OBJECTIVE BAYESIAN METHODS FOR MODEL SELECTION: INTRODUCTION AND COMPARISON [1]

James O. Berger and Luis R. Pericchi

Duke University
North Carolina, U.S.A.

Universidad Simón Bolívar
Caracas, Venezuela

Abstract and Index

The basics of the Bayesian approach to model selection are first presented, as well as the motivations for the Bayesian approach. We then review four methods of developing default Bayesian procedures that have undergone considerable recent development, the Conventional Prior approach, the Bayes Information Criterion, the Intrinsic Bayes Factor, and the Fractional Bayes Factor. As part of the review, these methods are illustrated on examples involving the normal linear model. The later part of the chapter focuses on comparison of the four approaches, and includes an extensive discussion of criteria for judging model selection procedures. As the chapter is lengthy, we include here an index to the sections.

[1] This research was supported by the National Science Foundation (USA), Grants DMS-9303556 and DMS-9802261, and by CONICIT-Venezuela G-97000592. The second author held a Guggenheim Fellowship during part of his research. An earlier version of this manuscript was presented at the workshop *Bayesian Model Selection*, held in Cagliari in June, 1997.

Contents

1 Introduction **3**
1.1 Bayes Factors and Posterior Model Probabilities . . . 3
1.2 Motivation for the Bayesian Approach to Model Selection . . . 4
1.3 Utility Functions and Prediction . . . 6
1.4 Motivation for Objective Bayesian Model Selection . . . 7
1.5 Difficulties in Objective Bayesian Model Selection . . . 7
1.6 Preview . . . 9

2 Objective Bayesian Model Selection Methods, with Illustrations in the Linear Model **10**
2.1 Conventional Prior Approach . . . 10
2.2 Intrinsic Bayes Factor (IBF) Approach . . . 13
2.3 The Fractional Bayes Factor (FBF) Approach . . . 16
2.4 Asymptotic Methods and BIC . . . 17

3 Evaluating Objective Bayesian Model Selection Methods **18**

4 Study Examples for Comparison of the Objective Methodologies **25**
4.1 Improper Likelihoods (Example 1) . . . 26
4.2 Irregular Models (Example 2) . . . 27
4.3 One-Sided Testing (Example 3) . . . 28
4.4 Increasing Multiplicity of Parameters (Example 4) . . . 31
4.5 Group Invariant Models (Example 5) . . . 33
4.6 When Neither Model is True (Example 6) . . . 35

5 Summary Comparisons **37**
5.1 Clarity of Definition . . . 37
5.2 Computational Simplicity . . . 39
5.3 Domain of Applicability . . . 39
5.4 Correspondence with Reasonable Intrinsic Priors . . . 41
5.5 Comparisons with Other Recent Approaches . . . 42

6 Recommendations **44**

1 Introduction

1.1 Bayes Factors and Posterior Model Probabilities

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

$$M_i\text{: } \mathbf{X} \text{ has density } f_i(\mathbf{x}|\boldsymbol{\theta}_i), \quad i = 1, \ldots, q,$$

where the $\boldsymbol{\theta}_i$ are unknown model parameters. Suppose that we have available prior distributions, $\pi_i(\boldsymbol{\theta}_i)$, $i = 1, \ldots, q$, for the unknown parameters. Define the marginal or predictive densities of $\mathbf{X}$,

$$m_i(\mathbf{x}) = \int f_i(\mathbf{x}|\boldsymbol{\theta}_i)\pi_i(\boldsymbol{\theta}_i)\, d\boldsymbol{\theta}_i.$$

The *Bayes factor* of M_j to M_i is given by

$$B_{ji} = \frac{m_j(\mathbf{x})}{m_i(\mathbf{x})} = \frac{\int f_j(\mathbf{x}|\boldsymbol{\theta}_j)\pi_j(\boldsymbol{\theta}_j)\, d\boldsymbol{\theta}_j}{\int f_i(\mathbf{x}|\boldsymbol{\theta}_i)\pi_i(\boldsymbol{\theta}_i)\, d\boldsymbol{\theta}_i}. \tag{1}$$

The Bayes factor is often interpreted as the "odds provided by the data for M_j versus M_i." Thus $B_{ji} = 10$ would suggest that the data favor M_j over M_i at odds of ten to one. Alternatively, B_{ji} is sometimes called the "weighted likelihood ratio of M_j to M_i," with the priors being the "weighting functions." These interpretations are particularly appropriate when, as here, we focus on conventional or default choices of the priors.

If prior probabilities $P(M_j)$, $j = 1, \ldots, q$, of the models are available, then one can compute the posterior probabilities of the models from the Bayes factors. Indeed, it is easy to see that *posterior probability* of M_i, given the data $\mathbf{x}$, is

$$P(M_i \mid \mathbf{x}) = \frac{P(M_i)m_i(\mathbf{x})}{\sum_{j=1}^{q} P(M_j)m_j(\mathbf{x})} = \left[\sum_{j=1}^{q} \frac{P(M_j)}{P(M_i)} B_{ji}\right]^{-1}. \tag{2}$$

A particularly common choice of the prior model probabilities is $P(M_j) = 1/q$, so that each model has the same initial probability. The posterior model probabilities are then the same as the renormalized marginal probabilities, given by

$$\bar{m}_i(\mathbf{x}) = \frac{m_i(\mathbf{x})}{\sum_{j=1}^{q} m_j(\mathbf{x})}. \tag{3}$$

Indeed, in scientific reporting, it is common to provide the $\bar{m}_i(\mathbf{x})$, rather than the $P(M_j \mid \mathbf{x})$, since the prior probabilities of models can be a contentious matter and anyone can use the $\bar{m}_i(\mathbf{x})$ to determine their personal posterior probabilities via (2), noting that $B_{ji} = \bar{m}_j(\mathbf{x})/\bar{m}_i(\mathbf{x})$. In most of this chapter we thus focus on determination of default $\bar{m}_i(\mathbf{x})$ or, equivalently, default Bayes factors.

1.2 Motivation for the Bayesian Approach to Model Selection

Reason 1: *Bayes factors and posterior model probabilities are easy to understand.* The interpretation of Bayes factors as odds and the direct probability interpretation of posterior model probabilities are readily understandable by even non-statisticians. For instance, in Example 5 (section 4.5) the three models under consideration will be seen to have (for the preferred default Bayesian analysis) posterior probabilities given by $\bar{m}_1(\mathbf{x}) = 0.18$, $\bar{m}_2(\mathbf{x}) = 0.48$ and $\bar{m}_3(\mathbf{x}) = 0.34$. In contrast, alternative model selection schemes are not based on readily interpretable quantities. For instance, many schemes are based directly or indirectly on p-values corresponding to various models, but the extreme difficulty of properly interpreting p-values is well known (see, e.g., Edwards, Lindeman, and Savage (1963), Berger and Sellke (1987), Berger and Delampady (1987), and Sellke, Bayarri and Berger (2000)).

Reason 2: *Bayesian model selection is consistent.* This means that, if one of the entertained models is actually the true model, then Bayesian model selection will (under very mild conditions) guarantee selection of the true model if enough data is observed. Rather surprisingly, use of most classical model selection tools, such as p-values, C_p, and AIC, does not guarantee consistency.

It is sometimes argued that consistency is not a highly relevant concept because none of the models being considered are likely to be exactly true. While the true model might indeed typically be outside the candidate set of models, use of a procedure that fails to be consistent in the 'nice' case is disturbing. Furthermore, even when the true model is not among those being considered, results in Berk (1966) and Dmochowski (1996) show that (asymptotically and under mild conditions) Bayesian model selection will choose that model among the candidates that is closest to the true model in terms of Kullback-Leibler divergence. While this is a very nice property, the situation is complicated by the fact that approximate Bayesian model selection procedures may not share the optimality properties of real Bayesian procedures. For instance, Shibata (1981) shows that BIC, a popular approximate Bayesian procedure (see section 2.4), is not optimal for certain situations in which the true model is not in the candidate set. Further developments along these lines can be found in Berger, Ghosh, and Mukhopadhyay (1999), and suggest caution in use of approximate Bayesian model selection procedures.

Reason 3. *Bayesian model selection procedures are automatic Ockham's razors,* favoring simpler models over more complex models when the data provides roughly comparable fits for the models. Overfitting is a continual problem in model selection, since more complex models will always provide a somewhat better fit to the data than will simpler models. In classical statistics, overfitting is addressed by introduction of an adhoc penalty term (as in AIC), which increases as the complexity (i.e., the number of unknown parameters) of the model increases. There is a huge literature discussing the (unanswerable) question of which penalty term is best (see, e.g., Shao (1997)). In contrast, Bayesian procedures naturally penalize model complexity, and need no adhoc penalty terms. For an interesting historical example and general discussion and references, see Jefferys and Berger (1992).

Reason 4. *The Bayesian approach to model selection is conceptually the same, regardless of the number of models under consideration.* In contrast, there is a significant distinction in the classical approach between consideration of two models and consideration of more than two models; the former case is approached with the tools of hypothesis testing, while the latter is approached with the often quite different tools of model selection. Besides requiring the learning of different statistical technologies for testing and model selection, having a distinction between the two cases is philosophically unappealing.

Reason 5. *The Bayesian approach does not require nested models, standard distributions, or regular asymptotics.* Essentially all of classical model selection is based on at least one of these assumptions. Example 3 (see section 4.3) is an example of model selection with irregular models.

Reason 6. *The Bayesian approach can account for model uncertainty.* Selecting an hypothesis or model on the basis of data, and then using the same data to estimate model parameters or make predictions based upon the model, is well known to yield (often severely) overoptimistic estimates of accuracy. In the classical approach it is often thus recommended to use part of the data to select a model and the remaining part of the data for estimation and prediction. When only limited data is available, this can be difficult. Furthermore, this approach still ignores the fact that the selected model might very well be wrong, so that predictions based on assuming the model is true could be overly optimistic.

The Bayesian approach takes a different tack: ideally, all models are left in the analysis with, say, prediction being done using a weighted average of the predictive distributions from each model, the weights being determined from the posterior probabilities of each model. This is known as 'Bayesian model averaging,' and is widely used today as the basic methodology of accounting for model uncertainty. See Geisser (1993), Draper (1995), Raftery, Madigan and Hoeting (1997), and Clyde (1999) for discussion and references.

Although keeping all models in the analysis is an ideal, this can be cumbersome for communication and descriptive purposes. If only one or two models receive substantial posterior probability, it would not be an egregious sin to eliminate the other models from consideration. Even if one must report only one model, the fact mentioned above - that Bayesian model selection acts as a strong Ockham's razor - means that at least the selected model will not be an overly complex model, and so estimates and predictions based on this model will not be quite so overly optimistic.

Reason 7. *The Bayesian approach can yield optimal conditional frequentist procedures.* As mentioned earlier, model selection when only two models are under consideration can be viewed as hypothesis testing. The standard frequentist testing procedure, Neyman-Pearson testing, has the disadvantage of requiring the report of a fixed error probability α, no matter what the data. The common data-adaptive versions of classical testing, namely p-values, are not true frequentist procedures and also suffer from the rather severe interpretational problems discussed earlier. Thus, until recently, frequentists did not have satisfactory data-adaptive testing procedures.

In Berger, Brown, and Wolpert (1994), Berger, Boukai, and Wang (1997), and Dass and Berger (1998) it is shown that tests based on Bayes factors can be constructed such that the posterior probabilities of the hypotheses have direct interpretations as conditional frequentist error probabilities. The reported error probabilities thus vary with the data, and yet have valid frequentist interpretations.

The necessary technical detail to make this work is the defining of suitable conditioning sets upon which to compute the conditional error probabilities. These sets necessarily include data in both the acceptance and the rejection regions, and can roughly be described as the sets which include data points providing equivalent strength of evidence (in terms of Bayes factors) for each of the hypotheses.

There are more surprises arising from this equivalence of Bayesian and conditional frequentist testing. One is that, in sequential testing using these tests, the stopping rule is largely irrelevant to the stated error probabilities. Thus there is no need to consider spending α or any of the difficult computations involved with unconditional sequential Neyman-Pearson testing. See Berger, Boukai and Wang (1999) for an illustration.

1.3 Utility Functions and Prediction

As with any statistical problem, one should, in principle, approach model selection from the perspective of decision analysis. See Bernardo and Smith (1994) for discussion of a variety of decision and utility-based approaches to model selection. Recent articles advocating particular such approaches include Dupuis and Robert (1998), Goutis and Robert (1998), and Key, Pericchi and Smith (1999). It should also be noted that posterior probabilities do not necessarily arise as components of such analyses. Frequently, however, the statistician's goal is not to perform a formal decision analysis, but to summarize information from a study in such a way that others can perform decision analyses (perhaps informally) based on this information. Also, models are typically used for a wide variety of subsequent purposes, making advance selection through problem-specific utility considerations difficult. (Note, however, that some of the works in this direction, including the above-mentioned articles, propose generic utility functions that are argued to be broadly reflective of the purposes of model selection.) Even though these issues are by no means settled, it is likely that posterior probabilities (and Bayes factors) will always remain important tools of Bayesian model selection.

The most common use of models is for prediction of future observables, and there is considerable model selection methodology that is specifically oriented towards this goal. One such methodology is Bayesian model averaging (mentioned above), that explicitly bases predictions on the posterior weighted average of all the model predictions. Indeed, this can be shown to yield optimal Bayesian predictions for a variety of loss functions. Since the posterior model probabilities are an integral part of Bayesian model averaging, the discussion in this chapter is of direct relevance to that approach.

Often, one is constrained to select a single model that will be used for subsequent prediction and, somewhat surprisingly, it is not always optimal to select the model with the largest posterior probability. The largest posterior probability model is optimal under very general conditions if only two models are being entertained (see Berger, 1999) and is often optimal for variable selection in linear models having orthogonal design matrices (cf. Clyde and George, 1998). For other cases, such as in nested linear models, the optimal single model for prediction is the 'median probability model,' defined and illustrated in Barbieri and Berger (2000). Again, however, this model is found through analysis of the posterior model probabilities, so that the developments in this chapter are of direct relevance.

1.4 Motivation for Objective Bayesian Model Selection

There has been a long debate in the Bayesian community as to the roles of subjective and objective Bayesian analysis. Few would disagree that subjective Bayesian analysis is an attractive ideal, but the objective Bayesians argue that it is frequently not a realistic possibility, either because of outside constraints (e.g., the appearance of objectivity is needed), or because it is simply not feasible to obtain the extensive needed elicitations from subject experts.

In model selection, this last argument is particularly compelling, because one often initially entertains a wide variety of models, and careful subjective specification of prior distributions for all the parameters of all the models is essentially impossible. Indeed, this would typically be an egregious waste of the (always limited) time for which subject experts are available; one would typically want to use this available expert time for model formulation and, possibly, prior elicitation for the model that is ultimately selected. There is actually little debate over this issue in model selection; virtually all the analyses one sees involve default methods.

1.5 Difficulties in Objective Bayesian Model Selection

There are four main difficulties with the development of default Bayesian methods of model selection.

Difficulty 1. *Computation can be difficult.* Calculation of the Bayes factor in (1) can be challenging when the parameter spaces are high dimensional. Also, the total number of models under consideration, for which computations need to be done, can be enormous, especially in model selection problems such as variable selection. We do not address computational issues here; some recent papers on the subject are Carlin and Chib (1995), Green (1995), Kass and Raftery (1995), Verdinelli and Wasserman (1995), Raftery, Madigan and Hoeting (1997) and Clyde (1999).

Difficulty 2. *When the models have parameter spaces of differing dimensions, use of improper noninformative priors yields indeterminate answers.* To see this, suppose that improper noninformative priors π_i^N and π_j^N are entertained for models M_i and M_j, respectively. The 'formal' Bayes factor using these priors, B_{ji}, would then be given by (1). But, because the priors are improper, one could have just as well used, as the noninformative priors, $c_i\pi_i^N$ and $c_j\pi_j^N$, in which case the Bayes factor would be $(c_j/c_i)B_{ji}$. Since the choice of c_j/c_i is arbitrary, the Bayes factor is clearly indeterminate.

When the parameters $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$ can be thought of as essentially similar, choosing $c_j = c_i$ is reasonable. Situations in which this can be justified (through group invariance arguments) are given in Berger, Pericchi and Varshavsky (1998). If the parameter spaces of M_i and M_j are the same, it is common practice to also choose $c_j = c_i$, although we know of no formal way of justifying the practice. When (as is typically the case) the parameter spaces are of differing dimensions, choosing $c_j = c_i$ can be a bad idea, although it may not be egregiously bad if the dimensions are close. For special situations, there have been efforts to assign reasonable values to the c_j based on an extrinsic argument; an example is the approach of Spiegelhalter and Smith (1982). Ghosh and Samanta (1999) present an interesting generalization.

Difficulty 3. *Use of 'vague proper priors' usually gives bad answers in Bayesian model selection.* It is virtually never the case in Bayesian analysis that use of vague proper priors is superior to use of improper noninformative priors, and this is especially so in model selection. To see the danger, consider the following example.

Example. Suppose we observe $\mathbf{X} = (X_1, \ldots, X_n)$, where the X_i are iid $\mathcal{N}(0,1)$ under M_1 and $\mathcal{N}(\theta,1)$ under M_2. Suppose θ under M_2 is given the $\mathcal{N}(0,K)$ prior, with variance K large; this is the usual vague proper prior for a normal mean. An easy calculation using (1) yields

$$B_{21} = (\frac{1}{n} + K)^{-1/2} \exp\,(\frac{Kn^2}{2(1+Kn)}\,\bar{x}^2). \tag{4}$$

For large K (or large n), this is roughly $K^{-1/2}\exp\,(z^2/2)$, where $z = n\bar{x}^2$. So B_{21} depends very strongly on the *arbitrarily chosen* 'large' value of K. Note that even popular hierarchical priors, with vague proper priors on the hyperparameters, can run afoul of this difficulty.

In contrast, the usual noninformative prior for θ in this situation is $\pi_2^N(\theta) = 1$. The resulting Bayes factor is $B_{21} = \sqrt{n/2\pi}\,\exp\,(z^2/2)$, which is a reasonable value. The short story here is thus *never use vague proper priors for model selection, but improper noninformative priors may give reasonable results.*

Difficulty 4. *Even 'common parameters' can change meaning from one model to another, so that prior distributions must change in a corresponding fashion.* Here is 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:

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

Thinking, first, about M_2, suppose the elicited prior density is of the form $\pi_2(\beta_1, \beta_2, \sigma_2) = \pi_{21}(\beta_1)\pi_{22}(\beta_2)\pi_{23}(\sigma_2)$. Since β_1 is 'common' to the two models, one frequently sees the same prior, $\pi_{21}(\beta_1)$, also used for this parameter in M_1. This is not reasonable, since β_1 has a very different meaning (and value) under M_1 than under M_2. Indeed, regressing fuel consumption on weight alone will clearly yield a larger coefficient than regressing on both weight and engine size, because of the considerable positive correlation between weight and engine size. Similarly, one often sees the variances σ_1^2 and σ_2^2 being equated and assigned the same prior, even though it is clear that σ_1^2 will typically be larger than σ_2^2.

It should be noted that this problem affects subjective, as well as default Bayesian model selection. Here, for instance, an automotive expert might center a subjective prior for β_1 under M_1 at 0.8, but might center the prior under M_2 at 0.5. Obtaining such properly compatible assessments is far from easy.

1.6 Preview

Four methods of developing default Bayesian model selection procedures have undergone considerable recent development. These methods are the Conventional Prior (CP) approach, the Bayes Information Criterion (BIC), the Intrinsic Bayes Factor (IBF) approach, and the Fractional Bayes Factor (FBF) approach. We will mention other more recent approaches, as appropriate, but our attention will focus on these four methods.

The most obvious default Bayesian method that can be employed for model selection is simply to choose 'conventional' proper prior distributions, priors that seem likely to be reasonable for typical problems. This was the approach espoused by Jeffreys (1961), who recommended specific proper priors for certain standard testing problems. We shall call this approach the *Conventional Prior Approach*, and discuss it in section 2.1, with application to selection from among linear models.

While this approach is arguably about the best that can be done from a default perspective, it is difficult to implement in general, requiring careful selection of a default prior for each specific situation. This difficulty has lead to use, in practice, of rather crude approximations to Bayes factors, typified by the *Bayesian Information Criterion (BIC)*. This is defined in section 2.2 and illustrated on the linear model.

Concerns over the accuracy and applicability of BIC have resulted in the recent development of alternative default methods of Bayesian model selection. The two most prominent of these methods are the *Fractional Bayes Factor* approach of O'Hagan (1995) and the *Intrinsic Bayes Factor* approach of Berger and Pericchi (1996a, 1996b), and

Berger, Pericchi and Varshavsky (1998). These are introduced in sections 2.3 and 2.4, respectively, and applied to the linear model.

The main purpose of this chapter is to review and compare these four approaches to default model selection. To this end, it is necessary to first discuss - in section 3 - our views as to how model selection methodologies should be evaluated. This is a mixture of application and theory. 'Testing' the various methodologies in specific applications is clearly important, as is evaluating their ease of use in application. On the theoretical side, we will argue that the most enlightening approach is to investigate correspondence of the procedures with actual Bayesian procedures.

Section 4 presents the evaluations of the four studied methodologies. The section is oriented around six important or challenging examples: a situation involving group invariance, the situation of improper likelihoods, irregular models, one-sided testing, an example in which the dimension grows with the sample size, and an example in which none of the entertained models is true. Section 5 presents a summary of the comparisons.

Section 6 gives some final recommendations. In brief, our personal view of the situation is that all four discussed methods have value and should be utilized under appropriate circumstances. The challenge is thus to outline the circumstances under which each should be used or, perhaps more importantly, when each should not be used. This chapter is thus a summary of our experience in this regard. Clearly, evaluations such as this will be continually evolving as practical experience with the procedures increases.

2 Objective Bayesian Model Selection Methods, with Illustrations in the Linear Model

In this section, we discuss the four default Bayesian approaches to model selection that will be considered in this chapter, the Conventional Prior (CP) approach, the Bayes Information Criterion (BIC), the Intrinsic Bayes Factor (IBF) approach, and the Fractional Bayes Factor (FBF) approach. These approaches will be illustrated through application to model selection in the linear model.

2.1 Conventional Prior Approach

Jeffreys (1961, Chapter 5) dealt with the issue of indeterminacy of noninformative priors by (i) using noninformative priors only 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 (but not vague proper priors) for parameters that would occur in one model but not the other. He presented arguments justifying certain default proper priors, but mostly on a case-by-case basis. This line of development has been successfully followed by many others (for instance, by Zellner and Siow, 1980; see Berger and Pericchi, 1996a, for other references.)

Illustration 1: Normal Mean, Jeffreys' Conventional Prior

Suppose the data is $\mathbf{X} = (X_1, \ldots, X_n)$, where the X_i are iid $\mathcal{N}(\mu, \sigma_2^2)$ under M_2. Under M_1, the X_i are $\mathcal{N}(0, \sigma_1^2)$. Note that, because of Difficulty 4 in section 1.5, we differentiate between σ_1^2 and σ_2^2. However, in this situation the mean and variance can be shown to be orthogonal parameters (i.e., the expected Fisher information matrix is diagonal), in which case Jeffreys argues that σ_1^2 and σ_2^2 do have the same meaning across models and can be identified as $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Because of this identification, Jeffreys suggests that the variances can be assigned the same (improper) noninformative prior $\pi^J(\sigma) = 1/\sigma$, since the indeterminate multiplicative constant for the prior would cancel in the Bayes factor. (See Sansó, Pericchi and Moreno, 1996, for a formal justification.)

As the unknown mean μ occurs in only M_2, it needs to be assigned a proper prior. Through a series of ingenious arguments, Jeffreys obtains the following desiderata that this proper prior should satisfy: i) it should be centered at zero (i.e., centered at M_1); ii) have scale σ; iii) be symmetric around zero; and iv) have no moments. He argues that the simplest distribution that satisfies these conditions is the Cauchy$(0, \sigma^2)$. In summary, Jeffreys's conventional prior for this problem is:

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

Although this solution appears to be rather ad hoc, it is quite reasonable; choosing the scale of the prior for μ to be σ_2 (the only available non-subjective 'scaling' in the problem) and centering it at M_1 are natural choices, and Cauchy priors are known to be robust in various ways. Although it is easy to object to having such choices imposed on the analysis, it is crucial to keep in mind that there is no real default Bayesian alternative here. Alternative objective methods either thenselves correspond to imposition of some (proper) default prior or, worse, end up not corresponding to *any* actual Bayesian analysis. (See, in this regard, the evaluation *Principle* in section 3.)

Illustration 2: Linear Model, Zellner and Siow Conventional Priors

In Zellner and Siow (1980), a generalization of the above conventional Jeffreys prior is suggested for comparing two nested models within the normal linear model. Let $\mathbf{X} = [\mathbf{1} : \mathbf{Z}_1 : \mathbf{Z}_2]$ be the design matrix for the 'full' linear model under consideration, where $\mathbf{1}$ is the vector of 1's, and (without loss of generality) it is assumed that the regressors are measured in terms of deviations from their sample means, so that $\mathbf{1}^t\mathbf{Z}_j = 0$, $j = 1, 2$. It is also assumed that the model has been parameterized in an orthogonal fashion, so that $\mathbf{Z}_1^t\mathbf{Z}_2 = 0$. (This can also be achieved without essential loss of generality.) The corresponding normal linear model, M_2, for n observations $\mathbf{y} = (y_1, \ldots, y_n)^t$ is

$$\mathbf{y} = \alpha\mathbf{1} + \mathbf{Z}_1\boldsymbol{\beta}_1 + \mathbf{Z}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon}$ is $\mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$, the n-variate normal distribution with mean vector $\mathbf{0}$ and covariance matrix σ^2 times the identity. (Because the regression parameters are orthogonal to the variance in linear models, one can again justify using a common noninformative prior for the variance; we will thus 'cheat' and not differentiate between the σ^2 in different models.) Here, the dimensions of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are $k_1 - 1$ and p, respectively, the odd notation chosen for compatibility with subsequent developments.

For comparison of M_2 with the model $M_1 : \boldsymbol{\beta}_2 = 0$, Zellner and Siow (1980) propose the following default conventional priors:

$$\pi_1^{ZS}(\alpha, \boldsymbol{\beta}_1, \sigma) = 1/\sigma,$$

$$\pi_2^{ZS}(\alpha, \boldsymbol{\beta}_1, \sigma, \boldsymbol{\beta}_2) = h(\boldsymbol{\beta}_2|\sigma)/\sigma,$$

where $h(\boldsymbol{\beta}_2|\sigma)$ is the $\text{Cauchy}_p(\mathbf{0}, \mathbf{Z}_2^t\mathbf{Z}_2/(n\sigma^2))$ density

$$h(\boldsymbol{\beta}_2|\sigma) = c\frac{|\mathbf{Z}_2^t\mathbf{Z}_2|^{1/2}}{(n\sigma^2)^{p/2}}\left(1 + \frac{\boldsymbol{\beta}_2^t\mathbf{Z}_2^t\mathbf{Z}_2\boldsymbol{\beta}_2}{n\sigma^2}\right)^{-(p+1)/2},$$

with $c = \Gamma[(p+1)/2]/\pi^{(p+1)/2)}$. Thus the improper priors of the "common" $(\alpha, \boldsymbol{\beta}_1, \sigma)$ are assumed to be the same for the two models (again justifiable by orthogonality), while the conditional prior of the (unique to M_2) parameter $\boldsymbol{\beta}_2$, given σ, is assumed to be the (proper) p-dimensional Cauchy distribution, with location at $\mathbf{0}$ (so that it is 'centered' at M_1) and covariance matrix $\mathbf{Z}_2^t\mathbf{Z}_2/(n\sigma^2)$, "...a matrix suggested by the form of the information matrix," to quote Zellner and Siow (1980).

Computation for these prior distributions cannot be done in closed form. However, using the fact that a Cauchy distribution can be written as a scale mixture of normal distributions, it is possible to compute the needed marginal distributions, $m_i(\mathbf{y})$, with one-dimensional numerical integration.

When there are more than two models, or the models are non-nested, there are various possible extensions of the above strategy. Zellner and Siow (1984) utilize what is often called the 'encompassing' approach, wherein one compares each submodel, M_i, to the *encompassing* model, M_0, that contains all possible covariates from the submodels. One then obtains, using the above priors, the pairwise Bayes factors B_{0i}, $i = 1, \ldots, q$. The Bayes factor of M_j to M_i is then *defined* to be

$$B_{ji} = B_{0i}/B_{0j}. \tag{5}$$

Illustration 2 (continued): Linear Model, Conjugate g-priors

Another common choice of prior for the normal linear model is the conjugate prior, called a g-prior in Zellner (1986). For a linear model (adopting a more compact notation)

$$M : \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \;\; \boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I}_n),$$

where σ^2 and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^t$ are unknown and $\mathbf{X}$ is an $(n \times k)$ given design matrix of rank $k < n$, the *g-prior* density is defined by

$$\pi(\sigma) = \frac{1}{\sigma}, \quad \pi(\boldsymbol{\beta} \mid \sigma) \text{ is } \mathcal{N}_k(\mathbf{0}, g\sigma^2(\mathbf{X}^t\mathbf{X})^{-1}).$$

Sometimes $g = n$ is chosen while sometimes g is estimated by empirical Bayes methods (see, e.g., George and Foster, 1997, and Clyde and George, 1998).

The key advantage of g-priors is that the marginal density, $m(\mathbf{y})$, is available in closed form and, indeed, is given by

$$m(\mathbf{y}) = \frac{\Gamma(n/2)}{2\pi^{n/2}(1+g)^{k/2}} \left(\mathbf{y}^t\mathbf{y} - \frac{g}{(1+g)} \mathbf{y}^t\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y} \right)^{-n/2}.$$

Thus the Bayes factors and posterior model probabilities for comparing any two linear models are available in closed form.

Unfortunately, g-priors have some some undesirable properties when used for model selection. For instance, suppose one is interested in comparing the linear model above with $M^* : \boldsymbol{\beta} = \mathbf{0}$. It can be shown that, as the least squares estimate $\hat{\boldsymbol{\beta}}$ goes to infinity, so that one becomes *certain* that M^* is wrong, the Bayes factor of M^* to M goes to the nonzero constant $(1+g)^{(k-n)/2}$. It was essentially this undesirable property that caused Jeffreys (1961) to reject g-priors for model selection, in favor of the priors discussed above (for which the Bayes factor will go to zero when the evidence is overwhelmingly against M^*).

We find the conventional prior approach to be appealing in principle, but the above discussion reveals that it is difficult to implement, even in a 'simple' situation such as the linear model. Of perhaps more concern is that there seems to be no general method for determining such conventional priors. (This is, in contrast, to the situation for, say, estimation problems, where general techniques for deriving default priors do exist; see, e.g., Berger and Bernardo, 1992.) The remaining three model comparison methods that are discussed in this section have the advantage of applying automatically to quite general situations.

2.2 Intrinsic Bayes Factor (IBF) Approach

For the q models $M_1, \ldots, M_q$, suppose that (ordinary, usually improper) noninformative priors $\pi_i^N(\boldsymbol{\theta}_i)$, $i = 1, \ldots, q$, are available. In general, we recommend that these be chosen to be 'reference priors' (see Berger and Bernardo, 1992), but other choices will also typically give excellent results. Define the corresponding marginal or predictive densities of $\mathbf{X}$,

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

The general strategy for defining IBF's starts with the definition of a proper and minimal 'training sample,' which is simply to be viewed as some subset of the entire data $\mathbf{x}$. Because we will consider a variety of training samples, we index them by l.

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.

The "standard" use of a training sample to define Bayes factors is to use $\mathbf{x}(l)$ to "convert" the improper $\pi_i^N(\theta_i)$ to proper posteriors, $\pi_i^N(\boldsymbol{\theta}_i|\mathbf{x}(l))$, and then use the latter to define Bayes factors for the remaining data. The result, for comparing M_j to M_i, can easily be seen (under most circumstances) to be

$$B_{ji}(l) = B_{ji}^N(\mathbf{x}) \cdot B_{ij}^N(\mathbf{x}(l)),$$

where

$$B_{ji}^N = B_{ji}^N(\mathbf{x}) = \frac{m_j^N(\mathbf{x})}{m_i^N(\mathbf{x})} \qquad \text{and} \qquad B_{ij}^N(l) = B_{ij}^N(\mathbf{x}(l)) = \frac{m_i^N(\mathbf{x}(l))}{m_j^N(\mathbf{x}(l))} \tag{6}$$

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, we "average" the $B_{ji}(l)$ over all possible training samples $\mathbf{x}(l)$, $l = 1, \ldots, L$. A variety of different averages are possible; here we consider only the *arithmetic IBF* (AIBF) and the *median IBF* (MIBF) defined, respectively, as

$$B_{ji}^{AI} = B_{ji}^N \cdot \frac{1}{L} \sum_{l=1}^{L} B_{ij}^N, \quad B_{ji}^{MI} = B_{ji}^N \cdot \text{Med}[B_{ij}^N(l)], \tag{7}$$

where "Med" denotes median. For the AIBF, it is typically necessary to place the more "complex" model in the numerator, i.e., to let M_j be the more complex model, and then *define* B_{ij}^{AI} by $B_{ij}^{AI} = 1/B_{ji}^{AI}$. The IBFs defined in (2) are *resampling* summaries of the evidence of the data for the comparison of models, since in the averages there is sample re-use. These are the only resampling methods systematically studied in the chapter.

These IBFs were defined in Berger and Pericchi (1996a) along with alternate versions, such as the *encompassing* IBF and the *expected* IBF, which we recommended for certain scenarios. We will refer to these other IBFs in some of the illustrations and examples. Originally, our focus was on finding the 'optimal' IBF for given scenarios. The MIBF was not optimal for any of the scenarios we considered, so that we did not give it much emphasis. We subsequently found, however, that the MIBF is the most robust and widely applicable IBF (see Berger and Pericchi, 1998) so that, for those who desire *one* simple default model selection tool, the MIBF is what we would recommend.

One additional aspect of the IBF approach should be mentioned. As part of the general evaluation strategy discussed in section 3, we propose investigation of so-called *intrinsic priors* corresponding to a model selection method. A strong argument can be made that, in nested models, intrinsic priors corresponding to the AIBF are very reasonable as *conventional priors* for model selection. Hence the IBF approach can also be thought of as the long sought device for generation of good conventional priors for model selection in nested scenarios.

Illustration 1 (continued): Normal Mean, AIBF and MIBF

We start with the noninformative priors $\pi_1^N(\sigma_1) = 1/\sigma_1$ and $\pi_2^N(\mu, \sigma_2) = 1/\sigma_2^2$. Note that π_2^N is not the reference prior that we recommend one use to begin computation of the IBF; but π_2^N yields simpler expressions for illustrative purposes. It turns out that minimal training samples consist of any two distinct observations $\mathbf{x}(l) = (x_i, x_j)$, and calculation shows that

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

Computation yields the following (unscaled) Bayes factor for data $\mathbf{x}$, when using π_1^N and π_2^N directly as the priors:

$$B_{21}^N = \sqrt{\frac{2\pi}{n}} \cdot \left(1 + \frac{n\overline{x}^2}{s^2}\right)^{n/2}, \tag{8}$$

where $s^2 = \sum_{i=1}^n (x_i - \overline{x})^2$. Using (7), the AIBF is then clearly equal to

$$B_{21}^{AI} = B_{21}^N \cdot \frac{1}{L} \sum_{l=1}^{L} \frac{(x_1(l) - x_2(l))^2}{2\sqrt{\pi}[x_1^2(l) + x_2^2(l)]}, \tag{9}$$

while the MIBF is given by

$$B_{21}^{MI} = B_{21}^N \cdot \text{Med}_{l=1,\ldots,L}[\frac{(x_1(l) -_2 (l))^2}{2\sqrt{\pi}[x_1^2(l) + x_2^2(l)]}].$$

Illustration 2 (continued): Linear Models, AIBF and MIBF

The IBF for linear and related models are studied in Berger and Pericchi (1996a, 1996b and 1997). Suppose, for $j = 1, ..., q$, that model M_j for $\mathbf{Y}$ $(n \times 1)$ is the linear model

$$M_j : \mathbf{y} = \mathbf{X}_j \boldsymbol{\beta}_j + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon}_j \sim \mathcal{N}_n(\mathbf{0}, \sigma_j^2 \mathbf{I}_n),$$

where σ_j^2 and $\boldsymbol{\beta}_j = (\beta_{j1}, \beta_{j2}, ..., \beta_{jk_j})^t$ are unknown, and $\mathbf{X}_j$ is an $(n \times k_j)$ given design matrix of rank $k_j < n$. We will consider priors of the form

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

Common choices of q_j are $q_j = 0$ (the reference prior, Berger and Bernardo, 1992), or $q_j = k_j$ (Jeffreys rule prior). When comparing model M_i nested in M_j, Berger and Pericchi (1996a) also consider a *modified Jeffreys* prior, having $q_i = 0$ and $q_j = k_j - k_i$. This is intermediate between reference and Jeffreys priors.

For these priors, a minimal training sample $\mathbf{y}(l)$, with corresponding design matrix $\mathbf{X}(l)$ (under M_j), is a sample of size $m = \max\{k_j\} + 1$ such that all $(\mathbf{X}_j^t \mathbf{X}_j)$ are nonsingular. Computation then yields

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

where R_i and R_j are the residual sums of squares under models M_i and M_j, respectively. Similarly, $B_{ij}^N(l)$ 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 $\mathbf{y}(l)$.

Inserting these expressions in (7) results in the Arithmetic and Median IBFs for the three default priors being considered. For instance, using the modified Jeffreys prior and defining $p = k_j - k_i > 0$, the AIBF is

$$B_{ji}^{AI} = \frac{|\mathbf{X}_i^t \mathbf{X}_i|^{1/2}}{|\mathbf{X}_j^t \mathbf{X}_j|^{1/2}} \cdot \left(\frac{R_i}{R_j}\right)^{(n-k_i)/2} \cdot \frac{1}{L} \sum_{l=1}^{L} \frac{|\mathbf{X}_j^t(l)\mathbf{X}_j(l)|^{1/2}}{|\mathbf{X}_i^t(l)\mathbf{X}_i(l)|^{1/2}} \cdot \left(\frac{R_j(l)}{R_i(l)}\right)^{(p+1)/2} . \tag{11}$$

To obtain the MIBF, simply replace the arithmetic average by the median. Note that the MIBF does not require M_i to be nested in M_j, as does the AIBF.

When multiple linear models are being compared, IBFs can have the unappealing feature of violating the basic Bayesian coherency condition $B_{jk} = B_{ji}B_{ik}$. To avoid this, one can utilize the *encompassing* approach, described in the paragraph preceding (5). This leads to what is called the Encompassing IBF. See Lingham and Sivaganesan (1997, 1999) and Kim and Sun (2000) for different applications.

2.3 The Fractional Bayes Factor (FBF) Approach

The fractional Bayes factor (developed in O'Hagan, 1995) is based on a similar intuition to that behind the IBF but, instead of using part of the *data* to turn noninformative priors into proper priors, it uses a fraction, b, of each *likelihood function*, $L_i(\boldsymbol{\theta}_i) = f_i(\mathbf{x}|\boldsymbol{\theta}_i)$, with the remaining $1-b$ fraction of the likelihood used for model discrimination. It is easy to show that the fractional Bayes factor of model M_j to model M_i is then given by

$$B_{ji}^F = B_{ji}^N(\mathbf{x}) \frac{\int L^b(\boldsymbol{\theta}_i)\pi_i^N(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i}{\int L^b(\boldsymbol{\theta}_j)\pi_j^N(\boldsymbol{\theta}_j)d\boldsymbol{\theta}_j} = B_{ji}^N(\mathbf{x}) \frac{m_i^b(\mathbf{x})}{m_j^b(\mathbf{x})}. \tag{12}$$

One common choice of b (see the examples in O'Hagan, 1995, and the discussion by Berger and Mortera of O'Hagan, 1995) is $b = m/n$, where m is the "minimal training sample size," i.e. the number of observations contained in a minimal training sample (as defined in section 2.2), assuming that this number is uniquely defined. O'Hagan (1995, 1997) also discuss other possible choices.

Illustration 1 (continued). Normal Mean

Assume, as in section 2.2, that $\pi_1^N(\sigma_1) = 1/\sigma_1$ and $\pi_2^N(\mu, \sigma_2) = 1/\sigma_2^2$. Consider $b = r/n$, where r is to be specified. Then the correction factor to B_{ji}^N in (12) can be computed to be

$$\mathrm{CF}_{21}^{F(b)} = \frac{m_1^b(\mathbf{x})}{m_2^b(\mathbf{x})} = \left(\frac{r}{2\pi}\right)^{1/2} \left(1 + \frac{n\overline{x}^2}{s^2}\right)^{-r/2} , \tag{13}$$

and thus

$$B_{21}^{F(b)} = B_{21}^{N}\mathrm{CF}_{21}^{F(b)} = \left(\frac{r}{n}\right)^{1/2}\left(1+\frac{n\overline{x}^2}{s^2}\right)^{(n-r)/2}, \tag{14}$$

where B_{21}^N and s^2 are as in subsection 2.2. Since minimal training samples are of size two, the usual choice of r, as mentioned above, is $r = 2$. This choice will thus be utilized in subsequent comparisons.

Illustration 2 (continued). Linear Model

Using the notation of subsection 2.2, the correction factor is

$$\mathrm{CF}_{ij}^{F(b)} = \frac{(b/2)^{(q_j-q_i)/2}}{\pi^{(k_j-k_i)/2}} \cdot \frac{\Gamma[(r-k_i+q_i)/2]}{\Gamma[(r-k_j+q_j)/2]} \cdot \frac{|\mathbf{X}_j^t\mathbf{X}_j|^{1/2}}{|\mathbf{X}_i^t\mathbf{X}_i|^{1/2}} \cdot \frac{R_j^{(r-k_j+q_j)/2}}{R_i^{(r-k_i+q_i)/2}}. \tag{15}$$

The FBF, found by multiplying B_{ji}^N by this correction factor, is thus

$$B_{ji}^{F(b)} = b^{(q_j-q_i)/2} \cdot \frac{\Gamma[(n-k_j+q_j)/2]\Gamma[(r-k_i+q_i)/2]}{\Gamma[(n-k_i+q_i)/2]\Gamma[(r-k_j+q_j)/2]} \cdot \left(\frac{R_i}{R_j}\right)^{(n-r)/2}. \tag{16}$$

Here, $r = m = \max\{k_j\} + 1$ will typically be chosen.

2.4 Asymptotic Methods and BIC

Laplace's asymptotic method (cf, Haughton, 1988, Gelfand and Dey, 1994, Kass and Raftery, 1995, Dudley and Haughton, 1997, and Pauler, 1998) yields, as an approximation to a Bayes factor, B_{ji}, with respect to two priors $\pi_j(\boldsymbol{\theta}_j)$ and $\pi_i(\boldsymbol{\theta}_i)$,

$$B_{ji}^{L} = \frac{f_j(\mathbf{x}|\hat{\boldsymbol{\theta}}_j)|\hat{\mathbf{I}}_j|^{-1/2}}{f_i(\mathbf{x}|\hat{\boldsymbol{\theta}}_i)|\hat{\mathbf{I}}_I|^{-1/2}} \cdot \frac{(2\pi)^{k_j/2}\pi_j(\hat{\boldsymbol{\theta}}_j)}{(2\pi)^{k_i/2}\pi_i(\hat{\boldsymbol{\theta}}_i)}, \tag{17}$$

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

$$B_{ji}^{S} = \frac{f_j(\mathbf{x}|\hat{\boldsymbol{\theta}}_j)}{f_i(\mathbf{x}|\hat{\boldsymbol{\theta}}_i)} \cdot n^{(k_i-k_j)/2}. \tag{18}$$

Discussion concerning this choice can be found in Kass and Wasserman (1995) and Pauler (1998).

The BIC approximation has the advantages of simplicity and an (apparent) freedom from prior assumptions. However, it is valid only for 'nice' problems. Among the problems for which it does not apply directly are models with irregular asymptotics (see, e.g., section 4.2) and problems in which the likelihood can concentrate at the boundary of the parameter space for one of the models. Dudley and Haughton (1997) and Kass and Vaidyanathan (1992) give extensions of (17) to such situations, but we will not formally consider such extensions here. We note, in passing, that the approximation (42) in Appendix 1 is both more accurate and widely applicable than (17).

Illustration 1 (continued). Normal Mean

Application of (18) yields,

$$B_{21}^S = \frac{B_{21}^N}{\sqrt{2\pi}} = \frac{1}{\sqrt{n}} \left(1 + \frac{n\overline{x}^2}{s^2}\right)^{n/2} .$$

Illustration 2 (continued). Linear Model

The usual BIC for linear models has the simple expression

$$B_{21}^S = \left(\frac{R_1}{R_2}\right)^{n/2} \cdot n^{(k_1-k_2)/2}.$$

3 Evaluating Objective Bayesian Model Selection Methods

There are many possible methods for evaluating default Bayes factors. An obviously important criterion is studying how they work in various examples; in particular, how often they fail to apply and how often they fail to give satisfactory answers. The discussion in later sections will focus, to a large extent, on these simple criteria.

One can also study formal properties of default Bayes factors, such as consistency, compatibility with sufficiency and the likelihood principle, and various types of coherence (cf, O'Hagan, 1995, 1997). These can be useful, but we feel that they are typically secondary, or can be subsumed within, discussion of the following basic principle.

Principle: *Testing and model selection methods should correspond, in some sense, to actual Bayes factors, arising from reasonable default prior distributions.*

That default Bayes factors should behave like real Bayes factors may seem almost tautological, but the principle seems to be questioned by some Bayesians. There are two major lines of argument against the principle. One is to question the value of Bayes factors as a general Bayesian measure of evidence in comparing hypotheses or models (e.g., to question whether Bayes factors have any value except as a mathematical component of posterior probabilities). For a defense of Bayes factors in this regard, see Berger (1999).

The second line of argument, popularized in O'Hagan (1995), is based on a type of decomposition of the Bayes factor into components, with the argument being that one might want to 'robustify' one component (typically that arising from the prior) by borrowing information from another component (typically that arising from the likelihood).

Our view of the situation is simply that only the end result, namely the overall Bayes factor, has any importance or significance, and we seek only to evaluate this end result.

One of the primary reasons that we (the authors) are Bayesians is that we believe that the best *discriminator between procedures* is study of the prior distribution giving rise to the procedures. Insights obtained from studying overall properties of procedures (e.g., consistency) are enormously crude in comparison (at least in parametric problems, where such properties follow automatically once one has established correspondence of a procedure with a real Bayesian procedure). Moreover, we believe that one of the best ways of studying any biases in a procedure is by examining the corresponding prior for biases.

A side issue, but of considerable relevance in practice, is that we know how to interpret Bayes factors. If an entirely different entity is created (i.e., a default measure that does not have an interpretation as a Bayes factor or a posterior probability), learning how it should be interpreted would be a formidable undertaking, even if it were somehow superior. We have already seen ample demonstration of this danger in our profession; for instance, regardless of the inherent value of p-values, they are a disaster in practice because the vast majority of practitioners incorrectly interpret them as posterior probabilities.

If one accepts the value (necessity?) of studying priors corresponding to a default Bayes factor, there are two hurdles in applying the above Principle. The first is that of interpreting the phrase "in some sense." Sometimes a default Bayes factor can be shown to exactly correspond to a real Bayes factor, but this can often only be done in an approximate sense. In Berger and Pericchi (1996a), we defined one useful sense of approximation, namely asymptotics: indeed, we defined an *intrinsic prior* as a prior distribution that would yield essentially the same answer as a default Bayes factor if there was a large amount of data. The reason for employing an asymptotic argument is technical: default Bayes factors often depend on "reuse" of the data, and corresponding priors will then only exist in an asymptotic sense. Appendix 1 outlines the development of asymptotic intrinsic priors.

One can, of course, question the use of an asymptotic argument here. Indeed, default Bayes factors, such as the IBF and FBF, will typically only approximate the "intrinsic prior Bayes factor" to order $O(1/\sqrt{n})$, and hence may differ substantially for smaller n. However, we have observed that any "biases" or unreasonable features of an intrinsic prior are typically also reflected in the small sample behavior of the corresponding default Bayes factor. Thus detecting unreasonable behavior of a default Bayes factor can usually be done more easily and more accurately through study of the intrinsic prior, than, say, by conducting a huge simulation study. Also, some properties, such as consistency, still follow automatically from the existence of a (proper) asymptotic intrinsic prior.

It should be noted that asymptotic determination of an intrinsic prior is related to, but distinct from, the asymptotics yielding BIC; such asymptotics at best correspond to utilization of an intrinsic prior evaluated solely at the null model (see Kass and Wasserman, 1995, and Pauler, 1998).

Illustration 3: As an illustration of some of the above concepts, consider the simple problem of testing a normal mean, with known variance. Suppose $X_1, ..., X_n$ are i.i.d. from the normal distribution with mean θ and variance one. It is desired to compare $M_1 : \theta = 0$ with $M_2 : \theta \neq 0$. Utilizing the standard noninformative prior, $\pi_2^N(\theta) = 1$, a minimal training sample is a single x_i, and

$$B_{21}^N = (2\pi/n)^{1/2} \exp(n\bar{x}^2/2), \;\; B_{12}^N(x_i) = (2\pi)^{-1/2} \exp(-{x_i}^2/2). \tag{19}$$

It follows that the AIBF and MIBF are given by

$$B_{21}^{AI} = B_{21}^N \cdot (2\pi)^{-1/2} \cdot \frac{1}{n}\sum_{i=1}^{n} \exp(-\frac{1}{2}{x_i}^2), \;\; B_{21}^{MI} = B_{21}^N \cdot (2\pi)^{-1/2} \exp(-\frac{1}{2}\mathrm{Med}[{x_i}^2]). \tag{20}$$

The FBF, with fraction b, can also easily be shown to be

$$B_{21}^F = \sqrt{b} \cdot \exp[n(1-b)\bar{x}^2/2]. \tag{21}$$

The first interesting feature to note is that B_{21}^F *exactly* equals the Bayes factor arising from a $\mathcal{N}(0, n^{-1}(b^{-1}-1))$ prior, so that this prior is the exact intrinsic prior corresponding to the FBF. Furthermore, note that, as b ranges from 0 to 1, the variance of this prior ranges from ∞ to 0. Hence there is a one-to-one correspondence between the choice of b and the choice of the intrinsic prior variance. Our view is thus that, in evaluating the FBF for this situation, one need only evaluate the suitability of the prior variance corresponding to b. A choice such as $b = 1/10$ would correspond to a prior variance of $9/n$. Then, as the sample size grew, the variance would be shrinking to 0, which seems very unreasonable from a Bayesian perspective. However, a choice such as $b = 1/n$ (the choice mentioned earlier, since the minimal training sample size here is 1) would correspond to a prior variance of $(1 - n^{-1})$, which at least has a stable behavior as n grows large.

The attractive property of the FBF in this situation, of having reasonable intrinsic priors (at least for suitable choices of b), is only approximately shared by the IBFs in (20). Indeed, note that, as $n \to \infty$,

$$\frac{1}{n}\sum_{i=1}^{n} \exp(-\frac{1}{2}{x_i}^2) \to E_\theta[\exp(-{X_i}^2)] = 2^{-1/2} \exp(-\frac{1}{4}\theta^2), \;\; \mathrm{Med}[{x_i}^2] \to (0.46) + \theta^2,$$

the last expression only being approximate for small or moderate θ. Using (45), (46) and (49) in Appendix 1, it is easy to see that the intrinsic priors for the AIBF and MIBF are, respectively, the $\mathcal{N}(0, 2)$ distribution and (approximately) $(1.26)\cdot\mathcal{N}(0, 1)$. Thus the AIBF does correspond to use of a sensible default prior for large n; and the relative error of the approximation (at $\theta = 0$, say) is of order $(0.4)/\sqrt{n}$, indicating that the AIBF will behave like a real Bayes factor even for quite modest n. The situation for the MIBF is different, in that it fails to behave like a true Bayes factor because of the constant (1.26) multiplying the (proper) $\mathcal{N}(0, 1)$ prior. We would view this as a defect of the MIBF here, amounting to a 26% "bias" in favor of M_2. Again, however, this must be kept in perspective; a 26% bias for a default Bayes factor should usually not be of much practical concern.

As a matter of definition, note that we use the term "intrinsic prior" to stand for any prior, proper or improper, to which a default Bayes factor corresponds (at least approximately). Of course, if the intrinsic prior turns out to be improper, some additional justification may be needed. Indeed, perhaps the most problematical aspect of the above Principle is interpreting the term "reasonable default prior distributions." Consider, for instance, Illustration 1 in section 2.1, namely deciding if a normal mean μ is zero or not, when the variance σ^2 is also unknown. As mentioned in section 2.1, Jeffreys (1961) devotes considerable discussion to the question of what makes a default prior "reasonable" in such a situation. His arguments involve concepts such as balance (the prior should give equal mass to the regions $\mu < 0$ and $\mu > 0$), symmetry, scaling (given σ^2, μ should have a distribution with scale σ), propriety (given σ^2, μ should have a proper distribution but the nuisance parameter, σ^2, can have an improper distribution) and appropriateness of the prior tails. While one might not agree with all of the conclusions of Jeffreys, his arguments as to what makes a default prior "reasonable" are classic.

One central issue in this regard is Difficulty 4 in section 1.5, namely the need to recognize when it is reasonable (and when it is not reasonable) to assign a 'common' parameter from different models the same prior distribution. Jeffreys proposed utilizing the same prior distribution for such a common parameter only when the parameter is orthogonal (in the sense of Fisher information) to the other parameters in the models under consideration. For a modern application of this principle, see Clyde and Parmigiani, 1996.

In part because of the difficulties inherent in orthogonal paramerization, and in part because of philosophical arguments, our own preference has been to think in terms of "predictive matching" priors, rather than priors based on orthogonalization. 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 $m(\mathbf{y}) = \int f_i(\mathbf{y}|\boldsymbol{\theta}_i)\pi_i(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i$, 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.

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 a "future" sample of data. Previously (Berger and Pericchi, 1997), we suggested that this should be essentially an "imaginary minimal training sample," which would typically be the smallest set of observations for which all model parameters are identifiable, but we have lately realized that a better choice would be the imaginary minimal training sample for the *smallest* model. (This work and its implications will be reported elsewhere.)

We will utilize the notion of predictive matching in examples later in the paper. Note that these ideas are related to ideas of elicitation through predictives (cf, Kadane, et.al., 1980). Also similar uses of predictive matching to define priors for model selection can be found in Suzuki (1983), Laud and Ibrahim (1995) and Ibrahim and Laud (1994).

While the above considerations are far from precise, they can frequently be implemented, at least in part. Furthermore, the idea is not to demand demonstration that a default Bayes factor has a reasonable intrinsic prior in a given situation, before it can be used (although this would certainly be reassuring); rather, the idea is to show that a default Bayes factor approach leads to reasonable intrinsic priors for a wide variety of situations, so that Bayesians will be willing to trust the methodology in new settings.

Before proceeding to the 'evaluation' examples of later sections, we look back at the illustrations from the previous section to demonstrate the use of the desiderata presented here.

Illustration 1 (continued). Normal Mean, Intrinsic Priors

Intrinsic priors for the various default approaches can be derived using the technique in Appendix 1. The needed mle's for application of that technique are $\hat{\mu} = \overline{x}$, $\hat{\sigma}_1 = (\sum_{i=1}^n x_i^2/n)^{1/2} = (\overline{x}^2 + s^2/n)^{1/2}$, and $\hat{\sigma}_2 = (s^2/n)^{1/2}$.

Conventional Prior Approach: Obviously the intrinsic prior would be the original conventional prior. The Zellner and Siow (1980) conventional prior is quite sensible, as has already been discussed, but the g-prior has the unpleasant feature, mentioned in section 2.1, of yielding a Bayes factor that does not converge to zero even when the evidence against M_1 is overwhelming.

Intrinsic Bayes Factor Approach: In Berger and Pericchi (1996a), it is shown that the intrinsic priors corresponding to the AIBF are given by the usual noninformative priors for the standard deviations ($\pi_i^I(\sigma_i) = 1/\sigma_i$) and

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

This last conditional distribution is proper (integrating to one over μ) and, furthermore, is virtually equivalent to the Cauchy$(0, \sigma_2)$ choice of $\pi_2(\mu|\sigma_2)$ suggested by Jeffreys (1961); indeed, the two prior densities never differ by more than 15%. Note that π_2^I also obeys all of the desiderata of Jeffreys for choice of a good conventional prior for this situation. That the AIBF has a conditionally proper intrinsic prior is not specific to this example, but holds generally for nested models, as discussed in Berger and Pericchi (1996a), Dmochowski (1996), and Moreno, Bertolino and Racugno (1998a). This is a rather remarkable property of AIBFs and leads to our recommending their use in the nested situation. A final curiosity is that, even though the original noninformative prior that was used for σ_2 was $\pi(\sigma_2) = 1/\sigma_2^2$, the intrinsic prior for σ_2 is the standard reference prior $\pi(\sigma_2) = 1/\sigma_2$.

An intrinsic prior can also be found for the MIBF. It is very similar to π^I above, but differs by a moderate constant and is hence not conditionally proper. We would interpret this constant as the amount of bias inherent in use of the MIBF.

Fractional Bayes Factor Approach: Selecting $r = m = 2$ as before, it is shown in De Santis and Spezzaferri (1997) that the intrinsic priors are given by the usual noninformative priors for the standard deviations ($\pi_i^I(\sigma_i) = 1/\sigma_i$) and

$$\pi_2^{F(2/n)}(\mu|\sigma_2) = \sqrt{\pi} \cdot \mathrm{Cauchy}(0, \sigma_2).$$

This conditional prior clearly integrates to $\sqrt{\pi}$, which we would interpret as indicating a bias of about 77% in favor of M_2. The prior is otherwise the same as that recommended by Jeffreys (1961), so that we would judge the FBF to be reasonable here, except for the bias. Note that increasing r will lead to a conditional intrinsic prior that is a Student-t density with r degrees of freedom, but will still have a biasing factor.

BIC Approach: It is easy to see that the intrinsic prior corresponding to BIC is constant in μ, which we do not feel is optimal. In Kass and Wasserman (1995), it is argued that B_{21}^S, the BIC Bayes factor, is approximately a real Bayes factor with respect to the usual noninformative prior for the variances (okay) and $\pi_2(\mu|\sigma_2) = \mathcal{N}(0, \sigma_2^2)$. For this prior, however, the Laplace approximation in (17) yields

$$B_{21}^L = B_{21}^S \cdot \exp[-\overline{x}^2/(2\hat{\sigma}_2^2)].$$

If M_1 is the true model and n is large, then $\overline{x}$ will be close to zero and B_{21}^L will be close to B_{21}^S. This is the basis of the Kass and Wasserman (1995) argument. Of course, ignoring $\exp[-\overline{x}^2/(2\hat{\sigma}_2^2)]$ can be bad if the sample size is not large or M_1 is not true. Indeed, we would view this quantity as the 'bias' against M_1 that is inherent in the use of BIC, and the amount of bias here could be very substantial. It is observed in Kass and Wasserman (1995) that, when n is large and the bias is extreme (because $|\overline{x}|$ is large), then this 'error' will not really matter, since B_{21}^S will overwhelm the bias term. This does require n to be large, however.

Illustration 2 (continued). Linear Models, Intrinsic Priors

Conventional Prior Approach: If only two models are under consideration in the Zellner and Siow (1980) approach, the intrinsic prior would obviously be the original conventional prior. If there are multiple models under consideration, however, intrinsic priors do not typically exist. For instance, suppose the encompassing approach (discussed in the paragraph preceding (5)) is utilized. The prior for the encompassing model, M_0, changes with the M_i being considered in the pairwise comparisons. This suggests that there is no single assignment of priors that would correspond to the Zellner and Siow procedure, and this can, indeed, be shown to be the case. We do not feel that this is a particularly serious flaw, however, just as we do not feel that a slight 'bias' in terms of the intrinsic prior mass is a serious flaw.

What about the conventional priors themselves? Following the arguments of Jeffreys (1961), we feel that it is very reasonable to assign 'common' parameters the usual noninformative priors and to use a proper conditional prior for the 'extra' parameters that is centered at $\mathbf{0}$ and has a Cauchy shape. Choosing the the prior scale matrix to be $\mathbf{Z}_2^t\mathbf{Z}_2/n$ is considerably less compelling, however. Indeed, in Nadal (1999), it is argued that, as the number of groups grows large in ANOVA models, this prior can become too concentrated about zero. Exploration of this important issue in other situations is needed.

For the conventional g-priors, there is no problem with multiple model coherency, as the priors are defined separately for each model. However, as indicated at the end of section 2.1, these priors have the undesirable property that the Bayes factor will not go to zero, even when the evidence against a model becomes overwhelming. These conventional priors also depend on $\mathbf{X}^t\mathbf{X}$, and can thus have the potential problem of 'over concentration about zero,' that was just mentioned.

Intrinsic Bayes Factor Approach: For the AIBF, pairwise intrinsic priors exist, and are discussed in Berger and Pericchi (1996b, 1997). We generally recommend use of the encompassing approach when dealing with linear models, so that all models are compared with the encompassing model, M_0, and Bayes factors are computed via (5). The main result is that the intrinsic prior for comparing M_1 (say) to M_0 is the usual reference prior for M_1, given by $\pi_1^I(\boldsymbol{\beta}_1, \sigma_1) = 1/\sigma_1$, while, under M_0, $\pi_0^I(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \sigma_0) = \pi_0^I(\boldsymbol{\beta}_0|\boldsymbol{\beta}_1, \sigma_0)/\sigma_0$, where $\pi_0^I(\boldsymbol{\beta}_0|\boldsymbol{\beta}_1, \sigma_0)$ is a proper density when the AIBF is derived using reference priors (i.e., $q_1 = q_0 = 0$); is proper when the AIBF is derived using modified Jeffreys priors (i.e., $q_1 = 0$, $q_0 = k_0 - k_1 = p$); but is not proper when the AIBF is derived using the Jeffreys rule noninformative priors (i.e., $q_1 = k_1, q_0 = k_0$). Thus the use of Jeffreys rule priors is not recommended for deriving the AIBF.

As an example, use of the modified Jeffreys priors, in deriving the AIBF, results in the (proper) conditional intrinsic prior

$$\pi_0^I(\boldsymbol{\beta}_0|\boldsymbol{\beta}_1, \sigma_0) = \frac{\sigma_0^{-p} c^*}{L} \cdot \sum_{l=1}^{L} \frac{|\mathbf{X}_0^t(l)\mathbf{X}_0(l)|^{1/2}}{|\mathbf{X}_1^t(l)\mathbf{X}_1(l)|^{1/2}} \cdot \psi(\lambda(l), \sigma_0),$$

where $c^* = (2\pi)^{-p/2}$, $\psi(\lambda(l), \sigma_0) = 2^{-p}\exp(-\lambda(l)/2)M((p+1)/2, p+1, \lambda(l)/2)$, M is the Kummer function $M(a, b, z) = \frac{\Gamma(b)}{\Gamma(a)} \cdot \sum_{j=0}^{\infty} \frac{\Gamma(a+j)}{\Gamma(b+j)} \cdot \frac{z^j}{j!}$ and $\lambda(l) = \sigma_0^{-2}\boldsymbol{\beta}_0^t\mathbf{X}_0^t(l)(I - \mathbf{X}_1(l)[\mathbf{X}_1^t(l)\mathbf{X}_1(l)]^{-1}\mathbf{X}_1^t(l))\mathbf{X}_0(l)\boldsymbol{\beta}_0$.

This (proper) conditional intrinsic prior behaves very much like a mixture of p-variate t-densities with one degree of freedom, location $\mathbf{0}$, and scale matrices

$$\boldsymbol{\Sigma}(l) = 2\sigma_0^2[\mathbf{X}_0^t(l)(I - \mathbf{X}_1(l)[\mathbf{X}_1^t(l)\mathbf{X}_1(l)]^{-1}\mathbf{X}_1^t(l))\mathbf{X}_0(l)]^{-1}.$$

Each individual training sample captures part of the structure of the design matrix, which thus gets reflected in the corresponding scale matrix above, but the overall mixture prior does not concentrate about the nested model to nearly as great an extent as do the earlier discussed conventional priors, which effectively choose the full $\sigma_0^2(\mathbf{X}_0^t\mathbf{X}_0)^{-1}$ as the scale matrix. In this regard, the intrinsic priors seem considerably more sensible.

Fractional Bayes Factor Approach: For ease of comparison, we assume that M_i is nested in M_j, use modified Jeffreys priors to derive the FBF, and assume that $b = m/n$, where $m = k_j + 1$ is the minimal training sample. Using the expressions in De Santis and Spezzaferri (1997), one can show that the FBF has an intrinsic prior of exactly the Zellner and Siow (1980) form, except that the conditional prior of the 'extra' parameters does not integrate to one. Indeed, the biasing factor is

$$\left(\frac{m}{2}\right)^{p/2} \cdot \frac{\sqrt{\pi}}{\Gamma[(p+1)/2]},$$

which can be quite large. Thus, although the FBF has an intrinsic prior with a sensible centering and Cauchy form, it has a potentially large (constant) bias. Futhermore, the concerns expressied earlier concerning the scale matrix in the Zellner and Siow (1980) prior also clearly apply to the intrinsic prior for the FBF.

BIC Approach: The situation here is the same as in Illustration 1. Priors can clearly be found for which the BIC Bayes factor equals the actual Bayes factor when the m.l.e of the 'extra parameters' in the more complex model is at $\mathbf{0}$. But, as this m.l.e. departs from $\mathbf{0}$, the BIC Bayes factor will have an increasingly dramatic bias towards the more complex model. This bias becomes even more severe as the dimensional difference between the two models increases. See Pauler (1998) for a related discussion.

4 Study Examples for Comparison of the Objective Methodologies

Construction of counterexamples to default model selection methodologies is a time-honored tradition. These can be useful in suggesting positive alterations of the theory, and in suggesting the limits of applicability of the methodology. The situation here is complicated, however, by the fact that, almost by definition, it is possible to create counterexamples to any default Bayesian methodology (at least if one believes basic Bayesian coherency arguments). This means that the value of a counterexample is proportional to how representative it is of statistical application. Most useful are examples that are representative of large classes of statistical problems, and that indicate difficulties with the various methodologies that likely apply to the entire class of problems.

Six such examples will be studied here. In most of these examples, conventional priors do not exist, and the asymptotic approaches are of limited use; hence these examples will tend to focus on comparison of FBFs and IBFs. Furthermore, the examples mostly consider situations in which FBFs encounter certain difficulties, while IBFs do not. The intent of the examples is not, however, to suggest that the FBF approach is questionable; indeed, we are convinced that the FBF approach is very useful in default model selection. Rather, the examples are primarily studies we conducted to attempt to cause the IBF approach to "fail," in order to understand its limitations. We did not succeed in causing the IBF approach to fail, however.

4.1 Improper Likelihoods (Example 1)

There are a variety of models for which the full likelihood has infinite mass (considered as a function of $\boldsymbol{\theta}$), and these models can pose challenges in use of default Bayes factors. In a number of situations, such as exponential regression models (see Ye and Berger, 1991), use of suitable noninformative priors (e.g., Jeffreys or reference priors) will correct the problem, yielding finite marginal densities (so that default Bayes factors based on utilization of noninformative priors will then exist). That Jeffreys and reference priors seem to yield finite marginal densities the great majority of the time is one of the unexplained mysteries of default Bayesian analysis, and is one of the signifiant attractions in their use.

Another class of models with improper likelihoods is considerably more problematical. These are models in which no improper prior can yield a finite marginal density. One common example is mixture models, of which the following is a simple illustration. Suppose $X_1, ..., X_n$ are i.i.d. from the mixture density

$$f(x_i|\boldsymbol{\theta}) = p \cdot (2\pi)^{-1/2} e^{-x_i{}^2/2} + (1-p) \cdot (2\pi)^{-1/2} e^{-(x_i-\mu)^2/2}, \tag{22}$$

where $\boldsymbol{\theta} = (p, \mu)$ is unknown. Suppose we want to compare $M_1 : \mu = 0$ versus $M_2 : \mu \neq 0$, although the following difficulty applies to virtually any Bayesian attempt to utilize this density. While p poses no problem, since either the uniform prior or Jeffreys (reference) prior for p are proper, μ poses a considerable problem. The likelihood is the product (over i from 1 to n) of the density in (22), and expanding the product leads to a sum of terms, the first of which is a positive constant not depending on μ. Thus using, say, the uniform prior on p, $m(\mathbf{x}) = \int f(\mathbf{x}|\mu, p)\pi(\mu)\, d\mu\, dp$ is clearly infinite, unless $\pi(\mu)$ is proper (or, at least, has finite mass). Since this is true for any sample size, it is clear that there is no minimal training sample for this problem, and hence IBFs cannot be defined. Likewise it is easy to see that FBFs do not exist.

As asymptotic methods do not operate through integration, one might hope that it would be possible to apply them to problems with improper likelihoods. While this might be so, the problem is complicated by the fact that, in situations such as this, the correct asymptotic expressions can be very difficult to compute. It would appear that the only option here is thus to use the conventional prior approach; for the mixture problem, for instance, one might choose a Cauchy$(0, 1)$ prior for μ. Interestingly, however, one can modify the IBF approach or FBF approach to deal with mixture models. The idea in the above example would be to "pretend" that a training sample arises from the second component of the mixture in the IBF approach. It seems unnatural, however, to give all training samples equal weight; rather it is appealing to choose them in accordance with the probability that the training sample arose from the second component (the "classification probability" in mixture model language). Since these classification probabilities must be estimated, the procedure must operate iteratively. An algorithm implementing this idea in general mixture models is developed in Shui (1996), and appears to work quite successfully.

The chief limitation of this modified IBF approach is its considerable computational expense; thus, an FBF version of the analysis was also developed in Shui (1996). The key idea in this development was to notice that certain fractions of parts of the likelihood (with differing fractions for different parts of the likelihood) would have effects similar to those of the probabilistically classified training samples. The fractions still have to be determined iteratively, but a fixed point algorithm can be used to do so, rather than the stochastic iterative algorithm needed for the IBF. The FBF version is thus much more computationally efficient. Another recent modification of the IBF strategy for dealing with mixture models is the 'expected posterior prior' approach, discussed in section 5.5.

4.2 Irregular Models (Example 2)

In a large class of statistical models, the parameter space becomes constrained by the data. Such models do not typically have regular asymptotics. One of the simplest such examples is the location exponential distribution. Thus, suppose $X_1, X_2, ..., X_n$ are i.i.d. with density

$$f(x_i \,|\, \theta) = \exp(-(x_i - \theta))\, 1_{(\theta,\infty)}(x_i),$$

where "1" denotes the indicator function. It is desired to compare

$$M_1 : \theta = \theta_0 \quad \text{versus} \quad M_2 : \theta > \theta_0,$$

employing the usual non-informative prior $\pi_2^N(\theta) = 1$. Computation yields

$$m_1^N(\mathbf{x}) = \exp(n\theta_0 - S) 1_{(\theta_0,\infty)}(x_{min}),$$

$$m_2^N(\mathbf{x}) = \frac{1}{n}[\exp(nx_{min} - S) - \exp(n\theta_0 - S)] \cdot 1_{(\theta_0,\infty)}(x_{min}),$$

where $S = \sum_{i=1}^n x_i$ and $x_{min} = min[x_1, ..., x_n]$. Hence,

$$B_{21}^N(\mathbf{x}) = \frac{1}{n}[\exp(n(x_{min} - \theta_0)) - 1].$$

Intrinsic Bayes Factors: Any single observation is a minimal training sample. The Arithmetic and Median IBFs are thus given by

$$B_{21}^{AI} = B_{21}^N \cdot \frac{1}{n} \sum_{i=1}^n [\exp(x_i - \theta_0) - 1]^{-1}, \quad B_{21}^{MI} = B_{21}^N \cdot [\exp(\text{Med}[x_i] - \theta_0) - 1]^{-1}, \qquad (23)$$

the last following from the monotonicity of $[\exp(x-\theta_0)-1]^{-1}$.

In Appendix 2, the intrinsic priors for the AIBF and the MIBF are computed. They are given, respectively, (on $\theta > \theta_0$) by

$$\pi_2^{AI}(\theta) = -e^{(\theta-\theta_0)}\log(1-e^{(\theta_0-\theta)})-1, \quad \pi_2^{MI}(\theta) = [2e^{(\theta-\theta_0)}-1]^{-1}. \tag{24}$$

While $\pi_2^{AI}(\theta)$ can be shown to integrate to 1, the mass of $\pi_2^{MI}(\theta)$ is only about 0.69. The MIBF thus appears to have a modest bias in favor of M_1. It should also be noted that the statement that $\pi_2^{AI}(\theta)$ is the intrinsic prior is something of a conjecture, due to a technical issue discussed in Appendix 2.

Fractional Bayes Factors: Application of the FBF approach to this problem, utilizing fraction b of the likelihood, results in

$$B_{21}^F = B_{21}^N \cdot bn[e^{bn(x_{min}-\theta_0)}-1]^{-1}. \tag{25}$$

This is completely unsatisfactory. As one indication of this, it is easy to show that $B_{21}^F > 1$ for any $0 < b < 1$. (The proof is trivial, using the fact that $b^{-1}[\exp(bv)-1]$ is increasing in b for any $v > 0$.) Always favoring the more complex model, no matter what the data and no matter what fraction b is used, is clearly unreasonable.

The difficulty with the FBF here is that the most important part of the likelihood is the indicator function $1_{(x_{min},\infty)}(\theta)$, giving the data-dependent region in which the parameter θ must lie. The FBF operates by attempting to use just a "fraction" of the likelihood to update the prior, but clearly a fraction of an indicator function is the indicator function itself. This difficulty with the FBF could well apply to virtually any non-regular problem in which there are data-dependent constraints on the parameter. (Note that the IBF overcomes this problem because most training samples only provide a mild constraint on the parameter space and one "averages" over these mild constraints.)

BIC and Conventional Priors: BIC is inapplicable in this problem because the asymptotics are non-regular, so that (17) does not apply. Note, however, that the alternative asymptotic approximation (42) in Appendix 1 does apply; asymptotic approximations to Bayes factors that are based on this approximation are under development. To the best of our knowledge, no conventional proper priors have been proposed for this problem.

4.3 One-Sided Testing (Example 3)

One-sided testing is, of course, a broad class of problems, especially if considered in multivariate contexts. It poses interesting issues for all default methodologies; see Berger and Mortera (1999), Lingham and Sivaganesan (1997, 1999) and Sun and Kim (1997) for general discussion of these issues. Here, we consider only one-sided testing in the scale exponential model, to illustrate some of the more basic points.

Suppose $X_1, ..., X_n$ are i.i.d. with density of the form

$$f(x_i|\theta) = \theta^{-1}e^{-x_i/\theta} \quad \text{for } x_i > 0 \text{ and } \theta > 0. \tag{26}$$

It is desired to test $M_1 : \theta < \theta_0$ versus $M_2 : \theta > \theta_0$.

For the IBF and FBF approaches, it is natural to utilize the standard non-informative prior $\pi_i^N(\theta) = 1/\theta$ (for both models). Computation then yields

$$B_{21}^N(\mathbf{x}) = \frac{\int_{\theta_0}^{\infty} \theta^{-(n+1)} \exp(-n\bar{x}/\theta) d\theta}{\int_0^{\theta_0} \theta^{-(n+1)} \exp(-n\bar{x}/\theta) d\theta} = ([\gamma(n, n\bar{x}/\theta_0)]^{-1} - 1)^{-1},$$

where $\gamma(\alpha, x) = (1/\Gamma(\alpha)) \int_0^x \xi^{\alpha-1} e^{-\xi} d\xi$ is the incomplete Gamma function.

IBFs and FBFs: A minimal training sample is any single observation, x_i, and

$$B_{12}^N(x_i) = (e^{x_i/\theta_0} - 1)^{-1}.$$

A basic limitation of IBFs in one-sided testing is that arithmetic IBFs typically do not have intrinsic priors (cf., Dmochowski, 1995, and Moreno, Bertolino, and Racugno, 1998a). The reason is that $B_{12}^N(X_i)$ typically does not have finite expectation (under one of the models), so that (45) in Appendix 1 may well fail. Two possible alternatives are the encompassing arithmetic IBF (and its expected version, both defined in Berger and Pericchi, 1996a), and the median IBF. These are extensively studied in Berger and Mortera (1999), with the advantage seemingly belonging to the encompassing versions. While we would thus recommend these IBFs for the one-sided testing problem (in tune with our overall philosophy that different default Bayes factors can be optimal for different situations), we herein consider only the median IBF (in part, to also show its limitations).

Since $B_{12}^N(x_i)$ is clearly monotonic in x_i, the MIBF is

$$B_{21}^{MI} = B_{21}^N \cdot (\exp(\text{Med}[x_i]/\theta_0) - 1)^{-1}. \quad (27)$$

It is easy to see that, with the choice $b = 1/n$, the FBF is given by

$$B_{21}^F = B_{21}^N \cdot (\exp(\bar{x}/\theta_0) - 1)^{-1}. \quad (28)$$

Proper intrinsic priors exist for the median IBF and the FBF. They are computed in Berger and Mortera (1999) and (up to an irrelevant normalizing constant) are given, respectively, by

$$\pi^{FI}(\theta) = \begin{cases} \theta^{-1}(e^{\theta/\theta_0} - 1) & \theta < \theta_0 \\ (e-1)\theta^{-1}(e^{\theta/\theta_0} - 1)^{-1} & \theta > \theta_0 \end{cases},$$

$$\pi^{MI}(\theta) = \begin{cases} \theta^{-1}(2^{\theta/\theta_0} - 1) & \theta < \theta_0 \\ \theta^{-1}(2^{\theta/\theta_0} - 1)^{-1} & \theta > \theta_0 \end{cases}. \quad (29)$$

This intrinsic prior for the FBF has the somewhat unappealing property of being discontinuous at θ_0, but this is unlikely to cause any serious problems in practice. A considerably more important property of intrinsic priors in one-sided testing is their degree of "balance between the hypotheses" as indicated by the prior odds ratio $Pr(M_1)/Pr(M_2)$. For the two intrinsic priors above, computations in Berger and Mortera (1999) yielded prior odds ratios of 2.67 and 1.46, respectively. That these are not equal to one indicates that both the FBF and median IBF are biased, here in favor of M_1. The 46% relative bias of the median IBF may be large enough to be of concern to some, but the situation with the FBF is quite problematical, as it effectively carries a 169% bias against M_2. One can easily see this strong bias in simple data examples also (see Berger and Mortera, 1999).

Conventional Priors: In one-sided testing, it is often felt to be legitimate to perform a default Bayesian analysis directly, with standard noninformative prior distributions. Thus, in the one-sided exponential testing problem above, it would be common to use the noninformative prior $\pi^N(\theta) = 1/\theta$, and directly compute the Bayes factor of M_1 to M_2. A variety of arguments can be given which suggest that this is reasonable from a Bayesian perspective, at least as an approximation with large sample sizes. And the resulting answer coincides with the classical p-value (this is typically true for location or scale invariant models), so that it would seem that the entire profession accepts this conventional prior analysis.

Note, however, that the legitimacy of using noninformative priors directly has only been suggested in simple (invariant) situations of one-sided testing, whereas the IBF and FBF appear to be very widely applicable. Perhaps more importantly, one can question whether the conventional prior Bayesian answer is actually suitable for small or moderate sample sizes. The usual Bayesian justification for direct use of the conventional prior here is that the resulting answer (also the p-value) is the lower bound of the posterior probability of the relevant M_i over reasonable classes of prior densities (cf, Casella and Berger, 1987). It is natural to ask, however, if this lower bound is really the best evidential summary to provide. In Bayesian terms, if it were the case that the posterior probability of M_1 were some number between 0.05 and 0.5, depending on assumptions, would it really be reasonable to report 0.05 as the evidential summary?

This point was dramatically illustrated in the discussion by Morris of the Casella and Berger article (Morris, 1987), wherein a reasonable practical situation was considered and it was demonstrated that the lower bound is an unreasonable measure of the evidence against M_1. From a Bayesian perspective, Morris's argument was essentially that typical prior beliefs will concentrate closer to the dividing line between the hypotheses (θ_0 in the one-sided testing problem mentioned above), and that using a prior distribution which is extremely diffuse is thus unreasonable, at least for small or moderate sample sizes.

Interestingly, the IBFs and FBF do produce answers which are not as extreme as the standard (classical or Bayesian) answers, often being 2 to 10 times larger than the p-values against a hypothesis (equivalently, the conventional posterior probability of the hypothesis) when the sample size is small; see Berger and Mortera (1999).

4.4 Increasing Multiplicity of Parameters (Example 4)

In numerous models in use today, the number of parameters is increasing with the amount of data. The original example of this is the Neyman-Scott problem, the testing version of which we will consider here. Suppose we observe

$$X_{ij} \sim \mathcal{N}(\mu_i,\ \sigma^2), \quad i = 1, \ldots, n;\ j = 1, 2. \tag{30}$$

We are interested in comparing the two models $M_1 : \sigma^2 = \sigma_0^2$ versus $M_2 : \sigma^2 \neq \sigma_0^2$. Defining $\bar{x}_i = (x_{i1} + x_{i2})/2,\ \bar{\mathbf{x}} = (\bar{x}_1, \ldots, \bar{x}_n),\ S^2 = \sum_{i=1}^{n}(x_{i1} - x_{i2})^2$, and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$, the likelihood function (under M_2) can be written

$$L(\boldsymbol{\mu}, \sigma) \propto \sigma^{-2n} \cdot \exp\,[-\frac{1}{\sigma^2}(|\bar{\mathbf{x}} - \boldsymbol{\mu}|^2 + \frac{S^2}{4})]. \tag{31}$$

The feature of this problem that is most relevant to the following analysis is that the information available concerning each μ_i is limited (only two observations for each), while the information concerning σ^2 is increasing with n. This is indicated by the Fisher information matrix: it is diagonal, with diagonal elements $(2/\sigma^2, \ldots, 2/\sigma^2, 2n/\sigma^2)$, the last entry corresponding to the information for σ^2. This great difference in information can pose difficulties for default Bayes factors. Furthermore, techniques such as BIC, which are based on a fixed number of parameters and asymptotics in the sample size, cannot be used here.

The reference prior for the problem is

$$\pi_1^N(\boldsymbol{\mu}) = 1, \quad \text{and} \quad \pi_2^N(\boldsymbol{\mu}, \sigma) = 1/\sigma.$$

Note that the noninformative Jeffreys rule prior for M_2, namely $\pi_2^N(\boldsymbol{\mu}, \sigma) = 1/\sigma^n$, typically gives very bad performance for any inferences involving the Neyman-Scott problem. For the reference prior, computation yields

$$B_{21}^N = \frac{1}{2}\Gamma(\frac{n}{2})\frac{\sigma_0^n \exp\,[S^2/(4\sigma_0^2)]}{(S^2/4)^{n/2}}. \tag{32}$$

Intrinsic Bayes Factors: It is easy to see that a minimal training sample consists of both observations corresponding to one of the μ_i, and one observation corresponding to each of the others. (Using all the observations corresponding to one of the μ_i for the training sample should be cause for worry, but the IBFs will be seen to be fine.) Computation for a training sample shows that $B_{12}^N(\mathbf{x}(l))$ depends only on the two observations corresponding to the same μ_i, and it follows easily that the AIBF is given by

$$B_{21}^{AI} = B_{21}^N \cdot \frac{1}{n}\sum_{i=1}^{n} \frac{|x_{i1} - x_{i2}|}{\sqrt{\pi}\sigma_0} \exp\,[-\frac{(x_{i1} - x_{i2})^2}{4{\sigma_0}^2}]. \tag{33}$$

In Appendix 3, it is shown that the intrinsic prior corresponding to the AIBF is

$$\pi_1^{AI}(\boldsymbol{\mu}) = 1, \text{ and } \pi_2^{AI}(\boldsymbol{\mu}, \sigma) = \frac{2}{\pi\sigma_0(1+\sigma^2/\sigma_0{}^2)}. \tag{34}$$

Thus (for M_2) given σ, $\boldsymbol{\mu}$ has the usual constant noninformative prior, while, marginally, σ has the half-Cauchy distribution, with median equal to σ_0. This density for σ is not only proper, but it has the appealing property of being "balanced," in the sense of giving equal weight to $[\sigma < \sigma_0]$ and $[\sigma > \sigma_0]$. (Since π_2^{AI} is improper, it is not technically correct to talk about conditional and marginal densities; but such statements can be justified here in various limiting senses. Also, because of the "impropriety in the intrinsic prior for $\boldsymbol{\mu}$," properties such as consistency do not automatically hold, but consistency, at least, is easy to prove directly.)

Fractional Bayes Factors: For $b \leq 1/2$, the fractional Bayes factor does not exist. For $b > 1/2$, an easy computation yields

$$B_{21}^{F,b} = \frac{b^{n(b-1/2)}\Gamma(n/2)}{\Gamma(n(b-1/2))} \cdot [\frac{S^2}{4\sigma_0{}^2}]^{n(b-1)} \cdot \exp[\frac{(1-b)S^2}{4\sigma_0{}^2}]. \tag{35}$$

This is not a satisfactory default Bayes factor. Indeed it need not even be consistent, as shown by the following lemma (whose proof is given in Appendix 3).

Lemma 1 *Consider any monotonic sequence of fractions $(b_1, b_2, \ldots)$. Then B_{21}^{F,b_n} is not necessarily consistent as $n \to \infty$; for instance, if $\sigma^2 = 2\sigma_0{}^2$, then $\lim_{n\to\infty} B_{21}^{F,b_n} \leq 1$, even though M_2 is true. Furthermore, if $n(1-b_n) \to \infty$, then $B_{21}^{F,b_n} \to 0$, so that the FBF will eventually be certain to select the wrong model.*

The failure of the FBF in this example is due to the fact that (in intuitive terms) it was forced to take the same fraction, b, of the likelihood for each μ_i and the likelihood for σ, even though the amount of information in the likelihood for each is vastly different. (An earlier example of this phenomenon, based on comparison of two exponential means with vastly different sample sizes, was given in Iwaki, 1997.) Note that the IBF was fine in this regard, effectively taking just one observation for training purposes for each of the $\boldsymbol{\mu}_i$ and σ. The obvious suggestion is thus to attempt to write the likelihood function in terms of likelihoods for $\boldsymbol{\mu}$ and for σ, and then take different "fractions" of each. The natural way to write the likelihood function in this regard is

$$L(\boldsymbol{\mu}, \sigma) \propto [\frac{1}{\sigma^n} e^{-|\bar{\mathbf{x}}-\boldsymbol{\mu}|^2/\sigma^2}] \cdot [\frac{1}{\sigma^n} e^{-S^2/(4\sigma^2)}]. \tag{36}$$

Taking fraction b_1 of the first component of the likelihood, and fraction b_2 of the second, leads to a "training likelihood"

$$L^{b_1,b_2}(\boldsymbol{\mu}, \sigma) \propto [\frac{1}{\sigma^n} e^{-|\bar{\mathbf{x}}-\boldsymbol{\mu}|^2/\sigma^2}]^{b_1} \cdot [\frac{1}{\sigma^n} e^{-S^2/(4\sigma^2)}]^{b_2}. \tag{37}$$

Inserting this into (12), in place of L^b. defines a more general class of fractional Bayes factors.

As an example, consider the choices $b_1 = (n-1)/n$ and $b_2 = 2/n$ (roughly reflecting the fact that there is n times as much information in the likelihood about σ^2 as about each μ_i). Computation yields that the resulting FBF is

$$B_{21}^{F,b_1,b_2} = B_{21}^N \cdot [\frac{2S^2}{n\sigma_0\pi}]^{1/2} e^{-S^2/(2n\sigma_0^2)}. \tag{38}$$

This is quite sensible; indeed, it is shown in Appendix 3 that there is a corresponding intrinsic prior, given by

$$\pi_1^{FI}(\boldsymbol{\mu}) = 1, \text{ and } \pi_2^{FI}(\boldsymbol{\mu}, \sigma) = \frac{2}{\sqrt{\pi}\sigma_0} \cdot e^{-\sigma^2/\sigma_0{}^2}, \tag{39}$$

which is constant in $\boldsymbol{\mu}$ and (under M_2) gives σ the half-Normal distribution, with scale $\sigma_0/\sqrt{2}$. This density for σ is proper, but is not quite as "balanced" as was π_2^{AI}, in that the median of the prior is roughly $\sigma_0/2$, not σ_0. Nevertheless, we would view this as a very satisfactory intrinsic prior, and hence an appealing default Bayes factor.

While this generalized "fractionalization" appears to solve the problem here, note that the solution comes with the price of introducing additional complexity into the FBF approach. Indeed, it is not at all apparent, in general, how one should go about choosing the factorization of the likelihood (as in (36)), or which fractions should be used for each component. For interesting ideas in this direction, see De Santis and Spezzaferri (1999b).

4.5 Group Invariant Models (Example 5)

A very "nice" class of model comparisons is that which consists of comparisons between models having a given invariance structure. For general discussion of this class of comparisons, see Berger, Pericchi, and Varshavsky (1998); here we content ourselves with a simple example, which is also the final example in Bertolino and Racugno (1997) (also discussed from a different perspective in Berger and Pericchi, 1997). Of interest is comparison between three separate scale models: thus assume $X_1, ..., X_n$ are i.i.d., with $(X_i - \mu_0)/\sigma$ (μ_0 known) having either a standard

$$M_1 : Normal, \quad M_2 : Laplace \text{ (double exponential)}, \quad \text{or } M_3 : Cauchy$$

density. For the three models, Bertolino and Racugno (1997) consider noninformative priors of the form $\pi_j(\sigma) = \sigma^{-\alpha_j}$, with various choices of α_1, α_2, and α_3.

We should begin by noting that, for comparing models with the same group invariance structure (the multiplicative group for the above scale invariant models), there is a very strong reason to use the reference prior (given by $\alpha_j = 1$ here). This reason can be stated in two ways. First, the reference prior is typically exactly predictively matched for imaginary minimal training samples (in the sense discussed in section 3) for *all* models with the same group structure. This is a powerful inherent justification for its use, suggesting that one can directly use the B_{ji}^N in such a situation. The second way in which this can be viewed is to note that, when one has exact predictive matching for all minimal training samples, then all versions of IBFs reduce to simply B_{ji}^N. To see this explicitly in the scale parameter case, note that the marginal density of a minimal training sample (a single x_i here) is

$$m_j^N(x_i) = \frac{1}{2|x_i - \mu_0|},\ j = 1, 2, 3.$$

Since these are equal for all three models, the "correction terms" in IBFs are all equal to one, and the IBFs reduce to B_{ji}^N. (See Berger, Pericchi and Varshavsky, 1998, Berger and Pericchi, 1997, and Sansó, Pericchi and Moreno, 1996, for general discussion of this phenomenon. Note, in particular, that the predictive matching actually occurs for the *right invariant Haar measure* of the group action on the parameter space; this virtually always equals the so-called one-at-a-time reference prior, however.)

For priors other than the reference prior, IBFs will not simplify and so, to further explore the example of Bertolino and Racugno (1997), we will have to utilize an IBF. The three models here are separate models of equal complexity, and so it is not clear which to place in the numerator of the AIBF. Indeed, Bertolino and Racugno (1997) show that this ambiguity can lead to problems; they chose different values of the hyperparameters to show that the AIBF changes with the priors and that, for certain values of the hyperparameters, the AIBF can be incoherent in the sense that, simultaneously, $B_{12}^{AI} > 1$, and $B_{21}^{AI} > 1$. This issue is explored in Berger and Pericchi (1997), with the resulting recommendation that the MIBF be employed for separate models. (Note that the AIBF was, from the beginning, only recommended for nested models.) Hence we consider the MIBF in the following.

The FBF does not simplify in this situation when reference priors are used; comparison of invariant models is thus a situation where the IBF is actually simpler than the FBF. Furthermore, the FBF utilizes a fraction of the likelihood for updating the prior even though the predictive matching argument suggests that this is unnecessary. In computing the FBF below we used the fraction $b = 1/n$, since the minimal training sample is of size one.

In spite of our strong preference for the reference prior in this problem, we put the MIBF and the FBF to the test, not only with the reference prior ($\alpha_1 = \alpha_2 = \alpha_3 = 1$), but also with the prior specified by ($\alpha_1 = 1,\ \alpha_2 = \alpha_3 = 1.5$). (The behavior we observed for the other priors considered in Bertolino and Racugno, 1997, was not substantially different, so we confined attention to the above two priors.) In Table 1, we present the Bayes factors arising from the data considered in Bertolino and Racugno (1997), namely $\{x_i - \mu_0\} = \{-1, -0.4, -0.2, 0.001, 0.01, 0.1, 0.3, 1\}$.

Notice that the MIBF is quite robust to the choice of noninformative prior. The FBF seems somewhat more sensitive to the prior but, of more concern, gives answers which

Table 1: MIBF and FBF for Separate Scale Models, with reference and other priors.

α_1	α_2	α_3	B^M_{21}	B^F_{21}	B^M_{31}	B^F_{31}	B^M_{32}	B^F_{32}
1	1	1	2.64	1.86	1.88	0.90	0.71	0.48
1	1.5	1.5	2.38	2.06	1.52	0.46	0.64	0.22

are considerably different than those provided by the IBF with the reference prior (which we are arguing is the correct default answer for this problem). Of course, this is but one example and one data set, but it does reinforce the earlier message that IBFs are most well suited to group invariant situations.

Finally, observe that we essentially argued above that there is a highly acceptable *conventional prior* for model comparisons among models with the same group invariance structure, namely the reference prior (or, more precisely, the right invariant Haar measure). We have not studied the accuracy of BIC in this situation, but no obvious difficulties in its application are apparent, as long as the models are regular. (BIC would not apply, for instance, if one of the models were the scale uniform model.)

4.6 When Neither Model is True (Example 6)

In practice, it is probably rare for *any* of the entertained models to be true. (See Key, Pericchi and Smith, 1999, for general discussion.) How this affects the various default Bayes factors is only in the preliminary stages of investigation, but the issue is important enough to deserve mention.

Note, first of all, the interesting result in Smith (1995), that the *geometric* intrinsic Bayes factor (which replaces the arithmetic average in the AIBF with a geometric average), is an estimated version of the optimal model selector under a prequential scenario in which neither of the models being considered is the true model. The geometric IBF (GIBF) surfaces again in the following example, arising from O'Hagan (1997).

Consider the scenario of the Illustration 3 in section 3, where $X_1, ..., X_n$ are i.i.d. from the normal distribution with mean θ and variance one, and it is desired to compare $M_1 : \theta = 0$ with $M_2 : \theta \neq 0$. Recall that the FBF (for $b = 1/n$) is given by

$$B^F_{21} = \frac{1}{\sqrt{n}} \exp\left(\frac{n-1}{2}\bar{x}^2\right).$$

Simple computation shows that the GIBF is given by

$$B^{GI}_{21} = \frac{1}{\sqrt{n}} \exp\left(\frac{n-1}{2}[\bar{x}^2 - \frac{S^2}{n}]\right),$$

where $S^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2/(n-1)$. Concerning S^2, O'Hagan (1997) says: "The appearance of the sample variance in this formula represents the effect of the variability in the partial Bayes factor. The more the variability in the data, the more the IBF favours the simpler model, which is not intuitively reasonable behaviour."

O'Hagan's conclusion is far from clear. Indeed, if S^2 is large, most statisticians would begin to doubt the model assumption that $\sigma^2 = 1$. To explore the situation, let us assume that the *true* model for the data is $M_3 : \mathcal{N}(0,\, \sigma^2)$, with $\sigma^2 > 1$. Note that M_1 is clearly closer to M_3 than is M_2, so that we should still prefer M_1 to M_2. (We are not viewing this problem from the viewpoint of diagnostics or model elaboration; the scenario is still that we must choose between M_1 and M_2 using a default Bayes factor, and all we are trying to ascertain is which default Bayes factors will do a good job of choosing the model closest to M_3.)

It is useful to define $Z = \sqrt{n}\bar{x}/\sigma$, which clearly has a $\mathcal{N}(0, 1)$ distribution under M_3. In terms of Z, we can write the FBF and the GIBF as

$$B_{21}^F = \frac{1}{\sqrt{n}} \exp\left(\frac{(n-1)}{2n} Z^2\sigma^2\right), \quad B_{21}^{GI} = \frac{1}{\sqrt{n}} \exp\left(\frac{(n-1)}{2n}[Z^2\sigma^2 - S^2]\right).$$

If σ^2 is large enough, B_{21}^F will clearly typically favor the "worst" model M_2. For instance, even in the moderate case that $\sigma = 2$ and $Z = 1$, it is easy to check that B_{21}^F will exceed 1 (thus favoring M_2) unless the sample size is greater than 50. In contrast, the GIBF appears to be trying to compensate for M_3 by subtracting S^2 from the exponent. (Note that $E[Z^2\sigma^2 - S^2] = 0$.) Even for very large σ^2, the GIBF will typically favor M_1.

A different view of the situation can be obtained by studying the intrinsic priors that result from this "wrong models" scenario. Since the FBF was earlier seen to be the exact Bayes factor for a $\mathcal{N}(0,\, 1 - n^{-1})$ prior, we have no work to do there. Currently, we have only preliminary results on intrinsic priors for IBFs under a "wrong models" scenario but, for the AIBF in this situation, it is easy to show that its intrinsic prior is $\mathcal{N}(0,\, \sigma^2 + 1)$ (assuming M_3 is true). Thus the AIBF is "obtaining the scale for the prior" from the true model, not the assumed (incorrect) models.

The phenomenon here also relates to the criticism of IBFs that they typically do not depend solely on the sufficient statistics for the models, in that the training samples are not typically sufficient statistics. We have previously answered this criticism with the responses "it usually makes little difference in practice;" and "if desired, the matter is often easy to resolve, by either employing the *expected* IBF (which only depends on the sufficient statistics), or simply by taking an appropriate expectation of any IBF, conditional on the sufficient statistics (since, by definition, such an expectation does not depend on the unknown parameters)." We have always been reluctant to apply either of these "corrections," however, precisely because it seemed quite possible that IBFs were somehow employing information about the true model that was contained in the training samples. Note that any default Bayes factors which depend only on sufficient statistics (e.g., the FBF, BIC, and even conventional prior Bayes factors) cannot adapt to the true model in a "wrong models" scenario. Obviously much more investigation needs to be done to ascertain the success with which IBFs adapt to the true model, and it would be foolish to suggest that IBFs can adapt to all possible model deviations, but the indications are tantalizing.

5 Summary Comparisons

We do not repeat all of the comparisons made in section 4, but it is useful to summarize certain basic points. Also, we include some comparative comments obtained from other studies or examples.

5.1 Clarity of Definition

Comparison of the approaches is made more difficult by the fact that the approaches are, in various senses, not uniquely defined. Furthermore, as we are searching for an "automatic" procedure, we deem it to be negative if a default methodology requires considerable user input.

Conventional Prior Approach: The conventional prior approach suffers from the obvious problem of having a large possible multiplicity of conventional priors for a given situation. Part of the problem here is the brutal honesty of the approach: the conventional prior is clearly visible for all to see and criticize. In contrast, while BIC, IBFs, and FBFs may (at best) correspond to Bayes factors with intrinsic priors, these priors are not immediately visible. While statisticians would generally prefer the honesty of the conventional prior approach, many users of statistics prefer to "sweep such issues under the rug" (to borrow from I. J. Good), so that the conventional prior approach may well suffer in practical usage. (Let us be clear here: we are not happy with this state of affairs, but would much prefer to see BIC, IBFs, or FBFs used in practice by someone who, for this reason alone, would reject the conventional prior approach and use, say, a p-value.)

Intrinsic Bayes Factors: For the intrinsic Bayes factor, indeterminacy arises partly because of the existence of a variety of different IBFs. Our view of the IBF approach is that it is a *strategy* for approaching model selection problems, and that different IBFs will be "optimal" in different situations. (It is arguably naive to think that any single default Bayes factor will be optimal for all situations.) Thus, in Berger and Pericchi (1996a), we recommended that one should use the "expected IBF" if the sample size is small; the "arithmetic IBF" for two nested models; and the "encompassing IBF" for multiple linear models. We gave no clear recommendation for non-nested models. Recently (Berger and Pericchi, 1997b), we added the "median IBF" to our recommended list, in part to fill the gap for non-nested models; in part to serve as an alternative to the encompassing approach for multiple models; and possibly to serve as a "single default tool" for users who do not wish to vary the tool with the problem. (While not necessarily always optimal from our perspective, the median IBF is almost always reasonable - though see Ghosh and Samanta, 1999.)

As pointed out in Berger and Pericchi (1996a) (and emphasized in O'Hagan, 1997), IBFs also have the indeterminacy of being dependent on the way in which the data is presented. For instance, in O'Hagan (1997) it is observed that whether one presents the data as $(x_1, x_2, \ldots, x_n)$ or as $(x_1 - \bar{x}, \ldots, x_n - \bar{x}, \bar{x})$ can have a pronounced effect on IBFs if one were to blindly apply the formulae. Of course, one can always reconstruct the original data from $(x_1 - \bar{x}, \ldots, x_n - \bar{x}, \bar{x})$ but this requires some additional knowledge on the part of the user about IBFs. An obvious suggestion is to recommend that IBFs be applied to raw, not processed, data (although if the data could be processed to make them more nearly exchangeable, this would probably be advantageous; see section 4.5 in Iwaki, 1997, for an artificial, but interesting, example of how non-exchangeable data can affect IBFs). Sometimes, however, the notion of 'raw' data is not even well-defined. If the data consists of a realization of a stochastic process, for instance, there is no natural notion of a single 'raw' piece of data. For suggestions concerning the application of IBFs in such situations, see Sivaganesan and Lingham (1998, 1999). Even more problematical is the situation in which only sufficient statistics are available. Then the only IBFs available are the *expected* IBF and direct use of the intrinsic prior, since these are not based on resampling the actual data. Note that FBFs are essentially immune to these difficulties.

Fractional Bayes Factor: For the fractional Bayes factor, the main lack of definition is in the choice of the fraction b. Indeed, without a "recipe" for choosing b, one might question whether the FBF approach is actually a "default" Bayes factor. In Illustration 3 in section 3, for instance, we saw that choice of b is equivalent to choice of the (normal) prior variance; thus, suggesting that the choice of b be left to the investigator is equivalent to suggesting that the investigator subjectively choose the prior variance, i.e., adopt the subjective Bayesian approach. In practice, the most commonly used choice of b is m/n, where m is the minimal training sample size (when it exists).

We also saw that the standard FBF approach was unable to deal with various important classes of problems, such as those discussed in sections 4.1 and 4.4, unless it was modified to allow for using differing "fractions" of parts of the likelihood. If one embarks upon this generalization of the FBF approach (as we have found it essential to do), then the definitional issue appears to become even more involved; we have found, however, that choosing the (multiple) fractions by following IBF insight still succeeds.

BIC: Indeterminacy in asymptotic approximations such as BIC arises from several sources. First, one can utilize different asymptotic approximations. For instance, expression (42) in Appendix 1 typically gives a much more accurate approximation to a Bayes factor than does (17), and can be used as the basis for asymptotic approximations. (This is being explored elsewhere.) Even with the standard Laplace approximation, there are a variety of possible choices of the constant which replaces $\pi_j(\hat{\boldsymbol{\theta}}_j)$ and $\pi_i(\hat{\boldsymbol{\theta}}_i)$ in (17). And, even with a supposedly well-specified method such as BIC, which depends only on the likelihood, the dimensions of the parameters, and the sample size, there is considerable uncertainty as to how to define the sample size once one departs from i.i.d. situations. Pauler (1998) has begun the enterprise of determining "effective sample size" to deal with more complex situations.

5.2 Computational Simplicity

The computational edge is typically with BIC, as it requires only standard m.l.e. computations in its calculation. (Note that computation of Bayes factors is not a domain in which MCMC algorithms have an edge over m.l.e. computations.)

In general, use of conventional priors and FBFs is roughly equal in computational complexity although, for many "standard" problems, FBFs are available in closed form while conventional prior Bayes factors may require numerical integration. Normal linear models is one such scenario, where FBFs (and even IBFs) are available in closed form, but the standard conventional priors are Cauchy or t-priors, which require numerical integration.

IBFs are typically the most difficult default Bayes factors to compute (comparison of invariant models being an exception). Also, with large sample sizes, computation of IBFs is only possible with use of suitable schemes for sampling from the training samples, since the number of training samples might be enormous. These issues are discussed in Berger and Pericchi (1996a) and Varshavksy (1995). An interesting result from the latter work, based on the theory of U-statistics, is that utilizing only mn training samples, where m is the minimal training sample size and n is the overall sample size, is essentially as accurate as utilizing all training samples. Of course, even computing mn of the $B_{12}^N(\mathbf{x}(l))$ might appear to be computationally imposing, but note that these are very frequently available in closed form, even when B_{21}^N is not. Finally, it should be mentioned that Expected IBFs and even direct use of the intrinsic prior may be computationally more attractive in certain scenarios since they involve no training sample computations (cf, Berger and Pericchi, 1996a).

5.3 Domain of Applicability

Conventional priors have been developed primarily on a case-by-case basis, starting with the situations considered in Jeffreys (1961). There are, however, two general classes of problems for which satisfactory conventional priors can be said to exist. The first is the comparison of models having the same invariance structure (as in section 4.5 above). Then, using the notion of predictive matching, Berger, Pericchi, and Varshavsky (1998) argue that model selection can be done using the right invariant Haar measure corresponding to the group action for the models. Typically, this prior is simply the reference prior. Note that these priors are usually improper, but use of such is eminently defensible in this situation.

The second class of problems in which general conventional priors can be said to exist is nested models. The idea here is simply to use the intrinsic prior corresponding to the arithmetic IBF (as given, say, by (45) and (49) in Appendix 1). Our experience is that this intrinsic prior is an excellent conventional prior, typically being predictively matched for nuisance parameters and proper for the parameter under test. Of course, one can question the utility of computing and using the intrinsic prior, compared with simply using the arithmetic IBF to compute the Bayes factor directly; but, for small sample sizes, or for those who feel more comfortable using a conventional prior, this approach to construction of conventional priors deserves very serious consideration. Note that this is the first general approach to the construction of conventional priors in nested models.

BIC is actually quite limited in applicability, requiring larger sample sizes, models with regular asymptotics, models in which the likelihood does not concentrate on a boundary (as in one-sided testing), and determination of "effective sample size" in non-i.i.d. situations. Again, however, there is ongoing work attempting to resolve some of these limitations.

FBFs are impressively general in applicability, especially if the modifications suggested in sections 4.1 and 4.4 are considered. Irregular models (section 4.2) and one-sided testing (section 4.3) seem to cause problems for which there is no clear remedy. Also, there are domains, such as comparison of invariant models, where FBFs seem to introduce an unneeded (and somewhat detrimental) complication.

IBFs have the widest range of applicability, modulo possible computational limitations. The main non-computational limitation of the more common IBFs is very small sample sizes; and "small" must be interpreted sensibly, remembering that the IBF is roughly "getting the prior from the data." Thus, if one were considering a parameterized outlier model, with a data set having only one outlier, it would not be reasonable to use an IBF based on training samples, since there is not enough data to effectively obtain the prior distribution for the parameters of the outlier distribution. (This was illustrated in Example 3 of O'Hagan, 1997.) For some "small sample" situations of this type, IBFs seem to work fine (e.g., the Neyman-Scott problem in section 4.4), but caution should be the rule. (See Beattie, Fong, and Lin, 1999, for another example.) For extremely small sample sizes, expected IBFs or direct use of the intrinsic prior are probably at least as sensible as anything else of a default nature, but it can be argued that no default Bayes factors are really likely to be sensible for extremely small sample sizes. (There will typically be extreme sensitivity to the conventional prior chosen, to the choice of b in the FBF, etc.).

The various types of IBFs each have their own range of applicability, with the AIBF possibly being the most limited, and the MIBF being the most general. The AIBF is essentially limited to nested models (wherein it seems to perform the best). We have not found any serious limitations to applicability of the MIBF, except for situations, such as that in section 4.1, where no sample will yield a finite marginal if improper priors are used. (Recall, however, that, even with mixture models, it was possible to alter the IBF strategy to yield successful default Bayes factors.)

5.4 Correspondence with Reasonable Intrinsic Priors

Conventional Prior Approach: It may seem somewhat odd to ask whether the conventional prior approach yields Bayes factors which correspond to reasonable intrinsic priors; indeed, what we are simply concerned with here are the implications of "reasonable." Some of the difficulties inherent in the approach were discussed in sections 2 and 3, and numerous - perhaps most - of the constructions of conventional priors in the literature have violated one or more of the basic considerations mentioned therein. This should serve as a warning that the conventional prior approach must be viewed with the same scrutiny as other methods. Again, it is instructive to read Jeffreys (1961), and see the care with which he constructed conventional priors.

BIC Approach: The most basic fact in consideration of asymptotic approaches, such as BIC, is that they cannot correspond to an actual Bayesian analysis, since they replace the $\pi_j(\hat{\boldsymbol{\theta}}_j)$ and $\pi_i(\hat{\boldsymbol{\theta}}_i)$ in (17) with a constant, independent of the data. An interesting argument can be given in defense of BIC, however, in nested model scenarios. Indeed, Kass and Wasserman (1995) and Pauler (1998) argue that BIC then does correspond to an actual Bayes factor under the nested model; the "constant" used in the approximation can essentially be chosen to be the constant that would arise from a default conventional prior (which they call the "unit information prior") in the Laplace approximation under the nested model. The argument proceeds by observing that, under the more complex model, the Bayes factor (of the complex to the nested model) will typically go to infinity at an exponential rate, and hence that the inaccuracy then induced by replacing $\pi_j(\hat{\boldsymbol{\theta}}_j)$ and $\pi_i(\hat{\boldsymbol{\theta}}_i)$ with the "wrong" constant is of limited practical concern.

Our view of this argument is positive, in that we feel it does cleverly justify use of BIC (or other such approximations), if a quick approximation is needed. One should not read too much into the argument, however. With moderate amounts of data, it is very frequently the case that the Bayes factor is not conclusively in favor of either hypothesis and, when this is the case, one cannot trust the approximation; virtually by definition, one is then unsure as to which model is correct and, if the complex model is correct, the approximation can be bad. (And this is even ignoring the question of the accuracy of the Laplace approximation for smaller data sets.)

Another set of issues surrounds the choice of the "unit information prior," which is the intrinsic prior behind the BIC justification. Our original plan to also discuss this set of issues has run aground due to the length of the chapter, and so we will have to delay discussion to another forum.

FBF Approach: FBFs do seem to typically have reasonable intrinsic priors, as long as b is chosen to roughly reflect the same fraction of the likelihood as is a minimal training sample of the data. Problems in which there is concern include irregular models (section 4.2), where FBFs can fail to behave at all like Bayes factors; one-sided testing (section 4.3), where FBFs often correspond to intrinsic priors which are quite biased in favor of one of the hypotheses; and problems of increasing parameter multiplicity or highly varying parameter information content (section 4.4), where only FBFs which are modified to allow for multiple fractionation may be sensible. In spite of these serious omissions, we are generally quite positive about the true Bayesian nature of FBFs, and do not hesitate to use them in straightforward situations (or "tune" them with IBF reasoning in more complicated situations).

IBF Approach: From the beginning, our interest in IBFs has been motivated by the amazing ability they seem to possess to correspond to Bayes factors with reasonable intrinsic priors in even highly challenging situations, such as those in sections 4.2, 4.3, and 4.4. And in "nice" situations, such as that of section 4.5 where the natural default priors are eminently sensible for direct use, IBFs do the "right thing" and leave the default priors alone. Time and again we have also observed strong indications that intrinsic priors from IBFs have the key property of predictive matching (discussed in section 3), but we have not been able to formulate a general result in this direction. (Part of the difficulty is the technical complexity of having to deal with improper intrinsic priors; see Moreno, Bertolino, and Racugno, 1998a, for possible tools to use in this regard.)

The situation with IBFs is certainly not perfect in regards to intrinsic priors, however. Only for comparing two nested models, and using the AIBF, are reasonable intrinsic priors almost guaranteed to exist. And the AIBF typically does not have intrinsic priors outside of nested models. The other IBFs (such as the median IBF) typically yield reasonable, but not ideal, intrinsic priors. Thus, in section 4.2, the intrinsic prior for the MIBF only had mass 0.69 while, in section 4.3, it had a modestly unbalanced prior odds ratio; such problems indicate that the MIBF is typically modestly "biased" in favor of one the models.

5.5 Comparisons with Other Recent Approaches

Expected posterior prior approach: This is a recent highly promising approach, based on use of 'imaginary training samples,' developed in Pérez (1998) and Pérez and Berger (2000) (and, independently in a special case, in Schluter, Deely and Nicholson, 1999). This approach utilizes imaginary training samples to directly develop default conventional priors for use in model comparison. Letting $\mathbf{x}^*$ denote the imaginary training sample (usually taken to be a minimal training sample) and starting with noninformative priors $\pi_i^N(\boldsymbol{\theta}_i)$ for M_i as usual, one first defines the posterior distributions, given $\mathbf{x}^*$,

$$\pi_i^*(\boldsymbol{\theta}_i|\,\mathbf{x}^*) = \frac{f_i(\mathbf{x}^*|\,\boldsymbol{\theta}_i)\pi_i^N(\boldsymbol{\theta}_i)}{\int f_i(\mathbf{x}^*|\,\boldsymbol{\theta}_i)\pi_i^N(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i}. \tag{40}$$

Since $\mathbf{x}^*$ is not actually observed, these posteriors are not available for use in model selection. However, we can let $m^*(\mathbf{x}^*)$ be a suitable predictive measure for the (imaginary) data $\mathbf{x}^*$, and define priors for model comparison by

$$\pi_i^*(\boldsymbol{\theta}_i) = \int \pi_i^N(\boldsymbol{\theta}_i|\,\mathbf{x}^*)m^*(\mathbf{x}^*)d\mathbf{x}^*. \tag{41}$$

These are called *expected posterior priors* (or *EP priors* for short) because this last integral can be viewed as the expectation (with respect to m^*) of the posteriors in (40).

EP priors can successfully deal with the four difficulties discussed in section 1.5. Of particular note is that, because they can be viewed as an expectation of training sample posteriors, it is usually possible to utilize MCMC computational schemes to compute Bayes factors. Also noteworthy is that, since the $\pi_i^*(\boldsymbol{\theta}_i)$ are actual conventional priors, they inherit the attractive properties of that approach.

The key issue in utilization of EP priors is that of appropriate choice of $m^*(\mathbf{x}^*)$. Two choices that are natural are (i) the empirical distribution of the actual data (so that the EP prior approach can then be viewed as a resampling approach); and (ii) the (usually improper) predictive arising from an *encompassed* model (i.e., a model that is nested within all others under consideration) and under the usual nonformative prior. Of considerable interest is that this latter approach results in the $\pi_i^*(\boldsymbol{\theta}_i)$ being the intrinsic priors corresponding to the AIBF for nested models. Thus the EP prior approach can be viewed as an outgrowth of the IBF approach (and, indeed, arose in that way). Alternatively, it can be viewed as a method of actually implementing use of IBF intrinsic priors.

Posterior expected marginal likelihood approach: This is an approach recently developed in Iwaki (1997), which also contains extensive discussion and comparison with other methods. This approach is very similar to the EP prior approach but, instead of using $m^*(\mathbf{x}^*)$ in (41), it uses the predictive distribution of $\mathbf{x}^*$, given the actual data $\mathbf{x}$. This approach thus also generates priors, but they are data-dependent. For instance, in comparison of nested models, these priors seem to (roughly) be centered at the m.l.e. under the more complex model, rather than centered at the nested model. We view this particular type of data-dependency in the prior as inducing a bias in favor of the more complex model, and hence prefer the EP prior approach. (Note, however, that the empirical version of the EP prior approach can be similarly criticized.) The degree of this bias, however, is modest compared to most other methodologies, and so we view the approach of Iwaki (1997) as promising and deserving serious study.

Robust Bayesian analysis approach: Another approach that can be used as a default model selection method is robust Bayesian analysis (see Berger, 1996, for a review and references). In particular, one can often find upper and/or lower bounds on Bayes factors over wide classes of prior distributions. These bounds can be the basis of a default analysis. The three limitations of the approach are: (i) often only an upper or a lower bound, but not both, are available; (ii) computations can be formidable; and (iii) conclusions often cannot be reached solely from a bound on the Bayes factor. In spite of these limitations, it should be recognized that, when the approach can be applied, one obtains an answer that is considerably more compelling than answers from other default methodologies. Note, also, that there have been several interesting recent developments utilizing robust Bayesian methodology in concert with either the fractional or intrinsic Bayes factor approaches. References include De Santis and Spezzaferri (1996, 1999a), Moreno (1997), Moreno, Bertolino, and Racugno (1998b), Sansó, Pericchi, and Moreno (1996).

6 Recommendations

We end with a (provisional) answer to the question that motivated this paper: When should each of the default Bayes factor methodologies be used, and when should they not? In one sense, this question cannot be answered without asking another: Who is to be the user of the methodology, and in what fashion will it be used? BIC, the FBF (with, say, $b = m/n$ and constant priors) and the MIBF (with constant priors) could each be used as a single fairly general tool by unsophisticated users. At the other extreme, highly sophisticated users can view all default model selection methods as simply a collection of available tools, with specific strengths and weaknesses, that can be modified or adapted to deal with highly complex situations. A middle ground, which we find it useful to imagine, is that of a good Bayesian statistical computer package, in which a set of rules could be encoded to employ the default model selection strategy most suited for the situation being analyzed. In this spirit, we conclude with a few such comments about each of the four approaches we have considered.

The conventional prior approach should be used if there is a reasonable (and well-studied) conventional prior available for the given situation, and if operating in an environment where use of a specific conventional prior is sociologically acceptable. The approach is particularly valuable if the sample size is small (in which case all of the other approaches become suspect), or in problems (such as mixture models) where marginal densities with respect to noninformative priors are not finite.

The asymptotic approaches (such as BIC) can be used with confidence in situations with large sample sizes (relative to the number of parameters). For moderate sample sizes, there is a justification for their use when calculational necessity precludes utilization of any of the other methodologies; if one of the other approaches can be implemented, however, that would typically be preferred. Note, also, that the standard BIC has considerable limitations in its domain of applicability.

The intrinsic Bayes factor approach is the most generally applicable approach, if it can be implemented computationally. IBFs based on training samples can be used with considerable confidence (unless the sample size is very small) in (i) nested model comparisons (AIBF or encompassing IBF preferred); (ii) comparison of non-nested models of roughly equal size (MIBF preferred); and (iii) situations where reference priors are available for computing the IBFs. If the models are of highly varying dimension and reference priors are not available, we know of nothing better to use than the MIBF, but we do not have enough experience to be overly comfortable. When the sample size is very small, we recommend use of the expected IBF (or direct use of the intrinsic prior as a conventional prior). One should also be aware that, viewed as a *strategy*, IBFs can often be modified to handle difficult situations, as in the situation with mixture models discussed in section

4.1.

The recommended domain of the fractional Bayes factor approach is a large region somewhere between the recommended domains of the IBF and BIC. The FBF is typically easier to compute than the IBF, but it is considerably more difficult to compute than BIC. On the other hand, the range of applicability of the FBF is considerably greater than that of BIC, but is more limited than that of IBFs. Note that, if one generalizes the FBF to allow use of varying fractions in different parts of the likelihood, with the fractions chosen to mimic the effect of minimal training samples on IBFs (or, alternatively, chosen to reflect the varying amount of information about the parameters in the likelihood), the generality of the approach can be considerably increased. FBFs are typically available even for very small sample sizes, but their utility in such situations is tempered by the fact that the answer will typically then be highly sensitive to the choice of the fraction b, and no reasonable automatic choices for this fraction seem possible in such situations.

With judicious application of the above four methodologies, we are, for the first time, at the point where we can tackle most of the model selection and hypothesis testing problems with which we are presented.

Appendix 1: Intrinsic Prior Equations

The formal definition of an intrinsic prior, given in Berger and Pericchi (1996a), was based on an asymptotic analysis, utilizing the following approximation to a Bayes factor associated with priors π_j and π_i:

$$B_{ji} = B_{ji}^N \cdot \frac{\pi_j(\hat{\boldsymbol{\theta}}_j)\pi_i^N(\hat{\boldsymbol{\theta}}_i)}{\pi_j^N(\hat{\boldsymbol{\theta}}_j)\pi_i(\hat{\boldsymbol{\theta}}_i)}(1+o(1)); \tag{42}$$

here $\hat{\boldsymbol{\theta}}_i$ and $\hat{\boldsymbol{\theta}}_j$ are the m.l.e.s under M_i and M_j. (The approximation in (42) holds in considerably greater generality than does the Schwarz approximation in (17).)

Most default Bayes factors (certainly FBFs and IBFs) can be written in the form

$$B_{ji} = B_{ji}^N \cdot \tilde{B}_{ij}, \tag{43}$$

where $\tilde{B}_{ij}$ is often called the *correction factor.* To define intrinsic priors, equate (42) with (43), yielding

$$\frac{\pi_j(\hat{\boldsymbol{\theta}}_j)\pi_i^N(\hat{\boldsymbol{\theta}}_i)}{\pi_j^N(\hat{\boldsymbol{\theta}}_j)\pi_i(\hat{\boldsymbol{\theta}}_i)}(1+o(1)) = \tilde{B}_{ij}. \tag{44}$$

We next need to make some assumptions about the limiting behavior of the quantities in (44). The following are typically satisfied, and will be assumed to hold as the sample size grows to infinity:

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

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

When dealing with the AIBF, it will typically be the case that, for $k = i$ or $k = j$,

$$B_k^*(\boldsymbol{\theta}_k) = \lim_{L\to\infty} E_{\boldsymbol{\theta}_k}^{M_k}\left[\frac{1}{L}\sum_{l=1}^{L} B_{ij}^N(l)\right]; \tag{45}$$

if the $\mathbf{X}(l)$ are exchangeable, then the limits and averages over L can be removed. For the MIBF, it will simply be the case that, for $k = i$ or $k = j$,

$$B_k^*(\boldsymbol{\theta}_k) = \lim_{L\to\infty} \mathrm{Med}[B_{ij}^N(l)]. \tag{46}$$

For the FBF, De Santis and Spezzaferri (1997) show, for the i.i.d. situation and with $b = m/n$ (where m is fixed) that, for $k = i$ or $k = j$,

$$\begin{aligned} B_k^*(\tilde{\boldsymbol{\theta}}_k) &= \lim_{n\to\infty} \frac{\int L^b(\boldsymbol{\theta}_i)\pi_i^N(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i}{\int L^b(\boldsymbol{\theta}_j)\pi_j^N(\boldsymbol{\theta}_j)d\boldsymbol{\theta}_j} \\ &= \frac{\int \exp\left(m\, E_{\tilde{\boldsymbol{\theta}}_k}^{M_k}[\log(f_i(X_l \mid \boldsymbol{\theta}_i))]\right)\pi_i^N(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i}{\int \exp\left(m\, E_{\tilde{\boldsymbol{\theta}}_k}^{M_k}[\log(f_j(X_l \mid \boldsymbol{\theta}_j))]\right)\pi_j^N(\boldsymbol{\theta}_j)d\boldsymbol{\theta}_j}. \end{aligned} \tag{47}$$

(We have written $\tilde{\boldsymbol{\theta}}_k$ instead of $\boldsymbol{\theta}_k$ above to distinguish this variable from the dummy variables of integration; also, X_l stands for a single observation.)

Passing to the limit in (44), 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):

$$\begin{aligned} \frac{\pi_j^I(\boldsymbol{\theta}_j)\pi_i^N(\psi_i(\boldsymbol{\theta}_j))}{\pi_j^N(\boldsymbol{\theta}_j)\pi_i^I(\psi_i(\boldsymbol{\theta}_j))} &= B_j^*(\boldsymbol{\theta}_j), \\ \frac{\pi_j^I(\psi_j(\boldsymbol{\theta}_i))\pi_i^N(\boldsymbol{\theta}_i)}{\pi_j^N(\psi_j(\boldsymbol{\theta}_i))\pi_i^I(\boldsymbol{\theta}_i)} &= B_i^*(\boldsymbol{\theta}_i). \end{aligned} \tag{48}$$

The motivation, again, is that priors which satisfy (48) would yield answers which are asymptotically equivalent to use of the given default Bayes factors. We note that solutions are not necessarily unique, do not necessarily exist, and are not necessarily proper (cf, Dmochowski, 1996, and Moreno, Bertolino, and Racugno, 1998a).

In the nested model scenario (M_i nested in M_j), and under mild assumptions, solutions to (48) are trivially given by

$$\pi_i^I(\boldsymbol{\theta}_i) = \pi_i^N(\boldsymbol{\theta}_i), \quad \pi_j^I(\boldsymbol{\theta}_j) = \pi_j^N(\boldsymbol{\theta}_j)B_j^*(\boldsymbol{\theta}_j). \tag{49}$$

Typically there are also many other solutions, but the solutions in (49) are the simplest. See Dmochowski (1996) and Moreno, Bertolino, and Racugno, (1998a), for interesting characterizations of intrinsic priors in nested problems.

Appendix 2: Technical details from Section 4.2

We seek to establish (24). Even though the situation is highly nonregular, (42) can still be shown to apply, providing $\pi_2^N(\theta)$ and $\pi_2(\theta)$ are continuous and bounded at θ_0. Furthermore, this can be viewed as a nested problem, so that the intrinsic priors are given by (49) after determining the limits in (45) and (46).

For the MIBF, note that $\text{Med}[X_i] \to \theta_0 + \log 2$. The intrinsic prior in (24) follows immediately. For the AIBF, computation yields

$$E_\theta^{M_2}[(e^{(X_i-\theta_0)} - 1)^{-1}] = -e^{(\theta-\theta_0)} \log(1 - e^{(\theta_0-\theta)}) - 1. \tag{50}$$

This would indeed seem to establish the validity of (24). There is, however, an interesting technical difficulty, namely that (45) does not hold under M_1 (the expectation being infinite). While this might seem to doom the analysis, the intrinsic prior can be seen to be unbounded at θ_0, behaving as $-(\log(\theta - \theta_0) + 1)$. This means that (42) does not directly apply either, but it seems quite possible that generalized versions of (42), (44), and (45) would hold, since all quantities seem to be going to infinity at the same rate at θ_0. Unfortunately, we have not yet completed this generalization. The issue is of more than technical interest in that proper behavior of the AIBF, in such a difficult situation, can yield considerably increased confidence in the procedure.

Appendix 3: Technical details from Section 4.4

Intrinsic Priors: The intrinsic prior equations in Appendix 1 do not apply directly to this situation, because their asymptotic motivation is clearly inapplicable to the μ_i. Note, however, that the priors in (34) and (39) are both constant in the μ_i. Hence we can directly integrate the likelihood in (31) over the μ_i, before applying the results in Appendix 1. The resulting marginal likelihood of σ is

$$L_n(\sigma) \propto \sigma^{-n} e^{(-S^2/(4\sigma^2))}. \tag{51}$$

Note also that S^2 is a sum of the i.i.d. $S_i = (X_{i1} - X_{i2})^2$, having density

$$f(s_i \mid \sigma) \propto \sigma^{-1} e^{(-s_i/(4\sigma^2))},$$

and that

$$B_{21}^N = \frac{\int L_n(\sigma) \cdot \sigma^{-1}\, d\sigma}{L_n(\sigma_0)}.$$

Hence the results in Appendix 1 apply directly to this "marginal" problem.

Since the problem is comparison of nested models, (49) applies. For the AIBF, (45) yields

$$B_2^*(\sigma) = E_\sigma^{M_2}\left[\frac{|X_{i1} - X_{i2}|}{\sqrt{\pi}\sigma_0} \exp\left(-\frac{(X_{i1} - X_{i2})^2}{4\sigma_0{}^2}\right)\right] = \frac{2\sigma}{\pi\sigma_0}(1 + \frac{\sigma^2}{\sigma_0{}^2})^{-1}.$$

Inserting this into (49) yields (34). For the FBF in (38), one can (in the marginal problem) simply observe that $S^2/(2n) \to \sigma^2$. Hence (49) directly yields (39) as the intrinsic prior.

Proof of Lemma 1: Note first that $S^2/(2\sigma^2)$ has a chi-squared distribution with n degrees of freedom, from which it follows that, when $\sigma^2 = 2\sigma_0^2$ and $n \to \infty$,

$$\frac{S^2}{4n\sigma_0^2} = 1 + Z(\frac{2}{n})^{1/2} + \gamma_n, \tag{52}$$

where Z is standard normal and $\gamma_n = O(1/n)$. Note next that

$$\begin{aligned}\log(B_{21}^{F,b_n}) = \quad & n(1-b_n)\cdot\tfrac{S^2}{4n\sigma_0^2} + n(b_n-1)[\log(\tfrac{S^2}{4n\sigma_0^2}) - \log n] \\ & + n(b_n - \tfrac{1}{2})\log b_n + \log\Gamma(\tfrac{n}{2}) - \log\Gamma(n(b_n-\tfrac{1}{2})).\end{aligned} \tag{53}$$

Case 1. Assume that $n(b_n - 1/2) \to \infty$, and that b_n is bounded away from 1.

Using (52) in (53), expanding $\log(1 + Z\sqrt{2/n} + \gamma_n)$, and using Stirling's approximation with the Γ functions, results in the expression

$$\begin{aligned}\log(B_{21}^{F,b_n}) = \quad & (1-b_n)Z^2 + \tfrac{1}{2}\log(2b_n - 1) + n[(b_n - \tfrac{1}{2})\log(b_n/(b_n-\tfrac{1}{2})) \\ & -\tfrac{1}{2}\log 2] + O(\tfrac{1}{\sqrt{n}}).\end{aligned} \tag{54}$$

It can be shown that $(b - 0.5)\log(b/(b-0.5))$ is an increasing function of b on $(0.5\,,\,1)$, so that the term in square brackets in (54) is negative (since b_n is bounded away from 1). It is immediate that the FBF is then tending to 0.

Case 2. Assume that $n(b_n - 1/2) \to \infty$ and that $b_n \to 1$.

Expanding the log terms in (54) and taking limits yields

$$\lim_{n\to\infty}\log(B_{21}^{F,b_n}) = \lim_{n\to\infty} n(1-b_n)(\frac{1}{2} - \log 2).$$

Since $(0.5 - \log 2)$ is negative, the conclusion that the FBF is ≤ 1 in the limit is immediate. Also, if $n(1-b_n) \to \infty$, the conclusion that the FBF $\to 0$ follows.

Case 3. Assume that $\rho_n = n(b_n - 1/2) < K < \infty$.

Inserting ρ_n into (53) and performing a similar expansion to that leading to (54), results in the expression

$$\log(B_{21}^{F,b_n}) = \frac{1}{2}Z^2 + \rho_n(\log\frac{n}{2} - 1) - \frac{\log n}{2} - \frac{(n-1)}{2}\log 2 + \frac{\log(2\pi)}{2} - \log\Gamma(\rho_n) + O(\frac{1}{\sqrt{n}}).$$

The only positive unbounded term, $\rho_n \log n$, is clearly dominated by $-(n/2)\log 2$, so that the expression goes to $-\infty$, establishing that the FBF goes to 0. The case where ρ_n is not bounded above (and does not go to ∞ in limit) is handled by a more tedious version of Cases 1 and 3. We omit the details.

REFERENCES

Barbieri, M. and Berger, J. (2000), "Optimal Predictive Variable Selection," ISDS Discussion Paper, Duke University.

Berger, J. (1999), "Bayes Factors," in the *Encyclopedia of Statistical Sciences*, Update Volume 3, S. Kotz, et. al. (Eds.). Wiley, New York, 20–29.

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., Boukai, B. and Wang, Y. (1997), "Unified Frequentist and Bayesian Testing of a Precise Hypothesis (with discussion)," *Statistical Science*, **12**, 133–160.

Berger, J., Boukai, B. and Wang, Y. (1999), "Simultaneous Bayesian - Frequentist Sequential Testing of Nested Hypotheses," *Biometrika*, **86**, 79–92.

Berger, J., Brown, L. and Wolpert, R. (1994), "A Unified Conditional Frequentist and Bayesian Test for Fixed and Sequential Hypothesis Testing," *Ann. Statist.* **22**, 1787-1807.

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

Berger, J., Ghosh, J.K. and Mukhopadhyay, N. (1999), "Approximations to the Bayes factor in model selection problems and consistency issues," Technical Report, Purdue University.

Berger, J. and Mortera, J. (1999), "Default Bayes Factors for One-Sided Hypothesis Testing," *Journal of the American Statistical Association*, **94**, 542–554.

Berger, J. and Pericchi, L. (1996a), "The Intrinsic Bayes Factor for Model Selection and Prediction," *Journal of the American Statistical Association*, **91**, 109–122.

Berger, J. and Pericchi, L. (1996b), "The Intrinsic Bayes Factor for Linear Models," in *Bayesian Statistics 5*, eds. J. M. Bernardo, et. al., London: Oxford University Press, pp. 23–42.

Berger, J. and Pericchi, L. (1997), "On the Justification of Default and Intrinsic Bayes Factors," in *Modeling and Prediction*, eds. J. C. Lee, et. al., New York: Springer-Verlag, pp. 276–293.

Berger, J. and Pericchi, L. (1998), "Accurate and Stable Bayesian Model Selection: the Median Intrinsic Bayes Factor," *Sankya B*, **60**, 1–18.

Berger, J., Pericchi, L., and Varshavsky, J. (1998), "Bayes Factors and Marginal Distributions in Invariant Situations", *Sankya A*, **60**, 307–321.

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.

Berk, R. (1966), "Limiting Behavior of Posterior Distributions when the Model is Incorrect," *Annals of Mathematical Statistics*, **37**, 51-58.

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

Beattie, S., Fong, D., and Lin, D. (1999), “A Two-Stage Bayesian Model Selection Strategy for Supersaturated Designs,” Technical Report, The Pennsylvania State University.

Bertolino, F. and Racugno, W. (1997), “Is the Intrinsic Bayes Factor Intrinsic?” *Metron*, **LIV**, 5–15.

Carlin, B.P. and Chib, S. (1995), “Bayesian Model Choice via Markov Chain Monte Carlo,” *Journal of Royal Statistical Society - Series B*, **57**, 473–484.

Casella, G., and Berger, R. (1987), “Reconciling Bayesian and Frequentist Evidence in the One-Sided Testing Problem (with discussion),” *Journal of the American Statistical Association*, **82**, 106–111.

Clyde, M. (1999), “Bayesian Model Averaging and Model Search Strategies,” In *Bayesian Statistics 6*, J.M. Bernardo, A.P. Dawid, J.O. Berger, and A.F.M. Smith eds., Oxford University Press, pp. 157–185.

Clyde, M., DeSimone, H. and Parmigiani, G. (1996), “Prediction via Orthogonalized Model Mixing,” *Journal of the American Statistical Association*, **91**, 1197–1208.

Clyde, M. and George, E.I. (1998), “Flexible Empirical Bayes Estimation for Wavelets,” ISDS Discussion Paper 98-21, Duke University.

Dass, S. and Berger, J. (1998), “Unified Bayesian and Conditional Frequentist Testing of Composite Hypotheses,” ISDS Discussion Paper 98-43, Duke University.

De Santis, F., and Spezzaferri, F. (1996), “Comparing Hierarchical Models Using Bayes Factor and Fractional Bayes Factors: a Robust Analysis,” in *Bayesian Robustness*, eds. J. Berger, et. al., Hayward: IMS Lecture Notes, Volume 29.

De Santis, F., and Spezzaferri, F. (1997), “Alternative Bayes Factors for Model Selection,” *The Canadian Journal of Statistics*, **25**, 503-515.

De Santis, F., and Spezzaferri, F. (1998), “Consistent Fractional Bayes Factor for Linear Models,” *Pubblicazioni Scientifiche del Dipartimento di Statistica, Probab. e Stat. Appl.,* Univ. di Roma, “La Sapienza”, Serie a, n.19.

De Santis, F., and Spezzaferri, F. (1999a), “Methods for Robust and Default Bayesian Model Selection: the Fractional Bayes Factor Approach,” *International Statistical Review*, **67**, 1–20.

De Santis, F., and Spezzaferri, F. (1999b), “Bayes Factors and Hierarchical Models,” *Journal of Statistical Planning and Inference*, to appear.

de Vos, A. F. (1993), “A Fair Comparison Between Regression Models of Different Dimension,” Technical Report, The Free University, Amsterdam.

Dmochowski, J. (1996), “Intrinsic Priors Via Kullback-Leibler Geometry,” in *Bayesian Statistics 5*, eds J.M. Bernardo et al., London: Oxford University Press, pp. 543-549.

Draper, D. (1995) ”Assessment and Propagation of Model Uncertainty”, *Jour. of the Roy. Statist. Soc.*, Series B, **57**, 45-98.

Dudley, R. and Haughton, D. (1997), “Information Criteria for Multiple Data Sets and Restricted Parameters,” *Statistica Sinica*, **7**, 265–284.

Dupuis, J.A. and Robert, C.P. (1998), “Bayesian Variable Selection in Qualitative Models by Kullback-Leibler Projections ,” In *Proceedings of the Workshop on Model Selection*,

W. Racugno, ed., CNR, Collana Atti di Congressi, Pitagora, Editrice, Bologna, pp. 275–305.

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

Findley, D.F. (1991), "Counterexamples to Parsimony and BIC," *Ann. Inst. Statist. Math.*, **43**, 505–514

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.

George, E. I. and Foster, D. P. (1997), "Calibration and Empirical Bayes Variable Selection," Technical Report, University of Texas.

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

Ghosh, J.K. and Samanta, T. (1999), "Nonsubjective Bayesian Testing - an Overview," Technical Report, Indian Statistical Institute, Calcutta.

Goutis, C. and Robert, C.P. (1998), "Model Choice in Generalized Linear Models: a Bayesian Approach via Kullback-Leibler Projections," *Biometrika*, **85**, 29–37.

Green, P. (1995), "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination," *Biometrika*, **82**, 711–732.

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

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. (1997), "Posterior Expected Marginal Likelihood for Testing Hypotheses," *J. of Economics, Asia University*, **21**, 105–134.

Jefferys, W. and Berger, J.O. (1992), "Ockham's Razor and Bayesian Analysis", *American Scientist*, **80**, 64–72.

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," *Journal of the American Statistical Association*, **90**, 773–795.

Kass, R. E. and Vaidyanathan, S. K. (1992), "Approximate Bayes Factors and Orthogonal Parameters, with Application to Testing Equality of Two Binomial Proportions," *Journal of the Royal Statistical Society*, **54**, 129–144.

Kass, R. E. and Wasserman, L. (1995), "A Reference Bayesian Test for Nested Hypotheses and Its Relationship to the Schwarz Criterion," *Journal of the American Statistical*

Association, **90**, 928–934.

Key, J. T., Pericchi, L. R., and Smith, A. F. M. (1999), "Bayesian Model Choice: What and Why?"*Bayesian Statistics 6*, Bernardo J.M. et al, eds., Oxford University Press, pp. 343-370.

Kim, S. and Sun, D. (2000), "Intrinsic Priors for Model Selection Using an Encompassing Model," *Life Time Data Analysis*, in press.

Laud, P.W. and Ibrahim, J. (1995), "Predictive Model Selection," *J. Roy. Statist. Soc. B* **57**, 247–262.

Lingham, R. and Sivaganesan, S. (1997), "Testing Hypotheses About the Power Law Process Under Failure Truncation Using Intrinsic Bayes Factors," *Ann. Inst. Statist. Math.*, **49**, 693–710.

Lingham, R. and Sivaganesan, S. (1999), "Intrinsic Bayes Factor Approach to a Test for the Power Law Process," *J. Statist. Planning and Inference*, to appear.

Moreno, E. (1997), "Bayes Factors for Intrinsic and Fractional Priors in Nested Models: Bayesian Robustness," Technical Report, University of Granada.

Moreno, E., Bertolino, F., and Racugno, W. (1998a), "An Intrinsic Limiting Procedure for Model Selection and Hypothesis Testing", *Journal of the American Statistical Association*, **93**, 1451–1460.

Moreno, E., Bertolino, F., and Racugno, W. (1998b), "Model Selection and Hypothesis Testing," Technical Report, University of Granada.

Morris, C. (1987), Comment to "Testing a Point Null Hypothesis: The Irreconcilability of p-values and Evidence," *J. American Statistical Association*, **82**, 112–139.

Nadal, N. (1999), "El Análisis de Varianza basado en los Factores de Bayes Intrínsecos," Ph.D Thesis, Universidad Simón Bolívar, Venezuela.

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

O'Hagan, A. (1997), "Properties of Intrinsic and Fractional Bayes Factors," *Test* **6**, 101–118.

Pauler, D. (1998), "The Schwarz Criterion and Related Methods for Normal Linear Models," *Biometrika*, 85, 1, pp. 13-27.

Pérez, J. M. (1998), " Development of Conventional Prior Distributions for Model Comparisons," Ph.D. Thesis, Purdue University.

Pérez, J. M. and Berger, J. (2000), " Expected Posterior Prior Distributions for Model Selection," ISDS Discussion Paper, Duke University.

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.

Raftery, A., Madigan, D. and Hoeting, J. (1997), "Bayesian Model Averaging for Linear Regression Models," *J. Amer. Statist. Assoc.* **92**, 179–191.

Sansó, B., Pericchi, L. R., and Moreno, E. (1996), "On the Robustness of the Intrinsic Bayes Factor for Nested Models," in *Bayesian Robustness*, eds. J. Berger, et. al., Hayward: IMS Lecture Notes, Volume 29, pp. 157-176.

Schluter, P.J., Deely, J.J. and Nicholson, A.J. (1999), "The Averaged Bayes Factor: A New Method for Selecting Between Competing Models," Technical Report, University of Canterbury.

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

Shao, J. (1997), "An Asymptotic Theory for Linear Model Selection," *Statistica Sinica*, **7**, 221–264.

Shibata, R. (1981), "An Optimal Selection of Regression Variables," *Biometrika*, **68**, 45-54.

Shui, C. (1996), "Default Bayesian Analysis of Mixture Models," Ph.D. Thesis, Department of Statistics, Purdue University.

Sivaganesan, S. and Lingham, R. (1998), "Bayes Factors for a Test About the Drift of Brownian Motion Under Noninformative Priors," Technical Report, Division of Statistics, Northern Illinois University.

Sivaganesan, S. and Lingham, R. (1998), "Bayes Factors for Model Selection for some Diffusion Processes Under Improper Priors," Technical Report, Division of Statistics, Northern Illinois University.

Smith, A. F. M. (1995), Discussion of O'Hagan, A., *Journal of the Royal Statistical Society*, Ser. B, **57**, 99–138.

Smith, A. F. M. and Spiegelhalter, D. J. (1980), "Bayes Factors and Choice Criteria for Linear Models," *Journal of the Royal Statistical Society*, Ser. B, **42**, 213–220.

Spiegelhalter, D. J. and Smith, A. F. M. (1982), "Bayes Factors for Linear and Log-Linear Models with Vague Prior Information," *Journal of the Royal Statistical Society*, Ser. B, **44**, 377–387.

Stone, M. (1979), "Comments on Model Selection Criteria of Akaike and Schwarz," *Joural of the Royal Statistical Society*, Ser. B, **41**, 276-278.

Sun, D. and Kim, S. (1997), "Intrinsic Priors for Testing Ordered Exponential Means," Technical Report, University of Missouri.

Suzuki, Y. (1983), "On Bayesian Approach to Model Selection," in *Proceedings of the International Statistical Institute*, Voorburg, ISI Publications, 288-291.

Varshavsky, J. (1995), "On the Development of Intrinsic Bayes Factors," Ph.D.Thesis, Purdue University.

Verdinelli, I. and Wasserman, L. (1995), "Computing Bayes Factors Using a Generalization of the Savage-Dickey Density Ratio," *Journal of the American Statistical Association*, **90**, 614-618.

Ye, K. and Berger, J. (1991), "Noninformative Priors for Inferences in Exponential Regression Models," *Biometrika*, **78**, 645–656.

Young, K. and Amiss, J. (1995), "Comparisons of Bayes Factors with Non-informative Priors," Technical Report, University of Surrey.

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.

Zellner, A. (1984), "Posterior Odds Ratios for Regression Hypothesis: General Considerations and Some Specific Results," in *Basic Issues in Econometrics*, Chicago: University of Chicago Press, pp. 275-305.

Zellner, A. (1986), "On Assessing Prior Distributions and Bayesian Regression Analysis with g-prior Distributions," in *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, eds. P.K. Goel and A. Zellner, Amsterdam: North-Holland, pp. 233-243.