

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

Sylvia Frühwirth-Schnatter<sup>a</sup>, Rudolf Frühwirth<sup>b</sup>,  
Leonhard Held<sup>c</sup> and Håvard Rue<sup>d</sup>

<sup>a</sup>*Department of Applied Statistics and Econometrics  
Johannes Kepler Universität Linz, Austria*

<sup>b</sup>*Institute of High Energy Physics  
Austrian Academy of Sciences, Vienna, Austria*

<sup>c</sup>*Institute of Social and Preventive Medicine  
University of Zurich, Switzerland*

<sup>d</sup>*Department of Mathematical Sciences  
Norwegian University of Science and Technology, Trondheim, Norway*

May 29, 2007

## ABSTRACT

The article proposes an improved method of auxiliary mixture sampling for count data, binomial data and multinomial data. In contrast to previously proposed samplers the method uses a limited 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 smaller 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 on two epidemiological case studies.

*Key words:* Count data, Binomial data, Disease mapping, Gaussian mixture, Log-Gamma distribution, Multinomial data

*Submitted to Journal of Computational and Graphical Statistics*

# 1 Introduction

During the past years, auxiliary mixture sampling has turned out to be a useful tool for the Bayesian analysis of hierarchical and parameter-driven models of non-Gaussian data. The method has been used first by Shephard (1994) for stochastic volatility models and has been applied in this context by a couple of authors (Kim et al., 1998; Chib et al., 2002; Omori et al., 2004). Recently, auxiliary mixture sampling has been extended to rather general hierarchical models for non-Gaussian data like state-space and random-effects models (Frühwirth-Schnatter and Wagner, 2005, 2006; Frühwirth-Schnatter and Frühwirth, 2007). For each dependent observation  $y_i$  latent variables are introduced the expectation of which depends on the unknown parameters in a linear way. The error distribution follows a type I extreme value distribution, which is then approximated by a Gaussian mixture distribution.

The number of these latent variables differs for the various distribution families. For binary data the (univariate) utility of choosing category 1 is introduced, whereas for data with  $m+1$  categories the utilities of choosing any category but one have dimension  $m$ . For data from the Poisson distribution,  $y_i + 1$  interarrival times are introduced for every observation  $y_i$ ; thus their number is increasing with the underlying intensity. For data from a binomial distribution with repetition parameter  $N_i$  a latent utility is introduced for each of  $N_i$  binary experiments, leading to dimension  $N_i$ . A similar method is applied for multinomial data, where the dimension is equal to  $mN_i$ , leading in both cases to an increasing number of latent variables with increasing number of repetitions.

In this note we propose an improved method of auxiliary mixture sampling that utilizes a limited number of latent variables per observation, namely at most two instead of  $y_i + 1$  for Poisson data, one instead of  $N_i$  for binomial data, and  $m - 1$  instead of  $(m - 1)N_i$  for multinomial data. This leads to a substantial increase in efficiency of auxiliary mixture sampling for highly structured models like random-effects models or other hierarchical models for repeated measurements and for state space modelling of non-Gaussian time series.

The latent variables of the improved method aggregate the latent variables used in Frühwirth-Schnatter and Wagner (2006) and Frühwirth-Schnatter and Frühwirth (2007) in such a way that their expectation is still a linear function of the unknown parameters. The deviation from the expectation,

however, follows a more general distribution, namely the distribution of the negative logarithm of a Gamma random variable with integer shape parameter  $\nu$  and unit scale. The shape parameter is equal to  $y_i$  for Poisson data and to  $N_i$  for data from the binomial and the multinomial distribution. For each latent variable this distribution is approximated by a Gaussian mixture distribution, and the component indicator is introduced as a further auxiliary variable. We discuss the computation of the Gaussian mixture distributions for arbitrary integer values of the shape parameter. Due to the Central Limit Theorem the number of required mixture components drops with rising  $\nu$ . 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, in contrast to the previously proposed sampler.

The modified sampler is applied to two epidemiological case studies, namely disease mapping and analyzing incidence cases of cervical cancer.

## 2 Auxiliary Mixture Sampling for Count Data

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

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

The precise model for  $\boldsymbol{\beta}_i$  is left unspecified at this stage; it could be a spatial, a temporal, or a spatio-temporal model, for example. We only assume that the joint distribution  $p(\boldsymbol{\alpha}, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_N|\boldsymbol{\theta})$  is a normal distribution, indexed by some unknown parameter  $\boldsymbol{\theta}$ . Furthermore we assume that, conditional on knowing  $\boldsymbol{\alpha}, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_N$ ,  $y_i$  and  $y_j$  are mutually independent. In the application presented below the prior model is a Gaussian Markov random field (Rue and Held, 2005).

### 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  $\lambda_i$ , having

occurred in the time interval  $0 \leq t \leq 1$ . In Frühwirth-Schnatter and Wagner (2006), 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. This kind of auxiliary mixture sampling seems to be infeasible for high intensity data or panels of count data with a high number of total observations.

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

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

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

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

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

$$-\log \tau_{i1}^* = \log \lambda_i + \varepsilon_{i1}, \quad (4)$$

$$-\log \tau_{i2}^* = \log \lambda_i + \varepsilon_{i2}, \quad (5)$$

where  $\varepsilon_{i1} = -\log \xi_{i1}$  with  $\xi_{i1} \sim \text{Ex}(1) = \text{Ga}(1, 1)$  and  $\varepsilon_{i2} = -\log \xi_{i2}$  with  $\xi_{i2} \sim \text{Ga}(y_i, 1)$ . For  $y_i = 0$  we are dealing only with equation (4).

The first step of improved auxiliary mixture sampling introduces the bivariate latent variable  $\tau_i = (\tau_{i1}^*, \tau_{i2}^*)$  for each nonzero observation  $y_i$  and the single latent variable  $\tau_i = \tau_{i1}^*$  for zero observations. In the second step the densities of  $\varepsilon_{i1}$  and  $\varepsilon_{i2}$  in (4) and (5) are approximated by Gaussian mixtures, and for both mixture distributions 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.

For the distribution of  $\varepsilon_{i1}$  the same mixture approximation is used as in Frühwirth-Schnatter and Wagner (2006). Finding a mixture approximation for  $\varepsilon_{i2}$  is more challenging because this is a negative log-Gamma distribution with integer shape parameter  $\nu$  equal to  $y_i$ . In Subsection 2.3 such an

approximation is derived for arbitrary integer shape parameters  $\nu$ ,

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

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

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

$$\begin{aligned} -\log \tau_{i1}^* &= \log \lambda_i + m_{r_{i1}}(1) + \varepsilon_{i1}, & \varepsilon_{i1} | r_{i1} &\sim N(0, s_{r_{i1}}^2(1)), \\ -\log \tau_{i2}^* &= \log \lambda_i + m_{r_{i2}}(y_i) + \varepsilon_{i2}, & \varepsilon_{i2} | r_{i2} &\sim N(0, s_{r_{i2}}^2(y_i)), \end{aligned}$$

with  $\log \lambda_i = (\mathbf{Z}_i^\alpha)^T \boldsymbol{\alpha} + (\mathbf{Z}_i^\beta)^T \boldsymbol{\beta}_i$ . For  $y_i = 0$  we are dealing only with the first equation. Consequently, the conditional posterior  $p(\boldsymbol{\alpha}, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_N | \boldsymbol{\theta}, \boldsymbol{\tau}, \boldsymbol{S}, \mathbf{y})$  is proportional to a multivariate normal density:

$$\begin{aligned} p(\boldsymbol{\alpha}, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_N | \boldsymbol{\theta}, \boldsymbol{\tau}, \boldsymbol{S}, \mathbf{y}) &\propto & (7) \\ & p(\boldsymbol{\alpha}, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_N | \boldsymbol{\theta}) \prod_{i=1}^N f_N(-\log \tau_{i1}^*; \log \lambda_i + m_{r_{i1}}(1), s_{r_{i1}}^2(1)) \\ & \prod_{i=1, y_i \neq 0}^N f_N(-\log \tau_{i2}^*; \log \lambda_i + m_{r_{i2}}(y_i), s_{r_{i2}}^2(y_i)). \end{aligned}$$

### 2.1.2 The sampling scheme

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

- (1) Sample  $\boldsymbol{\alpha}, \boldsymbol{\beta} = \{\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_N\}$  and  $\boldsymbol{\theta}$ , conditional on  $\boldsymbol{\tau}, \boldsymbol{S}$ , and  $\mathbf{y}$ .
- (2) Sample the interarrival times  $\boldsymbol{\tau}$  and the component indicators  $\boldsymbol{S}$  conditional on  $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}$  and  $\mathbf{y}$  by running the following steps, for  $i = 1, \dots, N$ .
  - (a) Sample  $\xi_i \sim \text{Ex}(\lambda_i)$ . If  $y_i = 0$ , set  $\tau_{i1}^* = 1 + \xi_i$ . If  $y_i > 0$ , sample  $\tau_{i2}^*$  from a Beta( $y_i, 1$ )-distribution and set  $\tau_{i1}^* = 1 - \tau_{i2}^* + \xi_i$ .
  - (b) Sample the component indicator  $r_{i1}$  from the following discrete distribution where  $k = 1, \dots, R(1)$ :

$$\text{pr}\{r_{i1} = k | \tau_{i1}, \boldsymbol{\theta}, \boldsymbol{\beta}_i, \boldsymbol{\alpha}\} \propto$$

$$\frac{w_k(1)}{s_k(1)} \exp \left\{ -\frac{1}{2} \left( \frac{-\log \tau_{i1} - \log \lambda_i - m_k(1)}{s_k(1)} \right)^2 \right\}.$$

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

$$\begin{aligned} \text{pr}\{r_{i2} = k | \tau_{i2}, \boldsymbol{\theta}, \boldsymbol{\beta}_i, \boldsymbol{\alpha}, y_i\} &\propto \\ \frac{w_k(y_i)}{s_k(y_i)} \exp \left\{ -\frac{1}{2} \left( \frac{-\log \tau_{i2} - \log \lambda_i - m_k(y_i)}{s_k(y_i)} \right)^2 \right\}. \end{aligned}$$

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  $\boldsymbol{S}$ . For model (1), for instance, we may implement a Gibbs type move, by first sampling  $(\boldsymbol{\alpha}, \boldsymbol{\beta})$  conditional on  $\boldsymbol{\theta}$  from the multivariate normal distribution (7), and then sampling  $\boldsymbol{\theta}$  conditional on  $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ .

To speed up convergence, it may be necessary to implement a joint move which updates  $\boldsymbol{\theta}$  and  $(\boldsymbol{\alpha}, \boldsymbol{\beta})$  jointly (Knorr-Held and Rue, 2002). We use the following construction. First, propose a new value for  $\boldsymbol{\theta}$ , say  $\boldsymbol{\theta}'$ , using for example the simple proposal  $\boldsymbol{\theta}' \sim q(\boldsymbol{\theta}' | \boldsymbol{\theta})$ . Then, conditioned on  $\boldsymbol{\theta}'$ , sample a new proposal  $(\boldsymbol{\alpha}', \boldsymbol{\beta}')$  from the full (Gaussian) conditional for  $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ . Finally, accept/reject  $(\boldsymbol{\alpha}', \boldsymbol{\beta}', \boldsymbol{\theta}')$  jointly. The rationale for such a construction is to break the strong dependency between  $\boldsymbol{\theta}$  and  $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ , and is discussed in great detail by Rue and Held (2005, Sec. 4.1). Since the full conditional for  $(\boldsymbol{\alpha}, \boldsymbol{\beta})$  is Gaussian, this joint step updates  $\boldsymbol{\theta}$  using the proposal  $q(\boldsymbol{\theta}' | \boldsymbol{\theta})$  from the joint posterior where the latent Gaussians  $(\boldsymbol{\alpha}, \boldsymbol{\beta})$  are integrated out (Rue and Held, 2005, p. 141).

Step 2 is an appropriate modification of the corresponding step in Frühwirth-Schnatter and Wagner (2006, Subsection 3.1), based on decomposing the joint posterior of  $(\boldsymbol{\tau}, \boldsymbol{S})$  as

$$p(\boldsymbol{\tau}, \boldsymbol{S} | \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = p(\boldsymbol{S} | \boldsymbol{\tau}, \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \cdot p(\boldsymbol{\tau} | \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}).$$

We first sample the arrival times  $\tau_1, \dots, \tau_N$  from the density  $p(\tau_i | y_i, \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta})$  as they are independent for different time points  $i$ , given  $\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\alpha}$  and  $\boldsymbol{y}$ . For any  $i$  with  $y_i > 0$  the joint distribution of  $(\tau_{i1}^*, \tau_{i2}^*)$  factorizes as

$$p(\tau_{i1}^*, \tau_{i2}^* | y_i, \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = p(\tau_{i1}^* | y_i, \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \tau_{i2}^*) \cdot p(\tau_{i2}^* | y_i).$$

Conditionally on  $y_i$ , only  $y_i$  jumps occur in  $[0, 1]$ , whereas the  $(y_i + 1)$ th jump occurs after  $t = 1$ . By well-known properties of a Poisson process, the arrival

time  $\tau_{i2}^*$  of the  $y_i$ th 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). As a result of the zero-memory property of the exponential distribution, the waiting time until the first jump after  $t = 1$  is distributed as  $\text{Ex}(\lambda_i)$ , and therefore  $\tau_{i1}^* = 1 - \tau_{i2}^* + \xi_i$ , where  $\xi_i \sim \text{Ex}(\lambda_i)$ . This justifies Step 2(a).

To sample the indicators  $\mathbf{S}$  from  $p(\mathbf{S}|\boldsymbol{\tau}, \mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ , we use the fact that all indicators are conditionally independent given  $\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}$  and  $\boldsymbol{\tau}$ :

$$p(\mathbf{S}|\boldsymbol{\tau}, \mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{i=1}^N \prod_{j=1}^{\min(y_i+1, 2)} p(r_{ij}|\tau_{ij}, \boldsymbol{\theta}, \boldsymbol{\beta}_i, \boldsymbol{\alpha}, \mathbf{y}).$$

Thus for each  $i = 1, \dots, N$ , the indicators  $r_{i1}$  and, if  $y_i > 0$ ,  $r_{i2}$ , are sampled independently from  $p(r_{i1}|\tau_{ij}, \boldsymbol{\theta}, \boldsymbol{\beta}_i, \boldsymbol{\alpha}, y_i)$  and  $p(r_{i2}|\tau_{ij}, \boldsymbol{\theta}, \boldsymbol{\beta}_i, \boldsymbol{\alpha}, y_i)$  which obviously are equal to the discrete densities given in step 2(b).

Starting values for  $\boldsymbol{\tau}$  and  $\mathbf{S}$  are obtained in the following way. The component indicator  $r_{i1}$  is drawn uniformly from 1 to  $R(1)$ , and, if  $y_i > 0$ , the component indicator  $r_{i2}$  is drawn uniformly from 1 to  $R(y_i)$ . Step 2(a) is used to sample starting values for  $\tau_{i2}^*$ . To obtain a starting value for  $\tau_{i1}^*$ , we sample  $\xi_i$  from  $\text{Ex}(\lambda_i)$  with  $\lambda_i = y_i$ , if  $y_i > 0$ . For all  $i$  where  $y_i = 0$ ,  $\lambda_i$  is set to a small value; in our examples we used  $\lambda_i = 0.1$ .

### 2.1.3 Adding a rejection step

A rejection step could be added as in Frühwirth-Schnatter and Wagner (2006, Subsection 3.2) to evaluate the accuracy of auxiliary mixture sampling. We have computed the acceptance rate of the improved auxiliary mixture sampler for the simple example discussed there, namely Bayesian inference for  $N$  independent observations  $y_1, \dots, y_N$  from the  $\text{Po}(\lambda)$  distribution under the prior  $\lambda \sim \text{Ga}(a_0, b_0)$ , in which case  $\lambda|y \sim \text{Ga}(a_0 + N\bar{y}, b_0 + N)$ , with  $\bar{y}$  being the sample mean.

Note that for the auxiliary mixture sampler of Frühwirth-Schnatter and Wagner (2006) on average  $N(1 + \lambda)$  mixture approximations take place, whereas for the new sampler this number is equal to  $N(2 - e^{-\lambda})$ . For  $\lambda = 10$  and  $N = 1000$ , for instance, the expected number of approximations is equal to 2000 instead of 11000.

Table 1 demonstrates that the approximation error of the new sampler is even smaller than the approximation error the sampler of Frühwirth-Schnatter and Wagner (2006), in particular for large values of  $\lambda$ .

## 2.2 Application to disease mapping

Bayesian hierarchical models with Poisson observations often arise in epidemiological applications. A typical example is the area of disease mapping, where a commonly used formulation assumes that the observed disease counts  $y_i$  in district  $i = 1, \dots, N$  are conditionally independent Poisson with mean  $e_i \exp(\eta_i)$ , where  $e_i$  are known expected counts and  $\eta_i$  are the unknown log relative risk parameters. The model proposed in Besag et al. (1991) now decomposes the log relative risk into spatially structured and unstructured heterogeneity. More specifically, in the first stage of the hierarchical model responses  $y_i$  are conditionally independent Poisson with mean  $e_i \exp(\eta_i)$ , in the second stage  $\boldsymbol{\eta} = (\eta_1, \dots, \eta_N)^T$  is multivariate Gaussian with mean  $\mathbf{u} = (u_1, \dots, u_N)^T$  and diagonal precision matrix  $\lambda \mathbf{I}$ , and in the third stage  $\mathbf{u}$  follows an intrinsic Gaussian Markov random field (GMRF)

$$\pi(\mathbf{u}|\kappa) \propto \kappa^{\frac{N-1}{2}} \exp\left(-\frac{\kappa}{2} \sum_{i \sim j} (u_i - u_j)^2\right), \quad (8)$$

see Rue and Held (2005). In (8),  $i \sim j$  denotes all pairs of adjacent districts  $i$  and  $j$ . This prior leaves the overall level of the GMRF unspecified, as only differences of log relative risk parameters enter in (8). For the unknown precision parameter  $\lambda$  and  $\kappa$  we adopt the usual (independent) Gamma hyperpriors, say  $\lambda \sim G(a, b)$  and  $\kappa \sim G(c, d)$ , we have used  $a = c = 1.0$  and  $b = d = 0.01$ .

Statistical inference via MCMC in this highly parametrized model is difficult, especially if the data are sparse. Joint block updating of  $\boldsymbol{\eta}$  and  $\mathbf{u}$ , 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 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  $\kappa$  and  $\lambda$ . Subsequently  $\boldsymbol{\eta}$  and  $\mathbf{u}$  are sampled based on the GMRF approximation, as described above. Finally, all updated parameters are accepted or rejected in a joint Metropolis-Hastings step. For

further details see Knorr-Held and Rue (2002).

Alternatively, the proposed auxiliary variable approach can be implemented in this setting. This has the distinct advantage that the conditional distribution of  $\boldsymbol{\eta}$  and  $\mathbf{u}$  is already a GMRF, so no approximation is necessary and  $\boldsymbol{\eta}$  and  $\mathbf{u}$  can be updated with a Gibbs step. In the joint update, this Gibbs proposal will replace the GMRF approximation.

We now report results from an empirical comparison of both algorithms based on two datasets. 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), while the second is more common with a total of 12,835 cases (median of 15).

Table 2 and 3 summarize the results for the Sardinia and Germany data respectively: reported is the effective sample size (ESS) (Kass et al., 1998) and the effective sample size per second for the two precision parameters  $\lambda$  and  $\kappa$  (both on a log scale) and the posterior deviance  $D$ , defined for example in Spiegelhalter et al. (2002). 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 which would be required to obtain a parameter estimate with the same precision as the MCMC estimate based on  $n$  dependent samples (here we used  $n = 2,000$  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  $n$  used from the Markov chain divided by the empirical autocorrelation time

$$\tau = 1 + 2 \cdot \sum_{s=1}^v \rho(s),$$

where  $\rho(s)$  is the empirical autocorrelation at lag  $s$ . The initial monotone sequence estimator by Geyer (1992) is used to determine  $v$  based on the sum of adjacent pairs of empirical autocorrelations

$$\Phi(s) = \rho(2 \cdot s) + \rho(2 \cdot s + 1).$$

Let  $k$  be the largest integer so that  $\Phi(s) > 0$  and  $\Phi(s)$  is monotone for  $s = 1, \dots, k$ , then  $v$  is defined as  $v = 2 \cdot k + 1$ .

First commenting on Table 2 we note that the auxiliary mixture sampling (AMS) 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 — auxiliary mixture sampling yields acceptance rates equal to unity, whereas the GMRF approximation has acceptance rates of approximately 70% for these data (Knorr-Held and Rue, 2002). 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 3, the results are even more in favour of the auxiliary mixture 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.

## 2.3 Approximation of the Negative Log-Gamma Distribution by Gaussian Mixtures

### 2.3.1 The negative log-Gamma distribution

Assume that  $x$  is Gamma-distributed with integer shape parameter  $\nu$  and unit scale,  $x \sim \text{Ga}(\nu, 1)$ . This distribution is the convolution of  $\nu$  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

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

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

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

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

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

This has the advantage that the effective support of the distribution is almost independent of  $\nu$ . Still, for small values of  $\nu$  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  $\nu$ . For large  $\nu$ , the distribution of  $u$  approaches the standard normal distribution. Approximation by Gaussian mixtures therefore requires less components for increasing  $\nu$ .

### 2.3.2 Approximation by Gaussian mixtures

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

$$\begin{aligned} D(\mathbf{w}, \mathbf{m}, \mathbf{s}^2) &= \int_{\mathcal{S}} f(u; \nu) \log \frac{f(u; \nu)}{f_{\text{N}}(u, \mathbf{w}(\nu), \mathbf{m}(\nu), \mathbf{s}^2(\nu))} du \quad (9) \\ &+ \lambda \left( \sum_{r=1}^{R(\nu)} w_r - 1 \right)^2, \end{aligned}$$

where  $f_{\text{N}}(u, \mathbf{w}(\nu), \mathbf{m}(\nu), \mathbf{s}^2(\nu))$  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  $\lambda = 10^9$ . Note that  $d_{\text{KL}}$  is invariant under affine transformations and in particular under standardization. The integral in (9) 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). Approximating mixtures were computed for the following values of  $\nu$ :

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

An approximation was accepted only if the Kullback-Leibler divergence  $d_{\text{KL}}$  of the mixture density from the target density was below a threshold  $t_{\text{KL}}$  and if the maximum absolute difference  $d_{\text{max}}$  between the two densities was below a threshold  $t_{\text{max}}$ . We chose  $t_{\text{KL}} = 10^{-5}$  and  $t_{\text{max}} = 5 \cdot 10^{-4}$ . 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  $d_{\text{KL}} \leq t_{\text{KL}}$  and  $d_{\text{max}} \leq t_{\text{max}}$ .
- (2) Save the estimated parameters.
- (3) Reduce the number of components by 1.
- (4) Compute new starting values by merging the smallest component with its smallest neighbour.
- (5) Minimize the objective function.
- (6) If  $d_{\text{KL}} \leq t_{\text{KL}}$  and  $d_{\text{max}} \leq t_{\text{max}}$ , go to step 2.
- (7) Otherwise, store the saved parameters.

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

### 2.3.3 Parametrization of the mixtures

For small values of  $\nu$  the mixture parameters change substantially when  $\nu$  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  $\nu$  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  $\nu$  summarized in Table 4. 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 2 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 implementation is available from the authors; an implementation in C is included in the GMRFLib library (Rue and Held, 2005, Appendix).

## 3 Auxiliary Mixture Sampling for Binomial and Multinomial Data

### 3.1 Dealing with Binomial Data

We present details for the following model. Let  $\mathbf{y} = (y_1, \dots, y_N)$  be a sequence of data from a binomial distribution and assume that

$$\begin{aligned} y_i | \pi_i &\sim \text{Bino}(N_i, \pi_i), \\ \log \frac{\pi_i}{1 - \pi_i} &= \log \lambda_i = (\mathbf{Z}_i^\alpha)^T \boldsymbol{\alpha} + (\mathbf{Z}_i^\beta)^T \boldsymbol{\beta}_i, \end{aligned} \tag{10}$$

with  $N_i$  being known. The precise model for  $\boldsymbol{\beta}_i$  is left unspecified at this stage; we only assume that the joint distribution  $p(\boldsymbol{\alpha}, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_N | \boldsymbol{\theta})$  is a normal distribution, indexed by some unknown parameter  $\boldsymbol{\theta}$ . Furthermore

we assume that conditional on knowing  $\alpha, \beta_1, \dots, \beta_N, y_i$  and  $y_j$  are mutually independent.

### 3.1.1 Data augmentation

For each  $i$ , the distribution of  $y_i|\pi_i$  is regarded as the distribution of the number of successes in  $N_i$  independent binary experiments with success probability  $\pi_i$ . As in Frühwirth-Schnatter and Frühwirth (2007), we recover the full binary experiment, involving the repeated binary measurements  $z_{ni}$ , where

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

and  $z_{ni}$  follows a binary logit model with the same log odds ratio as (10):

$$\text{pr}\{z_{ni} = 1|\pi_i\} = \pi_i = \frac{\lambda_i}{1 + \lambda_i}.$$

The first step of data augmentation in Frühwirth-Schnatter and Frühwirth (2007) introduces for each binary observation  $z_{ni}$  the utility  $y_{ni}^u$  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. This kind of auxiliary mixture sampling seems to be infeasible for data with a high number of total repetitions  $\sum_{i=1}^N N_i$ .

A more efficient method may be derived in the following way. First note that for any utility  $y_{ni}^u$  the following holds for  $n = 1, \dots, N_i$ :

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

where  $\varepsilon_{ni}$  follows a type I extreme value distribution and therefore the random variable  $\exp(-\varepsilon_{ni})$  follows a standard exponential distribution. If we consider the sum over all  $n$  we obtain:

$$\sum_{n=1}^{N_i} \exp(-y_{ni}^u) = \frac{1}{\lambda_i} \xi_i, \quad \xi_i = \sum_{n=1}^{N_i} \exp(-\varepsilon_{ni}). \quad (11)$$

Due to the independence of the binary experiments  $\xi_i$  follows a  $\text{Ga}(N_i, 1)$  distribution. By taking the negative logarithm in (11) we obtain:

$$y_i^* = \log \lambda_i + \varepsilon_i, \quad (12)$$

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

$$y_i^* = -\log \sum_{n=1}^{N_i} \exp(-y_{ni}^u). \quad (13)$$

The first step of improved auxiliary mixture sampling introduces for each binomial observation  $y_i$  the (univariate) aggregated utility  $y_i^*$  as a latent variable, rather than the entire vector of  $N_i$  individual utilities  $y_{1i}^u, \dots, y_{N_i i}^u$ . The second step is exactly the same as for Poisson data. For every  $i$ , the density of  $\varepsilon_i$  in (12), which follows a negative log-Gamma distribution with integer shape parameter  $N_i$ , is approximated by a mixture of normal distributions. 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  $\mathbf{y}^* = \{y_1^*, \dots, y_N^*\}$  and  $\mathbf{S} = \{r_1, \dots, r_N\}$ , the nonlinear non-Gaussian model (10) reduces to a linear Gaussian model where the mean of the observation equation is linear in  $\boldsymbol{\alpha}, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_N$  and the error term follows a normal distribution:

$$y_i^* = \log \lambda_i + m_{r_i}(N_i) + \varepsilon_i, \quad \varepsilon_i | r_i \sim N(0, s_{r_i}^2(N_i)),$$

with  $\log \lambda_i = (\mathbf{Z}_i^\alpha)^T \boldsymbol{\alpha} + (\mathbf{Z}_i^\beta)^T \boldsymbol{\beta}_i$ . Consequently, the conditional posterior  $p(\boldsymbol{\alpha}, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_N | \boldsymbol{\theta}, \mathbf{y}^*, \mathbf{S}, \mathbf{y})$  is multivariate normal:

$$p(\boldsymbol{\alpha}, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_N | \boldsymbol{\theta}, \mathbf{y}^*, \mathbf{S}, \mathbf{y}) \propto p(\boldsymbol{\alpha}, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_N | \boldsymbol{\theta}) \prod_{i=1}^N f_N(y_i^*; \log \lambda_i + m_{r_i}(N_i), s_{r_i}^2(N_i)).$$

If  $N_i \equiv 1$ , model (10) reduces to a binary logit model, and the improved method introduced in this section reduces to the one described for binary data in Frühwirth-Schnatter and Frühwirth (2007).

### 3.1.2 The sampling scheme

Select starting values for  $\mathbf{y}^*$  and  $\mathbf{S}$  and repeat the following steps.

- (1) Sample  $\boldsymbol{\alpha}, \boldsymbol{\beta} = \{\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_N\}$ , and  $\boldsymbol{\theta}$ , conditional on  $\mathbf{y}^*, \mathbf{S}$ , and  $\mathbf{y}$ .
- (2) Sample the aggregated utilities  $\mathbf{y}^*$  and the indicators  $\mathbf{S}$  conditional on  $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}$  and  $\mathbf{y}$ , by running the following steps, for  $i = 1, \dots, N$ .

(a) Sample the aggregated utility  $y_i^*$  conditional on  $\lambda_i$  and  $y_i$  as

$$y_i^* = -\log \left( \frac{U_i}{1 + \lambda_i} + \frac{V_i}{\lambda_i} \right), \quad (14)$$

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 the component indicator  $r_i$  from the following discrete distribution where  $j = 1, \dots, R(N_i)$ :

$$\begin{aligned} \text{pr}\{r_i = j | y_i^*, \boldsymbol{\theta}, \boldsymbol{\beta}_i, \boldsymbol{\alpha}, y_i\} &\propto \\ &\frac{w_j(N_i)}{s_j(N_i)} \exp \left\{ -\frac{1}{2} \left( \frac{y_i^* - \log \lambda_i - m_j(N_i)}{s_j(N_i)} \right)^2 \right\}. \end{aligned} \quad (15)$$

Again step 1 is model dependent, but standard for many models, as we are dealing with a Gaussian model, once we condition on  $\mathbf{y}^*$  and  $\mathbf{S}$ . Step 2 is a modification of the corresponding step in Frühwirth-Schnatter and Frühwirth (2007, Subsection 2.2).

To justify sampling of the aggregated utility  $y_i^*$  as in (14), we use (13) and draw the individual utilities  $y_{ni}^u$  from the posterior distribution  $p(y_{ni}^u | y_i, \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta})$  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:

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

Step 2(a) is justified by the facts that

$$\begin{aligned} \sum_{n=1}^{N_i} (-\log U_{ni}) &\sim \text{Ga}(N_i, 1), \\ y_i < N_i &\implies \sum_{n=y_i+1}^{N_i} (-\log V_{ni}) \sim \text{Ga}(N_i - y_i, 1). \end{aligned}$$

Evidently, the indicators  $r_i$  have to be sampled from the discrete density  $p(r_i | y_i^*, \boldsymbol{\theta}, \boldsymbol{\beta}_i, \boldsymbol{\alpha}, y_i)$  given in (15). Starting values for the component indicator  $r_i$  are drawn uniformly from 1 to  $R(N_i)$ ; to obtain a starting value for  $y_i^*$  we use (14) with  $\lambda_i = 1$ .

### 3.2 Application to cancer incidence data

We have reanalyzed Example 4.3.5 from Rue and Held (2005) with the auxiliary mixture sampler described in Subsection 3.1. The data analyzed are all incidence cases of cervical cancer in the former East German Republic (GDR) from 1979, stratified by district and age group. Each of the  $N = 6\,690$  cases has been classified into either a premalignant (3755 cases) or a malignant (2935 cases) stage. It is of interest to estimate the spatial variation of the incidence ratio of premalignant to malignant cases in the 216 districts, after adjusting for age effects. Age was categorized into  $J = 15$  age groups. For more background information and motivation see Knorr-Held et al. (2000).

Let  $y_i = 1$  denote a premalignant case and  $y_i = 0$  a malignant case. Rue and Held (2005) assume a logistic binary regression model  $y_i \sim \text{Bino}(1, \pi_i)$ ,  $i = 1, \dots, N$  with

$$\text{logit}(\pi_i) = \alpha + \beta_{j(i)} + \gamma_{k(i)},$$

where  $j(i)$  and  $k(i)$  denote age group and district of case  $i$ , respectively. The age group effects  $\beta$  are assumed to follow a random walk of second order,

$$\pi(\beta|\kappa_\beta) \propto \kappa_\beta^{\frac{N-2}{2}} \exp\left(-\frac{\kappa_\beta}{2} \sum_{j=2}^{J-1} (\beta_{j-1} - 2\beta_j + \beta_{j+1})^2\right),$$

see Rue and Held (2005, Section 3.4.1) for more details. For the spatial effect  $\gamma$  we assume that it is the sum of an IGMRF model plus additional unstructured variation:

$$\gamma_k = u_k + v_k.$$

Here,  $\mathbf{u}$  follows the IGMRF model (8) with precision  $\kappa_{\mathbf{u}}$  and  $\mathbf{v}$  is normal with zero mean and diagonal precision matrix with entries  $\kappa_{\mathbf{v}}$ . This model for the spatial effects is just a reparametrization of the one described in Subsection 2.2. For the corresponding precision parameters we assume a  $\text{Ga}(1.0, 0.01)$  for both  $\kappa_{\mathbf{u}}$  and  $\kappa_{\mathbf{v}}$  and a  $\text{Ga}(1.0, 0.0005)$  for  $\kappa_\beta$ . A diffuse prior is assumed for the overall mean  $\alpha$ , and sum-to-zero constraints are placed both on  $\beta$  and  $\mathbf{u}$ .

Let  $\boldsymbol{\kappa} = (\kappa_\beta, \kappa_{\mathbf{u}}, \kappa_{\mathbf{v}})^T$  denote the vector of all precision parameters in the model. Following Holmes and Held (2006), Rue and Held (2005) used auxiliary variables based on the representation of the logistic distribution by a scale-mixture of Gaussians. Due to the nature of the Holmes and Held (2006) algorithm, they had to introduce two auxiliary variables for each binary observation. They grouped all variables into two subblocks and updated all

variables in one subblock conditional on the rest. The first subblock consists of  $(\alpha, \beta, \mathbf{u}, \mathbf{v}, \kappa)$ , while the auxiliary variables  $(\mathbf{w}, \boldsymbol{\lambda})$  form the other block. Updating the first block was performed with a joint Metropolis-Hastings step as described in Subsection 2.2, where a common scaling factor  $f$  was used for all three precision parameters  $\kappa$ . Subsequently  $\alpha$ ,  $\beta$ ,  $\mathbf{u}$  and  $\mathbf{v}$  are sampled from their joint multivariate normal full conditional distribution. Finally, all parameters in this block are accepted or rejected in a joint Metropolis-Hastings step. The second block was updated with a simple Gibbs step.

We have now reanalyzed these data with the auxiliary mixture sampler. Aggregation of the binary to binomial observations was possible, with a moderate decrease in the number of observations (2578 binomial rather than 6690 binary observations). The blocking strategy was chosen just as above, replacing the auxiliary variables with the auxiliary mixture variables. The auxiliary mixture sampler was slightly faster (roughly 9%) in terms of pure computing time. Slightly lower acceptance rates have been observed for the auxiliary mixture sampler using the same scaling factor for the precision parameters as in the original algorithm. For example, for a scaling factor of 1.5, the acceptance rate was 42% rather than 53%. Slightly higher autocorrelation have been observed for some of the precision parameters, which sometimes outweighed the increase in computational speed.

### 3.3 Dealing with Multinomial Data

A similar method may be applied to data from a multinomial distribution which will be illustrated by the following model. Let  $\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_N)$  be a sequence of data arising from a multinomial distribution with  $m + 1$  categories:

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

with known repetition parameters  $N_i$ . Each observation is a discrete vector,  $\mathbf{y}_i = (y_{1i}, \dots, y_{mi})$ , where  $y_{ki}$  counts the number of times category  $k$  is observed on occasion  $i$ .

The precise model for  $\boldsymbol{\beta}_{ki}$  is left unspecified at this stage; we only assume that the joint distribution  $p(\boldsymbol{\alpha}, \boldsymbol{\beta}_{11}, \dots, \boldsymbol{\beta}_{mN} | \boldsymbol{\theta})$  is a normal distribu-

tion, indexed by some unknown parameter  $\boldsymbol{\theta}$ . Furthermore we assume that conditional on knowing  $\boldsymbol{\pi}_i$  and  $\boldsymbol{\pi}_j$ ,  $\mathbf{y}_i$  and  $\mathbf{y}_j$  are mutually independent.

### 3.3.1 Data augmentation

First, the random variable  $y_{ki}$  is regarded for each  $i$  as the number of times category  $k$  is observed when drawing  $N_i$  independent categorical random variables  $z_{ni}$  from the probability distribution

$$\boldsymbol{\pi}_i = (\pi_{0i}, \pi_{1i}, \dots, \pi_{mi}), \quad \text{pr}\{z_{ni} = k | \boldsymbol{\pi}_i\} = \pi_{ki}.$$

We recover  $z_{ni}$  for  $n = 1, \dots, N_i$  as

$$z_{ni} = \begin{cases} k, & \sum_{l=1}^{k-1} y_{li} < n \leq \sum_{l=1}^k y_{li}, \\ 0, & \sum_{l=1}^m y_{li} < n \leq N_i. \end{cases}$$

The first step of data augmentation in Frühwirth-Schnatter and Frühwirth (2007) introduces for each categorical observation  $z_{ni}$  the utilities  $y_{1ni}^u, \dots, y_{mni}^u$  of choosing categories 1 to  $m$  as latent variables. This leads to a total of  $2m(\sum_{i=1}^N N_i)$  latent variables.

A more efficient method may be derived by extending the improved auxiliary mixture sampler introduced in Subsection 3.1 for data from the binomial distribution in the following way. For any utility  $y_{kni}^u, k = 1, \dots, m$  the following holds for  $n = 1, \dots, N_i$ :

$$\exp(-y_{kni}^u) = \frac{1}{\lambda_{ki}} \exp(-\varepsilon_{kni}),$$

where  $\exp(-\varepsilon_{kni}) \sim \text{Ex}(1)$ . If we sum over all  $n = 1, \dots, N_i$  as in Subsection 3.1 and define for each category the following aggregated utility  $y_{ki}^*$ :

$$y_{ki}^* = -\log \sum_{n=1}^{N_i} \exp(-y_{kni}^u), \quad (17)$$

we obtain

$$y_{ki}^* = \log \lambda_{ki} + \varepsilon_{ki}, \quad (18)$$

where  $\varepsilon_{ki} = -\log \xi_{ki}$ , with  $\xi_{ki} = \sum_{n=1}^{N_i} \exp(-\varepsilon_{kni}) \sim \text{Ga}(N_i, 1)$ .

The first step of improved auxiliary mixture sampling introduces for each observation  $\mathbf{y}_i$  the  $m$  aggregated utilities  $\mathbf{y}_i^* = (y_{1i}^*, \dots, y_{mi}^*)$  as latent variables, rather than the entire sequence of  $mN_i$  individual utilities

$y_{11i}^u, \dots, y_{mN_i}^u$ . The second step is exactly the same as for data from the Poisson and the binomial distribution. The densities of  $\varepsilon_{ki}$  in (18) are approximated by Gaussian mixtures, and the indicators  $r_{ki}$  are introduced as additional latent variables. This leads to a total of  $2mN$  rather than  $2m(\sum_{i=1}^N N_i)$  latent variables.

Conditional on  $\mathbf{y}^* = \{\mathbf{y}_1^*, \dots, \mathbf{y}_N^*\}$  and  $\mathbf{S} = \{\mathbf{r}_1, \dots, \mathbf{r}_N\}$ , where  $\mathbf{r}_i = (r_{1i}, \dots, r_{mi})$ , the nonlinear non-Gaussian model (16) reduces to a linear Gaussian model where the mean of the observation equation is linear in  $\boldsymbol{\alpha}$  and  $\boldsymbol{\beta}_{11}, \dots, \boldsymbol{\beta}_{mN}$  and the error term follows a normal distribution:

$$\begin{aligned} y_{1i}^* &= \log \lambda_{1i} + m_{r_{1i}}(N_i) + \varepsilon_{1i}, & \varepsilon_{1i} | r_{1i} &\sim N(0, s_{r_{1i}}^2(N_i)), \\ &\vdots \\ y_{mi}^* &= \log \lambda_{mi} + m_{r_{mi}}(N_i) + \varepsilon_{mi}, & \varepsilon_{mi} | r_{mi} &\sim N(0, s_{r_{mi}}^2(N_i)), \end{aligned}$$

with  $\log \lambda_{ki} = (\mathbf{Z}_i^\alpha)^T \boldsymbol{\alpha}_k + (\mathbf{Z}_i^\beta)^T \boldsymbol{\beta}_{ki}$ . Consequently, the conditional posterior  $p(\boldsymbol{\alpha}, \boldsymbol{\beta}_{11}, \dots, \boldsymbol{\beta}_{mN} | \boldsymbol{\theta}, \mathbf{y}^*, \mathbf{S}, \mathbf{y})$  is multivariate normal.

If  $N_i \equiv 1$ , model (16) reduces to a multinomial logit model, and the improved method introduced in this subsection reduces to the one described for categorical data in Frühwirth-Schnatter and Frühwirth (2007).

### 3.3.2 The sampling scheme

Select starting values for  $\mathbf{y}^*$ ,  $\mathbf{S}$  and  $\boldsymbol{\theta}$ , and repeat the following steps.

- (1) Sample  $\boldsymbol{\alpha}$ ,  $\boldsymbol{\beta} = \{\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_N\}$ , and  $\boldsymbol{\theta}$ , conditional on  $\mathbf{y}^*$ ,  $\mathbf{S}$ , and  $\mathbf{y}$ .
- (2) Sample the aggregated utilities  $\mathbf{y}^*$  and the indicators  $\mathbf{S}$  conditional on  $\boldsymbol{\alpha}$ ,  $\boldsymbol{\beta}$ ,  $\boldsymbol{\theta}$  and  $\mathbf{y}$ , by running the following steps, for  $i = 1, \dots, N$ .
  - (a) Sample the aggregated utility  $\mathbf{y}_i^* = (y_{1i}^*, \dots, y_{mi}^*)$  as:

$$y_{ki}^* = -\log \left( \frac{U_i}{1 + \sum_{l=1}^m \lambda_{li}} + \frac{V_{ki}}{\lambda_{ki}} \right), \quad (19)$$

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

- (b) Sample the component indicators  $r_{ki}$  from the following discrete distribution where  $j = 1, \dots, R(N_i)$ :

$$\begin{aligned} \text{pr}\{r_{ki} = j | y_{ki}^*, \boldsymbol{\theta}, \boldsymbol{\beta}_{ki}, \boldsymbol{\alpha}, \mathbf{y}\} &\propto \\ \frac{w_j(N_i)}{s_j(N_i)} \exp \left\{ -\frac{1}{2} \left( \frac{y_{ki}^* - \log \lambda_{ki} - m_j(N_i)}{s_j(N_i)} \right)^2 \right\}. & \quad (20) \end{aligned}$$

Again step 1 is model dependent, but standard for many models, as we are dealing with a Gaussian model once we condition on  $\mathbf{y}^*$  and  $\mathbf{S}$ . Step 2 is a modification of the corresponding step in Frühwirth-Schnatter and Frühwirth (2007, Subsection 3.2). To justify sampling of  $y_{ki}^*$  as in (19) we use (17) and sample the individual utilities  $\mathbf{y}_{ni}^u = (y_{1ni}^u, \dots, y_{mni}^u)$  from the posterior distribution  $p(\mathbf{y}_{ni}^u | \mathbf{y}_i, \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta})$  as in Frühwirth-Schnatter and Frühwirth (2007):

$$y_{kni}^u = -\log \left( -\frac{\log U_{ni}}{1 + \sum_{l=1}^m \lambda_{li}} - \frac{\log V_{kni}}{\lambda_{ki}} I_{\{z_{ni} \neq k\}} \right), \quad (21)$$

using independent uniform random numbers  $U_{ni}, V_{1ni}, \dots, V_{mni}$ . This yields

$$\begin{aligned} y_{ki}^* &= -\log \sum_{n=1}^{N_i} \exp(-y_{kni}^u) \\ &= -\log \left( \frac{\sum_{n=1}^{N_i} (-\log U_{ni})}{1 + \sum_{l=1}^m \lambda_{li}} + \frac{\sum_{n: z_{ni} \neq k} (-\log V_{kni})}{\lambda_{ki}} \right). \end{aligned}$$

Step 2(a) is justified by the facts that

$$\begin{aligned} \sum_{n=1}^{N_i} (-\log U_{ni}) &\sim \text{Ga}(N_i, 1), \\ y_{ki} < N_i &\implies \sum_{n: z_{ni} \neq k} (-\log V_{kni}) \sim \text{Ga}(N_i - y_{ki}, 1). \end{aligned}$$

Evidently,  $r_{ki}$  has to be sampled independently from the discrete density  $p(r_{ki} | \mathbf{y}_i^*, \boldsymbol{\theta}, \boldsymbol{\beta}_i, \boldsymbol{\alpha}, \mathbf{y})$  given in (20). Starting values for the component indicator  $r_{ki}$  are drawn uniformly from 1 to  $R(N_i)$ ; to obtain a starting value for  $y_{ki}^*$ , we use (19) with  $\lambda_{ki} = 1$ .

## 4 Concluding Remarks

In this paper we have developed auxiliary mixture sampling algorithms for hierarchical models of Poisson, 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 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 (2006) and Frühwirth-Schnatter and Frühwirth (2007).

In our two case studies, auxiliary mixture sampling allowed us to approach fairly large models using joint updates of the hyperparameters and

the latent Gaussian field. In the first study, we found that auxiliary mixture sampling is comparable if not better than a Gaussian approximation to the non-normal likelihood. In the second study we found similar efficiency in terms of the effective sample size as the Holmes and Held (2006) algorithm for binary logistic regression. Presumably the advantage of the auxiliary mixture sampler over the Holmes and Held (2006) algorithm would become more apparent in an example where stronger aggregation to binomial counts is possible.

The main motivation for the development of auxiliary mixture sampling has not been to yield a uniformly better algorithm, but to simplify the implementation and to improve the computational performance of MCMC algorithms for non-Gaussian hierarchical models. In particular, auxiliary mixture sampling allows 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.

Furthermore, auxiliary mixture sampling allows simple implementation of Bayesian model selection for a broad class of non-Gaussian models like random-effect models and state space models. Frühwirth-Schnatter and Wagner (2007) investigate the computation of marginal likelihoods using auxiliary mixture sampling, while Tüchler (2006) suggests a simple stochastic variable selection scheme based on auxiliary mixture sampling for covariate and covariance selection in logistic regression and random-effects models.

## References

- Bernardo, J. M., and Smith, A. F. M. (1994), *Bayesian Theory*, Chichester: Wiley.
- 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.
- Chib, S., Nardari, F., and Shephard N. (2002), “Markov chain Monte Carlo methods for stochastic volatility models,” *Journal of Econometrics*, 108, 281–316.
- Frühwirth-Schnatter, S., and Frühwirth R. (2007), “Auxiliary mixture sam-

- pling with applications to logistic models,” *Computational Statistics and Data Analysis*, 51, 3509–3528.
- Frühwirth-Schnatter, S., and Wagner H. (2005), “Data augmentation and Gibbs sampling for regression models of small counts,” *Student*, 5, 207–220.
- Frühwirth-Schnatter, S., and Wagner H. (2006), “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 (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.
- 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.
- Holmes, C. C., and Held, L. (2006), “Bayesian auxiliary variable models for binary and multinomial regression,” *Bayesian Analysis*, 1, 145–168.
- 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.
- Kim, S., Shephard, N., and Chib, S. (1998), “Stochastic volatility: Likelihood inference and comparison with ARCH models,” *Review of Economic Studies*, 65, 361–393.
- Knorr-Held, L., and Raßer, G. (2000), “Bayesian detection of clusters and discontinuities in disease maps,” *Biometrics*, 56, 13–21.
- Knorr-Held, L., Raßer, G., and Becker, N. (2002), “Disease mapping of stage-specific cancer incidence data,” *Biometrics*, 58, 492–501.

- Knorr-Held, L., and Rue, H. (2002), “On block updating in Markov random field models for disease mapping,” *Scandinavian Journal of Statistics*, 29, 597–614.
- Nelder, J. A., and Mead, R. (1965), “A Simplex Method for Function Minimization,” *Computer Journal*, 7, 308–313.
- Omori, Y., Chib, S., Shephard, N., and Nakajima, J. (2004), “Stochastic volatility with leverage: Fast and efficient likelihood inference,” *Journal of Econometrics*, in press.
- Robert, C. P., and Casella, G. (1999), *Monte Carlo Statistical Methods*, Springer Series in Statistics, New York/Berlin/Heidelberg: Springer.
- Rue, H., and Held, L. (2005), *Gaussian Markov Random Fields: Theory and Applications*, Volume 104 of *Monographs on Statistics and Applied Probability*, Boca Raton, FL: Chapman & Hall/CRC.
- Shephard, N. (1994), “Partial non-Gaussian state space,” *Biometrika*, 81, 115–131.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002), “Bayesian measures of model complexity and fit,” *Journal of the Royal Statistical Society, Ser. B*, 64, 583–639.
- Tüchler, R. (2006), “Bayesian variable selection for logistic models using auxiliary mixture sampling,” Research Report Series Report 31, Department of Statistics and Mathematics, Wirtschaftsuniversität Wien, Vienna (Austria).

## Tables

Table 1: Expected acceptance rate (%), for a Metropolis-Hastings algorithm based on the old (left) and the new (right) method of auxiliary mixture sampling for  $N$  observations from the  $Po(\lambda)$  distribution

$N$	$\lambda=1$	$\lambda=3$	$\lambda=10$	$N$	$\lambda=1$	$\lambda=3$	$\lambda=10$
1	99.71	99.84	99.71	1	99.88	99.93	99.93
10	99.5	99.59	99.18	10	99.74	99.71	99.54
100	99.34	99.14	99.44	100	99.53	99.36	99.51
1000	99.48	99.33	99.15	1000	99.56	99.41	99.42

Table 2: Empirical comparison of the GMRF approximation and auxiliary mixture sampling (AMS) for the Sardinia data

Scaling factor	Method	Speed (it/sec)	Acc. rate	Parameter	ESS	ESS per sec
2.0	GMRF	42.3	61.1	$\lambda$	388.2	1.6
				$\kappa$	166.0	0.7
				$D$	459.3	1.9
	AMS	159.3	50.1	$\lambda$	200.1	3.2
				$\kappa$	164.5	2.6
				$D$	201.7	3.2
3.0	GMRF	43.0	46.4	$\lambda$	670.9	2.9
				$\kappa$	361.1	1.6
				$D$	709.8	3.0
	AMS	159.4	31.1	$\lambda$	334.5	5.3
				$\kappa$	150.6	2.4
				$D$	250.1	4.0
5.0	GMRF	42.7	29.8	$\lambda$	840.3	3.6
				$\kappa$	537.4	2.3
				$D$	914.4	3.9
	AMS	163.4	15.8	$\lambda$	370.8	6.1
				$\kappa$	134.5	2.2
				$D$	145.8	2.4

Table 3: Empirical comparison of the GMRF approximation and auxiliary mixture sampling (AMS) for the Germany data

Scaling factor	Method	Speed (it/sec)	Acc. rate	Parameter	ESS	ESS per sec
1.5	GMRF	27.9	33.4	$\lambda$	220.3	0.6
				$\kappa$	609.6	1.7
				$D$	1036.8	2.9
	AMS	102.5	41.9	$\lambda$	271.5	2.8
				$\kappa$	760.2	7.8
				$D$	1176.8	12.1
2.0	GMRF	27.7	19.9	$\lambda$	323.7	0.9
				$\kappa$	671.9	1.9
				$D$	837.1	2.3
	AMS	105.6	21.5	$\lambda$	415.4	4.4
				$\kappa$	529.2	5.6
				$D$	607.8	6.4
3.0	GMRF	28.1	10.3	$\lambda$	347.8	1.0
				$\kappa$	403.6	1.1
				$D$	274.8	0.8
	AMS	104.2	9.2	$\lambda$	282.2	2.9
				$\kappa$	426.0	4.4
				$D$	272.3	2.8

Table 4: 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

## Figures

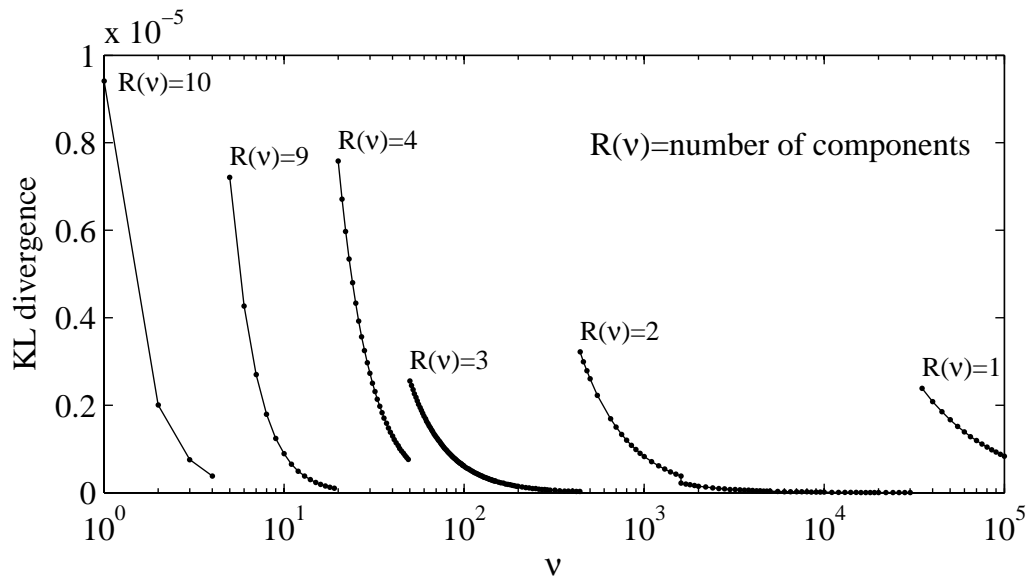


Figure 1: Kullback-Leibler divergence of the estimated mixtures from the standardized negative log-Gamma distribution as a function of the shape parameter  $\nu$ , for  $1 \leq \nu \leq 100000$ .  $R(\nu)$  is the number of components in the mixtures.

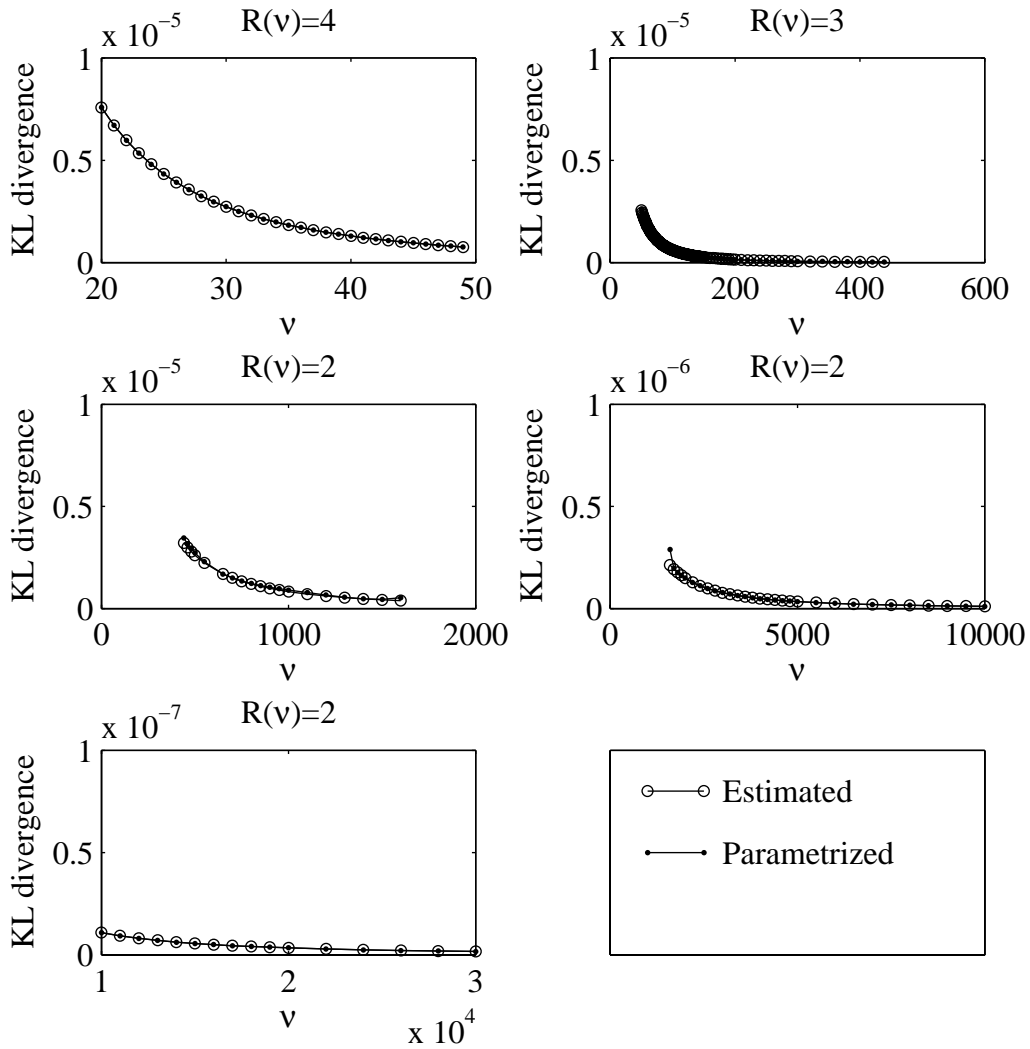


Figure 2: Kullback-Leibler divergence of the estimated and of the parametrized mixtures from the standardized negative log-Gamma distribution as a function of the shape parameter  $\nu$ , for  $20 \leq \nu \leq 100000$ .  $R(\nu)$  is the number of components in the mixtures.