On the Justification of
Default and Intrinsic Bayes Factors*

by

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

Technical Report #95-18C

Department of Statistics
Purdue University

April 1995

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

ABSTRACT:

In Bayesian model selection or hypothesis testing, it is difficult to develop default Bayes factors, since (improper) noninformative priors cannot typically be used. In developing such default Bayes factors, we feel that it is important to keep several principles in mind. The first is that the default Bayes factor should correspond, in some sense, to an actual Bayes factor with a (sensible) prior, which we call an intrinsic prior. The second principle is that such priors should be properly calibrated across models, in the sense of being "predictively matched." These notions will be described and illustrated, primarily using examples involving the intrinsic Bayes factor, a recently proposed default Bayes factor. It will be seen that intrinsic Bayes factors seem to correspond to actual Bayes factors with proper priors, at least for nested model scenarios. The corresponding intrinsic priors are specifically given for the normal linear model.

1. INTRODUCTION

There are a number of compelling reasons to consider use of Bayes factors in model selection and hypothesis testing. There are also a number of compelling reasons for development of 'default' or 'automatic' Bayes factors, especially in the preliminary stages of modelling when careful specification of subjective priors for all models under consideration is typically not feasible. For discussion of these issues, see Jeffreys (1961), Edwards, Lindman, and Savage (1963), Berger and Sellke (1987), Berger and Delampady (1987), Draper (1995), Kass and Raftery (1995), Madigan and Raftery (1995), and Berger and Pericchi (1993).

There are two main difficulties with the development of default Bayes factors. The first is the well-known difficulty that, when the models or hypotheses have parameter spaces of differing dimension, one cannot use only (improper) noninformative priors for computing the Bayes factors; improper priors are unaffected by multiplication by an arbitrary positive constant, but such arbitrary constants directly affect Bayes factors. The second difficulty in developing default (or even subjective) Bayes factors is that parameters do not typically have meaning independent of the model. Although this difficulty is also well-known, it is less often discussed, and is of enough importance to deserve emphasis through an example.

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

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

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

The first approach to overcoming these difficulties was that proposed by Jeffreys (1961). He proposed the use of orthogonal parameters (i.e., parameters for which the corresponding expected Fisher information matrix is diagonal, or block diagonal if the parameters are to be handled in blocks), presumably in an effort to overcome the type of difficulty illustrated in the above example. That use of orthogonal parameters overcomes this difficulty is a belief in the statistical folklore and is undoubtedly true in certain asymptotic senses, but we have not seen a clear Bayesian argument as to why this should be so. The other problems with orthogonalization are (i) it is frequently extremely difficult or impossible to find orthogonal parameters, and (ii) orthogonal parameters typically have no intuitive meaning, and so models expressed in terms of subsets of orthogonal parameters often have no meaning. Nevertheless, the use of orthogonal parameters, when possible, appears to be a quite effective tool. Jeffreys (1961) provides a number of convincing examples. For a modern successful use of the idea, see Clyde and Parmigiani (1995).

Jeffreys (1961) dealt with the issue of indeterminacy of noninformative priors by (i) only using noninformative priors for common (orthogonal) parameters in the models, so that the arbitrary multiplicative constant for the priors would cancel in all Bayes factors, and (ii) using default proper priors for parameters that would occur in one model but not the other. He presented arguments for appropriate default proper priors, but mostly on a case-by-case basis. This line of development has been successfully followed by several others, for instance by Zellner and Siow (1980).

Although use of particular default proper priors can be criticized for being somewhat arbitrary, one cannot be too demanding here. Any automatic procedure is going to contain some quite arbitrary features, and we feel that the Jeffreys approach is among those with the least objectionable arbitrary feartures. Indeed, we feel that any default Bayes factor should correspond (in some sense, perhaps asymptotic) to use of an actual Bayes factor with some proper prior distribution; if not, the Bayes factor is not compatible with Bayesian reasoning, and we feel that it is then probably uninterpretable. Furthermore, we feel that the best method of evaluating such 'good' default Bayes factors is to find the prior distribution to which they correspond, which we call the intrinsic prior, and to determine whether or not this distribution is sensible. In Section 3 we carry out this program for the intrinsic Bayes factor, a default Bayes factor that was proposed in Berger and Pericchi (1993, 1995).

Another general approach to overcoming the difficulties discussed above is the idea of selecting prior distributions that are somehow "matched" across models. Suzuki (1983) proposed matching the entropies of the prior distributions, or perhaps matching the entropies of an intermediate sequence of priors that converge to noninformative priors upon renormalization.

Perhaps more natural is to attempt to choose priors to match predictives. The underlying motivation is the foundational Bayesian view that one should concentrate on predictive distributions of observables; models and priors are, at best, convenient abstractions. According to this perspective, it is a predictive distribution $m(\mathbf{y})$ that describes reality, where $\mathbf{y}$ is a variable of predictive interest. We can choose to represent $m(\mathbf{y})$ as $m_i(\mathbf{y}) = \int f_i(\mathbf{y} \mid \theta_i)\pi_i(\theta_i)d\theta_i$, where f_i is a model and π_i a prior, but these are merely a convenient abstraction.

From this perspective, if one is comparing models $M_1 : f_1$ versus $M_2 : f_2$, then the priors π_1 and π_2 should be chosen so that $m_1(\mathbf{y})$ and $m_2(\mathbf{y})$ are as close as possible. Thus we think of π_1 and π_2 as being properly calibrated if, when filtered through the models M_1

and M_2, they yield similar predictives. This could be assessed by defining some distance measure, $d(m_1, m_2)$, and calling π_1 and π_2 calibrated if $d(m_1, m_2)$ is small. We explore this formal approach elsewhere, here being content simply with showing that intrinsic priors which arise from intrinsic Bayes factors seem to be well-calibrated.

One key issue in operationalizing this idea is that of choosing the variable $\mathbf{y}$ at which a predictive match is desired. It seems natural, in the exchangeable case, to choose $\mathbf{y}$ to be an "imaginary" minimal training sample, which is typically the smallest set of observations for which the various model parameters are identifiable.

The ideas here are related to ideas of elicitation through predictives (cf, Kadane, et.al., 1980). Also, a similar use of predictive matching to define priors for model selection can be found in Laud and Ibrahim (1993) and Ibrahim and Laud (1994).

2. THE INTRINSIC BAYES FACTOR

2.1 Definition of IBF's

Suppose that we are comparing q models for the data $\mathbf{x}$,

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

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

$$m_i^N(\mathbf{x}) = \int f_i(\mathbf{x}|\theta_\mathbf{i})\pi_i^N(\theta_\mathbf{i})d\theta_\mathbf{i}.$$

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

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

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

where

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

are the Bayes factors that would be obtained for the full data $\mathbf{x}$ and training sample $\mathbf{x}(l)$, 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 $\mathbf{x}(l)$. To eliminate this dependence and to increase stability, a natural idea is to average the $B_{ji}(l)$ over all possible training samples $\mathbf{x}(l)$, $l = 1, \ldots, L$. 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. We then *define* B_{ij}^{AI} by

$$
B_{ij}^{AI} = 1/B_{ji}^{AI}. \tag{5}
$$

2.2 The IBF for Two Non-Nested Examples

The following two scenarios will be used to illustrate several of the issues raised in the Introduction.

The IBF for Fixed Design Linear Models

Assume that we are considering the Linear Models

$$M_j : \mathbf{Y} = \mathbf{X}\boldsymbol{\beta}_j + \sigma_j \boldsymbol{\varepsilon}_j, \tag{6}$$

for $j = 1, \ldots, q$ alternative error models $\boldsymbol{\varepsilon}_j \sim g_j$; here $\mathbf{Y}$ is $n \times 1$, $\mathbf{X}$ is $n \times k, \boldsymbol{\beta}_j \epsilon R^k$ is $k \times 1, \sigma_j > 0$, and $\boldsymbol{\varepsilon}_j$ is $n \times 1$. Note that the design matrix, $\mathbf{X}$, is assumed to be fixed across models. We label the unknown $\boldsymbol{\beta}$ and σ by j, so as to emphasize that parameters can have different meanings within different models. We will use reference default priors, $\pi_j^N(\boldsymbol{\beta}_j, \sigma_j) = 1/\sigma_j$. A minimal training sample can be seen to be any $(k+1)$-vector $\mathbf{y}(l)$ with corresponding sub-matrix $\mathbf{X}(l)$, of $\mathbf{X}$, such that $\mathbf{X}^t(l)\mathbf{X}(l)$ is nonsingular. Let $|\mathbf{A}|$ denote the determinant of a matrix $\mathbf{A}$.

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

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

Lemma 1 is established in Berger, Pericchi and Varshavsky (1994). It is a quite surprising result because $m_j(\mathbf{y}(l))$ 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.

This provides our first illustration of the "predictive matching" idea described in the Introduction. Indeed, Lemma 1 suggests that the reference prior is properly calibrated for comparison of *any* models of the form (6), in that the predictives for a minimal sample are then identical. (Note that this will not be the case if other noninformative priors, e.g., the Jeffreys prior, are used.) 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))

$$B_{ji}^{AI} = B_{ji}^{GI} = B_{ji}^{N} \tag{8}$$

Comparison of Exponential and Lognormal Models

Suppose $X_1, \ldots, X_n$ are i.i.d. according to one of the following models:

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

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

For M_1 and M_2, the standard noninformative priors are $\pi_1^N(\theta_1) = 1/\theta_1$ and $\pi_2^N(\mu, \sigma) = 1/\sigma$. Calculation yields, for $\mathbf{x} = (x_1, \ldots, x_n)$,

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

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 $\mathbf{x}(l) = (x_i, x_j), x_i \neq x_j$, so that

$$m_1^N(\mathbf{x}(l)) = \frac{1}{(x_i + x_j)^2}, \qquad m_2^N(\mathbf{x}(l)) = \frac{1}{2x_i x_j |\log(x_i/x_j)|}.$$

The IBF can thus be computed using (2) and (3) or (4). We defer discussion of this example to Section 3.3.

2.3 The IBF for the Normal Linear Model

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

$$M_j : \mathbf{Y} = \mathbf{X_j}\boldsymbol{\beta}_\mathbf{j} + \boldsymbol{\varepsilon}_\mathbf{j}, \qquad \boldsymbol{\varepsilon}_\mathbf{j} \sim \mathcal{N}_\mathbf{n}(\mathbf{0}, \sigma_\mathbf{j}^\mathbf{2}\mathbf{I_n}),$$

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

$$\hat{\boldsymbol{\beta}}_j = (\mathbf{X}_j^t \mathbf{X}_j)^{-1}\mathbf{X}_j^t \mathbf{y} \ \textit{and}\ R_j = |\mathbf{y} - \mathbf{X_j}\hat{\boldsymbol{\beta}}_\mathbf{j}|^\mathbf{2}$$

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

We will consider default priors of the form

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

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 $\mathbf{y}(l)$, with corresponding design matrices $\mathbf{X}_j(l)$ (under the M_j), is a sample of size $m = \max\{k_j\} + 1$ such that all $(\mathbf{X}_j^t(l)\mathbf{X}_j(l))$ 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

$$B_{ji}^N = \frac{\pi^{(k_j-k_i)/2}}{2^{(q_i-q_j)/2}} \cdot \frac{\Gamma((n-k_j+q_j)/2)}{\Gamma((n-k_i+q_i)/2)} \cdot \frac{|\mathbf{X}_i^t\mathbf{X}_i|^{1/2}}{|\mathbf{X}_j^t\mathbf{X}_j|^{1/2}} \cdot \frac{R_i^{(n-k_i+q_i)/2}}{R_j^{(n-k_j+q_j)/2}}, \tag{10}$$

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

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

Inserting these expressions in (1) results in the following arithmetic IBF's in (3) 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:

$$B_{ji}^{AI} = \frac{|\mathbf{X}_i^t\mathbf{X}_i|^{1/2}}{|\mathbf{X}_j^t\mathbf{X}_j|^{1/2}} \cdot \left(\frac{R_i}{R_j}\right)^{n/2} \cdot \frac{1}{L}\sum_{l=1}^{L} \frac{|\mathbf{X}_j^t(l)\mathbf{X}_j(l)|^{1/2}}{|\mathbf{X}_i^t(l)\mathbf{X}_i(l)|^{1/2}} \cdot \left(\frac{R_j(l)}{R_i(l)}\right)^{m/2}. \tag{12}$$

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

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

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

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

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

For Known σ^2: If the σ_j^2 are known and equal σ^2, and the $\pi_j^N(\boldsymbol{\beta}_j) = 1$, then

$$B_{ji}^N = (2\pi\sigma^2)^{(k_j-k_i)/2} \cdot \frac{|\mathbf{X}_i^t \mathbf{X}_i|^{1/2}}{|\mathbf{X}_j^t \mathbf{X}_j|^{1/2}} \cdot \exp\left\{-\frac{1}{2\sigma^2}(R_j - R_i)\right\}. \tag{16}$$

Here, a minimal training sample is a sample of size $m = \max\{k_j\}$ such that all $(\mathbf{X}_j^t(l)\mathbf{X}_j(l))$ are nonsingular, and $B_{ji}^N(l)$ is as in (16) with $\mathbf{X}_i$, $\mathbf{X}_j$, R_i, and R_j replaced by $\mathbf{X}_i(l)$, $\mathbf{X}_j(l)$, $R_i(l)$, and $R_j(l)$. Thus the arithmetic intrinsic Bayes factor from (1) and (3) is

$$\begin{aligned} B_{ji}^{AI} &= \frac{|\mathbf{X}_i^t \mathbf{X}_i|^{1/2}}{|\mathbf{X}_j^t \mathbf{X}_j|^{1/2}} \cdot \exp\left\{-\frac{1}{2\sigma^2}(R_j - R_i)\right\} \\ &\quad \times \frac{1}{L} \sum_{l=1}^{L} \frac{|\mathbf{X}_j^t(l) \mathbf{X}_j(l)|^{1/2}}{|\mathbf{X}_i^t(l) \mathbf{X}_i(l)|^{1/2}} \cdot \exp\left\{-\frac{1}{2\sigma^2}(R_j(l) - R_i(l))\right\}. \end{aligned} \tag{17}$$

We discuss intrinsic priors and predictive matching for IBF's in Section 3.2.

3. INTRINSIC PRIORS FOR IBF's

3.1 Definition and Motivation.

Our major goal is to show that arithmetic IBF's correspond to actual Bayes factors with respect to what we call an *intrinsic prior*. We view the fact that IBFs tend to correspond to actual Bayes factors w.r.t. (sensible) intrinsic priors to be their strongest justification. Hence, determination of the intrinsic priors is of inherent theoretical interest, as well as providing the best insight into the behavior of IBFs.

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

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

The formal definition of an intrinsic prior, given in Berger and Pericchi (1993), was based on an asymptotic analysis, utilizing the following approximation to a Bayes factor:

$$B_{ji} = B_{ji}^N \cdot \frac{\pi_j(\hat{\theta}_j)\pi_i^N(\hat{\theta}_i)}{\pi_j^N(\hat{\theta}_j)\pi_i(\hat{\theta}_i)}(1 + o(1)); \tag{18}$$

here β_{ji} denotes the Bayes factor associated with priors π_j and π_i, π_i^N and π_j^N are the noninformative priors used to compute B_{ji}^N, and $\hat{\theta}_i$ and $\hat{\theta}_j$ are the MLEs under M_i and M_j. (The approximation in (18) holds more generally than the more standard Schwarz approximation that is discussed, for instance in Schwarz, 1978, Gelfand and Dey, 1994, and Kass and Raftery, 1995.)

To define intrinsic priors, equate (18) with (3) or (4), yielding

$$\frac{\pi_j(\hat{\theta}_j)\pi_i^N(\hat{\theta}_i)}{\pi_j^N(\hat{\theta}_j)\pi_i(\hat{\theta}_i)}(1 + o(1)) = \tilde{B}_{ij}^N, \tag{19}$$

where we define $\tilde{B}_{ij}^N$ to be either the arithmetic or geometric average of the $B_{ij}^N(l)$. We next need to make some assumptions about the limiting behavior of the quantities in (19). The following are typically satisfied, and will be assumed to hold as the sample size grows to infinity:

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

(ii) Under M_i, $\hat{\theta}_i \to \theta_i$, $\hat{\theta}_j \to \psi_j(\theta_i)$, and $\tilde{B}_{ij}^N \to B_i^*(\theta_i)$. (20)

(iii) For $k = i$ or $k = j$, the following limits exist:

$$B_k^*(\theta_k) = \begin{cases} \lim\limits_{L\to\infty} \quad E_{\theta_k}^{M_k}\left[\frac{1}{L}\sum\limits_{l=1}^{L} B_{ij}^N(l)\right] & \text{arithmetic case} \\ \lim\limits_{L\to\infty} \quad \exp\left\{E_{\theta_k}^{M_k}\left[\frac{1}{L}\sum\limits_{l=1}^{L} \log B_{ij}^N(l)\right]\right\} & \text{geometric case;} \end{cases} \tag{21}$$

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

Passing to the limit in (19), 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)

$$\frac{\pi_j^I(\theta_j)\pi_i^N(\psi_i(\theta_j))}{\pi_j^N(\theta_j)\pi_i^I(\psi_i(\theta_j))} = B_j^*(\theta_j), \tag{22}$$

$$\frac{\pi_j^I(\psi_j(\theta_i))\pi_i^N(\theta_i)}{\pi_j^N(\psi_j(\theta_i))\pi_i^I(\theta_i)} = B_i^*(\theta_i). \tag{23}$$

The motivation, again, is that priors which satisfy (22) and (23) would yield answers which are asymptotically equivalent to use of the intrinsic Bayes factors. We note that solutions are not necessarily unique, do not necessarily exist, and are not necessarily proper (cf, Dmochowski, 1994).

As a simple example of the above ideas, consider the fixed design linear model from Section 2.2. It is clear that $B_j^*(\theta_j) = B_i^*(\theta_i) = 1$; it follows trivially that solutions to (22) and (23) are given by

$$\pi_k^I(\theta_k) = \pi_k^N(\theta_k), \quad k = i, j.$$

Thus the intrinsic priors are merely the original noninformative priors. (Note, however, that this happens only because we used the reference noninformative priors; it would not happen, for instance, had the Jeffreys noninformative prior been used.)

3.2 Intrinsic Priors for Arithmetic IBF's in Nested Linear Models

Here we consider the normal linear model situation of Section 2.3. For the nested situation and use of arithmetic IBF's, it will be shown that proper intrinsic priors exist. Model M_i will be said to be nested in M_j if $\mathbf{X}_i$ consists of a subset of the columns of $\mathbf{X}_j$. (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 $\mathbf{X}_j = (\mathbf{X}_i \ \mathbf{X}^*)$ (the concatenation of the two matrices, not the product), and that the parameterization has been chosen so that $X_i^t X^* = 0$. Writing $\boldsymbol{\beta}_j^t = (\boldsymbol{\beta}_0^t, \boldsymbol{\beta}^{*t})$, it is convenient to write $\pi_j(\theta_j) = \pi_j(\boldsymbol{\beta}_j, \sigma_j)$ as

$$\pi_j(\boldsymbol{\beta}_j, \sigma_j) = \pi_j^1(\boldsymbol{\beta}^* | \boldsymbol{\beta}_0, \sigma_j) \cdot \pi_j^2(\boldsymbol{\beta}_0, \sigma_j). \tag{24}$$

Note that $(\boldsymbol{\beta}_0, \sigma_j)$ is the analogue, under M_j, of $(\boldsymbol{\beta}_i, \sigma_i)$ under M_i. As discussed in the Introduction, 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 $\pi_i^N(\boldsymbol{\beta}_i, \sigma_i) = \sigma_i^{-(1+q_i)}$, so that

$$\pi_i(\boldsymbol{\beta}_i, \sigma_i) = \sigma_i^{-(1+q_i)}, \quad \pi_j^2(\boldsymbol{\beta}_0, \sigma_j) = \sigma_j^{-(1+q_i)}. \tag{25}$$

If $(\boldsymbol{\beta}_i, \sigma_i)$ and $(\boldsymbol{\beta}_0, \sigma_j)$ 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 2.2 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 $\pi_j^2(\boldsymbol{\beta}_0, \sigma_j)$ is needed for the Jeffreys prior case.)

For this situation, the conditions in (20) can be shown to hold, with $\psi_i(\theta_j) = (\boldsymbol{\beta}_0, \sigma_j)$ and $\psi_j(\theta_i) = \theta_i = (\boldsymbol{\beta}_i, \sigma_i)$, providing the limits in (21) exist. There can be a certain ambiguity in defining this limit when the design matrix is unpatterned; we will thus assume that, as $n \to \infty$, the design matrix is patterned or replicated in such a way that the limits in (21) exist.

Next, observe that expectation in (21) under M_i is equivalent to expectation under M_j with $\theta_j = ((\boldsymbol{\beta}_0, 0), \sigma_j)$. It is then straightforward to show that (22) and (23) are both equivalent to the single equation.

$$\begin{aligned}\pi_j^1(\boldsymbol{\beta}^* | \boldsymbol{\beta}_0, \sigma_j) &= \sigma_j^{(q_i - q_j)} \cdot B_j^*(\theta_j) \\ &= \sigma_j^{(q_i - q_j)} \cdot \frac{1}{L} \sum_{l=1}^{L} E^{M_2}_{(\boldsymbol{\beta}_j, \sigma_j)}[B^N_{ij}(l)]. \end{aligned} \tag{26}$$

Interestingly, the expectations in (26) can be computed in closed form; see the Appendix. Using these expressions, the "intrinsic priors" in (26) can be written as follows. (We also include the result for the known variance case; the analogue of (26) for this case is easy to derive using Fact(v) from the Appendix.)

Unknown σ_i^2 and σ_j^2:

$$\pi_j^1(\boldsymbol{\beta}^* | \boldsymbol{\beta}_0, \sigma_j) = \frac{\sigma_j^{(q_i - q_j)} C^*}{L} \cdot \sum_{l=1}^{L} \frac{|\mathbf{X}_j^t(l)\mathbf{X}_j(l)|^{1/2}}{|\mathbf{X}_i^t(l)\mathbf{X}_i(l)|^{1/2}} \cdot \psi(\lambda(l), \sigma_j), \tag{27}$$

where C^* is defined in (A2) of the Appendix, $\psi(\lambda(l), \sigma_j)$ is either (A3), (A4), or (A5), depending on the default prior used, and

$$\lambda(l) = \sigma_j^{-2} \boldsymbol{\beta}^{*t} \mathbf{X}^{*t}(l)(\mathbf{I} - \mathbf{X}_i(l)[\mathbf{X}_i^t(l)\mathbf{X}_i(l)]^{-1}\mathbf{X}_i^t(l))\mathbf{X}^*(l)\boldsymbol{\beta}^*. \tag{28}$$

Known $\sigma_i^2 = \sigma_j^2 = \sigma^2$: Defining $p = k_j - k_i$ (recall that k_j is the dimension of β_j)

$$\pi_j^1(\boldsymbol{\beta}^* | \boldsymbol{\beta}_0) = \frac{1}{(4\pi\sigma^2)^{p/2}} \cdot \frac{1}{L} \sum_{l=1}^{L} \frac{|\mathbf{X}_j^t(l)\mathbf{X}_j(l)|^{1/2}}{|\mathbf{X}_i^t(l)\mathbf{X}_i(l)|^{1/2}} \cdot \exp\{-\lambda(l)/4\}. \tag{29}$$

Of course, we have not yet answered the big question: is $\pi_j^1(\boldsymbol{\beta}^* | \boldsymbol{\beta}_0, \sigma_j)$ 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

$$\boldsymbol{\Sigma}\,(l) \equiv (\mathbf{X}^*(l)^t(\mathbf{I} - \mathbf{X}_i(l)[\mathbf{X}_i^t(l)\mathbf{X}_i(l)]^{-1}\mathbf{X}_i^t(l))\mathbf{X}^*(l))^{-1}\sigma^2 \tag{30}$$

has determinant

$$|\mathbf{\Sigma}\,(l)| = \sigma^{2p}|\mathbf{X}_i^t(l)\mathbf{X}_i(l)|/|\mathbf{X}_j^t(l)\mathbf{X}_j(l)|.$$

Hence (29) can be written

$$\pi_j^1(\boldsymbol{\beta}^*|\boldsymbol{\beta}_0) = \frac{1}{L}\sum_{l=1}^{L}\pi_l(\boldsymbol{\beta}^*), \tag{31}$$

where the π_l are $\mathcal{N}_p(0, \frac{1}{2}\mathbf{\Sigma}\,(l))$ distributions. Thus π_j^1 is a mixture of normals, and is trivially a proper distribution. The following theorem deals with the unknown variance case.

Theorem 1. *For the reference prior and modified Jeffreys prior cases, $\pi_j^1(\boldsymbol{\beta}^*|\boldsymbol{\beta}_0, \sigma_j)$ in (27) is a proper density. For the Jeffreys prior case,*

$$\int \pi_j^1(\boldsymbol{\beta}^*|\boldsymbol{\beta}_0, \sigma_j)d\boldsymbol{\beta}^* = C_0 = \frac{\Gamma((k_i+1)/2)\Gamma((p+1)/2)}{\Gamma((k_j+1)/2)\Gamma(1/2)}.$$

Proof. We freely use notation and facts from the Appendix. For the Jeffreys prior case,

$$\pi_j^1(\boldsymbol{\beta}^*|\boldsymbol{\beta}_0, \sigma_j) = \frac{1}{L}\sum_{l=1}^{L} g_{_l}(\boldsymbol{\beta}^*),$$

$$g_{_l}(\boldsymbol{\beta}^*) = \frac{C^{**}}{(2\pi\sigma_j^2)^{p/2}} \cdot \frac{|\mathbf{X}_j^t(l)\mathbf{X}_j(l)|^{1/2}}{|\mathbf{X}_i^t(l)\mathbf{X}_i(l)|^{1/2}} \cdot e^{-\lambda(l)/2} \cdot M\left(\frac{p+1}{2}, \frac{p+k_j+2}{2}, \frac{\lambda(l)}{2}\right).$$

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

$$\frac{|\mathbf{X}_i^t(l)\mathbf{X}_i(l)|^{1/2}}{|\mathbf{X}_j^t(l)\mathbf{X}_j(l)|^{1/2}} \cdot \frac{(\pi\sigma_j^2)^{p/2}}{\Gamma(p/2)} \cdot \lambda(l)^{(p-2)/2},$$

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

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

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. $\square$

It is interesting that $\pi_j^1(\boldsymbol{\beta}^*|\boldsymbol{\beta}_0, \sigma_j)$ is proper for the reference prior and modified Jeffreys prior cases, but is *not* for the Jeffreys prior case. This suggests that our choice of $\pi_i(\boldsymbol{\beta}_i, \sigma_i) = \sigma_i^{-(1+q_i)}$ and $\pi_j^2(\boldsymbol{\beta}_0, \sigma_j) = \sigma_j^{-(1+q_i)}$ for the Jeffreys prior IBF are not properly "calibrated"; choosing $\pi_j^2(\boldsymbol{\beta}_0, \sigma_j) = C_0^{-1}\sigma_j^{-(1+q_i)}$ would ensure that $\pi_j^1(\boldsymbol{\beta}^*|\boldsymbol{\beta}_0, \sigma_j)$ is then proper, and is hence perhaps the correct calibration of π_j^2.

The nature of $\pi_j^1(\boldsymbol{\beta}^*|\boldsymbol{\beta}_0, \sigma_j)$ is of considerable interest in providing insight into the behavior of the associated IBF's. In the known variance case, $\pi_j^1(\boldsymbol{\beta}^*|\boldsymbol{\beta}_0)$ is rather simple, and clearly has mean 0 and covariance

$$\boldsymbol{\Sigma}^* = \frac{1}{2L}\sum_{l=1}^{L}\boldsymbol{\Sigma}\,(l). \tag{32}$$

Note that, in balanced cases where the $\boldsymbol{\Sigma}\,(l)$ are equal, $\pi_j^1(\boldsymbol{\beta}^*|\boldsymbol{\beta}_0)$ is just a single normal prior, and is similar to the prior used for model comparison by Zellner and Siow (1980). Seeing how $\boldsymbol{\Sigma}^*$ differs from the Zellner and Siow covariance matrix in unbalanced cases would be of considerable interest.

The behavior of $\pi_j^1(\boldsymbol{\beta}^*|\boldsymbol{\beta}_0, \sigma_j)$ 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 (A4). For instance, when $p = 1$, using (27) and (28) yields

$$\begin{aligned}\pi_j^1(\boldsymbol{\beta}^*|\boldsymbol{\beta}_0, \sigma_j) &= \frac{1}{\sqrt{2\pi\sigma_j^2}} \cdot \frac{1}{L}\sum_{l=1}^{L}\frac{|\mathbf{X}_j^t(l)\mathbf{X}_j(l)|^{1/2}}{|\mathbf{X}_i^t(l)\mathbf{X}_i(l)|^{1/2}} \cdot \frac{1}{\lambda(l)}(1 - e^{-\lambda(l)/2}) \\ &= \frac{1}{L}\sum_{l=1}^{L}\frac{1}{2\sqrt{\pi V(l)}} \cdot \frac{1}{(\boldsymbol{\beta}^{*2}/V(l))} \cdot (1 - e^{-\boldsymbol{\beta}^{*2}/V(l)}),\end{aligned} \tag{33}$$

where

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

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 $\pi_j^1(\boldsymbol{\beta}^*|\boldsymbol{\beta}_0, \sigma_j)$ is a mixture of densities that behave like $\mathcal{T}_p(1, 0, \boldsymbol{\Sigma}^*(l))$ densities: p-variate t-densities with 1 degree of freedom, location 0, and scale matrix

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

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 $\pi_j^1(\boldsymbol{\beta}^*|\boldsymbol{\beta}_0, \sigma_j)$ are considerably more involved, and also do not appear to be proper distributions.

3.3 Intrinsic Priors in Nonnested Models: An Example

For nonnested models, finding a solution to (22) and (23) is often more difficult. Consider comparison of the nonnested models M_1: Exponential (θ_1) and M_2: Lognormal (μ, σ), introduced in Section 2.2. Assumption (20) can be shown to be satisfied, since

$$\begin{aligned} \text{under } M_1, \quad \hat{\theta}_2 = (\hat{\mu}, \hat{\sigma}) &= (\overline{y}, (S_y^2/n)^{1/2}) \\ &\overset{(n\to\infty)}{\longrightarrow} (E_{\theta_1}^{M_1}[\overline{Y}], (\frac{1}{n}E_{\theta_1}^{M_1}[S_y^2])^{1/2}) \\ &\overset{(n\to\infty)}{\longrightarrow} \psi_2(\theta_1) \equiv (\log\theta_1 - 0.5772, 1.2825); \end{aligned} \tag{34}$$

$$\text{under } M_2, \quad \hat{\theta}_1 = \overline{x} \to \psi_1(\mu, \sigma) = E_{(\mu,\sigma)}^{M_2}[\overline{X}] = \exp\{\mu + \tfrac{1}{2}\sigma^2\}. \tag{35}$$

Also, (21) becomes

$$\begin{aligned} B_1^* &= \begin{cases} E_{\theta_1}^{M_1}\left[\frac{2X_iX_j|\log(X_i/X_j)|}{(X_i+X_j)^2}\right] & \text{arithmetic case} \\ \exp\left\{E_{\theta_1}^{M_1}\left[\log\left(\frac{2X_iX_j|\log(X_i/X_j)|}{(X_i+X_j)^2}\right)\right]\right\} & \text{geometric case} \end{cases} \\ &= \begin{cases} 0.2954 & \text{arithmetic case} \\ 0.2383 & \text{geometric case;} \end{cases} \end{aligned} \tag{36}$$

$$B_2^* = \begin{cases} E_{(\mu,\sigma)}^{M_2}\left[\frac{2X_iX_j|\log(X_i/X_j)|}{(X_i+X_j)^2}\right] & \text{arithmetic case} \\ \exp\left\{E_{(\mu,\sigma)}^{M_2}\left[\log\left(\frac{2X_iX_j|\log(X_i/X_j)|}{(X_i+X_j)^2}\right)\right]\right\} & \text{geometric case} \end{cases}$$

$$= \begin{cases} H^A(\sigma) \equiv E^Z\left[\frac{\sqrt{2}\sigma|Z|}{1+cosh(\sqrt{2}\sigma Z)}\right] & \text{arithmetic case} \\ H^G(\sigma) \equiv \frac{3\sigma}{2}\cdot\exp\left\{-2E^Z\left[\log\left(1+e^{\sqrt{2}\sigma Z}\right)\right]\right\} & \text{geometric case,} \end{cases} \tag{37}$$

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

For the arithmetic case, equations (22) and (23) thus become

$$\frac{\pi_2^I(\mu,\sigma)(1/\exp\{\mu+\frac{1}{2}\sigma^2\})}{(1/\sigma)\pi_1^I(\exp\{\mu+\frac{1}{2}\sigma^2\})} = H^A(\sigma), \tag{38}$$

$$\frac{\pi_2^I(\log\theta_1 - 0.5772, 1.2825)(1/\theta_1)}{(1/1.2825)\pi_1^I(\theta_1)} = (0.2954). \tag{39}$$

We have not attempted to characterize the solutions to (38) and (39) in general. The equations are fairly easy to solve, however, if one assumes that

$$\pi_2^I(\mu,\sigma) = \pi_{21}^I(\mu)\pi_{22}^I(\sigma). \tag{40}$$

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

$$\begin{aligned} \pi_1^I(\theta_1) &= 2/\theta_1^c \\ \pi_2^I(\mu,\sigma) &= \frac{1}{2\sigma}H^A(\sigma)\exp\{(1-c)(\mu+\frac{1}{2}\sigma^2)\}, \end{aligned} \tag{41}$$

where $c = 1.1291$. A similar analysis for the geometric IBF yields, as the intrinsic priors, the expressions in (41) with H^A replaced by H^G and $c = 1.2602$.

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

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

where H is either H^A or H^G. The point of this transformation is that θ_1 and ν are then both the mean parameters of their respective distributions, and are given the same improper prior. Curiously, however, it is not the usual inverse noninformative prior. We speculate that this noninformative prior might prove to provide a better predictive match for these "common" mean parameters.

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

$$\text{as } \sigma \to 0, \quad \pi_{22}^I(\sigma) \cong \begin{cases} 1/(2\sqrt{\pi}) & \text{arithmetic case,} \\ 3/16 & \text{exponential case,} \end{cases}$$
$$\text{as } \sigma \to \infty, \quad \pi_{22}^I(\sigma) \cong \begin{cases} 1/(\sqrt{\pi}\sigma^2) & \text{arithmetic case,} \\ \frac{3}{4}\exp(-2\sigma/\sqrt{\pi}) & \text{exponential case.} \end{cases}$$

It is thus clear that $\pi_{22}^I(\sigma)$ is integrable; indeed, we have normalized (41) so that, in the arithmetic case, $\pi_{22}^I(\sigma)$ is a proper density.

The pattern we have observed thus seems to be holding: for parameters that are in some sense "common," the intrinsic priors are the same and are of a noninformative type, while parameters that exist only in one of the models receive proper intrinsic priors. For a variety of other examples and characterizations of intrinsic priors, see Dmochowski (1994).

APPENDIX

Proof of Equation (27): Defining $p = k_j - k_i$, note that

$$\frac{1}{L}\sum_{l=1}^{L} E_{\boldsymbol{\beta}_j,\sigma_j}^{M_j}[B_{ij}^N(l)] = \frac{C^*}{L}\sum_{l=1}^{L} \frac{|\mathbf{X}_j^t(l)\mathbf{X}_j(l)|^{1/2}}{|\mathbf{X}_i^t(l)\mathbf{X}_i(l)|^{1/2}} E_{\boldsymbol{\beta}_j,\sigma_j}^{M_j}\left[\frac{(R_j(l))^{(q_j+1)/2}}{(R_i(l))^{(q_i+p+1)/2}}\right], \tag{A1}$$

where

$$C^* = \frac{\pi^{-p/2}}{2^{(q_j-q_i)/2}} \cdot \frac{\Gamma((q_i+p+1)/2)}{\Gamma((q_j+1)/2)}. \tag{A2}$$

The expectation in (A1) can be evaluated in closed form for the default priors we consider. The answers are in terms of Kummer's function, $M(a,b,c)$ (see Abramowitz and Stegun, 1970, Chapter 13).

In the proofs, the following standard facts will be repeatedly used ; all notation is taken from Sections 2.3 and 3.2.

(i) Under M_j,

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

where χ^2_ν denotes the central chi-square distribution with ν degrees of freedom, and $\chi^2_p(\lambda(l))$ is the noncentral chi-square distribution with $p = k_j - k_i$ degrees of freedom and noncentrality parameter $\lambda(l)$ which, in the nested case, is given by (28). 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)

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

(iv) With obvious abuse of notation,

$$E\left[\left(\frac{\chi^2_1}{\chi^2_1+\chi^2_\nu}\right)^s\right] = \frac{\Gamma(s+1/2)\Gamma((\nu+1)/2)}{\Gamma(1/2)\Gamma(s+(\nu+1)/2)},$$

providing χ^2_1 and χ^2_ν are independent.

(v)

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

(vi)

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

Lemma 2. For the various noninformative priors, the expectations in (A1) are given in the following expressions:

Using the Jeffreys prior: Here $q_i = k_i$ and $q_j = k_j$, and the expectation in (A1) becomes

$$E^{M_2}_{\beta_j,\sigma_j}\left[\left(\frac{R_j(l)}{R_i(l)}\right)^{(k_j+1)/2}\right] = C^{**}e^{-\lambda(l)/2}M\left(\frac{p+1}{2}, \frac{p+k_j+2}{2}, \frac{\lambda(l)}{2}\right), \qquad \text{(A3)}$$

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 (A1) becomes

$$E\left[\left(\frac{R_j(l)}{R_i(l)}\right)^{(p+1)/2}\right] = 2^{-p} e^{-\lambda(l)/2} M(\frac{p+1}{2},\ p+1,\ \frac{\lambda(l)}{2}) \tag{A4}$$

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

Using the Reference prior: Here $q_i = q_j = 0$, and the expectation in (A1) becomes

$$E\left[\frac{(R_j(l))^{1/2}}{(R_i(l))^{(p+1)/2}}\right] = \frac{\exp\{-\lambda(l)/2\}}{\sigma_j 2^{p/2}\Gamma((p+2)/2)} M(\frac{1}{2}, \frac{p+2}{2}, \frac{\lambda(l)}{2}). \tag{A5}$$

Proof of Lemma 2: 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), (A3) follows immediately. The proof of (A4) is almost identical, but also uses Fact (vi). The explicit forms given for $p = 1, 3, 5$ follow from representations of M.

To prove (A5) use, in order, Facts (i) and (iii) to obtain

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

Defining $c_j^{-1} = 2^{(p+2j+1)/2}\Gamma(1/2)\Gamma((p+2j)/2)$, it is clear that

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

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

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

References

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

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. (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.

Clyde, M. and Parmigiani, G. (1995), "Orthogonalizations and Prior Distributions for Orthogonalized Model Mixing," ISDS Discussion Paper 95-07, Duke University.

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.

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.

Ibrahim, J. and Laud, P. (1994), "A Predictive Approach to the Analysis of Designed Experiments," *Journal of the American Statistical Association*, 89, 309-319.

Iwaki, K. (1995), "Posterior Expected Marginal Likelihood for Comparison of Hypotheses," Technical Report, Purdue University.

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

Kadane, J.B., Dickey, J., Winkler, R., Smith, W., and Peters, S. (1980), "Interactive Elicitation of Opinion for a Normal Linear Model," *Journal of the American Statistical Association*, 75, 845-854.

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

Laud, P.W. and Ibrahim, J. (1993), "Predictive Model Selection," Technical Report 93-01, Northern Illinois University.

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.*

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.

Suzuki, Y. (1983), “On Bayesian Approach to Model Selection,” in Proceedings of the International Statistical Institute, Madrid Vol.1, 288-291.

Schwarz, G. (1978), “Estimating the Dimension of a Model,” *Annals of Statistics*, 6, 461–464.

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.