Statistica Sinica **13**(2003), 625-639

A NONITERATIVE SAMPLING METHOD FOR COMPUTING POSTERIORS IN THE STRUCTURE OF EM-TYPE ALGORITHMS

Ming Tan*, Guo-Liang Tian* and Kai Wang Ng†

*University of Maryland at Baltimore and †The University of Hong Kong

Abstract: We propose a noniterative sampling approach by combining the *inverse Bayes formulae* (IBF), sampling/importance resampling and posterior mode estimates from the Expectation/Maximization (EM) algorithm to obtain an i.i.d. sample approximately from the posterior distribution for problems where the EM-type algorithms apply. The IBF shows that the posterior is proportional to the ratio of two conditional distributions and its numerator provides a natural class of built-in *importance sampling functions* (ISFs) directly from the model specification. Given that the posterior mode by an EM-type algorithm is relatively easy to obtain, a best ISF can be identified by using that posterior mode, which results in a large overlap area under the target density and the ISF. We show why this procedure works theoretically. Therefore, the proposed method provides a novel alternative to perfect sampling and eliminates the convergence problems of Markov chain Monte Carlo methods. We first illustrate the method with a proof-of-principle example and then apply the method to hierarchical (or mixed-effects) models for longitudinal data. We conclude with a discussion.

Key words and phrases: Bayesian computation, data augmentation, EM algorithm, Gibbs sampler, inverse Bayes formulae, MCMC, sampling/importance resampling.

1. Introduction

The EM algorithm (Dempster, Laired and Rubin (1977)) is an iterative deterministic method for finding the *maximum likelihood estimate* (MLE), and is a powerful yet easy to use tool for likelihood inference. It has been especially useful in incomplete data problems. However, in applications with small samples, the likelihood may not be adequately summarized by the MLE and its asymptotic covariance matrix. Although finding the MLE itself is relatively easy, the standard errors in problems involving many parameters are difficult to obtain with the EM algorithm. For example, the method of Louis (1982) involves second order partial derivatives. Therefore, it is appealing to use a Bayesian model with non-informative priors to find the entire posterior of the parameters of interest. The key then is how to calculate the posterior.

Several approaches have been taken to calculate posteriors. Traditionally, accept-reject sampling and *sampling/importance resampling* (SIR) (Rubin (1988)) are used to obtain samples from the posterior. Although both sampling methods generate independent samples, they are problematic in many applications (especially those of high-dimension) in which the envelope function or the *importance sampling function* (ISF) is not readily available, thus limiting their scope of application. For example, although an ingenious development, SIR has rarely been used in practice, partly because of the lack of an efficient generic ISF directly from the model specification of a practical problem.

The great progress in Bayesian computation over the last decade has focused on the Gibbs sampler and in general *Markov chain Monte Carlo* (MCMC) methods (see, e.g., Chen, Shao and Ibrahim (2000)). The Gibbs sampler is appealing for its general applicability and ease of implementation. However, the burden of proof is shifted to the monitoring of stochastic convergence and mixing of the Markov chain, which so far can only be assessed with convergence diagnostics (Robert and Casella (1999), Jones and Hobert (2001)). As pointed out by Gelfand (2002), "in general, convergence can never be assessed, as comparison can be made only between different iterations of one chain or between different observed chains, but never with the true stationary distribution". Because of this problem, intense research has also focused on generating samples distributed *perfectly* as the stationary distribution of the Markov chain (Green and Murdoch (1999), Casella, Lavine and Robert (2001)). Unfortunately, such an algorithm is currently feasible only for limited low-dimensional problems, and the cost of obtaining multiple (n) samples is far greater than that of the usual MCMC, because essentially the entire algorithm must be repeated n times.

The purpose of this article is to develop a noniterative sampling method, as opposed to iterative sampling in MCMC, for computing posteriors based on the *inverse Bayes formulae* (IBF) and SIR to obtain i.i.d. samples approximately from the posterior distribution while utilizing the posterior mode and structure from the EM-type algorithm. The IBF include a pointwise, a sampling-wise and a function-wise version. The sampling-wise IBF (2.6) expresses the observed posterior density as the ratio of the complete-data posterior to the conditional predictive density, up to a normalizing constant, and can be derived readily from the fundamental Bayes Theorem, whereas the pointwise and function-wise IBF give explicit formulae for the observed posterior. Interestingly, the origin of the sampling-wise IBF can be traced back to Hammersley and Clifford (1970) (see also Besag (1974)). It is quite recent, however, that the pointwise version and its implications are explored by Ng (1997). Meng (1996) discussed the usefulness of the pointwise IBF in checking compatibility.

The idea of the IBF sampling in an EM framework is as follows. First we augment the observed data with latent data and obtain the structure of

augmented posterior/conditional predictive distributions as in the EM or the *data augmentation* (DA) algorithm (Tanner and Wong (1987)). Then, in the class of built-in ISFs provided by the sampling-wise IBF, we choose a best ISF by using preliminary estimates from the EM algorithm so that the overlap area under the target density and the ISF is large. Finally the sampling-wise IBF and SIR are combined to generate i.i.d. samples approximately from the observed posterior distribution. We show that the synergism of IBF, EM and SIR creates an attractive sampling approach for Bayesian computation. Since the sampling-wise IBF and the EM share the DA structure, the IBF sampling via the EM does not require extra derivations, and can be applied to problems where the EM is applicable while obtaining the whole posterior.

In Section 2, we propose the IBF sampling method and theoretically justify an optimal choice of ISF. The performance of the IBF sampling is first evaluated in a proof-of-principle example and is then demonstrated with the hierarchical (or mixed-effects) models for longitudinal data in Section 3. Some discussion is presented in Section 4.

2. The IBF Method

For ease of exposition, we adopt the familiar notations in the EM/DA algorithm and focus on the structure of augmented posterior/conditional predictive distributions. Let Y_{obs} denote the observed data and θ the parameter vector of interest. The observed data Y_{obs} is augmented with latent variables (or missing data) Z so that both the augmented (or complete-data) posterior distribution $f_{(\theta|Y_{\text{obs}},Z)}(\theta|Y_{\text{obs}},z)$ and the conditional predictive distribution $f_{(Z|Y_{\text{obs}},\theta)}(z|Y_{\text{obs}},\theta)$ are available. Let $\hat{\theta}_{\text{obs}}$ denote the mode of the observed posterior density $f_{(\theta|Y_{\text{obs}})}(\theta|Y_{\text{obs}})$ and $\mathcal{S}_{(\theta,Z|Y_{\text{obs}})}$, $\mathcal{S}_{(\theta|Y_{\text{obs}})}$ and $\mathcal{S}_{(Z|Y_{\text{obs}})}$ denote the joint and conditional supports of $(\theta,Z)|Y_{\text{obs}}$, $\theta|Y_{\text{obs}}$ and $Z|Y_{\text{obs}}$, respectively. Throughout this paper, we always make two basic assumptions: (a) the observed posterior mode $\hat{\theta}_{\text{obs}}$ (or the MLE $\tilde{\theta}$) is already obtained via the EM algorithm; and (b) the joint support is a *product space*, i.e., $\mathcal{S}_{(\theta,Z|Y_{\text{obs}})} = \mathcal{S}_{(\theta|Y_{\text{obs}})} \times \mathcal{S}_{(Z|Y_{\text{obs}})}$, or equivalently, $\mathcal{S}_{(\theta|Y_{\text{obs}},Z)} = \mathcal{S}_{(\theta|Y_{\text{obs}})}$ and $\mathcal{S}_{(Z|Y_{\text{obs}},\theta)} = \mathcal{S}_{(Z|Y_{\text{obs}})}$. Our goal is to obtain i.i.d samples from the observed posterior distribution $f_{(\theta|Y_{\text{obs}})}(\theta|Y_{\text{obs}})$. To achieve this aim, we propose two versions of IBF sampling because of their different implications in Bayesian computation.

2.1. IBF sampling

The basic idea is that if we can obtain m i.i.d. samples $\{z^{(k_1)},\ldots,z^{(k_m)}\}$ from the marginal predictive distribution $f_{(Z|Y_{\text{obs}})}(z|Y_{\text{obs}})$ and generate $\theta^{(i)}$ from the augmented posterior distribution $f_{(\theta|Y_{\text{obs}},Z)}(\theta|Y_{\text{obs}},z^{(k_i)})$ for $i = 1,\ldots,m$,

then $\theta^{(1)}, \ldots, \theta^{(m)}$ are i.i.d. samples from the observed posterior $f_{(\theta|Y_{\rm obs})}(\theta|Y_{\rm obs})$. Therefore, the key is to be able to generate i.i.d. samples from $f_{(Z|Y_{\rm obs})}(z|Y_{\rm obs})$. This can be achieved by using

$$f_{(Z|Y_{\rm obs})}(z|Y_{\rm obs}) \propto \frac{f_{(Z|Y_{\rm obs},\theta)}(z|Y_{\rm obs},\theta_0)}{f_{(\theta|Y_{\rm obs},Z)}(\theta_0|Y_{\rm obs},z)}, \qquad \begin{array}{l}\text{for some arbitrary } \theta_0 \in \mathcal{S}_{(\theta|Y_{\rm obs})} \\ \text{and all } z \in \mathcal{S}_{(Z|Y_{\rm obs})},\end{array} \tag{2.1}$$

where, in general, $\mathcal{S}_{(\theta|Y_{\rm obs})} = \mathcal{S}_\theta$, but $\mathcal{S}_{(Z|Y_{\rm obs})} \neq \mathcal{S}_Z$. Considering the conditional predictive distribution as an approximation to the marginal predictive distribution $f_{(Z|Y_{\rm obs})}(z|Y_{\rm obs})$, IBF sampling is realized via SIR: (i) Draw J independent samples of Z from $f_{(Z|Y_{\rm obs},\theta)}(z|Y_{\rm obs},\theta_0)$, denoted by $z^{(1)}, \ldots, z^{(J)}$; (ii) Calculate the reciprocals of the augmented posterior densities to obtain the weights

$$\omega_j = f^{-1}_{(\theta|Y_{\rm obs},Z)}(\theta_0|Y_{\rm obs},z^{(j)}) \Big/ \sum_{\ell=1}^{J} f^{-1}_{(\theta|Y_{\rm obs},Z)}(\theta_0|Y_{\rm obs},z^{(\ell)}), \qquad j = 1, \ldots, J; \tag{2.2}$$

(iii) Choose a subset from $\{z^{(1)}, \ldots, z^{(J)}\}$ via resampling *without replacement* from the discrete distribution on $\{z^{(j)}\}$ with probabilities $\{\omega_j\}$ to obtain an i.i.d. sample of size $m\,(< J)$ approximately from $f_{(Z|Y_{\rm obs})}(z|Y_{\rm obs})$, denoted by $\{z^{(k_1)}, \ldots, z^{(k_m)}\}$. It is worth noting that only one pre-specified θ_0 is needed for the whole IBF sampling process.

Clearly, the sampling-wise IBF (2.1) provides a natural class of ISFs: the conditional predictive distributions $\{f_{(Z|Y_{\rm obs},\theta)}(z|Y_{\rm obs},\theta) : \theta \in \mathcal{S}_{(\theta|Y_{\rm obs})}\}$, available from the model specification. However, the efficiency (but not the correctness) of IBF sampling depends on how well the ISF approximates the target function $f_{(Z|Y_{\rm obs})}(z|Y_{\rm obs})$. Since (2.1) holds for any given $\theta_0 \in \mathcal{S}_{(\theta|Y_{\rm obs})}$, it suffices to select a θ_0 such that $f_{(Z|Y_{\rm obs},\theta)}(z|Y_{\rm obs},\theta_0)$ best approximates $f_{(Z|Y_{\rm obs})}(z|Y_{\rm obs})$. Heuristically, if θ_0 is chosen to be the observed posterior mode $\hat\theta_{\rm obs}$, the overlap area under the two functions would be substantial since the approximation is accurate to the order of $O(1/n)$, as shown in the following theorem.

Theorem 1. *Let the observed posterior density $f_{(\theta|Y_{\rm obs})}(\theta|Y_{\rm obs})$ be a unimodal function with mode $\hat\theta_{\rm obs}$ and let n denote the sample size of the observed data $Y_{\rm obs}$. Then*

$$f_{(Z|Y_{\rm obs})}(z|Y_{\rm obs}) = f_{(Z|Y_{\rm obs},\theta)}(z|Y_{\rm obs},\hat\theta_{\rm obs})\{1 + O(1/n)\}. \tag{2.3}$$

Proof. Let $g(\theta)$ be an arbitrarily smooth, positive function for $\theta \in \mathcal{S}_{(\theta|Y_{\rm obs})} \subseteq I\!R^k$, $L(\theta|Y_{\rm obs})$ the likelihood function and $f_\theta(\theta)$ the prior. The posterior mean of $g(\theta)$ can be written as

$$\mathrm{E}\{g(\theta)|Y_{\rm obs}\} = \int_{\mathcal{S}_{(\theta|Y_{\rm obs})}} g(\theta) f_{(\theta|Y_{\rm obs})}(\theta|Y_{\rm obs})\, d\theta = \frac{\int g(\theta)\exp\{n\,\ell(\theta)\}\, d\theta}{\int \exp\{n\,\ell(\theta)\}\, d\theta}, \tag{2.4}$$

where $n\,\ell(\theta) = \log\{L(\theta|Y_{\rm obs})f_\theta(\theta)\} \propto \log\{f_{(\theta|Y_{\rm obs})}(\theta|Y_{\rm obs})\}$. Thus $\ell(\theta)$ has the same mode as $f_{(\theta|Y_{\rm obs})}(\theta|Y_{\rm obs})$, i.e., $\ell'(\hat\theta_{\rm obs}) = 0$. Applying Laplace's method to the numerator of (2.4), we have $\int g(\theta)\exp\{n\ell(\theta)\}d\theta \doteq g(\hat\theta_{\rm obs})\exp\{n\ell(\hat\theta_{\rm obs})\}$ $(2\pi/n)^{k/2}|\Sigma|^{1/2}$, where $\Sigma_{k\times k} = -(\partial^2\ell(\hat\theta_{\rm obs})/(\partial\theta\,\partial\theta^\top))^{-1}$. Similarly, for the denominator of (2.4), we have $\int \exp\{n\,\ell(\theta)\}\,d\theta \doteq \exp\{n\,\ell(\hat\theta_{\rm obs})\}(2\pi/n)^{k/2}|\Sigma|^{1/2}$. Tierney and Kadane (1986) showed that the resulting ratio is $g(\hat\theta_{\rm obs})$ up to error $O(1/n)$. Thus, $\mathrm{E}\{g(\theta)|Y_{\rm obs}\} = g(\hat\theta_{\rm obs})\{1 + O(1/n)\}$. Since $f_{(Z|Y_{\rm obs})}(z|Y_{\rm obs}) = \int f_{(Z|Y_{\rm obs},\theta)}(z|Y_{\rm obs},\theta)f_{(\theta|Y_{\rm obs})}(\theta|Y_{\rm obs})\,d\theta = \mathrm{E}\Big\{f_{(Z|Y_{\rm obs},\theta)}(z|Y_{\rm obs},\theta)\Big|Y_{\rm obs}\Big\}$, (2.3) follows immediately.

Remark 1. Since a subset of independent random variables is still independent, in IBF sampling $\{z^{(k_1)},\ldots,z^{(k_m)}\}$ is an independent sample. The second part of Step (iii), i.e., "resampling from the discrete distribution on $\{z^{(j)}\}$ with probabilities $\{\omega_j\}$", implies $\{z^{(k_1)},\ldots,z^{(k_m)}\}$ are approximately from $f_{(Z|Y_{\rm obs})}(z|Y_{\rm obs})$ with the approximation "improving" as J increases (Smith and Gelfand (1992)). However, resampling *with replacement* would result in dependent samples.

Remark 2. The weights in the IBF sampling differ fundamentally from those associated with the harmonic mean estimate of Newton and Raftery (1994) which, as pointed out by Gelfand and Dey (1994), is likely to suffer from numeric instability since the reciprocals of augmented posterior densities may approach infinity. However, in the proposed method, the weights $\{\omega_j\}$ in (2.2) are ratios and free from this kind of numeric instability. In fact, for some j_0 $(1 \le j_0 \le J)$, if $f^{-1}_{(\theta|Y_{\rm obs},Z)}(\theta_0|Y_{\rm obs},z^{(j_0)}) \to \infty$, we have

$$\omega_{j_0} = \left\{1 + \sum_{\ell=1,\,\ell\neq j_0}^{J} \frac{f^{-1}_{(\theta|Y_{\rm obs},Z)}(\theta_0|Y_{\rm obs},z^{(\ell)})}{f^{-1}_{(\theta|Y_{\rm obs},Z)}(\theta_0|Y_{\rm obs},z^{(j_0)})}\right\}^{-1} \longrightarrow 1.$$

When J is very large, say $J = 10^5$, some weights will be extremely small. According to our experience, the use of the exponent of the logarithm of the ratio in calculating weights $\{\omega_j\}$ helps enhance numeric accuracy.

2.2. IBF sampling: an alternative

An alternative sampling method can be derived by exchanging the role of θ and Z in sampling-wise IBF (2.1). The observed posterior can be expressed in three different ways:

$$f_{(\theta|Y_{\rm obs})}(\theta|Y_{\rm obs}) = \left\{\int_{\mathcal{S}_{(Z|Y_{\rm obs})}} \frac{f_{(Z|Y_{\rm obs},\theta)}(z|Y_{\rm obs},\theta)}{f_{(\theta|Y_{\rm obs},Z)}(\theta|Y_{\rm obs},z)}dz\right\}^{-1}, \text{for any given } \theta\in\mathcal{S}_{(\theta|Y_{\rm obs})}, \quad (2.5)$$

$$f_{(\theta|Y_{\rm obs})}(\theta|Y_{\rm obs}) \propto \frac{f_{(\theta|Y_{\rm obs},Z)}(\theta|Y_{\rm obs},z_0)}{f_{(Z|Y_{\rm obs},\theta)}(z_0|Y_{\rm obs},\theta)}, \quad \begin{array}{l}\text{for some arbitrary } z_0\in\mathcal{S}_{(Z|Y_{\rm obs})}\\ \text{and all } \theta\in\mathcal{S}_{(\theta|Y_{\rm obs})},\end{array} \quad (2.6)$$

$$f_{(\theta|Y_{\rm obs})}(\theta|Y_{\rm obs})=\frac{f_{(\theta|Y_{\rm obs},Z)}(\theta|Y_{\rm obs},z_0)}{f_{(Z|Y_{\rm obs},\theta)}(z_0|Y_{\rm obs},\theta)}\cdot\left\{\int_{\mathcal{S}_{(\theta|Y_{\rm obs})}}\frac{f_{(\theta|Y_{\rm obs},Z)}(\theta|Y_{\rm obs},z_0)}{f_{(Z|Y_{\rm obs},\theta)}(z_0|Y_{\rm obs},\theta)}\,d\theta\right\}^{-1}, \quad (2.7)$$
for some arbitrary $z_0 \in \mathcal{S}_{(Z|Y_{\rm obs})}$ and all $\theta \in \mathcal{S}_{(\theta|Y_{\rm obs})}$.

Equations (2.5) (called the pointwise IBF) and (2.7) (called the function-wise IBF) can sometimes be used to obtain the explicit expression of the observed posterior density (Tian, Ng and Geng (2003), Tian and Tan (2003)). Similarly, the sampling-wise IBF (2.6) can always be combined with SIR using $f_{(\theta|Y_{\rm obs},Z)}(\theta|Y_{\rm obs},z_0)$ as the ISF to generate i.i.d. samples approximately from the observed posterior.

Now the key is to be able to find a z_0 such that $f_{(\theta|Y_{\rm obs},Z)}(\theta|Y_{\rm obs},z_0)$ approximates $f_{(\theta|Y_{\rm obs})}(\theta|Y_{\rm obs})$ well. The idea is to simply take the z_0 at which the EM algorithm for finding the observed mode $\hat{\theta}_{\rm obs}$ converges. When Z is a continuous random variable (or vector), we choose

$$z_0 = E(Z|Y_{\rm obs},\hat{\theta}_{\rm obs}), \quad z_0 \in \mathcal{S}_{(Z|Y_{\rm obs})}. \quad (2.8)$$

When Z is discrete, z_0 obtained from (2.8) may not belong to $\mathcal{S}_{(Z|Y_{\rm obs})}$. Then, we choose the $z_0 \in \mathcal{S}_{(Z|Y_{\rm obs})}$ such that the distance between $\hat{\theta}_{\rm aug}(z)$ and $\hat{\theta}_{\rm obs}$ is minimized, i.e.,

$$z_0 = \arg\min_{z\in\mathcal{S}_{(Z|Y_{\rm obs})}} ||\hat{\theta}_{\rm aug}(z) - \hat{\theta}_{\rm obs}||, \quad (2.9)$$

where $\hat{\theta}_{\rm aug}(z)$ denotes the mode of the augmented posterior $f_{(\theta|Y_{\rm obs},Z)}(\theta|Y_{\rm obs},z)$. Note that both $f_{(\theta|Y_{\rm obs})}(\theta|Y_{\rm obs})$ and $f_{(\theta|Y_{\rm obs},Z)}(\theta|Y_{\rm obs},z_0)$ with z_0 given by (2.8) or (2.9) share the mode $\hat{\theta}_{\rm obs}$, thus there is substantial overlap in the areas under the target density curve and the ISF.

3. Applications

We start with the simple genetic linkage model where the posterior can be obtained exactly using pointwise IBF and thus the performance of IBF sampling can be evaluated unequivocally. This model was a useful proof-of-principle example for what are now well-known computational methods such as EM, DA and Gibbs sampler (Dempster, Laird and Rubin (1977), Tanner and Wong (1987) and Gelfand and Smith (1990)). Then we apply the proposed method to the hierarchical (or mixed-effects) model.

3.1. The genetic linkage model: an illustrative example

In this study, 197 animals are distributed according to a 4-cell multinomial distribution with cell probabilities: $(\theta+2)/4$, $(1-\theta)/4$, $(1-\theta)/4$, $\theta/4$, $0 \le \theta \le 1$. The observed data $Y_{\rm obs} = (y_1, y_2, y_3, y_4)^\top = (125, 18, 20, 34)^\top$ can be

augmented with latent data Z such that the complete-data is $\{Y_{\text{obs}}, Z\} = (Z, y_1 - Z, y_2, y_3, y_4)$. Using the usual Beta$(a, b)$ prior, we have $\theta|(Y_{\text{obs}}, Z = z) \sim$ Beta $(a + y_4 + z, b + y_2 + y_3)$, and $Z|(Y_{\text{obs}}, \theta) \sim$ Binomial $(y_1, \theta/(\theta + 2)$. Note that $\mathcal{S}_{(\theta, Z|Y_{\text{obs}})} = \mathcal{S}_{(\theta|Y_{\text{obs}})} \times \mathcal{S}_{(Z|Y_{\text{obs}})} = [0, 1] \times \{0, 1, \ldots, y_1\}$. Therefore, from (2.5), the posterior density of θ is

$$f_{(\theta|Y_{\text{obs}})}(\theta|Y_{\text{obs}}) = \left\{ \sum_{z=0}^{y_1} \frac{f_{(Z|Y_{\text{obs}},\theta)}(z|Y_{\text{obs}}, \theta)}{f_{(\theta|Y_{\text{obs}},Z)}(\theta|Y_{\text{obs}}, z)} \right\}^{-1} = \frac{(\theta + 2)^{y_1} \theta^{a+y_4-1} (1 - \theta)^{b+y_2+y_3-1}}{\sum_{z=0}^{y_1} \binom{y_1}{z} B(a + y_4 + z, b + y_2 + y_3) 2^{y_1 - z}}. \tag{3.1}$$

To implement IBF sampling, we first use the EM algorithm to find the best z_0. Both E-step and M-step have closed-form expressions:

$$z^{(t)} = \frac{y_1 \theta^{(t)}}{\theta^{(t)} + 2}, \qquad \theta^{(t+1)} = \frac{a + y_4 + z^{(t)} - 1}{(a + y_4 + z^{(t)} - 1) + (b + y_2 + y_3 - 1)}.$$

Setting $\theta^{(0)} = 0.5$ and $a = b = 1$ corresponding to the noninformative prior, the EM converged to $\hat{\theta}_{\text{obs}} = 0.6268$ after four iterations. Formula (2.8) yields $z_0 \approx 29.83$ and (2.9) results in $z_0 = 30$.

Figure 1(a) suggests that the augmented posterior (i.e., the chosen ISF) $f_{(\theta|Y_{\text{obs}},Z)}(\theta|Y_{\text{obs}}, z_0)$ with $z_0 = 30$ well-approximates $f_{(\theta|Y_{\text{obs}})}(\theta|Y_{\text{obs}})$ given by (3.1), both curves share the mode and have the maximum overlap. Figure 1(b) compares the exact observed posterior given by (3.1) and the approximate observed posterior given by the importance sampling method based on (2.7), that is,

$$f_{(\theta|Y_{\text{obs}})}(\theta|Y_{\text{obs}}) \doteq \frac{f_{(\theta|Y_{\text{obs}},Z)}(\theta|Y_{\text{obs}}, 30)}{f_{(Z|Y_{\text{obs}},\theta)}(30|Y_{\text{obs}}, \theta)} \cdot \left\{ \frac{1}{I} \sum_{i=1}^{I} \frac{1}{f_{(Z|Y_{\text{obs}},\theta)}(30|Y, \theta^{(i)})} \right\}^{-1}, \tag{3.2}$$

where $\{\theta^{(1)}, \ldots, \theta^{(I)}\}$, $I = 500$, were drawn from the best augmented posterior $f_{(\theta|Y_{\text{obs}},Z)}(\theta|Y_{\text{obs}}, 30)$. The two curves virtually coincide.

We implement IBF sampling based on (2.6) by drawing $J = 2500$ i.i.d. samples $\{\theta^{(j)} : j = 1, \ldots, J\}$ from $f_{(\theta|Y_{\text{obs}},Z)}(\theta|Y_{\text{obs}}, 30)$, and computing the weights $\{\omega_j\}$ according to (2.2). Then resample without replacement from the discrete distribution on $\{\theta^{(j)}\}$ with probabilities $\{\omega_j\}$ to obtain an i.i.d. sample of size $m = 2000$ approximately from $f_{(\theta|Y_{\text{obs}})}(\theta|Y_{\text{obs}})$. When the ISF $f_{(\theta|Y_{\text{obs}},Z)}(\theta|Y_{\text{obs}}, z_0)$ is very close to the objective function $f_{(\theta|Y_{\text{obs}})}(\theta|Y_{\text{obs}})$, Rubin (1988) suggested that $m \approx J$. The accuracy of the IBF sampling is remarkable as shown in Figure 1(c), where $f_{(\theta|Y_{\text{obs}})}(\theta|Y_{\text{obs}})$ is estimated by a kernel density smoother based on

the i.i.d. IBF samples. In addition, the histogram based on these samples is plotted in Figure 1(d), which shows that IBF sampling has recovered the density completely.

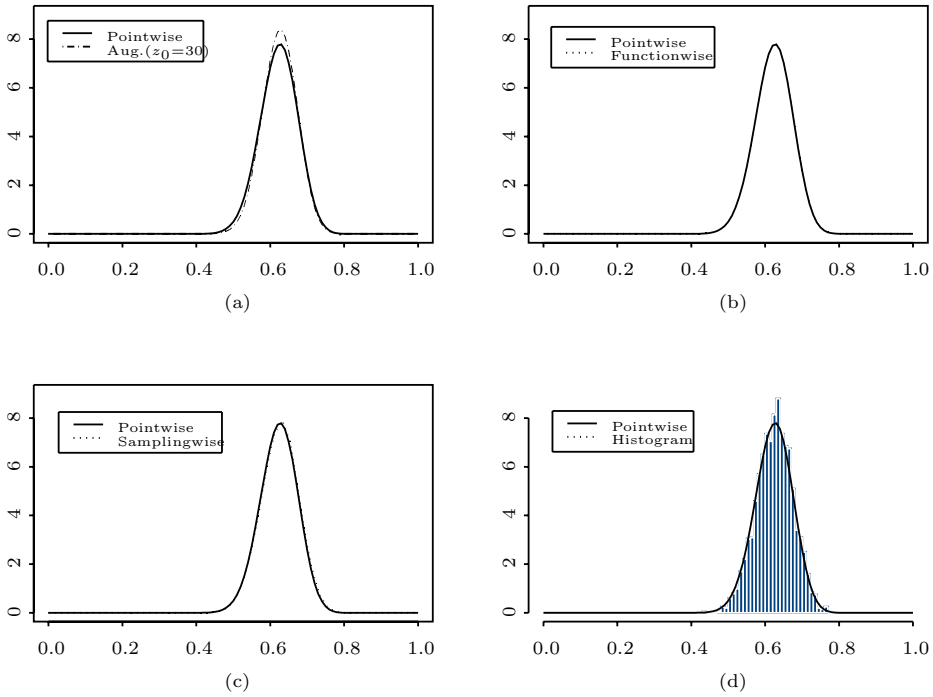

Figure 1. The pointwise IBF $f_{(\theta|Y_{\text{obs}})}(\theta|Y_{\text{obs}})$ given by (3.1): the solid line in all plots for the observed data $Y_{\text{obs}} = (125, 18, 20, 34)^\top$. (a) The augmented posterior density (·–·–) with $z_0 = 30$; (b) The function-wise IBF (· · · · · ·) given by (3.2); (c) The sampling-wise IBF (· · · · · ·) estimated by a kernel density smoother based on i.i.d. IBF samples ($J = 2500$, $m = 2000$); (d) The histogram based on i.i.d. IBF samples.

So far the observed and augmented posteriors are distributed nearly symmetrically. To see the performance of the method in highly skewed observed/augmented posterior distributions, let $Y_{\text{obs}} = (y_1, y_2, y_3, y_4)^\top = (14, 0, 1, 5)^\top$. From $\theta^{(0)} = 0$, the EM yields $\hat{\theta}_{\text{obs}} = 0.9034$ after five iterations. According to (2.9), the optimal choice is $z_0 = 4$, which fully restore the posterior density (figures not shown here).

3.2. Hierarchical (or mixed-effects) models

We consider the most common hierarchical model: the Bayesian version of the linear mixed effects model. Let Y_{ij} be the jth response for subject i, where $j = 1, \ldots, n_i$ and $i = 1, \ldots, N$. The normal linear mixed-effects model (Laird and Ware (1982), Liu and Rubin (1994)) is

$$Y_i = X_i^\top \beta + W_i^\top b_i + \varepsilon_i, \quad b_i \sim N_q(0,\ D), \quad \varepsilon_i \sim N_{n_i}(0,\ \sigma^2 R_i), \quad i = 1, \ldots, N, \tag{3.3}$$

where $Y_i\,(n_i \times 1)$ denotes the collection of responses for subject i, $X_i\,(p \times n_i)$ and $W_i\,(q \times n_i)$ are known design matrices relating to the covariates, β are the $p \times 1$ fixed effects, b_i are $q \times 1$ random effects, $D\,(q \times q)$ is an unknown positive definite matrix relating to the correlation structure of Y_i, σ^2 is an unknown variance parameter, the $R_i > 0$ are known $n_i \times n_i$ correlation matrices, and b_i is independent of ε_i. The model (3.3) can be rewritten in a hierarchical form:

$$Y_i | b_i \sim N_{n_i}(X_i^\top \beta + W_i^\top b_i,\ \sigma^2 R_i), \quad b_i \sim N_q(0,\ D), \quad i = 1, \ldots, N.$$

Let $Y_{\text{obs}} = \{(Y_i,\ X_i,\ W_i, R_i) :\ i = 1, \ldots, N\}$ denote the observed data. After treating $\boldsymbol{b} = \{b_1, \ldots, b_N\}$ as the missing data, the likelihood function for the complete-data $\{Y_{\text{obs}}, \boldsymbol{b}\}$ is

$$L(\beta, \sigma^2, D | Y_{\text{obs}}, \boldsymbol{b}) = \prod_{i=1}^{N} \Big\{ N_q(b_i | 0,\ D) * N_{n_i}(Y_i | X_i^\top \beta + W_i^\top b_i,\ \sigma^2 R_i) \Big\}.$$

Consider the independent prior distributions $\beta \sim N_p(\mu_0,\ \Sigma_0)$ with $\Sigma_0^{-1} \to 0$, $\sigma^2 \sim \text{IG}(q_0/2,\ \lambda_0/2)$ with inverse gamma density $\text{IG}(u | q_0/2, \lambda_0/2) = [(\lambda_0/2)^{q_0/2} / \Gamma(q_0/2)] \cdot u^{-1-q_0/2} \exp\{-\lambda_0/2u\}$, and $D \sim \text{IW}_q(\nu_0,\ \Lambda_0^{-1})$ with inverse Wishart density $\text{IW}_q(D | \nu_0, \Lambda_0^{-1}) \propto |D|^{-(\nu_0+q+1)/2} \exp\{-1/2 \text{tr}(\Lambda_0 D^{-1})\}$. For convenience, define $\theta \equiv (\beta, \sigma^2, D)$ and $\xi \equiv (\sigma^2, D)$. Then the complete-data posterior of θ is

$$\begin{aligned} & f_{(\theta | Y_{\text{obs}}, \boldsymbol{b})}(\theta | Y_{\text{obs}}, \boldsymbol{b}) \\ &= f_{(\beta | Y_{\text{obs}}, \boldsymbol{b}, \sigma^2)}(\beta | Y_{\text{obs}}, \boldsymbol{b}, \sigma^2) * f_{(\sigma^2 | Y_{\text{obs}}, \boldsymbol{b})}(\sigma^2 | Y_{\text{obs}}, \boldsymbol{b}) * f_{(D | Y_{\text{obs}}, \boldsymbol{b})}(D | Y_{\text{obs}}, \boldsymbol{b}) \\ &= N_p(\beta | \hat{\beta},\ \sigma^2 \hat{\Sigma}) * \text{IG}\Big(\sigma^2 \Big| \frac{q_0 + n - p}{2}, \frac{\lambda_0 + s}{2}\Big) * \text{IW}_q(D | \nu_0 + N, \Lambda^{-1}), \end{aligned} \tag{3.4}$$

where

$$\hat{\beta} = \hat{\Sigma} * \sum_{i=1}^{N} X_i R_i^{-1} (Y_i - W_i^\top b_i), \qquad \hat{\Sigma} = \Big(\sum_{i=1}^{N} X_i R_i^{-1} X_i^\top \Big)^{-1}, \qquad n = \sum_{i=1}^{N} n_i,$$

$$s = \sum_{i=1}^{N} (Y_i - X_i^\top \hat{\beta} - W_i^\top b_i)^\top R_i^{-1} (Y_i - X_i^\top \hat{\beta} - W_i^\top b_i), \qquad \Lambda = \Lambda_0 + \sum_{i=1}^{N} b_i b_i^\top.$$

The conditional predictive density is $f_{(\boldsymbol{b}|Y_{\rm obs},\theta)}(\boldsymbol{b}|Y_{\rm obs},\theta)=\prod_{i=1}^{N}N_q(b_i|\hat{b}_i(\theta),\Omega_i(\xi))$, where the mean vector $\hat{b}_i(\theta)$ and the covariance matrix $\Omega_i(\xi)$ have two alternative expressions:

$$\begin{aligned}\hat{b}_i(\theta) &= DW_i\Delta_i(\xi)(Y_i - X_i^\top\beta) = (\sigma^2 D^{-1} + W_i R_i^{-1} W_i^\top)^{-1} W_i R_i^{-1}(Y_i - X_i^\top\beta),\\ \Omega_i(\xi) &= D - DW_i\Delta_i(\xi)W_i^\top D \ = \sigma^2(\sigma^2 D^{-1} + W_i R_i^{-1} W_i^\top)^{-1},\end{aligned}$$

where $\Delta_i(\xi) \equiv (\sigma^2 R_i + W_i^\top D W_i)^{-1}$. Our objective is to obtain i.i.d. samples from $f_{(\theta|Y_{\rm obs})}(\theta|Y_{\rm obs})$. According to §2.1, we only need to obtain i.i.d. samples from $f_{(\boldsymbol{b}|Y_{\rm obs})}(\boldsymbol{b}|Y_{\rm obs})$. From (2.1), we have

$$f_{(\boldsymbol{b}|Y_{\rm obs})}(\boldsymbol{b}|Y_{\rm obs}) \propto \frac{f_{(\boldsymbol{b}|Y_{\rm obs},\theta)}(\boldsymbol{b}|Y_{\rm obs},\theta_0)}{f_{(\theta|Y_{\rm obs},\boldsymbol{b})}(\theta_0|Y_{\rm obs},\boldsymbol{b})}, \qquad \text{for some arbitrary } \theta_0. \tag{3.5}$$

We choose $\theta_0=\tilde{\theta}=(\tilde{\beta},\tilde{\sigma}^2,\tilde{D})$, the observed posterior mode of $f_{(\theta|Y_{\rm obs})}(\theta|Y_{\rm obs})$. Liu and Rubin (1994) consider the MLE of θ. We can similarly obtain $\tilde{\theta}$ as follows. Using the current estimates $\theta^{(t)} = (\beta^{(t)},\xi^{(t)}) = (\beta^{(t)},\sigma^{2(t)},D^{(t)})$, the E-step calculates $E(b_i|Y_{\rm obs},\theta^{(t)}) = \hat{b}_i(\theta^{(t)})$ and $E(b_i b_i^\top|Y_{\rm obs},\theta^{(t)}) = \Omega_i(\xi^{(t)}) + \hat{b}_i(\theta^{(t)})\hat{b}_i(\theta^{(t)})^\top$ for $i=1,\ldots,N$. The M-step is to find the posterior modes based on the complete-data. We have

$$\begin{aligned}\beta^{(t+1)} &= \hat{\Sigma} * \sum_{i=1}^{N} X_i R_i^{-1}\{Y_i - W_i^\top \hat{b}_i(\theta^{(t)})\},\\ \sigma^{2(t+1)} &= \frac{1}{q_0+2+n}\Big\{\lambda_0 + \sum_{i=1}^{N}\Big[r_i^{(t+1)\top} R_i^{-1} r_i^{(t+1)} + \sigma^{2(t)}\mathrm{tr}\Big(I_q - \sigma^{2(t)}\Delta_i(\xi^{(t)})R_i\Big)\Big]\Big\},\\ D^{(t+1)} &= \frac{1}{\nu_0+q+1+N}\Big\{\Lambda_0 + \sum_{i=1}^{N} E\Big(b_i b_i^\top|Y_{\rm obs},\theta^{(t)}\Big)\Big\},\end{aligned}$$

where $r_i^{(t+1)} \equiv Y_i - X_i^\top\beta^{(t+1)} - W_i^\top\hat{b}_i(\theta^{(t)})$, $i=1,\ldots,N$.

We now analyze the growth data introduced by Pothoff and Roy (1964). These data consist of growth measurements for 16 boys and 11 girls. For each subject, the distance from the center of the pituitary to the maxillary fissure was recored at ages 8, 10, 12 and 14. A regression model is fitted where the response is a linear function of age, with separate regressions for boys and girls. If $Y_{(s)ij}$ is the measurement for subject i in sex group s ($s=1$ for boys and $s=-1$ for girls) at age x_j, then $Y_{(s)ij} = \alpha_{si}^* + \gamma_{si}^* x_j + \varepsilon_{(s)ij}$, $i=1,\ldots,27$, $j=1,\ldots,4$, where α_{si}^* and γ_{si}^* are random intercept and slope for subject i in group s, $\boldsymbol{x} = (x_1,x_2,x_3,x_4)^\top = (8,10,12,14)^\top$ and $\varepsilon_{(s)ij}$ are errors. Furthermore, assume that $(\alpha_{si}^*,\gamma_{si}^*)^\top \sim N_2((\alpha_s,\gamma_s)^\top,\ D)$. Using matrix notation, we have two

models: $Y_{(s)i} = (\mathbf{1}_4, \boldsymbol{x})\begin{pmatrix}\alpha_s\\ \gamma_s\end{pmatrix} + (\mathbf{1}_4, \boldsymbol{x})b_i + \varepsilon_{(s)i}$, $s = 1, -1$. The original goal is to estimate α_1, γ_1, α_{-1}, γ_{-1}, σ^2 and D. A unified model can be written as

$$Y_{(s)i} = (\mathbf{1}_4, s\mathbf{1}_4, \boldsymbol{x}, s\boldsymbol{x})\beta + (\mathbf{1}_4, \boldsymbol{x})b_i + \varepsilon_{(s)i}, \quad b_i \sim N_2(0, D), \quad \varepsilon_{(s)i} \sim N_4(0, \sigma^2 I_4),$$

where $\beta = (\beta_1, \beta_2, \beta_3, \beta_4)^\top$. The relationship between (α_s, γ_s) and β is given by $\alpha_1 = \beta_1 + \beta_2$, $\gamma_1 = \beta_3 + \beta_4$, $\alpha_{-1} = \beta_1 - \beta_2$ and $\gamma_{-1} = \beta_3 - \beta_4$. Therefore, we only need to estimate β, σ^2 and D. We take noninformative priors, i.e., $q_0 = \lambda_0 = \nu_0 = 0$ and $\Lambda_0 = 0$. Using the MLEs $\hat{\beta} = (16.8566, -0.5160, 0.6319, 0.1524)^\top$, $\hat{\sigma}^2 = 1.7162$ and $\hat{D} = \begin{pmatrix} 4.5569 & -0.1983\\ -0.1983 & 0.0238\end{pmatrix}$ as the initial values (Verbeke and Molenberghs (2000, p.253)), the EM algorithm converged to the observed posterior mode $\tilde{\theta} = (\tilde{\beta}, \tilde{\sigma}^2, \tilde{D})$, where $\tilde{\beta} = (16.8578, -0.5271, 0.6331, 0.1541)^\top$, $\tilde{\sigma}^2 = 1.3049$ and $\tilde{D} = \begin{pmatrix} 2.2012 & -0.0091\\ -0.0091 & 0.0065\end{pmatrix}$. Based on (3.5), we implement IBF sampling using $J = 3000$ to obtain an i.i.d. sample of size $m = 2500$ approximately from $f_{(\boldsymbol{b}|Y_{\text{obs}})}(\boldsymbol{b}|Y_{\text{obs}})$, denoted by $\boldsymbol{b}^{(1)}, \ldots, \boldsymbol{b}^{(m)}$. Generating $\theta^{(i)} \sim f_{(\theta|Y_{\text{obs}}, \boldsymbol{b})}(\theta|Y_{\text{obs}}, \boldsymbol{b}^{(i)})$ for $i = 1, \ldots, m$, then $\theta^{(1)}, \ldots, \theta^{(m)} \overset{\text{i.i.d.}}{\sim} f_{(\theta|Y_{\text{obs}})}(\theta|Y_{\text{obs}})$. The corresponding posterior estimates of parameters are given in Table 1.

Table 1. Summary of posterior estimates of parameters.

Parameter	MLEs	Posterior mode	Posterior mean	Posterior sd	95% Posterior interval estimates
β_1	16.8566	16.8578	16.8656	0.5800	[15.7434, 18.0174]
β_2	−0.5160	−0.5271	−.5315	0.5839	[−1.6684, 0.6167]
β_3	0.6319	0.6331	0.6315	0.0515	[0.5315, 0.7314]
β_4	0.1524	0.1541	0.1537	0.0519	[0.0508, 0.2542]
σ^2	1.7162	1.3049	1.3631	0.1744	[1.0748, 1.7742]
d_{11}	4.5569	2.2012	2.4487	0.7398	[1.3438, 4.2482]
d_{22}	0.0238	0.0065	0.0067	0.0020	[0.0037, 0.0113]
d_{12}	−0.1983	−0.0091	−0.0087	0.0266	[−0.0645, 0.0413]

4. Discussion

We have proposed a noniterative sampling approach for obtaining i.i.d. samples approximately from an observed posterior by combining IBF with SIR and EM as an alternative to perfect sampling. Although both perfect sampling and the proposed alternative generate i.i.d. samples and eliminate problems in monitoring convergence and mixing of the Markov chain, we have shown that IBF sampling is a simple yet highly efficient algorithm that performs well in hierarchical models. Practically, IBF sampling is a method to quickly generate i.i.d. sam-

ples approximately from the posterior once the posterior mode is identified. For example, the hierarchical model took only about one minute on a Pentium 4 PC.

IBF sampling is applicable to Monte Carlo EM structures where the M-step is simple but the E-step is complicated. The posterior mode can be determined by using traditional methods such as the Newton-Raphson algorithm, the scoring algorithm, the Laplace method and so on. On the other hand, we can also use the MCEM algorithm with a Gibbs sampling/MCMC E-step to obtain the posterior mode. Such a technique retains the advantage of drawing samples from full conditional distributions but avoids the difficulty of monitoring convergence and mixing of the Markov chain in Gibbs sampling. In the Gibbs sampling E-step of MCEM, the convergence diagnosis can be ignored because the final convergence is controlled by the EM algorithm (McCulloch (1997)). In this case, the sampling-wise IBF (2.6) provides an alternative to (2.1) so long as $f_{(Z|Y_{\text{obs}},\theta)}(z|Y_{\text{obs}},\theta)$ is relatively easy to evaluate. For example, in the Bayesian analysis of multivariate probit models (Chib and Greenberg (1998)), $f_{(Z|Y_{\text{obs}},\theta)}(z|Y_{\text{obs}},\theta)$ is a multivariate normal density truncated to a specific region. Sometimes the missing data Z can further be augmented by another latent vector b such that all $f_{(\theta|Y_{\text{obs}},Z,b)}(\theta|Y_{\text{obs}},z,b)$, $f_{(Z|Y_{\text{obs}},b,\theta)}(z|Y_{\text{obs}},b,\theta)$ and $f_{(b|Y_{\text{obs}},Z,\theta)}(b|Y_{\text{obs}},z,\theta)$ are available. Then, by applying (2.1) twice, we can still obtain i.i.d. samples approximately from $f_{(\theta|Y_{\text{obs}})}(\theta|Y_{\text{obs}})$.

A difficult problem in MCMC is the high autocorrelation between $\theta|Y_{\text{obs}}$ and $Z|Y_{\text{obs}}$, which results in a slow moving chain. In missing data problems, a Gibbs sampler with such slow convergence corresponds to an EM with slow convergence. Hence, some fast EM-type algorithms such as Alternating ECM (Meng and van Dyk (1997)), PX-EM (Liu, Rubin and Wu (1998)) can be used to accelerate the convergence. That is, the slow convergence problem in the Gibbs sampler can be bypassed in IBF sampling by running a fast EM-type algorithm if $\text{Var}(Z|Y_{\text{obs}},\hat{\theta}_{\text{obs}})$ is not much less than $\text{Var}(Z|Y_{\text{obs}})$. This may explain why the proposed method usually does not have difficulty with autocorrelation and works well in hierarchical models without further model raparameterization or the centering which is usually needed with the Gibbs sampler (see, e.g., Qiu, Song and Tan (2002)).

The proposed IBF sampling combines the strengths of SIR and EM-type algorithms. For example, SIR generates independent samples but it does not provide an efficient ISF directly from the model specification of a practical problem. It is easy to check if the EM algorithm has converged to the posterior mode or not, but it is difficult to calculate its standard error. The EM/DA algorithm and the sampling-wise IBF (2.1) or (2.6) share the structure of augmented posterior/conditional predictive distributions, thus no extra derivations are needed for IBF sampling. This implies that IBF sampling is applicable to problems where any EM-type algorithms can be applied, a wide range of practical problems.

Finally, further developments of the method are of interest, for example, in constrained parameter problems where it is difficult to sample from hyperparameters with the MCMC method (see, e.g., Chapter 6, Chen, Shao and Ibrahim (2000) and Tan, Fang, Tian and Houghton (2002)). The method can also be used to check compatibility and the convergence in Gibbs sampling. Examples may also include those where $\mathrm{Var}(Z|Y_{\mathrm{obs}}, \theta_0)$ is likely to be much less than $\mathrm{Var}(Z|Y_{\mathrm{obs}})$, although we have found none in practice. The method is potentially useful in improving methods for model selection with Bayes factor or marginal likelihood (Chib (1995)) that has been based on Gibbs output with the posterior mode calculated. Further research is also needed for cases where the number of blocks of random variables is extremely large, e.g., spatial models (either on lattice or point process) and belief networks.

Acknowledgements

The authors wish to thank Dr. R. J. Carroll for helpful discussions. We are grateful to an associate editor and a referee for valuable comments and suggestions. Ming Tan and Guo-Liang Tian's research was supported in part by U.S. National Cancer Center support grant CA21765. The research of Kai Wang Ng was partially supported by a research grant of the University of Hong Kong.

References

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussions). *J. Roy. Statist. Soc. Ser. B* **36**, 192-236.

Casella, G., Lavine, M. and Robert, C. P. (2001). Explaining the Perfect Sampler. *Amer. Statist.* **55**, 299-305.

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

Chib, S. (1995). Marginal likelihood from the Gibbs output. *J. Amer. Statist. Assoc.* **90**, 1313-1321.

Chib, S. and Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika* **85**, 347-361.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B* **39**, 1-38.

Gelfand, A. E. (2002). Gibbs sampling. In *Statistics in the 21st Century* (Edited by A. E. Raftery, M. A. Tanner and M. T. Wells), 341-349. Chapman and Hall/CRC, Boca Raton.

Gelfand, A. E. and Dey, D. K. (1994). Bayesian model choice: asymptotic and exact calculations. *J. Roy. Statist. Soc. Ser. B* **56**, 501-514.

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

Green, P. J. and Murdoch, D. J. (1999). Exact sampling for Bayesian inference: towards general purpose algorithms (with discussion). In *Bayesian Statistics* **6** (Edited by J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), 301-321. Oxford University Press, Oxford.

Hammersley, J. M. and Clifford, M. S. (1970). Markov fields on finite graphs and lattices. Unpublished.

Jones, G. L. and Hobert, J. P. (2001). Honest exploration of intractable probability distribution via Markov chain Monte Carlo. *Statist. Sci.* **16**, 312-334.

Laird, N. M. and Ware, J. H. (1982). Random effects models for longitudinal data. *Biometrics* **38**, 963-974.

Liu, C. H. and Rubin, D. B. (1994). The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence. *Biometrika* **81**, 633-648.

Liu, C. H., Rubin, D. B. and Wu, Y. N. (1998). Parameter expansion to accelerate EM — the PX-EM algorithm. *Biometrika* **85**, 755-770.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **44**, 226-233.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *J. Amer. Statist. Assoc.* **92**, 162-170.

Meng, X. L. (1996). Comment on "statistical inference and Monte Carlo algorithms" by G. Casella. *Test* **5**, 310-318.

Meng, X. L. and van Dyk, D. A. (1997). The EM algorithm — an old folk song sung to a fast new tune (with discussion). *J. Roy. Statist. Soc. Ser. B* **59**, 511-567.

Newton, M. A. and Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap (with discussion). *J. Roy. Statist. Soc. Ser. B* **56**, 3-48.

Ng, K. W. (1997). Inversion of Bayes formula: explicit formulas for unconditional pdf. In *Advances in the Theory and Practice in Statistics* (Edited by N. L. Johnson and N. Balakrishnan), 571-584. Wiley, New York.

Pothoff, R. F. and Roy, S. N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika* **51**, 313-326.

Qiu, Z., Song, P. and Tan, M. (2002). Bayesian hierarchical models for multi-level ordinal data using WinBUGS. *J. Biopharm. Statist.* **12**, 121-135.

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

Rubin, D. B. (1988). Using the SIR algorithm to simulate posterior distributions (with discussion). In *Bayesian Statistics* **3** (Edited by J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith), 395-402. Oxford University Press, Oxford.

Smith, A. F. M. and Gelfand, A. E. (1992). Bayesian statistics without tears: a sampling-resampling perspective. *Amer. Statist.* **46**, 84-88.

Tan, M., Fang, H. B., Tian, G. L. and Houghton, P. J. (2002). Small-sample inference for incomplete longitudinal data with truncation and censoring in tumor xenograft models. *Biometrics* **58**, 612-620.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *J. Amer. Statist. Assoc.* **82**, 528-540.

Tian, G. L., Ng, K. W. and Geng, Z. (2003). Baysian computation for contingency tables with incomplete cell-counts. *Statist. Sinica* **13**, 189-206.

Tian, G. L. and Tan, M. (2003). Exact statistical solutions using the inverse Bayes formulae. *Statist. Probab. Lett.* **62**, 305-315.

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

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. Springer, New York.

Division of Biostatistics, University of Maryland Greenebaum Cancer Center, 22 South Greene Street, Baltimore, MD 21201, U.S.A.

E-mail: mtan@umm.edu

Division of Biostatistics, University of Maryland Greenebaum Cancer Center, 22 South Greene Street, Baltimore, MD 21201, U.S.A.

E-mail: gtian2@umm.edu

Department of Statistics and Actuarial Science, the University of Hong Kong, Pokfulam Road, Hong Kong.

E-mail: kaing@hku.hk

(Received April 2002; accepted February 2003)