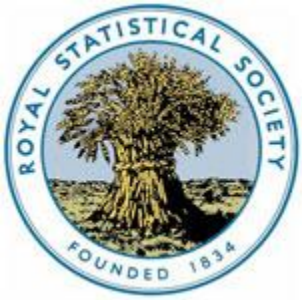

J. R. Statist. Soc. B (2008)
70, *Part* 3, *pp.* 589–607

Marginal likelihood estimation via power posteriors

N. Friel

University College Dublin, Republic of Ireland

and A. N. Pettitt

Queensland University of Technology, Brisbane, Australia, and Lancaster University, UK

[Received November 2005. Final revision November 2007]

Summary. Model choice plays an increasingly important role in statistics. From a Bayesian perspective a crucial goal is to compute the marginal likelihood of the data for a given model. However, this is typically a difficult task since it amounts to integrating over all model parameters. The aim of the paper is to illustrate how this may be achieved by using ideas from thermodynamic integration or path sampling. We show how the marginal likelihood can be computed via Markov chain Monte Carlo methods on modified posterior distributions for each model. This then allows Bayes factors or posterior model probabilities to be calculated. We show that this approach requires very little tuning and is straightforward to implement. The new method is illustrated in a variety of challenging statistical settings.

Keywords: Bayes factor; Hidden Markov model; Model choice; Regression; Survival analysis

1. Introduction

Suppose that data $\mathbf{y}$ are assumed to have been generated by one of M models indexed by k in the set $\{1, 2, \ldots, M\}$. An important goal of Bayesian model selection is to calculate $p(k|\mathbf{y})$—the posterior model probability for model k. Here the aim may be to obtain a single most probable model, or indeed a subset of likely models, *a posteriori*. Alternatively, posterior model probabilities may be synthesized from all competing models to calculate some quantity of interest that is common to all models, by using model averaging (Hoeting *et al.*, 2001).

We denote by $\boldsymbol{\theta}_k$, the parameters that are specific to model k, where $\boldsymbol{\theta}$ denotes the collection of all model parameters. Specifying prior distributions for within-model parameters $p(\boldsymbol{\theta}_k|k)$, priors for model indicators $p(k)$ and a likelihood for the data, $p(\mathbf{y}|\boldsymbol{\theta}_k, k)$, allows Bayesian inference to proceed by examining the posterior distribution

$$p(\boldsymbol{\theta}_k, k|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta}_k, k)\, p(\boldsymbol{\theta}_k|k)\, p(k). \tag{1}$$

Across-model strategies proceed by sampling from this joint posterior distribution of model indicators and parameters. The reversible jump Markov chain Monte Carlo (RJMCMC) algorithm of Green (1995) is a popular approach for this situation. Other across-model search strategies include those of Godsill (2001) and Carlin and Chib (1995). By contrast, within-model methods examine the posterior distribution within model k separately for each k. Here the within-model posterior appears as

$$p(\boldsymbol{\theta}_k|\mathbf{y}, k) \propto p(\mathbf{y}|\boldsymbol{\theta}_k, k)\, p(\boldsymbol{\theta}_k|k),$$

Address for correspondence: N. Friel, School of Mathematical Sciences, University College Dublin, Belfield, Dublin 4, Republic of Ireland.
E-mail: nial.friel@ucd.ie

 1369–7412/08/70589

where the constant of proportionality, which is often termed the marginal likelihood or integrated likelihood for model k, is written as

$$p(\mathbf{y}|k) = \int_{\boldsymbol{\theta}_k} p(\mathbf{y}|\boldsymbol{\theta}_k, k)\, p(\boldsymbol{\theta}_k|k)\, \mathrm{d}\boldsymbol{\theta}_k. \tag{2}$$

This, in general, is a difficult integral to compute, possibly involving high dimensional model parameters $\boldsymbol{\theta}_k$. However, if we could do so then we would readily be able to make statements about posterior model probabilities, by using Bayes's theorem,

$$p(k|\mathbf{y}) = p(\mathbf{y}|k)\, p(k) \Big/ \sum_{k=1}^{M} p(\mathbf{y}|k)\, p(k).$$

The marginal likelihoods can be used to compare two models by computing Bayes factors,

$$B_{ij} = \frac{p(\mathbf{y}|k=i)}{p(\mathbf{y}|k=j)},$$

without the need to specify prior model indicator probabilities. Note that the Bayes factor B_{ij} gives the evidence that is provided by the data in favour of model i compared with model j. It can also be seen that

$$B_{ij} = \frac{p(k=i|\mathbf{y})}{p(k=j|\mathbf{y})} \frac{p(k=j)}{p(k=i)}.$$

In other words, the Bayes factor is the ratio of posterior odds to prior odds.

An improper prior distribution $p(\boldsymbol{\theta}_k|k)$ leads necessarily to an improper marginal likelihood, which in turns implies that the Bayes factor is not well defined in this case. To circumvent the difficulty of using improper priors for model comparison, O'Hagan (1995) introduced a method that is termed the fractional Bayes factor. Here an approximate (proper) marginal likelihood is defined by the ratio

$$\frac{\displaystyle\int_{\boldsymbol{\theta}_k} p(\mathbf{y}|\boldsymbol{\theta}_k, k)\, p(\boldsymbol{\theta}_k|k)\, \mathrm{d}\boldsymbol{\theta}_k}{\displaystyle\int_{\boldsymbol{\theta}_k} p(\mathbf{y}|\boldsymbol{\theta}_k, k)^a\, p(\boldsymbol{\theta}_k|k)\, \mathrm{d}\boldsymbol{\theta}_k},$$

since any impropriety in the prior for $\boldsymbol{\theta}_k$ cancels above and below. Other approaches to this problem include the intrinsic Bayes factor (Berger and Pericchi, 1996) and expected posterior prior (Perez and Berger, 2002). In this paper we concentrate on the case where prior model distributions are proper.

Various methods have been proposed in the literature to estimate the marginal likelihood (2). For example Chib (1995) estimated the marginal likelihood $p(\mathbf{y}|k)$ by using output from a Gibbs sampler for $(\boldsymbol{\theta}_k|\mathbf{y}, k)$. It relies, however, on a block updating approach for $\boldsymbol{\theta}_k$. Clearly this is not always possible to do. This work was then extended in Chib and Jeliazkov (2001), where output from a Metropolis–Hastings algorithm for the posterior $(\boldsymbol{\theta}_k|\mathbf{y}, k)$ can be used to estimate the marginal likelihood. Annealed importance sampling (Neal, 2001) estimates the marginal likelihood by using ideas from importance sampling. Here an independent sample from the posterior is generated by defining a sequence of distributions that is indexed by a temperature parameter t from the prior through to the posterior. Importantly, however, the collection of importance weights can be used to estimate the marginal likelihood.

In this paper we propose a new method to compute the marginal likelihood that is based on samples from a distribution proportional to the prior multiplied by the likelihood raised

to a power t, which we term the *power posterior*. This method was inspired by ideas from path sampling or thermodynamic integration (Gelman and Meng, 1998). We find that the marginal likelihood $p(\mathbf{y}|k)$ can be expressed as an integral with respect to t from 0 to 1 of the expected deviance for model k, where the expectation is taken with respect to the power posterior distribution, at power t. We argue that this method requires very little tuning or bookkeeping, unlike other methods. It is easy to implement, requiring minor modification to computer code which samples from the posterior distribution. In this paper, in Section 2 we carry out a review of various approaches to across- and within-model marginal likelihood computation. Section 3 introduces the new method for estimating the marginal likelihood, which is based on sampling from the so-called power posterior distribution. Here we also outline how this method could be implemented in practice, while also giving some guidance about sensitivity of the estimate of the marginal likelihood (or Bayes factor) to the diffusivity of the prior model parameters. Three illustrations of how the new method performs in practice are given in Section 4. We conclude this paper with discussion and final remarks in Section 5.

2. Review of methods to compute Bayes factors

Numerous methods and techniques are available to estimate marginal likelihood (2). Generally speaking two approaches are possible—across-model or within-model computations. The former approach, in an MCMC setting, involves generating a single Markov chain which traverses the joint model and parameter space (1). A popular choice is the reversible jump sampler (Green, 1995). Godsill (2001) retained aspects of the reversible jump sampler but considered the case where parameters, for example, are shared between different models, as occurs for example in nested models. Other approaches include those of Stephens (2000), Carlin and Chib (1995) and Dellaportas *et al.* (2001). Within-model computations essentially aim to estimate the marginal likelihood (2) for each model k separately, and then if desired use this information to form Bayes factors (Chib, 1995; Chib and Jeliazkov, 2001). Neal (2001) combined aspects of simulated annealing and importance sampling to provide a method of gathering an independent sample from a posterior distribution of interest, but importantly also to estimate the marginal likelihood. Bridge sampling (Meng and Wong, 1996) offers the possibility of estimating the Bayes factor by linking the two posterior distributions by a bridge function. Bartolucci *et al.* (2006) used this approach, although it is based on an across-model reversible jump sampler.

Within-model approaches are disadvantageous when the cardinality of the model space is large such as variable selection in regression settings. However, as noted in chapter 6 of Green *et al.* (2003), the ideal situation for a within-model approach is one where the models are all reasonably heterogeneous. In effect, this is the case where it is difficult to choose proposal distributions when jumping between models—and indeed the situation where parameters across models of the same dimension have different interpretations. In some scientific situations it is important to obtain weights of evidence for different statistical models which represent different mechanistic processes such as in disease transmission (Forrester *et al.*, 2007). Here it may be required to decide whether new cases are entirely random or due to a specific transmission path. In other situations, as in mixture analysis, the main scientific focus may be on the number of components rather than obtaining a flexible distribution (Ridall *et al.*, 2007).

This short review is not intended to be exhaustive. A more complete picture can be found in Sisson (2005) and Han and Carlin (2001).

2.1. Reversible jump Markov chain Monte Carlo sampling

RJMCMC sampling (Green, 1995) offers the potential to carry out inference for all unknown parameters in the joint model and parameter space (1) in a single logical framework. A crucial innovation in the seminal paper by Green (1995) was to illustrate that detailed balance could be achieved for general state spaces. In particular this extends the Metropolis–Hastings algorithm to variable dimension state spaces of the type $(\boldsymbol{\theta}_k, k)$. To implement the algorithm, proposing to move from $(\boldsymbol{\theta}_k, k)$ to $(\boldsymbol{\theta}_l, l)$ proceeds by generating random numbers $\mathbf{u}$ from a distribution g and setting $(\boldsymbol{\theta}_l, l) = f_{kl}(\boldsymbol{\theta}_k, \mathbf{u}, k)$, for some deterministic function f_{kl}. Similarly to move from $(\boldsymbol{\theta}_l, l)$ to $(\boldsymbol{\theta}_k, k)$ requires random numbers $\mathbf{u}^*$ following some distribution g^*, and setting $(\boldsymbol{\theta}_k, k) = f_{lk}(\boldsymbol{\theta}_l, \mathbf{u}^*, l)$, for some deterministic function f_{lk}. However, it is important that both the transformation f_{kl} from $(\boldsymbol{\theta}_k, k)$ to $(\boldsymbol{\theta}_l, l)$ is a bijection and its differential invertible. A necessary condition for this to apply is if the so-called 'dimension matching' condition applies, i.e. if $\dim(\boldsymbol{\theta}_k) + \dim(\mathbf{u}) = \dim(\boldsymbol{\theta}_l) + \dim(\mathbf{u}^*)$. In this case, the probability of accepting such a move appears as

$$\min\left\{1, \frac{p(\boldsymbol{\theta}_l, l|\mathbf{y})\, p(l \to k)\, g^*(\mathbf{u}^*)}{p(\boldsymbol{\theta}_k, k|\mathbf{y})\, p(k \to l)\, g(\mathbf{u})} |J|\right\}$$

where $p(l \to k)$ is the probability of moving from model l to model k and in addition J is the Jacobian resulting from the transformation from $(\boldsymbol{\theta}_k, \mathbf{u}, k)$ to $(\boldsymbol{\theta}_l, \mathbf{u}^*, l)$. In practice, this may be simplified slightly by not insisting on stochastic moves in both directions, so that, for example, $\dim(\mathbf{u}^*) = 0$, whence the term $g^*(\cdot)$ disappears in the numerator above. Finally, for the case of nested models, a possible move type is $(\boldsymbol{\theta}_{k+1}, k+1) = ((\boldsymbol{\theta}_k, u), k+1)$ in which case the Jacobian term equals 1.

In some respects RJMCMC sampling is difficult to use for practically important problems. The main problem appears to be model mixing across dimensions. Typically, this is as a result of the difficulty in choosing both the mapping f_{lk} and then a suitable 'jump' proposal distribution. For the latter, it is unclear how reasonably to centre and scale the distribution to increase the chance of the move being accepted. However, recent work in Brooks *et al.* (2003) has tackled this problem to some extent.

2.2. Chib's method

An important method of marginal likelihood estimation within each model is that of Chib (1995). This method follows from noticing that, for any parameter configuration $\boldsymbol{\theta}^*$, Bayes's rule implies that the marginal likelihood of the data $\mathbf{y}$ for model k satisfies

$$p(\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta}^*)\, p(\boldsymbol{\theta}^*)}{p(\boldsymbol{\theta}^*|\mathbf{y})}.$$

Here and for the remainder of this paper, for ease of notation, we remove reference to the model indicator k, except where this is ambiguous. Each factor on the right-hand side above can be calculated immediately, with the exception of the posterior probability $p(\boldsymbol{\theta}^*|\mathbf{y})$. Typically $\boldsymbol{\theta}^*$ would be chosen as a point of high posterior probability to increase the numerical accuracy of the estimate. Chib illustrated that this probability can be estimated via Gibbs sampling provided that $\boldsymbol{\theta}^*$ can be partitioned into q non-overlapping blocks $\{\boldsymbol{\theta}_i^*\}$, say, where the full conditional of each block is amenable to Gibbs sampling. It is clear that

$$p(\boldsymbol{\theta}^*|\mathbf{y}) = p(\theta_1^*) \prod_{i=2}^{q} p(\theta_i^*|\theta_{i-1}^*, \ldots, \theta_1^*, \mathbf{y}).$$

Now each factor $p(\theta_j^*|\theta_{j-1}^*,\ldots,\theta_1^*,\mathbf{y})$ can be estimated from the Gibbs output by integrating out parameters $\theta_{j+1}^*,\ldots,\theta_q^*$:

$$p(\theta_j^*|\theta_{j-1}^*,\ldots,\theta_1^*,\mathbf{y}) = \frac{1}{I}\sum_{i=1}^{I} p(\theta_j^*|\theta_q^{(i)},\ldots,\theta_{j+1}^{(i)},\theta_{j-1}^*,\ldots,\theta_1^*,\mathbf{y}), \tag{3}$$

where the index i indicates iterations of the Markov chain at stationarity. Further the normalizing constant of each block must be known exactly for full conditional probabilities to be estimated. Chib and Jeliazkov (2001) extended this methodology to the case where equation (3) can be updated by using Metropolis–Hastings output, employing an identity that was based solely on the Metropolis–Hastings acceptance probabilities, but which does not require the normalizing constant of $p(\boldsymbol{\theta}^*|\mathbf{y})$. However, implementing both methods relies on judicious partitioning of the parameter $\boldsymbol{\theta}^*$, in addition to a considerable amount of bookkeeping. Clearly both methods increase in computational complexity as the dimension of $\boldsymbol{\theta}$ increases.

2.3. *Annealed importance sampling*

Estimating the marginal likelihood by using ideas from importance sampling is also possible as illustrated in Neal (2001). The idea is to define a sequence of distributions, starting from a sequence for which it is possible to generate perfect samples, e.g. the prior distribution, and ending at a target distribution. Neal (2001) defined a possible choice of this sequence geometrically as

$$p_{t_i}(\boldsymbol{\theta}|\mathbf{y}) = p(\boldsymbol{\theta})^{1-t_i}\, p(\boldsymbol{\theta}|\mathbf{y})^{t_i},$$

where $0 = t_0 < t_1 < \ldots < t_n = 1$. Thus p_{t_0} and p_{t_n} correspond to the prior and posterior distribution respectively. At iteration j, the algorithm begins by sampling $\boldsymbol{\theta}_{t_0}^{(j)}$ from the prior p_{t_0}. At the ith step of iteration j, $\boldsymbol{\theta}_{t_{i+1}}^{(j)}$ is generated from $p_{t_{i+1}}$ via a Markov chain transition kernel at $\boldsymbol{\theta}_{t_i}^{(j)}$, e.g. via Gibbs or Metropolis–Hastings updating. The final step n of iteration j yields $\boldsymbol{\theta}_{t_n}^{(j)}$ from the posterior. After R iterations, this scheme yields an independent sample $\boldsymbol{\theta}_{t_n}^{(1)},\ldots,\boldsymbol{\theta}_{t_n}^{(R)}$ from the posterior. In effect, distribution p_{t_i} is an importance distribution for $p_{t_{i+1}}$. An important by-product of this scheme is that the collection of R importance weights

$$w^{(j)} = \frac{p_{t_1}(\boldsymbol{\theta}_{t_0})}{p_0(\boldsymbol{\theta}_{t_0})}\frac{p_{t_2}(\boldsymbol{\theta}_{t_1})}{p_{t_1}(\boldsymbol{\theta}_{t_1})}\cdots\frac{p_{t_n}(\boldsymbol{\theta}_{t_n})}{p_{t_{n-1}}(\boldsymbol{\theta}_{t_n})}$$

(where the superscript j has been omitted from $\boldsymbol{\theta}_t^{(j)}$ for clarity) is such that

$$p(\mathbf{y}) = \sum_{j=1}^{R} w^{(j)} \Big/ R,$$

i.e. the marginal likelihood is obtained as the average of the importance weights.

3. Marginal likelihoods and power posteriors

Here we introduce a new approach to estimating the integrated likelihood that is based on ideas of thermodynamic integration or path sampling (Gelman and Meng, 1998). Consider introducing an auxiliary variable (or temperature parameter) $t \in [0, 1]$. Consider the *power posterior*, which is defined as

$$p_t(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta})^t\, p(\boldsymbol{\theta}). \tag{4}$$

Now, define

$$z(\mathbf{y}|t) = \int_{\theta} p(\mathbf{y}|\boldsymbol{\theta})^t \, p(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}.$$

By construction, $z(\mathbf{y}|t=0)$ is the integral of the prior for $\boldsymbol{\theta}$, which equals 1. Further, $z(\mathbf{y}|t=1)$ is the marginal likelihood of the data. Here we assume of course that $z(\mathbf{y}|t) < \infty$ for all $t \in [0, 1]$.

Now ideas from path sampling (Gelman and Meng, 1998) can be used to calculate the integral of interest $z(\mathbf{y}|t=1)$. The following identity is crucial to the problem in hand:

$$\log\{p(\mathbf{y})\} = \log\left\{\frac{z(\mathbf{y}|t=1)}{z(\mathbf{y}|t=0)}\right\} = \int_0^1 \mathbf{E}_{\boldsymbol{\theta}|\mathbf{y},t}[\log\{p(\mathbf{y}|\boldsymbol{\theta})\}] \, \mathrm{d}t. \tag{5}$$

Thus the marginal likelihood results as the integral over t of half of the mean deviance, where the expectation is taken with respect to the power posterior (4) at temperature t, where t moves from 0 to 1. The identity (5) can be derived easily as follows:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\log\{z(\mathbf{y}|t)\} &= \frac{1}{z(\mathbf{y}|t)}\frac{\mathrm{d}}{\mathrm{d}t}z(\mathbf{y}|t) \\
&= \frac{1}{z(\mathbf{y}|t)}\frac{\mathrm{d}}{\mathrm{d}t}\int_{\theta} p(\mathbf{y}|\boldsymbol{\theta})^t \, p(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} \\
&= \frac{1}{z(\mathbf{y}|t)}\int_{\theta} p(\mathbf{y}|\boldsymbol{\theta})^t \log\{p(\mathbf{y}|\boldsymbol{\theta})\} \, p(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} \\
&= \int_{\theta} \frac{p(\mathbf{y}|\boldsymbol{\theta})^t \, p(\boldsymbol{\theta})}{z(\mathbf{y}|t)} \log\{p(\mathbf{y}|\boldsymbol{\theta})\} \, \mathrm{d}\boldsymbol{\theta} \\
&= \mathbf{E}_{\boldsymbol{\theta}|\mathbf{y},t}[\log\{p(\mathbf{y}|\boldsymbol{\theta})\}].
\end{aligned}$$

Equation (5) now follows by integrating with respect to t. This approach shares some analogies with annealed importance sampling that was outlined in Section 2.3. However, here we estimate the marginal likelihood on the log-scale, ensuring increased numerical stability. Further our method estimates $\log\{p(\mathbf{y})\}$ by using expectations, again aiding the numerical stability.

It is interesting that the fraction $z(\mathbf{y}|t=1)/z(\mathbf{y}|t=a)$, where $0 < a < 1$, is precisely the approximation to the marginal likelihood that is used to compute the fractional Bayes factor (O'Hagan, 1995). In addition note that the 'likelihood' contribution, $p(\mathbf{y}|\boldsymbol{\theta})^t$, to the power posterior is generally not a proper likelihood since it may not always hold that $\int p(\mathbf{y}|\boldsymbol{\theta})^t \, \mathrm{d}\mathbf{y} = 1$. Finally, in common with simulated annealing and simulated tempering, the effect of the temperature parameter t is to flatten the likelihood contribution in the power posterior, so that it is approximately uniform for values of t that are close to 0, in which case the power posterior approximates the prior contribution.

Note that path sampling has been employed to compute high dimensional normalizing constants, most notably in estimation of parameters of Markov random fields (MRFs). In this context the technique has been used to calculate normalizing constants of model parameters, which are then used as a look-up table in the estimation process; see for example Green and Richardson (2002) and Dryden *et al.* (2003).

3.1. Sensitivity of $p(\mathbf{y}|k)$ to the prior

It is well understood that the Bayes factor is sensitive to the choice of prior model parameters. Here we outline for a simple example how this impacts on values of the marginal likelihood using identity (5).

Consider the simple situation where data $\mathbf{y} = \{y_i : i = 1, \ldots, N\}$ are independent and normally distributed with mean θ and unit variance. Assuming that $\theta \sim N(m, v)$, *a priori*, leads to a power

posterior, $\theta|\mathbf{y}, t \sim N(m_t, v_t)$, where

$$m_t = \frac{Nt\bar{y} + m/v}{Nt + 1/v},$$

$$v_t = \frac{1}{Nt + 1/v}.$$

It is straightforward to show that

$$\mathbf{E}_{\theta|\mathbf{y},t}[\log\{p(\mathbf{y}|\theta)\}] = -\frac{N}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{N}(y_i - \bar{y})^2 - \frac{N}{2}\frac{(m-\bar{y})^2}{(vmt+1)^2} - \frac{N}{2}\frac{1}{(Nt+1/v)}. \quad (6)$$

Recall that the logarithm of the marginal likelihood is obtained by integrating equation (6) with respect to t over $t \in [0, 1]$. Consider the situation when $t = 0$. In this case the final term on the left-hand side of equation (6) appears as $-Nv/2$. Clearly as $v \to \infty$, so also does $\mathbf{E}_{\theta|\mathbf{y},t=0}[\log\{p(\mathbf{y}|\theta)\}]$, and at the same speed. This illustrates the sensitivity of the marginal likelihood as defined in terms of the power posterior to the prior specification. Fig. 1 plots $\mathbf{E}_{\theta|\mathbf{y},t}[\log\{p(\mathbf{y}|\theta)\}]$ against t for prior variance $v = 1, 5, 10$ (for illustrative purposes the first two terms on the left-hand side take a constant value -1, $\bar{y} = 0$, $m = 0$ and $N = 10$). It is our experience that the behaviour of the mean deviance under each power posterior at temperature t, as illustrated in Fig. 1, is typical in more complex settings. This in turns impacts on how we estimate the marginal likelihood using identity (5), for which the next section offers some alternatives. Note in Fig. 1 that the case of $v = 1$ corresponds to the default unit information prior approach of Kass and Wasserman (1995).

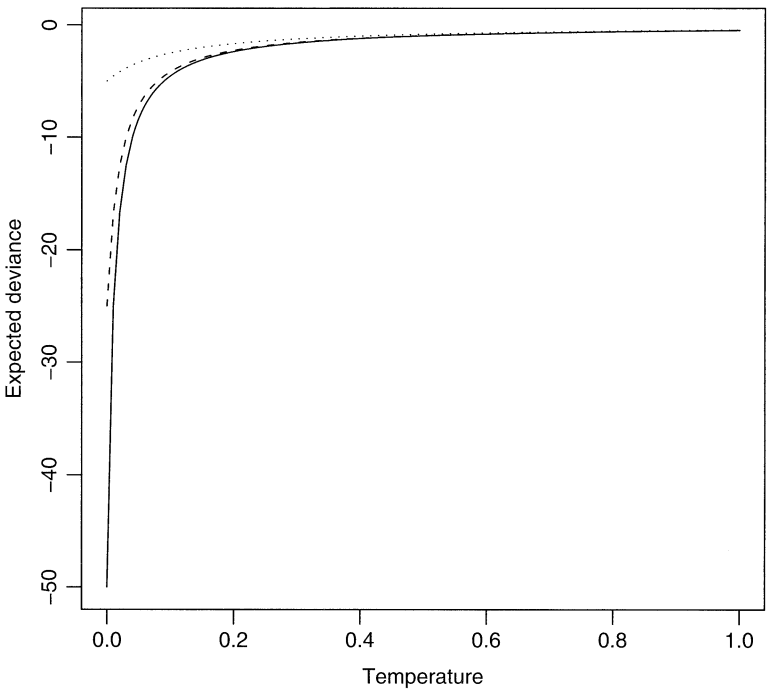

Fig. 1. Expected (half) deviance (6), under the distribution ($\theta|\mathbf{y}$, t), plotted against t for prior variance equal to 1 (·····), 5 (-------) and 10 (———): as v increases so also does the rate at which the mean deviance changes with t

3.2. Estimating the marginal likelihood

3.2.1. A single chain over $(\boldsymbol{\theta}, t)$

A natural idea in the Bayesian framework is to consider the temperature parameter t as a random variable with prior distribution $p(t)$. In this case identity (5) can be written as

$$\log\{p(\mathbf{y})\} = \mathbf{E}_{\boldsymbol{\theta},t|\mathbf{y}}\left[\frac{\log\{p(\mathbf{y}|\boldsymbol{\theta})\}}{p(t)}\right],$$

where the expectation is with respect to the joint distribution

$$p(\boldsymbol{\theta}, t|\mathbf{y}) = p(\boldsymbol{\theta}|t, \mathbf{y})\, p(t) = \frac{p(\mathbf{y}|\boldsymbol{\theta})^t\, p(\boldsymbol{\theta})}{z(\mathbf{y}|t)} p(t).$$

Now the full conditional distribution of $\boldsymbol{\theta}$ looks like

$$p(\boldsymbol{\theta}|\mathbf{y}, t) \propto p(\mathbf{y}|\boldsymbol{\theta})^t\, p(\boldsymbol{\theta})$$

whereas, if we assume that $p(t) \propto z(\mathbf{y}|t)$, then the full conditional distribution of t satisfies

$$p(t|\boldsymbol{\theta}, \mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta})^t.$$

Now a sample $\{(\boldsymbol{\theta}^{(1)}, t_1), \ldots, (\boldsymbol{\theta}^{(N)}, t_N)\}$ that is gathered from $p(\boldsymbol{\theta}, t|\mathbf{y})$ can be used to calculate identity (5) by ordering the t_is and calculating $\log\{p(\mathbf{y}|\boldsymbol{\theta})\}$, estimating the integral via quadrature. All of this hinges on the assumption that $p(t) \propto z(\mathbf{y}|t)$. It is our experience that $z(\mathbf{y}|t)$ varies by orders of magnitude with t. This is not surprising since by construction $z(\mathbf{y}|t=0) = 1$, whereas the marginal likelihood $z(\mathbf{y}|t=1)$ could be very large depending on the problem in hand. Thus, using a single chain, values of t that are close to 0 would tend not to be sampled with high frequency, leading to poor estimation of $p(\mathbf{y})$, rendering this approach unworkable.

3.2.2. A serial Markov chain Monte Carlo approach: discretizing $t \in [0, 1]$

As a more direct approach we suggest discretizing the integral (5) over $t \in [0, 1]$, running separate chains for each t, sampling from the power posterior to estimate the half mean deviance, $\mathbf{E}_{\boldsymbol{\theta}|\mathbf{y},t}[\log\{p(\boldsymbol{\theta}|\mathbf{y})\}]$. Numerical integration using, for example, a trapezoidal rule over t yields an estimate of the marginal likelihood. For example choosing a discretization $0 = t_0 < t_1 < \ldots < t_{n-1} < t_n = 1$ leads to an approximation

$$\log\{p(\mathbf{y})\} \approx \sum_{i=0}^{n-1} (t_{i+1} - t_i) \frac{\mathbf{E}_{\boldsymbol{\theta}|\mathbf{y},t_{i+1}}[\log\{p(\mathbf{y}|\boldsymbol{\theta})\}] + \mathbf{E}_{\boldsymbol{\theta}|\mathbf{y},t_i}[\log\{p(\mathbf{y}|\boldsymbol{\theta})\}]}{2}. \tag{7}$$

Note that the Monte Carlo standard error, s_i say, for each $\mathbf{E}_{\boldsymbol{\theta}|\mathbf{y},t_i}[\log\{p(\mathbf{y}|\boldsymbol{\theta})\}]$, can be pieced together to give an overall Monte Carlo standard error for $\log\{p(\mathbf{y})\}$ given as

$$\sqrt{\left\{\frac{(t_2 - t_1)^2}{2} s_1^2 + \sum_{i=2}^{n-1} \frac{(t_{i+1} - t_{i-1})^2}{2} s_i^2 + \frac{(t_n - t_{n-1})^2}{2} s_n^2\right\}}. \tag{8}$$

Bearing in mind the discussion in Section 3.1, we see that the choice of spacing for the t_is in approximation (7) is important. For example, a temperature schedule of the type $t_i = x_i^c$, where $x_i = i/n$ is an equal spacing of the n points in the interval [0, 1], and $c > 1$ is a constant, ensures that the t_is are chosen with high frequency close to $t = 0$. Prescribing the collection of t_is in this way should improve the efficiency of the estimate of $p(\mathbf{y})$. We offer some guidance on the choice of the discretization, n and the temperature parameter c in Section 4.1. Convergence of the collection of Markov chains is clearly crucial and we suggest the following strategy where

each chain is sampled in succession from $t_0 = 0, \ldots, t_n = 1$. Beginning at temperature $t_0 = 0$, immediate convergence to the power posterior is guaranteed if for example it is possible to initialize parameters to their prior mean values, or if perfect sampling from the prior is possible. Otherwise usual convergence diagnostics could be employed to estimate convergence to the prior distribution. However, the posterior mean estimate of $p(\boldsymbol{\theta}|\mathbf{y}, t)$ should give a reasonable initial value for the Markov chain with stationary distribution $p(\boldsymbol{\theta}|\mathbf{y}, t+1)$, effectively ensuring that little burn-in is needed. Here we make the assumption that the power posterior $p(\boldsymbol{\theta}, |\mathbf{y}, t-1)$ approximates $p(\boldsymbol{\theta}|\mathbf{y}, t)$. Effectively, this strategy should efficiently feed forward information from the current chain into the subsequent chain. We outline the algorithm below, where $\boldsymbol{\theta}_i^{(j)}$ denotes iteration j of parameters $\boldsymbol{\theta}$ from an MCMC sampler with stationary distribution $p(\boldsymbol{\theta}|\mathbf{y}, t_i)$.

Initialize $\boldsymbol{\theta}_0^{(0)}$.
For $i = 0, \ldots, n$:
  set the temperature parameter $t_i = (i/n)^c$;
  generate a sample $\{\boldsymbol{\theta}_i^{(j)}\}_{j=K+1}^{R}$ via MCMC sampling from $p(\boldsymbol{\theta}|\mathbf{y}, t_i)$;
  estimate

$$\mathbf{E}_{\boldsymbol{\theta}|\mathbf{y},t_i}[\log\{p(\mathbf{y}|\boldsymbol{\theta})\}] \approx \frac{1}{R-K} \sum_{j=K+1}^{R} p(\mathbf{y}|\boldsymbol{\theta}_i^{(j)});$$

  while $i < n$, initialize the next chain to

$$\boldsymbol{\theta}_{i+1}^{(0)} = \frac{1}{R-K} \sum_{j=K+1}^{R} \boldsymbol{\theta}_i^{(j)}.$$

Compute $\log\{p(\mathbf{y})\}$ via numerical integration, using the estimates of $\mathbf{E}_{\boldsymbol{\theta}|\mathbf{y},t_i}[\log\{p(\mathbf{y}|\boldsymbol{\theta})\}]$.

It is possible to modify this algorithm in several directions. For example, for models with diffuse priors, we suggest that chains for small values of t should be run for more iterations, since in this case, as t increases, the power posterior will typically become more concentrated on the actual posterior.

We remark that, if the likelihood follows an exponential family model, then raising the likelihood to a power t amounts to multiplying the exponent by t. We therefore expect that in many cases, if the posterior is amenable to Gibbs sampling, then so also would the power posterior. In terms of computation, as our algorithm shows, modifying existing MCMC code which samples from posterior model parameters is trivial. Essentially all that is needed is an extra iteration loop for the temperature parameter t, calculating the expected deviance under the power posterior at each temperature iteration.

3.2.3. A population Markov chain Monte Carlo approach

In contrast with Section 3.2.2, an alternative method to estimate identity (5) is to define a single Markov chain whose states are the collection of parameters $\{\boldsymbol{\theta}_{t_i} : i = 1, \ldots, n\}$ corresponding to each power posterior distribution and whose stationary distribution is $\Pi_{i=1}^{n} p(\cdot|\mathbf{y}, t_i)$. In effect we are tackling this problem by using a population MCMC approach, following for example Liang and Wong (2001).

Population MCMC is used in situations where the target distribution, $p(\mathbf{x})$ say, is difficult to sample from, for example, if it is multimodal, resulting in a slow mixing chain. The population MCMC solution is to consider an augmented target distribution consisting of a product of tempered versions of the target, $\Pi_{i=1}^{n} p(\mathbf{x})^{t_i}$, where, as in our situation, the t_is are an increasing sequence of temperatures with $t_n = 1$. The population MCMC algorithm proceeds by updat-

ing parameter values within each distribution $p(\mathbf{x})^{t_i}$ via usual MCMC methods. But, crucially, information is also shared between chains, for example, by proposing to swap parameter values at different temperatures. The intuition is that chains at lower temperatures may facilitate rapid movement through the parameter space. One criticism of this approach is that only that trace of the chain at temperature $t_n = 1$ is used, and hence the information at every other chain is discarded. However, in our situation the opposite is the case. Information at each chain is used to compute the marginal likelihood; moreover, information on the actual posterior $p(\boldsymbol{\theta}|\mathbf{y})$ is also available, if needed.

We propose the following population MCMC algorithm with two move types. The first is simply a sweep updating parameters at each temperature. The second move type sweeps through each chain by using a Metropolis–Hastings sampler, proposing to swap parameter values at the current chain with parameters at a neighbouring chain. Suppose that the chain is visiting power posterior with temperature t_i; then a neighbouring chain is proposed from a discrete Laplacian distribution defined as

$$p_i(j) \propto \exp(\beta \| i - j \|),$$

where $j \in \{1, \ldots, n\} \backslash i$. Setting $\beta = 0.5$ ensures that $j = i \pm 1$ is roughly three times more likely than a jump to $i \pm 3$, for example. Then we propose to swap $\boldsymbol{\theta}_{t_i}$ with $\boldsymbol{\theta}_{t_j}$. A typical iteration of a population MCMC sampler would update the current population, choosing one of the two move types that are described algorithmically below.

Step 1 (within-chain move)—for $i = 0, \ldots, n$ update $\boldsymbol{\theta}_{t_i}$ via an MCMC update from $p(\boldsymbol{\theta}|\mathbf{y}, t_i)$.
Step 2 (between-chain move)—for $i = 0, \ldots, n$:
sample j from $p_i(j) \propto \exp(\beta \| i - j \|)$, for $j = \{0, \ldots, n\} \backslash i$;
propose to move from $\{\boldsymbol{\theta}_{t_0}, \ldots, \boldsymbol{\theta}_{t_i}, \ldots, \boldsymbol{\theta}_{t_j}, \ldots, \boldsymbol{\theta}_{t_n}\}$ to $\{\boldsymbol{\theta}_{t_0}, \ldots, \boldsymbol{\theta}_{t_j}, \ldots, \boldsymbol{\theta}_{t_i}, \ldots, \boldsymbol{\theta}_{t_n}\}$ with probability

$$\min\left\{1, \frac{p(\boldsymbol{\theta}_{t_j}|\mathbf{y}, t_i)\, p(\boldsymbol{\theta}_{t_i}|\mathbf{y}, t_j)\, p_j(i)}{p(\boldsymbol{\theta}_{t_i}|\mathbf{y}, t_i)\, p(\boldsymbol{\theta}_{t_j}|\mathbf{y}, t_j)\, p_i(j)}\right\}.$$

Of course, more sophisticated move types are also possible to allow exchange of information between chains. See for example chapter 11 of Liu (2001).

3.3. Kullback–Leibler distances and power posteriors

Calculating marginal likelihoods via the power posterior method routinely allows estimation of the Kullback–Leibler distance between posterior and prior, $\mathrm{KL}\{p(\boldsymbol{\theta}|\mathbf{y}), p(\boldsymbol{\theta})\}$ say. We can formulate this distance as follows:

$$\begin{aligned}
\mathrm{KL}\{p(\boldsymbol{\theta}|\mathbf{y}), p(\boldsymbol{\theta})\} &= \int \log\left\{\frac{p(\boldsymbol{\theta}|\mathbf{y})}{p(\boldsymbol{\theta})}\right\} p(\boldsymbol{\theta}|\mathbf{y})\, \mathrm{d}\boldsymbol{\theta} \\
&= \int \log\left\{\frac{p(\mathbf{y}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})}{p(\boldsymbol{\theta})\, p(\mathbf{y})}\right\} p(\boldsymbol{\theta}|\mathbf{y})\, \mathrm{d}\boldsymbol{\theta} \\
&= \int \log\{p(\mathbf{y}|\boldsymbol{\theta})\}\, p(\boldsymbol{\theta}|\mathbf{y})\, \mathrm{d}\boldsymbol{\theta} - \int \log\{p(\mathbf{y})\}\, p(\boldsymbol{\theta}|\mathbf{y})\, \mathrm{d}\boldsymbol{\theta} \\
&= \mathbf{E}_{\boldsymbol{\theta}|\mathbf{y}}[\log\{p(\mathbf{y}|\boldsymbol{\theta})\}] - \log\{p(\mathbf{y})\}.
\end{aligned}$$

Each term on the right-hand side is available after using the power posterior method, although the posterior mean of the log-likelihood does not require the power posterior method. This

distance gives a measure of information gain in updating prior beliefs to posterior beliefs. It can be interpreted as a measure of mismatch between prior and posterior. In this sense it can be used to measure how diffuse a prior is. For example, if the Kullback–Leibler distance between prior and posterior is computed for a collection of priors with, for instance, increasing prior variances, then plotting the Kullback–Leibler distance against the variance would give an objective measure of the divergence between prior and posterior distributions.

4. Examples

4.1. Linear regression—non-nested models

The data set for this example was taken from Williams (1959). The data describe the maximum compression strength parallel to the grain y_i, the density x_i and the resin-adjusted density z_i for 42 specimens of *radiata* pine. This data set appears in Han and Carlin (2001), and was also examined in Carlin and Chib (1995) and Bartolucci *et al.* (2006), where several methods were compared to estimate the Bayes factor between two non-nested competing models.

The competing models are as follows:

(a) model $k=1$, $y_i=\alpha+\beta(x_i-\bar{x})+\varepsilon_i, \varepsilon_i \sim N(0,\sigma^2)$;
(b) model $k=2$, $y_i=\gamma+\delta(z_i-\bar{z})+\eta_i, \eta_i \sim N(0,\tau^2)$.

The following prior specification was used (which is identical to those in Han and Carlin (2001), Carlin and Chib (1995) and Bartolucci *et al.* (2006)): $N\{(3000,185)^{\mathrm{T}}, \mathrm{diag}(10^6,10^4)\}$ for both $(\alpha,\beta)^{\mathrm{T}}$ and $(\gamma,\delta)^{\mathrm{T}}$. An $\mathrm{IG}\{3,(2\times 300^2)^{-1}\}$ prior was chosen for both σ^2 and τ^2, where $\mathrm{IG}(a,b)$ is an inverse gamma distribution with density

$$f(x)=\frac{1}{\exp(1/bx)\,\Gamma(a)b^a x^{a+1}}.$$

Green and O'Hagan (1998) found, for the given prior specification, by numerical integration, that the Bayes factor $B_{21}=4862$. We aim to use this example to give us some empirical evidence to guide the choice of parameters of the power posterior algorithms, namely n, the amount of discretization in the interval $[0,1]$, and c, the temperature exponent. But, further, we aim to see, for this straightforward situation, what statistical efficiency can be achieved by using the power posterior method to compute the Bayes factor compared with using RJMCMC sampling.

4.1.1. Choice of c and n

In this study we examined the choice of temperature parameter c and amount of discretization n in the interval $[0,1]$. For various combinations of c and n we estimated the Bayes factor B_{21} by both the serial MCMC methods and the population MCMC methods that were described in Sections 3.2.2 and 3.2.3 respectively. For each combination of parameters we ran both algorithms 100 times, using 100 000 iterations, of which 30 000 were discarded as burn-in. In Tables 1 and 2 we display bias and standard errors of estimates of the Bayes factor B_{21}.

For the serial MCMC method with the prescribed data and model, reasonable estimates of B_{21} result when the temperature parameter $c=3$ or $c=5$ and n lies between 20 and 100. In effect this suggests that the estimate of B_{21} is not too sensitive to the amount of discretization n and to the choice of temperature c. In general the population MCMC approach perhaps does not perform as well as the serial MCMC approach. For this approach, again parameter values for $c=3$ or $c=5$ and n between 20 and 100 yielded the best results.

Table 1. Bias and (standard errors) of estimates of B_{21} by using the serial MCMC approach

c	*Results for the following values of n:*					
	10	*20*	*40*	*60*	*80*	*100*
2	4 (255)	8 (188)	0 (156)	−19 (167)	−13 (175)	−18 (165)
3	3 (186)	11 (134)	10 (132)	−4 (144)	2 (137)	−12 (136)
5	−23 (166)	−18 (181)	6 (152)	21 (140)	2 (157)	−11 (145)
7	59 (193)	6 (172)	21 (153)	−4 (162)	3 (169)	−17 (182)
10	54 (243)	−15 (200)	51 (179)	−17 (157)	1 (213)	−18 (183)

Table 2. Bias and (standard errors) of estimates of B_{21} by using the population MCMC approach

c	*Results for the following values of n:*					
	10	*20*	*40*	*60*	*80*	*100*
2	−103 (290)	18 (179)	0 (197)	−13 (187)	8 (179)	−7 (215)
3	−15 (198)	8 (182)	22 (154)	−36 (154)	−9 (182)	12 (197)
5	−2 (185)	2 (196)	−60 (174)	−28 (187)	15 (209)	−28 (205)
7	−30 (216)	−4 (180)	1 (210)	−32 (238)	−2 (214)	−1 (212)
10	25 (281)	−32 (220)	−20 (231)	7 (243)	−32 (280)	−53 (254)

4.1.2. Power posterior methods and reversible jump Markov chain Monte Carlo sampling

Here we compare the results of estimates by using the power posterior methods with those of RJMCMC sampling. To implement RJMCMC sampling we specified $p(k=1)=p(k=2)=0.5$. The reversible jump sampler was run 100 times for 100000 iterations each, with the first 30000 iterations removed as burn-in iterations. Within-model parameters were updated via Gibbs sampling from the full conditional distribution of each parameter. Across-model moves were proposed by simply setting $(\alpha, \beta, \sigma) = (\gamma, \delta, \tau)$, resulting in the Jacobian term taking the value 1. The frequency with which the sampler visits each model provides posterior model probabilities. These posterior model probabilities can be combined with prior model probabilities to estimate the Bayes factor as the ratio of posterior odds to prior odds. However, specifying equally weighted models, *a priori*, resulted in poor estimation of B_{21}. This is simply because the reversible jump sampler does not mix well and so does not visit model 1 very often, leading to a poor posterior estimate of $p(k=1|\mathbf{y})$. Running the reversible jump sampler with the prior model probability strongly weighted towards model 1 with $p(k=1)=0.9995$ and $p(k=2)=0.0005$ (method RJ corrected in Table 3) leads to estimates of B_{21} with similar efficiency to that of the power posterior methods that were presented in Section 4.1.1. Note also that the acceptance probability for moves between models in the RJMCMC algorithm with reweighted prior model probabilities was around 15%. Table 3 gives biases, standard errors and relative errors for estimates of B_{21} by using both power posterior methods when $c=3$ and $n=40$.

Table 3. Linear regression models: estimates of bias, standard errors and relative errors of B_{21} by using the method of power posteriors and RJMCMC sampling†

Method	*Bias*	*Standard error*
RJMCMC	67	2678
RJ corrected	9	124
Serial MCMC	10	132
Population MCMC	22	154

†'RJMCMC' entries correspond to prior model probabilities $p(k=1)=p(k=2)=0.5$, whereas 'RJ corrected' corresponds to $p(k=1)=0.9995$ and $p(k=2)=0.0005$.

4.2. Regression in survival analysis

McGilchrist and Aisbett (1991) analysed the time to first and second occurrence of infection for 38 kidney patients. The same data set is also analysed in examples, volume 1, in WinBUGS (Spiegelhalter *et al.*, 2003). We follow the analysis in WinBUGS and assume a parametric Weibull distribution for the survivor function with density

$$f(t_{ij}|r,\mu_{ij}) = r\mu_{ij}t_{ij}^{r-1}\exp(-\mu_{ij}t_{ij}^{r}) \tag{9}$$

where t_{ij} denotes the time of the jth occurrence of infection for patient i. In addition there are three covariates, namely age, sex and a diseases variable taking one of four levels. In this example we consider two models of interest:

(a) model $k=1$, $\log(\mu_{ij}) = X_i^{\mathrm{T}}\beta_j$;
(b) model $k=2$, $\log(\mu_{ij}) = X_i^{\mathrm{T}}\beta_j + \alpha_i$.

Model 1 is a fixed effects model, where the regression effect β_j, for the jth occurrence of kidney infection, remains constant for each individual i. A random-effects term α_i is, however, included in model 2, accounting for variability between individuals.

The shape parameter of the survival distribution, r, was given a uniform prior distribution, uniform(0.1, 10), whereas the regression coefficients (including a constant term) that are common to both models were given unit information independent normal priors (Kass and Wasserman, 1995). Finally, the random-effects term α_i was given an $N(0,\sigma^2)$ prior, where the hyperparameter σ was assigned a uniform(0.1, 10) prior.

Here samples from the power posteriors were collected by using the WinBUGS software (Spiegelhalter *et al.*, 2003). To implement approximation (7) we chose a temperature schedule $t_i = x_i^4$, where the x_is are 40 equally spaced points in the interval [0, 1]. Within each temperature t_i, 10 000 samples were collected from the stationary distribution of which the first 4000 were discarded. Fig. 2 displays a plot of the expected deviance under each power posterior for both models.

Applying the trapezoidal rule (7) yields $\log\{p(\mathbf{y}|k=1)\} = -347.49$ and $\log\{p(\mathbf{y}|k=2)\} = -348.26$, with associated Monte Carlo standard errors of 0.11 and 0.38 for the trapezoidal rule respectively. This yields a Bayes factor $B_{12} = 2.16$ and leads to the moderately strong conclusion that a fixed effects model is more probable. However, if we take into account the Monte Carlo sampling variability of the estimate of the Bayes factor there is no significant evidence that the Bayes factor is different from 1. Obviously, substantially more Monte Carlo simulations are

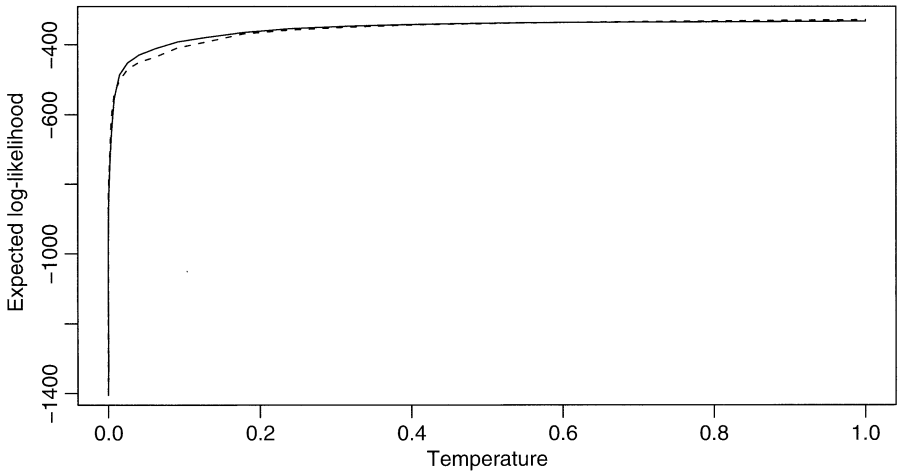

Fig. 2. Expected deviance against temperature for the non-random-effects model (———) and the random-effects model (-------)

required to reduce the standard error. The deviance information criterion values (Spiegelhalter *et al.*, 2002) for the fixed and random-effects models are 674.6 and 672.3 respectively, giving more weight to the random-effects model. The Monte Carlo sampling variability was not provided.

4.3. Hidden Markov random-field models

MRFs are often used to model binary spatially dependent data—the auto-logistic model (Besag, 1974) is a popular choice. Here the joint distribution of $\mathbf{x} = \{x_i : i = 1, 2, \ldots, N\}$ taking values $\{(-1, 1)\}$ on a regular lattice is defined as

$$p(\mathbf{x}|\beta) \propto \exp\left(\beta_0 \sum_i x_i + \beta_1 \sum_{i \sim j} x_i x_j\right) \tag{10}$$

conditional on parameters $\beta = (\beta_0, \beta_1)$. Positive values of β_0 encourage x_i to take the values 1, whereas positive values of β_1 encourage homogeneous regions of 1s or -1s. The notation $i \sim j$ denotes that x_i and x_j are neighbours. For this example we examine two models, which are defined via their neighbourhood structure:

(a) model $k = 1$, a first-order neighbourhood where each point x_i has as neighbours the four nearest adjacent points;
(b) model $k = 2$, a second-order neighbourhood structure where, in addition to the first-order neighbours, the four nearest diagonal points also belong to the neighbourhood.

Both neighbourhood structures are modified along the edges of the lattice. MRF models are difficult to handle in practice, owing to the computational burden of calculating the proportional constant, in expression (10), $c(\beta)$ say.

A hidden MRF $\mathbf{y}$ arises when an MRF $\mathbf{x}$ is corrupted by some noise process. The underlying MRF is essentially hidden and appears as parameters in the model. Typically it is assumed that conditional on $\mathbf{x}$ the y_is are independent, which gives the likelihood

$$p(\mathbf{y}|\mathbf{x}, \mu) = \prod_{i=1}^{N} p(y_i|x_i, \mu),$$

for some parameters μ.

Once prior distributions $p(\beta)$ and $p(\mu)$ have been specified for β and μ respectively, a complete Bayesian analysis proceeds by making inference on the posterior distribution

$$p(\mathbf{x}, \beta, \mu|\mathbf{y}, k) \propto p(\mathbf{y}|\mathbf{x}, \mu)\, p(\mathbf{x}|\beta, k)\, p(\beta)\, p(\mu).$$

It is relatively straightforward to sample from the full conditional distribution of each of $\mathbf{x}$ and μ. Sampling from the full conditional distribution of β is more problematic, owing to the difficulty of calculating the normalizing constant of the MRF, $c(\beta)$. However, provided that the minimum of the number of rows and the number of columns is not greater than 20 and that the larger of the number of rows and the number of columns does not exceed 50, then the forward recursion method that was presented in Reeves and Pettitt (2004) can be used to calculate $c(\beta)$. For a more complete description of the problem of Bayesian estimation of hidden MRFs, the reader is referred to Friel *et al.* (2005).

For this example, gene expression levels were measured for 34 genes in a cluster of 38 neighbouring genes at geographical neighbouring locations on the *Streptomyces coelcicolor* genome for 10 time points. The cluster of 38 neighbouring genes under study is responsible for the production of calcium-dependent antibodies. We define the observations on a 38×10 regular lattice, where log-expression-level y_{tg} corresponds to the gth gene at time point t. Fig. 3 displays the data $\mathbf{y}$, indicating gene locations for which there are no data. Here we assume that the data $\mathbf{y}$ mask an MRF process $\mathbf{x}$, where states $(-1, 1)$ correspond to 'up-regulation' and 'down-regulation' respectively. We assume that the MRF process follows a first-order neighbourhood structure ($k = 1$), or a second-order neighbourhood structure ($k = 2$). Finally we assume that the distribution of $\mathbf{y}$ given $\mathbf{x}$ is modelled as independent Gaussian noise with state-specific mean $\mu(x_i)$, and a known common variance σ^2. Wit and McClure (2004) showed that normality of log-expression-levels is a reasonable assumption for similar experimental set-ups.

It is straightforward to handle the missing data—in the full conditional distribution of the latent process $\mathbf{x}$, the likelihood function needs to be modified slightly to allow for the fact that four of the columns of $\mathbf{x}$ are not supported by any data.

An uniformative proper normal prior was chosen for each of the β-parameters. The prior distribution for μ was distributed uniformly from the set $\{(\mu(-1), \mu(+1)) : -2 \leqslant \mu(-1) \leqslant 2, \mu(-1) \leqslant \mu(+1) \leqslant 2\}$. The values -2 and 2 represent approximate minimum and maximum values which are found in similar data sets.

Here we chose a temperature schedule $t_i = x_i^4$, where the x_is are 40 equally spaced points in the interval $[0, 1]$. Within each temperature t_i, 10000 samples were collected from the station-

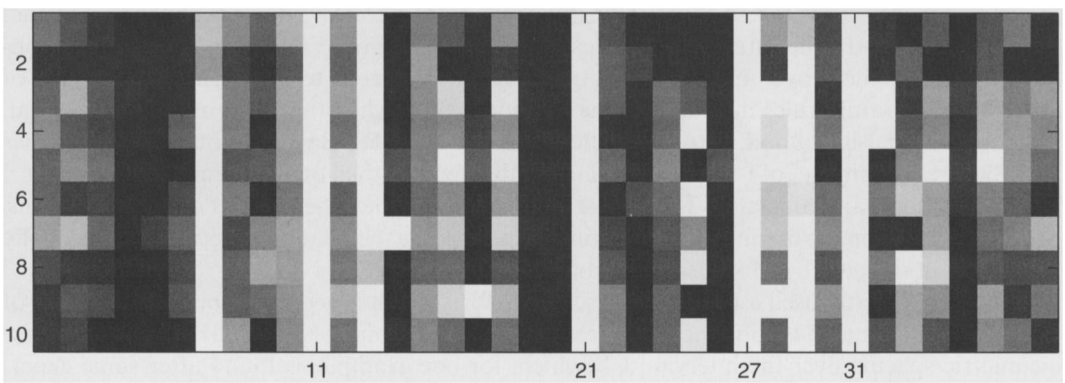

Fig. 3. Expression levels of 34 genes on the *Streptomyces* genome for 10 consecutive time points: the x-axis labels missing columns

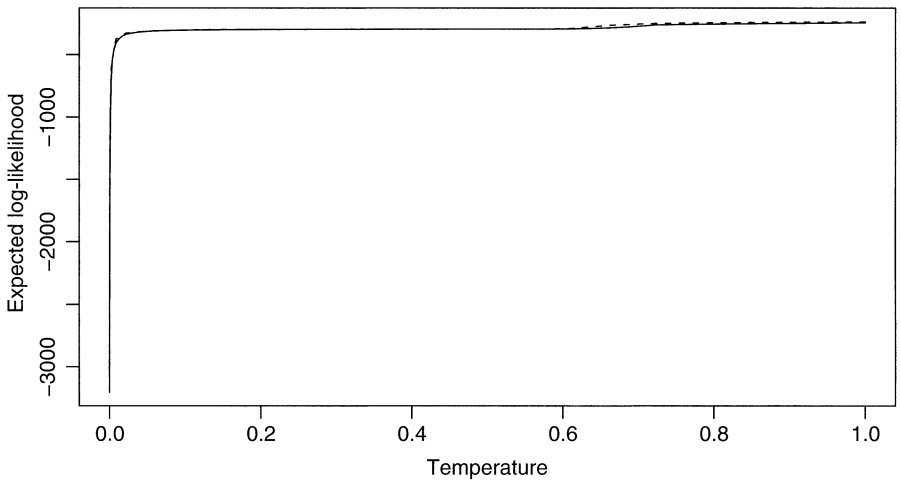

Fig. 4. Expected deviance against temperature for model 1 (———) and model 2 (-------)

ary distribution $p_{t_i}(\mathbf{x}, \beta, \mu|\mathbf{y}, k)$, for $k = 1, 2$. Fig. 4 shows the values of the expected deviance for the two models and demonstrates, at this scale, the similarity between these two functions. However, applying the trapezoidal rule (7) yields $\log\{p(\mathbf{y}|k = 1)\} = -284.27$ and $\log\{p(\mathbf{y}|k = 2)\} = -289.401$, with associated Monte Carlo standard errors of 0.019 and 0.021 respectively. Thus the first-order neighbourhood model is deemed substantially more probable *a posteriori*.

5. Discussion and concluding remarks

We first discuss choices of quadrature rule, choice of temperature schedule for t and use of simulation resources. These issues are obviously related to derive an efficient computational approach which provides marginal likelihood estimates with tolerably small standard errors.

Although there is a history of using quadrature in Bayesian statistics (e.g. Smith *et al.* (1987)), the problems that were encountered there of integrating unimodal likelihood functions are different from the problem here of integrating the expected log-likelihood under the power posterior. Consider first the choice of quadrature rule. We have chosen to use the trapezoidal rule which has an error which involves the second derivative of the integrand rather than, for example, Simpson's rule which has an error which involves the third derivative or Romberg's rule which involves a higher derivative. Additionally, if the integrand is concave down then the trapezoidal rule underestimates the integral. In the cases here the integrand has been found to be concave down but linear over most of the range $[0, 1]$. There is need to investigate whether use of different quadrature rules does reduce the statistical error when the usual use of the integral, the log-marginal-likelihood, is to take differences to obtain log-Bayes-factors. As in examples here, where the same set of t-values is used for both integrals, then quadrature errors may cancel when differences are calculated. The simple regression example of Section 4.1 demonstrates this. Additionally when the number of quadrature points is large the quadrature error is additionally small and dominated by Monte Carlo error.

We have explicitly used a temperature schedule for t which is geometric but investigation of other schedules may possibly lead to improvements. For example, Neal (2001) used a piecewise geometric spacing over the interval $[0, 1]$ which, for one example, is found after some experimentation. An important property of the temperature schedule is to make sure that power posterior distributions are reasonably overlapping so that the serial MCMC approach works

well with a small burn-in number of iterations. In the importance sampling literature, there is a similar problem to move smoothly from one distribution to another, and a larger number of values, n, of t is generally used than we have advocated here. For example Neal (2001) used 1000 values in examples. Our serial method, as demonstrated for the regression problem in Section 4.1, appears to work well for a range of values of n in [10,100]. However, it would be expected that with more uninformative choices of priors larger values of n would be required.

Any reasonable schedule for t must also reflect optimization of the resulting Monte Carlo standard errors, which are given by equation (8). When a fixed number of MCMC iterations for each t-value is used, as in the serial MCMC algorithm of Section 3.2.2, then the Monte Carlo standard errors for each estimate of the mean log-likelihood, s_i in equation (8), increase for smaller values of t. Optimally, the decreasing spacings of the t-values as $t \to 0$ should match the increasing values of the s_is. Some experimentation for a particular case could lead to an adaptation of the spacings on this basis.

Computation of the marginal likelihood requires a proper prior. The sensitivity of the value of the marginal likelihood to the choice of prior can be readily investigated by using our method. Various approaches have been proposed for the case where the prior is improper. As we mentioned above, the fractional Bayes factor is straightforwardly computed as a by-product of the marginal likelihood. For those seeking such approximations our method provides a straightforward solution.

Our limited studies suggested that there were few advantages to be obtained by using a population Monte Carlo approach. However, experience gained in investigations that are not reported here suggests that in models where the likelihood is multimodal the population Monte Carlo approach has a somewhat superior performance and should be considered in these cases.

In this paper, we have introduced a new method of estimating the marginal likelihood for complex hierarchical Bayesian models which involves a minimal amount of change to commonly used algorithms which compute the posterior distributions of unknown parameters. We have illustrated the technique for three examples. For the first, a simple regression example, the prior model probabilities needed to be tuned to estimate the Bayes factor well by using RJMCMC sampling. The second example involved a random-effects model for survival data and demonstrated the ease of computing the marginal likelihood with a standard software package such as WinBUGS. Here the results demonstrated little difference in marginal likelihoods for the two models that were considered. The third example involved a complex hidden Markov structure and the results demonstrated a difference in terms of marginal likelihood between the two models.

As illustrated by the hidden MRF example our method deals straightforwardly with missing and auxiliary data models. A further example arises with an MCMC algorithm which involves RJMCMC sampling for an essentially fixed dimension model. It was implemented in Forrester *et al.* (2007) to impute pathogen colonization times of patients for those patients whose colonization status is uncertain owing to the imperfect sensitivity of a test. Interest focuses here on Bayes factors for models which involve and do not involve transmission from patients to other patients via health care workers. It is straightforward to adapt the power posterior method to calculate within-model marginal likelihoods in this instance by using RJMCMC sampling to find the necessary simulation means and therefore to compute Bayes factors.

In conclusion, we have illustrated a method of computing the marginal likelihood which is straightforward to implement as an addition to existing algorithms which can be used for complex models.

Acknowledgements

The work of both authors was supported by the Australian Research Council. The authors kindly acknowledge Candice Hincksman for her assistance with computational aspects of this work. Nial Friel acknowledges the School of Mathematical Sciences, Queensland University of Technology, for its hospitality during June 2005. Both authors also acknowledge the help that was provided by the Joint Editor, an Associate Editor and a reviewer which led to a greatly improved paper.

References

Bartolucci, F., Scaccia, L. and Mira, A. (2006) Efficient Bayes factor estimation from the reversible jump output. *Biometrika*, **93**, 41–52.

Berger, J. O. and Pericchi, L. R. (1996) The intrinsic Bayes factor for linear models. In *Bayesian Statistics 5* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 25–44. Oxford: Oxford University Press.

Besag, J. (1974) Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. R. Statist. Soc.* B, **36**, 192–236.

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

Carlin, B. P. and Chib, S. (1995) Bayesian model choice via Markov chain Monte Carlo methods. *J. R. Statist. Soc.* B, **57**, 473–484.

Chib, S. (1995) Marginal likelihood from the Gibbs output. *J. Am. Statist. Ass.*, **90**, 1313–1321.

Chib, S. and Jeliazkov, I. (2001) Marginal likelihood from the Metropolis-Hastings output. *J. Am. Statist. Ass.*, **96**, 270–281.

Dellaportas, P., Forster, J. J. and Ntzoufras, I. (2001) On Bayesian model and variable selection using MCMC. *Statist. Comput.*, **12**, 27–36.

Dryden, I. L., Scarr, M. R. and Taylor, C. C. (2003) Bayesian texture segmentation of weed and crop images using reversible jump Markov chain Monte Carlo methods. *Appl. Statist.*, **52**, 31–50.

Forrester, M., Pettitt, A. N. and Gibson, G. (2007) Bayesian inference of hospital-acquired infections and control measures given imperfect surveillance data. *Biostatistics*, **8**, 383–401.

Friel, N., Pettitt, A. N., Reeves, R. and Wit, E. (2005) Bayesian inference in hidden Markov random fields for binary data defined on large lattices. *Technical Report*. University of Glasgow, Glasgow.

Gelman, A. and Meng, X.-L. (1998) Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Statist. Sci.*, **13**, 163–185.

Godsill, S. J. (2001) On the relationship between Markov Chain Monte Carlo methods for model uncertainty. *J. Computnl Graph. Statist.*, **10**, 230–248.

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

Green, P. J., Hjort, N. L. and Richardson, S. (eds) (2003) Trans-dimensional Markov chain Monte Carlo. In *Highly Structured Stochastic Systems*. Oxford: Oxford University Press.

Green, P. J. and O'Hagan, A. (1998) Model choice with MCMC on product spaces without using pseudo-priors. *Technical Report 98-13*. University of Nottingham, Nottingham.

Green, P. J. and Richardson, S. (2002) Hidden Markov models and disease mapping. *J. Am. Statist. Ass.*, **97**, 1055–1070.

Han, C. and Carlin, B. P. (2001) Markov chain Monte Carlo methods for computing Bayes factors: a comparative review. *J. Am. Statist. Ass.*, **96**, 1122–1132.

Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T. (2001) Bayesian model averaging: a tutorial. *Statist. Sci.*, **14**, 382–417.

Kass, R. E. and Wasserman, L. (1995) A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J. Am. Statist. Ass.*, **90**, 928–934.

Liang, F. and Wong, W. H. (2001) Real-parameter evolutionary sampling with applications in Bayesian mixture models. *J. Am. Statist. Ass.*, **96**, 653–666.

Liu, J. (2001) *Monte Carlo Strategies in Scientific Computing*. New York: Springer.

McGilchrist, C. A. and Aisbett, C. W. (1991) Regression with frailty in survival analysis. *Biometrics*, **47**, 461–466.

Meng, X.-L. and Wong, W. (1996) Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statist. Sin.*, **6**, 831–860.

Neal, R. M. (2001) Annealed importance sampling. *Statist. Comput.*, **11**, 125–139.

O'Hagan, A. (1995) Fractional Bayes factors for model comparison (with discussion). *J. R. Statist. Soc.* B, **57**, 99–138.

Perez, J. M. and Berger, J. (2002) Expected posterior prior distributions for model selection. *Biometrika*, **89**, 491–512.

Reeves, R. and Pettitt, A. N. (2004) Efficient recursions for general factorisable models. *Biometrika*, **91**, 751–757.
Ridall, P. G., Pettitt, A. N., Friel, N., McCombe, P. A. and Henderson, R. D. (2007) Motor unit number estimation using reversible jump Markov chain Monte Carlo methods (with discussion). *Appl. Statist.*, **56**, 235–269.
Sisson, S. A. (2005) Trans-dimensional Markov chains: a decade of progress and future perspectives. *J. Am. Statist. Ass.*, **100**, 1077–1089.
Smith, A. F. M., Skene, A. M., Shaw, J. E. H. and Naylor, J. C. (1987) Progress with numerical and graphical methods for practical Bayesian statistics. *Statistician*, **36**, 75–82.
Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002) Bayesian measures of model complexity and fit (with discussion). *J. R. Statist. Soc.* B, **64**, 583–639.
Spiegelhalter, D. J., Thomas, A. and Best, N. G. (2003) *WinBUGS: Bayesian Inference using Gibbs Sampling, Manual Version 1.4*. Cambridge: Medical Research Council Biostatistics Unit.
Stephens, M. (2000) Bayesian analysis of mixture models with an unknown number of components—an alternative to reversible jump methods. *Ann. Statist.*, **28**, 40–74.
Williams, E. (1959) *Regression Analysis*. Chichester: Wiley.
Wit, E. and McClure, J. (2004) *Statistics for Microarrays: Design, Analysis and Inference*. Chichester: Wiley.