happen to already be coherent (because the $B_{32}^N(\boldsymbol{x}(\ell))$ all equal 1).

A justification for B_{ij}^{MAI} is given by Lemma A2 in Appendix 1. Note that, computationally, use of B_{ij}^{MAI} is more complex than use of B_{ij}^{GI} if the number of models is large. This is because B_{ij}^{GI} can be rewritten as

$$B_{ij}^{GI} = \frac{m_i^N(\boldsymbol{x})}{[\prod_{\ell=1}^{L} m_i^N(\boldsymbol{x}(\ell))]^{1/L}} \cdot \frac{[\prod_{\ell=1}^{L} m_j^N(\boldsymbol{x}(\ell))]^{1/L}}{m_j^N(\boldsymbol{x})}, \tag{5.3}$$

so that one need only compute, say, the first factor in (5.3) for each of the p models, to determine all of the B_{ij}^{GI}. In contrast, use of (5.2) requires computation of all $q(q-1)/2$ of the B_{ij}^{AI} for $i > j$. (Note, however, that we are only discussing simple algebraic operations; often the most difficult step is computation of the $m_i^N(\boldsymbol{x})$ or $m_i^N(\boldsymbol{x}(\ell))$, and this difficulty is common to all IBFs.)

As a final computational note, observe that the encompassing model approach of Section 3.2 allows use of arithmetic IBFs while retaining the computational efficiency of geometric IBFs. This is clear from (3.9) (recalling that $B_{ji}^N = m_j^N(\boldsymbol{x})/m_i^N(\boldsymbol{x})$).

5.2 Posterior Probabilities

Because Bayes factors can easily be converted to posterior probabilities via (1.1), it would seem that determination of the B_{ij} suffices for a statistical analysis; readers could use their own model prior probabilities, p_i, to compute the $P(M_i|\boldsymbol{x})$, or could simply directly interpret the B_{ij}. For a default analysis, however, use of subjective p_i will often not be possible, and the entire collection of B_{ij} is too large to be digestible if the number of models, q, is large. Thus default choices of the p_i will often be used.

In many situations, the obvious default choice is $p_i = 1/q$, leading to the reporting of

$$P(M_i|\boldsymbol{x}) = 1/\sum_{j=1}^{q} B_{ji}. \tag{5.4}$$

It is somewhat amusing to note that then $B_{ij} = P(M_i|\boldsymbol{x})/P(M_j|\boldsymbol{x})$; thus, if one desires to SUTC the choice $p_i = 1/q$, one can instead present the $P(M_i|\boldsymbol{x})$ in (5.4) as the "relative model weightings" from which all Bayes factors can be reconstructed.

In nested model situations, other default choices of the p_i are often made. Suppose there are r_1 models of dimension k_1, r_2 of dimension $k_2, \ldots, r_s$ of dimension k_s. Then it is common to:

(i) assign a prior probability p_i^* to the class of models of dimension k_i;

(ii) give each model of dimension k_i equal prior probability p_i^*/r_i.

The most common choice of the p_i^* is $p_i^* = 1/s$, although Ockham's razor might suggest that decreasing choices, such as $p_i^* = i^{-1}/\sum_{j=1}^{s} j^{-1}$, are more reasonable.

Utilization of the above default choices can be important to counteract selection effects from searching among possible multitudes of submodels.

5.3 Prediction

Frequently, the ultimate goal of the statistical analysis is prediction of some variable $\boldsymbol{Y}$ which, under the model M_i and for given data $\boldsymbol{x}$, has density $g_i(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}_i)$. Then the predictive density of $\boldsymbol{Y}$, given $\boldsymbol{x}$, is

$$g(\boldsymbol{y}|\boldsymbol{x}) = \sum_{i=1}^{q} m_i^*(\boldsymbol{y}|\boldsymbol{x}) \cdot P(M_i|\boldsymbol{x}), \tag{5.5}$$

where $P(M_i|\boldsymbol{x})$ is given by (1.1) and

$$m_i^*(\boldsymbol{y}|\boldsymbol{x}) = \int g_i(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}_i)\pi_i(\boldsymbol{\theta}_i|\boldsymbol{x})d\boldsymbol{\theta}_i. \tag{5.6}$$

As mentioned in the introduction, one of the strengths of the Bayesian approach to model comparison is that it allows one to keep all models in the analysis, accounting for model uncertainty by weighting the effect of each model by its posterior probability, as in (5.5).

To utilize (5.5), we must determine the $P(M_i|\boldsymbol{x})$ and the $m_i^*(\boldsymbol{y}|\boldsymbol{x})$. We propose, as a default analysis, use of IBFs (and the prior probabilities in Section 5.2) to determine the $P(M_i|\boldsymbol{x})$, and use of the noninformative priors π_i^N (preferably, reference priors) to determine the $m_i^*(\boldsymbol{y}|\boldsymbol{x})$. (See Draper, 1994, for simpler — but cruder — approximations and other references.)

This proposal for default prediction has no glaring deficiencies; for instance, the "scaling" problem for improper π_i^N is irrelevant to computation of either $m_i^*(\boldsymbol{y}|\boldsymbol{x})$ or IBFs. There is, however, a subtle incoherency in the proposal. The incoherency arises from the fact that use of an IBF corresponds, at least roughly, to use of an intrinsic prior in computing $P(M_i|\boldsymbol{x})$, while $m_i^*(\boldsymbol{y}|\boldsymbol{x})$ would be computed using π_i^N. Hence the effective priors being used in each part of the analysis would differ.

An obvious coherent solution to the problem is to determine the intrinsic priors, and use them to compute the $m_i^*(\boldsymbol{y}|\boldsymbol{x})$ as well as the $P(M_i|\boldsymbol{x})$. While this "solution" deserves serious study, we do not recommend it at this time for two reasons. First, determination of intrinsic priors can be hard. Second, it is not clear that intrinsic priors are suitable for automatic use in computing the predictive distributions $m_i^*(\boldsymbol{y}|\boldsymbol{x})$. Use of the π_i^N for this purpose is known to be quite reasonable, but using the, often proper, intrinsic priors might well lead to an undesirable biasing of the $m_i^*(\boldsymbol{y}|\boldsymbol{x})$. To put this another way, we are forced to use (essentially) proper default priors to determine the $P(M_i|\boldsymbol{x})$, but such priors are typically considerably less robust than the π_i^N, so we turn to the latter to determine the $m_i^*(\boldsymbol{y}|\boldsymbol{x})$. Note, however, that, asymptotically, the incoherence of our suggestion disappears, since $m_i^*(\boldsymbol{y}|\boldsymbol{x})$ does not depend on the prior asymptotically.

6. CONCLUSIONS AND RECOMMENDATIONS

It is worthwhile to summarize the advantages and disadvantages of IBFs.

Advantages of IBFs:

1. They are completely automatic Bayes factors, in that they are based only on the data and standard noninformative priors. Note, however, that issues such as the "optimal" choice of training samples in dependent data situations are yet to be resolved.

2. They seem to correspond to actual Bayes factors for reasonable "intrinsic priors,"

thus attaining a type of "second order" Bayesian correspondence; in contrast, most other default methods achieve (at best) a first order correspondence with Bayesian methods, with many having a systematic bias in favor of the more complex model. Compared with other "second order" Bayesian methods, such as that of Jeffreys, IBFs have the advantage of being very generally applicable. (They also have the somewhat cynical advantage that they SUTC the choice of the default prior.)

3. IBFs apply to non-nested, as well as nested, model comparisons, and can be applied to any distributions.

4. They can be used for default Bayesian hypothesis testing, as well as model comparison.

5. They can be applied in situations in which even the usual Bayesian asymptotics (e.g., BIC) does not apply.

6. They can be used for default multiple model comparison and for default prediction in the face of model uncertainty.

7. They are invariant to univariate transformations of the data. If suitably invariant noninformative prior distributions are used, they are also invariant to choice of the parameterizations of the models.

Disadvantages of IBFs:

1. They can be computationally intensive. Recall, however, that the training sample adjustment factors are often available in closed form (such as when comparing normal linear models), and sampling from the collection of training samples can reduce the computational problem to a very manageable level.

2. The arithmetic IBFs may require adjustments to be coherent across multiple models.

3. The standard IBFs can be unstable if the sample size is small. At the extreme, when the sample size is only slightly larger than the size of a minimal training sample, the standard IBFs should probably not be used. However, the expected IBFs can still be used in many situations, regardless of the sample size; and, if the intrinsic priors can be found,

they can be used as priors in an ordinary Bayes factor computation.

4. IBFs are not invariant to multivariate transformations of the data that alter the nature of minimal training samples. Because IBFs average over all minimal training samples, however, the effect of multivariate transformations will be mitigated.

5. IBFs will be formally incoherent in a variety of ways, as are other default Bayesian methodologies. The standard IBFs (but not the expected IBFs or intrinsic prior Bayes factors) even have certain incoherent attributes that are not typical of default Bayesian methods, such as possible violation of sufficiency. While not pleased with these incoherencies, we feel that they tend to have a very minor effect in practice, and are the price that must be paid for performing a sensible default analysis.

Recommendations:

We have proposed a variety of IBFs in the paper; B^{AI}, B^{EAI}, $B^{\alpha AI}$, B^{0AI}, B^{MAI}, as well as their geometric analogues and Bayes factors arising from intrinsic priors. This is probably too large a collection, in that some subset of them will probably suffice to handle most practical situations. Considerably more practical experience (and perhaps theoretical investigation) is going to be necessary before a final set of IBFs can be definitively recommended, however, and these are the candidates that we feel should be studied. Note that we have already considered a multitude of other IBFs, and the above list is a refinement of the original huge list of possibilities.

Even though we feel that much study remains to be done among these "finalist" IBFs, we can give some tentative recommendations concerning their use in practice.

1. We recommend using reference noninformative priors (cf, Berger and Bernardo, 1992) to compute the IBFs. For large data sets, the effect of the initial noninformative priors is probably minor, but in small data sets we have found that reference priors seem to give the most stable and reasonable answers. Also, intrinsic priors for "common" parameters in the models tend to be reference priors, rather than, say, Jeffreys priors.

2. In general, we prefer the behavior of the arithmetic IBFs to that of the geometric IBFs. This came as a surprise and a disappointment, because geometric IBFs are considerably more appealing at first sight; for instance, they automatically combine across multiple

models as Bayes factors should, with resulting intuitive and computational advantages (see, also, Good, 1985). We found, however, that geometric IBFs seem to be less stable than the recommended arithmetic versions. Furthermore, it is the arithmetic IBFs that appear to correspond to actual Bayes factors with respect to intrinsic priors. This is not to say that geometric IBFs are bad, and our conclusions here are, admittedly, tentative. Indeed, Pericchi and Smith (1994) shows that geometric IBFs yield optimal model weights under a prequential type of utility function.

3. For comparing two nested models, we recommend B^{AI} or B^{EAI}, the latter being particularly recommended if the sample size is small.

4. For normal linear models, we recommend using the encompassing approach with B^{0AI}; for smaller sample sizes, the expectation versions are preferable (see Berger and Pericchi, 1994). Note that intrinsic priors are also available for linear models, and could be used directly to compute ordinary Bayes factors.

5. For other non-nested models, we have less clear recommendations. Arithmetic IBFs, with moderate trimming and multiple model adjustment, seem to work fine. Geometric IBFs are probably preferable to untrimmed and/or unadjusted arithmetic IBFs. And the encompassing approach is always appealing if it is easy to determine an encompassing model.

Acknowledgements. Many people provided helpful input into the development of these ideas. We are particularly grateful to Jacek Dmochowski, Paul Garthwaite, Alan Gelfand, J. K. Ghosh, Prakash Laud, Eglée Pérez, J. M. Pérez, Tom Sellke, Julia Varshavsky, and referees for an earlier version of the paper.

APPENDIX 1. Technical Lemmas

Lemma A1. *If X_1 and X_2 are independent observations from the location-scale density (w.r.t. Lebesgue measure) $\sigma^{-1}g((x_i-\mu)/\sigma)$ and $\pi^N(\mu,\sigma)=1/\sigma$, then, for $x_1 \neq x_2$,*

$$m(x_1,x_2)=\int_0^\infty \int_{-\infty}^\infty \frac{1}{\sigma^3} g(\frac{x_1-\mu}{\sigma}) g(\frac{x_2-\mu}{\sigma}) d\mu d\sigma = \frac{1}{2|x_1-x_2|}.$$

Proof. Assume, w.l.o.g., that $x_2 > x_1$, and make the change of variables $(\mu,\sigma) \to (v,w) \equiv$

$((x_1 - \mu)/\sigma, (x_2 - \mu)/\sigma)$. Then $m(x_1, x_2)$ becomes

$$\begin{aligned} m(x_1, x_2) &= \frac{1}{|x_1 - x_2|} \int_{-\infty}^{\infty} \int_{v}^{\infty} g(v)g(w)dwdv \\ &= \frac{1}{|x_1 - x_2|} \cdot P(V < W), \end{aligned}$$

where V and W are independent with density $g(\cdot)$. Clearly $P(V < W) = P(W < V) = 1/2$, completing the proof. □

Lemma A2. *The multiple intrinsic Bayes factors $B_{ij}^{MAI} = m_i^*/m_j^*$, where the $\{m_k^*\}$ are defined by (5.2), provide the best fit to the raw B_{ij}^{AI}, subject to nonnegativity and the coherency condition (5.1), when fit is measured by*

$$\sum_{i=2}^{q} \sum_{j=1}^{i-1} [\log(B_{ij}/B_{ij}^{AI})]^2.$$

(Measuring fit on a log-scale is natural for Bayes factors; cf, Good, 1985.)

Proof. Define $t_k = \log m_k^*$ and $c_{ij} = \log B_{ij}^{AI}$. Then we seek to minimize

$$\sum_{i=2}^{q} \sum_{j=1}^{i-1} (t_i - t_j - c_{ij})^2,$$

over choice of the $\{t_k\}$. Differentiating w.r.t. t_k and setting the result equal to zero yields

$$t_k = \bar{t} + \frac{1}{q} \left[\sum_{j=1}^{k-1} c_{kj} - \sum_{j=k+1}^{q} c_{jk} \right].$$

Thus

$$m_k^* = e^{t_k} = e^{\bar{t}} \left[\prod_{j=1}^{k-1} B_{kj}^{AI} / \prod_{j=k+1}^{q} B_{jk}^{AI} \right]^{1/q}.$$

Noting that the multiplicative constant $\exp\{\bar{t}\}$ is irrelevant to the definition of $B_{ij}^{MAI} = m_i^*/m_j^*$, the result is immediate. □

APPENDIX 2. Linear Models

Suppose, for $i = 1, \ldots, q$, that model M_i is the linear model

$$M_i\colon\ \boldsymbol{Y} = \boldsymbol{X}_i\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i, \quad \boldsymbol{\varepsilon}_i \sim \mathcal{N}_n(\boldsymbol{0}, \sigma_i^2 \boldsymbol{I}_n),$$

where σ_i^2 and $\boldsymbol{\beta}_i = (\beta_{i1}, \beta_{i2}, \ldots, \beta_{ik_i})^t$ are unknown, and $\boldsymbol{X}_i$ is an $(n \times k_i)$ given design matrix of rank $k_i < n$. Let

$$\hat{\boldsymbol{\beta}}_i = (\boldsymbol{X}_i^t \boldsymbol{X}_i)^{-1} \boldsymbol{X}_i^t \boldsymbol{y} \ \text{ and } \ R_i = |\boldsymbol{y} - \boldsymbol{X}_i \hat{\boldsymbol{\beta}}_i|^2$$

denote the least squares estimator of $\boldsymbol{\beta}_i$ and residual sum of squares, respectively.

Consider noninformative priors of the form

$$\pi_i^N(\boldsymbol{\beta}_i, \sigma_i) = \sigma_i^{-(1+q_i)}, \quad q_i > -1.$$

Common choices of q_i are $q_i = 0$ (the reference prior) or $q_i = k_i$ (the Jeffreys prior). It can be shown, for such priors, that a minimal training sample $\boldsymbol{y}(\ell)$, with corresponding design matrices $\boldsymbol{X}_i(\ell)$ (under the M_i), is a sample of size $m = \max\{k_i\} + 1$ such that all $(\boldsymbol{X}_i^t(\ell)\boldsymbol{X}_i(\ell))$ are nonsingular. (Note that if $q_i = -1$, i.e., constant noninformative priors are used, then one would instead need $m = \max\{k_i\} + 2$.)

Computation yields that

$$B_{ji}^N = \frac{\pi^{(k_j - k_i)/2}}{2^{(q_i - q_j)/2}} \cdot \frac{\Gamma((n - k_j + q_j)/2)}{\Gamma((n - k_i + q_i)/2)} \cdot \frac{(\det \boldsymbol{X}_i^t \boldsymbol{X}_i)^{1/2}}{(\det \boldsymbol{X}_j^t \boldsymbol{X}_j)^{1/2}} \cdot \frac{R_i^{(n-k_i+q_i)/2}}{R_j^{(n-k_j+q_j)/2}},$$

and that $B_{ij}^N(\ell)$ is given by the inverse of this expression with n, $\boldsymbol{X}_i$, $\boldsymbol{X}_j$, R_i, and R_j replaced by m, $\boldsymbol{X}_i(\ell)$, $\boldsymbol{X}_j(\ell)$, $R_i(\ell)$, and $R_j(\ell)$, respectively; here $R_i(\ell)$ and $R_j(\ell)$ are the residual sums of squares corresponding to the training sample $\boldsymbol{y}(\ell)$.

Verification of the above statements, together with derivation of expected IBFs and intrinsic priors for linear models, can be found in Berger and Pericchi (1994). Also, in that paper, IBFs are compared with the related methodology of de Vos (1993), which, for linear models, suggests an approximate weighted geometric average of the training sample Bayes factors, with the weights chosen so as to simplify the resulting computation.

References

Aitkin, M. (1991). Posterior Bayes factors (with discussion). *J. R. Statist. Soc. B*, **53**, 111–142.

Albert, J. H. (1990). A Bayesian test for a two-way contingency table using independence priors. *Canadian J. Statist.*, **14**, 1583–1590.

Atkinson, A. C. (1978). Posterior probabilities for choosing a regression model. *Biometrika*, **65**, 39–48.

Berger, J. and Bernardo, J. M. (1992). On the development of the reference prior method. In *Bayesian Statistics IV*, (eds., J. M. Bernardo, et. al.), London: Oxford University Press.

Berger, J. O. and Delampady, M. (1987). Testing precise hypotheses. *Statist. Sci.*, **3**, 317–352.

Berger, J. and Pericchi, L. (1994). Intrinsic Bayes factors for model selection and prediction in the general linear model. In preparation.

Berger, J. O. and Sellke, T. (1987). Testing a point null hypothesis: the irreconcilability of P values and evidence. *J. Amer. Statist. Assoc.*, **82**, 112–122.

Bernardo, J.M. (1979). Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. B*, **41**, 113–147.

Bertolino, F., Piccinato, L., and Racugno, W. (1992). Multiple Bayes factors for testing hypotheses. Technical Report N. 15 (1992), Dipartimento di Statistica, Università di Roma.

Delampady, M. and Berger, J. O. (1990). Lower bounds on Bayes factors for multinomial distribution, with application to chi-squared tests of fit. *The Annals of Statistics*, **18**, 1295–1316.

de Vos, A. F. (1993). A fair comparison between regression models of different dimension. Technical Report, The Free University, Amsterdam.

Draper, D. (1994). Assessment and propogation of model uncertainty. To appear in *J. Roy. Statist. Soc.* **B**, 56.

Edwards, W., Lindman, H. and Savage, L. J. (1963). Bayesian statistical inference for

psychological research. *Psych. Rev.*, **70**, 193–242.

Garthwaite, P.H. and Dickey, J.M. (1992). Elicitation of prior distributions for variable-selection problems. *The Annals of Statistics*, **20**, 4, 1697-1719.

Geisser, S. and Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, **74**, 153–160.

Gelfand, A. E. and Dey, D. K. (1992). Bayesian model choice: Asymptotics and exact calculations. Technical Report, Department of Statistics, University of Connecticut.

Gelfand, A. E., Dey, D. K. and Chang, H. (1992). Model determination using predictive distributions with implementations via sampling-based methods. In *Bayesian Statistics 4* (eds., J. M. Bernardo et al.), London: Oxford University Press, 147–167.

Geyer, C. and Thompson, E. (1992). Constrained Monte Carlo maximum likelihood for dependent data. *J. Roy. Statist. Soc.*, **B** 54, 657-683.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **88**, 881–889.

Good, I. J. (1985). Weight of evidence: a brief survey. In *Bayesian statistics 2*, (eds., J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith), New York: Elsevier, 249–269.

Haughton, D. (1988). On the choice of a model to fit data from an exponential family. *Ann. Statist.*, **16**, 342–355.

Haughton, D. and Dudley, R. (1992). Information criteria for multiple data sets and restricted parameters. Technical Report, Dept. of Mathematical Sciences, Bentley College, Waltham.

Jefferys, W. H. and Berger, J. O. (1992). Ockham's razor and Bayesian analysis. *American Scientist*, January–February, 64–72.

Jeffreys, H. (1961). *Theory of Probability*. London: Oxford University Press.

Kadane, J. B., Dickey, J., Winkler, R., Smith, W. and S. Peters. (1980). Interactive elicitation of opinion for a normal linear model. *J. Amer. Statist. Assoc.*, **75**, 845–854.

Kass, R. E. and Raftery, A. (1993). Bayes factors and model uncertainty. Technical Report #571, Department of Statistics, Carnegie-Mellon University, Pittsburgh.

Kass, R. E. and Wasserman, L. (1992). A reference Bayesian test for nested hypotheses with large samples. Technical Report No. 567, Department of Statistics, Carnegie Mellon University.

Laud, P. W. and Ibrahim, J. (1992). Predictive variable selection in Bayesian linear regression. Statistics Report No. 92–01, Northern Illinois University.

Lempers, F. B. (1971). *Posterior Probabilities of Alternative Linear Models.* Rotterdam: University of Rotterdam Press.

Lindley, D. V. (1957). A statistical paradox. *Biometrika*, **44**, 187–192.

Madigan, D. and Raftery, A. E. (1991). Model selection and accounting for model uncertainty in graphical models using Occam's window. Technical Report 213, Department of Statistics, University of Washington.

McCulloch, R. E. and Rossi, P. E. (1993). Bayes factors for nonlinear hypotheses and likelihood distributions. *Biometrika*, **79**, 663–676.

Mitchell, T. J., and Beauchamp, J. J. (1988). Bayesian variable selection in regression. *Journal of the American Statistical Association*, **83**, 1023–1032.

O'Hagan, A. (1994). Fractional Bayes factors for model comparisons. To appear in *J. Roy. Statist. Soc.* **B**, 56.

Pericchi, L. R. and Pérez, M. E. (1993). Posterior robustness with more than one sampling model. To Appear, *J. Statist. Planning and Inference.*

Pericchi, L. R. and Smith, A. F. M. (1994). Bayesian model selection without assuming a "true" model. (Work in progress).

Poirier, D. J. (1985). Bayesian hypothesis testing in linear models with continuously induced conjugate priors across hypotheses. In *Bayesian Statistics 2*, (eds., J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith), New York: Elsevier, 711–722.

Raftery, A. E. (1993). Approximate Bayes factors and accounting for model uncertainty in generalized linear models. Technical Report No. 255, Department of Statistics, University of Washington.

Rosenkranz, S. (1992). The Bayes factor for model evaluation in a hierarchical Poisson

model for area counts. Ph.D. dissertation, Department of Biostatistics, University of Washington.

San Martini, A. and Spezzaferri, F. (1984). A predictive model selection criterion. *The Journal of the Royal Statistical Society B*, **46**, 296–303.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.

Smith, A. F. M. and Spiegelhalter, D. J. (1980). Bayes factors and choice criteria for linear models. *J. R. Statist. Soc. B*, **42**, 213–220.

Smith, A. F. M., and Spiegelhalter, D. J. (1981). Bayesian approaches to multivariate structure. In *Interpreting Multivariate Data* (ed., V. Barnett), Chichester: Wiley.

Spiegelhalter, D. J. (1980). An omnibus test for Normality for small samples. *Biometrika* **67**, 493–496.

Spiegelhalter, D. J. and Smith, A. F. M. (1982). Bayes factors for linear and log-linear models with vague prior information. *J. Royal Statist. Soc. B* **44**, 377–387.

Stewart, L. (1987). Hierarchical Bayesian analysis using Monte Carlo integration: computing posterior distributions when there are many possible models. *The Statistician*, **36**, 211–219.

Verdinelli, I. and Wasserman, L. (1993). Bayes factors, nuisance parameters, and imprecise tests. Technical Report #570, Department of Statistics, Carnegie-Mellon University.

Zellner, A. (1984). Posterior odds ratios for regression hypotheses: general considerations and some specific results. In *Basic Issues in Econometrics*, Chicago: University of Chicago Press, 275–305.

Zellner, A. and Siow (1980). Posterior odds for selected regression hypotheses. In *Bayesian Statistics 1* (eds., J. M. Bernardo *et al.*), Valencia: Valencia University Press, 585–603.

James Berger
Department of Statistics
Purdue University
West Lafayette, IN 47907, USA

Luis R. Pericchi
CESMa and Departamento de Matemáticas
Universidad Simón Bolívar
Apartado 8900 Caracas 1080A,
Venezuela