Bayes Factors [1]

Robert E. Kass
Carnegie-Mellon University

Adrian E. Raftery
University of Washington

Technical Report no. 254
Department of Statistics, University of Washington
Technical Report no. 571
Department of Statistics, Carnegie-Mellon University
March 1993; Revision 3: July 6, 1994

[1]Robert E. Kass is Professor, Department of Statistics, Carnegie-Mellon University, Pittsburgh, PA 15213. Adrian E. Raftery is Professor of Statistics and Sociology, Department of Statistics, GN-22, University of Washington, Seattle, WA 98195. Kass's research was supported by the National Science Foundation under grant no. DMS-9005858 and by the National Institutes of Health under grant no. RO1-CA54852-01. Raftery's research was supported by ONR contract no. N-00014-91-J-1074, by the Ministère de la Recherche et de l'Espace, Paris, by the Université de Paris VI, and by INRIA, Rocquencourt, France. Raftery thanks the latter two institutions, Paul Deheuvels and Gilles Celeux for hearty hospitality during his Paris sabbatical in which part of this article was written. The authors are grateful to former editor Don Guthrie for encouraging them to write this article, to David Madigan and Larry Wasserman for many helpful comments and discussions, and to Jim Albert, James Dickey, Andrew Gelman, Julia Mortera, Michael Newton, Sue Rosenkranz, Michael Sobel, Mike Titterington, the Editor, the Associate Editor and two anonymous referees for very helpful comments on an earlier version of the paper.

Abstract

In a 1935 paper, and in his book *Theory of Probability*, Jeffreys developed a methodology for quantifying the evidence in favor of a scientific theory. The centerpiece was a number, now called the *Bayes factor*, which is the posterior odds of the null hypothesis when the prior probability on the null is one-half. Although there has been much discussion of Bayesian hypothesis testing in the context of criticism of P-values, less attention has been given to the Bayes factor as a practical tool of applied statistics. In this paper we review and discuss the uses of Bayes factors in the context of five scientific applications in genetics, sports, ecology, sociology and psychology.

The points we emphasize are:
— from Jeffreys' Bayesian point of view, the purpose of hypothesis testing is to evaluate the evidence in favor of a scientific theory;
— Bayes factors offer a way of evaluating evidence *in favor of* a null hypothesis;
— Bayes factors provide a way of incorporating external information into the evaluation of evidence about a hypothesis;
— Bayes factors are very general, and do not require alternative models to be nested;
— several techniques are available for computing Bayes factors, including asymptotic approximations which are easy to compute using the output from standard packages that maximize likelihoods;
— in "non-standard" statistical models that do not satisfy common regularity conditions, it can be technically simpler to calculate Bayes factors than to derive non-Bayesian significance tests;
— the Schwarz criterion (or BIC) gives a crude approximation to the logarithm of the Bayes factor, which is easy to use and does not require evaluation of prior distributions;
— when one is interested in estimation or prediction, Bayes factors may be converted to weights to be attached to various models so that a composite estimate or prediction may be obtained that takes account of structural or model uncertainty;
— algorithms have been proposed that allow model uncertainty to be taken into account when the class of models initially considered is very large;
— Bayes factors are useful for guiding an evolutionary model-building process;
— and, finally, it is important, and feasible, to assess the sensitivity of conclusions to the prior distributions used.

KEY WORDS: Bayesian hypothesis tests; BIC; Importance sampling; Laplace method; Markov chain Monte Carlo; Model selection; Monte Carlo integration; Posterior model probabilities; Posterior odds; Sensitivity analysis; Quadrature; Schwarz criterion; Strength of evidence.

Contents

1 Introduction

The Bayesian approach to hypothesis testing was developed by Jeffreys (1935, 1961) as a major part of his program for scientific inference. Although Jeffreys called his methods "significance tests", apparently borrowing the term from Fisher, this is misleading because Jeffreys's perspective and goals were quite different. Jeffreys was concerned with the comparison of predictions made by two competing scientific theories. In his approach, statistical models are introduced to represent the probability of the data according to each of the two theories and Bayes' theorem is used to compute the posterior probability that one of the theories is correct.

Considerable attention has been given to distinctions between the two approaches (e.g. Berger and Delampady, 1987; Berger and Berry, 1988, and the references therein). Often lost from the controversy, however, are the practical aspects of the Bayesian methods: how conclusions may be drawn from them, how they can provide answers when non-Bayesian methods are hard to construct, what their strengths and limitations are. These concerns are the focus of this article. We will also discuss the Bayesian approach to accounting for uncertainty in the model-building process, which is closely connected to the methodology for hypothesis testing.

In Section 2 we motivate the work with several applications from the areas of genetics, sports, ecology, sociology and psychology. These will help connect hypothesis testing with model selection and will introduce several problems that Bayesian methodology can solve, including the evaluation of the evidence in favor of a null hypothesis, the inclusion of other information in the weighing of evidence, the comparison of non-nested models, and accounting for uncertainty in the choice of models. In Section 3 we introduce the *Bayes factor*, which is the posterior odds of one hypothesis when the prior probabilities of the two hypotheses are equal.

Bayesian methods involve integrals and thus, often, numerical integration. Many integration techniques have been adapted to problems of Bayesian inference, including the computation of Bayes factors; this is discussed in Section 4. Bayes factors require priors on the parameters appearing in the models that represent the competing hypotheses. The choice of these priors and the extent to which Bayes factors are sensitive to this choice is discussed in Section 5.

In Section 6 we take up the problem of accounting for uncertainty about model form. The data analyst is often faced with many models which involve different assumptions,

distributional forms, or sets of covariates. Although he or she may wish to summarize findings with a single model, there are usually many choices to be made, and in estimating quantities of interest it is desirable to provide an assessment of the uncertainty that accounts for the model-building process itself. This can be done with Bayesian methods, in which Bayes factors are used to calculate the posterior probabilities of the models considered.

In Section 7 we return to the applications and show how the methods reviewed in Sections 3–6 may be used to solve the problems posed in Section 2.

There has been a lot of controversy about Bayes factors. In Section 8 we discuss several issues, including the purpose of testing sharp hypotheses, disagreements between Bayes factors and P-values, and alternative model selection criteria. In Section 9 we briefly mention some other work and we give references to the calculation of Bayes factors for some specific models. Finally, in Section 10, we conclude by summarizing the most important points, and by highlighting outstanding problems.

2 Applications

In this section we present five applications that pose problems usefully solved with Bayes factors. In the first two the goal is to evaluate the evidence *in favor of* a null hypothesis. The third involves irregular, non-nested models. The fourth has to do with drawing inferences while trying to account for the uncertainty in modeling, and the fifth with determination of which of two sets of alternative explanatory variables better predicts some binary repeated-measures data; the latter will lead to computational difficulties solved by some of the techniques reviewed in this paper. Here we describe the problems; in Section 7 we will describe solutions to them.

Application 1: *E. coli* mutagenesis.

In an experiment in molecular biology (Sklar and Strauss, 1980), the investigators hypothesized that in the uvrE strain of *E. coli* bacteria, mutations leading to "acetate utilization deficiency" would occur by an unusual error-prone DNA repair mechanism. As a consequence, this mutation would fail to be linked to mutations at neighboring loci. Specifically, they noted that if the acetate utilization deficiency mutation occurred during DNA replication, it would be linked to the relatively rare trait of "rifampin resistance", but if the error-prone repair mechanism were responsible there would be no such linkage. The investigators therefore created a pair of cell lines, of which one contained cells "selected"

for rifampin resistance, and the other contained "unselected" cells. The absence of linkage, predicted by the error-prone repair hypothesis, would imply that the proportions p_1 and p_2 of bacteria exhibiting acetate utilization deficiency in the "selected" and "unselected" cell lines would be equal. When the investigators took samples from each cell line and found $\hat{p}_1$ and $\hat{p}_2$ to be approximately equal they believed that this ought to have represented fairly strong evidence in favor of error-prone DNA repair.

Already there is an interesting problem here. The investigators understood that the hypothesis of no linkage corresponded to the statistical null hypothesis $H_0 : p_1 = p_2$, and they computed Pearson's chi-squared statistic to test it. They found that the chi-squared test did not reject the null hypothesis $p_1 = p_2$, but they were aware that the usual interpretation of significance tests is that they may be used only to *reject* hypotheses and do not offer an assessment of the strength of the evidence *in favor of* the null hypothesis. Thus, they were left in doubt about the question they thought their data should have been able to answer.

In fact, there is more to the story. They also had data on 12 other strains of *E. coli*, which showed a range of variation in the differences between p_1 and p_2. There was clearly considerable related information provided by the other strains, and it was desirable to try to use that information in the statistical analysis of the uvrE data. □

This first application poses several problems. First, the investigators wished to know the *strength of the evidence* provided by the data in favor of their scientific hypothesis, which was translated to a statistical hypothesis. Second, the hypothesis at issue was the *null*, and the failure of the Pearson chi-squared statistic to reject it did not indicate the strength of evidence in its favor. Third, there were other data that could be used to construct an alternative hypothesis, i.e., there was prior information. These facets of the problem suggest that the Bayesian approach could be useful.

In the next application, evidence in favor of the null is again at issue, but prior information is less well documented. This example concerns the evidence against extra-Binomial variability.

Application 2: The Hot Hand

It has been argued that belief in the "hot hand" in basketball is based on misperception of random sequences (Gilovich, Vallone, and Tversky, 1985). That is, erratic behavior of shooting may not reflect any real tendency for players to have good streaks or bad streaks, but may instead be consistent with a stable shooting percentage. If there were a strong tendency

for players to have good and bad streaks, then one would expect to see this expressed in good and bad days. In fact, many people who engage in athletic activity do seem to think that they themselves have good days and bad days. Fans often have similar beliefs about players. As a check on results reported by Gilovich, Vallone, and Tversky, one of us (Kass) collected data on Larry Bird's game-by-game performance during the 1986-1987 season. Bird's field goal shooting percentage ranged from 21% to 92% with an average of 53.5% over the 44 games that were reported in *The Boston Globe*. The null hypothesis was that Bird's shooting percentages resulted from 44 Binomial distributions all having the same probability of success, that is, from 44 batches of independent coin flips with a "Larry Bird coin".

It turned out that the null-hypothetical Binomial model fit these data fairly well. That is, there was not enough variability in these data to reject the null hypothesis that each shot was exactly analogous to a flip of a coin. Here, however, it was not clear whether the data failed to reject the null hypothesis because the null hypothesis was true or because the sample size was too small. To check this, an alternative to the Binomial could be constructed and then the posterior odds of the Binomial could be calculated. A particularly simple alternative is to assume that the Binomial parameters vary across games according to a Beta distribution. In addition, Kass collected more data, using Vinnie Johnson, the player who, at that time, seemed to be the most commonly cited example of a "streak shooter", as opposed to Larry Bird who was viewed as a relatively stable shooter. Kass obtained Johnson's shooting data for 380 games during the 1985-1989 seasons. □

Bayes factors allow easy comparison of non-nested models, and of irregular models. This is illustrated in the next application.

Application 3: Ozone Exceedances

As described by Smith (1989), ground-level ozone is a topic of concern because high levels of ozone indicate that the air is polluted. U.S. standards specify that a threshold level be exceeded on no more than three days in any three-year period. In a number of U.S. cities, including Houston, Texas, this standard is far from being met, and the task of regulatory bodies such as the Texas Air Control Board is to introduce measures to reduce the frequency and level of high exceedances.

The question of whether ozone exceedances were decreasing in Houston may be examined using the times of occurrence of clusters of exceedances above the threshold of 16 parts per 100 million (pphm) between 1974 and 1986, deseasonalized; the data are given in Raftery

(1989, Table 1). Exceedances may be modeled as a Poisson process with a rate that is perhaps varying over time. Raftery (1989) considered three models for the rate that represent (a) no change, (b) a gradual decrease, and (c) an abrupt decrease. There are several interesting features of this application. Models (b) and (c) are not nested. In addition, model (c) is not a regular statistical model in that it has a highly discontinuous likelihood function. As a result, frequentist methods for testing (a) against (c) are complex to develop, while Bayes factors are fairly simple to calculate. We know of no frequentist way of testing (b) against (c). The analysis suggested a more precise mechanism, and, potentially, a more developed statistical model, illustrating that Bayes factors are not restricted to the comparison of previously formulated hypotheses, but are also useful for guiding an evolving model-building process. □

In the foregoing applications, simple plausible models were used to represent competing hypotheses. In each case, though, one could phrase the question differently by asking which of the models fits the data better. That is, the problem of testing a hypothesis was identified with one of model selection. The connotation of the phrase "model selection", however, differs according to whether the model comes from underlying theory or from looking at the data. If the latter, Bayes factors can still be used, but choosing a model may no longer be the primary goal.

This arises, for example, when one is interested in the effect of one variable on another and where there are several other possible covariates to be included in a regression equation. One is often not sure of having chosen exactly the right set of covariates, and this is a source of uncertainty which should be taken into account. Similarly, other functional or distributional assumptions may lead to different estimates of quantities of interest and, again, one would like to take account of uncertainty about the assumptions within the estimation process. The Bayesian approach allows this to be done in a natural way by averaging over the candidate models with their posterior probabilities as weights. The following application raises this issue, where the competing models correspond to different link functions in a generalized linear model.

Application 4: Educational Transitions

Do social class background, ability, and type of school attended affect educational attainment? These questions were addressed in the context of Ireland by Greaney and Kelleghan (1984), who concluded that the Irish educational system is approaching the meritocratic

ideal. By this they meant that progress within the system is determined largely by educationally relevant attributes, and not by other, educationally irrelevant, attributes such as social class. One question of particular policy interest is whether and to what extent students in vocational (non-academic second-level) schools drop out of school earlier than their counterparts in secondary (academic second-level) schools with the same ability and social class origin.

The questions may be addressed by reanalyzing the longitudinal data of Greaney and Kelleghan (1984) using models based on logistic regression. They then reduce to questions about the presence or absence of the effects of interest in the regression model and their size. Questions about the size of these effects have to be answered in the presence of several competing models (in this case different link functions) between which the data do not distinguish clearly, but which nevertheless yield very different results. It is essential to take account of this model uncertainty. □

In the final application, simple approximate methods were applied in a modestly complicated setting (repeated-measures logistic regression with alternative predictors). It was desired to check the modeling assumptions and the accuracy of the asymptotics used. This raised problems of computation and the determination of prior distributions.

Application 5: Human Working Memory Failure in Computer-Based Tasks.

In human-computer interaction it is of interest to characterize tasks that tend to lead to human error, so that procedures may be written to avoid them. Carlin, Kass, Lerch, and Huguenard (1992) have described one such effort in an experimental study involving the database management system SQL, which presented subjects with query tasks of varying complexity (such as "Find all customers having an outstanding invoice of more than $200"). Understanding of errors is guided by cognitive psychological theory concerning overload of what is now called "working memory" (a concept which has replaced and refined what used to be called short-term memory). Thus, the notion is that errors tend to occur when human working memory is overloaded because of excessive demands in the task. At issue in this study was whether overload tended to be due to (i) the number of conditions in the query or (ii) the complexity of the query. To determine which characterization led to better predictions of error rates, two alternative statistical models were constructed based on alternative predictor variables, either the number of conditions or a pair of measures of

query complexity.

There were 20 experimental subjects, each of whom was given 50 tasks to complete. Half the subjects received a "cue", which was supposed to remind them of the essential feature of the problem and to prevent them from making the error under study. Carlin *et al.* used a logistic regression model with random subject ability effects (a single number for each subject) nested within the cue effect, together with the respective predictor variables. The two competing models had four parameters in common and the number-of-conditions model had one additional parameter (the coefficient of the predictor) while the query-complexity model had two additional parameters (the coefficients of the two predictors). Evaluation of the likelihood in either model required the numerical calculation of an integral due to the random subject effects. □

3 Bayes Factors

3.1 Definition

We begin with data D assumed to have arisen under one of the two hypotheses H_1 and H_2 according to a probability density $\mathrm{pr}(D|H_1)$ or $\mathrm{pr}(D|H_2)$. Given *a priori* probabilities $\mathrm{pr}(H_1)$ and $\mathrm{pr}(H_2) = 1 - \mathrm{pr}(H_1)$, the data produce *a posteriori* probabilities $\mathrm{pr}(H_1|D)$ and $\mathrm{pr}(H_2|D) = 1 - \mathrm{pr}(H_1|D)$. Since any prior opinion gets transformed to a posterior opinion through consideration of the data, the transformation itself represents the evidence provided by the data. In fact, the same transformation is used to obtain the posterior probability, regardless of the prior probability. Once we convert to the odds scale (odds = probability/(1 - probability)), the transformation takes a simple form. From Bayes' Theorem we obtain

$$\mathrm{pr}(H_k|D) = \frac{\mathrm{pr}(D|H_k)\mathrm{pr}(H_k)}{\mathrm{pr}(D|H_1)\mathrm{pr}(H_1) + \mathrm{pr}(D|H_2)\mathrm{pr}(H_2)} \quad (k = 1, 2),$$

so that

$$\frac{\mathrm{pr}(H_1|D)}{\mathrm{pr}(H_2|D)} = \frac{\mathrm{pr}(D|H_1)}{\mathrm{pr}(D|H_2)} \frac{\mathrm{pr}(H_1)}{\mathrm{pr}(H_2)},$$

and the transformation is simply multiplication by

$$B_{12} = \frac{\mathrm{pr}(D|H_1)}{\mathrm{pr}(D|H_2)}, \tag{1}$$

which is the *Bayes factor*. Thus, in words,

$$\text{posterior odds} = \text{Bayes factor} \times \text{prior odds},$$

and the Bayes factor is the ratio of the posterior odds of H_1 to its prior odds, regardless of the value of the prior odds. (The terminology is apparently due to I.J. Good, who attributes the method to Turing in addition to, and independently of, Jeffreys at about the same time; see Good, 1983.) When the hypotheses H_1 and H_2 are equally probable *a priori* so that $\mathrm{pr}(H_1) = \mathrm{pr}(H_2) = 0.5$, the Bayes factor is equal to the posterior odds in favor of H_1. The two hypotheses may well not be equally likely *a priori*, however.

In the simplest case, when the two hypotheses are single distributions with no free parameters (the case of "simple versus simple" testing), B_{12} is the likelihood ratio. In other cases, when there are unknown parameters under either or both of the hypotheses, the Bayes factor is still given by (1) and, in a sense, it continues to have the form of a likelihood ratio. Then, however, the densities $\mathrm{pr}(D|H_k)$ $(k = 1, 2)$ are obtained by *integrating* (not maximizing) over the parameter space, so that in equation (1),

$$\mathrm{pr}(D|H_k) = \int \mathrm{pr}(D|\theta_k, H_k)\pi(\theta_k|H_k)\mathrm{d}\theta_k, \tag{2}$$

where θ_k is the parameter under H_k, $\pi(\theta_k|H_k)$ is its prior density and $\mathrm{pr}(D|\theta_k, H_k)$ is the probability density of D given the value of θ_k, or the likelihood function of θ. Here, θ_k may be a vector and in what follows we will denote its dimension by d_k.

The prior distributions $\pi(\theta_k|H_k)$ $(k = 1, 2)$ are necessary. This may be considered both good and bad. Good, because it is a way of including other information about the values of the parameters (as in Application 1). Bad, because these prior densities may be hard to set when there is no such information. We discuss the problem of setting priors and assessing sensitivity to the choices in Section 5.

The quantity $\mathrm{pr}(D|H_k)$ given by equation (2) is the marginal probability of the data, since it is obtained by integrating the joint density of (D, θ_k) given D over θ_k. It is also the predictive probability of the data, i.e., the probability of seeing the data that actually were observed, calculated *before* any data became available. It is also sometimes called a marginal likelihood, or an integrated likelihood. Note that, as in computing the likelihood ratio statistic, but unlike in some other applications of likelihood, all constants appearing in the definition of the likelihood $\mathrm{pr}(D|\theta_k, H_k)$ must be retained when computing B_{12}. In fact, B_{12} is closely related to the likelihood ratio statistic, in which the parameters θ_k are eliminated by maximization rather than by integration. We discuss this relationship in Sections 4.1.2 and 4.1.3.

Other notations are often used for the Bayes factor. When many hypotheses are involved we will write B_{jk} as the Bayes factor for H_j against H_k. Often, one of the hypotheses

is considered the null and is therefore denoted by H_0. In this case, if there is only one alternative it will be denoted by H_1 so that, putting $\mathrm{pr}(D|H_0)$ in the numerator, the Bayes factor becomes B_{01}. In this situation, Jeffreys instead used K. When comparing results with standard likelihood ratio tests it is convenient to instead put the null hypothesis in the denominator of (1) and thus use B_{10} as the Bayes factor.

For the usual (non-Bayesian) large-sample distribution theory of likelihood ratio tests to be applicable, the null must be nested within the alternative. That is, there must be some parameterization under H_1 of the form $\theta = (\beta, \psi)$ such that H_0 is obtained from H_1 when $\psi = \psi_0$ for some ψ_0. Here, both β and ψ may be vectors. Although the expression (1) does not require the models to be nested, the case of nested models is of special interest in the Bayesian approach as well and we will refer to it frequently below.

3.2 Interpretation

The Bayes factor is a summary of the evidence provided by the data in favor of one scientific theory, represented by a statistical model, as opposed to another. Jeffreys (1961, Appendix B) suggested interpreting B_{10} in half-units on the $\log_{10}$ scale. Pooling two of his categories together for simplification we have:

$\log_{10}(B_{10})$	B_{10}	Evidence against H_0
0 to $\frac{1}{2}$	1 to 3.2	Not worth more than a bare mention
$\frac{1}{2}$ to 1	3.2 to 10	Substantial
1 to 2	10 to 100	Strong
> 2	> 100	Decisive

Probability itself provides a meaningful scale defined by betting, and so these categories are not a calibration of the Bayes factor, but rather a rough descriptive statement about standards of evidence in scientific investigation. We speak here in terms of B_{10} because weighing evidence *against* a null hypothesis is more familiar, but Bayes factors can equally well provide evidence *in favor of* a null hypothesis. Of course, the interpretation may depend on the context. For example, Evett (1991) has argued that for forensic evidence alone to be conclusive in a criminal trial one would require posterior odds for H_1 (guilt) against H_0 (innocence) of at least 1000 rather than the 100 suggested by Jeffreys.

It can be useful to consider twice the natural logarithm of the Bayes factor, which is on the same scale as the familiar deviance and likelihood ratio test statistics. Rounding, we then obtain a slight modification:

$2\log_e(B_{10})$	B_{10}	Evidence against H_0
0 to 2	1 to 3	Not worth more than a bare mention
2 to 5	3 to 12	Positive
5 to 10	12 to 150	Strong
> 10	> 150	Decisive

From our own experience, these categories seem to furnish appropriate guidelines.

The logarithm of the marginal probability of the data may also be viewed as a predictive score. This is of interest because it leads to an interpretation of the Bayes factor that does not depend on viewing one of the models as "true". Suppose that $D = \{y_1, \ldots, y_n\}$ and that, for each i, we form a predictive distribution $\hat{\mathrm{pr}}_i(\cdot)$ of y_i given the already available data $\{y_1, \ldots, y_{i-1}\}$. We use the logarithmic scoring rule, $\log \hat{\mathrm{pr}}_i(y_i)$ (Good, 1952), to assess performance. Then the overall score of any rule that generates such predictive distributions is $LS = \sum_i \log \hat{\mathrm{pr}}_i(y_i)$. In particular, if the prediction rule is derived from the model H_k (i.e. likelihood and prior), then $\log \mathrm{pr}(D|H_k) = \sum_i \log \mathrm{pr}(y_i|y_{i-1}, \ldots, y_1, H_k) = LS_k$. It follows that the log Bayes factor is $\log B_{10} = LS_1 - LS_0$, i.e. the difference in predictive scores. Thus the Bayes factor can be viewed as measuring the relative success of H_1 and H_0 at predicting the data. This is related to *prequential analysis* (Dawid, 1984) and also to *stochastic complexity* (Rissanen, 1987); the connections are discussed by Dawid (1992) and Hartigan (1992). Good (1985), and in many other publications, refers to the log Bayes factor as the "weight of evidence".

4 Calculating Bayes Factors

In some elementary cases the integral (2), which we will rewrite in this section as

$$I = \int \mathrm{pr}(D|\theta, H)\pi(\theta|H)\mathrm{d}\theta \tag{3}$$

may be evaluated analytically. More often, it is intractable and thus must be computed by numerical methods. Most available software developed by numerical analysts, however, is generally so inefficient for these integrals that it is of little use. One reason is that when sample sizes are moderate or large the integrand becomes highly peaked around its maximum, which may be found by other techniques, and quadrature methods that do not begin with knowledge of the maximum are likely to have difficulty finding the region where the integrand mass is accumulating. A second reason is that some problems are of high dimension. In this case, Monte Carlo methods may be used, but these, too, need to be adapted to the statistical context.

Exact analytic evaluation of the integral (3) is possible for exponential family distributions with conjugate priors, including normal linear models (DeGroot, 1970, Chapter 9; Zellner, 1971, Chapter 10).

4.1 Asymptotic Approximation

4.1.1 Laplace's method

A useful approximation to the marginal density of the data as given by (3) is obtained by assuming that the the posterior density, which is proportional to $(\mathrm{pr}(D|\theta,H)\pi(\theta|H))$, is highly peaked about its maximum $\tilde{\theta}$, which is the posterior mode. This will usually be the case if the likelihood function $\mathrm{pr}(D|\theta,H)$ is highly peaked near its maximum $\hat{\theta}$, which will be the case for large samples. Let $\tilde{\ell}(\theta) = \log\left(\mathrm{pr}(D|\theta,H)\pi(\theta|H)\right)$. Expanding $\tilde{\ell}(\theta)$ as a quadratic about $\tilde{\theta}$ and then exponentiating yields an approximation to $(\mathrm{pr}(D|\theta,H)\pi(\theta|H))$ that has the form of a Normal density with mean $\tilde{\theta}$ and covariance matrix $\tilde{\Sigma} = \left(-D^2\tilde{\ell}(\tilde{\theta})\right)^{-1}$, where $D^2\tilde{\ell}(\tilde{\theta})$ is the Hessian matrix of second derivatives. Integrating this approximation yields

$$\hat{I} = (2\pi)^{d/2}|\tilde{\Sigma}|^{\frac{1}{2}}\mathrm{pr}(D|\tilde{\theta},H)\pi(\tilde{\theta}|H), \tag{4}$$

where d is the dimension of θ.

This is Laplace's method of approximation (de Bruijn, 1970, Section 4.4; Tierney and Kadane, 1986). For many problems in which the sample size n is moderate it produces answers well within the accuracy required for drawing conclusions according to the scheme of Section 3.2. Formally, under conditions spelled out in Kass, Tierney, and Kadane (1990), as $n \to \infty$, $I = \hat{I}\,(1 + O(n^{-1}))$, i.e., the relative error is $O(n^{-1})$. Thus, when Laplace's method is applied to both the numerator and denominator of B_{10} in (1) the resulting approximation also has relative error of order $O(n^{-1})$.

The accuracy of Laplace's method in this context has been examined by Kass and Vaidyanathan (1992), Rosenkranz (1992) and Raftery (1993c), and is mentioned in Applications 4 and 5. In general, the method provides adequate approximations for well-behaved problems (those in which the likelihood functions are not grossly non-Normal) of modest dimensionality. We are unable to be much more specific than this, though our rough feeling is that samples of size less than $5d$ are worrisome (with d being the dimension of θ) while those of size greater than $20d$ are large enough for the method to work well in most familiar problems, provided that a reasonably good parameterization is used. See Slate (1994) for a detailed discussion of sample sizes required to obtain posterior Normality (which would

guarantee accuracy of Laplace's method) for various parameterizations of exponential families. Software for applying Laplace's method is available in S and in LispStat (Tierney, 1989, 1990; Raftery, 1993c). Approximations of this kind have been used by many authors, notably Jeffreys (1961), Lindley (1961), Mosteller and Wallace (1964), and Leonard (1982).

4.1.2 Variants on Laplace's method

Laplace's method may be applied in alternative forms by omitting part of the integrand from the exponent when performing the expansion. For the general formulation see Kass and Vaidyanathan (1992), who followed Tierney, Kass, and Kadane (1989), and Mosteller and Wallace (1964, Section 4.6). An important variant on (4) is

$$\hat{I}_{MLE} = (2\pi)^{d/2}|\hat{\Sigma}|^{\frac{1}{2}}\mathrm{pr}(D|\hat{\theta},H)\pi(\hat{\theta}|H) \tag{5}$$

where $\hat{\Sigma}^{-1}$ is the observed information matrix, i.e., the negative Hessian matrix of the log-likelihood evaluated at the MLE $\hat{\theta}$. This approximation again has relative error of order $O(n^{-1})$. Although it is likely to be less accurate than (4) when the prior is somewhat informative relative to the likelihood, it has the advantage that it is easily computed from any statistical software package that reports the MLE, the observed information matrix (or its inverse), and the value of the maximized likelihood.

Some software packages calculate the expected information matrix (the usual Fisher information matrix), or its inverse, as the asymptotic covariance matrix, rather than the observed information matrix. The inverse of the expected information matrix may be used in place of $\hat{\Sigma}$ in (5). The resulting approximation has a larger asymptotic relative error, of order $O(n^{-\frac{1}{2}})$, but it remains sufficiently accurate to be of use in many problems.

Now suppose that we have nested hypotheses, with parameter (β,ψ) having prior $\pi(\beta,\psi|H_1)$ under H_1 and then $H_0:\psi=\psi_0$ with prior $\pi(\beta|H_0)$. When (5) is applied we obtain

$$2\log B_{10} \approx \Lambda + \log|\hat{\Sigma}_1| - \log|\hat{\Sigma}_0| + \log\pi(\hat{\beta},\hat{\psi}|H_1) - \log\pi(\hat{\beta}^*,|H_0) + (d_1-d_0)\log(2\pi) \tag{6}$$

where $\Lambda = 2(\log\mathrm{pr}(D|(\hat{\beta},\hat{\psi}),H_1) - \log\mathrm{pr}(D|\hat{\beta},H_0))$ is the log likelihood ratio statistic having degrees of freedom (d_1-d_0) and $\hat{\beta}^*$ denotes the MLE under H_0. Again, either observed or expected information may be used in computing the covariance matrices $\hat{\Sigma}_k$ and (6) then has relative error of order $O(n^{-1})$ or $O(n^{-\frac{1}{2}})$, respectively. Jeffreys (1961, Section 5.31) gave an approximation for the case where $(d_1-d_0)=1$ in essentially the form of (6) but with expected rather than observed information, and Chow (1981) extended Jeffreys' result to higher dimensions.

Raftery (1993c) suggested approximating $\tilde{\theta}_k$ by a single Newton step starting from $\hat{\theta}_k$ and substituting the result into equation (4), which yields the approximation

$$2 \log B_{10} \approx \Lambda + (E_1 - E_0). \tag{7}$$

In equation (7),

$$E_k = 2\lambda_k(\hat{\theta}_k) + \lambda_k'(\hat{\theta}_k)^T (F_k + G_k)^{-1} \{2 - F_k (F_k + G_k)^{-1}\} \lambda_k'(\hat{\theta}_k) - \log |F_k + G_k| + d_k \log(2\pi), \tag{8}$$

where $\mathrm{Var}\,[\theta_k | H_k] = W_k$, $F_k = \hat{\Sigma}_k^{-1}$, $G_k = W_k^{-1}$, $\lambda_k(\theta_k) = \log \mathrm{pr}(\theta_k | H_k)$ is the log-prior density, and $\lambda_k'(\hat{\theta}_k)$ is the d_k-vector of derivatives of $\lambda_k(\theta_k)$ with respect to the elements of θ_k $(k = 0, 1)$. This often improves on the approximation of (5) but it does not require any additional information. It is implemented for generalized linear models in the GLIB software; see Section 9.

4.1.3 The Schwarz criterion

It is possible to avoid the introduction of the prior densities $\pi_k(\theta_k | H_k)$ in (1) by using

$$S = \log \mathrm{pr}(D | \hat{\theta}_1, H_1) - \log \mathrm{pr}(D | \hat{\theta}_2, H_2) - \tfrac{1}{2} (d_1 - d_2) \log(n)$$

where $\hat{\theta}_k$ is the MLE under H_k, d_k is the dimension of θ_k, and n is the sample size. As $n \to \infty$, this quantity, often called the Schwarz criterion, satisfies

$$\frac{S - \log B_{12}}{\log B_{12}} \to 0 \tag{9}$$

and thus may be viewed as a rough approximation to the logarithm of the Bayes factor. Minus twice the Schwarz criterion is often called the Bayes Information Criterion (BIC); sometimes an arbitrary constant is added. See Section 8.3 for additional discussion.

In contrast to the approximations furnished by (4) or (5) the relative error of $\exp(S)$ in approximating B_{12} is generally $O(1)$. Thus, even for very large samples it does not produce the correct value. On the other hand, keeping in mind the rough interpretation of B_{12} on the logarithmic scale suggested in Section 3.2, equation (9) shows that in large samples the Schwarz criterion should provide a reasonable indication of the evidence.

The Schwarz criterion is appealing in that it can be applied as a standard procedure even when the priors $\pi(\theta_k | H_k)$ are hard to set precisely. In this sense it provides an often-useful reference procedure for scientific reporting. Also in its favor is the following. For independent and identically distributed sampling in nested models with $H_0 : \psi = \psi_0$ it is

reasonable to take the prior on ψ to be $N(\psi_0, W)$ with $|W| = |J(\psi_0)|^{-1}$ where $J(\psi)$ is the Fisher information in the parameter ψ. This says that the amount of information in the prior is equal to the amount of information in one observation. In this case, under certain conditions (the assumptions of null-orthogonality and local alternative mentioned in Section 5.2), the Schwarz criterion furnishes an order $O(n^{-\frac{1}{2}})$ approximation to $\log B_{12}$ (Kass and Wasserman, 1992b). Thus, if one were willing to use this prior as a "reference prior" suitable for standardized reporting, the Schwarz criterion would be a reasonably good approximation to the log of the Bayes factor. As Kass and Wasserman (1992b) note, Jeffreys's method was essentially to use $S + c$ as an approximation to the log of the Bayes factor, with c being a constant determined by substituting a Cauchy prior in place of the Normal prior on ψ.

The sample size n in the definition of S needs to be figured carefully. It is apparent from the derivation of S (e.g., Kass, 1993) that n should be the rate at which the Hessian matrix of the loglikelihood function grows; thus, n becomes the number of data values contributing to the summation that appears in the formula for the Hessian. For instance, as Raftery (1986a) pointed out, in the case of loglinear models for contingency tables n is the sum of the counts and not the number of cells. Similarly, in models for binomial responses, n is the sum of the denominators, and not either the number of responses or the number of successes. In survival analysis, n is the number of *uncensored* observations, i.e., of deaths (Raftery, Madigan and Volinsky, 1995). In comparing models of multivariate Normal structure (e.g. structural equation models), n is the total number of *scalar* observations, i.e., the number of vector observations times their dimension (Raftery, 1993b). In two-stage hierarchical models the appropriate n depends on which parameters are being tested. We hope to elaborate on this subtle but important matter elsewhere.

Schwarz (1978) gave a rigorous derivation of the criterion for linear subfamilies of exponential families and Haughton (1988) extended Schwarz's result to curved exponential families. Heuristic derivations of (9) are quite easy: one applies (5) and then neglects constant-order terms. It is apparent from arguments such as those in Kass, Tierney, and Kadane (1990) that (9) holds much more generally than in the restricted setting of curved exponential families (e.g. Katz, 1981; Leonard, 1982; Raftery, 1986a; Kass, 1993). Rigorously demonstrating the assumptions for the validity of Laplace's method, however, seems to be enough of a chore that no very general precisely formulated result has been published.

4.2 Simple Monte Carlo, Importance Sampling, and Gaussian Quadrature

Dropping the notational dependence on H_k, equation (2) becomes

$$\mathrm{pr}(D) = \int \mathrm{pr}(D|\theta)\pi(\theta)\mathrm{d}\theta.$$

The simplest Monte Carlo integration estimate of this is

$$\hat{\mathrm{pr}}_1(D) = \frac{1}{m}\sum_{i=1}^{m} \mathrm{pr}(D|\theta^{(i)}),$$

where $\{\theta^{(i)} : i = 1, \ldots, m\}$ is a sample from the prior distribution; this is the average of the likelihoods of the sampled parameter values (e.g. Hammersley and Handscomb, 1964).

This possibility was mentioned by Raftery and Banfield (1991), and was investigated in some detail in particular cases by McCulloch and Rossi (1991). A major difficulty with $\hat{\mathrm{pr}}_1(D)$ is that most of the $\theta^{(i)}$ have small likelihood values if the posterior is concentrated relative to the prior, so that the simulation process will be quite inefficient. Thus the estimate is dominated by a few large values of the likelihood, and so the variance of $\hat{\mathrm{pr}}_1(D)$ is large and its convergence to a Gaussian distribution is slow. These problems were apparent in the examples studied in detail by McCulloch and Rossi (1991).

The precision of simple Monte Carlo integration can be improved by importance sampling. This consists of generating a sample $\{\theta^{(i)} : i = 1, \ldots, m\}$ from a density $\pi^*(\theta)$. Under quite general conditions, a simulation-consistent estimate of I is

$$\hat{I} = \frac{\sum_{i=1}^{m} w_i \mathrm{pr}(D|\theta^{(i)})}{\sum_{i=1}^{m} w_i}, \tag{10}$$

where $w_i = \pi(\theta^{(i)})/\pi^*(\theta^{(i)})$; the function $\pi^*(\theta)$ is known as the *importance sampling function*. For general discussion and references see Geweke (1989).

A more efficient scheme is based on adaptive Gaussian quadrature. Using well-established methods from the numerical analysis literature, Genz and Kass (1993) show how integrals that are peaked around a dominant mode may be evaluated. This approach is effective in such problems when the dimensionality of the parameter space is modest (roughly, less than 15) .

4.3 Simulating from the Posterior

There are now several methods available for simulating from posterior distributions. In the simplest cases these include direct simulation and rejection sampling. In more complex

cases, Markov chain Monte Carlo (MCMC) methods, particularly the Metropolis-Hastings algorithm and the Gibbs sampler, provide a general recipe (e.g. Smith and Roberts, 1993, and references therein). Another fairly general recipe is the weighted likelihood bootstrap (Newton and Raftery, 1994).

Any of these methods gives us a sample approximately drawn from the posterior density $\pi^*(\theta) = \mathrm{pr}(\theta|D) = \mathrm{pr}(D|\theta)\pi(\theta)/\mathrm{pr}(D)$. Substituting into equation (10) yields as an estimate for $\mathrm{pr}(D)$,

$$\hat{\mathrm{pr}}_2(D) = \left\{ \frac{1}{m} \sum_{i=1}^{m} \mathrm{pr}(D|\theta^{(i)})^{-1} \right\}^{-1}, \tag{11}$$

the harmonic mean of the likelihood values (Newton and Raftery, 1994). This converges almost surely to the correct value, $\mathrm{pr}(D)$, as $m \to \infty$, but it does not, in general, satisfy a Gaussian central limit theorem. This manifests itself by the occasional occurrence of a value of $\theta^{(i)}$ with small likelihood and hence large effect on the final result. However, it is very easy to calculate and experience to date suggests that while it is indeed unstable, it often gives results that are accurate enough for interpretation on the logarithmic scale of Section 3.2 (Rosenkranz, 1992; Carlin and Chib, 1993; Raftery, 1994).

Several modifications of the harmonic mean estimator (11) have been suggested to get around its instability. Newton and Raftery (1994) suggested using as importance sampling function in (10) a mixture of the prior and posterior densities, $\pi^*(\theta) = \delta\pi(\theta) + (1-\delta)\mathrm{pr}(\theta|D)$, where $0 < \delta < 1$. The resulting estimator, $\hat{\mathrm{pr}}_3(D)$, has the efficiency of $\hat{\mathrm{pr}}_2(D)$ due to being based on many values of θ with high likelihood, but avoids its instability and does satisfy a Gaussian central limit theorem. However, it has the irksome aspect that one must simulate from the prior as well as the posterior. This may be avoided by simulating all m values from the posterior distribution and imagining that a further $\delta m/(1-\delta)$ values of θ are drawn from the prior, all of them with likelihoods $\mathrm{pr}(D|\theta^{(i)})$ equal to their expected value $\mathrm{pr}(D)$. The resulting estimator, $\hat{\mathrm{pr}}_4(D)$, may be evaluated using a simple iterative scheme.

A simple modification of (11) is to instead calculate

$$\hat{\mathrm{pr}}_5(D) = \frac{1}{m} \sum_{i=1}^{m} f(\theta^{(i)}) / \left(\mathrm{pr}(D|\theta^{(i)})\pi(\theta^{(i)}) \right), \tag{12}$$

where $f(\cdot)$ is any d-dimensional probability density. This was mentioned by Gelfand and Dey (1994). It is an unbiased and consistent estimator of the marginal likelihood $\mathrm{pr}(D)$, and satisfies a Gaussian central limit theorem if the tails of $f(\cdot)$ are thin enough, specifically if $\int \{f(\theta)^2 / (\mathrm{pr}(D|\theta)\pi(\theta))\} \, \mathrm{d}\theta < \infty$. High efficiency would seem most likely to result if $f(\theta)$ were roughly proportional to $\mathrm{pr}(D|\theta)$. The very limited experience to date indicates that for

low-dimensional problems with a good choice of f, $\hat{\text{pr}}_5(D)$ can be very accurate. For high-dimensional problems, however, it may be hard to find an appropriate f, and the results can be very poor. Meng and Wong (1993) considered an optimal choice of f and showed how this could be computed iteratively from an initial guess. (Their framework is actually somewhat more general because it applies to the computation of a ratio of integrals of the form (3).) Their approach appears promising but has not yet been extensively tested.

Raftery (1994) suggested what he called the "Laplace-Metropolis" estimator of $\text{pr}(D)$, obtained by using the posterior simulation output to estimate the quantities needed to compute the Laplace approximation (4), namely the posterior mode, $\tilde{\theta}$, and minus the inverse Hessian at the posterior mode, $\tilde{\Sigma}$. The posterior mode can be estimated as the $\theta^{(i)}$ that maximizes $\left(\text{pr}(D|\theta^{(i)})\pi(\theta^{(i)})\right)$. This requires computing the likelihood for each simulated $\theta^{(i)}$; if this is takes too much computer time, an alternative is to use the multivariate or componentwise posterior median, or to estimate the posterior mode by nonparametric density estimation. The matrix $\tilde{\Sigma}$ can be estimated by the estimated posterior covariance matrix; it is wise to use a high-breakdown point robust estimator. The resulting estimator has performed well in some limited numerical experiments. A similar combination of simulation with Laplace's method was suggested by Kass and Wasserman (1992a), who provided a correction term as well.

In the case of nested models in which the priors satisfy (13) of Section 5.1 and β and ψ are independent, then

$$B_{10} = \frac{\pi(\psi_0|H_1)}{\text{pr}(\psi_0|D, H_1)}.$$

(Actually, instead of *a priori* independence of β and ψ it is enough to have $\pi(\beta|H_0) = \pi(\beta \mid \psi = \psi_0, H_1)$.) This result is apparently due to L.J. Savage and has been called the "Savage density ratio method" by Dickey (1971). It has been exploited in several applications, e.g., McCulloch and Rossi (1991). Verdinelli and Wasserman (1993a) point out that a method of generating marginal posterior densities discussed by Chen (1992) may then be applied advantageously in this case. This leads to a method similar to $\hat{\text{pr}}_5(D)$, and the choice of f is again crucial and difficult.

Carlin and Chib (1993) suggested including a model indicator variable in the MCMC scheme and defining "pseudo-priors" for $(\theta_1|H_2)$ and $(\theta_2|H_1)$. This involves designing and running a special MCMC algorithm to calculate Bayes factors. Similar suggestions have been made by Carlin and Polson (1991) and George and McCulloch (1993).

An alternative approach is available when many parameters, z, are present in all the models considered. These might be missing data, the values of a random effect in a hi-

erarchical model, or "latent data" chosen in such a way that the "complete data likelihood" $\mathrm{pr}(D, z|\theta)$ has a simple form (Tanner and Wong, 1987). Then the Bayes factor can be simulation-consistently estimated by the average of the quantities $B_{10}(z^{(i)}) = \mathrm{pr}(D, z^{(i)}|H_1)/\mathrm{pr}(D, z^{(i)}|H_0)$, where the $z^{(i)}$ are simulated from the posterior distribution of z under H_0; this is often possible using Markov chain Monte Carlo methods. The $B_{10}(z^{(i)})$ are then often easy to calculate, or at least to approximate fairly well, for example using the Laplace method. When z is present in H_1 but not in H_0, we again recover the harmonic mean estimator of $\mathrm{pr}(D|H_1)$ (Raftery, 1993a). This is related to previous work of Thompson and Wijsman (1990) on the calculation of likelihood ratios.

Finally, some general methods of calculating I have been considered in the statistical physics literature under the name "free energy estimation". The approaches are not automatic and require analytical effort to tailor them to statistical applications. However, they may be of use in certain problems. See Neal (1992) for references.

4.4 Comparison of Methods

The different methods for calculating and approximating Bayes factors have been compared by Rosenkrantz (1992) in the contexts of normal models and of Poisson-gamma models for counts with independent variables, unobserved heterogeneity, and outliers, and by Raftery (1993c) in the context of generalized linear models.

Exact analytic evaluation is best because it is most accurate and usually also most efficient computationally, but it is feasible only for a narrow class of models. The Laplace method for integrals yields accurate approximations and is usually quite computationally efficient. In particular, the approximation using the posterior mode in (4) or its one-step approximation given by equations (7) and (8) can be very accurate. The latter is easy to compute using the output from standard software. For cases of modest dimensionality the adaptive quadrature method of Genz and Kass (1993) is effective. The Monte Carlo integration and importance sampling methods are less precise and more computationally demanding, but they may well be the only ones available in complex models. The methods using MCMC seem promising but have not yet been applied in many demanding problems. In addition, they require large numbers of likelihood function evaluations, which in some cases is itself difficult.

The Schwarz criterion is the easiest approximation to compute and has the advantage of not requiring that the user specify the prior distributions. It is suitable for communicating results even when more detailed calculations are made using other methods (as in Application 5). As an asymptotic approximation it is admittedly crude but, as long as the number of

degrees of freedom involved in the comparison is reasonably small relative to sample size, it does not seem to be grossly misleading in a qualitative sense. It can be very poor, however, when the number of degrees of freedom involved in the comparison is large; McCulloch and Rossi (1991) gave an example of this with 115 degrees of freedom.

5 The Choice of Priors

In order to compute a Bayes factor, the prior distributions $\pi(\theta_k|H_k)$ on the parameters of each model must be specified. Sometimes, as in Application 1, there are closely related data with which to construct priors. This nicest situation is rare, however. More often, as in Application 5, some combination of relevant data, information from the literature and rough guesses must be used. In that case there will be doubts about the accuracy of the prior distribution. Thus, a first concern is how to choose prior distributions to represent the available information, but once this is done an important issue is the sensitivity of the Bayes factor to the choices of priors.

The easiest way to deal with the problem of prior choice is to ignore it and simply use the Schwarz criterion, or Jeffreys's variant of it (see Section 4.1.3). Although this will lead to appropriate conclusions in sufficiently large samples, there is not much available guidance as to the operational meaning of "sufficiently large". Also, in contrast with Bayesian point estimates such as the posterior mean, the Bayes factor does tend to be sensitive to the choices of priors on the model parameters. We discuss informative prior selection and the problem of sensitivity in Section 5.1, and in Section 5.2 we review results and methods for assessing sensitivity over a range of priors. In Section 5.3 we consider the use of improper priors and note the resulting difficulties.

Sensitivity analysis concerns distributional forms for models $\mathrm{pr}(D|\theta_k, H_k)$ as well as priors. When alternatives are introduced (e.g., the Student's t distribution in place of the Normal) Bayes factors may be used to determine which best fits the data. One may also assess the influence of individual data values by computing the Bayes factor after omitting each observation in turn (Pettit and Young, 1990). Asymptotic approximation makes the "leave-one-out" diagnostic approach easy; see Application 5.

5.1 Prior Information

The problem of determining a prior distribution from available information appears throughout Bayesian inference. The information may come from other data, from the subjective

knowledge of experts, or it may be too hard to express as a prior distribution. Even when there are other data, judgment must be used because a distribution must be chosen (as in Application 1), but this is familiar. The use of subjective opinion is different. For the most part, in both of these cases, the formulation of priors is problem-specific and there is not much general methodology for it. There is, however, a modest literature on "eliciting" probabilistic information from individuals. One psychologically appealing device is the method of imaginary observations (Good, 1950). Kadane *et al.* (1980) discuss a formal elicitation procedure in the context of linear regression and Garthwaite (1992) includes more recent references. Formal representation of opinion may be useful for private analysis and in verifying that a simpler public analysis leads to appropriate conclusions, as in Application 5. Prior distributions sometimes become communally accepted after much research and discussion; this happened in the estimation of bowhead whale population size (Raftery and Zeh, 1993).

In choosing priors, just as in choosing models for data distributions, simplifications are often made. This occurs notably when there are nested models in which a hypothesis $H_0 : \psi = \psi_0$ is being tested in the presence of an additional parameter β. For example, it is often assumed that

$$\pi(\beta|H_0) = \int \pi(\beta, \psi|H_1)d\psi. \tag{13}$$

If, in addition, it is assumed that β and ψ are independent *a priori* under H_1, then one needs only to choose one prior for β and another for ψ.

Sometimes, instead of (13), it is assumed that $\pi(\beta|H_0) = \pi(\beta|\psi = \psi_0, H_1)$. This prescription, however, depends on the choice of parameterization used: if an alternative parameterization (ξ, λ) is used under H_1, with H_0 then being specified by $\lambda = \lambda_0$, and if $\pi(\xi, \lambda|H_1)$ is obtained from $\pi(\beta, \psi|H_1)$ by the change-of-variables formula (introducing the appropriate Jacobian determinant) it may happen that $\pi(\xi|H_0)$ is *not* obtained by a change of variables from $\pi(\beta|H_0)$. (In a personal communication, J. Dickey has noted that L.J. Savage discussed this as an instance of the "Borel paradox" in a 1963 lecture; see also Dickey, 1985.) Variations on the marginal and conditional density methods are discussed by McCulloch and Rossi (1992) and Verdinelli and Wasserman (1993a).

Again as in data modeling, simplifications involving priors should be considered carefully since they may affect the results and yet not be justified. Here, testing is different from estimation. In frequentist theory, estimation and testing are complementary, but in the Bayesian approach the problems are completely different. In testing, Bayesians put positive prior probability on models that represent hypotheses, while in estimation continuous priors are used that would assign zero probability to special values such as $\psi = \psi_0$. In estimation,

priors are often picked for convenience, knowing that if the sample is fairly large the effect of the prior is small. In testing this is not so. As expression (5) shows, to order $O(n^{-1})$ the prior density may not be eliminated as the sample size increases, which contrasts with the case of estimation, where the MLE approximates the posterior mean to this order. As a result, the Bayes factor tends to be more sensitive to the choice of prior than is the posterior probability of an interval (e.g., Kass and Greenhouse, 1989; Kass, 1993).

In fact, it may happen that conclusions based on estimation seem to contradict those from a Bayes factor (as illustrated in Kass and Greenhouse, 1989). This occurs when an estimate, say of ψ, is found to be distant from a null-hypothetical value ψ_0. In this case, the data seem unlikely under H_0, but if the Bayes factor turns out to be *in favor of* H_0 then the data are *even more unlikely* under H_1 than they would have been under H_0. (This is one cause of discrepancies between conclusions reached with Bayes factors and those reached using P-values; see Section 8.2.) A consequence is that using a prior with a very large spread on ψ under H_1 in an effort to make it "noninformative" will force the Bayes factor to favor H_0. This was noted by Bartlett (1957) and is sometimes called "Bartlett's paradox". As Jeffreys recognized, to avoid this difficulty priors on parameters being tested (ψ in the discussion here) generally must be proper and not have too big a spread; thus, the standard improper priors that he used for estimation are not applicable to testing. We return to the use of improper priors on nuisance parameters in Section 5.3.

5.2 Sensitivity Analysis

Since Bayes factors can be sensitive to the prior, it is important to evaluate the Bayes factor over a range of possibilities. This involves specifying classes of priors to use under H_1 and H_2, and it also makes the issue of computation more urgent, since many multidimensional integrals (as in equation (2)) have to be calculated. We mention here several classes of priors that may be used. When there is enough information to yield initial priors with given hyperparameters (such as a $N(\nu, \phi^2)$ prior in which the hyperparameters are ν and ϕ), the hyperparameters may be perturbed (e.g., by halving and doubling ϕ or changing ν to $\nu \pm \phi$) and the Bayes factor recomputed, as in Applications 1, 4 and 5. See also McCulloch and Rossi (1991).

An important computational device is to use equation (5) as an approximation to (2) substituted in (1). From this, if we change from an original pair of priors under H_1 and H_2

to a new pair and thus compute a new Bayes factor $B_{12}^{(NEW)}$, we obtain

$$B_{12}^{(NEW)} \doteq B_{12} \cdot \left(\frac{\pi^{(NEW)}(\hat{\theta}_1|H_1)}{\pi^{(NEW)}(\hat{\theta}_2|H_2)} \right) \cdot \left(\frac{\pi(\hat{\theta}_2|H_2)}{\pi(\hat{\theta}_1|H_1)} \right), \tag{14}$$

with an error of order $O(n^{-1})$, which also holds if (5) is used to evaluate the original Bayes factor B_{12}. Since the ratio of prior ordinates at the MLE is easy to evaluate, (14) makes results for large numbers of priors easy to obtain; see Application 5. The accuracy of the method was explored in detail for the case of testing equality of two Binomial proportions by Kass and Vaidyanathan (1992).

The maximum of B_{10} (and thus the maximal evidence against H_0) over classes of priors was discussed by Edwards, Lindman, and Savage (1963) in contrasting Bayes factors with P-values, and was further developed by Berger and Delampady (1987), Raftery (1988b) and others (see Section 8.2). Suppose, first, that ψ is one-dimensional and that there is no additional parameter β. In this case, under H_0 we have $\psi = \psi_0$ and there is no prior. Under H_1, taking $\psi \sim N(\psi_0, \phi^2)$ leaves only the parameter ϕ^2 to be determined; sensitivity in the Bayes factor is then reflected in its sensitivity to the choice of ϕ. Edwards *et al.* (1963) computed the maximum of B_{10} over all choices of ϕ, which we will denote by $B_{10}^{(NORMAX)}$. There is no corresponding minimum: it is generally the case that $B_{10} \to 0$ as $\phi \to \infty$.

Edwards *et al.* (1963) also considered the maximum of B_{10} over all possible priors on ψ, which in simple cases occurs for the prior with all its mass at the MLE $\hat{\psi}$. This class is wider, but it can be too large and can yield bounds on B_{10} that are too big. As a result, Berger and Delampady (1987) examined the class of "symmetric unimodal" priors, which are symmetric about ψ_0 and nondecreasing in $|\psi - \psi_0|$, and obtained an expression for the maximum of B_{10}. They also considered other classes, which opens up the question of what class should be used. Note that $B_{10}^{(NORMAX)}$ is also the maximum of B_{10} over all priors that are scale mixtures of the $N(\psi_0, \phi^2)$ distribution. This seems a reasonably large and interesting class. Berger and Delampady (1987) showed that $B_{10}^{(NORMAX)}$ is not very different from the bound obtained using the symmetric unimodal class. A computationally simple approximation to $B_{10}^{(NORMAX)}$ may be obtained using Laplace's method (equation (5)) together with the argument leading to equation (15), below; the formula may be found in the Technical Report version of this paper.

When ψ may be a vector and there is also, under both H_0 and H_1, a nuisance parameter β (again possibly a vector), the problem is more complicated. Jeffreys (1961, pp. 249-250) showed that, under certain conditions, the prior on the nuisance parameter β is much less relevant than that on the parameter being tested ψ if the simplification (13) is used

together with *a priori* independence under H_1. Kass and Vaidyanathan (1992) extended and sharpened the result, which we now describe, by considering local alternatives (ψ near ψ_0).

In addition to the simplification of (13) together with *a priori* independence of β and ψ under H_1, assume that β and ψ are *null-orthogonal* in the sense that the Fisher information matrix $J(\beta, \psi)$ is block-diagonal when $\psi = \psi_0$. Kass and Vaidyanathan (1992) note that this condition often holds, at least approximately, and by a transformation of parameters it can always be made valid. They assume that $\hat{\psi} - \psi_0 = O(n^{-\frac{1}{2}})$ as would be the case if the "true" value of ψ were either ψ_0 or a neighboring alternative ψ_n such that $\psi_n - \psi_0 = O(n^{-\frac{1}{2}})$; when this situation does not hold, the Bayes factor will quickly become decisive and issues of asymptotic approximation will no longer be of concern. Now let B_{10} be the Bayes factor for one prior on β and B_{10}^* be the Bayes factor for a different prior. Then (under suitable regularity conditions),

$$B_{10} = B_{10}^* \cdot (1 + O(n^{-1})). \tag{15}$$

That is, up to order $O(n^{-1})$ the Bayes factor no longer depends on the choice of prior on β. Thus, when the parameters are null-orthogonal, sensitivity analysis may be confined to examination of priors on ψ. In practice if, under H_1, β and ψ are approximately uncorrelated (they are "observed-orthogonal", which may be easier to check) the relative insensitivity to choice of prior on β may be expected to hold (this was done in Application 5). Furthermore, the same argument (under the null, or for local alternatives) shows that (15) holds to order $O(n^{-1/2})$ even if the parameters are not null-orthogonal.

When there is little prior information, Raftery (1993c) argued that subjective priors will often be relatively flat in the region where the likelihood is large and that their impact on Bayes factors for the comparison of both nested and nonnested models should be small. Thus an alternative to careful and time-consuming prior elicitation in this case may be to use a set of priors constructed to have this property. Raftery (1993c) showed how this can be done for generalized linear models. These models relate a dependent variable y_i to independent variables $x_i = (x_{i1}, \ldots, x_{iJ})$ where $x_{i1} = 1$ in such a way that $E[y_i|x_i] = \mu_i$, $\mathrm{Var}\,[y_i|x_i] = \sigma^2 v(\mu_i)$ and $g(\mu_i) = x_i\beta$ where $\beta = (\beta_1, \ldots, \beta_J)^T$. The idea is to narrow down an initial class of baseline priors for β, here taken to be such that when $g(\mu) = \mu$, $v(\mu) = 1$ and the variables have been standardized to have mean zero and variance 1, the β_j are independent normal *a priori* with $\beta_1 \sim N(\nu, \eta^2)$ and $\beta_j \sim N(0, \phi^2)$ $(j = 2, \ldots, J)$. Bayes factors tend to be insensitive to ν and η (as indicated by the remarks following (15)) but they can be quite sensitive to the choice of ϕ.

One way of defining a reasonable range of priors is to require that the ratio of prior

ordinates at the MLE, given in equation (14), not be too far from 1 for any of the possible values of $\hat{\beta}$, namely $|\hat{\beta}_2| \leq 1$ when $J = 2$. We wish the same to be true when H_0 and H_1 are not nested. This corresponds to requiring that the prior not contribute much evidence in favor of either model, whether the models being compared are nested or not. These requirements are to some extent in conflict: for non-nested models it implies that ϕ be large, while for nested models it implies that ϕ not be too large. Balancing these two desiderata in a certain sense gives $\phi = e^{\frac{1}{2}} = 1.65$, and requiring that the priors not contribute evidence "worth more than a bare mention" beyond what is unavoidable leads to the range $1 \leq \phi \leq 5$. The resulting priors are then transformed back to the original scale for the variables; results for other choices of $g(\mu)$ and $v(\mu)$ are obtained by weighting the cases appropriately.

The result is what Raftery (1993c) calls a *reference set of proper priors* for generalized linear models. These are used in the GLIB software; see Section 9. While they are mildly data-dependent, they do have properties that one would associate with genuine subjective data-independent priors that represent a small amount of prior information. Similar reasoning can be applied to other classes of baseline priors and to other models.

5.3 Bayes Factors with Improper Priors

In Section 5.1 we indicated that improper priors on parameters of interest (ψ when we have $H_0 : \psi = \psi_0$) are problematic because, when used under H_1 and not under H_0 they force B_{10} to become zero. Jeffreys (1961), however, used improper priors on nuisance parameters appearing in both null and alternative models (e.g., in testing the value of a Normal mean $H_0 : \mu = \mu_0$, he took the prior on μ under H_1 to be proper but set $\pi(\sigma) = 1/\sigma$; p. 268). This leads to an improper predictive distribution specified by (2), but the value of (2) for the given data remains well-defined so this impropriety did not seem to bother Jeffreys, and others are untroubled by the procedure as well (Moulton, 1991; Robert and Caron, 1992; Robert, 1992). Equation (15) shows that in many cases the choice of prior on the nuisance parameter does not greatly affect the results.

Some authors have used improper priors for all parameters appearing in the models. This has the problem that flat priors are specified only up to an undefined multiplicative constant. Thus, the Bayes factor in this case also contains undefined constants. One effort to resolve this difficulty is the "imaginary training sample device" of Smith and Spiegelhalter (1981) and Spiegelhalter and Smith (1982). This consists of imagining that a data set is available which involves the smallest possible sample size permitting a comparison of H_0 and H_1 and provides maximum possible support for H_0, and then arguing that $B_{10} = (1 + \epsilon)^{-1}$, where

$\epsilon \geq 0$ is small; they take $\epsilon = 0$. This yields a value for the ratio of constants.

The authors of several published applications of the method found it useful (Racine *et al.*, 1986; Akman and Raftery 1986a; Raftery and Akman, 1986b; Raftery, 1987, 1988). The dimensionalities of the alternative models in these examples were not very different, however, and it is not clear how the method will perform in the more difficult case where this is not so.

Another solution is to set aside part of the data to use as a training sample which is combined with the improper prior distribution to produce a prior prior distribution. The Bayes factor is then computed from the remainder of the data. This idea was introduced by Lempers (1971), and other implementations have been suggested more recently under the names of partial Bayes factors (O'Hagan, 1991), intrinsic Bayes factors (Berger and Perrichi, 1993), and fractional Bayes factors (O'Hagan, 1994).

6 Accounting for Model Uncertainty

Practical model-building often involves far more than the comparison of two models; there are usually many other choices to be made. For example, in regression the analyst must choose the independent variables, decide which, if any, of the observations are outliers, and how, if at all, to transform the variables. Each possible combination of choices defines a different model, so that the model-building process consists of comparing many competing models. Strategies for doing this are commonly guided by a series of significance tests, often based on the approximate asymptotic distribution of a test statistic.

There are several problems with this. The sampling properties of the overall strategy, as distinct from those of the individual tests, are not well understood (Freedman, 1983; Miller, 1984, 1990). The models being compared are often not nested. Power considerations are usually not taken into account when setting significance levels; indeed, the power characteristics of the tests are often unknown. Any approach that selects a single model and then makes inference conditionally on that model ignores the uncertainty involved in model selection, which can be a big part of overall uncertainty. This leads to underestimation of the uncertainty about quantities of interest, sometimes to a dramatic extent. See Application 4 for an example of this.

All of these difficulties can be avoided, at least in principle, if one adopts a Bayesian approach and calculates the posterior probabilities of all the competing models, which follow directly from the Bayes factors (e.g., Leamer, 1978; Stewart, 1987). A composite inference

can then be made that takes account of model uncertainty in a simple and formally justifiable way. In Section 6.1 we review this approach, and in Sections 6.2–6.5 we discuss methods for handling the difficult situation where the number of models is very large.

6.1 Basic Ideas

When several models are being considered, Bayes factors yield their posterior probabilities, as follows. Suppose that $(K+1)$ models, $H_0, H_1, \ldots, H_K$ are being considered. Each of $H_1, \ldots, H_K$ is compared in turn with H_0, yielding Bayes factors $B_{10}, \ldots, B_{K0}$. Then the posterior probability of H_k is

$$\mathrm{pr}(H_k|D) = \alpha_k B_{k0} / \sum_{r=0}^{K} \alpha_r B_{r0}, \tag{16}$$

where $\alpha_k = \mathrm{pr}(H_k)/\mathrm{pr}(H_0)$ is the prior odds for H_k against H_0 $(k = 0, \ldots, K)$; here $B_{00} = \alpha_0 = 1$. Taking all the prior odds α_k equal to 1 is a natural choice, but other values of α_k may be used to reflect prior information about the relative plausibility of competing models.

The posterior model probabilities given by equation (16) lead directly to solutions of the prediction, decision-making and inference problems that take account of model uncertainty. For a quantity of interest Δ that is well-defined for every model, the posterior density given model H_k is $\mathrm{pr}(\Delta|D, H_k) = \int \mathrm{pr}(\Delta|D, \theta_k, H_k)\mathrm{pr}(\theta_k|D, H_k)\mathrm{d}\theta_k$. This can be used to make inferences about Δ conditionally on model H_k, but instead we may use the posterior density of Δ without conditioning, namely

$$\mathrm{pr}(\Delta|D) = \sum_{k=0}^{K} \mathrm{pr}(\Delta|D, H_k)\mathrm{pr}(H_k|D). \tag{17}$$

This accounts for the uncertainty about model form by weighting the conditional posterior densities according to the posterior probabilities of each model. The posterior mean and standard deviation of Δ are as follows (Raftery, 1993b):

$$E[\Delta|D] = \sum_{k=0}^{K} E[\Delta|D, H_k] \cdot \mathrm{pr}(H_k|D), \tag{18}$$

$$\mathrm{Var}[\Delta|D] = \sum_{k=0}^{K} \left(\mathrm{Var}[\Delta|D, H_k] + (E[\Delta|D, H_k])^2\right) \cdot \mathrm{pr}(H_k|D) - E[\Delta|D]^2. \tag{19}$$

Racine *et al.* (1986) showed how this method can be used to make inference about a treatment effect in the presence of uncertainty about the existence of a carryover effect.

The decision-making problem is solved by maximizing the posterior expected utility of each course of action considered. The latter is equal to a weighted average of the posterior expected utilities conditional on each of the models, with the weights equal to the posterior model probabilities $\mathrm{pr}(H_k|D)$. Smith (1991) discussed the situation where a model is to be chosen and then a decision made. Our view is that, if possible, a single model should not be selected before decision-making, and that model uncertainty should be accounted for in the calculation of posterior expected utilities.

Much of the literature on statistical analysis in the presence of a set of rival models has focussed on the selection of a single model. Equation (17) shows that selecting a single model and proceeding conditionally upon it may be reasonable if one of the $\mathrm{pr}(H_k|D)$ is close to unity, or if the sum is dominated by models for which the values of $\mathrm{pr}(\Delta|D, H_k)$ are similar. If not, analyses conditional on a single selected model fail to take account fully of uncertainty about structure, and so may well underestimate the uncertainty associated with their conclusions. This can lead, for example, to policy choices that are riskier than one thinks (Hodges, 1987).

6.2 Occam's Window

In spite of the importance of model uncertainty and the existence of a general strategy for dealing with it, at least since Leamer (1978), there have been three major obstacles to the widespread adoption of the method outlined in the previous section. The first is the difficulty of calculating Bayes factors, and we have shown in Section 4 that there is now a range of feasible computational strategies for doing this.

The second obstacle is that the number of terms in equation (17) can be enormous. For example, in regression with n cases and J candidate independent variables, considering all possible subsets, the possibility of outliers and four possible transformations of each variable gives an initial set of around $\left(2^J \times \binom{n}{O_{\max}} \times 4^{J+1}\right)$ models, where $O_{\max}$ is the maximum number of possible outliers envisaged. Even for a relatively small problem, with $n = 40$, $J = 12$ and $O_{\max} = 5$, this is on the order of 10^{16} models.

The third obstacle is that prior distributions for the parameters have to be specified for each model. Various possible ways around this now exist. One approach is to use the Schwarz criterion, relying on the result of Kass and Wasserman (1992b) that this gives an accurate approximation for a particular, reasonable prior. Another way is to specify prior distributions for one or several "big" models within which all or most of the models considered are nested,

and then to obtain the priors for the nested models by conditioning on the constaints that define them, as in Raftery (1993c).

In this section and the next we describe two general algorithmic approaches to solving this problem, which lead to feasible methods. The first, known as "Occam's Window", was proposed by Madigan and Raftery (1994) and selects a subset of the models initially considered. This involves averaging over a much smaller set of models than in (17), thereby facilitating effective communication of model uncertainty.

Those authors argue that if a model is far less likely *a posteriori* than the most likely model, then it has been discredited and should no longer be considered. Thus models not belonging to

$$\mathcal{A}' = \left\{ H_k : \frac{\max_l \{\mathrm{pr}(H_l|D)\}}{\mathrm{pr}(H_k|D)} \leq C \right\},$$

should be excluded from equation (17), where C is a fairly large number ($>> 1$) chosen by the data analyst, such as $C = 20$. Appealing to Occam's razor, they also exclude from (17) complex models which receive less support from the data than their simpler counterparts, namely those belonging to

$$\mathcal{B} = \left\{ H_k : \exists H_l \in \mathcal{A}, H_l \subset H_k, \frac{\mathrm{pr}(H_l|D)}{\mathrm{pr}(H_k|D)} > 1 \right\}.$$

Then equation (17) is replaced by

$$\mathrm{pr}(\Delta|D) = \frac{\sum_{H_k \in \mathcal{A}} \mathrm{pr}(\Delta|H_k, D)\mathrm{pr}(D|H_k)\mathrm{pr}(H_k)}{\sum_{H_k \in \mathcal{A}} \mathrm{pr}(D|H_k)\mathrm{pr}(H_k)},$$

where $\mathcal{A} = \mathcal{A}' \backslash \mathcal{B}$.

The search strategy to identify the models in $\mathcal{A}$ consists of a sequence of pairwise comparisons of nested models. If a model is rejected in favor of a larger one, then all the models nested within it are also rejected. Also, if there is evidence for H_0, then H_1 is rejected, but to reject H_0 we require strong evidence for the larger model, H_1. If the evidence is inconclusive (falling in "Occam's Window") neither model is rejected.

Typically the number of terms in (17) is reduced to 25 or less, and often to as few as one or two. This procedure mimics the evolutionary process of model selection which is typical of science. The final solution is fairly independent of the initial class considered, in the sense that most initial classes that contain $\mathcal{A}$ give the same result.

6.3 Markov Chain Monte Carlo Model Composition (MC^3)

Madigan and York (1992) proposed approximating (17) using Markov chain Monte Carlo methods, generating a process which moves through model space. They constructed an irreducible Markov chain $\{H(t)\}, t = 1, 2, \ldots$ with state space $\mathcal{H}$ and equilibrium distribution $\mathrm{pr}(H_i|D)$, where $\mathcal{H}$ is the space of models considered. Then for any function $u(H_i)$, if this Markov chain is simulated for $t = 1, \ldots, m$, the average, $\hat{U} = \frac{1}{m}\sum_{t=1}^{m} u(H(t))$, converges with probability one to $E[u(H)]$ as $m \to \infty$. To compute (17) in this way they set $u(H) = \mathrm{pr}(\Delta|H, D)$.

To construct the Markov chain, for each model H they defined a neighborhood $\mathrm{nbd}(H)$ consisting of H itself and the models which differ from H by just one parameter. They defined a transition matrix R by setting $R(H \to H') = 0$ for all $H' \notin \mathrm{nbd}(H)$ and $R(H \to H')$ constant for all $H' \in \mathrm{nbd}(H)$. H' is then drawn from $q(H \to H')$ and accepted with probability

$$\min\left\{1, \frac{\mathrm{pr}(H'|D)}{\mathrm{pr}(H|D)}\right\}.$$

Otherwise the chain stays in state H. Madigan and York (1992) reported that this process is highly mobile and that runs of 10,000 or less are typically adequate.

George and McCulloch (1993) proposed a similar method in which the chain moves through both model space and parameter space at once. To ensure that the chain is irreducible, they never actually eliminated a parameter from the model, but instead set it close to zero with high probability. If that probability is very high, the chain tends to move slowly, and if not the results can be hard to interpret.

6.4 Model Expansion

Draper (1994) proposed the model expansion method, in which one model is selected initally, and then this is generalized to a set of models that include the initially selected one as a special case, but that relax some of the structural assumptions underlying it. Equation (17) is then used for inference about quantities of interest, but restricted to the set of models obtained in the generalization step. In regression, for example, a set of variables and functional forms might be selected initially and a normal distribution assumed for the errors. This might be generalized by embedding the normal error distribution in the symmetric power-exponential family (Box and Tiao, 1962). Model expansion can be continuous, as in that example, or discrete.

Model expansion is useful for taking account of uncertainty about specific structural assumptions in a model, but is not designed to take account of the uncertainty inherent in model-building when many models are initially considered, as in variable selection in regression.

6.5 Evaluation of Methods

The efficacy of a modeling strategy can be judged by how well the resulting predictive distributions predict future observations (Dawid, 1984). Madigan and Raftery (1994) measured predictive performance by splitting complete data into two subsets, one of which (typically about 25–50% of the total) was used to calculate model probabilities with the other subset being used as a set of test cases. Predictive performance was then measured using the logarithmic scoring rule of Good (1952); see Section 3.2. This method can be used to assess the performance of any method that generates predictive distributions, Bayesian or not, model-based or not, statistical or not. With this scoring rule, equation (17) is guaranteed in a certain sense to give better predictions on average than those based on any individual selected model (Madigan and Raftery, 1994).

Model averaging, by both the Occam's window and MC^3 methods, have given consistently and substantially better predictions than those based on any one model alone, for several data sets. The differences between individual "good" models were smaller than the gain due to taking account of model uncertainty. The Markov chain Monte Carlo method had better predictive performance than Occam's window, but at the cost of greater computational expense and less easily interpretable results. This is true for discrete graphical models (Madigan and Raftery, 1994), linear regression models (Raftery, Madigan and Hoeting, 1993), and survival analysis (Raftery, Madigan and Volinsky, 1995).

7 Applications, Revisited

7.1 Application 1: *E. coli* Mutagenesis

The raw data for each strain of *E. coli* were two pairs of sample sizes and corresponding proportions $(n_{i1}, \hat{p}_{i1})$ and $(n_{i2}, \hat{p}_{i2})$. Here $i = 1, \ldots, 13$ with the 13th strain being the one in question, uvrE. The data were assumed to be distributed as Binomial proportions, and were transformed to the logit scale according to $Y_i = \log[\hat{p}_{i1}(1-\hat{p}_{i2})/(\hat{p}_{i2}(1-\hat{p}_{i1}))]$. Then Y_i was assumed to be Normally distributed, with σ_i^2 taken to be known and equal to the first-

order approximate variance based on Binomial sampling, namely $\sigma_i^2 = [n_{i1}\hat{p}_{i1}(1-\hat{p}_{i1})]^{-1} + [n_{i2}\hat{p}_{i2}(1-\hat{p}_{i2})]^{-1}$.

We write $\psi_i = \log(p_{i2}/(1-p_{i2})) - \log(p_{i1}/(1-p_{i1}))$ so that the null hypothesis is H_0 : $\psi_{13} = 0$ and under this hypothesis $Y_{13} \sim N(0, \sigma_{13}^2)$. For the alternative H_1, we begin with $Y_{13} \sim N(\psi_{13}, \sigma_{13}^2)$ and then, considering the data from the first twelve strains to be directly relevant, we use them to formulate a prior for ψ_{13} (taking the strains to be exchangeable). This may be done by assuming that $\psi_i \sim N(\mu, \tau^2)$, i.i.d., $i = 1, \ldots, 13$ so that $Y_{13} \sim N(\mu, \tau^2 + \sigma_{13}^2)$. The quantities μ and τ may then be estimated from the data on the first twelve strains. This was done by maximum likelihood as in Kass and Steffey (1989), where the transformed data may be found (the uvrE strain is number 12 in their Table 1). The Bayes factor is then

$$B_{10} = \frac{n(y_{13}; \hat{\mu}, \tau^2 + \sigma_{13}^2)}{n(y_{13}; 0, \sigma_{13}^2)},$$

where $n(x; m, v)$ denotes the normal density with mean m and variance v evaluated at x. The result was $B_{10} = 0.065$, indicating strong evidence in favor of H_0. □

7.2 Application 2: The Hot Hand.

Kass and a colleague K. Hsiao — hereafter KH — analyzed the data for Vinnie Johnson. Details may be found in Hsiao (1994). Under H_0, we have $Y_i \overset{\text{i.i.d.}}{\sim} B(n_i, p)$ $(i = 1, \ldots, 380)$, there being 380 games in the data set. Under the alternative, $Y_i \sim B(n_i, p_i)$ independently for $i = 1, \ldots, 380$, and $p_i \overset{\text{i.i.d.}}{\sim} Beta(a, b)$. KH began by reparameterizing the beta distribution according to $(a, b) = (\xi/\omega, (1-\xi)/\omega)$. Then, under the beta-binomial, $E(Y_i/n_i) = \xi$ and the binomial becomes a limiting case of the beta-binomial as $\omega \to 0$. Thinking of this as a nested model and using the notation of Section 5.1, ψ becomes ω and β becomes ξ which reduces to p when $\omega = 0$. The additional simplification mentioned there is then to take ξ and ω to be *a priori* independent, with the distribution on ξ being the same as that used on p under the binomial. KH put a Uniform(0,1) prior on ξ and p and then examined the Bayes factor for several values of ω.

To think about the values of ω that should be of interest, it is helpful to convert ω to the standard deviation s.d. $= \xi(1-\xi)/(1+\omega^{-1})$ which may be done by taking ξ equal to Johnson's overall shooting rate, $\hat{\xi} = .43$. Under H_1, Johnson's underlying game shooting ability varies from good games with high values of p_i to bad games with low values. To consider the variation sufficiently large to represent interesting swings in ability, it seems to us that the p_i's would have to vary frequently by about $\pm.05$. That is, values of s.d. of at

least .05 (roughly) would be needed to represent a meaningful amount of hot-handedness. With this in mind, here are several values of B_{10}:

ω	s.d.	B_{10}
.005	.035	.16
.01	.049	.017
.03	.085	3×10^{-7}

¿From these calculations the evidence does not rule out small degrees of game-to-game extra-Binomial variability, but it is decisively against substantial game-to-game extra-Binomial variability (as represented by the Beta-Binomial). In addition, KH used a Normal prior for which "the information in the prior equals the information in one observation" as described in Section 4.1.3 and obtained $B_{10} = 1/19$; the Schwarz criterion gave $B_{10} \approx 1/21$.

The uniform prior on ξ and p is a natural choice, but B_{10} may be sensitive to it. KH showed that this sensitivity is not enough to call the conclusions into question; indeed, the parameters are null orthogonal as discussed in Section 5.2.

Thus these data show decisively that Vinnie Johnson did not have a "hot hand", in the sense that that there was no substantial game-to-game extra-Binomial variability in his shooting percentage. (The data may be obtained by sending the e-mail message "send Vinnie.Johnson from data" to *statlib@stat.cmu.edu*.) □

7.3 Application 3: Ozone Exceedances

Success of the regulatory efforts of the Texas Air Control Board would be indicated by a decrease in the rate of occurrence of exceedances. If there were no trend, the data would be close to being from a homogeneous Poisson process; we denote this model by H_0. An alternative hypothesis is that the exceedance rate has been decreasing smoothly and gradually. This may be represented by the log-linear Poisson process, $H_1 : \lambda(s) = \rho\, e^{-\gamma s}$, where $\lambda(s)$ is the rate of occurrence at time s, and $\gamma > 0$. Another possibility is that the exceedance rate decreased fairly abruptly within a short time period. This may be represented by the change-point Poisson process, $H_2 : \lambda(s) = \lambda_1$ if $0 \leq s \leq \tau$ and $\lambda(s) = \lambda_2$ if $\tau < s \leq T$.

Raftery (1989) calculated Bayes factors for these hypotheses using improper reference priors and the imaginary training sample device of Spiegelhalter and Smith (1982). This is probably a reasonable approximation to vague prior information here because the numbers of degrees of freedom involved in the comparisons are small. The Bayes factor for a gradual change against no trend, B_{10}, is 0.02, indicating strong evidence (odds of 50) *against* a smooth

decrease; the Bayes factor corresponds to evidence *for* the null hypothesis. By contrast, the Bayes factor for an abrupt change against no trend, B_{20}, is 2.75, indicating evidence for an abrupt change that is positive but "worth no more than a bare mention". Note that *if* there was a change, it is much more likely to been abrupt than smooth by odds of $B_{21} = B_{20}/B_{10} = 135$. This is a non-nested model comparison, and one can see how easy it is with the Bayesian method.

The finding of some evidence for an abrupt decrease in the exceedance rate gave rise to the suggestion that this was due, not to a change in the underlying ozone levels, but to a change in measurement technology which led to lower variances and hence to fewer extreme values. Subsequent exploratory analyses of Smith (1989, rejoinder) suggested this to be plausible, especially since there were indeed changes in measurement instruments which could have led to such a change (Fairley, 1989). It would be possible to expand the model to take account of this, and then to test for it explicitly by calculating the Bayes factors for the expanded model against the previously considered ones. This example shows that Bayes factors are not restricted to tests of previously formulated hypotheses, but can also be used to guide an empirical model-building process.

In this example, Bayes factors have the advantage over frequentist approaches of being *technically simpler*. The comparisons between H_0 and H_1, and between H_0 and H_2 both lead to non-standard frequentist testing problems, even though their Bayesian solution is straightforward. Under H_1, the expected total number of events over all time is finite, so that the usual asymptotic arguments do not apply. Under H_2, the likelihood is highly discontinuous as a function of the change-point τ, so that the usual asymptotics based on smoothness of the likelihood do not apply at all. To see how much more cumbersome the frequentist approach to testing H_0 against H_2 is, compare the frequentist analyses of Worsley (1986) and Akman and Raftery (1986b) with the Bayesian analysis of Raftery and Akman (1986). And we know of no non-Bayesian testing procedures for comparing the non-nested and "irregular" models H_1 and H_2. In particular, the methods of Cox (1961, 1962) do not apply because the regularity conditions that they require do not hold.

7.4 Application 4: Educational Transitions

We address the question of interest in this application using part of the longtitudinal data collected by Greaney and Kelleghan (1984). In 1967, they selected a random sample of 11-year-old children in Irish elementary schools and followed them through the rest of their educational careers. 441 students entered second-level education and, of these, 230 completed

it by taking the Leaving Certificate Examination, while 211 did not. Social class background and ability were both measured prospectively at age 11 by continuous-valued variables, while type of school was a dichotomous variable, indicating whether a secondary or vocational school was attended.

We use logistic regression with the reference set of proper priors described in Section 5.2. It is clear that social class background does have an effect, with a Bayes factor of over 2000 for the model that includes it and ability against the model that includes only ability. Thus Greaney and Kelleghan's conclusion about meritocracy was unwarranted. Also, gender has no effect. A frequentist analysis would say merely that the gender effect is "not significant", but the Bayesian analysis provides strong evidence (odds of more than 20) *for* the simpler model from which the gender effect is absent; as in Applications 1 and 2 the Bayes factor provides evidence for a null hypothesis.

The main uncertainty is about the link function: the data provide some evidence (a Bayes factor of 3.7) for the complementary log-log (hereafter cloglog) link against the logit link, but this is not strong. This Bayes factor varied by less than 0.02 over the entire range of priors considered. The models corresponding to different links are not nested, so that a formal non-Bayesian comparison between these models would be cumbersome while the Bayesian comparison is straightforward.

A question of policy interest is the difference between the attainments of secondary and vocational school students, after controlling for the other variables. Here we consider the ratio of the odds on completion of second-level education for secondary school students to that for vocational school students when both social class and ability are high (71 and 127 respectively, corresponding to the 95th percentile in each case). We will refer to this quantity simply as the odds ratio. Even though the data do not distinguish clearly between the two link functions, they imply very different things about the odds ratio. With the logit link function, the maximum likelihood estimator of the odds ratio is 23, while with the cloglog link it is 91, about four times larger.

Approximate posterior distributions of the odds ratio under the two models are shown in Figure 1, as is the combined inference using the Bayesian mixture of equation (17); the corresponding quantiles are shown in Table 1. The posterior distributions under the two models are very different; that for the logit link is more concentrated and centered about a lower value. The combined posterior distribution has both the peak around 20 from the logit link, and the long tail from the cloglog link. The combined 95% interval in Table 1 is close to the union of the intervals corresponding to the individual models, but is somewhat

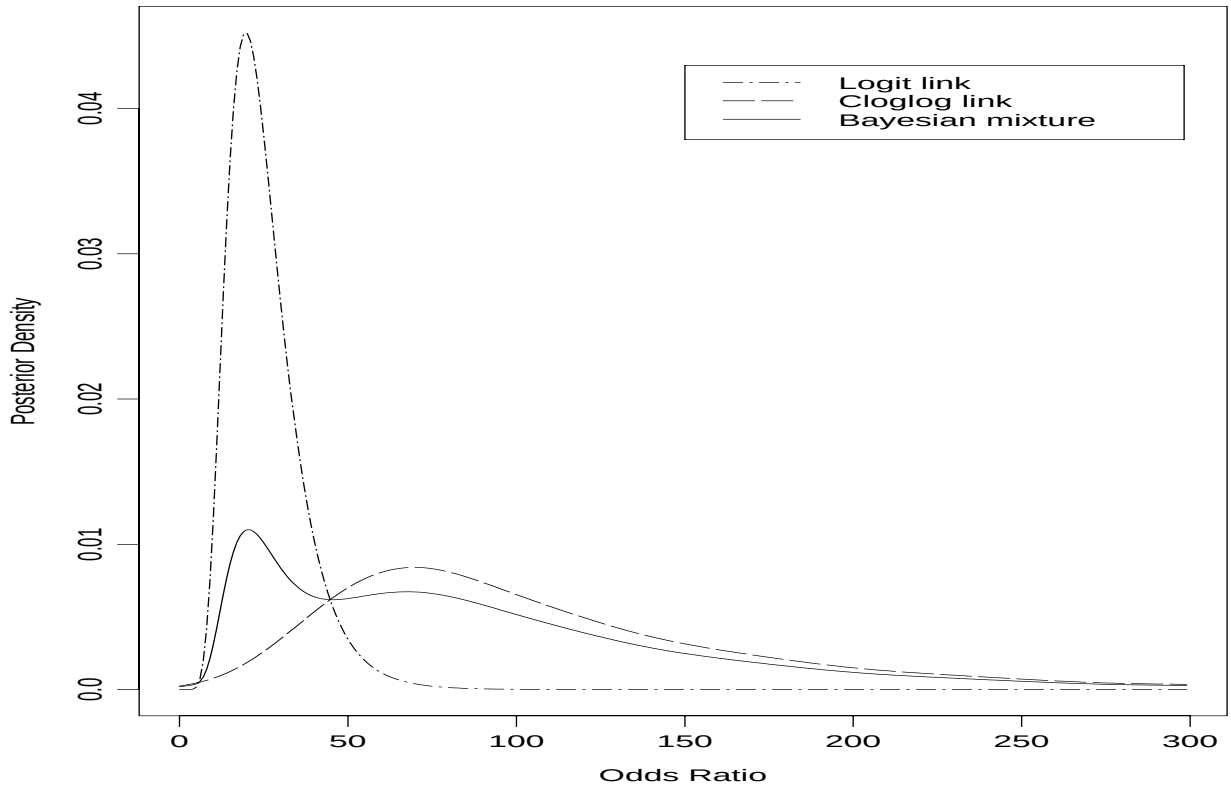

Figure 1: Posterior distributions of the odds ratio in the educational transitions application. The results are indistinguishable for all the prior distributions (i.e. values of ϕ) considered.

shorter; this is intuitively reasonable.

For other analyses of these and related data see Raftery and Hout (1985, 1993). The data may be obtained by sending the e-mail message "send irish.ed from data" to *statlib@stat.cmu.edu*.

7.5 Application 5: Predicting Working Memory Failure

The analysis, including prior elicitation, was reported in detail by Carlin *et al.* (1992). Here we discuss several important issues: the computational burden and the effectiveness of Laplace's method, the difficulty and effectiveness of eliciting the prior and assessing how sensitive the results are to it, the use of influence diagnostics, and the role of the Schwarz criterion.

Computing (2) involved evaluating a one-dimensional integral nested within a five-dimensional integral for the number-of-conditions model and a six-dimensional integral for the query-

Table 1: Posterior Quantiles of the Odds Ratio in the Educational Transitions Application.

Model	Quantile		
	.025	.5	.975
logit link (H_3)	10	23	52
cloglog link (H_4)	24	94	338
Bayesian mixture	13	77	314

complexity model. Laplace's method was programmed in the S language and the result was $\hat{B}_{21} = 8.0$ in favor of the query complexity model. This was checked with subregion-adaptive integration, which yielded $B_{21} = 10.8$. Thus, Laplace's method was somewhat inaccurate, but clearly accurate enough for the inferential purpose at hand. (Subregion-adaptive integration was more difficult and was done by Alan Genz.)

Simple forms were used for the priors, but there were still 15 hyperparameters to be determined. Some of the information for this came from another related experiment, but some was based on the experience of the investigators and was really rough guesswork. Sensitivity analysis was carried out by shifting all prior means by one prior standard deviation unit in each direction, and then doubling and halving each prior standard deviation, which resulted in approximately 10^7 alternative prior distributions. Even for this large number of priors, computations were performed quite easily using the Laplace approximation method of equation (5). After eliminating certain of the alternative priors as unreasonable, the minimal value obtained over all remaining priors was was $\hat{B}_{21} = 3.0$.

A concern was the influence that individual subjects might have on the results (via the Normal distributional assumption of subject ability effects). Leave-one-out influence diagnostics were computed using the Laplace approximation methods, which are again quite easy to obtain in S. A small number of subjects in the experiment did carry much of the comparative information. Taking these sensitivity analyses into account, the overall conclusion was that there was "some evidence, though not strong evidence" in favor of the query complexity model. The Schwarz criterion was $S = 3.6$ giving $B_{21} \approx 6$, which was consistent with the overall conclusion and thus could be used as a summary. □

8 Issues and Controversies

8.1 Why Test Sharp Hypotheses?

We introduced Bayes factors as a way of assessing the evidence in favor of a scientific theory. Our statement that Jeffreys's approach computes "the posterior probability that one of the theories is correct" invites argument, however. Some would say that theories are never correct, and thus that any approach that assumes they are must be flawed. Those who make this argument generally prefer the use of interval estimates.

Certainly we would agree with Jeffreys (1961, pp. 389-390) and Edwards, Lindman, and Savage (1963, p. 235) when they say that often hypothesis testing is not applicable and estimation is more appropriate. We also recognize a legitimate worry on the part of many statisticians that empirical models are often taken too seriously, and that poor models are sometimes accepted merely because they fit the data better than other models that are even worse.

Some authors, however, see the introduction of sharp hypotheses as "silly" (e.g., Gelfand, Dey, and Chung, 1992). This point of view ignores the way scientific investigations usually proceed: though one rarely believes a scientific law in an absolute sense, it is a great convenience to speak and to act as if laws are valid. When one says that a certain theory is correct, one means that deviations from it are sufficiently minor to be irrelevant for all practical purposes at hand. Thus, theories do become "accepted" for a period of time during which they are used to make predictions about new phenomena. Jeffreys (1961, p. 391) pointed out that the best available law gets used for future calculations even when discrepancies are found, noting that "there has not been a single date in the history of the law of gravitation when a modern significance test would not have rejected all laws and left us with no law."

The admission that minor discrepancies may exist between theory and data does not imply that estimation is more natural than testing. For there are many ways in which a theory might err, and in representing the theory by a statistical model one should not presume that possible errors are eliminated simply by letting some parameter ψ take values other than a particular ψ_0. Simple laws are preferred partly because one cannot be sure that by including more parameters to model one kind of error, the law will succeed better in some unanticipated new situation. In some cases it may be worthwhile to include additional parameters, but this is an empirical question. In fact, it is exactly this empirical question that Bayes factors are supposed to answer. They do so by comparing predictions made by the simpler and the more complicated theories.

For example, in Application 1, an alternative to testing equality of the Binomial proportions would be to assume that they are different and to ask how different they are. But how would this correspond to the scientific theory? One would have to say something like "we estimate the difference between the log odds to be less than .38 (with probability .95), which indicates that deviations from the error-prone DNA repair mechanism are quite small". But this is merely an elaboration of the more succinct "evidence in favor of the error-prone DNA repair mechanism." Careful examination of the data using either approach ought to lead an analyst to appropriate conclusions. We find that expressing results in terms of Bayes factors is simpler. It is also reassuring that, as Berger and Delampady (1987) noted, in testing a normal mean the point null is a good approximation to an interval null as long as the width of the interval is less than about one half of the standard error of the sample mean (see also Dickey, 1976, and Verdinelli and Wasserman, 1993a). For related discussion along these lines see Jeffreys, (1961, p. 367), Zellner (1987), and Raftery (1992).

It has been argued that a comprehensive account of Bayesian model selection requires decision theory (Kadane and Dickey, 1980; Smith, 1991). The approach discussed here avoids the introduction of utilities, which would bring with it another layer of sensitivity concerns.

8.2 Bayes Factors Versus Non-Bayesian Significance Testing

There is a substantial literature on the controversy between Bayesian and non-Bayesian testing procedures. This is not a central theme of our paper, but we do wish to briefly mention several points that have been made in the literature.

1. There is no reason to expect a P-value to be similar to the posterior probability that the null hypothesis is correct. However, partly because this misinterpretation of P-values is common among non-statisticians, it is of some interest to compare results. This was done by Jeffreys (1961, pp. 434-435) and many subsequent workers. There is a general feeling that Bayes factors are more conservative than P-values, mainly because when comparisons are made it becomes clear that a P-value of .05 cannot represent much evidence against the null (Edwards, Lindman, and Savage, 1963; Berger and Mortera, 1991, and the references cited there).

2. Frequentist tests tend to reject null hypotheses almost systematically in very large samples, whereas Bayes factors do not. This has been a real problem in sociology where data sets frequently have thousands of cases. A dramatic example with $n = 110,000$ was discussed by Raftery (1986b). There a substantively meaningful model which

explained 99.7% of the deviance was rejected by a standard chi-squared test with a P-value of about 10^{-120}, but was nevertheless favored by the Bayes factor. Faced with this problem, sociologists had taken to ignoring significance tests and using other criteria of reasonableness and common sense in comparing models (e.g. Fienberg and Mason, 1979; Grusky and Hauser, 1984). Bayes factors are now widely used in sociology, usually with BIC as an approximation.

3. Bayes factors, like Bayesian procedures generally, follow the likelihood principle (e.g., Berger and Wolpert, 1984). As a result, in settings such as clinical trials where cases may accrue sequentially, Bayes factors may be applied without concerns about unscheduled analysis of the data. See, e.g., Cornfield (1966a,b) and Berger and Berry (1988).

4. Bayes factors can be applied as easily to non-nested models as to nested ones. By contrast, the application of non-Bayesian significance tests to non-nested models is difficult. The approach of Cox (1961, 1962), which has spawned a large literature, tends to be cumbersome to implement and requires the often arbitrary designation of one of the two non-nested models as the null hypothesis. One way around this arbitrariness is to carry out two tests rather than one, with each model in turn as the null hypothesis. However, there is no guarantee of getting the standard kind of result of a test, namely rejection of one model and non-rejection of the other. Both models may fail to be rejected, in which case it is not clear how to make inferences about quantities of interest, especially if the two models lead to different conclusions. Both models may be rejected (as often happens with large samples), in which case the tests do not provide a comparison between the two models.

5. Non-Bayesian significance tests were developed for the comparison of two models, but practical data analysis often involves far more than two models, at least implicitly. In this case, carrying out multiple frequentist tests to guide a search for the best model can give very misleading results (e.g. Freedman, 1983). By allowing us to take account of model uncertainty, Bayes factors can avoid this problem (e.g. Raftery, Madigan and Hoeting, 1993).

With regard to Bayesian calibration of frequentist methods, for large samples the Schwarz criterion may be used to obtain the required value of an approximate t statistic in order for

Table 2: Approximate Minimum t-values for Different Grades of Evidence and Sample Sizes

Evidence for H_1	$2 \log B_{10}$ (Approximation (21))	$n =$			
		30	100	1000	10000
Positive	$0 - 5$	1.84	2.15	2.63	3.03
Strong	$5 - 10$	2.90	3.10	3.45	3.77
Decisive	> 10	3.66	3.82	4.11	4.38

it to represent strong or decisive evidence. Equation (9) implies that

$$2 \log B_{10} \approx \Lambda - (d_1 - d_0) \log(n). \tag{20}$$

Now if H_0 and H_1 are nested and differ by just one degree of freedom, so that $(d_1 - d_0) = 1$, then, approximately in large samples, $\Lambda \approx t^2$, where t is the t-statistic for (frequentist) testing of H_0 against H_1, obtained for example by dividing $(\hat{\psi} - \psi_0)$ by its large sample approximate standard error. Thus, in that case equation (20) implies that

$$2 \log B_{10} \approx t^2 - \log(n). \tag{21}$$

The approximate t-values corresponding to different grades of evidence (on the cruder scale in Section 3.2), and different sample sizes are shown in Table 2 (as in Raftery, 1993b; see also Berger, 1985). For "positive" evidence this is $t = \sqrt{\log n}$, for "strong" evidence it is $t = \sqrt{\log n + 5}$, while for "decisive" evidence it is $t = \sqrt{\log n + 10}$. Note that the critical values quickly become larger than the usual cutoffs based on P-values as n becomes large.

Atkinson (1978) has noted some instances of Bayes factors favoring the simpler model H_0 even when a more complex model H_1 is correct. However, Smith and Spiegelhalter (1980, p.216) showed that this occurs only when the two models are so close that there is nothing to be lost for predictive purposes by cutting back to the simpler model, so that the Bayes factor functions as a *fully automatic Occam's razor*.

I.J. Good has for many years advocated a Bayes/non-Bayes compromise, consisting essentially of calculating the Bayes factor and using it as a frequentist test statistic. In Good (1992) he reviewed these ideas and listed nearly 50 of his own publications on the topic, spanning 40 years. This is a frequentist testing proposal and there seems to be no reason why it should escape the difficulties that other frequentist tests have: these arise from the way the test statistic is calibrated and not from the choice of test statistic. However, Good has also pointed out that his compromise can be inverted, with P-values transformed to yield approximate Bayes factors; this is a sensible and useful point.

Goodness-of-fit tests are different from other frequentist tests because they do not aim to compare two competing models, but rather to detect departures from a null hypothesis even when no alternative has yet been formulated. One might question the use of "rejecting" a hypothesis if there is nothing to put in its place. Many Bayesians do, however, see a useful role for goodness-of-fit tests (Dempster, 1971; Box, 1980; Rubin, 1984). They can be useful for calibrating the diagnostic checks to which a model is subjected, and thus guiding the search for a better model. In Jeffreys' view, a model should not be abandoned until a better one (in the posterior model probability sense) is found.

8.3 Bayes Factors Versus the AIC

Akaike (1973) advocated that, given a class of competing models for a data set, one choose the model that minimizes

$$\text{AIC} = -2(\text{log maximized likelihood}) + 2(\text{number of parameters}). \tag{22}$$

Two main justifications for the AIC have been advanced. The first, due to Akaike (1973), is based on a predictive argument. Suppose that, given current data and a set of possible models, we want the predictive distribution of a future datum. Then, *if the predictive distibution is conditional on a single model and on its estimated parameters,* the AIC picks the model which gives the best approximation, asymptotically, in the Kullback-Leibler sense. However, such a predictive distribution is incorrect, because it does not incorporate the uncertainty about parameter values and model form (Aitchison and Dunsmore, 1975). Shibata (1976) and Katz (1981) have shown that the AIC tends to overestimate the number of parameters needed, even asymptotically. Thus, if one must ignore both parameter uncertainty and model uncertainty when making predictions, it may be worthwhile to have a model that is too big (Shibata, 1976; Stone, 1979). Related remarks appear in Zellner (1978) and Stone (1979).

A related argument is that the AIC picks the correct model asymptotically if the complexity of the true model grows with sample size (Shibata, 1980, 1981). Typically this is taken to mean that the model grows in one respect (e.g. the order of an autoregressive model) but remains fixed in all other respects (e.g. normality, linearity). Our experience with large data sets in sociology is that the AIC selects models that are too big even when the sample size is large, including effects that are counter-intuitive or not borne out by subsequent research.

The second main justification for the AIC, perhaps best described by Akaike (1983), is Bayesian. He wrote that model comparisons based on the AIC are asymptotically equivalent

to those based on Bayes factors. This is true, however, only if the precision of the prior is comparable to that of the likelihood, but not in the more usual situation where prior information is small relative to the information provided by the data. In the latter more usual situation, the Schwarz criterion indicates that the model with the highest posterior probability is the one that minimizes

$$\text{BIC} = -2(\log \text{maximized likelihood}) + (\log N)(\text{number of parameters}). \qquad (23)$$

Comparing equations (22) and (23) indicates that BIC tends to favor simpler models than those chosen by the AIC criterion.

Findley (1991) gave some cases in which BIC does not yield consistent model selection but AIC does. However, these are cases in which standard asymptotics do not apply and thus the theory in Section 4.1.3 leading to the approximation (9) also does not apply.

Linhart and Zucchini (1986) generalized Akaike's (1973) approach, replacing the Kullback-Leibler distance between true and estimated predictive distributions by any arbitrary distance, and replacing the quantity to be predicted (the next data point in Akaike's development) by any quantity of interest. Their work shows, for example, that the "2" by which the number of parameters is multiplied in equation (22) is arbitrary in that it depends crucially on the choices of distance and quantity of interest, and that other choices can lead to quite different multipliers, such as "4". However, their approach is open to the same general criticisms as the AIC. In particular, it provides no way of taking account of model uncertainty and so is somewhat at a loss when several models score almost equally well.

9 Bibliographical Remarks and Additional Work

Jeffreys (1961), Good (1952, 1983, 1985), Mosteller and Wallace (1964), and Zellner (1971) are important basic sources for applications of Bayes factors. The literature on model selection also contains many papers on Bayes factors, particularly in econometrics, e.g., Leamer (1983), Poirier (1985), Rossi (1985, 1988), McCulloch and Rossi (1991), Schotman and van Dijk (1991).

Several Bayesian methods of selecting models without Bayes factors have been proposed, often using some kind of sample-splitting as in cross-validation (Geisser and Eddy, 1979; Fornell and Rust, 1989; Gelfand, Dey, and Chang, 1992).

Aitkin (1991) describes so-called "posterior Bayes factors" in which the marginal likelihood $\text{pr}(D|H_k)$ defined by equation (2), which is the prior mean of the likelihood, is replaced

in equation (1) by the posterior mean of the likelihood, namely $\int \mathrm{pr}(D|\theta_k, H_k)\mathrm{pr}(\theta_k|D, H_k)\mathrm{d}\theta_k$. In spite of their name, these are quite different in practice from Bayes factors as defined in this article. Several discussants of Aitkin (1991) pointed out that the procedure has little Bayesian justification, does not have any known frequentist optimality properties, and yields counter-intuitive results (e.g. discussions by Dawid, Fearn, Goldstein, Lindley, and Whittaker).

We now give a noncomprehensive list of references in which explicit expressions for Bayes factors in different models appear.

Multinomials: Jeffreys (1961); Good (1967); Dickey and Lientz (1970); Gunel and Dickey (1974); Good and Crook (1987); Albert (1990).

Linear models: Zellner (1971); Dickey (1971); Dickey (1975); Zellner and Siow (1980); Dickey (1980); Smith and Spiegelhalter (1981); Zellner (1984); Broemeling (1985); Draper and Guttman (1987); Mitchell and Beauchamp (1988); Raftery, Madigan and Hoeting (1993).

Outliers: Pena and Guttman (1992); Pettit and Young (1990).

Logistic regression and Log-linear models: Raftery (1986a, 1988b, 1993c); Stewart (1987); Madigan and Raftery (1994); McCulloch and Rossi (1991); Madigan and York (1992); Kass and Vaidyanathan (1992). Also the GLIB software (see below).

Survival analysis: Raftery, Madigan and Volinsky (1995).

Multivariate analysis: Dickey (1971); Dickey (1975); Dayal and Dickey (1976); Smith and Spiegalhalter (1981) [multivariate normal models]; Cooper and Herskovits (1992), Cowell, Dawid and Spiegelhalter (1993), Madigan and Raftery (1994), Madigan *et al.* (1994) [discrete graphical models]; Banfield and Raftery (1993) [cluster analysis]; Raftery (1993b) [structural equation models (LISREL)].

Stochastic processes: Dickey and Lientz (1970), Katz (1981) [Markov chains]; Broemeling (1985) [autoregressive models]; Katz (1981) [Markov chains]; Akman and Raftery (1986a) [point processes]; Raftery and Akman (1986) [change-point Poisson processes]; Raftery (1987, 1988) [software reliability]; Le, Raftery and Martin (1990) [autoregressive models with outliers]. Polson and Roberts (1994) [diffusion processes].

Deterministic models: Raftery, Givens and Zeh (1994).

Other models: Berry, Evett and Pinchin (1992) [assessment of forensic evidence].

Software: There is little general purpose software for Bayes factors, although the BIC approximation can often be easily calculated from the output of standard statistical software. GLIB (Generalized LInear Bayesian modeling) is an S-PLUS function that returns accurate

Bayes factors and posterior probabilities for generalized linear models, as well as inference about the parameters that takes account of model uncertainty. It can be obtained by sending the e-mail message "send glib from S" to *statlib@stat.cmu.edu*.

10 Conclusion

We have reviewed applications and developments of a method introduced by Jeffreys over fifty years ago. The appeal of the concept was that it provided a simple and satisfying description of the process of accepting new scientific laws as replacements for older ones, which Jeffreys illustrated through the use of many examples in his book *Theory of Probability*.

We have tried to show that Jeffreys' methodology has worn well with time and that the Bayesian approach to hypothesis testing has evolved to fill a niche in modern computationally-intensive statistical practice. It applies to a limited but important class of problems in scientific inference, and also to the assessment of uncertainty when many models are considered initially. The points we wish to emphasize are:

- From Jeffreys' Bayesian point of view, the purpose of hypothesis testing is to evaluate the evidence in favor of a scientific theory.
- Bayes factors offer a way of evaluating evidence *in favor of* a null hypothesis.
- Bayes factors provide a way of including other information when assessing the evidence for a hypothesis.
- Bayes factors are very general. In particular, they do not require alternative models to be nested.
- Several techniques are available for computing Bayes factors, including asymptotic approximations which are easy to compute using the output from standard packages that maximize likelihoods.
- In "non-standard" statistical models that do not satisfy common regularity conditions, it can be technically simpler to calculate Bayes factors than to derive non-Bayesian significance tests.
- The Schwarz criterion (or BIC) gives a crude approximation to the logarithm of the Bayes factor, which is easy to use and does not require evaluation of prior distributions. It is well suited for summarizing results in scientific communication.

- When several models are considered initially, Bayes factors can be used to calculate posterior model probabilities, yielding composite estimates or predictions that take account of model uncertainty.
- Algorithms have been proposed that allow model uncertainty to be taken into account when the class of models initially considered is very large.
- Bayes factors are useful for guiding an evolutionary model-building process.
- It is important, and feasible, to assess the sensitivity of conclusions to the prior distributions used. In our applications we have found our conclusions to be robust to the prior in a qualitative sense, but this is not guaranteed to be the case.

Bayes factors have many of the strengths and limitations of the Bayesian approach generally. The essential strength is their solid logical foundation which offers great flexibility. Some of the practical advantages of this are noted above; another is invariance with respect to stopping rules in clinical trials, mentioned in Section 8.2. Recently, advances in computing and the development of methods that take advantage of additional computational power have greatly extended the usefulness of Bayesian methods. Bayes factors can now be computed for a wide variety of models.

The chief limitations of Bayes factors are their sensitivity to the assumptions in the parametric model and the choice of priors. We have discussed ways of doing sensitivity analysis in Section 5, which are illustrated in Applications 4 and 5. Application 4 illustrates the usefulness of a reference set of proper priors for sensitivity analysis, which has been implemented for generalized linear models in the GLIB software; this idea needs to be extended to other classes of models.

The more situation-specific sensitivity analysis in Application 5 is rather cumbersome. It requires close attention to the details of model and prior. It may be argued that this is appropriate: we should not be cavalier in making inferences that depend on our assumptions. The Schwarz criterion (or BIC) may be used for reporting scientific results with other analysis omitted but serving as background support. The question of how much effort should be made before conclusions are drawn arises in any data analysis problem. This is part of the art of applied statistics, but there is room for research that would help the applied statistician decide whether or not to proceed with the full Bayesian analysis.

References

Aitchison, J. and Dunsmore, I.R. (1975). *Statistical Prediction Analysis.* Cambridge, U.K.: Cambridge University Press.

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

Akaike (1973), Information Theory and an Extension of the Maximum Likelihood Principle, in *Second International Symposium on Information Theory,* eds. B.N. Petrox and F. Caski, Budapest: Akademiai Kiado, pp.267.

Akaike (1983), Information Measures and Model Selection, *Bulletin of the International Statistical Institute,* 50, 277-290.

Akman, V.E., and Raftery, A.E. (1986a). Bayes factors for non-homogeneous Poisson processes with vague prior information. *J. R. Statist. Soc. B,* 48, 322-329.

Akman, V.E. and Raftery, A.E. (1986b) Asymptotic analysis of a change-point Poisson process. *Ann. Statist.*, 14, 1583–1590.

Albert, J.H. (1990). A Bayesian test for a two-way contingency table using independence priors. *Canadian J. Statist.*, 4, 347–363.

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

Banfield, J.D. and Raftery, A.E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49, 803–821.

Bartlett, M.S. (1957). Comment on D.V. Lindley's statistical paradox, *Biometrika*, 44, 533-534.

Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis.* New York: Springer-Verlag.

Berger, J.O. and Berry, D.A. (1988). The relevance of stopping rules in statistical inference. In *Statistical Decision Theory and Related Topics IV*, (eds. S.S. Gupta and J.O. Berger), Vol. 1, 29-47.

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

Berger, J.O. and Mortera, J. (1991). Interpreting the stars in precise hypothesis testing. *International Statistical Review*, 59, 337-353.

Berger, J.O and Perrichi, L.R. (1993). The intrinsic Bayes factor for model selection and prediction. Technical Report, Department of Statistics, Purdue University.

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

Berger, J.O. and Wolpert, R.L. (1984). *The Likelihood Principle.* IMS Monograph, volume 6,

Hayward, CA: Institute of Mathematical Statistics.

Berry, D.A., Evett, I.W. and Pinchin, R. (1992). Statistical inference in crime investigations using Deoxyribonucleic acid profiling (with Discussion). *Applied Statistics*, 41, 499–531.

Box, G.E.P. (1980). Sampling and Bayes' inference in scientific modeling (with Discussion). *Journal of the Royal Statistical Society, series A*, 143, 383–430.

Box, G.E.P. and Tiao, G.C. (1962). A further look at robustness via Bayes's theorem. *Biometrika*, **49**, 419–432.

Broemeling, L.L. (1985) *Bayesian Inference for Linear Models.* New York: Marcel Dekker.

Carlin, B.P. and Chib, S. (1993). Bayesian model choice via Markov chain Monte Carlo. Research Report 93-006, Division of Biostatistics, University of Minnesota.

Carlin, B.P., Kass, R.E., Lerch, J. and Huguenard, B. (1992) Predicting working memory failure: A subjective Bayesian approach to model selection. *Journal of the American Statistical Association*, 87, 319-327.

Carlin, B.P. and Polson, N.G. (1991). Inference for nonconjugate Bayesian models using the Gibbs sampler. *Canadian J. Statist.*, 19, 399-405.

Chen, M-H. (1992). Importance weighted marginal Bayesian posterior density estimation. Technical Report, Department of Statistics, Purdue University.

Chow, G.C. (1981), A Comparison of the Information and Posterior Probability Criteria for Model Selection, *Journal of Econometrics,* 16, 21-33.

Cooper, G. and Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9, 309–347.

Cornfield, J. (1966a). A Bayesian test of some classical hypotheses – with applications to sequential clinical trials, *J. Am. Stat. Assoc.*, 61, 577-594.

Cornfield, J. (1966b). Sequential trials, sequential analysis, and the likelihood principle. *The American Statistician*, 20: 18-23.

Cowell, R.G., Dawid, A.P. and Spiegelhalter, D.J. (1993). Sequential model criticism in probabilistic expert systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15, 209–219.

Cox, D.R. (1961). Tests of separate families of hypotheses. *Proc. 4th Berkeley Symposium*, 1, 105–123.

Cox, D.R. (1962). Further results on tests of separate families of hypotheses. *Journal of the Royal Statistical Society, series B*, 24, 406–424.

Dawid, A.P. (1984). Present position and potential developments: some personal views. Statistical theory. The prequential approach (with Discussion). *J. R. Statist. Soc.* A, 147,

178–292.

Dawid, A.P. (1992). Prequential analysis, stochastic complexity and bayesian inference. In *Bayesian Statistics 4*, (Bernardo, J.M. et al.) Oxford Science Publications, pp. 109-125.

Dayal, H.H. and Dickey, J.M. (1976). Bayes factors for Behrens-Fisher problems. *Sankhya*, 38, 315-328.

De Bruijn, N.G. (1970) *Asymptotic Methods in Analysis.* Amsterdam: North-Holland.

DeGroot, M.H. (1970). *Optimal Statistical Decisions.* New York: McGraw-Hill.

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

Dempster, A.P. (1971). Model searching and estimation in the logic of inference. In *Foundations of Statistical Inference* (eds. V.P. Godambe and D.A. Sprott), Toronto: Holt, Rinehart and Winston of Canada.

Dickey, J.M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *Annals of Mathematical Statistics*, 42, 204–223.

Dickey, J.M. (1975). Bayesian alternatives to the F-test and least squares estimate in the normal linear model. In *Studies in Bayesian Econometrics and Statistics*, (eds. S.E. Fieberg and A. Zellner), Amsterdam: North-Holland Publishing Co., 515-554.

Dickey, J.M. (1976). Approximate posterior distributions. *J. Amer. Statist. Assoc.*, 71, 680-689.

Dickey, J.M. (1980). Approximate coherence for regression model inference. Chapter 22 in *Bayesian Analysis in Econometrics and Statistics* (A. Zellner, ed.), Amsterdam: North-Holland, pp. 334–354.

Dickey, J.M. and Kadane, J.B. (1980). Bayesian Decision Theory and the Simplification of Models. In *Evaluation of Econometric Models*, (J. Kmenta and J. Ramsey, eds.), Academic Press, (1980), 245-268.

Dickey, J.M. and Lientz, B.P. (1970). The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain. *Ann. Math. Statist.*, 41, 214–226.

Draper, D. (1994). Assessment and propagation of model uncertainty (with Discussion). *Journal of the Royal Statistical Society B*, **56**, to appear.

Draper, N.R. and Guttman, I. (1987). A common model selection criterion. *Probability and Bayesian Statistics*, (ed. R. Viertl), Plenum Publishing Corp., 139-150.

Edwards, W., Lindman, H. and Savage, L.J. (1963) Bayesian statistical inference for psychological research. *Psych. Rev.*, 70, 193–242.

Evett, I.W. (1991) Implementing Bayesian methods in forensic science. Paper presented at the Fourth Valencia International Meeting on Bayesian Statistics.

Fairley, D. (1989). Comment. *Statistical Science*, 4, 381–383.

Fienberg, S.E. and Mason, W.M.M. (1979). Identification and estimation of age-period-cohort effects in the analysis of discrete archival data. *Sociological Methodology 1979*, pp. 1–67.

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

Fornell, C. and Rust, R.T. (1989). Incorporating prior theory in covariance structure analysis: a Bayesian approach. *Biometrika*, 54, 249-259.

Freedman, D.A. (1983). A note on screening regression equations. *American Statistician*, 37, 152–155.

Garthwaite, P.H. (1992). Preposterior expected loss as a scoring rule for prior distributions. Technical Report, Department of Statistics, University of Aberdeen.

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

Gelfand, A.E. and Dey, D.K. (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society, series 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.*), Oxford University Press, pp. 147-167.

Genz, A. and Kass, R.E. (1993). Subregion adaptive integration of functions having a dominant peak. Technical Report, Department of Statistics, Carnegie Mellon University.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, 57, 1317-1340.

George, E.I. and McCulloch, R.E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88, 881–889.

Gilovich, T., Vallone, R., and Tversky, A. (1985) The hot hand in basketball: on the misperception of random sequences, *Cognitive Psychology, 17*, 295-314.

Good, I.J. (1950). *Probability and the Weighing of Evidence.* London: Griffin.

Good, I.J. (1952). Rational decisions. *Journal of the Royal Statistical Society (Series B)*, 14, 107–114.

Good, I.J. (1967). A Bayesian significance test for multinomial distributions. *Journal of the Royal Statistical Society, Series B*, 29, 399-431.

Good, I.J. (1983). Explicativity, corroboration, and the relative odds of hypotheses (#846).

In *Good Thinking – The Foundations of Probability and Its Applications*, University of Minnesota Press, Minneapolis, MN.

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

Good, I.J. (1992). The Bayes/non-Bayes compromise: A brief review. *Journal of the American Statistical Association*, 87, 597–606.

Good, I.J. and Crook, J.F. (1987). The robustness and sensitivity of the mixed Dirichlet Bayesian test for 'independence' in contingency tables. *Annals of Statistics*, 15, 670–693.

Greaney, V. and Kelleghan, T. (1984) *Equality of Opportunity in Irish Schools.* Dublin: Educational Company.

Groenewald, P.C.N. and de Waall, D.J. (1989). Bayesian tests for some precise hypotheses on multi-normal means. *Austral. J. Statist.*, 31, 393-408.

Grusky, D.B. and Hauser, R.M. (1984). Comparative social mobility revisited: Models of convergence and divergence in 16 countries. *Amer. Sociol. Rev.*, 49, 19-38.

Gunel, E. and Dickey, J. (1974). Bayes factors for independence in contingency tables. *Biometrika*, 61, 545-557.

Hammersley, J.M. and Handscomb, D.C. (1964). *Monte Carlo Methods.* London: Chapman and Hall.

Hartigan, J.A. (1992). Locally uniform prior distributions. Technical Report, Department of Statistics, Yale University.

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

Hodges, J.S. (1987) Uncertainty, policy analysis and statistics. *Statistical Science*, 2, 259–291.

Hsiao, K. (1994). *?* Ph.D. dissertation, Department of Statistics, Carnegie Mellon University.

Jeffreys, H. (1935) Some tests of significance, treated by the theory of probability. *Proc. Camb. Phil. Soc., 31:* 203-222.

Jeffreys, H. (1961). *Theory of Probability*, Third Edition. Oxford University Press.

Kadane, J.B. and Dickey, J.M. (1980). Bayesian Decision Theory and the Simplification of Models. In *Evaluation of Econometric Models*, (eds. J. Kmenta and J. Ramsey), Academic Press, 245-268.

Kadane, J.B., Dickey, J., Winkler, R., Smith, W. and S. Peters. (1980). Interactive Elicitation of Opinion for A Normal Linear Model, *J. Amer. Statist. Assoc.*, 75: 845-854.

Kass, R.E. (1993). Bayes factors in practice. *Statistician*, 42: 551-560.

Kass, R.E. and Greenhouse, J.B. (1989). Comment on "Investigating therapies of potentially great benefit: ECMO," by Ware (1989), *Statist. Science*, 4, 310-317.

Kass, R.E. and Steffey, D.L. (1989). Approximate Bayesian methods in conditionally independent hierarchical models (parametric empirical Bayes models). *J. Amer. Statist. Assoc.*, 84, 717-726.

Kass, R.E., Tierney, L., and Kadane, J.B. (1990). The validity of posterior asymptotic expansions based on Laplace's method, in *Bayesian and Likelihood Methods in Statistics and Econometrics*, ed. by S. Geisser, J.S. Hodges, S.J. Press, and A. Zellner. New York: North-Holland.

Kass, R.E. and Vaidyanathan, S. (1992) Approximate Bayes factors and orthogonal parameters, with application to testing equality of two Binomial proportions. *J. Royal Statist. Soc. B.*, 54: 129-144.

Kass, R.E. and Wasserman, L. (1992a). Improving the Laplace approximation using posterior simulation. Technical Report no. 566, Department of Statistics, Carnegie Mellon University.

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

Katz, R.W. (1981), On some criteria for estimating the order of a Markov chain, *Technometrics,* 23, 243-249.

Le, N.D., Raftery, A.E. and Martin, R.D. (1990). Robust order selection for autoregressive models using robust Bayes factors. Technical Report no. 194, Department of Statistics, University of Washington.

Leamer, E.E. (1978) *Specification Searches: Ad Hoc Inference with Nonexperimental Data.* New York: Wiley.

Leamer, E.E. (1983). Model choice and specification analysis. In *Handbook of econometrics*, Vol. 1, (eds. Z. Griliches and M.D. Intriligator), North-Holland, Amsterdam.

Leonard, T. (1982). Comment on "A simple predictive density function," *J. Amer. Statist. Assoc.*, 77, 657-658.

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

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

Lindley, D.V. (1961). The use of prior probability distributions in statistical inference and decisions. Fourth Berkeley Symposium. In *Proc. of Fourth Berkeley Symposium on Math. Stat. and Prob.*, (ed. J. Neyman), Berkeley: U. of California Press, 453-468.

Lindley, D.V. (1988). Statistical inference concerning Hardy-Weinberg equilibrium. In *Bayesian Statistics 3*, (eds. J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith), Oxford University Press, 307-326.

Linhart, H. and Zucchini, W. (1986). *Model Selection.* New York: Wiley.

Madigan, D. and Raftery, A.E. (1994) Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, to appear.

Madigan, D. and York, J. (1992) Bayesian graphical models for discrete data. Technical Report no. 259, Department of Statistics, University of Washington.

Madigan, D., Raftery, A.E., York, J.C., Bradshaw, J.M., and Almond, R.G. (1994). Strategies for graphical model selection. In *Selecting Models from Data: AI and Statistics IV*, (P. Cheeseman and R.W. Oldford, eds.), New York: Springer-Verlag, pp. 91-100.

McCulloch, R.E. and Rossi, P.E. (1991). A Bayesian approach to testing the arbitrage pricing theory. *Journal of Econometrics*, 49, 141–168.

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

Meng, X.L. and Wong, W.H. (1993). Simulating ratios of normalizing constants via a simple identity. Technical Report no. 365, Department of Statistics, University of Chicago.

Miller, A.J. (1984) Selection of subsets of regression variables (with Discussion). *J. R. Statist. Soc. (ser. A)*, 147, 389–425.

Miller, A.J. (1990). *Subset Selection in Regression.* London: Chapman and Hall.

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

Mosteller, F., and Wallace, D.L. (1964), *Inference and Disputed Authorship: The Federalist*, Reading, MA: Addison-Wesley. Reprinted in 1985 by Springer-Verlag.

Moulton, B.R. (1991). A Bayesian approach to regression selection and estimation, with application to a price index for radio services. *Journal of Econometrics*, 49, 169–193.

Neal, R.M. (1992). Probabilistic inference for artificial intelligence using Monte Carlo methods based on Markov chains. Technical Report, Department of Computer Science, University of Toronto.

Newton, M.A. and Raftery, A.E. (1994) Approximate Bayesian inference by the weighted likelihood bootstrap (with Discussion). *Journal of the Royal Statistical Society, series B*, **56**, 3–48.

O'Hagan, A. (1991). Contribution to the discussion of "Posterior Bayes factors". *Journal of*

the Royal Statistical Society, series B, 53, 137.

O'Hagan, A. (1994). Fractional Bayes factors for model comparison (with Discussion). *Journal of the Royal Statistical Society B*, **56**, to appear.

Parzen, E. (1974) Some recent advances in time series modeling. *IEEE Trans. Autom. Control,* 19, 723-730.

Pena, D. and Guttman, I. (1992). Comparing probabilistic methods for outlier detection in linear models. Technical Report, Department of Statistics, University of Toronto.

Pettit, L.I. and Young, K.D.S. (1990). Measuring the effect of observations on Bayes factors. *Biometrika*, 77, 455–466.

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

Polson, N.G. and Roberts, G.O. (1994). Bayes factors for discrete observations from diffusion processes. *Biometrika*, 81, 11–26.

Racine, A., Grieve, A.P., Fluhler, H., and Smith, A.F.M. (1986) Bayesian methods in practice: experiences in the pharmaceutical industry (with Discussion). *Appl. Statist.,* 35, 93-150.

Raftery, A.E. (1986a) A note on Bayes factors for log-linear contingency table models with vague prior information. *J. R. Statist. Soc. B,* 48, 249-250.

Raftery, A.E. (1986b). Choosing models for cross-classifications. *American Sociological Review*, 51, 145–146.

Raftery, A.E. (1987) Inference and prediction for a general order statistic model with unknown population size. *J. Amer. Statist. Ass.,* 82, 1163-1168.

Raftery, A.E. (1988a) Analysis of a simple debugging model. *Appl. Statist.*, 37, 12–22.

Raftery, A.E. (1988b) Approximate Bayes factors for generalized linear models. Technical Report 121, Department of Statistics, University of Washington.

Raftery, A.E. (1989). Are ozone exceedance rates decreasing? *Statistical Science*, 4, 378–381.

Raftery, A.E. (1992). Discussion of "Model determination using predictive distributions with implementations via sampling-based methods." In *Bayesian Statistics 4* (eds. J.M. Bernardo *et al*), Oxford University Press, pp. 160–163.

Raftery, A.E. (1993a). Discussion of three papers on the Gibbs sampler and other Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, series B*, 53, 85.

Raftery, A.E. (1993b). Bayesian model selection in structural equation models. In *Testing Structural Equation Models* (eds. K.A. Bollen and J.S. Long), Beverly Hills: Sage, pp. 163–180.

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

Raftery, A.E. (1994). Hypothesis testing and model selection with posterior simulation. In *Practical Markov Chain Monte Carlo* (W.R. Gilks, D.J. Spiegelhalter and S. Richardson, eds.), London: Chapman and Hall, to appear.

Raftery, A.E., and Akman, V.E. (1986) Bayesian analysis of a Poisson process with a change-point. *Biometrika,* 73, 85-89.

Raftery, A.E. and Banfield, J.D. (1990). Stopping the Gibbs sampler, the use of morphology and other issues in spatial statistics. *Annals of the Institute of Statistical Mathematics,* 43, 32–43.

Raftery, A.E., Givens, G.H. and Zeh, J.E. (1994). Inference from a deterministic population dynamics model about bowhead whale, *Balaena mysticetus,* replacement yield (with Discussion). *Journal of the American Statistical Association*, to appear.

Raftery, A.E., and Hout, M. (1985) Does Irish education approach the meritocratic ideal? A logistic analysis. *Economic and Social Review,* 16, 115-140.

Raftery, A.E. and Hout, M. (1993). Maximally maintained inequality: Expansion, reform and opportunity in Irish education, 1921-1975. *Sociology of Education*, 66, 41–62.

Raftery, A.E., Madigan, D. and Volinsky, C.T. (1995). Accounting for model uncertainty in survival analysis improves predictive performance. In *Bayesian Statistics 5* (J.M. Bernardo *et al.*, eds.), to appear.

Raftery, A.E. and Zeh, J.E. (1993). Estimation of Bowhead Whale, *Balaena mysticetus*, population size (with Discussion). In *Bayesian Statistics in Science and Technology: Case Studies*, (eds. C. Gatsonis, J.S. Hodges, R.E. Kass, N.D. Singpurwalla), New York: Springer-Verlag, pp. 163–240.

Raftery, A.E., Madigan, D.M. and Hoeting (1993). Model selection and accounting for model uncertainty in linear regression models. Technical Report no. 262, Department of Statistics, University of Washington.

Rissanen, J. (1987) Stochastic complexity. *J. R. Statist. Soc. B,* 49, 223-239.

Robert, C.P. (1992). A note on Jeffreys-Lindley paradox. *Statistica Sinica*, to appear.

Robert, C.P. and Caron, N. (1991). Noninformative Bayesian testing and neutral Bayes factors. Technical Report no. 148, Laboratoire de Statistique Théorique et Appliquée, Université de Paris VI.

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

for area counts. Ph.D. dissertation, Department of Biostatistics, University of Washington, 1992.

Rossi, P.E. (1985). Comparison of alternative functional forms in production. *Journal of Econometrics*, 30, 345-361.

Rossi, P.E. (1988). Comparison of dynamic factor demand models. Dynamic econometric modeling. Proceedings of the Third International Symposium in Economic Theory & Econometrics, (eds W.A. Barnett, E.R. Berndt and H. White), Cambridge University Press, 357-376.

Rubin, D.B. (1984) Bayesianly justifiable and relevant frequency calculations for the applied statistician, *Ann. Statist.*, 12, 1151-1172.

Schotman, P. and van Dijk, H. (1991). A Bayesian analysis of the unit root in real exchange rates. *J. Econometrics*, 49, 195-238.

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

Shibata, R. (1976), Selection of the order of an autoregressive model by Akaike's Information Criterion. *Biometrika,* 63, 117-126.

Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Annals of Statistics*, 8, 147–164.

Shibata, R. (1981). Optimal selection of regression variables. *Biometrika*, 68, 45-54.

Sklar, R. and Strauss, B. (1980). Role of the *uvrE* gene product and of inducible 0^*-methylguanine removal in the induction of mutations by N-methyl-N'-nitro-N-nitrosoguanidine in *Escherichia coli. J. Molec. Biol., 143*, 345-363.

Slate, E. (1994). Parameterizations for natural exponential families with quadratic variance functions. *J. Amer. Statist. Assoc.*, to appear.

Smith, A.F.M. (1991). Contribution to the Discussion of "Posterior Bayes factors". *Journal of the Royal Statistical Society, series B*, 53, 132–133.

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

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

Smith, A.F.M. and Roberts, G.O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods, *J. Royal Statist. Soc., B*, 55, 3-23.

Smith, R.L. (1989). Extreme value analysis of environmental time series: An application to trend detection in ground-level ozone (with discussion) *Statist. Science*, 4, 367–377.

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

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

Stone, M. (1979) Comments on model selection criteria of Akaike and Schwarz. *Journal of the Royal Statistical Society, series B*, 41, 276–278.

Tanner, M. and Wong, W. (1987). The calculation of posterior distributions by data augmentation (with discussion). *J. Amer. Statist. Assoc.*, 82, 528-550.

Thompson, E.A. and Wijsman, E.M. (1990) The Gibbs sampler on extended pedigrees: Monte Carlo methods for the genetic analysis of complex traits. Technical Report no. 193, Department of Statistics, University of Washington.

Tierney, L. (1989). Bayesian analysis in New-S and Lisp. *Proceedings of the Section on Statistical Computing*, American Statistical Association.

Tierney, L. (1990). *Lisp-Stat*. New York: Wiley.

Tierney, L., and Kadane, J.B. (1986) Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Ass.*, 81, 82-86.

Tierney, L., Kass, R.E. and Kadane, J.B. (1989). Fully exponential Laplace approximations to expectations and variances of nonpositive functions. *J. Amer. Statist. Assoc.*, 84, 710-716.

Verdinelli, I. and Wasserman, L. (1993a). Bayes factors, nuisance parameters and imprecise tests. Technical Report no. 570, Department of Statistics, Carnegie-Mellon University.

Verdinelli, I. and Wasserman, L. (1993b). A note on generalizing the Savage-Dickey density ratio. Technical Report no. 573, Department of Statistics, Carnegie-Mellon University.

Worsley, K.J. (1986). Confidence regions and tests for a change-point in a sequence of exponential family random variables. *Biometrika*, 73, 91–104.

Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics.* New York: Wiley.

Zellner, A. (1978). Jeffreys-Bayes posterior odds ratio and the Akaike Information criterion for discriminating between models. *Economics Letters 1*, North Holland Publishing Company, 337-342.

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

Zellner, A. (1987). Comment. *Statistical Science*, 2, 339–341.

Zellner, A. and Siow (1980). Posterior odds for selected regression hypotheses. In *Bayesian Statistics 1* (eds. J.M. Bernardo *et al.*), Valencia University Press, pp. 585–603. Reply, 638-643.