Philippine Statistician, 2005
Vol. 54, Nos. 1-4, pp. 21-38

An Introductory Tutorial To Bayesian Statistics

Balgobin Nandram[1]

ABSTRACT

Bayesian statistics is considerably different from non-Bayesian statistics, primarily because in Bayesian statistics the parameters are stochastic while in non-Bayesian statistics the parameters are not. This stochastic nature allows Bayesians to input prior information in a coherent manner using probabilistic methods, and to make interpretations that are desirable even by non-Bayesians. The extra step in Bayesian statistics is the assignment of distributions for the stochastic parameters. This can be obtained if there is prior information (informative prior) or no prior information (non-informative prior). The process that produces the data are the same in both frameworks. All inference is obtained by exploiting the posterior density, obtained through an application of Bayes' theorem. In most practical applications, the posterior distribution is complex, and there is a need for computation usually achieved through Markov chain Monte Carlo methods. This paper is an introductory tutorial in which I review basic procedures in Bayesian statistics.

Key Words: Bayes' theorem; Credible interval; Gibbs sampler; Model choice and assessment; Proper prior.

I. INTRODUCTION

Lindley (1983) stated that "Bayesian statistics is based on one, simple idea: the only satisfactory description of uncertainty is by means of probability. We are, all of us, surrounded by uncertainty: it plays a dominant role in all our lives. The Bayesian paradigm provides, in probability, a powerful tool for understanding, manipulating and controlling this pervasive, and often unpleasant, feature of our appreciation of our environment. The practical import is immediate: any unknown quantity should be described probabilistically." Indeed, this statement, made by one of the fathers of Bayesian statistics, is profound.

Suppose your interest is on a quantity, or set of quantities, θ and some data D are available, and D has some relationship on the uncertainty of θ, via a probability density (or mass) function, $p(D|\theta)$. (Here D is known but θ is unknown.) The Bayesian view says that the appropriate description of your knowledge of θ in the presence of D is by the probability of θ, given D, and written as $\pi(\theta|D)$. Lindley (1983) stated that the Bayesian paradigm can be thought of as the following recipe:

(1) What is uncertain and of interest to you? Call it θ and summarize the information about θ in a probability distribution, denoted by $\pi(\theta)$, called the *prior distribution.*

(2) What data do you have that bear a relationship with θ? Call it D, and specify the distribution D given θ, namely the conditional

[1] Professor at the Department of Mathematical Sciences, Worcester Polytechnic Institute, 100 Institute Road, Worcester MA 01609, USA. E-mail address : balnan@wpi.edu

distribution $p(D|\theta)$. Viewed as a function of θ this is the *likelihood function.*

(3) Then calculate $\pi(\theta|D)$, called the *posterior distribution.*

The calculation in (3) is done using the rules of probability. All these rules start with the Axioms of Probability (e.g., the Kolmorgorov probability space). Then all information about θ resides in $\pi(\theta|D)$, and all inference proceeds by examining this distribution using summary statistics or preferably graphical methods (i.e., plotting the posterior distribution).

An essential feature of the scientific method is the collection of data D, preferably by controlled experimentation or designed surveys, or alternatively by observational studies. Here randomization is of key importance to ensure data of high quality are obtained. Other data (e.g., observational studies) are also covered, but just as in non-Bayesian statistics inference is limited by the data collection mechanisms.

Bayes' theorem states that

$$\pi(\theta|D) \propto p(D|\theta)\pi(\theta).$$

That is, the *posterior* is proportional to *prior* × *likelihood.* There is nothing subjective about Bayes' theorem; any elementary text book in probability shows how Bayes' theorem follows directly from the Axioms of Probability together with the basic notion of conditional probability.

Of fundamental importance is the likelihood principle, and it states that the totality of information about θ provided by D, is given by the likelihood function of θ for the observed data, D. The principle requires consideration of a unique D, that is observed, but all possible values of θ. The problem of non-Bayesian method is in the violation of this principle. By considering data values that might have occurred but did not, as with a tail area (significance test or confidence interval), non-Bayesians become incoherent because they imagine "hypothetical" repetitions of an experiment. It is only after the data are observed that the principle applies. Indeed, almost all situations ultimately call for a judgment about a unique occasion and it is a great strength of the Bayesian view that it can handle them. Clearly, this is a weakness of the non-Bayesian view. Note that the notion of a sufficient statistic (non-Bayesian) is important because one automatically conditions on the sufficient statistic in the likelihood function.

Thus, in the Bayesian paradigm, there is a prior distribution on the parameters, say θ. This prior distribution is assigned by the user. There may be substantial prior information, and so *informative* prior distributions are used. In other situations, virtually no information might exist, and then the user chooses *non-informative* prior distributions. Strictly speaking, a non-informative prior is one that is constant on the support of the likelihood function. Non-informative priors are typically improper. Letting $p(\theta)$ be a prior distribution on $(-\infty,\infty)$, then $p(\theta)$ is *proper* if $\int_{-\infty}^{\infty} p(\theta)d\theta < \infty$; otherwise $p(\theta)$ is *improper.* It is important to check that the posterior density is proper when there are improper priors; otherwise inference may be unreliable.

I discuss five major aspects of Bayesian statistical procedures in this paper. First, *prior construction* is discussed. The most important issue in Bayesian statistics is what is the appropriate prior. Whether an informative prior or a noninformative prior is needed is an important issue. Once a prior is selected, by using the rules of probability, the posterior density follows mechanically. Second, how to summarize the posterior density using *credible intervals* is discussed. When a posterior density is obtained, the best way to proceed with the analysis is to present a picture of the posterior density. With many parameters this is generally inconvenient and difficult. An alternative procedure is to present the posterior mean and the posterior standard deviation, but these are not adequate when the posterior density is skewed. Thus, a good standard alternative is to present a 95% credible interval. Third, hypothesis testing using the *Bayes factor* is discussed. Bayesians use test of hypotheses, based on a probabilistic structure. Bayesians view hypotheses as models, and models are random (i.e., there is a probability mass function over the set of models). Fourth, the *Gibbs sampler*, the workhorse for Bayesian computations, is discussed. In most practical problems much computation is needed, and this computation is generally based on Markov chain Monte Carlo methods with algorithms such as the Gibbs sampler. These algorithms are iterative; they need a "burn-in" period and a Markov sequence of multivariate vectors is the output. The Gibbs sampler is routinely used for such computation, and it can be adopted to perform the computation in almost any application, although there are more efficient samplers in more complex problems. Finally, model choice and assessment are discussed. In any scientific investigation, it is usually important to select a model, and to assess how well the selected model works. Bayesians have modified non-Bayesian methods such as those based on *cross-validation* and *deviance* analyses.

This paper has six more sections. In Sections 2, 3, 4, 5, 6, I discuss prior construction, credible intervals, hypothesis testing, Markov chain Monte Carlo methods with special emphasis on the Gibbs sampler, and model choice and assessment. Section 7 has a discussion which highlights my experience in Bayesian statistics.

II. PRIOR CONSTRUCTION

In the absence of information, a Bayesian chooses a reference prior (i.e., a type of noninformative prior that any other Bayesian might use) because this choice reduces subjectivity. In fact, these prior distributions might be called objective or noninformative and are chosen with respect to the principle of invariance (e.g., Jeffreys' prior). One characteristic of these distributions is that the integral over the whole parameter space does not exist, thereby making them noninformative. The procedure for constructing Jeffreys' objective prior is to use a prior distribution $p(\theta) \propto \sqrt{|I(\theta)|}$ where $|I(\theta)|$ is the determinant of Fisher's information. Suppose $x_1,...,x_{n_k} \mid \mu,\sigma^2 \overset{i.i.d.}{\sim}$ Normal(μ,σ^2). If μ and σ^2 are treated independently, $p(\mu,\sigma^2) \propto 1/\sigma^2$ (i.e., $p(\mu)=1$ and $p(\sigma^2) \propto 1/\sigma^2$); otherwise $p(\mu,\sigma^2) \propto 1/(\sigma^2)^{3/2}$. The difference between these two priors is negligible, but generally one uses $p(\mu,\sigma^2) \propto 1/\sigma^2$.

One might use other kinds of noninformative priors, called (*proper diffuse priors*), in practice, especially in more complex problems when it may be difficult to calculate Fisher's information. Two simple examples are now described. In the simple problem,

$\mu \sim \text{Normal}(\mu_0, {\sigma_0}^2)$, where $\mu_0 = \bar{x} = \sum_{i=1}^n x_i/n$ and $\sigma_0^2 = 100s^2$ with $s^2 = \sum_{i=1}^n (x_i - \bar{x})^2/(n-1)$; the scale factor of 100 makes it close to a noninformative prior but still proper. Also one can take $\sigma^{-2} \sim \text{Gamma}(a,b)$, where $a = b = .001$ for a proper but almost noniformative prior distribution; the use of this prior distribution is controversial (Gelman 2004). However, one can use the prior density $p(\sigma^2) = a/(a+\sigma^2)^2, \sigma^2 > 0$, which is proper, has median at a, and has none of the moments. One needs to specify a, but typically one can take $a = 1$. These proper diffuse priors are generally used by WinBUGS (see Cowles, 2004).

In general, for scale parameters, say like σ^2, $p(\sigma^2) \propto 1/\sigma^2$ and for location parameter like μ, $p(\mu) = 1$. For a parameter like ρ, in the interval (0,1) one would take $\rho \sim \text{Uniform}(0,1)$, a proper prior. The notion here is that the probabilities of being in intervals on the support of the same widths are the same; hence such prior distributions are noniformative. However, when noniformative priors are used, one must be cautious about propriety in the posterior distribution. Although it is frequently the case that the posterior distribution is proper with improper prior, it is possible to have improper posterior distribution, unknown to the user. Thus, whenever noninformative (improper) priors are used, one needs to prove propriety, an important mathematical exercise.

One might have genuine prior information, and in that case such prior distributions lead to large gains in precision. Suppose that $x_1, ..., x_{n_k} \mid p \overset{i.i.d.}{\sim}$ Bernoulli(p). It is possible to choose a *conjugate* prior density for p, in which the prior and the posterior are in the same family (i.e., as a function of p the prior and the likelihood have the same forms). Thus, we take $\rho \sim \text{Beta}(a,b)$, where one must specify a and b, and the posterior density is $p \mid \mathbf{x} \sim \text{Beta}(s+a, n-s+b)$, where $\mathbf{x} = (x_1, ..., x_n)'$ with $s = \sum_{i=1}^n x_i$. With no prior information, one can take $a = b = 1/2$ for Jeffreys' prior or $a = b = 1$, a uniform prior, both proper. Although these are almost the same, there may be important difference in some situations (e.g., for calculating a Bayes factor or when integrating over the prior distribution). In an important practical situation, a scientist might expect p somewhere between p_0 and p_1 with high confidence. Letting $p_m = (p_1 + p_0)/2$ and $p_s = (p_1 - p_0)/2$, by equating moments with the Beta(a,b) density, one can take $a \approx p_m^2(1-p_m)/p_s^2$ and $b \approx p_m(1-p_m)^2/p_s^2$.

III. CREDIBLE INTERVALS

Let $f(\theta \mid \mathbf{d})$ denote the posterior density of a parameter θ given data $\mathbf{d}$. An interval (a,b) is called a $100(1-\alpha)\%$ credible interval if its posterior probability content is $1-\alpha$, that is, $\int_a^b f(\theta \mid \mathbf{d}) d\theta = 1-\alpha$. Thus, credible intervals are not unique, but they are still useful.

There are two ways to construct credible intervals: numerical and sampling-based. First, let $F(\theta \mid \mathbf{d}) = \int_{-\infty}^{\theta} f(t \mid \mathbf{d}) dt$ be the cumulative distribution function (cdf). Let $F^{-1}(. \mid \mathbf{d})$

be the inverse cdf. Then $a = F^{-1}(\frac{\alpha}{2} | \mathbf{d})$ and $b = F^{-1}(1-\frac{\alpha}{2} | \mathbf{d})$ give the $100(1-\alpha)\%$ credible interval (a,b). Second, draw a random sample of 1,000 values from $f(\theta | \mathbf{d})$. Arrange the values in ascending order, $\theta^{(1)} < \theta^{(2)} < ... < \theta^{(1000)}$. Then an estimator obtained from these order statistics of the 95% credible interval is $(\theta^{(25)}, \theta^{(976)})$. This method is usually used in complex problems, and it works well for large samples (i.e., about 1000). In Bayesian analysis, we can obtain a sample as large as we please (subject to the capability of the computer).

Not only should we be concerned with the probability content of the interval, but we wish to use the interval with the highest posterior density, whenever it is possible. A $100(1-\alpha)\%$ credible interval (a,b) is the highest posterior density (HPD) interval if for any $\theta_1 \in (a,b)$ and $\theta_2 \notin (a,b)$, $f(\theta_1 | \mathbf{d}) \geq f(\theta_2 | \mathbf{d})$. In other words, the height of any point of the density within the HPD interval is greater than for any point outside the interval.

The $100(1-\alpha)\%$ HPD interval is unique for any unimodal posterior density. If the mode is on a boundary of the parameter space, then that boundary is one of the end points of the interval. The $100(1-\alpha)\%$ HPD interval is the shortest interval with $100(1-\alpha)\%$ coverage. For a unimodal posterior density the $100(1-\alpha)\%$ HPD interval is obtained by using a single equation. Let $F(a | \mathbf{d}) = \int_{-\infty}^{a} f(\theta | \mathbf{d}) d\theta$. If $f(\theta | \mathbf{d})$ is a unimodal posterior density with its mode on the lower boundary B, the interval is $\int_B^a f(\theta | \mathbf{d}) d\theta = 1-\alpha$, or simply $(B, F^{-1}(a | \mathbf{d}))$. If $f(\theta | \mathbf{d})$ is a unimodal posterior density with mode not on the boundary, $f(a |)\mathbf{d}) = f(b | \mathbf{d})$, $F(a | \mathbf{d}) - F(b | \mathbf{d}) = \int_a^b f(\theta | \mathbf{d}) d\theta = 1-\alpha$, and solving in terms of a,

$$f(a | \mathbf{d}) = f(b | \mathbf{d}) = f(F^{-1}[F(a) + (1-\alpha)]).$$

For a symmetric density, the equal ordinate condition guarantees equal tails, and therefore, the HPD interval is the same as the credible interval with equal tails.

I provide a simple example with an illustration. Let $x_1, ..., x_{n_k} | \mu, \sigma^2 \overset{i.i.d.}{\sim}$ Normal(μ, σ^2), and consider the noninformative prior $p(\mu, \sigma^2) \propto \frac{1}{\sigma^2}$. For Bayesians, letting $\bar{x} = \sum_{i=1}^{n} x_i / n$ and $s^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 / (n-1)$, $\frac{\mu - \bar{x}}{s/\sqrt{n}} | \mathbf{x} \sim t_{n-1}$ and for non-Bayesian $\frac{\mu - \bar{x}}{s/\sqrt{n}} | \mu \sim t_{n-1}$, where t_{n-1} is the Student's t density on $n-1$ degrees of freedom. Then, in either case a $100(1-\alpha)\%$ HPD or confidence interval for μ is

$$\bar{x} \pm \frac{s}{\sqrt{n}} t_{n-1, \alpha/2},$$

where $t_{n-1, \alpha/2}$ is the $100(1-\alpha)$ percentile point of the Student's t density on $n-1$ degrees of freedom; see Box and Tiao (1973) for similar examples. As an illustration a sample of 60 young white males from Middlesex, Massachusetts, had their body mass indices (BMI) measured. Their average BMI is 20.38 and their standard deviation BMI is 5.24, and

assuming normality a 95% HPD (or confidence) interval for μ is (19.03,21.73); note that $t_{59,0.025} = 2.0003$. While Bayesians have $Pr(19.03 \le \mu \le 21.73 \mid \mathbf{x}) = 0.95$, the statement $Pr(19.03 \le \mu \le 21.73) = 0.95$ cannot be made by non-Bayesians although they would like to do so!

HDP regions can be constructed for multivariate parameters (e.g., see Box and Tiao 1973), and simultaneous intervals can be constructed for many parameters (e.g., see Besag, Green, Higdon, and Mengersen 1995).

IV. HYPOTHESIS TESTING

Bayesians think about hypotheses as models, and so for testing one hypothesis versus another, two models are actually compared. In non-Bayesian statistics, unknown to the investigator one hypothesis is correct and this is always true regardless of the data, and the null and the alternative hypotheses enter the problem asymmetrically. In Bayesian analysis, these hypotheses (or models) enter the problem symmetrically (i.e., it does not matter which is the null hypothesis and which is the alternative hypothesis). Bayesians have uncertainty about these models and a priori they put a "lump" of probability on each model (or hypothesis). The evidence for one model or the other is measured by the *Bayes factor*.

Let M_1 and M_2 be the two competing models. A priori $P(M_1) = 1 - P(M_2)$, and these probabilities are specified by the user. Then, by Bayes' theorem a posteriori, $P(M_1 \mid \mathbf{x}) = 1 - P(M_2 \mid \mathbf{x})$, the prior odds of M_1 to M_2 are $P(M_1)/P(M_2)$, and the posterior odds are $P(M_1 \mid \mathbf{x})/P(M_2 \mid \mathbf{x})$. Finally, the Bayes Factor (BF) is

$$Bf = \frac{\text{Posterior Odds}}{\text{Prior Odds}}.$$

Here BF is interpreted as the evidence provided by the data for M_1 beyond that provided by the prior. We note that the Bayes factor is related to the *marginal likelihoods* of M_k,

$$f(\mathbf{x} \mid M_k) = \int_{\Theta} f(\mathbf{x} \mid \theta)\pi_{M_k}(\theta)d\theta, \; k = 1,2,$$

and $\pi_{M_k}(\theta)$ must be proper for sensible calibration. Note that if θ is a point mass at θ, then $f(\mathbf{x}) = f(\mathbf{x} \mid \theta)$. Using Bayes' theorem, it is easy to show that the Bayes factor is the ratio of the marginal likelihoods, $f(\mathbf{x} \mid M_1)/f(\mathbf{x} \mid M_2)$. As you can see integrating over the parameter space could be an enormous task.

Kass and Raftery (1995) gave a "rule of thumb" to judge the strength of evidence for M_1 in Table 1.

Table 1: Rule of thumb of strength of evidence for M_1

Bf	$\log_e$(BF)	Evidence
$1 \le Bf < 3$	0 – 1.099	little
$3 \le Bf < 20$	1.099 – 2.996	positive
$20 \le Bf < 150$	2.996 – 5.011	strong
$Bf \ge 150$	5.011 –	very strong

Let us consider a simple example. Consider testing for association in a $r \times c$ categorical table with cell counts $n_{jk}, j = 1,...,r, k = 1,...,c$. We fit a multinomial-Dirichlet model under association (as) and under no association (nas), Both models have $\mathbf{n} \mid \mathbf{p} \sim$ Multinomial$(n, \mathbf{p})$. The model with association has $\mathbf{p} \sim$ Dirichlet(1,...,1). Note that by no association we mean that $p_{jk} = p_j^{(1)} p_k^{(2)}, j = 1,...,r, k = 1,...,c$, where $\sum_{j=1}^{r} p_j^{(1)} = 1$ and $\sum_{k=1}^{c} p_k^{(2)} = 1$. A priori, we take $\mathbf{p}^{(1)} \sim$ Dirichlet(1,...,1) and independently $\mathbf{p}^{(2)} \sim$ Dirichlet(1,...,1) where $\mathbf{p}^{(1)}$ and $\mathbf{p}^{(2)}$ have r and c components respectively. It is easy to show that the marginal likelihoods are

$$p_{\text{as}}(\mathbf{n}) = (rc-1)!n!/(n+rc-1)!$$

and

$$p_{\text{nas}}(\mathbf{n}) = p_{\text{as}}(\mathbf{n}) \frac{(r-1)!(c-1)!}{(rc-1)!} \frac{(n+rc-1)!}{(n+r-1)!(n+c-1)!} \frac{\prod_{j=1}^{r} n_{j\cdot}! \prod_{k=1}^{c} n_{\cdot k}!}{\prod_{j=1}^{r} \prod_{k=1}^{c} n_{jk}!}.$$

Of course, the standard Pearson chi-squared statistic can be used, but when the assumptions of the multinomial distribution are violated, one alternative is to construct appropriate models and use the Bayes factor.

Consider our data in Table 2 on a 3×3 categorical table of bone mineral density (BMD) and family income (FI). Under independence (i.e., no association) the observed chi-squared statistic is 12.7 on 4 degrees of freedom with a p-value of .013, and no association is rejected. The marginal likelihoods are $p_{\text{nas}}(\mathbf{n}) = -46.2$ and $p_{\text{as}}(\mathbf{n}) = -49.6$ resulting in a log Bayes factor of 3.40 for evidence of no association relative to association. Therefore, while the chi-squared test provides strong evidence against no association, the log Bayes factor provides strong evidence for no association. Thus, there is a contradictory evidence for no association. The Pearson chi-squared statistic is dominated by cells (3, 1) and (3, 3) with squares of the Pearson residuals being 4.61 and 6.15 respectively (the observed chi-squared statistic is 12.7). We have also collapsed the two categories, osteopenia and osteoporosis, into a single category forming a 2×3 categorical table. For this 2×3 categorical table, the chi-squared test statistic is 1.7 on 2 degrees of freedom with a p-value of .42. The marginal likelihoods are $p_{\text{nas}}(\mathbf{n}) = -28.2$ and $p_{\text{as}}(\mathbf{n}) = -32.0$ resulting in a log Bayes factor of −3.81. Therefore, both tests suggest no association for this 2×3 table.

Table 2. Classification of bone mineral density (BMD) and family income (FI) for 1,844 white females, at least 20 years old (20+)

	FI		
BMD	1	2	3
1	621	290	284
2	260	131	117
3	93	30	18

NOTE: BMD: $1(>0.82g/cm^2$; normal), $2(>0.64,\leq 0.82g/cm^2$; osteopenia), $3(\leq 0.64g/cm^2$; osteoporosis); FI: $1(<\$20{,}000)$, $2(\geq \$20{,}000,<\$45{,}000)$, $3(\geq \$45{,}000)$; BMD is only measured for age 20+.

Finally, we note that the Bayes factor may be sensitive to prior specifications, especially when there are not enough data to estimate the parameters under test; see Sinharay and Stern (2002) for an interesting discussion on nested models. How sensitive is the Bayes factor to the choice of the prior distributions? First, note that the prior density that any reasonable person might use in this problem is the Dirichlet distribution because the prior density and the posterior density are both Dirichlet (i.e., within a *conjugate family* the arithmetic is simple). For the model with association we have selected the prior distributions to be $\mathbf{p} \sim \text{Dirichlet}(\gamma)$ and for the model with no association $\mathbf{p}^{(1)} \sim \text{Dirichlet}(\alpha)$ and independently $\mathbf{p}^{(2)} \sim \text{Dirichlet}(\beta)$. Let $\mathbf{n}_{j\cdot}^{(1)} = \sum_{k=1}^{c} n_{jk}, j = 1,...,r$ and $\mathbf{n}_{\cdot k}^{(2)} = \sum_{j=1}^{r} n_{jk}, k = 1,...,c$. Then, it is easy to show that the Bayes factor for a test of association versus no association is

$$BF = \frac{D_{rc}(\mathbf{n}+\gamma)/D_r(\mathbf{n}_{\cdot}^{(1)}+\alpha)D_c(\mathbf{n}_{\cdot}^{(2)}+\beta)}{D_{rc}(\gamma)/D_r(\alpha)D_c(\beta)},$$

where $D_r(u)$ refers to the Dirichlet function with components, $u_1,...,u_r$ etc. Then, we choose each of the components of α, β and γ to be κ (e.g., in $p_{as}(\mathbf{n})$ and $p_{nas}(\mathbf{n})$, $\kappa = 1$). Sensitivity to the choice of prior distributions can be studied in terms of κ. Here $\kappa = 1$ corresponds to the prior distributions that are usually used in the multinomial-Dirichlet model, and $k = .50$, the Jeffreys' prior. Thus, we have chosen $\kappa = .25, 0.5, 1., 1.5, 2, 3$, and the corresponding Bayes factors (log scale) are $4.7, 3.6, 3.4, 3.9, 4.7, 6.6$. Thus, while the Bayes factor is sensitive to the choice of the prior distributions, it is not too sensitive. Of course, if there is informative prior information, in which κ is substantially large, it is a different issue.

V. MARKOV CHAIN MONTE CARLO METHODS

The Gibbs sampler is a Monte Carlo integration method which proceeds by a Markovian updating scheme. It was developed formally by Geman and Geman (1984) in the context of image restoration. In the statistical framework, Tanner and Wong (1987) used essentially this algorithm in their substitution sampling approach to missing data problems. Gelfand and Smith (1990) brought to the attention of Bayesians how applicable the Gibbs sampler is to general parametric Bayesian computations; but see Casella and George (1992) for a simple explanation.

We summarize the main features of the Gibbs sampler. Suppose that we have a collection of p (possibly vector-valued) random variables $\theta_1,...,\theta_p$, and samples may be generated by some method, given values of the appropriate conditioning random variables from the full conditional distributions, denoted generically by $f(\theta_s \mid \theta_1,...,\theta_{s-1},\theta_{s+1},...,\theta_p), s=1,...,p$. Under mild conditions (see Besag (1974)), these conditional distributions uniquely determine the full joint distribution $f(\theta_1,...,\theta_p)$, and hence all marginal distributions $f(\theta_s)$, $s=1,...,p$. The Gibbs sampler generates samples from the joint distribution as follows. Given an arbitrary starting set of values $\theta_1^{(0)},...,\theta_p^{(0)}$, draw $\theta_1^{(1)}$ from $f(\theta_1 \mid \theta_2^{(0)},...,\theta_p^{(0)})$, then $\theta_2^{(1)}$ from $f(\theta_2 \mid \theta_1^{(1)},\theta_3^{(0)},...,\theta_p^{(0)})$, and so on up to $\theta_p^{(1)}$ from $f(\theta_p \mid \theta_1^{(1)},...,\theta_{p-1}^{(1)})$ to complete one iteration of the scheme. After t such iterations one obtains $(\theta_1^{(t)},...,\theta_p^{(t)})$. Geman and Geman (1984) show under mild conditions that this p-tuple converges in distribution to a random observation from $f(\theta_1,...\theta_p)$ as $t \to \infty$.

The iterates before convergence occurs, called the "burn in," are discarded, because these iterates are not from a stable process. One can then "thin out" the iterates to obtain a random sample from the joint posterior density. This is monitored using the autocorrelation function. Thus, a random sample $\theta^{(h)}, h=1,...,M$, where M is chosen to give the highest precision to the estimates of the posterior densities, can be obtained. One can study posterior summaries (mean and standard deviation) to see what value of M one needs for convergence of the estimates, not the Gibbs sampler itself, or rely on a reasonable sample size calculation.

Gelfand and smith (1990) recommend a density estimate of the form

$$\hat{f}(\theta_s) = \sum_{h=1}^{M} f(\theta_s \mid \theta_r^{(h)}, r=1,...,p, r \neq s)/M.$$

This is a discrete mixture distribution, and is, in fact, a Monte Carlo integration to accomplish the desired marginalization. Gelfand and Smith (1990) call this process Rao-Blackwellization, after the well-known Rao-Blackwell theorem in non-Bayesian statistics. There may also be interest in a function of the parameters, say $W(\theta_1,...,\theta_p)$. Each p-tuple $(\theta_1^{(h)},...,\theta_p^{(h)})$ provides an observed $W^{(h)} \equiv W(\theta_1^{(h)},...,\theta_p^{(h)})$ whose marginal distribution is approximately $f(W)$. A Rao-Blackwellized density estimator can also be obtained. If θ_s actually appears as an argument of W, the complete conditional density of

$W(\theta_s \mid \theta_r, r=1,...,p, r \neq s)$ can be obtained by univariate transformation from that of $\theta_s \mid (\theta_r, r=1,...,p, r \neq s)$.

There is virtually no limit to the dimension, p, and there are generalizations of the Gibbs sampler. One important generalization is the Metropolis-Hastings sampler, which is useful when it is difficult to apply the Gibbs sampler; see Chib and Greenberg (1995) for a pedagogical review. There are other tricks that can be used (e.g., see Robert and Casella, 1999).

I will now present a simple example of the implementation of the Gibbs sampler. A useful and flexible type of prior distribution is a *hierarchical* prior; see Good (1980) for discussion and early references. It is often convenient to model the sampling process and one's prior belief (or experience) in stages. Lindley and Smith (1972) introduced the hierarchical Bayesian linear model with three stages for normal data with normal priors, and Smith (1973) extended this work. Consider the following simple model,

$$Y_{k1}, Y_{k2},..., Y_{km_k} \mid \mu_k, \sigma^2 \overset{i.i.d.}{\sim} N(\mu_k, \sigma^2), k=1,...,n,$$

$$\mu_1, \mu_2,..., \mu_n \mid \theta, \delta^2 \overset{iid}{\sim} \mathrm{N}(\theta, \delta^2).$$

This is a one-way random effects model in which there are n groups of individuals with the k^{th} having m_k individuals. This model expresses indifference among the individuals within a group and indifference among the groups. A priori, θ, σ^{-2}, and δ^{-2} are independent with

$$p(\theta)=1 \text{ and } \sigma^{-2}, \delta^{-2} \overset{i.i.d.}{\sim} \mathrm{Gamma}(a/2, b/2),$$

where $a=b=.002$.

Let $\bar{y}_k = \sum_{j=1}^{m_k} y_{kj}/m_k$ and $\lambda_k = \delta^2/(\delta^2 + \sigma^2/m_k)$, the conditional posterior densities required to run the Gibbs sampler are

$$\mu_k \mid \theta, \sigma^2, \delta^2, \mathbf{y}_5 \overset{ind}{\sim} \mathrm{Normal}\{\lambda_k \bar{y}_k + (1-\lambda_k)\theta, (1-\lambda_k)\delta^2\}, k=1,...,n, \quad (1)$$

$$\theta \mid \mu, \sigma^2, \delta^2, \mathbf{y}_5 \sim \mathrm{Normal}\{\sum_{k=1}^{n} \mu_k / n, \delta^2/n\}, \quad (2)$$

$$/n\ \sigma^{-2} \mid \mu, \theta, \delta^2, \mathbf{y}_5 \sim \mathrm{Gamma}\{(a+\sum_{k=5} m_k)/2, (b+\sum_{k=5} \sum_{j=1}^{m_k} (y_{kj}-\mu_k)^2)/2\}, \quad (3)$$

$$\delta^{-2} \mid \mu, \theta, \sigma^2, \mathbf{y}_5 \sim \mathrm{Gamma}\{(a+n)/2, (b+\sum_{k=5} (\mu_k-\theta)^2)/2\}, \quad (4)$$

By starting with any reasonable values of θ, σ^2 and δ^2, one can perform the Gibbs sampler to draw samples from (1), (2), (3) and (4), in turn, iterating the process until convergence. One will have to "burn in" and might have to "thin out" the Gibbs sampler to obtain a "random" sample $(\mu^{(h)}, \theta^{(h)}, \sigma^{2(h)}, \delta^{2(h)}), h=1,...,M$ for a reasonably large value of M ($M \approx 1000$).

VI. MODEL CHOICE AND ASSESSMENT

An important part of a scientific investigation is to select a model from a set of plausible competing models, and a scientist would need to assess how well the selected model fits the data (see Box 1976, 1980). To accomplish these tasks, Bayesians have modified non-Bayesian methods such as those based on cross-validation analyses and deviances.

Let $y_i, i = 1,...,n$ be a sample from $f(y_i \mid \theta)$ and let $f(\mathbf{y} \mid \theta)$ denote the joint probability density (or mass) function of $\mathbf{y}$. Suppose also that there is a prior density $\pi(\theta)$, and the posterior density $\pi(\theta \mid \mathbf{y})$ is proper. Note that for the measures discussed here $\pi(\theta)$ does not have to be proper, an advantage over the Bayes factor. To assess the goodness of fit, one can consider the posterior predictive ordinate

$$PPO_i = f(y_i \mid \mathbf{y}_{(i)}), i = 1,...,n,$$

where $\mathbf{y}_{(i)}$ is the vector of all observations excluding the i^{th} observation. Note that if the y_i are discrete, PPO_i is the posterior predictive probability that y_i takes the observed value given that it is omitted. Also, note that

$$PPO_i = f(\mathbf{y})/f(\mathbf{y}_{(i)}), i = 1,...,n,$$

where $f(\mathbf{y}) = \int f(\mathbf{y} \mid \theta)\pi(\theta)d\theta$ is the marginal likelihood for data $\mathbf{y}$. Thus, PPO_i is the ratio of the marginal likelihood of all the observed values to the marginal likelihood of all the observed values excluding the i^{th} observation. Thus, if the model fits well, a plot of PPO_i versus y_i should show a horizontal line; those points away from the horizontal line do not support the model, and are possible outliers.

One can use a summary measure to compare different models. A reasonable measure is the arithmetic mean of the logarithm of the PPO_i, called the *posterior predictive score*,

$$PPS = \frac{1}{n}\sum_{i=1}^{n} \log(PPO_i)$$

under each model. We can choose the model with the largest PPS.

It is now straight forward to compute the PPO_i using Monte Carlo methods (see Gelfand, Dey and Chang 1992 and Gelfand and Dey 1994). They recommend an estimator, $\tilde{P}PO_i$, of PPO_i as

$$\tilde{P}PO_i = \sum_{k=1}^{M} \omega_i^{(k)} f(y_i \mid \mathbf{y}_{(i)}, \theta^{(k)}), \quad \omega_i^{(k)} = \frac{f(\mathbf{y}_{(i)} \mid \theta^{(k)})/f(\mathbf{y} \mid \theta^{(k)})}{\sum_{k=1}^{M} f(\mathbf{y}_{(i)} \mid \theta^{(k)})/f(\mathbf{y} \mid \theta^{(k)})},$$

$i = 1,...,n$, $h = 1,...,M$. Here $\theta^{(h)}, h = 1,...,M$, are a random sample from the posterior density $\pi(\theta \mid \mathbf{y})$. When $y_1,...,y_n$ are independent,

$$\tilde{P}PO_i = \sum_{k=1}^{M} \omega_i^{(k)} f(y_i \mid \theta^{(k)}), \quad \omega_i^{(k)} = \frac{1/f(y_i \mid \theta^{(k)})}{\sum_{k=1}^{M} 1/f(y_i \mid \theta^{(k)})}, h = 1,...,M, \ i = 1,...,n.$$

One can also define the standardized cross-validation residual as

$$DRES_i = \{x_i - E(x_i \mid \mathbf{x}_{(i)})\}/SD(x_i \mid \mathbf{x}_{(i)}).$$

That is, the i^{th} observed x_i is "held out" and compared with its point estimator, $E(x_i \mid \mathbf{x}_{(i)})$, which is evaluated without using the observed x_i. Measures like $DRES_i$ are used as in non-Bayesian analyses. Nandram, Sedransk and Pickle (2000) use the $DRES_i$ in a summary form, to rank competing models by counting the number of values such that $|DRES_i| \geq q$ (e.g., $q = 3$), which they call "number of outliers." They also use $DRES_i$ to study the models individually.

Using the minimum posterior predictive loss approach (Gelfand and Ghosh 1998), under squared error loss, the deviance is given by

$$D = P + G,$$

where P is penalty for over-fitting or under-fitting and G is a goodness-of-fit measure,

$$P = \sum_i Var(x_i^{rep} \mid \mathbf{x}^{obs}), \quad G = \sum_i \left\{E(x_i^{rep} \mid \mathbf{x}^{obs}) - x_i^{obs}\right\}^2.$$

Here, expectations are taken under

$$f(x_i^{rep} \mid \mathbf{x}^{obs}) = \int f(x_i^{rep} \mid \theta)\pi(\theta \mid \mathbf{x}^{obs})d\theta,$$

where $\mathbf{x}^{obs}$ is the vector of observations and $\mathbf{x}^{rep}$ future draws; see Gelfand and Ghosh (1998) for further discussion. Other related measures based on *expected predictive deviance* are given by Spiegelhalter, Best, Carlin, and Linde (2002).

I now consider a simple illustrative example. Suppose it is required to discriminate between the two models, M_1 and M_2, where for M_1,

$$x_1,...,x_n \mid \mu,\sigma^2 \overset{iid}{\sim} \text{Normal}(\mu,\sigma^2), p(\mu,\sigma^2) \propto 1/\sigma^2$$

and for M_2,

$$x_1,...,x_n \mid \alpha,\beta \overset{iid}{\sim} \text{Gamma}(\alpha,\beta), p(\alpha,\beta) \propto 1/(1+\alpha)^2$$

Note that these two models are related with $\mu = \alpha/\beta$ and $\sigma^2 = \mu/\beta$.

Then, letting $\phi(\cdot)$ denote the standard normal density function, $\bar{x} = \sum_{i=1}^{n} x_i/n$ and $s^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2/(n-1)$, for M_1 it is easy to show that

$$PPO_i^{M_1} = \sum_{h=1}^{M} \omega_i^{(h)} \frac{1}{\sigma^{(h)}} \phi\{\frac{x_i - \mu^{(h)}}{\sigma^{(h)}}\}, \quad \omega_i^{(h)} = [\frac{1}{\sigma^{(h)}} \phi\{\frac{x_i - \mu^{(h)}}{\sigma^{(h)}}\}]^{-1} / \sum_{k=1}^{M} [\frac{1}{\sigma^{(k)}} \phi\{\frac{x_i - \mu^{(k)}}{\sigma^{(k)}}\}]^{-1},$$

where $(\mu^{(h)}, \sigma^{2(h)}), h = 1,...,M$ are a random sample from $\mu \mid \sigma^2, \bar{x} \sim \text{Normal}\left(\bar{x}, \frac{\sigma^2}{n}\right)$ and $\sigma^{-2} \mid s^2 \sim \text{Gamma}\left(\frac{n-1}{2}, \frac{(n-1)s^2}{2}\right)$. Note that the posterior density $p(\mu, \sigma^2 \mid \mathbf{x})$ is proper.

For M_2, letting $\bar{x}_a = \sum_{i=1}^{n} x_i/n$ and $\bar{x}_g = (\prod_{i=1}^{n} x_i)^{1/n}$, we have

$$PPO_i^{M_2} = \sum_{h=1}^{M} \omega_i^{(h)} \beta^{(h)\alpha^{(h)}} x_i^{\alpha^{(h)}-1} e^{-\beta^{(h)} x_i} / \Gamma(\alpha^{(h)}),$$

$$\omega_i^{(h)} = [\beta^{(h)\alpha^{(h)}} x_i^{\alpha^{(h)}-1} e^{-\beta^{(h)} x_i} / \Gamma(\alpha^{(h)})]^{-1} / \sum_{k=1}^{M} [\beta^{(k)\alpha^{(k)}} x_i^{\alpha^{(k)}-1} e^{-\beta^{(k)} x_i} / \Gamma(\alpha^{(k)})]^{-1},$$

where $(\alpha^{(h)}, \beta^{(h)}), h = 1,...,M$ are a random sample from $\beta \mid \alpha, \mathbf{x} \sim \text{Gamma}(n\alpha, n\bar{x}_a)$ and $p(\alpha \mid \mathbf{x}) \propto \{\Gamma(n\alpha)/\Gamma(\alpha)^n n^{n\alpha}\}(\bar{x}_g/\bar{x}_a)^{n\alpha}/(1+\alpha)^2, \alpha > 0$. Note that a sufficient condition for propriety of $p(\alpha, \beta \mid \mathbf{x})$ is that $\alpha^{1/\alpha} < \bar{x}_a/\bar{x}_g$, a condition that is met for many practical problems.

Finally, for M_1, $E(x_i \mid \mathbf{x}_{(i)}) = \sum_{h=1}^{M} \omega_i^{(h)} \mu^{(h)}$, and

$$DRES_i = \frac{x_i - E(x_i \mid \mathbf{x}_{(i)})}{[\sum_{k=1}^{M} \omega_i^{(k)} \{\sigma^{2(k)} + (\mu^{(k)} - E(x_i \mid \mathbf{x}_{(i)}))^2\}]^{1/2}},$$

$i = 1,...,n, h = 1,...,M$. A similar expression holds for M_2. Also, it is easy to write down similar formula for P and G.

In a simple example, the data consist of a sample of the tensile strengths of 33 steel bars; see Nandram (1995) for a description of the data. A 95% credible interval for μ under the normal model is $(89.4, 91.7)$ and under the gamma model is $(89.3, 91.7)$; also a 95% credible interval for α is $(397, 1098)$. The $DRES_i$ and the PPO_i for the two models are respectively very similar; the number of $DRES_i > 1$ is 10 for both models, and the number of $DRES_i > 2$ is 0 for both models. Thus, the models are very similar. Can any differences between these models be identified? In Table 3 we present the penalty (P), the goodness-of-fit (G), and the deviance (D).

Table 3: Comparison of the normal and the gamma models using the penalty (P), goodness-of-fit (G) and deviance (D)

Model	P	G	D
Normal	393	347	740
Gamma	421	347	768

NOTE: Three significant digits are used, and the analysis is based on a sample of 10,000. The goodness-of-fit measures for the two models are very similar, the penalty, and therefore the deviance, is smaller for the normal model. Also, PPS is -2.57 for the normal model and -2.62 for the gamma model. Thus, we conclude that the normal model is preferred.

VII. DISCUSSION

I have given a introductory tutorial to Bayesian statistics. In this final section, I will discuss my experience in Bayesian statistics.

Bayesian statistics is ubiquitous in applications too numerous to mention here. But, see Nandram, Sedransk and Smith (1997) for an application on the size of a fish stock, and Nandram, Sedransk and Pickle (2000) for an application on chronic obstructive pulmonary disease. Bayesian statistics is to be preferred over non-Bayesian statistics because once a sensible model is written down, it can be accurately analyzed, even if one needs to use Monte Carlo methods. Not too long from now, it is believed that non-Bayesian statistics will give way to Bayesian statistics at least in applications. Of course, non-Bayesian statistics, in the form of probability calculations, is necessary. Markov chain Monte Carlo methods, the state of the art, have emerged as standard Bayesian computational methods. Enormous progress has been made by Bayesians over the past fifteen years, and with the WinBUGS software (see Cowles 2004 for a review of WinBUGS 1.4) many applied statisticians and scientists are now invading the area. In my work, I use Fortran and SAS extensively, and I believe that a researcher in Bayesian statistics (methodology) should use a high level language such as Fortran for computations.

Bayesian methods can be learnt by taking courses in Mathematical Statistics (yes, non-Bayesian!), Numerical Analysis and a few courses in Bayesian Statistics. I believe that every graduate student in statistics should have a course in Bayesian statistics. Of course, one should take courses in linear models, categorical data, time series, survey sampling; the more you know, the better it is. As a graduate student, I took a two-course sequence in Bayesian Statistics, taught by Professor James Dickey when he was at the State University of New York at Albany. He used De Groot (1970), Box and Tiao (1973) and some of his own lecture notes. In one of the courses I made a presentation on Smith (1973). Earlier I took a Bayesian course, taught by Dr. Ann Mitchell of Imperial College, when I was in the Master's Program in Statistics; no text book was recommended. When I was a doctoral student at the University of Iowa, I made two lengthy Bayesian presentations. One was in an advanced course in Multivariate Analysis, taught by Professor George Woodworth, in which I made a presentation on "Some aspects of multivariate analysis," Box and and Tiao (1973, Ch 8). The other was in the seminar in Order Restricted Inference, taught by Professor Tim Robertson, in which I made a presentation on Sedransk, Monahan and Chiu (1985). But most of what I have learnt in Bayesian Statistics came from its use in my research and the Bayesian course I

teach at Worcester Polytechnic Institute (WPI). When I teach Bayesian Statistics, I use the text book by Gelman, Carlin, Stern and Rubin (2004), and those of Congdon (2001), Box and Tiao (1973), De Groot (1970) and others for reference. I recommend Box and Tiao (1973) highly, but it is not appropriate for my course at WPI.

Bayesian statistics is an area that has received much attention in the past fifteen years, and there is still a lot that can be done in the following areas: (a) Construction of appropriate prior distributions, objective ones when there is no prior information; (b) Full Bayesian methods to assess model fit accurately; (c) Nonparametric methods without relying heavily on likelihoods; and (d) Computation methods when there are very awkward likelihood functions and a Metropolis-Hastings sampler is not feasible. For example, in survey sampling (the area in which I work mostly), when there is nonresponse or selection bias, prior construction is not straightforward and the posterior densities cannot be sampled easily using a Metropolis-Hastings sampler. In this case (a)-(d) are useful. Nowadays data are collected with all kinds of problems (e.g., nonidentifiability issues, key information is missing, correlation in data values), and one needs to input extra information beyond the data. Good models can help.

ACKNOWLEDGMENT

I am very pleased to present this paper to the Philippine Statistician, responding to an invitation from Dr. Jose Ramon Albert, President, Philippine Statistical Association and past Editor, the Philippine Statistician. This invitation came through my work primarily with Dr. Corinne Burgos at De La Salle University and the University of the Philippines.

References

BESAG, J. (1974). "Spatial interaction and the statistical analysis of lattice systems" (with discussion). *Journal of the Royal Statistical Society*, Series B, 36, 192-326.

BESAG, J., GREEN, P., HIGDON, D., and MENGERSEN, K. (1995). "Bayesian computation and stochastic systems" (with discussion). *Statistical Science*, 10, 3-66.

BOX, G. E. P. (1980). "Sampling and Bayes' inference in scientific modeling and robustness." *Journal of the Royal Statistical Society*, Series A, 143, 383-430.

BOX, G. E. P. (1976). "Science and statistics." *Journal of the American Statistical Association*, 71, 791-799.

BOX, G.E.P and TIAO, G.C. (1973). *Bayesian Inference in Statistical Analysis.* Wiley: New York.

CASELLA, G. and GEORGE, E.I. (1992). "Explaining the Gibbs sampler." *The American Statistician*, 46, 167-174.

CHIB, S. and GREENBERG, E. (1995). "Understanding the Metropolis-Hastings algorithm." *The American Statistician*, 49, 327-335.

CONGDON, P. (2001). *Bayesian Statistical Modeling.* John Wiley, New York.

COWLES, M.K. (2004). "Review of WinBUGS 1.4." *The American Statistician*, 58, 330-336.

DE GROOT, M. H. (1970). *Optimal Statistical Decisions.* Mc Graw-Hill, New York.

GELFAND, A. E., and DEY, D. K. (1994). "Bayesian model choice: Asymptotics and exact calculations." *Journal of the Royal Statistical Society*, Series B, 56, 501-514.

GELFAND, A. E., DEY, D. K. and CHANG, H. (1992). "Model determination using predictive distributions with implementation via sampling-based methods." In *Bayesian Statistics 4*, (ed. J.M. bernardo, J.o. Berger, A.P. Dawid and A.F.M. Smith), pp. 147-167. Oxford: Oxford University Press.

GELFAND, A., and GHOSH, S. (1998). "Model choice: A minimum posterior predictive 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.

GELMAN, A. (2004). "Prior distribution for variance parameters in hierarchical models." *Bayesian Analysis.*

GELMAN, A.B., CARLIN, J.S., STERN, H.S. and RUBIN, D.B.(2004). *Bayesian Data Analyis.* Chapman and Hall, New York.

GEMAN, S. and GEMAN, G. (1984). "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721-741.

GOOD, I.J. (1980). "Some history of the hierarchical Bayesian methodology." In *Bayesian Statistics II*, J.M. Bernardo, M.H. DeGroot, D.V. Lindley, and A.F.M. Smith (eds.). University Press, Valencia, 489-504.

KASS, R. and RAFTERY, A. (1995). "Bayes factors." *Journal of the American Statistical Association*, 90, 773-795.

LINDLEY, D.V. (1983). "Theory and practice of Bayesian statistics." *The Statistician*, 32. 1-11.

LINDLEY, D.V. and SMITH, A.F.M. (1972). "Bayes estimates for the linear model" (with discussion). *Journal of the Royal Statistical Society*, Series B, 34, 1-41.

NANDRAM, B. (1995). "Bayesian cuboid prediction intervals: An application to tensile-strength prediction." *Journal of Statistical Planning and Inference*, 44, 2, 167-180.

NANDRAM, B., SEDRANSK, J. and PICKLE, L.W. (2000), "Bayesian analysis and mapping of mortality rates for chronic obstructive pulmonary disease," *Journal of the American Statistical Association*, 95, 1110-1117.

NANDRAM, B., SEDRANSK, J. and SMITH, S. J. (1997). "Order restricted Bayesian estimation of the age composition of a population of Atlantic cod." *Journal of the American Statistical Association*, 92, 33-40.

ROBERT, C. P. and CASELLA, G. (1999). *Monte Carlo Statistical Methods.* Springer, New York.

SEDRANSK, J., MONAHAN, J. and CHIU, H. Y. (1985). "Bayesian estimation of finite population parameters in categorical data models incorporating order restrictions." *Journal of the Royal Statistical Society*, Series B, 47, 519-527.

SINHARAY, S. and STERN, H. S. (2002). "On the sensitivity of Bayes factors to the prior distributions." *The American Statistician*, 56, 196-201.

SMITH, A.F.M. (1973). "A general Bayesian linear model." *Journal of the Royal Statistical Society*, Series B, 35, 67-75.

SPIEGELHALTER, D. J., BEST, N. G., CARLIN, B. P. and LINDE, A. V. D. (2002). "Bayesian measures of model complexity and fit." *Journal of the Royal Statistical Society*, Series B, 64, 1-34.

TANNER, M. and WONG, W. (1987). "The calculation of posterior distributions by data augmentation" (with discussion). *Journal of the American Statistical Association*, 82, 528-550.