Bayesian Analysis (2018) **13**, Number 2, pp. 437–459

Efficient Model Comparison Techniques for Models Requiring Large Scale Data Augmentation

Panayiota Touloupou[*], Naif Alzahrani[†], Peter Neal[‡], Simon E. F. Spencer[§], and Trevelyan J. McKinley[¶]

Abstract. Selecting between competing statistical models is a challenging problem especially when the competing models are non-nested. In this paper we offer a simple solution by devising an algorithm which combines MCMC and importance sampling to obtain computationally efficient estimates of the marginal likelihood which can then be used to compare the models. The algorithm is successfully applied to a longitudinal epidemic data set, where calculating the marginal likelihood is made more challenging by the presence of large amounts of missing data. In this context, our importance sampling approach is shown to outperform existing methods for computing the marginal likelihood.

Keywords: epidemics, marginal likelihood, model evidence, model selection.

1 Introduction

The central pillar of Bayesian statistics is Bayes' Theorem. That is, given a parameteric model $\mathcal{M}$ with parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$ and data $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, the joint distribution of $(\boldsymbol{\theta}, \mathbf{x})$ satisfies

$$\pi(\boldsymbol{\theta}|\mathbf{x})\pi(\mathbf{x}) = \pi(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}). \tag{1}$$

The four terms in (1) are the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{x})$, the marginal likelihood or evidence $\pi(\mathbf{x})$, the likelihood $\pi(\mathbf{x}|\boldsymbol{\theta})$ and the prior distribution $\pi(\boldsymbol{\theta})$. The terms on the right hand side of (1) are usually easier to derive than those on the left hand side. The statistician has considerable control over the prior distribution and this can be chosen pragmatically to reflect prior beliefs and to be mathematically tractable. For many statistical problems the likelihood can easily be derived. However, the quantity of primary interest is usually the posterior distribution. Rearranging (1) it is straightforward to obtain an expression for $\pi(\boldsymbol{\theta}|\mathbf{x})$ so long as the marginal likelihood can be computed. This involves computing

[*]Department of Statistics, University of Warwick, Coventry, CV4 7AL, UK
[†]Department of Mathematics and Statistics, Lancaster University, Lancaster, LA1 4YF, UK
[‡]Department of Mathematics and Statistics, Lancaster University, Lancaster, LA1 4YF, UK, p.neal@lancaster.ac.uk
[§]Department of Statistics, University of Warwick, Coventry, CV4 7AL, UK
[¶]College of Engineering, Mathematics and Physical Sciences, University of Exeter, Penryn Campus, Penryn, Cornwall, TR10 9EZ, UK

 https://doi.org/10.1214/17-BA1057

$$\pi(\mathbf{x}) = \int \pi(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}, \tag{2}$$

which is only possible analytically for a relatively small set of simple models.

A key solution to being unable to obtain an analytical expression for the posterior distribution is to obtain samples from the posterior distribution using Markov chain Monte Carlo (MCMC; Metropolis et al., 1953; Hastings, 1970). A major strength of MCMC is that it circumvents the need to compute $\pi(\mathbf{x})$ and this has led to its widespread use in Bayesian statistics over the last 25 years or so. However, Bayesian model choice typically requires the computation of Bayes Factors (Kass and Raftery, 1995) or posterior model probabilities, which are both functions of the marginal likelihoods for the competing models. In Chib (1995) a simple rewriting of (1) was exploited to obtain estimates of the marginal likelihood using output from a Gibbs sampler. This has been extended in Chib and Jeliazkov (2001) and Chen (2005) to be used with the general Metropolis-Hastings algorithm. Importance sampling approaches to estimating the marginal likelihood have also been suggested (Gelfand and Dey, 1994), along with generalisations such as bridge sampling (Meng and Wong, 1996), which 'bridges' information from posterior and importance samples. More recently approaches have exploited the 'thermodynamic integral' such as power posterior methods Friel and Pettitt (2008). Alternative methods such as Sequential Monte Carlo (e.g. Zhou et al., 2015) and nested sampling (Skilling, 2004) do not require any MCMC: computation of the marginal likelihood and samples from the posterior distribution are produced simultaneously. A potential drawback for many of the above approaches to marginal likelihood estimation is that it may not be obvious how to apply them efficiently to models incorporating large amounts of missing data.

It should be noted that there are model comparison techniques such as reversible jump (RJ)MCMC (Green, 1995) which can used to compare models without the need to compute the marginal likelihood. RJMCMC works well for nested models where it is straightforward to define a good transition rule for models with different parameters. However, in the case where we have large amounts of missing data it is often necessary to use some form of data augmentation technique, where the missing information is inferred alongside the other parameters of the model. Using RJMCMC becomes much harder in these cases since the dimension of the parameter space (including the augmented data) becomes large. This is exacerbated further when the missing information between the competing models has a different structure. In this latter case the use of intermediary (bridging) models (Karagiannis and Andrieu, 2013) to move between the models of interest is a possibility.

The aim of the current paper is to demonstrate a straightforward mechanism for estimating the marginal likelihood of models with large amounts of missing data. The idea combines MCMC and importance sampling in a natural and semi-automatic manner to produce marginal likelihood estimates. The details of the algorithm developed are given Section 2. In Section 3 we consider a linear mixed model example. This enables us to demonstrate two important facets of the approach. Firstly, for the special case of the linear model we can compare our estimates of the marginal likelihood with exact computations which show very good agreement. Secondly, we show that making the distinction between model parameters and augmented data (random effects terms) assists

tremendously in devising an efficient estimator of the marginal likelihood. In Section 4 we consider final outcome household epidemic data where in the special case of the Reed-Frost model exact, but expensive, computation of the marginal likelihood is possible for comparison purposes. In Section 5 we apply the methodology to an epidemic example, for the transmission of *Streptococcus pneumoniae* (Melegaro et al., 2004) comparing our algorithm to existing methods for computing the marginal likelihood demonstrating its simplicity and effectiveness in the presence of missing data. Finally in Section 6 we briefly discuss extensions and limitations of the algorithm.

2 Algorithm

Our starting point in the estimation of $\pi(\mathbf{x})$ is to note that we can rewrite (2) as

$$\pi(\mathbf{x}) = \int_{\boldsymbol{\theta}} \pi(\mathbf{x}|\boldsymbol{\theta}) \frac{\pi(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} q(\boldsymbol{\theta})\, d\boldsymbol{\theta}, \tag{3}$$

where $q(\boldsymbol{\theta})$ denotes a d-dimensional probability density function. We assume that if $\pi(\boldsymbol{\theta}) > 0$ then $q(\boldsymbol{\theta}) > 0$. Then an unbiased estimator, $\widehat{P}_q$ of $\pi(\mathbf{x})$ is obtained by sampling $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_N$ from $q(\boldsymbol{\theta})$ and setting

$$\widehat{P}_q = \frac{1}{N} \sum_{i=1}^{N} \pi(\mathbf{x}|\boldsymbol{\theta}_i) \frac{\pi(\boldsymbol{\theta}_i)}{q(\boldsymbol{\theta}_i)}. \tag{4}$$

Thus $\widehat{P}_q$ is an importance sampled (see, for example, Ripley, 1987) estimate of $\pi(\mathbf{x})$, and the effectiveness of the estimator given by (4) depends upon the variability of $\pi(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})/q(\boldsymbol{\theta})$.

The remainder of the paper and this Section, in particular, is focussed on how we can effectively exploit (4) in the estimation of $\pi(\mathbf{x})$. The first observation is that the optimal choice of $q(\boldsymbol{\theta})$ is $\pi(\boldsymbol{\theta}|\mathbf{x})$, the posterior density but if we knew this, then $\pi(\mathbf{x})$ would also be known. A simple solution is to use output from an MCMC algorithm to inform the proposal distribution (Clyde et al., 2007). For most statistical models the likelihood times the prior is unimodal for sufficiently large n. In these circumstances, the posterior distribution of $\boldsymbol{\theta}$ is almost always approximately Gaussian with mean $\widehat{\boldsymbol{\theta}}$, the posterior mode, and covariance matrix $\Sigma = -\mathcal{I}(\widehat{\boldsymbol{\theta}})^{-1}$, where $\mathcal{I}(\boldsymbol{\theta})$ denotes the Fisher information evaluated at $\boldsymbol{\theta}$. That is, we have a central limit theorem type behaviour similar to that observed for maximum likelihood estimators as $n \to \infty$. This central limit theorem approximation is implicitly behind the Laplace approximations of integrals used in Tierney and Kadane (1986), (2.2) and Gelfand and Dey (1994), (8). This underpins the simple suggestion in Clyde et al. (2007) of using a multivariate t-distribution as an importance sampling distribution with location and scale parameters estimated from MCMC output. Alternatively, we can use a "defense mixture" (Hesterberg, 1995),

$$q_D(\boldsymbol{\theta}) = p\phi(\boldsymbol{\theta}; \boldsymbol{\mu}, \Sigma) + (1-p)\pi(\boldsymbol{\theta}), \tag{5}$$

where $\phi(\cdot; \boldsymbol{\mu}, \Sigma)$ is the probability density function of a multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ, estimated from the MCMC output, and

p is a mixing proportion. This proposal ensures that the ratio of the prior density to the proposal density is bounded above by $1/(1-p)$ with p typically chosen to be 0.95. We found that a t-distribution proposal was preferable in Sections 3 and 4, whereas the defense mixture proposal was the preferred choice in Section 5.

We are now in position to outline the three step algorithmic procedure, which is implemented in the paper followed by highlighting the scope and limitation of the approach. The steps are as follows:

1. Obtain a sample $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_K$ from the (approximate) posterior distribution, $\pi(\boldsymbol{\theta}|\mathbf{x})$. Throughout this paper, and in practice, this will generally be achieved using MCMC with K chosen such that the sample is representative of the posterior distribution. However, any alternative method for obtaining an approximate sample from the posterior distribution could be used.

2. Use the sample $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_K$ to derive a parametric approximation of the posterior distribution and let $q(\cdot)$ denote the corresponding probability density function. For example, choosing $q(\cdot)$ either to be a multivariate t-distribution or a "defense mixture" will usually work well.

3. Sample $\tilde{\boldsymbol{\theta}}_1, \tilde{\boldsymbol{\theta}}_2, \ldots, \tilde{\boldsymbol{\theta}}_N$ from $q(\cdot)$. (The tilde notation is used to distinguish the sample obtained from $q(\cdot)$ from the sample used to estimate $q(\cdot)$.) For each $i = 1, 2, \ldots, N$, compute $\pi(\mathbf{x}|\tilde{\boldsymbol{\theta}}_i)$ and estimate $\pi(\mathbf{x})$ using (4).

In situations where $\pi(\mathbf{x}|\boldsymbol{\theta})$ is analytically available, the construction of an MCMC algorithm to sample from $\pi(\boldsymbol{\theta}|\mathbf{x})$ will be straightforward and implementation of the algorithm will be trivial. Then the procedure becomes a simple and fast appendage to a standard MCMC algorithm. However, assuming an independent and identically distributed sample from $q(\cdot)$, the variance of the importance sampling estimator given in (4) is given by

$$\begin{aligned}\mathrm{Var}(\widehat{P}_q) &= N^{-1} \int \left(\pi(\mathbf{x}|\boldsymbol{\theta}) \frac{\pi(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} - \pi(\mathbf{x}) \right)^2 q(\boldsymbol{\theta})\, d\boldsymbol{\theta} \\ &= N^{-1} \pi(\mathbf{x})^2 \int \left(\frac{\pi(\boldsymbol{\theta}|\mathbf{x})}{q(\boldsymbol{\theta})} - 1 \right)^2 q(\boldsymbol{\theta})\, d\boldsymbol{\theta},\end{aligned}$$

which again highlights the importance of the proposal $q(\boldsymbol{\theta})$ resembling the posterior $\pi(\boldsymbol{\theta}|\mathbf{x})$ as closely as possible. As the dimension of $\boldsymbol{\theta}$ increases the variance of the estimator will typically grow due to the curse of dimensionality (see Doucet and Johansen, 2011, page 671 for an explanation) and this is the main potential limitation. The examples in Section 5 show that the algorithm can be effectively used for moderate numbers of parameters with $d = 11$. In passing we remark that a dependent sample from $q(\cdot)$ in Step 3 of the algorithm can be exploited to reduce the variance of the estimator. A prime example is the defense mixture proposal where pN and $(1-p)N$ samples are drawn from the multivariate Gaussian distribution and the prior, respectively.

The motivation for the work are in circumstances where $\pi(\mathbf{x}|\boldsymbol{\theta}_i)$ is not readily available, see Sections 3 and 5, and further work is required to implement the algorithm.

When $\pi(\mathbf{x}|\boldsymbol{\theta})$ is not available, it is often possible, with the addition of augmented data $\mathbf{y}$, to obtain an analytical expression for $\pi(\mathbf{x}, \mathbf{y}|\boldsymbol{\theta})$. This can then be utilised within an MCMC algorithm to obtain samples from the joint posterior $\pi(\boldsymbol{\theta}, \mathbf{y}|\mathbf{x})$. Devising an importance sampling proposal distribution $q(\boldsymbol{\theta}, \mathbf{y})$ approximating $\pi(\boldsymbol{\theta}, \mathbf{y}|\mathbf{x})$ will not be practical if $\mathbf{y}$ is high-dimensional, for example, the dimension of $\mathbf{y}$ is equal to or greater to n, the dimension of $\mathbf{x}$. See, for example, Section 3 for limitations of this approach. The solution that we propose is to use the marginal MCMC output from $\pi(\boldsymbol{\theta}|\mathbf{x})$ to inform the proposal distribution $q(\boldsymbol{\theta})$ in the importance sampling above, and then to separately consider the computation of $\pi(\mathbf{x}|\boldsymbol{\theta})$, which will be largely problem specific. In the linear mixed model example in Section 3, the distribution of y_i (random effect term) is readily available given $\boldsymbol{\theta}$ and $\mathbf{x}_i$, and hence we can sample the random effects $\mathbf{y}$ from their full conditional distributions. This approach extends to the epidemic model in Section 5, where $\mathbf{y}$ represents the unobserved infectious status of individuals with respect to *Streptococcus pneumoniae* carriage and the Forward Filtering Backwards Sampling (FFBS) algorithm (Carter and Kohn, 1994) can be used to calculate $\pi(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$, and hence $\pi(\mathbf{x}|\boldsymbol{\theta})$. In future work we will show how particle filtering, (Gordon et al., 1993), can be applied to estimate $\pi(\mathbf{x}|\boldsymbol{\theta})$ extending the scope of the algorithm with particular reference to Poisson regression models (Zeger, 1988). The estimation of $\pi(\mathbf{x}|\boldsymbol{\theta})$ can be potentially computationally costly and thus the overall cost of the algorithm needs to be considered. However, the computation of the $\{\pi(\mathbf{x}|\tilde{\boldsymbol{\theta}}_i)\}$'s can, in contrast to the MCMC runs, be undertaken in parallel, which can ease the computational burden.

Our approach can be used to estimate Bayes' Factors in objective model selection where for two competing models, common model parameters are assigned improper, non-informative priors. Consider two competing models $\mathcal{M}_1$ and $\mathcal{M}_2$ with parameters $\boldsymbol{\theta}^1 = (\boldsymbol{\phi}, \boldsymbol{\omega}^1)$ and $\boldsymbol{\theta}^2 = (\boldsymbol{\phi}, \boldsymbol{\omega}^2)$, respectively, and let $\boldsymbol{\phi}$ denote parameters common to both models. Let $\Phi \subseteq \mathbb{R}^d$ denote the sample space for $\boldsymbol{\phi}$. Suppose that a common prior $\pi_0(\boldsymbol{\phi})$ is chosen for $\boldsymbol{\phi}$ in both models and that the prior for $\mathcal{M}_k$ $(k = 1, 2)$ factorises as $\pi_k(\boldsymbol{\theta}^k) = \pi_0(\boldsymbol{\phi})\pi_{k1}(\boldsymbol{\omega}^k)$, where $\pi_{k1}(\cdot)$ is assumed to be a proper probability density. We can then choose $\boldsymbol{\phi}_0 \in \Phi$ as a reference point and set $\tilde{\pi}_0(\boldsymbol{\phi}_0) = 1$ and for all $\boldsymbol{\phi} \in \Phi$, set $\tilde{\pi}_0(\boldsymbol{\phi}) = \pi_0(\boldsymbol{\phi})/\pi_0(\boldsymbol{\phi}_0)$. Let

$$\pi_k(\mathbf{x}) = \int \int \pi_k(\mathbf{x}|\boldsymbol{\phi}, \boldsymbol{\omega}^k)\pi_0(\boldsymbol{\phi})\pi_{k1}(\boldsymbol{\omega}^k)\, d\boldsymbol{\phi}\, d\boldsymbol{\omega}^k, \tag{6}$$

and

$$\tilde{\pi}_k(\mathbf{x}) = \int \int \pi_k(\mathbf{x}|\boldsymbol{\phi}, \boldsymbol{\omega}^k)\tilde{\pi}_0(\boldsymbol{\phi})\pi_{k1}(\boldsymbol{\omega}^k)\, d\boldsymbol{\phi}\, d\boldsymbol{\omega}^k. \tag{7}$$

Then letting $B_{12} = \pi_1(\mathbf{x})/\pi_2(\mathbf{x})$ denote the Bayes' Factor between models 1 and 2, it follows from (6) and (7) that

$$B_{12} = \frac{\pi_1(\mathbf{x})}{\pi_2(\mathbf{x})} = \frac{\tilde{\pi}_1(\mathbf{x})}{\tilde{\pi}_2(\mathbf{x})}. \tag{8}$$

Therefore it suffices to estimate $\tilde{\pi}_k(\mathbf{x})$ $(k = 1, 2)$ in order to estimate B_{12}. The estimation of $\tilde{\pi}_k(\mathbf{x})$ can proceed along the same lines as $\pi_k(\mathbf{x})$ in (4) by selecting a proper proposal

density $q_k(\cdot)$ and using samples $(\boldsymbol{\phi}_1^k, \boldsymbol{\omega}_1^k), \ldots, (\boldsymbol{\phi}_N^k, \boldsymbol{\omega}_N^k)$ from $q_k(\cdot)$ to estimate $\tilde{\pi}_k(\mathbf{x})$ by

$$\widetilde{P}_q^k = \frac{1}{N} \sum_{i=1}^{N} \pi_k(\mathbf{x}|\boldsymbol{\phi}_i^k, \boldsymbol{\omega}_i^k) \frac{\tilde{\pi}_0(\boldsymbol{\phi}_i^k)\pi_{k1}(\boldsymbol{\omega}_i^k)}{q_k(\boldsymbol{\phi}_i^k, \boldsymbol{\omega}_i^k)}. \tag{9}$$

In this case the "defense mixture" proposal is inappropriate but a multivariate t-distribution can be used as an effective proposal distribution.

In our approach each model is required to be analysed separately and the computational cost increases approximately linearly in the number of models to be compared. Therefore this approach is not competitive for comparing large numbers of nested models, for example, the inclusion or exclusion of p covariates in a generalised linear model, a situation where reversible jump MCMC (Green, 1995) can be effectively applied. Our approach is more suited to comparing a small number of competing models which potentially have rather different dynamics such as integer valued autoregressive (Neal and Subba Rao, 2007) and Poisson regression (Zeger, 1988) models for integer valued time series, an example which we will present in future work. The approach is particularly suited to situations which allow the posterior distribution of the parameters to be approximately Gaussian, assisting in the construction $q(\cdot)$, but this assumption can be relaxed. Furthermore, the appropriateness of a Gaussian, or t-distribution approximation of the posterior can easily be assessed from the MCMC output.

3 Linear mixed model

We illustrate our methodology on the linear mixed model. In particular we may wish to ask the model choice question of whether it is necessary to include a random effect in the model or not. This question would be extremely challenging to address using reversible jump MCMC because it would require an efficient proposal distribution for the complete set of random effects when jumping between models. However it is straightforward to fit both models using MCMC due to the availability of a Gibbs sampler. The full conditional distribution of the random effects then unlocks an efficient importance sampling algorithm for the calculation of the marginal likelihood.

The simplest linear mixed model takes the following form. Let the data be divided into m units or clusters, and assume that

$$x_{ij} = \mathbf{z}_{ij}^T \boldsymbol{\beta} + \delta_i + \epsilon_{ij}, \tag{10}$$

for $i = 1, \ldots, m$ and $j = 1, \ldots, n_i$, where $\epsilon_{ij} \sim N(0, \sigma^2)$ are independent and identically distributed errors. We assume that the random effects satisfy $\delta_i|\phi \sim N(0, \phi^2)$ and are independent conditional on the standard deviation parameter ϕ. The vector of unknown parameters for the model is given by $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma, \boldsymbol{\delta}, \phi)$. Let $\mathbf{Z}$ denote the design matrix of the fixed effects, with rows $\mathbf{z}_{ij}^T$ and let $\mathbf{W}$ be the design matrix for the random effects, so that $\mathbf{x} = \mathbf{Z}\boldsymbol{\beta} + \mathbf{W}\boldsymbol{\delta} + \boldsymbol{\epsilon}$. For a review of Bayesian approaches to generalized linear mixed models, see for example Fong et al. (2010).

3.1 Simulation study

To illustrate the application of our importance sampling technique we performed a simulation study where the true model is known. We simulated data for $m = 50$ clusters, each containing $n_i = 3$ observations, giving $n = 150$ in total. We generated 3 predictor variables for each cluster by drawing m values from a standard normal distribution. We fixed our true parameters to be $\boldsymbol{\beta}^T = (10, -20, 30)$, $\sigma = 1$ and $\phi = 2$. For every cluster we assumed the same predictors and drew a random effect δ_i from $\delta_i|\phi \sim N(0, \phi^2)$. Finally, the observed data $\mathbf{x} = [x_{ij}]$ were drawn from (10).

3.2 Model 1 – with random effects

For the fixed effects we chose Zellner's g-prior (Smith and Kohn, 1996), namely $\boldsymbol{\beta}|\sigma \sim N(\mathbf{0}, g\sigma^2(\mathbf{Z}^T\mathbf{Z})^{-1})$. In our application we chose $g = n$, known as the unit information prior (Kohn et al., 2001). For the variance parameters we used inverse gamma priors: $\sigma^2 \sim IG(a_\sigma, b_\sigma)$ and $\phi^2 \sim IG(a_\phi, b_\phi)$, setting these parameters equal to 1 in our implementation. These conjugate priors allow a Gibbs sampling algorithm to sample from the posterior distribution. The full conditional distributions are given by,

$$\boldsymbol{\beta}, \sigma^2|\mathbf{x}, \boldsymbol{\delta} \sim NIG(\mathbf{m}^*, \mathbf{V}^*, a^*, b^*), \tag{11}$$

$$\mathbf{m}^* = \frac{g}{1+g}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T(\mathbf{x} - \mathbf{W}\boldsymbol{\delta}),$$

$$\mathbf{V}^* = \frac{g}{1+g}(\mathbf{Z}^T\mathbf{Z})^{-1},$$

$$a^* = a_\sigma + \frac{n}{2}, \tag{12}$$

$$b^* = b_\sigma + \frac{1}{2}(\mathbf{x} - \mathbf{W}\boldsymbol{\delta})^T\left(\mathbf{I}_n - \frac{g}{1+g}\mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\right)(\mathbf{x} - \mathbf{W}\boldsymbol{\delta}), \tag{13}$$

$$\phi|\boldsymbol{\delta} \sim IG(a_\phi + \frac{m}{2}, b_\phi + \frac{1}{2}\boldsymbol{\delta}^T\boldsymbol{\delta}), \tag{14}$$

$$\delta_i|\mathbf{x}, \boldsymbol{\beta}, \sigma, \phi \sim N\left(\frac{1}{\kappa_i}\sum_{j=1}^{n_i} x_{ij} - \mathbf{z}_{ij}^T\boldsymbol{\beta}, \kappa_i^{-1}\right), \tag{15}$$

$$\kappa_i = \frac{1}{\phi^2} + \frac{n_i}{\sigma^2}.$$

After a burn-in of 1000 iterations, we drew 10000 samples from the MCMC. To demonstrate the increased efficiency provided by making use of the approach to handle missing data described in Section 2, we considered two importance sampling estimators.

Full posterior importance sampling

In the full posterior importance sampler, we estimate the mean and covariance matrix for the full parameter vector $\boldsymbol{\theta}$ including the random effects, giving $m+5$ parameters in total. We then used these as the centre and scale matrix for multivariate t-distributed

proposal distribution with 5 degrees of freedom, with density $q_1(\boldsymbol{\theta})$. We drew $N = 1000$ samples from this proposal, denoting them by $\{\boldsymbol{\theta}_i\}_{i=1}^N$. The importance sampling estimator is then $\widehat{P}_{q_1}$ given in (4).

Marginal posterior importance sampling

In the second importance sampling approach we make use of missing data technique described in Section 2. We calculate the marginal mean and covariance matrix for the (restricted) parameter vector $\boldsymbol{\psi} = (\boldsymbol{\beta}, \sigma, \phi)$. Again we use these as the centre and scale matrix for a multivariate t-distributed proposal with 5 degrees of freedom, but this time it has just 5 dimensions. For each $\boldsymbol{\psi}_i$ that is drawn from the proposal $q_2(\boldsymbol{\psi})$, we sample the random effects $\boldsymbol{\delta}_i$ from their full conditional distribution in (15). The importance proposal is therefore given by $q_2(\boldsymbol{\psi})\pi(\boldsymbol{\delta}|\mathbf{x}, \boldsymbol{\psi})$.

Note that in both of these estimators we have chosen to parameterise our proposal distribution in terms of the standard deviations (σ and ϕ) rather than the variances in order to lighten the tails. However since the priors are written in terms of the variance parameters we must multiply the proposal densities $q_1(\boldsymbol{\theta})$ and $q_2(\boldsymbol{\psi})$ by the Jacobian $\sigma\phi/4$ to obtain the correct marginal likelihood estimator.

3.3 Model 0 – without random effects

The model with no random effects is just a linear model, given by $x_{ij} = \mathbf{z}_{ij}^T\boldsymbol{\beta} + \epsilon_{ij}$. We assume the same conjugate priors for $\boldsymbol{\beta}$ and σ^2 described in Section 3.2 and in this case the marginal likelihood may be calculated analytically, namely,

$$\pi(\mathbf{x}) = \frac{b_\sigma^{a_\sigma}\Gamma(a^*)}{(2\pi)^{n/2}\Gamma(a_\sigma)}(b^*)^{-a^*},$$

where a^* and b^* are given by (12) and (13) evaluated at $\boldsymbol{\delta} = \mathbf{0}$. For comparison purposes, we also use our importance sampling approach based on 10000 samples drawn directly from the joint posterior for $\boldsymbol{\beta}$ and σ. The importance proposal q was again based on a t-distribution with 5 degrees of freedom.

3.4 Results

Figure 1 shows the variation in 50 Monte Carlo replicates of each importance sampling estimator, based on 1000 samples. The importance sampling estimates for the linear model fall close to the true log marginal likelihood value, indicated by a dashed vertical line. For the linear mixed model treating the random effects as missing data and drawing them from their full conditional distribution greatly reduces the variance of the importance sampling estimator. The Monte Carlo standard errors for the linear model and linear mixed model with missing data were 0.0123 and 0.0171, compared with 0.106 for the linear mixed model with an importance proposal for the full parameter vector.

When we increase the number of clusters m from 50 to 500 (Figure 2 of the supplementary material (Touloupou et al., 2017)) the Monte Carlo standard error for the

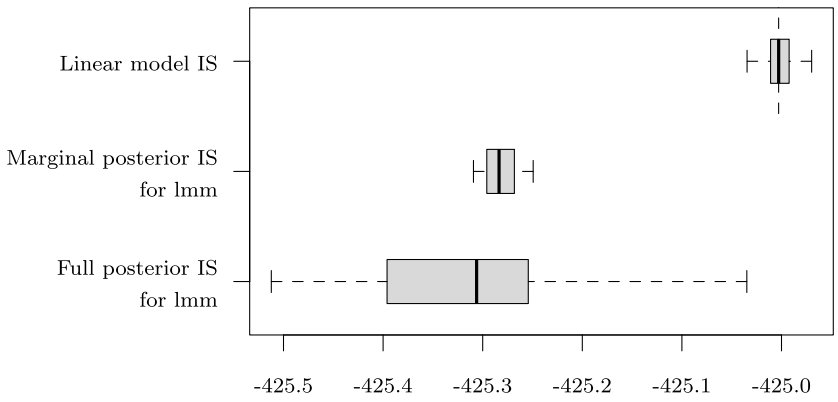


Figure 1: Variation of the log marginal likelihood estimates for the linear model and linear mixed model (lmm) over 50 replicates. For the lmm, in the full posterior approach the whole parameter vector (including random effects) was approximated in the importance proposal; in the marginal posterior approach the random effects were left out, and drawn from their full conditional distribution. A dashed vertical line indicates the true log marginal likelihood for the linear model.

marginal posterior approach was 0.015, showing no increase due to the increase in missing data. For comparison, the Monte Carlo standard error for the full parameter importance sampling approach increased to 0.825 for $m = 500$.

The precision of the marginal likelihood estimator is important when it comes to estimating the Bayes factor. With $m = 50$ clusters the 50 Monte Carlo estimates of Bayes factor in favour of the linear model (from the missing data approach) fall between 1.279 and 1.359 for the marginal posterior importance sampling estimator. When the full posterior importance sampling estimator is used the 50 Bayes factors fall between 1.032 and 1.664, and it's not hard to envisage situations in which this loss of precision could lead to an incorrect conclusion.

4 Final outcome epidemic data

In this Section, we look at applying the methodology developed in Section 2 to final outcome household epidemic data. Specifically, we assume that the data consist of the number of individuals infected during the course of an epidemic in a number of households of various sizes. We follow Addy et al. (1991) and Neal and Kypraios (2015) in assuming that the epidemics in each of the households are independent with each member of a household having probability p_G of being infected globally from the community at large and the within household epidemic spread emanating from the individuals infected globally. Within each household the disease dynamics are assumed to follow a homogeneously mixing, generalised stochastic epidemic model, where infectious individuals have independent and identically distributed infectious periods according to an

arbitrary, but specified, non-negative probability distribution Q with $E[Q] = 1$ and during their infectious period make contact with a given susceptible in their household at the points of a homogeneous Poisson point process with rate λ_L. The special case where $Q \equiv 1$ has the same final outcomes in terms of those infected as a household Reed-Frost model within household infection probability $p_L = 1 - \exp(-\lambda_L)$. Throughout we assign an Exp(1) prior to λ_L which corresponds to a $U(0,1)$ prior on p_L and also a $U(0,1)$ prior on p_G.

For the household Reed-Frost epidemic it is trivial to adapt the approach of Neal and Kypraios (2015), Section 3.3 to compute the marginal likelihood exactly. Details of how this can be done are given in the supplementary material (Touloupou et al., 2017). Therefore we are able to compare our estimation of the marginal likelihood with the exact marginal likelihood. The exact computation of the marginal likelihood grows exponentially in the total number of households whilst the MCMC algorithm for estimating the parameters and the algorithm for estimating the marginal likelihood have essentially constant computational cost for a given maximum household size. Therefore the exact computation of the marginal likelihood is only practical for data containing a small number of households and we apply it to the influenza data sets from Seattle, reported in Fox and Hall (1980), which contain approximately 90 households each.

Exact computation of the marginal likelihood is not possible for general Q. Therefore we use our approach to compare three different choices of infectious period $Q \equiv 1$, $Q \sim \text{Gamma}(2,2)$ and $Q \sim \text{Exp}(1)$ to study which infectious period distribution is most applicable for a given epidemic data set. This mimics analysis carried out in Addy et al. (1991) in a maximum likelihood framework where two infectious periods, a constant and a gamma with shape parameter 2, were compared for a combined data set of two influenza outbreaks in Tecumseh, Michigan, see Monto et al. (1985).

Let $\mathbf{x}$ denote the observed epidemic data. The recursive equations given in Ball et al. (1997), (3.12), can be used to compute P_k^h $(h = 1, 2, \ldots; k = 0, 1, \ldots, h)$, the probability of observing k individuals out of h being infected in a household of size h. Therefore it is straightforward to compute $\pi(\mathbf{x}|\lambda_L, p_G)$ or $\pi(\mathbf{x}|p_L, p_G)$ for the Reed-Frost model. Consequently, it is trivial to construct a random walk Metropolis algorithm to sample from $\pi(\lambda_L, p_G|\mathbf{x})$ (or $\pi(p_L, p_G|\mathbf{x})$ for the Reed-Frost model) and to estimate the marginal likelihood using samples from a proposal density $q(\lambda_L, p_G)$, which is a multivariate t distribution with 10 degrees of freedom and mean and covariance matrix obtained from the MCMC samples.

We applied the algorithm for estimating the marginal likelihood to the two Seattle data sets and the Tecumseh data set for each of the three infectious period distributions. In all cases we ran the MCMC algorithm for 11000 iterations discarding the first 1000 iterations as burn-in and then used 1000 samples to estimate the marginal likelihood. For the Seattle influenza A data set with maximum household size of 3 it took approximately 2.4 seconds to compute each marginal likelihood in R on a desktop PC with Intel i5 processor. For the Seattle influenza B data set and the Tecumseh data set with maximum household size of 5 it took approximately 5 seconds to compute each marginal likelihood. The log marginal likelihoods are given in Table 1, they show that is little information in the data to choose between different Q, agreeing with findings in Addy et al. (1991)

and Ball et al. (1997). Interestingly $Q \sim \text{Exp}(1)$ is preferred for the Seattle influenza A and Tecumseh data sets, whereas $Q \equiv 1$ is Seattle influenza B data set. For the two Seattle data sets we also computed the exact log marginal likelihoods, -15.08 and -24.98, for the household Reed-Frost model applied to the Seattle influenza A and influenza B data sets, respectively. For the Seattle data sets we repeated the estimation of the log marginal likelihood 100 times for the Reed-Frost model to obtain Monte Carlo standard errors for the estimated log marginal likelihoods of 0.0062 and 0.0073 for the Seattle influenza A and Seattle influenza B data sets, respectively, demonstrating good agreement between the estimated and exact log marginal likelihoods. The calculations of the exact marginal likelihood took approximately 0.5 and 14 seconds for the Seattle influenza A and influenza B data sets, respectively. The code for computing the exact marginal likelihood could not be applied to the combined Tecumseh data set, or even the separate Tecumseh data sets (see, for example Clancy and O'Neill, 2007) since enumerating over all possible augmented data states exceeded R's memory allocation. This problem could be circumvented to some extent using sufficient statistics as in Neal and Kypraios (2015) but the current approach offers a simple and fast alternative.

Data Set	$Q \sim \text{Exp}(1)$	$Q \sim \text{Gamma}(2,2)$	$Q \equiv 1$
Seattle A	-14.69	-14.86	-15.08
Seattle B	-25.27	-25.11	-24.99
Tecumseh	-45.58	-45.59	-45.87

Table 1: Estimated log marginal likelihood for the three influenza data sets using the three choices of infectious period distribution; $Q \sim \text{Exp}(1)$, $Q \sim \text{Gamma}(2,2)$ and $Q \equiv 1$.

5 Longitudinal epidemic model

5.1 Introduction

In this Section, we explore the application of the methodology developed in Section 2 to a scenario where $\pi(\mathbf{x}|\boldsymbol{\theta})$ is not readily available, and data augmentation is required both with in the MCMC algorithm and estimation of the marginal likelihood. The example used is based on a longitudinal household study of preschool children under 3 years old and all household members was conducted in the United Kingdom from October 2001 to July 2002 (Hussain et al., 2005). The size of the families varied from 2 to 7, although in most there were 3 or 4 members. All family members were examined for *Streptococcus pneumoniae* carriage (Pnc) using nasopharyngeal swabs once every 4 weeks over a 10-month period. The carriage status of each individual was recorded at each occasion as 1, if a carrier or 0, if a non-carrier.

Following Melegaro et al. (2004), we consider an Susceptible-Infected-Susceptible (SIS) epidemic model for the transmission of Pnc within a household. At any given time, an individual is assumed to be in either the susceptible non-carrier state 0, or the infectious carrier state 1. The population is divided into two age groups, children under 5 years old and everyone else greater than 5 years (whom for brevity we refer to

as 'adults'), denoted by $i = 1, 2$, respectively. Let $I_1(t)$ and $I_2(t)$ denote the numbers of carrier children and carrier adults in the household at time t. The transition between state 0 and 1 is referred to as an infection and the reverse transition is referred to as clearance. The transition probabilities between states in a short time interval δt are defined for an individual in the age group i:

$$P(\text{Infection in } (t, t+\delta t]) = 1 - \exp\left\{-\left(k_i + \frac{\beta_{1i} I_1(t) + \beta_{2i} I_2(t)}{(z-1)^w}\right)\delta t\right\}, \tag{16}$$

$$P(\text{Clearance in } (t, t+\delta t]) = 1 - \exp(-\mu_i \cdot \delta t), \tag{17}$$

where μ_i and k_i are the clearance and the community acquisition rates respectively for age group i and z is the household size. The rate β_{ij} is the transmission rate from an infected individual in age group i to an uninfected individual in age group j. The term $(z-1)^w$ in (16) represents a density correction factor, where w corresponds to the level of density dependence and $(z-1)$ is the number of other family members in a household size z. For example, $w = 1$ represents frequency dependent transmission, where the average number of contacts is equal for each individual in the population. Finally, the probability of infection at the initial swab is assumed to be π_i for age group i. We refer to this model as $\mathcal{M}_1$.

Given the dependence of the carriage status of all individuals in a household, the within household carriage dynamics in a household of size z can be modelled as a discrete time Markov chain with 2^z states (all possible binary vectors of infectious statuses of the z individuals). The presence of unobserved events, that may have occurred in between swabbing intervals, has been discussed previously (Auranen et al., 2000), and must be considered in setting up the model. The approach adopted in this paper to overcome this issue is to use Bayesian data augmentation methods. Model fitting is performed within a Bayesian framework using an MCMC algorithm, imputing the unobserved carriage states in each household. Let $O_j \subseteq \{1, 2, \ldots, T\}$ denote the set of prescheduled observation times of household $j = 1, 2, \ldots, J$, and let $U_j = \{1, 2, \ldots, T\} \setminus O_j$ denote the unobserved times. Let $\mathbf{x}_{j,t}$ be the binary vector of carriage states for individuals in household j at observation time t. The observed longitudinal data $\mathbf{X} = [\mathbf{x}_{j,t}]_{t \in O_j; j=1,\ldots,J}$ consists of the household carriage statuses $\mathbf{x}_{j,t}$ at the observation times. Similarly let $\mathbf{y}_{j,t}$ be the corresponding latent carriage status of household j at time $t \in U_j$, and form the corresponding missing data matrix $\mathbf{Y} = [\mathbf{y}_{j,t}]_{t \in U_j; j=1,\ldots,J}$. Let $\boldsymbol{\theta}$ denote the vector of model parameters, including the rates of acquiring and clearing carriage, the density correction w and the initial probabilities of carriage.

The remainder of this Section is structured as follows. In Sections 5.2 and 5.3, we introduce the MCMC algorithm and importance sampling algorithms, respectively, required to implement our approach. In particular, we introduce the Forward Filtering Backward Sampling algorithm (Carter and Kohn, 1994) to assist with dealing with the augmented data $\mathbf{y}$. In Sections 5.4 and 5.5 simulated data (where the true model is known) are used to illustrate the implementation, performance and applicability of the proposed method and its comparative performance against a range of alternatives. We demonstrate that for a fixed computational cost our approach performs at least as well as existing methods, and with the exception of bridge sampling (Meng and Wong, 1996), performs considerably better.

5.2 Markov chain Monte Carlo algorithm

In the Bayesian approach, the missing data is represented as a nuisance parameter and inferred from the observed data like any other parameter. The joint posterior density of the latent carriage states $\mathbf{y}$, and the model parameters $\boldsymbol{\theta}$ can be factorized as:

$$\begin{aligned}\pi(\mathbf{y}, \boldsymbol{\theta} \mid \mathbf{x}) &\propto P(\mathbf{y}, \mathbf{x} \mid \boldsymbol{\theta})\, \pi(\boldsymbol{\theta}) \\ &= \pi(\boldsymbol{\theta}) \prod_{j=1}^{J} \prod_{t=1}^{T} P(\mathbf{z}_{j,t} | \mathbf{z}_{j,t-1}, \boldsymbol{\theta}),\end{aligned}$$

where $\mathbf{z}_{j,t}$ equals $\mathbf{x}_{j,t}$ if $t \in O_j$; $\mathbf{y}_{j,t}$ if $t \in U_j$ and $\emptyset$ if $t = 0$. This factorization is based on the assumption that conditionally on the model parameters, the carriage process is assumed to be independent across households.

In order to simulate from the posterior distribution, we construct an MCMC algorithm that employs both Gibbs and Metropolis-Hastings updates. The main emphasis is on sampling the unobserved carriage process $\mathbf{y}$, which we do using a Gibbs step via the Forward Filtering Backward Sampling (FFBS) algorithm (Carter and Kohn, 1994). In the first part of this algorithm, recursive filtering equations (Anderson and Moore, 1979) are used to calculate $P(\mathbf{y}_{j,t} \mid \mathbf{z}_{j,t+1}, \mathbf{x}_{j,O_j \cap \{1:t\}}, \boldsymbol{\theta})$ for each $t \in U_j$ working forwards in time. The second part then works backwards through time, simulating $\mathbf{y}_{j,t}$ from these conditionals, starting with $t = \max(U_j)$ and ending with $t = \min(U_j)$. The model parameters π_1 and π_2 are updated using Gibbs updates and the remaining parameters are updated jointly using an adaptive Metropolis-Hastings random walk proposal (Roberts and Rosenthal, 2009).

5.3 Marginal likelihood estimation via importance sampling

The availability of the full conditional distribution of the missing data $P(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$ from the FFBS algorithm allows the missing data component $\mathbf{y}$ to be updated using a Gibb's step in the MCMC algorithm. This full conditional can be exploited further in the estimation of the marginal likelihood. We require $P(\mathbf{x}|\boldsymbol{\theta})$ in order to form the importance sampling estimator in (4). Using Bayes' Theorem we can rewrite this as

$$P(\mathbf{x}|\boldsymbol{\theta}) = \frac{P(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})\, P(\mathbf{y}|\boldsymbol{\theta})}{P(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})} = \frac{P(\mathbf{x}, \mathbf{y}|\boldsymbol{\theta})}{P(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})}, \tag{18}$$

for any $\mathbf{y}$ such that $P(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) > 0$. Therefore evaluation of $P(\mathbf{x}|\boldsymbol{\theta})$ at the point $\boldsymbol{\theta}$ can be done by evaluating the right-hand-side of (18) with any suitable $\mathbf{y}$. A suitable $\mathbf{y}$ is guaranteed if it is sampled from the full conditional distribution $\mathbf{y}|(\mathbf{x}, \boldsymbol{\theta})$.

Our approach proceeds as follows. In step 1 we use MCMC to obtain samples from the joint posterior of $\boldsymbol{\theta}$ and $\mathbf{y}$. In step 2 we fit a multivariate normal distribution to the posterior samples for $\boldsymbol{\theta}$ only, and use it to construct a normalised proposal density $q(\boldsymbol{\theta})$. In step 3, we obtain N samples from $q(\boldsymbol{\theta})$ and for each sample $\boldsymbol{\theta}_i$ we obtain a corresponding sample for the missing data $\mathbf{y}_i$ using the Forward Filtering Backward

Sampling algorithm. We then use these samples to calculate the importance sampling estimator of the marginal likelihood:

$$\widehat{P}_q(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} \frac{P(\mathbf{x}, \mathbf{y}_i|\boldsymbol{\theta}_i)}{P(\mathbf{y}_i|\mathbf{x}, \boldsymbol{\theta}_i)} \frac{\pi(\boldsymbol{\theta}_i)}{q(\boldsymbol{\theta}_i)}. \tag{19}$$

The choice of $q(\boldsymbol{\theta})$ is important for the accuracy and computational efficiency of the importance sampling approach. As discussed earlier, we want $q(\boldsymbol{\theta})$ to be a good approximation of $\pi(\boldsymbol{\theta}|\mathbf{x})$ but with heavier tails to ensure that the variance of $\widehat{P}_q$ is small. We therefore investigate a range of proposals distributions based on a fitted multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ based on the MCMC output. These include drawing $\boldsymbol{\theta}$ from $IS_{N_j} : N(\boldsymbol{\mu}, j\boldsymbol{\Sigma})$ $(j = 1, 2, 3)$, a multivariate Normal distribution with different variances; $IS_{t_d} : t_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ $(d = 4, 6, 8)$, a multivariate Student's t distribution with d degrees of freedom, mean $\boldsymbol{\mu}$ and covariance matrix $\frac{d}{d-2}\boldsymbol{\Sigma}$ (if $d > 2$) and $IS_{\text{mix}} : q(\boldsymbol{\theta}) = 0.95 \times N(\boldsymbol{\theta}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) + 0.05 \times \pi(\boldsymbol{\theta})$ (mixture of a multivariate Normal density and the prior).

5.4 Marginal likelihood estimation

We consider the problem of estimating the marginal likelihood under the model introduced in Section 5.1, using the methods described above. These estimators were evaluated on synthetic data analogous to the real data in Melegaro et al. (2004). More specifically, the parameter values were based on the maximum likelihood estimates from the analysis of Pnc data; parameters were chosen to be $k_1 = 0.012$, $k_2 = 0.004$, $\beta_{11} = 0.047$, $\beta_{12} = 0.005$, $\beta_{21} = 0.106$, $\beta_{22} = 0.048$, $\mu_1 = 0.020$, $\mu_2 = 0.053$, $w = 1.184$, $\pi_1 = 0.425$ and $\pi_2 = 0.095$. We set the time-interval $\delta t = 7$. Only complete family transitions, where the infection state of all household members was known on two consecutive observations, were used previously (Melegaro et al., 2004; 51% of the full dataset). Although our approach could easily handle the missing data, for comparability we match the number of complete transitions by family size and number of adults to generate our data set; a total of 66 families comprising 260 individuals including 94 children under 5 years. The simulations were designed so that real and simulated datasets have the same sampling times. The hidden variable $\mathbf{y}$ consists of 1650 $\mathbf{y}_{j,t}$'s, comprising 6500 unobserved binary variables in total.

We compare the proposed importance sampling approach for estimating the marginal likelihood (based on the 7 proposal densities) with bridge sampling (Meng and Wong, 1996) (using the importance samples from IS_{mix}), harmonic mean (Newton and Raftery, 1994), Chib's method (Chib, 1995; Chib and Jeliazkov, 2001) and the power posteriors method (Friel and Pettitt, 2008). Details of the computation of these estimators are given in the supplementary material (Touloupou et al., 2017). To compare the different methods on a fair basis, we chose to dedicate equivalent amounts of computational effort for estimation of the log marginal likelihood, instead of fixing the total number of samples.

Implementation details are given as follows. The construction of the importance density was based on 25000 MCMC samples after a burn-in of 5000, obtained from the

MCMC sampler described in Section 5.2. These posterior samples were used to estimate the reference parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ for a multivariate Student's t or normal proposal density. The marginal likelihood estimate was then based on 25000 importance sampling draws from the obtained proposal density $q(\boldsymbol{\theta})$, using the estimator in (19). To produce the bridge sampling estimate, the 25000 samples from IS_{mix} were combined with 250 thinned samples from the MCMC. In order to apply Chib's methods, the same posterior samples were used for computing the high posterior density point. The log marginal was estimated by generating 22000 draws in each complete and reduced MCMC run, with the first 2000 draws removed as burn-in. Harmonic mean analysis was based on 50000 posterior samples, following a 3000 iteration burn-in. For the power posterior method, it was necessary to specify the temperature scheme and a pilot analysis (not counted in the computation cost) was used to choose 20 partitions on the unit interval. The MCMC sampler was run for 2650 iterations for each temperature in the descending series, omitting the first 650 as burn-in, finishing with 2650 samples at $t = 0$ (the prior).

Each procedure was repeated 50 times to provide an empirical Monte Carlo estimate of the variation in each approach. We also vary the total running time in order to investigate the effect of this on the accuracy of the marginal likelihood estimates, see Table 1 in the supplementary material. For each analysis method we used the same priors: Gamma(0.01,0.01) for the density factor w; Beta(1,1) for the initial probabilities of infection π_1 and π_2 and Gamma(1,1) for the remaining parameters.

Figure 2 shows the variability of the eleven marginal likelihood estimators. Except for the harmonic mean, all the methods appear to have produced consistent estimates of the marginal likelihood. Chib's method produced better estimates of the marginal likelihood than the power posterior method, which is more computationally expensive than the other methods and therefore uses a small number of MCMC samples at each temperature, leading to large uncertainty. However as seen in Figure 2, the bridge sampling and the importance sampling methods offer significant improvements in precision over the other methods. Moreover, increasing the number of samples N, led to a decrease in the Monte Carlo standard errors of order $\mathcal{O}(\sqrt{N})$, see Table 1 in the supplementary material, indicating that the variances of the corresponding estimators are finite.

The success of the importance sampling approach is not surprising since it explores the posterior distribution of parameters more efficiently than the other methods due to the independence of the samples drawn from the proposal density. More surprisingly we were unable to use the bridge sampling technique to improve substantially on the standard errors, which dropped from 0.0196 for IS_{mix} to 0.0179 for BS_{mix}. The bridge sampling estimator attempts to combine information from the MCMC and importance samples, however the optimal estimator is derived assuming that independent samples from the posterior were available, which we approached by applying a thinning of 100 to the samples. With low levels of thinning (results shown in supplementary material Figure 4) we found that bridge sampling actually increased the standard error of the marginal likelihood estimate.

On the basis of this example, the lowest variance importance sampling estimator was obtained using the proposal density IS_{mix} – a mixture of the prior and the

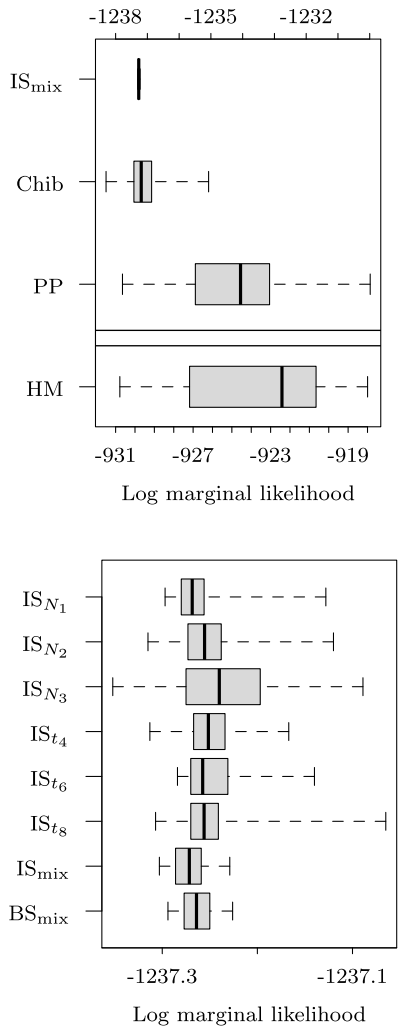


Figure 2: Top: Boxplots of the estimated log marginal likelihood for model $\mathcal{M}_1$ over 50 replicates for our importance sampling approach with the mixture proposal (IS_{mix}), Chib's method (Chib), power posteriors method (PP) and harmonic mean (HM) (note the different scales for the top and bottom plots). Bottom: Zoomed in boxplots of the estimated log marginal likelihood for model $\mathcal{M}_1$ over 50 replicates for each of our importance sampling approach ($IS_{N_1}, IS_{N_2}, IS_{N_3}, IS_{t_4}, IS_{t_6}, IS_{t_8}, IS_{mix}$) and bridge sampling ($BS_{mix}$).

normal fitted to the posterior samples. Therefore, in the next section we use this proposal density when estimating the log marginal likelihood via importance sampling.

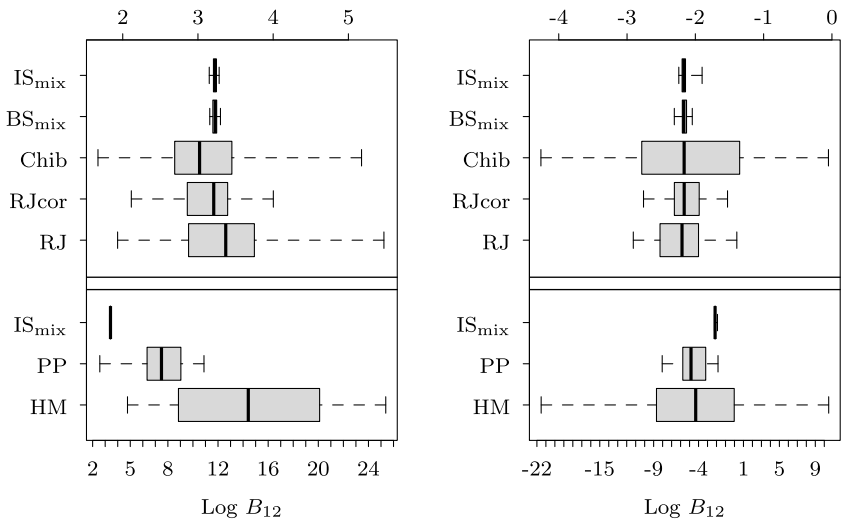
5.5 Model comparison

In this Section, we apply the marginal likelihood estimation approaches to the problem of Bayesian model choice. We focus on their ability to distinguish between biologically motivated hypotheses concerning the dynamics of Pnc transmission. In particular we compare their performance against the established technique of Reversible Jump Markov Chain Monte Carlo (RJMCMC) and then demonstrate that the importance sampling approach can solve problems that are extremely challenging with RJMCMC. We show that using our approach it is possible to answer the epidemiological important question of how household size is related to transmission with extended discussion given in the supplementary material.

Suppose that we wish to evaluate the evidence in favour of the community acquisition rates being equal for adults and children, in the hope of developing a more parsimonious model. We call the model described in Section 5.1, in which children have community acquisition rate k_1 and adults have rate k_2, model $\mathcal{M}_1$. The nested model, in which $k_1 = k_2$ is called $\mathcal{M}_2$. We generated realistic simulated datasets from each of these models and then used importance sampling, bridge sampling, Chib's method, power posteriors, the harmonic mean and reversible jump MCMC to estimate the Bayes factor in favour of $\mathcal{M}_1$, denoted by B_{12}. As before, we used approximately the same computational effort for each of these approaches. For $\mathcal{M}_1$ we assumed $k_1 = 0.012$ and $k_2 = 0.004$, whilst for $\mathcal{M}_2$ we assumed $k_1 = k_2 = 0.008$.

Details of the RJMCMC algorithm for selecting between models $\mathcal{M}_1$ and $\mathcal{M}_2$ are given in the supplementary material. We ran the RJMCMC chain with a 30000 burn-in followed by 76000 samples which ensured that similar computational effort was given to RJMCMC as to the other methods. When the evidence is strongly in favour of one model, the RJMCMC will not move between models very often and can provide poor estimates of the Bayes factor. A variant of the method, called RJMCMC corrected (RJcor), can tackle this issue by assigning higher prior probability to the model that is visited less often. This probability is estimated as $\pi(\mathcal{M}_m) = 1 - \widehat{\pi}(\mathcal{M}_m \mid \mathbf{x})$, where $\widehat{\pi}(\mathcal{M}_m \mid \mathbf{x})$ is obtained from a pilot run of RJMCMC with initial $\pi(\mathcal{M}_m) = 0.5$, for $m = 1, 2$. For RJcor we did 30000 pilot iterations and then another 76000 iterations, of which 30000 were discarded as a burn in.

Figure 3 provides a graphical representation of the variability in $\log(B_{12})$ over 50 repeats of each Monte Carlo approach. The plot highlights that the estimators based on importance sampling and bridge sampling were the most accurate in both scenarios. The left panel of Figure 3 gives results for data generated from $\mathcal{M}_1$. Importance sampling, bridge sampling, Chib and RJ methods lead to similar estimates, whereas power posterior and harmonic mean overestimated the log Bayes factor. Moreover, RJcor produced slightly more accurate estimates of the log Bayes factor than *vanilla* RJMCMC. All methods selected the correct model, with largest variation from the harmonic mean estimator. In the right panel of Figure 3, the results use data generated from model $\mathcal{M}_2$. Due to the huge variance in $\log(B_{12})$, the harmonic mean sometimes favoured the wrong model. Although the remaining methods correctly identified the true model, the importance and bridge sampling methods again produced the most precise estimates of the Bayes factor; the standard errors provided by the two methods are almost identical.



(a) Data simulated from model $\mathcal{M}_1$.

(b) Data simulated from model $\mathcal{M}_2$.

Figure 3: Variability of the log Bayes factor estimates based on 50 Monte Carlo repeats for the importance sampling method with mixture proposals (IS_{mix}), bridge sampling method with mixture proposals (BS_{mix}), Chib's method (Chib), reversible jump MCMC (RJ), corrected reversible jump MCMC (RJcor), power posteriors (PP) and harmonic mean (HM) methods (note the different scales for the top and bottom plots).

Figure 4 demonstrates the evolution of the log Bayes factor in favour of $\mathcal{M}_1$ as a function of computation time using data generated from $\mathcal{M}_1$. The importance sampling estimator (in blue) converges much more rapidly than the other estimators, showing very tight credible intervals. Chib's method (in green) and corrected RJMCMC (in red) appear to converge to the same value, but more slowly and have wider CIs. The power posterior method gradually approaches the consensus estimate, requiring significantly more samples to stabilize. The harmonic mean estimator was heavily unstable and also provided much wider credible intervals than the other methods.

In the supplementary material (Touloupou et al., 2017), further model comparison questions are considered and the strength of the importance sampling technique for answering these questions is further demonstrated. In particular, we consider heterogeneity in household transmission rates, density-dependence in within-household transmission and the amount of missing data.

6 Conclusions

In this paper we have introduced a simple three stage algorithm for efficiently estimating the marginal likelihood. The key components are an MCMC algorithm for obtaining samples from the posterior distribution, $\pi(\boldsymbol{\theta}|\mathbf{x})$, an approximating distribution $q(\boldsymbol{\theta})$ to

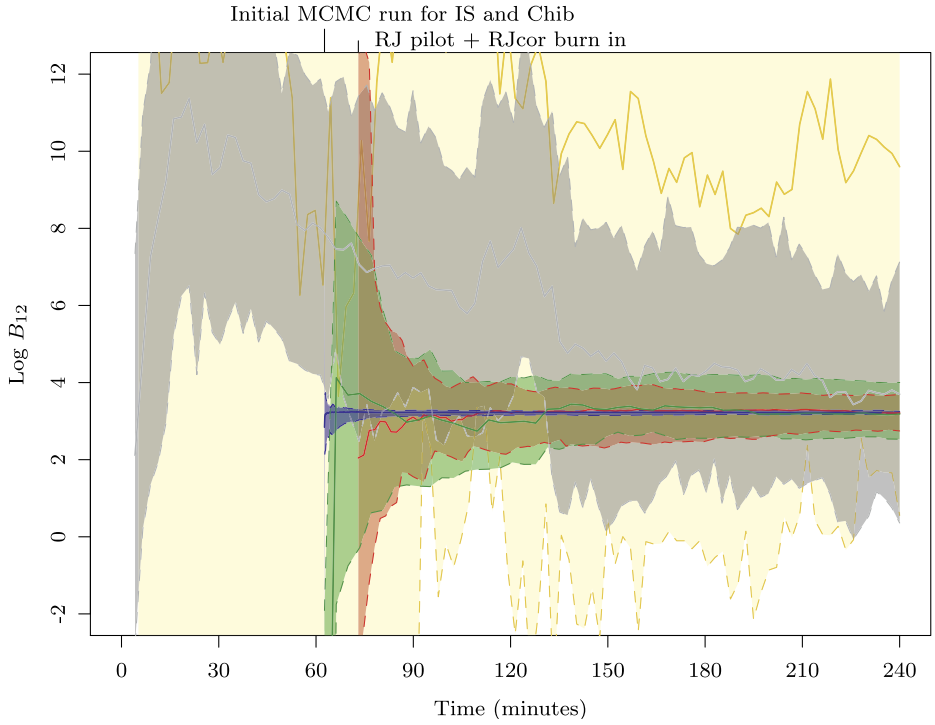


Figure 4: Evolution of log Bayes factor estimates in favour of model $\mathcal{M}_1$ as a function of computation time. The solid lines corresponds to the median and the shaded areas give the 95% credible intervals, estimated from 50 Monte Carlo replicates. Yellow represents the harmonic mean method, grey is for the power posterior, red and green correspond to RJMCMC corrected and Chib's methods respectively and blue represents the importance sampling approach with the mixture proposals.

sample from and an effective estimate of the likelihood $\pi(\mathbf{x}|\boldsymbol{\theta})$. The first observation is whilst an MCMC algorithm will often be relatively straightforward to construct, alternative methods for sampling from the posterior distribution could be equally considered. Moreover, it is not important if a sample from an approximate posterior distribution (for example, Monte Carlo within Metropolis; O'Neill et al., 2000) is used since all that is required for computation of the marginal likelihood is to be able to make a reasonable choice of $q(\cdot)$. The key limitation to using this approach is effective estimation of the likelihood $\pi(\mathbf{x}|\boldsymbol{\theta})$ in cases where it is not analytically tractable. In Section 5 of this paper the temporal nature of the data allowed the FFBS algorithm to be utilised to compute $\pi(\mathbf{x}|\boldsymbol{\theta})$ and more generally filtering methods are a promising avenue of research to explore in the estimation of $\pi(\mathbf{x}|\boldsymbol{\theta})$. The importance sampling and the associated estimation of the likelihood is trivially parallelisable which can be utilised to speed up implementation. Finally, in cases where the likelihood can easily be computed the algorithm becomes a simple add-on to MCMC to compute the marginal likelihood.

Supplementary Material

Supplementary material: Efficient model comparison techniques for models requiring large scale data augmentation (DOI: 10.1214/17-BA1057SUPP; .pdf).

References

Addy, C. L., Longini, I. M. and Haber, M. (1991). A generalized stochastic model for the analysis of infectious disease final size data. *Biometrics* **47**, 961–974. 445, 446

Anderson, B. D. O. and Moore, J. B. (1979). Optimal filtering. *Englewood Cliffs, New Jersey: Prentice Hall*. 449

Auranen K., Arjas E., Leino T. and Takala A. K. (2000). Transmission of pneummonococcal carriage in families: A latent Markov process model for binary longitudinal data. *Journal of the American Statistical Association* **95**, 1044–1053. MR1821713. doi: https://doi.org/10.2307/2669741. 448

Ball, F. G., Mollison, D. and Scalia-Tomba, G. (1997). Epidemics with two levels of mixing. *The Annals of Applied Probability* **7**, 46–89. MR1428749. doi: https://doi.org/10.1214/aoap/1034625252. 446, 447

Carter, C. K. and Kohn, R. (1994). On Gibbs sampling for state space models. *Biometrika* **81**, 541–553. MR1311096. doi: https://doi.org/10.1093/biomet/81.3.541. 441, 448, 449

Chen, M.-H. (2005). Computing marginal likelihoods from a single MCMC output. *Statistica Neerlandica* **59**, 16–29. MR2137379. doi: https://doi.org/10.1111/j.1467-9574.2005.00276.x. 438

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association* **90**, 1313–1321. MR1379473. 438, 450

Chib, S. and Jeliazkov, I. (2001). Marginal likelihood from the Metropolis-Hastings output. *Journal of the American Statistical Association* **96**, 270–281. MR1952737. doi: https://doi.org/10.1198/016214501750332848. 438, 450

Clancy, D. and O'Neill, P. (2007). Exact Bayesian inference and model selection for stochastic models of epidemics among a community of households. *Scandinavian Journal of Statistics* **34**, 259–274. MR2346639. doi: https://doi.org/10.1111/j.1467-9469.2006.00522.x. 447

Clyde, M. A., Berger, J. O., Bullard, F., Ford, E. B., Jefferys, W. H., Luo, R., Paulo, R. and Loredo, T. (2007). Current challenges in Bayesian model choice. In *Statistical Challenges in Modern Astronomy IV ASP Conference Series, Vol. 371, proceedings of the conference held 12–15 June 2006 at Pennsylvania State University, in University Park, Pennsylvania, USA. Edited by G. Jogesh Babu and Eric D. Feigelson.* p. 224. 439

Doucet, A. and Johansen, A. M. (2011). A tutorial on particle filtering and smoothing: fifteen years later. In *Handbook of Nonlinear Filtering* (eds D. Crisan and B. Rozovsky). Cambridge: Cambridge University Press. pp. 656–704. MR2884612. 440

Fong, Y., Rue, H. and Wakefield, J. (2010). Bayesian inference for generalized linear mixed models. *Biostatistics* **11**, 397–412. 442

Fox, J. P. and Hall, C. E. (1980). *Viruses in families.* PSG Publishing, Littleton, MA. 446

Friel, N. and Pettitt, A. N. (2008). Marginal likelihood estimation via power posteriors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**, 589–607. MR2420416. doi: https://doi.org/10.1111/j.1467-9868.2007.00650.x. 438, 450

Gelfand, A. E. and Dey, D. K. (1994). Bayesian model choice. Exact and asymptotic calculations. *Journal of the Royal Statistical Society. Series B (Methodological)* **56**, 501–514. MR1278223. 438, 439

Gordon, N. J., Salmond, D. J., and Smith, A. F. M. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *Radar and Signal Processing, IEE Proceedings F* **140**, 107–113. 441

Green P.J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732. MR1380810. doi: https://doi.org/10.1093/biomet/82.4.711. 438, 442

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109. MR3363437. doi: https://doi.org/10.1093/biomet/57.1.97. 438

Hesterberg, T. (1995). Weighted average importance sampling and defense mixture distributions. *Technometrics* **37**, 185–194. 439

Hussain M., Melegaro A., Pebody R. G., George R., Edmunds W. J., Talukdar R., Martin S. A., Efstratiou A. and Miller E. (2005). A longitudinal household study of *Streptococcus pneumoniae* nasopharyngeal carriage in a UK setting. *Epidemiology and Infection* **5**, 891–898. 447

Karagiannis, G., and Andrieu, C. (2013). Annealed importance sampling reversible jump MCMC algorithms. *Journal of Computational and Graphical Statistics* **22**, 623–648. MR3173734. doi: https://doi.org/10.1080/10618600.2013.805651. 438

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association* **90**, 773–795. MR3363402. doi: https://doi.org/10.1080/01621459.1995.10476572. 438

Kohn, R., Smith, M., Chan D. (2001). Nonparametric regression using linear combinations of basis functions. *Statistics and Computing* **11**, 313–322. MR1863502. doi: https://doi.org/10.1023/A:1011916902934. 443

Melegaro, A., Gay, N., and Medley, G. (2004). Estimating the transmission parameters of pneumococcal carriage in households. *Epidemiology and Infection* **132**, 433–441. 439, 447, 450

Metropolis N., Rosenbluth A., Rosenbluth M., Teller A. and Teller E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics* **21**, 1087–1092. 438

Meng, X.-L. and Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica* **6**, 831–860. MR1422406. 438, 448, 450

Monto, A. S., Koopman, J. S. and Longini, I. M. (1985). Tecumseh study of illness. XIII. Influenza infection and disease, 1976–1981. *American Journal of Epidemiology* **121**, 811–822. 446

Neal, P. and Kypraios, T. (2015). Exact Bayesian inference via data augmentation. *Statistics and Computing* **25**, 333–347 MR3306710. doi: https://doi.org/10.1007/s11222-013-9435-z. 445, 446, 447

Neal, P. and Subba Rao, T. (2007). MCMC for integer-valued ARMA processes. *Journal of Time Series Analysis* **28**, 92–110. MR2332852. doi: https://doi.org/10.1111/j.1467-9892.2006.00500.x. 442

Newton, M. A. and Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society. Series B (Methodological)* **56**, 3–48. MR1257793. 450

O'Neill, P. D., Balding, D. J., Becker, N. G., Eerola, M. and Mollison, D. (2000). Analyses of infectious disease data from household outbreaks by Markov Chain Monte Carlo methods. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **49**, 517–542. MR1824557. doi: https://doi.org/10.1111/1467-9876.00210. 455

Ripley, B. D. (1987). *Stochastic Simulation*. Wiley & Sons. MR0875224. doi: https://doi.org/10.1002/9780470316726. 439

Roberts, G. O. and Rosenthal, J. S. (2009). Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics* **18**, 349–367 MR2749836. doi: https://doi.org/10.1198/jcgs.2009.06134. 449

Skilling, J. (2004). Nested sampling. *AIP Conference Proceedings* **735**, 395–405. MR2266273. doi: https://doi.org/10.1063/1.1835238. 438

Smith, M. and Kohn, R. (1996). Nonparametric regression using Bayesian variable selection. *Journal of Econometrics* **75**, 317–344. 443

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association* **81**, 82–86. 590–599. MR0830567. 439

Touloupou, P., Alzahrani, N., Neal, P., Spencer, S. E. F., and McKinley, T. J. (2017). Supplementary material: Efficient model comparison techniques for models

requiring large scale data augmentation. *Bayesian Analysis*. doi: https://doi.org/10.1214/17-BA1057SUPP. 444, 446, 450, 454

Zeger, S. (1988). A regression model for time series of counts. *Biometrika* **75**, 621–629. MR0995107. doi: https://doi.org/10.1093/biomet/75.4.621. 441, 442

Zhou, Y., Johansen, A. M. and Aston, J. A. D. (2015). Towards automatic model comparison: An adaptive sequential Monte Carlo approach. *Journal of Computational and Graphical Statistics* **25**, 701–726. MR3533634. doi: https://doi.org/10.1080/10618600.2015.1060885. 438

Acknowledgments

PT was supported by a University of Warwick PhD scholarship. NA was supported by a PhD scholarship from the Saudi Arabian Government. PN, SS and TM would like to thank the organisers of the Design and Analysis of Infectious Disease Studies workshop at Oberwolfach (November 2013), where many helpful discussions took place.

We thank an associate editor and a referee for insightful comments which have helped in revising the paper.