Some Recent Developments in Bayesian Analysis, with Astronomical Illustrations

James O. Berger

ABSTRACT New developments in default Bayesian hypothesis testing and model selection are reviewed. As motivation, the surprising differences between Bayesian and classical answers in hypothesis testing are discussed, using a simple example. Next, an example of model selection is considered, and used to illustrate a new default Bayesian technique called the "intrinsic Bayes factor". The example involves selection of the order of an autoregressive time series model of sunspot data. Classification and clustering is next considered, with the default Bayesian approach being illustrated on two astronomical data sets. In part, Bayesian analysis is experiencing major growth because of the development of powerful new computational tools, typically called Markov Chain Monte Carlo methods. A brief review of these developments is given. Finally, some philosophical comments about reconciliation of Bayesian and classical schools of statistics are presented.

1 Introduction

Bayesian Analysis and Astronomy have a long history together. Although Bayesian analysis is named after Thomas Bayes, who wrote the first paper on the subject (Bayes, 1783), it was Laplace who, in the late eighteenth and early nineteenth centuries, extensively developed the Bayesian approach to statistics, culminating in his revolutionary statistics book, Laplace (1812). Much of Laplace's motivation in this development was the solution of problems in celestial mechanics.

The Bayesian approach to statistics, then called "inverse probability," dominated the statistical scene for most of the nineteenth century, but fell into disfavor in the first half of the twentieth century. It has recently staged a major revival, and is indeed today the dominant approach in several statistical areas of interest to astronomy, such as image processing. Use of maximum entropy methods, which is a type of Bayesian analysis, has also become very popular in some astronomical circles. Additional insight into the use of Bayesian analysis in the astronomical community is available in the excellent papers Loredo (1992) and Ripley (1992), from the previous conference.

Virtually any problem involving uncertainty can be approached from a Bayesian perspective, and the quantity of Bayesian methodological development being undertaken today is enormous. Hence a review of recent Bayesian developments or assessment of its future impact is virtually impossible. The goal of this paper is thus considerably more modest. We will simply try to illustrate some of the features of Bayesian analysis, with an eye towards helping astronomers to better judge whether or not they should consider using Bayesian methodology. There is a substantial learning curve involved in becoming adept at use of Bayesian methodology, and so some indications of the rewards and potential uses of the methodology can be helpful in deciding whether or not to invest effort in this direction.

A personal caveat is in order before proceeding. I have had little direct involvement in the analysis of astronomical problems, and so the illustrations I will consider are rather sterile, containing rather minimal astronomical content; indeed, the illustrations might even be "bad science" in terms of astronomical understanding. I hope readers will not be too distracted by this, and that at least some of the potential of Bayesian analysis is nevertheless apparent.

Related to this is the caveat that I will be focusing primarily on 'default' Bayesian methodology. This can best be described as a toolkit of Bayesian procedures which can be used automatically, much the same as many standard classical statistical procedures. In a sense, limiting the paper to this subject is somewhat unfortunate, because the Bayesian paradigm is much more; indeed, many argue that its greatest strength is to allow completely general interactive modelling between science, data, and opinions of the investigator, to reach a holistic end. My own experiences in this regard are not within astronomy, however, and hence I cannot provide an astronomical illustration of this possibility.

Even within default Bayesian analysis, I will primarily focus on one issue: use of Bayesian methods in hypothesis testing and model selection. This area is of considerable interest because it is an area in which Bayesian answers systematically differ from classical answers. (In contrast, when dealing with estimation or prediction problems for reasonably large data sets, there are typically only modest differences between classical and Bayesian answers.) Section 2 introduces the subject through an elementary, but surprising, example. Section 3 considers a more involved application to time series analysis. Section 4 deals with analysis of mixture problems, with applications to clustering.

One of the reasons for the recent upsurge in use of Bayesian methods is the advent of powerful computational engines based on Markov Chain Monte Carlo procedures. In Section 5 we give a brief introduction to these techniques. Section 6 concludes with some philosophical perspective.

2 Bayesian hypothesis testing and model selection

2.1 Motivation and a simple example

The Bayesian approach to hypothesis testing and model selection is conceptually straightforward. One assigns *prior* probabilities to all unknown hypotheses or models, and uses probability theory to compute the *posterior* probabilities of the hypotheses or models, given observed data. The key advantage of this approach is that the answers can be interpreted as the actual probabilities that the hypotheses or models are true, in contrast to standard significance testing which lacks any such interpretation. While there is a common intellectual appreciation of the difference, the impact of this intellectual appreciation upon practice seems to be quite limited; the following illustration of the difference is useful in helping to convince oneself that it is a serious matter.

Example 1. Suppose we observe independent $N(\theta, 1)$ data (i.e., data following the normal or Gaussian distribution with mean θ and variance 1), and that it is desired to test $H_1 : \theta = 0$ versus $H_2 : \theta \neq 0$. The typical classical procedure that is used is to compute the P-value or observed significance level, and to consider there to be significant evidence against H_1 if this P-value is small enough.

The following interesting simulation is easy to perform: repeatedly generate random data from H_1 and H_2, corresponding to a series of tests of H_1 versus H_2, and see where the series of P-values happens to fall. For data generated from H_1, this is no mystery; by definition, the fraction of P-values that would fall in a given interval, say $0.04 < P\text{-value} < 0.05$, would be the size of the interval (here 0.01). The *nonshaded* vertical bars in Figure 1 reflect this, showing the fraction of P-values that would fall in each of many intervals, were the data generated from H_1. (Three separate graphs are presented, because it is the smaller ranges of P-values that are typically of more interest.)

Now let us see where the series of P-values happen to fall if they arise from H_2. There is a certain ambiguity here, because what does it mean to generate data from H_2? One possibility is to simply pick some value of θ, say 2, and generate data from the $N(2, 1)$ distribution. Another possibility is to randomly select θ from some distribution $\pi(\theta)$, generate data from the selected θ, and then repeat the process with new θ arising from $\pi(\theta)$. This latter possibility better mirrors the actual daily use of significance testing, and so Figure presents the results of one such simulation, with $\pi(\theta)$ chosen to be a $N(0, 2)$ distribution and the sample size (for each P-value computation) chosen to be one. (The simulation itself repeats the generation of θ and data many times.) The *shaded* columns in Figure 1 present the fraction of P-values that fell in each interval. (Note that, while the data for the shaded columns were generated from H_2, we still just compute the P-value relative to H_1.)

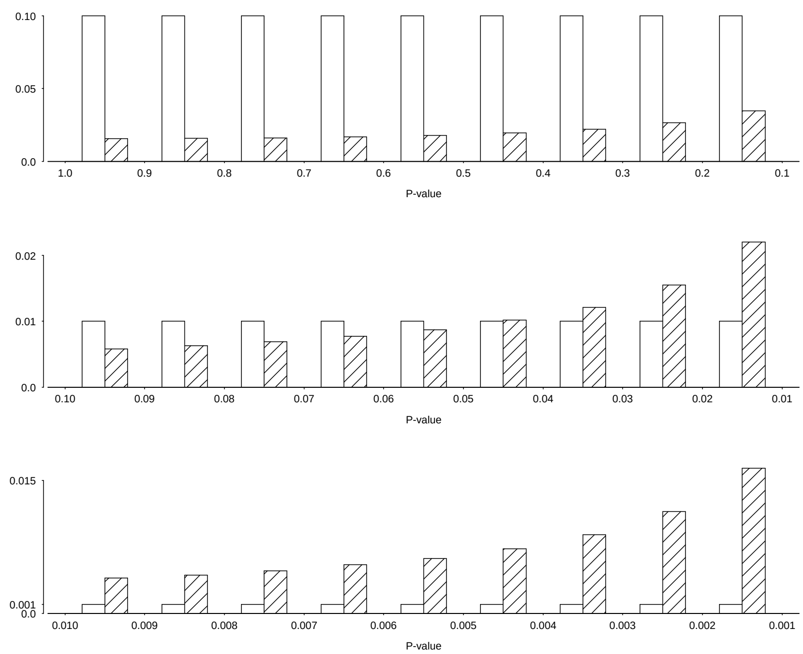

FIGURE 1. The fraction of normal observations that will fall in each given interval of p-values under $H_1 : \theta = 0$ (fraction given by unshaded colmun) and under $H_2 : \theta \neq 0$ (fraction given by shaded colmun).

The qualitative nature of the results is no surprise: the larger the P-value, the more likely that it arose from H_2 than from H_1. But the quantitative implications come as a considerable surprise to those who see this for the first time. For instance, suppose you observe a P-value in the interval $(0.04, 0.05)$. This is a rare event under H_1, happening only 1% of the time, but it is nearly as rare under H_2, happening only about 2% of the time. Thus, if one had to bet whether this P-value arose from H_1 or H_2, it would be prudent to bet on H_2 at no more than 2 to 1 odds (assuming the two hypotheses were judged to be equally likely a priori). This last is essentially the Bayesian conclusion from the problem, that a P-value of 0.05 provides at most weak evidence in favor of H_2.

Consider next the case in which the P-value is in the interval $(0.009, 0.01)$. In classical language this is typically termed "highly significant evidence against H_1," and yet we see that the likelihood that the data came from H_2 is only 5 times the likelihood that it came from H_1. Odds of 5 to 1 are meaningful, but hardly carry the level of conclusiveness that is usually accorded the phrase "highly significant evidence."

A final comment before leaving this example: one might well be suspicious that the startling nature of Figure 1 is due to the particular way in which we generated data from H_2. This is not the case. One can make *any* choices of the parameters under H_2, and use *any* sample size (for the P-value computation), and the story remains roughly the same or worse. For instance, the fraction of P-values that will fall in the interval $(0.04, 0.05)$, when the data is generated under H_2, can be shown to be *at most* 0.034, and so the odds of H_2 to H_1 can be *at most* 3.4 to 1 when the P-value is in this interval. Interestingly, the fraction of P-values in the interval has no lower bound, and decreases rapidly as the sample size increases; thus if one performed this simulation with a large sample size (for computing the P-value), it would typically indicate that a P-value in the interval $(0.04, 0.05)$ is evidence *in favor* of H_1.

The above example illustrates what is, perhaps, the most attractive feature of Bayesian analysis, namely that its conclusions carry a clear interpretation. While there is nothing "wrong" with the classical P-value here, in that its definition is precise and it does convey information of interest, learning how to interpret a P-value as a measure of the evidence for H_1 relative to H_2 is extremely difficult, requiring major adjustments for the sample size and the type of testing that is done (among other things).

The phenomenon in the above example is very general, applying to most testing problems in which the hypotheses are of differing dimensions, including typical chi-squared testing of fit. For more extensive discussion, see Edwards, Lindeman, and Savage (1963), Berger and Sellke (1987), Berger and Delampady (1987), and Delampady and Berger (1990).

2.2 Notation

Hypothesis testing and model selection are essentially the same from a Bayesian perspective; we will henceforth use notation that is more designed for model selection. Also, we will only consider analysis of parametric models.

The data, $\underset{\sim}{X}$, is assumed to have arisen from one of several possible models $M_1, \ldots, M_m$. Under M_i, the density of $\underset{\sim}{X}$ is $f_i(\underset{\sim}{x}|\underset{\sim}{\theta}_i)$, where $\underset{\sim}{\theta}_i$ is an unknown vector of parameters of f_i.

The Bayesian approach to model selection begins by assigning prior probabilities, $P(M_i)$, to each model; often, equal prior probabilities are used, i.e. $P(M_i) = 1/m$. It is also necessary to choose prior distributions $\pi(\underset{\sim}{\theta}_i)$ for the unknown parameters of each model; sometimes these can also be chosen in a "default" manner, as will be illustrated later.

The analysis then proceeds by computing the posterior probabilities of each model, which elementary probability theory (Bayes theorem) shows to be equal to

$$P(M_i|\underset{\sim}{x}) = \frac{P(M_i)m_i(\underset{\sim}{x})}{\sum\limits_{j=1}^{m} P(M_j)m_j(\underset{\sim}{x})}, \tag{1.1}$$

where $m_j(\underset{\sim}{x}) = \int f_j(\underset{\sim}{x}|\underset{\sim}{\theta}_j)\pi_j(\underset{\sim}{\theta}_j)d\underset{\sim}{\theta}_j$. Typically one selects the model (or models) with largest posterior probability. Note that, when the prior probabilities, $P(M_i)$, are equal, then the $P(M_i|\underset{\sim}{x})$ equal

$$P_i \equiv m_i(\underset{\sim}{x}) / \sum_{j=1}^{m} m_j(\underset{\sim}{x}). \tag{1.2}$$

It is common to simply report the P_i in the summary of an investigation, since someone with unequal $P(M_i)$ can easily recover *their* $P(M_i|\underset{\sim}{x})$ via the alternate expression

$$P(M_i|\underset{\sim}{x}) = P(M_i) \cdot P_i / \sum_{j=1}^{m} P(M_j) \cdot P_j.$$

Example 1 (continued). Here, the data is $\underset{\sim}{X} = (X_1, \ldots, X_n)$, where the X_i are independent $N(\theta, 1)$ observations. Model M_1 corresponds to assuming $\theta = 0$, in which case $f_1(\underset{\sim}{x}|0)$ is just the standard normal density. Model M_2 corresponds to assuming $\theta \neq 0$. Since M_1 has no unspecified parameters, $\pi_1(\theta_1)$ is not needed. For M_2, we assumed in the earlier simulation that θ has a $N(0, 2)$ distribution, which would thus be $\pi_2(\theta)$. Computation then yields

$$\begin{aligned} m_1(\underset{\sim}{x}) &= f_1(\underset{\sim}{x}|0) = \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi}}\, e^{-x_i^2/2} \right), \\ m_2(\underset{\sim}{x}) &= \int_{-\infty}^{\infty} f_2(\underset{\sim}{x}|\theta)\pi_2(\theta)d\theta \\ &= \int_{-\infty}^{\infty} \left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}}\, e^{-(x_i-\theta)^2/2} \right) \cdot \frac{1}{\sqrt{4\pi}} e^{-\theta^2/4} d\theta. \end{aligned}$$

Evaluation of the above integral, and use of (1.2) yields

$$P_1 = 1 - P_2 = \left[1 + (1+2n)^{-1/2}\, \exp\{n\overline{x}^2/(2+\frac{1}{n})\} \right]^{-1}. \tag{1.3}$$

Thus if $n = 1$ and $|\overline{x}| = 1.96$ (so that the P-value would be exactly 0.05), then $P_1 = 0.325$ and $P_2 = 0.675$. These mirror the conclusion from Figure 1 that, even though the P-value is 0.05, there is almost a (1/3) chance that $\theta = 0$ is true (assuming that we initially give M_1 and M_2 equal prior probabilities of being true).

2.3 Default implementation

The two difficulties in implementing Bayesian model selection are (i) choosing the prior distributions $\pi_i(\underset{\sim}{\theta_i})$, and (ii) computing the $m_i(\underset{\sim}{x})$. A variety of strategies exist for carrying out the integrations necessary to compute the $m_i(\underset{\sim}{x})$; see Kass and Raftery (1995) for discussion. Choosing the $\pi_i(\underset{\sim}{\theta_i})$ is more of a problem.

It may well be the case that subjective knowledge about the $\underset{\sim}{\theta_i}$ is available, and can be incorporated into subjective proper priors for the $\underset{\sim}{\theta_i}$. This is clearly desirable if it can be done. Thus, in Example 1, one might feel that, if $\theta \neq 0$, then θ will be near 5 with an uncertainty of ± 1. (Perhaps the alternative theory, M_2, would predict this.) Choosing $\pi_2(\theta)$ to then be, say, a $N(5, 1)$ distribution would be reasonable. If one had no specific alternatives in mind, one might at least want to specify a guess, τ, as to the likely amount of departure of θ from 0 under H_2. Interpreting τ as a prior "standard error", it would then be reasonable to use a $N(0, \tau^2)$ prior distribution for $\pi_2(\theta)$. An alternative to guessing τ is to simply specify a plausible range for τ, and see if the desired conclusion holds for the entire range. This is called *robust Bayesian analysis*; see Jeffreys and Berger (1992) for a simple but interesting astronomical example, and Berger (1994) for a recent review.

Subjective Bayesian analysis is often avoided, for a variety of reasons. The most common argument against subjective Bayesian analysis is that "scientific discourse demands objectivity, and hence one cannot use a subjective Bayesian analysis." The merits of this argument are highly debatable (cf, Berger and Berry, 1988), but one cannot deny that at least the appearance of objectivity can be helpful.

The other primary objection to subjective Bayesian analysis is that it is often too difficult. This is especially true of model selection problems, in which there may be several high-dimensional models and eliciting all the needed high-dimensional prior distributions would be a truly formidable undertaking. (We do not mean to imply that there is anything wrong with subjective Bayesian analysis - indeed it can be argued that one should always first try to implement such an analysis - but the difficulties are real.)

For these and other reasons, the most popular Bayesian methods tend to be default methods, which operate with default prior distributions. For estimation and prediction problems, the default Bayesian theories are well developed, and use prior distributions that are designed to be "noninformative" in some sense. The most famous of these is the *Jeffreys prior*, named after the famous geophysicist who, through his exemplary book Jeffreys (1961), was most responsible for the modern Bayesian revival. *Maximum entropy* priors are another well-known type of noninformative prior (although they often have certain features specified). The more recent statistical literature emphasizes what are called *reference priors*, which prove remarkably successful even in higher dimensional problems (cf., Berger and Bernardo, 1992, and Yang and Berger, 1995).

Testing and model selection has proved much more resistant to the development of default Bayesian methods. This is because the "noninformative" priors discussed above are typically improper distributions, meaning they do not have mass equal to one. This does not typically pose a problem in estimation and prediction, but it does for testing and model selection. The expressions in (1.1) and (1.2) really make sense only if the $\pi_i(\underset{\sim}{\theta}_i)$ (and hence the $m_i(\underset{\sim}{x}_i)$) are proper distributions.

Jeffreys faced this problem squarely, and considered the various (unappealing) choices. One choice is to use classical measures, such as the P-value; but their very misleading nature made that choice especially unattractive. Another choice is to demand subjective Bayesian analysis, but Jeffreys felt that this was too restrictive a requirement. The third possible choice is to simply pick some proper prior distributions that seem reasonable (for the given models), and *conventionally* use them for default hypothesis testing or model selection. This was the choice that Jeffreys adopted. Thus, in Example 1, he gave extensive arguments supporting use of a standard Cauchy distribution as the default prior. (The choice we made, of a $N(0,2)$ distribution, gives almost the same answers and is easier to handle computationally.)

In principle, this approach of Jeffreys is our favorite approach. Its major disadvantage is that there is no well-developed theory for determining default priors for hypothesis testing and model selection. Hence progress in this direction has been very sporadic, with only certain special models being treated on a case-by-case basis. Because of this limitation, Bayesian model selection has, instead, typically been performed using an asymptotic approach which is known as the BIC (Bayesian Information Criterion, see Kass and Raftery, 1995, for discussion). This approach, however, has the usual disadvantages of asymptotics, including the need for regular models and large sample sizes (though see Kass and Wasserman, 1995).

Recently, two very general default Bayesian methods have appeared: the *fractional Bayes factor* approach of O'Hagan (1995), and the *intrinsic Bayes factor* approach of Berger and Pericchi (1996a). This last approach appears to be applicable to virtually any hypothesis testing and model selection problem (even for nonregular models), and seems to closely correspond to the recommended analyses of Jeffreys for the cases he considered. Hence we feel that it holds great promise for solving the hypothesis testing and model selection problem.

There are several versions of intrinsic Bayes factors, with different versions argued to be particularly useful in certain settings. There is, however, one quite simple version that seems to work well across all problems, and that is what we will describe here.

The idea behind intrinsic Bayes factors is simple, and is based on an ad hoc method that has long been used to address the model selection problem. The ad hoc method is to use part of the data (typically as small a part as possible) as "training data", to convert the standard noninformative priors (used in estimation problems) to proper distributions. Then one uses these proper distributions, with the remainder of the data, to compute the posterior probabilities in (1.1) or (1.2).

Besides being an ad hoc approach with no clear justification, the method was rather arbitrary in that the choice of the training data is typically arbitrary. The simple idea in Berger and Pericchi (1996a) was to eliminate this arbitrariness by doing the computation for all possible choices of the training data (or some reasonably large random subset of such choices), and then picking an "average" of the ensuing answers. As a general purpose method, picking the median of the ensuing answers seems to work extremely well, and defines what we call the *median intrinsic Bayes factor.* (This is strictly defined only for pairwise comparisons of models, but pairwise answers can easily be adapted to deal with multiple models or hypotheses.)

Example 1 (continued). The common noninformative prior for θ, under M_2, is $\pi(\theta) = 1$. This is improper, but if we use just one of the observations, say x_1, as a training sample, then we can convert $\pi(\theta)$ to the *posterior distribution*

$$\pi(\theta|x_1) = \frac{f_2(x_1|\theta) \cdot \pi(\theta)}{\int f_2(x_1|\theta) \cdot \pi(\theta) d\theta} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\theta - x_1)^2}.$$

Using this as the (now proper) distribution for θ, and computing (1.2) for the remaining data $x_2, \ldots, x_n$, yields

$$P_1 = 1 - P_2 = \left[1 + \sqrt{n} \;\; e^{-n\overline{x}^2/2} \cdot e^{x_1^2/2}\right]^{-1}.$$

(Here, $\overline{x}$ is the average of all the data.) Since choice of x_1 as the training sample was arbitrary, one can do the computation for all possible choices, and then take the median of the resulting answers. The result is clearly

$$P_1^* = 1 - P_2^* = \left[1 + \sqrt{n} \;\; e^{-n\overline{x}^2/2} \cdot e^{x^{*2}/2}\right]^{-1}, \tag{1.4}$$

where x^{*^2} is defined to be the median of $x_1^2, \ldots, x_n^2$.

While simple to implement (modulo possible computational difficulties), the intrinsic Bayes factor approach at first appeared to be just another ad hoc approach. Interest grew enormously, however, when it was shown that the answers resulting from this approach correspond very closely to answers from actual proper default prior analyses, of the type recommended by Jeffreys. One could now obtain reasonable answers without going through the involved arguments needed for the Jeffreys-type implementation.

Detailed discussion of this approach, its limitations, and advice for its computational implementation, can be found in Berger and Pericchi (1996a, 1996b). Note, however, that the median intrinsic Bayes factor is not emphasized therein; we have only recently come to appreciate it, for its general applicability and stability; except for extremely small sample sizes, we have not found any serious contraindications to its use.

3 An application to time series analysis

A situation in which one typically is considering a multitude of models is in time series analysis. Consider the time series in Figure 2, for instance. This

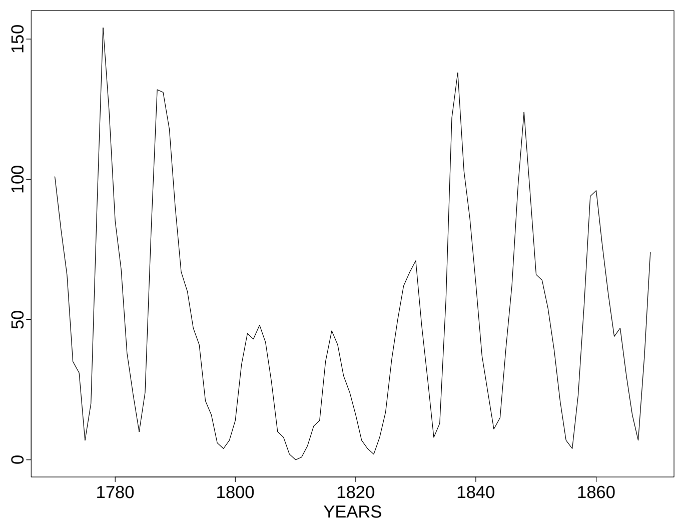

FIGURE 2. Wolf sunspot numbers.

is the famous Wolf sunspot series data, consisting of the number of sunspots observed each year from 1770 to 1869. A possibly reasonable model for the series is a stationary autoregressive process with drift. For instance, the AR(1) model with a linear drift would be described as

$$Y_t = \beta_1 t + \phi_1(Y_{t-1} - \beta_1(t-1)) + \epsilon_t, \tag{1.5}$$

where Y_t is the observation at time t, β_1 is the unknown linear coefficient, ϕ_1 is the unknown autocorrelation, and the ϵ_t are i.i.d. $\mathcal{N}(0, \sigma^2)$ errors, σ^2 also unknown.

It is decided to consider autoregressive models of order 1, 2, 3 and 4, and also to consider constant (C), linear (L), and quadratic (Q) drift. Thus the

TABLE 1.1. Posterior probabilities of models, assuming equal prior probabilities.

Model	$P(M_i \mid \underset{\sim}{x})$	Model	$P(M_i \mid \underset{\sim}{x})$
AR(1), C	~ 0	AR(3), C	0.740
AR(1), L	~ 0	AR(3), L	0.001
AR(1), Q	~ 0	AR(3), Q	0.001
AR(2), C	0.161	AR(4), C	0.076
AR(2), L	0.011	AR(4), L	0.006
AR(2), Q	0.001	AR(4), Q	0.001

AR(j) model with drift of polynomial order k ($k = 0, 1, 2$) can be written

$$Y_t = \sum_{\ell=0}^{k} \beta_\ell t^\ell + \sum_{r=1}^{j} \phi_r (Y_{t-r} - \sum_{\ell=0}^{k} \beta_\ell (t-r)^\ell) + \epsilon_t. \tag{1.6}$$

We are thus considering twelve models (any of the four AR models together with any of the three polynomial drifts).

The "intrinsic Bayes factor" approach applies directly to this problem. It utilizes only standard noninformative priors for the parameters (constant priors for the β_i and ϕ_i, and $1/\sigma^2$ for the variance, σ^2). Recall that one cannot use these noninformative priors directly, but must use them through the "intrinsic Bayes factor" algorithm. There is also a computational complication: because of the stationarity assumption, $\underset{\sim}{\phi} = (\phi_1, \phi_2, \ldots, \phi_j)$ is restricted to the "stationarity region," and so the integration in computation of the $m_i(\underset{\sim}{x})$ must be carried out over this region. Methods of doing this, as well as the relevant intrinsic Bayes factor algorithm, can be found in Varshavsky (1996). The results are summarized in Table 1.1.

The AR(3) model with no drift is clearly the preferred model, although the AR(2) model with no drift receives some support, and should not be ruled out on the basis of this data. Higher or lower autoregressive structures receive little support.

Note that the analysis very strongly discourages including any drift term in the model; none of the models with a drift term has posterior probability greater than 0.011. Indeed, there was no scientific reason to include drift terms in the model, but we did so as a pedagogical illustration. One of the highly attractive features of the Bayesian approach to model selection is that it acts as a natural "Ockham's razor", in the sense of favoring simpler models over more complex models, if the data provides roughly comparable fits for the models. Here, the models with drift will certainly fit the data slightly better than will the models without drift, but the Bayesian analysis automatically prefers the simpler non-drift models in the absence of a clearly superior fit. And this is without having to introduce any explicit penalty, such as reduced prior probabilities for the more complex models (although this is often also recommended - cf, Jeffreys, 1961).

In classical statistics, overfitting is avoided by introducing an ad hoc penalty term (as in AIC, Akaike Information Criterion), which increases as the complexity (i.e., the number of unknown parameters) of the model increases. Not only are such corrections ad hoc, but the standard ones do not sufficiently penalize complex models. For instance, a recent bias corrected version of AIC, designed for time series (see Hurvich and Tsai, 1989), when applied to the above data selects as the top four models the (AR(4), Q), (AR(4), C), (AR(4), L), and finally (AR(3), C) models. For an interesting historical example of Ockham's razor, and general discussion and references, see Jefferys and Berger (1992).

We have limited the discussion of this example to the model selection question. There is, of course, considerably more of interest to the problem. One might also want to estimate the parameters of the selected model (and provide associated standard errors), and/or use the selected model for optimal prediction. Any such inferences are readily available in the Bayesian approach; for instance, the noninformative priors mentioned earlier in the example could be used to compute the posterior distribution of the parameters, from which any inferences follow directly.

One item of special interest here is that the analysis would automatically incorporate the stationarity constraint, since the prior distribution is supported only on the stationarity region for the autocorrelation parameters. Incorporating this constraint in classical analysis is problematic; maximum likelihood estimates will often be on the boundary of the stationarity region, in which case the standard errors produced from likelihood asymptotics will usually be considerably too small.

Prediction also deserves special mention. It is a well-known problem that predictions based on selected models typically turn out to be much less accurate than the model would have suggested. This is especially so if the selected model is used in raw form, with estimated values of the parameters inserted. The obvious reason for this overly optimistic prediction is that one is pretending to know the model and the parameter values, when one really does not. To obtain reasonable estimates of predictive accuracy, one must not only incorporate the uncertainty in parameter values (as reflected by their posterior distributions), but also must incorporate the uncertainty in the model. In the above example, for instance, it would be reasonable to base the predictions on the appropriate weighted mixture of predictions from the (AR(3), C) and (AR (2), C) models. For discussion on this general issue, see Draper (1995).

4 An application to classification and mixture models

Figures 3 and 4 show two different sets of bivariate observations that arose in an astronomical setting. It was desired to (i) identify how many clusters are present in each set of data; (ii) determine which data set exhibits stronger clustering; (iii) provide a characterization of the clusters in each figure; and (iv) classify each observation in terms of cluster membership.

The natural Bayesian approach to this problem is to model the data as arising from a mixture distribution. We will talk as if each cluster can be identified with a separate distribution from this mixture, and that each distribution in the mixture corresponds to a separate population of observations. These interpretations are not strictly necessary, but they are useful conceptually and will often be reasonable as an explanation of the underlying process.

We will assume the observations arise from an additive mixture of bivariate normal populations; thus the overall density of an observation $\underset{\sim}{x} = (x_1, x_2)$ is

$$f(\underset{\sim}{x}) = \sum_{j=1}^{m} \gamma_j \; f_j(\underset{\sim}{x} | \underset{\sim}{u}_j, \underset{\sim}{\Sigma}_j),$$

where m is the number of populations (clusters) in the mixture; γ_j is the probability that an observation arises from population j (i.e., γ_j is the proportion of elements in cluster j); and f_j is a bivariate normal distribution with mean vector $\underset{\sim}{u}_j$ and covariance matrix $\underset{\sim}{\Sigma}_j$. Here we will assume that m and all the γ_j, $\underset{\sim}{u}_j$, and $\underset{\sim}{\Sigma}_j$ are completely unknown. Frequently, one might want to make further assumptions (such as assuming that the $\underset{\sim}{\Sigma}_j$ have a specified form or are all equal), but for analysis of the data in Figures 3 and 4 we make no such assumptions.

The mixture model is a difficult model for classical statistical analysis because standard asymptotics fails, and because maximum likelihood methods are often very unstable due to the presence of multiple modes in the likelihood. For Bayesian analysis there are also difficulties, primarily because standard (improper) noninformative priors cannot be used. (Subjective Bayesian analysis of mixture models is relatively straightforward, if one is willing to make the investment in prior specification; cf, Lavine and West, 1992).

We have recently modified the intrinsic Bayes factor algorithm to also enable analysis of mixture models. The idea is to, again, use (small) parts of the data as training samples; however, since we do not know which populations gave rise to which data, this has to be done as an iterative simulation involving the classification probabilities of the data. Details of the algorithm can be found in Shui (1996). The output of the algorithm is the posterior distribution of all unknown parameters. For m, only the

TABLE 1.2. The posterior probability of $m = 1, 2$, or 3 clusters, in each of the data sets in Figures 3 and 4.

	m		
	1	2	3
Data Set 1	8.1×10^{-14}	5.7×10^{-10}	~ 1
Data Set 2	1.3×10^{-6}	0.918	0.082

values 1, 2, and 3 were considered, and the posterior probabilities of each (assuming equal prior probabilities of 1/3) are given in Table 1.2 (for both the data set in Figure 3 and that in Figure 4).

For Data Set 2, there is overwhelming evidence that at least two clusters are present, and little evidence to support more than two clusters. This last is another illustration of the automatic Ockham's razor effect of Bayesian analysis; the simpler two-cluster model is preferred, because three clusters do not provide a markedly better fit.

The situation with Data Set 1 is more interesting. The two-cluster model is much preferred over the one-cluster model (the "odds" would be the ratio of the posterior probabilities, here 703), but three clusters is the overwhelmingly preferred model. The reason is that there are a number of obvious "outliers" in the lower right quadrant of Figure 3, and these outliers become their own "cluster". This effect is typical of normal mixture models, as normal distributions do not admit outliers. In a sense, this is not bad, since outliers often deserve special identification and treatment.

Table 1.3 gives the Bayesian estimates (here, posterior medians) of all parameters in the various populations for Data Set 1. (Standard errors of these estimates are also available, but are not reported here.) The populations, or clusters, have been labelled in terms of their size, with "1" being the largest.

There are no surprises in these results. For the preferred 3-cluster model, the third component indeed appears to be the outliers, and has the very small $\hat{\gamma}_3 = 0.025$. (There were 174 observations in all, so $\hat{\gamma}_3$ intuitively reflects about 4 outliers.) Note that the covariance matrices seem to be quite different, in both the 2-component and the 3-component models.

The final output available from the algorithm is the posterior probability that each observation arises from each cluster. This is given in Table 1.4 for the labeled observations from Figure 3. (As before, the clusters are labeled according to their size, with "1" being the largest and "3" the smallest.) There are no surprises in the table, although note that there is considerable uncertainty attached to the classification of some of the observations. Ordinary classification and clustering algorithms do not typically allow for reflection of such uncertainty.

TABLE 1.3. Estimates (posterior medians) of parameters in the mixture model for Data Set 1.

Parameter	Model: 1-Cluster	2-Cluster	3-Cluster
γ_1		0.774	0.749
$\underset{\sim}{u}_1$	$(0.892, -0.325)$	$(1.236, -0.408)$	$(1.266, -0.363)$
$\underset{\sim}{\Sigma}_1$	$\begin{pmatrix} 0.628 & -0.095 \\ -0.095 & 0.147 \end{pmatrix}$	$\begin{pmatrix} 0.223 & 0.010 \\ 0.010 & 0.157 \end{pmatrix}$	$\begin{pmatrix} 0.201 & 0.001 \\ 0.001 & 0.057 \end{pmatrix}$
γ_2		0.226	0.226
$\underset{\sim}{u}_2$		$(-0.309, \quad -0.075)$	$(-0.290, \quad -0.051)$
$\underset{\sim}{\Sigma}_2$		$\begin{pmatrix} 0.193 & 0.001 \\ 0.001 & 0.040 \end{pmatrix}$	$\begin{pmatrix} 0.166 & 0.000 \\ 0.000 & 0.037 \end{pmatrix}$
γ_3			0.025
$\underset{\sim}{u}_3$			$(1.195, \quad -1.729)$
$\underset{\sim}{\Sigma}_3$			$\begin{pmatrix} 0.203 & -0.011 \\ -0.011 & 0.265 \end{pmatrix}$

TABLE 1.4. Posterior probabilities that the labeled observations from Figure 3 belong to cluster 1, 2, or 3, in the favored three-cluster model.

Observation	Belongs to Cluster "1"	"2"	"3"
1	0.002	0.998	~ 0
2	0.429	0.571	~ 0
3	0.654	~ 0	0.346
4	~ 1	~ 0	~ 0
5	0.906	~ 0	0.094
6	~ 0	~ 0	~ 1

5 Advances in Bayesian computation

5.1 Introduction

Recent computational tools have allowed application of Bayesian methods to highly complex and nonstandard models. Indeed, for many complicated models, Bayesian analysis has now become the simplest (and often only possible) method of analysis.

Although other goals are possible, most Bayesian computation is focused on calculation of posterior expectations $E^*[g(\theta)]$, where E^* represents expectation with respect to the posterior distribution and $g(\theta)$ is some function of interest. For instance, if $g(\theta) = \theta$, then $E^*[g(\theta)] = E^*[\theta] \equiv \mu$, the posterior mean; if $g(\theta) = (\theta - \mu)^2$, then $E^*[g(\theta)]$ is the posterior variance of θ; and, if $g(\theta)$ is 1 if $\theta > C$ and 0 otherwise, then $E^*[g(\theta)]$ is the posterior probability that θ is greater than C. Another common type of Bayesian computation is calculation of the posterior mode (as in computation of MAP estimates in image processing); we do not formally discuss this here, although a number of techniques discussed below can also be useful in this regard.

5.2 Traditional numerical methods

The 'traditional' numerical methods for computing $E^*[g(\theta)]$ are numerical integration, Laplace approximation, and Monte Carlo Importance Sampling. Brief introductions to these methods can be found in Berger (1985). Here we say only a few words, to place the methods in context and provide references.

A successful general approach to **numerical integration** in Bayesian problems, using adaptive quadrature methods, was developed in Naylor and Smith (1982). This was very effective in moderate (e.g., 10) dimensional problems.

Extension of the *Laplace approximation* method of analytically approximating $E^*[g(\theta)]$, leading to a reasonably accurate general technique, was carried out in Tierney et al. (1989). The chief limitations of the method are the need for analytic derivatives, the need to redo parts of the analysis for each different $g(\theta)$, and the lack of an estimate of the error of the approximation. For many problems, however, the technique is remarkably successful.

Monte Carlo importance sampling [see Geweke (1989) and Wolpert (1991) for discussion] has been the most commonly used traditional method of computing $E^*[g(\theta)]$. The method can work in very large dimensions, and carries with it a fairly reliable accuracy measure. Although one of the oldest computational devices, it is still one of the best, being nearly 'optimal' in many problems. It does require determination of a good 'importance function', however, and this can be a difficult task. Current research continues to address the problem of choosing a good importance function; for instance, Oh and Berger (1993) developed a method of selecting an importance function for a multimodal posterior.

5.3 Markov chain simulation techniques

The newest techniques to be extensively utilized for numerical Bayesian computations are Markov chain simulation techniques, including the popular Gibbs Sampling. [Certain of these techniques are actually quite old — see, e.g., Hastings (1970); it is their application and adoption to Bayesian problems that is new.] A brief generic description of these methods is as follows:

Step 1. Select a 'suitable' Markov chain on the parameter space Θ, with $p(\cdot,\cdot)$ being the transition probability density (i.e., $p(\theta,\theta^*)$ gives the transition density for movement of the chain from θ to θ^*). Here 'suitable' means primarily that the posterior distribution of θ given the data x, $\pi(\theta|x)$, is a stationary distribution of the Markov chain, which can be assured in a number of ways.

Step 2. Starting at a point $\theta^{(0)} \in \Theta$, generate a sequence of points $\theta^{(1)}, \theta^{(2)}$, $\ldots, \theta^{(m)}$ from the chain.

Step 3. Then, for large m, $\theta^{(m)}$ is (approximately) distributed as $\pi(\theta|x)$ and

$$\frac{1}{m}\sum_{i=1}^{m} g(\theta^{(i)}) \cong E^*[g(\theta)]. \tag{1.7}$$

The main strengths of Markov chain methods for computing $E^*[g(\theta)]$ are:

(1) Many different g can simultaneously be handled via Step 3, once the sequence $\theta^{(1)}, \ldots, \theta^{(m)}$ has been generated.

(2) Programming tends to be comparatively simple.

(3) Methods of assessing convergence and accuracy exist and/or are being developed.

The main weaknesses of the Markov chain methods are:

(1) They can be quite slow. It is not uncommon in complicated problems to need m to be in the hundreds of thousands, requiring millions of random variable generations if the dimension of θ is appreciable.

(2) One can be misled into prematurely judging that convergence has obtained.

The more common Markov chain methods, corresponding to different choices of $p(\cdot,\cdot)$, will briefly be discussed. A recent general guide to these methods, and their use in practice, is Gelman, et. al. (1995). See also Smith (1991) and Besag, et. al. (1995).

Metropolis-Hastings algorithm: One generates a new θ^* based on a 'probing' distribution, and then moves to the new θ^* or stays at the old θ according to a certain 'accept-reject' probabilities, see Hastings (1970).

Gibbs sampling: The Markov chain moves from $\theta^{(i)}$ to $\theta^{(i+1)}$ one coordinate at a time (or one group of coordinates at a time), the transition density being the conditional posterior density of the coordinate(s) being moved given the other coordinates. This is a particularly attractive procedure in many Bayesian scenarios, such as analysis of hierarchical models, because the conditional posterior density of one parameter given the others is often relatively simple (or can be made so with the introduction of auxiliary variables). Extensive discussion and illustration of Gibbs sampling can be found in Gelfand and Smith (1990), Gelman and Rubin (1992), Raftery (1992), and Smith and Gelfand (1992). We confine ourselves here to a very elementary example, but one which illustrates the basic technique.

Example 3. The following posterior density is a very simplified version of posterior densities which occur commonly in Bayesian analysis, and which are particularly amenable to Gibbs sampling. Suppose the posterior density is

$$\pi(\theta_1, \theta_2|\text{data}) = \frac{1}{\pi}\exp\{-\theta_1(1+\theta_2^2)\} \tag{1.8}$$

on the domain $\theta_1 > 0$, $-\infty < \theta_2 < \infty$. Many posterior expectations involving this density cannot be done in closed form. Gibbs sampling, however, can easily be applied to this distribution to compute all integrals of interest.

Note, first, that the conditional distribution of θ_2, given θ_1, is Normal with mean zero and variance $1/2\theta_1$; and, given θ_2, θ_1 has an exponential distribution with mean $1/(1+\theta_2^2)$. Hence the Gibbs sampling algorithm can be given as follows:

Step 0.	Choose an initial value for θ_2; for instance, the maximizer of the posterior, $\theta_2^{(0)} = 0$.
Step i(a).	Generate $\theta_1^{(i)} = \mathcal{E}/(1+[\theta_2^{(i-1)}]^2$, where $\mathcal{E}$ is a standard exponential random variable.
Step i(b).	Generate $\theta_2^{(i)} = Z/\sqrt{2\theta_1^{(i)}}$, where Z is a standard normal random variable.
Repeat	Steps i(a) and i(b) for $i = 1, 2, \ldots, m$.
Final Step.	Approximate the posterior expectation of $g(\theta_1, \theta_2)$ by

$$\begin{aligned} E[g(\theta_1, \theta_2)] &= \int_{-\infty}^{\infty}\int_{1}^{\infty} g(\theta_1,\theta_2)\pi(\theta_1,\theta_2|\text{data})d\theta_1 d\theta_2 \\ &\cong \frac{1}{m}\sum_{i=1}^{m} g(\theta_1^{(i)}, \theta_2^{(i)}). \end{aligned} \tag{1.9}$$

For instance, the typical estimate of θ_1 would be its posterior mean, approximated by $\hat{\theta}_1 = (1/m)\sum_{i=1}^{m}\theta_1^{(i)}$. Table 1.5 presents the results of this computation for various values of m. Note that the true posterior mean here is 0.5.

TABLE 1.5. Approximate values of posterior mean of θ_1 from Gibbs Sampling.

m	100	500	1,000	10,000	50,000
$\hat{\theta}_1$	0.43761	0.53243	0.48690	0.49857	0.50002

Hit and run sampling: The idea here is roughly that one moves from $\theta^{(i)}$ to $\theta^{(i+1)}$ by choosing a random direction and then moving in that direction according to the appropriate conditional posterior distribution. This method is particularly useful when Θ is a sharply constrained parameter space. Extensive discussion and illustration can be found in Belisle et al. (1993), and Chen and Schmeiser (1993).

Hybrid methods: Complex problems will typically require a mixture of the above (and other) methods. Here is an example, from Müller (1991), the purpose of which is to do Gibbs sampling when the posterior conditionals [e.g., $\pi(\theta_i|x$, other $\theta_k)$] are not 'nice'.

Step 1. Each step of the Markov chain will either
- generate $\theta_j^{(i)}$ from $\pi(\theta_j|x$, other $\theta_k^{(i)})$ if the conditional posterior is 'nice' or
- generate $\theta_j^{(i)}$ by employing one or several steps of the Metropolis-Hastings algorithm if the conditional is not nice.

Step 2. For the probing function in the Metropolis-Hastings algorithm, use the relevant conditional distribution from a global multivariate normal (or t) importance function, as typically developed in Monte Carlo importance sampling.

Step 3. Adaptively update the importance function periodically, using estimated posterior means and covariance matrices.

Other discussions or instances of use of hybrid methods include Geyer (1992, 1995), Gilks and Wild (1992), Tanner (1991), Smith and Roberts (1993), Berger and Chen (1993), and Tierney (1994).

5.4 *Software existence and development*

Availability of general user-friendly Bayesian software would rapidly advance use of Bayesian methods. A number of software packages do exist, and are very useful for particular scenarios. An example is BATS (cf., Pole, West, and Harrison, 1994, and West and Harrison, 1989), which is designed for Bayesian time series analysis. A listing and description of pre-1990 Bayesian software can be found in Goel (1988) and Press (1989).

Four recent software developments are BAIES, a Bayesian expert system (see Cowell, 1992); [B/D], an 'expectation based' subjective Bayesian system (see Wooff 1992); BUGS, designed to analyze general hierarchical models via Gibbs sampling (see Thomas et. al. 1992); and XLISP-STAT, a general system with excellent interactive and graphics facilities, but limited computational power (see Tierney 1990). Numerous other Bayesian software developments are currently underway.

Two of the major strengths of the Bayesian approach create certain difficulties in developing generic software. One is the extreme flexibility of Bayesian analysis, with virtually any constructed model being amenable to analysis. Classical packages need contend with only a few well-defined models or scenarios for which a classical procedure has been determined. Another strength of Bayesian analysis is the possibility of extensive utilization of subjective prior information, and Bayesians tend to feel that software should include an elaborate expert system for prior elicitation. This is hard, in part because much remains to be done empirically to determine optimal ways to elicit priors. Note that such an expert system is not, by any means, a strict need for Bayesian software; it is possible to base a system on use of noninformative priors.

6 Conclusions

Papers on Bayesian analysis frequently tout the advantages of Bayesian over classical methods, and this paper has been no exception. In a sense, this is unavoidable since, for a scientist to try Bayesian methods, a considerable retooling and investment of effort may be required, and the case must be made that this effort is worthwhile. At the same time, criticism of classical statistics is rather unfortunate, because Bayesian statistics and classical statistics share a great deal in common, and have much the same aims. Indeed, the two schools of statistics have been drawing closer together of late, so much so that one can envisage at least a philosophical unification sometime in the near future.

As an example of this, consider the situation of Bayesian testing, as illustrated by Example 1. We were quite critical of the use of P-values in that example, but P-values are also criticized by classical frequentist statisticians, in part because they are not true frequentist procedures having an interpretation in terms of a long-run error rate. In a recent surprising development (based on ideas of Kiefer, 1977), Berger, Brown, and Wolpert (1995) and Berger, Boukai, and Wang (1994) show for simple versus simple testing and for testing a precise hypothesis, respectively, that Bayesian tests (with, say, equal prior probabilities of the hypotheses) yield posterior probabilities which have direct interpretations as conditional frequentist error probabilities. In Example 1, for instance, P_1 in (1.3) can be interpreted as the conditional Type I frequentist error probability, and P_2 can be interpreted as an average conditional Type II error probability. Note that the reported error probabilities thus vary with the data, in contrast with the usual Neyman-Pearson error probabilities. Also, use of these conditional error probabilities is arguably greatly superior to use of the usual Neyman-Pearson error probabilities, even from the frequentist perspective.

The necessary technical detail to make this work is the defining of suitable conditioning sets upon which to compute the conditional error probabilities. These sets necessarily include data in both the acceptance and the rejection regions, and can roughly be described as the sets which include data points providing equivalent strength of evidence for and against H_1. Note that computation of these sets is not necessary for practical implementation of the procedures.

The primary limitation of this Bayesian - frequentist equivalence is that there will typically be a region, which is called the no-decision region, in which frequentist and Bayesian interpretations are incompatible. Hence this region is excluded from the decision space. In Example 1, for instance, and if $n = 20$, then the no-decision region is the set of all points where the usual z-statistic ($\sqrt{n}|\overline{x}|$) is between 1.18 and 1.95. In all examples we have studied, the no-decision region is similarly a region where both frequentists and Bayesian would feel indecisive, and hence its presence in the procedure is not detrimental from a practical perspective.

While a philosophical reconciliation of the statistical schools appears to be within the realm of possibility, the ease of interpretation of Bayesian answers and the comparative simplicity in implementing the (default) Bayesian techniques will still argue in favor of their use.

Acknowledgments: This research was supported by the National Science Foundation, under Grant DMS - 9303556. Chimei Shui and Dejun Tang were instrumental in carrying out the computations.

7 References

[1] Bayes, T. (1783). An essay towards solving a problem in the doctrine of chances. *Phil. Trans. Roy. Soc.*, **53**, 370-418.

[2] Belisle, C., Romeijn, H. E. and Smith, R. (1993). Hit-and-run algorithms for generating multivariate distributions. *Mathematics of Operation Research*, **18**, 255–266.

[3] Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis* (2nd edition). Springer-Verlag, NY.

[4] Berger, J. (1994). An overview of robust Bayesian analysis. *Test*, **3**, 5–124.

[5] Berger, J. and Bernardo, J. (1992). On the development of the reference prior method. In J. Bernardo, J. Berger, A. Dawid and A. F.

M. Smith (editors), *Bayesian Statistics*, **4**, Oxford University Press, London.

[6] Berger, J. and Berry, D. (1988). Analyzing data: Is objectivity possible? *American Scientist*, **76**, 159-165.

[7] Berger, J., Brown, L. and Wolpert, R. (1994). A unified conditional frequentist and Bayesian test for fixed and sequential hypothesis testing. *Annals of Statistics*, **22**, 1787–1807.

[8] Berger, J., Boukai, B., and Wang, Y. (1994). Unified frequentist and Bayesian testing of a precise hypothesis. Technical Report 94-25C, Purdue University, West Lafayette.

[9] Berger, J. and Chen, M. H. (1993). Determining retirement patterns: prediction for a multinomial distribution with constrained parameter space. *The Statistician*, **42**, 427–443.

[10] Berger, J. and Delampady, M. (1987). Testing precise hypotheses (with discussion). *Statist. Science*, **2**, 317–352.

[11] Berger, J., and Pericchi, L. R. (1996a). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, **91**, 109-122.

[12] Berger, J., and Pericchi, L. R. (1996b). The intrinsic Bayes factor for linear models. *Bayesian Statistics*, **5**. J. M. Bernardo, et. al. (eds.), pp. 23-42, Oxford University Press, London.

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

[14] Besag, J., Green, P., Higdon, D., and Mengersen, K. (1995). Bayesian computation and stochastic systems. *Statistical Science*, **10**, 1-58.

[15] Chen, M. H. and Schmeiser, B. (1993). Performance of the Gibbs, hit-and-run, and Metropolis samplers. *Journal of Computational and Graphical Statistics*, **2**, 1–22.

[16] Delampady, M. and Berger, J. (1990). Lower bounds on posterior probabilities for multinomial and chi-squared tests. *Annals of Statistics*, **18**, 1295–1316.

[17] Draper, D. (1995). Assessment and propogation of model uncertainty. *J. Roy. Statist. Soc. B*, **57**, 45–98.

[18] Cowell, R. G. (1992). BAIES: A probabilistic expert system shell with qualitative and quantitative learning. In: *Bayesian Statistics*, **4** (J. Bernardo, J. Berger, A. Dawid and A. F. M. Smith, Eds.). Oxford University Press, Oxford.

[19] Edwards, W., Lindman, H. and Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, **70**, 193-242.

[20] Gelfand, A. E. and Smith, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.*, **85**, 398–409.

[21] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian Data Analysis*. Chapman and Hall, London.

[22] Gelman, A. and Rubin, D. B. (1992). On the routine use of Markov Chains for simulation. In J. Bernardo, J. Berger, A. Dawid, and A. F. M. Smith (editors), *Bayesian Statistics*, **4**, Oxford University Press, London.

[23] Geweke, J. (1989). Bayesian inference in econometrics models using Monte Carlo integration. *Econometrica*, **57**, 1317–1340.

[24] Geyer, C. (1992). Practical Markov chain Monte Carlo. *Statistical Science*, **7**, 473–483.

[25] Geyer, C. (1995). Conditioning in Markov Chain Monte Carlo. *J. Comput. Graph. Statist.*, **4**, 148-154.

[26] Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. In J. Bernardo, J. Berger, A. Dawid, and A. F. M. Smith (editors), *Bayesian Statistics*, **4**, Oxford University Press, London.

[27] Goel, P. (1988). Software for Bayesian analysis: current status and additional needs. In: *Bayesian Statistics*, **3**, J. M. Bernardo, M. DeGroot, D. Lindley and A. Smith, (Eds.). Oxford University Press, Oxford.

[28] Hastings, W. K. (1970). Monte-Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

[29] Hurvich, C. M. and Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika*, **76**, 297-307.

[30] Jeffreys, H. (1961). *Theory of Probability* (3rd edition), Oxford University Press, London.

[31] Jefferys, W. and Berger, J. (1992). Ockham's razor and Bayesian analysis. *American Scientist*, **80**, 64–72.

[32] Kass, R. and Raftery, A. (1995). Bayes factors and model uncertainty. *J. Amer. Statist. Assoc.*, **90**, 773–795.

[33] Kass, R. E., and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, **90**, 928-934.

[34] Kiefer, J. (1977). Conditional confidence statements and confidence estimators. *Journal of the American Statistical Association*, **72**, 789–827.

[35] Laplace, P. S. (1812). *Theorie Analytique des Probabilites.* Courcier, Paris.

[36] Lavine, M. and West, M. (1992). A Bayesian method for classification and discrimination. *Canadian J. of Statistics*, **20**, 421-461.

[37] Loredo, T. (1992). Promise of Bayesian inference for astophysics. In: *Statistical Challenges in Modern Astronomy*, E. Feigelson and G. J. Babu (Eds.). Springer-Verlag, New York.

[38] Naylor, J. and Smith, A. F. M. (1982). Application of a method for the efficient computation of posterior distributions. *Appl. Statist.*, **31**, 214–225.

[39] O'Hagan, A. (1995). Fractional Bayes factors for model comparisons. *J. Roy. Statist. Soc. B*, **57**, 99–138.

[40] Oh, M. S. and Berger, J. (1993). Integration of multimodal functions by Monte Carlo importance sampling. *J. Amer. Statist. Assoc.*, **88**, 450–456.

[41] Raftery, A. (1992). How many iterations in the Gibbs sampler? In J. Bernardo, J. Berger, A. P. Dawid, and A. F. M. Smith (editors), *Bayesian Statistics 4*, Oxford University Press.

[42] Ripley, B. D. (1992). Bayesian methods of deconvolution and shape classification. In: *Statistical Challenges in Modern Astronomy*, E. Feigelson and G. J. Babu (Eds.). Springer-Verlag, New York.

[43] Shui, C. (1996). Default Bayesian Analysis of Mixture Models. Ph.D. Thesis, Purdue University.

[44] Smith, A. (1991). Bayesian computational methods. *Phil. Trans. Roy. Soc.*, **337**, 369-386.

[45] Smith, A. F. M. and Gelfand, A. E. (1992). Bayesian statistics without tears: a sampling-resampling perspective. *American Statistician*, **46**, 84–88.

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

[47] Tanner, M. A. (1991). *Tools for Statistical Inference: Observed Data and Data Augmentation Methods*, Lecture Notes in Statistics **67**, Springer Verlag, New York.

[48] Thomas, A., Spiegelhalter, D. J. and Gilks, W. (1992). BUGS: A program to perform Bayesian inference using Gibbs sampling. In: *Bayesian Statistics*, **4** (J. Bernardo, J. Berger, A. Dawid and A. F. M. Smith, Eds.). Oxford University Press, Oxford.

[49] Tierney, L. (1994). Markov chains for exploring posterior distributions. *Ann. Statist.*, **22**, 1701-1762.

[50] Tierney, L. (1990). *Lisp-Stat, an Object-Oriented Environment for Statistical Computing and Dynamic Graphics.* Wiley, New York.

[51] Tierney, L., Kass, R. and Kadane, J. (1989). Fully exponential Laplace approximations to expectations and variances of non-positive functions. *J. Amer. Statist. Assoc.*, **84**, 710–716.

[52] Varshavsky, J. (1996). Intrinsic Bayes factors for model selection with autoregressive data. To appear in J. Bernardo et. al. (editors), *Bayesian Statistics*, **5**, Oxford University Press, London.

[53] Wolpert, R. L. (1991). Monte Carlo importance sampling in Bayesian statistics. In: *Statistical Multiple Integration* (N. Flournoy and R. Tsutakawa, Eds.). *Contemporary Mathematics*, Vol. 115.

[54] Wooff, D. A. (1992). [B/D] works. In: *Bayesian Statistics*, **4** (J. Bernardo, J. Berger, A. Dawid and A. F. M. Smith, Eds.). Oxford University Press, Oxford.

[55] Yang, R. and Berger, J. (1996). A catalogue of noninformative priors. Technical Report, Purdue University.