Bayesian Inference in the Multinomial Logit Model

Frühwirth-Schnatter, Sylvia; Frühwirth, Rudolf

Published in:
Austrian Journal of Statistics

DOI:
10.17713/ajs.v41i1.186

Published: 01/01/2012

Document Version
Publisher's PDF, also known as Version of record

AUSTRIAN JOURNAL OF STATISTICS
Volume 41 (2012), Number 1, 27–43

Bayesian Inference in the Multinomial Logit Model

Sylvia Frühwirth-Schnatter[1] and Rudolf Frühwirth[2]

[1]University of Economics and Business, Vienna
[2]Austrian Academy of Sciences, Vienna

Abstract: The multinomial logit model (MNL) possesses a latent variable representation in terms of random variables following a multivariate logistic distribution. Based on multivariate finite mixture approximations of the multivariate logistic distribution, various data-augmented Metropolis-Hastings algorithms are developed for a Bayesian inference of the MNL model.

Zusammenfassung: Das multinomiale logistische (MNL) Regressionsmodell besitzt eine latente Variablendarstellung, die einen zufälligen Fehlerterm beinhaltet, der einer multivariaten logistischen Verteilung folgt. Aufbauend auf einer finiten Mischungsapproximation der multivariaten logistischen Verteilung werden mehrere Metropolis-Hastings-Verfahren für eine Bayes-Analyse im MNL Regressionsmodell entwickelt.

Keywords: Bayesian Inference, Finite Mixture Distributions, Markov Chain Monte Carlo, Metropolis-Hastings Algorithm, Multinomial Logit Model, Multivariate Logistic Distribution.

1 Introduction

In the past decades, finite mixture modeling became a rapidly developing area with numerous applications in biometrics, economics, genetics, medicine, among many others; see Frühwirth-Schnatter (2006) for a review. An early application of finite normal mixture models has been modeling aberrant observations in astronomical data of transit of Mercury (Newcomb, 1886). One of the pioneering papers discussing a Bayesian approach to outlier analysis based on finite normal mixtures is the work by Guttman, Dutter, and Freeman (1978).

Finite normal mixture distributions are useful for practical data analysis because they capture many specific properties of real data such as multimodality, skewness, and kurtosis. Also, they arise in a natural way as marginal distribution for statistical models involving clustering or unobserved heterogeneity. Moreover, they are useful for developing efficient estimation procedures for non-Gaussian models, early examples being Sorenson and Alspach (1971) and Alspach and Sorenson (1972).

Finite mixture distributions possess the following approximation property (Titterington, Smith, and Makov, 1985). Let $g(\varepsilon)$ be an arbitrary probability density function and let $q_K(\varepsilon)$ be a mixture of normals:

$$q_K(\varepsilon) = \sum_{k=1}^{K} w_k f_N(\varepsilon; m_k, s_k^2). \tag{1}$$

For sufficiently large K the Kullback-Leibler (KL-)distance between $g(\varepsilon)$ and $q_K(\varepsilon)$,

$$d_{\mathrm{KL}} = \int_{\Re} g(\varepsilon) \log \frac{g(\varepsilon)}{q_K(\varepsilon)} \mathrm{d}\varepsilon$$

can be made arbitrarily small. To approximate $g(\varepsilon)$ for a fixed K, one has to select the weights $w_1, \ldots, w_K$, the means $m_1, \ldots, m_K$ and the variances $s_1^2, \ldots, s_K^2$ such that d_{KL} is minimized. It should be noted that this is not a parameter estimation problem, but a problem of numerical optimization.

It has been noted by several authors that Bayesian inference is considerably simpler for many non-Gaussian models if a certain density $g(\varepsilon)$ is replaced by an accurate finite mixture approximation $q_K(\varepsilon)$. This is in particular true for Markov chain Monte Carlo (MCMC) estimation, where substitution of $g(\varepsilon)$ by $q_K(\varepsilon)$ leads to simple Gibbs-type sampling schemes; see Gamerman and Lopes (2006) for a review of MCMC methods. Typically, the density $g(\varepsilon)$ is independent of any parameters or depends only on an integer parameter; hence the optimal parameters in (1) can be obtained beforehand.

Applications of this auxiliary mixture sampling approach include stochastic volatility modeling (Shephard, 1994; Chib, Nardari, and Shephard, 2002; Omori, Chib, Shephard, and Nakajima, 2007), where $g(\varepsilon)$ is the density of the log of a χ_1^2-distributed random variable. A series of recent papers applies this approach to modeling discrete-valued data such as count data (Frühwirth-Schnatter and Wagner, 2006; Frühwirth-Schnatter, Frühwirth, Held, and Rue, 2009) and binary and categorical data (Frühwirth-Schnatter and Frühwirth, 2007, 2010). In these cases, $g(\varepsilon)$ is, respectively, the density of the negative logarithm of an $\mathcal{E}(1)$- or a $\mathcal{G}(\nu, 1)$-distributed random variable with integer ν, or the logistic distribution.

In all of these papers, $g(\varepsilon)$ is a univariate density, hence even moderate values of K yield a good approximation. The present work is a first attempt at taking the idea of auxiliary mixture sampling to higher dimensions, which requires that a multivariate density $g(\boldsymbol{\varepsilon})$ is approximated by a multivariate mixture distribution. As an example, we consider Bayesian inference for multinomial logit regression modeling of discrete outcome variables with $m + 1$ categories. Data augmentation leads to an error term possessing an m-variate logistic distribution which is independent of any parameters and has a quite rigid structure. We will approximate this distribution by both multivariate normal and multivariate Student-t mixtures, minimizing again the KL distance. However, due to the curse of dimensionality, we do not expect to obtain perfect approximations. Nevertheless, these mixture approximations may be used to construct a joint proposal for all regression parameters within a Metropolis-Hastings (MH) algorithm.

2 Data Augmentation for the Multinomial Logit Regression Model

Let $\{y_i\}$ be a sequence of categorical data, $i = 1, \ldots, N$, where y_i is equal to one of $m+1$ unordered categories, labeled by $L = \{0, \ldots, m\}$. Very often it is of interest to model the probability that y_i takes the value k, for each $k \in \{1, \ldots, m\}$, in terms of covariate information. A popular choice is the multinomial logit (MNL) model for which the choice

probabilities are easily computed. Usually, the MNL model takes the following somewhat restricted form:

$$\Pr(y_i = k|\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m) = \frac{\exp(\mathbf{x}_i\boldsymbol{\beta}_k)}{1 + \sum_{l=1}^{m} \exp(\mathbf{x}_i\boldsymbol{\beta}_l)}\,, \tag{2}$$

where $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m$ are category specific unknown regression coefficients of dimension d and $\mathbf{x}_i$ is a $(1 \times d)$ row vector containing covariates which are not category specific.

However, for many important applications the MNL model takes a more general form, where the choice probabilities contain regression coefficients that are not category specific. Examples include discrete choice models in marketing (Rossi, Allenby, and McCulloch, 2005) and the partial credit model, used in large educational assessment programs such as PISA (Fox, 2010). In its most general form, the probability that y_i takes the value k is modeled for $k = 1, \ldots, m$ in the following way:

$$\Pr(y_i = k|\boldsymbol{\beta}) = \frac{\exp(\mathbf{x}_{ki}\boldsymbol{\beta})}{1 + \sum_{l=1}^{m} \exp(\mathbf{x}_{ki}\boldsymbol{\beta})}\,, \tag{3}$$

where $\mathbf{x}_{ki}$ is a $(1 \times r)$-dimensional, category specific covariate vector and $\boldsymbol{\beta}$ are unknown regression coefficients of dimension r.

2.1 The RUM and the dRUM Representation

Following McFadden (1974), the MNL model (3) may be written as the following random utility model (RUM):

$$y^u_{0i} = \epsilon_{0i}\,, \tag{4}$$

$$y^u_{ki} = \mathbf{x}_{ki}\boldsymbol{\beta} + \epsilon_{ki}\,, \quad k = 1, \ldots, m\,, \tag{5}$$

$$y_i = k \;\Leftrightarrow\; y^u_{ki} = \max_{l \in L} y^u_{li}\,. \tag{6}$$

Thus the observed category is equal to the category with maximal utility. If the random utilities $\epsilon_{0i}, \epsilon_{1i}, \ldots, \epsilon_{mi}$ appearing in (4) and (5) are i.i.d. following an extreme value distribution, then the MNL model (3) results as the marginal distribution of y_i.

An alternative way to write the MNL model (3) is a difference random utility model (dRUM), which is obtained by choosing a baseline category k_0, typically $k_0 = 0$, and considering the model in terms of the differences of the utilities. From (4) to (6) we obtain the following dRUM representation:

$$z_{ki} = \mathbf{x}_{ki}\boldsymbol{\beta} + \varepsilon_{ki}\,, \qquad \varepsilon_{ki} \sim \mathcal{LO}\,, \qquad k = 1, \ldots, m\,, \tag{7}$$

$$y_i = \begin{cases} 0, & \text{if } \max_{l \in L_{-0}} z_{li} < 0, \\ k > 0, & \text{if } z_{ki} = \max_{l \in L_{-0}} z_{li} > 0, \end{cases}$$

where $z_{ki} = y^u_{ki} - y^u_{0i}$ and $\varepsilon_{ki} = \epsilon_{ki} - \epsilon_{0i}$, and L_{-0} is the set of all categories but 0.

The dRUM representation is the standard choice for the multinomial probit model (see e.g. McCulloch, Polson, and Rossi (2000) and Imai and van Dyk (2005)), but is less commonly used for the multinomial logit model, exceptions being Holmes and Held (2006) and Frühwirth-Schnatter and Frühwirth (2010).

Whereas in the multinomial probit model the error term follows a multivariate normal distribution, the vector $\boldsymbol{\varepsilon}_i$ that appears in the dRUM representation (7) of the MNL model has a multivariate logistic distribution. The multivariate logistic distribution was introduced by Malik and Abraham (1973) as a generalization of Gumbel's bivariate logistic distribution (Gumbel, 1961). If m denotes the number of variates, its pdf reads

$$f_{\mathcal{LO}_m}(\boldsymbol{\varepsilon}) = f_{\mathcal{LO}_m}(\varepsilon_1, \ldots, \varepsilon_m) = m! \frac{\exp(-\sum_{l=1}^m \varepsilon_l)}{(1 + \sum_{l=1}^m \exp(-\varepsilon_l))^{m+1}} .$$

As shown by Balakrishnan (1992, Section 11.2), a multivariate logistic distribution results if an i.i.d. sequence of $(m+1)$ random variables $\boldsymbol{\epsilon} = (\epsilon_0, \ldots, \epsilon_m)'$ from the extreme value distribution is transformed into a sequence of m random variables $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_m)$ by setting $\varepsilon_k = \epsilon_k - \epsilon_0$ for $k = 1, \ldots, m$. This is exactly the transformation of the error $\boldsymbol{\epsilon}$ in the RUM representation to the error $\boldsymbol{\varepsilon}$ in the dRUM representation.

Kotz, Johnson, and Balakrishnan (2000, Chapter 51) provides a comprehensive review of further properties of the multivariate logistic distribution. For instance, while the errors in the RUM representation (5) are i.i.d., the errors in the dRUM representation (7) are no longer independent across categories, but correlated. The variance-covariance matrix $\mathbf{R}$ of $\boldsymbol{\varepsilon}$ is given by

$$\mathbf{R} = \frac{\pi^2}{6}(\mathbf{I}_m + \mathbf{e}_m \mathbf{e}_m^{'}) = \frac{\pi^2}{3} \begin{pmatrix} 1 & 0.5 & \cdots & 0.5 \\ 0.5 & 1 & \cdots & 0.5 \\ \vdots & \vdots & \ddots & \vdots \\ 0.5 & 0.5 & \cdots & 1 \end{pmatrix} .$$

Since the correlation coefficient is equal to 0.5 for all pairs $(\varepsilon_k, \varepsilon_l)$, $\mathbf{R}$ is a uniform covariance matrix. It follows immediately that the inverse $\mathbf{R}^{-1}$ can be computed explicitly:

$$\mathbf{R}^{-1} = \frac{6}{\pi^2}(\mathbf{I}_m - \frac{1}{m+1} \mathbf{e}_m \mathbf{e}_m^{'}) . \tag{8}$$

2.2 Bayesian Inference

Subsequently, we pursue a Bayesian approach and assume that *a priori* the regression coefficient $\boldsymbol{\beta}$ follows a normal distribution $\mathcal{N}_r(\mathbf{b}_0, \mathbf{B}_0)$ with known hyperparameters $\mathbf{b}_0$ and $\mathbf{B}_0$.

Since the posterior distribution $p(\boldsymbol{\beta}|\mathbf{y})$ of the regression coefficient $\boldsymbol{\beta}$ in the MNL model does not have any closed form, it is usual to apply data augmentation and Markov chain Monte Carlo estimation; see Frühwirth-Schnatter and Frühwirth (2010) for a recent review. Data augmentation has been based both on the RUM representation (Frühwirth-Schnatter and Frühwirth, 2007; Scott, 2011) and on the dRUM representation (Holmes and Held, 2006; Frühwirth-Schnatter and Frühwirth, 2010). The case studies in Frühwirth-Schnatter and Frühwirth (2010, Section 4) reveal that the corresponding MCMC samplers are much more efficient for the dRUM representation than for the RUM representation.

However, the presence of the multivariate logistic distribution complicates MCMC sampling for the dRUM representation. If only category specific coefficients are present, as in (2), then it is possible to derive a partial dRUM representation of the MNL model. For each category k, the corresponding latent variable representation is a dRUM representation of a binary logit model, with category k being one outcome and all alternative categories being the second outcome. This allows to apply any of the efficient samplers that have been developed in Frühwirth-Schnatter and Frühwirth (2010, Section 3.2) from the dRUM representation of a binary logit model. The sampler developed by Holmes and Held (2006) also works with the partial dRUM representation, but is much more involved in terms of computing time and therefore less efficient.

Regrettably, the partial dRUM representation does not lead to simple MCMC sampling for the more general model (3), which contains regression coefficients that are not category specific. Alternative sampling methods for this case have been developed and will be presented in the following section.

3 Data Augmented Metropolis-Hastings Algorithms in the dRUM Representation

The data augmented MH algorithm operates in the dRUM representation (7) of the MNL model. Following the MCMC literature on the multinomial probit model, the latent variables $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_N)$, where $\mathbf{z}_i = (z_{1i}, \ldots, z_{mi})'$ are introduced as missing data. The sampler iterates between sampling from $\boldsymbol{\beta}|\mathbf{z}$ and sampling from $\mathbf{z}|\boldsymbol{\beta}, \mathbf{y}$:

(a) Sample from $\boldsymbol{\beta}|\mathbf{z}$;

(b) Sample from $\mathbf{z}|\boldsymbol{\beta}, \mathbf{y}$.

A closed form Gibbs step is available for joint sampling of $\mathbf{z}|\boldsymbol{\beta}, \mathbf{y}$. To sample $\mathbf{z}_i$, for $i = 1, \ldots, N$, we sample the latent utilities $\mathbf{y}_i^u = (y_{0i}^u, \ldots, y_{mi}^u)$ in the RUM model (4) to (6) from the posterior $\mathbf{y}_i^u|\boldsymbol{\beta}, \mathbf{y}$, which is given by:

$$y_{ki}^u = -\log\left(-\frac{\log(U_i)}{\sum_{l=0}^m \lambda_{li}} - \frac{\log(V_{ki})}{\lambda_{ki}} I\{y_i \neq k\}\right),$$

where U_i and $V_{1i}, \ldots, V_{mi}$ are $m+1$ independent uniform random numbers in $[0,1]$, $\lambda_{li} = \exp(\mathbf{x}_{li}\boldsymbol{\beta})$ for $l = 1, \ldots, m$, and $\lambda_{0i} = 1$. Then we define $\mathbf{z}_i = (z_{1i}, \ldots, z_{mi})'$ as the differences in utility, i.e. $z_{ki} = y_{ki}^u - y_{0i}^u$, $k = 1, \ldots, m$.

To sample from $\boldsymbol{\beta}|\mathbf{z}$, we rewrite the dRUM model (7) as multivariate regression model:

$$\mathbf{z}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i\,, \tag{9}$$

where $\mathbf{X}_i$ is a $(m \times r)$-matrix with the kth row being equal to $\mathbf{x}_{ki}$. However, whereas sampling from $\boldsymbol{\beta}|\mathbf{z}$ is straightforward for the multinomial probit model, because $\boldsymbol{\varepsilon}_i$ is multivariate normal, this step is non-standard in the MNL model because $\boldsymbol{\varepsilon}_i$ is multivariate logistic. Subsequently, we suggest and compare various MH algorithms for joint sampling from $\boldsymbol{\beta}|\mathbf{z}$.

3.1 A Data-augmented Independence Metropolis-Hastings Sampler

First, we construct an independence MH step, by sampling $\boldsymbol{\beta}^{\text{new}}$ from a proposal density $q(\boldsymbol{\beta}|\mathbf{z})$ which is independent of the previous draw of $\boldsymbol{\beta}$. As usual, $\boldsymbol{\beta}^{\text{new}}$ is accepted with probability $P = \min(\alpha, 1)$, where:

$$\alpha = \frac{p(\mathbf{z}|\boldsymbol{\beta}^{\text{new}})p(\boldsymbol{\beta}^{\text{new}})q(\boldsymbol{\beta}|\mathbf{z})}{p(\mathbf{z}|\boldsymbol{\beta})p(\boldsymbol{\beta})q(\boldsymbol{\beta}^{\text{new}}|\mathbf{z})} .$$

The proposal density $q(\boldsymbol{\beta}|\mathbf{z})$ is based on approximating the distribution of $\boldsymbol{\varepsilon}_i$ in (9) by a multivariate normal distribution with the expectation (which is equal to $\mathbf{0}$) and the variance-covariance matrix $\mathbf{R}$, given in (8), of the m-variate logistic distribution. This leads to a multivariate regression model with homoscedastic, equi-correlated errors, which reads for $i = 1, \ldots, N$:

$$\mathbf{z}_i = \mathbf{X}_i\boldsymbol{\beta} + \tilde{\boldsymbol{\varepsilon}}_i , \qquad \tilde{\boldsymbol{\varepsilon}}_i \sim \mathcal{N}_m(\mathbf{0}, \mathbf{R}) .$$

Under the prior distribution $\boldsymbol{\beta} \sim \mathcal{N}_r(\mathbf{b}_0, \mathbf{B}_0)$, the posterior of this approximate model is equal to the multivariate normal distribution $\mathcal{N}_r(\mathbf{b}_N, \mathbf{B}_N)$ with moments:

$$\mathbf{b}_N = \mathbf{B}_N \left(\mathbf{B}_0^{-1}\mathbf{b}_0 + \sum_{i=1}^{N} \mathbf{X}_i^{'}\mathbf{R}^{-1}\mathbf{z}_i \right) , \qquad \mathbf{B}_N = \left(\mathbf{B}_0^{-1} + \sum_{i=1}^{N} \mathbf{X}_i^{'}\mathbf{R}^{-1}\mathbf{X}_i \right)^{-1} .$$

This posterior is then used as proposal $q(\boldsymbol{\beta}|\mathbf{z})$. By using the explicit expression for $\mathbf{R}^{-1}$ in (8) we obtain:

$$\mathbf{B}_N^{-1} = \mathbf{B}_0^{-1} + \frac{6}{\pi^2} \left(\sum_{i=1}^{N} \mathbf{X}_i^{'}\mathbf{X}_i - \frac{1}{m+1} \sum_{i=1}^{N} \mathbf{w}_i\mathbf{w}_i^{'} \right) ,$$
$$\mathbf{B}_N^{-1}\mathbf{b}_N = \mathbf{B}_0^{-1}\mathbf{b}_0 + \frac{6}{\pi^2} \left(\sum_{i=1}^{N} \mathbf{X}_i^{'}\mathbf{z}_i - \frac{1}{m+1} \sum_{i=1}^{N} \mathbf{w}_i c_i \right) ,$$

where $\mathbf{w}_i = \mathbf{X}_i^{'}\mathbf{e}_m = (\mathbf{e}_m^{'}\mathbf{X}_i)^{'}$ is a $(r \times 1)$ column vector containing the column sums of $\mathbf{X}_i$ and $c_i = \mathbf{e}_m^{'}\mathbf{z}_i$ is a scalar containing the column sum of $\mathbf{z}_i$.

For the special case of model (2) where $r = d \cdot m$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m)$ contains only category specific covariates, $\mathbf{X}_i = \mathbf{I}_m \otimes \mathbf{x}_i$, hence

$$\mathbf{X}_i^{'}\mathbf{X}_i = \mathbf{I}_m \otimes (\mathbf{x}_i^{'}\mathbf{x}_i) , \quad \mathbf{w}_i = \mathbf{e}_m \otimes \mathbf{x}_i^{'} , \quad \mathbf{w}_i\mathbf{w}_i^{'} = \mathbf{e}_m \otimes (\mathbf{x}_i^{'}\mathbf{x}_i) ,$$
$$\mathbf{X}_i^{'}\mathbf{z}_i = \mathbf{z}_i \otimes \mathbf{x}_i^{'} , \quad \mathbf{w}_i c_i = c_i\mathbf{e}_m \otimes \mathbf{x}_i^{'} .$$

Therefore:

$$\mathbf{B}_N^{-1} \sum_{i=1}^{N} = \mathbf{B}_0^{-1} + \mathbf{R}^{-1} \otimes \left(\sum_{i=1}^{N} \mathbf{x}_i^{'}\mathbf{x}_i \right) , \quad \mathbf{B}_N^{-1}\mathbf{b}_N = \mathbf{B}_0^{-1}\mathbf{b}_0 + \frac{6}{\pi^2} \left(\sum_{i=1}^{N} (\mathbf{z}_i - c_i\mathbf{e}_m) \otimes \mathbf{x}_i^{'} \right) .$$

If each coefficient $\boldsymbol{\beta}_k$ has the same normal prior $\boldsymbol{\beta}_k \sim \mathcal{N}_d\left(\tilde{\mathbf{b}}_0, \tilde{\mathbf{B}}_0\right)$, then $\mathbf{B}_0^{-1}\mathbf{b}_0 = \mathbf{e}_m \otimes (\tilde{\mathbf{B}}_0^{-1}\tilde{\mathbf{b}}_0)$ and $\mathbf{B}_0^{-1} = \mathbf{I}_m \otimes \tilde{\mathbf{B}}_0^{-1}$.

3.2 Doubly Data-augmented Metropolis-Hastings Samplers

In this section we construct rather general MH samplers for the multinomial logit model by approximating the distribution of ε_i in (9) by a mixture of multivariate distributions, where the pdf $\tilde{p}(\varepsilon_i)$ results as the marginal density of an $(m+1)$-variate random variable $(\varepsilon_i, \lambda_i)$, with $\varepsilon_i|\lambda_i$ following a normal distribution, i.e. $\varepsilon_i|\lambda_i \sim \mathcal{N}_m(\mathbf{m}_i, \mathbf{R}_i)$, and $\lambda_i \sim p(\lambda_i)$. We focus on mixture distributions where it is easy to sample from the conditional density $\lambda_i|\varepsilon_i$, such as the multivariate Student-t distribution, which is a scale mixture of multivariate normal distributions, finite multivariate normal mixtures, and finite multivariate Student-t mixtures. Once the approximate distribution $\tilde{p}(\varepsilon_i)$ has been chosen, no further tuning parameters appear in this MH-sampler.

3.2.1 Sampling Scheme

The advantage of approximating the distribution of ε_i in (9) by a mixture of multivariate normals is that double data augmentation, i.e. conditioning on the latent variables $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_N)$ in addition to the latent variables $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_N)$, leads to a multivariate regression model with normally distributed errors which reads for $i = 1, \ldots, N$:

$$\mathbf{z}_i = \mathbf{X}_i\boldsymbol{\beta} + \tilde{\boldsymbol{\varepsilon}}_i\,, \qquad \tilde{\boldsymbol{\varepsilon}}_i \sim \mathcal{N}_m(\mathbf{m}_i, \mathbf{R}_i)\,. \tag{10}$$

The posterior $q(\boldsymbol{\beta}|\boldsymbol{\lambda}, \mathbf{z})$ of this model is given by

$$q(\boldsymbol{\beta}|\boldsymbol{\lambda}, \mathbf{z}) = \frac{p(\boldsymbol{\beta}) \prod_{i=1}^N f_N(\mathbf{z}_i|\boldsymbol{\beta}, \lambda_i)}{p(\mathbf{z}|\boldsymbol{\lambda})}\,,$$

and is equal to a normal distribution $\mathcal{N}_r(\mathbf{b}_N, \mathbf{B}_N)$. All approximate error distributions discussed below have in common that the error covariance matrix $\mathbf{R}_i$ in the approximate model (10) is a uniform covariance matrix, i.e., $\mathbf{R}_i = \sigma_i^2\boldsymbol{C}_i$, where $\boldsymbol{C}_i = (1-\rho_i)\mathbf{I}_m + \rho_i\mathbf{e}_m\mathbf{e}_m'$. This leads to a straightforward way to compute the moments $\mathbf{b}_N$ and $\mathbf{B}_N$:

$$\mathbf{B}_N^{-1} = \mathbf{B}_0^{-1} + \left(\sum_{i=1}^N a_i\mathbf{X}_i'\mathbf{X}_i - \sum_{i=1}^N b_i\mathbf{w}_i\mathbf{w}_i'\right)\,,$$

$$\mathbf{B}_N^{-1}\mathbf{b}_N = \mathbf{B}_0^{-1}\mathbf{b}_0 + \left(\sum_{i=1}^N a_i\mathbf{X}_i'(\mathbf{z}_i - \mathbf{m}_i) - \sum_{i=1}^N b_i\mathbf{w}_i d_i\right)\,,$$

where the $(r \times 1)$-vector $\mathbf{w}_i = \mathbf{X}_i'\mathbf{e}_m$ contains the column sums of $\mathbf{X}_i$, the scalar $d_i = \mathbf{e}_m'(\mathbf{z}_i - \mathbf{m}_i)$ contains the sum of all elements of $(\mathbf{z}_i - \mathbf{m}_i)$, and a_i and b_i are given by:

$$a_i = \frac{1}{\sigma_i^2(1-\rho_i)}\,, \qquad b_i = \frac{\rho_i}{\sigma_i^2(1-\rho_i)(1+(m-1)\rho_i)}\,.$$

Since $q(\boldsymbol{\beta}|\boldsymbol{\lambda}, \mathbf{z})$ is an important building block of our MH-algorithm, we call the resulting sampler a doubly data-augmented MH sampler. Conditional on the utilities $\mathbf{z}$, the proposal $q(\boldsymbol{\beta}^{\text{new}}|\boldsymbol{\beta}^{\text{old}})$ is constructed in the following way:

$$q(\boldsymbol{\beta}^{\text{new}}|\boldsymbol{\beta}^{\text{old}}) = q(\boldsymbol{\beta}^{\text{new}}|\boldsymbol{\lambda}, \mathbf{z}) \prod_{i=1}^N q(\lambda_i|\boldsymbol{\beta}^{\text{old}}, \mathbf{z}_i)\,,$$

where $q(\boldsymbol{\beta}^{\text{new}}|\boldsymbol{\lambda}, \mathbf{z})$ is the posterior of model (10) and $q(\lambda_i|\boldsymbol{\beta}^{\text{old}}, \mathbf{z}_i)$ is equal to the conditional posterior of λ_i given $\boldsymbol{\varepsilon}_i = \mathbf{z}_i - \mathbf{X}_i\boldsymbol{\beta}^{\text{old}}$:

$$q(\lambda_i|\boldsymbol{\beta}^{\text{old}}, \mathbf{z}_i) = \frac{f_N(\mathbf{z}_i|\boldsymbol{\beta}^{\text{old}}, \lambda_i)p(\lambda_i)}{\tilde{p}(\mathbf{z}_i|\boldsymbol{\beta}^{\text{old}})},$$

which is available in closed form according to our assumption.

It is easy to show that

$$\begin{aligned} q(\boldsymbol{\beta}^{\text{new}}|\boldsymbol{\beta}^{\text{old}}) &= \frac{p(\boldsymbol{\beta}^{\text{new}})}{p(\mathbf{z}|\boldsymbol{\lambda})} \prod_{i=1}^{N} f_N(\mathbf{z}_i|\boldsymbol{\beta}^{\text{new}}, \lambda_i) \frac{f_N(\mathbf{z}_i|\boldsymbol{\beta}^{\text{old}}, \lambda_i)p(\lambda_i)}{\tilde{p}(\mathbf{z}_i|\boldsymbol{\beta}^{\text{old}})} \\ &= \frac{p(\boldsymbol{\beta}^{\text{new}})}{p(\boldsymbol{\beta}^{\text{old}})} \prod_{i=1}^{N} \frac{\tilde{p}(\mathbf{z}_i|\boldsymbol{\beta}^{\text{new}})}{\tilde{p}(\mathbf{z}_i|\boldsymbol{\beta}^{\text{old}})} q(\boldsymbol{\beta}^{\text{old}}|\boldsymbol{\lambda}, \mathbf{z}) \prod_{i=1}^{N} q(\lambda_i|\boldsymbol{\beta}^{\text{new}}, \mathbf{z}_i) \\ &= \frac{p(\boldsymbol{\beta}^{\text{new}})}{p(\boldsymbol{\beta}^{\text{old}})} q(\boldsymbol{\beta}^{\text{old}}|\boldsymbol{\beta}^{\text{new}}) \prod_{i=1}^{N} \frac{\tilde{p}(\mathbf{z}_i|\boldsymbol{\beta}^{\text{new}})}{\tilde{p}(\mathbf{z}_i|\boldsymbol{\beta}^{\text{old}})}. \end{aligned}$$

Therefore, the acceptance probability $P = \min(\alpha, 1)$ may be expressed entirely in terms of likelihood ratios between the exact multivariate logistic distribution and the approximate distribution $\tilde{p}(\boldsymbol{\varepsilon})$:

$$\alpha = \frac{p(\mathbf{z}|\boldsymbol{\beta}^{\text{new}})p(\boldsymbol{\beta}^{\text{new}})q(\boldsymbol{\beta}^{\text{old}}|\boldsymbol{\beta}^{\text{new}})}{p(\mathbf{z}|\boldsymbol{\beta}^{\text{old}})p(\boldsymbol{\beta}^{\text{old}})q(\boldsymbol{\beta}^{\text{new}}|\boldsymbol{\beta}^{\text{old}})} = \prod_{i=1}^{N} \frac{f_{\mathcal{LO}_m}(\mathbf{z}_i - \mathbf{X}_i\boldsymbol{\beta}^{\text{new}})\tilde{p}(\mathbf{z}_i - \mathbf{X}_i\boldsymbol{\beta}^{\text{old}})}{\tilde{p}(\mathbf{z}_i - \mathbf{X}_i\boldsymbol{\beta}^{\text{new}})f_{\mathcal{LO}_m}(\mathbf{z}_i - \mathbf{X}_i\boldsymbol{\beta}^{\text{old}})}.$$

Hence, if $\tilde{p}(\boldsymbol{\varepsilon})$ is a good approximation to $f_{\mathcal{LO}_m}(\boldsymbol{\varepsilon})$ over a wide range of $\boldsymbol{\varepsilon}$, then the acceptance rate will be close 1. Subsequently, we consider several error distributions $\tilde{p}(\boldsymbol{\varepsilon})$, obtained by approximating $f_{\mathcal{LO}_m}(\boldsymbol{\varepsilon})$ in various ways.

Finally, note that $\boldsymbol{\lambda}$ is an auxiliary variable sampled only in order to construct the proposal. Since the utilities are sampled from $\mathbf{z}|\boldsymbol{\beta}$ using the exact dRUM model, the latent variables $\boldsymbol{\lambda}$ may not be stored and used in any subsequent MCMC sweep, see van Dyk and Park (2008) for a theoretical justification.

3.2.2 Using a Multivariate Student-*t* Distribution

Several authors (Albert and Chib, 1993; Liu, 2004) approximate the binary logit model by a binary discrete choice models based on the cdf of a univariate t_ν-distribution with ν in the range of 7 to 8, because the cdfs are very similar over a wide range. Since all univariate marginal distributions of the multivariate logistic distribution are logistic distributions, this suggests to approximate the multivariate logistic distribution by a multivariate Student $t_\nu(\mathbf{0}, \boldsymbol{\Sigma})$-distribution. As the multivariate logistic distribution is invariant to permuting the elements of $\boldsymbol{\varepsilon}$, $\boldsymbol{\Sigma}$ has to be a uniform covariance matrix: $\boldsymbol{\Sigma} = \sigma^2((1-\rho)\mathbf{I}_m + \rho\,\mathbf{e}_m\mathbf{e}_m')$.

The multivariate t_ν-distribution is a scale mixture of normal distribution, i.e., $\boldsymbol{\varepsilon}_i|\lambda_i \sim \mathcal{N}_m(\mathbf{0}, \boldsymbol{\Sigma}/\lambda_i)$ with $\lambda_i \sim \mathcal{G}(\nu/2, \nu/2)$ being a scale variable taking values in $\Re^+$. Hence, $\mathbf{m}_i = \mathbf{0}$ and $\mathbf{R}_i = \boldsymbol{\Sigma}/\lambda_i$ in the approximate model (10). The conditional posterior $q(\lambda_i|\boldsymbol{\beta}, \mathbf{z}_i)$ is given by:

$$\lambda_i|\boldsymbol{\beta}, \mathbf{z}_i \sim \mathcal{G}\left((\nu + m)/2, (\nu + (\mathbf{z}_i - \mathbf{X}_i\boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}(\mathbf{z}_i - \mathbf{X}_i\boldsymbol{\beta}))/2\right).$$

Table 1: Parameters of the optimal multivariate Student-t distribution.

m	2	3	4	5	6
ν	8.3930	9.0449	10.0821	10.9011	11.8589
σ^2	2.5060	2.5746	2.6515	2.6968	2.7463
ρ	0.5032	0.5041	0.5044	0.5061	0.5051
d_{KL}	0.0421	0.1020	0.1719	0.2439	0.3219

It remains to choose ν, σ^2, and ρ. We have determined them by minimizing the KL-distance. The minimization was performed by the MATLAB implementation of the simplex algorithm according to Nelder and Mead (1965). The corresponding optimal parameters for $m = 2, \ldots, 6$ are reported in Table 1, along with the KL-distance to the target distribution.

3.2.3 Using Multivariate Finite Normal Mixture Distributions

In the univariate case the approximation can be made accurate enough to have an acceptance rate of virtually 1. Hence, a Gibbs-type sampler can be run without a rejection step. In the multivariate case a finite mixture approximation has to be found in m variates. Due to the curse of dimensionality, it is not to be expected that an MH-step can be implemented without a rejection step.

The density of the multivariate finite normal mixture approximation reads

$$\tilde{p}(\boldsymbol{\varepsilon}_i) = \sum_{k=1}^{K} w_k f_N(\boldsymbol{\varepsilon}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\,.$$

The corresponding latent variable representation involves the discrete random variable $\lambda_i \sim \text{MulNom}\,(w_1, \ldots, w_K)$ taking values in the set $\{1, \ldots, K\}$ and $\boldsymbol{\varepsilon}_i|\lambda_i \sim \mathcal{N}_m\left(\boldsymbol{\mu}_{\lambda_i}, \boldsymbol{\Sigma}_{\lambda_i}\right)$. Hence, $\mathbf{m}_i = \boldsymbol{\mu}_{\lambda_i}$ and $\mathbf{R}_i = \boldsymbol{\Sigma}_{\lambda_i}$ in the approximate model (10). The conditional posterior $q(\lambda_i|\boldsymbol{\beta}, \mathbf{z}_i)$ is given by $\lambda_i \sim \text{MulNom}\,(p_{i1}, \ldots, p_{iK})$, where

$$p_{ik} \propto w_k f_N(\mathbf{z}_i - \mathbf{X}_i\boldsymbol{\beta}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\,,$$

and $\sum_{k=1}^{K} p_{ik} = 1$ for all $i = 1, \ldots, N$.

Finding an approximation with acceptance rate close to 1 is difficult. However, the multivariate logistic distribution has a rigid structure, which allows to impose restrictions on the means and covariance matrices of the mixture components. In particular, the multivariate logistic distribution in m variates is invariant under a permutation of the variates and therefore has m-fold symmetry with respect to the axis vector $\mathbf{a} = (1, \ldots, 1)$. Consequently, each component k of the mixture consists of m copies. Each copy has the same weight w_k/m and the same covariance matrix $\sigma_k^2\,\mathbf{R}_k$, with

$$\mathbf{R}_k = \begin{pmatrix} 1 & \rho_k & \cdots & \rho_k \\ \rho_k & 1 & \cdots & \rho_k \\ \vdots & \vdots & \ddots & \vdots \\ \rho_k & \rho_k & \cdots & 1 \end{pmatrix}\,.$$

The mean vectors of the m copies are arranged symmetrically according to:

$$\mathbf{m}_{k,i} = \lambda_k \mathbf{a} + \mu_k \mathbf{b}_i \,, \qquad i = 1, \ldots, m \,,$$

with $\mathbf{b}_i = (b_{i,1}, \ldots, b_{i,m})'$, where

$$b_{i,j} = \left(\delta_{ij} \sqrt{m} - \frac{1}{\sqrt{m}} \right) \,, \qquad j = 1, \ldots, m \,.$$

It is easy to see that every $\mathbf{b}_i$ is orthogonal to $\mathbf{a}$ and that the angle φ between any two $\mathbf{b}_i$ is equal to

$$\varphi = \arccos\left(-\frac{1}{m-1} \right) \,.$$

The set of all vectors $\mathbf{b}_i$ is invariant under a permutation of the variates, and so are therefore the mean vectors $\mathbf{m}_{k,i}$.

The length of vector $\mathbf{a}$ is equal to $\sqrt{m}$. The vectors $\mathbf{b}_i$ are confined to a subspace of dimension $m-1$, therefore they are scaled to a length of $\sqrt{m-1}$. With this convention we expect the coefficients λ_k and μ_k to stabilize for increasing values of m. This is borne out by our results.

Each mixture component is parameterized by its weight w_k, its variance σ_k^2, its correlation ρ_k, and the two coefficients λ_k and μ_k. The total number of parameters to be estimated is therefore 5 times the number of components, irrespective of the dimension m.

We have computed the approximating mixtures for $m = 2, \ldots, 12$ and $K = 1, \ldots, 6$, by minimizing the KL-distance. The integral was computed by averaging over a sample of up to 250,000 simulated data points. As an example, Figure 1 shows the contour lines of the bivariate logistic distribution and the approximating normal mixture with five components ($m = 2$, $K = 5$). The univariate marginal distributions are also shown.

Figure 2 shows the KL-distances as a function of m and K. Note that for $m > 12$ the parameters of the mixtures for $m = 12$ have been used. We find that the KL-distance rapidly increases as we move from the bivariate to higher-dimensional distributions.

Figure 3 shows the development of the coefficients λ_k and μ_k for $K = 5$. Above $m = 10$ both sets of coefficients are fairly stable, as are the other parameters.

3.2.4 Using Multivariate Finite Student-t Mixture Distributions

Finally, we consider the density of the finite multivariate Student-t mixture as the approximate error distribution:

$$\tilde{p}(\boldsymbol{\varepsilon}_i) = \sum_{k=1}^{K} w_k f_{St}(\boldsymbol{\varepsilon}_i; \nu_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \,.$$

The corresponding latent variable representation involves a bivariate random variable $\boldsymbol{\lambda}_i = (\lambda_{1i}, \lambda_{2i})$, where $\lambda_{1i} \sim \text{MulNom}\,(w_1, \ldots, w_K)$ is a discrete random variable taking values in the set $\{1, \ldots, K\}$ and $\lambda_{2i} | \lambda_{1i} \sim \mathcal{G}\,(\nu_{\lambda_{1i}}/2, \nu_{\lambda_{1i}}/2)$ is a scale variable taking values in $\Re^+$. The conditional distribution of $\boldsymbol{\varepsilon}_i$ given $\boldsymbol{\lambda}_i$ reads $\boldsymbol{\varepsilon}_i | \boldsymbol{\lambda}_i \sim \mathcal{N}_m\left(\boldsymbol{\mu}_{\lambda_{1i}}, \boldsymbol{\Sigma}_{\lambda_{1i}}/\lambda_{2i}\right)$, hence $\mathbf{m}_i = \boldsymbol{\mu}_{\lambda_{1i}}$ and $\mathbf{R}_i = \boldsymbol{\Sigma}_{\lambda_{1i}}/\lambda_{2i}$ in the approximate model (10).

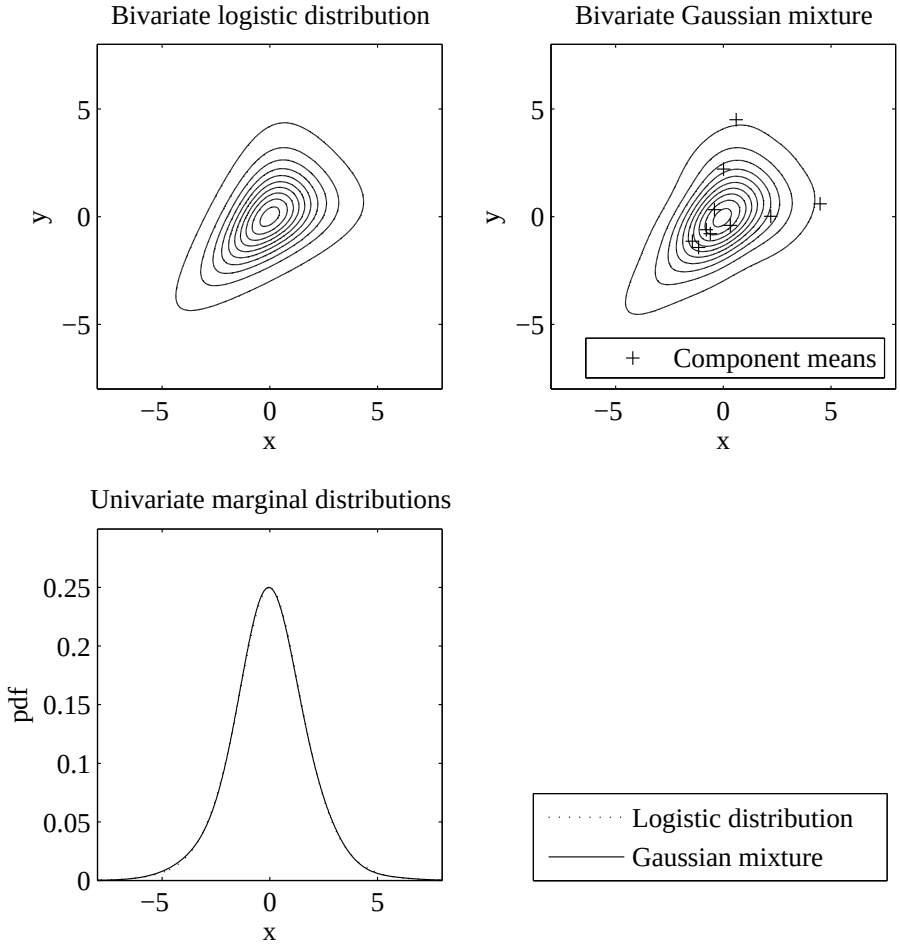

Figure 1: Top: Contour lines of the bivariate logistic distribution and the approximating Gaussian mixture with five components ($m = 2$, $K = 5$). The corresponding contour lines are at the same height. The plus signs indicate the locations of the ten component means. Bottom: the univariate marginal distributions. They are virtually identical.

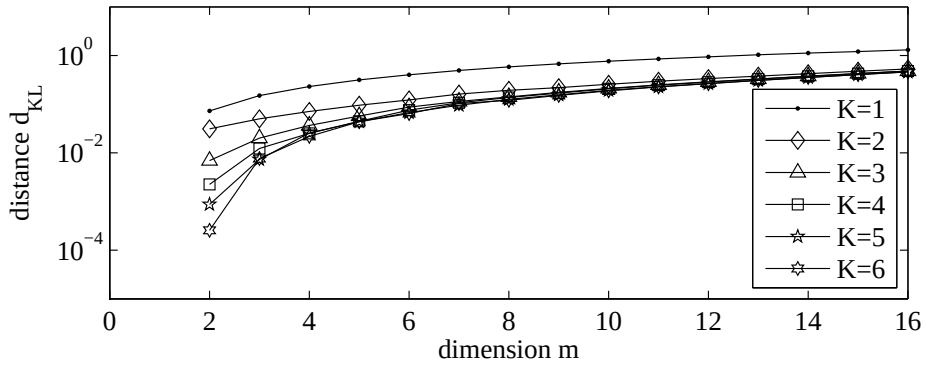

Figure 2: The KL-distance between the Gaussian mixtures and the multivariate logistic distribution as a function of the dimension m and the number K of mixture components.

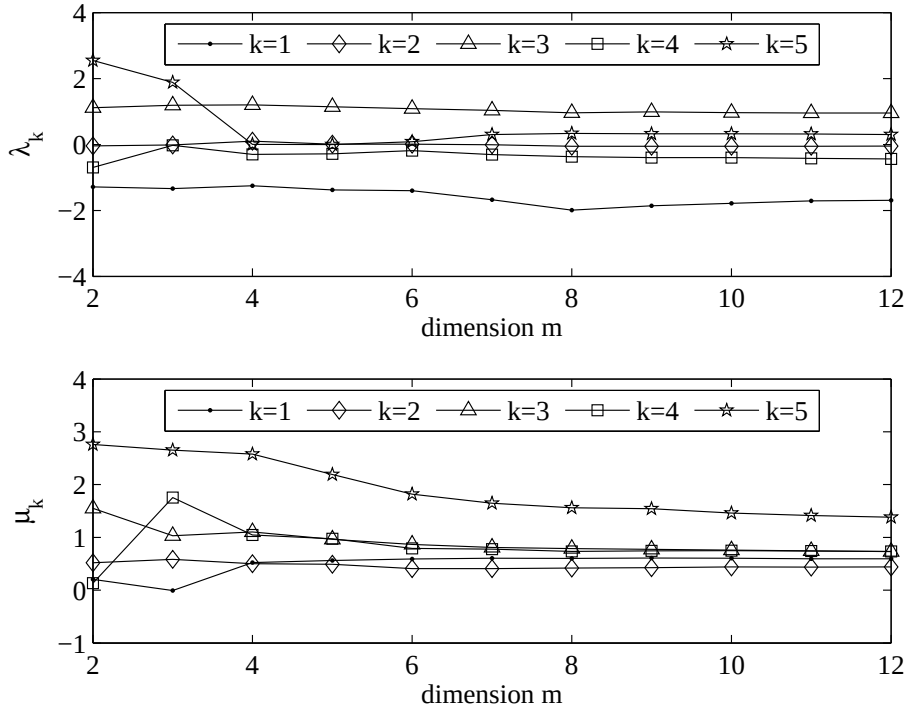

Figure 3: Top: the coefficients λ_k for $K = 5$, as a function of the dimension m. Bottom: the coefficients μ_k for $K = 5$, as a function of the dimension m.

The conditional posterior $q(\boldsymbol{\lambda}_i|\boldsymbol{\beta}, \mathbf{z}_i) = q(\lambda_{1i}|\boldsymbol{\beta}, \mathbf{z}_i)q(\lambda_{2i}|\lambda_{1i}\boldsymbol{\beta}, \mathbf{z}_i)$, where λ_{1i} is sampled marginally from $\lambda_{1i} \sim$ MulNom $(p_{i1}, \ldots, p_{iK})$ with

$$p_{ik} \propto w_k f_{St}(\boldsymbol{\varepsilon}_i; \nu_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\,,$$

and $\sum_{k=1}^{K} p_{ik} = 1$ for all $i = 1, \ldots, N$, and $q(\lambda_{2i}|\lambda_{1i}\boldsymbol{\beta}, \mathbf{z}_i)$ is given by:

$$\lambda_{2i}|\lambda_{1i}, \boldsymbol{\beta}, \mathbf{z}_i \sim \mathcal{G}\left((\nu_{\lambda_{1i}}+m)/2, (\nu_{\lambda_{1i}}+(\mathbf{z}_i-\mathbf{X}_i\boldsymbol{\beta}-\boldsymbol{\mu}_{\lambda_{1i}})'\boldsymbol{\Sigma}_{\lambda_{1i}}^{-1}(\mathbf{z}_i-\mathbf{X}_i\boldsymbol{\beta}-\boldsymbol{\mu}_{\lambda_{1i}}))/2\right).$$

The approximating Student-t mixtures were obtained in the same way as the Gaussian mixtures, with the number of degrees of freedom as an additional parameter, which was assumed to be the same for all components. Figure 4 shows the KL-distances between the target distribution and the mixtures, as a function of the dimension m and the number K of mixture components. Again, for $m > 12$ the parameters of the mixtures for $m = 12$ have been used.

4 Illustrative Applications

For all examples, we take an independent standard normal prior for each regression coefficient, i.e., $\mathbf{b}_0 = \mathbf{0}$ and $\mathbf{B}_0 = \mathbf{I}_r$. We use each of the four MH methods presented above to produce $M = 10000$ draws from the posterior distribution after running burn-in for 2000 iterations. For any proposals based on a finite mixture approximation, we increase the number of mixture components K from 2 to 6.

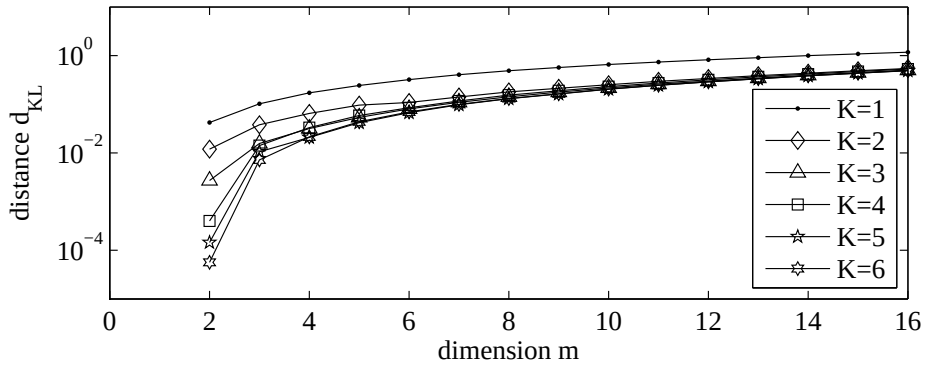

Figure 4: The KL-distance between the Student-t mixtures and the multivariate logistic distribution as a function of the dimension m and the number K of mixture components.

4.1 Simulated Data Sets

To evaluate the acceptance rate of the four MH samplers with increasing dimension, we consider a simple example, namely N i.i.d. observations $y_1, \ldots, y_N$ with $m+1$ categories, drawn from $\Pr(y_i = k|\boldsymbol{\beta}) = \pi_k = \exp(\beta_k)/(1 + \exp(\beta_k))$ for $k = 1, \ldots, m$ and $\Pr(y_i = 0|\boldsymbol{\beta}) = \pi_0 = 1/(1 + \exp(\beta_k))$. $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)$ is therefore a vector of dimension m. The latent equation in the corresponding dRUM model reads for $i = 1, \ldots, N$

$$\mathbf{z}_i = \mathbf{I}_m \boldsymbol{\beta} + \tilde{\boldsymbol{\varepsilon}}_i \,. \tag{11}$$

Note that the mixture approximation is applied to equation (11) not only once, but N times.

We made various comparisons with respect to the true parameter $\boldsymbol{\beta}$ and investigated both balanced distributions, where the π_k are roughly the same, and unbalanced distributions, where some of the π_k are very small. We found, however, that the acceptance rate of the samplers were insensitive to these change. Hence, we decided to assume that $\pi_0, \ldots, \pi_m$ is a uniform distribution over the $(m + 1)$ categories.

To evaluate the quality of the various mixture approximations, we start with a single observation, i.e., $N = 1$. Note that the proper prior distribution $p(\boldsymbol{\beta})$ guarantees that the posterior $p(\boldsymbol{\beta}|y_1)$ is proper although the dimension of $\boldsymbol{\beta}$ is larger than $N = 1$. Table 2 shows the acceptance rate as a function of the number K of components ($K = 1, \ldots, 6$) and the dimension m ($m = 2, \ldots, 6$). We see the expected behavior, as the acceptance rate drops with increasing m and rises with increasing K. For $N = 100$ observations the acceptance rate drops by about 15 to 20 percentage points, and the effect is more pronounced for small K. If N is increased to 1000, the acceptance rate drops only slightly, by a couple of percentage points in the worst case.

4.2 Real Data Sets

Furthermore, we consider two real data sets. First, the car data (Scott, 2011), which is a medium sized data set ($N = 263$) with 3 categories and 4 regressors, i.e. $r = 8$; second,

Table 2: Acceptance rate of the various mixture approximations for a single observation ($N = 1$) as well as for $N = 100$ and $N = 1000$ observations; K varies between 2 and 6; the dimension m increases from 2 to 6.

	K					K				
	2	3	4	5	6	2	3	4	5	6
$N = 1$	Normal					Student-t				
$m = 2$	92.7	95.1	97.7	98.4	98.4	94.5	96.9	98.5	98.4	98.9
3	89.9	92.4	94.1	95.9	95.8	90.7	93.1	93.7	94.6	95.5
4	87.4	89.1	90.8	90.5	92.1	86.7	89.4	89.7	92	91.6
5	84.7	86.1	88.2	87.6	88.5	82.9	86.4	86.4	87.2	87.9
6	81.8	83.2	84.1	84.4	84.6	81.2	83.2	82.9	84.8	83.5
$N = 100$	Normal					Student-t				
$m = 2$	79.4	88.7	92.3	94.1	94.7	88.4	92.3	95.6	94.8	96
3	75.1	82.2	85.6	88.8	88.8	79.8	84.8	85.2	87.8	89.3
4	69.5	76.2	80.7	80.6	81.6	71.2	78.1	77.6	81.7	81.5
5	64.3	69.7	72.8	72.9	73.1	63.6	71.4	69.5	75.1	73.5
6	57.8	63.4	64.7	67.1	67.8	61.9	65.6	64.7	67.4	67.6
$N = 1000$	Normal					Student-t				
$m = 2$	77.6	85.3	88	89.7	87.6	84.6	85.5	87.9	87.9	90.7
3	74.3	81.4	84.5	87.5	87.9	77.9	83.4	83.2	86.2	86.6
4	68.3	75.8	79.5	77.4	78.6	70.7	77.1	76.5	82.2	80.7
5	63.1	69.9	72.7	68.6	66.5	63.4	71	69.6	74.3	72.2
6	57.3	58.1	61.9	59.7	67.4	61.2	65	63.5	67.1	66.6

the Caesarean birth data (Fahrmeir and Tutz, 2001, Table 1.1), where $N = 251$, the outcome variable has 3 categories, and a model with 8 regressors is fitted, i.e. $r = 16$.

Table 3 and Table 4 evaluate and compare the various MH samplers using common measures such as the average acceptance rate after burn-in (in percent, Acc) and runtime in terms of the CPU time T_{CPU} after burn-in (in seconds, CPU). In addition, for each regression coefficient β_k, $k = 1, \ldots, r$, the inefficiency factor $\tau = 1 + 2 \cdot \sum_{h=1}^{H} \rho(h)$ is computed, where $\rho(h)$ denotes the empirical autocorrelation of the MCMC draws at lag h, and H is determined by the initial monotone sequence estimator (Geyer, 1992), the effective sampling size (Kass, Carlin, Gelman, and Neal, 1998), defined by ESS=M/τ, as well as the effective sampling rate defined as the ratio ESS/T_{CPU}. Table 3 and Table 4 report the median inefficiency factor (Ineff) and the median effective sample rate (ESR) over all regression coefficients.

In both examples, the acceptance rate rises with increasing number K of components, and the inefficiency factor drops. These effects, however, are not strong enough to compensate for the rise of the computational load with increasing K, so that we observe no net gain in terms of the effective sample rate, and even a net loss in three out of the four cases studied.

Table 3: Evaluating the various MH-algorithms for the car data.

	Normal mixtures				Student-t mixtures			
K	Acc	Ineff	ESR	T_{CPU}	Acc	Ineff	ESR	T_{CPU}
1	39.4	10.4	74.4	12.9	55.9	8.76	80.2	14.2
2	59.2	9.18	47.2	23.1	75.3	10.4	36.4	26.5
3	77.1	6.92	55.9	25.8	84.6	8.17	41.5	29.5
4	84.1	7.32	48.8	28	89.1	7.03	43.3	32.9
5	88.5	8.02	41.3	30.2	90.2	8.28	33.8	35.7
6	89.6	7.89	40.1	31.6	91.3	8.8	30.5	37.3

Table 4: Evaluating the various MH-algorithms for the Caesarean birth data.

	Normal mixtures				Student-t mixtures			
K	Acc	Ineff	ESR	T_{CPU}	Acc	Ineff	ESR	T_{CPU}
1	31.4	11	66.2	13.7	48.7	7.7	89.6	14.5
2	51.5	8.09	51.8	23.9	70.5	8.16	46.5	26.3
3	72.4	6.14	60.2	27	82.3	6.47	52.5	29.4
4	81.4	5.69	65.3	26.9	88	5.2	61.6	31.2
5	85.8	6.14	55.5	29.3	88.5	6.22	46.6	34.5
6	88.6	5.49	58	31.4	91.3	6.06	46	35.9

5 Concluding Remarks

We have shown how to construct data-augmented Metropolis-Hastings samplers for the general multinomial logistic model. The data augmentation relies on two mixture approximations to the multivariate logistic error distribution, which is characteristic for the dRUM representation of the model. We have studied the corresponding MH samplers on simulated and on real data sets. The results show that the sampling scheme is sound, but that the approximations are not yet precise enough to yield an acceptance rate and an inefficiency factor that more than offsets the rise in the computational load inherent to the evaluation of the multivariate mixtures.

References

Albert, J. H., and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, *88*, 669–679.

Alspach, D. L., and Sorenson, H. W. (1972). Nonlinear Bayesian estimation using Gaussian sum approximations. *IEEE Transactions on Automatic Control*, *17*, 439–448.

Balakrishnan, N. (1992). *Handbook of the Logistic Distribution*. New York: Marcel Dekker.

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

Fahrmeir, L., and Tutz, G. (2001). *Multivariate Statistical Modelling based on Generalized Linear Models* (2nd ed.). New York/Berlin/Heidelberg: Springer.

Fox, J. (2010). *Bayesian Item Response Modeling*. New York: Springer.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. New York: Springer.

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

Frühwirth-Schnatter, S., and Frühwirth, R. (2010). Data augmentation and MCMC for binary and multinomial logit models. In T. Kneib and G. Tutz (Eds.), *Statistical Modelling and Regression Structures – Festschrift in Honour of Ludwig Fahrmeir* (pp. 111–132). Heidelberg: Physica-Verlag. (Also available at `http://www.ifas.jku.at/ifas/content/e114480`, IFAS Research Paper Series 2010-48)

Frühwirth-Schnatter, S., Frühwirth, R., Held, L., and Rue, H. (2009). Improved auxiliary mixture sampling for hierarchical models of non-Gaussian data. *Statistics and Computing*, *19*, 479-492.

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

Gamerman, D., and Lopes, H. F. (2006). *Markov Chain Monte Carlo. Stochastic Simulation for Bayesian Inference* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.

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

Gumbel, E. J. (1961). Bivariate logistic distributions. *Journal of the American Statistical Association*, *56*, 335–349.

Guttman, I., Dutter, R., and Freeman, P. R. (1978). Care and handling of univariate outliers in the general linear model to detect spuriosity — A Bayesian approach. *Technometrics*, *20*, 187–193.

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

Imai, K., and van Dyk, D. A. (2005). A Bayesian analysis of the multinomial probit model using marginal data augmentation. *Journal of Econometrics*, *124*, 311–334.

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.

Kotz, S., Johnson, N. L., and Balakrishnan, N. (2000). *Continous Multivariate Distributions: Models and Applications*. Wiley.

Liu, C. (2004). Robit regression: a simple robust alternative to logistic and probit regression. In A. Gelman and X.-L. Meng (Eds.), *Applied Bayesian Modeling and Casual Inference from Incomplete-Data Perspectives* (pp. 227–238). Chichester: Wiley.

Malik, H. J., and Abraham, B. (1973). Multivariate logistic distributions. *The Annals of Statistics*, *1*, 588–590.

McCulloch, R. E., Polson, N. G., and Rossi, P. E. (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.

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

Newcomb, S. (1886). A generalized theory of the combination of observations so as to obtain the best result. *American Journal of Mathematics*, *8*, 343–366.

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

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

Scott, S. L. (2011). Data augmentation, frequentist estimation, and the Bayesian analysis of multinomial logit models. *Statistical Papers*, *52*, 87–109.

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

Sorenson, H. W., and Alspach, D. L. (1971). Recursive Bayesian estimation using Gaussian sums. *Automatica*, *6*, 465–479.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. New York: Wiley.

van Dyk, D. A., and Park, T. (2008). Partially collapsed Gibbs samplers: Theory and methods. *Journal of the American Statistical Association*, *103*, 790–796.

Authors' addresses:

Sylvia Frühwirth-Schnatter
Institute for Statistics and Mathematics
Vienna University of Economics and Business
Augasse 2-6
1090 Wien
Austria

Rudolf Frühwirth
Institute of High Energy Physics
Austrian Academy of Sciences
Nikolsdorfer Gasse 18
1050 Wien
Austria