Department for Applied Statistics
Johannes Kepler University Linz

IFAS Research Paper Series
2010-48

Data Augmentation and MCMC for Binary and Multinomial Logit Models

Sylvia Frühwirth-Schnatter and Rudolf Frühwirth[a]

January 2010

[a]Institut für Hochenergiephysik der Österreichischen Akademie der Wissenschaften, Wien, Austria

Published as:

Frühwirth-Schnatter, S. and Frühwirth, R. (2010): Data augmentation and MCMC for binary and multinomial logit models. In Kneib, T. and Tutz, G. (Eds.): *Statistical Modelling and Regression Structures – Festschrift in Honour of Ludwig Fahrmeir*, pp. 111–132, Heidelberg: Physica-Verlag.

Abstract

The paper introduces two new data augmentation algorithms for sampling the parameters of a binary or multinomial logit model from their posterior distribution within a Bayesian framework. The new samplers are based on rewriting the underlying random utility model in such a way that only differences of utilities are involved. As a conseqence, the error term in the logit model has a logistic distribution. If the logistic distribution is approximated by a finite scale mixture of normal distributions, auxiliary mixture sampling can be implemented to sample from the posterior of the regression parameters. Alternatively, a data augmented Metropolis–Hastings algorithm can be formulated by approximating the logistic distribution by a single normal distribution. A comparative study on five binomial and multinomial data sets shows that the new samplers are superior to other data augmentation samplers and to Metropolis–Hastings sampling without data augmentation.

Keywords: Binomial data, multinomial data, data augmentation, Markov chain Monte Carlo, logit model, random utility model

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.

Estimation of these models is quite challenging, in particular if latent components are present, such as in random-effects modelling or in state space modelling of discrete data. Fahrmeir and Tutz (2001) provide a review of likelihood-based estimation methods; see also Fahrmeir and Kaufmann (1986a) and Fahrmeir and Kaufmann (1986b) for a rigorous mathematical treatment.

Zellner and Rossi (1984) were the first to perform Bayesian inference for a logit model using importance sampling based on a multivariate Student-t distribution, with mean and scale matrix being equal to the posterior mode and the asymptotic covariance matrix. Starting with Zeger and Karim (1991), many Markov chain Monte Carlo (MCMC) methods have been developed for the Bayesian estimation of the binary and the multinomial logit model. MCMC estimation has been based on single-move adaptive rejection Gibbs sampling Dellaportas and Smith (1993), Metropolis–Hastings

(MH) sampling Gamerman (1997); Lenk and DeSarbo (2000); Rossi et al. (2005), data augmentation and Gibbs sampling Holmes and Held (2006); Frühwirth-Schnatter and Frühwirth (2007), and data augmented Metropolis–Hastings sampling Scott (2009).

In the present article we focus on practical Bayesian inference for binary and multinomial logit models using data augmentation methods. For these models, data augmentation relies on the interpretation of the logit model as a random utility model McFadden (1974). Frühwirth-Schnatter and Frühwirth (2007) and Scott (2009) base data augmentation directly on this random utility model (RUM) by introducing the utilities as latent variables. Holmes and Held (2006) choose the differences of utilities as latent variables, which is the standard data augmentation method underlying MCMC estimation of probit models, see e.g. Albert and Chib (1993) and McCulloch et al. (2000). We call this interpretation the difference random utility model (dRUM).

In the following we show how to implement data augmentation based on the dRUM representation for the binary and the multinomial logit model. We introduce yet two other data augmentation MCMC samplers by extending the ideas underlying Frühwirth-Schnatter and Frühwirth (2007) and Scott (2009) to the dRUM representation. The extension of the data augmented MH algorithm of Scott (2009) is straightforward, while the extension of the auxiliary mixture sampling approach of Frühwirth-Schnatter and Frühwirth (2007) involves approximating the logistic distribution by a finite scale mixture of normal distributions Monahan and Stefanski (1992).

We compare the two new data augmentation samplers with the three existing ones for several well-known case studies. This exercise reveals that data augmentation samplers based on the dRUM representation are considerably more efficient in terms of reducing autocorrelation in the resulting MCMC draws than data augmentation based on the RUM. Under the dRUM representation, both auxiliary mixture sampling and data augmented MH sampling are considerably faster than the sampler suggested by Holmes and Held (2006), making the two new samplers an attractive alternative to other data augmentation methods.

Since it is often believed that MCMC sampling without data augmentation can be even more efficient than MCMC sampling with data augmentation, we include several MH algorithms into our comparison, namely the independence MH sampler suggested in Rossi et al. (2005), a multivariate random walk MH with asymptotically optimal scaling chosen as in Roberts and Rosenthal (2001), and the DAFE-R MH algorithm suggested by Scott (2009). While the independence MH sampler of Rossi et al. (2005) turns out to be superior to any other MH sampler without data augmentation, we find for all but one (very well-behaved) case study that our two new dRUM data augmentation samplers are superior to the independence MH sampler both in terms of efficiency and in terms of the effective sampling rate.

2 MCMC Estimation Based on Data Augmentation for Binary Logit Regression Models

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

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

where $\mathbf{x}_i$ is a row vector of regressors, including 1 for the intercept, and $\boldsymbol{\beta}$ is an unknown regression parameter of dimension d. 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 $\boldsymbol{\beta}$ is a normal distribution, $\mathrm{No}_d(\mathbf{b}_0, \mathbf{B}_0)$ with known hyperparameters $\mathbf{b}_0$ and $\mathbf{B}_0$. The posterior density $p(\boldsymbol{\beta}|\mathbf{y})$ of $\boldsymbol{\beta}$ given all observations $\mathbf{y} = (y_1, \ldots, y_N)$ does not have a closed form:

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

Hence Bayesian estimation relies either on data augmentation, to be discussed in this section, or on MH sampling, as in Section 4.

2.1 Writing the Logit Model as a Random Utility Model

The interpretation of a logit model as a random utility (RUM) model was introduced by McFadden (1974). Two representations of the logit model as a RUM are common.

Let y^u_{ki} be the utility of choosing category k, which is assumed to depend on covariates $\mathbf{x}_i$. The RUM representation corresponding to the logit model reads:

$$\begin{aligned} y^u_{0i} &= \mathbf{x}_i\boldsymbol{\beta}_0 + \delta_{0i}, \qquad \delta_{0i} \sim \mathrm{EV}, && (2)\\ y^u_{1i} &= \mathbf{x}_i\boldsymbol{\beta}_1 + \delta_{1i}, \qquad \delta_{1i} \sim \mathrm{EV}, && (3)\\ y_i &= I\{y^u_{1i} > y^u_{0i}\}, \end{aligned}$$

where $I\{\cdot\}$ is the indicator function and δ_{0i} and δ_{1i} are i.i.d. random variables following a type I extreme value (EV) distribution with density:

$$f_{\mathrm{EV}}(\delta) = \exp\left(-\delta - e^{-\delta}\right), \tag{4}$$

with expectation $\mathbb{E}(\delta) = \gamma$ and variance $\mathbb{V}(\delta) = \pi^2/6$, where $\gamma = 0.5772$ is Euler's constant.

Thus category 1 is observed, i.e. $y_i = 1$, iff $y^u_{1i} > y^u_{0i}$; otherwise $y_i = 0$. To achieve identifiability, it is assumed that $\boldsymbol{\beta}_0 = \mathbf{0}$, i.e. $\boldsymbol{\beta} = \boldsymbol{\beta}_1$, because only the difference $\boldsymbol{\beta} = \boldsymbol{\beta}_1 - \boldsymbol{\beta}_0$ can be identified.

An alternative way to write the logit model as an augmented model involving random utilities is the difference random utility model (dRUM), which is obtained by choosing a baseline category, typically 0, and to consider the model involving the differences of the utilities:

$$\begin{aligned} z_i &= \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i, \qquad \epsilon_i \sim \mathrm{Lo}, && (5)\\ y_i &= I\{z_i > 0\}, \end{aligned}$$

where $z_i = y^u_{1i} - y^u_{0i}$. The error term $\epsilon_i = \delta_{1i} - \delta_{0i}$, being the difference of two i.i.d. EV random variables, follows a logistic (Lo) distribution, with density:

$$f_{\mathrm{Lo}}(\epsilon) = \frac{e^{\epsilon}}{(1 + e^{\epsilon})^2},$$

with $\mathbb{E}(\epsilon) = 0$ and $\mathbb{V}(\epsilon) = \pi^2/3$.

For both representations the binary logit regression model (1) results as the marginal distribution of y_i.

2.2 Data Augmentation Based on the Random Utility Model

Several data augmentation algorithms have been suggested for the logit model, all of which are based on the interpretation of a logit model as a random utility model. However, depending on whether the RUM or the dRUM is considered, different data augmentation algorithms result.

Frühwirth-Schnatter and Frühwirth (2007) and Scott (2009) consider the RUM representation (2) for data augmentation and introduce for each i, $i = 1, \ldots, N$, the latent utility of choosing category 1, i.e. $\mathbf{z} = (y^u_{11}, \ldots, y^u_{1N})$, as missing data. Holmes and Held (2006) use the dRUM representation (5) and introduce the differences in utilities, i.e. $\mathbf{z} = (z_1, \ldots, z_N)$, as missing data. For both representations, data augmentation leads to a two-step MCMC sampler which draws from the conditional densities $p(\mathbf{z}|\boldsymbol{\beta}, \mathbf{y})$ and $p(\boldsymbol{\beta}|\mathbf{z}, \mathbf{y})$, respectively.

For both representations it is possible to sample all components of $\mathbf{z}|\boldsymbol{\beta}, \mathbf{y}$ simultaneously in a simple manner. For the RUM this step reads:

$$y^u_{1i} = -\log\left(\mathrm{Ex}\left(1 + \lambda_i\right) + \mathrm{Ex}\left(\lambda_i\right)\left(1 - y_i\right)\right), \tag{6}$$

where $\mathrm{Ex}\left(\lambda\right)$ denotes a random variable from an exponential distribution with density equal to $\lambda \exp(-\lambda y)$. For the dRUM this step reads:

$$z_i = \log(\lambda_i U_i + y_i) - \log(1 - U_i + \lambda_i(1 - y_i)), \tag{7}$$

where $U_i \sim \mathrm{Un}\left[0, 1\right]$. In both cases $\lambda_i = \exp(\mathbf{x}_i\boldsymbol{\beta})$.

In contrast to sampling from $p(\mathbf{z}|\boldsymbol{\beta}, \mathbf{y})$, sampling from $p(\boldsymbol{\beta}|\mathbf{z}, \mathbf{y})$ is not possible in closed form, regardless of the underlying representation. Conditional on $\mathbf{z}$, the posterior of $\boldsymbol{\beta}$ is independent of $\mathbf{y}$ and can be derived from regression models (3) or (5), respectively, which are linear in $\boldsymbol{\beta}$, but have a non-normal error term. Various methods have been suggested to cope with this non-normality when sampling the regression parameter $\boldsymbol{\beta}$.

Scott (2009) uses an independence MH algorithm where a normal proposal distribution $\mathrm{No}_d(\mathbf{b}_N, \mathbf{B}_N)$ for $\boldsymbol{\beta}$ is constructed by approximating the non-normal error δ_{1i} appearing in (3) by a normal error with same mean and variance:

$$\mathbf{b}_N = \mathbf{B}_N\left(\mathbf{B}_0{}^{-1}\mathbf{b}_0 + \frac{6}{\pi^2}\mathbf{X}'(\mathbf{z} - \gamma)\right), \qquad \mathbf{B}_N = \left(\mathbf{B}_0{}^{-1} + \frac{6}{\pi^2}\mathbf{X}'\mathbf{X}\right)^{-1}, \tag{8}$$

where row i of the $(N \times d)$ matrix $\mathbf{X}$ is equal to the regressor $\mathbf{x}_i$ of the logit model (1). This leads to a very fast sampler, because $\mathbf{B}_N$ is fixed while running MCMC; however, the acceptance rate might be low in higher dimensional problems.

Frühwirth-Schnatter and Frühwirth (2007) approximate the density of the EV distribution in (3) by the density of a finite normal mixture distribution with 10 components with optimized, but fixed parameters (m_r, s^2_r, w_r) in component r:

$$y^u_{1i} = \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i, \quad \epsilon_i|r_i \sim \mathrm{No}\left(m_{r_i}, s^2_{r_i}\right), \quad r_i \sim \mathrm{MulNom}(w_1, \ldots, w_{10}). \tag{9}$$

To perform MCMC estimation they add the latent indicators $\mathbf{r} = (r_1, \ldots, r_N)$ as missing data. The advantage of this additional data augmentation is that conditional on $\mathbf{z}$ and $\mathbf{r}$, the regression parameter $\boldsymbol{\beta}$ may be sampled from regression model (9), leading

to a normal conditional posterior. To complete MCMC, each indicator r_i has to be sampled from the discrete posterior $r_i|z_i, \boldsymbol{\beta}$ which is a standard step in finite mixture modelling.

Holmes and Held (2006) represent the logistic distribution appearing in (5) as an infinite scale mixture of normals Andrews and Mallows (1974):

$$z_i = \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i, \quad \epsilon_i|\omega_i \sim \mathrm{No}\left(0, \omega_i\right), \quad \sqrt{\omega_i}/2 \sim \mathrm{KS}, \tag{10}$$

where KS is the Kolmogorov–Smirnov distribution. To perform MCMC estimation they add the latent scaling factors $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_N)$ as missing data. Conditional on $\mathbf{z}$ and $\boldsymbol{\omega}$, the regression parameter $\boldsymbol{\beta}$ is sampled from regression model (10), leading to a normal conditional posterior. To complete MCMC, each scaling factor ω_i has to be sampled from the posterior $\omega_i|\boldsymbol{\beta}, z_i$ which has no closed form, the density of the KS distribution having no closed form either, but only a representation involving an infinite series. To sample ω_i, Holmes and Held (2006) implement a single move rejection sampling method based on deriving upper and lower squeezing functions from a truncated series representation of the density of the KS distribution. However, as will be illustrated by the case studies in Section 5, this rejection sampling step makes the algorithm computationally intensive and therefore quite slow.

2.3 Two New Samplers Based on the dRUM Representation

The case studies to be discussed in Section 5 demonstrate a remarkable advantage of Holmes and Held (2006) compared to Frühwirth-Schnatter and Frühwirth (2007), namely that the autocorrelations of the MCMC draws are in general much smaller, making the sampler more efficient. This increase in efficiency turns out to be closely related to using the dRUM rather than the RUM representation of the logit model.

In this paper, we propose two new samplers based on the dRUM representation of the logit model. They are constructed by applying the ideas underlying Frühwirth-Schnatter and Frühwirth (2007) and Scott (2009). As will be illustrated by the case studies, these samplers are much faster than the approach of Holmes and Held (2006), while the efficiency is about the same. Both are much more efficient than the corresponding ones in the RUM representation.

To apply the ideas underlying Scott (2009) to the dRUM representation, we construct a proposal density for $\boldsymbol{\beta}$ by approximating the error term in (5) by a normal error with zero mean and variance equal to $\pi^2/3$. Because a logistic error is closer to the normal distribution than an error following the EV distribution, it is to be expected that the acceptance rate for the resulting independence MH algorithm is much higher than in the RUM model. This expectation is confirmed by our case studies. Details of this sampler are given in Algorithm 1.

Algorithm 1 *Independence Metropolis–Hastings algorithm in the dRUM representation of a logit model.*

Choose starting values for $\boldsymbol{\beta}$ and $\mathbf{z} = (z_1, \ldots, z_N)$ and repeat the following steps:

(a) Propose $\boldsymbol{\beta}^{\mathrm{new}}$ from the proposal $q(\boldsymbol{\beta}^{\mathrm{new}}|\mathbf{z}) = \mathrm{No}_d\left(\mathbf{b}_N, \mathbf{B}_N\right)$ with moments:

$$\mathbf{b}_N = \mathbf{B}_N\left(\mathbf{B}_0{}^{-1}\mathbf{b}_0 + \frac{3}{\pi^2}\mathbf{X}'\mathbf{z}\right), \qquad \mathbf{B}_N = \left(\mathbf{B}_0{}^{-1} + \frac{3}{\pi^2}\mathbf{X}'\mathbf{X}\right)^{-1}.$$

Accept $\boldsymbol{\beta}^{\text{new}}$ with probability $\min(\alpha, 1)$, where:

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

and $p(\mathbf{z}|\boldsymbol{\beta})$ is the likelihood of model (5):

$$p(\mathbf{z}|\boldsymbol{\beta}) = \prod_{i=1}^{N} f_{\text{Lo}}(z_i - \mathbf{x}_i\boldsymbol{\beta}).$$

(b) Sample from $z_i|\boldsymbol{\beta}, \mathbf{y}$ for $i = 1, \ldots, N$ as in (7). □

To apply the ideas underlying Frühwirth-Schnatter and Frühwirth (2007) to the dRUM representation, we approximate in (5) the density of the logistic distribution $f_{\text{Lo}}(\epsilon_i)$ by the density of a normal mixture distribution. As $f_{\text{Lo}}(\epsilon_i)$ is symmetric around 0, it is sensible to use a finite scale mixture of normal distributions with all component means being equal to 0. For a fixed number H of components this mixture is characterized by component specific variances s_r^2 and weights w_r :

$$f_{\text{Lo}}(\epsilon_i) \approx \sum_{r=1}^{H} w_r f_{\text{No}}(\epsilon_i; 0, s_r^2). \tag{11}$$

The contribution of Monahan and Stefanski (1992) to the handbook of the logistic distribution Balakrishnan (1992) contains such an approximation. As they use a different parameterization, the correct weights and variances in (11) are given by $w_r = p_r$ and $s_r^2 = 1/(s_r^\star)^2$, where p_r and $s_r^\star$ are the values published in Monahan and Stefanski (1992, Table 18.4.1). The corresponding parameters are reproduced in Table 1. We investigate the accuracy of this approximation as well as an alternative approximation in Subsection 2.4.

In general, we expect the number of components necessary to approximate the logistic distribution to be smaller than in Frühwirth-Schnatter and Frühwirth (2007), because the logistic distribution is much closer to the normal distribution than the EV distribution. In fact, the results in Subsection 2.4 show that the 3-component approximation of Monahan and Stefanski (1992) gives about the same acceptance rates as the 10-component approximation in the RUM representation, see Frühwirth-Schnatter and Frühwirth (2007, Table 2), while choosing $H = 6$ leads to an extremely accurate approximation. Thus we recommend choosing $H = 3$ in larger applications, where computing time matters, and to work with $H = 6$ whenever possible.

Having approximated the density of the logistic distribution by a scale mixture of H normal densities, we obtain a representation of the dRUM similar to (10), but ω_i is drawn with fixed probabilities $w_1, \ldots, w_H$ from the set $\{s_1^2, \ldots, s_H^2\}$:

$$\begin{aligned} z_i &= \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i, \qquad \epsilon_i|\omega_i \sim \text{No}\,(0, \omega_i)\,, \\ \omega_i &= s_{r_i}^2, \qquad r_i \sim \text{MulNom}(w_1, \ldots, w_H). \end{aligned} \tag{12}$$

Note that in this way we approximate the logit model by a very accurate finite scale mixture of probit models.

Like in Holmes and Held (2006), we add the scaling factors $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_N)$ as missing data. However, an advantage compared to Holmes and Held (2006) is that instead of sampling ω_i directly, we sample an indicator r_i from the discrete posterior $r_i|z_i, \boldsymbol{\beta}$, which can be done in a very efficient manner, and define $\omega_i = s^2_{r_i}$. Details of this sampler are given in Algorithm 2.

Algorithm 2 *Auxiliary mixture sampling in the dRUM representation of a logit model.*

Choose starting values for $\mathbf{z} = (z_1, \ldots, z_N)$ and $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_N)$ and repeat the following steps:

(a) Sample the regression coefficient $\boldsymbol{\beta}$ conditional on $\mathbf{z}$ and $\boldsymbol{\omega}$ based on the normal regression model (12) from $\mathrm{No}_d(\mathbf{b}_N, \mathbf{B}_N)$ with moments:

$$\mathbf{b}_N = \mathbf{B}_N \left({\mathbf{B}_0}^{-1} \mathbf{b}_0 + \sum_{i=1}^{n} {\mathbf{x}_i}' z_i / \omega_i \right), \qquad \mathbf{B}_N = \left({\mathbf{B}_0}^{-1} + \sum_{i=1}^{n} {\mathbf{x}_i}' \mathbf{x}_i / \omega_i \right)^{-1}.$$

(b) For $i = 1, \ldots, N$, sample from $z_i|\boldsymbol{\beta}, \mathbf{y}$ as in (7). Sample the indicator r_i conditional on z_i from the discrete density:

$$\Pr(r_i = j|z_i, \boldsymbol{\beta}) \propto \frac{w_j}{s_j} \exp\left[-\frac{1}{2} \left(\frac{z_i - \log \lambda_i}{s_j} \right)^2 \right],$$

and set $\omega_i = s^2_{r_i}$. The quantities $(w_j, s^2_j), j = 1, \ldots, H$ are the parameters of the H component finite mixture approximation tabulated in Table 1. □

2.4 Finite Mixture Approximations to the Logistic Distribution

Monahan and Stefanski (1992) obtained their finite scale mixture approximation by minimizing the KS-distance between the true and the approximate distribution function. The results are given in Table 1.

Because the approximation in Frühwirth-Schnatter and Frühwirth (2007) is based on minimizing the Kullback–Leibler distance between the densities, we redid a related analysis for the logistic distribution. The fitted components are reported in Table 2.

Similarly as in Frühwirth-Schnatter and Frühwirth (2007), we evaluate the effect of using different distance measures and different numbers of mixture components for a simple example, namely Bayesian inference for N i.i.d. binary observations $y_1, \ldots, y_N$, drawn with $\Pr(y_i = 1|\beta) = \pi = e^\beta/(1 + e^\beta)$.

First we run the data augmented MH algorithm as in Algorithm 2, which corresponds to approximating the logistic distribution by the single normal distribution $\mathrm{No}(0, \pi^2/3)$, i.e. $H = 1$. Then the data augmented MH algorithm is refined by proposing β from an approximate model, where the logistic distribution is approximated by a scale mixture of H normal distributions with H ranging from 2 to 6. Similarly as in Frühwirth-Schnatter and Frühwirth (2007), we use numerical integration methods to compute the corresponding expected acceptance rate for various values of π and N. Table 3 and Table 4 report, respectively, the expected acceptance rate for the mixture

Table 1: Approximation of the density of the logistic distribution by finite scale mixtures of normal distributions with H components, based on Monahan and Stefanski (1992).

	$H=2$		$H=3$		$H=4$		$H=5$		$H=6$	
r	s_r^2	$100w_r$	s_r^2	$100w_r$	s_r^2	$100w_r$	s_r^2	$100w_r$	s_r^2	$100w_r$
1	1.6927	56.442	1.2131	25.22	0.95529	10.65	0.79334	4.4333	0.68159	1.8446
2	5.2785	43.558	2.9955	58.523	2.048	45.836	1.5474	29.497	1.2419	17.268
3			7.5458	16.257	4.4298	37.419	3.012	42.981	2.2388	37.393
4					9.701	6.0951	5.9224	20.759	4.0724	31.697
5							11.77	2.3291	7.4371	10.89
6									13.772	0.90745

Table 2: Approximation of the density of the logistic distribution by finite scale mixtures of normal distributions with H components, based on minimizing the K-L distance.

	$H=2$		$H=3$		$H=4$		$H=5$		$H=6$	
r	s_r^2	$100w_r$	s_r^2	$100w_r$	s_r^2	$100w_r$	s_r^2	$100w_r$	s_r^2	$100w_r$
1	1.9658	68.966	1.4418	38.834	1.1509	20.638	0.95132	10.159	0.84678	5.8726
2	6.2324	31.034	3.7181	52.719	2.6072	52.008	1.9567	40.842	1.61	28.74
3			9.1139	8.4469	5.6748	25.032	3.8969	36.99	2.8904	36.756
4					11.884	2.3212	7.5025	11.233	5.0772	22.427
5							14.163	0.7753	8.9109	5.8701
6									15.923	0.33466

approximation based on Monahan and Stefanski (1992) and the mixture approximation based on the Kullback–Leibler distance.

As expected, by increasing the number of components the expected acceptance rate approaches 100% for both distances. The expected acceptance rates are rather similar for both distance measure; however, the approximations obtained by Monahan and Stefanski (1992) are slightly better than the approximations based on the Kullback–Leibler distance. Both approximations are already very good for H as small as 3 and are extremely accurate for $H = 6$.

Note that the mixture approximation is applied not only once, but N times. Both tables show how the approximation error accumulates with increasing N. Again, we find that the mixture approximations derived by Monahan and Stefanski (1992) are slightly more reliable in this respect than the mixture approximations based on the Kullback–Leibler distance.

Table 3: Expected acceptance rate in percent for a Metropolis–Hastings algorithm, based for $H = 1$ on the normal distribution $\mathrm{No}\,(0, \pi^2/3)$ and for $H > 1$ on the scale mixture approximations of Monahan and Stefanski (1992). N is the number of i.i.d. binary observations, and π is the probability of observing 1.

π	N	H					
		1	2	3	4	5	6
0.05	1	90.990	99.165	99.889	99.984	99.998	100.00
	10	89.508	98.628	99.778	99.961	99.994	99.999
	100	88.562	97.956	99.630	99.932	99.986	99.997
	1000	88.267	97.850	99.549	99.906	99.980	99.996
0.20	1	90.787	99.188	99.889	99.980	99.997	100.00
	10	88.491	98.408	99.740	99.957	99.992	99.999
	100	88.273	97.831	99.611	99.927	99.986	99.997
	1000	88.139	97.697	99.518	99.900	99.979	99.995
0.50	1	90.966	99.207	99.897	99.984	99.998	100.00
	10	88.748	98.321	99.724	99.950	99.991	99.998
	100	88.289	97.883	99.630	99.929	99.986	99.997
	1000	88.236	97.678	99.520	99.899	99.978	99.995

Table 4: Expected acceptance rate in percent for a Metropolis–Hastings algorithm, based on a mixture approximation with H components minimizing the Kullback–Leibler distance. N is the number of i.i.d. binary observations, and π is the probability of observing 1.

π	N	H				
		2	3	4	5	6
0.05	1	98.788	99.786	99.958	99.992	99.992
	10	97.996	99.580	99.908	99.981	99.988
	100	97.745	99.499	99.879	99.973	99.987
	1000	97.732	99.470	99.875	99.972	99.986
0.20	1	98.750	99.791	99.958	99.992	99.992
	10	97.909	99.548	99.903	99.979	99.988
	100	97.696	99.475	99.875	99.973	99.986
	1000	97.618	99.457	99.873	99.972	99.986
0.50	1	98.818	99.798	99.960	99.992	99.992
	10	97.846	99.534	99.896	99.979	99.988
	100	97.654	99.477	99.873	99.971	99.986
	1000	97.625	99.463	99.873	99.971	99.986

3 MCMC Estimation Based on 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. The categories are labeled by $L = \{0, \ldots, m\}$, and for any

k the set of all categories but k is denoted by $L_{-k} = L \setminus \{k\}$.

We assume that the observations are mutually independent and that for each $k \in L$ the probability of y_i taking the value k depends on covariates $\mathbf{x}_i$ in the following way:

$$\Pr(y_i = k|\boldsymbol{\beta}_0, \ldots, \boldsymbol{\beta}_m) = \pi_{ki}(\boldsymbol{\beta}_0, \ldots, \boldsymbol{\beta}_m) = \frac{\exp(\mathbf{x}_i\boldsymbol{\beta}_k)}{\sum_{l=0}^{m} \exp(\mathbf{x}_i\boldsymbol{\beta}_l)}, \tag{13}$$

where $\boldsymbol{\beta}_0, \ldots, \boldsymbol{\beta}_m$ are category specific unknown parameters of dimension d. To make the model identifiable, the parameter $\boldsymbol{\beta}_{k_0}$ of a baseline category k_0 is set equal to $\mathbf{0}$: $\boldsymbol{\beta}_{k_0} = \mathbf{0}$. Thus the parameter $\boldsymbol{\beta}_k$ is in terms of the change in log-odds relative to the baseline category k_0. In the following, we assume without loss of generality that $k_0 = 0$. To pursue a Bayesian approach, we assume that the prior distribution $p(\boldsymbol{\beta}_k)$ of each $\boldsymbol{\beta}_k$ is a normal distribution $\mathrm{No}_d(\mathbf{b}_0, \mathbf{B}_0)$ with known hyperparameters $\mathbf{b}_0$ and $\mathbf{B}_0$.

3.1 Data Augmentation in the RUM

As for the binary model, data augmentation is based on writing the multinomial logit model as a random utility model McFadden (1974):

$$y^u_{ki} = \mathbf{x}_i\boldsymbol{\beta}_k + \delta_{ki}, \quad k = 0, \ldots, m, \tag{14}$$

$$y_i = k \Leftrightarrow y^u_{ki} = \max_{l \in L} y^u_{li}. \tag{15}$$

Thus the observed category is equal to the category with maximal utility. If the random variables $\delta_{0i}, \ldots, \delta_{mi}$ appearing in (14) are i.i.d. following an EV distribution, then the multinomial logit model (13) results as the marginal distribution of y_i.

Frühwirth-Schnatter and Frühwirth (2007) and Scott (2009) use this RUM formulation of the multinomial logit model to carry out data augmentation based on introducing the latent utilities as missing data, i.e. $\mathbf{z} = ((y^u_{k1}, \ldots, y^u_{kN}), k = 1, \ldots, m)$. As for the binary RUM it is possible to sample the latent utilities $\mathbf{z}|\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m, \mathbf{y}$ simultaneously:

$$y^u_{ki} = -\log\left(-\frac{\log(U_i)}{1 + \sum_{l=1}^{m} \lambda_{li}} - \frac{\log(V_{ki})}{\lambda_{ki}} I\{y_i \neq k\}\right), \tag{16}$$

where U_i and $V_{1i}, \ldots, V_{mi}$ are $m + 1$ independent uniform random numbers in $[0, 1]$, and $\lambda_{li} = \exp(\mathbf{x}_i\boldsymbol{\beta}_l)$ for $l = 1, \ldots, m$.

3.2 Data Augmentation in the dRUM

An alternative way to write a multinomial model is as a difference random utility model (dRUM) which is obtained by choosing a baseline category k_0 and considering the model involving the differences of the utilities. This representation is the standard choice in the MCMC literature on the multinomial probit model, see e.g. McCulloch et al. (2000) and Imai and van Dyk (2005).

If we write the multinomial logit model as a dRUM, we obtain the following representation:

$$z_{ki} = \mathbf{x}_i\boldsymbol{\beta}_k + \epsilon_{ki}, \qquad \epsilon_{ki} \sim \text{Lo}, \qquad k = 1, \ldots, m, \tag{17}$$
$$y_i = \begin{cases} 0, & \text{if} \quad \max_{l\in L_{-0}} z_{li} < 0, \\ k > 0, & \text{if} \quad z_{ki} = \max_{l\in L_{-0}} z_{li} > 0, \end{cases}$$

where $z_{ki} = y^u_{ki} - y^u_{0i}$ and $\epsilon_{ki} = \delta_{ki} - \delta_{0i}$. The regression parameters appearing in (17) are identical to the ones appearing in (13), because $\boldsymbol{\beta}_{k_0} = \boldsymbol{\beta}_0 = \mathbf{0}$.

In contrast to the multinomial probit model, where $\boldsymbol{\epsilon}_i = (\epsilon_{1i}, \ldots, \epsilon_{mi})'$ follows a multivariate normal distribution, the vector $\boldsymbol{\epsilon}_i$ appearing in the dRUM representation of the multinomial logit model has a multivariate logistic distribution with logistic marginals (Balakrishnan, 1992, Section 11.2). While the errors in the RUM representation (14) are i.i.d., the errors ϵ_{ki} in the dRUM representation (17) are no longer independent across categories.

This complicates MCMC sampling to a certain degree. Following the MCMC literature on the multinomial probit model, we could introduce $\mathbf{z} = ((z_{k1}, \ldots, z_{kN}), k = 1, \ldots, m)$ as missing data and sample $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m|\mathbf{z}$ and $\mathbf{z}|\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m, \mathbf{y}$. However, while sampling $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m|\mathbf{z}$ is trivial in the multinomial probit model because $\boldsymbol{\epsilon}_i$ is multivariate normal, this step is non-standard in the multinomial logit model because $\boldsymbol{\epsilon}_i$ is multivariate logistic.

In the present paper we consider a different way of representing a multinomial model by differences in utilities. Note that equation (15) may be written as

$$y_i = k \Leftrightarrow y^u_{ki} > y^u_{-k,i}, \qquad y^u_{-k,i} = \max_{l\in L_{-k}} y^u_{li}. \tag{18}$$

Thus category k is observed iff y^u_{ki} is bigger than the maximum of all other utilities. Now we define for each (fixed) value of $k \in L_{-0}$ the latent variables w_{ki} as the difference between y^u_{ki} and $y^u_{-k,i}$ and construct binary observations $d_{ki} = I\{y_i = k\}$. Then it is possible to rewrite (18) as a binary model in the dRUM representation:

$$w_{ki} = y^u_{ki} - y^u_{-k,i}, \qquad d_{ki} = I\{w_{ki} > 0\}. \tag{19}$$

We term (19) the partial dRUM representation, because d_{ki} uses only partial information from the original data, namely whether y_i is equal to k or not.

It should be mentioned that the partial dRUM representation is not restricted to the multinomial logit model, but holds for arbitrary error distributions in the RUM representation (14). However, while the distribution of w_{ki} is in general unfeasible, it has an explicit form for the multinomial logit model. First of all,

$$\exp\left(-y^u_{-k,i}\right) \sim \text{Ex}\left(\lambda_{-k,i}\right), \qquad \lambda_{-k,i} = \sum_{l\in L_{-k}} \lambda_{li}, \tag{20}$$

because $\exp(-y^u_{-k,i}) = \min_{l\in L_{-k}} \exp(-y^u_{li})$, and $\exp(-y^u_{li}) \sim \text{Ex}\left(\lambda_{li}\right)$. We recall that $\lambda_{li} = \exp(\mathbf{x}_i\boldsymbol{\beta}_l)$. (20) may be rewritten as $y^u_{-k,i} = \log(\lambda_{-k,i}) + \delta_{-k,i}$, where $\delta_{-k,i}$ follows an EV distribution. Therefore

$$w_{ki} = y^u_{ki} - y^u_{-k,i} = \mathbf{x}_i\boldsymbol{\beta}_k - \log\left(\lambda_{-k,i}\right) + \delta_{ki} - \delta_{-k,i},$$

where $\delta_{-k,i}$ and $\delta_{k,i}$ are i.i.d. following an EV distribution. Thus the multinomial logit model has the following partial dRUM representation:

$$w_{ki} = \mathbf{x}_i \boldsymbol{\beta}_k - \log(\lambda_{-k,i}) + \epsilon_{ki}, \qquad d_{ki} = I\{w_{ki} > 0\}, \tag{21}$$

where $\epsilon_{ki} \sim \text{Lo}$. Evidently, for $m = 1$, (21) reduces to the dRUM given by (5).

The constant $\log(\lambda_{-k,i})$ appearing in (21) is independent of $\boldsymbol{\beta}_k$ and depends only on the regression parameters $\boldsymbol{\beta}_{-k}$ of the remaining categories. Thus given $\mathbf{z}_k = (w_{k1}, \ldots, w_{kN})$ and $\boldsymbol{\beta}_{-k}$, the regression parameter $\boldsymbol{\beta}_k$ corresponding to category k appears only in a linear regression model with logistic errors, given by (21).

Thus the partial dRUM is very useful when implementing MCMC for a multinomial model. At each MCMC draw we iterate over the categories for $k = 1, \ldots, m$. For each k, the partial dRUM actually is a binary dRUM and we may proceed as in Subsection 2.3 to sample $\mathbf{z}_k|\boldsymbol{\beta}_k, \mathbf{y}$ and $\boldsymbol{\beta}_k|\boldsymbol{\beta}_{-k}, \mathbf{z}_k$.

Evidently, $w_{ki}|\boldsymbol{\beta}_k, y_i$ is distributed according to a logistic distribution, truncated to $[0, \infty)$ if $y_i = k$, and truncated to $(-\infty, 0]$ otherwise. Thus w_{ki} is sampled as:

$$w_{ki} = \log\left(\lambda^\star_{ki} U_{ki} + I\{y_i = k\}\right) - \log\left(1 - U_{ki} + \lambda^\star_{ki}\, I\{y_i \neq k\}\right),$$

where $U_{ki} \sim \text{Un}[0, 1]$ and $\lambda^\star_{ki} = \lambda_{ki}/\lambda_{-k,i}$.

Then $\boldsymbol{\beta}_k$ is sampled from the non-normal regression model (21), where the constant $\log(\lambda_{-k,i})$ is added to both sides of equation (21) to obtain a zero mean error. To deal with the non-normality of ϵ_{ki}, one can apply any of the sampling strategies discussed in Subsection 2.3 for the dRUM representation of the logit model.

Actually, Holmes and Held (2006) sample $\boldsymbol{\beta}_k$ for a multinomial logit model using the partial dRUM representation, but do not provide a rigorous derivation from a random utility model as we did above. They represent the logistic distribution of ϵ_{ki} in (21) as an infinite scale mixture of normals and introduce and sample scaling factors ω_{ki}, $i = 1, \ldots, N$, for all $k = 1, \ldots, m$. As for the logit model, this sampler is rather demanding from a computational point of view.

Alternatively, we can apply the ideas underlying Scott (2009) to the partial dRUM representation (21). This involves sampling $\boldsymbol{\beta}_k$ by an independence MH algorithm, where the proposal is constructed from regression model (21) by replacing the logistic error term ϵ_{ki} by a normal error with the same variance, i.e. $\pi^2/3$.

Finally, the finite scale mixture approximation of the logistic distribution introduced in Subsection 2.3 may be applied to (21). This involves introducing and sampling indicators r_{ki}, $i = 1, \ldots, N$, for all $k = 1, \ldots, m$. Because this sampling step can be implemented in a very efficient way, auxiliary mixture sampling in the partial dRUM representation turns out to be much more efficient than the related sampler of Holmes and Held (2006).

4 MCMC Sampling without Data Augmentation

It is generally believed that MCMC samplers based on data augmentation are less efficient than MCMC samplers without data augmentation. However, we will demonstrate in Section 5 that the new data augmentation samplers introduced in this paper are more efficient than commonly used MH algorithms.

For our comparison we consider the two MH algorithms suggested in Rossi et al. (2005) and the DAFE-R MH algorithm suggested by Scott (2009). We assume without loss of generality that the baseline is chosen equal to 0 and provide details for the multinomial model. We use $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m)$ to denote the vector of all unknown regression parameters. The binary model results with $m = 1$.

Rossi et al. (2005, Section 3.11) discuss various MH algorithms based on the expected Hessian of the negative log-posterior $-\log p(\boldsymbol{\beta}|\mathbf{y})$. The elements of this matrix read:

$$-\mathrm{E}\left(\frac{\partial^2 \log p(\boldsymbol{\beta}|\mathbf{y})}{\partial \boldsymbol{\beta}_k^2}\right) = {\mathbf{B}_0}^{-1} + \sum_{i=1}^{N} \mathbf{x}_i{}' \mathbf{x}_i \, \pi_{ki}(\boldsymbol{\beta})(1 - \pi_{ki}(\boldsymbol{\beta})), \tag{22}$$

$$-\mathrm{E}\left(\frac{\partial^2 \log p(\boldsymbol{\beta}|\mathbf{y})}{\partial \boldsymbol{\beta}_k \partial \boldsymbol{\beta}_l}\right) = -\sum_{i=1}^{N} \mathbf{x}_i{}' \mathbf{x}_i \, \pi_{ki}(\boldsymbol{\beta}) \pi_{li}(\boldsymbol{\beta}).$$

An alternative approach uses the expected Hessian of the negative log-likelihood $-\log p(\mathbf{y}|\boldsymbol{\beta})$; however, this matrix is rank deficient if for a certain category k, $\pi_{ki} = 0$ for all $i = 1, \ldots, N$. Thus, adding the prior information matrix ${\mathbf{B}_0}^{-1}$ in (22) helps to stabilize the inverse of the expected Hessian in cases where for a certain k the probabilities π_{ki} are equal or close to 0 for most of the observations.

To obtain a proposal variance-covariance matrix that is independent of $\boldsymbol{\beta}$, the probabilities $\pi_{ki}(\boldsymbol{\beta})$ are substituted by some estimator, for instance $\hat{\pi}_{ki} = \pi_{ki}(\hat{\boldsymbol{\beta}})$, with $\hat{\boldsymbol{\beta}}$ being the posterior mode. It is useful to write the expected Hessian matrix as:

$$\mathbf{H} = \mathbf{I}_m \otimes {\mathbf{B}_0}^{-1} + \sum_{i=1}^{N} (\mathrm{Diag}\,(\hat{\boldsymbol{\pi}}_i) - \hat{\boldsymbol{\pi}}_i \hat{\boldsymbol{\pi}}_i{}') \otimes \mathbf{x}_i{}' \mathbf{x}_i,$$

where $\hat{\boldsymbol{\pi}}_i = (\hat{\pi}_{1i} \cdots \hat{\pi}_{mi})'$.

Rossi et al. (2005) construct two kinds of MH algorithms based on the matrix $\mathbf{H}$, namely an independence MH algorithm with a multivariate Student-t proposal $t_\nu(\hat{\boldsymbol{\beta}}, \mathbf{H}^{-1})$ with a small number of degrees of freedom ν, and a random walk MH algorithm with proposal $\boldsymbol{\beta}^{\text{new}}|\boldsymbol{\beta}^{\text{old}} \sim \mathrm{No}_{md}(\boldsymbol{\beta}^{\text{old}}, s^2\mathbf{H}^{-1})$ with scaling factor s^2. Roberts and Rosenthal (2001) prove that for a (md)-variate normal posterior distribution with variance-covariance equal to the identity matrix an asymptotically optimal scaling is given by $s^2 = 2.38^2/(md)$, with the corresponding optimal acceptance rate being equal to 0.234. Since the posterior $p(\boldsymbol{\beta}|\mathbf{y})$ is asymptotically normal with variance-covariance matrix equal to $\mathbf{H}^{-1}$, we use the following random walk proposal for $\boldsymbol{\beta}$:

$$\boldsymbol{\beta}^{\text{new}}|\boldsymbol{\beta}^{\text{old}} \sim \mathrm{No}_{md}\left(\boldsymbol{\beta}^{\text{old}}, \frac{2.38^2}{md}\mathbf{H}^{-1}\right). \tag{23}$$

Rossi et al. (2005, p.95) suggest to use the scaling factor $s^2 = 2.93^2/(md)$; however, it turns out that this scaling is inferior to the asymptotically optimal scaling.

Scott (2009) introduces the so-called DAFE-R MH algorithm which is based on computing the asymptotic variance-covariance matrix of the augmented posterior $p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{z})$ from the augmented random utility model (3). This variance-covariance matrix is used as a proposal in a multivariate random walk MH algorithm for the *marginal* model.

For the binary model this proposal reads:

$$\boldsymbol{\beta}^{\text{new}}|\boldsymbol{\beta}^{\text{old}} \sim \text{No}_d\left(\boldsymbol{\beta}^{\text{old}}, \left(\mathbf{B}_0{}^{-1} + \frac{6}{\pi^2}\mathbf{X}'\mathbf{X}\right)^{-1}\right). \tag{24}$$

The DAFE-R MH algorithm is applied to a multinomial logit model by using the proposal $\boldsymbol{\beta}_k^{\text{new}}|\boldsymbol{\beta}_k^{\text{old}} \sim \text{No}_d(\boldsymbol{\beta}_k^{\text{old}}, (\mathbf{B}_0{}^{-1} + 6/\pi^2\mathbf{X}'\mathbf{X})^{-1})$ for single-move sampling of $\boldsymbol{\beta}_k$ from $p(\boldsymbol{\beta}_k|\boldsymbol{\beta}_{-k}, \mathbf{y})$.

The proposal used in the DAFE-R MH algorithm has the advantage that the variance-covariance matrix depends only on $\mathbf{X}$ and consequently is very easily computed prior to MCMC sampling, while determining the Hessian $\mathbf{H}$ requires estimators of all unknown probabilities π_{ki}. However, since the DAFE-R is a random walk MH algorithm, it is likely to be inferior to the asymptotically optimal random walk (23), which is confirmed by the case studies in Section 5.

For a binary model, for instance, the proposal of the asymptotically optimal random walk simplifies to:

$$\boldsymbol{\beta}^{\text{new}}|\boldsymbol{\beta}^{\text{old}} \sim \text{No}_d\left(\boldsymbol{\beta}^{\text{old}}, \left(\frac{d}{2.38^2}\mathbf{B}_0{}^{-1} + \mathbf{X}'\text{Diag}\left(a_1, \ldots, a_N\right)\mathbf{X}\right)^{-1}\right),$$

where $a_i = \hat{\pi}_i(1-\hat{\pi}_i)d/2.38^2$. This proposal looks rather similar to the DAFE-R proposal (24), the main difference being the weight attached to $\mathbf{x}_i{}'\mathbf{x}_i$, which is equal to $6/\pi^2 = 0.6079$ rather than a_i for the DAFE-R algorithm. Thus if, on average, $6/\pi^2 > a_i$, the scaling of the DAFE-R algorithm is too small, causing the acceptance rate to be too high. For instance, if $\hat{\pi}_i = 0.5$ this happens if $d < 14$, while for $\hat{\pi}_i = 0.1$ this happens if $d < 38$. Thus we expect that the acceptance rate of the DAFE-R algorithm is too high in small regression models.

5 Comparison of the Various MCMC Algorithms

We apply nine different MCMC samplers to five well-known data sets. The (binary) nodal involvement data Chib (1995) is a small data set ($N = 53$) with a small set of regressors ($d = 5$). The (binary) heart data Holmes and Held (2006) is a medium sized data set ($N = 270$) with a larger set of regressors ($d = 14$). The (binary) German credit card data Holmes and Held (2006) is a large data set ($N = 1000$) with a large number of regressors ($d = 25$). The (multinomial) car data Scott (2009) is a medium sized data set ($N = 263$) with 3 categories and a small set of regressors ($d = 4$).

Finally, we consider the (multinomial) Caesarean birth data of Fahrmeir and Tutz (2001, Table 1.1), where the outcome variable has 3 categories (no infection and two type of infections) and $N = 251$. The data are organized as a three-way contingency table with eight factor combinations. The table is very unbalanced with a few cells containing a large fraction of the data, while other cells are empty. This makes statistical inference quite a challenge, and for illustration we fit a saturated logit model, i.e. $d = 8$.

For all examples, we take an independent standard normal prior for each regression coefficient and use each MCMC method to produce $M = 10000$ draws from the posterior distribution after running burn-in for 2000 iterations. All implementations are carried out using MATLAB (Version 7.3.0) on a notebook with a 2.0 GHz processor.

Naturally, we prefer fast samplers being nearly as efficient as i.i.d. sampling from the posterior $p(\boldsymbol{\beta}|\mathbf{y})$. Thus in Tables 5– 9 we summarize for each data set the performance of the various samplers in CPU time T_{CPU} (in seconds) needed to obtain the M draws (excluding burn-in) and the efficiency compared to i.i.d. sampling.

To evaluate the loss of efficiency, we compute for each regression coefficient β_{kj}, $k = 1, \ldots, m$, $j = 1, \ldots, d$ the inefficiency factor

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

where $\rho(h)$ is the empirical autocorrelation of the MCMC draws of that particular regression parameter at lag h. The initial monotone sequence estimator of Geyer (1992) is used to determine K, based on the sum of adjacent pairs of empirical autocorrelations $\Phi(s) = \rho(2s) + \rho(2s + 1)$. If n is the largest integer so that $\Phi(s) > 0$ and $\Phi(s)$ is monotone for $s = 1, \ldots, n$, then K is defined by $K = 2n + 1$. We determine for each regression coefficient the effective sample size ESS Kass et al. (1998) according to ESS $= M/\tau$. The closer ESS is to M, the smaller is the loss of efficiency. In Tables 5– 9 we report the median ESS for all regression coefficients, as well as the minimum and the maximum.

To compare a slow, but efficient sampler with a fast, but inefficient sampler, we consider for each regression coefficient the effective sampling rate ESR (per second), defined as ESR $= \text{ESS}/T_{\text{CPU}}$, and report the median ESR for all regression coefficients, as well as the minimum and the maximum. The median ESR is the most significant number in comparing the different MCMC samplers: the higher the median, the better the sampler.

We analyze three samplers using data augmentation in the dRUM, namely the sampler of Holmes and Held (2006) (dRUM-HH), our new auxiliary mixture sampler which substitutes the logistic distribution by the finite scale mixture approximation of Monahan and Stefanski (1992) with $H = 3$ and $H = 6$ (dRUM-FSF), and the new data augmented MH sampler which uses the posterior of the approximate standard linear regression model as proposal in the spirit of Scott (2009) (dRUM-Scott). We consider two samplers using data augmentation in the RUM, namely the auxiliary mixture sampler of Frühwirth-Schnatter and Frühwirth (2007) (RUM-FSF) and the original data augmented MH sampler of Scott (2009) (RUM-Scott). Finally, we consider the various random walk MH algorithms discussed in Section 4, namely the independence MH sampler of Rossi et al. (2005) (MH-Rossi), the asymptotically optimal random walk MH sampler of Roberts and Rosenthal (2001) (MH-RR), and the DAFE-R algorithm of Scott (2009) (MH-Scott).

We start the various MCMC samplers in the following way. All MH algorithms (with and without data augmentation) as well as all partial dRUM samplers for the multinomial logit model need a starting value for $\boldsymbol{\beta}_k$, $k = 1, \ldots, m$, which is set to $\mathbf{0}$. All data augmentation samplers need starting values for $\mathbf{z}$. For binary models starting values for $\mathbf{z}$ are sampled under the RUM representation from (6) and under the dRUM representation from (7) using $\lambda_i = \log \hat{\pi} - \log(1 - \hat{\pi})$, where $\hat{\pi} = \min(\max(\sum_{i=1}^{N} y_i/N, 0.05), 0.95)$. For multinomial models starting values for $\mathbf{z}$ are sampled from (16) with $\lambda_{0i} = 1$ and $\lambda_{li} = \log \hat{\pi}_l - \log(1 - \hat{\pi}_l)$, where $\hat{\pi}_l = \min(\max(\sum_{i=1}^{N} I\{y_i = l\}/N, 0.05), 0.95)$ for $l = 1, \ldots, m$. These values are transformed according to (19) to obtain starting values

Table 5: Comparing MCMC samplers for the nodal involvement data ($N = 53$, $d = 5$, $m = 1$); based on $M = 10\,000$ draws after burn-in of $2\,000$ draws.

	a	T_{CPU}	ESS (total draws)			ESR (draws/s)		
Sampler	(%)	(s)	min	med	max	min	med	max
dRUM-HH		25.4	3459.0	3883.5	4948.7	136.2	152.9	194.8
dRUM-FSF ($H = 3$)		5.1	3616.2	4025.1	4162.7	707.8	787.8	814.8
dRUM-FSF ($H = 6$)		5.5	3862.3	3986.1	4298.3	708.3	731.0	788.2
dRUM-Scott	71.5	2.9	3035.6	3156.4	3229.4	1061.8	1104.0	1129.6
RUM-FSF		8.7	213.6	233.4	305.7	24.6	26.9	35.3
RUM-Scott	32.9	3.6	459.6	533.8	593.5	126.2	146.6	163.0
MH-Rossi	14.5	3.7	837.8	884.8	1042.5	225.3	237.9	280.3
MH-RR	29.8	3.1	552.5	652.9	754.6	181.3	214.3	247.7
MH-Scott	54.4	3.0	339.5	450.6	477.4	111.5	147.9	156.7

Table 6: Comparing MCMC samplers for the heart data ($N = 270$, $d = 14$, $m = 1$); based on $M = 10\,000$ draws after burn-in of $2\,000$ draws.

	a	T_{CPU}	ESS (total draws)			ESR (draws/s)		
Sampler	(%)	(s)	min	med	max	min	med	max
dRUM-HH		94.3	863.2	1379.6	6225.6	9.2	14.6	66.0
dRUM-FSF ($H = 3$)		12.1	808.8	1432.4	5569.0	66.7	118.1	459.3
dRUM-FSF ($H = 6$)		14.7	931.6	1432.0	6196.7	63.4	97.4	421.5
dRUM-Scott	43.7	6.2	446.0	778.6	2037.7	72.3	126.2	330.2
RUM-FSF		31.0	57.2	94.0	868.5	1.84	3.03	28.0
RUM-Scott	5.6	7.5	17.0	30.7	156.1	2.3	4.1	20.8
MH-Rossi	18.0	5.5	320.6	421.2	588.1	58.6	77.0	107.5
MH-RR	27.0	4.8	212.1	255.2	300.7	44.5	53.6	63.1
MH-Scott	43.1	4.6	129.8	194.9	500.5	28.1	42.2	108.2

for w_{ki} in the partial dRUM representation. Finally, all elements of the latent scaling factors $\boldsymbol{\omega}$ are initialized with 1 for dRUM-HH, with $\pi^2/3$ for dRUM-FSF, and with $\pi^2/6$ for RUM-FSF.

Not surprisingly, we find for all data sets that MH sampling without data augmentation is faster than any data augmentation sampler in terms of CPU time T_{CPU}. To evaluate any MH sampler (with or without data augmentation) we report additionally the acceptance rate a, which is averaged over the categories for MH-Scott for multinomial models. For both random walk MH samplers a should be close to the asymptotically optimal rate of 0.234, which is actually the case for MH-RR with the exception of the Caesarean birth data in Table 9. The acceptance rate of MH-Scott deviates from the asymptotically optimal rate for all examples but the German credit card data in Table 7, which causes the effective sample size and the effective sampling rate to be smaller than for MH-RR.

With the exception of the Caesarean birth data, MH-Rossi outperforms the other MH samplers without data augmentation in terms of effective sample size and effective

Table 7: Comparing MCMC samplers for the German credit card data ($N = 1000$, $d = 25$, $m = 1$); based on $M = 10\,000$ draws after burn-in of $2\,000$ draws.

	a	T_{CPU}	ESS (total draws)			ESR (draws/s)		
Sampler	(%)	(s)	min	med	max	min	med	max
dRUM-HH		333.4	1556.0	2325.7	3494.4	4.7	7.0	10.5
dRUM-FSF ($H = 3$)		41.8	1573.5	2313.5	3780.1	37.7	55.4	90.4
dRUM-FSF ($H = 6$)		61.5	1666.9	2268.3	3872.2	27.1	36.9	63.0
dRUM-Scott	30.4	21.6	592.6	824.4	1090.8	27.4	38.2	50.5
RUM-FSF		134.2	91.5	133.7	261.5	0.68	1.00	1.95
RUM-Scott	0.8	25.0	9.7	11.8	26.1	0.39	0.47	1.0
MH-Rossi	7.1	11.2	117.3	178.5	290.9	10.4	15.9	25.9
MH-RR	25.0	11.1	92.5	138.2	188.0	8.3	12.5	17.0
MH-Scott	22.0	10.4	103.9	149.8	189.2	10.0	14.4	18.1

Table 8: Comparing MCMC samplers for the car data ($m = 2$, $N = 263$, $d = 3$); based on $M = 10\,000$ draws after burn-in of $2\,000$ draws.

	a	T_{CPU}	ESS (total draws)			ESR (draws/s)		
Sampler	(%)	(s)	min	med	max	min	med	max
dRUM-HH		182.5	1716.8	2558.2	3020.5	9.4	14.0	16.6
dRUM-FSF ($H = 3$)		20.9	1831.7	2535.8	3200.5	87.6	121.2	153.0
dRUM-FSF ($H = 6$)		27.20	1570.9	2307.6	2942.4	57.8	84.8	108.2
dRUM-Scott	70.2	13.0	1468.3	2101.1	2662.8	113.4	162.2	205.6
RUM-FSF		46.6	111.3	171.8	253.2	2.4	3.7	5.4
RUM-Scott	33.8	9.5	289.8	388.7	472.8	30.6	41.1	49.9
MH-Rossi	57.4	6.0	3158.0	3323.7	3899.0	526.3	554.0	649.8
MH-RR	27.2	5.3	366.2	397.2	499.1	69.3	75.2	94.5
MH-Scott	61.7	10.2	269.4	365.1	527.0	26.5	35.8	51.7

sampling rate. This is true even for the German credit data where the acceptance rate is as low as 7.1%. In general, the acceptance rate of MH-Rossi varies considerably across the various case studies, being pretty high for the car data in Table 8 and being extremely small for the Caesarean birth data in Table 9.

For the Caesarean birth data the Hessian matrix is very ill-conditioned due to the unbalanced data structure mentioned earlier, leading to a very low acceptance rate both for MH-Rossi and MH-RR. For this particular data set MH-Scott outperforms the other MH samplers, because it avoids the Hessian when constructing the variance-covariance matrix of the proposal density.

When comparing the various data augmentation samplers in the RUM and in the dRUM representation, we find for all case studies that both the effective sample size and the effective sampling rate are considerably higher for the dRUM representation than for the RUM representation, leading to the conclusion that data augmentation in the RUM should be avoided.

Among the data augmentation samplers in the dRUM representation, data aug-

Table 9: Comparing MCMC samplers for the Caesarean birth data ($m = 2$, $N = 251$, $d = 8$); based on $M = 10\,000$ draws after burn-in of $2\,000$ draws.

	a	T_{CPU}	ESS (total draws)			ESR (draws/s)		
Sampler	(%)	(s)	min	med	max	min	med	max
dRUM-HH		177.8	1153.4	2643.1	4553.1	6.5	14.9	25.6
dRUM-FSF ($H = 3$)		21.0	1195.2	2587.8	4621.8	56.8	123.0	219.8
dRUM-FSF ($H = 6$)		26.4	1125.5	2777.4	4765.0	42.6	105.1	180.2
dRUM-Scott	63.9	12.3	714.9	1790.8	3084.8	58.4	146.2	251.8
RUM-FSF		42.1	148.6	344.1	899.6	3.5	8.2	21.4
RUM-Scott	23.4	10.2	213.5	389.9	729.2	21.0	38.3	71.6
MH-Rossi	2.0	5.7	37.0	89.1	120.0	6.5	15.5	20.9
MH-RR	3.9	4.9	22.8	50.0	83.5	4.7	10.3	17.1
MH-Scott	39.8	9.7	254.7	354.1	486.7	26.3	36.6	50.3

mented MH based on the approximate normal proposal (dRUM-Scott) is the fastest. As expected, the acceptance rate a, which should be as high as possible, is considerably larger for dRUM-Scott than under the RUM representation (RUM-Scott), because the logistic distribution underlying the dRUM is much closer to a normal distribution than the extreme value distribution underling the RUM. For the German credit data in Table 7, for instance, the acceptance rate increases from 0.8% for RUM-Scott to 30.4% for dRUM-Scott.

Compared to dRUM-Scott, the other two dRUM data augmentation samplers are slower, because both dRUM-HH and dRUM-FSF introduce the latent scaling factors $\boldsymbol{\omega}$ as a second set of auxiliary variables. We find that dRUM-HH requires much more computation time than dRUM-FSF, even if the latter uses the very accurate mixture approximation with six components, while the efficiency in terms of effective sample size is more or less the same. This makes our new dRUM auxiliary mixture sampler much more efficient in terms of effective sampling rate than the sampler of Holmes and Held (2006).

Interestingly the effective sample size of dRUM-HH and dRUM-FSF is larger than dRUM-Scott. Introducing the latent scaling factors $\boldsymbol{\omega}$ allows dRUM-HH and dRUM-FSF to accept $\boldsymbol{\beta}$ at each sweep of the MCMC sampler, because a conditional Gibbs step is implemented. In contrast to that dRUM-Scott uses an MH update for $\boldsymbol{\beta}$, meaning that the sampler is stuck at the current value with probability $1 - a$, which increases the autocorrelation in the MCMC sample.

When we compare our new dRUM data augmentation samplers, we find that they outperform any other data augmentation sampler in terms of the effective sampling rate. With the exception of the car data in Table 8, the samplers even outperform the independence MH sampler of Rossi et al. (2005). The relatively high acceptance rate of MH-Rossi for the car data explains its superiority for this particular example.

Finally, we discuss the performance of our new samplers in relation to each other. While dRUM-Scott is faster, dRUM-FSF has a higher effective sample size. The effective sampling rate is higher for dRUM-Scott with the exception of the German credit card data in Table 7, where the acceptance rate of dRUM-Scott is smaller than in the other examples. It appears from the various tables that an acceptance rate of dRUM-

Scott above 40% makes the sampler more efficient in terms of the effective sampling rate than dRUM-FSF.

Because the coding of dRUM-Scott is extremely simple, we recommend to make this new data augmented MH sampler the first choice. However, while there is no tuning in the proposal of dRUM-Scott — which makes it easy to implement — there is, on the other hand, no control over the acceptance rate. Thus the acceptance rate may be arbitrarily small, depending on the particular application. Thus, if the acceptance rate turns out to be considerably smaller than say 40%, it is to be expected that dRUM-FSF is more efficient and should be the method of choice.

6 Concluding Remarks

In this paper we have introduced yet two other data augmentation algorithms for sampling the parameters of a binary or a multinomial logit model from their posterior distribution within a Bayesian framework. They are based on rewriting the underlying random utility model in such a way that only differences of utilities appear in the model. Applications to five case studies reveal that these samplers are superior to other data augmentation samplers and to Metropolis–Hastings sampling without data augmentation.

We have confined our investigations to the standard binary and multinomial logit regression model; however, we are confident that our new samplers will be of use for the MCMC estimation of more general latent variable models such as analyzing discrete-valued panel data using random-effects models, or analyzing discrete-valued time series using state space models. For latent variable models, auxiliary mixture sampling in the dRUM representation is of particular relevance, because introducing the auxiliary latent variables $\mathbf{z}$ and $\boldsymbol{\omega}$ leads to a conditionally Gaussian model, which allows efficient sampling of the random effects or the state vector.

Furthermore, dRUM auxiliary mixture sampling could be useful for Bayesian variable selection in binary data analysis simply by replacing less efficient samplers such as the Holmes and Held (2006) sampler, which was used in the same paper for variable selection in logistic regression models, and the RUM auxiliary mixture sampling, which was used in Tüchler (2008) for covariance selection in panel data models with random effects. Furthermore, it could be applied to the stochastic variable selection approach of Frühwirth-Schnatter and Wagner (2010) for state space modelling of binary time series.

It remains an open issue whether representations comparable to the dRUM exist for more general discrete-valued distributions. Frühwirth-Schnatter et al. (2009) improve auxiliary mixture sampling for data from a binomial or multinomial distribution by using an aggregated RUM representation instead of the RUM representation of the underlying individual binary experiments. It seems worth investigating whether auxiliary mixture sampling for such data can be improved further using an aggregated version of the dRUM representation; however, we leave this issue for further research.

Acknowledgement

The first author's research is supported by the Austrian Science Foundation (FWF) under the grant S 10309-G14 (NRN "The Austrian Center for Labor Economics and the Analysis of the Welfare State", Subproject "Bayesian Econometrics").

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.

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

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

Chib, S. (1995). Marginal likelihood from the Gibbs output, *Journal of the American Statistical Association* **90**: 1313–1321.

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

Fahrmeir, L. and Kaufmann, H. (1986a). Asymptotic inference in discrete response models, *Statistical Papers* **27**: 179–205.

Fahrmeir, L. and Kaufmann, H. (1986b). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models, *The Annals of Statistics* **13**: 342–368.

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

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., 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. (2010). Stochastic model specification search for Gaussian and partially non-Gaussian state space models, *Journal of Econometrics*, **154**: 85–100.

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

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

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.

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

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*, Academic, New York, pp. 105–142.

Monahan, J. F. and Stefanski, L. A. (1992). Normal scale mixture approximations to $F^{\star}(z)$ and computation of the logistics normal integral, *in* N. Balakrishnan (ed.), *Handbook of the Logistic Distribution*, Marcel Dekker, New York, pp. 529–549.

Roberts, G. O. and Rosenthal, J. S. (2001). Optimal scaling for various Metropolis–Hastings algorithms, *Statistical Science* **16**: 351–367.

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

Scott, S. L. (2009). Data augmentation and the Bayesian analysis of multinomial logit models, *Statistical Papers*, forthcoming.

Tüchler, R. (2008). Bayesian variable selection for logistic models using auxiliary mixture sampling, *Journal of Computational and Graphical Statistics* **17**: 76–94.

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

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