Computational Statistics and Data Analysis 71 (2014) 392–401

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Computation of marginal likelihoods with data-dependent support for latent variables[☆]

Sarah E. Heaps[*], Richard J. Boys, Malcolm Farrow

School of Mathematics & Statistics, Herschel Building, Newcastle University, Newcastle upon Tyne, NE1 7RU, United Kingdom

ARTICLE INFO

Article history:
Received 9 June 2012
Received in revised form 22 July 2013
Accepted 23 July 2013
Available online 31 July 2013

Keywords:
Annealed importance sampling
Latent variables
Linked importance sampling
Marginal likelihood
Power posterior method
Spatial count data

ABSTRACT

Several Monte Carlo methods have been proposed for computing marginal likelihoods in Bayesian analyses. Some of these involve sampling from a sequence of intermediate distributions between the prior and posterior. A difficulty arises if the support in the posterior distribution is a proper subset of that in the prior distribution. This can happen in problems involving latent variables whose support depends upon the data and can make some methods inefficient and others invalid. The correction required for models of this type is derived and its use is illustrated by finding the marginal likelihoods in two examples. One concerns a model for competing risks. The other involves a zero-inflated over-dispersed Poisson model for counts of centipedes, using latent Gaussian variables to capture spatial dependence.

1. Introduction

The marginal likelihood, also known as the integrated likelihood or the evidence, plays an important role in Bayesian inference, particularly in model selection and model averaging, where it is used in the computation of Bayes factors and posterior model probabilities.

Consider data $\mathbf{y}$ and a statistical model $p(\mathbf{y}|\boldsymbol{\theta})$ which depends on unknowns $\boldsymbol{\theta}$. The marginal likelihood is defined as $p(\mathbf{y}) = \int p(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}$, where $\pi(\boldsymbol{\theta})$ is the prior density. Typically this integral cannot be evaluated in closed form and so we turn to numerical approximation; see, for example, Friel and Wyse (2012), for a recent review. It is convenient to use methods which involve Markov chain Monte Carlo (MCMC) sampling. In particular, this allows auxiliary or latent variables to be included in the unknowns $\boldsymbol{\theta}$ and sampled along with the model parameters. We focus primarily on latent variable problems in this paper.

Amongst the Monte Carlo methods particularly suitable for latent variable problems are Chib's method (Chib, 1995; Chib and Jeliazkov, 2001) and techniques which we term *intermediate-density-methods*. Chib's method is based on a rearrangement of Bayes Theorem to express the marginal likelihood in terms of the prior density, likelihood and posterior density. Evaluating or approximating each term in the resulting identity at a single point in the parameter space then yields the marginal likelihood approximation. Intermediate-density-methods connect the unnormalised prior and posterior densities through a sequence of intermediate densities labelled by an index $t \in [0, 1]$. They are derived from more general approaches for computing ratios of normalising constants and include the power posterior method (Friel and Pettitt, 2008; Friel et al., 2012), annealed importance sampling (AIS) (Neal, 2001) and linked importance sampling (LIS) (Neal, 2005). The

[☆] This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

[*] Corresponding author. Tel.: +44 0 191 222 7245.
E-mail addresses: sarah.heaps@newcastle.ac.uk (S.E. Heaps), richard.boys@newcastle.ac.uk (R.J. Boys), malcolm.farrow@newcastle.ac.uk (M. Farrow).

0167-9473/$ – see front matter © 2013 The Authors. Published by Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.07.033

unnormalised intermediate density with index t may then be, for example, the product of the unnormalised prior and the likelihood raised to the power t. An advantage of Chib's and, in particular, intermediate-density-methods, is the ease with which they can be programmed, often simply by rearranging code for sampling from the posterior distribution. Both types of methods can also be very effective. For instance, Germain (2010) found them to provide an easily implemented and accurate approximation to the marginal likelihoods of hidden Markov models with different numbers of states.

Latent variable problems can have the property that the support of the prior and posterior distributions do not coincide because the support for the latent variables changes when data are observed. For models with this property, some of the intermediate-density-methods, such as the power posterior approach, cannot be directly applied whilst others, like AIS, are likely to be very inefficient in cases where the prior probability of the posterior support is small, such as in multivariate probit models. Data-dependent support can also present problems for Chib's method if the likelihood ordinate (typically the observed data likelihood) is difficult to evaluate. This paper addresses the former of these issues and describes a general two-stage procedure to correct, or improve the efficiency of, intermediate-density-methods in problems involving data-dependent support, whilst also highlighting the situations in which implementation of the proposed approach is likely to be simpler than Chib's method.

We review intermediate-density-methods for computing marginal likelihoods in Section 2. In Section 3 we discuss the change of support problem and derive our two-stage approximation procedure. Next, in Section 4 we consider two examples. Section 4.1 concerns a simple model for competing risks and Section 4.2 applies our two-stage procedure to a zero inflated–over-dispersed Poisson model for a set of centipede count data. In this model, latent Gaussian variables capture the spatial dependences between the presence and the abundance of centipedes and we compare three variants of the model which use different parametric forms for the covariance matrix. Finally Section 4.3 provides a numerical comparison between our proposed method and other, related methods for marginal likelihood approximation.

2. Computing marginal likelihoods using sequences of densities

Consider a pair of density functions $p_t(\boldsymbol{\theta})$, $t = 0, 1$, with $p_t(\boldsymbol{\theta}) = q_t(\boldsymbol{\theta})/z_t$ for $\boldsymbol{\theta} \in \Theta_t$, where $q_t(\boldsymbol{\theta})$ is the unnormalised density, z_t is a normalising constant and Θ_t is the support of p_t. Several techniques for computing marginal likelihoods are special cases of more general methods for computing ratios of normalising constants, $r = z_1/z_0$. Let $p_0(\boldsymbol{\theta})$ be the prior density, $\pi(\boldsymbol{\theta})$, and let $p_1(\boldsymbol{\theta})$ be the posterior density, $\pi(\boldsymbol{\theta}|\mathbf{y})$. Then, if $q_1(\boldsymbol{\theta}) = p(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$, where $p(\mathbf{y}|\boldsymbol{\theta})$ is the likelihood, z_1 is the marginal likelihood, $p(\mathbf{y})$. Typically, the normalising constant of the prior distribution will be known and we can assume that $z_0 = 1$. Then the ratio $r = z_1/z_0 = z_1 = p(\mathbf{y})$.

2.1. Computing ratios of normalising constants

Provided that $\Theta_1 \subseteq \Theta_0$, it appears that we might approximate the ratio z_1/z_0 using *simple importance sampling*:

$$\frac{z_1}{z_0} = \mathrm{E}_{p_0}\left\{\frac{q_1(\boldsymbol{\theta})}{q_0(\boldsymbol{\theta})}\right\} \simeq \frac{1}{M}\sum_{i=1}^{M}\frac{q_1(\boldsymbol{\theta}^{[i]})}{q_0(\boldsymbol{\theta}^{[i]})}, \tag{1}$$

where E_{p_0} denotes expectation with respect to p_0 and $\boldsymbol{\theta}^{[1]}, \ldots, \boldsymbol{\theta}^{[M]}$ are a sample drawn from p_0. However this method works poorly when the overlap of p_0 and p_1 is small, as will typically be the case if p_0 and p_1 represent the prior and posterior and the posterior is very concentrated relative to the prior.

In response to this problem, *bridge sampling* (Meng and Wong, 1996) uses an unnormalised density $q_{0.5}$, with support $\Theta_0 \cap \Theta_1$, to provide a "bridge" between p_0 and p_1. This leads to the identity

$$\frac{z_1}{z_0} = \frac{\mathrm{E}_{p_0}\left\{\frac{q_{0.5}(\boldsymbol{\theta})}{q_0(\boldsymbol{\theta})}\right\}}{\mathrm{E}_{p_1}\left\{\frac{q_{0.5}(\boldsymbol{\theta})}{q_1(\boldsymbol{\theta})}\right\}}, \tag{2}$$

in which the ratios in the numerator and denominator are each approximated using simple importance sampling, as in (1). Whereas simple importance sampling requires $\Theta_1 \subseteq \Theta_0$, bridge sampling only requires $\int_{\Theta_0\cap\Theta_1} p_0(\boldsymbol{\theta})p_1(\boldsymbol{\theta})\, d\boldsymbol{\theta} > 0$.

When there is little overlap between p_0 and p_1, bridge sampling with a single intermediate density will perform poorly. However we can improve performance by introducing a sequence of intermediate densities, $p_{t_i}(\boldsymbol{\theta}) = q_{t_i}(\boldsymbol{\theta})/z_{t_i}$, $\boldsymbol{\theta} \in \Theta_{t_i}$, $i = 0, \ldots, n$, between p_0 and p_1, with $0 = t_0 < t_1 < \cdots < t_n = 1$. Then the ratio z_1/z_0 can be expressed as

$$\frac{z_1}{z_0} = \prod_{i=1}^{n}\frac{z_{t_i}}{z_{t_{i-1}}}. \tag{3}$$

Each of the ratios $z_{t_i}/z_{t_{i-1}}$ can then be approximated by simple importance sampling or by bridge sampling using an unnormalised bridging density $q_{t_{i-0.5}}$. Provided that each pair $p_{t_{i-1}}, p_{t_i}$ displays sufficient overlap, this can provide substantial improvement over standard importance or bridge sampling. In the remainder of this paper, methods based on these ideas will be called *extended importance sampling* and *extended bridge sampling* techniques.

Path sampling (Gelman and Meng, 1998) is based on the construction of a continuous path $q(\boldsymbol{\theta}|t)$ between the unnormalised densities q_0 and q_1. Writing $p(\boldsymbol{\theta}|t)$, $q(\boldsymbol{\theta}|t)$, $z(t)$ and $\Theta(t)$ in place of $p_{t_i}(\boldsymbol{\theta})$, $q_{t_i}(\boldsymbol{\theta})$, z_{t_i} and Θ_{t_i}, we have $p_t(\boldsymbol{\theta}) = q(\boldsymbol{\theta}|t)/z(t)$, $\boldsymbol{\theta} \in \Theta(t)$, for $t \in [0, 1]$ with $p_0(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|t = 0)$, $p_1(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|t = 1)$ and $r = z(1)/z(0)$. Provided that $\Theta(t)$ does not depend on t, Gelman and Meng (1998) show that the identity

$$\log\left\{\frac{z(1)}{z(0)}\right\} = \int_0^1 \mathrm{E}_{\boldsymbol{\theta}|t}\{U(\boldsymbol{\theta}, t)\}\,dt, \tag{4}$$

where $U(\boldsymbol{\theta}, t) = \frac{d}{dt}\log q(\boldsymbol{\theta}|t)$ and the expectation is taken with respect to the density $p(\boldsymbol{\theta}|t)$, can be obtained from (3) by approximating each ratio by bridge sampling, taking logarithms and then considering the limit as the number of intermediate densities approaches infinity.

2.2. Computing marginal likelihoods

If q_0 and q_1 are the unnormalised prior and posterior densities and $z_0 = 1$ then z_1/z_0 is the marginal likelihood. Several methods including the power posterior approach, AIS and LIS have been developed which tailor the techniques described in Section 2.1 to this special case. Consider first the power posterior approach. This is based on path sampling in which the unnormalised intermediate densities are defined by $q(\boldsymbol{\theta}|\mathbf{y}, t) = p(\mathbf{y}|\boldsymbol{\theta})^t q(\boldsymbol{\theta}|t = 0)$ for $\boldsymbol{\theta} \in \Theta(t)$ and $t \in [0, 1]$. Here $q(\boldsymbol{\theta}|t = 0)$ is the unnormalised prior and $p(\boldsymbol{\theta}|\mathbf{y}, t) = q(\boldsymbol{\theta}|\mathbf{y}, t)/z(\mathbf{y}|t)$ is termed the *power posterior* at temperature t. It follows from (4) that the log marginal likelihood can be expressed as

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

where the expectation is with respect to $p(\boldsymbol{\theta}|\mathbf{y}, t)$.

Friel and Pettitt (2008) (henceforth FP) suggest a serial MCMC approach to compute the integral in (5). The integral is discretised over $t \in [0, 1]$ as $0 = t_0 < t_1 < \cdots t_{n-1} < t_n = 1$, and then approximated by

$$\log p(\mathbf{y}) \simeq \sum_{i=0}^{n-1} \frac{1}{2}(t_{i+1} - t_i)\left[\mathrm{E}_{\boldsymbol{\theta}|\mathbf{y},t_{i+1}}\{\log p(\mathbf{y}|\boldsymbol{\theta})\} + \mathrm{E}_{\boldsymbol{\theta}|\mathbf{y},t_i}\{\log p(\mathbf{y}|\boldsymbol{\theta})\}\right]. \tag{6}$$

By separately sampling from the power posterior at each temperature t_i, the expectations $\mathrm{E}_{\boldsymbol{\theta}|\mathbf{y},t_i}\{\log p(\mathbf{y}|\boldsymbol{\theta})\}$ in (6) can be approximated. A theoretical advantage of this method is that it computes the marginal likelihood on the log-scale, thereby offering numerical stability.

AIS and LIS use a sequence of unnormalised intermediate distributions between the prior and posterior and apply (3). For example, the sequence employed in the power posterior method can be used, that is $q_{t_i}(\boldsymbol{\theta}|\mathbf{y}) = p(\mathbf{y}|\boldsymbol{\theta})^{t_i} q_0(\boldsymbol{\theta})$, where $0 = t_0 < t_1 < \cdots < t_n = 1$ and $q_0(\boldsymbol{\theta})$ is proportional to the prior. First consider AIS, which requires that $\Theta_{t_i} \subseteq \Theta_{t_{i-1}}$ for each $i = 1, \ldots, n$. At iteration j, $\boldsymbol{\theta}_{t_1}^{[j]}$ is sampled (preferably independently) from the prior p_{t_0}, then each of a series of Markov chain transitions, $T_{t_i}(\boldsymbol{\theta}_{t_i}^{[j]}, \cdot)$, generates a single draw, $\boldsymbol{\theta}_{t_{i+1}}^{[j]}$, from p_{t_i} for $i = 1, \ldots, n-1$, allowing iteration j to yield an importance weight

$$w^{[j]} = \frac{q_{t_1}(\boldsymbol{\theta}_{t_1}^{[j]})}{q_{t_0}(\boldsymbol{\theta}_{t_1}^{[j]})}\frac{q_{t_2}(\boldsymbol{\theta}_{t_2}^{[j]})}{q_{t_1}(\boldsymbol{\theta}_{t_2}^{[j]})}\cdots\frac{q_{t_n}(\boldsymbol{\theta}_{t_n}^{[j]})}{q_{t_{n-1}}(\boldsymbol{\theta}_{t_n}^{[j]})}.$$

The arithmetic mean of the weights computed from M iterations, $\sum w^{[j]}/M$, then provides an approximation to the marginal likelihood. An n-th Markov chain transition $T_{t_n}(\boldsymbol{\theta}_{t_n}^{[j]}, \cdot)$ at the end of each iteration j can be used to generate a draw $\boldsymbol{\theta}_{t_{n+1}}^{[j]}$ from the posterior p_{t_n}. AIS can be viewed as an importance sampler on an extended state space with points $(\boldsymbol{\theta}_{t_n}, \ldots, \boldsymbol{\theta}_{t_1})$ (whence the name "importance weight" for $w^{[j]}$) or as the average of M approximations, $w^{[j]}$, by an extended importance sampler, in which each ratio $z_{t_i}/z_{t_{i-1}}$ in every approximation is based on a single draw from $p_{t_{i-1}}$. When the posterior is multimodal, an advantage of AIS is that the intermediate densities allow the Markov chain to move more freely around the state space. Therefore if the additional Markov transition T_{t_n} is applied at the end of each iteration j, the resulting posterior sampler may be able to reach isolated modes which would otherwise be missed. Moreover, if the draws from the prior are independent, these draws, $\boldsymbol{\theta}_{t_{n+1}}^{[1]}, \ldots, \boldsymbol{\theta}_{t_{n+1}}^{[M]}$, from the posterior will also be independent.

LIS is similar to AIS except that the approximation of each of the ratios $z_{t_i}/z_{t_{i-1}}$ is akin to bridge sampling, rather than simple importance sampling; see Neal (2005). Note that, within every iteration of LIS, the approximation of each $z_{t_i}/z_{t_{i-1}}$ is based on multiple samples (from p_{t_i} and $p_{t_{i-1}}$) rather than a single sample (from $p_{t_{i-1}}$) in the case of AIS. Therefore when $\Theta_{t_i} \subset \Theta_{t_{i-1}}$ for some $i = 1, \ldots, n$, the approximation of $z_{t_i}/z_{t_{i-1}}$ on any iteration of LIS will only be zero if *all* the samples from $p_{t_{i-1}}$ lie in $\Theta_{t_{i-1}} \setminus \Theta_{t_i}$ on that iteration.

3. Data-dependent support for latent variables

Let $\boldsymbol{\theta} = (\boldsymbol{\theta}', \mathbf{z})$ where $\mathbf{z}$ are latent variables and $\boldsymbol{\theta}'$ are "model parameters". Let $\Theta_0,\ \Theta_1$ and Θ_t denote the supports of $\boldsymbol{\theta}$ in the prior, posterior and an intermediate density respectively. Note that $\Theta_1 \subseteq \Theta_0$. When $\Theta_1 = \Theta_0$ the methods described in Section 2.2 can be applied directly. This is the case for many problems in which $\boldsymbol{\theta}$ includes latent variables, such as the hidden Markov random field discussed in FP. For some models, however, the data can only be consistent with a constrained range of values for the latent variables $\mathbf{z}$. That is, the likelihood $p(\boldsymbol{\theta}|\mathbf{y})$ is zero for some $\mathbf{z}$ where $\boldsymbol{\theta} = (\boldsymbol{\theta}', \mathbf{z}) \in \Theta_0$. Consequently the posterior support of the latent variables depends on the data, so that $\Theta_1 = \Theta_1(\mathbf{y}) \subset \Theta_0$. For example, latent variables have data-dependent support in multivariate probit models (e.g. see Chib and Greenberg, 1998) or multinomial probit models (e.g. see McCulloch et al., 2000) because the components of the latent multivariate normal random vector have support over $(-\infty, \infty)$ in the prior, but only over some truncated region, such as $(-\infty, 0)$ or $[0, \infty)$, in the posterior. In the remainder of the paper we are concerned with this situation.

3.1. The change-of-support problem

Derivation of (5) for the power posterior method relies on an interchange of integration and differentiation which is not generally valid when Θ_t depends on t. In problems involving data-dependent support $\Theta_t = \Theta_1$ for $t \in (0, 1]$ but $\Theta_0 \neq \Theta_1$. Therefore the power posterior method cannot be applied directly. More generally, (4) in path sampling also depends on the legitimacy of this operation.

There is no such invalidity with extended importance or bridge sampling methods such as AIS or LIS. However, difficulties arise in approximating the ratio z_{t_1}/z_{t_0} if there is little overlap between $\Theta_{t_1} = \Theta_1$ and $\Theta_{t_0} = \Theta_0$, as is typically the case in data-dependent support problems. For instance, for the three models in Section 4.2, only 0.41%–6.23% of the prior probability lies within Θ_1. In AIS, this would make a high proportion of the weights $w^{[j]}$ equal to zero. The situation is slightly improved in LIS because approximation of z_{t_1}/z_{t_0} in each iteration j uses multiple samples from p_{t_0}, all of which would need to lie in $\Theta_0 \setminus \Theta_1$ for the ensuing weight $w^{[j]}$ to be zero. Nevertheless, the method will still be inefficient for problems where the prior probability of the posterior support is small.

We might therefore classify intermediate-density-methods for computing marginal likelihoods into two groups. Some, such as the power posterior method, require $\Theta_1 = \Theta_0$. Others, like those based on extended bridge or importance sampling, do not but may be inefficient in cases where there is a large difference between Θ_0 and Θ_1. In examples such as multivariate probit models, where the support for some or all of the latent variables becomes restricted when data are observed, the difference in support can be very large. Intermediate-density-methods can be effective when the change-of-support problem does not arise and specific methods offer their own benefits. Therefore we propose the following general modification to this collection of methods which facilitates their efficient use in problems involving data-dependent support.

3.2. Two-stage approximation of the marginal likelihood

Writing $\Theta_1(\mathbf{y})$ to show the dependence of the posterior support on $\mathbf{y}$, let $p^*(\mathbf{y}) = \int_{\Theta_1(\mathbf{y})} \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}$. The marginal likelihood can be factorised as $p(\mathbf{y}) = p^*(\mathbf{y})\bar{p}(\mathbf{y}|\boldsymbol{\theta} \in \Theta_1(\mathbf{y}))$, where $\bar{p}(\mathbf{y}|\boldsymbol{\theta} \in \Theta_1(\mathbf{y})) = \int_{\Theta_1(\mathbf{y})} p(\mathbf{y}|\boldsymbol{\theta})\pi^*(\boldsymbol{\theta})\, d\boldsymbol{\theta}$ and the prior density of $\boldsymbol{\theta}$ truncated to $\Theta_1(\mathbf{y})$ is $\pi^*(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta})/p^*(\mathbf{y})$. In some problems where $\Theta_1(\mathbf{y}) \subset \Theta_0$, $p(\mathbf{y}|\boldsymbol{\theta}) = C(\mathbf{y})$, a constant with respect to $\boldsymbol{\theta}$ for $\boldsymbol{\theta} \in \Theta_1(\mathbf{y})$. For example, in a simple probit model, each observed binary variable depends deterministically on the sign of a latent Gaussian variable and $C(\mathbf{y}) = 1$. In such a case $p(\mathbf{y}) = p^*(\mathbf{y})C(\mathbf{y})$. In this paper we focus on the more interesting situation where $p(\mathbf{y}|\boldsymbol{\theta})$ is *not* a constant for $\boldsymbol{\theta} \in \Theta_1(\mathbf{y})$. For this case we propose the following two-stage procedure.

1. Approximate $\bar{p}(\mathbf{y}|\boldsymbol{\theta} \in \Theta_1(\mathbf{y})) = p(\mathbf{y})/p^*(\mathbf{y})$ using one of the Monte Carlo techniques described in Section 2. These methods work better when independent samples from the prior are available. However independent sampling from $\pi^*(\boldsymbol{\theta})$ will not always be feasible so additional care may be required to ensure all important parts of $\Theta_1(\mathbf{y})$ are reached.
2. Approximate the adjustment $p^*(\mathbf{y})$.

Typically only the support of the latent variables $\mathbf{Z}$ changes and

$$p^*(\mathbf{y}) = \int_{\Theta'} \Pr(\mathbf{Z} \in \mathcal{S}(\mathbf{y})|\boldsymbol{\theta}')\pi(\boldsymbol{\theta}')\, d\boldsymbol{\theta}',$$

where Θ' is the prior (and posterior) support for $\boldsymbol{\theta}'$ and $\mathcal{S}(\mathbf{y})$ is the data-dependent posterior support for $\mathbf{Z}$. Thus $p^*(\mathbf{y})$ is the marginal likelihood for a simpler model with likelihood function $\Pr(\mathbf{Z} \in \mathcal{S}(\mathbf{y})|\boldsymbol{\theta}')$ and prior $\pi(\boldsymbol{\theta}')$. It should therefore be possible to compute $p^*(\mathbf{y})$ using an existing marginal likelihood method and, crucially, the approximation of $p^*(\mathbf{y})$ will be simpler than direct approximation of $p(\mathbf{y})$. For example, Chib's method could be used in conjunction with a MCMC scheme which includes the latent variables $\mathbf{Z}$. In many cases, however, computation of $p^*(\mathbf{y})$ will be much more straightforward since it typically involves integration over a subset of the support of standard distributions such as multivariate normal distributions. Often this can be performed using standard cubature software. This is especially true if conjugate priors are chosen as these often allow some of the integration to be performed in closed form.

4. Examples

In this section, we illustrate our two-stage procedure in application to a simple model for competing risks and to a novel spatial model for the presence and abundance of centipedes. In each case we first use the power posterior approach to compute the ratio $p(\mathbf{y})/p^*(\mathbf{y})$ and then use a straightforward method to compute the adjustment $p^*(\mathbf{y})$. Unlike the power posterior approach, Chib's method, AIS and LIS can all, in principle, be applied directly to problems involving data-dependent support. Section 4.3 therefore provides a numerical comparison with these other methods to illustrate the situations in which each method might be more or less efficient.

4.1. Example 1: survival of radiation exposed mice

The data, taken from Hoel and Walburg (1972), give the lifetimes of mice after exposure to radiation. There are two groups of mice: 95 conventional mice and 82 germ-free mice. The time of death, in days, is recorded along with the cause of death classified as follows: (i) thymic lymphoma; (ii) reticulum cell sarcoma; (iii) other causes. Each mouse has a single cause of death and there is no censoring. We analyse these data using a *competing risks* model and *latent lifetimes* (see, e.g. Crowder, 2001).

For mouse i ($i = 1, \dots, N = 177$), we observe the lifetime Y_i, the cause of death $c_i \in \{1, 2, 3\}$ and the group $g_i \in \{1, 2\}$ to which the mouse belongs. We suppose that $Y_i = \min\{Z_{i,1}, Z_{i,2}, Z_{i,3}\}$ where $Z_{i,1}, Z_{i,2}, Z_{i,3}$ are three potential “lifetimes”, one for each cause of death. Thus, for each mouse, we observe $Y_i = Z_{i,c_i}$ but all we know about $Z_{i,k}$, where $k \neq c_i$, is that $Z_{i,k} > Y_i$. We consider the three “lifetimes” for a given mouse to be conditionally independent given the model parameters and adopt a Gamma lifetime distribution

$$Z_{i,k}|\alpha, \lambda_{g_i,k} \sim \mathrm{Ga}(\alpha, \lambda_{g_i,k}).$$

Letting $\eta_{j,k} = \log \lambda_{j,k}$ for all j and k, we adopt a prior distribution in which $\boldsymbol{\eta} = (\eta_{1,1}, \dots, \eta_{2,3})^T$ is independent of α. We use the prior specification $\alpha \sim \mathrm{Ga}(4, 1)$ and $\boldsymbol{\eta} \sim \mathrm{N}_6(\mathbf{e}, E)$, where $\mathbf{e} = (-5, \dots, -5)^T$ and E is such that $\mathrm{Var}(\eta_{j,k}) = 0.1$, $\mathrm{Cov}(\eta_{j,k}, \eta_{j,m}) = \mathrm{Cov}(\eta_{j,k}, \eta_{\ell,k}) = 0.07$ and $\mathrm{Cov}(\eta_{j,k}, \eta_{\ell,m}) = 0.05$ for all other pairs $(\eta_{j,k}, \eta_{\ell,m})$.

Let $\boldsymbol{\theta} = (\boldsymbol{\theta}', \mathbf{z})$ where $\boldsymbol{\theta}' = (\alpha, \boldsymbol{\eta})$ and $\mathbf{z}$ comprises the two latent lifetimes for each mouse. The latent variable $Z_{i,k}$, where $k \neq c_i$, has non-zero support over $(0, \infty)$ in the prior, but only over (y_i, ∞) in the posterior. This means that $\Theta_1 \subset \Theta_0$. In this example $p(\mathbf{y}|\boldsymbol{\theta})$ is not a constant for $\boldsymbol{\theta} \in \Theta_1(\mathbf{y})$ so our two-stage procedure is applicable.

Note that the contribution of mouse i, $i = 1, \dots, N$, to the observed data likelihood is simply

$$\Gamma(\alpha)^{-1}\lambda_{g_i,c_i}^{\alpha} y_i^{\alpha-1} \exp(-\lambda_{g_i,c_i} y_i) \prod_{k \neq c_i} \int_{y_i}^{\infty} \Gamma(\alpha)^{-1}\lambda_{g_i,k}^{\alpha} z_{i,k}^{\alpha-1} \exp(-\lambda_{g_i,k} z_{i,k})\, dz_{i,k}.$$

The integral here is easily computed so we could avoid sampling the latent lifetimes and compute the log marginal likelihood in a single run of the unmodified power posterior method. A long run gave $\log p(\mathbf{y}) = -1402.8$ with Monte Carlo standard error 0.0292. However methods involving sampling latent variables are popular and can be conveniently implemented in standard MCMC software and used with our two-stage approach. This example thus provides an illustrative comparison.

To apply the two-stage method, we first computed $\log\{p(\mathbf{y})/p^*(\mathbf{y})\}$ using the power posterior method. We followed the recommendation of FP and chose a geometric spacing of the temperatures, $t_i = (i/n)^c$, for $i = 0, \dots, n$ where $n = 40$ and $c = 4$. At each temperature, 100 000 samples were generated, omitting the first 40 000 as burn-in. Again following the advice in FP, we integrated $\mathrm{E}_{\boldsymbol{\theta}|\mathbf{y},t}\{\log p(\mathbf{y}|\boldsymbol{\theta})\}$ over t numerically using the trapezoidal rule, as in (6). This produced an approximation of -1397.3 with Monte Carlo standard error 0.1261. To compute the adjustment $p^*(\mathbf{y})$ we used the draws of the model parameters from the power posterior at temperature $t_0 = 0$. These constitute a sample from the posterior with density proportional to $\Pr(\mathbf{Z} \in \mathcal{S}(\mathbf{y})|\boldsymbol{\theta}')\pi(\boldsymbol{\theta}')$ and with marginal likelihood equal to $p^*(\mathbf{y})$. After transforming to a new set of parameters $\boldsymbol{\theta}^\dagger = (\log\alpha, \boldsymbol{\eta})$, we constructed a normal approximation $\phi_7(\boldsymbol{\theta}^\dagger|\mathbf{m}, V)$ to the posterior density, with inflated variance, and then based our calculation of the marginal likelihood on the importance sampling identity

$$p^*(\mathbf{y}) = \int_{\mathbb{R}^7} \frac{\Pr(\mathbf{Z} \in \mathcal{S}(\mathbf{y})|\boldsymbol{\theta}^\dagger)\pi(\boldsymbol{\theta}^\dagger)}{\phi_7(\boldsymbol{\theta}^\dagger|\mathbf{m}, V)} \phi_7(\boldsymbol{\theta}^\dagger|\mathbf{m}, V)\, d\boldsymbol{\theta}^\dagger,$$

using 100 000 draws from the normal density $\phi_7(\boldsymbol{\theta}^\dagger|\mathbf{m}, V)$ in the calculation. This produced an approximation of $\log p^*(\mathbf{y}) = -5.8$ with Monte Carlo error 0.0138, obtained using the delta method. Overall this gave a marginal likelihood approximation of $\log p(\mathbf{y}) = -1403.1$ with Monte Carlo standard error 0.1269 which is consistent with our yardstick value. The size of the Monte Carlo standard error can be explained by the need to sample a large number of latent variables during computation of $\log\{p(\mathbf{y})/p^*(\mathbf{y})\}$.

4.2. Example 2: centipede presence and abundance

Blackburn et al. (2002) describe a study in which the numbers of centipedes, of several different species, in each of four microhabitats in small areas at each of $N = 30$ sites, were counted. For the purpose of this example we consider just one species, *Lithobius forficatus*, and a single microhabitat, rotting wood. The population density λ, in centipedes per square metre, will vary from place to place. Furthermore there may be some areas where the species is completely absent, i.e. $\lambda = 0$. We model the presence or absence of the species using a multivariate probit model. For each site i, we introduce a latent variable $Z_{0,i}$ and a presence indicator D_i where $d_i = 1$ if $z_{0,i} \geq 0$ and $d_i = 0$ if $z_{0,i} < 0$. Note that lower case letters are used for realisations. Let $\mathbf{Z}_0 \sim \mathrm{N}_N(\boldsymbol{\mu}, \Sigma)$ where $\mathbf{Z}_0 = (Z_{0,i})$, $\boldsymbol{\mu} = (\mu_i)$ and Σ is constrained to be a correlation matrix to ensure identifiability of $\boldsymbol{\mu}$ and Σ in the observed data likelihood. We let $\mu_i = \boldsymbol{\beta}^T \mathbf{x}_i$ where $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_5)^T$ and $\mathbf{x}_i = (1, x_{i,1}, \ldots, x_{i,5})^T$. Here $x_{i,1}, \ldots, x_{i,4}$ are centred covariates for site i, namely the logarithm of the percentage of organic matter in the soil, the logarithm of the site's altitude and the air and soil temperatures. Finally $x_{i,5}$ is a habitat-type variable taking the values 1 or -1 to designate whether site i is "Synanthropic" or "Deciduous" respectively.

To model the non-zero population densities we introduce a second latent variable $Z_{1,i}$ at site i with $Z_{1,i} = \gamma_0 + \gamma_1 Z_{0,i} + \varepsilon_i$ where $\varepsilon_i \sim \mathrm{N}(0,\ 1/\tau)$ with ε_i and ε_j independent given τ for $i \neq j$. The expected count for site i, given d_i and $z_{1,i}$, is then $\lambda_i = d_i \exp(z_{1,i})$. The count of *L. forficatus* at site i is Y_i, where, given λ_i, $Y_i \sim \mathrm{Po}(\lambda_i h_i)$, in which h_i is the area (in m^2) sampled at that site, and each h_i is small compared to the scale over which population density is likely to vary. Thus we have a zero-inflated, over-dispersed Poisson model where dependence between occurrences and abundances given occurrences is captured through latent Gaussian variables. This model is similar to one described in the discussion of Schmidt and Rodríguez (2011). When $y_i = 0$, D_i is not observed and so there is no restriction on the support of $Z_{0,i}$ but when $y_i > 0$, we observe $d_i = 1$ and so must have $z_{0,i} \geq 0$. In other words, *a posteriori*, $\mathbf{Z}_0$ is constrained to lie in the space $\mathcal{S}(\mathbf{y}) = \prod_{i=1}^N \mathcal{S}_i(y_i)$ where

$$\mathcal{S}_i(y_i) = \mathbb{R}^+ \quad \text{if } y_i > 0 \quad \text{or} \quad \mathcal{S}_i(y_i) = \mathbb{R} \quad \text{if } y_i = 0. \tag{7}$$

We wish to compare three models $\mathcal{M}_1,\ \mathcal{M}_2,\ \mathcal{M}_3$ with different forms $\Sigma_1, \Sigma_2, \Sigma_3$ respectively for Σ. Let the (i,j)th element of Σ_m be $\Sigma_{m,i,j}$. The three models are as follows:

$\mathcal{M}_1$: *spatial independence*. Here $\Sigma_{1,i,j} = 0$ for $i \neq j$.

$\mathcal{M}_2$: 2-*d exponential covariance*. Here $\Sigma_{2,i,j} = \exp(-\delta r_{i,j})$, $\delta \geq 0$, where $r_{i,j}$ is the distance (in km) between sites i and j.

$\mathcal{M}_3$: 3-*d exponential covariance*. Here we set $\beta_2 = 0$ and instead include altitude as a third spatial coordinate, using a simple case of the projection models of Schmidt et al. (2011). We set $\Sigma_{3,i,j} = \exp\{-(\mathbf{r}_{i,j}^T M \mathbf{r}_{i,j})^{1/2}\}$ where the three elements of the vector $\mathbf{r}_{i,j}$ are the differences in Easting (km), Northing (km) and altitude (m) respectively between sites i and j. We take $M = \mathrm{diag}(\delta, \delta, \delta_A)^2$. Schmidt and Rodríguez (2011) used a similar structure in a model for counts of fish, in which the third coordinate was depth of a lake.

Both $\mathcal{M}_2$ and $\mathcal{M}_3$ use special cases of the powered exponential family of spatial correlation functions (with shape parameter 1); see, for example, Section 3.4.2 of Diggle and Ribeiro (2007).

Let $\boldsymbol{\theta}_1' = (\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T, \tau)^T$ where $\boldsymbol{\gamma} = (\gamma_0, \gamma_1)^T$, and $\boldsymbol{\theta}_2' = \left((\boldsymbol{\theta}_1')^T, \delta\right)^T$ and $\boldsymbol{\theta}_3' = \left((\boldsymbol{\theta}_2')^T, \delta_A\right)^T$ so that $\boldsymbol{\theta}_m'$ denotes the collection of parameters for model $\mathcal{M}_m$. Our priors take the form

$$\begin{aligned} \pi(\boldsymbol{\theta}_1') &= \pi(\boldsymbol{\beta}|B_1)\,\pi(\boldsymbol{\gamma})\,\pi(\tau), \\ \pi(\boldsymbol{\theta}_2') &= \pi(\boldsymbol{\beta}|B_2)\,\pi(\boldsymbol{\gamma})\,\pi(\tau)\,\pi(\delta) \\ \text{and } \pi(\boldsymbol{\theta}_3') &= \pi(\boldsymbol{\beta}|B_3)\,\pi(\boldsymbol{\gamma})\,\pi(\tau)\,\pi(\delta)\,\pi(\delta_A). \end{aligned}$$

Denoting the standard normal cumulative distribution function by $\Phi(\cdot)$, our prior beliefs regarding the probability $\Phi(\boldsymbol{\beta}^T \mathbf{x}_j)$ that centipedes are present at various (hypothetical) sites j are summarised by $\boldsymbol{\beta}|B_m \sim \mathrm{N}_6(\mathbf{b}, B_m)$ where $\mathbf{b} = (0, \ldots, 0)^T$ and $B_m = \mathrm{diag}(0.630, 0.134, s_{m,2}, 0.056, 0.056, 0.540)^2$ for model $\mathcal{M}_m$ with $s_{1,2} = s_{2,2} = 0.112$ and $s_{3,2} = 0$. Consideration of our beliefs about the centipede density λ_i when centipedes are present led to the prior specification $\tau \sim \mathrm{Ga}(a_1, a_2)$ with $a_1 = 2.1$ and $a_2 = 0.21$, and $\boldsymbol{\gamma} \sim \mathrm{N}_2(\mathbf{g}, G)$ where $\mathbf{g} = (1.0, 0.2)^T$ and G is diagonal, $G = \mathrm{diag}(1.5, 0.1)^2$. Using the quantile method (Garthwaite et al., 2005) we chose $\delta \sim \mathrm{Ga}(c_1, c_2)$ for $\mathcal{M}_2$ and $\mathcal{M}_3$ and finally $\delta_A \sim \mathrm{Ga}(c_3, c_4)$ for $\mathcal{M}_3$ where $c_1 = c_3 = 1.56$, $c_2 = 54.2$ and $c_4 = 67.5$.

Let $\mathbf{d}_{\text{miss}} = \{d_i : y_i = 0\}$, $\mathbf{d}_{\text{obs}} = \{d_i : y_i > 0\}$ and $\mathbf{d} = \mathbf{d}_{\text{miss}} \cup \mathbf{d}_{\text{obs}}$. The values of $d_i \in \mathbf{d}_{\text{miss}}$ are unobserved while, if $d_i \in \mathbf{d}_{\text{obs}}$, then $d_i = 1$. Let $\mathbf{y} = (y_1, \ldots, y_N)^T$, $\mathbf{z}_1 = (z_{1,1}, \ldots, z_{1,N})^T$ and let X be a $(N \times 6)$ design matrix whose i-th row is $\mathbf{x}_i^T$. We denote the k-dimensional multivariate normal, $\mathrm{N}_k(\mathbf{m}, V)$, density function by $\phi_k(\cdot|\mathbf{m}, V)$, and the Poisson, $\mathrm{Po}(\lambda)$, mass function by $\mathrm{Po}(\cdot|\lambda)$. For model $\mathcal{M}_m$, the posterior distribution of interest is

$$\pi(\boldsymbol{\theta}_m', \mathbf{z}_0, \mathbf{z}_1, \mathbf{d}_{\text{miss}}|\mathbf{y}, \mathbf{d}_{\text{obs}}) \propto p(\mathbf{y}|\mathbf{z}_1, \mathbf{d})p(\mathbf{z}_1|\mathbf{z}_0, \boldsymbol{\theta}_m')p(\mathbf{d}|\mathbf{z}_0)p(\mathbf{z}_0|\boldsymbol{\theta}_m')\pi(\boldsymbol{\theta}_m'),$$

where the prior $\pi(\boldsymbol{\theta}_m')$ is as defined above,

$$\begin{aligned} p(\mathbf{z}_0|\boldsymbol{\theta}_m') &= \phi_N(\mathbf{z}_0|X\boldsymbol{\beta}, \Sigma_m), \\ p(\mathbf{z}_1|\mathbf{z}_0, \boldsymbol{\theta}_m') &= \prod_{i=1}^N p(z_{1,i}|z_{0,i}, \boldsymbol{\theta}_m') = \prod_{i=1}^N \phi_1(z_{1,i}|\gamma_0 + \gamma_1 z_{0,i},\ 1/\tau), \\ p(\mathbf{y}|\mathbf{z}_1, \mathbf{d}) &= \prod_{i:d_i=1} p(y_i|z_{1,i}, d_i) = \prod_{i:d_i=1} \mathrm{Po}\{y_i|h_i \exp(z_{1,i})\}, \\ p(\mathbf{d}|\mathbf{z}_0) &= \prod_{i=1}^N p(d_i|z_{0,i}) = \mathbb{I}\{d_i = \mathbb{I}(z_{0,i} > 0)\} \end{aligned} \tag{8}$$

and $\mathbb{I}(A)$ is an indicator function which is 1 if A is true and 0 otherwise.

Table 1
Approximation of $\log\{p(\mathbf{y}, \mathbf{d}_{\text{obs}})/p^*(\mathbf{y}, \mathbf{d}_{\text{obs}})\}$ by the power posterior approach and of $\log p^*(\mathbf{y}, \mathbf{d}_{\text{obs}})$. These are added to produce the overall approximation of the log marginal likelihood, $\log p(\mathbf{y}, \mathbf{d}_{\text{obs}})$. Shown in parentheses are the Monte Carlo standard errors. Note that the errors in approximating $\log p^*(\mathbf{y}, \mathbf{d}_{\text{obs}})$ were negligible. The times taken in R to produce the overall approximations are also indicated.

Model	$\log\left\{\frac{p(\mathbf{y},\mathbf{d}_{\text{obs}})}{p^*(\mathbf{y},\mathbf{d}_{\text{obs}})}\right\}$	$\log p^*(\mathbf{y}, \mathbf{d}_{\text{obs}})$	$\log p(\mathbf{y}, \mathbf{d}_{\text{obs}})$	Time (min)
$\mathcal{M}_1$	−74.9 (0.0521)	−5.5	−80.5 (0.0521)	784
$\mathcal{M}_2$	−77.3 (0.0691)	−2.8	−80.1 (0.0691)	850
$\mathcal{M}_3$	−75.8 (0.0656)	−3.7	−79.5 (0.0656)	976

Let $\boldsymbol{\theta}_m = (\boldsymbol{\theta}'_m, \mathbf{z}_0, \mathbf{z}_1, \mathbf{d}_{\text{miss}})$. For the parameters $\boldsymbol{\theta}'_m$ and the latent variables $(\mathbf{z}_1, \mathbf{d}_{\text{miss}})$, the set of values which has non-zero support in the prior is the same as that in the posterior. However, $\mathbf{z}_0$ has non-zero support over $\mathbb{R}^N$ in the prior but only over $\mathcal{S}(\mathbf{y})$, as defined in (7), in the posterior. It follows that $\Theta_1 \subset \Theta_0$ for all three models. Now, for any actually observed centipede count dataset $\mathbf{y}$, with associated $\mathbf{d}_{\text{obs}}$, the likelihood is given by $p(\mathbf{y}, \mathbf{d}_{\text{obs}}|\boldsymbol{\theta}_m) = p(\mathbf{y}|\mathbf{z}_1, \mathbf{d})$, where $p(\mathbf{y}|\mathbf{z}_1, \mathbf{d})$ is as defined in (8), whilst $p(\mathbf{y}, \mathbf{d}_{\text{obs}}|\boldsymbol{\theta}_m) = 0$ if $\boldsymbol{\theta}_m \notin \Theta_1$. Eq. (8) depends on $\mathbf{z}_1$ and $\mathbf{d}_{\text{miss}}$ and so is not a constant with respect to $\boldsymbol{\theta}_m$. Therefore we can use our two-stage procedure to compute the log marginal likelihood.

We computed $\log\{p(\mathbf{y}, \mathbf{d}_{\text{obs}})/p^*(\mathbf{y}, \mathbf{d}_{\text{obs}})\}$ using the power posterior method. At temperature $t_0 = 0$, we sampled from $\pi^*(\boldsymbol{\theta})$ by Gibbs sampling which allowed the latent variables $z_{0,i}$ to be updated one-at-a-time from their univariate truncated normal conditionals. Straightforward Metropolis within Gibbs sampling was used at each temperature $t > 0$. Only the full conditional distributions of $Z_{0,i}$ (when $y_i = 0$) and $Z_{1,i}$ actually depend on t. Again we chose a geometric spacing of the temperatures, $t_i = (i/n)^c$, for $i = 0, \ldots, n$ where $n = 40$ and $c = 4$, and generated 100 000 samples from the power posterior at each temperature, omitting the first 40 000 as burn-in. The integration over the temperature variable t was, again, carried out using the trapezoidal rule.

Now consider the integral of the prior over the support of the posterior, $p^*(\mathbf{y}, \mathbf{d}_{\text{obs}})$. For model $\mathcal{M}_1$, $p^*(\mathbf{y}, \mathbf{d}_{\text{obs}}) = \Pr\{\mathbf{Z}_0 \in \mathcal{S}(\mathbf{y})\}$ whilst for models $\mathcal{M}_2$ and $\mathcal{M}_3$,

$$p^*(\mathbf{y}, \mathbf{d}_{\text{obs}}) = \int_{\mathcal{S}_{\Sigma m}} \Pr\{\mathbf{Z}_0 \in \mathcal{S}(\mathbf{y})|\Sigma_m\} p(\Sigma_m) d\Sigma_m, \tag{9}$$

where $\mathbf{Z}_0|\Sigma_m \sim \mathrm{N}_N(X\mathbf{b}, XB_mX^T + \Sigma_m)$. It is straightforward to approximate each of these integrals numerically, for example, using standard software available in R (R Development Core Team, 2012). Specifically we used `pmvnorm()` from the `mvtnorm` package (Genz et al., 2012; Genz and Bretz, 2009) to define a function which (approximately) computes the integrand in (9). Then we used `integrate()` or `adaptIntegrate()`, the latter from the `cubature` package (Johnson and Narasimhan, 2011), to integrate numerically over the one-dimensional spaces, $\mathcal{S}_{\Sigma_2} = \{\mathcal{S}_{\Sigma_2} : \delta \geq 0\}$ or the two-dimensional space $\mathcal{S}_{\Sigma_3} = \{\mathcal{S}_{\Sigma_3} : \delta \geq 0, \delta_A \geq 0\}$.

Table 1 presents the approximations of the log marginal likelihoods together with their numerical standard errors. For each model, approximation of the correction term by numerical integration makes a negligible contribution to the numerical error. However it is clear that each of these correction terms makes an appreciable contribution to the overall marginal likelihood.

We can compute Bayes factors $B_{2,1} = 1.48$, $B_{3,1} = 2.62$ and $B_{3,2} = 1.78$ which remain greater than 1 even allowing for the numerical errors. Therefore, given our choice of priors for $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$ and $\boldsymbol{\theta}_3$, the data seem to support the two models which allow spatial correlation in the presence and abundance of centipedes, particularly when altitude is included as a third spatial coordinate (model $\mathcal{M}_3$). Assuming equal prior model probabilities, the posterior probabilities are 0.196, 0.289 and 0.515 for models $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_3$ respectively. However, the magnitude of the Bayes factors is not overwhelming, which illustrates the importance of accurately approximating the marginal likelihood.

Model $\mathcal{M}_1$ does noticeably worse in terms of $\log p^*(\mathbf{y}, \mathbf{d}_{\text{obs}})$, which is the log of the prior predictive probability that $\mathbf{D}_{\text{obs}} = \mathbf{d}_{\text{obs}}$. The models which include spatial correlation seem to be better able to account for the observed configuration of presence/absence of the centipedes. However, once we have allowed for this, Model $\mathcal{M}_1$ does better than the others, particularly $\mathcal{M}_2$, in terms of the remainder of the log marginal likelihood $\log\{p(\mathbf{y}, \mathbf{d}_{\text{obs}})/p^*(\mathbf{y}, \mathbf{d}_{\text{obs}})\}$ which suggests that the spatial correlation of the latent variables is now a handicap.

4.3. Numerical comparison

Data-dependent support presents no theoretical problem for extended importance and bridge sampling techniques, such as AIS and LIS, or for Chib's method. However, when applied to problems of this type, AIS and LIS may be inefficient whilst Chib's method can be difficult to implement (see Sections 1 and 2). To investigate the relative advantage of our two-stage procedure, we compared the results from the previous section with the results of applying each of these other approximation methods to the centipede data. For AIS, we chose a temperature schedule $t_i = (i/n)^c$, $i = 0, \ldots, n$, for $n = 400$ and $c = 4$. At temperature t, the parameters were sampled from their full conditional distributions, repeating the sequence of updates 10 times to give the overall transition kernel. The runs were repeated until the number of draws from powered posterior distributions was roughly equal to the number used during implementation of the power posterior approximation

Table 2
Approximation of the log marginal likelihood $\log p(\mathbf{y}, \mathbf{d}_{\text{obs}})$ using AIS, LIS and Chib's method. The Monte Carlo standard errors are shown in parentheses. The times taken in R to produce the approximations are also indicated.

Model	$\log p(\mathbf{y}, \mathbf{d}_{\text{obs}})$			Time (min)		
	AIS	LIS	Chib	AIS	LIS	Chib
$\mathcal{M}_1$	−80.5 (0.0602)	−80.5 (0.0673)	−80.3 (0.0360)	2771	1379	435
$\mathcal{M}_2$	−79.9 (0.0831)	−79.9 (0.1215)	−79.8 (0.0452)	862	875	507
$\mathcal{M}_3$	−79.5 (0.0662)	−79.5 (0.1029)	−79.4 (0.0474)	910	913	506

($41 \times 100\,000 = 4\,100\,000$ draws). For LIS, we chose the same form for the temperature schedule, this time taking $n = 40$ and $c = 4$ with a *geometric bridge*, $q_{t_{i+0.5}}(\boldsymbol{\theta}) = \sqrt{q_{t_i}(\boldsymbol{\theta})q_{t_{i+1}}(\boldsymbol{\theta})}$, as the unnormalised bridging density for each $i = 0, \dots, n-1$. We generated 100 samples from the powered posterior at each temperature and repeated the runs until the number of draws from powered posterior distributions was roughly equal to 4 100 000. In both cases, if the draws at temperature $t_0 = 0$ would guarantee an importance weight of zero for run j, the weight was recorded and then run j was terminated immediately.

In general, when data augmentation is used, the state $\boldsymbol{\theta}$ of the MCMC chain comprises latent variables $\mathbf{z}$ as well as parameters $\boldsymbol{\theta}'$. In such situations, Chib's method is ideally based on the identity

$$p(\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta}'^*)p(\boldsymbol{\theta}'^*)}{p(\boldsymbol{\theta}'^*|\mathbf{y})}, \tag{10}$$

for some high density point $\boldsymbol{\theta}'^*$ in the posterior support. Although the identity

$$p(\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{z}^*, \boldsymbol{\theta}'^*)p(\mathbf{z}^*, \boldsymbol{\theta}'^*)}{p(\mathbf{z}^*, \boldsymbol{\theta}'^*|\mathbf{y})}, \tag{11}$$

also holds, it is generally best avoided because the posterior ordinate can have huge dimension, making the approximation inefficient (Chib and Jeliazkov, 2001). In our case the observed data likelihood term in (10) is given by

$$p(\mathbf{y}, \mathbf{d}_{\text{obs}}|\boldsymbol{\theta}'_m) = \int_{\mathbb{R}^N} \int_{\mathcal{S}(\mathbf{y})} p(\mathbf{y}|\mathbf{z}_1, \mathbf{d})p(\mathbf{z}_0, \mathbf{z}_1|\boldsymbol{\theta}'_m)d\mathbf{z}_0 d\mathbf{z}_1, \tag{12}$$

where $p(\mathbf{z}_0, \mathbf{z}_1|\boldsymbol{\theta}'_m)$ is multivariate normal and the term $p(\mathbf{y}|\mathbf{z}_1, \mathbf{d})$ was defined in (8). If the term $p(\mathbf{y}|\mathbf{z}_1, \mathbf{d})$ had been omitted, an effective method for computing the integral would have been to transform to an integral over the unit hypercube and then to employ a recursive Monte Carlo algorithm (see, for example, Genz, 1992; Chib and Greenberg, 1998). For this example, however, the latter approach worked poorly because the density $p(\mathbf{y}|\mathbf{z}_1, \mathbf{d})$ was concentrated relative to $p(\mathbf{z}_0, \mathbf{z}_1|\boldsymbol{\theta}'_m)$ and so the approximation was dominated by a few large values of $p(\mathbf{y}|\mathbf{z}_1, \mathbf{d})$ and failed to converge even after one billion Monte Carlo iterations. Fortunately, in the centipede example, the dimension of the latent variable space was relatively small (60 latent variables in total), and so the problem of computing the observed data likelihood could be sidestepped by basing the approximation on (11). The approximations in Table 2 were produced by generating 100 000 draws in each complete and reduced MCMC run, of which the first 40 000 were discarded as burn-in.

Table 2 shows the marginal likelihood approximations obtained using AIS, LIS and Chib's method, together with the Monte Carlo standard errors and the time taken to produce the approximations in R. Considering also the results from Table 1, all four methods produce consistent marginal likelihood approximations, given the numerical errors. The Monte Carlo standard errors for AIS are similar to those for the modified power posterior approach but those for LIS are up to twice as large. For models $\mathcal{M}_2$ and $\mathcal{M}_3$, compared with the computation time for the two-stage power posterior approach, the times taken for both AIS and LIS were similar. However, for model $\mathcal{M}_1$, where the prior probability of the posterior support was smallest, the computation time was around 3.5 and 1.8 times larger for AIS and LIS, respectively. This was due to the time spent generating samples from the prior which fell outside the posterior support. As predicted in Section 2.2, the problem was less pronounced for LIS where only one of the 100 samples generated from the prior on each run needed to lie within the posterior support to produce a non-zero weight.

Although both the computation time and the numerical standard errors for Chib's method were smaller than those for our two-stage approach, it is unlikely that the approximation based on (11) would have scaled well with the dimension of the latent variable space. In this case successful application of Chib's method would have required use of the identity (10) which, in turn, would demand a method for accurately approximating the observed data likelihood (12), which we had been unable to find. In contrast, for our two-stage procedure, although computation of the adjustment term $p^*(\mathbf{y}, \mathbf{d}_{\text{obs}})$ by the proposed cubature method may become infeasible with increasing dimension of the latent variable space, the term could readily be computed by, for example, an application of Chib's method for multivariate probit models, as this would only require integration with respect to $\mathbf{Z}_0$ (to give $\Pr(\mathbf{Z}_0 \in \mathcal{S}(\mathbf{y})|\boldsymbol{\theta}'_m)$) to be carried out once. Our two-stage procedure therefore breaks the problem down into two separate integral approximations, each of which is feasible.

5. Discussion

The incorporation of latent variables is a common and useful technique in model building. Sometimes the posterior support of the latent variables depends on the data and is a proper subset of the prior support. Sampling from a sequence of intermediate distributions which connect the prior to the posterior is the basis for several Monte Carlo methods for approximating marginal likelihoods. These methods can be very effective when the support does not change but difficulties arise in problems involving data-dependent support. In these cases, some methods, such as the power posterior approach, are not valid because they require that the supports of the intermediate densities are equal. Other methods, such as AIS, present no theoretical problem but are likely to be inefficient if the spaces with non-zero prior and posterior density differ substantially. In order to make these intermediate-density-methods valid (or more efficient, as appropriate) in problems involving data-dependent support, we have proposed a general and usually straightforward adjustment.

First the ratio $p(\mathbf{y})/p^*(\mathbf{y})$ is computed using one of the existing techniques, modified so that the prior is truncated to the support of the posterior. This is then corrected through a second step which approximates $p^*(\mathbf{y})$. For the applications in Section 4 it was straightforward to compute the term $p^*(\mathbf{y})$ by standard importance sampling or cubature, and this would be the case for many other models. It may be possible to define sequences of intermediate densities which allow computation of the marginal likelihood in one step but this requires further investigation. In fact Section 4.2 suggests that it may be useful for model criticism to be able to see the two components of the log marginal likelihood separately.

Section 4.3 provided a numerical comparison between various intermediate-density-methods as well as Chib's method to help inform the following general guidelines for marginal likelihood approximation in problems involving data-dependent support. A one-step application of AIS or LIS can work well if the prior probability of the posterior support is not too small. However, as shown for model $\mathcal{M}_1$ in the centipede example, these methods can be very slow when this condition is not met. To know how small this integral will be we typically need to compute it and its computation is the second step of our proposed two-stage procedure. Chib's method is likely to work well if evaluation of the observed data likelihood $p(\mathbf{y}|\boldsymbol{\theta}'^*)$ is straightforward or, as seen in the centipede example, if the dimension of the space of latent variables is not overly large, allowing the approximation to be based on (11). However, when these conditions are not met, accurate approximation of the likelihood ordinate can be a much more difficult problem than computation of the prior probability of the posterior support. In this case our two-stage approach is likely to provide a more feasible solution by breaking the problem into two simpler integrations. Finally, in problems where computation of $p^*(\mathbf{y})$ requires only standard numerical integration software, our approach provides a viable alternative to Chib's method.

Section 4.2 introduced a zero-inflated over-dispersed Poisson model for spatially correlated count data in which the spatial dependence is carried by the latent variable governing presence. Using our two-stage procedure, we approximated the marginal likelihood for three models which assumed different parametric forms for the correlation matrix. In a later paper we plan to extend this model to describe more than one species, by introducing more latent Gaussian random variables with between-species correlation. The model could also be extended to handle spatio-temporal data by allowing the latent variables, $\mathbf{Z}_0$, at every time point to follow a temporal process.

Acknowledgements

This paper arose from work in the Ph.D. project of Sarah E. Heaps (née Germain) who was funded by the Engineering and Physical Sciences Research Council. We are grateful to Prof. Wallace Arthur of the National University of Ireland, Galway, for permission to use the centipede data and to the Editor and two Reviewers for comments which have led to improvements in the paper.

References

Blackburn, J., Farrow, M., Arthur, W., 2002. Factors influencing the distribution, abundance and diversity of geophilomorph and lithobiomorph centipedes. J. Zool. 256, 221–232.
Chib, S., 1995. Marginal likelihood from the Gibbs output. J. Amer. Statist. Assoc. 90, 1313–1321.
Chib, S., Greenberg, E., 1998. Analysis of multivariate probit models. Biometrika 85, 347–361.
Chib, S., Jeliazkov, I., 2001. Marginal likelihood from the Metropolis–Hastings output. J. Amer. Statist. Assoc. 96, 270–281.
Crowder, M.J., 2001. Classical Competing Risks. Chapman and Hall/CRC, London.
Diggle, P.J., Ribeiro, P.J., 2007. Model-Based Geostatistics. In: Springer Series in Statistics, Springer-Verlag, New York.
Friel, N., Hurn, M., Wyse, J., 2012. Improving power posterior estimation of statistical evidence. arXiv Pre-print arXiv:1209.3198 [stat.CO].
Friel, N., Pettitt, A.N., 2008. Marginal likelihood estimation via power posteriors. J. R. Stat. Soc. Ser. B 70, 589–607.
Friel, N., Wyse, J., 2012. Estimating the evidence—a review. Stat. Neerl. 66, 288–308.
Garthwaite, P.H., Kadane, J.B., O'Hagan, A., 2005. Statistical methods for eliciting probability distributions. J. Amer. Statist. Assoc. 100, 680–700.
Gelman, A., Meng, X.L., 1998. Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. Statist. Sci. 13, 163–185.
Genz, A., 1992. Numerical computation of multivariate normal probabilities. J. Comput. Graph. Statist. 1, 141–149.
Genz, A., Bretz, F., 2009. Computation of Multivariate Normal and t Probabilities. In: Lecture Notes in Statistics, vol. 195. Springer-Verlag, Heidelberg.
Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., Hothorn, T., 2012. mvtnorm: multivariate normal and t distributions. R package version 0.9-9992.
Germain, S.E., 2010. Bayesian spatio-temporal modelling of rainfall through non-homogeneous hidden Markov models. Ph.D. Thesis. School of Mathematics & Statistics, Newcastle University, UK.
Hoel, D.G., Walburg, H.E., 1972. Statistical analysis of survival experiments. J. Natl. Cancer Inst. 49, 361–372.
Johnson, S.G., Narasimhan, B., 2011. cubature: adaptive multivariate integration over hypercubes. R package version 1.1-1.
McCulloch, R.E., Polson, N.G., Rossi, P.E., 2000. A Bayesian analysis of the multinomial probit model with fully identified parameters. J. Econometrics 99, 173–193.

Meng, X.L., Wong, W.H., 1996. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. Statist. Sinica 6, 831–860.

Neal, R.M., 2001. Annealed importance sampling. Stat. Comput. 11, 125–139.

Neal, R.M., 2005. Estimating ratios of normalizing constants using linked importance sampling. Technical Report 0511. Department of Statistics, University of Toronto.

R Development Core Team 2012. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN: 3-900051-07-0.

Schmidt, A.M., Guttorp, P., O'Hagan, A., 2011. Considering covariates in the covariance structure of spatial processes. Environmetrics 22, 487–500.

Schmidt, A.M., Rodríguez, M.A., 2011. Modelling multivariate counts varying continuously in space (with discussion). In: Bernardo, J.M., Bayarri, M.J., Berger, J.O., Dawid, A.P., Heckerman, D., Smith, A.F.M., West, M. (Eds.), Bayesian Statistics, Vol. 9. Oxford University Press, pp. 611–638.