Department for Applied Statistics
Johannes Kepler University Linz

IFAS Research Paper Series
2008-34

Improved Auxiliary Mixture Sampling for Hierarchical Models of Non-Gaussian Data

Sylvia Frühwirth-Schnatter[a], Rudolf Frühwirth[b],
Leonhard Held[c] and Håvard Rue[d]

March 2008

[a]Department of Applied Statistics and Econometrics, Johannes Kepler Universität Linz, Austria
[b]Institute of High Energy Physics, Austrian Academy of Sciences, Vienna, Austria
[c]Institute of Social and Preventive Medicine, University of Zurich, Switzerland
[d]Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway

Abstract

The article proposes an improved method of auxiliary mixture sampling for count, binomial and multinomial data. 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.

Key words: binomial data, count data, finite mixture models, Gaussian mixture, log-gamma distribution, multinomial data, negative binomial distribution

1 Introduction

During the past years, auxiliary mixture sampling (AMS) has turned out to be a useful tool for the Bayesian analysis of non-Gaussian models. The method has been used first by Shephard (1994) for stochastic volatility models, see also Omori et al. (2007). Recently, it has been extended to hierarchical models for non-Gaussian data like state-space and random-effects models by Frühwirth-Schnatter and Wagner (2006b, 2006a) and Frühwirth-Schnatter and Frühwirth (2007). The main motivation for the development of auxiliary mixture sampling has been to simplify Markov chain Monte Carlo (MCMC) estimation, see LeSage et al. (2007), Fahrmeir and Steinert (2006) and Gschlößl and Czado (2005). Rather recently, auxiliary mixture sampling has been shown to facilitate Bayesian model selection for non-Gaussian models (Holmes and Held, 2006; Frühwirth-Schnatter and Wagner, 2007; Tüchler, 2007).

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 Bino (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 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 4 for Poisson data, 2 for binomial data, and $2\,m$ 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 expecta-

tion 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 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 Section 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 Section 4 for finite mixtures of generalized linear models and a Bayesian hierarchical model for 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 $\boldsymbol{\beta}_i$:

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

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 derived in the following way. First note that for any observation $y_i > 0$ the arrival time of the last jump before $t = 1$, denoted by $\tau_{i2}^\star$, follows a $\text{Ga}(y_i, \lambda_i)$ distribution:

$$\tau_{i2}^\star = \frac{\xi_{i2}}{\lambda_i}, \quad \xi_{i2} \sim \text{Ga}(y_i, 1). \tag{2}$$

The $\mathrm{Ga}(a,b)$ distribution is defined as in Bernardo and Smith (1994), with density $f_{\mathrm{Ga}}(y;a,b) = b^a y^{a-1} e^{-by}/\Gamma(a)$. Second, the interarrival time between the last jump before and the first jump after $t = 1$, denoted by $\tau_{i1}^\star$, follows an exponential distribution:

$$\tau_{i1}^\star = \frac{\xi_{i1}}{\lambda_i}, \quad \xi_{i1} \sim \mathrm{Ex}(1). \tag{3}$$

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 $\varepsilon_{i1} = -\log \xi_{i1}$ with $\xi_{i1} \sim \mathrm{Ex}(1) = \mathrm{Ga}(1,1)$ and $\varepsilon_{i2} = -\log \xi_{i2}$ with $\xi_{i2} \sim \mathrm{Ga}(y_i, 1)$. For $y_i = 0$ we are dealing only with equation (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 this 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 with 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:

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

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 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 $\xi_i \sim \text{Ex}(\lambda_i)$. 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 $k = 1, \ldots, R(1)$:

$$\text{pr}\{r_{i1} = k | \tau_{i1}, \lambda_i\} \propto w_k(1)\, \varphi(-\log \tau_{i1} - \log \lambda_i; m_k(1), s_k^2(1)).$$

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

$$\text{pr}\{r_{i2} = k | \tau_{i1}, \lambda_i\} \propto w_k(y_i)\, \varphi(-\log \tau_{i2} - \log \lambda_i; m_k(y_i), s_k^2(y_i)).$$

To obtain starting values for $\boldsymbol{\tau}$ and $\boldsymbol{R}$ we use step 2 with $\lambda_i = \max(y_i, 0.1)$. Step 2 is based on decomposing the joint posterior of $(\boldsymbol{\tau}, \boldsymbol{R})$ as

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

Thus for each $i = 1, \ldots, N$, we may first sample the arrival 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 $p(\tau_{i1}^\star, \tau_{i2}^\star | y_i, \lambda_i) = p(\tau_{i1}^\star | y_i, \lambda_i) \cdot p(\tau_{i2}^\star | y_i)$. 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), while the waiting time until the first jump after $t = 1$ is distributed as $\text{Ex}(\lambda_i)$, and therefore $\tau_{i1}^\star = 1 - \tau_{i2}^\star + \xi_i$, where $\xi_i \sim \text{Ex}(\lambda_i)$.

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 from standard densities, only.

2.2 Evaluation of the Improved Sampler

To compare the improved with the original sampler of Frühwirth-Schnatter and Wagner (2006a) we have reanalyzed several data sets.

First, we performed state space modelling of road safety data as in Frühwirth-Schnatter and Wagner (2006a, Section 4), using both samplers. The data are monthly counts of children aged 6–10 years and senior people above 65 years killed or injured in Linz (Austria) from 1987–2005[1], i.e. $N = 228$. Since the counts are fairly small the original sampler is very efficient. Nevertheless, for the children data the CPU time of the new sampler is only about 54% of the CPU time of the original sampler, see Table 1. Part of this reduction results from reducing the total number of interarrival times used for data augmentation from 634 to 418. In addition, the aggregated interarrival times $-\log \tau_{i2}^\star$ appearing in (5) are closer to a normal distribution than the interarrival times in the original sampler, and less components are

Table 1: Evaluation of improved auxiliary mixture sampling. "ratio CPU" is the CPU time of the improved sampler over the CPU time of the original one.

Data	N	latent variables (without indicators) original	improved	ratio CPU
Children	228	634	418	54.1%
Senior people	228	1393	453	49.0%
Air polution	1147	18452	2294	6.3%

needed in the mixture approximation (6), see Appendix A.2. A similar reduction results for the senior people, see Table 1.

The larger the observed counts the greater, of course, is the reduction in computing time. For a further illustration we perform 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 observation 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 10μm, see Chiogna and Gaetan (2002) for more details.

Under a multivariate normal prior on $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_4)$, the improved sampler is implemented as described in Subsection 2.1.2. The conditional posterior $p(\boldsymbol{\alpha}|\boldsymbol{\tau}, \boldsymbol{R}, \boldsymbol{y})$ in Step 1 is a multivariate normal distribution. We observe a dramatic reduction in computing time, the CPU time of the improved sampler being only 6.3% of the CPU time of the original sampler. This is mainly due to the tremendous reduction of interarrival times used for data augmentation, see Table 1.

For illustration, Figure 1 shows the posterior draws of α_2, α_3 and α_4. Evidently, the sampler converges quickly to the stationary distribution and mixing is pretty good although the number of latent variables still is equal to 2294. Humidity has no significant effect. The minimum temperature, however, has a significant effect: the lower the minimum temperature, the higher the mortality rate. Concerning PM10, we obtain from the posterior draws that $\text{pr}\{\alpha_4 > 0|\boldsymbol{y}\} = 0.9511$. Thus the higher PM10, the higher the mortality rate.

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 distri-

[1]The time series in Frühwirth-Schnatter and Wagner (2006a) are shorter versions ranging from 1987–2002.

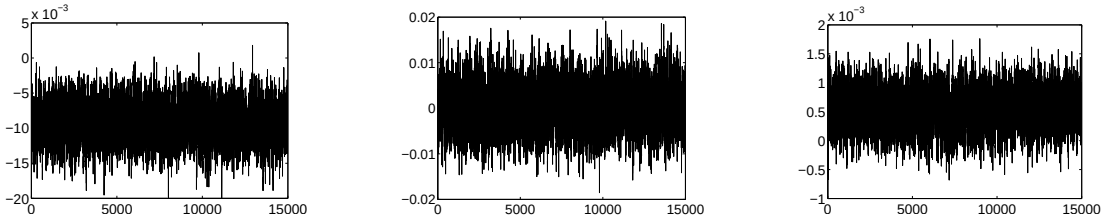

Figure 1: Birmingham mortality data; posterior draws ($M = 15000$) obtained for α_1 (left hand side), α_2 (middle) and α_3 (right hand side).

bution 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),

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

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 $p(y_i|\lambda_i^\mu, \rho)$ reads:

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

The sampling scheme in Subsection 2.1.2 has to be modified in the following way. In step 1, the random intercept $\gamma_1, \ldots, \gamma_N$ is 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 $\text{Ga}(\rho+y_i, \rho+\lambda_i^\mu)$.

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:

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

3.1.1 Data augmentation

To implement auxiliary mixture sampling, Frühwirth-Schnatter and Frühwirth (2007) recover the underlying repeated binary measurements $z_{1i}, \ldots, z_{N_i,i}$, where $z_{ni} = 1$ for $1 \leq n \leq y_i$ and $z_{ni} = 0$ for $y_i < n \leq N_i$, and introduce for each binary observation z_{ni} the utility y^u_{ni} of choosing category 1 as latent variable, 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 is derived in the following way. First note that for any utility y^u_{ni} the following holds for $n = 1, \ldots, N_i$:

$$\exp(-y^u_{ni}) = \frac{1}{\lambda_i} \exp(-\varepsilon_{ni}),$$

where $\exp(-\varepsilon_{ni})$ follows a standard exponential distribution. Taking the sum over all n we obtain:

$$\sum_{n=1}^{N_i} \exp(-y^u_{ni}) = \frac{1}{\lambda_i}\xi_i, \qquad \xi_i = \sum_{n=1}^{N_i} \exp(-\varepsilon_{ni}), \tag{10}$$

where ξ_i follows a $\mathrm{Ga}(N_i, 1)$ distribution due to the independence of the binary experiments. By taking the negative logarithm in (10) we obtain:

$$y^\star_i = \log \lambda_i + \varepsilon_i, \tag{11}$$

where $\varepsilon_i = -\log \xi_i$ with $\xi_i \sim \mathrm{Ga}(N_i, 1)$, and $y^\star_i$ is the following aggregated utility:

$$y^\star_i = -\log \sum_{n=1}^{N_i} \exp(-y^u_{ni}). \tag{12}$$

The first step of the improved sampler introduces for each binomial observation y_i the aggregated utility $y^\star_i$ as a latent variable, rather than the sequence of individual utilities $y^u_{1i}, \ldots, y^u_{N_i i}$. In the second step, the density of ε_i in (11), which follows 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 $\boldsymbol{y}^\star = \{y^\star_1, \ldots, y^\star_N\}$ and $\boldsymbol{R} = \{r_1, \ldots, r_N\}$, the nonlinear non-Gaussian model (9) reduces to a linear Gaussian model:

$$y^\star_i = \log \lambda_i + m_{r_i}(N_i) + \varepsilon_i, \quad \varepsilon_i | r_i \sim \mathrm{No}(0, s^2_{r_i}(N_i)),$$

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

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 utilities $\boldsymbol{y}^\star$ 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 $\lambda_i = \min(\max(y_i/N_i, 0.05), 0.95)$. To justify sampling of the aggregated utility $y_i^\star$, we use (12) and represent the individual utilities y_{ni}^u as in Frühwirth-Schnatter and Frühwirth (2007):

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

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

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

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

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 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 $N = 7$ groups having non-survivor by fitting various binomial logit regression models with appropriate design matrices $\boldsymbol{Z}_i$ and unknown regression parameter $\boldsymbol{\alpha}$:

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

Under the multivariate normal prior $\boldsymbol{\alpha} \sim \text{No}(\mathbf{a}_0, \mathbf{A}_0)$ the improved sampler is implemented as described in Subsection 3.1.2. The conditional posterior $p(\boldsymbol{\alpha}|\boldsymbol{\tau}, \boldsymbol{R}, \boldsymbol{y})$ in Step 1 is again a multivariate normal distribution.

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_7 defines the survival rate of the baseline, chosen to be an adult male in the first class, whereas α_j, $1 \leq j \leq 6$, captures the difference in the survival rate of group j compared to the baseline.

15,000 posterior draws were generated after a burn-in of 5,000 draws, using the prior $\mathbf{a}_0 = \mathbf{0}$ and $\mathbf{A}_0 = 4 \cdot \mathbf{I}$. Posterior inference is summarized in Table 2 showing 95% HPD regions of all regression coefficients along with inefficiency factors for the improved and the original sampler. For each regression coefficient the HPD regions were computed marginally as the shortest interval containing 95% of the posterior draws. The inefficiency factor $\tau = 1 + 2 \cdot \sum_{s=1}^{v} \rho(s)$, where $\rho(s)$ is the empirical autocorrelation at lag s, is computed as in Geyer (1992).

We observe a considerable reduction in computing time, the improved sampler being 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, see 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 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 meaning that they did not have a higher chance to survive than an adult male in the first class. Most likely, in the third class children stayed with their parents, thus had the same chance to survive as an adult in this class. We fitted several reduced regression models to test this hypothesis and computed marginal likelihoods as in Frühwirth-Schnatter and Wagner (2007), see Table 3. For the model with the largest marginal likelihood α_1 equals α_3 and α_2 equals α_6 had meaning that for a girl in the third class her chance was equal to the chance of the mother, while for a boy is was equal to the chance of his father.

Table 2: Posterior inference for the Titanic data

Group i	y_i/N_i	95% HPD region for α_i	inefficiency factors 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

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

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

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)

of times category k is observed on occasion i. Model (9) is modified accordingly:

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

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 utilities $\boldsymbol{y}_i^\star = (y_{1i}^\star, \ldots, y_{mi}^\star)$ as latent variables, in a similar manner as for binomial data, see also (11):

$$y_{ki}^\star = \log \lambda_{ki} + \varepsilon_{ki}, \tag{17}$$

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 $k = 1, \ldots, m$:

$$y_{ki}^\star = \log \lambda_{ki} + m_{r_{ki}}(N_i) + \varepsilon_{ki},$$

with $\log \lambda_{ki} = (\boldsymbol{Z}_i^\alpha)^T \boldsymbol{\alpha}_k + (\boldsymbol{Z}_i^\beta)^T \boldsymbol{\beta}_{ki}$.

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 utilities $\boldsymbol{y}^\star$ 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:

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

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

$$\text{pr}\{r_{ki} = j | y^\star_{ki}, \lambda_{ki}\} \propto w_j(N_i)\, \varphi(y^\star_{ki} - \log \lambda_{ki}; m_j(N_i), s^2_j(N_i)). \qquad (19)$$

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

4 Statistical Modeling Based on Auxiliary Mixture Sampling

Auxiliary mixture smapling allows straightforward statistical modeling of non-Gaussian data as demonstrated for non-Gaussian state-space and random-effects in Frühwirth-Schnatter and Wagner (2006a) and Frühwirth-Schnatter and Frühwirth (2007). Further illustration is provided by LeSage et al. (2007) who analyze knowledge spillovers across Europe through a Poisson spatial interaction model, by Fahrmeir and Steinert (2006) who evaluate post war human security in Cambodia using a geoadditive Bayesian latent variable model for Poisson indicators, and by Gschlößl and Czado (2005) who model the expected number of claims for policy holders of a German car insurance company using spatial regression modelling. 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 count data with a Gaussian 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, Section 9.4) for a review. A finite mixture of Poisson regression models, for instance, reads:

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

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

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 is 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, Section 3.5). In Step 1 in Subsections 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

We reconsider regression analysis of the fabric fault data (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 ($\mathcal{M}_1^P$) and a negative binomial regression model ($\mathcal{M}_1^{NB}$), as well as finite mixtures of Poisson ($\mathcal{M}_K^P$) and negative binomial regressions models ($\mathcal{M}_K^{NB}$) with up to $K = 4$ groups. Furthermore we consider mixtures of regression models, where the intercept is group specific, while the slope is fixed, both for the Poisson ($\mathcal{M}_K^{P,F}$) and the negative binomial distribution ($\mathcal{M}_K^{NB,F}$).

Bayesian analysis was carried out under a Dirichlet $\mathrm{Di}(4, \ldots, 4)$ prior for the weights $(\eta_1, \ldots, \eta_K)$ and a $\mathrm{No}(0, 4)$ prior both for fixed as well as for group specific regression coefficient. For the negative binomial distribution the number of degrees of freedom ρ_k was assumed to be group specific with following the prior: $p(\rho_k) \propto 2d\rho_k/(\rho_k + d)^3$ with median $d(1 + \sqrt{2}) = 10$. We sampled 10,000 posterior draws, after a burn-in of 2,000, using the improved sampler, requiring between 60 and 90 CPU seconds per model. For each model, marginal likelihoods were computed as in Frühwirth-Schnatter and Wagner (2007).

The posterior distribution of ρ shown in Figure 2 for a negative binomial regression model clearly indicates overdispersion compared to the Poisson distribution. This is confirmed by the marginal likelihoods $p(\boldsymbol{y}|\mathcal{M}_K^P)$ and $p(\boldsymbol{y}|\mathcal{M}_K^{NB})$ shown in Table 4. Furthermore, the marginal likelihoods lead to selecting a negative binomial regressions model rather than a mixture of two Poisson regression models as claimed by several authors (Aitkin, 1996; McLachlan and Peel, 2000). The estimated parameters are given in Table 5.

4.2 Gaussian Markov Random Field Models

We discuss now another extension of (1), where the joint distribution of $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_N$ is a hierarchical Gaussian Markov random field (GMRF). More specifically, in the first stage of the model responses y_i are conditionally independent Poisson with mean $e_i \exp(\beta_i)$, where e_i are exposures, in the second stage $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_N)^T$ is multivariate Gaussian with mean $\boldsymbol{u} = (u_1, \ldots, u_N)^T$ and diagonal precision matrix $\omega\mathbf{I}$,

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

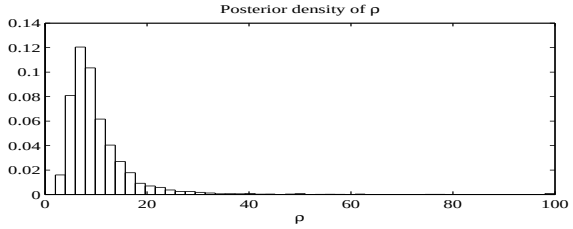

Figure 2: Fabric fault data; posterior density of ν under a negative binomial regression model.

Table 4: Log marginal likelihoods of various regression models for the fabric data (standard errors are given in parenthesis)

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 (fixed slope)	–101.79	**–97.46**	–97.65	–98.60
	(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 (fixed slope)	**–96.04**	–97.25	–98.76	–99.97
	(0.007)	(0.044)	(0.046)	(0.060)

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
ν	(2.9,22.2)	8.9

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

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

see Rue and Held (2005). In (22), $i \sim j$ denotes all pairs of adjacent observations i and j. This prior leaves the overall level of the GMRF unspecified, as only differences of log relative risk parameters enter in (22). For the unknown precision parameter ω and κ we adopt the usual (independent) Gamma hyperpriors, $\omega \sim \text{Ga}(a, b)$ and $\kappa \sim \text{Ga}(c, d)$,with $a = c = 1.0$ and $b = d = 0.01$.

4.2.1 Parameter estimation via MCMC

Statistical inference via MCMC in this highly parametrized model is difficult, especially if the data are sparse.

Joint block updating of $\boldsymbol{\beta}$ and $\boldsymbol{u}$, as proposed in Knorr-Held and Rue (2002), is based on the GMRF approximation as described in detail in Rue and Held (2005, Subsection 4.4.1). Basically a GMRF Metropolis-Hastings proposal is computed based on a quadratic Taylor approximation to the Poisson likelihood. 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 parameter. This specific choice has the advantage that the proposal ratio in the Metropolis-Hastings acceptance probability equals one. The proposal is used for both κ and ω. Subsequently $\boldsymbol{\beta}$ and $\boldsymbol{u}$ are sampled based on the GMRF approximation, as described above. 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 conditional distribution of $\boldsymbol{\beta}$ and $\boldsymbol{u}$ given κ and ω is already a GMRF, so no approximation is necessary. Step 1 of the auxiliary sampler presented in Subsection 2.1.2 may be implemented as a two-block Gibbs steps by first updating $\boldsymbol{\beta}$ and $\boldsymbol{u}$ conditional on κ and ω and than updating κ and ω conditional on $\boldsymbol{\beta}$ and $\boldsymbol{u}$.

To speed up convergence, it may be necessary to implement Step 1 as a Metropolis-Hastings move which updates $\boldsymbol{\beta}$, $\boldsymbol{u}$, κ and ω jointly with the same proposal for κ and ω as described above. Since the full conditional for $\boldsymbol{\beta}$ and $\boldsymbol{u}$ is Gaussian, this joint step updates κ and ω from the marginal posterior where the latent Gaussians $\boldsymbol{\beta}$ and $\boldsymbol{u}$ are integrated out (Rue and Held, 2005, p. 141).

4.2.2 Application to disease mapping

A typical application of a hierarchal GMRF model is disease mapping. This model assumes 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 the unknown log relative risk parameters. The model proposed in Besag et al. (1991) decomposes the log relative risk into spatially structured and unstructured heterogeneity, by assuming the hierarchal prior (21) and (22).

We now report results from an empirical comparison of improved auxiliary mixture sampling using the joint Metropolis-Hastings move and the GMRF approximation based on two data sets. 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 fairly common with a total of 12,835 cases (median of 15), whereas the second one is sparse with a total of 619 cases (median of 1 per district).

Tables 6 and 7 summarize the results for the Sardinia and Germany data, respectively, showing the effective sample 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 simplicity, we have used the same factor for both precision parameters, although this could be changed easily. ESS is an estimate of the number of independent samples

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
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
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
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
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

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 $M = 2,000$ 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.

Concerning Table 6 we note 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. 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 favour of the improved sampler with up to four times as large effective sample sizes per second. Interestingly, the acceptance rates are now higher for the auxiliary mixture sampler, except for the third case where the scaling parameter is quite large. 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 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 Subsections 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.

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, $x \sim \text{Ga}(\nu, 1)$. 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

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

and the characteristic function

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

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

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

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

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

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:

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

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 $7 \cdot 10^{-10}$. 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

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

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 $t_{\max} = 5 \cdot 10^{-4}$, 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 3 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.

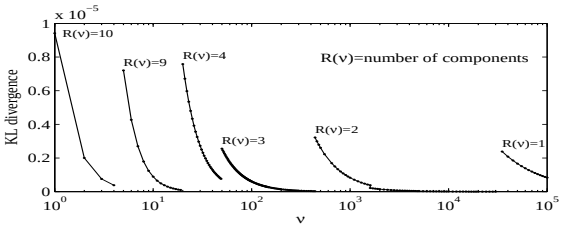

Figure 3: Kullback-Leibler divergence of the estimated mixtures from the standardized negative log-Gamma distribution as a function of the shape parameter ν, for $1 \leq \nu \leq 100000$. $R(\nu)$ is the number of components in the mixtures.

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

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 4 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, 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. (1996). A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and Computing 6*, 251–262.

Besag, J., York, J. C., and Mollié A. (1991), "Bayesian image restoration with two applications in spatial statistics" (with discussion), *Annals of the Institute of Statistical Mathematics,* 43, 1–59.

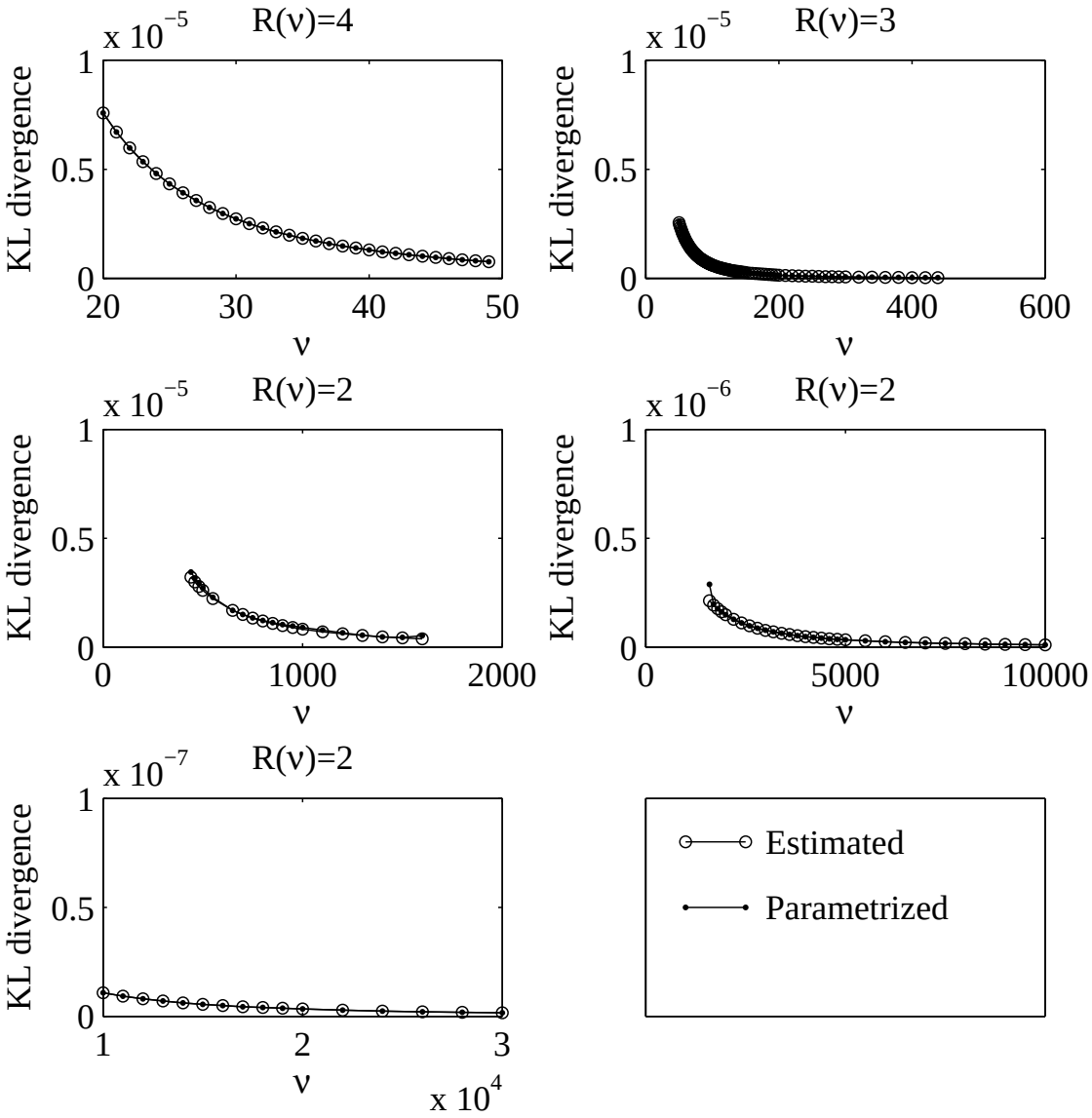

Figure 4: Kullback-Leibler divergence of the estimated and of the parametrized mixtures from the standardized negative log-Gamma distribution as a function of the shape parameter ν, for $20 \leq \nu \leq 100000$. $R(\nu)$ is the number of components in the mixtures.

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

Chiogna, M. and C. Gaetan (2002). Dynamic generalized linear models with application to environmental epidemiology. *Applied Statistics 51*, 453–468.

Fahrmeir, L. and S. Steinert (2006). A geoadditive Bayesian latent variable model for Poisson indicators. Technical report, University of Munich.

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

Frühwirth-Schnatter, S. and R. Frühwirth (2007). Auxiliary mixture sampling with applications to logistic models. *Computational Statistics and Data Analysis 51*, 3509–3528.

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

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

Frühwirth-Schnatter, S. and H. Wagner (2007). Marginal likelihoods for non-Gaussian models using auxiliary mixture sampling. Research Report IFAS 2007-24, http://www.ifas.jku.at/.

Geyer, C. (1992). Practical Markov chain Monte Carlo. *Statistical Science 7*, 473–511.

Gschlößl, S. and C. Czado (2005). Does a Gibbs sampler approach to spatial Poisson regression models outperform a single site MH sampler? Technical report, Center of Mathematical Sciences, University of Technology.

Held, L., Natario, I., Fenton, S., Rue, H., and Becker, N. (2005), "Towards joint disease mapping," *Statistical Methods in Medical Research*, 14, 61–82.

Held, L., Graziano, G., Frank, C., and Rue, H. (2006), "Joint spatial analysis of gastrointestinal infectious diseases," *Statistical Methods in Medical Research*, 15, 465–480.

Hilbe, J. M. (2007). *Negative binomial regression*. Cambridge: Cambridge University Press.

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

Hurn, M., A. Justel, and C. P. Robert (2003). Estimating mixtures of regressions. *Journal of Computational and Graphical Statistics 12*, 55–79.

Kass, R. E., Carlin, B., Gelman, A., and Neal, R. (1998), "Markov chain Monte Carlo in practice: A roundtable discussion," *The American Statistician*, 52, 93–100.

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

Knorr-Held, L. and H. Rue (2002). On block updating in Markov random field models for disease mapping. *Scandinavian Journal of Statistics 29*, 597–614.

LeSage, J. P., M. M. Fischer, and T. Scherngell (2007). Knowledge spillovers across Europe: Evidence from a Poisson spatial interaction model with spatial effects. *Papers in Regional Science 86*, 393–421.

Omori, Y., S. Chib, N. Shephard, and J. Nakajima (2007). Stochastic volatility with leverage: Fast likelihood inference. *Journal of Econometrics 140*, 425–449.

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

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

Nelder, J. A., and Mead, R. (1965), "A Simplex Method for Function Minimization," *Computer Journal*, 7, 308–313.

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

Shephard, N. (1994). Partial non-Gaussian state space. *Biometrika 81*, 115–131.

Tüchler, R. (2007). Bayesian variable selection for logistic models using auxiliary mixture sampling. *Journal of Computational and Graphical Statistics*, to appear.

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