Stat Comput
DOI 10.1007/s11222-013-9435-z

Exact Bayesian inference via data augmentation

Peter Neal · Theodore Kypraios

Received: 13 August 2012 / Accepted: 6 November 2013

Abstract Data augmentation is a common tool in Bayesian statistics, especially in the application of MCMC. Data augmentation is used where direct computation of the posterior density, $\pi(\boldsymbol{\theta}|\mathbf{x})$, of the parameters $\boldsymbol{\theta}$, given the observed data $\mathbf{x}$, is not possible. We show that for a range of problems, it is possible to augment the data by $\mathbf{y}$, such that, $\pi(\boldsymbol{\theta}|\mathbf{x},\mathbf{y})$ is known, and $\pi(\mathbf{y}|\mathbf{x})$ can easily be computed. In particular, $\pi(\mathbf{y}|\mathbf{x})$ is obtained by *collapsing* $\pi(\mathbf{y},\boldsymbol{\theta}|\mathbf{x})$ through integrating out $\boldsymbol{\theta}$. This allows the exact computation of $\pi(\boldsymbol{\theta}|\mathbf{x})$ as a mixture distribution without recourse to approximating methods such as MCMC. Useful byproducts of the exact posterior distribution are the marginal likelihood of the model and the exact predictive distribution.

Keywords Bayesian statistics · Data augmentation · Multinomial distribution · Reed-Frost epidemic · Integer valued autoregressive process

1 Introduction

A key aim of parametric Bayesian statistics is, given a model $\mathcal{M}$, with parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_d)$ and observed data $\mathbf{x}$, to derive the posterior distribution of $\boldsymbol{\theta}$, $\pi(\boldsymbol{\theta}|\mathbf{x})$. From $\pi(\boldsymbol{\theta}|\mathbf{x})$, it is then possible to obtain key summary statistics, such as the marginal posterior mean, $\mathbb{E}[\theta_1|\mathbf{x}]$, and variance, $var(\theta_1|\mathbf{x})$, of the parameter θ_1, or problem specific quantities, such as $\mathbb{E}[1_{\{\boldsymbol{\theta}\in A\}}|\mathbf{x}]$, for some $A \subset \mathbb{R}^d$. By Bayes' Theorem,

$$\pi(\boldsymbol{\theta}|\mathbf{x}) = \frac{\pi(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\mathbf{x})} \propto \pi(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}). \tag{1.1}$$

There are two potentially problematic components in (1.1). Firstly, computation of $\pi(\mathbf{x})$ is rarely possible. This is often circumvented by recourse to Markov chain Monte Carlo methods. Secondly, it is common that the likelihood function, $L(\boldsymbol{\theta}|\mathbf{x}) = \pi(\mathbf{x}|\boldsymbol{\theta})$, is not in a convenient form for statistical analysis. For the analysis of many statistical problems, both Bayesian and frequentest, data augmentation (Dempster et al. 1977; Gelfand and Smith 1990) has been used to assist in evaluating $\pi(\boldsymbol{\theta}|\mathbf{x})$ or computing the maximum likelihood estimate of $\boldsymbol{\theta}$. That is, additional, unobserved data, $\mathbf{y}$, is imputed so that $\pi(\mathbf{x},\mathbf{y}|\boldsymbol{\theta}) = L(\boldsymbol{\theta}|\mathbf{x},\mathbf{y})$ is of a convenient form for analysis. In a Bayesian framework $\mathbf{y}$ is often chosen so that the (conditional) posterior distribution of $\pi(\boldsymbol{\theta}|\mathbf{x},\mathbf{y})$ is known.

There are a number of algorithms such as the EM algorithm (Dempster et al. 1977) and the Metropolis-Hastings (MCMC) algorithm (Metropolis et al. 1953; Hastings 1970), which exploit $\pi(\boldsymbol{\theta}|\mathbf{x},\mathbf{y})$ and iterate between the following two steps.

1. Update $\boldsymbol{\theta}$ given $\mathbf{x}$ and $\mathbf{y}$. *i.e.* Use $\pi(\boldsymbol{\theta}|\mathbf{x},\mathbf{y})$.
2. Update $\mathbf{y}$ given $\mathbf{x}$ and $\boldsymbol{\theta}$. *i.e.* Use $\pi(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})$.

The above algorithmic structure often permits a Gibbs sampler algorithm (Geman and Geman 1984), an important special case of the Metropolis-Hastings algorithm. Examples

P. Neal (✉)
School of Mathematics, University of Manchester, Alan Turing Building, Oxford Rd, Manchester, M13 9PL, UK
e-mail: p.neal@lancaster.ac.uk

Present address:
P. Neal
Department of Mathematics and Statistics, Lancaster University, Lancaster, LA1 4YF, UK

T. Kypraios
School of Mathematical Sciences, University of Nottingham, University Park, Nottingham, NG7 2RD, UK

Published online: 03 December 2013

of Gibbs sampler algorithms that use data augmentation are the genetics linkage data (Dempster et al. 1977; Tanner and Wong 1987), mixture distributions (Diebolt and Robert 1994; Frühwirth-Schnatter 2006), censored data (Smith and Roberts 1993) and Gaussian hierarchical models (Papaspoliopoulos et al. 2003), to name but a few. The Gibbs sampler exploits the conditional distributions of the components of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$ and $\mathbf{y} = (y_1, \ldots, y_m)$, by at each iteration, successively drawing θ_i from $\pi(\theta_i|\boldsymbol{\theta}_{i-}, \mathbf{y}, \mathbf{x})$ $(i = 1, \ldots, d)$ and y_j from $\pi(y_j|\boldsymbol{\theta}, \mathbf{y}_{j-}, \mathbf{x})$ $(j = 1, \ldots, m)$, where $\boldsymbol{\theta}_{i-}$ and $\mathbf{y}_{j-}$ denote the vectors $\boldsymbol{\theta}$ and $\mathbf{y}$ with the ith and jth component omitted, respectively. Typically the sampling distributions are well known probability distributions. The Gibbs sampler generates dependent samples from $\pi(\boldsymbol{\theta}, \mathbf{y}|\mathbf{x})$, where usually only the $\boldsymbol{\theta}$ values are stored, with $\mathbf{y}$ being discarded as nuisance parameters. In many problems there is conditional independence between the elements of the parameter vector $\boldsymbol{\theta}$ given $\mathbf{y}$. That is, $\pi(\theta_i|\boldsymbol{\theta}_{i-}, \mathbf{y}, \mathbf{x}) = \pi(\theta_i|\mathbf{y}, \mathbf{x})$. This is seen, for example, in simple Normal mixture models with known variance (Diebolt and Robert 1994) or Poisson mixture models (Fearnhead 2005) and for the infection and recovery parameters of the general stochastic epidemic model with unknown infection times $(\mathbf{y})$ and observed recovery times $(\mathbf{x})$, see for example, O'Neill and Roberts (1999), Neal and Roberts (2005) and Kypraios (2007). Thus, in this case,

$$\pi(\boldsymbol{\theta}, \mathbf{y}|\mathbf{x}) = \pi(\mathbf{y}|\mathbf{x}) \prod_{i=1}^{d} \pi(\theta_i|\mathbf{y}, \mathbf{x}). \tag{1.2}$$

For joint densities which factorise in the form of (1.2), it is possible in certain circumstances to *collapse* the Gibbs sampler (Liu 1994) by integrating out $\boldsymbol{\theta}$. That is, compute $\pi(\mathbf{y}|\mathbf{x})$, up to a constant of proportionality, and construct a Metropolis-Hastings algorithm for sampling from $\pi(\mathbf{y}|\mathbf{x})$. Note that after collapsing it is rarely possible to construct a straightforward Gibbs sampler to sample from $\pi(\mathbf{y}|\mathbf{x})$. In Liu (1994), the benefits of collapsing the Gibbs sampler are highlighted. Collapsing can equally be applied to any Metropolis-Hastings algorithm, see, for example, Neal and Roberts (2005) and Kypraios (2007). In Fearnhead (2005) and Fearnhead (2006), the idea of collapsing is taken a stage further in the case where the number of possibilities of $\mathbf{y}$ is finite by computing $\pi(\mathbf{y}|\mathbf{x})$ exactly, and hence, express the posterior distribution, $\pi(\boldsymbol{\theta}|\mathbf{x})$, as a finite mixture distribution. Specifically, Fearnhead (2005) and Fearnhead (2006) consider mixture models and change-point models, respectively, with the main focus of both papers perfect simulation from the posterior distribution as an alternative to MCMC.

In this paper we present a generic framework for obtaining the exact posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{x})$ using data augmentation. The generic framework covers mixture models (Fearnhead 2005) and change-point models (Fearnhead 2006) as important special cases, but is applicable to a wide range of problems including two-level mixing Reed-Frost epidemic model (Sect. 3) and integer valued autoregressive time series models (Sect. 4). The key observation, which is developed in Sect. 2, is that there are generic classes of models, where augmented data, $\mathbf{y}$, can be identified such that, $\pi(\mathbf{y}|\mathbf{x})$ can be computed exactly, and $\pi(\boldsymbol{\theta}|\mathbf{y}, \mathbf{x})$ is a well known probability density. In such circumstances we can express the exact posterior distribution, $\pi(\boldsymbol{\theta}|\mathbf{x})$, as a finite mixture distribution satisfying

$$\pi(\boldsymbol{\theta}|\mathbf{x}) = \sum_{\mathbf{y}} \pi(\boldsymbol{\theta}|\mathbf{y}, \mathbf{x})\pi(\mathbf{y}|\mathbf{x}). \tag{1.3}$$

Throughout this paper we use the term, exact posterior (distribution), to refer to using (1.3). In Sect. 2, we show how in general $\mathbf{y}$ can be chosen and how sufficient statistics can be exploited to extend the usefulness of the methodology. Thus we extend the methodology introduced in Fearnhead (2005) beyond mixture models by identifying the key features which make (1.3) practical to use.

A natural alternative to computing the exact posterior distribution is to use MCMC, and in particular, the Gibbs sampler, to obtain a sample from the posterior distribution. There are a number of benefits from obtaining the exact posterior distribution. Firstly, any summary statistic $\mathbb{E}[h(\theta)|\mathbf{x}]$, for which $\mathbb{E}[h(\theta)|\mathbf{x}, \mathbf{y}]$ is known, can be computed without Monte Carlo error. Moreover, knowing the exact posterior distribution enables the use of importance sampling to efficiently estimate $\mathbb{E}[h(\theta)|\mathbf{x}]$, even when $\mathbb{E}[h(\theta)|\mathbf{x}, \mathbf{y}]$ is unknown. Secondly, there are none of the convergence issues of MCMC, such as determining the length of the burn-in period, and the total number of iterations required for a given effective sample size. Thirdly, the marginal likelihood (evidence) of the model, $\pi(\mathbf{x}) = \int \pi(\mathbf{x}|\theta)\pi(\theta)\, d\theta$ can easily be computed. This can be used for model selection, see, for example, Fearnhead (2005), Sect. 3. Model selection using MCMC samples is non-trivial, see, for example, Han and Carlin (2001), and often requires the construction of a reversible jump MCMC, Green (1995), adding an extra level of complexity to the MCMC procedure. Fourthly, obtaining the exact posterior distribution enables straightforward perfect simulation to obtain independent and identically distributed samples from the posterior distribution (Fearnhead 2005). This provides an efficient alternative to 'coupling from the past', Propp and Wilson (1996), perfect simulation MCMC algorithms, with the key computational cost, computing the exact posterior distribution, incurred only once (Fearnhead 2005). Finally, we can obtain the exact predictive distribution for forward prediction (see Liu 1994), although MCMC samples can easily be used to obtain good samples from the predictive distribution of future observations.

There are also important drawbacks of the proposed methodology compared with MCMC. Only a small number of relatively simple, albeit important, models can be analysed using (1.3), compared to the vast array of models that can increasingly be analysed using MCMC. MCMC algorithms are generally easier to program than algorithms for computing the exact posterior distribution. The computational cost of the exact posterior distribution grows exponentially in the total number of observations, whereas the computational cost of many MCMC algorithms will be linear or better in the total number of observations. However, for moderate size data sets computation of the exact posterior distribution and the MCMC alternative, obtaining a sufficiently large MCMC sample, can often take comparable amounts of time. We discuss computational costs in more detail in Sects. 3.3 and 4 for the household Reed-Frost epidemic model and the integer autoregressive (INAR) model, respectively.

The remainder of the paper is structured as follows. A generic framework is presented in Sect. 2 for computing the exact posterior distribution for a wide class of models, including Poisson mixture models, models for multinomial-beta data (Sect. 3) and integer valued autoregressive (INAR) processes (Sect. 4). We outline how sufficient statistics can be exploited to extend the usefulness of the methodology from small to moderate sized data sets. This involves giving details of an efficient mechanism for identifying the different sufficient statistics, **s**, compatible with **x** and for computing $\pi(\mathbf{s}|\mathbf{x})$. In Sect. 3, we consider a general model for multinomial-beta data motivated by the genetic linkage data studied in Dempster et al. (1977) and Tanner and Wong (1987) and applicable to a two-level mixing Reed-Frost epidemic model (see Addy et al. 1991). The exact posterior distribution of the parameters is obtained with computation of key summary statistics of the posterior distribution, such as the posterior mean and standard deviation. In Sect. 4, we obtain the exact posterior distribution for the parameters of the integer valued autoregressive (INAR) processes (McKenzie 2003; Jung and Tremayne 2006; Neal and Subba Rao 2007). In addition to computing the key summary statistics of the posterior distribution, we use the marginal likelihood for selecting the most parsimonious model for the data and we derive the exact predictive distribution for forecasting purposes. All the models considered in Sects. 3 and 4 have previously only been analysed using MCMC and/or by numerical methods that are only applicable for very small data sets. Throughout we compare analysing the data sets using the exact posterior distribution with using MCMC. Finally, in Sect. 5 we conclude with a few final remarks upon other models that can be analysed using (1.3).

2 Generic setup

2.1 Introduction to the generic setup

In this Section we consider the generic situation where $\pi(\boldsymbol{\theta}|\mathbf{x})$ is intractable, but there exists augmented data, **y**, such that, $\pi(\boldsymbol{\theta}|\mathbf{y},\mathbf{x})$ belongs to a well known probability distribution. Specifically, we assume that **y** is such that the components of $\boldsymbol{\theta}$ are conditionally independent given **y**. That is,

$$\pi(\boldsymbol{\theta}|\mathbf{y},\mathbf{x}) = \prod_{i=1}^{d} \pi(\theta_i|\mathbf{y},\mathbf{x}), \tag{2.1}$$

with $\theta_i|\mathbf{y},\mathbf{x}$ belonging to a well known probability distribution. In particular, suppose that we have

$$\pi(\mathbf{x}|\boldsymbol{\theta}) = \sum_{j=1}^{N} \pi(\mathbf{x},\mathbf{y}_j|\boldsymbol{\theta}), \tag{2.2}$$

where for $j = 1, 2, \ldots, N$, $\pi(\boldsymbol{\theta}|\mathbf{y}_j,\mathbf{x})$ satisfies (2.1). That is, there is a finite number N of possible augmented data values that are consistent with the data. In each case the conditional posterior distribution of the parameters given the augmented data is easily identifiable.

To exploit the data augmentation introduced in (2.2), we consider the joint posterior density $\pi(\boldsymbol{\theta},\mathbf{y}_j|\mathbf{x})$, $j = 1, 2, \ldots, N$. The first step is to rewrite $\pi(\boldsymbol{\theta}|\mathbf{x})$ as

$$\begin{aligned}\pi(\boldsymbol{\theta}|\mathbf{x}) &= \sum_{j=1}^{N} \pi(\boldsymbol{\theta},\mathbf{y}_j|\mathbf{x}) \\ &= \sum_{j=1}^{N} \pi(\boldsymbol{\theta}|\mathbf{y}_j,\mathbf{x})\pi(\mathbf{y}_j|\mathbf{x}).\end{aligned} \tag{2.3}$$

Since $\mathbf{y}_j$ is chosen such that $\pi(\boldsymbol{\theta}|\mathbf{y}_j,\mathbf{x})$ is known, we only need to compute $\{\pi(\mathbf{y}_j|\mathbf{x})\}$ to be able to express $\pi(\boldsymbol{\theta}|\mathbf{x})$ as a mixture density. The key step for obtaining $\pi(\mathbf{y}_j|\mathbf{x})$ is to observe that

$$\begin{aligned}\pi(\boldsymbol{\theta},\mathbf{y}_j|\mathbf{x}) &= \pi(\mathbf{y}_j|\mathbf{x})\pi(\boldsymbol{\theta}|\mathbf{y}_j,\mathbf{x}) \\ &\propto K_{\mathbf{y}_j} \prod_{i=1}^{d} \pi(\theta_i|\mathbf{y}_j,\mathbf{x}),\end{aligned}$$

for some $K_{\mathbf{y}_j}$ independent of $\boldsymbol{\theta}$. We can then integrate out $\boldsymbol{\theta}$, which since $\int \pi(\theta_i|\mathbf{y},\mathbf{x})d\theta_i = 1$, leaves $\pi(\mathbf{y}_j|\mathbf{x}) \propto K_{\mathbf{y}_j}$. Therefore, for $j = 1, 2, \ldots, N$,

$$\pi(\mathbf{y}_j|\mathbf{x}) = \frac{K_{\mathbf{y}_j}}{\sum_{k=1}^{N} K_{\mathbf{y}_k}}. \tag{2.4}$$

This is a 'brute-force' approach enumerating all the N possibilities and computing the normalising constant. The main complication is that computing the normalization constant in (2.4) is only be practical if N is sufficiently small.

A partial solution to the above brute-force enumeration is to use sufficient statistics to significantly reduce the number of computations required. This is successfully employed in Fearnhead (2005) for Poisson mixture models (Sect. 2.2), where a simple, sequential algorithmic approach is given to compute sufficient statistics. We develop this approach detailing how the sequential computation of the sufficient statistics can be efficiently implemented in the generic setup.

In general, let $S(\mathbf{y}, \mathbf{x})$ denote the sufficient statistics of the model under consideration, such that $\pi(\boldsymbol{\theta}|\mathbf{y}, \mathbf{x}) = \pi(\boldsymbol{\theta}|S(\mathbf{y}, \mathbf{x}))$. Let $\mathcal{S} = \{S(\mathbf{y}_1, \mathbf{x}), S(\mathbf{y}_2, \mathbf{x}), \dots, S(\mathbf{y}_N, \mathbf{x})\}$, the set of all possible sufficient statistics compatible with the data. Then we can write

$$\begin{aligned}\pi(\boldsymbol{\theta}|\mathbf{x}) &= \sum_{k=1}^{S} \pi(\boldsymbol{\theta}|\mathbf{s}_k)\left\{\sum_{j:S(\mathbf{y}_j,\mathbf{x})=\mathbf{s}_k} \pi(\mathbf{y}_j|\mathbf{x})\right\}\\ &= \sum_{k=1}^{S} \pi(\boldsymbol{\theta}|\mathbf{s}_k)\pi(\mathbf{s}_k|\mathbf{x}),\end{aligned} \tag{2.5}$$

where $S = |\mathcal{S}|$ and $\mathcal{S} = \{\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_S\}$. For (2.5) to be useful, a systematic approach to computing $\mathcal{S}$ and $\pi(\mathbf{s}_k|\mathbf{x}) = \sum_{j:S(\mathbf{y}_j,\mathbf{x})=\mathbf{s}_k} \pi(\mathbf{y}_j|\mathbf{x})$ is needed that requires far fewer than N computations.

The remainder of this section is structured as follows. In Sect. 2.2, we introduce the Poisson mixture model, (Fearnhead 2005) as an illustrative example of how the exact posterior distribution can be computed. Then in Sect. 2.3 we develop the general framework referencing the Poisson mixture model as an example. This culminates in a mechanism for computing sufficient statistics and associated probability weights in order to fully exploit (2.5). Finally, in Sect. 2.4 we make some final remarks about the generic framework and its limitations.

2.2 Illustrative example

A prime example of a model where (2.2) can be exploited is the Poisson mixture model (Fearnhead 2005). For the Poisson mixture model the data $\mathbf{x}$ arises as independent and identically distributed observations from X, where for $x = 0, 1, \dots, \mathbb{P}(X = x) = \sum_{k=1}^{m} p_k \lambda_k^x \exp(-\lambda_k)/x!$. That is, for $k = 1, 2, \dots, m$, the probability that X, is drawn from $\text{Po}(\lambda_k)$ is p_k. In this case, $y_i = 1, 2, \dots, m$ with $y_i = k$ denoting x_i is drawn from $\text{Po}(\lambda_k)$, see, for example, (Fearnhead 2005) for details. Thus $N = m^n$. Although the exact posterior distribution for this model has been derived in Fearnhead (2005), it is good illustrative example for the generic setup presented in this paper and it is a useful stepping stone before moving onto the more complex Reed-Frost epidemic model and integer valued autoregressive (INAR) model in Sects. 3 and 4, respectively.

As noted in Fearnhead (2005), $\pi(\boldsymbol{\theta}|\mathbf{y}, \mathbf{x})$ depends upon $\mathbf{y}$ and $\mathbf{x}$ through the $2m$ dimensional sufficient statistic $S(\mathbf{y}, \mathbf{x}) = (\mathbf{a}, \mathbf{z})\ (= \mathbf{s})$, where $a_k = \frac{1}{n}\sum_{i=1}^{n} 1_{\{y_i=k\}}$ and $z_k = \frac{1}{n}\sum_{i=1}^{n} 1_{\{y_i=k\}} x_i$ denote the total number of observations and the sum of the observations from the kth Poisson component, respectively. Note that $2(m-1)$ sufficient statistics suffice as a_m and z_m can be recovered from $\mathbf{x}$ and $(a_1, z_1, \dots, z_{m-1})$, see Fearnhead (2005). The sequential method for constructing $\mathcal{S}$ given in Fearnhead (2005) ensures that the amount of computations required are far fewer than N. In Fearnhead (2005), Sect. 3.1, for a Poisson mixture model comprising 2 components and 1096 observations, $N = 2^{1096} = 8.49 \times 10^{329}$ with $S = |\mathcal{S}| = 501501$. Whilst in this example S is still large, the computation of $\{\pi(\mathbf{s}_k|\mathbf{x})\}$ can be done relatively quickly, see Fearnhead (2005) and an example with S approximately 2.5×10^7 is considered in Sect. 4.2 of this paper.

2.3 Sufficient statistics

In this section we outline how (2.5) can be derived for a broad class of models that covers Poisson mixture models, the Reed-Frost household epidemic model (Sect. 3) and the INAR model (Sect. 4). First note that we can write,

$$\pi(\mathbf{x}|\boldsymbol{\theta}) = \prod_{t=1}^{n} \pi(x_t|\mathbf{x}_{t-1}, \boldsymbol{\theta}), \tag{2.6}$$

where $\mathbf{x}_t = (x_1, x_2, \dots, x_t)$ $(t = 1, \dots, n)$, with $\pi(x_1|\mathbf{x}_0, \boldsymbol{\theta}) = \pi(x_1|\boldsymbol{\theta})$. We consider models for which the augmented data, $\mathbf{y}$, can be chosen such that

$$\pi(\mathbf{x}|\boldsymbol{\theta}) = \prod_{t=1}^{n}\left\{\sum_{k=1}^{m_t} \pi(x_t, y_{t,k}|\mathbf{x}_{t-1}, \boldsymbol{\theta})\right\} \tag{2.7}$$

with $\prod_{t=1}^{n} m_t (= N)$ denoting the total number of possibilities for $\mathbf{y}$. That is, we assume that there are m_t possibilities, in terms of the choice of the augmented data, of how the observation x_t could have arisen given $\mathbf{x}_{t-1}$, with each $y_{t,k}$ corresponding to a different possibility. For the Poisson mixture model, the observations are assumed to be independent and identically distributed (x_t is independent of $\mathbf{x}_{t-1}$) and $m_t = m$ $(t = 1, 2, \dots, n)$ with $y_{t,k}$ corresponding to x_t arising from the kth Poisson component. The extension of (2.7) to the more general case $\pi(\mathbf{x}|\boldsymbol{\theta}) = \prod_{t=1}^{n}\{\sum_{k=1}^{m_t} \pi(x_t, y_{t,k}|\mathbf{x}_{t-1}, \mathbf{y}_{t-1}, \boldsymbol{\theta})\}$, where the allocation $y_{t,k}$ depends on both the observations $\mathbf{x}_{t-1}$ and the allocations $\mathbf{y}_{t-1}$ is straightforward.

The key step in using (2.7), is to identify appropriate augmented data, $\mathbf{y}$. The choice of $\mathbf{y} = (y_1, y_2, \dots, y_n)$ has to be made such that $\pi(\boldsymbol{\theta}|\mathbf{y},\mathbf{x})$ is from a well known probability distribution. Thus $\mathbf{y}$ is chosen such that

$$\pi(\mathbf{x},\mathbf{y}|\boldsymbol{\theta}) = \prod_{t=1}^{n}\left\{c_{t,y_t}\prod_{i=1}^{d}h_{ty_ti}(\theta_i)\right\}, \tag{2.8}$$

where $\pi(x_t|\mathbf{x}_{t-1},\boldsymbol{\theta}) = \sum_{k=1}^{m_t}\pi(x_t, y_{t,k}|\mathbf{x}_{t-1},\boldsymbol{\theta})$ and $h_{tki}(\cdot)$ is an integrable function throughout this paper $h_{tki}(q)$ will be $q^{A-1}(1-q)^{B-1}$ $(0 \le q \le 1)$ or $q^{A-1}\exp(-Bq)$ $(q \ge 0)$, although other choices of $h_{tki}(q)$ are possible. Note that both $q^{A-1}(1-q)^{B-1}$ $(0 \le q \le 1)$ and $q^{A-1}\exp(-Bq)$ $(q \ge 0)$ are proportional to probability density functions from an exponential family of distributions, namely, beta(A, B) and gamma(A, B), respectively, which arise naturally as the posterior distributions of discrete parametric distributions (for example, the binomial, negative binomial and Poisson distributions). We require that for a fixed i, $h_{sji}(\cdot)$ and $h_{tki}(\cdot)$ $(j, k, s, t \in \mathbb{N})$ belong to the same family of functions (probability distributions) and that a conjugate prior, $\pi_i(\cdot)$, for θ_i exists. This allows for easy identification of the conditional posterior distribution of θ_i, $\pi(\theta_i|\mathbf{y},\mathbf{x})$. Therefore, we have that

$$\pi(\mathbf{x},\mathbf{y},\boldsymbol{\theta}) = \left\{\prod_{t=1}^{n}c_{t,y_t}\right\}\prod_{i=1}^{d}\left(\left\{\prod_{t=1}^{n}h_{ty_ti}(\theta_i)\right\}\pi(\theta_i)\right). \tag{2.9}$$

Then

$$\pi(\mathbf{y}|\mathbf{x}) \propto \prod_{t=1}^{n}c_{t,y_t}\prod_{i=1}^{d}B_i(\mathbf{y}),$$

where $B_i(\mathbf{y}) = \int\prod_{t=1}^{n}h_{ty_ti}(\theta_i)d\theta_i$ $(i = 1, 2, \ldots, d)$. This in turn gives

$$\pi(\mathbf{y}_j|\mathbf{x}) = \prod_{t=1}^{n}c_{t,y_{t,j}}\prod_{i=1}^{d}B_i(\mathbf{y}_j) \Bigg/ \left\{\sum_{k=1}^{N}\left(\prod_{t=1}^{n}c_{t,y_{t,k}}\prod_{i=1}^{d}B_i(\mathbf{y}_k)\right)\right\}. \tag{2.10}$$

Note that for (2.10) it suffices to know $c_{t,y_{t,j}}$ up to a constant of proportionality. Hence we can replace $c_{t,y_{t,j}}$ by $\tilde{c}_{t,y_{t,j}} = k_{x_t}c_{t,y_{t,j}}$, where k_{x_t} is a constant independent of $y_{t,j}$.

Suppose that for $t = 1, 2, \ldots, n$, $h_{ty_ti}(\theta_i) = \theta_i^{e_{ti}}(1-\theta_i)^{f_{ti}}$ $(0 \le \theta_i \le 1)$ with a beta(C_i, D_i) prior on θ_i. Then setting $E_i = C_i + \sum_{t=1}^{n}e_{ti}$ and $F_i = D_i + \sum_{t=1}^{n}f_{ti}$, we have that the (conditional) posterior distribution for θ_i is beta(E_i, F_i) with

$$B_i(\mathbf{y}) = \frac{\Gamma(C_i + D_i)}{\Gamma(C_i)\Gamma(D_i)} \times \frac{\Gamma(E_i)\Gamma(F_i)}{\Gamma(E_i + F_i)}. \tag{2.11}$$

Alternatively suppose that for $t = 1, 2, \ldots, n$, $h_{ty_ti}(\theta_i) = \theta_i^{e_{ti}}\exp(-f_{ti}\theta_i)$ $(\theta_i \ge 0)$ with a gamma(C_i, D_i) prior on θ_i. Then setting $E_i = C_i + \sum_{t=1}^{n}e_{ti}$ and $F_i = D_i + \sum_{t=1}^{n}f_{ti}$, we have that the (conditional) posterior distribution for θ_i is gamma(E_i, F_i) with

$$B_i(\mathbf{y}) = \frac{D_i^{C_i}}{\Gamma(C_i)} \times \frac{\Gamma(E_i)}{F_i^{E_i}}. \tag{2.12}$$

The expressions for $B_i(\mathbf{y})$ in (2.11) and (2.12) are particularly straightforward if E_i and F_i are integers. This will be the case when the data arises from discrete parametric distributions with $C_i, D_i \in \mathbb{N}$. In both cases E_i and F_i are sufficient statistics (the dependence upon $\mathbf{y}$ is suppressed) and the sums of $n+1$ terms, with each term depending upon a given observation or the prior. This is key for constructing the sufficient statistics in an efficient manner and we shall discuss this shortly.

We illustrate the above with the 2 component Poisson mixture model, the extension to the m component Poisson mixture model is trivial. Assign a $beta(1, 1)$ prior on p $(p_1 = p, p_2 = 1 - p)$ and gamma(C_k, D_k) prior on λ_k $(k = 1, 2)$, giving

$$\begin{aligned}\pi(\mathbf{x},\mathbf{y},\boldsymbol{\theta}) &= \prod_{t=1}^{n}\left(p^{1_{\{y_t=1\}}}(1-p)^{1_{\{y_t=2\}}} \times \frac{\lambda_{y_t}^{x_t}}{x_t!}\exp(-\lambda_{y_t})\right) \\ &\quad \times 1 \times \prod_{k=1}^{2}\frac{D_k^{C_k}}{\Gamma(C_k)}\lambda_k^{C_k-1}\exp(-\lambda_k D_k) \\ &= \prod_{t=1}^{n}\Bigg(p^{1_{\{y_t=1\}}}(1-p)^{1_{\{y_t=2\}}} \\ &\quad \times \prod_{k=1}^{2}\Bigg\{\left(\prod_{t:y_t=k}\frac{\lambda_k^{x_t}}{x_t!}\exp(-\lambda_k)\right) \\ &\quad \times \frac{D_k^{C_k}}{\Gamma(C_k)}\lambda_k^{C_k-1}\exp(-\lambda_k D_k)\Bigg\}\Bigg),\end{aligned} \tag{2.13}$$

where $\theta = (p, \lambda_1, \lambda_2)$. Thus,

$$\begin{aligned}\pi(\mathbf{y},\boldsymbol{\theta}|\mathbf{x}) &= \left(\prod_{t=1}^{n}\frac{1}{x_t!}\right)p^{a_1}(1-p)^{a_2}\prod_{k=1}^{2}\lambda_k^{z_k+C_k-1} \\ &\quad \times \exp\bigl(-\lambda_k(a_k + D_k)\bigr) \\ &\propto p^{a_1}(1-p)^{a_2}\prod_{k=1}^{2}\lambda_k^{z_k+C_k-1} \\ &\quad \times \exp\bigl(-\lambda_k(a_k + D_k)\bigr),\end{aligned} \tag{2.14}$$

giving

$$\pi(\mathbf{y}|\mathbf{x}) \propto \frac{a_1!a_2!}{(a_1+a_2+1)!} \times \prod_{k=1}^{2}\frac{\Gamma(z_k + C_k)}{(a_k + D_k)^{z_k+C_k}}$$

$$\propto \prod_{k=1}^{2} \frac{a_k!\Gamma(z_k + C_k)}{(a_k + D_k)^{z_k + C_k}}, \tag{2.15}$$

since $a_1 + a_2 + 1 = n + 1$, independent of $\mathbf{y}$. Finally, $\pi(\mathbf{y}|\mathbf{x})$ can be computed by finding the normalising constant by summing the right hand side of (2.15) over the $N = 2^n$ possibilities. This is only practical if N is sufficiently small, hence the need to exploit sufficient statistics to reduce the number of calculations that are necessary.

Suppose that the vector of sufficient statistics, $S(\mathbf{y}, \mathbf{x}) = \mathbf{s}$, is D-dimensional and that $s_l = \sum_{t=1}^n g_{t,l}(y_t, x_t)$ ($l = 1, 2, \ldots, D$). This form of sufficient statistic is consistent with the beta and gamma conditional posterior distributions for the components of $\boldsymbol{\theta}$ outlined above. For example, for the Poisson mixture model with $\mathbf{s} = (\mathbf{a}, \mathbf{z})$, $a_l = \sum_{t=1}^n g_l(y_t, x_t)$ with $g_l(y, x) = 1_{\{y=l\}}$ and $z_l = \sum_{t=1}^n g_l(y_t, x_t)$ with $g_l(y, x) = 1_{\{y=l\}} x$. Now for $p = 1, 2, \ldots, n$, let $S^p(\mathbf{y}, \mathbf{x}) = \mathbf{s}^p$ denote the sufficient statistics for the first p observations, *i.e.* using $\mathbf{y}_p$ and $\mathbf{x}_p$ only, with $s_l^p = \sum_{t=1}^p g_{t,l}(y_t, x_t)$ denoting the corresponding partial sum. In the obvious notation, let $\mathcal{S}^p = \{S^p(\mathbf{y}_1, \mathbf{x}), S^p(\mathbf{y}_2, \mathbf{x}), S^p(\mathbf{y}_n, \mathbf{x})\} = \{\mathbf{s}_1^p, \mathbf{s}_2^p, \ldots, \mathbf{s}_{S^p}^p\}$ denote the set of possible sufficient statistics for the first p observations with $S^p = |\mathcal{S}^p|$. Let $C_j^p = \sum_{\mathbf{y}:S^p(\mathbf{y},\mathbf{x})=\mathbf{s}_j^p} \prod_{t=1}^p c_{t,y_t}$, the relative weight associated, with $\mathbf{s}_j^p$. It is trivial to obtain $\mathcal{S}^1$ and $\{C_j^1\}$. For $p = 2, 3, \ldots, n$, we can construct $\mathcal{S}^p$ and corresponding weights $\{C_j^p\}$ iteratively, using $\mathcal{S}^{p-1}$, $\{C_j^{p-1}\}$ and $\{(y_{p,k}, x_p)\}$ as follows. Let $\sigma^p(y_p, x_p) = (g_{p,1}(y_p, x_p), g_{p,2}(y_p, x_p), \ldots, g_{p,D}(y_p, x_p))$, the vector of terms that need to be added to the sufficient statistics $\mathbf{s}^{p-1}$ to construct $\mathbf{s}^p$, given that the pth observation and augmented data are x_p and y_p, respectively. Then $\mathcal{S}^p = \{\mathbf{s}^p = \mathbf{s}^{p-1} + \sigma^p(y_{p,k}, x_p) : \mathbf{s}^{p-1} \in \mathcal{S}^{p-1}, k \in \{1, 2, \ldots, m_t\}\}$, with $C_j^p = \sum_{\{\mathbf{s}_l^{p-1} + \sigma^p(y_{t,k}, x_t) = \mathbf{s}_j^p\}} C_l^{p-1} c_{p,y_{p,k}}$. That is, we consider all possible allocations for the pth observation combined with the sufficient statistics constructed from the first $p - 1$ observations. Finally, as noted after (2.10), for the calculation of the weights, we can multiply the $c_{t,y_{t,j}}$'s by arbitrary constants that do not depend upon $\mathbf{y}$.

The above construction of sufficient statistics is a generalisation of that given in Fearnhead (2005), Sect. 2, for mixture models, where the construction of sufficient statistics for data $\mathbf{x} = (1, 1, 2, 1)$ arising from a 2 component Poisson mixture model is discussed. For $p = 1, 2, 3, 4$, we take $c_{p,y_p} = 1$ regardless of the allocation of the pth observation, which ignores the constant $1/x_p!$ present in the first line of (2.14). Then in our notation with $\mathbf{s}^p = (a_1^p, z_1^p)$ since $a_2^p = p - a_1^p$ and $z_2^p = \sum_{i=1}^p x_i - z_1^p$, we have that $\mathcal{S}^3 = \{(3, 4), (2, 3), (2, 2), (1, 2), (1, 1), (0, 0)\}$ with $\mathcal{C}^3 = \{1, 2, 1, 1, 2, 1\}$ denoting the corresponding set of weights. It is straightforward to add $\sigma^p(1, 1) = (1, 1)$ and $\sigma^p(2, 1) = (0, 0)$ to the elements of $\mathcal{S}^3$ and the weights to $\mathcal{C}^3$ to get $\mathcal{S}^4 = \{(4, 5), (3, 4), (3, 3), (2, 3), (2, 2), (1, 2), (1, 1), (0, 0)\}$ with corresponding weights $\mathcal{C}^4 = \{1, 3, 1, 3, 3, 1, 3, 1\}$. For example, $\mathbf{s}^4 = (3, 4)$ can arise from $\mathbf{s}^3 = (3, 4)$ and $y_4 = 2$ or $\mathbf{s}^3 = (2, 3)$ and $y_4 = 1$. See Fearnhead (2005), Sect. 2, for a diagram of the construction.

2.4 Remarks

We make a few concluding comments about the generic setup before implementing the methodology in Sects. 3 and 4. For ease of exposition, throughout this section we have focused upon $\pi(\boldsymbol{\theta}|\mathbf{y}, \mathbf{x}) = \prod_{i=1}^d \pi(\theta_i|\mathbf{y}, \mathbf{x})$, (2.1), where θ_i is a single parameter. It is trivial to extend to the case where θ_i is a vector of parameters, provided that $\pi(\theta_i|\mathbf{y}, \mathbf{x})$ belongs to a well known probability distribution. A common example of this is where $\pi(\theta_i|\mathbf{y}, \mathbf{x})$ is the probability density function of a Dirichlet distribution and we will consider an example of this is Sect. 3.2. All the other examples considered in this paper satisfy (2.1).

The generic set up allows us to identify classes of models for which the methodology described in this paper are applicable and also models for which the methodology is not appropriate. The key requirement is that the augmented data $\mathbf{y}$ is discrete. Note that, in principle, we do not require $\mathbf{x}$ to be discrete with the methodology readily applicable to a mixture of m Gaussian distribution with unknown means and known variances with y_i denoting the allocation of an observation x_i to a particular Gaussian distribution. The main limitation in applying the methodology to mixtures of Gaussian distributions, or more generally mixtures of continuous distributions, is that all m^n possible values for $\mathbf{y}$ need to be evaluated as almost surely each combination of $(\mathbf{y}, \mathbf{x})$ yields a different sufficient statistic. Therefore for the methodology to be practically useful we require $\mathbf{x}$ and $\mathbf{y}$ to be discrete. For the examples considered in this paper and the Poisson mixture model, the probability of observing x_t can be expressed as a sum over m_t terms, where each term in the sum consists either of a single discrete random variable or the sum of independent discrete random variables. The well known discrete probability distributions are Bernoulli-based distributions (binomial, negative binomial, geometric distributions), the Poisson distribution and the multinomial distribution which give rise to beta, gamma and Dirichlet posterior distributions. Hence the emphasis on the discussion of these probability distributions above.

A final remark is that we have discussed the data $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ in terms of n observations. This is an applicable representation of the data for the models considered in Sects. 3 and 4. However, it will be convenient in Sect. 3 to give a slightly different representation of the data with x_i denoting the total number of observations in the data that belong to a given category i. For example, for data from a Poisson mixture model category i would correspond to observations equal to i. The four observations 1, 1, 2, 1 would be recorded as three 1 s and a 2 and we can then characterise the data augmentation by how many 1 s and 2 s are assigned to each Poisson distribution. This representation can greatly reduce N. For example, for the genetics linkage data in Sect. 3.2 it reduces the number of possibilities from 2^{125} to 126. This alternative representation can then negate the need for computing sufficient statistics or can speed up the computation of the sufficient statistics.

3 Multinomial-beta data

3.1 Introduction to multinomial-beta data

In this section we study in detail a special case of the generic model introduced in Sect. 2. This case is motivated by the classic genetic linkage data from (Rao 1965, pp. 368–369), popularized by Dempster et al. (1977), where the data are assumed to arise from a multinomial distribution with the components of θ having independent beta distributed posterior distributions, conditional upon $\mathbf{x}$ and $\mathbf{y}$. Other models that give rise to multinomial-beta data are the Reed-Frost epidemic model (Longini and Koopman 1982; Addy et al. 1991; O'Neill and Roberts 1999) and (O'Neill et al. 2000) and with minor modifications the *INAR*(p) model with Geometrically distributed innovations, see Sect. 4.3 and (McCabe and Martin 2005). As noted at the end of Sect. 2, it is convenient to use an alternative representation of the data than that given for the generic model and we give details of the generic form of the data below.

Suppose that there are n independent observations that are classified into t types with n_h observations of type $h = 1, 2, \dots, t$. For the genetic linkage data $t = 1$ and the subscript h can be dropped. The n_h observations of type h are divided into k_h categories with each observation independently having probability $p_{hi}(\boldsymbol{\theta})$ of belonging to category (h, i) $(i = 1, 2, \dots, k_h)$. Let $\mathbf{p}_h(\boldsymbol{\theta}) = (p_{h1}(\boldsymbol{\theta}), p_{h2}(\boldsymbol{\theta}), \dots, p_{hk_h}(\boldsymbol{\theta}))$ and let $\mathbf{x}_h = (x_{h1}, x_{h2}, \dots, x_{hk_h})$, where $\boldsymbol{\theta}$ is a d-dimensional vector of parameters and x_{hi} denotes the total number of observations in category (h, i). For $i = 1, 2, \dots, k_h$, we assume that there exists $m_{hi} \in \mathbb{N}$ such that $p_{hi}(\boldsymbol{\theta}) = \sum_{l=1}^{m_{hi}} c_{hil} \prod_{j=1}^{d} \theta_j^{a_{hilj}} (1 - \theta_j)^{b_{hilj}}$, where a_{hilj}, b_{hilj} and c_{hil} are constants independent of $\boldsymbol{\theta}$. Typically, a_{hilj} and b_{hilj} will be non-negative integers. Thus we have that

$$\pi(\mathbf{x}|\boldsymbol{\theta}) = \prod_{h=1}^{t} \frac{n_h!}{\prod_{i=1}^{k_h} x_{hi}!} \prod_{i=1}^{k_h} \sum_{l=1}^{m_{hi}} c_{hil} \prod_{j=1}^{d} \theta_j^{a_{hilj}} (1 - \theta_j)^{b_{hilj}},$$

which is of the same form as (2.7). If all of the m_{hi}'s are equal to 1, then assuming independent beta priors for the components of $\boldsymbol{\theta}$, the posterior distribution of $\boldsymbol{\theta}$ consist of d independent beta distributions, one for each component. (Throughout the remainder of this section, for ease of exposition, we assume a uniform prior on $\boldsymbol{\theta}$, that is, $\pi(\boldsymbol{\theta}) = 1$ for $\theta \in [0, 1]^d$.) However, if at least one of the m_{hi}'s is greater than 1 then the posterior distribution of θ is not readily available from $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_t)$ alone. Therefore we augment the data by dividing the type h data into $K_h = \sum_{i=1}^{k_h} m_{hi}$ categories, classified $(h, 1, 1), \dots, (h, 1, m_1), (h, 2, 1), \dots, (h, k_h, m_{k_h})$. For $i = 1, 2 \dots, k_h$ and $l = 1, 2, \dots, m_{hi}$, let y_{hil} and $q_{hil}(\boldsymbol{\theta}) = c_{hil} \prod_{j=1}^{d} \theta_j^{a_{hilj}} (1 - \theta_j)^{b_{hilj}}$ denote the total number of observations in category (h, i, l) and the probability that an observation belongs to category (h, i, l), respectively, with $x_{hi} = \sum_{l=1}^{m_i} y_{hil}$ and $p_{hi}(\boldsymbol{\theta}) = \sum_{l=1}^{m_{hi}} q_{hil}(\boldsymbol{\theta})$. Then

$$\pi(\mathbf{y}|\boldsymbol{\theta}) = \prod_{h=1}^{t} \frac{n_h!}{\prod_{i=1}^{k_h} \prod_{l=1}^{m_{hi}} y_{hil}!} \prod_{i=1}^{k_h} \prod_{l=1}^{m_{hi}} q_{hil}(\boldsymbol{\theta})^{y_{hil}},$$

with $\pi(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}) = \pi(\mathbf{x}|\mathbf{y}) = \prod_{h=1}^{t} \prod_{i=1}^{k_h} 1_{\{x_{hi} = \sum_{l=1}^{m_{hi}} y_{hil}\}}$, *i.e.* the augmented data $\mathbf{y}$ is consistent with the observed data $\mathbf{x}$. Therefore

$$\begin{aligned}\pi(\theta, \mathbf{y}|\mathbf{x}) \propto \prod_{h=1}^{t} \Bigg\{ & \left(\prod_{i=1}^{k_h} 1_{\{x_{hi} = \sum_{l=1}^{m_i} y_{hil}\}} \right) \frac{n_h!}{\prod_{i=1}^{k} \prod_{l=1}^{m_{hi}} y_{hil}!} \\ & \times \prod_{i=1}^{k_h} \prod_{l=1}^{m_{hi}} q_{hil}(\theta)^{y_{hil}} \Bigg\} \\ \propto \prod_{h=1}^{t} \Bigg\{ & \left(\prod_{i=1}^{k_h} 1_{\{x_{hi} = \sum_{l=1}^{m_i} y_{hil}\}} \right) \prod_{i=1}^{m_{hi}} \left(\frac{c_{hil}^{y_{hil}}}{y_{hil}!} \right) \\ & \times \prod_{j=1}^{d} \theta_j^{E_j(\mathbf{y})} (1 - \theta_j)^{F_j(\mathbf{y})} \Bigg\},\end{aligned} \tag{3.1}$$

where $E_j(\mathbf{y}) = \sum_{h=1}^{t} \sum_{i=1}^{k_h} \sum_{l=1}^{m_{hi}} a_{hilj} y_{il}$ and $F_j(\mathbf{y}) = \sum_{h=1}^{t} \sum_{i=1}^{k_h} \sum_{l=1}^{m_{hi}} b_{ilj} y_{hil}$. The sufficient statistics for the model are $\mathbf{S}(\mathbf{y}, \mathbf{x}) = (E_1(\mathbf{y}), F_1(\mathbf{y}), \dots, F_d(\mathbf{y}))$ in agreement with the observations made in Sect. 2. Hence, integrating $\boldsymbol{\theta}$ out of (3.1) yields

$$\begin{aligned}\pi(\mathbf{y}|\mathbf{x}) \propto & \prod_{h=1}^{t} \prod_{i=1}^{k_h} 1_{\{x_{hi} = \sum_{l=1}^{m_i} y_{hil}\}} \prod_{i=1}^{m_{hi}} \left(\frac{c_{hil}^{y_{hil}}}{y_{hil}!} \right) \\ & \times \prod_{j=1}^{d} \frac{E_j(\mathbf{y})! F_j(\mathbf{y})!}{(E_j(\mathbf{y}) + F_j(\mathbf{y}) + 1)!},\end{aligned} \tag{3.2}$$

with the components of $\boldsymbol{\theta}$, conditional upon $\mathbf{y}$, having independent beta posterior distributions.

For $h = 1, 2, \dots, t$ and $i = 1, 2, \dots, k_h$, the total number of possible states, $\mathbf{y}_{hi}$, satisfying $x_{hi} = \sum_{l=1}^{m_{hi}} y_{hil}$ is $\binom{x_{hi} + m_{hi} - 1}{x_{hi}}$ and thus the total number of possibilities for

$\mathbf{y}$ is $N = \prod_{h=1}^{t}\prod_{i=1}^{k_h}\binom{x_{hi}+m_{hi}-1}{x_{hi}}$. This is considerably fewer than the total number of data augmented states, $\prod_{h=1}^{t}\prod_{i=1}^{k_h} m_{hi}^{x_{hi}}$, that would be needed if we considered each of the n observations one by one rather than by category. Given $\pi(\mathbf{y}|\mathbf{x})$, the posterior distribution of $\pi(\boldsymbol{\theta}|\mathbf{x})$ can then easily be obtained. If N is small, we can work directly with (3.2), otherwise we can construct $\mathcal{S} = \{S(\mathbf{y}_1, \mathbf{x}), \dots, S(\mathbf{y}_N, \mathbf{x})\}$, the set of sufficient statistics consistent with $\mathbf{x}$ and the corresponding probability weights, as outlined in Sect. 2.

3.2 Genetic linkage data

The data consists of the genetic linkage of 197 animals and the data are divided into four genetic categories, labeled 1 through to 4, see Rao (1965), Dempster et al. (1977) and Tanner and Wong (1987). The probabilities that an animal belongs to each of the four categories are $\frac{1}{2}+\frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4}$, respectively. Let $\mathbf{x} = (x_1, x_2, x_3, x_4) = (125, 18, 20, 34)$ denote the total number of animals in each category. In the notation of Sect. 3.1, $d = 1$, $t = 1$, $m_1 = 2$, $m_2 = m_3 = m_4 = 1$, so data augmentation is required to gain insight into the posterior distribution of θ, see Tanner and Wong (1987). Let $q_{11}(\theta) = 1/2$, $q_{12}(\theta) = \theta/4$ and $y_{12} = z$ with $y_{11} = x_1 - z$. Then $N_{\mathbf{x}} = 126$ with the possible values of $z = 0, 1, \dots, x_1 (= 125)$ and

$$\pi(z|\mathbf{x}) \propto \frac{1}{(x_1-z)!z!} 2^{-z} \frac{\Gamma(z+x_4+1)}{\Gamma(z+x_2+x_3+x_4+2)}. \tag{3.3}$$

Then (3.3) and $\pi(\theta|z,\mathbf{x}) \sim \text{beta}(z+x_4+1, x_2+x_3+1)$ together imply that $\pi(\theta|\mathbf{x})$ is a mixture of beta distributions with

$$\mathbb{E}[\theta|\mathbf{x}] = \sum_{z=0}^{125} \frac{z+x_4+1}{z+x_2+x_3+x_4+2}\pi(z|\mathbf{x}) = 0.6228$$

and $var(\theta|\mathbf{x}) = (0.05094)^2 = 0.002595$. The posterior distribution of θ is plotted in Fig. 1.

In Gelfand and Smith (1990), Sect. 3.1, the genetic linkage example is extended to a model where the data $\mathbf{x}$ is assumed to arise from a Multinomial$(n, (a_1\theta + b_1, a_2\theta + b_2, a_3\eta + b_3, a_4\eta + b_4, c(1-\theta-\eta)))$, where (θ, η) are the unknown parameters of interest, $a_i, b_i \geq 0$ are known and $0 < c = 1 - \sum_{j=1}^{4} b_j = a_1 + a_2 = a_3 + a_4 < 1$. In Gelfand and Smith (1990), a Gibbs sampler is described for obtaining samples from the joint distribution of (θ, η) and this is applied to $\mathbf{x} = (14, 1, 1, 1, 5)$ $(n = 22)$ with probabilities $(\theta/4 + 1/8, \theta/4, \eta/4, \eta/4 + 3/8, (1-\theta-\eta)/2)$. It is easy to see that $\pi(\mathbf{x}|\theta, \eta)$ does not satisfy (3.1) but it is straightforward to extend the above arguments to obtain $\pi(\theta, \eta|\mathbf{x})$. We split categories 1 and 4 into two subcategories leaving the other three categories unaffected. For category 1 (4), observations have probabilities $q_{11} = \theta/4$ $(q_{41} = \eta/4)$ and $q_{12} = 1/8$ $(q_{42} = 3/8)$ of belonging to subcategories 11 (41) and 12 (42), respectively. We augment $\mathbf{x}$ by (z, w), where z and w denote the total number of observations in subcategories 11 and 41, respectively. Note that there are $(14 + 1) \times (1 + 1) = 30$ possibilities for (z, w). With a Dirichlet$(1, 1, 1)$ prior on $(\theta, \eta, 1-\theta-\eta)$, we have that

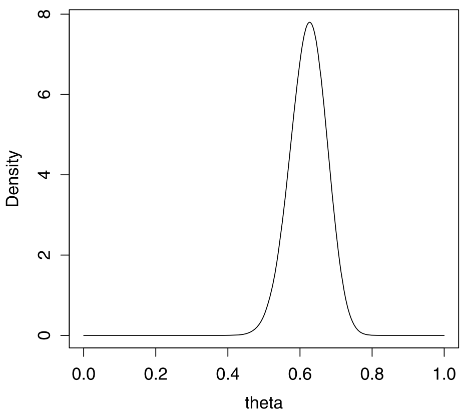

Fig. 1 Exact posterior density of $\pi(\theta|\mathbf{x})$ for the genetics linkage data

$$\pi\big(\theta, \eta, (z, w)|\mathbf{x}\big) \propto \frac{1}{(x_1-z)!z!(x_4-w)!w!} \times \theta^{z+x_2} 2^z \eta^{x_3+w} \left(\frac{2}{3}\right)^w (1-\theta-\eta)^{x_5},$$

which upon integrating out (θ, η) yields

$$\pi\big((z, w)|\mathbf{x}\big) \propto \frac{1}{(x_1-z)!z!(x_4-w)!w!} 2^z \left(\frac{2}{3}\right)^w \times \frac{(z+x_2)!(x_3+w)!}{(z+x_2+x_3+w+x_5+2)!}.$$

Given (z, w), the (conditional) posterior distribution of $(\theta, \eta, 1-\theta-\eta)$ is Dirichlet$(z+x_2+1, x_3+w+1, x_5+1)$. The marginal posterior means (standard deviations) of θ and η are 0.5200 (0.1333) and 0.1232 (0.0809), respectively, with the posterior correlation between θ and η equal to -0.1049.

3.3 Household Reed-Frost epidemic model

The model presented here is the specialization of the household epidemic model of Addy et al. (1991) to a constant infectious period, the Reed-Frost epidemic model. The data consists of the final outcome of an epidemic amongst a population of individuals partitioned into households. Let t denote the maximum household size. Then there are t types of observations corresponding to households with $h = 1, 2, \dots, t$ members. For type h, there are $k_h = h + 1$ categories corresponding to $i = 0, 1, \dots, h$ members of the household ultimately being infected with the disease. The data is assumed to arise as follows. Each individual in the population has independent probability $1 - q_G$ of being infected from outside their household, which we shall term a global infection. (Thus q_G is the probability that an individual avoids a global infection.) If an individual is ever infected, it attempts to infect susceptible members of its household, whilst infectious, before recovering from the disease and entering the removed class. Each infectious individual has probability $1 - q_L$ of infecting each given member of their household, and therefore makes $\mathrm{Bin}(H-1, 1-q_L)$ local infectious contacts if they belong to a household of size H. Infectious contacts with susceptibles result in infection, whereas infectious contacts with non-susceptibles have no impact on the recipient. Thus $\theta = (q_G, q_L)$.

It is trivial to show that, for $h = 0, 1, \ldots$, $p_{h0} = q_G^h$ and $p_{h1} = h(1-q_G)q_G^{h-1}q_L^{h-1}$, and hence, $m_{h0} = m_{h1} = 1$. For $i \geq 2$, p_{hi} can be obtained recursively using Addy et al. (1991), Theorem 2 or more conveniently for the above setting using Ball et al. (1997), (3.12). For any i and $h \geq i$, $m_{hi} = m_{ii}$ with $m_{22} = 2$, $m_{33} = 5$, $m_{44} = 13$ and $m_{55} = 33$. To see how m_{hi} is obtained it is useful to construct the within-household epidemic on a generation basis. Let generation 0 denote those individuals in the household infected globally, and for $l \geq 1$, let generation l denote those individuals infected by the infectives in generation $l - 1$. Then $\{a_0, a_1, \ldots, a_b\}$ denotes that there are a_c infectives in generation c with $a_{b+1} = 0$, *i.e.* the epidemic in the household has ended after generation b. The final size of the epidemic in the household is $\sum_{c=0}^{b} a_c$. For example, there are four ways that $\sum_{c=0}^{b} a_c = 3$, $\{1, 1, 1\}$, $\{1, 2\}$, $\{2, 1\}$ and $\{3\}$. For $\{2, 1\}$, this can arise from either one or both of the generation 0 individuals attempting to infect the generation 1 individual and we denote these two possibilities as $\{2, 1^1\}$ and $\{2, 1^2\}$, respectively. Thus the category $(h, 3)$ is split into five sub-categories corresponding to infection chains $\{1, 1, 1\}$, $\{1, 2\}$, $\{2, 1^1\}$, $\{2, 1^2\}$ and $\{3\}$. For the category $(h, 4)$, the 13 sub-categories are $\{1, 1, 1, 1\}$, $\{1, 1, 2\} \cup \{1, 2, 1^1\}$, $\{1, 2, 1^2\}$, $\{1, 3\}$, $\{2, 2^2\}$, $\{2, 2^3\}$, $\{2, 2^4\}$, $\{2, 1^1, 1\}$, $\{2, 1^2, 1\}$, $\{3, 1^1\}$, $\{3, 1^2\}$, $\{3, 1^3\}$ and $\{4\}$, where, for example, $\{2, 2^3\}$ denotes that there are 3 attempted infections between the 2 infectives in generation 0 and the 2 infectives in generation 1. Note that the outcomes $\{1, 1, 2\}$ and $\{1, 2, 1^1\}$ are combined into a single sub-category. This is because they have probabilities $(h!/(2(h-4)!))(1-q_G)q_G^{h-1}(1-q_L)^3 q_L^{4h-14}$ and $(h!/(h-4)!)(1-q_G)q_G^{h-1}(1-q_L)^3 q_L^{4h-14}$, respectively. In all cases the sub-category probabilities are of the form $c(1-q_G)^{b_G} q_G^{a_G} (1-q_L)^{b_L} q_L^{a_L}$ as required for Sect. 3.1. The full probabilities for $i \leq 4$ are given in Table 5 in the Appendix.

We applied the household Reed-Frost model to four influenza outbreaks, two outbreaks (Influenza A and Influenza B) from Seattle, Washington in the 1970's, reported in Fox and Hall (1980) and two outbreaks (1977-78, 1980-81) from Tecumseh, Michigan, reported in Monto et al. (1985). In addition, following Addy et al. (1991), Ball et al. (1997) and O'Neill et al. (2000), we also consider the two Tecumseh outbreaks as a combined data set. The data for the four influenza outbreaks are available in both (Clancy and O'Neill 2007) and (Neal 2012). The posterior means and standard deviations of q_L and q_G for each data set are recorded in Table 1 along with N, $S = |\mathcal{S}|$ and the time taken to compute the exact posterior distribution using Fortran95 on a computer with a dual 1 GHz Pentium III processor. The differences between S and N are dramatic showing a clear advantage for computing sufficient statistics. Note that for the Tecumseh data sets it is not feasible to compute the exact posterior distribution without computing $\mathcal{S}$. For all the data sets except the combined Tecumseh data set computation of the posterior distribution is extremely fast. In Fig. 2, contour plots of the exact joint posterior distribution of (q_G, q_L) are given for the Seattle, Influenza A and Influenza B outbreaks.

There are alternatives to computing the exact posterior distribution, for example, MCMC, either the Gibbs sampler or the random walk Metropolis algorithm, (O'Neill et al. 2000), or rejection sampling, (Clancy and O'Neill 2007). The rejection sampler has the advantage over MCMC of producing independent and identically distributed (perfect) samples from the posterior distribution, but incurs considerable overheads in finding an efficient proposal distribution and bounding constant. A Gibbs sampler algorithm can easily be constructed using the data augmentation scheme given above. However, the simplest approach is to construct a random walk Metropolis algorithm to obtain a sample from $\pi(\boldsymbol{\theta}|\mathbf{x})$ without data augmentation using the recursive equations for $\{p_{hi}\}$ given by Ball et al. (1997), (3.12).

We assessed the performance of computing the exact posterior distribution with both the Gibbs sampler and the random walk Metropolis algorithms. The MCMC algorithms were run for 11000 iterations for the Gibbs sampler and 110000 iterations for the random walk Metropolis algorithm with the first 1000 and 10000 iterations discarded as burn-in, respectively. This was sufficient for accurate estimates of all the summary statistics recorded in Table 1. For a fair comparison the MCMC algorithms were also run in Fortran95 on the same computer with the run times taking between two and thirty seconds. Therefore, with the exception of the full Tecumseh data set, the computation time of the exact posterior distribution compares favourably with obtaining moderate sized MCMC samples. However, since the computation of the exact posterior distribution grows exponentially with the total number of observations (households), its applicability is limited by the size of the data. In contrast to the MCMC algorithms which are not affected so long as $|\mathbf{x}|$ remains constant with the number of computations per iteration for the MCMC algorithms depending upon $|\mathbf{x}|$, the total number of categories. It should be noted though, that the individual Tecumseh data sets are not small consisting

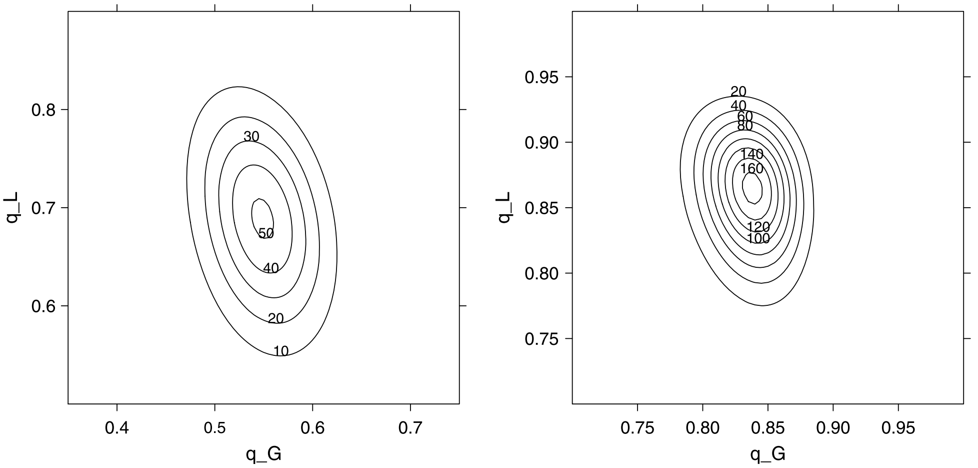

Fig. 2 Exact posterior density contour plot of $\pi(q_L, q_G|\mathbf{x})$ for the Seattle, Influenza A outbreak (*left*) and the Seattle, Influenza B outbreak (*right*)

Table 1 Posterior means and standard deviations for q_G and q_L for the Seattle and Tecumseh influenza data sets

Data set	N	S	Time	$E[q_G|\mathbf{x}]$ $(sd(q_G|\mathbf{x}))$	$E[q_L|\mathbf{x}]$ $(sd(q_L|\mathbf{x}))$
Seattle, A	13860	651	<1 s	0.5476 (0.0420)	0.6859 (0.0742)
Seattle, B	157500	448	<1 s	0.8354 (0.0247)	0.8585 (0.0390)
Tecumseh, 77-78	3.105×10^{10}	20073	7 s	0.8534 (0.0142)	0.8590 (0.0251)
Tecumseh, 80-81	9.081×10^{9}	26504	9 s	0.8720 (0.0132)	0.8439 (0.0269)
Tecumseh, Combined	2.820×10^{20}	519531	57 m	0.8681 (0.0099)	0.8498 (0.0186)

of nearly 300 households each, so the exact posterior distribution offers a good alternative to MCMC, for small-to-moderate data sets, for computing posterior summary statistics, in addition to the advantages of the exact distribution listed in Sect. 1.

4 Integer valued AR processes

4.1 Introduction

The second class of model is the integer-valued autoregressive (INAR) process, see McKenzie (2003), McCabe and Martin (2005) and Neal and Subba Rao (2007). An integer-valued time series $\{X_t; -\infty < t < \infty\}$ is an *INAR*(p) process if it satisfies the difference equation:

$$X_t = \sum_{i=1}^{p} \alpha_i \circ X_{t-i} + Z_t, \quad t \in \mathbb{Z},$$

for some generalized, Steutel and van Harn operators $\alpha_i \circ$ $(i = 1, 2, \ldots, p)$, see Latour (1997), and Z_t $(-\infty < t < \infty)$ are independent and identically distributed according to an arbitrary, but specified, non-negative integer-valued random variable. Generally, see, for example, Neal and Subba Rao (2007), $Z_t \sim \text{Po}(\lambda)$, and the operators $\alpha_i \circ$ are taken to be binomial operators with

$$\alpha_i \circ w = \begin{cases} \text{Bin}(w, \alpha_i) & w > 0, \\ 0 & w = 0. \end{cases} \tag{4.1}$$

This is the situation we primarily focus on here, with $\boldsymbol{\theta} = (\alpha_1, \alpha_2, \ldots, \alpha_p, \lambda)$. In Sect. 4.3, we also consider $Z_t \sim \text{Geom}(\beta)$, see, McCabe and Martin (2005). In Neal and Subba Rao (2007) and Enciso-Mora et al. (2009a), MCMC is used to obtain samples from the posterior distribution of the parameters of *INARMA*(p, q), whereas in McCabe and Martin (2005) numerical integration is used to compute the posterior distribution of *INAR*(1) models with Poisson, binomial and negative binomial models. Although there are similarities between the current work and McCabe and Martin (2005) in computing the exact posterior distribution, it should be noted that McCabe and Martin (2005) is only practical for an *INAR*(1) model with very low counts (for the data considered in McCabe and Martin (2005), $\max_t x_t = 2$) and depends upon the gridpoints used for the numerical integration. Maximum likelihood estimation and model selection for *INAR*(p) models and their extensions are considered in Bu and McCabe (2008), Bu et al. (2008) and Enciso-Mora et al. (2009a), Sect. 4.

We follow Neal and Subba Rao (2007), Sect. 3.1, in the data augmentation step and construction of the likelihood. The observed data consists of $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{x}_I = (x_{1-p}, x_{2-p}, \ldots, x_0)$, and we compute $\pi(\boldsymbol{\theta}|\mathbf{x}_I, \mathbf{x})$.

(The representation of the data is the same as in Sect. 2.) For $t \in \mathbb{Z}$, we represent $\alpha_i \circ X_{t-i}$ by $Y_{t,i}$. Note that given $\mathbf{Y}_t = (Y_{t,1}, Y_{t,2}, \dots, Y_{t,p})$, we have that $Z_t = X_t - \sum_{i=1}^p Y_{t,i}$. For $t \ge 1$, let $\mathbf{y}_t = (y_{t,1}, y_{t,2}, \dots, y_{t,p})$ denote a realization of $\mathbf{Y}_t$. (If $p = 1$, we simply write Y_t and y_t in place of $Y_{t,1}$ and $y_{t,1}$, respectively.) Then it is shown on p. 96 of Neal and Subba Rao (2007), that for $t \ge 1$,

$$\begin{aligned}
&\mathbb{P}(X_t = x_t, \mathbf{Y}_t = \mathbf{y}_t | \mathbf{x}_{t-1}, \mathbf{x}_I, \boldsymbol{\theta}) \\
&\quad \propto 1_{\{\sum_{i=1}^p y_{t,i} \le x_t\}} \prod_{i=1}^p \left\{ \binom{x_{t-i}}{y_{t,i}} \alpha_i^{y_{t,i}} (1-\alpha_i)^{x_{t-i}-y_{t,i}} \right\} \\
&\qquad \times \lambda^{x_t - \sum_{i=1}^p y_{t,i}} \exp(-\lambda) \\
&\quad \propto \left\{ \left\{ 1_{\{\sum_{i=1}^p y_{t,i} \le x_t\}} \prod_{i=1}^p \frac{1}{y_{t,i}!(x_{t-i} - y_{t,i})!} \right\} \right. \\
&\qquad \left. \times \frac{1}{(x_t - \sum_{i=1}^p y_{t,i})!} \right\} \\
&\qquad \times \prod_{i=1}^p \alpha_i^{y_{t,i}} (1-\alpha_i)^{x_{t-i}-y_{t,i}} \lambda^{x_t - \sum_{i=1}^p y_{t,i}} \exp(-\lambda) \\
&\quad = c_{\mathbf{y}_t, \mathbf{x}} \prod_{i=1}^p \alpha_i^{y_{t,i}} (1-\alpha_i)^{x_{t-i}-y_{t,i}} \lambda^{x_t - \sum_{i=1}^p y_{t,i}} \exp(-\lambda), \\
&\quad \text{say}, \qquad (4.2)
\end{aligned}$$

with

$$\begin{aligned}
&\mathbb{P}(\mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y} | \theta, \mathbf{x}_I) \\
&\quad = \prod_{t=1}^n \mathbb{P}(X_t = x_t, \mathbf{Y}_t = \mathbf{y}_t | \mathbf{x}_{t-1}, \mathbf{x}_I, \theta). \qquad (4.3)
\end{aligned}$$

The form of (4.2) and (4.3) satisfies (2.8) with $h_{t y_t i}(\alpha_i) = \alpha_i^{y_{t,i}} (1-\alpha_i)^{x_{t-i}-y_{t,i}}$ $(i = 1, 2, \dots, p)$ and $h_{t y_t (p+1)}(\lambda) = \lambda^{x_t - \sum_{i=1}^p y_{t,i}} \exp(-\lambda)$. Therefore we assign beta distributed priors for the α_i's, and a gamma distributed prior for λ. We differ slightly from Neal and Subba Rao (2007), in assuming independent uniform priors for the α_i's, instead of including the stationary condition that $\sum_{i=1}^p \alpha_i < 1$. The inclusion of the constraint $\sum_{i=1}^p \alpha_i < 1$ had no affect on the examples considered in Neal and Subba Rao (2007) which were all clearly stationary, and without the constraint we can easily obtain $\pi(\mathbf{y}|\mathbf{x}, \mathbf{x}_I)$. We take a gamma$(a_\lambda, b_\lambda)$ prior for λ and for ease of exposition set $a_\lambda = 1$.

We now depart from Neal and Subba Rao (2007), who used the above data augmentation within a Gibbs sampler, by integrating out $\boldsymbol{\theta}$ and identifying the sufficient statistics. For $i = 0, 1, \dots, p$, let $K_i = \sum_{t=1}^n x_{t-i}$ and for $i = 1, 2, \dots, p$, let $G_i(\mathbf{y}) = \sum_{t=1}^n y_{t,i}$. Then

$$\begin{aligned}
&\pi(\theta, \mathbf{y}|\mathbf{x}, \mathbf{x}_I) \\
&\quad = \prod_{t=1}^n \left\{ 1_{\{\sum_{i=1}^p y_{t,i} \le x_t\}} \prod_{i=1}^p \left\{ \binom{x_{t-i}}{y_{t,i}} \alpha_i^{y_{t,i}} (1-\alpha_i)^{x_{t-i}-y_{t,i}} \right\} \right. \\
&\qquad \left. \times \frac{\lambda^{x_t - \sum_{i=1}^p y_{t,i}}}{(x_t - \sum_{i=1}^p y_{t,i})!} \exp(-\lambda) \right\} \times e^{-b_\lambda \lambda} \\
&\quad = \prod_{t=1}^n \left\{ \left\{ 1_{\{\sum_{i=1}^p y_{t,i} \le x_t\}} \prod_{i=1}^p \frac{x_{t-i}!}{y_{t,i}!(x_{t-i} - y_{t,i})!} \right\} \right. \\
&\qquad \left. \times \frac{1}{(x_t - \sum_{i=1}^p y_{t,i})!} \right\} \\
&\qquad \times \prod_{i=1}^p \alpha_i^{G_i(\mathbf{y})} (1-\alpha_i)^{K_i - G_i(\mathbf{y})} \lambda^{K_0 - \sum_{i=1}^p G_i(\mathbf{y})} \\
&\qquad \times \exp\bigl(-(n + b_\lambda)\lambda\bigr).
\end{aligned}$$

Integrating out $\boldsymbol{\theta}$ yields

$$\begin{aligned}
\pi(\mathbf{y}|\mathbf{x}, \mathbf{x}_I) &\propto \prod_{t=1}^n \left\{ \left\{ 1_{\{\sum_{i=1}^p y_{t,i} \le x_t\}} \prod_{i=1}^p \frac{x_{t-i}!}{y_{t,i}!(x_{t-i} - y_{t,i})!} \right\} \right. \\
&\qquad \left. \times \frac{1}{(x_t - \sum_{i=1}^p y_{t,i})!} \right\} \\
&\qquad \times \prod_{i=1}^p \frac{G_i(\mathbf{y})!(K_i - G_i(\mathbf{y}))!}{(K_i + 1)!} \\
&\qquad \times \frac{(K_0 - \sum_{i=1}^p G_i(\mathbf{y}))!}{(n + b_\lambda)^{K_0 - \sum_{i=1}^p G_i(\mathbf{y}) + 1}}. \qquad (4.4)
\end{aligned}$$

We note that for $i = 1, 2, \dots, p$, $\alpha_i | \mathbf{y}, \mathbf{x}, \mathbf{x}_I \sim \text{beta}(G_i(\mathbf{y}) + 1, K_i - G_i(\mathbf{y}) + 1)$ and $\lambda | \mathbf{y}, \mathbf{x}, \mathbf{x}_I \sim \text{gamma}(K_0 - \sum_{i=1}^p G_i(\mathbf{y}) + 1, n + b_\lambda)$. Thus $\mathbf{G}(\mathbf{y}) = (G_1(\mathbf{y}), G_2(\mathbf{y}), \dots, G_p(\mathbf{y}))$ are sufficient statistics for the *INAR*(p) model.

For $\mathbf{g} \in \prod_{i=1}^p [0, K_i]$, let $C_\mathbf{g} = \sum_{\mathbf{y} \in \mathcal{S}_\mathbf{g}} \{\prod_{t=1}^n c_{t, \mathbf{y}_t, \mathbf{x}}\}$, where from (4.4),

$$\begin{aligned}
c_{t, \mathbf{y}_t, \mathbf{x}} = &\left\{ \left\{ 1_{\{\sum_{i=1}^p y_{t,i} \le x_t\}} \prod_{i=1}^p \frac{x_{t-i}!}{y_{t,i}!(x_{t-i} - y_{t,i})!} \right\} \right. \\
&\left. \times \frac{1}{(x_t - \sum_{i=1}^p y_{t,i})!} \right\} \qquad (4.5)
\end{aligned}$$

and $\mathcal{S}_\mathbf{g} = \{\mathbf{y} : \mathbf{G}(\mathbf{y}) = \mathbf{g}\}$. The construction of $\mathcal{S}_\mathbf{g}$ and $\{C_{\mathbf{g}_j}\}$ is then straightforward following the sequential procedure outlined in Sect. 2. Thus, for $j = 1, 2, \dots, S$, with $\mathbf{g}_j =$

Table 2 Posterior means (standard deviations) of the parameters and marginal log-likelihood for *INAR*(p) model for Westgren's data set

Model	α	λ	marginal log-likelihood	No. of categories
INAR(0)	–	1.5462 (0.0648)	−573.7400	1
INAR(1)	0.5302 (0.0360)	0.7262 (0.0636)	−521.5827	409
INAR(2)	0.4616,0.1861 (0.0478,0.0531)	0.5527 (0.0724)	−513.1185	120264
INAR(3)	0.4558,0.1489,0.1010 (0.0481,0.0568,0.0499)	0.4608 (0.0778)	−511.2693	25263253

$(g_{j1}, g_{j2}, \ldots, g_{jp})$, we have that

$$\pi(S_{\mathbf{y},\mathbf{x}} = \mathbf{g}_j|\mathbf{x}, \mathbf{x}_I) \propto C_{\mathbf{g}_j} \prod_{i=1}^{p} g_{ji}!(K_i - g_{ji})! \times \frac{(K_0 - \sum_{i=1}^{p} g_{ji})!}{(n+b_\lambda)^{K_0 - \sum_{i=1}^{p} g_{ji}+1}}.$$

For $p = 1$, $S = (K_1 + 1)$ and $\pi(S(\mathbf{y}, \mathbf{x}) = \mathbf{g}_j|\mathbf{x}, \mathbf{x}_I)$ $(j = 1, 2, \ldots, S)$ can be computed very quickly using the ordering of the outcomes $\mathbf{g}_1 = 0, \mathbf{g}_2 = 1, \ldots, \mathbf{g}_S = K_1$. For $p > 1$, computation of $S \leq \prod_{i=1}^{p}(K_i + 1)$ and $\pi(S(\mathbf{y}, \mathbf{x}) = \mathbf{g}_j|\mathbf{x}, \mathbf{x}_I)$ $(j = 1, 2, \ldots, N)$ is more time consuming but still feasible for $p = 2, 3$ as demonstrated in Sect. 4.2.

4.2 Westgren's gold particle data set

This data set consists of 380 counts of gold particles at equidistant points in time. The data, originally published in Westgren (1916), has been analysed in Jung and Tremayne (2006) and Neal and Subba Rao (2007) using an *INAR*(2) model. In Enciso-Mora et al. (2009a), it was shown using reversible jump MCMC that the *INAR*(2) model is the most appropriate model of *INARMA*(p, q) type for the data.

We obtain the exact posterior distribution for the parameters of the *INAR*(p) model ($p = 0, 1, 2, 3$). (The *INAR*(0) corresponds to independent and identically distributed Poisson data.) We also computed the marginal log-likelihood (evidence) for the models for comparison. The results are presented in Table 2. There is clear evidence from the marginal log-likelihood for $p = 2$. Applying the BIC-based penalisation prior used in Enciso-Mora et al. (2009a), where for $p = 0, 1, 2, 3$ the prior on *INAR*(p) was set proportional to $n^{-p/2}$ gives posterior probabilities of 0.0030, 0.7494 and 0.2476, for the *INAR*(1), *INAR*(2) and *INAR*(3) models, respectively. The total number of categories grows rapidly with the order p of the model, and to compute $\{\pi(\mathbf{y}|\mathbf{x})\}$ using Fortran95 for the *INAR*(1), *INAR*(2) and *INAR*(3) models, took less than a second, 8 seconds and 45 minutes, respectively. It should be noted that the *INAR*(3) model was at the limits of what is computationally feasible requiring over 1500 MB of computer memory. The memory limitation is due to the total number of categories, and for smaller data sets, either in terms of n, or the magnitude of x_t's, it would be possible to study *INAR*(p) models with $p > 3$. However, most interest in *INAR*(p) models is when $p \leq 3$.

For comparison the MCMC algorithms of Neal and Subba Rao (2007) and Enciso-Mora et al. (2009a) were run on the gold particle data set with findings similar to those presented above and reported in Neal and Subba Rao (2007) and Enciso-Mora et al. (2009a). For fixed $p = 1, 2, 3$, the MCMC algorithm of Neal and Subba Rao (2007) was run for 110000 iterations with the first 10000 iterations discarded as burn-in. The algorithm took about 30 seconds to run, regardless of the value of p, although the mixing of the MCMC algorithm gets worse as p increases. Therefore for analysis of the *INAR*(1) and *INAR*(2) models, it is quicker to compute the exact posterior distribution than to obtain a sufficiently large sample from an MCMC algorithm. For model selection the reversible jump MCMC algorithm of Enciso-Mora et al. (2009a) was used, restricted to the *INAR*(p) models with $p = 1, 2$ or 3. The algorithm took approximately twice as long, per iteration, as the standard algorithm, due to the model switching step. Also for accurate estimation of the posterior model probabilities longer runs (10^6 iterations) of the algorithm are required due the low acceptance rate (0.4 %) of the model switching move. This is particularly the case for the estimation of the posterior model probability of the *INAR*(1) model, with typically only 1 or 2 excursions to this model in 10^6 iterations. Hence for model selection, using the exact posterior distribution is preferable to reversible jump MCMC, especially for computing the posterior probability for rarer models as noted in Sect. 1.

The exact one-step ahead predictive distribution $\mathbb{P}(X_{n+1} = x_{n+1}|\mathbf{x})$ can easily be obtained and then the r-step $(r = 2, 3, \ldots)$ ahead predictive distribution can be obtained recursively. In particular for the *INAR*(2) model, integrating over the posterior distribution of (α, λ), we get that

$$\mathbb{P}(X_{n+1} = x_{n+1}|\mathbf{x}, \mathbf{x}_I) = \sum_{j=1}^{S} C_{\mathbf{g}_j} \sum_{y_{n+1,1}} \sum_{y_{n+1,2}} \left(1_{\{y_{n+1,1}+y_{n+1,2} \leq x_{n+1}\}} \binom{x_n}{y_{n+1,1}} \times \binom{x_{n-1}}{y_{n+1,2}} \right.$$

Table 3 Marginal predictive distribution of X_{371} and X_{372} for the *INAR*(2) model for Westgren's data set

Observation	$P(X_t = x)$									
	0	1	2	3	4	5	6	7	8	≥9
X_{371}	0.0509	0.1870	0.2927	0.2566	0.1415	0.0532	0.0144	0.0030	0.0005	0.0001
X_{372}	0.0934	0.2340	0.2792	0.2114	0.1143	0.0471	0.0154	0.0041	0.0009	0.0002

Table 4 Posterior means (standard deviations) of the parameters and marginal log-likelihood for the models for US polio data set

Model	α	λ	β	marginal log-likelihood
PINAR	0.1880 (0.0467)	1.0926 (0.0949)	–	−287.8742
GINAR	0.0977 (0.0494)	–	0.4528 (0.0284)	−263.6767
IID Geometric data	–	–	0.4289 (0.0249)	−270.4720

$$\times \prod_{i=1}^{2}\left\{\frac{(K_i+1)!}{g_{ji}!(K_i-g_{ji})!}\right.$$

$$\left.\times \frac{(g_{ji}+y_{n+1,i})!(K_i+x_{n+1-i}-y_{n+1,i}-g_{ji})!}{(K_i+x_{n+1-i}+1)!}\right\}$$

$$\times \frac{(n+b_\lambda)^{K_0-\sum_{i=1}^{2} g_{ji}+1}}{(K_0-\sum_{i=1}^{2} g_{ji})!}$$

$$\left.\times \frac{(K_0-\sum_{i=1}^{2} g_{ji}+x_{n+1}-\sum_{l=1}^{2} y_{n+1,l})!}{(n+1+b_\lambda)^{K_0-\sum_{i=1}^{2} g_{ji}+x_{n+1}-\sum_{l=1}^{2} y_{n+1,l}+1}}\right)$$

$$\Bigg/ \sum_{k=1}^{S} C_{\mathbf{g}_k}.$$

We compute the predictive distributions of X_{371} and X_{372} with the results given in Table 3. For X_{372}, we have to sum over all possibilities for X_{371}, of which there are infinitely many. However, we restrict the summation to $X_{371} = 0, 1, \ldots, 10$, since $\mathbb{P}(X_{371} > 10) \leq 10^{-6}$ and restricting the sum does not affect the computation of the probabilities to four decimal places as presented in Table 3. Given the focus of this paper, we did not evaluate the predictions or the adequacy of the *INAR*(p) model.

4.3 US polio data set

The US polio data set, from Zeger (1988), consists of the total number of monthly polio (Poliomyelitis) cases in the USA from January 1970 to December 1983. The data has been studied by Zeger (1988) and Davis et al. (2000) using Poisson regression models, and (Enciso-Mora et al. 2009b) using an *INAR*(1) model with covariates. We consider two simple *INAR*(1) models for the polio data. The first model has Poisson innovations ($Z_t \sim \text{Po}(\lambda)$) as in Sect. 4.2 and the second model has geometric innovations ($Z_t \sim \text{Geom}(\beta), \mathbb{P}(Z_t = k) = (1-\beta)^k \beta (k = 0, 1, \ldots)$). We denote the models by *PINAR* (Poisson) and *GINAR* (Geometric), respectively. Although the more complex models, taking into account seasonality and trend, studied in Zeger (1988), Davis et al. (2000) and Enciso-Mora et al. (2009b), are probably more appropriate for this data, this example allows us to further demonstrate the scope of the approach taken in this paper.

A geometric, or more generally a negative binomial or binomial, innovation distribution falls under the auspices of a model with multinomial-beta data studied in Sect. 3 with

$$\mathbb{P}(X_t = x_t, Y_t = y_t | \alpha, \beta, X_{t-1} = x_{t-1})$$
$$= 1_{\{y_t \leq x_t\}} \binom{x_{t-1}}{y_t} \alpha^{y_t} (1-\alpha)^{x_{t-1}-y_t} (1-\beta)^{x_t - y_t} \beta,$$

and

$$\pi(\mathbf{y}|\mathbf{x}, \mathbf{x}_I)$$
$$\propto \prod_{t=1}^{n} \left\{ 1_{\{y_t \leq x_t\}} \binom{x_{t-1}}{y_t} \right\} \frac{G_1(\mathbf{y})!(K_1 - G_1(\mathbf{y}))!}{(K_1+1)!}$$
$$\times \frac{n!(K_0 - G_1(\mathbf{y}))!}{(n + K_0 + 1 - G_1(\mathbf{y}))!}.$$

The total number of possibilities for $G_1(\mathbf{y})$ for the *INAR*(1) models is 101, and results were obtained instantaneously using, either Fortran95 or R. The results are summarised in Table 4, and show that there is far stronger evidence for the *GINAR* model than the *PINAR* model. This is not surprising given the sudden spikes in the data with extreme values far more likely from a Geometric distribution than a Poisson distribution with the same mean. The posterior mean of α for the *GINAR* model is only 0.0977 and this suggests that an appropriate model could be to assume that the data are independent and identically distributed according to a Geometric distribution. However, as shown in

Table 4, such a model has a significantly lower marginal log-likelihood than the *GINAR* model, so the dependence in the data can not be ignored.

5 Conclusions

We have presented a generic framework for obtaining the exact posterior distribution $\pi(\theta|\mathbf{x})$ using data augmentation, $\mathbf{y}$. The models which can be analysed in this way are generally fairly simple, Poisson mixtures (Fearnhead 2005), household Reed-Frost epidemic model and $INAR(p)$, but have been widely applied, and offer a useful benchmark for comparing more complex models against. The computation time for the exact posterior distribution compares favourably to MCMC especially for smaller data sets with all the computations taking less than 15 seconds in Fortran95 on a computer with a dual 1 GHz Pentium III processor with the exception of the combined Tecumseh data set and the $INAR(3)$ model. Moreover, the genetics linkage, Seattle data sets and $INAR(1)$ models took less than 15 seconds to compute the posterior distribution in R on a standard desktop computer. The key elements in the feasibility of the method are the identification of sufficient statistics $S(\mathbf{y}, \mathbf{x})$ and the easy computation of $\pi(S(\mathbf{y}, \mathbf{x})|\mathbf{x})$.

This paper has focused upon the case where the data $\mathbf{x}$ correspond to discrete outcomes, for reasons discussed in Sect. 2.4. Thus, throughout the examples in Sects. 3 and 4, it has been assumed that the data have arisen from mixtures or sums of discrete distributions, such as the binomial, negative binomial and Poisson distributions. However, the approach taken can be extended to models based upon other discrete distributions, such as the discrete uniform distribution. The key requirement is that an appropriate choice of $\mathbf{y}$ can be found such that, $\pi(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y})$ belongs to a well-known probability distribution, typically the product of independent univariate densities, and that $\pi(\mathbf{y}|\mathbf{x})$ can easily be obtained by integrating out $\boldsymbol{\theta}$.

The methodology presented in this paper is not restricted to models satisfying (2.7) with change-point models, (Fearnhead 2006), being a prime example. A similar approach to Sect. 3 can be used to obtain the exact (joint) posterior distribution of the parameters and change-point for the Markov change-point model of Carlin et al. (1992), Sect. 5. Ongoing research involves extending the work in Sect. 4 to $INARMA(p, q)$ processes.

Acknowledgements We would like to thank the two anonymous referees and associate editor for their helpful comments which helped in preparing the revisions of this paper.

Table 5 Household epidemic outcomes for up to 4 infected individuals

Sub-category	c	a_G	b_G	a_L	b_L
$\{0\}$	1	h	0	0	0
$\{1\}$	h	$h-1$	1	$h-1$	0
$\{1, 1\}$	$\frac{h!}{(h-2)!}$	$h-1$	1	$2h-4$	1
$\{2\}$	$\frac{h!}{2!(h-2)!}$	$h-2$	2	$2h-4$	1
$\{1, 1, 1\}$	$\frac{h!}{(h-3)!}$	$h-1$	1	$3h-8$	2
$\{1, 2\}$	$\frac{h!}{2(h-3)!}$	$h-1$	1	$3h-9$	2
$\{2, 1^1\}$	$\frac{h!}{(h-3)!}$	$h-2$	2	$3h-8$	1
$\{2, 1^2\}$	$\frac{h!}{2(h-3)!}$	$h-2$	2	$3h-9$	2
$\{3\}$	$\frac{h!}{3!(h-3)!}$	$h-3$	3	$3h-9$	0
$\{1, 1, 1, 1\}$	$\frac{h!}{(h-4)!}$	$h-1$	1	$4h-13$	3
$\{1, 1, 2\} \cup \{1, 2, 1^1\}$	$\frac{3}{2} \times \frac{h!}{(h-4)!}$	$h-1$	1	$4h-14$	3
$\{1, 2, 1^2\}$	$\frac{h!}{2(h-4)!}$	$h-1$	1	$4h-15$	4
$\{1, 3\}$	$\frac{h!}{3!(h-4)!}$	$h-1$	1	$4h-16$	3
$\{2, 2^2\}$	$\frac{h!}{(h-4)!}$	$h-2$	2	$4h-14$	2
$\{2, 2^3\}$	$\frac{h!}{(h-4)!}$	$h-2$	2	$4h-15$	3
$\{2, 2^4\}$	$\frac{h!}{4(h-4)!}$	$h-2$	2	$4h-16$	4
$\{2, 1^1, 1\}$	$\frac{h!}{(h-4)!}$	$h-2$	2	$4h-13$	2
$\{2, 1^2, 1\}$	$\frac{h!}{2(h-4)!}$	$h-2$	2	$4h-14$	3
$\{3, 1^1\}$	$\frac{h!}{2(h-4)!}$	$h-3$	3	$4h-14$	1
$\{3, 1^2\}$	$\frac{h!}{2(h-4)!}$	$h-3$	3	$4h-15$	2
$\{3, 1^3\}$	$\frac{h!}{3!(h-4)!}$	$h-3$	3	$4h-16$	3
$\{4\}$	$\frac{h!}{4!(h-4)!}$	$h-4$	4	$4h-16$	0

Appendix: Household epidemic probabilities

We summarise the sub-categories and associated probabilities for up to and including 4 individuals being infected in a household with h denoting the household size. All the probabilities are of the form $c(1-q_G)^{b_G} q_G^{a_G} (1-q_L)^{b_L} q_L^{a_L}$ and we list c, a_G, b_G, a_L and b_L. Note that $\{0\}$ denotes nobody infected in the household. ,

References

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

Ball, F.G., Mollison, D., Scalia-Tomba, G.: Epidemics with two levels of mixing. Ann. Appl. Probab. **7**, 46–89 (1997)

Bu, R., McCabe, B.: Model selection, estimation and forecasting in $INAR(p)$ models: a likelihood-based Markov chain approach. Int. J. Forecast. **24**, 151–162 (2008)

Bu, R., McCabe, B., Hadri, K.: Maximum likelihood estimation in the $INAR(p)$ model. J. Time Ser. Anal. **29**, 973–994 (2008)

Carlin, B.P., Gelfand, A.E., Smith, A.F.M.: Hierarchical Bayesian analysis of change-point problems. J. R. Stat. Soc. C **41**, 389–405 (1992)

Clancy, D., O'Neill, P.: Exact Bayesian inference and model selection for stochastic models of epidemics among a community of households. Scand. J. Stat. **34**, 259–274 (2007)

Davis, R.A., Dunsmuir, W.T., Wang, Y.: On autocorrelation in a Poisson regression model. Biometrika **87**, 491–505 (2000)

Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. B **39**, 1–38 (1977)

Diebolt, J., Robert, C.P.: Estimation of finite mixture distributions through Bayesian sampling. J. R. Stat. Soc. B **56**, 363–375 (1994)

Enciso-Mora, V., Neal, P., Subba Rao, T.: Efficient order selection algorithms for integer valued ARMA processes. J. Time Ser. Anal. **30**, 1–18 (2009a)

Enciso-Mora, V., Neal, P., Subba Rao, T.: Integer valued AR processes with explanatory variables. Sankhya, Ser. B **71**, 248–263 (2009b)

Fearnhead, P.: Direct simulation for discrete mixture distributions. Stat. Comput. **15**, 125–133 (2005)

Fearnhead, P.: Exact and efficient Bayesian inference for multiple changepoint problems. Stat. Comput. **16**, 203–213 (2006)

Fox, J.P., Hall, C.E.: Viruses in Families. PSG, Littleton (1980)

Frühwirth-Schnatter, S.: Finite Mixture and Markov Switching Models. Springer, Berlin (2006)

Gelfand, A.E., Smith, A.F.M.: Sampling-based approaches to calculating marginal densities. J. Am. Stat. Assoc. **85**, 398–409 (1990)

Geman, S., Geman, D.: Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. **6**, 721–741 (1984)

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

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

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

Jung, R.C., Tremayne, A.R.: Coherent forecasting in integer time series models. Int. J. Forecast. **22**, 223–238 (2006)

Kypraios, T.: Efficient Bayesian inference for partially observed stochastic epidemics and a new class of semi-parametric time series models. PhD thesis, Lancaster University (2007)

Latour, A.: The multivariate GINAR(p) process. Adv. Appl. Probab. **29**, 228–248 (1997)

Liu, J.S.: The collased Gibbs sampler in Bayesian computations with applications to a gene regulation problem. J. Am. Stat. Assoc. **89**, 958–966 (1994)

Longini, I.M., Koopman, J.S.: Househol and community transmission parameters from final distributions of infections in households. Biometrics **38**, 115–126 (1982)

McCabe, B.P.M., Martin, G.M.: Bayesian predictions of low count time series. Int. J. Forecast. **22**, 315–330 (2005)

McKenzie, E.: Discrete variate time series. In: Shanbhag, D.N., Rao, C.R. (eds.) Stochastic Processes: Modelling and Simulation, pp. 573–606. Elsevier, Amsterdam (2003)

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., Teller, E.: Equations of state calculations by fast computing machines. J. Chem. Phys. **21**, 1087–1092 (1953)

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

Neal, P.: Efficient likelihood-free Bayesian computation for household epidemics. Stat. Comput. **22**, 1239–1256 (2012)

Neal, P.J., Roberts, G.O.: A case study in non-centering for data augmentation: stochastic epidemics. Stat. Comput. **15**, 315–327 (2005)

Neal, P.J., Subba Rao, T.: MCMC for integer valued ARMA processes. J. Time Ser. Anal. **28**, 92–110 (2007)

O'Neill, P.D., Roberts, G.O.: Bayesian inference for partially observed stochastic epidemics. J. R. Stat. Soc. A **162**, 121–129 (1999)

O'Neill, P.D., Balding, D.J., Becker, N.G., Eerola, M., Mollison, D.: Analyses of infectious disease data from household outbreaks by Markov chain Monte Carlo methods. J. R. Stat. Soc., Ser. C **49**, 517–542 (2000)

Papaspoliopoulos, O., Roberts, G.O., Sköld, M.: Non-centered parameterisations for hierarchical models and data augmentation. 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. 7, pp. 307–326. Oxford University Press, Oxford (2003)

Propp, J.G., Wilson, D.B.: Exact sampling with coupled Markov chains and applications to statistical mechanics. Random Struct. Algorithms **9**, 223–252 (1996)

Rao, C.R.: Linear Statistical Inference and Its Applications. Wiley, New York (1965)

Smith, A.F.M., Roberts, G.O.: Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. J. R. Stat. Soc. B **55**, 3–23 (1993)

Tanner, M.A., Wong, W.H.: The calculation of posterior distributions by data augmentation. J. Am. Stat. Assoc. **82**, 528–550 (1987)

Westgren, A.: Die Veränderungsgeschwindigkeit der lokalen Teilchenkonzentration in kollioden Systemen (Erste Mitteilung). Ark. Mat. Astron. Fys. **11**, 1–24 (1916)

Zeger, S.: A regression model for time series of counts. Biometrika **75**, 621–629 (1988)