Bayesian parsimonious covariance estimation for hierarchical linear mixed models

Sylvia Frühwirth-Schnatter · Regina Tüchler

Abstract We consider a non-centered parameterization of the standard random-effects model, which is based on the Cholesky decomposition of the variance-covariance matrix. The regression type structure of the non-centered parameterization allows us to use Bayesian variable selection methods for covariance selection. We search for a parsimonious variance-covariance matrix by identifying the non-zero elements of the Cholesky factors. With this method we are able to learn from the data for each effect whether it is random or not, and whether covariances among random effects are zero. An application in marketing shows a substantial reduction of the number of free elements in the variance-covariance matrix.

Keywords Covariance selection · Random-effects models · Markov chain Monte Carlo · Fractional prior · Variable selection

1 Introduction

This article addresses various problems associated with estimating the variance-covariance matrix **Q** of the random effects within the framework of hierarchical linear models.

A computational challenge in estimating hierarchical linear models is to select a suitable parameterization of the variance-covariance matrix, which typically has a large number of parameters related by the very complex constraint that the resulting matrix needs to be positive definite.

A particularly useful parameterization of variance-covariance matrices is based on the Cholesky decomposition of either **Q** or $\mathbf{Q}^{-1}$. As pointed out by Pinheiro and Bates (1996), this parameterization is of considerable numerical convenience as it involves unconstrained parameters only. Within the framework of hierarchical models, it is usual to work with the Cholesky decomposition of **Q**, as illustrated by Lindstrom and Bates (1988), Meng and van Dyk (1998), and Chen and Dunson (2003), among others. Following this tradition, we will consider a parameterization of the variance-covariance matrices based on the Cholesky decomposition $\mathbf{Q} = \mathbf{CC}'$ with a lower triangular matrix **C**.

We aim at a parsimonious representation of the random-effects covariance matrix, rather than estimating a fully unrestricted variance-covariance matrix, as it is usually done.

To some extent, we follow the seminal work of Smith and Kohn (2002), who model data arising from a multivariate normal distribution with unknown variance-covariance matrix **Q**. They identify zero off-diagonal elements in the inverse $\mathbf{Q}^{-1}$ of the variance-covariance matrix, whereas the diagonal is not the object of model selection. As opposed to Smith and Kohn (2002), we identify zeros in the Cholesky factors of **Q** rather than of $\mathbf{Q}^{-1}$. It will be shown that our approach allows us to shrink random effects toward fixed ones, a feature that would not result from a direct application of Smith and Kohn (2002) to the matrix $\mathbf{Q}^{-1}$ appearing in a hierarchical model.

In general, little work has been done on parsimonious variance-covariance selection for hierarchical models, exceptions being Albert and Chib (1993) and Chen and Dunson (2003). The covariance selection approach of Chen

S. Frühwirth-Schnatter (✉)
Department of Applied Statistics and Econometrics, Johannes Kepler Universität Linz, Linz, Austria
e-mail: sylvia.fruehwirth-schnatter@jku.at

R. Tüchler
Department of Statistics and Mathematics, Vienna University of Economics and Business Administration, Vienna, Austria
e-mail: regina.tuechler@wu-wien.ac.at

and Dunson (2003) selects fixed effects in the variance-covariance matrix. However, it is not possible to select the finer structure of zeros in the off-diagonal elements of $\mathbf{Q}$ with their approach. In contrast we consider variable selection on all free elements in the matrix $\mathbf{C}$, appearing in the Cholesky decomposition $\mathbf{Q} = \mathbf{CC}'$, as this leads to a more parsimonious representation of $\mathbf{Q}$ for highly correlated random effects. This Cholesky decomposition of $\mathbf{Q}$ leads in a natural way to a non-centered parameterization of a random-effects model, where all free elements of $\mathbf{C}$ appear as unknown coefficients in a regression type model, which is related to but different from the non-centered parameterization considered by Meng and van Dyk (1998) and Chen and Dunson (2003).

The rest of the article is organized as follows. In Sect. 2 we define a parsimonious representation of the random-effects model. In Sect. 3 we show that a straightforward MCMC scheme for joint variable selection and parameter estimation is available, that involves sampling from standard densities, only. In Sect. 4 we discuss parsimonious covariance selection for simulated data and apply the algorithm to real data coming from marketing in Sect. 5.

2 Parsimonious linear mixed models and variable selection

2.1 Parsimonious linear mixed models

It is common to write the random-effects model in the following centered parameterization for subjects $i = 1, \ldots, N$:

$$\mathbf{y}_i = \mathbf{X}_i^f \boldsymbol{\alpha} + \mathbf{X}_i^r \boldsymbol{\beta}_i^s + \boldsymbol{\varepsilon}_i, \quad \boldsymbol{\varepsilon}_i \sim \mathcal{N}_{T_i}(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}_{T_i}), \tag{1}$$

$$\boldsymbol{\beta}_i^s = \boldsymbol{\beta} + \mathbf{w}_i, \quad \mathbf{w}_i \sim \mathcal{N}_d(\mathbf{0}, \mathbf{Q}). \tag{2}$$

The vector $\mathbf{y}_i$ contains T_i observations and $\mathbf{X}_i^f$ is a design matrix of dimension $T_i \times d_f$ for the d_f-dimensional vector $\boldsymbol{\alpha}$ containing the fixed effects. $\mathbf{X}_i^r$ is the $T_i \times d$-dimensional design matrix for the d-dimensional vector of random effects $\boldsymbol{\beta}_i^s$, which are normally distributed with mean parameter $\boldsymbol{\beta}$ and variance-covariance matrix $\mathbf{Q}$.

While in a standard linear mixed model $\mathbf{Q}$ is an unrestricted variance-covariance matrix depending on $d(d + 1)/2$ unknown parameters, our goal is to represent $\mathbf{Q}$ by a lower dimensional parameter whose dimensionality is determined by the data. It is not straightforward how to do that when working directly with $\mathbf{Q}$ because reducing dimensionality could be achieved only by imposing complicated constraints on the elements Q_{kj} of $\mathbf{Q}$, in order to ensure that the resulting matrix is not negative definite.

To achieve parsimony while avoiding complicated constraints, it is useful to represent $\mathbf{Q}$ by

$$\mathbf{Q} = \mathbf{CC}', \tag{3}$$

where $\mathbf{C}$ is a $(d \times d)$ matrix, because the resulting co-variance matrix will be positive semi-definite for arbitrary matrices $\mathbf{C}$. From (3) we obtain the element Q_{kj} by multiplying the kth and the jth row of $\mathbf{C}$:

$$Q_{kj} = \sum_{l=1}^{d} C_{kl} C_{jl}, \tag{4}$$

hence any diagonal element is given by:

$$Q_{kk} = \sum_{l=1}^{d} C_{kl}^2. \tag{5}$$

In general, the elements of the matrix $\mathbf{C}$ are not uniquely defined. A necessary condition to identify $\mathbf{C}$ is that $\mathbf{C}$, like an unrestricted variance-covariance matrix $\mathbf{Q}$, depends on at most $d(d + 1)/2$ free parameters. One way to achieve this is to set at least $d(d - 1)/2$ elements of $\mathbf{C}$ equal to zero. As it turns out, a more parsimonious representation of $\mathbf{Q}$ results, if more than $d(d - 1)/2$ elements of $\mathbf{C}$ are equal to zero, leading to less than $d(d + 1)/2$ free parameters to represent $\mathbf{C}$ and consequently $\mathbf{Q}$.

However, only for a few special cases is the position of these zero elements uniquely defined by the structure of $\mathbf{Q}$. This is the case, for instance, if $\mathbf{Q}$ is a diagonal matrix. Then $\mathbf{C}$ will be diagonal, too, hence the position of the zeros in $\mathbf{C}$ is unique. Without further assumptions the position of the zero elements of $\mathbf{C}$ is not unique for general matrices $\mathbf{Q}$.

Uniqueness of the position of the zero elements, however, is achieved by considering in (3) the Cholesky decomposition of $\mathbf{Q}$, where $\mathbf{C}$ is a lower triangular matrix, hence for each $j = 2, \ldots, d$ $C_{kj} = 0$ by definition for all $1 \le k < j$. The lower triangular elements of $\mathbf{C}$ are determined recursively from:

$$Q_{kj} = \sum_{l=1}^{\min(j,k)} C_{kl} C_{jl}. \tag{6}$$

First, for each column k, $k = 1, \ldots, d$, the diagonal element C_{kk} is obtained by solving

$$Q_{kk} - \sum_{l=1}^{k-1} C_{kl}^2 = C_{kk}^2, \tag{7}$$

whereas the remaining elements C_{jk} of column k are given for $j = k + 1, \ldots, d$ by:

$$Q_{kj} - \sum_{l=1}^{k-1} C_{kl} C_{jl} = C_{kk} C_{jk}. \tag{8}$$

A careful investigation of recursions (7) and (8) reveals that the positions of zero and non-zero elements of $\mathbf{C}$ is uniquely defined. However, any nonzero element of $\mathbf{C}$ is not unique in

a strict sense. Recursion (7) leads to a diagonal element C_{kk} which is unique iff C_{kk} is equal to 0. If C_{kk} is different from 0, then the sign of C_{kk} evidently is not unique and could be switched without affecting (7), causing all remaining elements of column k, defined in (8), to change sign as well. Evidently, all zero elements in column k are unaffected by sign switching. Furthermore, if recursion (7) leads to a diagonal element C_{kk} which is equal to 0, then the remaining elements C_{jk} of column k are undefined, see again (8), and may be set to any arbitrary value. To achieve further parsimony, we assume in this case that C_{jk} is equal to 0 for $j = k + 1, \ldots, d$.

To distinguish between zero and free elements, we introduce for each element C_{lm}, $m = 1, \ldots, d$, $l = 1, \ldots, d$, an indicator γ_{lm} which takes the value 0, if $C_{lm} = 0$, and 1 if C_{lm} is unconstrained:

$$\gamma_{lm} = 0, \quad \text{iff } C_{lm} = 0,$$
$$\gamma_{lm} = 1, \quad \text{otherwise.} \tag{9}$$

Let $\boldsymbol{\gamma} = \{\gamma_{lm}, m = 1, \ldots, d, l = 1, \ldots, d\}$ denote the matrix containing all indicators. Then, for a given value of $\boldsymbol{\gamma}$, the number $q_{\boldsymbol{\gamma}}$ of free elements in $\mathbf{C}$ is given by:

$$q_{\boldsymbol{\gamma}} = \sum_{l=1}^{d} \sum_{m=1}^{d} \gamma_{lm}. \tag{10}$$

The indicator matrix $\boldsymbol{\gamma}$ turns out to be very useful when analyzing the structural properties of $\mathbf{Q}$. An interesting side effect of using the Cholesky decomposition is that the indicators provide information about the rank of $\mathbf{Q}$, which is given by

$$\mathrm{rg}(\mathbf{Q}) = \mathrm{rg}(\mathbf{C}) = \sum_{k=1}^{d} \gamma_{kk}, \tag{11}$$

because $\mathbf{C}$ is a lower triangular matrix. Rank reduction occurs for two reasons which are mirrored nicely by the zero pattern in $\mathbf{C}$.

First, if one of the random effects, say the kth effect $\beta_{i,k}^{s}$, reduces to a fixed effect then the kth column and the kth row of $\mathbf{Q}$ are necessarily equal to 0, reducing the rank of $\mathbf{Q}$ by one. Because $Q_{kk} = 0$, (5) implies immediately that the kth row (and consequently the kth column) of $\mathbf{C}$ must be equal to 0 as well. Thus a random effect $\beta_{i,k}^{s}$ reduces to a fixed effect iff the kth row of $\mathbf{C}$ contains only zeros. This allows us to identify fixed effects immediately from the indicator matrix $\boldsymbol{\gamma}$. For each effect k, $k = 1, \ldots, d$, we introduce a fixed effect indicator δ_{k}^{F}, which is defined by

$$\delta_{k}^{F} = \prod_{l=1}^{k} (1 - \gamma_{kl}), \tag{12}$$

and takes the value 1 iff the kth effect $\beta_{i,k}^{s}$ reduces to a fixed effect. Whenever any element in the kth row of $\mathbf{C}$ is different from zero, then $\beta_{i,k}^{s}$ is a truly random effect which deviates randomly from β_{k}. If $\mathrm{rg}(\mathbf{Q})$ is equal to $d - \sum_{k=1}^{d} \delta_{k}^{F}$, the total number of truly random effects, then the presence of fixed effects is the only reason for rank deficiency in $\mathbf{Q}$.

Further rank reduction occurs if the $d - \sum_{k=1}^{d} \delta_{k}^{F}$ truly random effects are linearly dependent and are related to some common factor of dimension $d - \sum_{k=1}^{d} \delta_{k}^{F} - p$ with $p > 0$. This will cause p columns of $\mathbf{C}$ to contain only zeros, while the corresponding lines contain non-zero elements. The number p of such columns is related to the indicators γ_{kl} by

$$p = \sum_{k=1}^{d} (1 - \gamma_{kk} - \delta_{k}^{F}). \tag{13}$$

If all effects are truly random ($\delta_{k}^{F} = 0$ for all $k = 1, \ldots, d$) while $\mathrm{rg}(\mathbf{Q}) = d - p < d$, the Cholesky decomposition leads to a kind of factor analytical representation of $\mathbf{Q}$ in terms of a $(d \times p)$ matrix $\mathbf{C}^{\star}$ which is obtained from $\mathbf{C}$ simply by deleting all zero columns. In general, $\mathbf{C}$ contains $d - \mathrm{rg}(\mathbf{Q})$ zero columns and $d - \mathrm{rg}(\mathbf{Q}) - p$ zero lines, where the position of the lines indicates which effects are fixed.

While many of the zeros in $\mathbf{C}$ will be caused by these two structural properties of the random effects, additional zeros might be present outside zero lines and zero columns, leading to a further reduction of $q_{\boldsymbol{\gamma}}$, the number of free parameters needed to represent $\mathbf{Q}$. We may think of these zero elements as achieving parsimony in describing the finer structure of $\mathbf{Q}$, like independence only of specific components of $\boldsymbol{\beta}_{i}^{s}$, say β_{ij}^{s} and β_{ik}^{s}. It has to be noted that the Cholesky decomposition does not necessarily lead to the most parsimonious representation of $\mathbf{Q}$ in terms of being, among all possible decompositions $\mathbf{Q} = \mathbf{C}\mathbf{C}'$, where $\mathbf{C}$ is an arbitrary $(d \times d)$ matrix, the one which yields the smallest number of nonzero elements. Nevertheless, as our case study from marketing will show, $q_{\boldsymbol{\gamma}}$ is typically much smaller than $d(d + 1)/2$, even if we stay within the framework of the Cholesky decomposition.

A final aspect which needs to be discussed is how the ordering of the random effects influences the representation of $\mathbf{Q}$ we obtain. As opposed to Smith and Kohn (2002), who considered longitudinal data, there is no natural ordering of the variables in a random effects model, and there exist $d!$ different ways of arranging the components of $\boldsymbol{\beta}_{i}^{s}$. Several characteristics of the corresponding matrices $\mathbf{C}$ will remain unaffected by reordering, while others change. First of all, the rank of $\mathbf{Q}$, and consequently, the rank of $\mathbf{C}$ as well as the number $\sum_{k=1}^{d} \delta_{k}^{F}$ of fixed effects and the number p of zero columns resulting from linear dependence among the truly random effects obviously are invariant under reordering.

The zero pattern, however, will change under reordering according to a permutation ρ. Assume that for a given ordering of the variables the variance-covariance matrix $\mathbf{Q}$ is represented by $\mathbf{Q} = \mathbf{C}\mathbf{C}'$, with $\mathbf{C}$ being a Cholesky factor containing q_γ free elements. When permuting the ordering of the random effects, we obtain

$$\mathbf{Q}_\rho = \mathbf{\Pi}_\rho \mathbf{Q} \mathbf{\Pi}_\rho' = (\mathbf{\Pi}_\rho \mathbf{C})(\mathbf{\Pi}_\rho \mathbf{C})' \tag{14}$$

for a suitable permutation matrix $\mathbf{\Pi}_\rho$, interchanging rows and columns. Note that $\mathbf{\Pi}_\rho \mathbf{C}$ is a decomposition of $\mathbf{Q}$ with the same number q_γ of free elements as $\mathbf{C}$, however, in general it is no longer a Cholesky decomposition.

If we reorder the variables before deriving the Cholesky decomposition, we obtain

$$\mathbf{Q}_\rho = \mathbf{C}_\rho \mathbf{C}_\rho', \tag{15}$$

where certain aspects of the zero pattern in $\mathbf{C}$ and $\mathbf{C}_\rho$ will match. Zero lines and zero columns in $\mathbf{C}$ corresponding to a fixed effect will change their positions according to the selected permutation and will correspond to the appropriate zero lines and zero columns in $\mathbf{C}_\rho$. Thus identifying fixed effects by searching for zero lines (and zero columns) in the Cholesky factor is invariant to reordering the variables and the fixed effect indicator δ_k^F will be zero iff the fixed effect indicator $\delta_{k'}^F$ where $k' = \rho(k)$ is zero. The position, but not the number of additional zero columns corresponding to linear dependence, will change.

However, the total number of free elements, q_γ, may vary for different permutations ρ. To give an example, we define the following variance-covariance matrix $\mathbf{Q}$ of rank two:

$$\mathbf{Q} = \begin{pmatrix} 4 & -2 & 2 & 0 \\ -2 & 5 & 0 & 0 \\ 2 & 0 & 1.25 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

$$\mathbf{C} = \begin{pmatrix} 2 & 0 & 0 & 0 \\ -1 & 2 & 0 & 0 \\ 1 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{16}$$

Rank reduction stems from one fixed effect at position four and from linear dependence of the truly random effects. Both is reflected by the pattern of zeros in the Cholesky factor $\mathbf{C}$: the fourth row and the third and the fourth column are zero rows or columns, respectively. Thus the number of free elements q_γ in $\mathbf{C}$ equals five.

If we reverse the order of the random effects, so that $\rho = [4\,3\,2\,1]$, we obtain the following $\mathbf{Q}_\rho$ with Cholesky fac-

tor $\mathbf{C}_\rho$:

$$\mathbf{Q}_\rho = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1.25 & 0 & 2 \\ 0 & 0 & 5 & -2 \\ 0 & 2 & -2 & 4 \end{pmatrix},$$

$$\mathbf{C}_\rho = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1.12 & 0 & 0 \\ 0 & 0 & 2.24 & 0 \\ 0 & 1.79 & -.89 & 0 \end{pmatrix} \tag{17}$$

As expected, the position of the zero columns and the zero line in $\mathbf{C}_\rho$ is changed according to the permutation. Note, however, that the rank of $\mathbf{Q}$ which is defined by the number of non-zero diagonal elements in $\mathbf{C}_\rho$ and the number of fixed effects, defined by the number of zero lines, stays unaffected. Concerning the finer structure of $\mathbf{Q}$, we find that $\beta_{i,2}^s$ and $\beta_{i,3}^s$ are uncorrelated, causing an additional zero element in $\mathbf{C}_\rho$. The total number of free elements q_γ in the Cholesky matrix $\mathbf{C}_\rho$ equals four, rather than five, after reordering.

This suggests that reordering the random effects may yield a more parsimonious representation of the variance-covariance $\mathbf{Q}$.

2.2 Variable selection in the non-centered parameterization of the linear mixed model

Parameterization (1) and (2) of the linear mixed model is known as the *centered* parameterization. The parameterization of the random-effects model turns out to be of enormous importance for the convergence behavior of various estimation methods, in particular for MCMC estimation, as analyzed by Gelfand et al. (1995) and Papaspiliopoulos et al. (2003). Convergence often improves, when the random-effects model is formulated in the *non-centered* parameterization, introduced by Meng and van Dyk (1998), and studied in great detail in Van Dyk and Meng (2001). In these papers, the non-centered parameterization reads

$$\mathbf{y}_i = \mathbf{X}_i^f \boldsymbol{\alpha} + \mathbf{X}_i^r \boldsymbol{\beta} + \mathbf{X}_i^r \mathbf{L}\mathbf{D}\mathbf{z}_i^s + \boldsymbol{\varepsilon}_i,$$

$$\boldsymbol{\varepsilon}_i \sim \mathcal{N}_{T_i}(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}_{T_i}), \tag{18}$$

$$\mathbf{z}_i^s \sim \mathcal{N}_d(\mathbf{0}, \mathbf{I}_d), \tag{19}$$

and is based on the Cholesky decomposition $\mathbf{Q} = \mathbf{L}\mathbf{D}\mathbf{L}'$, where $\mathbf{L}$ is a lower triangular matrix with ones in the diagonal, and $\mathbf{D}$ is a diagonal matrix.

The Cholesky decomposition (3) of the random-effects covariance matrix $\mathbf{Q}$, introduced in the previous subsection, leads to a similar, if slightly different, non-centered parameterization. Since the random effects $\boldsymbol{\beta}_i^s$ may be represented as $\boldsymbol{\beta}_i^s = \boldsymbol{\beta} + \mathbf{C}\mathbf{z}_i^s$, where $\mathbf{z}_i^s \sim \mathcal{N}_d(\mathbf{0}, \mathbf{I}_d)$, we obtain

the following non-centered parameterization of a hierarchical linear mixed model:

$$\mathbf{y}_i = \mathbf{X}_i^f \boldsymbol{\alpha} + \mathbf{X}_i^r \boldsymbol{\beta} + \mathbf{X}_i^r \mathbf{C} \mathbf{z}_i^s + \boldsymbol{\varepsilon}_i,$$

$$\boldsymbol{\varepsilon}_i \sim \mathcal{N}_{T_i}(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}_{T_i}), \tag{20}$$

$$\mathbf{z}_i^s \sim \mathcal{N}_d(\mathbf{0}, \mathbf{I}_d), \tag{21}$$

which has computational advantages over the centered parameterization similar to the non-centered parameterization introduced by Meng and van Dyk (1998).

An even more important advantage of using the Cholesky decomposition (3), together with this non-centered parameterization, is that it suggests a quite natural way of finding parsimonious variance-covariance matrices for a hierarchical linear mixed model. Combining these two techniques reduces the problem of variance-covariance selection for a hierarchical linear mixed model to the more common problem of Bayesian variable selection in multiple regression models, as reviewed, for instance, in George and McCulloch (1997).

This relation becomes more evident by rewriting the observation equation (20) as follows. Depending on the indicators $\boldsymbol{\gamma}$, various elements of $\mathbf{C}$ will be restricted to 0, whereas the remaining free elements of $\mathbf{C}$ are treated as an unknown parameter denoted by $\mathbf{C}^{\boldsymbol{\gamma}}$. The parameter vector $\mathbf{C}^{\boldsymbol{\gamma}}$ is constructed from $\mathbf{C}$ by stacking the free elements of $\mathbf{C}$ column by column. For known random effects $\mathbf{z}_i^s$, observation equation (20) may be regarded as following regression model in $\mathbf{C}^{\boldsymbol{\gamma}}$:

$$\mathbf{y}_i = \mathbf{X}_i^f \boldsymbol{\alpha} + \mathbf{X}_i^r \boldsymbol{\beta} + \mathbf{W}_i^{\boldsymbol{\gamma}} \mathbf{C}^{\boldsymbol{\gamma}} + \boldsymbol{\varepsilon}_i, \quad \boldsymbol{\varepsilon}_i \sim \mathcal{N}_{T_i}(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}_{T_i}), \tag{22}$$

where the predictor matrix $\mathbf{W}_i^{\boldsymbol{\gamma}}$ depends on the design matrix $\mathbf{X}_i^r$, and on the latent random effects $\mathbf{z}_i^s$. We will provide details of how $\mathbf{W}_i^{\boldsymbol{\gamma}}$ is constructed at the end of this subsection. Like in standard Bayesian variable selection, elements in the predictor matrix $\mathbf{W}_i^{\boldsymbol{\gamma}}$ will be included or deleted, depending on $\boldsymbol{\gamma}$. As a notable difference, however, variable selection in (22) is carried out with respect to predictors that are latent rather than directly observed.

For a fixed value of $\boldsymbol{\gamma}$, $\mathbf{W}_i^{\boldsymbol{\gamma}}$ is constructed from the design matrix $\mathbf{X}_i^r$ and the latent random effects $\mathbf{z}_i^s = (z_{i1}^s, \ldots, z_{id}^s)'$ in the following way: The matrix $\mathbf{W}_i^{\boldsymbol{\gamma}}$ consists of d submatrices with different number of columns,

$$\mathbf{W}_i^{\boldsymbol{\gamma}} = \left(\mathbf{W}_i^{\boldsymbol{\gamma} \cdot 1} z_{i1}^s \quad \cdots \quad \mathbf{W}_i^{\boldsymbol{\gamma} \cdot d} z_{id}^s \right),$$

where for all $m = 1, \ldots, d$, $\mathbf{W}_i^{\boldsymbol{\gamma} \cdot m}$ is the regressor matrix of all non-zero elements of the column $\mathbf{C}_{\cdot m}$ of $\mathbf{C}$. To account for zero elements in column $\mathbf{C}_{\cdot m}$, the lth columns of $\mathbf{X}_i^r$ has to be deleted whenever $\gamma_{lm} = 0$, in order to obtain $\mathbf{W}_i^{\boldsymbol{\gamma} \cdot m}$.

2.3 Related work

Our approach of finding a parsimonious variance-covariance matrix through Bayesian variable selection is related to Smith and Kohn (2002) and Chen and Dunson (2003), but differs from these papers in various important aspects.

By performing variable selection on the Cholesky decomposition of $\mathbf{Q}$, our approach is substantially different from Smith and Kohn (2002), who use the Cholesky decomposition $\mathbf{Q}^{-1} = \mathbf{LDL}'$ of the inverse of $\mathbf{Q}$, where $\mathbf{L}$ is a lower triangular matrix with ones in the diagonals and $\mathbf{D}$ is a diagonal matrix of full rank. Smith and Kohn (2002) introduce only $d(d-1)/2$ indicators γ_{lm} to perform variable selection on the strictly lower diagonal elements of $\mathbf{L}$, whereas the elements of $\mathbf{D}$ are assumed to be positive. If all indicators are equal to 1, all $d(d-1)/2$ elements of $\mathbf{L}$ are unconstrained, leading to the estimation of an arbitrary positive definite variance-covariance matrix $\mathbf{Q}$ as in our approach. If all indicator are equal to 0, however, $\mathbf{Q}$ is shrunk toward the diagonal matrix $\mathbf{D}^{-1}$. Thus a direct application of the Smith and Kohn (2002) approach to the inverse of the variance-covariance matrix of a random-effects model would not allow us to reduce any of the random effects to a fixed one.

Our approach is related to Chen and Dunson (2003), who apply a similar but more specific approach to the Cholesky decomposition $\mathbf{Q} = \mathbf{DLL}'\mathbf{D}$, where $\mathbf{L}$ is a lower triangular matrix with ones in the diagonal and $\mathbf{D}$ is a diagonal matrix. In order to reduce random effects to fixed ones, they allow the diagonal elements of $\mathbf{D}$ to have a positive probability of being zero, whereas no variable selection is performed for the elements of $\mathbf{L}$. Thus our approach is more general than theirs, as we introduce variable selection also on the lower diagonal elements of the Cholesky factor, and therefore are able to capture the finer structure of $\mathbf{Q}$, which is especially important in higher dimensional problems.

3 Bayesian estimation

3.1 Prior distributions

3.1.1 Prior for the indicator matrix $\boldsymbol{\gamma}$

For Bayesian estimation one has to select the prior of the indicator matrix $\boldsymbol{\gamma}$. Conditional on a known value $\tau \in [0, 1]$ we assume priori independence indicator variables γ_{lm} with $\Pr(\gamma_{lm} = 1|\tau) = \tau$. This implies that the number of non-zero elements in $\mathbf{C}$ follows the binomial distribution $\text{BiNom}(d_s, \tau)$, where $d_s = d(d+1)/2$ is the total number of free parameters in $\mathbf{C}$. For variance-covariance matrices $\mathbf{Q}$ of moderate size, this density is fairly non-informative on the number of non-zero elements, whereas with increasing

number of elements this density approaches a normal distribution with mean $d_s\tau$ and variance $d_s\tau(1-\tau)$ and the prior distribution of $q_{\boldsymbol{\gamma}}$, the number of non-zero elements in $\mathbf{C}$, will be crucially dependent on τ.

To reduce the sensitivity with respect to choosing τ, we consider it as a hyperparameter and use a uniform prior for τ on $[0, 1]$ as in Smith and Kohn (2002). If we integrate the hyperparameter τ out of the analysis, we obtain:

$$p(\boldsymbol{\gamma}) = \int p(\boldsymbol{\gamma}|\tau)p(\tau)d\tau = \text{Beta}(q_{\boldsymbol{\gamma}}+1, d_s-q_{\boldsymbol{\gamma}}+1). \tag{23}$$

Here, $\text{Beta}(\cdot, \cdot)$ is the beta function and $q_{\boldsymbol{\gamma}}$ is the number of non-zero elements in $\mathbf{C}$, see also (10). Note that the marginal prior (23) implies *a priori* dependence between the elements of the indicator matrix $\boldsymbol{\gamma}$.

Within our MCMC sampling scheme we need the conditional prior for one element γ_{lm} given the remaining elements of $\boldsymbol{\gamma}$, denoted $\boldsymbol{\gamma}_{\backslash lm}$. Let us first consider the case where $\gamma_{lm}^{\text{old}} = 1$. We derive the following conditional priors:

$$p(\gamma_{lm} = 0|\boldsymbol{\gamma}_{\backslash lm}) = h_1/(h_1+1),$$
$$p(\gamma_{lm} = 1|\boldsymbol{\gamma}_{\backslash lm}) = 1/(h_1+1).$$

Here

$$h_1 = \frac{d_s - q_{\boldsymbol{\gamma}} + 1}{q_{\boldsymbol{\gamma}}},$$

and $q_{\boldsymbol{\gamma}}$ is the number of non-zero elements in $\mathbf{C}^{\text{old}}$. Note that $1/(h_1 + 1) \approx \hat{\tau}$, where $\hat{\tau} = q_{\boldsymbol{\gamma}}/d_s$ is the estimated fraction of non-zero elements in $\mathbf{C}$. If $\gamma_{lm}^{\text{old}} = 0$, then

$$p(\gamma_{lm} = 0|\boldsymbol{\gamma}_{\backslash lm}) = h_0/(h_0+1),$$
$$p(\gamma_{lm} = 1|\boldsymbol{\gamma}_{\backslash lm}) = 1/(h_0+1),$$

where

$$h_0 = \frac{d_s - q_{\boldsymbol{\gamma}}}{q_{\boldsymbol{\gamma}} + 1}.$$

3.1.2 Fractional prior for the variance-covariance matrix of the random effects

Like in variable selection problems for the standard regression model, the specific choice of a prior for the Cholesky factor $\mathbf{C}$ is likely to be rather influential on the posterior of the indicator matrix $\boldsymbol{\gamma}$, see O'Hagan (1995) and George and McCulloch (1997). We extend the fractional prior approach introduced by O'Hagan (1995) to the present context of selecting the prior for the variance-covariance matrix of the random effects in hierarchical linear models. Fractional priors were first introduced to Bayesian estimation of variance-covariance matrices by Smith and Kohn (2002),

who use a fractional prior for the non-zero elements of the off-diagonal elements of $\mathbf{L}$ in the Cholesky decomposition $\mathbf{Q}^{-1} = \mathbf{LDL}'$.

It is, however, not at all clear how to define a fractional prior for a hierarchical linear model which is a latent variable model. As common in Gibbs sampling for latent variable models, we are using the conditional likelihood $p(\mathbf{y}|\boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{z}^s, \sigma_\varepsilon^2, \mathbf{C}^{\boldsymbol{\gamma}})$, where the random effects $\mathbf{z}^s = (\mathbf{z}_1^s, \ldots, \mathbf{z}_N^s)$ are fixed, for defining the fractional prior. Hence our prior is a conditionally fractional prior depending on the random effects.

We construct a fractional prior for $\mathbf{C}^{\boldsymbol{\gamma}}$ from a part of the conditional likelihood $p(\mathbf{y}|\boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{z}^s, \sigma_\varepsilon^2, \mathbf{C}^{\boldsymbol{\gamma}})$, namely a fraction $b \in (0, 1)$. The fractional prior is easily shown to be the density of a multivariate normal distribution,

$$p(\mathbf{C}^{\boldsymbol{\gamma}}|\mathbf{z}^s, \sigma_\varepsilon^2, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{y}^{TN \times b}) \sim \mathcal{N}_{q_{\boldsymbol{\gamma}}}(\mathbf{a}_N, \sigma_\varepsilon^2 \mathbf{A}_N/b), \tag{24}$$

where $\mathbf{a}_N$ and $\mathbf{A}_N$ are given by

$$\mathbf{a}_N = \mathbf{A}_N \left(\sum_{i=1}^{N} (\mathbf{W}_i^{\boldsymbol{\gamma}})'(\mathbf{y}_i - \mathbf{X}_i^f \boldsymbol{\alpha} - \mathbf{X}_i^r \boldsymbol{\beta}) \right) \tag{25}$$

$$\mathbf{A}_N^{-1} = \sum_{i=1}^{N} (\mathbf{W}_i^{\boldsymbol{\gamma}})' \mathbf{W}_i^{\boldsymbol{\gamma}}.$$

Details are given in Appendix 1. Following Smith and Kohn (2002), we choose the fraction b for the fractional prior equal to $1/\sum_{i=1}^{N} T_i$.

We finally note that the marginal prior $p(\mathbf{C}^{\boldsymbol{\gamma}}|\sigma_\varepsilon^2, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{y}^{TN \times b})$ where the random effects $\mathbf{z}^s$ are integrated out is an infinite mixture of these conditionally fractional priors, but not a fractional prior with respect to the marginal likelihood $p(\mathbf{y}|\mathbf{C}^{\boldsymbol{\gamma}}, \sigma_\varepsilon^2, \boldsymbol{\alpha}, \boldsymbol{\beta})$, where the random effects are integrated out.

This prior is related to the one introduced in Smith and Kohn (2002), who realized that for data from a multivariate normal distribution with unknown variance-covariance matrix $\mathbf{Q}$, the normal distribution is a natural conjugate conditional prior for the free elements of the lower triangular matrix L in the Cholesky decomposition $\mathbf{Q}^{-1} = \mathbf{LDL}'$. In the context of random-effects models, a conditionally conjugate normal prior for the Cholesky factors of the variance-covariance matrix $\mathbf{Q}$ was independently suggested by Tüchler and Frühwirth-Schnatter (2003) and Chen and Dunson (2003).

It is worth mentioning that the prior we consider in this article is different from the prior in Chen and Dunson (2003), who considered the Cholesky decomposition $\mathbf{Q} = \mathbf{DLL}'\mathbf{D}$, in various aspects. Chen and Dunson (2003) use a conditionally normal prior on the free elements of the lower triangular matrix $\mathbf{L}$, and consider a zero-inflated half-normal distribution for the free elements in the diagonal

matrix $\mathbf{D}$, consisting of a mass point at zero (with probability $1 - \tau$) and a normal density with mean $\mathbf{a}_0$ and variance $\mathbf{A}_0$ truncated below zero. Their prior may be formulated in terms of d variable indicators γ_l, $l = 1, \ldots, d$, for the d free elements of $\mathbf{D}$, in which case τ is found to be the prior probability of $\gamma_l = 1$. Chen and Dunson (2003) hold τ fixed for posterior inference. As discussed above, fixing τ will be of considerable influence on posterior inference within increasing size of $\mathbf{Q}$, whereas our prior is more flexible. Second, we include the diagonal into the Cholesky decomposition, which allows us to define a normal prior on all non-zero elements of $\mathbf{C}$, not only on the lower triangular matrix $\mathbf{L}$.

3.1.3 Remaining priors

A prior for the mean parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ and the observation error variance σ_ε^2 remains to be chosen. For the mean parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ we assume a joint normal prior distribution $\mathcal{N}_{d+d_f}(\mathbf{b}_0, \mathbf{B}_0)$, whereas the observation error variance σ_ε^2 is a priori inverted gamma distributed with $\mathcal{G}^{-1}(s_0/2, S_0/2)$.

3.2 MCMC sampling

3.2.1 The MCMC scheme

We introduce an MCMC scheme which simultaneously carries out model selection and estimation of all unknown parameters. The non-centered parameterization based on the Cholesky decomposition, together with the priors defined in Sect. 3.1, give way to the following convenient sampling scheme involving standard densities, only:

(i) Sample $\gamma_{lm} | \boldsymbol{\gamma}_{\setminus lm}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{z}^s \sigma_\varepsilon^2, \mathbf{y}$, where $\boldsymbol{\gamma}_{\setminus lm}$ denotes all elements of the indicator matrix $\boldsymbol{\gamma}$ but the element γ_{lm}, from a discrete density with two realizations.
(ii) Sample $\mathbf{C}^{\boldsymbol{\gamma}} | \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{z}^s, \sigma_\varepsilon^2, \mathbf{y}$ from a normal distribution.
(iii) Sample $\boldsymbol{\alpha}, \boldsymbol{\beta} | \mathbf{C}^{\boldsymbol{\gamma}}, \sigma_\varepsilon^2, \mathbf{y}$ from a normal distribution.
(iv) Sample $\mathbf{z}^s | \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{C}^{\boldsymbol{\gamma}}, \sigma_\varepsilon^2, \mathbf{y}$ from a normal distribution.
(v) Sample $\sigma_\varepsilon^2 | \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{z}^s, \mathbf{y}$ from an inverted Gamma distribution.

Subsequently, we will discuss each step in more detail.

3.2.2 Sampling the indicators and the Cholesky factors

The most crucial part of our algorithm is sampling the parsimonious variance-covariance matrix of the random effects. Based on the non-centered parameterization, we sample the Cholesky factor $\mathbf{C}$ of the variance-covariance matrix $\mathbf{Q}$ rather than the matrix itself in two steps. First, we sample the indicator for each of the $d(d + 1)/2$ free elements of the Cholesky factor from the marginal conditional density $p(\gamma_{lm} | \boldsymbol{\gamma}_{\setminus lm}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{z}^s, \sigma_\varepsilon^2, \mathbf{y})$, where $\boldsymbol{\gamma}_{\setminus lm}$ denotes all elements of the indicator matrix $\boldsymbol{\gamma}$ but the element γ_{lm}. Then,

conditional on knowing $\boldsymbol{\gamma}$, all non-zero elements $\mathbf{C}^{\boldsymbol{\gamma}}$ of $\mathbf{C}$ are sampled from the appropriate distribution.

Note that the density $p(\gamma_{lm} | \boldsymbol{\gamma}_{\setminus lm}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{z}^s, \sigma_\varepsilon^2, \mathbf{y})$ is marginalized over the Cholesky factors in order to avoid the computational problems discussed e.g. in George and Mc-Culloch (1997). To implement this step, the marginal likelihood $p(\mathbf{y} | \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{z}^s, \sigma_\varepsilon^2)$ where $\mathbf{C}^{\boldsymbol{\gamma}}$ is integrated out is required.

The marginal likelihood function under fractional priors
To combine the fractional prior with the information in the data in a variable selection context there are basically two routes to follow. The first approach, pursued by Smith and Kohn (2002), is to combine the fractional prior with the complete likelihood $p(\mathbf{y} | \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{z}^s, \sigma_\varepsilon^2, \mathbf{C}^{\boldsymbol{\gamma}})$. This means, however, using a fraction of the data, namely $100b$ percent, twice (both in the prior and in the likelihood).

Following O'Hagan (1995), we pursue the alternative approach, where information used for constructing the prior does not reappear in the likelihood. We define what could be called a fractional marginal likelihood for model selection in random-effects models, by combining the fractional prior with the remaining likelihood $p(\mathbf{y} | \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{z}^s, \sigma_\varepsilon^2, \mathbf{C}^{\boldsymbol{\gamma}})^{1-b}$:

$$p(\mathbf{y} | \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{z}^s, \sigma_\varepsilon^2)$$
$$= \int p(\mathbf{y} | \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{z}^s, \sigma_\varepsilon^2, \mathbf{C}^{\boldsymbol{\gamma}})^{1-b}$$
$$\times p(\mathbf{C}^{\boldsymbol{\gamma}} | \mathbf{z}^s, \sigma_\varepsilon^2, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{y}^{TN \times b}) d\mathbf{C}^{\boldsymbol{\gamma}}, \qquad (26)$$

where $p(\mathbf{C}^{\boldsymbol{\gamma}} | \mathbf{z}^s, \sigma_\varepsilon^2, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{y}^{TN \times b})$ is equal to the fractional prior (24). As only quadratic forms in $\mathbf{C}^{\boldsymbol{\gamma}}$ are involved both in the fractional prior as well as in the conditional likelihood, it is possible to carry out integration with respect to $\mathbf{C}^{\boldsymbol{\gamma}}$ explicitly in (26):

$$p(\mathbf{y} | \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{z}^s, \sigma_\varepsilon^2) = b^{q_{\boldsymbol{\gamma}}/2} \left(\frac{1}{2\pi \sigma_\varepsilon^2} \right)^{\sum_{i=1}^{N} T_i(1-b)/2}$$
$$\times \exp\left(-\frac{(1-b)}{2\sigma_\varepsilon^2} S^{\boldsymbol{\gamma}} \right), \qquad (27)$$

where $q_{\boldsymbol{\gamma}} = \dim(\mathbf{C}^{\boldsymbol{\gamma}})$ and

$$S^{\boldsymbol{\gamma}} = \sum_{i=1}^{N} \| \mathbf{y}_i - \mathbf{W}_i^{\boldsymbol{\gamma}} \mathbf{a}_N - \mathbf{X}_i^f \boldsymbol{\alpha} - \mathbf{X}_i^r \boldsymbol{\beta} \|_2. \qquad (28)$$

Sampling the indicator matrix To sample the indicator matrix, we proceed columnwise for $l = 1, \ldots, d$ and sample the diagonal elements γ_{ll} first. If the indicator of the lth diagonal element is zero we set all elements of the lth column to zero to preserve identification. If the indicator of the lth diagonal element equals one the remaining column-elements γ_{ml} are sampled for $m = l + 1, \ldots, d$.

To sample the indicators γ_{lm} we could apply a Gibbs sampler. However, efficient sampling of this step is essential for the speed of the algorithm. Smith and Kohn (2002) have developed a fast algorithm for sampling the indicators and we apply their scheme to our model here. Details may be found in Appendix 2.

Sampling $\mathbf{C}^{\gamma}$ We generate $\mathbf{C}^{\gamma}|\gamma, \alpha, \beta, \mathbf{z}^s, \sigma_{\varepsilon}^2, \mathbf{y}$ from the following normal posterior distribution:

$$\mathbf{C}^{\gamma}|\gamma, \mathbf{z}^s, \alpha, \beta, \sigma_{\varepsilon}^2, \mathbf{y} \sim \mathcal{N}_{q_{\gamma}}(\mathbf{a}_N, \sigma_{\varepsilon}^2 \mathbf{A}_N),$$

where $\mathbf{a}_N$ and $\mathbf{A}_N$ are given in (25).

Note that we obtain equal likelihoods if we change the sign of whole columns of $\mathbf{C}$, see Sect. 2.1. Inference about $\mathbf{Q}$, like, for example, estimation of the rank and the number of fixed effects is independent of these sign-switches. The number of free elements q_{γ} in $\mathbf{C}$ is also not affected by sign-switching. Therefore we do not need to identify a unique matrix out of all 2^d possible sign-combinations for the estimation problems of our paper. However, inference about quantities of $\mathbf{C}$, like for example estimation of the posterior mean, would not yield a meaningful result, and unique identification of the column signs would be necessary. Sign switching could be avoided by posing a formal non-negativity constraint on C_{kk}, like $C_{kk} > 0$.

3.2.3 Sampling the remaining parameters

Conditional on knowing γ and $\mathbf{C}^{\gamma}$, we are dealing with a random-effects model with known variance-covariance matrix $\mathbf{Q} = \mathbf{CC}'$, where $\mathbf{C}$ is the Cholesky factor with those elements set to zero which were indicated by γ. Consequently, one could use any of the available MCMC schemes from the literature in order to sample α, β, σ_{ε}^2, and the non-centered random effects $\mathbf{z}^s$. Here we use the partially marginalized sampler of Frühwirth-Schnatter et al. (2004), which samples the fixed effects and the mean parameters efficiently without conditioning on the random effects. We provide details in Appendix 3.

4 Simulation study

We include the variance-covariance matrices from our example of Sect. 2.1 in our simulation study. We generate a data set from a random-effects model with variance-covariance matrix $\mathbf{Q}$ given in equation (16) and obtain a second data set by simple reordering of the effects according to permutation $\rho = [4\,3\,2\,1]$, which yields the covariance matrix $\mathbf{Q}_{\rho}$ of (17) for this new data set. Details of the design and other parameters are given in Appendix 4. We base our analysis on 30,000 iterations after a burn-in of 5,000 iterations.

Table 1 Simulation study: relative frequency for the number of free elements q_{γ} in $\mathbf{C}$ ($\mathbf{C}_{\rho}$)

	4	5	6	7	8
Equation (16)	.18	.46	.24	.09	.02
Equation (17)	.54	.33	.09	.03	.01

Table 2 Simulation study: Posterior modes and in parentheses their relative frequency for: the number of fixed effects (column 1) and the rank of $\mathbf{Q}$ ($\mathbf{Q}_{\rho}$) (column 2)

	No. fixed eff.	Rank
Equation (16)	1 (.72)	2 (.59)
Equation (17)	1 (.84)	2 (.59)

In Table 1 and Table 2 we compare posterior modes for the number of free elements q_{γ} in the Cholesky factors, the rank of the variance-covariance matrices, and the number of fixed effects for the two data sets. In Sect. 2.1 we stated that the rank of the variance-covariance matrix as well as the number of fixed effects is invariant towards a reordering of the random effects. This is also reflected by the posterior modes of these measures, which are equal for both MCMC simulations. Furthermore, we noted in Sect. 2.1 that the number of free elements q_{γ} might vary for different permutations of the effects and that (17) yields a more parsimonious Cholesky factor than (16). This may again be observed for the MCMC simulations, where the posterior mode of q_{γ} for the latter case equals five, whereas this value is four for the ordering $\rho = [4\,3\,2\,1]$, see Table 1.

It is worth mentioning that the difference in q_{γ} also affects the posterior number of free elements in the variance-covariance matrix $\mathbf{Q}$. Note that indicators for unrestricted elements in $\mathbf{Q}$ may easily be derived at each MCMC iteration by multiplication of the indicator matrices of $\mathbf{C}$: $\gamma\gamma'$. In this new matrix, non-zero elements indicate unrestricted elements in $\mathbf{Q}$. In Table 3 we give posterior probabilities for the elements of $\mathbf{C}_{\rho}$ and of $\mathbf{Q}_{\rho}$ to be unrestricted. We observe that these posterior probabilities match the true pattern of non-zeros in (17) very well. But if we compare the posterior probabilities for $\mathbf{Q}$ to be unrestricted with $\mathbf{Q}$ of (16) we find that the zero element Q_{23} stays unrestricted throughout the MCMC simulations, see Table 4. There is an obvious explanation for this. The indicator matrix matching $\mathbf{C}$ of (16) would be

$$\gamma = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

The indicator matrix for the elements of $\mathbf{Q}$, $\gamma\gamma'$, has only unrestricted elements in the upper-left three times three

block. However, the posterior mean for element Q_{23} is close to zero and equals $-.35$, with a standard deviation of $.43$.

5 Application to real data

Our application comes from a brand-price trade-off study in the Austrian mineral water market. These data are challenging due to the high dimension of the variance-covariance matrix and the power of the new method may be demonstrated here. 213 consumers stated their likelihood to buy mineral water products on a 20 point rating scale. Five different brands were offered at three different prices levels. Therefore our data consist of 15 observations per consumer.

Table 3 Simulation study with $\mathbf{Q}_\rho$ from (17): posterior probabilities for the elements of $\mathbf{C}_\rho$ and $\mathbf{Q}_\rho$ to be unrestricted (rounded)

$\mathbf{C}_\rho$				$\mathbf{Q}_\rho$			
.16	0	0	0	.16	.02	.02	.02
.02	1	0	0	–	1	.1	1
.02	.1	1	0	–	–	1	1
.02	1	1	.32	–	–	–	1

Table 4 Simulation study with $\mathbf{Q}$ from (16): posterior probabilities for the elements of $\mathbf{C}$ and $\mathbf{Q}$ to be unrestricted (rounded)

$\mathbf{C}$				$\mathbf{Q}$			
1	0	0	0	1	1	1	.04
1	1	0	0	–	1	1	.08
1	.7	.29	0	–	–	1	.11
.04	.04	.05	.19	–	–	–	.28

The design matrices were defined in a way that effects of brands, prices, quadratic prices as well as interaction effects between brands and prices could be investigated. Details on this brand-price trade off study from the marketing point of view may be found in Otter et al. (2004). The design matrix $\mathbf{X}_i^r$ consists of 15 rows for the 15 observations per consumer and of 15 columns: 5 brand columns (one brand as the baseline), one price and one quadratic price column, four brand-by-linear-price and four brand-by-quadratic-price columns.

We reanalyzed these data, starting with a general model structure where all effects were specified as random effects and ran 50,000 iterations of our new procedure. The first 20,000 iterations were discarded for burn-in.

Without variable selection there would be 120 free elements in the Cholesky matrix $\mathbf{C}$. With our new procedure this number may be reduced substantially. In Fig. 1 we see the posterior distribution of the number of free elements q_γ in $\mathbf{C}$. We have to estimate only 32 parameters, on average. This is also reflected by the posterior estimates of the indicators $\boldsymbol{\gamma}$ in Table 5. These posterior probabilities for the elements of $\mathbf{C}$ to be different from zero are very small for many elements.

In Table 6 we give posterior probabilities for the elements of the variance-covariance matrix $\mathbf{Q}$ to be unrestricted. For the first nine effects of the design we obtain unrestricted elements for the diagonal as well as for the off-diagonal elements, whereas the structure is more sparse with many elements restricted to zero especially for the last four effects.

Let us now look at selected random and fixed effects. The diagonal of Table 6 may be interpreted as posterior probabilities for the effects to be random effects. Here only the 14th effect has a low probability of $.21$ for being a random effect.

Fig. 1 Posterior distribution of the number of free elements q_γ in $\mathbf{C}$, based on 30,000 iterations after burn-in
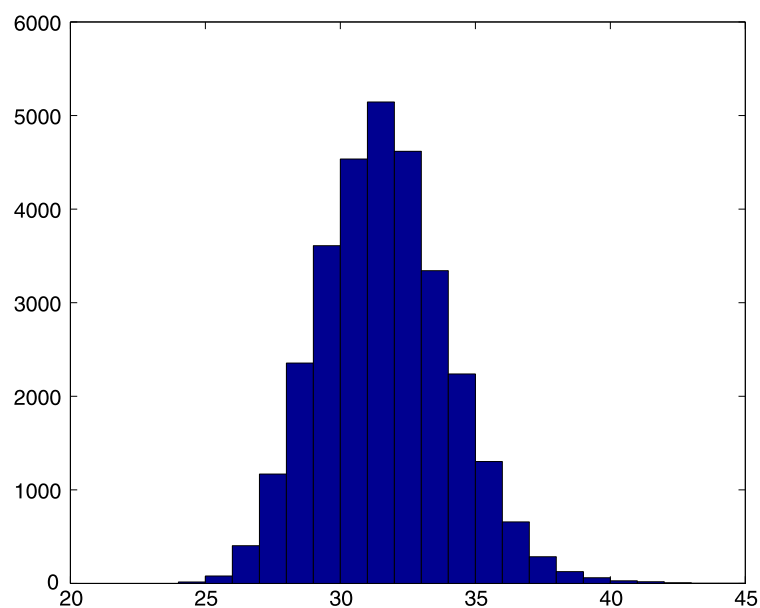
Table 5 Posterior probabilities for the elements of the Cholesky factor matrix **C** to be unrestricted (rounded)

1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
1	.97	1	0	0	0	0	0	0	0	0	0	0	0	0
1	.03	.21	1	0	0	0	0	0	0	0	0	0	0	0
1	.2	.04	1	1	0	0	0	0	0	0	0	0	0	0
1	.16	.53	1	.07	1	0	0	0	0	0	0	0	0	0
1	.75	.95	.12	.37	1	1	0	0	0	0	0	0	0	0
.97	.02	.08	.02	.03	.01	.1	1	0	0	0	0	0	0	0
.96	.04	.19	.01	.02	.04	.02	1	.03	0	0	0	0	0	0
.03	.16	.37	.96	.02	.02	.03	1	0	.25	0	0	0	0	0
.37	.02	.07	.01	1	.09	.02	.04	0	.04	.12	0	0	0	0
.01	.67	.05	.01	.06	.01	.02	.02	0	.01	0	.02	0	0	0
.05	.29	.15	.06	.03	.02	.01	.01	0	0	0	0	.03	0	0
.01	.06	.02	.04	.02	.01	.02	.01	0	0	0	0	0	.03	0
.01	.35	.01	.01	.22	.01	.01	.16	0	.01	.01	0	0	0	.02

Table 6 Posterior probabilities for the elements of the variance-covariance matrix **Q** to be unrestricted (rounded)

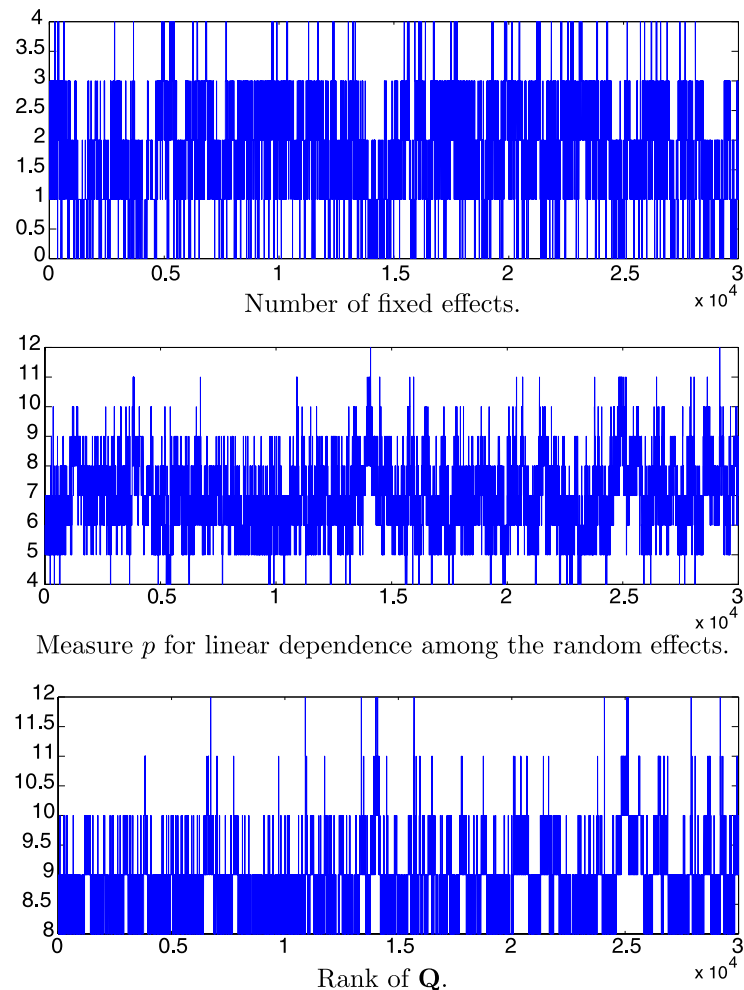|														
1	1	1	1	1	1	1	.97	.96	.03	.37	.01	.05	.01	.01
–	1	1	1	1	1	1	.97	.96	.18	.38	.67	.33	.06	.35
–	–	1	1	1	1	1	.97	.96	.48	.42	.66	.44	.08	.36
–	–	–	1	1	1	1	.97	.96	.96	.38	.05	.13	.05	.03
–	–	–	–	1	1	1	.97	.96	.96	1	.23	.17	.07	.23
–	–	–	–	–	1	1	.97	.96	.96	.49	.2	.3	.09	.08
–	–	–	–	–	–	1	.97	.96	.54	.67	.48	.28	.08	.4
–	–	–	–	–	–	–	1	1	1	.41	.04	.07	.02	.18
–	–	–	–	–	–	–	–	1	1	.41	.06	.12	.03	.18
–	–	–	–	–	–	–	–	–	1	.14	.2	.17	.07	.21
–	–	–	–	–	–	–	–	–	–	1	.1	.06	.02	.23
–	–	–	–	–	–	–	–	–	–	–	.72	.31	.06	.11
–	–	–	–	–	–	–	–	–	–	–	–	.52	.06	.03
–	–	–	–	–	–	–	–	–	–	–	–	–	.21	.01
–	–	–	–	–	–	–	–	–	–	–	–	–	–	.59

The first eleven effects are random throughout all MCMC iterations, whereas the remaining four effects are estimated as fixed at least in some iterations. In the upper row of Fig. 2 we give the sample path for the number of fixed effects. During 40 percent of the iterations there are two fixed effects. It is worth looking at Table 7 showing the five models which are selected with the highest probability. All these five models include the 14th effect as fixed effects. However, this fixed effect appears in different combinations with other fixed effects from the last four columns of the design.

Whereas most of the effects are random in this particular application, which is a common finding in marketing research, see, for instance Rossi et al. (2005), they seem to be highly correlated and linearly dependent. Apparently, strong rank reduction occurs in **Q**, see the posterior distrib-

Table 7 Model choice with respect to fixed effects

Effect no.				Prob
12	13	14	15	
Rand	Rand	Fixed	Fixed	.18
Rand	Rand	Fixed	Rand	.16
Fixed	Fixed	Fixed	Rand	.14
Rand	Fixed	Fixed	Rand	.12
Rand	Fixed	Fixed	Fixed	.11

ution of rg(**Q**) in the lower plot of Fig. 2. The middle-plot of Fig. 2 shows the posterior draws of p, the number of zero columns in **C**, not corresponding to zero lines, which has been derived in (13) from the indicator matrix γ. As ex-

Fig. 2 Details about rank reduction in **Q**

Number of fixed effects.

Measure p for linear dependence among the random effects.

Rank of **Q**.

plained in Sect. 2.1, p is a measure of linear dependence among the truly random effects, which may be explained by rg(**Q**) linearly independent factors which are random across the population. For the present application we find that the posterior mode of rg(**Q**) is given by eight, which means strong linear dependence among the random effects.

6 Concluding remarks

In this paper, we consider a non-centered parameterization of the standard random-effects model, which is based on the Cholesky decomposition of the variance-covariance matrix. This parameterization automatically delivers variance-covariance matrices without the need to introduce any constraints, as the Cholesky factors of variance-covariance matrices are unconstrained. This feature is rather desirable from a computational point of view.

Based on the non-centered parameterization, we are able to search for a parsimonious variance-covariance matrix by identifying the non-zero elements of the Cholesky factors using well-known Bayesian variable selection methods.

It turns out that the pattern of zeros in the Cholesky factors gives way to several important implications about the effects. First, we are able to learn from the data for each effect, whether it is random or not. Second, we are able to derive zero covariances among the random effects. This feature is of special importance for higher dimensional data, where determination of the finer structure in the covariance matrix often yields a substantial reduction of the number of parameters in the model. Finally, we may derive at each iteration the rank of the variance-covariance matrix. Again, this is of importance in many real applications, where the truly random effects are likely to depend linearly on a lower-dimensional common factor.

This method is potentially of great interest in many areas of applied statistics.

Appendix 1: Fractional prior for $\mathbf{C}^{\gamma}$

The basic idea of the fractional prior is to use part of the likelihood $p(\mathbf{y}|\boldsymbol{\gamma}, \mathbf{z}^s, \sigma_\varepsilon^2, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{C}^{\gamma})$, where $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$, to construct a proper prior for covariance selection under the improper prior $p(\mathbf{C}^{\gamma}|\sigma_\varepsilon^2) \propto \text{constant}$:

$$p(\mathbf{y}|\boldsymbol{\gamma}, \mathbf{z}^s, \sigma_\varepsilon^2, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{C}^{\gamma})^{1-b} p(\mathbf{y}|\boldsymbol{\gamma}, \mathbf{z}^s, \sigma_\varepsilon^2, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{C}^{\gamma})^b$$

$$\propto p(\mathbf{y}|\boldsymbol{\gamma}, \mathbf{z}^s, \sigma_\varepsilon^2, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{C}^{\gamma})^{1-b}$$

$$\times p(\mathbf{C}^{\gamma}|\mathbf{z}^s, \sigma_\varepsilon^2, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{y}^{TN \times b}), \tag{29}$$

where b lies between 0 and 1. $p(\mathbf{C}^{\gamma}|\mathbf{z}^s, \sigma_\varepsilon^2, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{y}^{TN \times b})$ is the fractional prior obtained from normalizing $p(\mathbf{y}|\boldsymbol{\gamma}, \mathbf{z}^s, \sigma_\varepsilon^2, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{C}^{\gamma})^b$:

$$p(\mathbf{C}^{\gamma}|\mathbf{z}^s, \sigma_\varepsilon^2, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{y}^{TN \times b})$$

$$= p(\mathbf{y}|\boldsymbol{\gamma}, \mathbf{z}^s, \sigma_\varepsilon^2, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{C}^{\gamma})^b / p(\mathbf{y}^{TN \times b}),$$

$$p(\mathbf{y}^{TN \times b}) = \int p(\mathbf{y}|\boldsymbol{\gamma}, \mathbf{z}^s, \sigma_\varepsilon^2, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{C}^{\gamma})^b d\mathbf{C}^{\gamma}.$$

Appendix 2: Details on sampling the indicators

Let γ_{lm}^{old} denote the current value of γ_{lm}. Generate u from a uniform distribution on $[0, 1]$. Then,

(i-1) If $\gamma_{lm}^{\text{old}} = 1$ and $u > p(\gamma_{lm} = 0|\boldsymbol{\gamma}_{\setminus lm})$, set $\gamma_{lm}^{\text{new}} = 1$.

(i-2) If $\gamma_{lm}^{\text{old}} = 0$ and $u > p(\gamma_{lm} = 1|\boldsymbol{\gamma}_{\setminus lm})$, set $\gamma_{lm}^{\text{new}} = 0$.

(i-3) If $\gamma_{lm}^{\text{old}} = 1$ and $u \leq p(\gamma_{lm} = 0|\boldsymbol{\gamma}_{\setminus lm})$, generate $v \sim U[0, 1]$ and set $\gamma_{lm}^{\text{new}} = 0$, iff $v \leq l(\gamma_{lm} = 0)/(l(\gamma_{lm} = 0) + l(\gamma_{lm} = 1))$; otherwise $\gamma_{lm}^{\text{new}} = 1$.

(i-4) If $\gamma_{lm}^{\text{old}} = 0$ and $u \leq p(\gamma_{lm} = 1|\boldsymbol{\gamma}_{\setminus lm})$, generate $v \sim U[0, 1]$ and set $\gamma_{lm}^{\text{new}} = 1$, iff $v \leq l(\gamma_{lm} = 1)/(l(\gamma_{lm} = 0) + l(\gamma_{lm} = 1))$; otherwise $\gamma_{lm}^{\text{new}} = 0$.

Here $p(\gamma_{lm} = i|\boldsymbol{\gamma}_{\setminus lm}), i = 0, 1$ is the conditional prior of γ_{lm}, see Sect. 3.1.1. $l(\gamma_{lm} = i)$ denotes the marginal likelihood $p(\mathbf{y}|\boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{z}^s, \sigma_\varepsilon^2)$ defined in (27) where γ_{lm} either takes the value $i = 0$ or $i = 1$. If the fraction $q_{\boldsymbol{\gamma}}/d_s$ of non-zero elements in $\mathbf{C}$ is small, step (i-1) will occur most often, whereas step (i-2) will occur most often, if this fraction is large. The other steps occur frequently if this fraction is about 0.5. Note that in cases (i-1) and (i-2) only the prior has to be calculated, which is computationally cheap compared to the likelihood appearing in the other two steps.

Appendix 3: Sampling the parameters of the random-effects model

Sampling $\boldsymbol{\alpha}, \boldsymbol{\beta}$

From models (1) and (2) we derive the marginal heteroscedastic model:

$$\mathbf{y}_i \sim \mathcal{N}_{T_i}(\mathbf{X}_i^f \boldsymbol{\alpha} + \mathbf{X}_i^r \boldsymbol{\beta}, \mathbf{X}_i^r \mathbf{Q}(\mathbf{X}_i^r)' + \sigma_\varepsilon^2 \mathbf{I}_{T_i}) \tag{30}$$

for $i = 1, \ldots, N$.

We sample the fixed effects $\boldsymbol{\alpha}$ and the mean parameter $\boldsymbol{\beta}$ together in one block from model (30) with the random effects being integrated out. This yields the following posterior distribution:

$$p(\boldsymbol{\alpha}, \boldsymbol{\beta}|\mathbf{C}^{\gamma}, \sigma_\varepsilon^2, \mathbf{y}) \sim \mathcal{N}_{d+d_f}(\mathbf{B}_N \mathbf{b}_N, \mathbf{B}_N),$$

where

$$\mathbf{b}_N = \sum_{i=1}^{N} [\mathbf{X}_i^f \mathbf{X}_i^r]'(\mathbf{X}_i^r \mathbf{Q}(\mathbf{X}_i^r)' + \sigma_\varepsilon^2 \mathbf{I}_{T_i})^{-1} \mathbf{y}_i + \mathbf{B}_0^{-1} \mathbf{b}_0,$$

$$\mathbf{B}_N^{-1} = \sum_{i=1}^{N} [\mathbf{X}_i^f \mathbf{X}_i^r]'(\mathbf{X}_i^r \mathbf{Q}(\mathbf{X}_i^r)' + \sigma_\varepsilon^2 \mathbf{I}_{T_i})^{-1} [\mathbf{X}_i^f \mathbf{X}_i^r] + \mathbf{B}_0^{-1}.$$

Sampling $\mathbf{z}_i^s$

To generate from $\mathbf{z}^s|\boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{C}^{\gamma}, \sigma_\varepsilon^2, \mathbf{y}$ we first observe that the various components of $\mathbf{z}^s = (\mathbf{z}_1^s, \ldots, \mathbf{z}_N^s)$ are conditionally independent. The conditional distribution of $\mathbf{z}_i^s|\boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{C}^{\gamma}, \sigma_\varepsilon^2, \mathbf{y}$ is a normal distribution, obtained by combining the prior $\mathbf{z}_i^s \sim \mathcal{N}_d(\mathbf{0}, \mathbf{I}_d)$ with the likelihood $p(\mathbf{y}_i|\mathbf{z}_i^s, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{C}^{\gamma}, \sigma_\varepsilon^2)$ through Bayes' theorem:

$$\mathbf{z}_i^s|\boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{C}^{\gamma}, \sigma_\varepsilon^2, \mathbf{y} \sim \mathcal{N}_d(\mathbf{P}_i \mathbf{p}_i, \mathbf{P}_i),$$

where

$$\mathbf{p}_i = \sigma_\varepsilon^{-2}(\mathbf{X}_i^r \mathbf{C})'(\mathbf{y}_i - \mathbf{X}_i^f \boldsymbol{\alpha} - \mathbf{X}_i^r \boldsymbol{\beta}),$$

$$\mathbf{P}_i^{-1} = \sigma_\varepsilon^{-2}(\mathbf{X}_i^r \mathbf{C})'(\mathbf{X}_i^r \mathbf{C}) + \mathbf{I}_d.$$

Sampling σ_ε^2

We sample $\sigma_\varepsilon^2|\boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{z}^s, \mathbf{y}$ from the inverted Gamma posterior density:

$$\sigma_\varepsilon^2|\boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{z}^s, \mathbf{y} \sim \mathcal{G}^{-1}(s_N/2, S_N/2),$$

with $s_N = \sum_{i=1}^{N} T_i + s_0$ and

$$S_N = S_0 + S^{\gamma} + (\mathbf{a}_N - \mathbf{a}_0)'\mathbf{A}_0^{-1}(\mathbf{a}_N - \mathbf{a}_0),$$

with S^{γ} being the sum of squared errors defined in (28).

Appendix 4: Design of the simulation study

We simulate data for $i = 1, \ldots, 200$ subjects from the random-effects models (1) and (2) with design matrix $\mathbf{X}_i^r$ equal to

$$\mathbf{X}_i^r = \begin{pmatrix} 1 & u_{i1} & 0 & n_{i1} \\ 1 & u_{i2} & 0 & n_{i2} \\ 1 & u_{i3} & 0 & n_{i3} \\ 1 & 0 & u_{i4} & n_{i4} \\ 1 & 0 & u_{i5} & n_{i5} \\ 1 & 0 & u_{i6} & n_{i6} \end{pmatrix},$$

where u_{ik} come from a uniform distribution on the interval $[1, 2]$ and n_{ik} are standard normally distributed random numbers, for $k = 1, \ldots, 6$. We include no fixed effects ($\boldsymbol{\alpha} = \mathbf{0}$), and the random effects have mean parameter $\boldsymbol{\beta} = (1 - 2\,1.5\,.8)'$ and variance-covariance matrices defined in (16). The model error variance σ_ε^2 equals 1.

References

Albert, J.H., Chib, S.: Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts. J. Bus. Econ. Stat. **11**, 1–15 (1993)

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

Frühwirth-Schnatter, S., Tüchler, R., Otter, T.: Bayesian analysis of the heterogeneity model. J. Bus. Econ. Stat. **22**, 2–15 (2004)

Gelfand, A.E., Sahu, S.K., Carlin, B.P.: Efficient parametrisations for normal linear mixed models. Biometrika **82**, 479–488 (1995)

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

Lindstrom, M., Bates, D.: Newton-Raphson and the EM-Algorithm for linear mixed-effects models for repeated measures data. J. Am. Stat. Assoc. **83**, 1014–1022 (1988)

Meng, X.-L., van Dyk, D.: Fast EM-type implementations for mixed effects models. J. Roy. Stat. Soc. B **60**, 559–578 (1998)

O'Hagan, A.: Fractional Bayes factors for model comparison. J. Roy. Stat. Soc. B **57**, 99–118 (1995)

Otter, T., Tüchler, R., Frühwirth-Schnatter, S.: Capturing consumer heterogeneity in metric conjoint analysis using Bayesian mixture models. Int. J. Mark. Res. **21**, 285–297 (2004)

Papaspiliopoulos, O., Roberts, G., Skold, M.: Non-centered parameterizations for hierarchical models and data augmentation. In: Bernardo, J.M., Bayarri, M.J., Berger, J.O., Dawid, A.P., Heckerman, D., Smith, A.F.M., West, M. (eds.) Bayesian Statistics 7, pp. 307–326. Oxford University Press, Oxford (2003)

Pinheiro, J., Bates, D.: Unconstrained parameterizations for variance-covariance matrices. Stat. Comput. **6**, 289–296 (1996)

Rossi, P.E., Allenby, G.M., McCulloch, R.: Bayesian Statistics and Marketing. Wiley Series in Probability and Statistics. Wiley, New York (2005)

Smith, M., Kohn, R.: Parsimonious covariance matrix estimation for longitudinal data. J. Am. Stat. Assoc. **97**, 1141–1153 (2002)

Tüchler, R., Frühwirth-Schnatter, S.: Bayesian parsimonious estimation of observed and unobserved heterogeneity. In: Verbeeke, G., Molenberghs, G., Aerts, M., Fieuws, S. (eds.) Proceedings of the 18th International Workshop on Statistical Modelling, pp. 427–431. Leuven, Belgium (2003)

Van Dyk, D., Meng, X.-L.: The art of data augmentation. J. Comput. Graph. Stat. **10**, 1–50 (2001)