Bayesian Variable Selection for Logistic Models Using Auxiliary Mixture Sampling

Tüchler, Regina

Published: 01/10/2006

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

Bayesian Variable Selection for Logistic Models Using Auxiliary Mixture Sampling

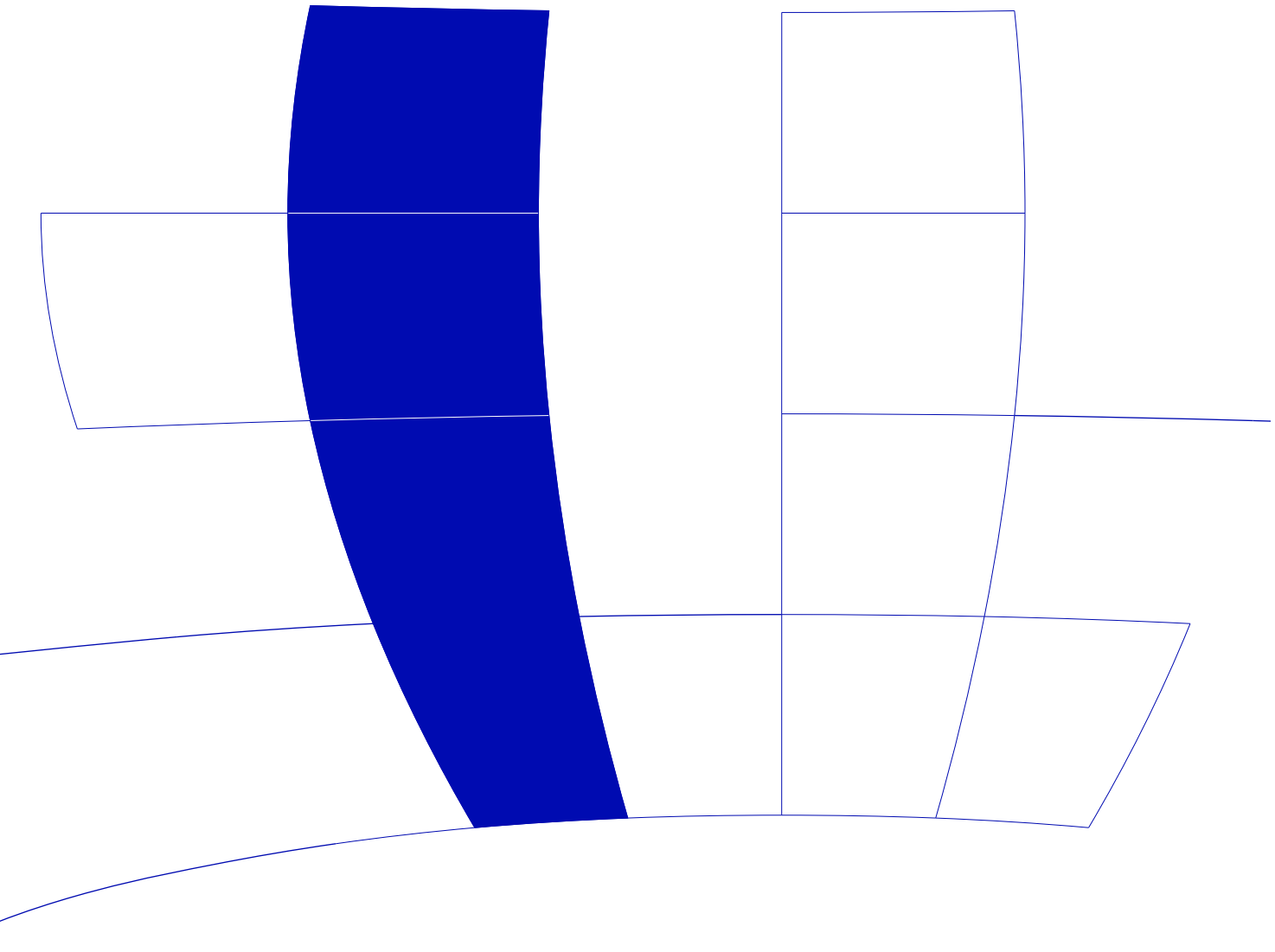
Regina Tüchler

Department of Statistics and Mathematics
Wirtschaftsuniversität Wien

Research Report Series

Report 31
March 2006

http://statistik.wu-wien.ac.at/



Bayesian Variable Selection for Logistic Models Using Auxiliary Mixture Sampling

Regina Tüchler*

January 19, 2006

Abstract

The paper presents an Markov Chain Monte Carlo algorithm for both variable and covariance selection in the context of logistic mixed effects models. This algorithm allows us to sample solely from standard densities, with no additional tuning being needed. We apply a stochastic search variable approach to select explanatory variables as well as to determine the structure of the random effects covariance matrix.

For logistic mixed effects models prior determination of explanatory variables and random effects is no longer prerequisite since the definite structure is chosen in a data-driven manner in the course of the modeling procedure. As an illustration two real-data examples from finance and tourism studies are given.

*Regina Tüchler is an Assistant Professor, Department of Statistics and Mathematics, Vienna University of Economics and Business Administration, Austria (E-mail: regina.tuechler@wu-wien.ac.at)

KEY WORDS: Covariance Selection, Markov Chain Monte Carlo, Mixed Effects Model, Parsimony

1 INTRODUCTION

In this paper we present an MCMC algorithm for variable and covariance selection in logistic mixed effects models. Bayesian variable selection methods make it possible to start with a very general model specification. The special structure is then chosen in the course of the modeling procedure, according to the principle of parsimony. For the estimation of the logistic mixed effects models presented in the paper this has two aspects. Firstly, we select explanatory variables in the regression from a possibly high number of covariates. Secondly, we determine zero and non-zero elements in the variance-covariance matrix of the Gaussian random effects. This also comprises model selection with regard to fixed versus random effects. Our algorithm samples from *standard densities* only. This is obtained by combining methods for variable selection (George and McCulloch (1993, 1997)) and covariance selection (Smith and Kohn (2002)) with recent suggestions for estimating logistic models (Scott (2005), Frühwirth-Schnatter and Frühwirth (2005)).

Stochastic search variable selection was introduced by George and McCulloch (1993, 1997) for normal linear regression models. We transfer their approach to logistic models. Indicators are used to determine zero and non-zero regression coefficients. These indicators are estimated together with the non-zero coefficients and all other model parameters. Stochastic search

variable procedures may be viewed in the context of Bayesian model averaging, see Raftery et al. (1997) and Hoeting et al. (1999), where the posterior estimates of the indicator vectors take the position of posterior model probabilities. An alternative approach is taken by Nott and Green (2004) and Nott and Leonte (2004) who combine the Swendson-Wang algorithm with a reversible jump method to select variables in linear and generalized linear models, respectively.

To carry out covariance selection we need to decompose the random effects covariance matrix. Estimation of variance-covariance matrices on the basis of different decompositions with varying estimation properties are an extensive field of research. One convenient way, which we also follow in this paper, is to choose a Cholesky decomposition. Such an approach may be found for example in Pinheiro and Bates (1996) or in Pourahmadi (1999, 2000), who apply a Cholesky decomposition to the inverse of the variance-covariance matrix. Cholesky decompositions in the context of hierarchical models are used in Lindstrom and Bates (1988) and Meng and van Dyk (1998), among others. Parsimonious representations of variance-covariance matrices were studied by Dempster (1972) and within a Bayesian framework by Smith and Kohn (2002). Covariance selection for hierarchical models was realized in Albert and Chib (1997) and Chen and Dunson (2003).

Bayesian covariance selection may be achieved by first decomposing the variance-covariance matrix and then by applying the variable selection steps of George and McCulloch (1993, 1997) to components of this decomposition. This was first observed by Smith and Kohn (2002) who applied these ideas to the inverse of the variance-covariance matrix of Gaussian time series data.

They selected zero and non-zero *off-diagonal* elements, whereas the diagonal elements were always fixed to non-zeros. Rank reduction is neither possible nor desired for the time series context of their approach. Chen and Dunson (2003) included covariance selection in mixed effects models for the Gaussian data case. They selected *whole rows (and corresponding columns)* as zero or non-zero in the random effects covariance matrix. Non-zero rows (columns) correspond to random effects then, whereas effects with zero rows (columns) in the variance-covariance matrix are determined as fixed effects. For non-zero diagonal elements all the off-diagonal elements are automatically included as non-zeros in the model. It is not possible to determine a finer structure of zeros and non-zeros for these off-diagonal elements with their method.

In the present paper we choose an approach, which makes it possible to select *all* elements of the variance-covariance matrix. If we obtain a zero diagonal element the rest of this row (column) is automatically determined as zero by the algorithm. The corresponding effect is a fixed effect like in Chen and Dunson (2003). But additionally to that it is also possible to obtain zero off-diagonal elements (covariances) for effects which have a non-zero diagonal element (variance) like in Smith and Kohn (2002). The Cholesky factors of the decomposition of the random effects covariance matrix appear as regression coefficients in the observation equation of the model. Covariance selection therefore amounts to selection of zero and non-zero elements in these Cholesky factors and is carried out together with the selection of the other explanatory variables by means of common variable selection tools, see George and McCulloch (1993, 1997). Only the non-zero parameters are

included in the resulting parsimonious representation of the logistic mixed effects model.

Until recently Bayesian estimation of logistic mixed effects models could not be done by pure sampling from standard densities. Some of the references are the seminal work of Zeger and Karim (1991) or Lenk and DeSarbo (2000) for mixtures of generalized linear mixed effects models. Scott (2005) made an important contribution to the sampling of logistic models by augmenting the model parameter vector by latent utilities for choosing the categories. The resulting augmented model is then a *linear* model. Frühwirth-Schnatter and Frühwirth (2005) took this up and added indicators of an auxiliary normal mixture distribution in a second data augmentation step, thus obtaining a linear model with *normally* distributed data. For the resulting augmented model they could build a sampling algorithm which involves standard densities only. Recently Holmes and Held (2006) presented another auxiliary variable approach to sample logistic models fully automatically. In our paper we adapt the ideas of Scott (2005) and Frühwirth-Schnatter and Frühwirth (2005) to the parsimonious logistic mixed effects model and thereby obtain a sampling scheme which samples from standard densities only.

With the present paper we contribute to research in various respects. Firstly, our algorithm makes it possible to carry out variable selection for logistic mixed effects models by sampling from standard densities only. This is especially convenient since no additional tuning is needed. Secondly, we include random effects in the logistic model and determine their covariance structure, again by sampling from standard densities only. Thereby we also

include model selection with respect to fixed versus random effects.

The paper is structured as follows. Section 2 shows how variable selection can be incorporated in the auxiliary mixture sampler of the logistic model. Here only the special case of binary data and *fixed* effects is being dealt with, whereas Section 3 introduces *random* effects and the covariance selection. Section 4 extends the algorithm to logistic models with more that two data categories. Section 5 gives two real-data examples. Section 6 summarizes the results.

2 BAYESIAN VARIABLE SELECTION FOR THE BINARY LOGIT REGRESSION MODEL

2.1 The Model

Let the binary data y_{it} take two possible values labelled $\{0, 1\}$. We observe data for subjects $i = 1, \ldots, N$ and for $t = 1, \ldots, T_i$ repetitions per subject. The binary logit model with fixed effects parameter $\theta = (\theta_1, \ldots, \theta_d)$ and design vector w_{it} writes

$$P(y_{it} = 1|\theta) = \frac{\exp(w_{it}\theta)}{1 + \exp(w_{it}\theta)}. \tag{1}$$

To select non-zero elements of θ it is common to introduce an indicator vector $\delta = (\delta_1, \ldots, \delta_d)$ (George and McCulloch (1997)). The indicators δ_k $(k = 1, \ldots, d)$ take the value 1, if the corresponding effect θ_k is unequal to 0, and takes the value 0 otherwise. Let θ^δ include only those effects, which are selected as non-zero effects. w_{it}^δ denotes the corresponding design

matrix, i.e. w_{it}^{δ} is derived by eliminating those elements from w_{it}, which have regression parameters equal to zero.

The logit model with variable selection is given by

$$P(y_{it} = 1|\theta^{\delta}) = \frac{\exp(w_{it}^{\delta}\theta^{\delta})}{1 + \exp(w_{it}^{\delta}\theta^{\delta})}, \tag{2}$$

where the observations are assumed to be independent conditional on knowing θ^{δ}.

2.2 Data Augmentation

We introduce two data augmentation steps into the binary logit regression model (2) to obtain a linear regression model with normally distributed data.

The first data augmentation step removes the non-linearity of model (2) by defining for each observation latent utilities for choosing category 0 and 1, respectively. This step was first introduced by Scott (2005) and involves the interpretation of logit models in terms of latent utilities by McFadden (1974). The utilities for choosing category 0 are denoted y_{it}^{0}. They follow a standard type I extreme value distribution and are therefore independent of any model parameters. The utilities for choosing category 1 are denoted y_{it}^{u}. These utilities depend on the covariates and model parameters through

$$y_{it}^{u} = w_{it}^{\delta}\theta^{\delta} + \varepsilon_{it}, \tag{3}$$

where the error term ε_{it} follows a standard type I extreme value distribution. The following relationship holds: $y_{it} = 1$ iff $y_{it}^{u} > y_{it}^{0}$, whereas $y_{it} = 0$ iff $y_{it}^{u} \leq y_{it}^{0}$.

Model (3) is linear with respect to the regression parameters. To remove the non-normality of the error term we approximate the type I extreme value

Table 1: Components of the normal mixture approximation of the type I extreme value error ε_{it}.

r	1	2	3	4	5	6	7	8	9	10
η_r	.004	.040	.168	.147	.125	.101	.104	.116	.107	.088
m_r	5.09	3.29	1.82	1.24	.76	.39	.04	-.31	-.67	-1.06
s_r^2	4.5	2.02	1.1	.42	.2	.11	.08	.08	.09	.15

error ε_{it} by the following mixture of normal distributions:

$$p(\varepsilon_{it}) = \exp(-\varepsilon_{it} - e^{-\varepsilon_{it}}) \approx \sum_{r=1}^{10} \eta_r f_{\mathrm{N}}(\varepsilon_{it}; m_r, s_r^2). \tag{4}$$

Such an approximation was proposed by Frühwirth-Schnatter and Frühwirth (2005) and applied in Frühwirth-Schnatter and Wagner (2005) in the context of log linear models. The values of the ten mixture components are given in Table 1.

In the second data augmentation step we introduce for each error term ε_{it} an indicator $r_{it} \in \{1, \ldots, 10\}$. Given these indicators the model errors are normally distributed with $\varepsilon_{it} \sim \text{Normal}\left(m_{r_{it}}, s_{r_{it}}^2\right)$.

After adding all utilities $y^u = \{y_{it}^u\}_{i=1,\ldots,N,t=1,\ldots,T_i}$ and all indicators $R = \{r_{it}\}_{i=1,\ldots,N,t=1,\ldots,T_i}$ we obtain the following multiple regression model with heteroscedastic errors

$$y_{it}^u = w_{it}^\delta \theta^\delta + m_{r_{it}} + \varepsilon_{it}, \quad \varepsilon_{it} \sim \text{Normal}\left(0, s_{r_{it}}^2\right). \tag{5}$$

To rewrite this model in matrix notation we denote for all $i = 1, \ldots, N, t = 1, \ldots, T_i$: $W^\delta = \{w_{it}^\delta\}$, $\varepsilon = \{\varepsilon_{it}\}$, $m = \{m_{it}\}$ and the diagonal matrix $\Sigma = \text{diag}\{s_{it}^2\}$. The augmented model takes the form

$$y^u = W^\delta \theta^\delta + m + \varepsilon, \quad \varepsilon \sim \text{Normal}\left(0, \Sigma\right). \tag{6}$$

2.3 The Sampler

2.3.1 MCMC Scheme

To estimate the model parameters we iterate between the following MCMC sampling steps:

(i) Sample each element δ_k of the indicator vector δ separately conditional on $\delta_{\backslash k}$ (all other elements of δ), R and y^u, for $k = 1, \ldots, d$.

(ii) Sample the non-zero elements θ^δ conditional on δ, R and y^u.

(iii) Sample R conditional on θ^δ and y^u.

(iv) Sample y^u conditional on θ^δ and the binary data $y = \{y_{it}\}_{i=1,\ldots,N, t=1,\ldots,T_i}$.

Conditional on the utilities y^u and the indicators R steps (i) and (ii) amount to including variable selection in the estimation of the Gaussian linear model with heteroscedastic errors (6).

Conditional on knowing θ^δ steps (iii) and (iv) amount to sampling the utilities y^u and the indicators R in a binary logit regression model.

In what follows we give details for these sampling steps.

2.3.2 Sampling the non-zero regression parameters θ^δ

Fractional prior for θ^δ One possible choice of a prior for the non-zero regression parameters θ^δ is a conditionally conjugate normal prior: $p(\theta^\delta) \sim \text{Normal}(a_0, A_0)$. The specific choice of this prior is known to be rather influential on the posterior estimates of the model indicators δ (O'Hagan (1995), George and McCulloch (1997)). We want to avoid this

and choose a fractional prior approach here. Fractional priors were introduced by O'Hagan (1995). They were applied to Bayesian estimation of variance-covariance matrices by Smith and Kohn (2002). We take their ideas up and take a fraction $b = 1/\sum_{i=1}^{N} T_i$ from the likelihood of model (6) to derive a fractional prior. By normalization we obtain the fractional prior $p(\theta^\delta|R, y^{uxb})$:

$$\begin{aligned} p(\theta^\delta|R, y^{uxb}) &= p(y^u|R, \theta^\delta)^b / \int p(y^u|R, \theta^\delta)^b d\theta^\delta \\ &\sim \text{Normal}\left(a_N, A_N \frac{1}{b}\right), \end{aligned} \tag{7}$$

where

$$A_N^{-1} = (W^\delta)' \Sigma^{-1} W^\delta, \tag{8}$$

$$a_N = A_N (W^\delta)' \Sigma^{-1} (y^u - m). \tag{9}$$

Posterior of θ^δ To obtain the posterior we combine the fractional prior (7) with the remaining part of the likelihood $p(y^u|\theta^\delta, R)^{(1-b)}$. This yields the following normal posterior for θ^δ:

$$p(\theta^\delta|R, y^u) \sim \text{Normal}\left(a_N, A_N\right), \tag{10}$$

with the posterior moments given in (8) and (9).

2.3.3 Sampling the Indicators δ

Prior for the Indicators δ A straightforward choice of a prior for the indicators would be to consider the δ_k as independent a priori and define

$P(\delta_k = 1|\tau) = \tau$ for some fixed $\tau \in (0,1)$. We wish to reduce the prior influence of τ and choose the following Beta function as prior for the indicators δ (Smith and Kohn (2002)):

$$p(\delta) = \text{Beta}(p_\delta, d - p_\delta + 1), \tag{11}$$

where p_δ is the number of non-zero parameters in θ. The reasoning behind this choice is that we introduce τ as a hyperparameter and integrate the density $p(\delta|\tau)p(\tau)$ with respect to τ to obtain (11).

The Marginal Likelihood To sample the indicators we need the marginal likelihood with respect to θ^δ. Therefore we combine (7) with the remaining part of the likelihood $p(y^u|\theta^\delta, R)^{(1-b)}$ and integrate this with respect to θ^δ. The marginal likelihood equals

$$p(y^u|\delta, R) = b^{p_\delta/2} \left(\frac{1}{2\pi}\right)^{NT(1-b)/2} \exp\left(-\frac{(1-b)}{2} S\right), \tag{12}$$

where $p_\delta = \dim(\theta^\delta)$ and

$$S = (y^u - W^\delta a_N - m)' \Sigma^{-1} (y^u - W^\delta a_N - m). \tag{13}$$

Fast Algorithm for Sampling δ To sample the indicators δ_k, we use the fast algorithm of Smith and Kohn (2002).

Let δ_k^{old} denote the current value and δ_k^{new} denote the new value of δ_k. We generate u from a uniform distribution on $[0,1]$. Then,

(i-1) if $\delta_k^{old} = 1$ and $u > p(\delta_k = 0)$, set $\delta_k^{new} = 1$.

(i-2) if $\delta_k^{old} = 0$ and $u > p(\delta_k = 1)$, set $\delta_k^{new} = 0$.

(i-3) if $\delta_k^{old} = 1$ and $u \leq p(\delta_k = 0)$, generate $u^* \sim U[0,1]$ and set $\delta_k^{new} = 0$, if $u^* \leq l(\delta_k = 0)/(l(\delta_k = 0) + l(\delta_k = 1))$.

(i-4) if $\delta_k^{old} = 0$ and $u \leq p(\delta_k = 1)$, generate $u^* \sim U[0,1]$ and set $\delta_k^{new} = 1$, if $u^* \leq l(\delta_k = 1)/(l(\delta_k = 0) + l(\delta_k = 1))$.

Here $p(\delta_k = i) = p(\delta_k = i|\delta_{\setminus k}), i = 0, 1$ is the conditional prior of δ_k (see below). $l(\delta_k = i)$ equals the marginal likelihood $p(y^u|\delta, R)$ defined in (12) where δ_k either takes the value $i = 0$ or $i = 1$.

The Conditional Prior of the Indicators To generate from $\delta_k|\delta_{\setminus k}, R, y^u$ we need the conditional prior of δ_k given the remaining elements $\delta_{\setminus k}$. Let p_δ be the number of elements of θ, which are non-zero (before sampling δ_k^{new}).

If $\delta_k^{old} = 1$, then

$$p(\delta_k = 0) = (d - p_\delta + 1)/(d + 1), \quad p(\delta_k = 1) = p_\delta/(d + 1).$$

If $\delta_k^{old} = 0$, then

$$p(\delta_k = 0) = (d - p_\delta)/(d + 1), \quad p(\delta_k = 1) = (p_\delta + 1)/(d + 1).$$

2.3.4 Sampling the Latent Indicators R

Sampling of R amounts to sampling the component indicators of the finite normal mixture with fixed parameters m_j, s_j^2 and η_j from Table 1. Conditional on the utilities y_{it}^u and the exponential of the linear predictor $\lambda_{it} = \exp(w_{it}^\delta \theta^\delta)$ we sample the component indicators r_{it} from the discrete density for j=1,...,10 :

$$\log P(r_{it} = j|y_{it}^u, \lambda_{it}) \propto -\log s_j - \frac{1}{2}\left(\frac{y_{it}^u - \log \lambda_{it} - m_j}{s_j}\right)^2 + \log \eta_j. \quad (14)$$

2.3.5 Sampling the Latent Utilities y_{it}^u

Conditional on $\lambda_{it} = \exp(w_{it}^{\delta}\theta^{\delta})$ the latent utilities y_{it}^u are derived from exponential distributions (details are given in the Appendix):

$$\begin{aligned} \exp(-y_{it}^u) &\sim \text{Exponential}\,(\lambda_{it}+1) && \text{if} \quad y_{it}=1, \\ \exp(-y_{it}^u) &\sim \text{Exponential}\,(\lambda_{it}+1) + \text{Exponential}\,(\lambda_{it}) && \text{if} \quad y_{it}=0. \end{aligned} \tag{15}$$

Therefore we sample the utilities from

$$y_{it}^u = -\log\left(-\frac{\log(U_{it})}{1+\lambda_{it}} - \frac{\log(U_{it}^*)}{\lambda_{it}} I_{\{y_{it}=0\}}\right), \tag{16}$$

where U_{it} and U_{it}^* are uniform random variables and $I_{\{\cdot\}}$ denotes the indicator function.

3 BAYESIAN VARIABLE AND COVARIANCE SELECTION FOR BINARY LOGIT MIXED EFFECTS MODELS

3.1 The Model

We now extend the binary logistic regression model and include additional random and fixed effects in model (2). Let $x_{it} = [x_{it}^f\, x_{it}^r]$ denote the design vector of these additional effects. x_{it}^f has dimension d_f and contains those elements which correspond to the fixed effects α, whereas x_{it}^r has dimension d_r and corresponds to the random effects vector β_i. The binary logit mixed effects model is given by

$$P(y_{it}=1|\theta^{\delta},\alpha,\beta_i) = \frac{\exp(w_{it}^{\delta}\theta^{\delta} + x_{it}^f\alpha + x_{it}^r\beta_i)}{1+\exp(w_{it}^{\delta}\theta^{\delta} + x_{it}^f\alpha + x_{it}^r\beta_i)}. \tag{17}$$

We assume normally distributed random effects with
$\beta_i \sim \text{Normal}_{d_r}(\beta^G, Q)$. The observations are assumed to be independent conditional on knowing θ^δ, α and $\beta^N = (\beta_1, \ldots, \beta_N)$.

We include two variable selection steps when estimating model (17). Firstly, following the ideas of Section 2, we select the non-zero elements θ^δ. This can be done by slight modifications of the steps from the previous section. Secondly, we want to select the non-zero elements of the variance-covariance matrix Q. Estimation of the structure of Q includes model selection with respect to fixed and random effects. If for example all elements of column s in Q (and therefore also of row s) are zero we would obtain the s-th effect of β_i as fixed effect.

In the case of normally distributed data linear mixed effects models are defined for two different parameterizations (e.g. Meng and van Dyk (1998)). The parameterization where the mean and the variance-covariance matrix Q appear in the latent equation, is called centered, whereas these parameters appear in the observation equation, if the non-centered parameterization is used.

We adopt these terms and call the representation (17) of the logit mixed effects model *centered*. This model is equivalent to the model in the so-called *non-centered* parameterization:

$$P(y_{it} = 1|\theta^\delta, \alpha, \beta^G, C, \tilde{z}_i) = \frac{\exp(w_{it}^\delta \theta^\delta + x_{it}^f \alpha + x_{it}^r \beta^G + x_{it}^r C\tilde{z}_i)}{1 + \exp(w_{it}^\delta \theta^\delta + x_{it}^f \alpha + x_{it}^r \beta^G + x_{it}^r C\tilde{z}_i)}, \quad (18)$$

where the Cholesky decomposition with Cholesky factors C was applied to the variance-covariance matrix Q: $Q = CC'$ (C lower triangular). Therefore the individual effects $\tilde{z}_i$ are standard normally distributed: $\tilde{z}_i \sim \text{Normal}_{d_r}(0, I)$

and $\beta_i = \beta^G + C \cdot \tilde{z}_i$. We denote $\tilde{z}^N = (\tilde{z}_1, \ldots, \tilde{z}_N)$.

In what follows we describe the selection of the elements of the variance-covariance matrix.

3.2 Covariance Selection

To include covariance selection in the modeling procedure we first observe, that conditional on $\tilde{z}^N$ the Cholesky factors C of the variance-covariance matrix appear as regression coefficients in model equation (18). Therefore the problem of selecting the form of the variance-covariance matrix may be treated as a variable selection problem of elements of C. Conditional on knowing the random effects $\tilde{z}^N$ we simply have to adapt those selection steps, which we already know for the selection of elements of θ from Section 2.

Technically this may be realized in the following way. We rewrite the $x^r_{it} C \tilde{z}_i$ part of (18) to obtain a design matrix for the regression coefficients C. C is a lower triangular matrix of dimension d_r. Let $\tilde{z}_i = (\tilde{z}_{i1}, \ldots, \tilde{z}_{id_r})$ denote the individual effects for subject i. Conditional on $\tilde{z}_i$ the design vector for the first column of C is constructed from all d_r elements of x^r_{it} and the first element of $\tilde{z}_i$ and equals $x^r_{it(1:d_r)} \cdot \tilde{z}_{i1}$. To construct the design vector for the lower triangular part of the second column of C we have to combine only the last $d_r - 1$ elements of x^r_{it} with the second element of $\tilde{z}_i$: $x^r_{it(2:d_r)} \cdot \tilde{z}_{i2}$. We proceed in the that way till the design vectors for all lower triangular columns of C are constructed. Finally we stack all these vectors and obtain the new design vector

$$v_{it} = [x^r_{it(1:d_r)} \tilde{z}_{i1} \; x^r_{it(2:d_r)} \tilde{z}_{i2} \; \ldots \; x^r_{it(d_r)} \tilde{z}_{id_r}].$$

The vector of the regression coefficients which belongs to v_{it} has dimension $d_r(d_r+1)/2$ and consists of the lower triangular elements of C stacked columnwise. We define an indicator vector $\gamma = (\gamma_1, \ldots, \gamma_{d_r(d_r+1)/2})$ for this parameter vector. γ_m takes the value 1, if the corresponding element of C is unequal to zero, and the value 0 otherwise. Let C^γ denote the column vector of non-zero regression coefficients, and v_{it}^γ consists only of those elements of v_{it}, which have indicators γ equal to 1.

Therefore the binary logit mixed effects model with variable and covariance selection reads:

$$P(y_{it}=1|\theta^\delta, \alpha, \beta^G, C^\gamma, \tilde{z}_i) = \frac{\exp(w_{it}^\delta\theta^\delta + x_{it}^f\alpha + x_{it}^r\beta^G + v_{it}^\gamma C^\gamma)}{1+\exp(w_{it}^\delta\theta^\delta + x_{it}^f\alpha + x_{it}^r\beta^G + v_{it}^\gamma C^\gamma)}. \quad (19)$$

3.3 Data Augmentation and Sampling

3.3.1 Data Augmentation

We apply again the same two data augmentation steps as in Section 2.2 and add the vector of utilities y^u and the indicators R to obtain a normal linear model:

$$y_{it}^u = w_{it}^\delta\theta^\delta + x_{it}^f\alpha + x_{it}^r\beta^G + v_{it}^\gamma C^\gamma + m_{r_{it}} + \varepsilon_{it}, \quad \varepsilon_{it} \sim \text{Normal}\left(0, s_{r_{it}}^2\right). \quad (20)$$

We include all observations in the design matrices W^δ, X^f, X^r and V^γ, respectively, and write the augmented model in matrix notation:

$$y^u = W^\delta\theta^\delta + X^f\alpha + X^r\beta^G + V^\gamma C^\gamma + m + \varepsilon, \quad \varepsilon \sim \text{Normal}\left(0, \Sigma\right). \quad (21)$$

3.3.2 MCMC Scheme

To build the MCMC scheme for variable and covariance selection in binary logistic mixed effects models we modify the scheme of Section 2.3.1.

(i-a) Sample each element δ_k of the indicator vector δ separately conditional on $\delta_{\setminus k}$ (all other elements of δ), γ, α, β^G, $\tilde{z}^N$, R and y^u.

(i-b) Sample each element γ_m of the indicator vector γ separately conditional on $\gamma_{\setminus m}$ (all other elements of γ), δ, α, β^G, $\tilde{z}^N$, R and y^u.

(ii-a) Sample the non-zero elements θ^δ and the non-zero elements of the Cholesky factor C^γ together in one block conditional on γ, δ, α, β^G, $\tilde{z}^N$, R and y^u.

(ii-b) Sample the fixed effects α and the mean parameter β^G together in one block conditional on C^γ, θ^δ, R and y^u and marginally with respect to $\tilde{z}^N$.

(ii-c) Sample the individual effects $\tilde{z}^N$ conditional on α, β^G, C^γ, θ^δ, R and y^u.

(iii) Sample R conditional on α, β^G, C^γ, θ^δ, $\tilde{z}^N$ and y^u.

(iv) Sample y^u conditional on α, β^G, C^γ, θ^δ, $\tilde{z}^N$ and the binary data y.

In what follows we give details of these steps.

3.3.3 Sampling the Non-zero Elements of θ^δ and C^γ

We sample the non-zero elements θ^δ and C^γ together in one block and derive their joint fractional prior with fraction $b = 1/\sum_{i=1}^{N} T_i$:

$$p(\theta^\delta, C^\gamma|\alpha, \beta^G, \tilde{z}^N, R, y^{uxb}) =$$

$$p(y^u|\alpha, \beta^G, \tilde{z}^N, R, \theta^\delta, C^\gamma)^b / \int p(y^u|\alpha, \beta^G, \tilde{z}^N, R, \theta^\delta, C^\gamma)^b d(\theta^\delta, C^\gamma) \sim$$

$$\text{Normal}\left(a_N, A_N \frac{1}{b}\right),$$

where

$$A_N^{-1} = [W^\delta \; V^\gamma]' \Sigma^{-1} [W^\delta \; V^\gamma],$$

$$a_N = A_N \left([W^\delta \; V^\gamma]' \Sigma^{-1} (y^u - X^f \alpha - X^r \beta^G - m) \right).$$

By combining the fractional prior with the remaining $(1 - b)$ proportion of the likelihood we obtain the normally distributed joint posterior for θ^δ, C^γ:

$$p(\theta^\delta, C^\gamma|\alpha, \beta^G, \tilde{z}^N, R, y^u) \sim \text{Normal}\left(a_N, A_N\right).$$

3.3.4 Sampling the Indicators δ and γ

By simple extension of the findings from Section 2.3 we obtain the prior for the indicators: $p(\delta) = \text{Beta}(p_\delta, d - p_\delta + 1)$ and
$p(\gamma) = \text{Beta}(p_\gamma, d_r(d_r + 1)/2 - p_\gamma + 1)$, where d and $d_r(d_r + 1)/2$ are the number of total free elements, whereas p_δ and p_γ denote the number of non-zero elements in θ and C, respectively.

We apply the fast sampling scheme of Section 2.3.3 to sample the indi-

cators. The marginal likelihood (12) is therefore modified to:

$$p(y^u|\delta,\gamma,\alpha,\beta^G,\tilde{z}^N,R) = b^{(p_\delta+p_\gamma)/2}\left(\frac{1}{2\pi}\right)^{NT(1-b)/2}\exp\left(-\frac{(1-b)}{2}\,S\right),$$

where

$$\begin{aligned} S = (y^u - [W^\delta\ V^\gamma]a_N - X^f\alpha - X^r\beta^G - m)'\cdot\Sigma^{-1} \\ \cdot(y^u - [W^\delta\ V^\gamma]a_N - X^f\alpha - X^r\beta^G - m). \end{aligned}$$

3.3.5 Sampling α, β^G and $\tilde{z}^N$

Steps (ii-b) and (ii-c) are simply equal to sampling regression parameters and random effects in a normal mixed effects model, see for example Chib and Carlin (1999). We sample α and β^G together in one block marginally with respect to the individual effects $\tilde{z}^N$ from model

$$y^u \sim \text{Normal}\left(X^f\alpha + X^r\beta^G + W^\delta\theta^\delta + m, D\right),$$

where D is block-diagonal with N blocks $D_i = X_i^r Q (X_i^r)' + \Sigma_i$, with $X_i^r = \{x_{it}^r\}_{t=1,\ldots,T_i}$ and $\Sigma_i = \text{diag}\{s_{it}^2\}_{t=1,\ldots,T_i}$ for $i = 1,\ldots,N$. We assume a joint conditionally conjugate normal prior for α and β^G: Normal (b_0, B_0). Posterior sampling amounts to sampling from the multivariate normal distribution

$$p(\alpha,\beta^G|C^\gamma,\theta^\delta,R,y^u) \sim \text{Normal}\,(b_N, B_N)\,,$$

$$B_N^{-1} = [X^f\ X^r]' D^{-1}[X^f\ X^r] + B_0^{-1},$$

$$b_N = B_N\left([X^f\ X^r]' D^{-1}(y^u - W^\delta\theta^\delta - m) + B_0^{-1}b_0\right).$$

The individual effects $\tilde{z}_i$ are conditionally independent for subjects $i = 1, \ldots, N$ and according to our model assumptions they follow a standard normal distribution a priori. Therefore the posterior is multivariate normal with

$$p(\tilde{z}_i|\alpha, \beta, C^\gamma, \Theta^\delta, R, y^u) \sim \text{Normal}\,(p_i, P_i)\,,$$

$$P_i^{-1} = (X_i^r C)'\Sigma^{-1}(X_i^r C) + I$$

$$p_i = P_i\left((X_i^r C)'\Sigma^{-1}(y_i^u - W^\delta\theta^\delta - X_i^f\alpha - X_i^r\beta^G - m_i)\right),$$

where y_i^u, X_i^f, X_i^r and m_i are those parts of y^u, X^f, X^r and m, which correspond to subject i.

3.3.6 Sampling the Latent Indicators R and the Latent Utilities y^u

Steps (iii) and (iv) are the same as in Section 2.3, only the exponential of the linear predictor λ_{it} has to be adapted to the new model: $\lambda_{it} = \exp(w_{it}^\delta\theta^\delta + x_{it}^f\alpha + x_{it}^r\beta^G + v_{it}^\gamma C^\gamma)$. By including these new λ_{it} in equations (14) and (16) we sample R and y^u, respectively.

4 THE MULTINOMIAL LOGISTIC MIXED EFFECTS MODEL

4.1 The Model

Let us now extend the binomial logit mixed effects model with variable selection to the case of more than two categories. The multi-categorical

data y_{it} take values in one of $L+1$ categories, labelled $\{0,\ldots,L\}$. The probability for y_{it} to take the category l depends on the model covariates and the parameters in the following way:

$$P(y_{it}=l|\theta_l^{\delta_l},\alpha_l,\beta_l^G,C_l^{\gamma_l},\tilde{z}_{li})=\frac{\exp(w_{it}^{\delta_l}\theta_l^{\delta_l}+x_{it}^f\alpha_l+x_{it}^r\beta_l^G+v_{lit}^{\gamma_l}C_l^{\gamma_l})}{1+\sum_{l'=1}^{L}\exp(w_{it}^{\delta_{l'}}\theta_{l'}^{\delta_{l'}}+x_{it}^f\alpha_{l'}+x_{it}^r\beta_{l'}^G+v_{l'it}^{\gamma_{l'}}C_{l'}^{\gamma_{l'}})}. \quad (22)$$

Note that the regression parameters as well as the indicators for the non-zero elements are category specific. Since the random effects $\tilde{z}_{li}$ are also category specific, the design matrix for the Cholesky factors C_l changes for each category: v_{lit}.

To make the model identifiable we assume that $l=0$ is the baseline category with $\theta_0^{\delta_0}=0$, $\alpha_0=0$, $\beta_0^G=0$ and $C_0^{\gamma_0}=0$. The observations are independent conditional on knowing $\theta_l^{\delta_l}$, α_l, β_l^G, C_l and $\tilde{z}_{li}^N$.

4.2 Data Augentation and Gibbs Sampling

4.2.1 Data Augmentation

The first data augmentation step introduces for each subject i latent utilities y_{lit}^u for choosing category $l=1,\ldots,L$ at observation t. We denote the vector of all these utilities y^u. After introducing the latent utilities this yields the following augmented model:

$$\begin{aligned} y_{1it}^u &= w_{it}^{\delta_1}\theta_1^{\delta_1}+x_{it}^f\alpha_1+x_{it}^r\beta_1^G+v_{1it}^{\gamma_1}C_1^{\gamma_1}+\varepsilon_{1it},\\ &\ldots \\ y_{Lit}^u &= w_{it}^{\delta_L}\theta_L^{\delta_L}+x_{it}^f\alpha_L+x_{it}^r\beta_L^G+v_{Lit}^{\gamma_L}C_L^{\gamma_L}+\varepsilon_{Lit}. \end{aligned} \quad (23)$$

The errors ε_{lit} follow a standard type I extreme value distribution. They

are again approximated by the mixture of ten normal distributions with parameters given in Table 1. In the second data augmentation step we introduce the group indicators r_{lit} for each subject i at each observation t and for each category $l = 1, \ldots, L$. We subsume all indicators for category l under R_l.

The augmented model takes the form

$$y_{lit}^u = w_{it}^{\delta_l}\theta_l^{\delta_l} + x_{it}^f\alpha_l + x_{it}^r\beta_l^G + v_{lit}^{\gamma_l}C_l^{\gamma_l} + m_{r_{lit}} + \varepsilon_{lit}, \tag{24}$$

$$\varepsilon_{lit} \sim \text{Normal}\left(0, s_{r_{lit}}^2\right), \quad l = 1, \ldots, L.$$

4.2.2 MCMC Scheme

Carry out steps (i-a) - (iii) for each category $l = 1, \ldots, L$ separately:

(i-a) Sample each element δ_{lk} of the indicator vector δ_l separately conditional on $\delta_{l\,\backslash k}$ (all other elements of δ_l), γ_l, α_l, β_l^G, $\tilde{z}_l^N$, R_l and y_l^u.

(i-b) Sample each element γ_{lm} of the indicator vector γ_l separately conditional on $\gamma_{l\,\backslash m}$ (all other elements of γ_l), δ_l, α_l, β_l^G, $\tilde{z}_l^N$, R_l and y_l^u.

(ii-a) Sample the non-zero elements $\theta_l^{\delta_l}$ and the non-zero elements of the Cholesky factor $C_l^{\gamma_l}$ together in one block conditional on γ_l, δ_l, α_l, β_l^G, $\tilde{z}_l^N$, R_l and y_l^u.

(ii-b) Sample the fixed effects α_l and the mean parameter β_l^G together in one block conditional on $C_l^{\gamma_l}$, $\theta_l^{\delta_l}$, R_l and y_l^u and marginally with respect to $\tilde{z}_l^N$.

(ii-c) Sample the individual effects $\tilde{z}_l^N$ conditional on α_l, β_l^G, $C_l^{\gamma_l}$, $\theta_l^{\delta_l}$, R_l and y_l^u.

(iii) Sample R_l conditional on α_l, β_l^G, $C_l^{\gamma_l}$, $\theta_l^{\delta_l}$, $\tilde{z}_l^N$ and y_l^u.

After having sampled the model parameters for all L categories:

(iv) Sample y^u conditional on the parameters α_l, β_l^G, $C_l^{\gamma_l}$, $\theta_l^{\delta_l}$, $\tilde{z}_l^N$ of all L categories and the binary data y.

Steps (i-a) - (iii) are based on the L equations (24), which resemble equation (20) for the binary data case. Therefore the corresponding steps of Section 3.3.2 may easily be transferred and steps (i-a) - (iii) are simply carried out L times for each category $l = 1, \ldots, L$ separately.

Step (iv) has to be modified to sample the utilities for multi-categorical data. We define $\lambda_{lit} = \exp(w_{it}^{\delta_l}\theta_l^{\delta_l} + x_{it}^1\alpha_l + x_{it}^2\beta_l^G + v_{lit}^{\gamma_l}C_l^{\gamma_l})$ and sample the utilities from

$$y_{kit}^u = -\log\left(-\frac{\log(U_{kit})}{1+\sum_{l=1}^L \lambda_{lit}} - \frac{\log(U_{kit}^*)}{\lambda_{kit}} I_{\{y_{it}\neq k\}}\right), \tag{25}$$

where U_{kit} and U_{kit}^* are uniform random variables. Details are given in the Appendix.

5 EXAMPLES

5.1 Selecting Risk-Factors in Credit-Scoring Data

This data set consists of 1000 binary observations of credit clients of a south German bank. We observe 700 "credit worthy" clients with $y_i = 1$, whereas $y_i = 0$ for 300 "not credit worthy" clients. Credit worthy clients did pay

back the credit on time. The bank is interested in estimating the risk that a client will not pay back the credit as agreed in the contract with the help of additionally available risk-factors. Such risk-factors are variables representing the economic and social situation of the clients, as for example the amount of the credit, whether the client has currently an account, his/her age, marital status, etc. These covariates are partly metrical and partly categorical. After introducing dummy design variables for the categorical variables we obtain a design matrix with 37 variables. The data are also analysed in Fahrmeir and Tutz (1994) and are available in electronic form from *http://www.stat.uni-muenchen.de/*.

The fraction for the fractional prior for the regression parameters θ^δ equals $b = 1/N = 1/1000$. We ran $S = 40\,000$ iterations (after $10\,000$ iterations burn-in) and give estimates of the posterior probabilities $\hat{P}(\delta = 1|y) = \frac{1}{S}\sum_{s=1}^{S}\delta^{(s)}$ together with the estimates of the significant parameters $\hat{\theta}$ in Table 2. 16 effects have posterior probability greater than 0.5 to be included in the model and are identified as being important for explaining whether a client is credit-worthy.

5.2 Selecting the Covariance Structure of Individual Effects in Travel Data

These data come from a conjoint study about packaged city trips and are described in Hatzinger and Mazanec (2005). The study was carried out in Vienna and 499 consumers were asked to rate 9 different city trip packages on a 10 point rating scale, which indicated the likelihood of booking such a package. Given the questionable metric properties of the rating data

Table 2: Selection and estimation of risk-factors for credit-scoring.

risk-factor	$\hat{P}(\delta = 1\|y) = 1$	$\hat{\theta}$
no running account	0.97	-0.6
good running account	1	1.27
duration of credit	1	-0.03
credit-worthy in the past	0.98	0.87
private purpose	0.97	0.53
amount of credit	0.31	
amount of credit, quadratic	0.88	-0.1
some savings	0.67	0.3
higher savings	1	0.85
same employer		
≤ 1 year	0.55	-0.27
$1 < .. \leq 4$ years	0.35	
> 4 years	0.74	0.36
rate (% of income)		
$20 \leq .. < 35$	0.33	
< 20	0.91	-0.51
male, not single	0.63	0.24
female, single	0.22	
other debtors	0.3	
surety	0.7	0.63
same home		
$1 < .. \leq 7$ years	0.63	-0.27
$.. > 7$ years	0.29	
car owner	0.27	
life insurance	0.25	
real estate	0.47	
age	0.42	
other credits		
bank	0.27	
others	0.41	
rented flat	0.82	0.42
freehold flat	0.34	
no. of credits in the past		
$1, 3$	0.21	
≥ 4	0.22	
unskilled, resident	0.21	
skilled	0.25	
manager/self-employed	0.21	
> 3 persons entitled to maintenance	0.25	
telephone	0.49	
no foreign worker	0.98	1.71

(Hatzinger and Mazanec (2005)) these data were transformed into binary categorical data. Altogether 1856 packages were approved by the consumers ($y_{it} = 1$), whereas 2635 were not approved ($y_{it} = 0$). Combinations of four package attributes were chosen to set up an orthogonal design with the 9 artificial trip packages. These attributes are:

destination:	close distance to Vienna (reference category); medium distance to Vienna
transport mode:	bus (reference category); train; plane
duration of stay:	2 days (reference category); 3 days; 4 days
price in Euro:	258 (reference category); 372 ; 482

We used dummy coding for the design variables in the same order as above. To account for individual differences we included all eight effects as random effects and carried out covariance selection in the logistic mixed effects model. The fraction for the fractional prior for C^γ equals $b = 1/(9 \cdot 499)$ and we choose the multivariate standard normal distribution as prior for the mean parameter vector β^G. We base our estimates again on $S = 40\,000$ iterations (after 10 000 iterations burn-in). The posterior probability for the elements of the Cholesky factors C to be different from zero $\hat{P}(\gamma = 1|y)$ are given in Table 3. Indicators for Q may easily be obtained at each MCMC iteration from the indicators γ of the Cholesky factors C and the identity $Q = CC'$. Posterior estimates for the probability of the elements of the variance-covariance matrix Q to be different from zero are given in Table 4.

We observe that the posterior estimates for the indicators of C are smaller than the estimates of indicators for Q. There are 11 elements of C with

Table 3: Travel data: estimates $\hat{P}(\gamma = 1|y)$ of indicators for C.

1	0	0	0	0	0	0	0
0.95	1	0	0	0	0	0	0
1	0.98	1	0	0	0	0	0
1	0.84	0.62	1	0	0	0	0
0.83	0.36	0.33	0.3	1	0	0	0
0.56	0.8	0.44	0.41	1	0.64	0	0
0.43	0.35	0.4	0.42	1	0.64	0.63	0
0.39	0.45	0.77	0.57	1	0.53	0.52	0.65

Table 4: Travel data: estimates of indicators for Q.

1	0.95	1	1	0.83	0.56	0.43	0.39
0.95	1	1	0.99	0.86	0.89	0.61	0.63
1	1	1	1	0.9	0.92	0.74	0.89
1	0.99	1	1	0.91	0.9	0.77	0.88
0.83	0.86	0.9	0.91	1	1	1	1
0.56	0.89	0.92	0.9	1	1	1	1
0.43	0.61	0.74	0.77	1	1	1	1
0.39	0.63	0.89	0.88	1	1	1	1

$\hat{P}(\gamma = 1|y)$ smaller than 0.5, whereas this is the case only for 2 elements of the variance-covariance matrix Q. Specification of the model in the non-centered parameterization in combination with covariance selection therefore yields a reduction of parameters and a more parsimonious representation of the model.

6 CONCLUSIONS

In real applications researchers would rather start off with a general model specification, with no prior information about the definite form of the model having to be at hand. In the context of logistic mixed effects models the presented algorithm allows to include many explanatory variables and many random effects at the beginning. The appropriate structure is then determined in a data driven manner.

The challenging problem of covariance selection is solved by applying a Cholesky decomposition to the random effects covariance matrix and choosing among all free elements of the matrix. Thereby we also determine random versus fixed effects in our model. For both the variable and the covariance selection we applied a stochastic search variable approach. Our method is especially convenient since no additional tuning is necessary for both the variable selection part and the logistic part of the model.

7 ACKNOWLEDGEMENT

I thank Sylvia Frühwirth-Schnatter for many helpful discussions.

APPENDIX: STEP (iv) - SAMPLING THE UTILITIES

Here we derive the basic properties of the distribution of the utilities y^u_{lit}. We assume categorical data for $L+1$ categories. The simplification for binary data ($L = 1$) is straightforward.

In what follows we denote $\lambda_{lit} = \exp(w_{it}^{\delta_l}\theta_l^{\delta_l} + x_{it}^f\alpha_l + x_{it}^r\beta_l^G + v_{lit}^{\gamma_l}C_l^{\gamma_l})$, (with obvious simplifications, if no random or fixed effects are present). As the errors in (23) follow a type I extreme value distribution we obtain

$$\exp(-y_{lit}^u) \sim \text{Exponential}\left(\lambda_{lit}\right), \quad l = 1, \ldots, L,$$

$$\exp(-y_{0it}^u) \sim \text{Exponential}\left(1\right).$$

Given, that the categorical observation belongs to category k, i.e. $y_{it} = k$, the utility y_{kit}^u is the maximum of all utilities:

$$y_{kit}^u = \max_{l=0,\ldots,L} y_{lit}^u \quad \Leftrightarrow \quad \exp(-y_{kit}^u) = \min_{l=0,\ldots,L} \exp(-y_{lit}^u).$$

Since $\exp(-y_{kit}^u)$ is the *minimum* of the exponentially distributed random variables, its parameter has to be equal to $1 + \sum_{l=1}^{L} \lambda_{lit}$.

For all other utilities $y_{\bar{k}it}^u$, $\bar{k} = 1, \ldots, L, \bar{k} \neq k$ the following relationship holds:

$$\exp(-y_{\bar{k}it}^u) \sim \text{Exponential}\left(1 + \sum_{l=1}^{L} \lambda_{lit}\right) + \text{Exponential}\left(\lambda_{\bar{k}it}\right).$$

Sampling the utilities from (16) and (25) follows immediately.

References

Albert, J. H. and S. Chib (1997). Bayesian tests and model diagnostics in conditionally independent hierarchical models. *Journal of the American Statistical Association 92*, 916–925.

Chen, Z. and D. B. Dunson (2003). Random effects selection in linear mixed models. *Biometrics 59*, 762–769.

Chib, A. and B. P. Carlin (1999). On mcmc sampling in hierarchical longitudinal models. *Statistics and Computing 9*, 17–26.

Dempster, A. P. (1972). Covariance selection. *Biometrics 28*, 157–175.

Fahrmeir, L. and G. Tutz (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer Series in Statistics. New York: Springer-Verlag Inc.

Frühwirth-Schnatter, S. and R. Frühwirth (2005). Auxiliary mixture sampling with applications to logistic models. *IFAS Research Paper Series, Department of Applied Statistics, Johannes Kepler University Linz*.

Frühwirth-Schnatter, S. and H. Wagner (2005). Auxiliary mixture sampling for parameter-driven models of time series of counts with applications to state space modelling. *IFAS Research Paper Series, Department of Applied Statistics, Johannes Kepler University Linz 11*.

George, E. I. and R. E. McCulloch (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association 88*(423), 881–889.

George, E. I. and R. E. McCulloch (1997). Approaches for Bayesian variable selection. *Statistica Sinica 7*, 339–373.

Hatzinger, R. and J. Mazanec (2005). Measuring the part worth of the mode of transport in a trip package: An extended Bradley-Terry model for paired-comparison conjoint data. *Working paper Vienna University of Economics and Business Administration*.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky (1999). Bayesian model averaging: A tutorial. *Statistical Science 14* (4), 382–417.

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

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

Lindstrom, M. and D. M. Bates (1988). Newton-Raphson and the EM-algorithm for linear mixed-effects models for repeated measures data. *Journal of the American Statistical Association 83*, 1014–1022.

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

Meng, X.-L. and D. A. van Dyk (1998). Fast EM-type implementations for mixed effects models. *Journal of the Royal Statistical Society, Series B 60*, 559–578.

Nott, D. J. and P. J. Green (2004). Bayesian variable selection and the Swendsen-Wang algorithm. *Journal of Computational and Graphical Statistics 13* (1), 141–157.

Nott, D. J. and D. Leonte (2004). Sampling schemes for Bayesian variable selection in generalized linear models. *Journal of Computational and Graphical Statistics 13* (2), 362–382.

O'Hagan, A. (1995). Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society, Series B 57*, 99–118.

Pinheiro, J. C. and D. M. Bates (1996). Unconstrained parametrizations for variance-covariance matrices. *Statistics and Computing 6*, 289–296.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterizaion. *Biometrika 86*(3), 677–690.

Pourahmadi, M. (2000). Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. *Biometrika 87*(2), 425–435.

Raftery, A. E., D. Madigan, and J. A. Hoeting (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association 92*(437), 179–191.

Scott, S. L. (2005). Data augmentation, frequentist estimation, and the Bayesian analysis of multinomial data. *Bayesian Analysis submitted*, http://www–rcf.usc.edu/∼sls/.

Smith, M. and R. Kohn (2002). Parsimonious covariance matrix estimation for longitudinal data. *Journal of the American Statistical Association 97*, 1141–1153.

Zeger, S. L. and M. R. Karim (1991). Generalized linear models with random effects; a Gibbs sampling approach. *Journal of the American Statistical Association 86*, 79–85.