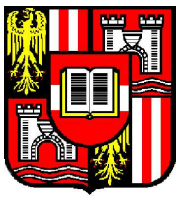

Department for Applied Statistics
Johannes Kepler University Linz

IFAS Research Paper Series
2006-14

Auxiliary Mixture Sampling with Applications to Logistic Models

Sylvia Frühwirth-Schnatter and Rudolf Frühwirth [a]

November 2006

[a]Austrian Academy of Sciences, Institute of High Energy Physics

Abstract

We describe a new method of data augmentation for binary and multinomial logit models. First, the latent utilities (McFadden, 1974) are introduced as auxiliary latent variables, leading to a latent model which is linear in the unknown parameters, but involves errors from the type I extreme value distribution. Second, for each error term the density of this distribution is approximated by a mixture of normal distributions, and the component indicators in these mixtures are introduced as further latent variables. This leads to Markov chain Monte Carlo estimation based on a convenient auxiliary mixture sampler that draws from standard distributions like normal or exponential distributions and, in contrast to more common Metropolis-Hastings approaches, does not require any tuning.

We show how the auxiliary mixture sampler is implemented for binary or multinomial logit models, and demonstrate how to extend the sampler to mixed effect models and time-varying parameter models for binary and categorical data. Finally, we discuss an application to Austrian labor market data.

Key words: binary data, categorical data, Markov chain Monte Carlo, random effect models, state space models, utilities

1 Introduction

Applied statisticians and econometricians commonly have to deal with modelling a binary or multinomial response variable in terms of covariates. Examples include modelling the probability of unemployment in terms of risk factors and modelling choice probabilities in marketing in terms of product attributes. A widely used tool for analyzing such data are binary or multinomial regression techniques using generalized linear models (McCullagh and Nelder, 1999), either based on the logit or the probit link function. In this paper we focus on computationally simple Markov chain Monte Carlo (MCMC) techniques for practical Bayesian inference of the standard binary and multinomial logit regression model and some of its extensions, like time-varying parameter models and regression models including random effects.

Seminal papers on the Bayesian estimation of logistic regression models are Zellner and Rossi (1984) who performed importance sampling based on a multivariate Student-t distribution, with mean and covariance matrix being equal to the posterior mode and the asymptotic covariance matrix, and Zeger and Karim (1991) who where the first to use Markov chain Monte Carlo (MCMC) methods. Several other authors have contributed to MCMC estimation of logistic models (Gamerman, 1997; Chib et al., 1998; Lenk and DeSarbo, 2000; Hurn et al., 2003; Scott, 2004); see also Dey et al. (2000) for a review. These techniques usually involve a Metropolis-Hastings algorithm for at least part of the unknown parameters, which in turn makes it necessary to define suitable proposal densities. For a routine application of logistic model is seems preferable to apply MCMC methods which run without any tuning. One such technique is single-move adaptive rejection sampling applied by Dellaportas and Smith (1993) which may lead to a poorly mixing sampler, since the posterior correlation in the coefficients of a logistic regression model may be extremely high; see Zellner and Rossi (1984).

Whereas it has been known for quite a while how to implement Gibbs sampling for the probit model (Albert and Chib, 1993; McCulloch and Rossi, 1994), Gibbs sampling seemed unfeasible under the logit link until very recently. Holmes and Held (2006) realized that Gibbs sampling is feasible also for logistic models through data augmentation using two sequences of auxiliary latent variables of the same size as the data. The first augmentation step is exactly the same as for the probit link and leads to a linear model with an error term that is non-normal but may be expressed as a scale mixture of normal distributions, where one-half of the square root of the scaling factor follows the Kolmogoroff-Smirnov distribution (Andrews and Mallows, 1974). Holmes and Held (2006) introduced for each error term the scaling factor as a second auxiliary latent variable, which conditionally leads to a linear normal regression model. The conditional posterior distribution of the scaling factors, however, does not have a closed form, and rejection sampling has to be used.

The main goal of the present article is to develop another Gibbs type sampling scheme for the Bayesian estimation of logistic models. It is similar to the approach of Holmes and Held (2006) but offers the advantage that the conditional posterior distribution of all auxiliary latent variables has closed form. This Gibbs type sampling scheme results from applying auxiliary mixture sampling, which has been introduced for a Bayesian analysis of stochastic volatility models by Shephard (1994) and has been applied in this context by a couple of authors (Kim et al., 1998; Chib et al., 2002; Omori et al., 2004). Frühwirth-Schnatter and Wagner (2006, 2005) introduced auxiliary mixture sampling for a Bayesian analysis of parameter-driven models for count data based on the Poisson distribution.

To extend the auxiliary mixture sampling approach to logistic models we first recall the interpretation of a logit-model in terms of utilities (McFadden, 1974) and introduce as in Scott (2004) the latent utilities as missing variables in a first data augmentation step. The introduction of this first set of latent variables eliminates non-linearity from the regression analysis and conditionally leads to a linear regression model. The non-normality of the error term which follows a type I extreme value distribution, however, remains. Whereas Scott (2004) uses a Metropolis-Hastings algorithm to sample the parameters, we eliminate the non-normality of the error term by a second sequence of latent variables. To this aim, the extreme value distribution is approximated by a mixture of normal distributions in a similar way as in Kim et al. (1998) and Chib et al. (2002) who used a normal mixture approximation to the density of a $\log \chi^2_1$-distribution in the context of stochastic volatility models. By introducing the component indicator of this normal mixture as a second sequence of missing data, a Gibbs sampling type algorithm is developed. This will be shown to be particularly useful for random effects models and for state space models for binary and categorical time series, as multi-move-sampling of all effects becomes feasible.

The rest of the paper is organized as follows. In Section 2, we discuss data augmentation and auxiliary mixture sampling for a binary logit regression model, which will be extended to multinomial logit models in Section 3. More complex models like state space modelling of binary time series and random effect models for binary and multinomial panel data are discussed in Section 4, where we also consider data from the binomial distribution. An application to capturing unobserved heterogeneity in the Austrian labor market is described in Subsection 4.4. Finally, Section 5 contains

the concluding remarks.

2 Data Augmentation and Auxiliary Mixture Sampling for the Binary Logit Regression Model

Given a sequence $y_1, \ldots, y_N$ of binary data, the binary logit regression model reads:

$$\Pr(y_i = 1|\boldsymbol{\beta}) = \frac{\exp(\mathbf{x}_i\boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i\boldsymbol{\beta})}, \tag{1}$$

where $\mathbf{x}_i$ is a row vector of regressors, including 1 for the intercept, and $\boldsymbol{\beta}$ is an unknown regression parameter. Furthermore we assume that conditional on knowing $\boldsymbol{\beta}$, the observations are mutually independent.

To pursue a Bayesian approach, we assume that the prior distribution $p(\boldsymbol{\beta})$ of $\boldsymbol{\beta}$ follows a normal distribution, $\mathcal{N}_d(\mathbf{b}_0, \mathbf{B}_0)$ with known hyperparameters $\mathbf{b}_0$ and $\mathbf{B}_0$. It is then possible to derive the posterior density $p(\boldsymbol{\beta}|\mathbf{y})$ by Bayes' theorem, given all observations $\mathbf{y} = (y_1, \ldots, y_N)$:

$$p(\boldsymbol{\beta}|\mathbf{y}) \propto p(\boldsymbol{\beta}) \prod_{i=1}^{N} \frac{(\exp(\mathbf{x}_i\boldsymbol{\beta}))^{y_i}}{1 + \exp(\mathbf{x}_i\boldsymbol{\beta})}.$$

Zellner and Rossi (1984) showed that improper priors like $p(\boldsymbol{\beta}) \propto$ constant do not necessarily lead to a proper posterior density $p(\boldsymbol{\beta}|\mathbf{y})$. Sufficient conditions on the likelihood involve conditions on the number of observed zeros and ones (both need to be positive), as well as conditions on the regressor; see Zellner and Rossi (1984, p.389) for details.

2.1 Data Augmentation for the Binary Logit Regression Model

The first data augmentation step was suggested by Scott (2004) in the context of multinomial logit models and involves the well-known interpretation of a logit-model in terms of utilities as introduced by McFadden (1974). Let y_{0i}^u be the utility of choosing category 0, which is assumed to be independent of any covariates for identifiability reasons. Let y_i^u be the utility of choosing category 1, which is modelled as depending on covariates $\mathbf{x}_i$:

$$y_i^u = \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i. \tag{2}$$

Then category 1 is observed, i.e. $y_i = 1$, iff $y_i^u > y_{0i}^u$, otherwise $y_i = 0$. If y_{0i}^u and ε_i follow a type I extreme value distribution, the binary logit regression model (1) results as the marginal distribution of y_i.

The first step of data augmentation introduces for each i, $i = 1, \ldots, N$, the latent utility y_i^u of choosing category 1 as missing data, with two desirable effects. First, the conditional posterior distribution $p(\boldsymbol{\beta}|\mathbf{y}^u, \mathbf{y})$ of $\boldsymbol{\beta}$, where in addition to $\mathbf{y}$ the latent utilities $\mathbf{y}^u = (y_1^u, \ldots, y_N^u)$ appear as a conditioning argument, is independent of $\mathbf{y}$: $p(\boldsymbol{\beta}|\mathbf{y}^u, \mathbf{y}) = p(\boldsymbol{\beta}|\mathbf{y}^u)$. Second, conditional on $\mathbf{y}^u$, the posterior of $\boldsymbol{\beta}$ can be

derived from regression model (2), which is non-normal, but linear in the unknown model parameters $\boldsymbol{\beta}$. Thus, the first augmentation step eliminates the non-linearity of the logit model; the non-normality of the error term ε_i, however, remains. Scott (2004) uses a Metropolis-Hastings algorithm based on various approximations to this regression model, to sample the regression parameters $\boldsymbol{\beta}$.

In the present paper we go a step further and eliminate also the non-normality of the error term through a second step of data augmentation. Note that the error term ε_i in (2) follows a type I extreme value distribution and that the density of this distribution is independent of any unknown model parameters:

$$p_\varepsilon(\varepsilon) = \exp\{-\varepsilon - e^{-\varepsilon}\}. \tag{3}$$

To obtain a model that is conditionally Gaussian, we approximate the non-normal density $p_\varepsilon(\varepsilon)$ by a normal mixture of M components with parameters m_r and s_r^2 for the r-th component:

$$p_\varepsilon(\varepsilon) = \exp\{-\varepsilon - e^{-\varepsilon}\} \approx q_{M,\varepsilon}(\varepsilon), \qquad q_{M,\varepsilon}(\varepsilon) = \sum_{r=1}^{M} w_r f_{\mathcal{N}}(\varepsilon; m_r, s_r^2). \tag{4}$$

This idea is influenced by the related articles of Shephard (1994), Kim et al. (1998), Chib et al. (2002) and Omori et al. (2004) who used a normal mixture approximation of the density of a $\log \chi_1^2$-distribution in the context of stochastic volatility models. The appropriate parameters $(w_r, m_r, s_r^2), r = 1, \ldots, M$, however, are different for our problem and are given in Table 1 for $M = 10$. The choice of these parameters will be discussed in Subsection 2.3.

The generation of ε from the mixture distribution (4) may be viewed as first drawing one of the M normal distributions, by drawing the components indicator r from the discrete probability distribution $w_1, \ldots, w_M$, and then drawing ε from the normal distribution $\mathcal{N}(m_r, s_r^2)$; see Frühwirth-Schnatter (2006) for more detail.

The second step of our data augmentation scheme approximates the density $p_\varepsilon(\varepsilon_i)$ in the regression model (2) by the normal mixture $q_{M,\varepsilon}(\varepsilon_i)$ and introduces for each ε_i the latent component indicator r_i as missing data. Conditionally on knowing r_i, the regression model (2) reduces to a Gaussian regression model with heteroscedastic errors with known variance:

$$y_i^u = \mathbf{x}_i\boldsymbol{\beta} + m_{r_i} + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}\left(0, s_{r_i}^2\right). \tag{5}$$

For such a model it is well known that the conditional posterior of $\boldsymbol{\beta}$ is a multivariate normal density (Zellner, 1971). This result is the basis for the two-block auxiliary mixture sampler that will be described in Subsection 2.2.

2.2 A Two-Block Auxiliary Mixture Sampler

Assume that data augmentation as described in Subsection 2.1 has been applied for all observations by introducing two sequences of latent auxiliary variables, namely the latent utilities $\mathbf{y}^u = \{y_1^u, \ldots, y_N^u\}$ and the latent component indicators $\mathbf{R} = \{r_1, \ldots, r_N\}$.

To implement a two-block auxiliary mixture sampler select starting values for $\mathbf{R}$ and $\mathbf{y}^u$ and repeat the following steps:

Table 1: Normal mixture approximation to the density of the type I extreme value distribution (10 components) obtained by minimizing the Kullback-Leibler distance

r	1	2	3	4	5	6	7	8	9	10
w_r	0.00397	0.0396	0.168	0.147	0.125	0.101	0.104	0.116	0.107	0.088
m_r	5.09	3.29	1.82	1.24	0.764	0.391	0.0431	-0.306	-0.673	-1.06
s_r^2	4.5	2.02	1.1	0.422	0.198	0.107	0.0778	0.0766	0.0947	0.146

(a) Sample the regression coefficient $\boldsymbol{\beta}$ conditional on knowing $\mathbf{y}^u$ and $\mathbf{R}$ based on the normal regression model (5).

(b) Sample the latent utilities $\mathbf{y}^u$ and the latent indicators $\mathbf{R}$ conditional on $\boldsymbol{\beta}$ and $\mathbf{y}$ by running steps (b1) and (b2) independently for $i = 1, \ldots, N$ with $\lambda_i = \exp(\mathbf{x}_i\boldsymbol{\beta})$:

(b1) Sample the latent utility y_i^u conditional on λ_i and y_i as

$$y_i^u = -\log\left(-\frac{\log(U_i)}{1+\lambda_i} - \frac{\log(V_i)}{\lambda_i} I_{\{y_i=0\}}\right), \tag{6}$$

where U_i and V_i are two independent uniform random numbers.

(b2) Sample the component indicators r_i conditional on y_i^u and λ_i from the following discrete density:

$$\Pr(r_i = j | y_i^u, \boldsymbol{\beta}) \propto \frac{w_j}{s_j} \exp\left\{-\frac{1}{2}\left(\frac{y_i^u - \log\lambda_i - m_j}{s_j}\right)^2\right\}. \tag{7}$$

The quantities $(w_j, m_j, s_j^2), j = 1, \ldots, M$ are the parameters of the M component finite mixture approximation tabulated in Table 1.

Note that step (b) involves only draws from standard densities. Step (b) can be used to sample starting values for y_i^u and r_i for each i, given the observed binary data y_i, by choosing starting values for $\lambda_i = \exp(\mathbf{x}_i\boldsymbol{\beta})$.

Details on the Sampling Steps

Conditionally on knowing $\mathbf{y}^u$ and $\mathbf{R}$, the binary logit model (1) reduces to the linear normal regression model (5). Therefore, in step (a), the conditional posterior of $\boldsymbol{\beta}$ is given by the $\mathcal{N}_d(\mathbf{b}_N, \mathbf{B}_N)$-distribution, where

$$\mathbf{b}_N = \mathbf{B}_N\left(\sum_{i=1}^N \mathbf{x}_i'(y_i^u - m_{r_i})/s_{r_i}^2 + \mathbf{B}_0^{-1}\mathbf{b}_0\right), \tag{8}$$

$$\mathbf{B}_N^{-1} = \mathbf{B}_0^{-1} + \sum_{i=1}^N \mathbf{x}_i'\mathbf{x}_i/s_{r_i}^2.$$

To verify the sampling steps (b1) and (b2), the posterior $p(\mathbf{R}, \mathbf{y}^u|\mathbf{y}, \boldsymbol{\beta})$ is decomposed as:

$$p(\mathbf{R}, \mathbf{y}^u|\mathbf{y}, \boldsymbol{\beta}) = p(\mathbf{R}|\mathbf{y}^u, \mathbf{y}, \boldsymbol{\beta})p(\mathbf{y}^u|\mathbf{y}, \boldsymbol{\beta}).$$

The latent utilities y_i^u are independent, given $\mathbf{y}$ and $\boldsymbol{\beta}$:

$$p(\mathbf{y}^u|\mathbf{y}, \boldsymbol{\beta}) = \prod_{i=1}^{N} p(y_i^u|y_i, \boldsymbol{\beta}).$$

To sample y_i^u from the conditional distribution $p(y_i^u|y_i, \boldsymbol{\beta})$, we use some well-known properties of the exponential distribution. First, from the relation between the type I extreme value distribution and the exponential distribution, we obtain

$$\exp(-y_{0i}^u) \sim \mathcal{E}(1), \quad \exp(-y_i^u) \sim \mathcal{E}(\lambda_i), \tag{9}$$

where y_{0i}^u is the utility of choosing category 0 and $\lambda_i = \exp(\mathbf{x}_i\boldsymbol{\beta})$. Second, as the minimum of exponential random variables follows again an exponential distribution, we obtain:

$$\min(\exp(-y_{0i}^u), \exp(-y_i^u)) \sim \mathcal{E}(1 + \lambda_i). \tag{10}$$

Third, knowing the minimum, the other random variable has a translated exponential distribution. If $y_i = 1$, then $y_i^u > y_{0i}^u$, or equivalently, $\exp(-y_i^u) < \exp(-y_{0i}^u)$. Therefore we obtain from (10):

$$\exp(-y_i^u) \sim \mathcal{E}(1 + \lambda_i). \tag{11}$$

On the other hand, if $y_i = 0$, then $y_i^u < y_{0i}^u$, or equivalently, $\exp(-y_{0i}^u) < \exp(-y_i^u)$, and:

$$\exp(-y_{0i}^u) \sim \mathcal{E}(1 + \lambda_i), \quad \exp(-y_i^u) = \exp(-y_{0i}^u) + \xi_i, \quad \xi_i \sim \mathcal{E}(\lambda_i). \tag{12}$$

By the help of two uniform random numbers U_i and V_i, (11) and (12) can be written immediately as in formula (6) in step (b1).

The component indicators r_i are mutually independent, given $\mathbf{y}^u$, $\boldsymbol{\beta}$ and $\mathbf{y}$:

$$p(\mathbf{R}|\mathbf{y}^u, \mathbf{y}, \boldsymbol{\beta}) = \prod_{i=1}^{N} p(r_i|y_i^u, \boldsymbol{\beta}).$$

The posterior of each component indicator r_i depends on the data only through y_i^u, thus step (b2) follows immediately.

2.3 Finding the Mixture Approximation

Auxiliary mixture sampling is based on approximating the type I extreme value distribution in (4) by a normal mixture distribution. To find an appropriate mixture distribution we have considered a whole set of normal mixture approximations with different number M of components, which were fitted to the type I extreme value distribution using different distance measures. The use of distance-based methods

to fit mixtures of normals to a given density is discussed in great detail in Titterington et al. (1985, Section 4.5). Among the distance measures suggested there, we considered the L_2 distance $\delta_{LB}(p_\varepsilon, q)$ defined by:

$$\delta_{LB}(p_\varepsilon, q_{\varepsilon,M}) = \int_{\Re} \left(p_\varepsilon(\varepsilon) - q_{\varepsilon,M}(\varepsilon)\right)^2 d\varepsilon \tag{13}$$
$$= \int_{\Re} \left(\exp\{-\varepsilon - e^{-\varepsilon}\} - \sum_{r=1}^{M} w_r f_{\mathcal{N}}(\varepsilon; m_r, s_r^2) \right)^2 d\varepsilon,$$

and the Kullback-Leibler distance $\delta_{KL}(p_\varepsilon, q_{\varepsilon,M})$ defined by:

$$\delta_{KL}(p_\varepsilon, q_{\varepsilon,M}) = \int_{\Re} p_\varepsilon(\varepsilon) \log \frac{p_\varepsilon(\varepsilon)}{q_{\varepsilon,M}(\varepsilon)} d\varepsilon \tag{14}$$
$$= \int_{\Re} \exp\{-\varepsilon - e^{-\varepsilon}\} \left\{ -\varepsilon - e^{-\varepsilon} - \log \left(\sum_{r=1}^{M} w_r f_{\mathcal{N}}(\varepsilon; m_r, s_r^2) \right) \right\} d\varepsilon.$$

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

$$w_r' = \ln(w_r) - \ln(1 - w_r), \quad s_r^{2\prime} = \ln s_r^2. \tag{15}$$

For M fixed, the unknown parameters were determined by minimizing either $\delta_{LB}(p_\varepsilon, q_{\varepsilon,M})$ or $\delta_{KL}(p_\varepsilon, q_{\varepsilon,M})$, using the function `fminsearch` in the optimization toolbox of MATLAB (Version 7.0.1). The function `fminsearch` uses a direct search method, the Nelder-Mead simplex algorithm (Nelder and Mead, 1965). It evaluates the objective function at the vertices of a simplex, then rejects the worst vertex and looks for a better one. It iteratively shrinks the simplex until the improvement falls below some bound or the maximum number of iterations is exceeded. As the Nelder-Mead method finds the nearest local optimum and convergence tends to be slow, good initial values of the mixture parameters are important. The initial values of the M-component mixture were determined by splitting the largest component of the $(M-1)$-component mixture into two new components with slightly shifted means and leaving the remaining components unchanged. Integration in (13) and (14) was carried out numerically by a simple trapezoidal rule on a grid of 45000 points in the interval $[-3, 15]$ which was considered to be the effective range of the type I extreme value distribution. Figure 1 shows two quality indicators of the approximating mixtures obtained by minimizing δ_{KL}, namely the Kullback-Leibler distance of the estimated mixture from the exact distribution and the maximum absolute deviation of the estimated mixture density from the exact density, both as a function of the number M of components.

The minimization procedure with `fminsearch` is quite lengthy. For instance, the mixture with 10 components and 30 parameters requires nearly 38,000 iterations to converge. On a mobile PC with a 1.73 GHz Pentium M processor this takes about 5000 seconds. Although it is clearly not excluded to go beyond ten components from the point of view of computing time, we have found that mixtures with 14 or more components are plagued by numerical instabilities, resulting in negative values of

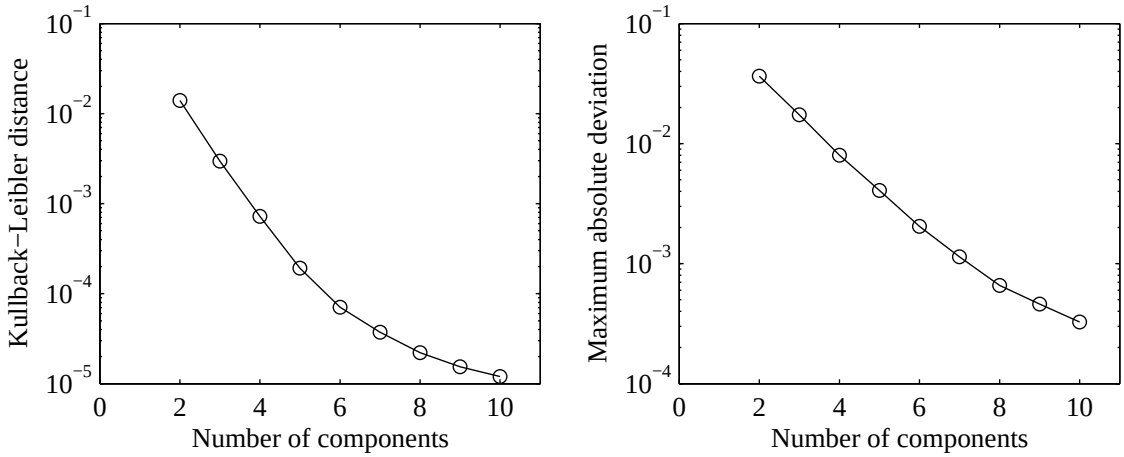

Figure 1: Left: Kullback-Leibler distance δ_{KL} of the estimated mixtures from the exact distribution; right: maximum absolute deviation of the estimated mixture densities from the exact density. The mixtures were estimated by minimizing δ_{KL}.

the Kullback-Leibler divergence. These could probably be cured by a denser grid and a more sophisticated integration rule, but the results show that this is not worth the effort.

To study the effect of using different distance measures and different numbers of mixture components, we consider a Metropolis-Hastings algorithm, based on proposing the unknown model parameter $\boldsymbol{\beta}$ from an approximate model, where in (2) the density $p_\varepsilon(\varepsilon_i)$ is substituted by the mixture approximation $q_{M,\varepsilon}(\varepsilon_i)$. The acceptance rate in the Metropolis-Hastings algorithm depends on the ratio

$$r(\boldsymbol{\beta}, \mathbf{y}^u) = \frac{p(\mathbf{y}^u|\boldsymbol{\beta})}{q_M(\mathbf{y}^u|\boldsymbol{\beta})}, \tag{16}$$

where $p(\mathbf{y}^u|\boldsymbol{\beta})$ is the likelihood of the exact model (2):

$$p(\mathbf{y}^u|\boldsymbol{\beta}) = \prod_{i=1}^{N} p_\varepsilon(y_i^u - \mathbf{x}_i\boldsymbol{\beta}), \tag{17}$$

and $q_M(\mathbf{y}^u|\boldsymbol{\beta})$ is the likelihood of the approximate model obtained by substituting $p_\varepsilon(\varepsilon_i)$ by the mixture approximation $q_{\varepsilon,M}(\varepsilon_i)$:

$$q_M(\mathbf{y}^u|\boldsymbol{\beta}) = \prod_{i=1}^{N} q_{\varepsilon,M}(y_i^u - \mathbf{x}_i\boldsymbol{\beta}). \tag{18}$$

By computing the expected acceptance rate in this Metropolis-Hastings algorithm, we will gain some insight into the accuracy of the various mixture approximations in a sampling based context. The closer the expected acceptance rate is to 100 percent, the more accurate is the corresponding mixture approximation.

We have evaluated this acceptance rate for a simple example, namely Bayesian inference for N iid binary observations $y_1, \ldots, y_N$, drawn with $\Pr(y_i = 1|\beta) = \pi = \exp(\beta)/(1 + \exp(\beta))$. Under the prior $\pi \sim \mathcal{B}(a_0, b_0)$ the posterior of π is known to

arise from the $\mathcal{B}(a_0 + S_N, b_0 + N - S_N)$-distribution with $S_N = \sum_{i=1}^{N} y_i$ counting the number of ones. The augmented model, obtained after the first step of data augmentation, reads:

$$y_i^u = \beta + \varepsilon_i, \tag{19}$$

with $\beta = \log \pi - \log(1 - \pi)$. To evaluate how the approximation error introduced through the mixture approximation influences the acceptance rate we consider a marginal two-step sampler without introducing the indicators, where in a first step we sample the utilities $y_1^u, \ldots, y_N^u$ as in step (b1) and propose β^{new} from the proposal density $q_M(\beta|\mathbf{y}^u) \propto q_M(\mathbf{y}^u|\beta)p(\beta)$, with $q_M(\mathbf{y}^u|\beta)$ being the likelihood of the approximate model defined in (18). The acceptance rate defined in (16) is random, depending both on β^{new} and on $\mathbf{y}^u$. Since in equilibrium $\mathbf{y}^u$ is drawn from the stationary distribution, we determine the expectation of the acceptance rate with respect to $q_M(\beta^{new}|\mathbf{y}^u)p(\mathbf{y}^u|\beta, \mathbf{y})p(\beta|\mathbf{y})$:

$$\int \left\{ \int \min\left(1, \frac{r(\beta^{new}, \mathbf{y}^u)}{r(\beta, \mathbf{y}^u)}\right) q_M(\beta^{new}|\mathbf{y}^u) d\beta^{new} \right\} p(\mathbf{y}^u|\beta, y)p(\beta|\mathbf{y}) d\mathbf{y}^u d\beta.$$

Since the stationary distribution $p(\beta|\mathbf{y})$ is known explicitly for this example, the outer integral is evaluted through Monte Carlo integration, whereas the inner integral is evaluated numerically.

Table 2 and Table 3 report the expected acceptance rate with simulated data for various values of π and N. The number of components rises from 2 two 10 both for the L_2 and for the Kullback-Leibler distance. As expected, by increasing the number of components the acceptance rate approaches 100% for both distances. The performance of the Kullback-Leibler distance is considerably better than the performance of the L_2 distance. The parameters of the best performing 10-component mixture approximation based on the Kullback-Leibler distance are given in Table 1.

Note that the mixture approximation is applied to equation (19) not only once, but N times. Both tables show how the approximation error accumulates when N increases. For smaller number of components the acceptance rate of the mixture approximations derived from the L_2 distance rapidly decrease with rising N. The mixture approximations derived from the Kullback-Leibler distance are much more reliable in this respect.

Rather than deriving new mixture approximations, we could have used the closely related normal mixture approximations suggested by Carter and Kohn (1997) for semiparametric Bayesian inference for time series with mixed spectra, and by Frühwirth-Schnatter and Wagner (2005) for a Bayesian analysis of regression models for small count data. In Carter and Kohn (1997), two different normal mixture approximations with 5 components were fitted to the density of a random variable defined as $\log(\frac{1}{2}X)$ with $X \sim \chi_2^2$. As the random variable $-\log(\frac{1}{2}X)$ follows a type I extreme value distribution, we could have used the mixture approximation published in Carter and Kohn (1997, Table 1 and 2) for our purpose, after switching the signs of the means given in both tables. In Frühwirth-Schnatter and Wagner (2005), a normal mixture approximation with 5 components was fitted to the density of a random variable defined as $\log X$ with $X \sim \mathcal{E}(1)$. As the random variable $-\log X$ follows a type I extreme value distribution, we could have used the mixture

Table 2: Expected acceptance rate (in percent) for a Metropolis-Hastings algorithm based on a mixture approximation with M components minimizing the Kullback-Leibler distance

π	N	2	3	4	5	6	7	8	9	10
0.05	1	90.0	96.5	98.3	99.2	99.5	99.7	99.9	99.9	99.9
	10	88.5	94.1	97.0	98.2	99.0	99.4	99.5	99.7	99.8
	100	86.3	91.3	94.4	96.9	98.4	98.9	99.1	99.3	99.4
	1000	83.2	90.1	94.5	96.8	97.8	98.7	99.2	99.3	99.4
0.20	1	90.0	96.5	98.6	99.2	99.6	99.7	99.9	99.9	99.9
	10	85.3	93.7	96.3	98.1	98.7	99.3	99.5	99.6	99.7
	100	83.4	90.4	94.5	96.4	98.0	98.5	98.8	99.0	99.2
	1000	84.6	92.4	94.6	97.2	98.2	98.5	98.9	99.0	99.2
0.50	1	90.7	96.2	98.3	99.1	99.5	99.7	99.8	99.9	99.9
	10	84.2	90.5	95.9	97.6	98.9	99.2	99.4	99.5	99.6
	100	83.1	91.0	95.3	97.7	98.3	99.1	99.3	99.4	99.5
	1000	83.2	90.5	94.4	97.1	97.8	98.3	98.6	98.9	99.0

Table 3: Expected acceptance rate (in percent) for a Metropolis-Hastings algorithm based on a mixture approximation with M components minimizing the L_2 distance

π	N	2	3	4	5	6	7	8	9	10
0.05	1	91.0	96.5	98.6	99.3	99.6	99.8	99.8	99.9	99.9
	10	86.6	94.3	97.4	98.7	99.3	99.5	99.7	99.7	99.8
	100	53.3	79.6	91.7	96.0	97.6	98.2	98.6	98.8	98.9
	1000	5.06	45.6	76.3	88.5	93.1	95.2	96.2	96.6	97.0
0.20	1	90.7	96.4	98.6	99.4	99.6	99.7	99.8	99.9	99.9
	10	81.5	93.0	96.9	98.6	99.2	99.5	99.6	99.7	99.8
	100	51.3	80.6	91.5	96.3	97.8	98.4	98.7	98.8	98.9
	1000	6.72	50.7	78.9	90.1	94.6	96.4	97.3	97.6	98.1
0.50	1	88.1	95.8	98.2	99.2	99.6	99.8	99.8	99.9	99.9
	10	81.6	90.8	95.4	97.8	98.7	99.1	99.3	99.4	99.5
	100	51.7	81.3	91.9	96.2	97.7	98.5	98.9	99.0	99.2
	1000	5.88	48.8	78.2	89.7	94.1	95.9	96.8	97.2	97.6

Table 4: Expected acceptance rate (in percent) for a Metropolis-Hastings algorithm based on previously published mixture approximations with 5 components; CK1=Carter and Kohn (1997, Table 1), CK2=Carter and Kohn (1997, Table 2), FW=Frühwirth-Schnatter and Wagner (2005, Table 1)

π	N	CK1	CK2	FW
0.05	1	96.6	94.8	99.1
	10	94.2	93.3	97.9
	100	89.8	89.7	96.3
	1000	79.4	86.0	94.6
0.20	1	96.6	94.8	99.1
	10	92.9	91.1	97.1
	100	91.1	89.5	96.0
	1000	88.3	85.3	95.7
0.50	1	96.3	95.2	99.2
	10	92.3	90.7	97.8
	100	79.0	89.4	97.1
	1000	76.3	87.1	95.0

approximation published in Frühwirth-Schnatter and Wagner (2005, Table 1) for our purpose, again after switching the signs of the means given in the table.

Since in both papers auxiliary mixture sampling was implemented without exploring the effect of using the approximate normal mixture distribution rather than the exact distribution, we included these mixture approximations into the simulation experiment described above. Table 4 reveals that the approximations derived by Carter and Kohn (1997) are rather imprecise, whereas the approximation given in Frühwirth-Schnatter and Wagner (2005) performs considerably better. Nevertheless, the 10-component mixture approximation derived in this paper improves the accuracy of auxiliary mixture sampling even further.

3 Data Augmentation and Auxiliary Mixture Sampling for the Multinomial Logit Regression Model

Let $\{y_i\}$ be a sequence of categorical data, $i = 1, \ldots, N$, where each y_i is assumed to take a value in one of $m+1$ unordered categories labelled by $\{0, \ldots, m\}$. For each category k, with $1 \leq k \leq m$, the probability that y_i takes the value k depends on covariates $\mathbf{x}_i$ in the following way:

$$\Pr(y_i = k|\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m) = \frac{\exp(\mathbf{x}_i\boldsymbol{\beta}_k)}{1 + \sum_{l=1}^{m} \exp(\mathbf{x}_i\boldsymbol{\beta}_l)}, \tag{20}$$

where $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m$ are category specific, unknown parameters. Furthermore we assume that conditional on knowing $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m$ the observations are mutually independent.

To make the model identifiable, the parameter $\boldsymbol{\beta}_0$ of the baseline category $k = 0$ is set equal to 0: $\boldsymbol{\beta}_0 = 0$. Thus the parameter $\boldsymbol{\beta}_k$ is in terms of the change in log-odds relative to the baseline category $k = 0$. To pursue a Bayesian approach, we assume that the prior distribution $p(\boldsymbol{\beta}_k)$ of $\boldsymbol{\beta}_k$ follows a normal distribution, $\mathcal{N}_d(\mathbf{b}_{k0}, \mathbf{B}_{k0})$ with known hyperparameters $\mathbf{b}_{k0}$ and $\mathbf{B}_{k0}$.

3.1 Data Augmentation for the Multinomial Logit Regression Model

As for the binary model, we consider two data augmentation steps. The first data augmentation step involves the well-known interpretation of a multinomial logit-model in terms of utilities (McFadden, 1974). The latent utility y_{ki}^u of observing the category k for observation y_i is modelled as being dependent on covariates:

$$\begin{aligned} y_{1i}^u &= \mathbf{x}_i \boldsymbol{\beta}_1 + \varepsilon_{1i}, \\ &\dots \\ y_{mi}^u &= \mathbf{x}_i \boldsymbol{\beta}_m + \varepsilon_{mi}, \end{aligned} \tag{21}$$

whereas the latent utility y_{0i}^u of observing the category 0 for observation y_i is independent of any covariates for reasons of identifiability. The observed category is equal to the category with maximal utility:

$$y_i = k \Leftrightarrow y_{ki}^u = \max_l y_{li}^u.$$

If $\varepsilon_{ki}, k = 1, \dots, m$, and y_{0i}^u follow a type I extreme value distribution, the multinomial logit model (20) results as the marginal distribution of y_i (McFadden, 1974).

The first data augmentation step introduces for each categorical observation y_i the latent utilities $\mathbf{y}_i^u = (y_{1i}^u, \dots, y_{mi}^u)$ as missing data as in Scott (2004). Conditional on $\mathbf{y}_i^u$, we are dealing with the linear model (21), rather than with the non-linear model (20). Scott (2004) uses this result to define multivariate proposal densities within a Metropolis-Hastings algorithm. In this paper, we obtain a model that is conditionally Gaussian by approximating the non-normal density of ε_{ki}, for each category $k = 1, \dots, m$, by a normal mixture as above. The second step of our data augmentation scheme introduces for each ε_{ki} the latent component indicator r_{ki} as missing data.

3.2 Auxiliary Mixture Sampling

Let $\mathbf{y}^u = \{y_{1i}^u, \dots, y_{mi}^u, i = 1, \dots, N, \}$ denote the collection of all latent utilities, and let $\mathbf{R} = \{r_{1i}, \dots, r_{mi}, i = 1, \dots, N\}$ denote the collection of all latent component indicators. Then conditional on $\mathbf{y}^u$ and $\mathbf{R}$ we are dealing for each $k = 1, \dots, m$ with the following linear regression model:

$$y_{ki}^u = \mathbf{x}_i \boldsymbol{\beta}_k + m_{r_{ki}} + s_{r_{ki}} \varepsilon_{ki}, \quad \varepsilon_{ki} \sim \mathcal{N}(0, 1). \tag{22}$$

Again it is easy to implement a two-block sampler that consists of the following steps:

(a) Independent sampling of $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m$ conditional on knowing $\mathbf{y}^u$ and $\mathbf{R}$, based on the Gaussian regression model (22).

(b) Sampling of the utilities $\mathbf{y}^u$ and the indicators $\mathbf{R}$ conditional on knowing $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m$ and $\mathbf{y}$.

Step (a) is carried out in the same manner as in Subsection 2.2. Step (b) extends the results of Subsection 2.2 to more than two categories. The joint posterior density $p(\mathbf{R}, \mathbf{y}^u | \mathbf{y}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m)$ is decomposed as:

$$p(\mathbf{R}, \mathbf{y}^u | \mathbf{y}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m) = \prod_{i=1}^{N} \prod_{k=1}^{m} p(r_{ki} | y_{ki}^u, \boldsymbol{\beta}_k) p(y_{1i}^u, \ldots, y_{mi}^u | y_i, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m).$$

To sample from $p(y_{1i}^u, \ldots, y_{mi}^u | y_i, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m)$, we consider first the augmented posterior $p(y_{0i}^u, \ldots, y_{mi}^u | y_i, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m)$ where the utility y_{0i}^u of choosing category 0 has been added. For fixed i, the latent utilities $y_{0i}^u, \ldots, y_{mi}^u$, are stochastically dependent, and the joint distribution factorizes as (Scott, 2004):

$$\begin{aligned} &p(y_{0i}^u, \ldots, y_{mi}^u | y_i = k, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m) \\ &= p(y_{ki}^u | y_i = k, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m) \prod_{l=0,\ldots,m, l \neq k} p(y_{li}^u | y_i = k, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m). \end{aligned}$$

As $\varepsilon_{ki}, k = 1, \ldots, m$, and y_{0i}^u follow a Type I extreme value distribution, we obtain:

$$\begin{aligned} &\exp(-y_{0i}^u) \sim \mathcal{E}(1), \\ &\exp(-y_{1i}^u) \sim \mathcal{E}(\lambda_{1i}), \\ &\ldots \\ &\exp(-y_{mi}^u) \sim \mathcal{E}(\lambda_{mi}), \end{aligned} \tag{23}$$

where $\lambda_{ki} = \exp(\mathbf{x}_i \boldsymbol{\beta}_k)$, for $1 \leq k \leq m$. Given $y_i = k$, y_{ki}^u is known to be the maximal utility. Thus $\exp(-y_{ki}^u)$ is the minimum among all random variables appearing in (23), and therefore:

$$\exp(-y_{ki}^u) \sim \mathcal{E}\left(1 + \sum_{l=1}^{m} \lambda_{li}\right). \tag{24}$$

Given the minimum, all other utilities are conditionally independent:

$$\exp(-y_{li}^u) = \exp(-y_{ki}^u) + \xi_{li}, \quad \xi_{li} \sim \mathcal{E}(\lambda_{li}), \tag{25}$$

where $l = 1, \ldots, m, l \neq k$. Therefore to sample $y_{1i}^u, \ldots, y_{mi}^u$ for $i = 1, \ldots, N$, we need $m + 1$ independent uniform random numbers U_i and $V_{1i}, \ldots, V_{mi}$:

$$y_{li}^u = -\log\left(-\frac{\log(U_i)}{1 + \sum_{k=1}^{m} \lambda_{ki}} - \frac{\log(V_{li})}{\lambda_{li}} I_{\{y_i \neq l\}}\right), \tag{26}$$

where $l = 1, \ldots, m$. Note that the same random number U_i is used for all categories, since this generates $\exp(-y_{ki}^u)$ by (24).

Conditional on the recent draw of y_{ki}^u, the component indicator r_{ki} is sampled from:

$$\Pr(r_{ki} = j | y_{ki}^u, \boldsymbol{\beta}_k) \propto \frac{w_j}{s_j} \exp\left\{-\frac{1}{2}\left(\frac{y_{ki}^u - \log \lambda_{ki} - m_j}{s_j}\right)^2\right\}.$$

4 Extension to More Complex Models

The basic logistic regression model has been modified in a number of ways. To account for the dependency likely to be present in sequences of binary data, past observations $y_{i-1}, y_{i-2}, \ldots$ have been introduced as covariates (Zeger and Qaqish, 1988). A couple of extensions deal with overdispersion due to omitted covariates, like mixtures of binary regressions models (Wang et al., 1996; Hurn et al., 2003), binary regression models with additive random effects (Aitkin, 1996), and mixtures of binary regression models with random effects (Lenk and DeSarbo, 2000).

To illustrate the great flexibility of auxiliary mixture sampling, we consider in detail MCMC estimation of two specific extensions of the standard logit model, namely binary state space models and multinomial logit models with random effects.

4.1 State Space Modelling of Binary Data

4.1.1 Background

Let $\{y_t\}$ be a time series of binary observations, observed for $t = 1, \ldots, T$. Each y_t is assumed to take one of two possible values, labelled by $\{0, 1\}$. The probability that y_t takes the value 1 depends on covariates $\mathbf{x}_t = (\mathbf{x}_t^f \ \mathbf{x}_t^r)$ through fixed parameters $\boldsymbol{\alpha}$ and time-varying parameters $\boldsymbol{\beta}_t^s$ in the following way:

$$\Pr(y_t = 1 | \boldsymbol{\beta}_1^s, \ldots, \boldsymbol{\beta}_T^s, \boldsymbol{\alpha}) = \frac{\exp(\mathbf{x}_t^f \boldsymbol{\alpha} + \mathbf{x}_t^r \boldsymbol{\beta}_t^s)}{1 + \exp(\mathbf{x}_t^f \boldsymbol{\alpha} + \mathbf{x}_t^r \boldsymbol{\beta}_t^s)}. \tag{27}$$

We assume that conditional on knowing $\boldsymbol{\beta}_1^s, \ldots, \boldsymbol{\beta}_T^s, \boldsymbol{\alpha}$, the observations are mutually independent. A commonly used model for describing the time-variation of $\boldsymbol{\beta}_t^s$ reads:

$$\boldsymbol{\beta}_t^s = \boldsymbol{\beta}_{t-1}^s + \mathbf{w}_t, \quad \mathbf{w}_t \sim \mathcal{N}_d(\mathbf{0}, \mathbf{Q}), \tag{28}$$

with $\boldsymbol{\beta}_0^s \sim \mathcal{N}_d(\boldsymbol{\beta}, \mathbf{B}_0)$. $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are unknown location parameters, and $\mathbf{Q}$ is an unknown covariance matrix.

Markov chain Monte Carlo estimation of logit-type state space models has been considered by various authors, in particular by Shephard and Pitt (1997) and Gamerman (1998). A characteristic feature of any existing MCMC approach for binary state space models, however, is that practical implementation requires the use of a Metropolis-Hastings algorithm for sampling the state process, which in turn makes it necessary to define a suitable proposal density in a rather high-dimensional parameter space. Single-move sampling for this type of models is known to be potentially very inefficient; see e.g. Shephard and Pitt (1997). We are going to show in the following subsection how to implement the auxiliary mixture sampler for a binary regression model with time-varying parameters, which is easily extended to more general state space models.

4.1.2 Data Augmentation and Gibbs Sampling

The data augmentation scheme introduced in Section 2 for the standard regression model can be applied to a time series without any changes. A latent utility y_t^u of

choosing category 1 is introduced for each y_t, to eliminate the non-linearity of the model:

$$y_t^u = \mathbf{x}_t^f \boldsymbol{\alpha} + \mathbf{x}_t^r \boldsymbol{\beta}_t^s + \varepsilon_t, \tag{29}$$

where ε_t follows a type I extreme value distribution. To eliminate non-normality, this distribution is approximated by a mixture of normals as in Subsection 2.1, and a latent indicator r_t is introduced for each y_t. Let $\mathbf{y}^u = \{y_1^u, \ldots, y_T^u\}$ denote the collection of all latent utilities, and let $\mathbf{R} = \{r_1, \ldots, r_T\}$ denote the collection of all latent component indicators. If we condition on the latent variables $\mathbf{y}^u$ and $\mathbf{R}$, we obtain a linear Gaussian state space model with heteroscedastic errors with known error variance:

$$\boldsymbol{\beta}_t^s = \boldsymbol{\beta}_{t-1}^s + \mathbf{w}_t, \quad \mathbf{w}_t \sim \mathcal{N}_d(\mathbf{0}, \mathbf{Q}), \tag{30}$$

$$y_t^u = \mathbf{x}_t^f \boldsymbol{\alpha} + \mathbf{x}_t^r \boldsymbol{\beta}_t^s + m_{r_t} + s_{r_t}\varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0,1), \tag{31}$$

for $t = 1, \ldots, T$. Thus it is easy to implement a three block auxiliary mixture sampler that consists of the following steps:

(a) Multi-move sampling of $\boldsymbol{\beta}_0^s, \ldots, \boldsymbol{\beta}_T^s, \boldsymbol{\beta}, \boldsymbol{\alpha}$ conditional on knowing $\mathbf{y}^u$, $\mathbf{R}$, and $\mathbf{Q}$, based on the conditional linear Gaussian state space model (30) and (31).

(b) Sampling of $\mathbf{Q}$ conditional on knowing $\boldsymbol{\beta}_0^s, \ldots, \boldsymbol{\beta}_T^s$, based on the transition equation (30) of the conditionally linear Gaussian state space model.

(c) Sampling of the utilities $\mathbf{y}^u$ and the indicators $\mathbf{R}$ conditional on knowing $\boldsymbol{\beta}_1^s, \ldots, \boldsymbol{\beta}_T^s$, $\boldsymbol{\alpha}$, and $\mathbf{y}$.

The most important aspect of our data augmentation scheme is that, conditional on y_t^u and the indicators r_t, we are dealing with a linear Gaussian state space model when sampling $\boldsymbol{\alpha}, \boldsymbol{\beta}$ and $\boldsymbol{\beta}_i^s$ in step (a) and sampling $\mathbf{Q}$ in step (b), where the binary observation y_t is substituted by the conditionally normal random variable y_t^u, and the error term follows a $\mathcal{N}\left(m_{r_t}, s_{r_t}^2\right)$-distribution. Thus for any state space model for binary data based on a logit link, step (a) and (b) in the sampling scheme introduced above are as simple as for the corresponding *linear Gaussian* state space model. In step (a), for instance, joint multi-move sampling of all location parameters $\boldsymbol{\beta}_1^s, \ldots, \boldsymbol{\beta}_T^s, \boldsymbol{\beta}, \boldsymbol{\alpha}$ is possible (Frühwirth-Schnatter, 1994; Carter and Kohn, 1994; De Jong and Shephard, 1995; Durbin and Koopman, 2002).

Step (c) is implemented by writing the posterior $p(\mathbf{R}, \mathbf{y}^u | \mathbf{y}, \boldsymbol{\beta}_1^s, \ldots, \boldsymbol{\beta}_T^s, \boldsymbol{\alpha})$ as:

$$p(\mathbf{R}, \mathbf{y}^u | \mathbf{y}, \boldsymbol{\beta}_1^s, \ldots, \boldsymbol{\beta}_T^s, \boldsymbol{\alpha}) = \prod_{t=1}^{T} p(r_t | y_t^u, \boldsymbol{\beta}_t^s, \boldsymbol{\alpha}) p(y_t^u | y_t, \boldsymbol{\beta}_t^s, \boldsymbol{\alpha}).$$

Sampling of the latent utility y_t^u and the component indicator r_t is carried out exactly as in Subsection 2.2:

$$y_t^u = -\log\left(-\frac{\log(U_t)}{1+\lambda_t} - \frac{\log(V_t)}{\lambda_t} I_{\{y_t=0\}}\right),$$

$$\Pr(r_t = j | y_t^u, \boldsymbol{\alpha}, \boldsymbol{\beta}_t^s) \propto \frac{w_j}{s_j} \exp\left\{-\frac{1}{2}\left(\frac{y_t^u - \log\lambda_t - m_j}{s_j}\right)^2\right\},$$

where U_t and V_t are two independent uniform random numbers, and $\lambda_t = \exp(\mathbf{x}_t^f \boldsymbol{\alpha} + \mathbf{x}_t^r \boldsymbol{\beta}_t^s)$.

4.2 Multinomial Logit Models with Random-Effects

4.2.1 Background

Let $\{y_{it}\}$ be repeated categorical data observed for N subjects i, $i = 1, \ldots, N$ on T_i occasions $t = 1, \ldots, T_i$. Each y_{it} is assumed to take a value in one of $m+1$ categories labelled as $\{0, \ldots, m\}$.

For category k, with $1 \leq k \leq m$, the probability that y_{it} takes the value k depends on covariates $\mathbf{x}_{it} = (\mathbf{x}_{it}^f \mathbf{x}_{it}^r)$ through fixed category specific parameters $\boldsymbol{\alpha}_k$ and subject-specific random category parameters $\boldsymbol{\beta}_{ki}^s$ in the following way:

$$\Pr(y_{it} = k|\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_m, \boldsymbol{\beta}_{1i}^s, \ldots, \boldsymbol{\beta}_{mi}^s) = \frac{\exp(\mathbf{x}_{it}^f \boldsymbol{\alpha}_k + \mathbf{x}_{it}^r \boldsymbol{\beta}_{ki}^s)}{1 + \sum_{l=1}^m \exp(\mathbf{x}_{it}^f \boldsymbol{\alpha}_l + \mathbf{x}_{it}^r \boldsymbol{\beta}_{li}^s)}. \quad (32)$$

Furthermore we assume that conditionally on knowing all $\boldsymbol{\beta}_{ki}^s$ and $\boldsymbol{\alpha}_k$ the observations are mutually independent.

To make the model identifiable, the parameters of the baseline category $k = 0$ are set equal to 0: $\boldsymbol{\alpha}_0 = 0$, $\boldsymbol{\beta}_{0i}^s = 0, i = 1, \ldots, N$. A commonly used distribution for modelling heterogeneity in $\boldsymbol{\beta}_{ki}^s$ reads:

$$\boldsymbol{\beta}_{ki}^s \sim \mathcal{N}_d(\boldsymbol{\beta}_k, \mathbf{Q}_k), \quad (33)$$

where $\boldsymbol{\beta}_k$ is an unknown location parameter, whereas $\mathbf{Q}_k$ is an unknown covariance matrix.

For $m = 1$ we are dealing with binary data, and the binary logit random effects model results:

$$\begin{aligned} &\Pr(y_{it} = 1|\boldsymbol{\alpha}, \boldsymbol{\beta}_i^s) = \frac{\exp(\mathbf{x}_{it}^f \boldsymbol{\alpha} + \mathbf{x}_{it}^r \boldsymbol{\beta}_i^s)}{1 + \exp(\mathbf{x}_{it}^f \boldsymbol{\alpha} + \mathbf{x}_{it}^r \boldsymbol{\beta}_i^s)}, \quad (34) \\ &\boldsymbol{\beta}_i^s \sim \mathcal{N}_d(\boldsymbol{\beta}, \mathbf{Q}). \end{aligned}$$

4.2.2 Data Augmentation and Gibbs Sampling

The data augmentation scheme introduced in the previous sections for standard regression models is easily extended to deal with repeated measurements. The first data augmentation step introduces for each subject i the latent utilities $y_{kit}^u, k = 1, \ldots, m$, of choosing category k at time t to eliminate the non-linearity of the model:

$$\begin{aligned} y_{1it}^u &= \mathbf{x}_{it}^f \boldsymbol{\alpha}_1 + \mathbf{x}_{it}^r \boldsymbol{\beta}_{1i}^s + \varepsilon_{1it}, \quad (35) \\ &\ldots \\ y_{mit}^u &= \mathbf{x}_{it}^f \boldsymbol{\alpha}_m + \mathbf{x}_{it}^r \boldsymbol{\beta}_{mi}^s + \varepsilon_{mit}, \end{aligned}$$

where $\varepsilon_{kit}, k = 1, \ldots, m$ follows a type I extreme value distribution. To eliminate the non-normality this distribution is approximated by a mixture of normals as in Subsection 2.1, and in the second step of our data augmentation scheme a latent indicator r_{kit} is introduced for each ε_{kit} as missing data.

Let $\mathbf{R} = \{r_{kit}, i = 1, \ldots, N, t = 1, \ldots, T_i, k = 1, \ldots, m\}$ denote the collection of all component indicators and let $\mathbf{y}^u = \{y_{1it}^u, \ldots, y_{mit}^u, i = 1, \ldots, N, t = 1, \ldots, T_i\}$

denote the collection of all latent utilities. If we condition on the latent variables $\mathbf{y}^u$ and $\mathbf{R}$, we obtain independently for each category $k, k = 1, \ldots, m$, a linear Gaussian random-effects model with heteroscedastic errors with known error variance:

$$\boldsymbol{\beta}^s_{ki} \sim \mathcal{N}_d\left(\boldsymbol{\beta}_k, \mathbf{Q}_k\right), \tag{36}$$

$$y^u_{kit} = \mathbf{x}^f_{it}\boldsymbol{\alpha}_k + \mathbf{x}^r_{it}\boldsymbol{\beta}^s_{ki} + m_{r_{kit}} + s_{r_{kit}}\varepsilon_{kit}, \quad \varepsilon_{kit} \sim \mathcal{N}\left(0, 1\right), \tag{37}$$

for $t = 1, \ldots, T_i$, $i = 1, \ldots, N$. Thus it is easy to implement a three block auxiliary mixture sampler that consists of the following steps:

(a) Multi-move sampling of $\boldsymbol{\beta}^s_{k1}, \ldots, \boldsymbol{\beta}^s_{kN}, \boldsymbol{\beta}_k, \boldsymbol{\alpha}_k$ conditional on knowing $\mathbf{y}^u$, $\mathbf{R}$, and $\mathbf{Q}$, independently for each category $k = 1, \ldots, m$, based on the conditionally linear Gaussian random-effects model (36) and (37).

(b) Sampling of $\mathbf{Q}_k$ independently for each category $k = 1, \ldots, m$ from (36), conditionally on knowing $\boldsymbol{\beta}^s_{k1}, \ldots, \boldsymbol{\beta}^s_{kN}, \boldsymbol{\beta}_k$.

(c) Sampling of the utilities $\mathbf{y}^u$ and the indicators $\mathbf{R}$ conditionally on knowing $\mathbf{y}$ and $(\boldsymbol{\beta}^s_{k1}, \ldots, \boldsymbol{\beta}^s_{kN}, \boldsymbol{\beta}_k, \boldsymbol{\alpha}_k)$, $k = 1, \ldots, m$.

An important aspect of our data augmentation scheme is that conditional on $\mathbf{y}^u$ and $\mathbf{R}$, we are dealing with a linear Gaussian random effects model when sampling $\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k$ and $\boldsymbol{\beta}^s_{ki}$ in step (a) and $\mathbf{Q}_k$ in step (b), where the categorical observation y_{it} is substituted by a conditionally normal random variable y^u_{kit}, and the error term follows a $\mathcal{N}\left(m_{r_{kit}}, s^2_{r_{kit}}\right)$-distribution. Thus for a binary or multinomial logit model with random effects, step (a) and (b) in the sampling scheme introduced above are as simple as for the corresponding *linear Gaussian* random-effects model. In step (a), joint multi-move sampling of all location parameters $\boldsymbol{\beta}^s_{k1}, \ldots, \boldsymbol{\beta}^s_{kN}, \boldsymbol{\beta}_k, \boldsymbol{\alpha}_k$ is possible by sampling $(\boldsymbol{\beta}_k, \boldsymbol{\alpha}_k)$ from the marginal model, where the random effects are integrated out (Frühwirth-Schnatter and Otter, 1999; Sahu and Roberts, 1999; Frühwirth-Schnatter et al., 2004); see also Subsection 4.2.3.

Step (c) is implemented as above by observing that:

$$p(\mathbf{R}, \mathbf{y}^u | \mathbf{y}, \boldsymbol{\beta}^s_{11}, \ldots, \boldsymbol{\beta}^s_{1N}, \ldots, \boldsymbol{\beta}^s_{m1}, \ldots, \boldsymbol{\beta}^s_{mN}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_m, \mathbf{Q}_1, \ldots, \mathbf{Q}_m) =$$
$$\prod_{i=1}^{N}\prod_{t=1}^{T} p(y^u_{1it}, \ldots, y^u_{mit} | y_{it}, \boldsymbol{\beta}^s_{1i}, \ldots, \boldsymbol{\beta}^s_{mi}, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_m) \prod_{k=1}^{m} p(r_{kit} | y^u_{kit}, \boldsymbol{\beta}^s_{ki}, \boldsymbol{\alpha}_k).$$

To sample $y^u_{1it}, \ldots, y^u_{mit}$, we need $m + 1$ independent uniform random numbers U_{it} and $V_{1it}, \ldots, V_{mit}$:

$$y^u_{kit} = -\log\left(-\frac{\log(U_{it})}{1 + \sum_{l=1}^{m}\lambda_{lit}} - \frac{\log(V_{kit})}{\lambda_{kit}} I_{\{y_{it} \neq k\}}\right), \tag{38}$$

where $\lambda_{kit} = \exp(\mathbf{x}^f_{it}\boldsymbol{\alpha}_k + \mathbf{x}^r_{it}\boldsymbol{\beta}^s_{ki})$, whereas each component indicator r_{kit} is sampled from a discrete distribution with $j = 1, \ldots, M$ categories:

$$\Pr(r_{kit} = j | y^u_{kit}, \boldsymbol{\beta}^s_{ki}, \boldsymbol{\alpha}_k) \propto \frac{w_j}{s_j} \exp\left\{-\frac{1}{2}\left(\frac{y^u_{kit} + \log\lambda_{kit} - m_j}{s_j}\right)^2\right\}.$$

4.2.3 Multi-move Sampling of all Regression Parameters

In this subsection we provide details on multi-move sampling of the category specific location parameters $\boldsymbol{\beta}^s_{k1}, \ldots, \boldsymbol{\beta}^s_{kN}, \boldsymbol{\beta}_k, \boldsymbol{\alpha}_k$ from the posterior

$$p(\boldsymbol{\beta}^s_{k1}, \ldots, \boldsymbol{\beta}^s_{kN}, \boldsymbol{\beta}_k, \boldsymbol{\alpha}_k | \mathbf{y}^u, \mathbf{R}, \mathbf{Q}_k) = \prod_{k=1}^{m} \prod_{i=1}^{N} p(\boldsymbol{\beta}^s_{ki} | \boldsymbol{\alpha}_k, \boldsymbol{\beta}_k, \mathbf{y}^u, \mathbf{R}) p(\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k | \mathbf{y}^u, \mathbf{R}, \mathbf{Q}_k), \quad (39)$$

which is carried out independently for each $k = 1, \ldots, m$.

First we sample $\boldsymbol{\alpha}_k$ and $\boldsymbol{\beta}_k$ from the marginal posterior $p(\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k | \mathbf{y}^u, \mathbf{R}, \mathbf{Q}_k)$ without conditioning on the random effects. For each unit i, the marginal model, where in (37) the random effects are integrated out is equal to a multivariate regression model with regression parameter $(\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k)$:

$$\mathbf{y}^u_{ki} = \mathbf{X}^f_i \boldsymbol{\alpha}_k + \mathbf{X}^r_i \boldsymbol{\beta}_k + \boldsymbol{m}_{ki} + \boldsymbol{\varepsilon}_{ki}, \quad \boldsymbol{\varepsilon}_{ki} \sim \mathcal{N}_{T_i}(\mathbf{0}, \mathbf{V}_{ki}), \quad (40)$$

using the matrix notation

$$\mathbf{X}^f_i = \begin{pmatrix} \mathbf{x}^f_{i1} \\ \vdots \\ \mathbf{x}^f_{i,T_i} \end{pmatrix}, \; \mathbf{X}^r_i = \begin{pmatrix} \mathbf{x}^r_{i1} \\ \vdots \\ \mathbf{x}^r_{i,T_i} \end{pmatrix},$$

$$\mathbf{y}^u_{ki} = \begin{pmatrix} y^u_{ki1} \\ \vdots \\ y^u_{ki,T_i} \end{pmatrix}, \; \boldsymbol{m}_{ki} = \begin{pmatrix} m_{r_{ki1}} \\ \vdots \\ m_{r_{ki,T_i}} \end{pmatrix}, \; \boldsymbol{\varepsilon}_{ki} = \begin{pmatrix} \varepsilon_{ki1} \\ \vdots \\ \varepsilon_{ki,T_i}, \end{pmatrix},$$

and error variance-covariance matrix $\mathbf{V}_{ki}$ given by:

$$\mathbf{V}_{ki} = \mathbf{X}^r_i \mathbf{Q}_k (\mathbf{X}^r_i)' + \boldsymbol{D}_{ki}, \quad \boldsymbol{D}_{ki} = \text{Diag}\left(s^2_{r_{ki1}}, \ldots, s^2_{r_{ki,T_i}}\right). \quad (41)$$

Assume a joint normal prior $\mathcal{N}_{d+r}(\mathbf{b}_{k0}, \mathbf{B}_{k0})$ for $(\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k)$ where $r = \dim(\boldsymbol{\alpha}_k)$. Then the posterior $p(\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k | \mathbf{y}^u, \mathbf{R}, \mathbf{Q}_k)$ is a multivariate normal distribution $\mathcal{N}_{d+r}(\mathbf{b}_{kN}, \mathbf{B}_{kN})$, where

$$\mathbf{B}^{-1}_{kN} = \mathbf{B}^{-1}_{k0} + \sum_{i=1}^{N} (\mathbf{X}_i)' \mathbf{V}^{-1}_{ki} \mathbf{X}_i, \quad \mathbf{b}_{kN} = \mathbf{B}_{kN} \left(\mathbf{B}^{-1}_{k0} \mathbf{b}_{k0} + \sum_{i=1}^{N} (\mathbf{X}_i)' \mathbf{V}^{-1}_{ki} (\mathbf{y}^u_{ki} - \boldsymbol{m}_{ki}) \right),$$
$$\mathbf{X}_i = \begin{pmatrix} \mathbf{X}^f_i & \mathbf{X}^r_i \end{pmatrix}.$$

The computation of these moments involves for each $i = 1, \ldots, N$ the inversion of the $(T_i \times T_i)$ matrix $\mathbf{V}_{ki}$. If the dimension d of $\boldsymbol{\beta}^s_{ki}$ is smaller than T_i the following inversion formula can be applied:

$$\begin{aligned} \mathbf{V}^{-1}_{ki} &= \boldsymbol{D}^{-1}_{ki} - \boldsymbol{D}^{-1}_{ki} \mathbf{X}^r_i \left(\mathbf{Q}^{-1}_k + (\mathbf{X}^r_i)' \boldsymbol{D}^{-1}_{ki} \mathbf{X}^r_i \right)^{-1} (\mathbf{X}^r_i)' \boldsymbol{D}^{-1}_{ki} \\ &= \boldsymbol{D}^{-1}_{ki} - \boldsymbol{D}^{-1}_{ki} \mathbf{X}^r_i \boldsymbol{A}_{ki} (\mathbf{X}^r_i)' \boldsymbol{D}^{-1}_{ki} \end{aligned} \quad (42)$$

which requires for each $i = 1, \ldots, N$ the inversion of a $(d \times d)$ matrix:

$$\boldsymbol{A}_{ki} = \left(\mathbf{Q}^{-1}_k + (\mathbf{X}^r_i)' \boldsymbol{D}^{-1}_{ki} \mathbf{X}^r_i \right)^{-1}. \quad (43)$$

Having drawn $\boldsymbol{\alpha}_k$ and $\boldsymbol{\beta}_k$, we sample for each $i = 1, \ldots, N$ the random effects $\boldsymbol{\beta}_{ki}^s$ conditional on $\boldsymbol{\alpha}_k$ and $\boldsymbol{\beta}_k$. The conditional posteriors $p(\boldsymbol{\beta}_{ki}^s|\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k, \mathbf{y}^u, \mathbf{R})$ is easily derived to be equal to the normal density $\mathcal{N}_d\left(\boldsymbol{a}_{ki}(\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k), \boldsymbol{A}_{ki}\right)$, where $\boldsymbol{A}_{ki}$ is the same as in (43), and the posterior mean reads:

$$\boldsymbol{a}_{ki}(\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k) = \boldsymbol{A}_{ki}\left(\mathbf{Q}_k^{-1}\boldsymbol{\beta}_k + (\mathbf{X}_i^r)'\boldsymbol{D}_{ki}^{-1}(\mathbf{y}_{ki}^u - \mathbf{X}_i^f\boldsymbol{\alpha}_k - \boldsymbol{m}_{ki})\right).$$

4.3 Modelling Data from the Binomial and the Multinomial Distribution

Auxiliary mixture sampling is easily extended to data from a binomial distribution. Consider a sequence of observations $z_1, \ldots, z_N$, where

$$z_i \sim \text{BiNom}\left(T_i, \pi_i\right), \qquad \log\frac{\pi_i}{1-\pi_i} = \mathbf{x}_i\boldsymbol{\beta},$$

with T_i being known. The binomial distribution is regarded as the marginal distribution of the number of successes y_{it} in T_i binary experiments with success probability π_i. Auxiliary mixture sampling is based on the full binary experiment, involving the repeated binary measurements y_{it}, where

$$y_{it} = \begin{cases} 1, & 1 \le t \le z_i, \\ 0, & z_i < t \le T_i. \end{cases}$$

Evidently,

$$\Pr(y_{it} = 1|\boldsymbol{\beta}) = \frac{\exp(\mathbf{x}_i\boldsymbol{\beta})}{1+\exp(\mathbf{x}_i\boldsymbol{\beta})},$$

and the application of auxiliary mixture sampling is straightforward.

A similar method may be applied to data from a multinomial distribution, which are regarded as repeated measurement of a categorical variable.

4.4 Application to the Austrian Labor Market

4.4.1 The Data and the Model

We consider a panel of Austrian employees who were observed between 1986 and 1998 on May 31st of each year (Weber, 2001). The data were obtained from the social security authority which collects detailed data for all employees. Here we use only a random sample of $N = 4376$ individuals. We reanalyze these data with the same wage categories as in Weber (2001). The wage of individual i in year t is modelled as a categorical variable y_{it} with states $k \in \{0, 1, \ldots, 5\}$, where category 0 corresponds to the no-income class. Non-zero wage data were categorized according to the quintiles of the yearly wage distribution into 5 income classes, coded as 1 to 5. For $t = 0, \ldots, T$, y_{it} takes the value k if person i belonged to wage category k at time t.

The number of available individual characteristics is rather small and incomplete; in particular there is no information on education, working time or family affiliation. The covariates $\boldsymbol{z}_{it}$ that are available are:

$$\boldsymbol{z}_{it} = \left(\begin{array}{cccc} \text{age}_i & \text{fem}_i & \text{change}_{it} & \text{whcollar}_{it} \end{array}\right),$$

where:

age_i	age of a person in 1986
fem_i	binary, 1 iff the person is female
change_{it}	binary, 1 iff the person's employers in years t and $t-1$ are different
whcollar_{it} ...	binary, 1 iff in year t the person is a white-collar employee

To analyze these data, Weber (2001) considered a multinomial logit regression model which captures overdispersion due to omitted covariates through a category specific regression intercept β^s_{ki} that varies between the units (Aitkin, 1996).

The multinomial model is based on an inhomogeneous Markov chain model which includes dependence of the occurrence probability of wage category y_{it} both on the wage category $y_{i,t-1}$ of the previous year and on the covariates $\boldsymbol{z}_{it}$. The model reads for $k = 1, \ldots, 5$ and $j = 0, \ldots, 5$:

$$\Pr(y_{it} = k | \boldsymbol{z}_{it}, y_{i,t-1} = j, \beta^s_{ki}) = \frac{\exp(\boldsymbol{z}_{it}\boldsymbol{\alpha}^z_k + \gamma_{jk} + \beta^s_{ki})}{1 + \sum_{l=1}^{5} \exp(\boldsymbol{z}_{it}\boldsymbol{\alpha}^z_l + \gamma_{jl} + \beta^s_{li})}. \tag{44}$$

The feedback parameter γ_{jk} captures the dynamic pattern of the Markov chain model and determine the transition matrix. Since for each k one of the feedback parameters γ_{jk} has to be assumed to be 0 for identifiability reasons, we set $\gamma_{0k} = 0$ for $k = 1, \ldots, 5$. Note that

$$\Pr(y_{it} = 0 | \boldsymbol{z}_{it}, y_{i,t-1} = j, \beta^s_{ki}) = \frac{1}{1 + \sum_{l=1}^{5} \exp(\boldsymbol{z}_{it}\boldsymbol{\alpha}^z_l + \gamma_{jl} + \beta^s_{li})},$$

therefore (44) also holds for the baseline category $k = 0$ with $\beta^s_{0i} = 0$, $\boldsymbol{\alpha}^z_0 = 0$, and $\gamma_{j0} = 0$.

The coefficients in this model have the following meaning. The elements of $\boldsymbol{\alpha}^z_k$ capture the effect of the corresponding covariate on the log odds ratio of moving to wage category k instead of moving to the no-income category. Thus the difference $\boldsymbol{\alpha}^z_k - \boldsymbol{\alpha}^z_{k'}$ captures the effect of the corresponding covariate on the log odd's ratio of moving to wage category k instead of moving to wage category k'. The feedback parameter γ_{jk} measures the difference in the log odds ratio of moving to wage category k instead of moving to the no-income category between a person coming from wage category j as opposed to a person coming from the no-income category. Thus the difference $\gamma_{jk} - \gamma_{j'k}$ measures the difference in the log odds ratio of moving to wage category k instead of moving to the no-income category between a person coming from wage category j as opposed to a person coming from wage category j'.

4.4.2 Parameter Estimation

It is well known that individual fixed parameters $\beta^s_{1i}, \ldots, \beta^s_{mi}$ cannot be estimated consistently within a maximum likelihood approach, since T_i is a typically small number of repetitions whereas only the number N of units is usually large. For this reason many researchers follow the random effects approach and assume that the individual parameters β^s_{ki} are drawn from a certain distribution of heterogeneity, commonly a normal distribution:

$$\beta^s_{ki} \sim \mathcal{N}(\beta_k, Q_k), \tag{45}$$

see, for instance, Verbeke and Lesaffre (1996) and Rossi et al. (2005, Chapter 5). If $Q_k = 0$ for all $k = 1, \ldots, 5$, then $\beta^s_{ki} = \beta_k$ and the model reduces to a standard multinomial regression model with fixed intercept β_k.

To obtain a model formulation as in (32), we introduce a design matrix for the feedback parameters and define $\boldsymbol{\alpha}_k = (\boldsymbol{\alpha}^z_k \, \gamma_{1k} \cdots \gamma_{5k})$ and

$$\mathbf{x}^f_{it} = \left(\begin{array}{ccccccc} \text{age}_i & \text{fem}_i & \text{change}_{it} & \text{whcollar}_{it} & I_{\{y_{i,t-1}=1\}} & \cdots & I_{\{y_{i,t-1}=5\}} \end{array} \right),$$

where $I_{\{y_{i,t-1}=k\}}$ captures the immediate income history of each person and takes 1 iff $y_{i,t-1} = k$. Note that $\mathbf{x}^r_{it} = 1$, since we are only dealing with a random intercept.

Marginally, model (44) is a mixed multinomial logit model with no closed form for the marginal probability $\Pr(y_{it} = k | \boldsymbol{z}_{it}, y_{i,t-1} = j)$ (McFadden and Train, 2000). Aitkin (1996) suggested, for the special case of binary data, to approximate the marginal distribution by a mixture of logit regression models using Gaussian-Hermite quadrature, and to use the resulting approximate likelihood function for inference, whereas McFadden and Train (2000) exploit maximum simulated likelihood estimation.

It is straightforward to implement a Bayesian approach for this model using the three block auxiliary mixture sampler described in Subsection 4.2. Our data augmentation scheme leads for each category to the normal random-effects regression model

$$\begin{gathered} \beta^s_{ki} \sim \mathcal{N}(\beta_k, Q_k), \\ y^u_{kit} = \mathbf{x}^f_{it} \boldsymbol{\alpha}_k + \beta^s_{ki} + m_{r_{kit}} + s_{r_{kit}} \varepsilon_{kit}, \quad \varepsilon_{kit} \sim \mathcal{N}(0,1), \end{gathered}$$

where the whole sequence $(\beta^s_{k1}, \ldots, \beta^s_{kN}, \beta_k, \boldsymbol{\alpha}_k)$ can be sampled simultaneously in an efficient manner.

To pursue the Bayesian approach we assume the following priors: for each category $k = 1, \ldots, 5$, the elements of $\boldsymbol{\alpha}^z_k$, β_k, and all parameters γ_{jk} are assumed to be independent apriori, each following a standard normal distribution. By choosing proper priors on the feedback parameters γ_{jk} we are able to avoid nonidentifiability due to unobserved transitions between certain wage categories. Finally, we assume an inverted Gamma $\mathcal{G}^{-1}(4, 3)$ prior on the variances Q_k, which implies a prior expectation of 1.

The auxiliary mixture sampler was run for 25000 iterations and Bayesian posterior inference is based on the last 12500 draws after discarding the first 12500 simulations.

4.4.3 Bayesian Posterior Inference

Parameter estimates and 95% credible regions for the parameters $\boldsymbol{\alpha}^z_k = (\alpha^z_{k1}, \ldots, \alpha^z_{k4)}$, $\gamma_{jk}, j = 1, \ldots, 5$, β_k, and Q_k are given for each category $k = 1, \ldots, 5$ in Table 5. The k-th column of this table corresponds to the effect of a certain covariate on the log odds of moving to wage category k as opposed to moving to the zero wage category.

Since a direct interpretation of the parameters is not easy, we estimated the transition matrix for 6 types of individuals born in 1971, who differ in some of their characteristics, namely male or female, white collar or blue collar worker, and having or not having changed employer. The corresponding transition matrices are

Table 5: Parameters estimated by the mean of the posterior distribution, 95% credible intervals in parentheses

Parameter	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
α^z_{k1} (age)	-0.004 $(-0.01, 0.0004)$	-0.0185 $(-0.023, -0.013)$	-0.016 $(-0.021, -0.011)$	-0.023 $(-0.029, -0.018)$	-0.023 $(-0.028, -0.018)$
α^z_{k2} (fem)	1.35 (1.23,1.48)	0.0393 $(-0.068, 0.14)$	-0.569 $(-0.69, -0.43)$	-0.722 $(-0.82, -0.61)$	-1.21 $(-1.36, -1.06)$
α^z_{k3} (change)	7.98 (7.79,8.16)	7.79 (7.61,7.97)	7.47 (7.32,7.63)	7.2 (7.01,7.38)	7.0 (6.83,7.17)
α^z_{k4} (whcollar)	-0.292 $(-0.43, -0.15)$	-0.197 $(-0.33, -0.074)$	0.27 (0.13,0.39)	0.73 (0.61,0.85)	1.7 (1.57,1.81)
γ_{1k} $(I_{\{y_{i,t-1}=1\}})$	5.19 (5.06,5.32)	4.29 (4.13,4.46)	3.29 (3.05,3.53)	2.0 (1.71,2.34)	1.16 (0.65,1.59)
γ_{2k} $(I_{\{y_{i,t-1}=2\}})$	3.59 (3.41,3.76)	6.45 (6.27,6.62)	5.77 (5.59,5.95)	3.7 (3.36,4.0)	0.484 (0.028,1.11)
γ_{3k} $(I_{\{y_{i,t-1}=3\}})$	2.16 (1.8,2.52)	4.82 (4.66,4.99)	7.38 (7.22,7.61)	6.33 (6.11,6.54)	2.96 (2.62,3.37)
γ_{4k} $(I_{\{y_{i,t-1}=4\}})$	1.09 (0.69,1.45)	2.63 (2.3,2.9)	5.6 (5.4,5.8)	8.07 (7.83,8.32)	6.47 (6.29,6.64)
γ_{5k} $(I_{\{y_{i,t-1}=5\}})$	0.785 (0.3,1.35)	1.04 (0.41,1.69)	2.51 (1.94,3.01)	5.68 (5.41,5.94)	8.57 (8.44,8.74)
β_k	-5.03 $(-5.21, -4.84)$	-4.42 $(-4.61, -4.24)$	-4.97 $(-5.21, -4.78)$	-5.27 $(-5.54, -5.02)$	-6.05 $(-6.23, -5.85)$
Q_k	0.77 (0.63,0.95)	0.63 (0.51,0.77)	0.40 (0.31,0.48)	0.44 (0.34,0.52)	0.24 (0.18,0.34)

estimated from (44) by averaging over the MCMC draws and are plotted in Figure 2. Independent of the other characteristics we find a typical difference between male and female employees. First the risk of moving to the no-income wage category is in general larger for women than for men. Second, the persistence probability in the lowest (positive) wage quintile ($k = 1$) is much larger for women than for men. Finally, for women the chance of moving to the next highest wage category is about the same as the risk of moving the next lowest one, whereas for men the chance of moving upwards is higher than the risk of moving downwards. This will exercise a strong cumulative effect on the wage of a woman as opposed to the wage of a man over the years.

Being a white collar rather than a blue collar worker has an effect in particular for workers having had an income in the upper wage categories in the past year, for both men and women. Thus the working status increases the chance of moving to the next highest wage category, and it also reduces the risk of moving to the no-income category.

Finally, changing the employer has a strong effect on the chance of moving out of the no-income category, for both women and men. For female and male workers being in the lowest (positive) wage quintile in the past years, changing the employee reduces dramatically the risk of moving to the no income category. While for women starting from the lowest (positive) wage quintile this mainly means improving the chance of staying in the same wage category, changing the employer improves for men starting from the lowest (positive) wage quintile also the chance of moving upwards. The effect of changing the employer disappears for workers starting in a higher wage quintile.

It is interesting to study the amount of unobserved heterogeneity in each wage category, measured by Q_k. The posterior mean of Q_k given in Table 5 decreases with increasing k, meaning that unobserved heterogeneity gets progressively smaller in the higher wage categories; see also Figure 3.

Finally, it is worthwhile to take a closer look at the expected deviation $\mathrm{E}(\beta_{ki}^s - \beta_k|\mathbf{y})$ of β_{ki}^s from β_k, which is estimated for each category $k = 1, \ldots, 5$ and for each individual by averaging over the MCMC draws $(\beta_{ki}^s)^{(m)} - \beta_k^{(m)}$. The box-plots in Figure 3 shows the empirical distributions of the estimates of $\mathrm{E}(\beta_{ki}^s - \beta_k|\mathbf{y})$ over all individuals for each wage category. Since these distributions are skewed to the right, covariates are missing that have a positive effect on the odds of being in this wage category as opposed to be in the no-wage category, which could be factors like education or working time.

5 Concluding Remarks

In this paper we have introduced a new data augmentation algorithm for sampling the parameters of a binary or a multinomial logit model from their posterior distribution within a Bayesian framework. The algorithm leads to a convenient Gibbs type sampler that draws from standard distributions like normal or exponential distributions and does not require any tuning. The auxiliary mixture sampler can be easily implemented for any binary or multinomial logit model, where the predictor is linear in the unknown parameters, with covariates being categorical as well as

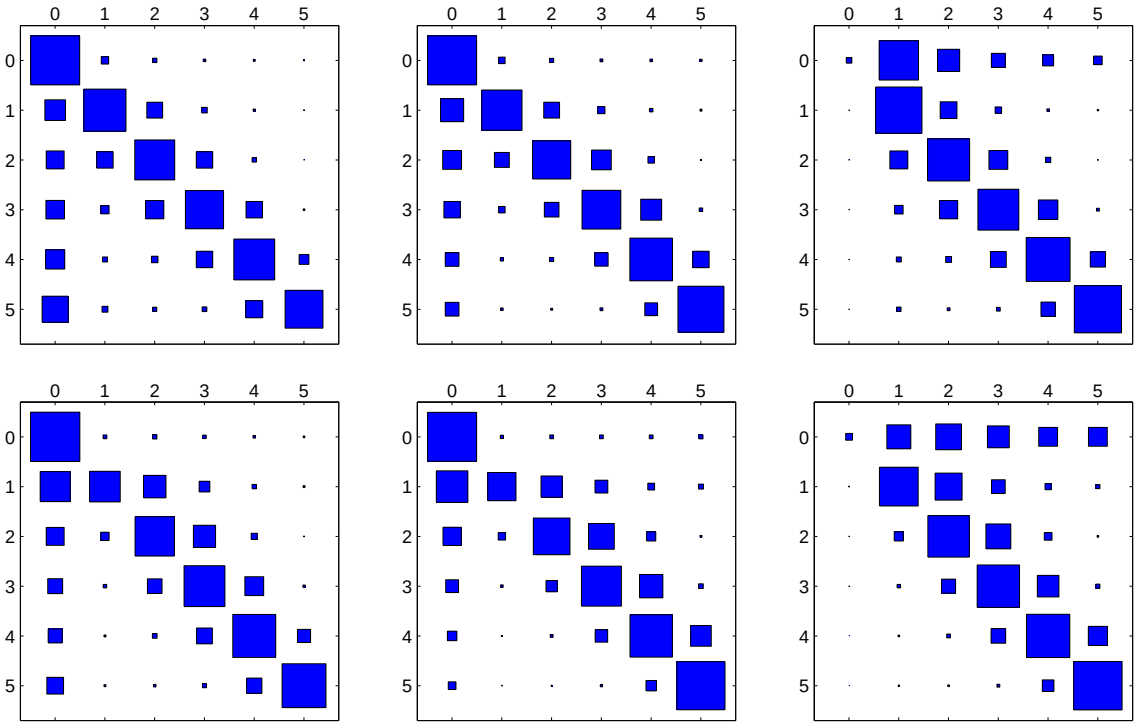

Figure 2: Estimated transition matrices between two years for six types of individuals born in 1971. Top: female, bottom: male; left: blue collar worker/no change of employee; center: white collar worker/no change of employee; right: white collar worker/change of employee. The area of each square is proportional to the transition probability.

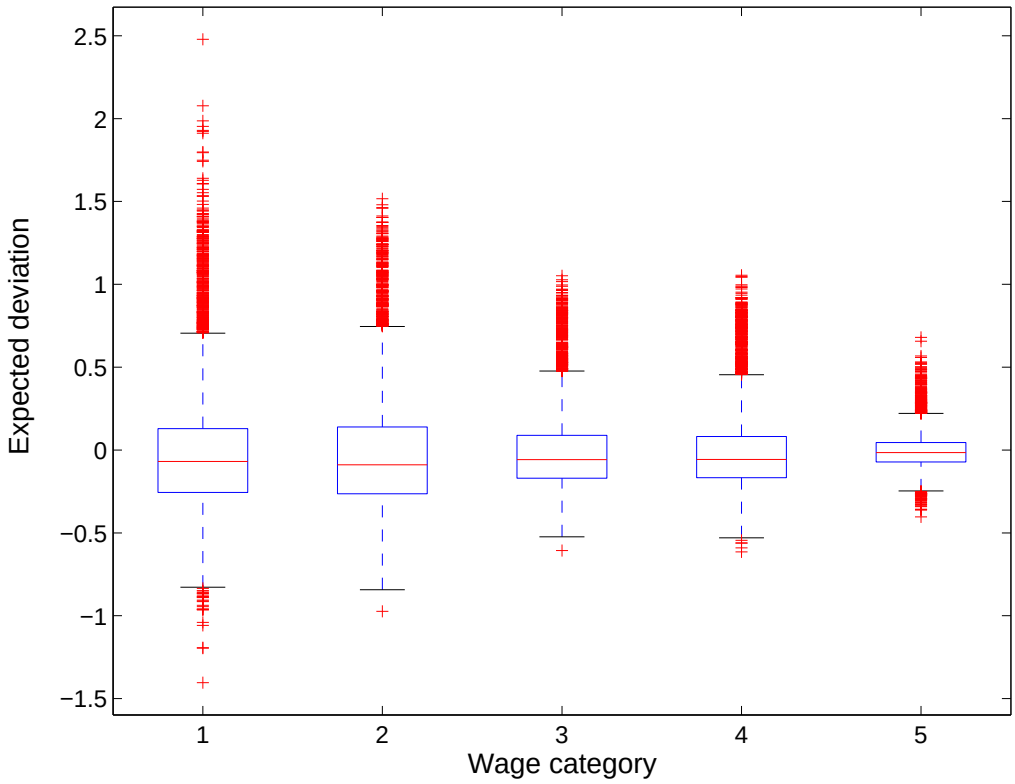

Figure 3: Box-plot of the distribution of the expected deviation $\mathrm{E}(\beta_{ki}^s - \beta_k|\mathbf{y})$ over all individuals, plotted for each wage category $k = 1, \ldots, 5$.

continuous.

Some care must be exercised when a weakly identified model is fitted to binary and multinomial data as data augmentation may lead to a poorly mixing sampler. In various applications we found that mixing improves when data augmentation is based on the scaled utilities $y^{u,\star}_{kit} = \delta y^u_{kit}$, with δ unknown. Although δ is unidentified from the data, it was noted by McCulloch and Rossi (1994) that introducing δ into the MCMC scheme speeds up convergence. A theoretical justification for this kind of data augmentation was developed by van Dyk and Meng (2001).

Note that auxiliary mixture sampling approximates the logit model by a mixture of probit models as in Geweke and Keane (1999). In our approach, however, the means, variances and weights of the mixture distribution are fixed rather than unknown. As suggested by one of the referees, the information in Table 1 could be used to set up a prior in the context of the Geweke and Keane (1999) approach to shrink the probit mixture model toward a logistic distribution.

Because of this relationship to mixtures of probit models, auxiliary mixture sampling could have been based on well-known methods for the Bayesian estimation of multinomial probit models (McCulloch and Rossi, 1994; McCulloch et al., 2000). We follow this approach in drawing $\boldsymbol{\beta}$ or, in more complex models, other regression parameters. There is, however, a minor inconsistency in our sampler, because we sample the utilities from the exact logistic distribution rather than from the probit regression, conditional on knowing the indicators. The reason for doing so is that our approach leads to very simple conditional densities, which are (translated) exponential distributions, and avoids the cumbersome sampling of the utilities from truncated m-dimensional normal densities as required in a multinomial probit model.

Acknowledgements

We thank Andrea Weber for providing the data analyzed in Subsection 4.4 and for helpful discussions. We also thank the anonymous referees for insightful and stimulating comments and suggestions.

References

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

Albert, J. H. and S. Chib (1993). Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts. *Journal of Business & Economic Statistics 11*, 1–15.

Andrews, D. F. and C. L. Mallows (1974). Scale mixtures of normal distributiond. *Journal of the Royal Statistical Society, Ser. B 36*, 99–102.

Carter, C. K. and R. Kohn (1994). On Gibbs sampling for state space models. *Biometrika 81*, 541–553.

Carter, C. K. and R. Kohn (1997). Semiparametric Bayesian inference for time series with mixed spectra. *Journal of the Royal Statistical Society, Ser. B 59*, 255–268.

Chib, S., E. Greenberg, and R. Winkelmann (1998). Posterior simulation and Bayes factors in panel count data models. *Journal of Econometrics 86*, 33–54.

Chib, S., F. Nardari, and N. Shephard (2002). Markov chain Monte Carlo methods for stochastic volatility models. *Journal of Econometrics 108*, 281–316.

De Jong, P. and N. Shephard (1995). The simulation smoother for time series models. *Biometrika 82*, 339–350.

Dellaportas, P. and A. F. M. Smith (1993). Bayesian inference for generalized linear and proportional hazards models via Gibbs sampling. *Applied Statistics 42*, 443–459.

Dey, D., S. K. Ghosh, and B. K. Mallick (Eds.) (2000). *Generalized Linear Models: a Bayesian Perspective*. NewYork/Basel: Marcel Dekker.

Durbin, J. and S. J. Koopman (2002). A simple and efficient simulation smoother for state space time series analysis. *Biometrika 89*, 603–615.

Frühwirth-Schnatter, S. (1994). Data augmentation and dynamic linear models. *Journal of Time Series Analysis 15*, 183–202.

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

Frühwirth-Schnatter, S. and T. Otter (1999). Conjoint-analysis using mixed effect models. In H. Friedl, A. Berghold, and G. Kauermann (Eds.), *Statistical Modelling. Proceedings of the Fourteenth International Workshop on Statistical Modelling*, Graz, pp. 181–191.

Frühwirth-Schnatter, S., R. Tüchler, and T. Otter (2004). Bayesian analysis of the heterogeneity model. *Journal of Business & Economic Statistics 22*, 2–15.

Frühwirth-Schnatter, S. and H. Wagner (2005). Data augmentation and Gibbs sampling for regression models of small counts. *Student 5*, 207–220. Available at http://www.ifas.jku.at/ as Research Report IFAS 2004-04.

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

Gamerman, D. (1997). Efficient sampling from the posterior distribution in generalized linear mixed models. *Statistics and Computing 7*, 57–68.

Gamerman, D. (1998). Markov chain Monte Carlo for dynamic generalized linear models. *Biometrika 85*, 215–227.

Geweke, J. and M. Keane (1999). Mixture of normals probit models. In C. Hsiao, M. H. Pesaran, K. L. Lahiri, and L. F. Lee (Eds.), *Analysis of Panels and Limited Dependent Variables*, pp. 49–78. Cambridge: Cambridge University Press.

Holmes, C. C. and L. Held (2006). Bayesian auxiliray 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.

Kim, S., N. Shephard, and S. Chib (1998). Stochastic volatility: Likelihood inference and comparison with ARCH models. *Review of Economic Studies 65*, 361–393.

Lenk, P. J. and W. S. DeSarbo (2000). Bayesian inference for finite mixtures of generalized linear models with random effects. *Psychometrika 65*, 93–119.

McCullagh, P. and J. A. Nelder (1999). *Generalized linear models*. Chapman & Hall Ltd.

McCulloch, R. and P. E. Rossi (1994). An exact likelihood analysis of the multinomial probit models. *Journal of Econometrics 64*, 207–240.

McCulloch, R. E., N. G. Polson, and P. E. Rossi (2000). A Bayesian analysis of the multinomial probit model with fully identified parameters. *Journal of Econometrics 99*, 173–193.

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

McFadden, D. and K. Train (2000). Mixed MNL models of discrete responses. *Journal of Applied Econometrics 15*, 447–470.

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

Omori, Y., S. Chib, N. Shephard, and J. Nakajima (2004). Stochastic volatility with leverage: fast likelihood inference. Research Report.

Rossi, P. E., G. M. Allenby, and R. McCulloch (2005). *Bayesian Statistics and Marketing*. Chichester: John Wiley & Sons Ltd.

Sahu, S. K. and G. O. Roberts (1999). On convergence of the EM algorithm and the Gibbs sampler. *Statistics and Computing 9*, 55–64.

Scott, S. L. (2004). Data augmentation, frequentistic estimation, and the Bayesian analysis of multinomial logit models.

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

Shephard, N. and M. K. Pitt (1997). Likelihood analysis of non-Gaussian measurement time series. *Biometrika 84*, 653–667.

Titterington, D. M., A. F. M. Smith, and U. E. Makov (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley Series in Probability and Statistics. New York: John Wiley & Sons Ltd.

van Dyk, D. and X.-L. Meng (2001). The art of data augmentation. *Journal of Computational and Graphical Statistics 10*, 1–50.

Verbeke, G. and E. Lesaffre (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association 91*, 217–221.

Wang, P., M. L. Puterman, I. Cockburn, and N. Le (1996). Mixed Poisson regression models with covariate dependent rates. *Biometrics 52*, 381–400.

Weber, A. (2001). State dependence and wage dynamics: A heterogeneous Markov chain model for wage mobility in Austria. Research report Institute for Advanced Sudies.

Zeger, S. and M. Karim (1991). Generalized linear models with random effects: a Gibbs sampling approach. *Journal of the American Statistical Association 86*, 79–86.

Zeger, S. L. and B. Qaqish (1988). Markov regression models for time series: A quasi-likelihood approach. *Biometrics 44*, 1019–1031.

Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. New York: John Wiley & Sons.

Zellner, A. and P. E. Rossi (1984). Bayesian analysis of dichotomous quantal response models. *Journal of Econometrics 25*, 365–393.