Stat Comput (2009) 19: 479–492
DOI 10.1007/s11222-008-9109-4

Improved auxiliary mixture sampling for hierarchical models of non-Gaussian data

Sylvia Frühwirth-Schnatter · Rudolf Frühwirth · Leonhard Held · Håvard Rue

Received: 29 February 2008 / Accepted: 3 November 2008 / Published online: 4 December 2008

Abstract The article considers Bayesian analysis of hierarchical models for count, binomial and multinomial data using efficient MCMC sampling procedures. To this end, an improved method of auxiliary mixture sampling is proposed. In contrast to previously proposed samplers the method uses a bounded number of latent variables per observation, independent of the intensity of the underlying Poisson process in the case of count data, or of the number of experiments in the case of binomial and multinomial data. The bounded number of latent variables results in a more general error distribution, which is a negative log-Gamma distribution with arbitrary integer shape parameter. The required approximations of these distributions by Gaussian mixtures have been computed. Overall, the improvement leads to a substantial increase in efficiency of auxiliary mixture sampling for highly structured models. The method is illustrated for finite mixtures of generalized linear models and an epidemiological case study.

S. Frühwirth-Schnatter (✉)
Department of Applied Statistics and Econometrics, Johannes Kepler Universität Linz, Linz, Austria
e-mail: sylvia.fruehwirth-schnatter@jku.at

R. Frühwirth
Institute of High Energy Physics, Austrian Academy of Sciences, Vienna, Austria

L. Held
Institute of Social and Preventive Medicine, University of Zurich, Zurich, Switzerland

H. Rue
Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway

Keywords Binomial data · Count data · Finite mixture models · Gaussian mixture · log-Gamma distribution · Multinomial data · Negative binomial distribution

1 Introduction

The article considers Bayesian analysis of hierarchical models for count, binomial and multinomial data using data augmentation and Markov chain Monte Carlo (MCMC) sampling procedures. As first shown by Albert and Chib (1993) in the context of probit regression models with binary outcomes, the introduction of appropriate auxiliary variables allows to recast non-Gaussian problems as conditionally linear Gaussian regression models. The complexity introduced by the auxiliary variables is more than compensated by the ability to sample the parameters in the conditionally Gaussian model in an efficient way.

Recently, this approach has been extended to count data from the Poisson distribution by Frühwirth-Schnatter and Wagner (2006a, 2006b) and to logit models with binary and categorical outcomes by Frühwirth-Schnatter and Frühwirth (2007). The resulting algorithm is called auxiliary mixture sampling, because data augmentation is based partly on approximating the probability density function of the type I extreme value distribution by a mixture of normal distributions. The main motivation for the development of auxiliary mixture sampling has been to simplify MCMC estimation of hierarchical models such as state-space and random-effects models for non-Gaussian data, see also Gschlößl and Czado (2005), Fahrmeir and Steinert (2006), and LeSage et al. (2007). Furthermore, auxiliary mixture sampling has been shown to facilitate Bayesian model selection for non-Gaussian models (Holmes and Held 2006; Frühwirth-Schnatter and Wagner 2008; Tüchler 2008).

Similar in spirit to Albert and Chib (1993), auxiliary mixture sampling uses data augmentation by introducing for each dependent observation y_i latent variables that lead to a conditionally Gaussian model. The number of latent variables per observation differs for the various distribution families, being $2(y_i+1)$ for Poisson data, $2N_i$ for data from a binomial distribution Bin(N_i,π_i), and $2mN_i$ for multinomial data with $m+1$ categories. Thus auxiliary mixture sampling seems to be infeasible for high intensity Poisson data, in particular for panels with a high number N of total observations, or for binomial and multinomial data with a high number of total repetitions $\sum_{i=1}^{N} N_i$.

In this paper we propose an improved method of auxiliary mixture sampling that utilizes a bounded number of latent variables per observation, namely at most four for Poisson data, two for binomial data, and $2m$ for multinomial data. This leads to a substantial increase in efficiency for highly structured hierarchical models. The latent variables of the improved sampler are constructed in such a way that their expectation is a linear function of the unknown parameters as for the original sampler. The deviation from the expectation, however, follows a more general distribution than in Frühwirth-Schnatter and Wagner (2006a) and Frühwirth-Schnatter and Frühwirth (2007), namely the distribution of the negative logarithm of a Gamma random variable with integer shape parameter ν and unit scale. The shape parameter is equal to y_i for Poisson data and to N_i for binomial and multinomial data. For each latent variable this distribution is approximated by a Gaussian finite mixture distribution, and the component indicator is introduced as a further auxiliary variable. Due to the Central Limit Theorem the number of required mixture components drops with rising ν. From the computational point of view, a larger intensity (in the case of count data) or a larger repetition number (in the case of binomial or multinomial data) is therefore an additional advantage.

The improved sampler for count data is described in Sect. 2. It is also shown how to extend it to data from the negative binomial distribution. Section 3 describes the improved sampler for binomial and multinomial data. The sampler is illustrated in Sect. 4 for finite mixtures of generalized linear models and a Bayesian hierarchical model for spatial count data with a Gaussian Markov random field prior. The technical details of the mixture approximation are given in the Appendix.

2 Auxiliary mixture sampling for count data

We present details for the following model. Let $\boldsymbol{y}=(y_1,\ldots,y_N)$ be a sequence of count data, and assume that $y_i|\lambda_i$ is Poisson distributed with parameter λ_i, where λ_i depends on covariates $\boldsymbol{Z}_i=(\boldsymbol{Z}_i^{\alpha},\boldsymbol{Z}_i^{\beta})$ through fixed coefficients $\boldsymbol{\alpha}$ and varying coefficients $\boldsymbol{\beta}_i$:

$$y_i|\lambda_i \sim \mathrm{Po}(\lambda_i), \quad \lambda_i=\exp((\boldsymbol{Z}_i^{\alpha})^T\boldsymbol{\alpha}+(\boldsymbol{Z}_i^{\beta})^T\boldsymbol{\beta}_i). \tag{1}$$

Furthermore, the data are conditionally independent given $\lambda_1,\ldots,\lambda_N$. The precise model for $\boldsymbol{\beta}_i$ is left unspecified at this stage; it could be a spatial or a temporal model, for example. We only assume that the joint distribution $p(\boldsymbol{\alpha},\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_N|\boldsymbol{\theta})$ is known and indexed by some unknown parameter $\boldsymbol{\theta}$.

2.1 Improved auxiliary mixture sampling

2.1.1 Data augmentation

For each i, the distribution of $y_i|\lambda_i$ is regarded as the distribution of the number of jumps of an unobserved Poisson process with intensity λ_i, having occurred in the time interval $0\le t\le 1$. In Frühwirth-Schnatter and Wagner (2006a), the first step of data augmentation creates such a Poisson process for each y_i and introduces the (y_i+1) interarrival times of this Poisson process as latent variables, yielding a total of $2(N+\sum_{i=1}^{N} y_i)$ latent variables once the mixture approximation has been applied.

A more efficient method is based on introducing only two latent variables, namely the arrival time $\tau_{i2}^{\star}$ of the y_ith jump and the interarrival time $\tau_{i1}^{\star}$ between the y_ith jump and the next one. Conditionally on knowing that exactly $y_i>0$ jumps have taken place in the time interval $0\le t\le 1$, $\tau_{i2}^{\star}$ is the arrival time of the last jump before $t=1$, while $\tau_{i2}^{\star}+\tau_{i1}^{\star}$ is the arrival time of the first jump after $t=1$. This result will be used in Sect. 2.1.2 to sample from the posterior distribution of $\tau_{i1}^{\star}$ and $\tau_{i2}^{\star}$ given y_i.

To formulate the augmented model the prior distribution of $\tau_{i1}^{\star}$ and $\tau_{i2}^{\star}$ is required. Evidently, the interarrival time $\tau_{i1}^{\star}$ follows an exponential distribution with mean $1/\lambda_i$:

$$\tau_{i1}^{\star}=\frac{\xi_{i1}}{\lambda_i}, \quad \xi_{i1}\sim \mathrm{Ex}(1), \tag{2}$$

while the arrival time $\tau_{i2}^{\star}$, being the sum of y_i independently exponentially distributed interarrival times, follows a Ga(y_i,λ_i) distribution:

$$\tau_{i2}^{\star}=\frac{\xi_{i2}}{\lambda_i}, \quad \xi_{i2}\sim \mathrm{Ga}(y_i,1). \tag{3}$$

The Ga(a,b) distribution is defined as in Bernardo and Smith (1994), with density $f_{\mathrm{Ga}}(y;a,b)=b^a y^{a-1}e^{-by}/\Gamma(a)$.

Equations (2) and (3) can be reformulated in the following way:

$$-\log\tau_{i1}^{\star}=\log\lambda_i+\varepsilon_{i1}, \tag{4}$$

$$-\log\tau_{i2}^{\star}=\log\lambda_i+\varepsilon_{i2}, \tag{5}$$

where $\varepsilon_{i1} = -\log \xi_{i1}$ and $\varepsilon_{i2} = -\log \xi_{i2}$. For $y_i = 0$ we are dealing only with (4).

The first step of the improved sampler introduces the bivariate latent variable $\tau_i = (\tau_{i1}^\star, \tau_{i2}^\star)$ for each nonzero observation y_i and the single latent variable $\tau_i = \tau_{i1}^\star$ for each zero observation. In the second step, the densities of ε_{i1} and ε_{i2} in (4) and (5) are approximated by Gaussian mixtures, and the latent component indicators $r_i = (r_{i1}, r_{i2})$ are introduced as missing data. For a zero observation this is done only for (4), so that $r_i = r_{i1}$ in this case.

The distribution of ε_{i1} is a type I extreme value distribution and the same mixture approximation as in Frühwirth-Schnatter and Wagner (2006a) can be used. Finding a mixture approximation for ε_{i2} is more challenging because its distribution is a negative log-Gamma distribution with integer shape parameter ν equal to y_i. In Appendix A such an approximation is described for arbitrary integer shape parameters ν:

$$p_\varepsilon(\varepsilon;\nu) = \frac{\exp(-\nu\varepsilon - e^{-\varepsilon})}{\Gamma(\nu)} \approx \sum_{r=1}^{R(\nu)} w_r(\nu)\,\varphi(\varepsilon; m_r(\nu), s_r^2(\nu)), \tag{6}$$

where $\varphi(\varepsilon; m_r(\nu), s_r^2(\nu))$ denotes a normal density. The number of components $R(\nu)$ depends on ν, as do the weights $w_r(\nu)$, the means $m_r(\nu)$ and the variances $s_r^2(\nu)$. For $\nu = 1$ (6) is identical to the mixture approximation derived in Frühwirth-Schnatter and Frühwirth (2007).

Conditional on the latent variables $\boldsymbol{\tau} = \{\tau_1, \ldots, \tau_N\}$ and $\boldsymbol{r} = \{r_1, \ldots, r_N\}$, the nonlinear non-Gaussian model (1) reduces to a linear Gaussian model where the mean of the observation equation is linear in $\boldsymbol{\alpha}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_N$ and the error term follows a normal distribution:

$$\begin{aligned} -\log \tau_{i1}^\star &= \log \lambda_i + m_{r_{i1}}(1) + \varepsilon_{i1}, \\ \varepsilon_{i1}|r_{i1} &\sim \mathrm{No}(0, s_{r_{i1}}^2(1)), \\ -\log \tau_{i2}^\star &= \log \lambda_i + m_{r_{i2}}(y_i) + \varepsilon_{i2}, \\ \varepsilon_{i2}|r_{i2} &\sim \mathrm{No}(0, s_{r_{i2}}^2(y_i)), \end{aligned}$$

with $\log \lambda_i = (\boldsymbol{Z}_i^\alpha)^T \boldsymbol{\alpha} + (\boldsymbol{Z}_i^\beta)^T \boldsymbol{\beta}_i$. For $y_i = 0$ we are dealing only with the first equation.

2.1.2 The sampling scheme

Select starting values for $\boldsymbol{\tau}$ and $\boldsymbol{r}$ and repeat the following steps.

(1) Sample $\boldsymbol{\alpha}$, $\boldsymbol{\beta} = \{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_N\}$, and $\boldsymbol{\theta}$ conditional on $\boldsymbol{\tau}$ and $\boldsymbol{r}$.

(2) Sample the arrival and interarrival times $\boldsymbol{\tau}$ and the component indicators $\boldsymbol{r}$ conditional on $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}$ and $\boldsymbol{y}$ by running the following steps, for $i = 1, \ldots, N$.

(a) Sample $\xi_i \sim \mathrm{Ex}(\lambda_i)$. If $y_i = 0$, set $\tau_{i1}^\star = 1 + \xi_i$. If $y_i > 0$, sample $\tau_{i2}^\star$ from a Beta$(y_i, 1)$-distribution and set $\tau_{i1}^\star = 1 - \tau_{i2}^\star + \xi_i$.

(b) Sample r_{i1} from the following discrete distribution where $k = 1, \ldots, R(1)$:

$$\begin{aligned} &\mathrm{pr}\{r_{i1} = k|\tau_{i1}^\star, \lambda_i\} \\ &\quad \propto w_k(1)\,\varphi(-\log \tau_{i1}^\star - \log \lambda_i; m_k(1), s_k^2(1)). \end{aligned}$$

If $y_i > 0$, sample r_{i2} from the following discrete distribution where $k = 1, \ldots, R(y_i)$:

$$\begin{aligned} &\mathrm{pr}\{r_{i2} = k|\tau_{i2}^\star, \lambda_i\} \\ &\quad \propto w_k(y_i)\,\varphi(-\log \tau_{i2}^\star - \log \lambda_i; m_k(y_i), s_k^2(y_i)). \end{aligned}$$

To obtain starting values for $\boldsymbol{\tau}$ and $\boldsymbol{r}$ we use Step 2 with appropriate starting values $\lambda_1, \ldots, \lambda_N$. If $y_i > 0$, we use the moment estimator $\lambda_i = y_i$; if $y_i = 0$, we set $\lambda_i = \lambda_0$, where $0 < \lambda_0 < 1$. In our applications we have chosen $\lambda_0 = 0.1$.

Step 2 is based on decomposing the joint posterior of $(\boldsymbol{\tau}, \boldsymbol{r})$ as

$$\begin{aligned} &p(\boldsymbol{\tau}, \boldsymbol{r}|\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \\ &= \prod_{i=1}^N \left(\prod_{j=1}^{\min(y_i+1,2)} p(r_{ij}|\tau_{ij}^\star, \lambda_i) \right) p(\tau_i|y_i, \lambda_i). \end{aligned}$$

Thus for each $i = 1, \ldots, N$, we first sample the arrival and interarrival times $\tau_i = (\tau_{i1}^\star, \tau_{i2}^\star)$ without conditioning on the indicators, and then sample the indicators r_{i1} and, if $y_i > 0$, r_{i2} independently conditional on τ_i. For any i with $y_i > 0$ the joint distribution of $\tau_i = (\tau_{i1}^\star, \tau_{i2}^\star)$ factorizes as $p(\tau_{i1}^\star, \tau_{i2}^\star|y_i, \lambda_i) = p(\tau_{i1}^\star|y_i, \lambda_i, \tau_{i2}^\star) \cdot p(\tau_{i2}^\star|y_i)$. Conditionally on y_i, the arrival time $\tau_{i2}^\star$ of the y_ith jump is the maximum of y_i Un$[0, 1]$ random variables and follows a Beta$(y_i, 1)$-distribution; see Robert and Casella (1999, p. 47). The waiting time until the first jump after $t = 1$ is distributed as $\mathrm{Ex}(\lambda_i)$, and therefore $\tau_{i1}^\star = 1 - \tau_{i2}^\star + \xi_i$, where $\xi_i \sim \mathrm{Ex}(\lambda_i)$.

Step 1 is model dependent, but standard for many models, as we are dealing with a Gaussian model once we condition on $\boldsymbol{\tau}$ and $\boldsymbol{r}$. For many models this leads to an easily implemented algorithm which samples only from standard densities.

2.2 Evaluation of the improved sampler

The reduction in computing time obtained by using the improved sampler introduced in the previous subsection rather than the original sampler of Frühwirth-Schnatter and Wagner (2006a) is of course the greater, the larger the observed counts are.

As an example, we have performed a similar analysis as Chiogna and Gaetan (2002), who evaluated the relationship

between mortality and air pollution for the city of Birmingham, Alabama (US). The observations y_i are daily counts from August 3, 1985 to December 31, 1988, i.e. $N = 1147$. The counts range between 3 and 32, the median being equal to 15. We explain y_i by the Poisson regression model $y_i|\lambda_i \sim \text{Po}(\lambda_i)$ where

$$\log \lambda_i = \alpha_1 + Z_{i,2}\alpha_2 + Z_{i,3}\alpha_3 + Z_{i,4}\alpha_4.$$

$Z_{i,2}$ is the minimum temperature and $Z_{i,3}$ is the humidity on day i, while $Z_{i,4}$ is equal to PM10 on day $i-1$. PM10 is defined as particle matter with a mass median aerodynamic diameter less than 10 µm.

The improved sampler is implemented as described in Sect. 2.1.2. Under a multivariate normal prior on $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_4)$, the conditional posterior $p(\boldsymbol{\alpha}|\boldsymbol{\tau}, \boldsymbol{r}, \boldsymbol{y})$ in Step 1 is a multivariate normal distribution. 15000 posterior draws were generated after a burn-in of 5000 draws. Using MATLAB (Version 7.3.0) on a notebook with a 2.0 GHz processor, we observe a dramatic reduction in computing time, the CPU time of the improved sampler being only 10 CPU minutes instead of 243 CPU minutes for the original sampler. This is mainly due to the tremendous reduction of arrival and interarrival times used for data augmentation, namely 2294 instead of 18452.

The inefficiency factors of the improved sampler are reported in Table 1. They show that mixing is pretty good, although the total number of latent variables is still equal to 4588. The inefficiency factor is defined by $\tau = 1 + 2 \cdot \sum_{h=1}^{H} \rho(h)$, where $\rho(h)$ is the empirical autocorrelation at lag h. The initial monotone sequence estimator by Geyer (1992) is used to determine H, based on the sum of adjacent pairs of empirical autocorrelations $\Phi(s) = \rho(2s) + \rho(2s+1)$. If k is the largest integer so that $\Phi(s) > 0$ and $\Phi(s)$ is monotone for $s = 1, \ldots, k$, then H is defined by $H = 2k+1$.

Table 1 also shows 95% HPD regions for α_1 to α_4. For each regression coefficient the HPD regions were computed marginally as the shortest interval containing 95% of the posterior draws. Evidently, humidity has no significant effect. The minimum temperature, however, has a significant negative effect: the lower the minimum temperature, the higher the mortality rate. While the 95% HPD region for α_4 covers 0, we obtain from the posterior draws that $\text{pr}\{\alpha_4 > 0|\boldsymbol{y}\} = 0.9511$. Thus higher PM10 implies a higher mortality rate.

Table 1 Inefficiency factors and HPD regions for the air pollution data

Regression parameter	95% HPD region	Inefficiency factor τ
α_1 (constant)	(2.6962, 2.7259)	5.7
α_2 (temperature)	(−0.0145, −0.0042)	5.8
α_3 (humidity)	(−0.0090, 0.0093)	5.9
α_4 (PM10)	(−0.0001, 0.0012)	5.7

2.3 Auxiliary mixture sampling for data from the negative binomial distribution

A model commonly applied to capture overdispersion in count data is the Poisson-Gamma model which leads to the negative binomial distribution as marginal distribution for the data; see Hilbe (2007) for a recent review.

Auxiliary mixture sampling for data from the negative binomial distribution has not been considered before, but is easily implemented by observing that such a model corresponds to the following modification of model (1),

$$\begin{aligned} &y_i|\lambda_i \sim \text{Po}(\lambda_i), \\ &\lambda_i = \lambda_i^{\mu}\gamma_i,\ \lambda_i^{\mu} = \exp((\boldsymbol{Z}_i^{\alpha})^T\boldsymbol{\alpha} + (\boldsymbol{Z}_i^{\beta})^T\boldsymbol{\beta}_i), \end{aligned} \tag{7}$$

where a random intercept deviating from the average intercept by $\log \gamma_i$ is present. The negative binomial distribution results if one assumes that γ_i follows a $\text{Ga}(\rho, \rho)$-distribution with degrees of freedom ρ, and that $\gamma_1, \ldots, \gamma_N$ are independent. The model converges to a Poisson model as ρ goes to infinity. For finite ρ, the marginal distribution $p(y_i|\lambda_i^{\mu}, \rho)$ is

$$p(y_i|\lambda_i^{\mu}, \rho) = \binom{\rho + y_i - 1}{\rho - 1} \left(\frac{\rho}{\rho + \lambda_i^{\mu}}\right)^{\rho} \left(\frac{\lambda_i^{\mu}}{\rho + \lambda_i^{\mu}}\right)^{y_i}. \tag{8}$$

The sampling scheme in Sect. 2.1.2 has to be modified in the following way. In Step 1, the random intercepts $\gamma_1, \ldots, \gamma_N$ are assumed to be known when sampling $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. A third step is added to draw $(\rho, \gamma_1, \ldots, \gamma_N)$ jointly. First, the number of degrees of freedom ρ is sampled marginally using a random walk Metropolis-Hastings algorithm without conditioning on $\gamma_1, \ldots, \gamma_N$, by combining the likelihood $p(\boldsymbol{y}|\boldsymbol{\alpha}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_N, \rho)$ constructed from (8) with a prior $p(\rho)$. Then $\gamma_1, \ldots, \gamma_N|\rho, \boldsymbol{y}$ are drawn independently from the conditional Gamma distribution $\text{Ga}(\rho + y_i, \rho + \lambda_i^{\mu})$. For an application to data, see Sect. 4.1.

3 Extension to binomial and multinomial data

3.1 Improved auxiliary mixture sampling

We start with the following modification of model (1), where $\boldsymbol{y} = (y_1, \ldots, y_N)$ are conditionally independent data from a binomial distribution with known repetition parameter N_i:

$$\begin{aligned} &y_i|\pi_i \sim \text{Bin}(N_i, \pi_i), \\ &\log \frac{\pi_i}{1-\pi_i} = \log \lambda_i = (\boldsymbol{Z}_i^{\alpha})^T\boldsymbol{\alpha} + (\boldsymbol{Z}_i^{\beta})^T\boldsymbol{\beta}_i. \end{aligned} \tag{9}$$

3.1.1 Data augmentation

Frühwirth-Schnatter and Frühwirth (2007) introduced auxiliary mixture sampling for logit models of binary data by considering the distribution of $y_i|\pi_i \sim \text{Bin}(1,\pi_i)$, following Scott (2008), as the marginal distribution of an augmented model involving a latent variable y_i^u. In an economic context, y_i^u is equal to the utility of choosing category 1 (McFadden 1974).

This sampler is applied to the binomial model (9) by considering each observation y_i as the total number of successes of N_i independent binary experiments with outcomes $z_{1i},\ldots,z_{N_i,i}$, where z_{ni} follows a binary logit model with the same log odds ratio as (9), i.e. $\text{pr}\{z_{ni}=1|\pi_i\}=\pi_i$. Since it is straightforward to reconstruct the binary outcomes $z_{1i},\ldots,z_{N_i,i}$ from the binomial observation y_i through

$$z_{ni}=\begin{cases}1, & 1\le n\le y_i,\\ 0, & y_i<n\le N_i,\end{cases}$$

it is possible to apply auxiliary mixture sampling to the binary outcomes $z_{1i},\ldots,z_{N_i,i}$ in order to draw inference about the original model (9). The resulting sampler introduces for each binary observation z_{ni} a latent variable y_{ni}^u, leading to a total of $2(\sum_{i=1}^N N_i)$ latent variables once the mixture approximation has been applied.

A more efficient sampler may be derived by introducing for each binomial observation y_i only a single aggregated latent variable, rather than the entire sequence $y_{1i}^u,\ldots,y_{N_i,i}^u$. This is similar in spirit to aggregating interarrival times for data from the Poisson distribution. To aggregate the latent variables $y_{1i}^u,\ldots,y_{N_i,i}^u$ we use the fact that in a binary experiment the "utility" y_{ni}^u has the following distribution (McFadden 1974):

$$y_{ni}^u=\log\lambda_i+\varepsilon_{ni},$$

where $\varepsilon_{ni}=-\log\xi_{ni}$ and ξ_{ni} follows a standard exponential distribution. Therefore the following holds for $n=1,\ldots,N_i$:

$$\exp(-y_{ni}^u)=\frac{1}{\lambda_i}\xi_{ni}.$$

Taking the sum over all $n=1,\ldots,N_i$ we obtain:

$$\sum_{n=1}^{N_i}\exp(-y_{ni}^u)=\frac{1}{\lambda_i}\xi_i, \tag{10}$$

where $\xi_i=\sum_{n=1}^{N_i}\xi_{ni}$ follows a $\text{Ga}(N_i,1)$ distribution because of the independence of the binary experiments. By taking the negative logarithm in (10) we obtain:

$$y_i^\star=\log\lambda_i+\varepsilon_i, \tag{11}$$

where $\varepsilon_i=-\log\xi_i$ with $\xi_i\sim\text{Ga}(N_i,1)$, and $y_i^\star$ is the following aggregated latent variable:

$$y_i^\star=-\log\sum_{n=1}^{N_i}\exp(-y_{ni}^u). \tag{12}$$

The first step of the improved sampler introduces for each binomial observation y_i the aggregated latent variable $y_i^\star$. In the second step, the distribution of ε_i in (11), which is a negative log-Gamma distribution with integer shape parameter N_i, is approximated by a mixture of normal distributions as before. The indicator r_i of this finite mixture is introduced as an additional latent variable. This leads to a total of $2N$ rather than $2(\sum_{i=1}^N N_i)$ latent variables.

Conditional on $\boldsymbol{y}^\star=\{y_1^\star,\ldots,y_N^\star\}$ and $\boldsymbol{r}=\{r_1,\ldots,r_N\}$, the nonlinear non-Gaussian model (9) reduces to a linear Gaussian model:

$$y_i^\star=\log\lambda_i+m_{r_i}(N_i)+\varepsilon_i,\qquad \varepsilon_i|r_i\sim\text{No}(0,s_{r_i}^2(N_i)),$$

with $\log\lambda_i=(\boldsymbol{Z}_i^\alpha)^T\boldsymbol{\alpha}+(\boldsymbol{Z}_i^\beta)^T\boldsymbol{\beta}_i$.

3.1.2 The sampling scheme

Select starting values for $\boldsymbol{y}^\star$ and $\boldsymbol{r}$ and repeat the following steps.

(1) Sample $\boldsymbol{\alpha}$, $\boldsymbol{\beta}=\{\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_N\}$, and $\boldsymbol{\theta}$ conditional on $\boldsymbol{y}^\star$ and $\boldsymbol{r}$.

(2) Sample the aggregated latent variable $\boldsymbol{y}^\star$ and the indicators $\boldsymbol{r}$ conditional on $\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\theta}$ and $\boldsymbol{y}$, by running the following steps, for $i=1,\ldots,N$.

(a) Sample $y_i^\star$ conditional on λ_i and y_i as

$$y_i^\star=-\log\left(\frac{U_i}{1+\lambda_i}+\frac{V_i}{\lambda_i}\right), \tag{13}$$

where $U_i\sim\text{Ga}(N_i,1)$, and $V_i\sim\text{Ga}(N_i-y_i,1)$, independently, if $y_i<N_i$, whereas $V_i=0$ if $y_i=N_i$.

(b) Sample r_i from the following discrete distribution where $j=1,\ldots,R(N_i)$:

$$\text{pr}\{r_i=j|y_i^\star,\lambda_i\}\propto w_j(N_i)\,\varphi(y_i^\star-\log\lambda_i;m_j(N_i),s_j^2(N_i)). \tag{14}$$

To obtain starting values for $y_i^\star$ and r_i we use Step 2 with $\pi_i=\min(\max(y_i/N_i,0.05),0.95)$. To justify sampling of the aggregated latent variable $y_i^\star$, we use (12) and represent the individual "utilities" y_{ni}^u as in Frühwirth-Schnatter and Frühwirth (2007):

$$y_{ni}^u=-\log\left(-\frac{\log U_{ni}}{1+\lambda_i}-\frac{\log V_{ni}}{\lambda_i}I_{\{z_{ni}=0\}}\right),$$

Table 2 Posterior inference for the Titanic passenger data

Group i	y_i/N_i	95% HPD region for α_i	Inefficiency factor τ	
			Improved	Original
child/female/class 3	14/31	(−0.272, 1.248)	9.5	40.4
child/male/class 3	13/48	(−0.997, 0.396)	13.7	52.1
adult/female/class 3	76/165	(0.117, 0.966)	8.4	52.2
adult/female/class 2	80/93	(1.833, 3.121)	13.1	70.6
adult/female/class 1	140/144	(3.213, 5.158)	53.8	201.8
adult/male/class 3	75/462	(−1.339, −0.561)	10.1	54.4
adult/male/class 2	14/168	(−2.360, −1.086)	19.9	123.5
adult/male/class 1	57/175	(−1.014, −0.403)	7.6	60.2

where U_{ni} and V_{ni} are independent uniform random numbers. This yields:

$$y_i^\star = -\log \sum_{n=1}^{N_i} \exp(-y_{ni}^u) = -\log\left(\frac{\sum_{n=1}^{N_i}(-\log U_{ni})}{1+\lambda_i} + \frac{\sum_{n=y_i+1}^{N_i}(-\log V_{ni})}{\lambda_i} \right).$$

Step 2(a) is justified by the facts that $\sum_{n=1}^{N_i}(-\log U_{ni}) \sim \mathrm{Ga}(N_i, 1)$ and, for $y_i < N_i$, $\sum_{n=y_i+1}^{N_i}(-\log V_{ni}) \sim \mathrm{Ga}(N_i - y_i, 1)$.

3.2 Evaluation of the improved sampler

In order to compare the improved sampler with the original sampler we have reanalyzed the Titanic data (Hilbe 2007, Table 6.11), reporting the number y_i of survivals of passengers in each of 12 groups corresponding to all combinations of class (first/second/third), age (child/adult) and gender. The number of exposures N_i in each group ranges from 1 to 462, the median being equal to 71.

While all children in the first and second class survived, this was not the case for the children in the third class. To compare their chance of survival with that of the adults in the various groups, we perform an ANOVA for the survival rates of all $N = 8$ groups having non-survivors by fitting various binomial logit regression models with appropriate design matrices $\boldsymbol{Z}_i$ and unknown regression parameter $\boldsymbol{\alpha}$:

$$y_i \sim \mathrm{Bin}(N_i, \pi_i), \qquad \log \frac{\pi_i}{1-\pi_i} = \log \lambda_i = \boldsymbol{Z}_i^T \boldsymbol{\alpha}. \tag{15}$$

First, we have fitted a saturated model with $\boldsymbol{Z}_i^T = (\delta_{i,1} \cdots \delta_{i,7}\ 1)$, with $\delta_{i,j} = 1$ iff $j = i$. Thus α_8 defines the survival rate of the baseline, chosen to be an adult male in the first class, whereas α_j, $1 \le j \le 7$, captures the difference in the survival rate of group j compared to the baseline.

The improved sampler is implemented as described in Sect. 3.1.2. Under the multivariate normal prior $\boldsymbol{\alpha} \sim \mathrm{No}(\mathbf{0}, 4 \cdot \mathbf{I})$ the conditional posterior $p(\boldsymbol{\alpha}|\boldsymbol{\tau}, \boldsymbol{r}, \boldsymbol{y})$ in Step 1 is again a multivariate normal distribution.

15000 posterior draws were generated after a burn-in of 5000 draws. Using MATLAB (Version 7.3.0) on a notebook with a 2.0 GHz processor we observe a considerable reduction in computing time. The improved sampler is more than six times faster than the original sampler (57 versus 358 CPU seconds). In addition to that, the new sampler also reduces inefficiency considerably, as shown by the inefficiency factors τ in Table 2. Only for α_5, which corresponds to a survival rate very close to 1, inefficiency is still rather high.

From the 95% HPD regions reported in Table 2 we find that an adult female had a significantly higher chance to survive than an adult male in the first class, with the chance increasing with class. In contrast, chance of survival was significantly lower for adult men in classes 2 and 3. Surprisingly, for children in the third class the regression coefficient α_i is not significant, which means that they did not have a higher chance to survive than an adult male in the first class.

For the purpose of model comparison, we fitted several reduced regression models and computed marginal likelihoods as in Frühwirth-Schnatter and Wagner (2008). Some results are shown in Table 3. The model with the largest marginal likelihood is the one with $\alpha_1 = \alpha_3$ and $\alpha_2 = \alpha_6$. Thus for a girl in the third class the survival chance was equal to that of an adult female in this class, while for a boy it was equal to that of an adult male.

3.3 Dealing with multinomial data

A similar method can be applied to data $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N)$ from a multinomial distribution with $m+1$ categories, where $\boldsymbol{y}_i = (y_{1i}, \ldots, y_{mi})$ and y_{ki} counts the number of times category k is observed on occasion i. Model (9) is

Table 3 Log marginal likelihoods $\log p(\boldsymbol{y}|\mathcal{M})$ of various regression models for the Titanic passenger data. Standard errors are given in parentheses

Model $\mathcal{M}$	Unrestricted	$\alpha_1=\alpha_2=\alpha_3$	$\alpha_1=\alpha_3, \alpha_2=\alpha_6$	$\alpha_1=\alpha_2=\alpha_6$
$\log p(\boldsymbol{y}|\mathcal{M})$	−38.82(0.006)	−47.58(0.009)	**−36.76**(0.004)	−47.56(0.008)

modified accordingly:

$$\begin{aligned}
&\boldsymbol{y}_i|\boldsymbol{\pi}_i \sim \text{MulNom}\,(N_i, \pi_{0i}, \pi_{1i}, \ldots, \pi_{mi}),\\
&\pi_{ki} = \frac{\lambda_{ki}}{1+\sum_{l=1}^{m}\lambda_{li}},\\
&\log\lambda_{ki} = (\boldsymbol{Z}_i^{\alpha})^T\boldsymbol{\alpha}_k + (\boldsymbol{Z}_i^{\beta})^T\boldsymbol{\beta}_{ki},\ k=1,\ldots,m, \qquad (16)
\end{aligned}$$

with known repetition parameters $N_i \geq 1$.

3.3.1 Data augmentation

The first step of the improved sampler introduces for each observation $\boldsymbol{y}_i$ m aggregated latent variables $\boldsymbol{y}_i^{\star} = (y_{1i}^{\star}, \ldots, y_{mi}^{\star})$, in a similar manner as for binomial data, see also (11):

$$y_{ki}^{\star} = \log\lambda_{ki} + \varepsilon_{ki}, \qquad (17)$$

where $\varepsilon_{ki} = -\log\xi_{ki}$, with $\xi_{ki} = \sum_{n=1}^{N_i}\exp(-\varepsilon_{kni}) \sim \text{Ga}(N_i, 1)$. In the second step the density of ε_{ki} in (17) is approximated by a Gaussian mixture, and the indicator r_{ki} is introduced as an additional latent variable. This leads to a total of $2mN$ rather than $2m(\sum_{i=1}^{N} N_i)$ latent variables.

Conditional on $\boldsymbol{y}^{\star} = \{\boldsymbol{y}_1^{\star}, \ldots, \boldsymbol{y}_N^{\star}\}$ and $\boldsymbol{r} = \{\boldsymbol{r}_1, \ldots, \boldsymbol{r}_N\}$, where $\boldsymbol{r}_i = (r_{1i}, \ldots, r_{mi})$, the nonlinear non-Gaussian model (16) reduces to m linear Gaussian models, reading for $k = 1, \ldots, m$:

$$\begin{aligned}
&y_{ki}^{\star} = \log\lambda_{ki} + m_{r_{ki}}(N_i) + \varepsilon_{ki},\\
&\text{with } \log\lambda_{ki} = (\boldsymbol{Z}_i^{\alpha})^T\boldsymbol{\alpha}_k + (\boldsymbol{Z}_i^{\beta})^T\boldsymbol{\beta}_{ki}.
\end{aligned}$$

3.3.2 The sampling scheme

Select starting values for $\boldsymbol{y}^{\star}$, $\boldsymbol{r}$ and $\boldsymbol{\theta}$, and repeat the following steps.

(1) Sample $\boldsymbol{\alpha}$, $\boldsymbol{\beta} = \{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_N\}$, and $\boldsymbol{\theta}$ conditional on $\boldsymbol{y}^{\star}$ and $\boldsymbol{r}$.

(2) Sample the aggregated latent variables $\boldsymbol{y}^{\star}$ and the indicators $\boldsymbol{r}$ conditional on $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}$ and $\boldsymbol{y}$, by running the following steps, for $i = 1, \ldots, N$.

(a) Sample $\boldsymbol{y}_i^{\star} = (y_{1i}^{\star}, \ldots, y_{mi}^{\star})$ as:

$$y_{ki}^{\star} = -\log\left(\frac{U_i}{1+\sum_{l=1}^{m}\lambda_{li}} + \frac{V_{ki}}{\lambda_{ki}}\right), \qquad (18)$$

where $U_i \sim \text{Ga}(N_i, 1)$ and, for $k = 1, \ldots, m$, $V_{ki} \sim \text{Ga}(N_i - y_{ki}, 1)$, if $y_{ki} < N_i$, with all random variables being independent, and $V_{ki} = 0$ if $y_{ki} = N_i$.

(b) Sample r_{ki} for $k = 1, \ldots, m$ from the following discrete distribution where $j = 1, \ldots, R(N_i)$:

$$\begin{aligned}
&\text{pr}\{r_{ki} = j|y_{ki}^{\star}, \lambda_{ki}\}\\
&\quad \propto w_j(N_i)\,\varphi(y_{ki}^{\star} - \log\lambda_{ki}; m_j(N_i), s_j^2(N_i)). \qquad (19)
\end{aligned}$$

The justification of Step 2 is similar to the one for binomial data.

4 Statistical modeling based on auxiliary mixture sampling

Auxiliary mixture sampling allows for straightforward statistical modeling of non-Gaussian data, as for instance demonstrated for non-Gaussian state-space and random-effect models in Frühwirth-Schnatter and Wagner (2006a) and Frühwirth-Schnatter and Frühwirth (2007). Further illustration is provided by LeSage et al. (2007), who have analyzed knowledge spillovers across Europe through a Poisson spatial interaction model, by Fahrmeir and Steinert (2006) who have evaluated post-war human security in Cambodia using a geoadditive Bayesian latent variable model for Poisson indicators, and by Gschlößl and Czado (2005) who have modeled the expected number of claims for policy holders of a German car insurance company, using spatial regression modeling. In this section, we apply auxiliary mixture sampling to two further classes of non-Gaussian models, namely finite mixtures of generalized linear models (GLMs) and a Bayesian hierarchical model for spatial count data with a Gaussian Markov random field prior.

4.1 Finite mixtures of GLMs

Finite mixtures of generalized linear models (GLMs) based on the Poisson, the binomial, the negative binomial, or the multinomial distribution, have found numerous applications in biology, medicine and marketing in order to deal with overdispersion and unobserved heterogeneity; see Frühwirth-Schnatter (2006, Sect. 9.4) for a review. A finite mixture of Poisson regression models, for instance, reads:

$$p(y_i|\boldsymbol{\theta}) = \sum_{k=1}^{K}\eta_k\, f_{\text{Po}}(y_i; \mu_{ki}), \qquad (20)$$

where $f_{\text{Po}}(y_i; \cdot)$ is the Poisson density with mean $\mu_{ki} = \exp(\boldsymbol{Z}_i^T\boldsymbol{\alpha}_k)$, and $\boldsymbol{\theta} = (\eta_1, \ldots, \eta_K, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_K)$ is an unknown parameter.

Table 4 Log marginal likelihoods of various regression models for the fabric fault data. Standard errors are given in parentheses

Model	$K=1$	$K=2$	$K=3$	$K=4$
Poisson	−101.79	**−99.21**	−100.74	−103.21
	(0.002)	(0.01)	(0.05)	(0.14)
Poisson	−101.79	**−97.46**	−97.65	−98.60
(fixed slope)	(0.002)	(0.073)	(0.073)	(0.062)
Negative Binomial	**−96.04**	−99.05	−102.21	−104.95
	(0.007)	(0.027)	(0.038)	(0.142)
Negative Binomial	**−96.04**	−97.25	−98.76	−99.97
(fixed slope)	(0.007)	(0.044)	(0.046)	(0.060)

4.1.1 *Parameter estimation using auxiliary mixture sampling*

Various proposals have been put forward on how to estimate the unknown parameter $\boldsymbol{\theta}$ for finite mixtures of GLMs using MCMC under the assumption of a multivariate normal prior for the group specific regression parameters $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_K$ and a Dirichlet prior for the weight distribution $(\eta_1, \ldots, \eta_K)$.

As the likelihood $p(\boldsymbol{y}|\boldsymbol{\theta})$ is available in closed form, one may use a single-move random walk Metropolis–Hastings algorithm as in Viallefont et al. (2002) or a multivariate random walk Metropolis–Hastings algorithm as in Hurn et al. (2003) to sample from the marginal posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$.

To avoid time-consuming tuning of the underlying proposal densities, a three step auxiliary mixture sampler is applied, based on introducing a latent group indicator S_i for each observation pair $(\boldsymbol{Z}_i, y_i)$ as missing data; see Frühwirth-Schnatter (2006, Sect. 3.5). In Step 1 in Sects. 2.1.2, 3.1.2 and 3.3.2, respectively, $\boldsymbol{\theta}$ is sampled by adding the latent group indicators $\boldsymbol{S} = (S_1, \ldots, S_N)$ as conditioning argument. This leads to a conditional multivariate normal posterior for $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_K$ and a conditional Dirichlet posterior for $(\eta_1, \ldots, \eta_K)$. A third step has to be added to sample the latent indicators $\boldsymbol{S} = (S_1, \ldots, S_N)$ conditional on knowing $\boldsymbol{\theta}$ and $\boldsymbol{y}$. This last step is based on the original finite mixture regression model rather than the augmented model as it utilizes for each observation y_i only η_k and $p(y_i|\boldsymbol{\alpha}_k)$, for each $k = 1, \ldots, K$.

4.1.2 *Application to fabric fault data*

For illustration, we consider regression analysis of data on fabric faults (Aitkin 1996). The response variable y_i is the number of faults in a bolt of length l_i. Based on the regressor matrix $\boldsymbol{Z}_i^T = (1 \log l_i)$, we fitted a Poisson and a negative binomial regression model as well as finite mixtures of Poisson and negative binomial regression models with increasing number of groups ($K = 2$ up to $K = 4$). Furthermore we consider mixtures of regression models, where the intercept is group specific, while the slope is fixed, both for the Poisson and the negative binomial distribution.

For each model Bayesian analysis was carried out under a normal $\mathrm{No}(0, 4)$ prior for both fixed and group specific regression coefficients, and a Dirichlet $\mathrm{Di}(4, \ldots, 4)$ prior for the weights $(\eta_1, \ldots, \eta_K)$. For the choice of priors in finite mixture models see e.g. Frühwirth-Schnatter (2006). For the negative binomial distribution the degrees of freedom parameter ρ_k was assumed to be group specific with prior $p(\rho_k) \propto 2d\rho_k/(\rho_k + d)^3$, having a median of $d(1 + \sqrt{2}) = 10$. We sampled 10000 posterior draws, after a burn-in of 2000, using the improved sampler. This required between 60 and 90 CPU seconds per model using MATLAB (Version 7.3.0) on a notebook with a 2.0 GHz processor.

To select the number of groups and the appropriate observation distribution, marginal likelihoods were computed for each model as in Frühwirth-Schnatter and Wagner (2008). They are reported in Table 4. Under the assumption that the data arise from a Poisson distribution, a finite mixture of two regression models with fixed slope is selected by Bayesian model choice, which confirms results obtained by Aitkin (1996) and McLachlan and Peel (2000) using alternative methods of model selection.

However, this model has a smaller marginal likelihood than a negative binomial regression model without any mixture structure, which is the model with the largest marginal likelihood in Table 4. The posterior distribution of ρ, shown in Fig. 1 for the negative binomial regression model, clearly indicates overdispersion compared to the Poisson distribution and confirms the results obtained from model selection. For this model, 95% HPD regions and inefficiency factors are given in Table 5.

4.2 Disease mapping

We now discuss a spatial extension of (1), a hierarchical model for disease mapping as introduced by Besag et al.

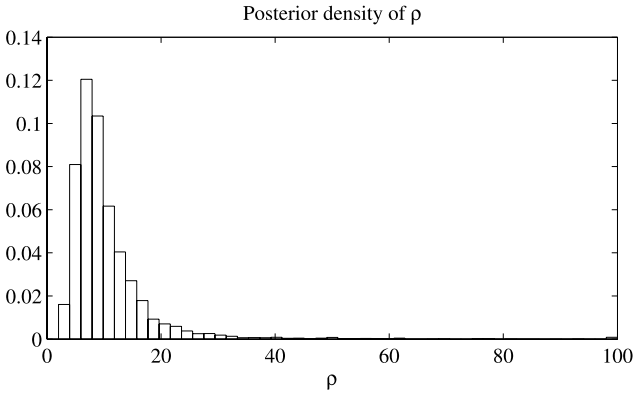

Fig. 1 Fabric fault data; posterior density of the degrees of freedom parameter ρ under a negative binomial regression model

Table 5 Posterior inference for the fabric fault data for the negative binomial regression model

	95% HPD region	Inefficiency factor τ
intercept	$(-5.35, -1.42)$	3.9
log(length)	$(0.57, 1.19)$	3.9
ρ	$(2.9, 22.2)$	8.9

(1991). The model decomposes the log relative risk in a specific region into a spatially structured and an unstructured component, assuming a Gaussian Markov random field (GMRF) and an exchangeable prior, respectively, with two additional unknown precision parameters.

This formulation falls into the class of latent Gaussian models, see Rue and Held (2005) for a general account. Following Knorr-Held and Rue (2002), we reparametrize the Besag et al. (1991) formulation and assume that the observed disease counts y_i in district $i = 1, \ldots, N$ are conditionally independent Poisson with mean $e_i \exp(\beta_i)$, where e_i are known expected counts and β_i are unknown log relative risk parameters. In the second stage of the hierarchical model, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_N)^T$ follows a multivariate Gaussian distribution with mean $\boldsymbol{u} = (u_1, \ldots, u_N)^T$ and diagonal precision matrix $\omega \mathbf{I}$,

$$\boldsymbol{\beta}|\omega \sim \mathrm{No}(\boldsymbol{u}, \omega \mathbf{I}), \tag{21}$$

and in the third stage $\boldsymbol{u}$ follows an intrinsic GMRF:

$$p(\boldsymbol{u}|\kappa) \propto \kappa^{\frac{N-1}{2}} \exp\left(-\frac{\kappa}{2} \sum_{i \sim j} (u_i - u_j)^2\right), \tag{22}$$

see Rue and Held (2005, Sect. 3.3.2). In (22), $i \sim j$ denotes all pairs of adjacent districts i and j. For the unknown precision parameters ω and κ we adopt the usual (independent) Gamma hyperpriors $\omega \sim \mathrm{Ga}(1, 0.01)$ and $\kappa \sim \mathrm{Ga}(1, 0.01)$.

4.2.1 Parameter estimation via MCMC

Statistical inference via MCMC in this highly parametrized model is difficult, especially if the disease counts are low. A single-site algorithm, as implemented for example in WinBugs, may fail miserably, as demonstrated in Knorr-Held and Rue (2002). These authors propose to use block updating of $\boldsymbol{\beta}$ and $\boldsymbol{u}$, based on the GMRF approximation as described in detail in Rue and Held (2005, Sect. 4.4.1). The idea is to construct a multivariate Metropolis-Hastings proposal based on a quadratic Taylor approximation to the full conditional distribution of $\boldsymbol{\beta}$ and $\boldsymbol{u}$. This can be combined with updates of the two precision parameters κ and ω to a joint Metropolis-Hastings proposal for all unknown parameters. Knorr-Held and Rue (2002) use a specific proposal, multiplying the current value of the precision parameter with a random variable z proportional to $1 + 1/z$ on $[1/f, f]$, where $f > 1$ is a constant scaling factor. This specific choice has the advantage that the proposal ratio in the Metropolis-Hastings acceptance probability equals one and is used for both κ and ω. Subsequently $\boldsymbol{\beta}$ and $\boldsymbol{u}$ are sampled based on the GMRF approximation. Finally, all updated parameters are accepted or rejected in a joint Metropolis-Hastings step.

Alternatively, auxiliary mixture sampling can be implemented. This has the distinct advantage that the full conditional distribution of $\boldsymbol{\beta}$ and $\boldsymbol{u}$ given κ and ω is already a GMRF, so no Taylor approximation is necessary. Step 1 of the auxiliary sampler presented in Sect. 2.1.2 may be implemented in two Gibbs steps by first updating $\boldsymbol{\beta}$ and $\boldsymbol{u}$ conditional on κ and ω and then updating κ and ω conditional on $\boldsymbol{\beta}$ and $\boldsymbol{u}$. To speed up convergence, we have implemented Step 1 as a Metropolis-Hastings move which updates $\boldsymbol{\beta}$, $\boldsymbol{u}$, κ and ω jointly in the same way as for the GMRF approximation.

4.2.2 Applications

We now report results from an empirical comparison of improved auxiliary mixture sampling using the joint Metropolis-Hastings move and the GMRF approximation. Both algorithms have been implemented in `C` using the library `GMRFLib` (Rue and Held 2005, Appendix B). All analyses were run on an Intel Core 2 Duo T2500 processor with 2 GHz under Ubuntu. Two data sets have been analysed: The first one gives the number of cases of Insulin dependent Diabetes Mellitus (IDDM) in Sardinia ($N = 366$), as analyzed in Knorr-Held and Rue (2002). The second one gives the number of deaths of oral cavity cancer in Germany ($N = 544$), as analyzed in Knorr-Held and Raßer (2000). The first disease is sparse with a total of 619 cases (median of 1 per district), whereas the second disease is fairly common with a total of 12,835 cases (median of 15).

Tables 6 and 7 summarize the results for the Sardinia and Germany data, respectively, showing the effective sample

Table 6 Empirical comparison of the GMRF approximation and improved auxiliary mixture sampling (IAMS) for the Sardinia data

Scaling factor	Method	Speed (it/sec)	Acc. rate	Parameter	ESS	ESS per sec
$f = 2.0$	GMRF	42.3	61.1	ω	388.2	1.6
				κ	166.0	0.7
	IAMS	159.3	50.1	ω	200.1	3.2
				κ	164.5	2.6
$f = 5.0$	GMRF	42.7	29.8	ω	840.3	3.6
				κ	537.4	2.3
	IAMS	163.4	15.8	ω	370.8	6.1
				κ	134.5	2.2

Table 7 Empirical comparison of the GMRF approximation and improved auxiliary mixture sampling (IAMS) for the Germany data

Scaling factor	Method	Speed (it/sec)	Acc. rate	Parameter	ESS	ESS per sec
$f = 1.5$	GMRF	27.9	33.4	ω	220.3	0.6
				κ	609.6	1.7
	IAMS	102.5	41.9	ω	271.5	2.8
				κ	760.2	7.8
$f = 3.0$	GMRF	28.1	10.3	ω	347.8	1.0
				κ	403.6	1.1
	IAMS	104.2	9.2	ω	282.2	2.9
				κ	426.0	4.4

size (ESS) (Kass et al. 1998) and the effective sample size per second for the two precision parameters ω and κ. Also given is the acceptance rate of the two algorithms for different choices of the scaling factor f. For the sake of simplicity, the same factor was used for both precision parameters, although this could be changed easily. ESS is an estimate of the number of independent samples which would be required to obtain a parameter estimate with the same precision as the MCMC estimate based on M dependent samples (here we used $M = 2000$ samples obtained by storing every fifth iteration of the MCMC algorithm). The effective sample size of a parameter is calculated as the number of samples M used from the Markov chain divided by the inefficiency factor τ.

Table 6 shows that the improved sampler is nearly four times as fast as the GMRF approximation, despite the large number of additional auxiliary variables. However, for the same values of the scaling parameters, the acceptance rates for the auxiliary mixture sampling are generally lower than the ones based on the GMRF approximation. At first sight this is surprising as—without the update of the precision parameters—the improved sampler yields acceptance rates equal to unity, whereas the GMRF approximation has acceptance rates of approximately 70% for these data. However, the auxiliary mixture sampler conditions on a particular mixture component, so the target distribution has smaller variance, and lower acceptance rates are possible for the joint Metropolis-Hastings update. The effective sample size is somewhat better for the GMRF approximation, since the samples are less autocorrelated. However, adjusting for computation time, the order is reversed and the auxiliary variable method is roughly twice as good in terms of ESS per second, if the acceptance rates are not too low.

For the Germany data, see Table 7, the results are even more in favor of the improved sampler, with effective sample sizes up to four times as large per second. Interestingly, for $f = 1.5$ the auxiliary mixture sampler has higher acceptance rates than the GMRF approximation. Presumably, for larger counts the mixture approximation will be dominated by one component, so the reduction of the conditional variance, compared to the GMRF approximation, will be minor.

5 Concluding remarks

In this paper we have developed improved auxiliary mixture sampling algorithms for hierarchical models of Poisson, binomial, negative binomial or multinomial data. In contrast to methods previously suggested in the literature, the number of auxiliary variables is independent of the

number of counts y_i in the Poisson and the negative binomial case, and independent of the number of repetitions N_i in the binomial and multinomial case. This is a clear improvement compared with the auxiliary mixture sampling algorithms proposed in Frühwirth-Schnatter and Wagner (2006a) and Frühwirth-Schnatter and Frühwirth (2007). Empirical evidence of this has been reported in Sects. 2.2 and 3.2.

The main motivation for the development of the improved sampler has not been to yield a uniformly better algorithm, but to simplify the implementation and to improve the computational performance of MCMC algorithms for fairly complex non-Gaussian hierarchical models. This was illustrated for a finite mixture of GLMs and an application to disease mapping. In particular, auxiliary mixture sampling allows us to construct good samplers with reasonable acceptance rates for block-updating a large or very large number of parameters, as in the spatial and spatio-temporal analysis of several health outcomes Held et al. (2005, 2006), where count and binomial data are commonplace.

Acknowledgements We thank the anonymous reviewers and the associate editor for numerous useful suggestions and comments.

Appendix A: Approximation of the negative log-Gamma distribution by Gaussian mixtures

A.1 The negative log-Gamma distribution

Assume that x is Gamma-distributed with integer shape parameter ν and unit scale, $x \sim \mathrm{Ga}(\nu, 1)$. This distribution is the convolution of ν exponential distributions with mean equal to one. Then $y = -\log x$ is distributed according to the negative of a log-Gamma distribution, with the probability density function

$$g(y;\nu) = \frac{\exp(-\nu y - e^{-y})}{\Gamma(\nu)},$$

and the characteristic function

$$\phi(t;\nu) = -\frac{\Gamma(it+\nu)}{\Gamma(\nu)}.$$

The moments can be computed explicitly in terms of polygamma functions. In particular, the expectation μ and the variance σ^2 are given by

$$\mu(\nu) = -\psi(\nu), \qquad \sigma^2(\nu) = \psi'(\nu),$$

where $\psi(\cdot)$ is the digamma function, and $\psi'(\cdot)$ is the trigamma function. In the following, only the standardized variate $u = (y-\mu)/\sigma$ will be used, with the density

$$f(u;\nu) = \frac{\sigma(\nu)\exp\{-\nu[\sigma(\nu)u+\mu(\nu)] - e^{-[\sigma(\nu)u+\mu(\nu)]}\}}{\Gamma(\nu)}.$$

Using the standardized variates has the advantage that the effective support of the distribution is almost independent of ν. For small values of ν, however, there is a noticeable tail to the right, so that the interval $\mathcal{S} = [-6, 10]$ has been used as the support for all values of ν. For large ν, the distribution of u approaches the standard normal distribution. Approximation by Gaussian mixtures therefore requires fewer components for increasing ν.

A.2 Approximation by Gaussian mixtures

The approximating Gaussian mixtures were estimated by minimizing the Kullback-Leibler divergence d_{KL} plus a penalty term that forces the sum of the weights to one:

$$\begin{aligned} &D(\boldsymbol{w}, \boldsymbol{m}, \boldsymbol{s}^2) \\ &\quad = \int_{\mathcal{S}} f(u;\nu)\log\frac{f(u;\nu)}{\varphi(u, \boldsymbol{w}(\nu), \boldsymbol{m}(\nu), \boldsymbol{s}^2(\nu))}\,\mathrm{d}u \\ &\qquad + \omega\left(\sum_{r=1}^{R(\nu)} w_r(\nu) - 1\right)^2, \end{aligned} \tag{23}$$

where $\varphi(u, \boldsymbol{w}(\nu), \boldsymbol{m}(\nu), \boldsymbol{s}^2(\nu))$ is the density of a Gaussian mixture with $R(\nu)$ components, weights $w_r(\nu)$, means $m_r(\nu)$, and variances $s_r^2(\nu)$. The penalty factor was set to $\omega = 10^9$. As a consequence, the sum of the weights differs from one by at most 7×10^{-10}. Note that d_{KL} is invariant under affine transformations and in particular under standardization. The integral in (23) was computed by the trapezoidal rule on a grid of size 32000.

As the component weights w_r are constrained to the interval $(0, 1)$ and the variances s_r^2 have to be positive, the mixture was rewritten in terms of the unconstrained transformed parameters

$$w_r' = \log(w_r) - \log(1-w_r), \qquad (s_r')^2 = \log s_r^2.$$

The modified objective function was minimized using the function `fminsearch` in the optimization toolbox of MATLAB (Version 7.0.1). This function implements a direct search method, the Nelder-Mead simplex algorithm (Nelder and Mead 1965).

The starting point was the 10-component approximation of the log-exponential distribution, corresponding to $\nu = 1$, described in Frühwirth-Schnatter and Frühwirth (2007). As it is neither feasible nor necessary to compute the approximating mixtures for all values of ν up to, say, $\nu = 100{,}000$,

a sequence of values of ν was chosen with ever increasing gaps above $\nu = 100$:

$$\begin{aligned}\boldsymbol{\nu} = \{&2, 3, \ldots, 100, 102, \ldots, 150, 155, \ldots, 200,\\ &220, \ldots, 300, 320, 340, \ldots, 500, 550, \ldots, 1000,\\ &1100, \ldots, 2000, 2200, 2400, \ldots, 5000,\\ &5500, \ldots, 10000, 11000, \ldots, 20000, 22000,\\ &24000, \ldots, 30000, 35000, \ldots, 100000\}.\end{aligned}$$

An approximation was accepted only if the Kullback-Leibler divergence d_{KL} of the mixture density from the target density was below a threshold t_{KL} and if the maximum absolute difference $d_{\max}$ between the two densities was below a threshold $t_{\max}$. We chose $t_{\mathrm{KL}} = 10^{-5}$ and $t_{\max} = 5 \times 10^{-4}$, which are the approximate values of d_{KL} and $d_{\max}$ for $\nu = 1$. Thus all approximations are at least as good as the one for $\nu = 1$, which has been shown to be excellent (Frühwirth-Schnatter and Frühwirth 2007). At the same time, we tried to find the smallest number of components required. The mixture approximation for $\nu = \nu_i$ was therefore computed in the following way:

(1) Take the parameters of the mixture for $\nu = \nu_{i-1}$ as starting values and minimize the objective function for $\nu = \nu_i$. If necessary, restart the minimization until $d_{\mathrm{KL}} \leq t_{\mathrm{KL}}$ and $d_{\max} \leq t_{\max}$.
(2) Save the estimated parameters.
(3) Reduce the number of components by 1.
(4) Compute new starting values by merging the component with the smallest weight and its neighbour with the smaller weight.
(5) Minimize the objective function.
(6) If $d_{\mathrm{KL}} \leq t_{\mathrm{KL}}$ and $d_{\max} \leq t_{\max}$, go to Step 2.
(7) Otherwise, store the saved parameters.

In order to achieve optimal precision for small values of ν, at least nine components were kept for $\nu < 20$. Figure 2 shows the Kullback-Leibler divergence d_{KL} in the range $1 \leq \nu \leq 100000$. For $\nu > 30000$ a single Gaussian passes the acceptance criteria.

A.3 Parametrization of the mixtures

For small values of ν the mixture parameters change substantially when ν is increased. The parameters are therefore stored individually for $1 \leq \nu \leq 19$. For $\nu \geq 20$ it is possible to parametrize the mixtures as a function of ν without sacrificing the accuracy of the approximation. This allows a more compact representation of the mixture parameters as well as the computation of mixtures that have not been estimated explicitly, including approximations to log-Gamma distributions with non-integer shape parameter.

The parametrization was performed separately in the five ranges of ν summarized in Table 8. A second-order polynomial was fitted to the mixture weights, and a rational function with quadratic numerator and linear denominator to the means and variances. Figure 3 shows the Kullback-Leibler divergence of the parametrized and of the original estimated mixtures from the respective target distributions. It can be seen that there is virtually no loss in accuracy when using the parametrization. A MATLAB function implementing the parametrization has been written and is available from the authors. The unstandardized mixture for shape parameter ν is obtained in the following way:

```
[p,m,v,nc]=compute_mixture(nu);
```

the input `nu` being the desired value of ν. The output vectors `p,m,v` return the weights, the means, and the variances,

Table 8 The five ranges of parametrization of the mixtures

Range	$\nu_{\min}$	$\nu_{\max}$	Components
1	20	49	4
2	50	439	3
3	440	1599	2
4	1600	10000	2
5	10000	30000	2

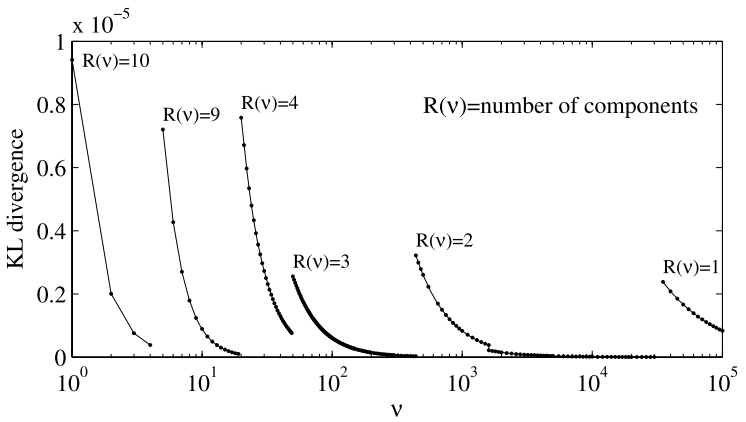

Fig. 2 Kullback-Leibler divergence of the estimated mixtures from the standardized negative log-Gamma distribution as a function of the shape parameter ν, for $1 \leq \nu \leq 100000$. $R(\nu)$ is the number of components in the mixtures

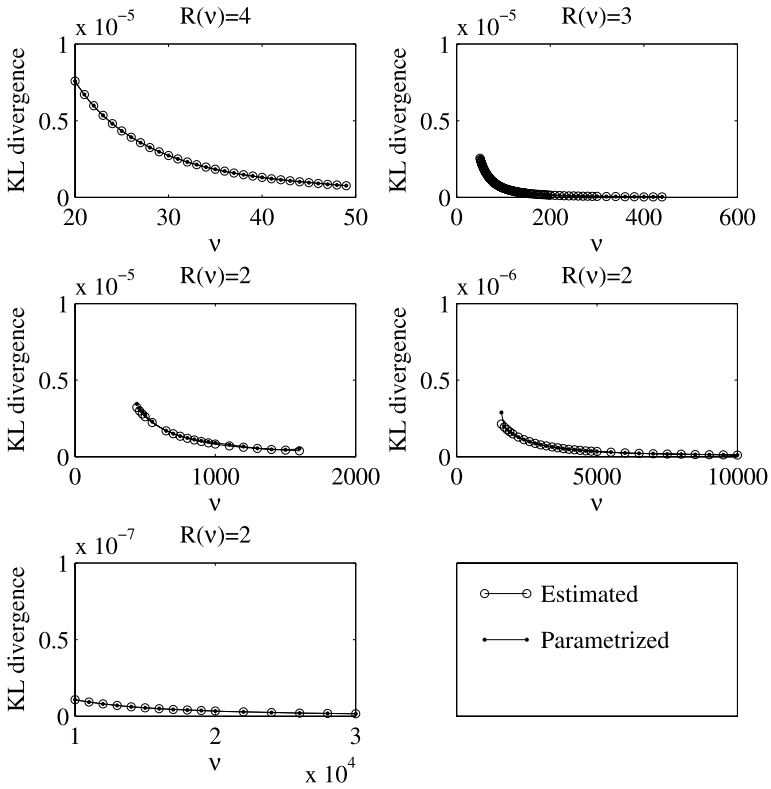

Fig. 3 Kullback-Leibler divergence of the estimated and of the parametrized mixtures from the standardized negative log-Gamma distribution as a function of the shape parameter ν, for $20 \leq \nu \leq 100{,}000$. $R(\nu)$ is the number of components in the mixtures

respectively, of the unstandardized mixture, and `nc` returns the number of mixture components. A similar implementation in the C language is included in the GMRFLib library (Rue and Held 2005, Appendix).

References

Aitkin, M.: A general maximum likelihood analysis of overdispersion in generalized linear models. Stat. Comput. **6**, 251–262 (1996)

Albert, J.H., Chib, S.: Bayesian analysis of binary and polychotomous response data. J. Am. Stat. Assoc. **88**, 669–679 (1993)

Bernardo, J.M., Smith, A.F.M.: Bayesian Theory. Wiley, Chichester (1994)

Besag, J., York, J.C., Mollié, A.: Bayesian image restoration with two applications in spatial statistics (with discussion). Ann. Inst. Stat. Math. **43**, 1–59 (1991)

Chiogna, M., Gaetan, C.: Dynamic generalized linear models with application to environmental epidemiology. Appl. Stat. **51**, 453–468 (2002)

Fahrmeir, L., Steinert, S.: A geoadditive Bayesian latent variable model for Poisson indicators. Technical Report, Department of Statistics, University of Munich (2006)

Frühwirth-Schnatter, S.: Finite Mixture and Markov Switching Models. Springer Series in Statistics. Springer, New York (2006)

Frühwirth-Schnatter, S., Frühwirth, R.: Auxiliary mixture sampling with applications to logistic models. Comput. Stat. Data Anal. **51**, 3509–3528 (2007)

Frühwirth-Schnatter, S., Wagner, H.: Auxiliary mixture sampling for parameter-driven models of time series of small counts with applications to state space modeling. Biometrika **93**, 827–841 (2006a)

Frühwirth-Schnatter, S., Wagner, H.: Data augmentation and Gibbs sampling for regression models of small counts. Student **5**, 221–234 (2006b). Also Research Report IFAS 2004-04, available at http://www.ifas.jku.at/

Frühwirth-Schnatter, S., Wagner, H.: Marginal likelihoods for non-Gaussian models using auxiliary mixture sampling. Comput. Stat. Data Anal. **52**, 4608–4624 (2008)

Geyer, C.: Practical Markov chain Monte Carlo. Stat. Sci. **7**, 473–511 (1992)

Gschlößl, S., Czado, C.: Does a Gibbs sampler approach to spatial Poisson regression models outperform a single site MH sampler? Comput. Stat. Data Anal. **52**, 4184–4202 (2005)

Held, L., Natario, I., Fenton, S., Rue, H., Becker, N.: Towards joint disease mapping. Stat. Methods Med. Res. **14**, 61–82 (2005)

Held, L., Graziano, G., Frank, C., Rue, H.: Joint spatial analysis of gastrointestinal infectious diseases. Stat. Methods Med. Res. **15**, 465–480 (2006)

Hilbe, J.M.: Negative Binomial Regression. Cambridge University Press, Cambridge (2007)

Holmes, C.C., Held, L.: Bayesian auxiliary variable models for binary and multinomial regression. Bayesian Anal. **1**, 145–168 (2006)

Hurn, M., Justel, A., Robert, C.P.: Estimating mixtures of regressions. J. Comput. Graph. Stat. **12**, 55–79 (2003)

Kass, R.E., Carlin, B., Gelman, A., Neal, R.: Markov chain Monte Carlo in practice: a roundtable discussion. Am. Stat. **52**, 93–100 (1998)

Knorr-Held, L., Raßer, G.: Bayesian detection of clusters and discontinuities in disease maps. Biometrics **56**, 13–21 (2000)

Knorr-Held, L., Rue, H.: On block updating in Markov random field models for disease mapping. Scand. J. Stat. **29**, 597–614 (2002)

LeSage, J.P., Fischer, M.M., Scherngell, T.: Knowledge spillovers across Europe: evidence from a Poisson spatial interaction model with spatial effects. Pap. Reg. Sci. **86**, 393–421 (2007)

McFadden, D.: Conditional logit analysis of qualitative choice behaviour. In: Zarembka, P. (ed.) Frontiers of Econometrics, pp. 105–142. Academic, New York (1974)

McLachlan, G.J., Peel, D.: Finite Mixture Models. Wiley Series in Probability and Statistics. Wiley, New York (2000)

Nelder, J.A., Mead, R.: A simplex method for function minimization. Comput. J. **7**, 308–313 (1965)

Robert, C.P., Casella, G.: Monte Carlo Statistical Methods. Springer Series in Statistics. Springer, New York (1999)

Rue, H., Held, L.: Gaussian Markov Random Fields: Theory and Applications. Monographs on Statistics and Applied Probability, vol. 104. Chapman & Hall/CRC, London (2005)

Scott, S.L.: Data augmentation and the Bayesian analysis of multinomial logit models. Stat. Pap., (2008, accepted for publication)

Tüchler, R.: Bayesian variable selection for logistic models using auxiliary mixture sampling. J. Comput. Graph. Stat. **17**, 76–94 (2008)

Viallefont, V., Richardson, S., Green, P.J.: Bayesian analysis of Poisson mixtures. J. Nonparametric Stat. **14**, 181–202 (2002)