THE INTRINSIC BAYES FACTOR FOR
MODEL SELECTION AND PREDICTION

by

James O. Berger and Luis R. Pericchi*

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.

KEYWORDS: Posterior probability; Noninformative prior; Hypothesis testing; Training sample; Asymptotic Bayes factors.

---

* James Berger is the Brumfield Distinguished Professor, Department of Statistics, Purdue University, West Lafayette, IN 47907. Luis R. Pericchi is Professor, CESMa and Departamento de Matemáticas, Universidad Simón Bolívar, Apartado 8900 Caracas 1080A, Venezuela. This research was supported by the National Science Foundation, Grants DMS-8923071 and DMS-9303556, and by BID-CONICIT (Venezuela). The authors thank Dipak Dey, Jacek Dmochowski, Paul Garthwaite, Alan Gelfand, J. K. Ghosh, Prakash Laud, Maria Eglée Pérez, Jose Miguel Pérez, Tom Sellke, and Julia Varshavsky for very helpful input. An Associate Editor and referees also provided valuable suggestions.

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 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).

(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, Bayesian methods are natural in that they can keep all models under consideration, weighted by their posterior probabilities (cf, Draper, 1995, for a review and earlier references, and also Section 5).

Among the many fine discussions of these issues are Jeffreys (1961), Edwards, Lindman, and Savage (1963), and Kass and Raftery (1995).

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 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.

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 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. This approach is arguably about the best that can be done from a default perspective, but it is very difficult to implement in general, requiring careful selection of a default prior for each specific situation.

In this paper, we introduce a technique for developing default Bayes factors, which we call *intrinsic Bayes Factors.* These can be constructed in very general situations — nested, non-nested, and even irregular problems — and they seem to correspond to actual Bayes factors, at least asymptotically.

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

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

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

$$B_{ji} = \frac{m_j(\boldsymbol{x})}{m_i(\boldsymbol{x})} = \frac{\int f_j(\boldsymbol{x}|\boldsymbol{\theta}_j)\pi_j(\boldsymbol{\theta}_j)d\boldsymbol{\theta}_j}{\int f_i(\boldsymbol{x}|\boldsymbol{\theta}_i)\pi_i(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i}; \tag{2}$$

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 (1995) for discussion and references concerning this viewpoint.) Although we formally discuss only model selection, our development will also apply to hypothesis testing.

Computing B_{ji} requires specification of $\pi_i(\boldsymbol{\theta}_i)$ and $\pi_j(\boldsymbol{\theta}_j)$. Often in Bayesian analysis, one can use noninformative (or default) priors $\pi_i^N(\boldsymbol{\theta}_i)$. 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 (2) would yield

$$B_{ji}^N = \frac{m_j^N(\boldsymbol{x})}{m_i^N(\boldsymbol{x})} = \frac{\int f_j(\boldsymbol{x}|\boldsymbol{\theta}_j)\pi_j^N(\boldsymbol{\theta}_j)d\boldsymbol{\theta}_j}{\int f_i(\boldsymbol{x}|\boldsymbol{\theta}_i)\pi_i^N(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i}. \tag{3}$$

The difficulty with (3) 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. The idea is that $\boldsymbol{x}(\ell)$ will be used to convert the $\pi_i^N(\boldsymbol{\theta}_i)$ to proper posterior distributions $\pi_i^N(\boldsymbol{\theta}_i|\boldsymbol{x}(\ell)) = f_i(\boldsymbol{x}(\ell)|\boldsymbol{\theta}_i)\pi_i^N(\boldsymbol{\theta}_i)/m_i^N(\boldsymbol{x}(\ell))$, where (slightly abusing notation) $f_i(\boldsymbol{x}(\ell)|\boldsymbol{\theta}_i)$ is the marginal density of $\boldsymbol{X}(\ell)$ under M_i and

$$m_i^N(\boldsymbol{x}(\ell)) = \int f_i(\boldsymbol{x}(\ell)|\boldsymbol{\theta}_i)\pi_i^N(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i. \tag{4}$$

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

Lemma 1. *The Bayes factor of model j to model i, conditional on $\boldsymbol{x}(\ell)$ and assuming that the $\pi_i^N(\boldsymbol{\theta}_i|\boldsymbol{x}(\ell))$ are proper, is given by*

$$B_{ji}(\boldsymbol{x}(\ell)) = B_{ji}^N \cdot B_{ij}^N(\boldsymbol{x}(\ell)), \tag{5}$$

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

Proof. Denoting the remaining data by $\boldsymbol{x}(-\ell)$, the result follows from the definition

$$B_{ji}(\boldsymbol{x}(\ell)) = \frac{\int f_j(\boldsymbol{x}(-\ell)|\boldsymbol{\theta}_j, \boldsymbol{x}(\ell))\pi_j^N(\boldsymbol{\theta}_j|\boldsymbol{x}(\ell))d\boldsymbol{\theta}_j}{\int f_i(\boldsymbol{x}(-\ell)|\boldsymbol{\theta}_i, \boldsymbol{x}(\ell))\pi_i^N(\boldsymbol{\theta}_i|\boldsymbol{x}(\ell))d\boldsymbol{\theta}_i}. \qquad \square$$

Clearly (5) 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 $B_{ij}^N(\boldsymbol{x}(\ell))$. This use of a training sample makes sense, however, only if the $m_i^N(\boldsymbol{x}(\ell))$ in (4) 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, but that training samples of two distinct observations are proper. Thus a training sample such as $\boldsymbol{x}(\ell) = (x_i, x_j)$ is minimal, and, indeed, then

$$m_1^N(\boldsymbol{x}(\ell)) = \frac{1}{2\pi(x_i^2 + x_j^2)}, \quad m_2^N(\boldsymbol{x}(\ell)) = \frac{1}{\sqrt{\pi}(x_i - x_j)^2}. \tag{6}$$

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 accommodating the dependence structure in the choice of a minimal training sample.

1.3 Relationship to Other "Automatic" Bayesian Methods

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

$$\pi_1(\sigma_1) = \frac{1}{\sigma_1}, \quad \pi_2(\mu, \sigma_2) = \frac{1}{\sigma_2} \cdot \frac{1}{\pi\sigma_2(1 + \mu^2/\sigma_2^2)}, \tag{7}$$

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 Section 2.1.

Jeffreys' argument for (7) 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), Poirier (1985), Stewart (1987), Mitchell and Beauchamp (1988), Albert (1990), Madigan and Raftery (1995), George and McCulloch (1993), McCulloch and Rossi (1993), Raftery (1993), and Verdinelli and Wasserman (1995). 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 Asymptotic Methods and BIC

Laplace's asymptotic method (cf, Haughton, 1988, and Kass and Raftery, 1995) yields, as an approximation to B_{ji},

$$B_{ji}^L = \frac{f_j(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_j)(\det\hat{\boldsymbol{I}}_j)^{-1/2}}{f_i(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_i)(\det\hat{\boldsymbol{I}}_i)^{-1/2}} \cdot \frac{(2\pi)^{k_j/2}\pi_j(\hat{\boldsymbol{\theta}}_j)}{(2\pi)^{k_i/2}\pi_i(\hat{\boldsymbol{\theta}}_i)}, \tag{8}$$

where $\hat{\boldsymbol{I}}_i$ and and $\hat{\boldsymbol{\theta}}_i$ are the observed information matrix and MLE, respectively, under model M_i. As the sample size goes to infinity, the first factor of B_{ji}^L typically goes to 0 or ∞, while the second factor stays bounded. The BIC criterion of Schwarz (1978) arises from replacing this second factor by an appropriate constant.

Since $\pi_j(\hat{\boldsymbol{\theta}}_j)$ and $\pi_i(\hat{\boldsymbol{\theta}}_i)$ in (8) cannot typically be constant (i.e., independent of the data) in model selection problems, use of BIC cannot typically correspond to an actual Bayesian analysis, and hence violates Principle 1. This violation is often not severe, however. Indeed Kass and Wasserman (1994) argue that, for nested models and contiguous alternatives, BIC does correspond to a Bayesian analysis; see Section 2.4.

The asymptotic expression (8) is also useful theoretically. It can be used to help develop default proper priors (see Section 4), and can also be used to compare the degree to which model selection criterion are approximately Bayes. In this regard, it is worth emphasizing that we seek a method which is Bayesian even up to the second factor in (8).

As a final comment, there are many model selection problems for which the asymptotic answer differs from (8); see Haughton and Dudley (1992). Interestingly, we will see that the intrinsic Bayes factor often succeeds in being Bayesianly compatible even in these nonstandard problems.

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 factors. Efforts have been made to conventionally specify the constants, c_i. For instance, Smith and Spiegelhalter (1980) 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. This method 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 (1994).

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 (8), 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 (5), 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 (5) 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 (1995) 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 Berger and Pericchi (1995) for discussion.

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. Section 3 discusses the IBF for nonnested models or hypotheses, and illustrates the method with several examples. Section 4 introduces the notion of the *intrinsic prior*; this is the (often proper) conventional prior that would give answers similar to the IBF. That IBFs correspond to Bayes factors for sensible conventional priors is a cornerstone of the justification for the approach. 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

$$f_1(\boldsymbol{x}|\boldsymbol{\theta}_1) = f_2(\boldsymbol{x}|\boldsymbol{\xi} = \boldsymbol{\theta}_1,\ \boldsymbol{\eta} = \boldsymbol{\eta}_0), \tag{9}$$

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). 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. Here, $\hat{\boldsymbol{\theta}}_2$ denotes the MLE under M_2.

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

$$\hat{\boldsymbol{\theta}}_2 \overset{\text{under } M_1}{\longrightarrow} \boldsymbol{\theta}_2^* = (\boldsymbol{\theta}_1, \boldsymbol{\eta}_0). \tag{10}$$

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

$$\mathcal{X}_T = \{\boldsymbol{x}(1), \boldsymbol{x}(2), \ldots, \boldsymbol{x}(L)\} \tag{11}$$

denote the set of all minimal training samples, $\boldsymbol{x}(\ell)$). Clearly the Bayes factor $B_{21}(\boldsymbol{x}(\ell))$, defined in (5), will depend on choice of the minimal training sample. To eliminate this dependence and increase stability, a natural idea is to average the $B_{21}(\boldsymbol{x}(\ell))$ 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

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

$$B_{21}^{GI} = \left(\prod_{\ell=1}^{L} B_{21}(\boldsymbol{x}(\ell))\right)^{1/L} = B_{21}^{N} \cdot \left(\prod_{\ell=1}^{L} B_{12}^{N}(\boldsymbol{x}(\ell))\right)^{1/L}, \tag{13}$$

where the $B_{12}^{N}(\boldsymbol{x}(\ell))$ are defined following (5). 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 (12) 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 (13) 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

$$B_{21}^N = \sqrt{\frac{2\pi}{n}} \cdot (1 + \frac{n\bar{x}^2}{s^2})^{n/2}, \tag{14}$$

where $s^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2$. Using (8) noting that $\mathcal{X}_T$ consists of all $L = n(n-1)/2$ pairs of (different) observations, it follows that

$$\begin{aligned} B_{21}^{AI} &= B_{21}^N \cdot \frac{1}{L} \sum_{\ell=1}^{L} \frac{(x_1(\ell) - x_2(\ell))^2}{2\sqrt{\pi}[x_1^2(\ell) + x_2^2(\ell)]} \\ &= B_{21}^N \cdot \frac{1}{n(n-1)} \sum_{i<j} \frac{(x_i - x_j)^2}{\sqrt{\pi}[x_i^2 + x_j^2]}, \end{aligned} \tag{15}$$

$$B_{21}^{GI} = B_{21}^N \cdot \left(\prod_{\ell=1}^{L} \frac{(x_1(\ell) - x_2(\ell))^2}{2\sqrt{\pi}[x_1^2(\ell) + x_2^2(\ell)]} \right)^{1/L}. \tag{16}$$

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 2. 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: $\theta \in \mathbb{R}^1$. 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

$$B_{21}^N = 1/\Phi(-\sqrt{n}\bar{x}), \tag{17}$$

where Φ is the standard normal c.d.f. A minimal training sample is any x_i, since

$$m_1^N(x_i) = \int_{-\infty}^0 \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_i-\theta_1)^2} d\theta_1 = \Phi(-x_i) \tag{18}$$

and $m_2^N(x_i) = 1$ are both finite for a single observation. Thus (12) and (13) become

$$B_{21}^{AI} = \frac{1}{\Phi(-\sqrt{n}\bar{x})} \cdot \frac{1}{n} \sum_{i=1}^n \Phi(-x_i), \tag{19}$$

$$B_{21}^{GI} = \frac{1}{\Phi(-\sqrt{n}\bar{x})} \cdot \left[\prod_{i=1}^n \Phi(-x_i)\right]^{1/n} . \tag{20}$$

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 (8) (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 (8):

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

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 (12) and (13) 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 (12) and (13) by their expectations, evaluated at the MLE. Formally, we define the *expected arithmetic intrinsic Bayes factor* and *expected geometric intrinsic Bayes factor* by, respectively,

$$B_{21}^{EAI} = B_{21}^N \cdot \frac{1}{L} \sum_{i=1}^L E_{\hat{\boldsymbol{\theta}}_2}^{M_2}[B_{12}^N(\boldsymbol{X}(\ell))], \tag{22}$$

$$B_{21}^{EGI} = B_{21}^N \cdot \exp\{\frac{1}{L} \sum_{i=1}^L E_{\hat{\boldsymbol{\theta}}_2}^{M_2}[\log B_{12}^N(\boldsymbol{X}(\ell))]\}, \tag{23}$$

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 (22) and (23) are justified as approximations to (12) and (13) 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 (9), shows that the expectations in (22) and (23) are equivalent to those under M_1.

Example 1 (continued). Here the $\boldsymbol{X}(\ell)$ are exchangeable, so (22) yields

$$B_{21}^{EAI} = B_{21}^{N} \cdot E_{\hat{\boldsymbol{\theta}}_2}^{M_2}\left[\frac{(X_i - X_j)^2}{2\sqrt{\pi}(X_i^2 + X_j^2)}\right] = B_{21}^{N} \cdot \left(\frac{1 - \exp\{-n\bar{x}^2/s^2\}}{2\sqrt{\pi}[n\bar{x}^2/s^2]}\right); \tag{24}$$

see Berger and Pericchi (1994) for computation of the expectation. B_{21}^{EGI} can only be evaluated as an infinite series (see Berger and Pericchi, 1994); but numerical computation is straightforward, as discussed in Section 2.5.

Example 2 (continued). Using (19), (22) becomes (since the $\boldsymbol{X}(\ell)$ are exchangeable)

$$B_{21}^{EAI} = \frac{1}{\Phi(-\sqrt{n}\bar{x})} \cdot E_{\hat{\theta}_2}^{M_2}[\Phi(-X_i)] = \frac{1}{\Phi(-\sqrt{n}\bar{x})} \cdot \Phi(-\bar{x}/\sqrt{2}); \tag{25}$$

here $X_i \sim \mathcal{N}(\theta, 1)$ under M_2 and $\hat{\theta}_2 = \bar{x}$. Again, B_{21}^{EGI} cannot be found 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 2) $E_{\hat{\boldsymbol{\theta}}_2}^{M_2}[B_{21}^{N}(\boldsymbol{X}(\ell))] = \infty$. This also explains the definition of $B_{12}^{AI} = 1/B_{21}^{AI}$; although B_{12}^{AI} could be defined as in (12) 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 (8), the Schwarz approximation, 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 2; 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: (8) becomes

$$B_{21}^L = B_{21}^N \cdot \frac{\hat{\sigma}_2^2}{\hat{\sigma}_1} \cdot \frac{\pi_2(\hat{\mu}, \hat{\sigma}_2)}{\pi_1(\hat{\sigma}_1)}. \tag{26}$$

Schwarz Approximation: This is given by $B_{21}^S = B_{21}^N/\sqrt{2\pi}$.

The Jeffreys Approach: For the priors in (7), Jeffreys (1961) approximated B_{21} by

$$B_{21}^J = B_{21}^N \cdot \hat{\sigma}_2 \cdot \frac{1}{\hat{\sigma}_2 \pi[1 + \hat{\mu}^2/\hat{\sigma}_2^2]}. \tag{27}$$

Smith and Spiegelhalter (1980): Their Bayes factor is $B_{21}^{SS} = B_{21}^N/\sqrt{\pi}$.

It is useful to rewrite $\pi_2(\mu, \sigma_2)$ as $\pi_2(\mu, \sigma_2) = \pi_2(\mu|\sigma_2)\pi_2(\sigma_2)$. If, now, the common noninformative choices $\pi_1(\sigma_1) = 1/\sigma_1$ and $\pi_2(\sigma_2) = 1/\sigma_2$ are made, then (26) becomes

$$B_{21}^L = B_{21}^N \cdot \hat{\sigma}_2 \cdot \pi_2(\hat{\mu}|\hat{\sigma}_2). \tag{28}$$

Note that B_{21}^J is of this form with $\pi_2(\mu|\sigma_2)$ being *Cauchy*$(0, \sigma_2)$, consistent with (7). Likewise, rewrite (24) as

$$B_{21}^{EAI} = B_{21}^N \cdot \hat{\sigma}_2 \cdot \left(\frac{1 - \exp\{-\hat{\mu}^2/\hat{\sigma}_2^2\}}{2\sqrt{\pi}[\hat{\mu}^2/\hat{\sigma}_2]} \right). \tag{29}$$

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,

$$\pi_2^*(\mu|\sigma_2) = \frac{1 - \exp\{-\mu^2/\sigma_2^2\}}{2\sqrt{\pi}[\mu^2/\sigma_2]} \tag{30}$$

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 $\sigma_2 = 1$.

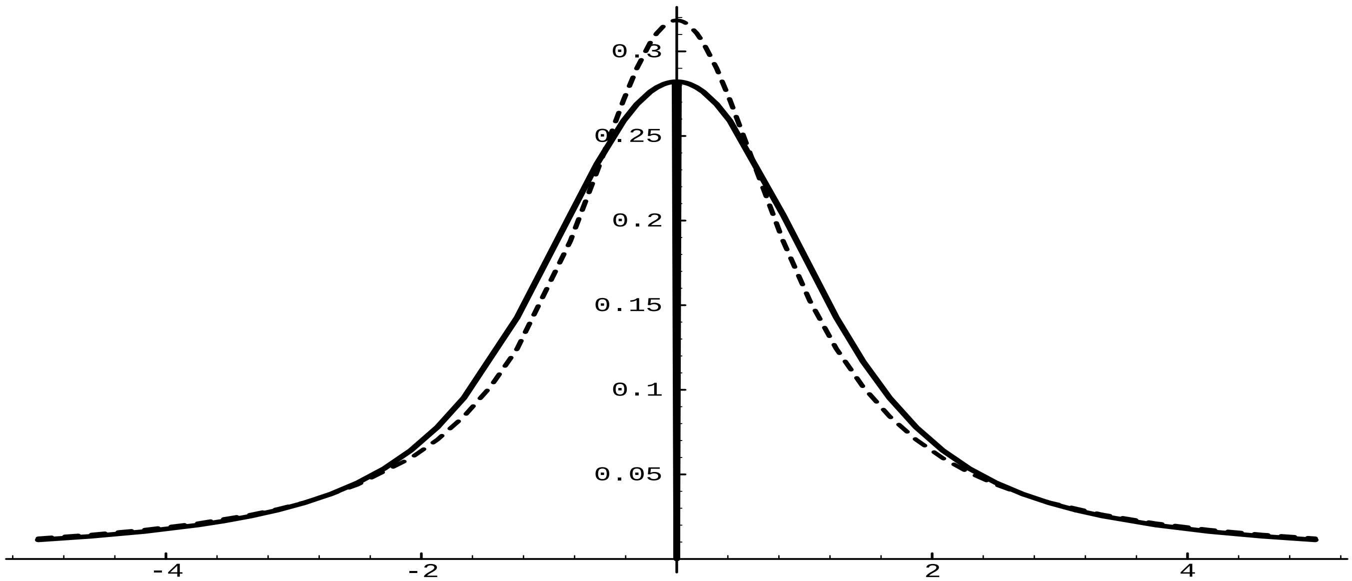

Figure 1. Comparison of the Cauchy prior (dashed) and $\pi_2^*(\mu|\sigma_2)$ (solid) for Example 1.

This remarkable property of B_{21}^{EAI} is, unfortunately, not shared by B_{21}^{EGI}. It does not correspond to a Bayes factor from a proper prior, differing by a constant multiple.

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 (1994) 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.

The story for B_{21}^{AI} and B_{21}^{GI} is similar. Remarkably, the average in (12) again corresponds to a proper prior, even though it is a data-dependent proper prior (see Section 4). The corresponding term of (13) is qualitatively similar, but again is not proper.

2.5 Computational Issues

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

There are three aspects to the computation of (12) or (13): 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 problem common to most Bayesian methods; see Rosenkrantz (1992), Kass and Raftery (1995), Raftery (1993) and Chib (1995) 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 (see Berger and Pericchi, 1995); and see Examples 3 and 5 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 (4) 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 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 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, (22) and (23) 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}; see Berger and Pericchi, 1995). In the exchangeable case, however, (22) and (23) 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 (22) becomes

$$E_{\hat{\boldsymbol{\theta}}_2}^{M_2}[B_{12}^r(\boldsymbol{X}(\ell))] \cong \frac{1}{r} \sum_{i=1}^{r} B_{12}^N(\boldsymbol{x}_i). \tag{31}$$

Note that use of (31) 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 3 (Location-Scale). Suppose the M_i are location-scale densities

$$f_i(\boldsymbol{x}|\mu_i, \sigma_i) = \prod_{j=1}^{n} \sigma_i^{-1} g_i((x_j - \mu_i)/\sigma_i), \tag{32}$$

where $\boldsymbol{x} = (x_1, \ldots, x_n)$, and $\mu_i \in \mathbb{R}^1$ and $\sigma_i > 0$ are unknown. The standard noninformative priors here are $\pi_i^N(\mu_i, \sigma_i) = 1/\sigma_i$. Remarkably, it can be shown that

$$m_i^N(x_j, x_k) = 1/(2|x_j - x_k|), \tag{33}$$

for *any* such M_i (see Berger, Pericchi, and Varshavsky, 1995). It is easy to see that $m_i^N(x_j) = \infty$, so that minimal training samples, $\boldsymbol{x}(\ell)$, are any pair of (different) observations. Using (33), it is clear that $B_{21}^N(\boldsymbol{x}(\ell)) = B_{12}^N(\boldsymbol{x}(\ell)) = 1$. It follows that

$$B_{21}^{AI} = B_{21}^{GI} = B_{21}^N, \quad B_{12}^{AI} = B_{12}^{GI} = B_{12}^N = 1/B_{21}^N, \tag{34}$$

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. See Pericchi and Pérez (1994) 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. A second difficulty is that, in nonnested situations,

$$\bar{B}_{ij}^N \equiv \frac{1}{L}\sum_{\ell=1}^{L} B_{ij}^N(\boldsymbol{x}(\ell)) \tag{35}$$

can be extremely unstable. For instance, its expectation can be infinite.

Example 4. 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 (35) becomes $\bar{B}_{12}^N = \frac{1}{n}\sum\limits_{\ell=1}^{n} 2|x_\ell|$, which is well-behaved under either M_1 or M_2. In contrast, the corresponding $\bar{B}_{21}^N = \frac{1}{n}\sum\limits_{\ell=1}^{n} 1/(2|x_\ell|)$ 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 $B_{21}^{AI} = B_{21}^N \cdot \frac{2}{n}\sum_{\ell=1}^n |x_\ell|$. 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 (12) or (13), 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 (33) would clearly cause problems. Note that one could trim *all* $B_{12}^N(\boldsymbol{x}(\ell))$ except the median value, B_{12}^{med}. This would result in

$$B_{21}^{\text{med}AI} = B_{21}^{\text{med}GI} = B_{21}^N \cdot B_{12}^{\text{med}} = \frac{1}{B_{12}^{\text{med}AI}} = \frac{1}{B_{12}^{\text{med}GI}}. \tag{36}$$

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.

As a final observation, note that one could provide the empirical distribution of the entire set $\{B_{21}(\boldsymbol{x}(\ell))\}$, displayed in, say, a box plot. See the discussion by Dey of Berger and Pericchi (1995) for elaboration of this very interesting idea. We finish this section with a real-data example, indicating the severity of the types of problems discussed above.

Example 5. 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 $\boldsymbol{x} = (x_1 \ldots, x_n)$,

$$m_1^N(\boldsymbol{x}) = \frac{\Gamma(n)}{(\Sigma x_i)^n}, \quad m_2^N(\boldsymbol{x}) = \frac{\Gamma((n-1)/2)}{(\prod_{i=1}^{n} x_i)\pi^{(n-1)/2} 2\sqrt{n} S_y^{(n-1)}}, \tag{37}$$

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

$$m_1^N(\boldsymbol{x}(\ell)) = \frac{1}{(x_i + x_j)^2}, \quad m_2^N(\boldsymbol{x}(\ell)) = \frac{1}{2x_i x_j |\log(x_i/x_j)|}. \tag{38}$$

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) 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 incoherency 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 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 3, 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

$$B_{ji}^{0AI} = B_{0i}^{AI}/B_{0j}^{AI} = B_{ji}^{N} \cdot (\bar{B}_{i0}^{N}/\bar{B}_{j0}^{N}), \tag{39}$$

where $\bar{B}_{i0}^{N}$ and $\bar{B}_{j0}^{N}$ are defined as in (35). (Note, from the second expression in (39), 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 (39) 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 (39) are that one must be able to determine a (minimal) encompassing model M_0 (which itself will not necessarily be unique and which may introduce 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.

Example 4 (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

$$\begin{aligned} \bar{B}_{10}^N &= \frac{1}{n(n-1)} \sum_{i<j} (x_i - x_j)^2 \exp\{-(x_i - x_j)^2/4\}, \\ \bar{B}_{20}^N &= \frac{1}{n(n-1)} \sum_{i<j} \frac{1}{\sqrt{\pi}} (x_i - x_j)^2/(x_i^2 + x_j^2). \end{aligned} \tag{40}$$

Then (39) 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 this larger training sample is beneficial or harmful. A larger training sample provides more stability, but seems to correspond to somewhat less attractive proper Bayes factors.

3.3 Expected Intrinsic Bayes Factors

For nonnested models, there is no obvious analogue of B_{21}^{EAI} in (22) or B_{21}^{EGI} in (23). This is because, in replacing the averages by expectations, one does not know whether to take the expectations under M_1 or M_2. (One could, of course, compute expectations under both models; if the two resulting expected IBFs yielded essentially the same conclusion, this would suffice.)

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)

$$B_{ji}^{E0AI} = \frac{B_{0i}^{EAI}}{B_{0j}^{EAI}}, \quad B_{ji}^{E0GI} = \frac{B_{0i}^{EGI}}{B_{0j}^{EGI}}. \tag{41}$$

Example 4 (continued). Computation yields

$$B_{01}^{EAI} = B_{01}^{N} \cdot E_{(\hat{\mu},\hat{\sigma})}^{M_0}\left[\frac{1}{2}(X_i - X_j)^2 e^{-(X_i - X_j)^2/4}\right] = B_{01}^{N} \cdot \frac{\hat{\sigma}^2}{(\hat{\sigma}^2+1)},$$

$$B_{02}^{EAI} = B_{02}^{N} \cdot E_{(\hat{\mu},\hat{\sigma})}^{M_0}\left[\frac{(X_i - X_j)^2}{2\sqrt{\pi}(X_i^2 + X_j^2)}\right] = B_{02}^{N} \cdot \frac{(1-\exp\{-\hat{\mu}^2/\hat{\sigma}^2\})}{2\sqrt{\pi}(\hat{\mu}^2/\hat{\sigma}^2)}, \tag{42}$$

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

$$B_{21}^{E0AI} = \frac{B_{01}^{EAI}}{B_{02}^{EAI}} = B_{21}^{N} \cdot \frac{2\sqrt{\pi}\hat{\mu}^2}{(\hat{\sigma}^2+1)(1-\exp\{-\hat{\mu}^2/\hat{\sigma}^2\})}. \tag{43}$$

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

4. INTRINSIC PRIORS

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 (35). 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 develops conventional priors for estimation problems. 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, (2) can be approximated by

$$B_{ji} = B_{ji}^{N} \cdot \frac{\pi_j(\hat{\boldsymbol{\theta}}_j)\pi_i^N(\hat{\boldsymbol{\theta}}_i)}{\pi_j^N(\hat{\boldsymbol{\theta}}_j)\pi_i(\hat{\boldsymbol{\theta}}_i)} \cdot (1+o(1)), \tag{44}$$

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 (8) but also holds in nonstandard situations such as that of Example 2 (see (17) and (21)).

To define intrinsic priors we begin by equating (44) with (12) or (13), yielding

$$\frac{\pi_j(\hat{\boldsymbol{\theta}}_j)\pi_i^N(\hat{\boldsymbol{\theta}}_i)}{\pi_j^N(\hat{\boldsymbol{\theta}}_j)\pi_i(\hat{\boldsymbol{\theta}}_i)}(1+o(1)) = \tilde{B}_{ij}^N, \tag{45}$$

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 (45). 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 $k = j$,

$$B_k^*(\boldsymbol{\theta}_k) = \begin{cases} \lim\limits_{L\to\infty} \quad E_{\boldsymbol{\theta}_k}^{M_k}\left[\frac{1}{L}\sum\limits_{\ell=1}^{L} B_{ij}^N(\boldsymbol{X}(\ell))\right] & \text{arithmetic case} \\ \lim\limits_{L\to\infty} \quad \exp\left\{E_{\boldsymbol{\theta}_k}^{M_k}\left[\frac{1}{L}\sum\limits_{\ell=1}^{L}\log B_{ij}^N(\boldsymbol{X}(\ell))\right]\right\} & \text{geometric case;} \end{cases} \tag{46}$$

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 (45), 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)

$$\frac{\pi_j^I(\boldsymbol{\theta}_j)\pi_i^N(\psi_i(\boldsymbol{\theta}_j))}{\pi_j^N(\boldsymbol{\theta}_j)\pi_i^I(\psi_i(\boldsymbol{\theta}_j))} = B_j^*(\boldsymbol{\theta}_j), \tag{47}$$

$$\frac{\pi_j^I(\psi_j(\boldsymbol{\theta}_i))\pi_i^N(\boldsymbol{\theta}_i)}{\pi_j^N(\psi_j(\boldsymbol{\theta}_i))\pi_i^I(\boldsymbol{\theta}_i)} = B_i^*(\boldsymbol{\theta}_i). \tag{48}$$

The motivation, again, is that priors which satisfy (47) and (48) 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.

In the nested model scenario of Section 2.1 and under Assumption N, solutions to (47) and (48) are trivially given by

$$\pi_1^I(\boldsymbol{\theta}_1) = \pi_1^N(\boldsymbol{\theta}_1), \qquad \pi_2^I(\boldsymbol{\theta}_2) = \pi_2^N(\boldsymbol{\theta}_2) B_2^*(\boldsymbol{\theta}_2). \tag{49}$$

Typically there are also many other solutions, perhaps even solutions that are proper distributions, but the solutions in (49) are the simplest. See Dmochowski (1994) for interesting characterizations of intrinsic priors in nested problems.

Example 1 (continued). For B_{21}^{AI}, it follows from (24) and (46) that $B_2^*(\boldsymbol{\theta}_2) = \sigma_2 \cdot \pi_2^*(\mu|\sigma_2)$, where $\pi_2^*(\mu|\sigma_2)$ was defined in (30). Hence the intrinsic prior is

$$\begin{aligned} &\pi_1^I(\sigma_1) = \pi_1^N(\sigma_1) = 1/\sigma_1, \\ &\pi_2^I(\mu, \sigma_2) = \pi_2^N(\sigma_2) B_2^*(\mu, \sigma_2) = \frac{1}{\sigma_2} \cdot \pi_2^*(\mu|\sigma_2). \end{aligned} \tag{50}$$

Thus B_{21}^{AI} behaves (asymptotically) like the actual Bayes factor which uses reference noninformative priors for σ_1 and σ_2, and the proper $\pi_2^*(\mu|\sigma_2)$ for the conditional prior of μ given σ_2. Besides the propriety of $\pi_2^*(\mu|\sigma_2)$, 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 2 (continued). For B_{21}^{AI}, we see from (19), (26), and (46) that $B_2^*(\boldsymbol{\theta}_2) = \Phi(-\theta_2/\sqrt{2})$. Hence, (49) becomes

$$\pi_1^I(\theta_1) = \pi_1^N(\theta_1) = 1, \qquad \pi_2^I(\theta_2) = 1 \cdot \Phi(-\theta_2/\sqrt{2}). \tag{51}$$

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: $\theta_2 \in \mathbb{R}^1$, as opposed to M_1: $\theta < 0$, M_2: $\theta \in \mathbb{R}^1$, 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 $\int_0^\infty \pi_2^I(\theta_2) d\theta_2 = \int\limits_0^\infty \Phi(-\theta_2/\sqrt{2}) d\theta_2 = 1/\sqrt{\pi}$. 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 (9)) 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. (See Berger and Pericchi, 1995, for illustration of this in nested normal linear models.)

For nested problems in which (46) holds (which is typically the case), the above observations can be formalized in a quite interesting fashion.

Theorem 1. *For the arithmetic IBF, suppose that (46) holds and that* $\pi_1^N(\boldsymbol{\theta}_1)$ *is proper. Then* $\pi_2^I(\boldsymbol{\theta}_2)$*, defined in (49), is also proper.*

Proof. Because the limit is assumed to exist in (46), it is true that

$$\begin{aligned}\int \pi_2^I(\boldsymbol{\theta}_2)d\boldsymbol{\theta}_2 &= \int \pi_2^N(\boldsymbol{\theta}_2)\left(\lim_{L\to\infty} E_{\boldsymbol{\theta}_2}^{M_2}\left[\frac{1}{L}\sum_{\ell=1}^{L} B_{12}^N(\boldsymbol{X}(\ell))\right]\right)d\boldsymbol{\theta}_2 \\ &= \lim_{L\to\infty}\frac{1}{L}\sum_{\ell=1}^{L}\int m_2^N(\boldsymbol{x}(\ell))\cdot[m_1^N(\boldsymbol{x}(\ell))/m_2^N(\boldsymbol{x}(\ell))]d\boldsymbol{x}(\ell) \\ &= \lim_{L\to\infty}\frac{1}{L}\sum_{\ell=1}^{L}(1) = 1,\end{aligned} \tag{52}$$

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

The great generality of this theorem is tempered by the requirement that $\pi_1^N(\boldsymbol{\theta}_1)$ be proper. Elsewhere we hope to show that improper π_1^N can be handled by an appropriate limiting argument. Finally, note that there is no analogue of Theorem 1 for B_{21}^{GI}.

Intrinsic priors can also be developed from (47) and (48) for nonnested models. See Berger and Pericchi (1993) and Dmochowski (1994) for examples.

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

$$B_{ij} = \frac{1}{B_{ji}}; \qquad \frac{B_{ij}}{B_{kj}} = B_{ik}. \tag{53}$$

For IBFs to satisfy (53), 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 (53). This is a very appealing.

Arithmetic IBFs will typically not satisfy (53). An exception is when the encompassing model approach of Section 3.2 is used, in which case it is, again, trivial to verify (53). 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 (53).

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

$$m_k^* = \left[\prod_{j=1}^{k-1} B_{kj}^{AI} / \prod_{j=k+1}^{q} B_{jk}^{AI}\right]^{1/p} . \tag{54}$$

It is obvious that the *multiple arithmetic intrinsic Bayes factors*, B_{ij}^{MAI}, then satisfy (53).

Example 5 (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 (54) 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 in the following lemma.

Lemma 2. *The multiple intrinsic Bayes factors $B_{ij}^{MAI} = m_i^*/m_j^*$, where the $\{m_k^*\}$ are defined by (54), provide the best fit to the raw B_{ij}^{AI}, subject to nonnegativity and the coherency condition (53), 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 $c_{ij} = \log B_{ij}^{AI}$. We seek to minimize $\sum\limits_{i=2}^{q} \sum\limits_{j=1}^{i-1} (t_i - t_j - c_{ij})^2$, 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. □

5.2 Posterior Probabilities

Because Bayes factors can easily be converted to posterior probabilities via (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

$$P(M_i|\boldsymbol{x}) = 1/\sum_{j=1}^{q} B_{ji}. \tag{55}$$

In nested model situations, other default choices of the p_i are often made, in part to counteract selection effects from searching among multitudes of models. 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 (see Jeffreys, 1961).

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

$$g(\boldsymbol{y}|\boldsymbol{x}) = \sum_{i=1}^{q} m_i^*(\boldsymbol{y}|\boldsymbol{x}) \cdot P(M_i|\boldsymbol{x}), \tag{56}$$

where $P(M_i|\boldsymbol{x})$ is given by (1) and $m_i^*(\boldsymbol{y}|\boldsymbol{x}) = \int g_i(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}_i)\pi_i(\boldsymbol{\theta}_i|\boldsymbol{x})d\boldsymbol{\theta}_i$. 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 (56).

To utilize (56), 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, 1995, 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. This 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. Note, however, that, asymptotically, the incoherence disappears, since $m_i^*(\boldsymbol{y}|\boldsymbol{x})$ does not depend on the prior asymptotically.

6. CONCLUSIONS AND RECOMMENDATIONS

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.

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

4. They can also be used for default Bayesian hypothesis testing.

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 prediction.

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 for multiple model coherency.

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:

Even though we feel that much study remains to be done to select from among the various IBFs that have been proposed in the paper, 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 (work in progress) have shown 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 (1995) for formulas and discussion in the linear model case.

5. For other non-nested models, we have less clear recommendations. Arithmetic IBFs, with moderate trimming and multiple model adjustment, seem to work fine. And the encompassing approach is always appealing if it can be implemented.

References

Aitkin, M. (1991), "Posterior Bayes Factors," *Journal of the Royal Statistical Society*, Ser. B, 53, 111–142.

Albert, J. H. (1990), "A Bayesian Test for a Two-Way Contingency Table Using Independence Priors. *Canadian Journal of Statistics*, 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, pp. 35–60.

Berger, J. and Delampady, M. (1987), "Testing Precise Hypotheses," *Statistical Science*, 3, 317–352.

Berger, J. and Pericchi, L. (1993), "The Intrinsic Bayes Factor for Model Selection and Prediction," Technical Report 93-43C, Purdue University, Department of Statistics.

Berger, J. and Pericchi, L. (1994), "The Intrinsic Bayes Factor for Linear Models," Technical Report 94-10C, Purdue University, Department of Statistics.

Berger, J. and Pericchi, L. (1995), "The Intrinsic Bayes Factor for Linear Models," in *Bayesian Statistics V*, eds. J. M. Bernardo, et. al., London: Oxford University Press, pp. 23–42.

Berger, J., Pericchi, L., and Varshavsky, J. (1995), "An Identity for Linear and Invariant Models, with Application to Non-Gaussian Model Selection," Technical Report 95-7C, Purdue University, Department of Statistics.

Berger, J. and Sellke, T. (1987), "Testing a Point Null Hypothesis: the Irreconcilability of P-Values and Evidence," *Journal of the American Statististical Association*, 82, 112–122.

Bernardo, J. M. (1979), "Reference Posterior Distributions for Bayesian Inference," *Journal of the Royal Statistical Society*, Ser. B, 41, 113–147.

Chib, S. (1995), "Marginal Likelihood from the Gibbs Output," to appear in *Journal of the American Statistical Association*.

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.

Dmochowski, J. (1994), "Intrinsic Priors Via Kullback-Liebler Geometry," Technical Report 94-15, Purdue University, Department of Statistics.

Draper, D. (1995), "Assessment and Propogation of Model Uncertainty. *Journal of the Royal Statistical Society*, Ser. B, 57, 45–98.

Edwards, W., Lindman, H. and Savage, L. J. (1963), "Bayesian Statistical Inference for Psychological Research," *Psychological Review*, 70, 193–242.

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. (1994), "Bayesian Model Choice: Asymptotics and Exact Calculations," *Journal of the Royal Statistical Society*, Ser. B, 56, 501–514.

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, pp. 147–167.

Geyer, C. and Thompson, E. (1992), "Constrained Monte Carlo Maximum Likelihood for Dependent Data," *Journal of the Royal Statistitical Society*, Ser. B, 54, 657-683.

George, E. I. and McCulloch, R. E. (1993), "Variable Selection Via Gibbs Sampling," *Journal of the American Statistical Association*, 88, 881–889.

Good, I. J. (1985), "Weight of Evidence: a Brief Survey," in *Bayesian Statistics 2*, eds. J. M. Bernardo, et. al., New York: Elsevier, pp. 249–269.

Haughton, D. (1988), "On the Choice of a Model to Fit Data ¿From an Exponential Family," *Annals of Statistics*, 16, 342–355.

Haughton, D. and Dudley, R. (1992), "Information Criteria for Multiple Data Sets and

Restricted Parameters," Technical Report, Bentley College, Department of Mathematics.

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

Kass, R. E. and Raftery, A. (1995), "Bayes Factors," to appear in the *Journal of the American Statistical Association*.

Kass, R. E. and Wasserman, L. (1994), "A Reference Bayesian Test for Nested Hypotheses and Its Relationship to the Schwarz Criterion," Technical Report, Carnegie Mellon University, Department of Statistics.

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

Madigan, D. and Raftery, A. E. (1995), "Model Selection and Accounting for Model Uncertainty In Graphical Models Using Occam's Window," to appear in the *Journal of the American Statistical Association.*

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. (1995), "Fractional Bayes Factors for Model Comparisons," *Journal of the Royal Statistical Society*, Ser. B, 57, 99–138.

Pericchi, L. R. and Pérez, M. E. (1994), "Posterior Robustness with More than One Sampling Model," *Journal of Statistical Planning and Inference*, 40, 279–294.

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, et. al., New York: Elsevier, pp. 711–722.

Proschan, F. (1963), "Theoretical Explanation of Observed Decreasing Failure Rate," *Technometrics*, 5, 375–383.

Raftery, A. E. (1993), "Approximate Bayes Factors and Accounting for Model Uncertainty in Generalized Linear Models," Technical Report No. 255, University of Washington, Department of Statistics.

Rosenkrantz, S. (1992), "The Bayes Factor for Model Evaluation in a Hierarchical Poisson Model for Area Counts," Ph.D. dissertation, University of Washington, Department of Biostatistics.

San Martini, A. and Spezzaferri, F. (1984), "A Predictive Model Selection Criterion," *The Journal of the Royal Statistical Society*, Ser. 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," *Journal of the Royal Statistical Society*, Ser. B, 42, 213–220.

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," *Journal of the Royal Statistical Society*, Ser. 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. (1995), "Bayes Factors, Nuisance Parameters, and Imprecise Tests," in *Bayesian Statistics V*, eds. J.M. Bernardo, et. al., London: Oxford University Press.

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, pp. 585–603.