Department for Applied Statistics
Johannes Kepler University Linz

IFAS Research Paper Series
2004-06

Data Augmentation and Gibbs Sampling for Logistic Models

Sylvia Frühwirth-Schnatter and Helmut Waldl

December 2004

Abstract

In this article we consider logit-type models, like the standard binary logistic regression, multinomial models with random effects, and state space models for binary data. Estimation of these models is carried out within a Bayesian framework using data augmentation and MCMC methods. We suggest a new MCMC sampler, which possesses a Gibbs transition kernel, where we draw from full conditional distributions belonging to standard distribution families, only.

Key words: binary data, data augmentation, generalized linear models, Gibbs sampling, multinomial data, utilities

1 Introduction

Applied statisticians commonly have to deal with the problem of modelling binary or categorical data in terms of covariates. Examples are modelling the probability of disease occurrences in terms of risk factors, or modelling choice probabilities in marketing in terms of product attributes. The logistic regression model, discussed for instance in McCullagh and Nelder (1999) in the framework of generalized linear models, is an important and extensively used tool for analyzing the effect of covariates on the occurrence probabilities of a certain event. The basic logistic regression model has been modified in a number of ways. To account for the dependency likely to be present in sequences of binary data, past observations $y_{i-1}, y_{i-2}, \dots$ have been introduced as covariates, see for instance Zeger and Qaqish (1988). A couple of extensions deal with overdispersion due to omitted covariates, like mixtures of binary regressions models (Wang et al. 1996; Hurn et al. 2003), binary regression models with additive random effects (Aitkin 1996), and mixtures of binary regression models with random effects (Lenk and DeSarbo 2000).

In this paper we consider Bayesian estimation of binary and multinomial logit models, using data augmentation as in Tanner and Wong (1987) and Markov chain Monte Carlo methods, as illustrated first by Zeger and Karim (1991) for generalized linear models with random effects. Since this seminal paper, numerous authors have contributed to MCMC estimation of logit-type models. We mention here in particular Lenk and DeSarbo (2000) for mixtures of logit-models with random effects, and Hurn et al. (2003) for mixtures of binary regression. A major difficulty with any of the existing MCMC approaches, however, is that practical implementation requires the use of a Metropolis-Hastings algorithm at least for part of the unknown parameter vector, which in turns makes it necessary to define suitable proposal densities.

The main contribution of the present article is to show that straightforward Gibbs sampling of all parameters, requiring only random draws from standard distributions such as multivariate normals, inverse Gamma, exponential and discrete distributions with a few categories is feasible for logit models. This rather unexpected result is achieved by introducing two sequences of latent variables through data augmentation. The first data augmentation step is based on Scott (2004), who introduced the latent utilities as missing variables. As shown by Scott (2004), the introduction of this first sequence eliminates the non-linearity of the observation equation, whereas the non-normality of the error term, which follows a type I extreme value distribution, remains. Whereas Scott (2004) uses a Metropolis-Hastings algorithm to sample the parameters, we eliminate the non-normality of the error term by a second sequence of latent variables. To this aim, the log of the extreme value distribution is approximated by a mixture of normal distributions in a similar way as in Kim, Shephard, and Chib (1998) and Chib, Nardari, and Shephard (2002) who used a normal mixture approximation to the density of a $\log \chi^2$-distribution in the context of stochastic volatility models. By introducing the component indicator of this normal mixture as a second sequence of missing data, a logistic regression model may be thought of as a partially Gaussian model as in Shephard (1994), and Gibbs sampling becomes feasible. This will be shown to be particularly useful for random effects models and for state space models for binary and categorical time series, as multi-move-sampling of the whole state process through forward-filtering backward sampling as in Frühwirth-Schnatter (1994), Carter and Kohn (1994), De Jong and Shephard (1995) and Durbin and Koopman (2002) becomes feasible.

The rest of the paper is organized as follows. In Section 2, we discuss in detail data augmentation and Gibbs sampling for binary logit regression models, which will be extended to more complex binary models, like time series models and panel data models in Section 3. In Section 4 we extend data augmentation and Gibbs sampling to multinomial logit models. Section 5 concludes.

2 Data Augmentation and Gibbs Sampling for the Binary Logit Regression Models

2.1 Background

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

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

where $\mathbf{x}_i$ is a row vector of regressors, including 1 for the intercept, and $\boldsymbol{\beta}$ is an unknown regression parameter.

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

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

The resulting posterior density, however, in general does not belong to a density from a well-known distribution family. Markov chain Monte Carlo methods to sample from the posterior distribution of a logit model were applied by Zeger and Karim (1991), Albert (1992), Chib et al. (1998), Lenk and DeSarbo (2000), Hurn et al. (2003), and Scott (2004), among many others. As mentioned in the introduction, any of these methods is based on Metropolis-Hastings sampling. We are going to demonstrate in the following subsection, that the introduction of two sequences of artificially missing data within a data augmentation scheme leads to a conditional posterior distribution for $\boldsymbol{\beta}$ that, in contrast to $p(\boldsymbol{\beta}|\mathbf{y})$, is a joint normal distribution, once we conditioned on the artificially missing data. Thus the whole regression parameter $\boldsymbol{\beta}$ could be sampled in one sweep from a normal distribution.

2.2 Data Augmentation for the Binary Logit Regression Model

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

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

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

The first step of data augmentation introduces for each i, $i = 1, \ldots, N$, the latent utility y_i^u of choosing category 1 as missing data, with two desirable effects. First, the full-conditional posterior distribution $p(\boldsymbol{\beta}|\mathbf{y}^u, \mathbf{y})$ of $\boldsymbol{\beta}$, where additionally to $\mathbf{y}$ the latent utilities $\mathbf{y}^u = (y_1^u, \ldots, y_N^u)$ appear as conditioning argument, is independent of $\mathbf{y}$, $p(\boldsymbol{\beta}|\mathbf{y}^u, \mathbf{y}) = p(\boldsymbol{\beta}|\mathbf{y}^u)$. Second, conditional on $\mathbf{y}^u$, the posterior of $\boldsymbol{\beta}$ could be derived from regression model (2), which is non-normal, but linear in the unknown model parameters $\boldsymbol{\beta}$. Thus, the first augmentation step eliminates the non-linearity of the logit model, the non-normality of the error term ε_i, however, remains. Scott (2004) uses a Metropolis-Hastings algorithm based on various approximations to this regression model, to sample the regression parameters $\boldsymbol{\beta}$.

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

$$p(\varepsilon_i) = \exp\{-\varepsilon_i - e^{-\varepsilon_i}\}. \tag{3}$$

To obtain a model that is conditionally Gaussian, we approximate the non-normal density $p(\varepsilon_i)$ by a normal mixture of 5 components with parameters m_r and s_r for the r-th component:

$$p(\varepsilon_i) = \exp\{-\varepsilon_i - e^{-\varepsilon_i}\} \approx \sum_{r=1}^{5} w_r f_{\mathcal{N}}(\varepsilon_i; m_r, s_r^2). \tag{4}$$

This idea is influenced by the related articles of Kim et al. (1998) and Chib et al. (2002), who used a normal mixture approximation of the density of a $\log \chi^2$-distribution in the context of stochastic volatility models. The appropriate parameters $(w_r, m_r, s_r^2), r = 1, \ldots, 5$, however, are different for our problem and are

Table 1: Normal mixture approximation of the density of the type I extreme value distribution (5 components)

r	1	2	3	4	5
w_r	0.2924	0.2599	0.2480	0.1525	0.0472
m_r	-0.0982	1.5320	0.7433	-0.8303	3.1428
s_r^2	0.2401	1.1872	0.3782	0.1920	3.2375

tabulated in Table 1 for 5 components, a number that we found to be sufficiently large in practice.[1]

Following Kim et al. (1998) and Chib et al. (2002), the mixture distribution (4) is regarded as the marginal distribution of a problem where additional to ε_i the component indicators r_i are observed. The second step of our data augmentation scheme introduces for each ε_i the latent component indicator r_i as missing data. Conditional on knowing the latent utility y_i^u and the latent indicator r_i, the binary logit regression model (1) reduces to a Gaussian regression model with heteroscedastic errors with known variance:

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

For such a model it is well known, that the conditional posterior of $\boldsymbol{\beta}$ is a multivariate normal density, see for instance Zellner (1971). This result is the basis for our new two-block Gibbs sampler, that will be described in the next subsection.

2.3 A Two-Block Gibbs Sampler

A two-block Gibbs sampler results, if data augmentation as described in the previous section is applied for all observations. This leads to two sequences of latent variables, the component indicators $\mathbf{R} = \{r_1, \ldots, r_N\}$, and the latent utilities $\mathbf{y}^u = \{y_1^u, \ldots, y_N^u\}$. Within our Gibbs sampling scheme, we select a starting value for $\mathbf{R} =$ and $\mathbf{y}^u$, and repeat the following steps:

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

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

(b1) Sample the latent utility y_i^u conditional on $\boldsymbol{\beta}$ and $\mathbf{y}$ as

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

where U_i and V_i are two independent uniform random numbers.

[1] This table is derived from a related table appearing in Frühwirth-Schnatter and Wagner (2004), by observing that $-\varepsilon_i$ has the same density as the log of an exponentially distributed random variable.

(b2) Sample the component indicators r_i conditional on y_i^u and $\boldsymbol{\beta}$ from the following discrete density:

$$\log \Pr(r_i = j | y_i^u, \boldsymbol{\beta}) \propto -\log s_j - \frac{1}{2}\left(\frac{y_i^u - \mathbf{x}_i\boldsymbol{\beta} - m_j}{s_j}\right)^2 + \log w_j. \qquad (7)$$

The quantities $(w_j, m_j, s_j^2), j = 1, \ldots, 5$ are the parameters of the finite mixture approximation tabulated in Table 1.

Note that step (b) involves only draws from standard densities. Thus sampling scheme (a) and (b) is actually a Gibbs sampler without any tuning. Step (b1) could be used to sample starting values for y_i^u for each i, given the observed binary data y_i, by choosing a starting values for $\lambda_i = \exp(\mathbf{x}_i\boldsymbol{\beta})$. Starting values for each component indicator r_i are obtained as random draws from 1 to 5.

2.3.1 Details on the Sampling Steps

Conditionally on knowing $\mathbf{y}^u = (y_1^u, \ldots, y_N^u)$ and $\mathbf{R} = (r_1, \ldots, r_N)$, the binary logit model (1) reduces to the linear normal regression model (5). Therefore, in step (a), the conditional posterior of $\boldsymbol{\beta}$ is given by the $\mathcal{N}_d(\mathbf{b}_N, \mathbf{B}_N)$-distribution, where

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

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

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

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

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

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

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

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

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

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

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

where $\lambda_i = \exp(\mathbf{x}_i\boldsymbol{\beta})$. Second, as the minimum of exponential random variables follows again an exponential distribution, we obtain:

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

Third, knowing the minimum, the other random variable has a translated exponential distribution. In particular, if $\exp(-y_{0i}^u) < \exp(-y_i^u)$, then

$$\exp(-y_i^u) = \exp(-y_{0i}^u) + \xi_i, \quad \xi_i \sim \mathcal{E}\left(\lambda_i\right). \tag{11}$$

These results enable sampling of the latent utility y_i^u, knowing y_i. If $y_i = 1$, then $y_i^u > y_{0i}^u$, or equivalently, $\exp(-y_i^u) < \exp(-y_{0i}^u)$. Therefore we obtain from (10):

$$\exp(-y_i^u) \sim \mathcal{E}\left(1 + \lambda_i\right). \tag{12}$$

On the other hand, if $y_i = 0$, then $y_i^u < y_{0i}^u$, or equivalently, $\exp(-y_{0i}^u) < \exp(-y_i^u)$. Therefore we obtain from (10) and (11):

$$\exp(-y_{0i}^u) \sim \mathcal{E}\left(1 + \lambda_i\right), \quad \exp(-y_i^u) = \exp(-y_{0i}^u) + \xi_i, \quad \xi_i \sim \mathcal{E}\left(\lambda_i\right). \tag{13}$$

By the help of two uniform random numbers U_i and V_i, (12) and (13) could be written immediately as in formula (6) in step (b1).

3 Extension to Complex Binary Logit Models

To illustrate the great flexibility of our Gibbs sampling scheme, we consider in detail more complex binary logit models, like binary state space models and binary logit models with random effects.

3.1 Binary Regression Models with Time-Varying Parameters

3.1.1 Background

Let $\{y_t\}$ be a time series of binary observations, observed for $t = 1, \ldots, T$. Each y_t is assumed to take one of two possible values, labelled by $\{0, 1\}$. The probability that y_t takes the value 1 depends on covariates $\mathbf{x}_t = (\mathbf{x}_t^1 \ \mathbf{x}_t^2)$ through fixed parameters $\boldsymbol{\alpha}$ and a time-varying parameters $\boldsymbol{\beta}_t^s$ in the following way:

$$\Pr(y_t = 1|\boldsymbol{\beta}_1^s, \ldots, \boldsymbol{\beta}_T^s, \boldsymbol{\alpha}) = \frac{\exp(\mathbf{x}_t^1\boldsymbol{\alpha} + \mathbf{x}_t^2\boldsymbol{\beta}_t^s)}{1 + \exp(\mathbf{x}_t^1\boldsymbol{\alpha} + \mathbf{x}_t^2\boldsymbol{\beta}_t^s)}. \tag{14}$$

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

$$\boldsymbol{\beta}_t^s = \boldsymbol{\beta}_{t-1}^s + \mathbf{w}_t, \quad \mathbf{w}_t \sim \mathcal{N}_d\left(\mathbf{0}, \mathbf{Q}\right), \tag{15}$$

with $\boldsymbol{\beta}_0^s \sim \mathcal{N}\left(\boldsymbol{\beta}, \mathbf{B}_0\right)$. $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are unknown location parameters, $\mathbf{Q}$ is an unknown covariance matrix. Note that this model may be regarded as a special case of a more general state space model for binary data.

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

3.1.2 Data Augmentation and Gibbs Sampling

The data augmentation scheme introduced in Section 2 for the standard regression model is actually identical when we are dealing with a time series. A latent utility y_t^u of choosing category 1 is introduced for each y_t, to eliminate the non-linearity of the model:

$$y_t^u = \mathbf{x}_t^1 \boldsymbol{\alpha} + \mathbf{x}_t^2 \boldsymbol{\beta}_t^s + \varepsilon_t, \tag{16}$$

where ε_t follows a type I extreme value distribution. To eliminate non-normality, this distribution is approximated by a mixture of normals as in Subsection 2.2, and a latent indicator r_t is introduced for each y_t. Let $\mathbf{y}^u = \{y_1^u, \dots, y_T^u\}$ denote the collection of all latent utilities, and let $\mathbf{R} = \{r_1, \dots, r_T\}$ denote the collection of all latent component indicators. If we condition on the latent variables $\mathbf{y}^u$ and $\mathbf{R}$, we obtain a linear Gaussian state space model with heteroscedastic errors with known error variance:

$$\boldsymbol{\beta}_t^s = \boldsymbol{\beta}_{t-1}^s + \mathbf{w}_t, \quad \mathbf{w}_t \sim \mathcal{N}_d(\mathbf{0}, \mathbf{Q}), \tag{17}$$

$$y_t^u = \mathbf{x}_t^1 \boldsymbol{\alpha} + \mathbf{x}_t^2 \boldsymbol{\beta}_t^s + m_{r_t} + s_{r_t}\varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, 1), \tag{18}$$

for $t = 1, \dots, T$, and $\boldsymbol{\beta}_0^s \sim \mathcal{N}(\boldsymbol{\beta}, \mathbf{B}_0)$. Thus it is easy to implement a three block Gibbs sampler, which consists of the following steps:

(a) Multi-move sampling of $\boldsymbol{\beta}_0^s, \dots, \boldsymbol{\beta}_T^s, \boldsymbol{\beta}, \boldsymbol{\alpha}$ conditional on knowing $\mathbf{y}^u$, $\mathbf{R}$, and $\mathbf{Q}$, based on the conditional linear Gaussian state space model (17) and (18).

(b) Sampling of $\mathbf{Q}$ conditional on knowing $\boldsymbol{\beta}_0^s, \dots, \boldsymbol{\beta}_T^s$, based on the transition equation (17) of the conditionally linear Gaussian state space model.

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

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

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

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

Sampling of the latent utility y_t^u and the component indicator r_t is carried out exactly as in Subsection 2.3:

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

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

where U_t and V_t are two independent uniform random numbers, and $\lambda_t = \exp(\mathbf{x}_t^1 \boldsymbol{\alpha} + \mathbf{x}_t^2 \boldsymbol{\beta}_t^s)$.

3.2 The Binary Logit Random Effects Model

3.2.1 Background

Let $\{y_{it}\}, t = 1, \ldots, T_i$ be repeated binary measurements, observed for N subjects $i = 1, \ldots, N$. Each y_{it} is assumed to take one of two possible values labelled by $\{0, 1\}$. The probability that y_{it} takes the value 1 depends on covariates $\mathbf{x}_{it} = (\mathbf{x}_{it}^1 \, \mathbf{x}_{it}^2)$ through fixed parameters $\boldsymbol{\alpha}$ and subject-specific parameters $\boldsymbol{\beta}_i^s$ in the following way:

$$\Pr(y_{it} = 1 | \boldsymbol{\beta}_1^s, \ldots, \boldsymbol{\beta}_N^s, \boldsymbol{\alpha}) = \frac{\exp(\mathbf{x}_{it}^1 \boldsymbol{\alpha} + \mathbf{x}_{it}^2 \boldsymbol{\beta}_i^s)}{1 + \exp(\mathbf{x}_{it}^1 \boldsymbol{\alpha} + \mathbf{x}_{it}^2 \boldsymbol{\beta}_i^s)}. \tag{19}$$

We assume that conditional on knowing $\boldsymbol{\beta}_1^s, \ldots, \boldsymbol{\beta}_N^s, \boldsymbol{\alpha}$, the observations are mutually independent. A commonly used prior for $\boldsymbol{\beta}_i^s$ reads $\boldsymbol{\beta}_i^s \sim \mathcal{N}_d(\boldsymbol{\beta}, \mathbf{Q})$. $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are unknown location parameters, whereas $\mathbf{Q}$ is an unknown covariance matrix.

3.2.2 Data Augmentation and Gibbs Sampling

The data augmentation scheme introduced in Section 2 for the standard regression model is easily extended to deal with repeated measurements. A latent utility y_{it}^u is introduced for each y_{it}, to eliminate the non-linearity of the model:

$$y_{it}^u = \mathbf{x}_{it}^1 \boldsymbol{\alpha} + \mathbf{x}_{it}^2 \boldsymbol{\beta}_i^s + \varepsilon_{it}, \tag{20}$$

where ε_{it} follows a type I extreme value distribution. To eliminate non-normality, this distribution is approximated by a mixture of normals as in Subsection 2.2, and an latent indicator r_{it} is introduced for each y_{it}.

Let $\mathbf{y}^u = \{(y^u_{i1}, \ldots, y^u_{i,T_i}), i = 1, \ldots, N\}$ denote the collection of all latent utilities, and let $\mathbf{R} = \{(r_{i1}, \ldots, r_{i,T_i}), i = 1, \ldots, N\}$ denote the collection of all latent component indicators. If we condition on the latent variables $\mathbf{y}^u$ and $\mathbf{R}$, we obtain a Gaussian linear random-effects model with heteroscedastic errors with known error variance:

$$\boldsymbol{\beta}^s_i \sim \mathcal{N}_d(\boldsymbol{\beta}, \mathbf{Q}), \tag{21}$$

$$y^u_{it} = \mathbf{x}^1_{it}\boldsymbol{\alpha} + \mathbf{x}^2_{it}\boldsymbol{\beta}^s_i + m_{r_{it}} + s_{r_{it}}\varepsilon_{it}, \quad \varepsilon_{it} \sim \mathcal{N}(0, 1), \tag{22}$$

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

(a) Multi-move sampling of $\boldsymbol{\beta}^s_1, \ldots, \boldsymbol{\beta}^s_N, \boldsymbol{\beta}, \boldsymbol{\alpha}$ conditional on knowing $\mathbf{y}^u$, $\mathbf{R}$, and $\mathbf{Q}$, based on the conditionally linear Gaussian random-effects model (22).

(b) Sampling of $\mathbf{Q}$ conditional on knowing $\boldsymbol{\beta}^s_1, \ldots, \boldsymbol{\beta}^s_N, \boldsymbol{\beta}$, based on (21).

(c) Sampling of the utilities $\mathbf{y}^u$ and the indicators $\mathbf{R}$ conditional on knowing $\boldsymbol{\beta}^s_1, \ldots, \boldsymbol{\beta}^s_N, \boldsymbol{\alpha}$, and $\mathbf{y}$.

An important aspect of our data augmentation scheme is that conditional on $\mathbf{y}^u$ and $\mathbf{R}$, we are dealing with a linear Gaussian random effects model, when sampling $\boldsymbol{\alpha}, \boldsymbol{\beta}$ and $\boldsymbol{\beta}^s_i$ in step (a) and $\mathbf{Q}$ in step (b), where the binary observation y_{it} is substituted by a conditionally normal random variable y^u_{it}, and the error term follows a $\mathcal{N}(m_{r_{it}}, s_{r_{it}})$-distribution. Thus for a binary logit model with random effects, step (a) and (b) in the Gibbs sampling scheme introduced above are as simple as for the corresponding *linear Gaussian* random-effects model. In step (a), joint multi-move sampling of all location parameters $\boldsymbol{\beta}^s_1, \ldots, \boldsymbol{\beta}^s_N, \boldsymbol{\beta}, \boldsymbol{\alpha}$ is possible along the lines indicated by Frühwirth-Schnatter et al. (2004), see also Frühwirth-Schnatter and Otter (1999) and Sahu and Roberts (1999), by sampling $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ from the marginal model, where the random effects are integrated out. We provide details in the next subsection.

Step (c) is implemented by writing the joint posterior $p(\mathbf{R}, \mathbf{y}^u|\mathbf{y}, \boldsymbol{\beta}^s_1, \ldots, \boldsymbol{\beta}^s_N, \boldsymbol{\alpha})$ as:

$$p(\mathbf{R}, \mathbf{y}^u|\mathbf{y}, \boldsymbol{\beta}^s_1, \ldots, \boldsymbol{\beta}^s_N, \boldsymbol{\alpha}) = \prod_{i=1}^{N}\prod_{t=1}^{T_i} p(y^u_{it}|y_{it}, \boldsymbol{\beta}^s_i, \boldsymbol{\alpha})p(r_{it}|y^u_{it}, \boldsymbol{\beta}^s_i, \boldsymbol{\alpha}).$$

Sampling of y^u_{it} is possible in terms of two uniform random variables U_{it} and V_{it}:

$$y^u_{it} = -\log\left(\frac{-\log(U_{it})}{1 + \lambda_{it}} - \frac{\log(V_{it})}{\lambda_{it}} I_{\{y_{it}=0\}}\right),$$

with $\lambda_{it} = \exp(\mathbf{x}^1_{it}\boldsymbol{\alpha} + \mathbf{x}^2_{it}\boldsymbol{\beta}^s_i)$, whereas each component indicator r_{it} is sampled from following discrete distribution:

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

3.2.3 Multi-move Sampling of all Regression Parameters

In this subsection we provide details on multi-move sampling of $\boldsymbol{\beta}_1^s, \ldots, \boldsymbol{\beta}_N^s, \boldsymbol{\beta}, \boldsymbol{\alpha}$ from the posterior

$$p(\boldsymbol{\beta}_1^s, \ldots, \boldsymbol{\beta}_N^s, \boldsymbol{\beta}, \boldsymbol{\alpha}|\mathbf{y}^u, \mathbf{R}, \mathbf{Q}) = \prod_{i=1}^{N} p(\boldsymbol{\beta}_i^s|\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{y}^u, \mathbf{R}) p(\boldsymbol{\alpha}, \boldsymbol{\beta}|\mathbf{y}^u, \mathbf{R}, \mathbf{Q}). \tag{23}$$

First we sample $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ from the marginal posterior $p(\boldsymbol{\alpha}, \boldsymbol{\beta}|\mathbf{y}^u, \mathbf{R}, \mathbf{Q})$, where the random effects are integrated out, whereas we sample the random effects conditional on $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

For a fixed unit i, the marginal model is equal to a multivariate regression model,

$$\mathbf{y}_i^u = \mathbf{X}_i^1 \boldsymbol{\alpha} + \mathbf{X}_i^2 \boldsymbol{\beta} + \boldsymbol{m}_i + \boldsymbol{\varepsilon}_i, \quad \boldsymbol{\varepsilon}_i \sim \mathcal{N}_{T_i}(\mathbf{0}, \mathbf{V}_i), \tag{24}$$

using the matrix notation

$$\mathbf{y}_i^u = \begin{pmatrix} y_{i1}^u \\ \vdots \\ y_{i,T_i}^u \end{pmatrix}, \mathbf{X}_i^1 = \begin{pmatrix} \mathbf{x}_{i1}^1 \\ \vdots \\ \mathbf{x}_{i,T_i}^1 \end{pmatrix}, \mathbf{X}_i^2 = \begin{pmatrix} \mathbf{x}_{i1}^2 \\ \vdots \\ \mathbf{x}_{i,T_i}^2 \end{pmatrix}, \boldsymbol{m}_i = \begin{pmatrix} m_{r_{i1}} \\ \vdots \\ m_{r_{i,T_i}} \end{pmatrix}, \boldsymbol{\varepsilon}_i = \begin{pmatrix} \varepsilon_{i1} \\ \vdots \\ \varepsilon_{i,T_i} \end{pmatrix}$$

with regression parameter $(\boldsymbol{\alpha}, \boldsymbol{\beta})$, and error variance-covariance matrix $\mathbf{V}_i$ given by:

$$\mathbf{V}_i = \mathbf{X}_i^2 \mathbf{Q} (\mathbf{X}_i^2)' + \boldsymbol{D}_i, \quad \boldsymbol{D}_i = \mathrm{Diag}\left(s_{r_{i1}}^2, \ldots, s_{r_{i,T_i}}^2\right).$$

Assume a joint normal prior $\mathcal{N}_d(\mathbf{b}_0, \mathbf{B}_0)$ for $(\boldsymbol{\alpha}, \boldsymbol{\beta})$. Then the posterior $p(\boldsymbol{\alpha}, \boldsymbol{\beta}|\mathbf{y}^u, \mathbf{R}, \mathbf{Q})$ of $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is a joint normal distribution $\mathcal{N}_d(\mathbf{b}_N, \mathbf{B}_N)$, where

$$\mathbf{B}_N^{-1} = \mathbf{B}_0^{-1} + \sum_{i=1}^{N} (\mathbf{X}_i)' \mathbf{V}_i^{-1} \mathbf{X}_i, \quad \mathbf{b}_N = \mathbf{B}_N \left(\mathbf{B}_0^{-1} \mathbf{b}_0 + \sum_{i=1}^{N} (\mathbf{X}_i)' \mathbf{V}_i^{-1} (\mathbf{y}_i^u - \boldsymbol{m}_i) \right),$$
$$\mathbf{X}_i = \begin{pmatrix} \mathbf{X}_i^1 & \mathbf{X}_i^2 \end{pmatrix}.$$

For each $i = 1, \ldots, N$, the conditional posteriors $p(\boldsymbol{\beta}_i^s|\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{y}^u, \mathbf{R})$ are easily derived to be equal to normal density $\mathcal{N}(\boldsymbol{a}_i(\boldsymbol{\alpha}, \boldsymbol{\beta}), \boldsymbol{A}_i)$ with following posterior moments:

$$\boldsymbol{A}_i^{-1} = \mathbf{Q}^{-1} + (\mathbf{X}_i^2)' \boldsymbol{D}_i^{-1} \mathbf{X}_i^2, \quad \boldsymbol{a}_i(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \boldsymbol{A}_i \left(\mathbf{Q}^{-1} \boldsymbol{\beta} + (\mathbf{X}_i^2)' \boldsymbol{D}_i^{-1} (\mathbf{y}_i^u - \mathbf{X}_i^1 \boldsymbol{\alpha} - \boldsymbol{m}_i) \right).$$

3.3 Application to Austrian Wage Data

3.3.1 The Data

We consider a panel of Austrian employees who are observed between 1986 and 1998 on May 31st of each year. The data were obtained from the social security records in Austria (Weber 2001). The social security authority collects detailed data for all worker, but we use here only a random sample of $N = 4376$ individuals. We consider the variable y_{it}, which observes if individual i has zero-income in year t ($y_{it} = 0$) or not ($y_{it} = 1$), as dependent variable. Thus in this subsection we will consider only two states, namely wether an individual i has any income in year t or not, a wage variable y_{it} with more categories will be considered in Section 4.3.

The number of available individual characteristics is rather small and incomplete. In particular there is no information on education, working time or family affiliation. The covariates that are available are $\mathbf{x}_{it}^1 = (byearstd_i\, fem_i\, change_{it}\, whcollar_{it}\, y_{i,t-1})$ were:

$byearstd_i$	...	year of birth of the person (standardized over all observations)
fem_i	...	binary, 1 if the person is female, 0 otherwise
$change_{it}$	...	binary, 1 if the persons employers in year t and $t-1$ are different, 0 otherwise
$whcollar_{it}$	...	binary, 1 if the person is white-collar employee, 0 otherwise
$y_{i,t-1}$	...	binary, 1 if person i had nonzero income in year $t-1$

3.3.2 A Binary Logistic Model with Overdispersion

To analyze these data, we will consider a binary logit regression model which captures overdispersion due to omitted covariates. A common way of dealing with this kind of overdispersion is the individual effects model, see Aitkin (1996), where the regression intercept varies between the units:

$$\text{logit}\Pr(y_i = 1) = \beta_i^s + \mathbf{x}_{it}^1 \boldsymbol{\alpha}, \tag{25}$$

where $\beta_i^s \sim \mathcal{N}(\beta, \sigma_\alpha^2)$. Thus overdispersion is modelled on the same level as the linear predictor. Note that $\mathbf{x}_{it}^2 = 1$. Marginally, this model is an infinite mixture of logistic regression models with no closed form. Aitkin (1996) suggested to approximate the marginal distribution by a mixture of logit regression models using Gaussian-Hermite quadrature. Our data augmentation scheme leads to a normal random-effects regression model, where the whole sequence $(\beta_1^s, \dots, \beta_N^s, \beta, \boldsymbol{\alpha})$ could be sampled simultaneously in an efficient manner.

3.3.3 Bayesian Posterior Inference

To show that multi-move sampling has considerable effect on the efficiency of the MCMC sampler, we compare the multi-move Gibbs sampler introduced in Subsection 3.2.2 with an alternative Gibbs sampler, where $\boldsymbol{\alpha}$ and β are sampled conditional on knowing $\beta_1^s, \dots, \beta_N^s$, rather than from the marginal density. In our example both Gibbs-samplers, i.e. the 2-step sampler as described in Subsection 3.2.2 and the marginal sampler, where the random effects are integrated out, perform quite well. Anyway the marginal Gibbs-sampler has better mixing properties and a shorter burn-in phase (see Figure 1). Furthermore the autocorrelation of the marginal Gibbs-sampler is clearly less than the autocorrelation of the 2-step Gibbs-sampler (see Figure 2).

The parameter-estimates, standard deviations and 95% credible regions have been computed for both samplers after cutting off the first 1000 simulations. The results for the fixed parameters $\boldsymbol{\alpha}$ are given in Table 2. The estimates are very similar both for the 2-step- and the marginal Gibbs-sampler. Apart from age, all other covariates have a significant influence on the probability of having a non-zero income. The strongest influence on the probability of having a non-zero income is given by a person's immediate income history. For two persons with different income history, which otherwise share identical values of $(byearstd_i\, fem_i\, change_{it}\, whcollar_{it})$, the

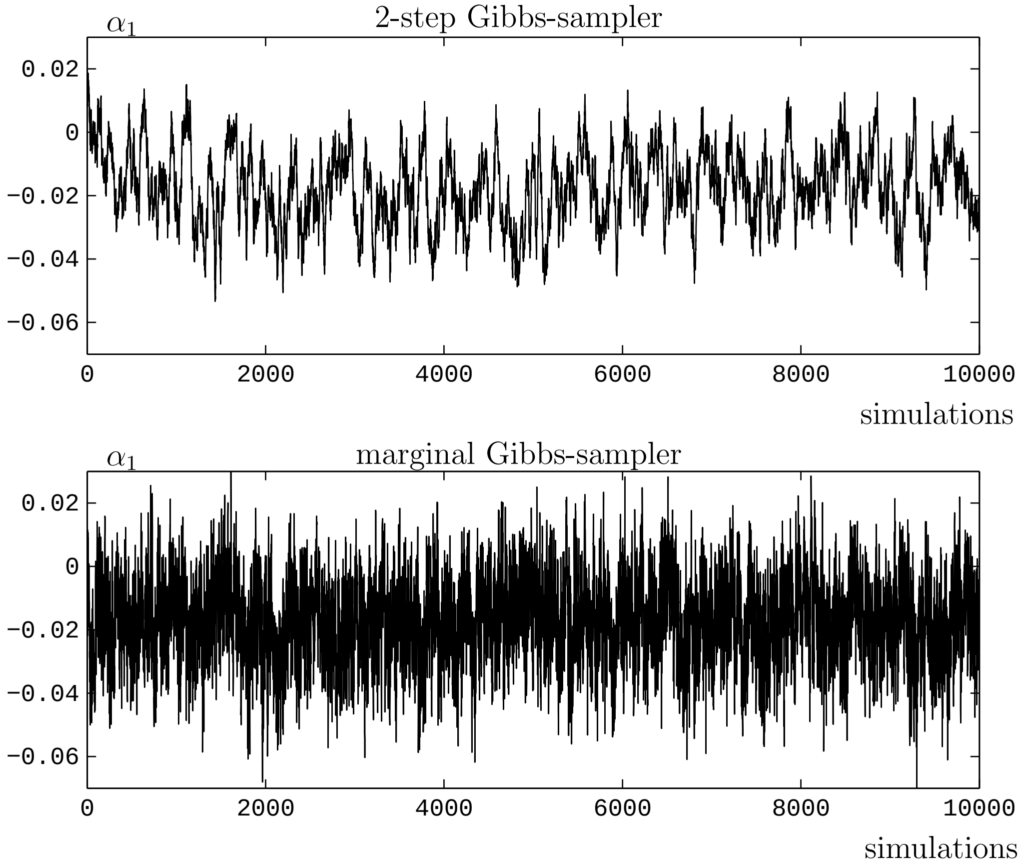

Figure 1: Simulated values of α_1 with 2-step- and marginal Gibbs-sampler

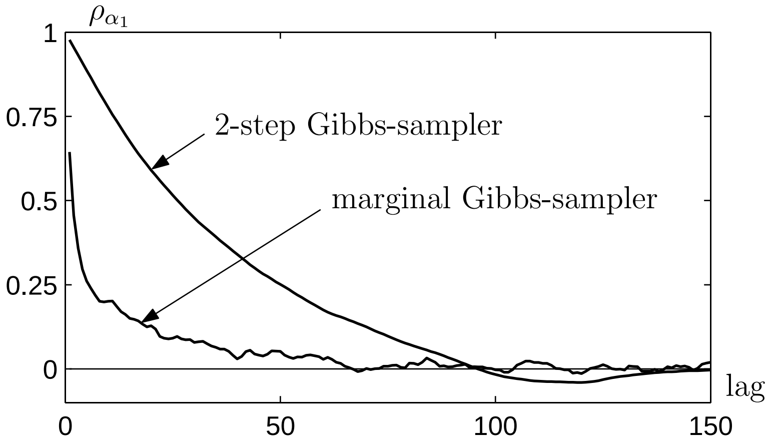

Figure 2: Autocorrelation of the simulated α_1-values of 2-step- and marginal Gibbs-sampler

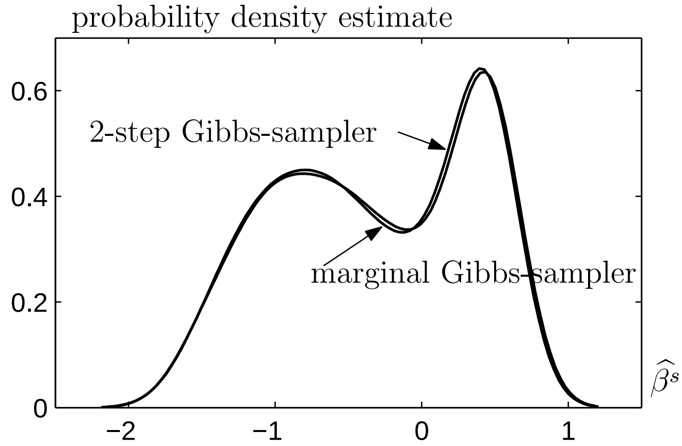

Figure 3: kernel density estimate of $\widehat{\beta}^s$ with 2-step- and marginal Gibbs-sampler

Table 2: Parameter estimates of the 2-step- and marginal Gibbs-sampler

	2-step sampler			marginal sampler		
	mean	std.dev.	95% credible region	mean	std.dev.	95% credible region
α_1	-.01829	0.01162	[-0.04107; 0.003517]	-.01742	0.01384	[-0.04472; 0.009462]
α_2	-0.5208	0.02119	[-0.5617; -0.4797]	-0.5165	0.02391	[-0.5638; -0.4698]
α_3	-0.3494	0.02756	[-0.4016; -0.2913]	-0.3416	0.02612	[-0.3934; -0.2918]
α_4	-0.3427	0.02312	[-0.3886; -0.2962]	-0.3358	0.02587	[-0.3868; -0.2864]
α_5	3.416	0.02106	[3.379; 3.461]	3.364	0.0327	[3.301; 3.43]

odd ratio of having income versus having no income in year t is $e^{3.364} \approx 29$ times larger for a person with non-zero income in year $t-1$ than for a person with no income in year $t-1$. For two persons with different gender, which otherwise share identical values of $(byearstd_i\, change_{it}\, whcollar_{it}\, y_{i,t-1})$, being a women rather than a man reduces the odd ratio of having income versus having no income by the factor $e^{-0.5165} \approx 0.6$.

Figure 3 shows the empirical distribution of $\widehat{\beta}_i^s$, which for each person is estimated as the mean over all MCMC draws, after cutting off the first 1000 simulations. Interestingly the posterior distribution of the subject-specific parameter-estimates over the population is a mixture distribution, which two groups of employee. Given identical covariates $\mathbf{x}_{it}^1$, for one group the expected value of β_i^s lies significantly above zero, whereas for a second group the expected value of β_i^s lies significantly below zero.

4 Multinomial Logit Models

4.1 The Multinomial Logit Regression Model

4.1.1 Background

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

$$\Pr(y_i = k | \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m) = \frac{\exp(\mathbf{x}_i \boldsymbol{\beta}_k)}{1 + \sum_{l=1}^{m} \exp(\mathbf{x}_i \boldsymbol{\beta}_l)}, \tag{26}$$

where $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m$ are category specific, unknown parameters. To make the model identifiable, the parameter $\boldsymbol{\beta}_0$ of the baseline category $k = 0$ is set equal to 0: $\boldsymbol{\beta}_0 = 0$. Finally, we assume that conditional on knowing $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m$, the observations are mutually independent.

4.1.2 Data Augmentation for Multinomial Logit Models

As for binary models, we consider two data augmentation steps. The first data augmentation step was suggested by Scott (2004) and involves the well-known interpretation of a logit-model in terms of utilities as introduced by McFadden (1974). The latent utility y_{ki}^u of observing the category k for observation y_i is modelled as being dependent on covariates:

$$\begin{aligned} y_{1i}^u &= \mathbf{x}_i \boldsymbol{\beta}_1 + \varepsilon_{1i}, \\ &\cdots \\ y_{mi}^u &= \mathbf{x}_i \boldsymbol{\beta}_m + \varepsilon_{mi}, \end{aligned} \tag{27}$$

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

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

It was shown by McFadden (1974), that if $\varepsilon_{ki}, k = 1, \ldots, m$, and y_{0i}^u follow a type I extreme value distribution, the multinomial logit model (26) results as the marginal distribution of y_i.

The first data augmentation step introduces for each categorical observation y_i the latent utilities $\mathbf{y}_i^u = (y_{1i}^u, \ldots, y_{mi}^u)$ as missing data as in Scott (2004). Conditional on $\mathbf{y}_i^u$, we are dealing with the linear model (27), rather than with the non-linear model (26). Scott (2004) uses this result to define multivariate proposal densities with a Metropolis-Hastings algorithm. In this paper, we obtain a model that is conditionally Gaussian by approximating the non-normal density of ε_{ki}, for $k = 1, \ldots, m$, by a normal mixture as above. The second step of our data augmentation scheme introduces for each ε_{ki} the latent component indicator r_{ki} as missing data.

4.1.3 Gibbs Sampling

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

$$y^u_{ki} = \mathbf{x}_i \boldsymbol{\beta}_k + m_{r_{ki}} + s_{r_{ki}} \varepsilon_{ki}, \quad \varepsilon_{ki} \sim \mathcal{N}(0, 1). \tag{28}$$

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

(a) Independent sampling of $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_m$ conditional on knowing $\mathbf{y}^u$ and $\mathbf{R}$, based on the Gaussian regression model (28).

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

Step (a) is carried out in an obvious manner. Step (b) extends the results of Subsection 2.3 to more than two categories. The joint posterior $p(\mathbf{R}, \mathbf{y}^u | \mathbf{y}, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_m)$ is decomposed as:

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

The component indicator r_{ki} is sampled from:

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

To sample from $p(y^u_{1i}, \dots, y^u_{mi} | y_i, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_m)$, we sample from the augmented posterior $p(y^u_{0i}, \dots, y^u_{mi} | y_i, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_m)$. For fixed i, the latent utilities $y^u_{0i}, \dots, y^u_{mi}$, are stochastically dependent, and the joint distribution factorizes as, see Scott (2004):

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

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

$$\begin{aligned} \exp(-y^u_{0i}) &\sim \mathcal{E}(\lambda_{0i}), \\ \exp(-y^u_{1i}) &\sim \mathcal{E}(\lambda_{1i}), \\ &\cdots \\ \exp(-y^u_{mi}) &\sim \mathcal{E}(\lambda_{mi}), \end{aligned} \tag{29}$$

where $\lambda_{0i} = 1$, and $\lambda_{ki} = \exp(\mathbf{x}_i \boldsymbol{\beta}_k)$, for $1 \leq k \leq m$. Given $y_i = k$, y^u_{ki} is known to the maximal utility. Thus $\exp(-y^u_{ki})$ is the minimum among all random variables appearing in (29), and therefore:

$$\exp(-y^u_{ki}) \sim \mathcal{E}\left(1 + \sum_{l=1}^{m} \lambda_{li}\right). \tag{30}$$

Given the minimum, all other utilities are conditionally independent:

$$\exp(-y^u_{li}) = \exp(-y^u_{ki}) + \xi_{li}, \quad \xi_{li} \sim \mathcal{E}\left(\lambda_{li}\right), \tag{31}$$

where $l = 1, \ldots, m, l \neq k$. Therefore to sample y^u_{li}, we simply need two independent uniform random numbers U_{li} and V_{li}:

$$y^u_{li} = -\log\left(-\frac{\log(U_{li})}{1 + \sum_{k=1}^m \lambda_{ki}} - \frac{\log(V_{li})}{\lambda_{li}} I_{\{y_i \neq l\}}\right), \tag{32}$$

where $l = 1, \ldots, m$, and $i = 1, \ldots, N$.

4.2 Multinomial Logit Models with Random-Effects

4.2.1 Background

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

For category k, with $1 \leq k \leq m$, the probability that y_{it} takes the category k depends on covariates $\mathbf{x}_{it} = (\mathbf{x}^1_{it}\,\mathbf{x}^2_{it})$ through fixed category specific parameters $\boldsymbol{\alpha}_k$ and subject-specific random category parameters $\boldsymbol{\beta}^s_{ki}$ in the following way:

$$\Pr(y_{it} = k|\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_m, \boldsymbol{\beta}^s_{1i}, \ldots, \boldsymbol{\beta}^s_{mi}) = \frac{\exp(\mathbf{x}^1_{it}\boldsymbol{\alpha}_k + \mathbf{x}^2_{it}\boldsymbol{\beta}^s_{ki})}{1 + \sum_{l=1}^m \exp(\mathbf{x}^1_{it}\boldsymbol{\alpha}_l + \mathbf{x}^2_{it}\boldsymbol{\beta}^s_{li})}. \tag{33}$$

To make the model identifiable, the parameters of the baseline category $k = 0$ are set equal to 0: $\boldsymbol{\alpha}_0 = 0$, $\boldsymbol{\beta}^s_{0i} = 0, i = 1, \ldots, N$. Finally, we assume that conditional on knowing all $\boldsymbol{\beta}^s_{ki}$ and $\boldsymbol{\alpha}_k$, the observations are mutually independent. A commonly used prior for $\boldsymbol{\beta}^s_{ki}$ reads:

$$\boldsymbol{\beta}^s_{ki} \sim \mathcal{N}_d\left(\boldsymbol{\beta}_k, \mathbf{Q}\right). \tag{34}$$

4.2.2 Data Augmentation and Gibbs Sampling

The first data augmentation step introduces for each subject i the latent utilities $y^u_{kit}, k = 1, \ldots, m$, of choosing category k at time t. Then

$$\begin{aligned} y^u_{1it} &= \mathbf{x}^1_{it}\boldsymbol{\alpha}_1 + \mathbf{x}^2_{it}\boldsymbol{\beta}^s_{1i} + \varepsilon_{1it}, \\ &\cdots \\ y^u_{mit} &= \mathbf{x}^1_{it}\boldsymbol{\alpha}_m + \mathbf{x}^2_{it}\boldsymbol{\beta}^s_{mi} + \varepsilon_{mit}, \end{aligned} \tag{35}$$

where $\varepsilon_{kit}, k = 1, \ldots, m$ follows a type I extreme value distribution. The second step of our data augmentation scheme, approximates the type I extreme value distribution by a mixture of univariate normal distributions, and introduces for each ε_{kit} the latent component indicator r_{kit} as missing data.

Let $\mathbf{R} = \{r_{kit}, i = 1, \ldots, N, t = 1, \ldots, T, k = 1, \ldots, m\}$ denote the collection of all component indicators and the $\mathbf{y}^u = \{y^u_{1it}, \ldots, y^u_{mit}, i = 1, \ldots, N, t = 1, \ldots, T\}$ denote the collection of all latent propensities. Select a starting value for the unknown model parameter $\mathbf{Q}$, the component indicators $\mathbf{R}$ and the latent propensities $\mathbf{y}^u$. A three block Gibbs sampler can easily implemented, which consists of the following steps:

(a) Multi-move sampling of $(\boldsymbol{\beta}^s_{k1}, \dots, \boldsymbol{\beta}^s_{kN}, \boldsymbol{\beta}_k, \boldsymbol{\alpha}_k), k = 1, \dots, m$, conditional on knowing $\mathbf{y}^u$, $\mathbf{R}$, and $\mathbf{Q}$, based on the conditional Gaussian linear random-effects model (35).

(b) Sampling of $\mathbf{Q}$ conditional on knowing $(\boldsymbol{\beta}^s_{k1}, \dots, \boldsymbol{\beta}^s_{kN}, \boldsymbol{\beta}_k), k = 1, \dots, m$, based on (34).

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

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

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

To sample y^u_{kit}, we simply need two independent uniform random numbers U_{kit} and V_{kit}:

$$y^u_{kit} = -\log \left(-\frac{\log(U_{kit})}{1 + \sum_{l=1}^{m} \lambda_{lit}} - \frac{\log(V_{kit})}{\lambda_{kit}} I_{\{y_{it} \neq k\}} \right), \tag{36}$$

where $\lambda_{kit} = \exp(\mathbf{x}^1_{it}\boldsymbol{\alpha}_k + \mathbf{x}^2_{it}\boldsymbol{\beta}^s_{ki})$, whereas each component indicator r_{kit} is sampled from a discrete distribution with $j = 1, \dots, M$ categories:

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

4.3 Application to the Austrian Labor Market

4.3.1 The Data

We reanalyze the data of Subsection 3.3, with the same wage categories as in Weber (2001). The wage of individual i in year t is modelled as a categorical variable y_{it} with states $k \in \{0, 1, \dots, 5\}$, where category 0 corresponds to the no-income class. Non-zero wage data were categorized according to the quintiles of the yearly wage distribution into 5 income classes, coded as 1 to 5. For $t = 0, \dots, T$, y_{it} takes the value k, if person i belonged to wage category k at time t. The covariates are $\mathbf{x}^1_{it} = (byearstd_i,\ fem_i,\ change_{it},\ whcollar_{it},\ I_{\{y_{i,t-1}=1\}}, \dots, I_{\{y_{i,t-1}=5\}})$ where the first four covariates have the same meaning as in Subsection 3.3, whereas $I_{\{y_{i,t-1}=l\}}$ captures the immediate income history of each person and takes 1 iif $y_{i,t-1} = l$. To account for unobserved heterogeneity, we fit the multinomial logit model with random effects defined in (33), where $\mathbf{x}^2_{it} = 1$. Thus a random intercept β^s_{ki} is introduced for each employee for each wage category $k = 1, \dots, 5$.

4.3.2 Bayesian Posterior Inference

As in the binary example both Gibbs-samplers, i.e. the 2-step sampler as described in Subsection 4.2.2 and the marginal sampler, where the random effects are integrated out, were applied. Again the marginal Gibbs-sampler has better mixing properties and a shorter burn-in phase. Furthermore the autocorrelation of the marginal Gibbs-sampler is clearly less than the autocorrelation of the 2-step Gibbs-sampler.

The parameter-estimates, standard deviations and 95% credible regions have been computed for both samplers after cutting off the first 1000 simulations. The results for the fixed parameters $\boldsymbol{\alpha}$ obtained from the marginal sampler are given in Table 3. The k-th column of this table corresponds to the effect of the different covariates on the probability to be in wage category k. Again we find a strong influence of a person's wage history on the odds of being in wage category k as opposed to be in any other wage category. For two persons, which share identical values of $(byearstd_i\, fem_i\, change_{it}\, whcollar_{it})$, the odds of being in wage category k as opposed to be in any other wage category in year t, is between $e^{1.26} \approx 3.5(k = 1)$ and $e^{1.66} \approx 5.3(k = 5)$ times larger for a person with the same wage category in year $t-1$ than for a person with a different wage category in year $t-1$. This indicates considerable wage immobility in the Austrian labor market. Again gender has a considerable effect. For each non-zero wage category, being a women rather than being a man reduces the chance of belonging to this wage category. This negative effect of gender increase with increasing wage category. For two persons with different gender, which otherwise share identical values of $(byearstd_i\, change_{it}\, whcollar_{it}\, y_{i,t-1})$, being a women rather than a man reduces the odd ratio of belonging to the highest income class versus belonging to any other income class by the factor $e^{-0.665} \approx 0.51$.

Also in this example it is worth while to take a closer look at the distributions of the estimates $\widehat{\beta^s_{ki}}$. First, we show the mean $\widehat{\beta^s_k}$ of all $\widehat{\beta^s_{ki}}$ for $k = 1, \dots, 5$ in Table 3. Next, Figure 4 estimates the empirical distributions of $\widehat{\beta^s_{ki}}$, $k = 1 \dots 5$, over the individuals, by a histogram, whereas the scatter plots in Figure 4 show all 10 2-dimensional empirical distributions of $(\widehat{\beta^s_{ki}}, \widehat{\beta^s_{mi}})$, $1 \leq k < m \leq 5$ over the individuals. Apparently these distributions not are very different across the categories, which suggest that a simplified models, where $\beta^s_{ki} \equiv \beta^s_i$ for all wage categories, might be a sensible simplification of this model.

5 Concluding Remarks

In this paper we introduced a new data augmentation algorithm for sampling the parameters of a binary or multinomial logit model from their posterior distribution within a Bayesian framework. The algorithm leads to a convenient Gibbs sampler that draws from standard distributions like normal or exponential distributions and does not require any tuning. This Gibbs sampler can be easily implemented for any binary or multinomial logit model, where the predictor is linear in the unknown parameters, with covariates being categorical as well as continuous. We gave details for standard regression models as well as for random effect and time-varying parameter models. Extension to more complex models including logistic components are straightforward.

Whereas to our knowledge, so far Gibbs sampling has been unfeasible for logit models, it has been known for a long while how to implement Gibbs sampling for the alternative probit model, see in particular Albert and Chib (1993) and McCulloch and Rossi (1994). This technical advantage of the probit over the logit model partly explain why most of the Bayesian analysis of binary and categorical data is based on the probit model. With the new Gibbs sampler for logit model discussed in this paper, the technical superiority of the probit model is no longer prevalent, and we

Table 3: Parameter estimates of the marginal Gibbs-sampler

mean (std.dev.)	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
α_{k1}	0.04695	0.07234	0.03718	0.01291	-0.0427
$(byearstd)$	(0.01422)	(0.01413)	(0.01421)	(0.01432)	(0.01414)
α_{k2}	-0.1367	-0.383	-0.54	-0.5854	-0.6654
(fem)	(0.02703)	(0.02811)	(0.02615)	(0.027)	(0.02717)
α_{k3}	0.5949	0.3623	0.1006	-0.01627	-0.04776
$(change)$	(0.02779)	(0.0268)	(0.02858)	(0.02897)	(0.02887)
α_{k4}	-0.4846	-0.495	-0.4236	-0.3033	-0.1319
$(whcollar)$	(0.02769)	(0.0283)	(0.02807)	(0.027)	(0.02913)
α_{k5}	1.26	0.3586	0.1782	0.1134	0.08114
$(I_{\{y_{i,t-1}=1\}})$	(0.02974)	(0.03447)	(0.03686)	(0.03838)	(0.03864)
α_{k6}	0.0148	1.439	0.5886	0.1404	0.07939
$(I_{\{y_{i,t-1}=2\}})$	(0.03584)	(0.03176)	(0.0338)	(0.03705)	(0.03622)
α_{k7}	-0.1052	0.2683	1.511	0.5532	0.07822
$(I_{\{y_{i,t-1}=3\}})$	(0.03688)	(0.03751)	(0.03097)	(0.03609)	(0.03913)
α_{k8}	-0.08991	0.0838	0.385	1.601	0.3921
$(I_{\{y_{i,t-1}=4\}})$	(0.042)	(0.03887)	(0.03608)	(0.03209)	(0.03593)
α_{k9}	-0.04503	0.125	0.2105	0.3465	1.661
$(I_{\{y_{i,t-1}=5\}})$	(0.04369)	(0.04206)	(0.03997)	(0.04118)	(0.03612)
$\widehat{\beta}_k^s$	-0.2755	-0.2293	-0.2372	-0.2566	-0.3149

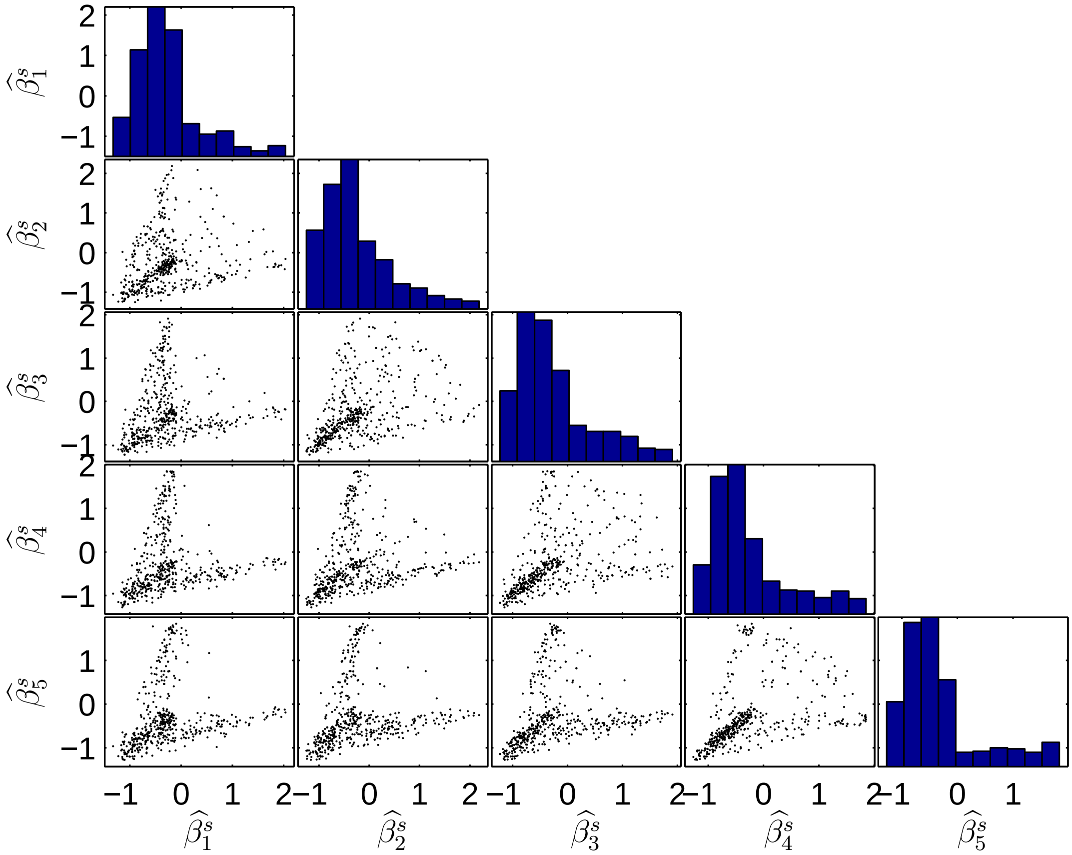

Figure 4: histograms and scatter plots of $\widehat{\beta}_i^s$

hope that more principled approaches of comparing logit and probit models, like Bayes factors, will lead to more data orienting decision concerning the choice of the appropriate link function.

References

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

Albert, J. H. (1992). A Bayesian analysis of a Poisson random-effects model. *American Statistician 46*, 246–253.

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

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

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

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

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

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

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

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

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

Frühwirth-Schnatter, S. and H. Wagner (2004). Gibbs sampling for parameter-driven models of time series of small counts with applications to state space modelling. Research Report IFAS, http://www.ifas.jku.at/.

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

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

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

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

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

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

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

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

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

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

Tanner, M. A. and W. H. Wong (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association 82*, 528–540.

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

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

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

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

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