10.1098/rsta.2003.1263

Bayesian computation: a statistical revolution

By Stephen P. Brooks

Statistical Laboratory, University of Cambridge, Centre for Mathematical Sciences, Wilberforce Road, Cambridge CB3 0WB, UK (steve@statslab.cam.ac.uk)

Published online 3 November 2003

The 1990s saw a statistical revolution sparked predominantly by the phenomenal advances in computing technology from the early 1980s onwards. These advances enabled the development of powerful new computational tools, which reignited interest in a philosophy of statistics that had lain almost dormant since the turn of the century. In this paper we briefly review the historic and philosophical foundations of the two schools of statistical thought, before examining the implications of the reascendance of the Bayesian paradigm for both current and future statistical practice.

Keywords: computer packages; Markov chain Monte Carlo; model discrimination; population ecology; prior beliefs; posterior distribution

1. Introduction

Though several key concepts were developed earlier—such as mathematical expectation (Huygens 1657), significance testing (Arbuthnot 1711), and the approximation of the binomial by the normal distribution (de Moivre 1718)—many of the earliest statistical methods were developed in the latter part of the nineteenth century. For example, the concept of linear regression first appeared in the work of Galton (1889), building on the concept of least squares introduced by Legendre (1805) and arguably also by Gauss (who claimed that he had been using the method since 1795 (Gauss 1809)). Galton (1888) also introduced the concept of statistical correlation, which was later developed by Edgeworth (1893), Yule (1897) and Pearson (1896), who also began the development of goodness-of-fit measures (e.g. Pearson 1900) at around the turn of the century. The field really took off in the 1920s and 1930s when Fisher (1922, 1925) developed the notion of likelihood for general estimation; Neyman & Pearson (1933) developed the basis for frequentist (often termed classical) hypothesis testing; and Yates & Cochran (1938) established the principles for the analysis of variance. Since then, the field has continued to grow and statistical inference now plays a key role in nearly every area of scientific research.

Bayesian methods date to the original paper by the Reverend Thomas Bayes, read to the Royal Statistical Society in 1763 (the paper was in fact read by Richard Price several years after Bayes' death). The area generated interest from Laplace (1774), Gauss (1809) and Pearson (Pearson & Filon 1898) amongst others and dominated statistical thinking throughout the nineteenth century. The Bayesian approach fell out of favour at the beginning of the twentieth century due, in part, to the domination

One contribution of 22 to a Triennial Issue 'Mathematics, physics and engineering'.

of the field by staunch opponents such as Neyman and Fisher, who held philosophical objections to the subjectivity of the Bayesian approach. Nonetheless, Bayesian stalwarts such as Jeffreys (1939), Savage (1954), Lindley (1965) and de Finetti (1970) kept the Bayesian flame alive by continuing to advocate Bayesian methods as remedies for certain deficiencies in the frequentist approach. It was not until the late 1980s that the Bayesian approach began to re-emerge, motivated both by rapid recent developments in computing and by the growing desire to describe increasingly complex scientific phenomena that older sampling theories were ill equipped to address. Since then the Bayesian approach has begun to dominate statistical research once more with new computational tools providing a far more flexible framework for statistical inference matching exactly the increasing complexity of scientific research as we move into the twenty-first century.

2. Two approaches to statistical inference

The two competing approaches to statistical inference are perhaps best explained in the context of a simple example. Suppose that we wish to toss a single coin, X. Our statistical model places a probability, say θ, on it coming down heads ($X = 1$) and a probability, $1 - \theta$, of it coming down tails ($X = 0$).[†] This is a statistical model parametrized by the unknown probability θ which simply says that $\text{Prob}(X = 1) = \theta$ and $\text{Prob}(X = 0) = 1 - \theta$. In fact, we can write down a general formula which covers both cases, i.e.

$$\text{Prob}(X = x) = \theta^x (1-\theta)^{1-x} = f(x \mid \theta), \quad \text{for } x = 0 \text{ and } x = 1.$$

Given a whole series of n coin tosses $X_1, \dots, X_n$ we can write down a joint probability for all n outcomes: $f(x_1, \dots, x_n \mid \theta)$,[‡] which is read aloud as 'the joint probability distribution, f, of x_1 up to x_n given the value of θ' and explicitly states the dependence of the series of observed coin tosses on the value of the parameter θ. This joint probability distribution then forms the basis for statistical inference, whichever inferential approach is used.

(a) *The frequentist approach*

Statistical inference is essentially an inversion problem. Given a parametric model and associated parameter values, it is possible to predict potential outcomes. For example, given the probability distribution above and the value of θ it would be possible to simulate or predict the outcomes of the coin-tossing experiment. However, in practice, it is the outcome of the experiment that is observed and the values of the parameters that are unknown. For example, we might observe three coin tosses each of which were heads and wish to infer from these observations what the value of θ (i.e. the probability of a head) might be. In this case it seems reasonable to consider

[†] Note that a tossed coin caught in mid-air will always have probability $\theta = \frac{1}{2}$. Coins can only be biased if allowed to land and bounce after being tossed. Alternatively, coins can be spun rather than tossed (see, for example, Gelman & Nolan 2002, § 7.6).

[‡] With independent coin tosses,

$$f(x_1, \dots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta).$$

inverting the joint probability distribution and to think of it not as a function of x given the value of θ, but as a function of θ given the observed values of x, i.e. to set $l(\theta \mid x_1, \dots, x_n) = f(x_1, \dots, x_n \mid \theta)$. This function l is referred to as the likelihood function and is the basis for frequentist statistical inference.

The likelihood is used to compare different candidate values for the model parameters. The parameter values which lead to the highest value of the likelihood function are considered to be the best and are known as the maximum likelihood estimates (MLEs) for the 'true values' of the model parameters.

In practical terms, this means that the frequentist approach to inference involves the maximization of a function of the model parameters. In some simple cases, this maximization can be performed analytically: by taking derivatives of the likelihood function, for example. However, for more complex examples numerical optimization procedures are required, which treat the likelihood function as a multi-dimensional surface where the height of the surface at any vector point $\boldsymbol{\theta}$ is given by the likelihood at that point. Such methods can involve substantial computational expense and often suffer from a lack of reliability in that many optimization routines are liable to become stuck in local rather than global optima. Though comparable problems also occur with the corresponding Bayesian tools, these can often be more easily overcome within the Bayesian paradigm using the powerful new computational tools that we shall introduce in § 3.

An important issue in parameter estimation is the quantification of the uncertainty surrounding the ability of our estimate to reflect the truth. The frequentist approach to this problem relies on the *strong repeated sampling principle* and assesses the performance of our statistical estimation procedure for a single sample of data on the basis of the expected long-run performance given a hypothetical series of datasets collected under identical conditions. An obvious consequence of this is the frequentist 95% confidence interval, which, for a model parameter θ, represents an interval on the real line $(a[\boldsymbol{x}], b[\boldsymbol{x}])$ constructed on the basis of data $\boldsymbol{x} = x_1, \dots, x_n$ which, if they were repeated for a series of new datasets, would cover the true value of θ 95% of the time. Thus, the uncertainty we attribute to our frequentist estimate is based upon our beliefs about the ability of our procedure to estimate the truth under hypothetical repetitions of the single experiment that we observe. To many, this seems a rather unnatural concept and appears to be addressing the wrong question. For example, the fact that a drug is successful 99% of the time is not of direct interest to any individual patient, who is more likely to be interested only in the chances of his/her own recovery (which is not necessarily 0.99).

The method by which these confidence intervals are usually constructed, for all but the simplest problems, is based upon the observation that the MLE has an asymptotic normal distribution with known mean and variance both of which are functions of the observed data.† The asymptotic aspect of this result refers to the size of our dataset, with increasing accuracy observed as the size of our dataset increases. However, in many cases the size of our dataset will be fixed. Thus, the dataset may not always be sufficiently large for these asymptotic results to hold, in which case the results should be viewed with some scepticism.

† Obvious exceptions include non-parametric inference and cases where exact distributions can be calculated, as in the case of linear models with homogeneous normal errors, for example.

Figure 1. Reverend Thomas Bayes (1702–1761). British mathematician and Presbyterian minister, famous for both the theorem in probability and the statistical philosophy that now bear his name.

(*b*) *The Bayesian approach*

The Bayesian approach to statistical inference is based upon Bayes' theorem (Bayes 1763), which, for continuous random variables, states that†

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{x}) \propto f(\boldsymbol{x} \mid \boldsymbol{\theta})p(\boldsymbol{\theta}).$$

The distribution $\pi(\boldsymbol{\theta} \mid \boldsymbol{x})$ is referred to as the *posterior* distribution and represents our beliefs about the model parameters after having observed the data $\boldsymbol{x}$. The distribution $p(\boldsymbol{\theta})$ is referred to as the *prior* and is concerned with our original beliefs about $\boldsymbol{\theta}$ before observing the data. Thus, Bayes' theorem allows us to update our prior beliefs on the basis of the observed data to obtain our posterior beliefs, which form the basis for our inference.

As in the frequentist case above, the Bayesian approach ends up with a function of the model parameters given the observed data, but this time our inference is in the form of a probability distribution for $\boldsymbol{\theta}$ rather than a simple point estimate. This illustrates one of the fundamental differences between the beliefs of the Bayesian and frequentist statisticians. The frequentist statistician believes that there exists a fixed true value for the model parameters and tries to estimate this value by maximizing the likelihood. The Bayesian believes that the unknown parameters have a fixed but unknown distribution which represents their beliefs about those parameters having observed the data. Note that the Bayesian may also believe that a true value exists, but since this can never be known with absolute certainty they prefer to think instead in terms of a distribution which reflects this uncertainty. As their level of information increases, the width of this distribution decreases and, in the limit (i.e. omnipotence),

† Note that statistical distributions need only be specified up to proportionality, since it is the shape of the distribution that determines the relative probabilities of different values under that distribution. If distribution $\pi(\theta) \propto g(\theta)$, then π is 'normalized' by dividing by $\int g(\theta)\, \mathrm{d}\theta$.

the Bayesian would, in theory, end up with a point mass (i.e. a single value with probability one) on a particular value for the parameter of interest.

From a practical perspective, this complex multi-dimensional posterior distribution may be difficult to interpret in terms of the physical process under study. In practice, we often summarize the distribution by calculating a variety of interpretable univariate summary statistics such as posterior means and variances for example.

Point estimates for parameters of interest are obtained via the specification of a loss function. This approach stems from the gaming roots of the Bayesian approach. The Bayesian approach to parameter estimation is to choose the value which minimizes their expected loss under the posterior distribution for some loss function ascribing a hypothetical cost (financial or otherwise) to getting the estimate wrong. For example, the quadratic error loss function for estimating the true value θ by ϕ is given by $\mathcal{L}(\phi - \theta)^2$, which increases in size the further our estimate lies from the truth. With this particular loss function it turns out that for any problem the estimate (i.e. value of ϕ) which minimizes the expected value of this loss function over values of θ simulated from the posterior distribution is simply the posterior mean. Thus, if we chose the quadratic error loss function, our best estimate of the true parameter value would be the posterior mean. Other loss functions would lead to other estimates for the parameter of interest. For example, under the zero–one loss function, under which we incur a penalty of 1 if we choose the incorrect value and no penalty at all if we are correct, the best estimate is the posterior mode.

This provides us with an interesting comparison between the Bayesian and frequentist approaches to parameter estimation. If we adopt a flat prior distribution for our parameter of interest, the corresponding posterior distribution for that parameter will be of identical shape to the joint probability distribution $f(\boldsymbol{x} \mid \theta)$. Therefore, locating the mode of the posterior distribution (i.e. adopting the zero–one loss function) with a flat prior provides an identical estimate of the parameter of interest as the frequentist approach. Many Bayesians have used this simple point to criticize the frequentist approach as being inherently Bayesian, but with an unjustifiable restriction to flat prior distributions and zero–one loss functions.

The Bayesian approach to statistical inference provides a far more natural framework to consider the concept of parameter uncertainty and the posterior variance provides us with a direct measure of the uncertainty associated with any given parameter. A more detailed uncertainty indicator is the credible interval, the Bayesian equivalent of the frequentist confidence interval. A 95% credible interval for a particular parameter θ is an interval on the real line within which θ lies with probability 0.95. There is no assumption here of the properties of our estimation process under hypothetical replications of our original experiment and the interpretation of the interval is arguably more natural than the frequentist analogue. Of course, the interval will depend upon the prior distributions adopted, though this influence typically decreases with increasing size of the original dataset.

(*c*) *Tossing the euro*

To illustrate the difference between the frequentist and Bayesian approaches let us return to our coin-tossing example and suppose that we decide to toss a newly minted euro coin 10 times and observe 8 heads. Given that the probability of a head is θ (the estimation of which is the object of our exercise), the probability distribution

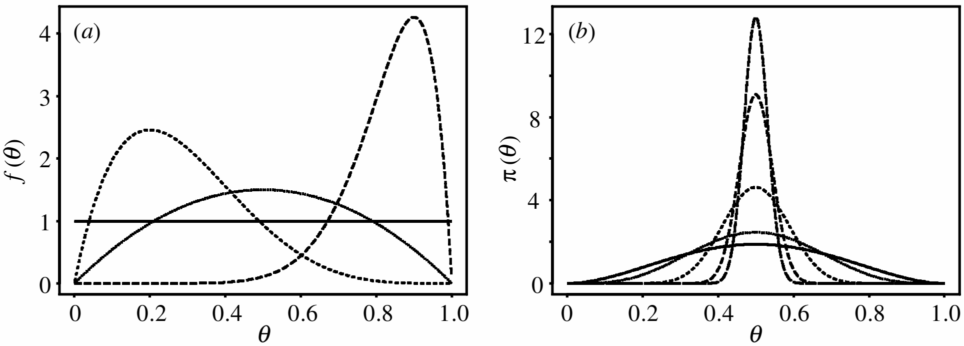

Figure 2. (a) Illustration of different possible shapes of the beta distribution. (b) Posterior distribution for θ under a flat prior distribution and for $n = 4, 8, 32, 128, 256$, with $x = n/2$. Note that the larger the value of n, the higher and narrower the peak.

for this outcome ($x = 8$ heads out of a total of $n = 10$ trials) is given by the binomial distribution, i.e.

$$f(x \mid \theta) = \frac{n!}{x!(n-x)!}\theta^x(1-\theta)^{n-x},$$

where $x!$, pronounced 'x factorial', is simply x multiplied by $x - 1$ multiplied by $x - 2$ and so on until we reach 1. As described above, the likelihood function $l(\theta \mid x)$ therefore takes the same form and it is fairly easy to see that it is maximized at $\theta = x/n$.†

To conduct a Bayesian analysis, we begin by placing a prior distribution on the parameter θ, which, since it is a probability, lies between zero and one. This prior can take many different forms and is intended to reflect our own personal beliefs about what parameter values we believe to be likely before observing the outcome of our experiment. In the absence of any *a priori* information, we might adopt a flat prior distribution for θ which makes all possible values of θ equally likely, i.e. $p(\theta) = 1$. A more general prior distribution which provides us with additional flexibility in terms of the range of beliefs that can be expressed is to take the beta distribution so that $p(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$. The variables a and b are our prior parameters and can be chosen so as to reflect our own personal beliefs. Figure 2a illustrates the shape of the beta distribution for different values of a and b. Taking $a = b = 1$, we obtain the flat (or uniform) distribution, and whenever $a = b$ we obtain a distribution which is symmetric about $\theta = \frac{1}{2}$.

Taking a beta prior distribution, we obtain the following posterior distribution:

$$\pi(\theta \mid x) \propto f(x \mid \theta)p(\theta) \propto \theta^{x+a-1}(1-\theta)^{n-x+b-1}.$$

Thus, the posterior distribution for θ is also in the form of a beta distribution with parameters $x + a$ and $n + b - x$. This has mean

$$\frac{x+a}{n+a+b},$$

† Taking the derivative of $\log l(\theta \mid x)$ with respect to θ and setting to zero, we obtain

$$\frac{x}{\theta} - \frac{n-x}{1-\theta} = 0.$$

Taking the second derivative confirms that this is indeed a maximum.

mode

$$\frac{x+a-1}{n+a+b-2}$$

and variance

$$\frac{(x+a)(n+b-x)}{(n+a+b)^2(n+a+b+1)}.$$

Note the clear difference between the posterior estimates and the MLE but that if we take a flat prior distribution (i.e. $a = b = 1$), then the posterior mode is exactly the MLE, as we should expect. Note also that as the sample size increases, the posterior mean and mode both converge to the MLE† and the posterior variance tends to zero. Figure 2*b* provides an illustration of the posterior distribution for p under a uniform prior distribution for different sample sizes observing $x = n/2$ heads in each sample of size n. The posterior variance clearly decreases with sample size and, as n increases, the posterior distribution puts increasing weight on the MLE, which is 0.5 here.

A practical problem with the Bayesian approach, the solution to which was one of the main catalysts for the re-emergence of the paradigm into mainstream statistics, is the difficulty associated with exploring and summarizing realistically complex posterior distributions. Unlike the coin-tossing example, in most practical problems, summary statistics, such as the posterior mean and variance of parameters of interest, will not be available analytically and computational tools are required to gain meaningful inference from the posterior distribution. The introduction to the statistical literature of the technique known as Markov chain Monte Carlo (MCMC) in the late 1980s overcomes this problem and greatly simplifies the Bayesian analysis of even the most complex data.

3. Bayesian computation

MCMC methods date from the original work of Metropolis *et al.* (1953), who were interested in methods for the efficient simulation of the energy levels of atoms in a crystalline structure. The original idea was subsequently generalized by Hastings (1970), but its true potential was not fully realized within the statistical literature until Gelfand & Smith (1990) demonstrated its application to the estimation of integrals commonly occurring in the context of Bayesian statistical inference.‡

Basically, MCMC methods provide a mechanism for generating observations from arbitrarily complex probability distributions, such as the posterior distribution arising from a Bayesian statistical analysis. This provides a practical means for obtaining inference by drawing a series of observations from our posterior distribution and then calculating sample estimates of any quantities of interest. In practice, this usually means calculating the sample means and variances of all model parameters, though posterior modes, inter-parameter correlations and quantiles may also be of interest, for example. Based upon a sample from the posterior distribution, these empirical

† This is perhaps most easily seen by reformulating the posterior mean as $(1-w)[a/(a+b)] + w[x/n]$, where $w = n/(a+b+n)$. Thus, the posterior mean is a weighted average of the MLE and the prior mean. As the sample size increases, the value of w tends to 1, giving all weight to the MLE.

‡ The calculation of posterior means and variances, for example, are essentially integration problems.

estimates approximate the true values under the posterior distribution and arbitrary accuracy can be obtained by increasing the sample size.

Generic methods for drawing samples from specific or even general statistical distributions have been around for some time. However, until the advent of MCMC methods there was no way of efficiently generating samples from arbitrarily complex multivariate distributions. MCMC methods work by constructing what is known as a Markov chain, which is essentially a series of observations generated one after the other so that the value at any point in the sequence depends only upon the preceding value. The key to MCMC methods is to construct a chain which has the property that certain key characteristics of the sequence of generated values are exactly the same as those from a sample of observations from the statistical distribution of interest. In particular, the marginal distributions of the sampled values is exactly that corresponding to the posterior distribution of interest, so that the sample mean of the sequence approximates the true mean under the posterior distribution, for example.

The MCMC algorithm itself can be likened to an explorer who wishes to map a complex and uncharted landscape. The landscape corresponds to the high-dimensional surface representing the probability distribution of interest, in this case a posterior distribution. The aim is to explore this surface and to learn about certain key properties such as the proportion of the total volume beneath a particular area on the surface, for example. The surface itself is 'normalized' so that the total volume beneath it is 1. Thus the volume beneath any particular area will be bounded between 0 and 1 and corresponds to the probability that a random draw from the corresponding probability distribution lies within that area. By estimating such volumes, we can answer questions such as 'what is the probability that parameter p is positive', for example. Similarly, by roaming the surface in such a way that the time spent in any area is proportional to the volume beneath that area, we can estimate the mean of our distribution by simply recording our position on the surface at regular intervals and calculating the corresponding sample mean.

Continuing our analogy, these surfaces are typically so complex that the explorer cannot simply plan an entire route around the landscape but must instead explore the surface constructing a suitable path as he goes. The MCMC algorithm provides a very simple framework for randomly determining where to travel next given the current position. By breaking the exploration process into a series of simple steps, arbitrarily complex surfaces can be mapped. The more complex the surface, the longer the route and the greater the computational expense. However, since the explorer is attracted to the higher ground, the peaks tend to be identified at a very early stage and the overall shape of the surface therefore quickly becomes apparent. As we continue the mapping process we add finer and finer detail until an adequate resolution is obtained. As an illustration, figure 3 provides several stages in the exploration of a simple Beta$(5, 5)$ distribution. We can clearly see the increase in accuracy as the exploration process (number of sampled observations) increases.

In contrast, the frequentist approach to statistical inference can be likened to a mountain climber given the task of scaling a complex surface, which, in this case, corresponds to the likelihood function. The climber is less interested in the overall shape of the surface, and more interested in the location and shape of the highest peak (corresponding to the MLE), which is a far more specific goal and requires far higher resolution (at least at the peak) than that of the Bayesian explorer. The climber's path is based upon sequential rules, much like the MCMC algorithm, and

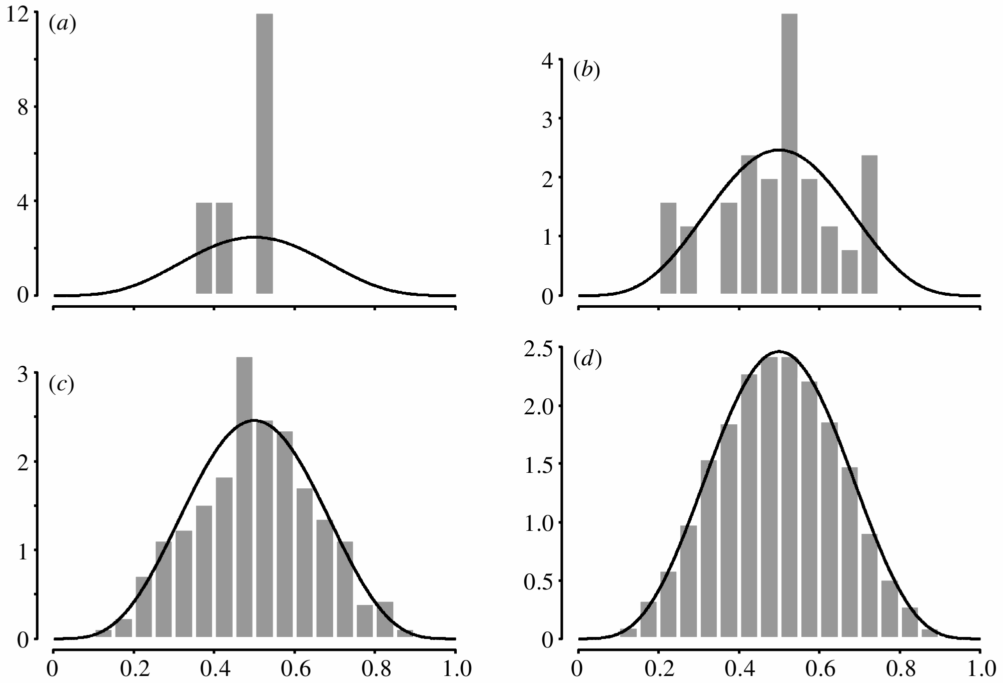

Figure 3. Approximations to the Beta(5, 5) density (solid line) from samples of size (*a*) 5, (*b*) 50, (*c*) 500 and (*d*) 5000.

the simplest approach is for the climber to consider the gradient in all directions at his current position, travelling in the direction of the steepest uphill gradient until he can ascend no further. The problem with this rule is that if he starts somewhere on the slope of a minor hill, he will immediately ascend that hill until he reaches its peak. He will then be unable to descend the other side (because of his uphill-only rule) to reach the next or any other peak which may be higher than his current position. Thus, the climber easily becomes stuck in local rather than global maxima.

There have been many suggestions for improving on this basic rule, many of which automatically move uphill but allow the occasional random downhill move whose probability decreases with the depth of the descent. This allows the climber to descend local peaks so as to locate more prominent peaks elsewhere. However, even these are still not guaranteed to locate the correct peak within a finite time. Various more technical improvements to the basic approach are also available, and these are designed to make the chances of locating the correct peak increase as the simulation continues. However, these rules depend upon the shape of the surface, with a more-complex (or high-dimensional) surface demanding more-complex rules so that there exists a practical limit to the complexity of the surface that the climber can successfully cover.

In practice, simplifying assumptions or approximations are often made to more complex surfaces in order to facilitate the frequentist exploration process. In contrast, the advent of MCMC methods means that the Bayesian need not concentrate solely on simple models and now has the tools to explore the increasingly complex model structures that are the focus of many areas of modern scientific research.

4. An average model?

Statistical inference is not limited to parameter estimation. Hypothesis testing, in which we discriminate between competing models in order to gain a better understanding of the structure of the stochastic system under study, is also a common goal.

The frequentist approach to model selection uses methods such as the likelihood ratio test to compare nested models to determine whether or not a particular parameter (or set of parameters) should be included within the model. To perform a likelihood ratio test, the value of the likelihood is calculated at the corresponding MLEs for both models. The ratio of these two likelihood values is then compared with appropriate statistical tables to determine whether or not the ratio is 'significantly' large. If it is, we reject the hypothesis that the parameters in question may be excluded from the model. As with the construction of frequentist confidence intervals and standard errors, the distribution of the likelihood ratio test statistic is asymptotic and is only approximately true even for moderate sample sizes.

Alternatively, procedures based upon information criteria, such as the AIC (Akaike 1974), can be used. Such criteria are derived from information theory and are based upon the evaluation of the log of the likelihood at the MLE, combined with a penalty term accounting for the number of parameters in the model. This penalty term ensures that parsimonious models are preferred so that if two models provide the same likelihood value at their corresponding MLEs, then the model with fewer parameters is selected. The inclusion of the number of parameters as an appropriate penalty term derives from an approximation to the bias inherent in the procedure of simply comparing the likelihoods alone. The AIC provides an approximation to a formal information-theoretic comparison procedure but performs well only when the observed dataset is large and when 'good' models are compared (Burnham & Anderson 1998).

In practice, when there are many plausible models to compare, the frequentist approaches to model selection may be computationally very costly. Exhaustive evaluation of all plausible models may be prohibitively expensive since the MLE must be calculated for each model in turn. Thus, various model exploration procedures have been proposed which attempt to identify the best models without having to evaluate them all. Perhaps the most common approach in the context of nested models is to evaluate the most-complex model first and then evaluate all models which can be derived from that model by the removal of a single model parameter. If an improvement can be found (according to some comparison criterion), the original model is replaced by the best alternative and the procedure repeated by considering all submodels obtained by deleting a second model parameter. This process is repeated until no further improvement is obtained. Various alternative implementations are available, such as starting with the smallest rather than the largest model and adding rather than deleting parameters, but all such procedures exhibit a tendency to get stuck at locally optimal models and none are guaranteed to find the best model.

The Bayesian approach to model discrimination is based upon an extension to the posterior distribution to include not only uncertainty as to the model parameters, but also the model itself. Suppose we observe data $\boldsymbol{x}$ and have a series of plausible models indexed by $m = 1, \ldots, M$. Suppose also that under model m we have a joint probability distribution for the observed data $\boldsymbol{x}$ given by $f_m(\boldsymbol{x} \mid \boldsymbol{\theta}_m)$, where $\boldsymbol{\theta}_m$ denotes the vector of parameters associated with model m. Then, by specifying a prior distribution $p_m(\boldsymbol{\theta}_m)$ for the model parameters under each model and a prior probability for each model, $p(m)$, we can derive a posterior distribution over both parameter and model space given by

$$\pi(\boldsymbol{\theta}_m, m \mid \boldsymbol{x}) \propto f_m(\boldsymbol{x} \mid \boldsymbol{\theta}_m) p_m(\boldsymbol{\theta}_m) p(m).$$

This joint posterior distribution can then be broken down into two components, since (by Bayes' theorem)

$$\pi(\boldsymbol{\theta}_m, m \mid \boldsymbol{x}) = \pi(\boldsymbol{\theta}_m \mid m, \boldsymbol{x}) \pi(m \mid \boldsymbol{x}),$$

where $\pi(\boldsymbol{\theta}_m \mid m, \boldsymbol{x})$ denotes the posterior distribution under model m and $\pi(m \mid \boldsymbol{x})$ denotes what we call the 'posterior model probability' representing our beliefs, having observed data $\boldsymbol{x}$, of the probability that model m is the true model given that one of models $1, \ldots, M$ is true.

Having obtained these posterior model probabilities we have two choices. We can either use them to discriminate between the competing models by computing the Bayes factor, which is simply the ratio of the posterior odds (the ratio of the posterior to the prior model probability) for any two competing models (Kass & Raftery 1995), or, if we are interested in prediction for example, we can use them to combine the predictions from the different models by weighting them according to their posterior model probability. In this way, we obtain predictive inference which properly reflects both sources of uncertainty (i.e. that associated with the model parameters and with the model itself) and provides a more realistic prediction than anything based upon any one of the models under consideration. The ability to perform model-averaged predictive inference is one of the great advantages of the Bayesian approach and ensures that predictive inference fully accounts for both parameter and model uncertainty.

The second substantial advantage of the Bayesian approach to model discrimination, as with parameter estimation, is the availability of computational tools capable of undertaking analyses that are either impossible or at least prohibitively computationally expensive to undertake within the frequentist paradigm. Such problems are commonly encountered when very large numbers of models need to be explored. These Bayesian tools, known as trans-dimensional (TD)MCMC methods (Green 1995), are a fairly straightforward extension to the MCMC algorithm, but allow for the construction of Markov chains capable of exploring the more complex posterior distribution described above.

As with the benefits of MCMC methods over the frequentist alternatives, TDMCMC methods allow us to explore almost arbitrarily large model spaces. This removes the need to ignore plausible models simply to reduce the model space, as is often the case for frequentist analyses. However, TDMCMC methods are a relatively new addition to the statistical literature, and the construction of the necessary between-model transitions, though well understood in theory, can be rather difficult to implement in practice. Ultimately, such techniques will only be routinely applied in practice once a suitable implementational framework has been established and incorporated into widely available computer packages. We are still some way off from developing this framework and this is the focus of a great deal of current methodological research.

5. Modelling bird survival

As an example of a model-selection problem highlighting the potential advantages of the Bayesian approach using TDMCMC, we consider a problem from population ecology.

Data collection for wildlife populations is often performed using capture–recapture and/or tag–recovery techniques (Schwarz & Seber 1998) which involve marking animals (often at birth) and then recording subsequent resightings either alive (recapture) or dead (recovery). The primary aim of such studies is to learn about the survival dynamics of the population under study, though recovery and recapture rates may also be of interest. In particular, we often wish to study the dependence of the annual survival rate on a variety of factors such as age or year as well as any number of environmental or individual covariates.

We consider recapture–recovery data from a population of Scottish shags† (*Phalacrocorax aristotelis*) previously analysed by Catchpole *et al.* (1998). Shags move through three distinct life stages and are referred to as *pulli* in their first year, *immature* in their next two years, finally becoming *adult* in their fourth year. The survival, recapture and recovery parameters are assumed to depend on year and/or age, so that the full model has three groups of parameters:

$$\begin{aligned}\phi_a(t) &= \mathbb{P}(\text{an animal of age } a \text{ alive in year } t \text{ survives until year } t+1);\\ p_a(t) &= \mathbb{P}(\text{an animal of age } a \text{ alive in year } t \text{ is resighted at that time});\\ \lambda_a(t) &= \mathbb{P}(\text{an animal of age } a \text{ which dies in } [t, t+1)\text{, has its band returned}),\end{aligned}$$

for $a \in \{1, 2, 3, \mathrm{A}\}$, where A represents adult. Different models can then be represented by placing different restrictions upon these parameters.

Here, we have no covariates, and are interested in determining the structure describing the dependence of survival upon age and/or year. The data were collected annually over a nine-year period and so the various combinations of year and age dependence provide us with almost 500 000 different models. For example, one model allows all parameters to depend upon both age and sex, while another may group the survival probability for ages 2 and 3 and set it constant for all time, etc. This provides an extremely difficult model-selection problem. From the frequentist perspective, it is computationally infeasible to exhaustively evaluate all models and so it is necessary to restrict the model space by 'guessing' which models are likely to describe the data best. This is the approach taken by Catchpole *et al.* (1998), who identified the model with the following restrictions as that which minimized the AIC statistic (with a value of 12 197.7) amongst a small selection of alternatives:

$$\begin{gathered}\phi_2(t) = \phi_3(t) = \phi_{\mathrm{imm}}, \qquad \phi_{\mathrm{A}}(t) = \phi_{\mathrm{A}};\\ p_1(t) = p_1, \qquad p_2(t) = p_2, \qquad p_3(t) = p_3;\\ \lambda_1(t) = \lambda_2(t) = \lambda_3(t) = \lambda_{\mathrm{A}}(t) = \lambda(t).\end{gathered}$$

The model here is described in terms of the restrictions imposed on the full model described earlier. In terms of survival, for example, we have time-dependent survival for pulli and a common survival rate for immatures which is constant over time, as

† A seabird somewhat similar to, but smaller than, a cormorant.

is the adult survival rate. Similarly, the model suggests that the recovery rates are the same for all ages, but change over time.

The Bayesian analysis using TDMCMC does not need to restrict attention to only a small selection of models as it is capable of exploring a very large model space with reasonable efficiency. Adopting a uniform prior distribution for all model parameters and a flat prior distribution over models, the Bayesian analysis gives highest posterior model probability to the following model:

$$\phi_2(t) = \phi_3(t) = \phi_{\text{imm}};$$
$$p_1(t) = p_1, \qquad p_2(t) = p_2;$$
$$\lambda_1(t) = \lambda_1, \qquad \lambda_2(t) = \lambda_3(t) = \lambda_{\text{imm}}, \qquad \lambda_{\text{A}}(t) = \lambda_{\text{A}}.$$

This model suggests that adult survival is year dependent rather than constant across all years and is more consistent with current ecological thinking in that we would expect that annual environmental fluctuations at the breeding site would almost certainly affect their survival. This would not be the case for immatures, which remain at sea throughout the year and would be less vulnerable to the influence of such variations. This model also suggests that the recovery probabilities are age rather than year dependent. Again, this is more consistent with expert opinion since

> the less experienced juveniles and first winter birds seem to die in ways associated with man (e.g. caught in nets or lobster pots) or in places where they are more likely to be found (e.g. blown inland or into estuaries, which rarely happens to older birds).
>
> M. Harris (2001, personal communication)

Thus, we should expect a larger recovery rate for these shags and a corresponding age-dependent structure for recoveries, which was not observed in the earlier model.

It is worth noting that this model has an associated AIC statistic of 12 138.72 and therefore improves (from the frequentist perspective) upon the model discussed by Catchpole *et al.* (1998). The reason that this model was not identified in the original analysis is that it simply was not considered. If it had been possible to consider all plausible models within the frequentist analysis, then both Bayesian and frequentist analyses would have identified the same model and would therefore have provided the same conclusions. Thus, it is the tool rather than the philosophy that leads us to the right result, as is often the case.

6. Looking to the future

To many, the Bayesian approach exhibits a natural affinity with recent scientific philosophy, stressing the advantages of cumulative evidence-based practice. Though opponents criticize the subjectivity in that prior distributions need to be used, others see this as an advantage which, in many cases, provides an essential extra source of information and invariably improves the precision of parameter estimates. It is worth noting that the subjective choice of prior is not too dissimilar from the problem of choosing a significance level for the frequentist hypothesis test, for example. In fact, most criticisms aimed at one or other of the two approaches seem to have an analogue in the alternative approach. Computationally, for example, the problem of locating the global maximum in the frequentist setting is comparable with determining the number of observations necessary for the corresponding MCMC simulation.

In practice, many researchers will be less interested in the philosophical foundation of the two distinct statistical approaches to inference than they are in getting the right answer to the specific questions arising from their own data. In many cases, the questions of interest will be partly directed by the statistical tools that have traditionally been available as well as the traditions of the discipline itself. For example, the automatic recourse to analysis of variance, t tests and χ^2 tests to analyse almost any dataset is standard in many disciplines. Such techniques are extremely easy to apply using standard and commonly available statistical packages and, if these techniques provide an answer to the question of interest, then the pragmatic approach is simply to use them. There is certainly little, if any, advantage in applying the Bayesian machinery to such problems, particularly given the fact that very few of the corresponding Bayesian techniques can be so easily implemented. Thus, if your primary question of interest can be simply expressed in a form amenable to a t test, say, there really is no need to try and apply the full Bayesian machinery to so simple a problem.

However, the many practitioners who find themselves tailoring their questions to the tools they have available; those who find themselves being forced to use unrealistic approximations or assumptions in their analysis; those who feel uncomfortable adopting essentially arbitrary critical values for testing hypotheses; or those who wish to incorporate additional information that could be conveniently expressed in the form of a prior distribution may find that the Bayesian approach provides an alternative which overcomes these difficulties. This may involve the construction of problem-specific code (as, in fact, is often necessary for frequentist analyses of 'non-standard' datasets) in order to conduct the necessary simulations and this may present a barrier to many, the removal of which is perhaps one of the biggest challenges facing statistical methodologists today.

Statistical tools to undertake standard frequentist statistical techniques are widely available in an enormous range of computer packages. As restrictive as some of these tools may be in terms of the analyses they provide, they are generally quick and easy to use. Bayesian analyses, on the other hand, generally require the development of problem-specific computer code because the implementation of many of the Bayesian tools requires at least some element of 'tuning' in order to achieve optimal performance. The development of techniques for automating this tuning process so that computer packages can begin to be developed is one of the biggest challenges facing statistical methodologists today, since, without them, many researchers will be unable to implement these methods themselves.

A recent success story is the development of the BUGS† project (Spiegelhalter *et al.* 1996) by the MRC Biostatistics Unit in Cambridge. This impressive suite of programs was originally developed to tackle a variety of standard problems encountered in medical statistics, but is now widely used in areas as diverse as ecology, sociology and geology. This powerful and very easy to use package enables statisticians and practitioners alike to analyse a wide variety of standard models as well as allowing a reasonable range of problem-specific models to be constructed and analysed.

† The BUGS package is freely available from the BUGS website at http://www.mrc-bsu.cam.ac.uk/bugs/.

However, the methodological framework is not yet in place to allow the construction of a general package capable of constructing arbitrary model structures. Nor do we yet have the methods in place to automate the model discrimination process via TDMCMC. Though some of the foundation work has already been laid (Brooks *et al.* 2003; Green 2003), there remains a great deal of work to be done in determining how to construct algorithms capable of moving efficiently between different models and, in particular, automating the selection of the various tuning parameters necessary to make these algorithms work well. It is the absence of these general implementation frameworks that has prevented the development of Bayesian statistical packages to rival the myriad of well-established frequentist packages already available. However, once this framework has been developed, the BUGS package will provide a template for the development of the next generation of statistical packages aimed at meeting the increasing demands of researchers working at the coalface of scientific research.

References

Akaike, H. 1974 A new look at the statistical identification model. *IEEE Trans. Automatic Control* **19**, 716–723.

Arbuthnot, J. 1711 An argument for divine providence, taken from the regularity observed in the births of both sexes. *Phil. Trans. R. Soc. Lond.* **27**, 186–190.

Bayes, T. 1763 An essay towards solving a problem in the doctrine of chances. *Phil. Trans. R. Soc. Lond.* **53**, 370–418.

Brooks, S. P., Giudici, P. & Roberts, G. O. 2003 Efficient construction of reversible jump MCMC proposal distributions (with discussion). *J. R. Statist. Soc.* B **65**, 3–55.

Burnham, K. P. & Anderson, D. R. 1998 *Model selection and inference: a practical information-theoretic approach.* Springer.

Catchpole, E. A., Freeman, S. N., Morgan, B. J. T. & Harris, M. P. 1998 Integrated recovery/recapture data analysis. *Biometrics* **54**, 33–46.

de Finetti, B. 1970 *Teoria della probabilitia* vols 1 and 2. Einaudi: Turin. (Transl. 1974, 1975 *Theory of probability*, vols 1 and 2. Wiley.)

de Moivre, A. 1718 *The doctrine of chances.* London: W. Pearson.

Edgeworth, F. Y. 1893 Statistical correlation between correlated averages. *J. R. Statist. Soc.* **56**, 563–568.

Fisher, R. A. 1922 On the mathematical foundations of theoretical statistics. *Phil. Trans. R. Soc. Lond.* A **222**, 309–368.

Fisher, R. A. 1925 *Statistical methods for research workers.* Edinburgh: Oliver & Boyd.

Galton, F. 1888 Co-relations and their measurement, chiefly from anthropological data. *Proc. R. Soc. Lond.* **45**, 135–145.

Galton, F. 1889 *Natural inheritance.* London: Macmillan.

Gauss, C. F. 1809 *Theoria motus corporum celestium.* Pertheset Besser: Hamburg. (Transl. 1857 *Theory of motion of the heavenly bodies moving about the Sun in conic sections*, C. H. Davis. Boston, MA: Brown.)

Gelfand, A. E. & Smith, A. F. M. 1990 Sampling-based approaches to calculating marginal densities. *J. Am. Statist. Assoc.* **85**, 398–409.

Gelman, A. & Nolan, D. 2002 *Teaching statistics: a bag of tricks.* Oxford University Press.

Green, P. J. 1995 Reversible jump MCMC computation and Bayesian model determination. *Biometrika* **82**, 711–732.

Green, P. J. 2003 Trans-dimensional Markov chain Monte Carlo. In *Highly structured stochastic systems.* Oxford University Press.

Hastings, W. K. 1970 Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.

Huygens, C. 1657 De ratiociniis in ludo aleae. In *Exercitationum mathematicarum* (ed. F. van Schooten), pp. 517–534. Leidein: Johannis Elsevirii.

Jeffreys, H. 1939 *Theory of probability.* Oxford University Press.

Kass, R. E. & Raftery, A. E. 1995 Bayes factors. *J. Am. Statist. Assoc.* **90**, 773–795.

Laplace, P. S. 1774 Memoire sur la probabilite des causes par les evenemen. In *Memoires de l'Academie Royale des Sciences presentes par divers savans*, pp. 621–656.

Legendre, A. M. 1805 Nouvelles méthodes pour la détermination des orbites des comètes. (Appendix: sur la méthode des moindres carrés). Paris: Courcier.

Lindley, D. V. 1965 *Introduction to probability and statistics from a Bayesian viewpoint.* Cambridge University Press.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. 1953 Equations of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087–1091.

Neyman, J. & Pearson, E. S. 1933 On the problem of the most efficient tests of statistical hypotheses. *Phil. Trans. R. Soc. Lond.* A **231**, 289–337.

Pearson, K. 1896 Mathematical contributions to the theory of evolution. III. Regression, heredity and panmixia. *Phil. Trans. R. Soc. Lond.* A **187**, 253–318.

Pearson, K. 1900 On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Phil. Mag.* **50**, 157–175.

Pearson, K. & Filon, L. N. G. 1898 Mathematical contributions to the theory of evolution. IV. On the probable errors of frequency constants and on the influence of random selection on variation and correlation. *Phil. Trans. R. Soc. Lond.* A **191**, 229–311.

Savage, L. J. 1954 *The foundations of statistics*. Wiley.

Schwarz, C. J. & Seber, G. A. F. 1998 A review of estimating animal abundance. III. *Statist. Sci.* **14**, 427–456.

Spiegelhalter, D. J., Thomas, A., Best, N. G. & Gilks, W. R. 1996 *BUGS: bayesian inference using Gibbs sampling*, version 0.50. Cambridge: MRC Biostatistics Unit.

Yates, F. & Cochran, W. G. 1938 The analysis of groups of experiments. *J. Agric. Sci.* **28**, 556–580.

Yule, G. U. 1897 On the theory of correlation. *J. R. Statist. Soc.* **60**, 812–854.

AUTHOR PROFILE

S. P. Brooks

Born in Crawley, Sussex, Steve Brooks studied Mathematics at the University of Bristol before moving to the University of Kent, Canterbury (UKC), to take his MSc in Statistics. He worked as a research assistant at UKC for a further year before moving to the University of Cambridge, where he gained his PhD in 1996. He returned to the University of Bristol as a lecturer before moving to the University of Surrey as a senior lecturer in 1999. Aged 33, he is now a reader in statistics at the University of Cambridge and an EPSRC advanced research fellow. Scientific interests include Bayesian and computational statistics and their applications in biology, ecology, archaeology, economics and engineering. His main recreation activity is scuba diving both in the UK and abroad.