Application of a predictive distribution formula to Bayesian computation for incomplete data models

By TREVOR J. SWEETING and SAMER KHARROUBI

University College London and University of Sheffield

SUMMARY

We consider exact and approximate Bayesian computation in the presence of latent variables or missing data. Specifically we explore the application of a posterior predictive distribution formula derived in Sweeting & Kharroubi (2003), which is a particular form of Laplace approximation, both as an importance and a proposal distribution. We show that this formula provides a stable importance function for use within poor man's data augmentation schemes and that it can also be used as a proposal distribution within a Metropolis-Hastings algorithm for models that are not analytically tractable. We illustrate both uses in the case of a censored regression model and a normal hierarchical model, with both normal and Student t distributed random effects. Although the predictive distribution formula is motivated by regular asymptotic theory, it is not necessary that the likelihood has a closed form or that it possesses a local maximum.

Keywords: asymptotic approximation, Bayesian computation, data augmentation, importance sampling, Laplace approximation, latent variables, Metropolis-Hastings, missing data, predictive distribution, simulation

1 Introduction

Widely used strategies for Bayesian computation include numerical integration, stochastic simulation and analytic approximation. With regard to the latter, modern methods of asymptotic analysis can provide extremely accurate approximations in practice in problems involving parameter spaces of small to moderate dimension. This is partly because these approximations have a non-local character and therefore often reflect the shape of the important region of the posterior density much better than asymptotic expansions based on high-order derivatives at the likelihood maximum. However, such approximations are degraded when the dimension of the parameter

space is high or when the data are insufficiently informative. In such cases the most widely used computational methods are based on simulation. A number of researchers have investigated the use of hybrid methods that combine the best features of simulation, or numerical integration, and asymptotics. See for example Evans & Swartz (1995), Sweeting (1996), DiCiccio et al. (1997). Possible roles for asymptotics within simulation include the construction of importance distributions for importance sampling and proposal distributions for Metropolis-Hastings schemes, both of which are explored here.

In this paper we will be concerned with exact and approximate Bayesian computation in the presence of latent variables or missing data. Such data structures are ubiquitous and there exist many likelihood-based methods of analysis. Data augmentation techniques are particularly useful in applications where the posterior analysis of the complete data is analytically tractable. In certain applications maximum likelihood estimation may be accomplished via the EM algorithm, which has been extended for analytically intractable likelihoods to various versions of Monte Carlo EM (MCEM). For full Bayesian computation in general, one can incorporate the latent data within a suitable Markov chain Monte Carlo (MCMC) scheme.

Here we will explore the application of a posterior predictive distribution formula obtained in Sweeting & Kharroubi (2003), both as an importance and a proposal distribution for missing data problems. This formula is a particular form of Laplace approximation for a posterior expectation and can be regarded as an alternative to the Tierney and Kadane formula (Tierney & Kadane, 1986). This formula expresses a predictive distribution as a finite mixture, and we shall refer to it generically as the 'p-bar' formula. As pointed out in Sweeting & Kharroubi (2003), this mixture structure makes the p-bar formula an attractive proposition for use within a Bayesian simulation scheme. We review this formula in §2.

In §3, after reviewing the basic data augmentation algorithm we consider poor man's data augmentation algorithms of the type described by Wei & Tanner (1990a). These algorithms provide non-iterative exact (*i.e.* simulation-consistent) or approximate computation of the posterior distribution. One attractive feature of these data augmentation schemes is that they can provide functional forms for marginal densities and other quantities of interest in a computationally efficient manner. As pointed out

by Tanner (1996), the simplest approximate scheme based on a first-order maximum likelihood approximation may provide a poor estimate of the posterior distribution and importance sampling implementations are preferable. However, we will show that such algorithms may exhibit poor convergence behavior due to a poor choice of the importance function.

In §4 we investigate the p-bar formula as an importance function for use within these poor man's data augmentation schemes. Since it has an asymptotically higher order of accuracy than a simple first-order approximation, it would be expected to possess superior operating characteristics. Furthermore, unlike the Tierney and Kadane Laplace approximation, the mixture form of p-bar makes it trivial to simulate from. Its superior convergence behavior is demonstrated in §4.1 for the censored regression model discussed by Wei & Tanner (1990a) and Tanner (1996).

More generally, and especially for models that are not analytically tractable, the p-bar formula may be used as a proposal distribution within a Metropolis-Hastings scheme. We illustrate such a use of p-bar in §5.1 for a censored regression model. Although in the present paper we do not attempt to make detailed comparisons with alternative computational schemes, we note that the resulting algorithm gives better mixing than a block Gibbs sampler in this case. We further illustrate the use of the p-bar formula in both data augmentation and MCMC in §6 with a normal hierarchical model. In the simplest case of normally distributed random effects exact data augmentation is possible. It is not possible, however, in the case of Student t distributed random effects since the likelihood does not have a closed form and instead we use MCMC methods. One feature of this example is that, although we make use of a predictive distribution formula motivated by regular asymptotic theory, the likelihood function here does not even possess a local maximum. This example illustrates the important device of *tilting*, which allows the application of p-bar in such cases.

A general discussion and summary is given in §7. Finally some definitions and technical details on asymptotic error rates are given in an appendix.

2 The p-bar formula

In this section we will review the higher-order asymptotic formula for a posterior predictive distribution derived in Sweeting & Kharroubi (2003). This formula is a particular form of Laplace approximation for a posterior expectation and can be regarded as an alternative to the Tierney and Kadane formula.

Sweeting (1996) derives an expression (formula (20) in that paper) for the posterior expectation of a real-valued function $v(\theta)$ that has the same asymptotic order of accuracy as that derived by Tierney & Kadane (1986). Since a posterior predictive distribution can be expressed as a posterior expectation, this formula leads to a new formula for a predictive distribution. However, this formula is not in a suitable form for direct sampling. Sweeting & Kharroubi (2003) obtain a modification of the formula in Sweeting (1996) that allows for direct sampling. In this section we present the necessary notation and give the result; the reader is referred to Sweeting & Kharroubi (2003) for the detailed derivation.

Suppose that the observed data Y consist of n independent and identically distributed observations. Let $p(Y|\theta)$ be the density of Y given a parameter $\theta \in \Theta \subset \mathcal{R}^d$, where $d \geq 1$. Suppose that a prior density $p(\theta)$ for θ is available and denote by $p(\theta|Y)$ the associated posterior density of θ given Y. Further, let $p(Z|Y)$ denote the posterior predictive density of latent unobserved data $Z \in \mathcal{Z}$ given Y and let $p(Z|Y,\theta)$ denote the conditional predictive density of Z given Y and θ.

Write $\theta = (\theta^1, \ldots, \theta^d)$ and let $l(\theta) = \log p(Y|\theta)$, $l_i(\theta) = \partial l(\theta)/\partial\theta^i$ and $j(\theta) = -d^2 l/d\theta^2$, the matrix of second-order partial derivatives of $-l(\theta)$. Assume that the maximum likelihood estimate (MLE) $\hat{\theta}$ exists and write $J = j(\hat{\theta})$. Let $\theta_i = (\theta^1, \ldots, \theta^i)$ and $\theta^{(i)} = (\theta^i, \ldots, \theta^d)$, the vectors of the first i and last $(d - i + 1)$ components of θ. Define $\hat{\theta}^{(i)}(\theta_{i-1})$ to be the MLE of $\theta^{(i)}$ conditional on θ_{i-1} and, for $j \geq i$, let $\hat{\theta}^j(\theta_{i-1})$ be the jth component of $(\theta_{i-1}, \hat{\theta}^{(i)}(\theta_{i-1}))$. For any function $g(\theta)$, when $i < d$ we use the short-hand $g(\theta_i)$ to denote $g(\theta_i, \hat{\theta}^{(i+1)}(\theta_i))$. Further, let $j^{(i)}(\theta)$ be the submatrix of $j(\theta)$ corresponding to $\theta^{(i)}$, setting $|j^{(d+1)}(\theta)| = 1$. Finally, define the *signed-root loglikelihood ratio* transformation

$$r^i(\theta_i) = \text{sign}\{\theta^i - \hat{\theta}^i(\theta_{i-1})\}[2\{l(\theta_{i-1}) - l(\theta_i)\}]^{1/2}\,,\ i = 1, \ldots, d\,. \tag{1}$$

Now, for $i = 1, \ldots, d$, define the scalars θ^{i-} and θ^{i+} as the solutions to the equa-

tions

$$r^i(\hat{\theta}_{i-1}, \theta^{i-}) = -\sqrt{d}\,, \quad r^i(\hat{\theta}_{i-1}, \theta^{i+}) = \sqrt{d} \tag{2}$$

and write $\theta_i^- = (\hat{\theta}_{i-1}, \theta^{i-})$ and $\theta_i^+ = (\hat{\theta}_{i-1}, \theta^{i+})$. Assuming that the conditional predictive density $p(Z|Y,\theta)$ is $O(1)$ in n, Sweeting & Kharroubi (2003) show that the formula

$$\bar{p}(Z|Y) = \sum_{i=1}^{d} \pi_i \left\{ \alpha_i^- p(Z|Y,\theta_i^-) + \alpha_i^+ p(Z|Y,\theta_i^+) \right\} \tag{3}$$

is an $O(n^{-2})$ approximation to the posterior predictive density $p(Z|Y)$, where the weights $\pi_i, \alpha_i^{\pm}$, $i = 1, \ldots, d$, are defined in the Appendix. Notice that the function $\bar{p}(Z|Y)$ is a proper density function since the weights do not depend on Z and satisfy $\sum_{i=1}^d \pi_i = 1$ and $\alpha_i^- + \alpha_i^+ = 1$. Since (3) expresses the predictive distribution as a finite mixture, we shall refer to it generically as the *p-bar* formula.

Since it is based on the signed-root loglikelihood (1), formula (3) is invariant under reparameterisation. However, the computation of θ^{i-} and θ^{i+} involves the solution of a set of nonlinear equations, which in turn involve conditional maximisation. Sweeting & Kharroubi (2003) also obtain an alternative asymptotically equivalent formula, which replaces the definition (2) of θ^{i-} and θ^{i+} by

$$\theta^{i-} = \hat{\theta}^i - \sqrt{d}(k^i)^{-1/2}\,, \quad \theta^{i+} = \hat{\theta}^i + \sqrt{d}(k^i)^{-1/2}\,, \tag{4}$$

where k^i is the reciprocal of the first entry in $(J^{(i)})^{-1}$. The disadvantage of this formulation is that (3) is no longer invariant to reparameterisation, so some care may be needed in the choice of parameterisation. However, the computational burden associated with the quantities θ^{i-} and θ^{i+} is greatly reduced. In the case of a single parameter this approximation is reminiscent of a suggestion by Tom DiCiccio reported in Kass's discussion of Sweeting (1996) in which the $\theta^{\pm}$ values are defined by $\theta^{\pm} = \hat{\theta} \pm \epsilon J^{-1/2}$ for some $\epsilon > 0$. However, using the present approach and positioning $\theta^{\pm}$ at the specific values as in (4) leads to the mixture formula (3) useful for direct simulation of the latent data Z (see below).

Although the p-bar formula is derived from a higher-order asymptotic analysis, formula (3) accounts for the shape of the likelihood function and prior in the neighborhood of $\hat{\theta}$ and would be expected to give reasonable approximations even in relatively small samples and in problems with dependent observations. This expectation is

borne out in our application of these formulae. Indeed, in §6 we successfully apply these formulae to a case where the likelihood function does not possess a local maximum.

We briefly compare the p-bar formula (3) for a predictive density with the formula arising from equation (2.4) in Tierney & Kadane (1986). Wei & Tanner (1990a) use this equation to obtain the alternative $O(n^{-2})$ approximation

$$p(Z|Y) \propto \frac{p(\theta^*)L^*(\theta^*;Z)p(Z|Y,\hat{\theta})}{|J^*|^{1/2}L^*(\hat{\theta};Z)}\,, \tag{5}$$

where θ^* maximizes $p(\theta|Y,Z)$, J^* is the negative Hessian of $\log p(\theta|Y,Z)$ evaluated at θ^* and $L^*(\theta;Z)$ is the augmented likelihood function associated with (Y,Z). Since it is a finite mixture of the conditional predictive densities $p(Z|Y,\theta)$, a key advantage of formula (3) over (5) is that it is in a convenient form from which to sample, making it an attractive proposition for use within a Bayesian simulation scheme, as discussed in §§4 - 6. Furthermore, since $\bar{p}(Z|Y)$ is a proper density function there is no need to renormalize. Finally, the abscissae $\theta_i^{\pm}$ and weights $\pi_i, \alpha_i^{\pm}$ only need to be calculated once, since they do not depend on Z. In contrast, in the approximation (5) θ^* must be recomputed for each Z (although the error continues to be $O(n^{-2})$ if one-step Newton approximations to θ^* are used). However, when the log-conditional predictive density $\log p(Z|Y,\theta)$ is not $O(1)$ then (5) may give a superior approximation to (3). In particular, in the extreme case in which $\log p(Z|Y,\theta)$ is $O(n)$ the approximation (3) is no better than using $p(Z|Y,\hat{\theta})$. On the other hand it is shown in the Appendix that in this case the relative error in (5) turns out to be $O(n^{-1})$.

3 Data augmentation algorithms

In this section we review the full data augmentation algorithm (Tanner & Wong, 1987), which allows for exact (*i.e.* simulation-consistent) computation of the entire posterior distribution. We then review the poor man's data augmentation (PMDA) algorithms (Wei & Tanner, 1990a, Tanner, 1996). These non-iterative algorithms, which may be exact or approximate, reduce the computational burden of the full data augmentation algorithm.

Recalling the definitions given at the beginning of §2, further let $p(\theta|Y,Z)$ be the augmented posterior density of θ given Y and Z. The data augmentation algorithm is

motivated by the following two equations. As in Tanner & Wong (1987), the *posterior equation* is

$$p(\theta|Y) = \int_{\mathcal{Z}} p(\theta|Y, Z)p(Z|Y)dZ\,, \tag{6}$$

and the *predictive equation* is

$$p(Z|Y) = \int_{\Theta} p(Z|Y, \theta)p(\theta|Y)d\theta\,.$$

Given a current approximation $g_i(\theta)$ to $p(\theta|Y)$, the algorithm proceeds by (a) generating θ from $g_i(\theta)$ followed by z from $p(Z|Y, \theta)$ and (b) repeating step (a) m times to yield $z_1, \ldots, z_m$ and the update

$$g_{i+1}(\theta) = m^{-1} \sum_{j=1}^{m} p(\theta|Y, z_j)\,.$$

The data augmentation algorithm consists of iterating between the two steps (a) and (b). Tanner & Wong (1987) present regularity conditions for the algorithm to converge and discuss the practical issues of monitoring convergence of the algorithm and the selection of m.

To motivate the simplest poor man's data augmentation algorithm, PMDA 1, suppose that the conditional predictive density $p(Z|Y, \theta)$ of Z is $O(1)$ in n. Then, since $p(Z|Y) = E\{p(Z|Y, \theta)|Y\}$, we have

$$p(Z|Y) = p(Z|Y, \hat{\theta}) \left\{1 + O(n^{-1})\right\}\,.$$

Given the MLE $\hat{\theta}$, the algorithm proceeds as follows.

a. Generate $z_1, \ldots, z_m$ from $p(Z|Y, \hat{\theta})$.

b. Approximate the posterior density by

$$m^{-1} \sum_{j=1}^{m} p(\theta|Y, z_j)\,.$$

Wei & Tanner (1990a) illustrate the PMDA 1 algorithm for a single parameter genetic linkage model in which 20 animals are distributed multinomially into four categories as $Y = (14, 0, 1, 5)$. They demonstrate that, although a first-order normal approximation is poor, PMDA 1 successfully recovers the skewed shape of the posterior density. However, although PMDA 1 may give a reasonable approximation to the posterior density when n is not too small and the dimension of the parameter space

not too large, it can produce a poor approximation if the log-conditional predictive density $\log p(Z|Y,\theta)$ is not $O(1)$ in n, as will be seen in §3.1.

As mentioned in §2, Wei & Tanner (1990a) use (5) based on equation (2.4) of Tierney & Kadane (1986) as an approximation to $p(Z|Y)$. Sampling from this approximation is possible using importance sampling (Ripley, 1987), resulting in the PMDA 2 algorithm, which is of the following form.

a1. Generate $z_1, \ldots, z_m$ from $p(Z|Y,\hat{\theta})$.

a2. Calculate the weights

$$w_j = \frac{p(\theta_j^*)L^*(\theta_j^*; z_j)}{|J_j^*|^{1/2}L^*(\hat{\theta}; z_j)} \ , \ j = 1, \ldots, m \, ,$$

where θ_j^* and J_j^* are the posterior mode and negative Hessian matrix corresponding to the jth augmented data set.

b. Approximate the posterior density by

$$\sum_{j=1}^{m} w_j p(\theta|Y, z_j) \Big/ \sum_{j=1}^{m} w_j \, . \tag{7}$$

For the genetic linkage model mentioned earlier, Wei & Tanner show that PMDA 2 provides an excellent approximation to the posterior distribution of θ. As with the PMDA 1 algorithm, if the log-conditional predictive density $\log p(Z|Y,\theta)$ is not $O(1)$ then approximation (5) will be less accurate. However, as discussed in the §2, this approximation may still be useful.

Since the sampling of the latent data in both PMDA 1 and PMDA 2 is based on an approximation to $p(Z|Y)$, these algorithms produce approximations to the posterior density. However, as pointed out by Wei & Tanner, if $p(Z|Y)$ is straightforward to calculate as a function of Z then one can use importance sampling to sample from the exact predictive distribution $p(Z|Y)$. Wei & Tanner refer to the resulting algorithm as PMDA-Exact. We further note here that one can always compute $p(Z|Y)$ up to proportionality whenever the constant of proportionality for the augmented posterior density is available analytically. This follows since

$$p(Z|Y) = \frac{p(\theta|Y)p(Z|Y,\theta)}{p(\theta|Y,Z)} \propto \frac{p(Z|Y,\theta)}{p(\theta|Y,Z)} \, , \tag{8}$$

which may be evaluated at $\theta = \hat{\theta}$, for example. In this form, the PMDA-Exact algorithm is of the same form as the PMDA 2 algorithm but with weights now given

by

$$w_j = \left\{ p(\hat{\theta}|Y, z_j) \right\}^{-1} \ , \ j = 1, \ldots, m \, .$$

This algorithm will provide a consistent estimator of $p(\theta|Y)$ as $m \to \infty$ since, from (8), we have

$$w_j \propto \frac{p(z_j|Y)}{p(z_j|Y, \hat{\theta})} \, .$$

3.1 Example: censored regression

We consider a censored failure dataset (Crawford, 1970) that has been analysed by several authors. In particular, Wei & Tanner (1990a, 1990b) carry out a full data augmentation analysis of these data, as well as the poor man's data augmentation algorithms PMDA 1 and PMDA 2.

The data arise from temperature accelerated life tests on electrical insulation in 40 motorettes. Ten motorettes were tested at each of four temperature levels (150°, 170°, 190° and 220°), resulting in a total of 17 failed units and 23 unfailed (censored) units. A model of the form (Schmee & Hahn, 1979)

$$y_i = \beta_0 + \beta_1 v_i + \sigma \epsilon_i$$

is fit, where y_i is the $\log_{10}(i$th failure time), $v_i = 1000/(\text{temperature} + 273.2)$ and the errors ϵ_i are assumed to be independent and follow a standard normal distribution. Failure times are in hours.

Re-ordering the data so that the first m observations are uncensored, with observed log-failure times y_i, and the remaining $n - m$ are censored at times c_i, the log-likelihood function is

$$-m \log\sigma - \frac{1}{2} \sum_{i=1}^{m} \left(\frac{y_i - \beta_0 - \beta_1 v_i}{\sigma} \right)^2 + \sum_{i=m+1}^{n} \log \left\{ 1 - \Phi \left(\frac{c_i - \beta_0 - \beta_1 v_i}{\sigma} \right) \right\} \, ,$$

where $\Phi(x)$ is the standard normal distribution function. For computational convenience here and later in §4.1 we work with the parameterisation $\theta = (\beta_0, \beta_1, \lambda)$, where $\lambda = \log \sigma$. The MLE is found to be $\hat{\theta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\lambda}) = (-6.0193, 4.3112, -1.3502)$ and the observed information matrix is

$$\begin{pmatrix} 427.66 & 931.31 & -65.39 \\ 931.31 & 2033.55 & -145.49 \\ -65.39 & -145.49 & 41.29 \end{pmatrix} \, .$$

We further assume the noninformative prior density $p(\theta) \propto 1$.

In order to implement the PMDA algorithms we require the functional forms of the augmented posterior density $p(\theta|Y,Z)$ and the conditional predictive density $p(Z|Y,\theta)$, where Z is the vector of unobserved failure times. The latter is simply obtained as a normal density conditional on $Z \geq c$, where c is the vector of censoring times. For an augmented data set $X = (Y,Z)$, the augmented posterior density $p(\beta_0,\beta_1,\sigma|X)$ can be factorised exactly as

$$p(\beta_0,\beta_1|\sigma,X)p(\sigma|X)\,,$$

where (Lee, 1989)

$$\begin{aligned}
\sigma^{-2}|X &\sim S^{-1}\chi^2(n-2) \\
\beta_1|\sigma,X &\sim N(b,\sigma^2/S_{vv}) \\
\alpha|\sigma,X &\sim N(a,\sigma^2/n)
\end{aligned}$$

and α,β_1 are conditionally independent given σ, where $\alpha = \beta_0 + \beta_1\bar{v}$. Here $S = S_{xx} - S_{xv}^2/S_{vv}$, the residual sum of squares associated with the augmented data set, where $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$, $S_{vv} = \sum_{i=1}^n (v_i - \bar{v})^2$, $S_{xv} = \sum_{i=1}^n (x_i - \bar{x})(v_i - \bar{v})$, $a = \bar{x}$ and $b = S_{xv}/S_{vv}$.

In order to illustrate the PMDA algorithms we examine the marginal posterior density of σ. In all cases it follows from equation (7) that the approximation to this density is proportional to

$$\sigma^{-39} \sum_{j=1}^m w_j S_j^{19} \exp\left(-\frac{1}{2}\sigma^{-2} S_j\right) \tag{9}$$

with the appropriate weights w_j, where S_j is the residual sum of squares for the jth augmented data set. Following Wei & Tanner (1990a), 5000 samples were drawn from $p(Z|Y,\hat{\theta})$. Fig. 1 shows the PMDA 1 approximation (dotted line), three different runs of PMDA 2 (dashed lines) and the exact (solid line) σ marginal (the latter obtained from a long run of the simulation-consistent algorithm to be given in §4). As can be seen from these plots, there is a noticeable discrepancy between PMDA 1 and the other curves. PMDA 2 represents an improvement, but also indicates that the convergence of the process is quite poor. Of course, in this example there is actually no need to use the asymptotic approximation in PMDA 2 since the complete augmented

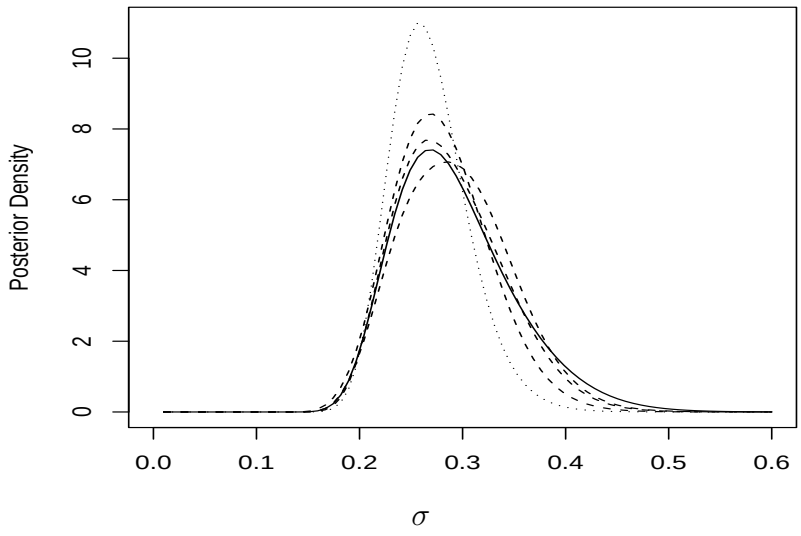

Figure 1: Censored regression. Marginal posterior density of σ: PMDA 1 (dotted line); PMDA 2 (dashed lines); exact (solid line).

posterior density is available so that we can use PMDA-Exact. Fig. 2 presents again the exact marginal density of σ (solid line) along with three different runs of the PMDA-Exact approximation, again with $m = 5000$ (dashed lines). Notice that the PMDA-Exact estimate is still subject to considerable sampling variability.

It is seen from this example that an important practical issue concerns the choice of importance function on which both PMDA 2 and PMDA-Exact are based. In the single parameter genetic linkage example the conditional predictive distribution $P(Z|Y,\hat{\theta})$ turns out to be a reasonable choice of importance function (Wei & Tanner, 1990a). In the censored regression example, however, convergence is poor using this importance function. Thus some doubt has been cast on the validity of $P(Z|Y,\hat{\theta})$ as an appropriate general-purpose importance function. In the next section we will investigate the use of an alternative importance function based on the formulae in §2.

4 Modified poor man's data augmentation algorithms

In this section we will investigate versions of the PMDA 1, PMDA 2 and PMDA-Exact algorithms in which $p(Z|Y,\hat{\theta})$ is replaced by the p-bar formula $\bar{p}(Z|Y)$ given

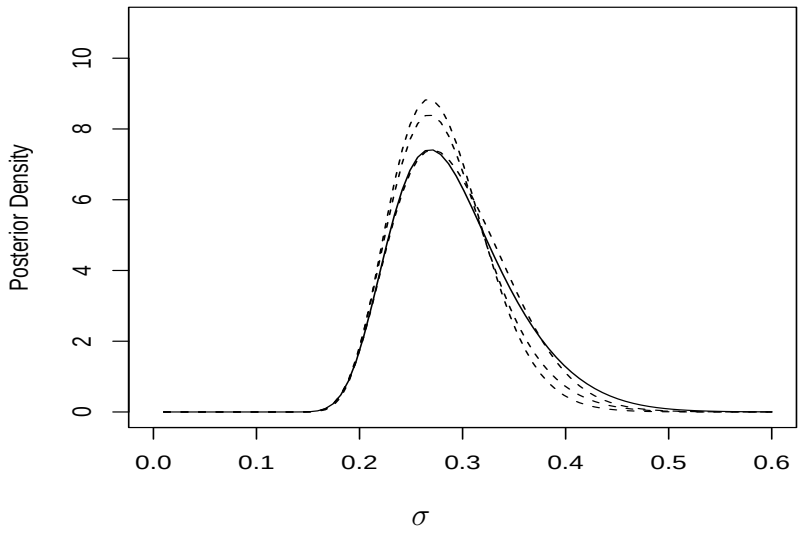

Figure 2: Censored regression. Marginal posterior density of σ: PMDA-Exact (dashed lines); exact (solid line).

by formula (3) of §2. We will denote the modified versions by PMDA 1′, PMDA 2′ and PMDA-Exact′.

The modified algorithm PMDA 1′ is obtained by simply replacing $p(Z|Y,\hat{\theta})$ by $\bar{p}(Z|Y)$ at step a in the PMDA 1 algorithm. Since we can easily sample from (3) there is no need for an importance sampling step in this algorithm, unlike when using the Tierney and Kadane approximation (5).

In situations where $p(Z|Y)$ is difficult to compute, the Tierney and Kadane approximation (5) is available, as in §3. PMDA 2′ is obtained by using (3) instead of $p(Z|Y,\hat{\theta})$ as the importance function and takes the following form.

a1. Generate $z_1,\ldots,z_m$ from $\bar{p}(Z|Y)$.

a2. Calculate

$$w_j = \frac{p(\theta_j^*)L^*(\theta_j^*;z_j)p(z_j|Y,\hat{\theta})}{|J_j^*|^{1/2}L^*(\hat{\theta};z_j)\bar{p}(z_j|Y)}\ ,\ j=1,\ldots,m\,.$$

b. Approximate the posterior density by

$$\sum_{j=1}^{m} w_j p(\theta|Y,z_j)\Big/\sum_{j=1}^{m} w_j\,.$$

Finally suppose that, as in §3, we are able to evaluate the constant of proportionality in the augmented posterior density. Then the posterior density $p(\theta|Y)$ may

be obtained by applying importance sampling based on the importance function (3). The PMDA-Exact′ algorithm has the same form as the PMDA 2′ algorithm but with weights given by

$$w_j = \frac{p(z_j|Y,\hat{\theta})}{p(\hat{\theta}|Y,z_j)\bar{p}(z_j|Y)} \ , \ j = 1,\ldots,m\,.$$

4.1 Example: censored regression (continued)

We first obtain all the quantities required for computation of the p-bar formula (3). First, solution of the equations (2) produces the quantities $\theta^{i\pm}$, where $\theta = (\beta_0, \beta_1, \lambda)$. Next, again assuming the non-informative prior $p(\theta) \propto 1$, we obtain the values of π_i and α_i^- associated with $(\beta_0, \beta_1, \lambda)$ as $(0.372, 0.331, 0.297)$ and $(0.593, 0.376, 0.406)$ respectively.

As in §3.1 we examine the marginal posterior density of σ. Based on equation (9), Fig. 3 presents posterior density approximations for σ arising from PMDA 1′ (dotted line) and three runs of the PMDA 2′ algorithm based on $m = 5000$ (dashed lines), along with the exact density. These graphs, which should be compared with those for PMDA 1 and PMDA 2 in Fig. 1, demonstrate the superiority of the PMDA 1′ approximation over the PMDA 1 approximation and show that the PMDA 2′ algorithm produces stable and accurate estimates in this case. Fig. 4 presents three runs of the PMDA-Exact′ algorithm with $m = 5000$. Again these graphs, which should be compared with those in Fig. 2, indicate that the PMDA-Exact′ algorithm produces very stable estimates. Generally, the p-bar formula provides a much more stable importance function than $p(Z|Y,\hat{\theta})$. The above results were obtained using the signed-root version of equation (3), but very similar results were obtained using the alternative form based directly on the parameter θ. We note that the exact form of the density plotted in all the previous figures was obtained using the PMDA-Exact′ algorithm with $m = 20000$, for which the Monte Carlo error was negligible.

Finally note that, if desired, independent simulations of (Z, θ) from its posterior distribution may be obtained by resampling the z_j values according to their importance weights w_j and sampling from the augmented posterior distribution of θ given the sampled value Z.

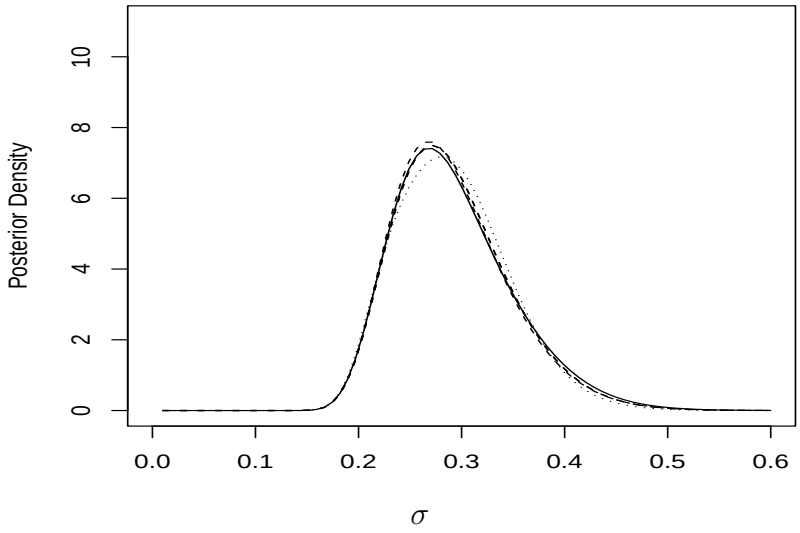

Figure 3: Censored regression. Marginal posterior density of σ: PMDA $1'$ (dotted line); PMDA $2'$ (dashed lines); exact (solid line).

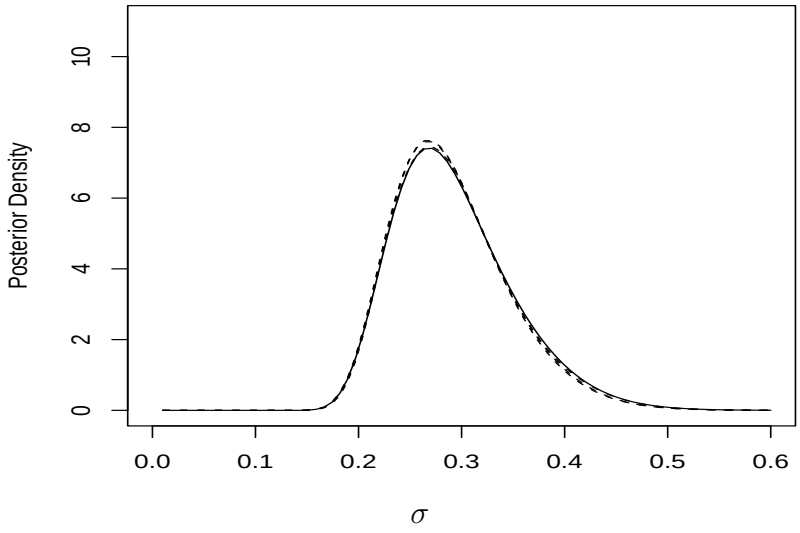

Figure 4: Censored regression. Marginal posterior density of σ: PMDA-Exact$'$ (dashed lines); exact (solid line).

5 Metropolis-Hastings implementations

In the previous section we have shown that the p-bar formula provides a stable importance function for use within poor man's data augmentation algorithms. However, it also has potential application as a proposal distribution within Metropolis-Hastings schemes for models that are not analytically tractable. See Hastings (1970) or, for example, Tanner (1996) for a description of the Metropolis-Hastings algorithm. We will give an example in which the conditional predictive density $p(Z|Y,\theta)$ is available for computation of the p-bar formula $\bar{p}(Z|Y)$ and one in which $p(Z|Y,\theta)$ is unavailable analytically but where an approximation $q(Z|Y,\theta)$ to $p(Z|Y,\theta)$ is used to produce an approximation $\bar{q}(Z|Y)$ to $\bar{p}(Z|Y)$. One possible form of Metropolis-Hastings algorithm incorporating $\bar{q}(Z|Y)$ as a proposal distribution is as follows. Here $q(\theta'|Y,Z',\theta)$ is an arbitrary proposal distribution for θ' given the previous value θ and current value Z'.

a. Given the current value (Z,θ), generate Z' from $\bar{q}(Z'|Y)$ and θ' from $q(\theta'|Y,Z',\theta)$.

b. Accept the new value (Z',θ') with probability

$$\rho = \begin{cases} \min\left\{\frac{u\{(Z',\theta'),(Z,\theta)\}}{u\{(Z,\theta),(Z',\theta')\}},1\right\} , & u\{(Z,\theta),(Z',\theta')\}>0 \\ 1 , & u\{(Z,\theta),(Z',\theta')\}=0 \end{cases} , \tag{10}$$

where

$$u\{(Z,\theta),(Z',\theta')\} = p(\theta)p(Y,Z|\theta)\bar{q}(Z'|Y)q(\theta'|Y,Z',\theta)\,.$$

This construction defines a Markov chain with equilibrium distribution $p(Z,\theta|Y)$.

In the following examples we use an independence sampler of the form $q(\theta'|Y,Z')$, giving rise to an independence chain. Clearly, if $\bar{q}(Z|Y)$ and $q(\theta|Y,Z)$ are good approximations to $p(Z|Y)$ and $p(\theta|Y,Z)$ respectively then we would expect a high level of acceptance and a low level of dependence in the chain.

5.1 Example: censored regression (continued)

We illustrate the above use of p-bar for the censored regression model of §3.1. We do not make detailed comparisons here with alternative computational schemes, but we will show that the resulting algorithm is competitive with a block Gibbs sampler and in fact achieves better mixing. Recall that $X=(Y,Z)$. Since the exact posterior density $p(\theta|X)$ of θ based on the augmented data is available analytically here and

is straightforward to sample from (see §3.1), we can take $q(\theta|X) = p(\theta|X)$, the exact posterior density of θ based on the augmented data. Similarly, since $p(Z|Y,\theta)$ is available analytically and is straightforward to sample from, we take $\bar{q}(Z|Y) = \bar{p}(Z|Y)$. Clearly some cancellation now occurs in (10) and the formula for ρ just depends on the ratio of the p-bar formula and the exact conditional density of Z given Y.

Fig. 5 shows the trace plot for the parameter λ for the final 1000 iterations of a run of length $m = 5000$ of the Metropolis-Hastings sampler using the p-bar formula based on the signed-root r. Table 1 gives the corresponding acceptance rate, estimated posterior mean of λ and its estimated Monte Carlo standard error after discarding the first 50 values. The performance of the Metropolis-Hastings sampler was very stable over repeated runs and the figures in Table 1 represent a typical result. Although there was no problem with this scheme in the present example, for general implementation we recommend that the proposal distribution of Z be made slightly wider than the p-bar formula. This avoids the possible occurrence of occasional high importance weights with associated small probabilities of acceptance at the next draw, leading to the chain becoming stuck for several iterations. This modification is easily accomplished by replacing $\sqrt{d}$ in formulae (2) and (4) by $f\sqrt{d}$, where $f > 1$ is a suitable scaling factor. If f is taken to be too large then the overall acceptance rate will be too small, leading to an inefficient algorithm. We found the value $f = 1.2$ to be a good compromise in general. This has been implemented here for the p-bar formula based directly on the parameter θ and the relevant figures appear in Table 1. Although there is a small reduction in the acceptance rate, the overall Monte Carlo precision is of a similar order to that for the $r-$based version.

Table 1 also gives the corresponding quantities for a run with $\bar{p}(Z|Y)$ replaced by $p(Z|Y,\hat{\theta})$ (referred to as p-hat in the table). However, we note that this algorithm is quite unreliable with different runs of length 5000 producing quite different results, arising from the chain occasionally becoming stuck in states with high importance weights. Finally, for comparison Table 1 gives the results for a run of length 5000 of a block Gibbs sampler based on the conditional densities $p(Z|Y,\theta)$ and $p(\theta|Y,Z)$. The performance of the Gibbs sampler was fairly stable in repeated runs, but within runs it suffered from relatively high autocorrelations; Fig. 6 shows the trace plot for λ for

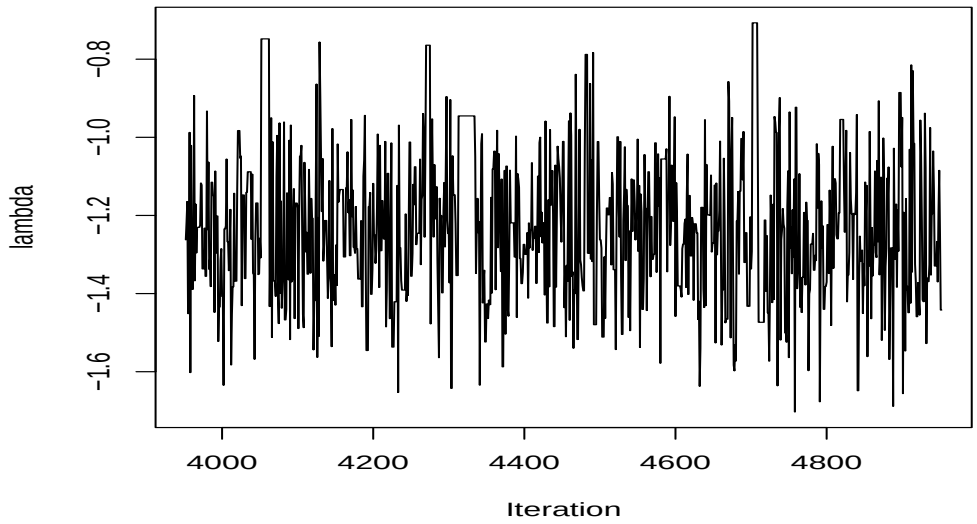

Figure 5: Censored regression. Trace plot of Metropolis-Hastings algorithm incorporating the p-bar formula.

Table 1: MCMC results for the censored regression example ($m = 5000$)

Algorithm	Acceptance rate	Posterior mean of λ	Monte Carlo s.e.
Metropolis-Hastings (p-bar, r)	0.77	-1.257	0.0044
Metropolis-Hastings (p-bar, θ)	0.69	-1.258	0.0045
Metropolis-Hastings (p-hat)	0.50	-1.286	0.0100
Block Gibbs	1.00	-1.242	0.0078

the final 1000 iterations. Overall, the Metropolis-Hastings scheme based on the p-bar formula provides a stable algorithm with good mixing properties here. All standard errors were computed by considering the initial sequence adaptive estimators based on sums of pairs of autocorrelations (Geyer, 1992). For comparison with asymptotic results, the MLE of λ is -1.350 while the approximation to the posterior mean of λ obtained from the higher-order formula derived by Sweeting & Kharroubi (2003) is -1.251.

6 Example: normal hierarchical model

As a second illustration of the use of the p-bar formula within both data augmentation and MCMC we consider a normal hierarchical model. Exact poor man's data aug-

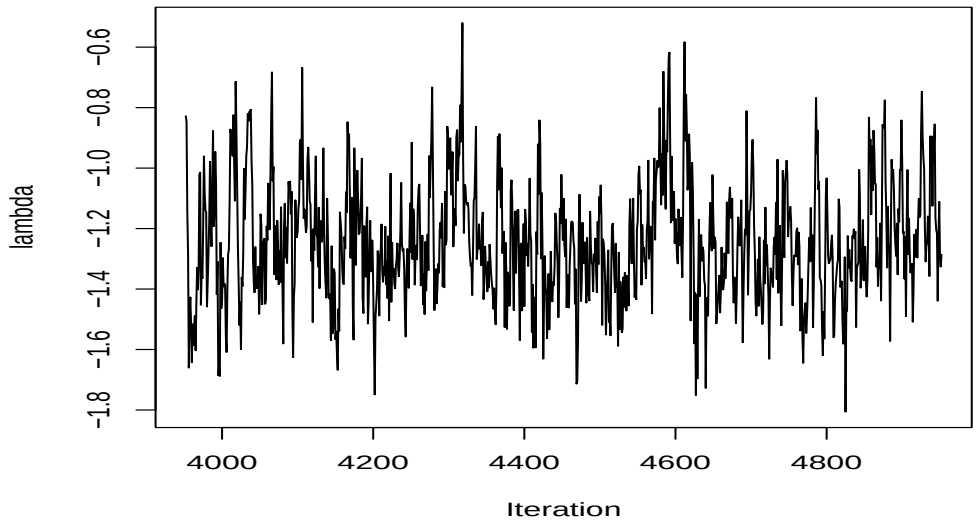

Figure 6: Censored regression. Trace plot of block Gibbs algorithm.

mentation is available in the simple case of normally distributed random effects, which we consider in §6.1. However, for Student t distributed random effects considered in §6.2 the likelihood does not have a closed form and we use MCMC methods. For the data we consider in neither case does there exist a local maximum of the likelihood function. However, we can still employ the p-bar approximation, even though it is motivated by regular asymptotic theory. These examples will illustrate the important device of exponential tilting, which here allows use of the p-bar formula.

We consider the random effects model

$$y_{ij} = \mu_i + \epsilon_{ij} \ , j = 1, \ldots, J_i \, , \, i = 1, \ldots, I \, ,$$

where y_{ij} is the jth observation from the ith group, μ_i is the effect of the ith group and ϵ_{ij} is the error. Suppose, assuming conditional independence throughout, that $\epsilon_{ij}|\mu_i \sim N(0, \sigma^2)$ so that $y_{ij}|\mu_i \sim N(\mu_i, \sigma^2)$. Let

$$y_{i.} = \sum_{j=1}^{J_i} y_{ij}/J_i \quad \text{and} \quad y_{..} = \frac{1}{I} \sum_{i=1}^{I} y_{i.} \, .$$

6.1 Normal random effects

Suppose further that $\mu_i \sim N(\mu, \tau^2)$. Then, unconditionally, (Box & Tiao, 1973)

$$y_{i.} \sim N(\mu, \phi_i^2/J_i)\,,$$

where $\phi_i^2 = \sigma^2 + J_i\tau^2$ is the between groups expected mean square (E.M.S) and

$$S_\sigma = \sum_{i=1}^{I}\sum_{j=1}^{J_i}(y_{ij} - y_{i.})^2 \sim \sigma^2\chi^2(W - I)\,,$$

where $W = \sum_{i=1}^{I} J_i$ and S_σ is the within groups sum of squares (S.S.).

Since S_σ is sufficient for σ^2, the likelihood function is proportional to

$$(\sigma^2)^{-\nu_\sigma/2}\left(\prod_{i=1}^{I}\phi_i^2\right)^{-1/2}\exp\left[-\frac{1}{2}\left\{\sum_{i=1}^{I}\frac{J_i(y_{i.}-\mu)^2}{\phi_i^2} + \frac{\nu_\sigma m_\sigma}{\sigma^2}\right\}\right]\,,$$

where $\nu_\sigma = \sum_{i=1}^{I}(J_i - 1) = W - I$ is the within groups degrees of freedom (d.f.) and $m_\sigma = S_\sigma/\nu_\sigma$ is the within groups mean square (M.S.). Further define $\nu_\tau = I-1$, $S_\tau = \sum_{i=1}^{I} J_i(y_{i.} - y_{..})^2$ and $m_\tau = S_\tau/\nu_\tau$ to be the between groups d.f., S.S. and M.S..

Before proceeding with the PMDA algorithms, we need to obtain the augmented posterior and conditional predictive distributions. Let $Z = (\mu_1, \ldots, \mu_I)$, $Y = (y_{11}, \ldots, y_{IJ_I})$ and assume that (τ^2, μ) and σ^2 are a priori independent with priors specified by $p(\tau^2, \mu) \propto 1$ and $p(\sigma^2) \propto 1/\sigma^2$. This particular noninformative specification ensures propriety of the posterior distribution (Tanner, 1996). For an augmented data set $X = (Y, Z)$, the augmented posterior density $p(\tau^2, \mu, \sigma^2|X)$ can be factorised as

$$p(\tau^2|X)p(\mu|\tau^2, X)p(\sigma^2|X)\,,$$

where (Lee, 1989)

$$\begin{aligned}
\tau^2|X &\sim S'_\tau/\chi^2(I-3)\\
\mu|\tau^2, X &\sim N(\bar{\mu}, \tau^2/I)\\
\sigma^2|X &\sim S'_\sigma/\chi^2(W)
\end{aligned}$$

with (τ^2, μ) and σ^2 a posteriori independent, where $S'_\sigma = \sum_{i=1}^{I}\sum_{j=1}^{J_i}(y_{ij} - \mu_i)^2$, $S'_\tau = \sum_{i=1}^{I}(\mu_i - \bar{\mu})^2$ and $\bar{\mu} = \sum_{i=1}^{I}\mu_i/I$. The conditional predictive density $p(\mu_i|\tau^2, \mu, \sigma^2, Y)$ is normal with mean given by the weighted average

$$\mu\left(\frac{1/\tau^2}{1/\tau^2 + J_i/\sigma^2}\right) + y_{i.}\left(\frac{J_i/\sigma^2}{1/\tau^2 + J_i/\sigma^2}\right)$$

Table 2: Box and Tiao's generated data for the two-component model

Batch	1	2	3	4	5	6
Individual	7.298	5.220	0.110	2.212	0.282	1.722
observations	3.846	6.556	10.386	4.852	9.014	4.782
	2.434	0.608	13.434	7.092	4.458	8.106
	9.566	11.788	5.510	9.288	9.446	0.758
	7.990	-0.892	8.166	4.980	7.198	3.758
$y_{i.}$	6.2268	4.6560	7.5212	5.6848	6.0796	3.8252

Table 3: Analysis of variance table for the generated data

Source	S.S.	d.f.	M.S.	E.M.S.
Between batches	$S_\tau = 41.6816$	$\nu_\tau = 5$	$m_\tau = 8.3363$	$\sigma^2 + 5\tau^2$
Within batches	$S_\sigma = 358.7014$	$\nu_\sigma = 24$	$m_\sigma = 14.9459$	σ^2
Total	400.3830	29		

of μ and $y_{i.}$ and with variance $(1/\tau^2 + J_i/\sigma^2)^{-1}$.

To illustrate the PMDA algorithms, we will focus on the marginal posterior density of τ. In all cases it follows from equation (7) that an estimate of $p(\tau|Y)$ is proportional to

$$\tau^{-(I-2)} \sum_{j=1}^{m} w_j (S'_{\tau j})^{(I-3)/2} \exp\left(-\frac{1}{2}\tau^{-2} S'_{\tau j}\right),$$

where $S'_{\tau j}$ is the between groups S.S. for the jth augmented data set.

We illustrate the use of the PMDA-Exact methods for the generated data set reported in Box & Tiao (1973). The data, with $I = 6$ and $J_i = J = 5$, are displayed in Table 2 and an analysis of variance table is provided in Table 3. It is seen from Table 3 that the usual unbiased estimate for the variance component τ^2 is negative, which is clearly objectionable. Furthermore, in this case the MLE does not exist and we therefore need to find a way to circumvent this problem. The method used here is the device of *exponential tilting*. In general, for $c \in R^d$ define $l_c(\theta) = l(\theta) - c^T\theta$. Then the turning points of l_c satisfy $l'(\theta) = c$. In particular, if we choose $c = l'(\bar{\theta})$ then $\bar{\theta}$ will be a local maximum of l_c, provided that $j(\bar{\theta}) > 0$. Furthermore, it follows from the more general treatment in Sweeting & Kharroubi (2003) that the p-bar formula (3) will continue to be of asymptotic accuracy $O(n^{-2})$ on replacing $l(\theta)$ by $l_c(\theta)$ and $p(\theta)$ by $p_c(\theta) \equiv p(\theta)e^{c^T\theta}$ provided that $c = O(1)$. In any case, we will use (3) as a proposal density for a suitably tilted version of l.

Write $\gamma = \log\tau$ and $\lambda = \log\sigma$. For the data in Table 2, using the parameterisation

$\theta = (\gamma, \mu, \lambda)$ the choice $c = (-2.5, 0, 0)$ produced a modified loglikelihood l_c that was reasonably symmetric around the (modified) MLE $\hat{\theta} = (0.5695, 5.6656, 1.2996)$ with (modified) observed information matrix

$$\begin{pmatrix} 3.098 & 0.000 & -1.558 \\ 0.000 & 1.032 & 0.000 \\ -1.558 & 0.000 & 56.178 \end{pmatrix}.$$

Here the marginal posterior density of σ is extremely well estimated by both the PMDA-Exact and PMDA-Exact$'$ algorithms, even with run lengths of less than 1000. Fig. 7 presents the marginal posterior density of τ based on three different runs of PMDA-Exact with $m = 1000$ (dashed lines) and the exact density (solid line) while Fig. 8 presents the corresponding plots for PMDA-Exact$'$. As can be seen, the output from PMDA-Exact$'$ is less variable than that from PMDA-Exact. In these figures the exact density was obtained using PMDA-Exact$'$ with $m = 50000$, for which the Monte Carlo error was negligible. The above results were obtained using the signed-root version of equation (3), but very similar results were obtained using the alternative form based directly on the parameter θ. The very good results obtained from using importance samplers based on the modified MLE $\hat{\theta}$ are quite remarkable given our use of an initial exponential tilt and in view of the relatively small amount of information in the sample about (γ, μ) (see, for example, the modified observed information matrix).

6.2 Student t random effects

In the previous example all the conditional distributions required for the poor man's data augmentation schemes were available analytically. However, if we elaborate the model by assuming Student t random effects then closed form expressions are no longer available. Suppose then that $\mu_i \sim t_\nu(\mu, \tau^2)$, the Student t distribution with location parameter μ, scale parameter τ and (known) degrees of freedom ν. Since it is not possible to write down an expression for the likelihood function in closed form, some form of Monte Carlo maximum likelihood is needed to obtain the necessary ingredients for computation of the p-bar formula. We do not review the literature on this topic here as this is not the main focus of the present paper. It suffices to mention that there are a number of such routines, such as the MCEM algorithm Wei & Tanner (1990a), variants of this algorithm and simulated maximum likelihood. Here a simple

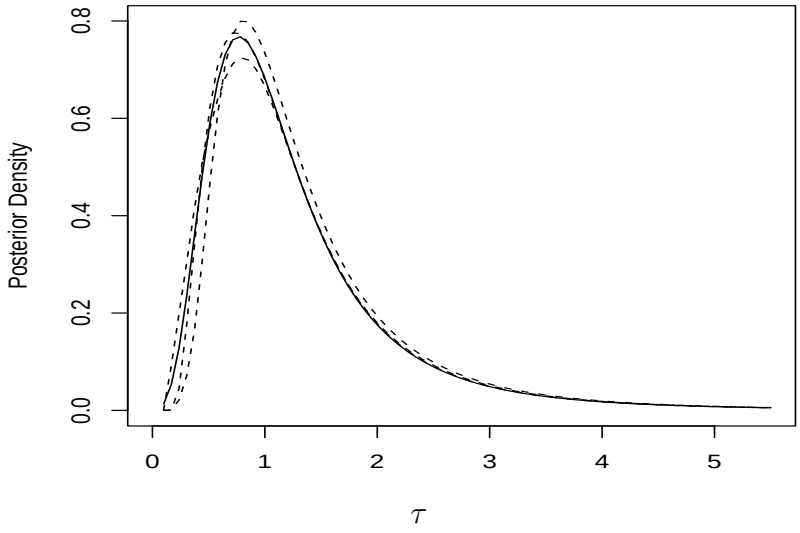

Figure 7: Variance components (normal model). Marginal density of τ. PMDA-Exact (dashed lines); exact (solid line).

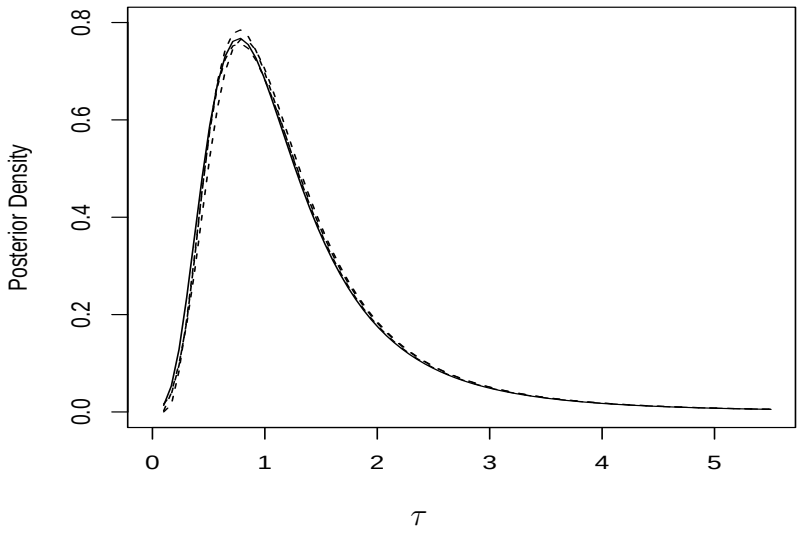

Figure 8: Variance components (normal model). Marginal density of τ. PMDA-Exact$'$ (dashed lines); exact (solid line).

Monte Carlo likelihood approximation worked well in the present example, producing the MLE and an approximation to the observed information matrix very quickly. In general, since the output from this initial phase is only used to form a proposal distribution accuracy is not of the utmost importance. For example, one could obtain a rough estimate of $\hat{\theta}$ followed by exponential tilting based on this estimate. Note that formula (3) based on the versions of θ^{i-} and θ^{i+} given in (4) is more attractive here than that based on signed-roots since it avoids excessive conditional maximization. We have therefore used this variant of the p-bar formula here, although we were also able to compute the signed-root version without too much difficulty; the final results were very similar.

We illustrate use of the Metropolis-Hastings algorithm as described in §5 in the case of a t_4 random effects model for Box & Tiao's data set given in Table 2. As in §6.1 the MLE does not exist here so again we apply an exponential tilt, as described in that section. Here we took $c = (-1.5, 0, 0)$, which produced the (modified) MLE $\hat{\theta} = (0.1016, 5.6572, 1.2966)$ of the parameter $\theta = (\gamma, \mu, \lambda)$ with (modified) observed information matrix

$$\begin{pmatrix} 2.653 & -0.025 & -1.024 \\ -0.025 & 1.461 & 0.244 \\ -1.024 & 0.244 & 56.720 \end{pmatrix}.$$

For this model neither the augmented posterior density nor the conditional predictive density of the latent data given θ and Y are available in closed form. We therefore based the proposal densities $q(\theta|Y, Z)$ and $\bar{q}(Z|Y)$ on the corresponding densities under the normal model of §6.1, matching the quartile deviation of the normal random effects distribution with that of the t_4 distribution.

Using the p-bar formula based on the parameter θ and with no modification to the definitions of θ^{i-} and θ^{i+}, the Metropolis-Hastings sampler was run for $m = 5000$ iterations. After discarding the first 100 values, the acceptance rate was $\rho = 0.53$ and the estimated posterior mean of γ was 1.474 with associated Monte Carlo standard error 0.018. Although inferior, use of $p(Z|Y, \hat{\theta})$ in place of $\bar{p}(Z|Y)$ was actually not too bad in this case, giving an acceptance rate of $\rho = 0.37$ and a Monte Carlo standard error of 0.027. However, we would not recommend using this algorithm in general as it has a tendency to get stuck in a state, as discussed in §3.1. Finally Fig. 9 shows the marginal posterior density of τ obtained from thinning and smoothing the MCMC

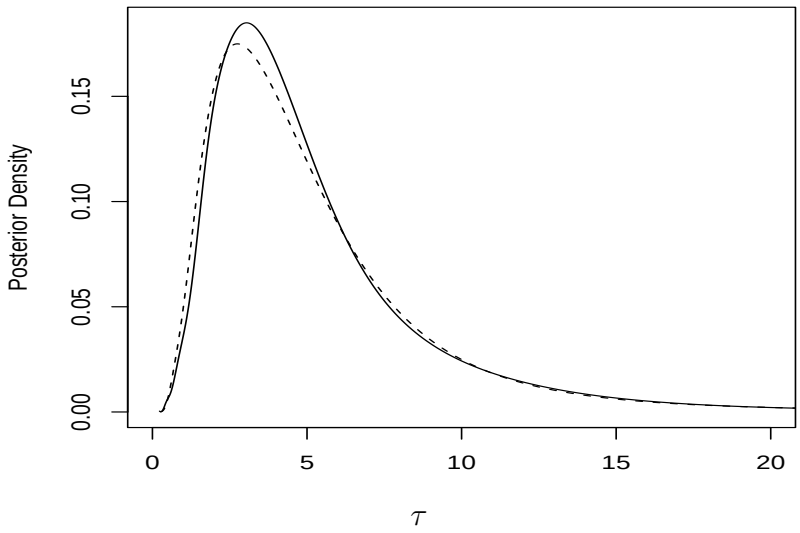

Figure 9: Variance components (Student t model). Marginal posterior density of τ based on Metropolis-Hastings incorporating the p-bar formula (dashed line); exact (solid line).

output for τ, along with the corresponding density obtained from a much longer run for which the Monte Carlo error is negligible.

7 Discussion

In this paper we have explored the use of hybrid methods involving data augmentation, importance sampling, MCMC and asymptotics and have provided some numerical illustrations. In the case of poor man's data augmentation algorithms that incorporate importance sampling, the validity of the estimated conditional predictive distribution $p(Z|Y,\hat{\theta})$ as an importance function is in doubt as convergence is seen to be quite poor. However, we have shown that the p-bar formula provides a stable and effective importance function in such schemes. The new non-iterative algorithms PMDA $2'$ and PMDA-Exact$'$ have been shown to yield accurate approximations to the true posterior distribution. These algorithms are easy to implement and greatly reduce the computational burden of the data augmentation algorithm. Additionally they are capable of generating independent draws of (Z,θ). Furthermore, PMDA $1'$ may provide a reasonable approximation in its own right. Notice that we could define

a variant of PMDA $2'$ by replacing formula (5) by formula (3). We do not recommend this when the dimension of Z is high, however, since the asymptotic accuracy of (3) is less than that of (5) in such a case. Thus (3) and (5) play complementary roles, with (3) being most suitable as an importance function, and (5) as an asymptotic approximation when Z is high-dimensional.

The version of the p-bar formula based directly on the abscissae (4) requires less computation than the signed-root version based on the equations (2). However, as discussed in §2, formula (3) is no longer invariant to reparameterisation so some care may be needed in selecting an appropriate parameterisation. Note however that the choice of parameterisation will only affect computational efficiency in the simulation schemes described here. As mentioned in Sweeting & Kharroubi (2003), we have noticed little difference in the quality of the expectation approximations under alternative reasonable parameterisations. Generally, we would expect any parameterisation that is approximately linear in the signed-root loglikelihood ratio statistic to perform well. Some further discussion is given in Sweeting & Kharroubi (2003).

The p-bar formula can be used more widely as an importance sampler. Here we have demonstrated its use as a proposal distribution within a Metropolis-Hastings scheme. This approach may be used in cases where the integration in (6) is intractable, through the initial use of a suitable Monte Carlo maximum likelihood algorithm. Furthermore, it is not confined to situations where the observed likelihood possesses a maximum, since we may be able to apply a suitable exponential tilting. The methods described in this paper therefore have potential application to other models which incorporate latent variables or missing data, such as models for general censored survival data, errors-in-variables models, longitudinal models and generalised linear and nonlinear random effects models.

The use of the p-bar formula requires the availability of suitable programmes to compute the quantities $\theta^{i\pm}, \pi_i$ and $\alpha_i^{\pm}$ needed for formula (3). However, generic programmes may be written to carry out these computations. Once these are available there will be relatively little additional computational effort over competing algorithms since, for a specific application, routines will always be required to compute the likelihood, augmented likelihood and conditional densities.

Appendix

Definition of the weights in equation (3)

Let $\nu_i(\theta) = p(\theta)|j^{(i+1)}(\theta)|^{-1/2}$. Write $l_i^- = l_i(\theta_i^-)$, $l_i^+ = l_i(\theta_i^+)$, $\nu_i^- = \nu_i(\theta_i^-)$, $\nu_i^+ = \nu_i(\theta_i^+)$, where $\theta_i^- = (\hat{\theta}_{i-1}, \theta^{i-})$, $\theta_i^+ = (\hat{\theta}_{i-1}, \theta^{i+})$ and θ^{i-}, θ^{i+} are obtained from either equation (2) or equation (4). Now let $\tau^i = (\nu_i^-/l_i^-) + (-\nu_i^+/l_i^+)$ and define the weights

$$\alpha_i^- = (\tau^i)^{-1}(\nu_i^-/l_i^-), \;\; \alpha_i^+ = (\tau^i)^{-1}(-\nu_i^+/l_i^+)\,.$$

Finally define

$$\pi_i = \left\{ \sum_{j=1}^{d} |J^{(j)}|^{1/2} (\omega^j)^{-1} \tau^j \right\}^{-1} |J^{(i)}|^{1/2} (\omega^i)^{-1} \tau^i\,,$$

where

$$\omega^i = \{-r^i(\theta_i^-)\}^{-1} + \{r^i(\theta_i^+)\}^{-1}\,.$$

Note that when θ^{i-} and θ^{i+} are obtained from equation (2), $\omega^i = 2/\sqrt{d}$ and hence cancels in the expression for π_i.

Asymptotic error of equation (5)

Wei & Tanner (1990a) use equation (2.4) of Tierney & Kadane (1986) to obtain the approximation (5) to the posterior predictive distribution $p(Z|Y)$. When the conditional predictive distribution $p(Z|Y,\theta)$ is of order $O(1)$, it follows from Tierney & Kadane (1986) that (5) is accurate to $O(n^{-2})$. Here we further investigate the error in this approximation.

It follows (Sweeting, 1996) that the augmented posterior density $p(\theta|Y,Z)$ may be approximated by

$$|J^*|^{1/2} \left\{ (2\pi)^{d/2} s^* \right\}^{-1} \left\{ \frac{p(\theta) L^*(\theta; Z)}{p(\theta^*) L^*(\theta^*; Z)} \right\} \{1 + \eta^{*T}(\theta - \theta^*)\} \tag{11}$$

to $O(n^{-2})$, where η^* is $O(n^{-1})$ and the normalising constant $s^* = 1 + O(n^{-1})$. Substituting (11) in expression (8), we obtain the $O(n^{-2})$ approximation

$$p(Z|Y) = (2\pi)^{d/2} |J^*|^{-1/2} s^* \left\{ \frac{p(\theta|Y) p(Z|Y,\theta) p(\theta^*) L^*(\theta^*; Z)}{p(\theta) L^*(\theta; Z)} \right\}.$$

On taking $\theta = \hat{\theta}$ we see that, to $O(n^{-2})$,

$$p(Z|Y) \propto s^* \left\{ \frac{p(\theta^*) L^*(\theta^*; Z) p(Z|Y,\hat{\theta})}{|J^*|^{1/2} L^*(\hat{\theta}; Z)} \right\}. \tag{12}$$

Now from Sweeting (1996) the correction term s^* may be taken as $1 + b^*$ to $O(n^{-2})$, where $b^* = O(n^{-1})$. When $\log p(Z|Y,\theta)$ is $O(1)$ it follows (Tierney & Kadane, 1986) that the error in (5) is $O(n^{-2})$. This can also be seen from (12) by showing that $b^* = b + O(n^{-2})$, where b is the corresponding quantity for the original data set Y and therefore does not depend on Z. However, if $\log p(Z|Y,\theta)$ is $O(n)$ then in general $b^* - b = O(n^{-1})$ and so in this case we see from (12) that the relative error in (5) is $O(n^{-1})$.

One reason for using the representation (11) to analyse the error in (5) is that (11) reveals clearly that the accuracy of (5) depends on the form of the augmented posterior, which is likely to be better suited to asymptotic approximation than $p(\theta|Y)$, especially when the dimension of Z is high.

References

Box, G. E. P. and G. Tiao (1973). *Bayesian Inference in Statistical Analysis*. Addison-Wesley.

Crawford, D. E. (1970). Analysis of incomplete life test data on motorists. *Insulation Circuits* **16**, 43–48.

DiCiccio, T. J., Kass, R. E., Raftery, A. and Wasserman, L. (1997). Computing Bayes factors by combining simulation and asymptotic approximations. *J. Amer. Statist. Assoc.* **92**, 903–915.

Evans, M. and Swartz, T. (1995). Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems. *Stat. Sci.* **10**, 254–272.

Geyer, C. J. (1992). Practical Markov Chain Monte Carlo. *Stat. Sci.* **7**, 473-483.

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

Lee, P. M. (1989). *Bayesian Statistics*. Wiley.

Ripley, B. (1987). *Stochastic Simulation*. Wiley.

Schmee, J. and Hahn, G. J. (1979). A simple method for regression analysis with censored data. *Technometrics* **21**, 417–432.

Sweeting, T. J. (1996). Approximate Bayesian computation based on signed roots of log-density ratios (with Discussion). In J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, editors, *Bayesian Statistics 5*. Oxford University Press.

Sweeting, T. J. and Kharroubi, S. A. (2003). Some new formulae for posterior expectations and Bartlett corrections. *Test* **12**, 497–521.

Tanner, M. A. (1996). *Tools for Statistical Inference*. Springer- Verlag.

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

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

Wei, G. C. G. and Tanner, M. A. (1990a). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm. *J. Amer. Statist. Assoc.* **85**, 699–704.

Wei, G. C. G. and Tanner, M. A. (1990b). Posterior computations for censored regression data. *J. Amer. Statist. Assoc.* **85**, 829–839.