THE INTRINSIC BAYES FACTOR FOR LINEAR MODELS

by

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

Technical Report #94-10C

Department of Statistics
Purdue University

May 1994

◇

The Intrinsic Bayes Factor for Linear Models *

James O. Berger
Department of Statistics
Purdue University
West Lafayette
Indiana, U.S.A.

and

Luis R. Pericchi
Departamento de Matemáticas and
Centro de Estadística y Software Matemático
Universidad Simón Bolívar
Caracas, Venezuela

Abstract

In Berger and Pericchi (1993) a general automatic Bayesian method for comparing models, the *Intrinsic Bayes Factor* (IBF) was proposed. One version, the *Arithmetic IBF*, was shown to essentially correspond to an actual Bayes factor for a reasonable *Intrinsic Prior*. A second version, the *Geometric IBF*, is justified in Pericchi and Smith (1993), using a prequential type of loss function, without assuming that one of the entertained models is the true sampling model. Here we analyze the general normal linear model, determining the intrinsic Bayes factors for any model comparisons, nested or separate, as well as for multiple model comparisons. In these situations we also discuss determination of intrinsic priors. We also generalize model elaboration ideas to linear models with fixed mean structure but arbitrary error distributions. The method is illustrated on examples and compared with other model selection methods.

Key Words: Automatic Bayes Factor; Bayesian Model Comparisons; Intrinsic Bayes Factors; Model Robustness; Tests of Hypotheses.

1 Introduction

1.1 Background

Berger and Pericchi (1993) argued that Bayesian thinking in Model Selection is necessary, as opposed to estimation problems where Bayesian methods

*This work has been supported by the National Science Foundation, Grants DMS-8923071 and DMS-9303556 and by BID-CONICIT, Venezuela

may be convenient but not always necessary. In regular low dimensional estimation problems, Bayesian and classical methods typically cohere as the sample size grows. However, for Model Selection and Hypothesis Testing problems, frequentist measures of evidence increasingly differ from Bayesian measures as the sample size grows. Paradoxically, estimation is much better understood and developed within the Bayesian paradigm than are model selection and hypothesis testing. Thus research in the latter areas is timely and exciting. In fact, in the Bayesian arena we are witnessing a healthy upsurge of interesting new proposals and new looks at old methods; see Bernardo and Smith (1994), Kass and Wasserman (1994), Kass and Raftery (1994), Berger and Pericchi (1993), Gelfand and Dey (1994), O'Hagan (1994) and de Vos (1993), among others.

To elaborate on this general issue, classical tests of hypotheses, and particularly P-values, are being increasingly regarded as providing suspect measures of evidence in favor or against statistical models. The critical literature covers both philosophical and practical issues. The former includes the difficulties in interpretation of P-values and their lack of compliance with the Likelihood Principle. The latter includes the tendency, in practice, of P-values to over reject simpler models, particularly with large samples; see, for example, Allenby (1990). In contrast, posterior probabilities act as an automatic Ockham's razor, discounting more complex models when there is no cost in doing so; see Smith and Spiegelhalter (1981) and Jefferys and Berger (1992).

Robust Bayes Theory has shed further light on the inadequacy of P-values. The infimum of posterior probabilities, over enormous classes of priors, has been found to be far larger than P-values; see Edwards, Lind-

man and Savage (1963) and Berger and Sellke (1987). In particular, the ubiquitous significance level of 0.05 should only be considered to be at most mild evidence against the null hypothesis.

A second reason for considering the Bayesian approach to hypothesis testing and model selection is that prediction is often the eventual real goal. In allowing 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; see Draper (1994) for discussion and review.

The main disadvantage of Bayesian Hypothesis Testing, in practice, is that there is no accepted general strategy for producing Bayes Factors that are "automatic" or "quasi-automatic"; i.e., that do not require substantial prior input from the user. Although the ideal Bayesian solution would be based on a careful subjective analysis, preferably in a Robust Bayesian fashion, routine information processing by non-specialists is not likely to follow this road. Even accomplished subjectivists frequently wish to utilize automatic or default methods.

It has long been known that the usual "quasi-automatic" approach in Bayesian statistics, based on noninformative or default improper priors, simply does not work in hypothesis testing or model selection problems. The arbitrary constants appearing in the improper prior specifications do not vanish, since model comparisons are based on the marginal likelihood of the observations, an unnormed quantity; the resulting Bayes Factors are thus undetermined. A number of "solutions" to this difficulty have been proposed, and are discussed below. In evaluating these solutions, our primary criterion is — How closely does the solution correspond to actual Bayes fac-

tors with respect to (reasonable) proper priors? For instance, one can ask whether or not the solution has the same asymptotic behavior as that of some proper Bayes factor.

One of the general implementations of quasi-automatic Bayes Factors is that of Geisser and Eddy (1979). However, as the same authors showed, the asymptotic behaviour of their Bayes Factors is different than that of proper Bayes factors. This is a consequence of their use of training samples with lengths that depend on the sample size. The same concern applies to the method of Aitkin (1991) and to some versions (though not all) of the Fractional Bayes factors of O'Hagan (1994). (See Gelfand and Dey, 1994, for asymptotic analysis of these and other methods.)

Other automatic Bayes Factors arise from asymptotic considerations, such as the Schwarz (1978) Criterion or BIC. Such asymptotic arguments nominally ignore a multiplicative constant term, however, and hence do not clearly satisfy our goal of corresponding to actual Bayes factors. Note, on the other hand, that Kass and Wasserman (1992) show that BIC corresponds, approximately, to a proper Bayes Factor in nested model situations when the simpler model is true. In general, however, BIC seems to overstate the evidence in favor of the complex model, which is a "bias" that, if anything, goes in the wrong direction. We will see examples of this.

For normal linear models there have been several specific "conventional" proper priors suggested. Jeffreys (1961) forcefully advocated use of such for certain specific situations. His approach was extended to quite general normal linear models in Zellner and Siow (1980). We will see that arithmetic IBF's appear to be closely related to this method, in that they seem to correspond to use of "intrinsic priors" that are quite similar to the conventional

priors of Jeffreys and Zellner and Siow.

Another reasonable automatic Bayes factor for nested hypotheses is that of Smith and Spiegelhalter (1981), based on Good's device of "imaginary training samples." We will present comparisons of this approach and the IBF approach later. The general finding is that, as with BIC, the Smith and Spiegelhalter approach tends to somewhat overstate the evidence in favor of the more complex model.

In Smith (1983) and Spiegelhalter (1980), a quasi-automatic procedure for Model Elaboration for location-scale alternatives was suggested. The goal was to allow for robustification of the sampling model, by comparison with non-normal location-scale models. Their suggested procedure has been found to be quite successful and robust; see Pericchi and Pérez (1993). However, their suggestion was not fully justified, since it was based on an improper prior. In Berger and Pericchi (1993) it is shown that, for exchangeable observations and any location-scale likelihood, Intrinsic Bayes Factors produce a correction of the improper procedure which is equal to one. Thus the Spiegelhalter and Smith procedure for location-scale alternatives is actually well-scaled. In Section 3 we will generalize the procedure just described to quite general Linear Non-Normal models with fixed mean structure.

A number of other default Bayes factors have been proposed, but are less related to our method. See Berger and Pericchi (1993) and Kass and Raftery (1994) for discussion and references.

1.2 Preview

The ideas behind the Intrinsic Bayes Factor are quite general and natural. Start with standard noninformative priors, and use a minimal part of the

data as a "training sample" to obtain proper posterior densities for the parameters under each model. Then use the obtained proper posteriors and the remaining data to compute the Bayes Factor. This procedure is well scaled, but depends on the arbitrary choice of the training sample. We thus symmetrize with respect to all possible minimal training samples. There are numerous potential ways to symmetrize. Here we consider arithmetic and geometric averages of the Bayes factors obtained from the training samples; these averages will be called the AIBF and espectively.

Several motivations for AIBF's and GIBF's can be given. For instance, in Berger and Pericchi (1993) it is shown that an (often) proper prior distribution exists that yields essentially the same result as the AIBF. These "Intrinsic Priors" may be thought of as reference priors for model comparison. On the other hand, within the general Decision Theoretic perspective of Bernardo and Smith (1994), it is shown in Pericchi and Smith (1993) and in Smith (1994) that the GIBF is justified using a prequential type of loss function in the open model perspective, that is when it is not assumed that one of the entertained models is true. This argument is briefly outlined in Section 2.2.

The paper is organized as follows. Section 2 reviews the Intrinsic Bayes Factor (IBF) and certain of its justifications. In Section 3, we study comparison of Linear Models having the same mean structure but differing error distributions. An example is provided. Section 4 develops the IBF for the General Normal Linear model, including Regression and ANOVA. A Hierarchical model is also considered. We illustrate IBF's with various examples and compare with other approaches. Also in Section 4, an Expectation form of the IBF is developed for use with small samples, and the intrinsic priors

for the Linear Model are determined. Finally, in Section 5 we provide some conclusions.

2 The Intrinsic Bayes Factor

2.1 Definition of IBF's

Suppose that we are comparing q models for the data x,

$$M_i\text{: } X \textit{ has density } f_i(x|\theta_i), \quad i = 1, \ldots, q,$$

and that we only have available default priors $\pi_i^N(\theta_i)$, $i = 1, \ldots, q$. The general strategy for defining IBF's starts with the definition of a proper and minimal training sample. The entire sample x is divided into two subsamples: $x(l)$, which is the training sample, and $x(-l)$ the remaining observations used for discrimination. Define the marginal or predictive densities of X,

$$m_i^N(x) = \int f_i(x|\theta_i)\pi_i^N(\theta_i)d\theta_i.$$

Definition. A training sample, $x(l)$, is called *proper* if $0 < m_i^N(x(l)) < \infty$ for all M_i, and *minimal* if it is proper and no subset is proper. (Note that, if $x(l)$ is proper, then all posteriors, $\pi_i^N(\theta_i|x(l))$, are proper.)

The "standard" use of a training sample to define a Bayes factor is based on using $x(l)$ to "convert" the improper $\pi_i^N(\theta_i)$ to proper posteriors, $\pi_i^N(\theta_i|x(l))$, and using the latter to define a Bayes factor for the remaining data $x(-\ell)$. The result, for comparing M_j to M_i, is (with obvious notation)

$$\begin{aligned} B_{ji}(l) &= \frac{\int f_j(x(-l))|\theta_j, x(l))\pi_i^N(\theta_j|x(l))d\theta_j}{\int f_i(x(-l))|\theta_i, x(l))\pi_i^N(\theta_i|x(l))d\theta_i} \\ &= B_{ji}^N \cdot B_{ij}^N(-\ell), \end{aligned} \tag{1}$$

where

$$B_{ji}^N = \frac{m_j^N(x)}{m_i^N(x)} \quad \text{and} \quad B_{ij}^N(l) = \frac{m_i^N(x(l))}{m_j^N(x(l))} \tag{2}$$

are the Bayes factors that would be obtained for the full data x and training sample $x(l)$, respectively, if one were to blindly use π_i^N and π_j^N.

While $B_{ji}(l)$ no longer depends on the scales of π_j^N and π_i^N, it does depend on the arbitrary choice of the (minimal) training sample $x(l)$. To eliminate this dependence and to increase stability, a natural idea is to average the $B_{ji}(l)$ over all possible training samples $x(l), l = 1, \ldots, L$. Thus, in Berger and Pericchi (1993), we defined the *arithmetic IBF* (AIBF) and *geometric IBF* (GIBF) as, respectively,

$$B_{ji}^{AI} = \frac{1}{L}\sum_{l=1}^{L} B_{ji}(l) = B_{ji}^N \cdot \frac{1}{L}\sum_{l=1}^{L} B_{ij}^N(l), \tag{3}$$

$$B_{ji}^{GI} = \left(\prod_{l=1}^{L} B_{ji}(l)\right)^{1/L} = B_{ji}^N \cdot \left(\prod_{l=1}^{L} B_{ij}^N(l)\right)^{1/L}. \tag{4}$$

An important point, observed in Berger and Pericchi (1993), is that the average of the correction factors, $B_{ij}^N(l)$, must converge (for large samples) in order for B_{ji}^{AI} to correspond to a proper Bayes factor. To this end, it is typically necessary to place the more "complex" model in the numerator of the AIBF, i.e., to let M_j be the more complex model. (See Section 4.2 for the recommended analysis when M_i is not nested in M_j.) We then *define* B_{ij}^{AI} by

$$B_{ij}^{AI} = 1/B_{ji}^{AI}.$$

2.2 Justification and Comparison of IBF's

IBF's apply to very general problems, even nonnested and highly irregular problems. They also possess a number of desirable invariance properties. See

Berger and Pericchi (1993) for discussion of these and related justifications.

Here we focus on the extent to which IBF's behave like actual Bayes factors, since we suggested in Section 1 that this should be of primary concern in evaluation of default methods of model selection. In this regard, a cornerstone of the justification for the AIBF is that it corresponds to use of a plausible default (proper) prior, that we call the Intrinsic prior. This is discussed in Section 4.5, with intrinsic priors for the normal linear model being derived.

As an illustration, consider comparing M_1, the i.i.d. $\mathcal{N}(0, \sigma_1^2)$ model, with M_2, the i.i.d. $\mathcal{N}(\beta, \sigma_2^2)$ model. For the AIBF derived from a "modified Jeffreys" prior (see Section 4), the conditional intrinsic prior for β is shown, in Section 4.5, to be

$$\pi(\beta|\sigma_2) = \frac{1}{2\sqrt{\pi}\sigma_2} \cdot \frac{(1 - \exp\{-\beta^2/\sigma_2^2\})}{(\beta^2/\sigma_2^2)}. \tag{5}$$

This (new) prior satisfies the desiderata in Jeffreys (1961) for a prior in this problem, namely: i) Location at zero, ii) Scale σ, iii) Symmetry around zero, and iv) It should have no moments. Jeffreys observed that the simplest distribution he knew that satisfied his desiderata was the Cauchy(0,σ_2^2) density, and hence he recommended its use. It is interesting that the intrinsic prior in this problem also satisfies i-iv; also appealing is that it is not as sharply peaked at zero as the Cauchy.

We regard this justification of the AIBF to be of paramount importance. However, the justification essentially assumes that the true sampling model is one of the entertained models. What happens if this is not so?

To answer this question, let us draw from Pericchi and Smith (1993) and Smith (1994), in which we entertain two models, M_1 and M_2, but do

not necessarily believe that either is the true sampling model. Suppose the data $x = (x_1, x_2, \ldots, x_n)$ actually arrived sequentially, and write $x^{(i)} = (x_1, \ldots, x_i)$ and $x_{(i+1)} = (x_{i+1}, \ldots, x_n)$.

Assume the prequential type of utility function, for model evaluation,

$$U(M_j) = \sum_{i=s+1}^{n} \log m_j(x_i|x^{(i-1)}),$$

where s is the minimum value so that $m_j(x_{s+1}|x^{(s)})$, for $j = 1, 2$, are proper predictive densities. Notice that, in Bernardo and Smith's words, we have a model when we have a predictive; thus before we reach the value s, we do not yet have proper models to compare.

It is easy to see that $U(M_j)$ can be rewritten as

$$U(M_j) = \log(m_j(x_{(s+1)}|x^{(s)})).$$

Thus the formal solution of the model selection problem is to prefer M_2 if

$$\int \log\left(\frac{m_2(x_{(s+1)}|x^{(s)})}{m_1(x_{(s+1)}|x^{(s)})}\right) m_A(x_{(s+1)}|x^{(s)}) dx_{(s+1)} > 0,$$

where $m_A(x_{(s+1)}|x^{(s)})$ corresponds to the actual sampling model.

For exchangeable observations, a Monte Carlo approximation to this integral leads to the approximate solution: prefer M_2 if

$$\prod_{l=1}^{K} [B_{21}(x_{(s+1)}(l)|x^{(s)}(l))]^{1/K} > 1,$$

based on K random selections from the $l = 1, \ldots, L$ partitions $x = (x^{(s)}(l), x_{(s+1)}(l))$. Therefore the (approximate) solution is equivalent to selection according to the GIBF. Notice, also, that this scenario motivates the use of minimal training samples; they define the first points at which the utility function is computable.

Another pleasant feature of the GIBF is that it is automatically coherent across models in the sense that $B_{ki}^{GI} = B_{kj}^{GI} B_{ji}^{GI}$. The AIBF is not automatically coherent across models, but Berger and Pericchi (1993) develop a modification that is coherent; for linear models, the "encompassing" model approach, recommended in Section 4.2, also results in coherence across models.

It is easy to see that the GIBF favors simpler models to a greater extent than does the AIBF, and that they can differ markedly in some instances. As they each have separate, reasonable motivations, we do not at this time recommend one over the other. Indeed, the problem of model selection is so complex that it is perhaps unreasonable to expect a simple "good for everything" recipe. We thus put forward a general strategy, the Intrinsic strategy, which has several possible valuable implementations, and will trust experience to reveal which should be used, and when. See Berger and Pericchi (1993) for more discussion of possible evaluations within the Intrinsic strategy.

There is an important class of problems for which the AIBF and the GIBF coincide, and it is to this class that we now turn.

3 The IBF for Fixed Design Linear Models but Different Error Distributions

Assume that we are considering the Linear Models

$$M_j\colon\ Y = X\beta_j + \sigma_j \epsilon_j, \tag{6}$$

for $j = 1, \ldots, q$ alternative error models $\epsilon_j \sim g_j$; here Y is $n \times 1$, X is $n \times k$, $\beta_j \in R^k$ is $k \times 1$, $\sigma_j > 0$, and ϵ_j is $n \times 1$. Note that the design matrix, X,

is assumed to be fixed across models. We label the unknown β and σ by j, so as to emphasize that parameters can have different meanings within different models. We will use reference default priors, $\pi_j^N(\beta_j, \sigma_j) = 1/\sigma_j$. A minimal training sample can be seen to be any $(k+1)$-vector $y(l)$ with corresponding sub-matrix $X(l)$, of X, such that $X^t(l)X(l)$ is nonsingular. Let $|A|$ denote the determinant of a matrix A.

Lemma 1. *In the above situation, if $g_j(v) = g_j(-v)$, then the marginal density of the minimal training sample $y(l)$ is*

$$m_j^N(y(l)) = [2|X^t(l)X(l)|^{1/2}|y(l) - X(l)(X^t(l)X(l))^{-1}X^t(l)y(l)|]^{-1}. \quad (7)$$

Lemma 1 is established in Berger, Pericchi and Varshavsky (1994). It is a quite surprising result because $m_j(y(l))$ does not depend in any way on g_j. For instance, it holds when g_j is any $\mathcal{N}_n(0, \Sigma_j)$ distribution, regardless of Σ_j. It also holds for nonnormal distributions.

Since "calibration" of priors for different models by "matching predictives" is an acknowledged technique (cf., Laud and Ibrahim, 1992, and Berger and Pericchi, 1993), Lemma 1 suggests that the reference prior is properly calibrated for comparison of *any* models of the form (6). This result greatly simplifies the model elaboration for linear models, since then all $B_{ij}^N(l)$ clearly equal one (see (2)) and hence (from (3) and (4))

$$B_{ji}^{AI} = B_{ji}^{GI} = B_{ji}^N. \quad (8)$$

The latter is often available in closed form, as in the following example.

Example 1. The data y_{ij}, $i = 1, 2$; $j = 1, \ldots, n_i$, is felt to arise from the "two group" model $y_{ij} = \mu_i + \sigma\epsilon_{ij}$. It is desired to compare the two models M_N: the ϵ_{ij} are i.i.d. $\mathcal{N}(0, 1)$, and M_U: the ϵ_{ij} are i.i.d. $\mathcal{U}(-\sqrt{3}, \sqrt{3})$.

	$n_i = 2$	$n_i = 3$	$n_i = 4$	$n_i = 5$	$n_i = 10$	$n_i = 15$	$n_i = 20$
$B^N_{N,U}$	0.992	0.95	0.799	0.646	0.297	0.018	0.007

Table 1: Simulation of two Uniform groups. Behaviour of the IBF with respect to sample size

Let r_i denote the range of the data for group i; let $r_T = \text{Max}\{r_1, r_2\}$; let $n = n_1 + n_2$; and let R denote the usual residual sum of squares after fitting the Normal model. Calculation shows that the marginal densities of the observations for the Uniform and Normal models are

$$m^N_U(y) = \frac{1}{r_T^{n-2}}\left(\frac{r_1 r_2}{n r_T^2} - \frac{r_1 + r_2}{(n-1) r_T} + \frac{1}{n-2}\right),$$

$$m^N_N(y) = \frac{\Gamma((n-2)/2)}{2\pi^{(n-2)/2}(n_1 n_2)^{1/2} R^{(n-2)/2}}.$$

Thus $B^N_{N,U} = m^N_N(y)/m^N_U(y)$.

In Table 1 we present the average of the IBF's for ten simulations of n_i observations from each of the two Uniform densities: $U(-2,0)$ and $U(0,2)$. At $n_i = 2$, the IBF's average is nearly one, as would be expected for minimal data. The IBF's in favor of the Uniform Model grows reasonably fast as n_i increases, being overwhelming for $n_i = 15$ or 20.

For other common error distributions in this example, the marginals of the observations are also either available in closed form or involve at most a one dimensional numerical integration (over σ).

4 The IBF for the Normal Linear Model

4.1 Basic Formulas

Suppose, for $j = 1, \ldots, q$, that model M_j for the data $Y(n \times 1)$ is the linear model

$$M_j\colon\ Y = X_j\beta_j + \epsilon_j, \qquad \epsilon_j \sim \mathcal{N}_n(0, \sigma_j^2 I_n),$$

where σ_j^2 and $\beta_j = (\beta_{j1}, \beta_{j2}, \ldots, \beta_{jk_j})^t$ are unknown, and X_j is an $(n \times k_j)$ given design matrix of rank $k_j < n$. Let

$$\hat{\beta}_j = (X_j^t X_j)^{-1} X_j^t y \text{ and } R_j = |y - X_j\hat{\beta}_j|^2$$

denote the least squares estimator β_j and residual sum of squares, respectively.

We will consider default priors of the form

$$\pi_j^N(\beta_j, \sigma_j) = \sigma_j^{-(1+q_j)}, \quad q_j > -1. \tag{9}$$

Common choices of q_j are $q_j = 0$ (the reference prior; cf., Bernardo, 1979, and Berger and Bernardo, 1992) or $q_j = k_j$ (the Jeffreys prior). When comparing model M_i nested in M_j, we will also consider a *modified Jeffreys prior*, having $q_i = 0$ and $q_j = k_j - k_i$. This is intermediate between the reference and Jeffreys priors.

It is easy to show, for these priors, that a minimal training sample $y(l)$, with corresponding design matrices $X_j(l)$ (under the M_j), is a sample of size $m = \max\{k_j\} + 1$ such that all $(X_j^t(l)X_j(l))$ are nonsingular. (Note that if $q_j = -1$, i.e., constant noninformative priors are used, then one would instead need $m = \max\{k_j\} + 2$.)

Computation yields that

$$B_{ji}^N = \frac{\pi^{(k_j-k_i)/2}}{2^{(q_i-q_j)/2}} \cdot \frac{\Gamma((n-k_j+q_j)/2)|}{\Gamma((n-k_i+q_i)/2)} \cdot \frac{|X_i^t X_i|^{1/2}}{|X_j^t X_j|^{1/2}} \cdot \frac{R_i^{(n-k_i+q_i)/2}}{R_j^{(n-k_j+q_j)/2}}, \qquad (10)$$

and that $B_{ij}^N(l)$ is given by the inverse of this expression with n, X_i, X_j, R_i, and R_j replaced by m, $X_i(l)$, $X_j(l)$, $R_i(l)$, and $R_j(l)$, respectively; here $R_i(l)$ and $R_j(l)$ are the residual sums of squares corresponding to the training sample $y(l)$, i.e.,

$$R_j = |y(l) - X_j(l)\hat{\beta}_j(l)|^2, \qquad \hat{\beta}_j(l) = (X_j^t(l)X_j(l))^{-1}X_j^t(l)y(l). \qquad (11)$$

Inserting these expressions in (1) results in the following arithmetic IBF's for the three default priors being considered. (For the corresponding geometric IBF's, simply replace the arithmetic averages by geometric averages.)

Using the Jeffreys prior:

$$B_{ji}^{AI} = \frac{|X_i^t X_i|^{1/2}}{|X_j^t X_j|^{1/2}} \cdot \left(\frac{R_i}{R_j}\right)^{n/2} \cdot \frac{1}{L}\sum_{l=1}^{L} \frac{|X_j^t(l)X_j(l)|^{1/2}}{|X_i^t(l)X_i(l)|^{1/2}} \cdot \left(\frac{R_j(l)}{R_i(l)}\right)^{m/2}. \qquad (12)$$

Using the Modified Jeffreys prior: Defining $p = k_j - k_i$,

$$B_{ji}^{AI} = \frac{|X_i^t X_i|^{1/2}}{|X_j^t X_j|^{1/2}} \cdot \left(\frac{R_i}{R_j}\right)^{(n-k_i)/2} \cdot \frac{1}{L}\sum_{l=1}^{L} \frac{|X_j^t(l)X_j(l)|^{1/2}}{|X_i^t(l)X_i(l)|^{1/2}} \cdot \left(\frac{R_j(l)}{R_i(l)}\right)^{(p+1)/2}. \qquad (13)$$

Using the Reference prior: Defining $p = k_j - k_i$ and

$$C = \frac{\Gamma((n-k_j)/2)\Gamma((k+1)/2)}{\Gamma((n-k_i)/2)\Gamma(1/2)}, \qquad (14)$$

$$B_{ji}^{AI} = \frac{|X_i^t X_i|^{1/2}}{|X_j^t X_j|^{1/2}} \cdot \frac{R_i^{(n-k_i)/2}}{R_j^{(n-k_j)/2}} \cdot \frac{C}{L}\sum_{l=1}^{L} \frac{|X_j^t(l)X_j(l)|^{1/2}}{|X_i^t(l)X_i(l)|^{1/2}} \cdot \frac{(R_j(l))^{1/2}}{(R_i(l))^{(p+1)/2}}. \qquad (15)$$

For Known σ^2: If the σ_j^2 are known and equal σ^2, and the $\pi_j^N(\beta_j) = 1$, then

$$B_{ji}^N = (2\pi\sigma^2)^{(k_j-k_i)/2} \cdot \frac{|X_i^t X_i|^{1/2}}{|X_j^t X_j|^{1/2}} \cdot \exp\left\{-\frac{1}{2\sigma^2}(R_j - R_i)\right\}. \qquad (16)$$

Here, a minimal training sample is a sample of size $m = \max\{k_j\}$ such that all $(X_j^t(l)X_j(l))$ are nonsingular, and $B_{ji}^N(l)$ is as in (16) with X_i, X_j, R_i, and R_j replaced by $X_i(l)$, $X_j(l)$, $R_i(l)$, and $R_j(l)$. Thus the arithmetic intrinsic Bayes factor in (1) is

$$B_{ji}^{AI} = \frac{|X_i^t X_i|^{1/2}}{|X_j^t X_j|^{1/2}} \cdot \exp\left\{-\frac{1}{2\sigma^2}(R_j - R_i)\right\} \times \frac{1}{L}\sum_{l=1}^{L} \frac{|X_j^t(l)X_j(l)|^{1/2}}{|X_i^t(l)X_i(l)|^{1/2}} \cdot \exp\left\{-\frac{1}{2\sigma^2}(R_j(l) - R_i(l))\right\}. \tag{17}$$

Independently of our work, deVos (1993) suggested, for linear models, an approximate weighted geometric average of the training sample Bayes factors, with the weights chosen so as to simplify the resulting computation. In our averages, we use only equal weights $w(l) = 1/L$. (An exception is when L is so large that only a subset of size L^* of the collection of training samples can be used; this would correspond to choosing $w(l) = 1/L^*$ on this subset, and $w(l) = 0$ otherwise.)

4.2 The Encompassing Model Approach

In order for IBF's to correspond to actual Bayes factors (asymptotically), it is necessary for the average of the $B_{ij}^N(l)$ to converge (asymptotically) to a nonzero value. It was mentioned, earlier, that comparing the more complex model to the simpler model will tend to ensure this convergence, whereas reversing the order can result in a lack of convergence (at least for the AIBF). When comparing nonnested linear models, however, it is often not clear which model (if either) is more complex. In Berger and Pericchi (1993) a solution to this dilemma was proposed: create a model, M_0, in which all M_j are nested, and then define all IBF's relative to M_0.

Definition. For the linear models M_j, $j = 1, \ldots, q$, given in (6), define the *encompassing model*, M_0, as the linear model with design matrix X_0 consisting of all covariates from all the M_j. Then define the *arithmetic encompassing IBF* of M_j to M_i as

$$B_{ji}^{0AI} = B_{0i}^{AI} / B_{0j}^{AI}. \tag{18}$$

Thus the encompassing model approach first compares all models to the encompassing model M_0, and then uses the multiplicative property of Bayes factors to perform other model comparisons. Note that the minimal training sample will now be a sample of size $m = k_0 + 1$, where k_0 is the dimension of β_0, assuming the resulting $(X_0^t(l)X_0(l))$ is nonsingular. Note, also, that the geometric encompassing IBF can be similarly defined; actually, however, it is clear from the multiplicative nature of geometric IBF's that the result will just be the usual geometric IBF, but with the minimal training sample size being $k_0 + 1$, rather than $\max\{k_j\} + 1$.

Example 2 (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 **1,2,3,4** and a constant term (included in all models) which we denote by **c**. This data set is somewhat extreme because of the very small sample size (n=13) and because the design matrix is nearly singular. The encompassing model is clearly that defined by **1,2,3,4,c**. The minimal training samples are of size 6 in this example and there are a total of 1715 such training samples.

In Table 2, we give the Bayes factors of the encompassing model to the submodels for the AIBF with reference and modified Jeffreys priors (AIBF1 and AIBF2, respectively), for the GIBF with reference and modified Jeffreys

Model	AIBF1	AIBF2	GIBF1	GIBF2	ZS	BIC	P-value
1,2,3,c	0.29	0.29	0.16	0.13	0.3	0.3	0.84
1,2,4,c	0.26	0.26	0.15	0.12	0.3	0.313	0.896
1,3,4,c	0.31	0.32	0.19	0.16	0.36	0.438	0.501
2,3,4,c	1.2	1.2	0.73	0.6	1.11	5.09	0.071
1,2,c	0.18	0.19	0.08	0.04	0.26	0.08	0.47
1,3,c	8242.3	15873.3	3965.9	1844.2	2439.0	7971320.2	0.0
1,4,c	0.46	0.45	0.2	0.09	0.56	0.45	0.168
2,3,c	216.4	361.4	96.3	44.8	90.91	32307.5	0.0002
2,4,c	2774.4	5071.1	1272.2	591.6	833.33	2383719.0	0.0
3,4,c	13.1	13.8	4.6	2.1	7.14	130.6	0.0055
1,c	4158.8	8530.9	1386.0	300.3	3125.0	775803.8	0.0
2,c	1909.8	3564.4	539.9	117.0	1176.5	234150.7	0.0
3,c	22842.1	52083.6	6327.2	1371.1	11494.2	13534582.1	0.0
4,c	851.3	1705.3	405.0	87.8	1086.9	213945.2	0.0
c	19721.5	37830.2	8316.2	749.3	11235.9	5442341.9	0.0

Table 2: Hald's data; Bayes Factors of the encompassing model to all possible sub-models

priors (GIBF1 and GIBF2, respectively), for Zellner and Siow's conventional prior (denoted by ZS), for BIC, and we also give the P-values corresponding to the submodels. We do not present the IBF's for the Jeffreys priors, partly due to a lack of space but mainly because our practical and also theoretical analysis (see Section 4.5) suggest that the Jeffreys priors are less reliable in this set up.

Discussion The AIBF's are reasonably stable with respect to the priors considered. Also they are quite similar to the actual Bayes factors of Zellner and Siow. The GIBF's are somewhat more favorable to the simpler models; GIBF1 seems to be roughly half the AIBF's, while GIBF2 is half again smaller. (We have found that GIBF2 tends to be rather unstable and

"too small," hence we do not recommend its use.) Considering the small sample sizes and high colinearity, the stability of the IBF's is encouraging. BIC values for most models are reasonable, particularly for the seemingly best model (**1,2,c**). In general, however, BIC does seem to favor the more complex model, as was mentioned in the Introduction.

Using (18), any pair of models can, of course, be compared. For instance, suppose that we wish to compare M_1: **1,2,c**; M_2: **3,4,c**; and M_3: **1,4,c**. Assume, say, reference priors. From Table 2 we obtain $B_{12}^{0AI} = 72.8$, $B_{13}^{0AI} = 2.54$, and $B_{23}^{0AI} = 0.035$ which are reasonably close to the respective GIBF's: 56.1, 2.45 and 0.044 and are also close to the Zellner and Siow values which are, respectively, 27.46, 2.15 and 0.078. Thus M_1 is moderately preferred to M_3 and quite strongly preferred to M_2. The corresponding values for BIC are 1637.2, 5.64, and 0.003. The first and last values seems too high and too low respectively, as compared with the other criteria. This is not surprising since, in this example, we are far from being in a large sample situation.

4.3 ANOVA and Hierarchical Models

We now specialize the formulae for the IBF in the general linear model to ANOVA1 models. Also, we compare these models with a (nonlinear) hierarchical model. This is an interesting comparison in its own right, since we are discriminating between versions of a "fixed effects" and a "random effects" model.

For simplicity assume that the design is balanced. The observations are

$$y_{ij} \sim \mathcal{N}(\beta_i,\ \sigma^2),\ i = 1, \ldots, I; j = 1, \ldots, J.$$

We employ the parametrization $\beta_i = \mu + \alpha_i$, with $\sum_{i=1}^{I} \alpha_i = 0$.

Suppose that, initially, we wish to compare the usual models,

$$M_1 : \alpha_1 = \ldots = \alpha_I = 0 \text{ vs. } M_2 : \text{ the } \alpha_i \text{ are unequal.} \tag{19}$$

Assume, as before, either reference priors ($q_1 = q_2 = 0$) or modified Jeffreys priors ($q_1 = 0, q_2 = I - 1$). Then the AIBF is seen to be

$$B_{21}^{AI} = \frac{\Gamma((n-I+q_2)/2)\Gamma((I+q_1)/2)}{\Gamma((n-1+q_1)/2)\Gamma((1+q_2)/2)} \left(\frac{2n}{J^I(I+1)}\right)^{1/2} \times \frac{R_1^{(n-1+q_1)/2}}{R_2^{(n-I+q_2)/2}} \frac{1}{L} \sum_{l=1}^{L} \frac{R_2^{(1+q_2)/2}(l)}{R_1^{(I+q_1)/2}(l)}, \tag{20}$$

and the GIBF has the same expression except that the arithmetic average is replaced by a geometric average.

We next analyze two data sets which appeared in Box and Tiao (1973, pp. 246-247). The first is the *dyestuff data* and the second a simulated data set where the within residual sum of squares is smaller than the between residual sum of squares so that the usual frequentist method for estimation of variances would yield a negative estimated variance. In both examples there are I=6 potentially different groups and J=5 replicates, so that n=30. Thus, in both examples, the minimal training sample size is m=7. As in Table 2, we denote by AIBF1 and AIBF2 the Arithmetic IBF with reference and modified Jeffreys priors, respectively. GIBF1 is the recommended GIBF (that with reference priors). We also give the Bayes factors arising from use of Zellner and Siow's (ZS) and Smith and Spiegelhalter's (SS) criteria as well as BIC and the P-values. We also include the Fractional Bayes Factor (O'Hagan, 1994) with the choice of the fraction $b = m/n$, denoted by Fm.

Example 3. Box and Tiao Dyestuff Data

See Table 3.

Example 4. Box and Tiao Simulated Data

Models	AIBF1	AIBF2	GIBF1	ZS	Fm	SS	BIC	P-value
M_2/M_1	5.36	4.20	1.15	5.1319	4.42	558.104	1044.12	0.0044
M_2/M_3	0.5		0.23					

Table 3: The Box and Tiao dyestuff data. Bayes Factors of M_2 to M_1 and of M_2 to M_3

Models	AIBF1	AIBF2	GIBF1	ZS	Fm	SS	BIC	P-value
M_2/M_1	0.0145	0.0038	0.0034	0.008	0.007	0.122	0.2279	0.7311
M_2/M_3	0.088		0.025					

Table 4: Box and Tiao simulated data. Bayes Factors of M_2 to M_1 and of M_2 to M_3

See Table 4

As in Example 2, the AIBF's and ZS are reasonably close. Also they are close to Fm in examples 3 and 4. Note, however, that this is so with the choice of the fraction $b = m/n$. Other choices favored by O'Hagan (1994), particularly those which do not go to zero with the sample size, are likely to yield very different numbers.

On the other hand, BIC favors the more complex model to a much greater extent, as does SS to a somewhat lesser degree. The version of BIC used for Tables 2, 3 and 4 is the Likelihood Ratio of the more complex over the simpler model multiplied by $|X_1^t X_1|^{1/2}/|X_2^t X_2|^{1/2}$; see Berger and Pericchi (1993). Another plausible version of BIC in these ANOVA examples is to multiply the Likelihood Ratio by $n^{*(k_1-k_2)/2}$, where n^* is the "effective" sample size, here $n^* = J$. In Tables 3 and 4 this alternate BIC is 426.23 and .0930, respectively.

That the BIC's (and SS) are unreasonably high is indicated by computing an upper bound on the Bayes factor over *all* i.i.d. $\mathcal{N}(\mu, \rho\sigma^2)$ priors for the β_i (under M_2), with reference priors used for the unknown μ and σ^2 (under both models). The upper bound (over ρ) for the dyestuff data is $\overline{B} = 24.15$. As this is an upper bound over a broad class of priors, BIC and SS are clearly suspect.

For both Examples 3 and 4 a plausible alternative model is the following hierarchical model that we denote by M_3:

$$\begin{gathered} y_{ij} \sim \mathcal{N}(\beta_i, \sigma^2), i = 1, \ldots, I; j = 1, \ldots, J, \\ \pi_1(\beta|\mu_\pi, \sigma_\pi^2) = \mathcal{N}_I(\mu_\pi 1, \sigma_\pi^2 I_{I\times I}),\ \pi_1(\sigma^2) = 1/\sigma^2, \\ \pi_2(\mu_\pi, \sigma_\pi^2) = \pi_{2,1}(\mu_\pi)\pi_{2,2}(\sigma_\pi^2), \end{gathered} \tag{21}$$

where $1 = (1, \ldots, 1)^t$, $\pi_{2,1}(\mu_\pi) = 1$, and $\pi_{2,2}(\sigma_\pi^2) = 1/(\frac{\sigma^2}{J} + \sigma_\pi^2)$. See Berger and Strawderman (1994) for discussion and justification of these and related priors for hierarchical models.

It is convenient to first make the change of variables $\rho = \sigma_\pi^2/\sigma^2$. In determining $m_3^N(y)$, all parameters, except ρ, can be integrated out, resulting in

$$\begin{aligned} m_3^N(y, \rho) &= K_1 \frac{(1/J + \rho)^{(n-I-2)/2}}{(R_w(1/J + \rho) + R_B)^{(n-1)/2}}, \\ K_1 &= \frac{2^{(n-1)/2}}{(2\pi)^{(n-1)/2} J^{I/2} I^{1/2}} \Gamma((n-1)/2), \end{aligned} \tag{22}$$

where R_w, R_B are the usual within and between residual sum of squares respectively. Hence $m_3^N(y)$ is obtained by a one dimensional integration of (22) over ρ.

Let us turn now to the correction factors. The more complex model here is M_2. So the minimal training samples are of size $I + 1$; one observation

from each group (to be denoted by $y_{i1}(l)$) plus one replica in, say, group k (to be denoted by $y_{i2}(l)$). Then, for the same prior specifications as above, a similar computation yields,

$$m_3^N(y(l),\rho) = \frac{\Gamma(I/2)2^{-(I+1)/2}\pi^{-I/2}}{(1+\rho)^{(I/2-1)}(1/J+\rho)((I+1)/2+I\rho)^{1/2}(H(\rho,l))^{I/2}}, \quad (23)$$

where, defining $\overline{y}_k(l) = (y_{k1}(l) + y_{k2}(l))/2$,

$$H(\rho,l) = \frac{1}{2}\left[\frac{1}{(1+\rho)}\sum_{i\neq k} y_{i1}^2(l) + (\rho+\frac{1}{2})^{-1}\overline{y}_k^2(l) + \frac{1}{2}(y_{k1}(l)-y_{k2}(l))^2 - \frac{(1+\rho)(\rho+1/2)}{(I\rho+(I+1)/2)}\cdot\left(\frac{1}{(1+\rho)}\sum_{i\neq k} y_{i1}(l) + (\rho+\frac{1}{2})^{-1}\overline{y}_k(l)\right)^2\right].$$

Thus computing the $m_3^N(y(l))$ will also involve a one dimensional numerical integration over ρ.

Tables 3 and 4 also give the IBF's for comparing M_2 (with the reference prior) to M_3 (with the prior stated above). Note that it is not clear how to define ZS, SS, BIC, or even the P-value for this comparison; hence the corresponding entries of Tables 3 and 4 are left blank.

Consider, first, the Bayes factors for the simulated data in Table 4. Here M_3 is substantially favored over M_2 (by factors of 11 and 40 for the AIBF and GIBF, respectively). Not surprisingly, the "correct" M_1 is favored by factors of 6 or 7 over M_3.

The Bayes factors for the dyestuff data in Table 3 show that M_3 is favored over M_2 by factors of 2 and 4 for the AIBF and GIBF, respectively, with M_1 being least favored.

These two examples support the value of using a hierarchical model. The hierarchical model, M_3, is not unreasonable for either of the two extremes;

indeed, somewhat surprisingly, it is the best model for the dyestuff data.

The computations involving M_3 were done by averaging over only 1/3 of the training samples. This cost saving device was used because numerical integration was required for each training sample.

4.4 The Expected Intrinsic Bayes Factors

For small sample sizes, the training sample averages in (12), (13), and (15) can have large variances (as statistics in a frequentist sense), which indicates a possible instability of IBF's when the sample size is small. An attractive solution to this problem, if M_i is nested in M_j, is to replace the averages by their expectations under M_j, evaluated at the parameter MLE under M_j. Since M_i is nested in M_j, the minimal training sample size is $m = k_j + 1$. Defining $p = k_j - k_i$ as before, the *expected arithmetic intrinsic Bayes factor* is given by

$$B_{ji}^{EAI} = B_{ji}^N \cdot \frac{1}{L}\sum_{l=1}^{L} E_{\hat{\beta}_j,\hat{\sigma}_j}^{M_j}[B_{ij}^N(l)] \tag{24}$$

$$= B_{ji}^N \cdot \frac{C^*}{L}\sum_{l=1}^{L} \frac{|X_j^t(l)X_j(l)|^{1/2}}{|X_i^t(l)X_i(l)|^{1/2}} E_{\hat{\beta}_j,\hat{\sigma}_j}^{M_j}\left[\frac{(R_j(l))^{(q_j+1)/2}}{(R_i(l))^{(q_i+p+1)/2}}\right], \tag{25}$$

where

$$C^* = \frac{\pi^{-p/2}}{2^{(q_j-q_i)/2}} \cdot \frac{\Gamma((q_i+p+1)/2)}{\Gamma((q_j+1)/2)} \tag{26}$$

and the expectation is under M_j with (β_j, σ_j) replaced by the standard estimates $(\hat{\beta}_j, \hat{\sigma}_j)$, where $\hat{\sigma}_j^2 = R_j/(n-k_j)$.

Under mild conditions, the law of large numbers can be applied to show that $(B_{ji}^{AI}/B_{ji}^{EAI}) \to 1$, as the sample size grows under M_j. But since M_i is nested in M_j, convergence is also achieved when M_i is true. It is thus reasonable to use B_{ji}^{EAI} in place of B_{ji}^{AI}.

If the training samples can be chosen to be exchangeable, so that the ratios of determinants in (25) are equal, then the summation in (25) disappears and one has a computationally quite simple IBF.

One can define the geometric analogue of (25), but we will not consider this because the expectations cannot be given in closed form. For nonnested models, one can define the expected arithmetic intrinsic Bayes factors w.r.t. an encompassing model, and proceed as in Section 4.2.

The expectation in (25) can be evaluated in closed form for the default priors we consider. We present the results here; see the Appendix for derivations. The answers are in terms of Kummer's function, $M(a,b,c)$ (see the Appendix and Abramowitz and Stegun, 1970, Chapter 13). Note that Kummer's function is available in standard software packages such as Mathematica. We evaluate the expectations under M_j at β_j. To use (25), one simply replaces (β_j, σ_j) in the expressions below by $(\hat{\beta}_j, \hat{\sigma}_j)$. The expectations depend on the "noncentrality" parameters

$$\lambda(l) = \sigma_j^{-2}\beta_j^t X_j^t(l)(I - X_i(l)[X_i^t(l)X_i(l)]^{-1}X_i^t(l))X_j(l)\beta_j. \qquad (27)$$

Using the Jeffreys prior: Here $q_i = k_i$ and $q_j = k_j$, and the expectation in (25) becomes

$$E^{M_2}_{\beta_j,\sigma_j}\left[\left(\frac{R_j(l)}{R_i(l)}\right)^{(k_j+1)/2}\right] = C^{**}e^{-\lambda(l)/2}M\left(\frac{p+1}{2}, \frac{p+k_j+2}{2}, \frac{\lambda(l)}{2}\right), \quad (28)$$

where

$$C^{**} = \frac{\Gamma((k_j+2)/2)\Gamma((p+1)/2)}{\Gamma((k_j+p+2)/2)\Gamma(1/2)}.$$

Using the Modified Jeffreys prior: Here $q_i = 0$ and $q_j = k_j - k_i = p$, and the expectation in (25) becomes

$$E\left[\left(\frac{R_j(l)}{R_i(l)}\right)^{(p+1)/2}\right] = 2^{-p}e^{-\lambda(l)/2}M(\frac{p+1}{2}, p+1, \frac{\lambda(l)}{2}) \qquad (29)$$

$$= \begin{cases} \frac{1}{\lambda(l)}[1 - e^{-\lambda(l)/2}] & \text{if } p = 1 \\ \frac{3}{\lambda(l)^2}[(1 - \frac{4}{\lambda(l)}) + (1 + \frac{4}{\lambda(l)})e^{-\lambda(l)/2}] & \text{if } p = 3 \\ \frac{15}{\lambda(l)^3}[(1 - \frac{12}{\lambda(l)} + \frac{48}{\lambda(l)^2}) - (1 + \frac{12}{\lambda(l)} + \frac{48}{\lambda(l)^2})e^{-\lambda(l)/2}] & \text{if } p = 5. \end{cases}$$

Examples 3 and 4 Continued: For the Dyestuff and Simulated data, the expected AIBF yields 3.82084 and 0.00308, respectively. These values are reasonably close to the values appearing in Tables 3 and 4, which are 4.20 and 0.0038. Note that the expected AIBF is here computationally far cheaper than the corresponding AIBF. The latter involves 187,500 training samples as compared with only six different terms in the former (corresponding to six different non-centrality parameters). Of course, one can always just compute the AIBF using a moderate subset of training samples.

Using the Reference prior: Here $q_i = q_j = 0$, and the expectation in (25) becomes

$$E\left[\frac{(R_j(l))^{1/2}}{(R_i(l))^{(p+1)/2}}\right] = \frac{\exp\{-\lambda(l)/2\}}{\sigma_j^p 2^{p/2}\Gamma((p+2)/2)} M(\frac{1}{2}, \frac{p+2}{2}, \frac{\lambda(l)}{2}). \qquad (30)$$

Known σ^2: For the case where the σ_j^2 are known and equal σ^2, and $\pi_j^N(\beta_j) = 1$, the analogue of (25) is (see (16))

$$B_{ji}^{EAI} = \frac{|X_i^t X_i|^{1/2}}{|X_j^t X_j|^{1/2}} \cdot \exp\left\{-\frac{1}{2\sigma^2}(R_j - R_i)\right\} \cdot \frac{1}{L}\sum_{l=1}^{L} \frac{|X_j^t(l)X_j(l)|^{1/2}}{|X_i^t(l)X_i(l)|^{1/2}} \cdot E_{\hat{\beta}_j}^{M_j}\left[\exp\left\{-\frac{1}{2\sigma^2}(R_j(l) - R_i(l))\right\}\right]. \qquad (31)$$

Here, the expectation can be evaluated (see the Appendix) as

$$E\left[\exp\left\{-\frac{1}{2\sigma^2}(R_j(l) - R_i(l))\right\}\right] = 2^{-p/2}\exp\{-\lambda(l)/4\}. \qquad (32)$$

4.5 Intrinsic Priors for Nested Linear Models

In Berger and Pericchi (1993), it was argued that arithmetic IBF's for nested models typically correspond to actual Bayes factors, with respect to reason-

able prior distributions. This was advanced as one of the strong arguments for use of arithmetic IBF's. The prior distribution that would correspond to an arithmetic IBF was called the *intrinsic prior.*

The formal definition of an intrinsic prior was based on an asymptotic analysis, utilizing the following approximation to a Bayes factor:

$$B_{ji} = B_{ji}^N \cdot \frac{\pi_j(\hat{\theta}_j)\pi_i^N(\hat{\theta}_i)}{\pi_j^N(\hat{\theta}_j)\pi_i(\hat{\theta}_i)}(1 + o(1)); \tag{33}$$

here π_i^N and π_j^N are the noninformative priors used to compute B_{ji}^N, π_i and π_j are the actual priors that determine B_{ji}, and $\hat{\theta}_i = (\hat{\beta}_i, \hat{\sigma}_i)$ and $\hat{\theta}_j = (\hat{\beta}_j, \hat{\sigma}_j)$ are the MLEs under M_i and M_j. The idea was to compare (3) with (33), and observe that, if the IBF in (3) corresponds to an actual Bayes factor, then

$$\frac{1}{L}\sum_{l=1}^{L} B_{ij}^N(l) \cong \frac{\pi_j(\hat{\theta}_j)\pi_i^N(\hat{\theta}_i)}{\pi_j^N(\hat{\theta}_j)\pi_i(\hat{\theta}_i)}.$$

For nested models, it was argued in Berger and Pericchi (1993) that this holds asymptotically if

$$\lim_{L\to\infty} \frac{1}{L}\sum_{l=1}^{L} B_{ij}^N(l) = \frac{\pi_j(\theta_j)\pi_i^N(\psi(\theta_j))}{\pi_j^N(\theta_j)\pi_i(\psi(\theta_j))}, \tag{34}$$

where it is assumed that $\hat{\theta}_i \to \psi(\theta_j)$ asymptotically, *under model* M_j.

Model M_i will be said to be nested in M_j if X_i consists of a subset of the columns of X_j. (More general types of nesting can be reduced to this by transformation.) In fact, we will assume that the covariates have been ordered so that $X_j = (X_i X^*)$ (the concatenation of the two matrices, not the product). Writing $\beta_j^t = (\beta_0^t, \beta^{*t})$, it is convenient to write $\pi_j(\theta_j) = \pi_j(\beta_j, \sigma_j)$ as

$$\pi_j(\beta_j, \sigma_j) = \pi_j^1(\beta^*|\beta_0, \sigma_j) \cdot \pi_j^2(\beta_0, \sigma_j). \tag{35}$$

Note that (β_0, σ_j) is the analogue, under M_j, of (β_i, σ_i) under M_i. We do not make the common mistake of identifying these parameters as being equal but, as they are related "nuisance" location-scale parameters, it is natural to assign them the same noninformative prior. We in fact will *choose* this common prior to be the same as $\pi_i^N(\beta_i, \sigma_i) = \sigma_i^{-(1+q_i)}$, so that

$$\pi_i(\beta_i, \sigma_i) = \sigma_i^{-(1+q_i)}, \quad \pi_j^2(\beta_0, \sigma_j) = \sigma_j^{-(1+q_i)}. \tag{36}$$

If (β_i, σ_i) and (β_0, σ_j) really were the same parameters, this choice would be noncontroversial. As they are not necessarily the same parameters, however, it could be argued that π_i and π_j^2 may not be properly "calibrated." If, however, $q_i = 0$ (i.e., the original π_i^N is the reference prior), then π_i and π_j^2 are themselves the reference priors, and we saw in Section 3 that this seems to provide a type of predictive "calibration" for *any* location-scale models. Thus our argument that IBF's correspond to sensible real Bayes factors is strongest if the IBF is defined for reference π_i^N, which occurs in either the "reference prior case" or the "modified Jeffreys prior case." (In fact, we will see that an "adjustment" of $\pi_j^2(\beta_0, \sigma_j)$ is needed for the Jeffreys prior case.)

Using (35) and (36) and recalling that $\pi_j^N(\beta_j, \sigma_j) = \sigma_j^{-(1+q_j)}$, (34) can be rewritten

$$\pi_j^1(\beta^* | \beta_0, \sigma_j) = \sigma_j^{(q_i - q_j)} \cdot \lim_{L \to \infty} \frac{1}{L} \sum_{l=1}^{L} B_{ij}^N(l). \tag{37}$$

Unfortunately there is ambiguity in taking this limit when the y_i are not exchangeable, but one can consider replications of the full y (i.e., imagine i.i.d. vector replications of Y from M_2, with the given X_2). Then, under mild conditions, the limit in (37) can be replaced by expectation, yielding

the final defining equation

$$\pi_j^1(\beta^*|\beta_0,\sigma_j) = \sigma_j^{(q_i-q_j)} \cdot \frac{1}{L}\sum_{l=1}^{L} E^{M_2}_{(\beta_j,\sigma_j)}[B_{ij}^N(l)]. \tag{38}$$

Note that expectations in (38) were computed in Section 4.4; indeed, one can see that the above argument would also apply to providing an "intrinsic prior" rationale for the expected intrinsic Bayes factor. Using the expressions in Section 4.4, the "intrinsic priors" in (38) can be written as follows. (We also include the result for the known variance case; the analogue of (38) for this case is easy to derive.)

Unknown σ_i^2 and σ_j^2:

$$\pi_j^1(\beta^*|\beta_0,\sigma_j) = \frac{\sigma_j^{(q_i-q_j)}C^*}{L} \cdot \sum_{l=1}^{L} \frac{|X_j^t(l)X_j(l)|^{1/2}}{|X_i^t(l)X_i(l)|^{1/2}} \cdot \psi(\lambda(l),\sigma_j), \tag{39}$$

where C^* is defined in (26) and $\psi(\lambda(l),\sigma_j)$ is either (28), (29), or (30), depending on the default prior used. From (27), it is straightforward to show, in the nested case, that

$$\lambda(l) = \sigma_j^{-2}\beta^{*t}X^{*t}(l)(I - X_i(l)[X_i^t(l)X_i(l)]^{-1}X_i^t(l))X^*(l)\beta^*. \tag{40}$$

Known $\sigma_i^2 = \sigma_j^2 = \sigma^2$: Defining $p = k_j - k_i$,

$$\pi_j^1(\beta^*|\beta_0) = \frac{1}{(4\pi\sigma^2)^{p/2}} \cdot \frac{1}{L}\sum_{l=1}^{L} \frac{|X_j^t(l)X_j(l)|^{1/2}}{|X_i^t(l)X_i(l)|^{1/2}} \cdot \exp\{-\lambda(l)/4\}. \tag{41}$$

Of course, we have not yet answered the big question: is $\pi_j^1(\beta^*|\beta_0,\sigma_j)$ a proper distribution? If so, we have established the Bayesian correspondence of IBF's.

Consider, first, the case of known $\sigma_i^2 = \sigma_j^2 = \sigma^2$. It is straightforward to show that

$$\Sigma(l) \equiv (X^*(l)^t(I - X_i(l)[X_i^t(l)X_i(l)]^{-1}X_i^t(l))X^*(l))^{-1}\sigma^2 \tag{42}$$

has determinant

$$|\Sigma(l)| = \sigma^{2p}|X_i^t(l)X_i(l)|/|X_j^t(l)X_j(l)|.$$

Hence (41) can be written

$$\pi_j^1(\beta^*|\beta_0) = \frac{1}{L}\sum_{l=1}^{L}\pi_l(\beta^*), \tag{43}$$

where the π_l are $\mathcal{N}_p(0, \frac{1}{2}\Sigma(l))$ distributions. Thus π_j^1 is a mixture of normals, and is trivially a proper distribution.

Establishing that (39) defines a proper distribution is considerably more difficult, and is considered in the Appendix. Interestingly, (39) is proper for the reference prior and modified Jeffreys prior cases, but is *not* for the Jeffreys prior case; it is off by a constant, C_0 (see the Appendix). This suggests that our choice of $\pi_i(\beta_i, \sigma_i) = \sigma_i^{-(1+q_i)}$ and $\pi_j^2(\beta_0, \sigma_j) = \sigma_j^{-(1+q_i)}$ for the Jeffreys prior IBF are not properly "calibrated"; choosing $\pi_j^2(\beta_0, \sigma_j) = C_0^{-1}\sigma_j^{-(1+q_i)}$ would ensure that $\pi_j^1(\beta^*|\beta_0, \sigma_j)$ is then proper, and is hence perhaps the correct calibration of π_j^2.

The nature of $\pi_j^1(\beta^*|\beta^0, \sigma_j)$ is of considerable interest in providing insight into the behavior of the associated IBF's. In the known variance case, $\pi_j^1(\beta^*|\beta^0)$ is rather simple, and clearly has mean 0 and covariance

$$\Sigma^* = \frac{1}{2L}\sum_{l=1}^{L}\Sigma(l). \tag{44}$$

Note that, in balanced cases where the $\Sigma(l)$ are equal, $\pi_j^1(\beta^*|\beta_0)$ is just a single normal prior, and is similar to the prior used for model comparison by Zellner and Siow (1980). Seeing how Σ^* differs from the Zellner and Siow covariance matrix in unbalanced cases would be of considerable interest.

The behavior of $\pi_j^1(\beta^*|\beta_0, \sigma_j)$ in the unknown variance case is more difficult to ascertain. For the modified Jeffreys prior case and $p = (k_j - k_i)$ an odd integer, simple closed form expressions are available, as shown following (29). For instance, when $p = 1$, using (39) and (40) yields

$$\begin{aligned} \pi_j^1(\beta^*|\beta_0, \sigma_j) &= \frac{1}{\sqrt{2\pi\sigma_j^2}} \cdot \frac{1}{L} \sum_{l=1}^{L} \frac{|X_j^t(l)X_j(l)|^{1/2}}{|X_i^t(l)X_i(l)|^{1/2}} \cdot \frac{1}{\lambda(l)} (1 - e^{-\lambda(l)/2}) \\ &= \frac{1}{L} \sum_{l=1}^{L} \frac{1}{2\sqrt{\pi V(l)}} \cdot \frac{1}{(\beta^{*2}/V(l))} \cdot (1 - e^{-\beta^{*2}/V(l)}), \end{aligned} \quad (45)$$

where

$$V(l) = 2\sigma_j^2 / [X^*(l)^t (I - X_1(l)(X_1^t(l)X_1(l))^{-1} X_1^t(l)) X^*(l)].$$

Each of the densities in this mixture is very similar to a Cauchy $(0, \sqrt{V(l)})$ density (never differing by more than 15%). This Cauchy density is similar to that recommended by Jeffreys (1961) or Zellner and Siow (1980).

In general, it can be shown (for the reference and modified Jeffreys cases) that $\pi_j^1(\beta^*|\beta_0, \sigma_j)$ is a mixture of densities that behave like $\mathcal{T}_p(1, 0, \Sigma^*(l))$ densities: p-variate t-densities with 1 degree of freedom, location 0, and scale matrix

$$\Sigma^*(l) = 2\sigma_j^2 [X^*(l)^t (I - X_1(l)(X_1^t(l)X_1(l))^{-1} X_1^t(l)) X^*(l)]^{-1}.$$

The fact that the degree of freedom here is minimal, seems related to the fact that minimal training samples were used.

As a final comment, note that an analogous derivation of intrinsic priors for geometric IBF's can be performed. However, the analogous expressions for $\pi_j^1(\beta^*|\beta_0, \sigma_j)$ are considerably more involved, and also do not appear to be proper distributions.

5 Conclusions

(i) The IBF seems to be a general, Bayesianly justifiable tool for dealing with normal linear models and generalizations. In particular, we saw that it is effective in difficult scenarios of high colinearity and moderately small sample sizes, and for non-normal error structures (with fixed linear structure) and hierarchical models. Most other automatic Bayes factors are comparatively limited in applicability.

(ii) We recommend use of the GIBF only with reference priors ($q_i = q_j = 0$). It appears to be quite unstable for other priors and small sample sizes. The AIBF is fine for either reference or modified Jeffreys priors ($q_i = 0, q_j = k_j - k_i$). We do not recommend use of the AIBF with the Jeffreys prior ($q_i = k_i,\ q_j = k_j$), however.

(iii) For very small sample sizes, we recommend use of the expected intrinsic Bayes factors. Indeed, these are fine for any sample size, and are only slightly more difficult to use. (It is necessary to compute (29) or (30).)

(iv) If comparing multiple models, we recommend use of the encompassing model approach in Section 4.2. This approach is well-defined for linear models.

(v) The existence of intrinsic priors provides one of the central justifications of AIBF's. There are, in addition, several potential uses of the intrinsic priors. One is to provide insight into the comparison of AIBF's with other Bayesianly justifiable methods, such as that of Zellner and Siow (1980). Another possibility is to actually use the intrinsic prior as a default prior distribution to compute an actual Bayes factor for model comparison. Both of these uses deserve further exploration.

Appendix

For the technical proofs of results in Section 4, the resulting standard facts are repeatedly used; all notation is taken from Section 4.

(i) Under M_j,

$$\begin{aligned} W &= \frac{R_j(l)}{\sigma_j^2} \sim \chi_1^2, \\ V &= \frac{R_i(l) - R_j(l)}{\sigma_j^2} \sim \chi_p^2(\lambda(l)), \end{aligned}$$

where χ_ν^2 denotes the central chi-square distribution with ν degrees of freedom, and $\chi_p^2(\lambda(l))$ is the noncentral chi-square distribution with $p = k_j - k_i$ degrees of freedom and noncentrality parameter $\lambda(l)$. Also, W and V are independent.

(ii)

$$M(a,b,z) = \frac{\Gamma(b)}{\Gamma(a)} \sum_{j=0}^{\infty} \frac{\Gamma(a+j)}{\Gamma(b+j)} \cdot \frac{z^j}{j!}.$$

(iii)

$$E[h(\chi_\nu^2(\lambda))] = \sum_{j=0}^{\infty} \frac{(\lambda/2)^j \exp\{-\lambda/2\}}{j!} \cdot E[h(\chi_{\nu+2j}^2)].$$

(iv) With obvious abuse of notation,

$$E\left[\left(\frac{\chi_1^2}{\chi_1^2 + \chi_\nu^2}\right)^s\right] = \frac{\Gamma(s+1/2)\Gamma((\nu+1)/2)}{\Gamma(1/2)\Gamma(s+(\nu+1)/2)},$$

providing χ_1^2 and χ_ν^2 are independent.

(v)

$$E[\exp\{-\frac{1}{2}\chi_p^2(\lambda)\}] = 2^{-p/2} e^{-\lambda/4}.$$

(vi)

$$\frac{\Gamma(1+p/2)\Gamma((p+1)/2)}{\Gamma(1/2)\Gamma(p+1)} = 2^{-p}.$$

Proof of (28): Using, in order, Facts (i), (iii), and (iv), we obtain

$$E^{M_2}\left[\left(\frac{R_j(l)}{R_i(l)}\right)^{(k_j+1)/2}\right] = E\left[\left(\frac{W}{W+V}\right)^{(k_j+1)/2}\right]$$
$$= \sum_{j=0}^{\infty} \frac{(\lambda(l)/2)^j \exp\{-\lambda(l)/2\}}{j!} \cdot E\left[\left(\frac{\chi_1^2}{\chi_1^2+\chi_{p+2j}^2}\right)^{(k_j+1)/2}\right]$$
$$= e^{-\lambda(l)/2} \sum_{j=0}^{\infty} \frac{(\lambda(l)/2)^j}{j!} \cdot \frac{\Gamma((k_j+2)/2)\Gamma((p+2j+1)/2)}{\Gamma(1/2)\Gamma((p+k_j+2j+2)/2)}.$$

Using Fact (ii), (28) follows immediately.

Proof of (29): Identical to that of (28), but also using Fact (vi). The explicit forms given for $p = 1, 3, 5$ follow from representations of M.

Proof of (30): Using, in order, Facts (i) and (iii) yields

$$E\left[\frac{(R_j(l))^{1/2}}{(R_i(l))^{(p+1)/2}}\right] = \sum_{j=0}^{\infty} \frac{(\lambda(l)/2)^j \exp\{-\lambda(l)/2\}}{j!\sigma_j^p} \cdot E\left[\frac{\sqrt{\chi_1^2}}{(\chi_1^2+\chi_{p+2j}^2)^{(p+1)/2}}\right].$$

Defining

$$c_j^{-1} = 2^{(p+2j+1)/2}\Gamma(1/2)\Gamma((p+2j)/2),$$

$$E\left[\frac{\sqrt{\chi_1^2}}{(\chi_1^2+\chi_{p+2j}^2)^{(p+1)/2}}\right] = \int_0^\infty \int_0^\infty \frac{c_j y^{(j-1+p/2)} e^{-(x+y)/2}}{(x+y)^{(p+1)/2}} dxdy$$
$$= c_j 2^{(j+1/2)}\Gamma(j+1/2)/(j+p/2).$$

Algebra, together with Fact (ii), yields the result.

Proof of (31): This is Fact (v).

Lemma A1. *For the reference prior and modified Jeffreys prior cases, $\pi_j^1(\beta^*|\beta^0, \sigma_j)$ in (39) is a proper density. For the Jeffreys prior case,*

$$\int \pi_j^1(\beta^*|\beta^0, \sigma_j)d\beta^* = C_0 = \frac{\Gamma((k_i+1)/2)\Gamma((p+1)/2)}{\Gamma((k_j+1)/2)\Gamma(1/2)}.$$

Proof. For the Jeffreys prior case,

$$\pi_j^1(\beta^*|\beta^0, \sigma_j) = \frac{1}{L}\sum_{l=1}^{L} g_l(\beta^*),$$

$$g_l(\beta^*) = \frac{C^{**}}{(2\pi\sigma_j^2)^{p/2}} \cdot \frac{|X_j^t(l)X_j(l)|^{1/2}}{|X_i^t(l)X_i(l)|^{1/2}} \cdot e^{-\lambda(l)/2} \cdot M\left(\frac{p+1}{2}, \frac{p+k_j+2}{2}, \frac{\lambda(l)}{2}\right).$$

The transformation $\beta^* \to \lambda(l)$ has Jacobian

$$\frac{|X_i^t(l)X_i(l)|^{1/2}}{|X_j^t(l)X_j(l)|^{1/2}} \cdot \frac{(\pi\sigma_j^2)^{p/2}}{\Gamma(p/2)} \cdot \lambda(l)^{(p-2)/2},$$

so that (writing $\lambda = \lambda(l)$)

$$\int g_l(\beta^*)d\beta^* = \frac{C^{**}}{2^{p/2}\Gamma(p/2)} \int_0^\infty \lambda^{(p-2)/2} e^{-\lambda/2} M(\frac{p+1}{2}, \frac{p+k_j+2}{2}, \frac{\lambda}{2})d\lambda.$$

Using Fact (ii), and integrating term by term yields

$$\begin{aligned}
&\int_0^\infty \lambda^{(p-2)/2} e^{-\lambda/2} M(\frac{p+1}{2}, \frac{p+k_j+2}{2}, \frac{\lambda}{2})d\lambda \\
&= \frac{\Gamma((p+k_j+2)/2)}{\Gamma((p+1)/2)} \sum_{j=0}^{\infty} \frac{\Gamma(j+(p+1)/2)}{\Gamma(j+(p+k_j+2)/2)(j!)2^j} \cdot \int_0^\infty \lambda^{(j-1+p/2)} e^{-\lambda/2} d\lambda \\
&= \frac{\Gamma((p+k_j+2)/2)}{\Gamma((p+1)/2)} \sum_{j=0}^{\infty} \frac{\Gamma(j+(p+1)/2)\Gamma(j+p/2)2^{p/2}}{\Gamma(j+(p+k_j+2)/2)(j!)} \\
&= 2^{p/2}\Gamma(p/2)F(\frac{p+1}{2}, \frac{p}{2}, \frac{p+k_j+2}{2}, 1) \\
&= \frac{2^{p/2}\Gamma(p/2)\Gamma((p+k_j+2)/2)\Gamma((k_i+1)/2)}{\Gamma((k_j+1)/2)\Gamma((k_j+2)/2)},
\end{aligned}$$

where F is the hypergeometric function, and we have used 15.1.20 of Abramowitz and Stegun (1970). Combining terms and simplifying yields C_0.

The identical argument works for the reference and modified Jeffreys prior cases, but now the integral equals 1.

References

Abramowitz, M. and Stegun, I. (1970). *Handbook of Mathematical Functions.* National Bureau of Standards Applied Mathematics Series 55.

Allenby, G. M. (1990). Hypothesis testing with scanner data: the advantage of Bayesian methods. *J. of Marketing Research* **27**, 379–389.

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. O. and Pericchi, L. R. (1993). The intrinsic Bayes factor for model selection and prediction. Technical Report #93-43C, Department of Statistics, Purdue University.

Berger, J., Pericchi, L. and Varshavsky, J. (1994). An identity for linear and invariant models, with application to non-Gaussian model selection. Technical Report, Purdue University.

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.

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 propagation 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.

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. *J. Roy. Statist. Soc.* **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, 147–167.

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.

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.

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. (1993). Bayesian model selection without assuming a "true" model. (Work in progress).

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.

Smith, A. F. M. (1994). In the discussion of O'Hagan, A., "Fractional Bayes factors for model comparisons." To appear in *J. Roy. Statist. Soc.* **B**, 56.

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.

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.