Applications of formal model choice to archaeological chronology building

Sujit K. Sahu,
Faculty of Mathematical Studies,
University of Southampton
S.K.Sahu@maths.soton.ac.uk

January 9, 2003

SUMMARY

Statistical methods are now an essential part of the archaeological inference making process. Nowhere is this more important than in the analysis and interpretation of chronological data, especially when information from several sources must be drawn together. Different statistical models may, however, provide widely different interpretations of the same data. Thus it is often possible to make conflicting re-constructions of archaeological past using different models.

Bayesian predictive model choice criteria can be used as possible solutions to this problem. A particular advantage of Bayesian techniques lies in their ability to compare widely different models based on different assumptions and prior information. In this paper, we discuss recent developments in applying formal model choice techniques in archaeological chronology building. We illustrate the methods with two examples one each from the absolute and the relative chronology building problems.

KEY WORDS: BAYESIAN METHODS, BAYES FACTOR, CORRESPONDENCE ANALYSIS, DATING METHODS, MCMC, PREDICTIVE INFERENCE, RADIOCARBON DATING, SERIATION.

1 Introduction

Statistical methods are now an essential part of the archaeological inference making process as illustrated in the books by Shennan (1988), Baxter (1994) and Buck *et al.* (1996). The statistical techniques we discuss here are based on the Bayesian paradigm which provides a natural and convenient way to incorporate prior information in practical problems. Although some authors criticize the Bayesian view (see for example Reece, 1994), it is regarded as *the* most general and coherent statistical inference procedure capable of solving practical problems. Bayesian methods have many advantages, see e.g.

the book by Buck *et al.* (1996). One particular advantage lies in their ability to compare widely different models based on different assumptions and prior information.

Prior information, although quite valuable, cannot build chronologies for *sure*. Often such information comes from expert archaeologists working on particular problems of interpretation of archaeological data. However, experts often disagree and they may provide different archaeological dates or explanations of the data. Moreover, the adopted statistical models are also liable to be uncertain. An evaluation of model uncertainties is required before making final inference. Thus in practical problems both the prior information and the assumed statistical models need to be thoroughly examined as part of the model evaluation process. The Bayesian methods that we are going to describe allow us to address and measure the uncertainties arising due to the possible mis-specification of the statistical model for the data and the assumed prior distribution for the parameters.

Archaeologists often seek two types of chronological evidence: absolute and relative. Absolute techniques provide estimates of the true calendar date of archaeological events. Relative techniques, on the other hand, simply allow estimates of the chronological order in which events took place. Of course, if absolute dates were available for all events of interest, relative dating would not be needed. Typically, however, this is not the case and ways are sought to combine both relative and absolute chronological information in order to enhance temporal understanding.

Nicholls and Jones (2001) consider two alternative prior distributions for the boundary parameters dividing the excavated layers for the purposes of absolute dating. Using one set of prior assumptions they obtain a much tighter posterior distribution for the span of the absolute dates. The span is the difference in age between the most recent layer and the deepest layer containing the oldest material. The Bayesian model choice methods help decide between the two prior models giving rise to two completely different posterior distributions. The interpretations obtained from two different posterior distributions are quite different and hence one must choose a model from the two alternatives considered.

On some archaeological excavations there are no reliable relationships between vertical location in the ground and relative date of deposition of the artefacts found. In some others, as in the following example, it is not possible to link excavated layers in one area with those in another. Statistical methodologies based upon the artefacts excavated are sometimes employed in an attempt to derive relative chronological information. Such methodologies, which identify temporal sequence on the basis of the number of different types of artefacts, are commonly referred to as *seriation* techniques.

Buck and Sahu (2000) consider seriation of a data set relating to the numbers of seven types of mesolithic flint tools (known as microliths) from six different sites, numbered $1, \ldots, 6$, in southern England. The objective is to identify the relative chronological order of the sites by studying the changes in the numbers of the seven types of microliths found at them. Two widely used competing methodologies suggest completely different orders for the sites: 2, 5, 3, 6, 1, 4 and 3, 6, 5, 2, 1, 4. Clearly these are likely to give rise to quite different archaeological conclusions. The problem in focus here is to choose between the two using model choice methods.

The remainder of this article is organized as follows. Section 2 provides the *model choice* framework within which statistical solutions are proposed. The model choice criteria are illustrated using a simple theoretical example in Section 3. Further, a radiocarbon dating example is provided in Section 4 and a well known example on relative chronology building is discussed in Section 5. Finally, few summary remarks are made in Section 6.

2 Bayesian methods

2.1 MCMC model fitting

Currently Markov chain Monte Carlo (MCMC) simulation techniques are used in a wide variety of statistical problems with relative ease and great success. These methods allow critical re-examination of existing model based approaches and are flexible enough to posit and develop more realistic models.

Let $\mathbf{y}$ denote the observed data to be modeled and let $\boldsymbol{\zeta}$ denote the unknown parameters in the model. Let $\pi(\boldsymbol{\zeta}|\mathbf{y})$ denote the posterior distribution of the parameters $\boldsymbol{\zeta}$ under the assumed Bayesian model. In order to implement the MCMC method known as Gibbs sampler (Gelfand and Smith, 1990) one writes down the complete conditional posterior distribution of all the parameters. These distributions have densities which are all proportional to the joint posterior density $\pi(\boldsymbol{\zeta}|\mathbf{y})$. The Gibbs sampler then simulates from each conditional distribution in turn for a large number of times, B say, starting from an arbitrary point. For large values of B, the effect of the starting point is forgotten and one obtains random samples from the joint posterior distribution. Features of the posterior distribution are then estimated accurately using appropriate averages of samples so obtained. For a general introduction to MCMC methods see the book by Gilks *et al.* (1996).

Due to the complexity of archaeological problems, however, many authors have shown that some ingenuity is needed in devising sampling schemes. Buck and Sahu (2000), for example, document several different attempts at implementation before a successful sampling scheme was devised. Once efficient algorithms for fitting statistical models for large and complex archaeological data sets have been implemented, we move to check the validity and adequacy of the fitted models. We propose to use predictive Bayesian model choice techniques both to facilitate model comparison and assess goodness-of-fit. Bayesian model checking serves the latter purpose and is important because in model selection we run the risk of selecting from a set of badly fitting alternatives.

2.2 Predictive Distributions

Bayesian model choice methods are based on Bayesian predictive distributions. In simple terms, these are distributions of future replicate data sets obtained by eliminating the parameter uncertainties. Different types of predictive distributions arise by considering different methods of eliminating the uncertain parameters. We list a few predictive densities below. Let $\mathbf{y}_{\text{obs}}$ denote the observed data with individual data points $y_{r,\text{obs}}, r = 1, \ldots, n$, and $\mathbf{y}_{\text{rep}}$ with components $y_{r,\text{rep}}$ (abbreviation for replicate) denote a future set of observables under the assumed model.

The *prior predictive density* of a set of observations at the actual observed point $\mathbf{y}_{\text{obs}}$ is given by

$$\pi(\mathbf{y}_{\text{obs}}) = \int \pi(\mathbf{y}_{\text{obs}}|\boldsymbol{\zeta})\ \pi(\boldsymbol{\zeta})\ d\boldsymbol{\zeta}. \tag{1}$$

In the Bayesian inference setup the actual observations $\mathbf{y}_{\text{obs}}$ is fixed, the above is interpreted as the density of a set of observables evaluated at the observed point $\mathbf{y}_{\text{obs}}$. This is also known as the *marginal likelihood* of the data. The prior predictive density is only meaningful if the prior distribution $\pi(\boldsymbol{\zeta})$ is a proper distribution (i.e, $\int \pi(\boldsymbol{\zeta})d\boldsymbol{\zeta} = 1$), due to its involvement in the definition (1).

Let $\mathbf{y}_{(r),\text{obs}}$ denote the set of observations $\mathbf{y}_{\text{obs}}$ with rth component deleted. The *cross-validation predictive density* is defined by:

$$\pi(y_r|\mathbf{y}_{(r),\text{obs}}) = \int \pi(y_r|\boldsymbol{\zeta}, \mathbf{y}_{(r),\text{obs}})\ \pi(\boldsymbol{\zeta}|\mathbf{y}_{(r),\text{obs}})\ d\boldsymbol{\zeta}. \tag{2}$$

In the case of conditionally independent observations given $\boldsymbol{\zeta}$,

$$\pi(y_r|\boldsymbol{\zeta}, \mathbf{y}_{(r),\text{obs}}) = \pi(y_r|\boldsymbol{\zeta}).$$

The predictive density (2) then simplifies to

$$\pi(y_r|\mathbf{y}_{(r),\text{obs}}) = \int \pi(y_r|\boldsymbol{\zeta})\ \pi(\boldsymbol{\zeta}|\mathbf{y}_{(r),\text{obs}})\ d\boldsymbol{\zeta}. \tag{3}$$

This density is also known as the *conditional predictive ordinate* (CPO). These densities are meaningful even when improper prior distributions for $\boldsymbol{\zeta}$ are considered as long as the posterior distribution $\pi(\boldsymbol{\zeta}|\mathbf{y}_{(r),\text{obs}})$ is proper for each r.

The *posterior predictive density* of $\mathbf{y}_{\text{rep}}$, given by

$$\pi(\mathbf{y}_{\text{rep}}|\mathbf{y}_{\text{obs}}) = \int \pi(\mathbf{y}_{\text{rep}}|\boldsymbol{\zeta})\ \pi(\boldsymbol{\zeta}|\mathbf{y}_{\text{obs}})\ d\boldsymbol{\zeta}, \tag{4}$$

is the predictive density of a new independent set of observables, $\mathbf{y}_{\text{rep}}$ under the model, given the actual data $\mathbf{y}_{\text{obs}}$. The posterior predictive density is easier to work with than the previous two densities, because features of $\mathbf{y}_{\text{rep}}$ having density (4) can be estimated easily when MCMC samples from the posterior $\pi(\boldsymbol{\zeta}|\mathbf{y}_{\text{obs}})$ are available. A new set of observations drawn from $\pi(\mathbf{y}_{\text{rep}}|\boldsymbol{\zeta})$, the likelihood model conditional on $\boldsymbol{\zeta}$, is a sample from the predictive density (4).

2.3 The Bayes factor

A pure Bayesian approach to model selection is to report posterior probabilities of each model by comparing Bayes factors, see for example DiCiccio *et al.* (1997) and Kass and Raftery (1995). The *Bayes factor* for comparing two given models M_1 and M_2 is

$$\text{BF} = \frac{\pi(\mathbf{y}_{\text{obs}}|M_1)}{\pi(\mathbf{y}_{\text{obs}}|M_2)},$$

where $\pi(\mathbf{y}_{\text{obs}}|M_i)$ is the density (1) when M_i is the assumed model, $i = 1, 2$.

The BF gives a summary of the evidence for M_1 against M_2 provided by the data. Calibration tables for the BF are available for deciding how strong is the evidence, see e.g. Kass and Raftery (1995). Recall that $\pi(\mathbf{y}_{\text{obs}}|M_i)$ is the marginal likelihood of the data under model M_i. Hence the BF chooses a model for which the marginal likelihood of the data is maximum.

For improper priors the Bayes factor is not meaningful since it cannot be calibrated. This is because the predictive density (1) is improper when $\pi(\boldsymbol{\zeta})$ is. To overcome this problem of interpretation O'Hagan (1995) proposed the fractional Bayes factor by considering training samples. This idea has been investigated further, see for example the article by Key *et al.* (1999).

The cross-validation predictive densities are used to form a variant of the Bayes factor called the *pseudo-Bayes factor* (PsBF) (Geisser and Eddy, 1979). For comparing two models M_1 and M_2 the PsBF is defined as,

$$\text{PsBF} = \prod_{r=1}^{n} \frac{\pi(y_{r,\text{obs}}|\mathbf{y}_{(r),\text{obs}}, M_1)}{\pi(y_{r,\text{obs}}|\mathbf{y}_{(r),\text{obs}}, M_2)}.$$

This is a surrogate for the Bayes factor and its interpretations are similar, see e.g. Gelfand (1996). The CPOs are also useful for checking model adequacy. Instead of using a single summary measure alone, e.g. the PsBF, the individual CPOs can also be compared under any two models. This is to guard against any single highly influential observation concealing a general trend. One observation, $y_{r,\text{obs}}$, prefers model M_1 to M_2 if the rth CPO is higher under M_1. The CPOs are not illustrated in this paper since the primary issue here is model choice and not model checking.

2.4 A decision theoretic approach

Gelfand and Ghosh (1998) and Laud and Ibrahim (1995) propose model selection criteria based on the posterior predictive densities. The current model is a 'good' fit to the observed data, $\mathbf{y}_{\text{obs}}$, if $\mathbf{y}_{\text{rep}}$ is able to replicate the data well. Hence, many model choice criteria can be developed by considering different loss functions for measuring the divergence between $\mathbf{y}_{\text{obs}}$ and $\mathbf{y}_{\text{rep}}$ (see for example Rubin, 1984). If the data are assumed to be symmetrically distributed with a common variance then it is natural to adopt a squared error loss function

$$L(\mathbf{y}_{\text{rep}},\ \mathbf{y}_{\text{obs}}) = \sum_r (y_{r,\text{rep}} - y_{r,\text{obs}})^2. \tag{5}$$

In the unequal variance cases one may weight the individual terms in the loss function by the inverse variance of $y_{r,\text{obs}}$ if it is known. Other loss functions are also possible, for example Buck and Sahu (2000) use the following deviance loss function

$$L(\mathbf{y}_{\text{rep}},\ \mathbf{y}_{\text{obs}}) = 2\left(\sum_r y_{r,\text{obs}} \log \frac{y_{r,\text{obs}}}{y_{r,\text{rep}}}\right) \tag{6}$$

where the data are assumed to follow the multinomial distribution. The best model among a given set of models is the model for which the expected value of the adopted loss function is the minimum, where the expectation is to be taken with respect to the posterior predictive distribution (4).

2.5 The *DIC*

There are many other Bayesian methods available for model comparison. These methods use the posterior distribution of the likelihood to arrive at suitable model choice criteria. For example, Aitkin (1997) interprets the p-values by using the posterior distribution of the likelihood function.

Recently, Spiegelhalter *et al.* (2002) propose a model selection criterion for arbitrarily complex models called the deviance information criterion (*DIC*). They first define the deviance function as follows:

$$D(\boldsymbol{\zeta}) = -2\,\log\{\pi(\mathbf{y}|\boldsymbol{\zeta})\} + 2\,\log\{\pi(\mathbf{y})\}$$

where $\pi(\mathbf{y}|\boldsymbol{\zeta})$ is the likelihood function, and $\pi(\mathbf{y}) = \pi(\mathbf{y}|\mu(\boldsymbol{\zeta}) = \mathbf{y})$. Here $\mu(\boldsymbol{\zeta})$ is defined as the mean of the data, that is $\mu(\boldsymbol{\zeta}) = E(\mathbf{Y}|\boldsymbol{\zeta})$.

They define the penalty factor as

$$p_D = E\{D(\boldsymbol{\zeta})|\mathbf{y}\} - D\{E(\boldsymbol{\zeta}|\mathbf{y})\}.$$

Thus p_D is the expected deviance minus the deviance evaluated at the posterior expectations. The p_D is called the *effective number of parameters* in a complex model. Subsequently, they define the model choice criterion

$$DIC = D\{E(\boldsymbol{\zeta}|\mathbf{y})\} + 2\,p_D.$$

The model with the smallest *DIC* is chosen to be the best model for data.

3 A simple example

We first consider a simple example which reveals the Bayesian model choice criteria in closed form analytic expressions. Suppose that $y_1, \ldots, y_n$ are observations from the $N(\theta, 1)$ population and the prior for θ is $N(0, \tau^2)$ where τ^2 is known and finite. In this example we have $\theta = \boldsymbol{\zeta}$. Consider the following two models.

$$M_1 : \theta = 0, \quad \text{vs} \quad M_2 : \theta \neq 0.$$

This is perhaps over-simplification, but the setup will aid understanding of the Bayesian model choice criteria.

The posterior distribution of θ is given by

$$\pi(\theta|\mathbf{y}) = N\left(\frac{n\bar{y}}{n + 1/\tau^2}, \frac{1}{n + 1/\tau^2}\right).$$

Thus the observations $y_1, \ldots, y_n$ enter into the posterior distribution through $\bar{y}$ and the model assumption for data is equivalent to

$$\bar{Y} = \frac{1}{n} \sum Y_i \sim N\left(\theta,\ \frac{1}{n}\right).$$

This is also due to the fact that $\bar{Y}$ is the sufficient statistic for θ.

Suppose that Z is a future observation for which we wish to calculate the predictive distribution. To have simpler notation we use the notation $Z = \mathbf{y}_{\text{rep}}$ and $\mathbf{y}_{\text{obs}} = \bar{y}$. If θ is known we have, $\pi(Z|\theta) = N\left(\theta,\ \frac{1}{n}\right)$. The predictive distributions (prior or posterior) of Z has two different forms under the two models, M_1 and M_2. Under model M_1 there are no unknown parameters and both the prior and posterior predictive distributions are given by $N\left(0, \frac{1}{n}\right)$.

The prior predictive distribution (1) of Z under model M_2 is given by,

$$\pi(z) = N\left(0,\ \frac{1}{n} + \tau^2\right).$$

As expected, if $\tau^2 = 0$ this distribution reduces to the prior predictive under M_1. The posterior predictive distribution is calculated as,

$$\pi(z|\bar{y}) = N\left(\frac{n\tau^2}{n\tau^2 + 1}\bar{y},\ \frac{1}{n} + \frac{\tau^2}{n\tau^2 + 1}\right).$$

As expected this posterior predictive distribution has less variability than the prior predictive distribution for non-zero vales of τ^2. This fact will be further discussed in Section 4. Moreover, the center of the distribution is located near the center of the data $\bar{y}$, unlike the prior predictive distribution which is centered at zero. We do not consider the cross-validation predictive densities because effectively there is only one data point $\bar{y}$.

Assume the loss function to be

$$L(z, \bar{y}) = (z - \bar{y})^2. \tag{7}$$

Now we derive the following decision rules based on the three predictive model selection criteria. Select model M_1 if

$$\begin{aligned} n\bar{y}^2 &< (1 + n\tau^2)\, \tfrac{\log(1+n\tau^2)}{n\tau^2}, \text{ using the Bayes factor,} \\ &< (1 + n\tau^2)\, \tfrac{1}{2+n\tau^2}, \text{ using the squared error loss function,(7)} \\ &< (1 + n\tau^2)\, \tfrac{2}{2+n\tau^2}, \text{ using the DIC.} \end{aligned}$$

It is straightforward to see that

$$\frac{\log(1 + n\tau^2)}{n\tau^2} \geq \frac{2}{2 + n\tau^2}, \quad \tau^2 \geq 0.$$

We interpret the above results as follows. If the loss function based approach selects model M_1 then the Bayes factor will select the same as well. The loss function based approach is likely to reject the simpler model M_1 more often than the Bayes factor based approach. The last two predictive criteria criticize the simpler model too much. They require the models to both fit and predict the data well. Thus the Bayes factor is seen to be less stringent regarding the choice of the simpler model than the remaining two model choice criteria. We shall extrapolate this theoretical result for the practical example in Section 5.

Layer	Sample	CRA	lab sd	Sample ID
1	1	580	47	NZ 7758
2	1	600	50	NZ 7761
3	1	537	44	NZ 7757
4	1	670	47	NZ 7756
5	1	646	47	NZ 7755
5	2	630	35	WK 2589
6	1	660	46	NZ 7771

4 Example: absolute chronology building

Often absolute chronologies are built using what is called the radiocarbon dating method. The radiocarbon dating laboratories provide the CRA (Conventional Radiocarbon Age) and an estimate of the associated error for a given sample from a dead organism. Statistical methods together with internationally agreed high precision calibration data are then used to convert the CRA to usable calendar dates. A full set of calibration data is available from `http://depts.washington.edu/qil/` (see also Stuiver *et al.*, 1998).

By way of an example, consider a set of seven CRA determinations which is a subset of a large set of dates gathered at the mouth of the Shag river, in southern New Zeland. The data set, given in Table 1, consists of all charcoal dates from a single series of six layers. Interest here focuses on the actual dates of deposition of the samples and the length of time for which the site was occupied.

We now discuss statistical formulation of the above problem as described by Nicholls and Jones (2001). Here the problem is to simultaneously calibrate several radiocarbon determinations found in a vertical series of a number of abutting layers of earth, I say. Suppose n_i radiocarbon age determinations are made in layer i, making $n = \sum_{i=1}^{I} n_i$ dates in all. Let Y_{ij} denote the value of the jth CRA measured in the ith layer and let θ_{ij} denote the corresponding true calendar date. Associated with θ_{ij} is a unique radiocarbon age, $\mu(\theta_{ij})$, which relates to the amount of ^{14}C present in the sample when it is measured. It is often assumed that:

$$Y_{ij} = \mu(\theta_{ij}) + \epsilon_{ij}^{(Y)} + \epsilon_{ij}^{(\mu)}, \tag{8}$$

where $\epsilon_{ij}^{(Y)}$ and $\epsilon_{ij}^{(\mu)}$ are independent normal random variables with zero means. Let the variance of $\epsilon_{ij}^{(Y)}$ be $\sigma_{ij}^{(Y)^2}$ and the variance of $\epsilon_{ij}^{(\mu)}$ given θ_{ij} be $\sigma^{(\mu)^2}(\theta_{ij})$. Thus, given θ_{ij}, the variance of Y_{ij} is $\sigma_{ij}^{(Y)^2} + \sigma^{(\mu)^2}(\theta_{ij})$. The quantities $\mu(\theta_{ij})$ and $\sigma^{(\mu)}(\theta_{ij})$ are obtained using piecewise linear functions of calibration data. There are other methods of determining the calibration functions $\mu(\theta)$ and $\sigma^{(\mu)}(\theta)$ using Gaussian process prior models, see for example Gomez-Portugal-Aguilar *et al.* (2002).

In our setup vertical mixing of earth is assumed to occur within layers, but not between layers. Let ψ_i denote the calendar date associated with the boundary between layers i and $i+1$, $i = 0, 1, \ldots, I$. Moreover, assume that layer $i = 1$ is the topmost and most recent layer, while layer $i = I$ is the deepest layer containing the oldest material. Let P and A $(P \leq A)$ denote the lower and upper bounds on the unknown parameters $\boldsymbol{\psi}$. In any layer i, the n_i calendar dates, $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{in_i})$ are all assumed to be in the interval (ψ_i, ψ_{i-1}). No other constraints are put on the calendar dates within a layer. Thus the model parameters satisfy the stratigraphic constraints:

$$P \leq \psi_I \leq \boldsymbol{\theta}_I \leq \psi_{I-1} \leq \boldsymbol{\theta}_{I-1} \leq \cdots \leq \psi_0 \leq A,$$

where the inequalities hold elementwise. Based on the constraints, and without any other more specific prior information, it is reasonable to assume independent uniform prior distributions for each component of $\boldsymbol{\theta}_i$ in the interval (ψ_i, ψ_{i-1}). Thus the prior distribution for $\boldsymbol{\theta}$ conditional on $\boldsymbol{\psi}$ is

$$\pi(\boldsymbol{\theta}|\boldsymbol{\psi}) = \prod_{i=1}^{I}\prod_{j=1}^{n_i}(\psi_{i-1} - \psi_i)^{-1} I(\psi_i \leq \theta_{ij} \leq \psi_{i-1}),$$

where $I(\cdot)$ is the indicator function. To complete the prior specification it remains to consider suitable prior distributions for the boundary parameters $\boldsymbol{\psi}$.

A much used prior distribution is the prior distribution which comes from ignorance on the relative positions of individual ψ_i. Thus one may assume that the unordered ψs follow the uniform distribution in the interval (P, A). The ordered samples in increasing order will then be taken as the ψ parameters. Suppose that U_0, $U_1, \ldots, U_I$ is a random sample from the uniform distribution in the interval (P, A), then we set $\psi_i = U_{(i)}$ where $U_{(I)} \leq U_{(I-1)} \leq U_{(0)}$. The associated prior distribution for $\boldsymbol{\psi}$ has the prior density:

$$\pi^{(1)}(\boldsymbol{\psi}) = \frac{(I+1)!}{R^{I+1}}, \quad P \leq \psi_I \leq \cdots \leq \psi_0 \leq A.$$

Nichols and Jones (2001) suggest an alternative prior distribution which they call the reference prior distribution. They assume that the span $\delta = \psi_0 - \psi_I$ follows the uniform distribution in the interval $(0, R)$ where $R = A - P$. Given δ, $\psi_I \sim U(P, A - \delta)$. This defines a joint prior distribution for the two endpoints, ψ_I and ψ_0. Given the two endpoints, the remaining $(I-1)$ unordered boundary parameters are assumed to follow the uniform distribution in the interval (ψ_I, ψ_0) independently. Suppose that $U_1, \ldots, U_{I-1}$ is a random sample from the uniform distribution in the interval (ψ_I, ψ_0), then we set, $\psi_i = U_{(i)}, i = 1, \ldots, I-1$ where $U_{(I-1)} \leq \cdots \leq U_{(1)}$. The prior density of $\boldsymbol{\psi}$ is

$$\pi^{(2)}(\boldsymbol{\psi}) = \frac{(I-1)!}{(\psi_0 - \psi_I)^{I-1}} \frac{1}{R(R - \psi_0 + \psi_M)}, \quad P \leq \psi_I \leq \cdots \leq \psi_0 \leq A.$$

There are fundamental differences between the two prior distributions $\pi^{(1)}$ and $\pi^{(2)}$. Under $\pi^{(1)}$ the distribution of the span δ is the distribution of $U_{(0)} - U_{(I)}$ where U_0, $U_1, \ldots, U_I$ are random samples from the uniform distribution in the interval (P, A). As a result the density of δ is the density of the sample range where the sample is obtained from the uniform distribution. The density of δ is given by

$$\pi^{(1)}(\delta) = \frac{I(I+1)}{R^{I+1}}\delta^{I-1}(R - \delta), \quad 0 < \delta < R.$$

That is, $\delta^* = \delta/R$ follows the standard beta distribution with parameters I and 2. However, under $\pi^{(2)}$, δ follows the uniform distribution in $(0, R)$, consequently δ^* follows

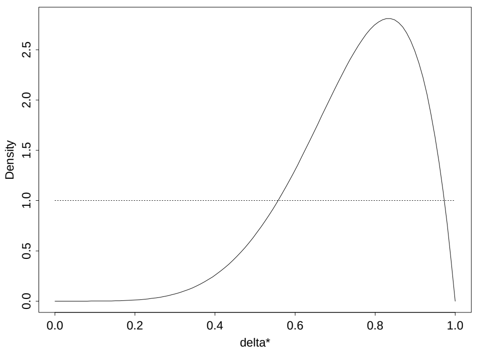

Figure 1: Prior densities for δ^*. Solid line is the density under $\pi^{(2)}$ and dotted line is the density under $\pi^{(1)}$.

the uniform distribution in $(0, 1)$. Nicholls and Jones (2001) point out that $\pi^{(1)}$ is more informative about δ and $\pi^{(2)}$ is not, see Figure 1.

The above authors formulate this problem as a Bayesian model choice problem and use the Bayes factor to decide between the prior distributions. Here we shall compare the two prior models using the predictive Bayesian methods discussed earlier.

4.1 Example 2: Model choice for absolute chronology

The Bayes factor for comparing the model 2 with prior $\pi^{(2)}$ against model 1 with prior $\pi^{(1)}$ has been reported to be 26 by Nicholls and Jones (2001). However, it is interesting to see what would have happened had we used the other model choice criteria described earlier.

In Table 1 we report all three model choice criteria for the two models. Both the expected loss criterion and the *DIC* criterion choose model 2 as well. For the non-linear models considered here we observed *negative* values of the penalty parameter p_D. The discussion paper by Spiegelhalter *et al.* (2002) explains why this can happen and it also suggests some possible remedies which we do not consider primarily because of the non-linear nature of the models. We choose to work with the overall *DIC* which is often used in comparing complex Bayesian models.

The differences between the two prior distributions are very strongly pronounced under the Bayes factor but not so strongly using either the expected loss or the *DIC*. Below we discuss the possible reasons for this.

The above phenomenon is explained by the fact that the Bayes factor uses the prior predictive distributions while the other two criteria use the posterior predictive distribution. The two different prior distributions induce different prior predictive distributions, hence the Bayes factor is large. However, the posterior predictive distributions under the two prior distributions are similar due to the fact that those have been smoothed by the knowledge of the data. Thus there are no big difference between the two posterior predictive distributions and as a result the difference between the expected losses is not

	Expected loss	*DIC*	BF
Model 1	13.97	4.51	1
Model 2	12.51	1.17	26

Table 1: Model choice for the Shag river data.

large. This is also confirmed by the insights gained in the simple theoretical example in Section 3. There it is seen that the posterior predictive distribution is smoother (i.e. has less variability) than the prior predictive distribution.

5 Example: relative chronology building

We return to the relative chronology building example mentioned in the introduction. Buck and Sahu (2000) have developed the loss function based model choice method for this example. Here we experiment with the other criteria as well.

5.1 The Robinson-Kendall model

Consider the following extension of the model originally proposed by Kendall (1971). Let y_{ij} denote the observed number of artefacts (for example, pottery or tools types) of type j ($j = 1, \ldots, J$) found in archaeological site, feature or context i ($i = 1, \ldots, I$). Let $N = \sum_{ij} y_{ij}$ denote the total number of artefacts. Also, let θ_{ij} denote the underlying proportion of artefact j available for deposition at i and let Θ denote the matrix with elements θ_{ij}. Let $\boldsymbol{\theta}$ denote the vector representation of Θ. Since θ_{ij}s are proportions it is assumed that $\theta_{ij} \geq 0$ and $\sum_i \sum_j \theta_{ij} = 1$. The problem then, is to estimate the true temporal order of the I rows which is a permutation of the indices $1, \ldots, I$. We represent this true permutation using $p(1)$, $p(2)$, $\ldots$, $p(I)$.

We assume that $\mathbf{y}$ has a multinomial distribution with parameters N and $\boldsymbol{\theta}$, that is, the probability of obtaining the observed configuration, y_{ij}, is given by

$$N! \prod_{ij} \theta_{ij}^{y_{ij}} / y_{ij}!. \tag{9}$$

A suitable way to represent prior information about θ_{ij} is to use the Dirichlet distribution, thus

$$\pi(\boldsymbol{\theta}) \propto \prod_{ij} \theta_{ij}^{\alpha_{ij}-1},$$

where $\alpha_{ij} > 0$ for all i and j. Note that we can use this to incorporate both informative and non-informative prior information quite successfully. For example, setting $\alpha_{11} = 5$ can be thought of as having 4 artefacts of type 1 in row 1. To specify non-informative prior information we simply set $\alpha_{ij} = 0.5$ for all i and j. The posterior distribution of $\boldsymbol{\theta}$ is given by

$$\pi(\boldsymbol{\theta}) \propto \prod_{ij} \theta_{ij}^{y_{ij}+\alpha_{ij}-1}. \tag{10}$$

We shall use this posterior distribution to make inference about the orders.

Suppose that the true chronological order is the given natural order, i.e. $p(1) = 1,\ p(2) = 2,\ \ldots,\ p(I) = I$. Then, for each j, the Robinson-Kendall (R-K) model assumes that there exist integers $1 \leq a_j \leq I$ such that:

$$\begin{array}{ll} \theta_{ij} \leq \theta_{i+1\,j} & \text{for } i = 1, \ldots, a_j - 1, \\ \theta_{i+1\,j} \leq \theta_{i\,j} & \text{for } i = a_j, \ldots, I - 1. \end{array} \tag{11}$$

Note that when a_j is either 1 or I only one set of inequalities in the above equations are required and the other set is redundant. A matrix Θ satisfying (11) is called a $Q-$matrix (for theoretical work on such matrices see for example, Kendall, 1971 and Laxton, 1976). In practice the true chronological order is unknown and one attempts to find an order $p(1), p(2), \ldots, p(I)$ such that Θ is a $Q-$matrix for a set of unknown integers $a_j, j = 1, \ldots, J$, where the matrix Θ is random and follows the posterior distribution (10).

The model in equation (11) is overly prescriptive for most real archaeological data since the strict, temporal, unimodal sequence assumed in the R-K model may be violated because of the nature of use and discard of objects in the past. To account for this type of violation consider the following extension. Suppose that the matrix Φ is a $Q-$matrix in the natural order and let $||\cdot||$ denote a suitable distance measure between two matrices Θ and Φ. For example, we may consider the Kullback-Leibler distance

$$||\Theta - \Phi|| = \sum_{ij} \theta_{ij} \log(\theta_{ij}/\phi_{ij}) \tag{12}$$

or the Euclidean distance

$$||\Theta - \Phi|| = \sqrt{\sum_{ij}}(\theta_{ij} - \phi_{ij})^2.$$

We adopt (12) in the following discussion, although it should be clear that any suitable measure can be used. The extended model is then that, for pre-specified $\epsilon > 0$, we have a matrix Θ which also satisfies the extended Robinson-Kendall model in the natural order if

$$||\Theta - \Phi|| \leq \epsilon.$$

It is clear that when ϵ is chosen to be zero the extended model reduces to the model in equation (11). In this sense the parameter ϵ dictates how much relaxation we want to allow our models to have over the strict and deterministic Robinson-Kendall model. A large value of ϵ will produce all possible permutations for plausible seriation of the data. On the other hand smaller values will typically produce only a few of the possible permutations of the rows for seriation.

5.2 Models for correspondence analysis

Correspondence analysis is viewed as an alternative to adopting the Robinson-Kendall model for seriation, see e.g. Baxter (1994, chap. 5) and Goodman (1986), but it has usually been used only in an exploratory fashion in archaeology. Following Buck and Sahu (2000) we adopt a model-based approach using hierarchical Bayesian models.

In the first stage of model building we assume that $\mathbf{y}$ has a multinomial distribution with parameters N and $\boldsymbol{\theta}$ as previously, see equation (9). We then assume that,

$$\theta_{ij} = \theta_{i+}\ \theta_{+j}\ \ (1 + \lambda\, u_i\, v_j)\,, \tag{13}$$

where $0 \leq \lambda \leq 1$ and u_i and v_j are unknown row and column scores satisfying the constraints

$$\sum_{i=1}^{I} u_i\, \theta_{i+} = \sum_{j=1}^{J} v_j\, \theta_{+j} = 0, \quad \sum_{i=1}^{I} u_i^2\, \theta_{i+} = \ \sum_{j=1}^{J} v_j^2\, \theta_{+j} = 1,$$

where $\theta_{i+} = \sum_{j=1}^{J} \theta_{ij}$ and $\theta_{+j} = \sum_{i=1}^{I} \theta_{ij}$. The above constraints orthogonalize and normalize the row and column scores, u_i and v_j. The parameter λ is called the canonical correlation and it is the principal eigenvalue (with the row score vector as the eigenvector) for the χ^2 distance matrix between the observed and the fitted cell counts in the contingency table. The chronological order produced by the CA is taken as the ordering of the score vector $u_1, u_2, \ldots, u_I$. Buck and Sahu (2000) detail how to specify prior distribution for the unknown parameters, λ, u_i, v_j, θ_{i+} and θ_{+j}.

5.3 Model choice for relative chronology

We return to the stone tools data example described in the introduction. The extended R-K model chooses the relative order $(2, 5, 3, 6, 1, 4)$ overwhelmingly while the CA model chooses the order $(3, 6, 5, 2, 1, 4)$. We can choose between the two models, hence the orders, using the Bayesian model choice methods.

We use the decision theoretic approach of model selection to choose between the two models. The expected values of the loss function under different models are presented in Table 2. The extended R-K model with any value of ϵ has substantially lower expected loss values than the model for CA. Hence the extended R-K model is quite emphatically selected using this criterion.

The DIC values for the R-K model with $\epsilon = 10^{-2}$ and the model for correspondence analysis are 51.2 and 395.5 respectively. Thus the DIC also selects the extended R-K model which is simpler than the model used for correspondence analysis. By extrapolating the theoretical results obtained in Section 3 we can intuitively conclude that the Bayes factor will also select the simpler R-K model. Although such extrapolation may not always hold, we do not recommend the calculation of the Bayes factor. The calculation is much more involved and can be numerically unstable because of the constrained nature of the parameter space under the above models. See Chapter 6 of Chen *et al.* (2000) for similar examples on calculation of the Bayes factor for models with constrained parameters.

	Expected Loss
R–K ($\epsilon = 10^{-2}$)	59.5
R–K ($\epsilon = 10^{-3}$)	57.0
R–K ($\epsilon = 10^{-4}$)	55.2
CA	427.1

Table 2: Model choice for the stone tools data.

6 Discussion

In this paper we have discussed and illustrated Bayesian model choice methods both for relative and absolute chronology building problems. Three different model choice methods have been compared and illustrated with practical examples. The paper also points out the pressing need for adopting formal model choice methods for more complex future models which seem appropriate for archaeological data interpretation. For many purists the Bayes factor is the most appropriate tool that conveys the inferential content of the data. In this paper, however, we have not taken such a strong view. Instead, we have presented three different competitive criteria for model choice.

There are other Bayesian model choice methods which can also be used for model comparison. For example, a Bayesian computation method known as the reversible jump MCMC (Green, 1995) can be used to obtain the posterior probabilities of a number of competing models belonging to a certain structured class of models. This method is not considered because the models compared here, e.g. the extended R-K model and the model for correspondence analysis, do not belong to any class of structured nested models. Also the two models compared in Section 4 corresponding two different prior distributions cannot be written as subsets of a super structured nested model.

Some remarks on the use of the Bayesian model averaging methods for prediction are also appropriate. These methods are perhaps ideal if the sole purpose is to predict using the models without selecting an intermediate model. In this article our primary focus is to discuss model choice methods for comparing arbitrary models. Prediction is not the focus of the current article and this eliminates the need for model averaging.

The proposed Bayesian model fitting and model choice methods are attractive because these do not rely on asymptotic arguments unlike many classical methods of statistical inference, e.g. the likelihood ratio test. Asymptotic arguments are often invalid for the archaeological inference problems since the associated data sets are often small.

The article illustrates the potential of the Bayesian model choice methods for statistical inference. The different Bayesian models (and prior distributions) may lead to different sets of conclusions which can be contradictory. The proposed model choice methods provide the justification for choosing one set of inferential conclusions over the others.

REFERENCES

Aitkin, M. (1997) The calibration of P-values, posterior Bayes factors and the AIC from the posterior distribution of the likelihood. *Statistics and Computing,* **7**, 253–261.

Baxter, M. J. (1994) *Exploring Multivariate Analysis in Archaeology.* Edinburgh University Press, Edinburgh.

Buck, C. E., Cavanagh, W. G. and Litton, C. D. (1996) *The Bayesian Approach to Interpreting Archaeological Data.* Wiley, Chichester.

Buck, C. E. and Sahu, S. K. (2000) Bayesian models for relative archaeological chronology building. *Applied Statistics*, **49**, 423–440.

Chen, M.-H., Shao, Q.-M. and Ibrahim, J. G. (2000) *Monte Carlo Methods in Bayesian Computation.* Wiley, New York.

DiCiccio, T. J., Kass, R. E., Raftery, A. Wasserman, L. (1997). Computing Bayes Factors by Combining Simulation and Asymptotic Approximations. *Journal of the American Statistical Association*, **92**, 903–915.

Gelfand, A. E. (1996) Model determination using sampling based methods. In *Markov Chain Monte Carlo in Practice.* (Eds. W. R. Gilks, S. Richardson and D. J. Spiegelhalter). London: Chapman and Hall, pp 145–161.

Gelfand, A. E. and Ghosh, S. (1998) Model Choice: a minimum posterior predictive loss approach. *Biometrika*, **85**, 1–11.

Gelfand, A. E. and Smith, A. F. M. (1990) Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398–409.

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

Gilks, W. R., Richardson, S. and Spiegelhalter, D. G. (1996) *Markov Chain Monte Carlo In Practice.* London: Chapman and Hall.

Gomez Portugal Aguilar, D., Litton, C. D. and O'Hagan, A. (2002) A new piece-wise linear radiocarbon calibration curve with more realistic variance. To appear in *Radiocarbon.*

Goodman, L. A. (1986) Some useful extensions of the usual correspondence analysis approach and the usual log linear model approach in the analysis of contingency tables. *International Statistical Review*, **54**, 243–309.

Green, P. J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika,* **82**, 711–732, 1995.

Jacobi, R. M., Laxton, R. R. and Switsur, V. R. (1980) Seriation and dating of mesolithic sites in southern England. *Rev. Archeom.* **4**, 165–173.

Kass, R. E. and Raftery, A. E. (1995) Bayes factors and model uncertainty. *Journal of the American Statistical Association*, **90**, 773–795.

Kendall, D. G. (1971) Seriation from abundance matrices. In *Mathematics in the Archaeological and Historical Sciences*, (eds D. G. Kendall, F. R. Hodson, and P. Tautu), pp 215–252. Edinburgh University Press, Edinburgh.

Key, J. T., Pericchi, L. R. and Smith, A. F. M. (1999) Bayesian Model Choice: What and Why? In *Bayesian Statistics 6* (Eds. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith). Oxford University Press, 343–370.

Laud, P. W. and Ibrahim, J. G. (1995) Predictive Model Selection. *Journal of the Royal Statistical Society* B, **57**, 247–262.

Laxton, R. R. (1976) A measure of pre-Q-ness with aplications to archaeology. *Journal of Archaeological Science*, **3**, 43–54.

Nicholls, G. and Jones, M. (2001) Radiocarbon dating with temporal order constraints. *Applied Statistics*, **50**, 503–521.

O'Hagan, A. (1995) Fractional Bayes factor for model comparison (with discussion). *Journal of the Royal Statistical Society* B, **57**, 99–138.

Reece, R. (1994) Are Bayesian statistics useful to archaeological reasoning? *Antiquity*, **68**, 848–850.

Rubin, D. B. (1984) Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, **12**, 1151–1172.

Shennan, S. (1988) *Quantifying Archaeology.* Edinburgh University Press, Edinburgh.

Spiegelhalter, D. J., Best, N. G. and Carlin, B. P. and van der Linde, A. (2002) Bayesian measures of model complexity and fit, (with discussion). *Journal of the Royal Statistical Society*, B, **64**, 583–639.

Stuiver, M. K., Reimer, P. J., and Braziunas, T. F. (1998) High-precision radiocarbon age calibration for terrestrial and marine samples. *Radiocarbon*, **40**, 1127–1151.