Bayesian parsimonious covariance estimation for hierarchical linear mixed models

Frühwirth-Schnatter, Sylvia; Tüchler, Regina

Published: 01/01/2004

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

Bayesian Parsimonious Covariance Estimation for Hierarchical Linear Mixed Models

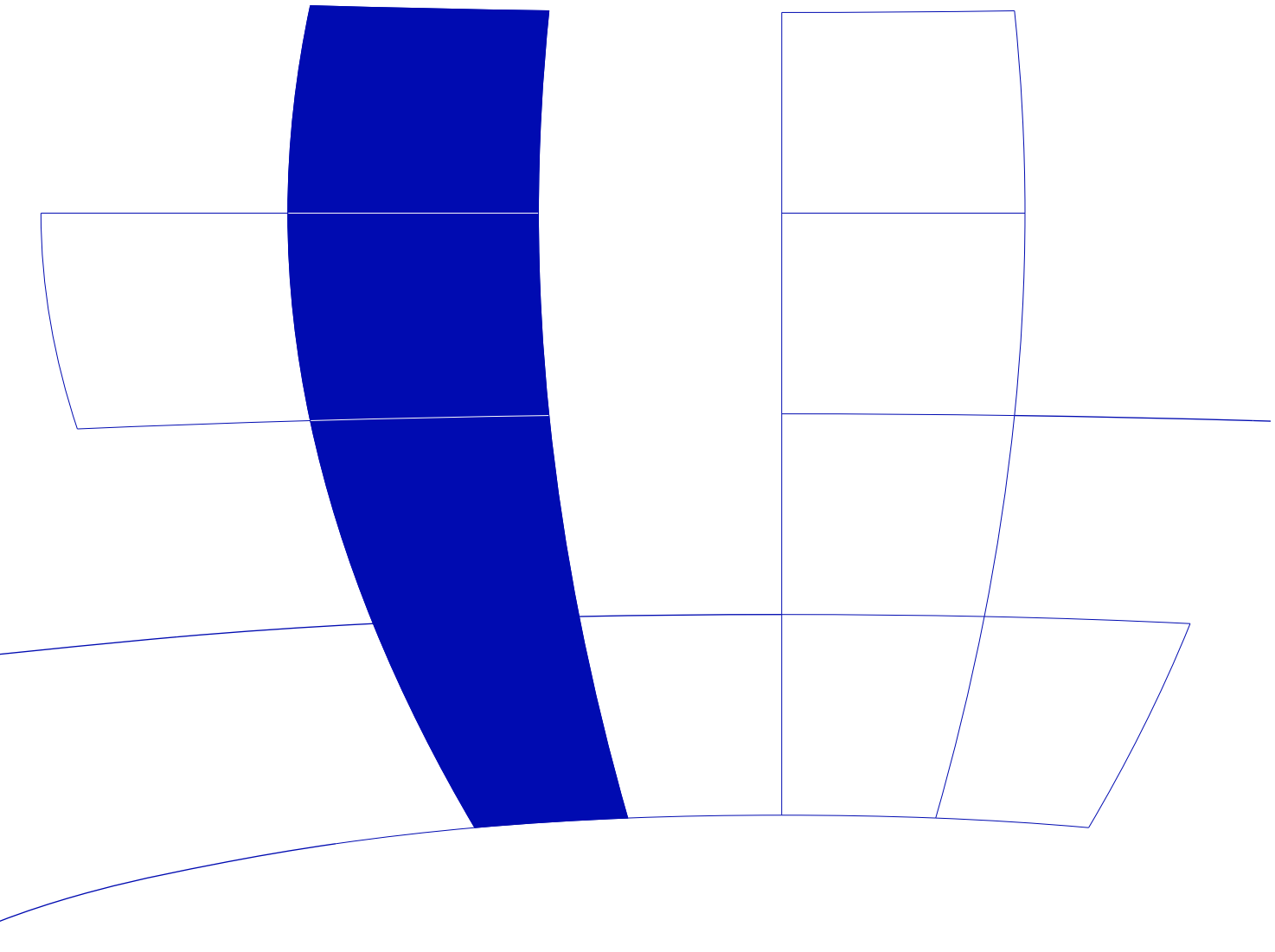
Sylvia Frühwirth-Schnatter, Regina Tüchler

Department of Statistics and Mathematics
Wirtschaftsuniversität Wien

Research Report Series

Report 11
November 2004

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



Bayesian Parsimonious Covariance Estimation for Hierarchical Linear Mixed Models

Sylvia Frühwirth-Schnatter

Johannes Kepler Universität Linz, Department of Applied Statistics and Econometrics (IFAS), Altenbergerstr. 69, 4040 Linz, Austria

Regina Tüchler

University of Economics and Business Administration Vienna, Department of Statistics and Mathematics, Augasse 2-6, 1090 Vienna, Austria

Summary.
We considered 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 to choose a simple, conditionally conjugate normal prior on the Cholesky factor. Based on the non-centered parameterization, we search for a parsimonious variance-covariance matrix by identifying the non-zero elements of the Cholesky factors using Bayesian variable selection methods. 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 or not. An application in marketing shows a substantial reduction of the number of free elements of the variance-covariance matrix.

Keywords: covariance selection, random effects models, Markov chain Monte Carlo, fractional prior, variable selection,
Corresponding author: Sylvia Frühwirth-Schnatter, email: `sylvia.fruehwirth-schnatter@jku.at`

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 is to select a suitable parameterization of the variance-covariance matrix, which typically has a large number of parameters, that are related by the very complex constraint that the resulting matrix needs to be positive definite. Within the Bayesian approach we pursue in the present article, a further important issue is prior elicitation for the variance-covariance matrix of the random effects. Finally, for practical applications of the random-effects model, model selection deserves consideration, as one would like to learn from the data, if actually all effects are random.

A particularly useful parameterization of variance-covariance matrices is based on the Cholesky decomposition of either Q or Q^{-1}. As pointed out by Pinheiro and Bates (1996), this parameterization is of considerable numerical convenience as is involves unconstrained parameters, only. For directly observed data, arising from a multivariate normal distribution with unknown variance-covariance matrix Q, it is usual to consider the Cholesky decomposition $Q^{-1} = CSC'$, with S being a diagonal matrix and C being a lower triangular matrix with ones on the diagonal. This parameterization is preferred, because zeros in the Cholesky factors of Q^{-1} may be interpreted as conditional independence between the corresponding variables, see Dempster (1972), Pourahmadi (1999), Smith and Kohn

(2002), and Wong et al. (2003), among others. Based on this parameterization, Smith and Kohn (2002) made two important contributions to variance-covariance estimation for multivariate, normally distributed data, first by realizing that the natural conjugate conditional prior for the Cholesky factors of Q^{-1} is a normal distribution, and second by introducing a Bayesian variable search method for parsimonious variance-covariance matrices. In this article we going to generalize the work of Smith and Kohn (2002), in order to deal with data arising from a hierarchical model, rather than the multivariate normal distribution.

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 in this article a parameterization of the variance-covariance matrices based on the Cholesky decomposition $Q = CC^{'}$ with a lower triangular matrix C. This Cholesky decomposition of Q leads in a natural way to a non-centered parameterization of a random-effects model, where all free elements of C appear as unknown coefficients in a regression type model. This parameterization is slightly different from the non-centered parameterization considered by Meng and van Dyk (1998), and Chen and Dunson (2003). With our parameterization, we are able to introduce a new prior for the variance-covariance matrix of the random effects, by choosing a conditionally conjugate normal prior for all elements of the Cholesky factors C in the decomposition $Q = CC^{'}$.

The choice of an appropriate prior on Q is rather challenging for hierarchical models, resulting from the need to estimate the variance-covariance matrix of latent, rather than directly observed variables. The most commonly used approach is to work with a conditionally conjugate inverted Wishart prior on Q, as this allows a straightforward implementation of a Gibbs sampling scheme for Bayesian estimation, and automatically leads to positive definite variance-covariance matrices. A problem with the conditional conjugate inverted Wishart prior, however, is that the prior parameters may be extremely influential on posterior inference, especially with increasing dimension of Q, see in particular Natarajan and McCulloch (1998) and Natarajan and Kass (2000). Another problem is that certain inverted Wishart priors, for instance the improper prior where both prior parameters are equal to zero, lead to improper posterior densities, see Hobert and Casella (1996), Natarajan and McCulloch (1998) and Sun et al. (2001).

Numerous alternatives to the inverted Wishart prior have been suggested in the literature, like selecting a uniform prior on the shrinkage factor appearing in the filtered estimate of each random effect, see Daniels (1999), Natarajan and Kass (2000) and Everson and Morris (2000). Many interesting non-conjugate priors have been constructed by considering different parameterizations of a variance-covariance matrix. Some approaches focus on parameterization in terms of eigenvalues and eigenvectors, and select non-conjugate priors involving these quantities, see in particular Leonard and Hsu (1992), Yang and Berger (1994), Chiu et al. (1996), and Daniels and Kass (1999). An alternative line of research focuses on the statistically motivated decomposition $Q = SRS$ of a variance-covariance matrix Q, with S being a diagonal matrix of standard deviations and R being the correlation matrix. Whereas it is possible to assume a conjugate inverted gamma prior for the standard deviations, non-conjugate priors have to be chosen for the correlation coefficients, see Daniels and Kass (1999), Barnard et al. (2000), Daniels and Kass (2001), and Liechty et al. (2004) for various suggestions.

Although lack of conjugacy nowadays is no problem in a Bayesian analysis, posterior simulations from hierarchical linear models with non-conjugate priors on Q may cause computational difficulties resulting from the need to produce positive definite matrices, see Liechty

et al. (2004) for a recent discussion. In contrast to most non-conjugate priors, the prior suggested in this paper automatically leads to non-negative definite variance-covariance matrices, and allows for Bayesian estimation using straightforward Gibbs sampling. Whereas this prior is as convenient as the inverted Wishart prior, we demonstrate by means of a simulation study that it is less influential on posterior inference than the inverted Wishart prior.

As a second contribution of the paper, we aim for parsimonious variance-covariance selection, rather than estimating a full rank variance-covariance matrix of the random effects, as is usually done. Little work, has been done for parsimonious variance-covariance selection for hierarchical models, exceptions being Albert and Chib (1993) and Chen and Dunson (2003). As in these papers, we propose a data-driven method to achieve parsimony in a variance-covariance matrix, by identifying zeros in the Cholesky decomposition of Q. Whereas Albert and Chib (1993) and Chen and Dunson (2003) perform variable selection only on the free elements of the diagonal matrix S in the Cholesky decomposition $Q = SLL'S$, where L is a lower triangular matrix with ones in the diagonal, we consider variable selection on all free elements in the matrix C appearing in the Cholesky decomposition $Q = CC'$. To some extend, we also follow the seminal work of Smith and Kohn (2002) and Wong et al. (2003), who identify zeros in the Cholesky factors C of the decomposition $Q^{-1} = CSC'$. However, we operate on the Cholesky factors of Q rather than on Q^{-1}. It will be shown, that our approach allows to shrink random effects toward fixed ones, a feature that would not result with a direct application of Smith and Kohn (2002) and Wong et al. (2003) to the matrix Q^{-1} appearing in a hierarchical model. We will show that a straightforward MCMC scheme for joint variable selection and parameter estimation is available, that involves sampling from standard densities, only.

The rest of the article is organized as follows. In Section 2 we define a parsimonious representation of the random-effects model. In Section 3 we specify the MCMC sampling steps. In Section 4 we describe the improvements of the new algorithm in comparison to existing algorithms for simulated data and we apply the algorithm to real data coming from marketing in Section 5.

2. Model Specification and Prior Distributions

2.1. The Non-centered Parameterization based on the Cholesky Decomposition

For each subject i, $i = 1, \ldots, N$ we write the random-effects model in the following way:

$$y_i = Z_i^1 \alpha + Z_i^2 \beta^G + Z_i^2 C \tilde{z}_i + \varepsilon_i, \quad \varepsilon_i \sim \text{Normal}_{T_i}\left(0, \sigma_\varepsilon^2 I\right), \tag{1}$$

$$\tilde{z}_i \sim \text{Normal}_d\left(0, I\right). \tag{2}$$

The vector y_i contains T_i observations and Z_i^1 is a design matrix of dimension $T_i \times d_f$ for the d_f-dimensional vector α containing the fixed . Z_i^2 is the $T_i \times d$-dimensional design matrix for the d-dimensional vector of random effects. C is a lower triangular square matrix with d rows. The quantities α, β^G, C, and σ_ε^2 are unknown parameters that need to be estimated from the data. By rewriting (1) as

$$y_i = Z_i^1 \alpha + Z_i^2 (\beta^G + C\tilde{z}_i) + \varepsilon_i,$$

it is easy to verify, that model (1) and (2) is equivalent to the well-known random-effects model:

$$y_i = Z_i^1 \alpha + Z_i^2 \beta_i + \varepsilon_i, \tag{3}$$

$$\beta_i = \beta^G + u_i, \quad u_i \sim \text{Normal}_d\left(0, Q\right), \tag{4}$$

where the random effects are normally distributed with mean parameter β^G and variance-covariance matrix $Q = CC'$. Evidently, parameterization (1) and (2) is based on the following Cholesky decomposition of the variance-covariance matrix Q:

$$Q = CC'.$$

Parameterization (3) and (4) is known as the *centered* parameterization, whereas in (1) and (2) the random-effects model is formulated in the *non-centered* parameterization, introduced by Meng and van Dyk (1998), and studied in much detail in van Dyk and Meng (2001). The non-centered parameterization defined in (1) and (2), however, is slightly different from the parameterization appearing in the work of Meng and van Dyk (1998), which reads

$$y_i = Z_i^1\alpha + Z_i^2\beta^G + Z_i^2 LS\tilde{z}_i + \varepsilon_i, \quad \varepsilon_i \sim \text{Normal}_{T_i}\left(0, \sigma_\varepsilon^2 I\right),$$

and is based on the Cholesky decomposition $Q = LSS'L'$, where L is a lower triangular matrix with ones in the diagonal and S is a diagonal matrix.

2.2. Parsimonious Variance-Covariance Matrices for Hierarchical Linear Mixed Models

Statistical inference for the variance-covariance matrix of a random-effects model is usually based on the estimation of a full rank variance-covariance matrix of the random effects. In contrast to that, we follow the principle of parsimony with respect to Q. Parsimony is achieved by restricting certain elements appearing in the matrix of the Cholesky factors C of Q to be 0. We let the data tell us which elements this should be.

2.2.1. Parsimonious Variance-Covariance Matrices through Variable Selection

Following the seminal work of Smith and Kohn (2002), we treat the problem of finding those elements of C that are non-zero as a variable selection problem and pursue a Bayesian approach. We introduce for each element C_{lm}, $m = 1, \ldots, d, l = m, \ldots, d$, an indicator γ_{lm} which takes the value 1, if $C_{lm} \neq 0$, and 0 otherwise:

$$\begin{array}{ll} C_{lm} = 0, & \text{iff } \gamma_{lm} = 0, \\ C_{lm} \neq 0, & \text{iff } \gamma_{lm} = 1. \end{array} \tag{5}$$

Note that C_{lm} is 0 by definition for all $1 \leq l < m$. Thus we actually need only a total of $d(d+1)/2$ indicators to represent all possible variance-covariance matrices. We will use γ to denote the collection of all $d(d+1)/2$ indicators γ_{lm}. If all indicators are equal to 1, all $d(d+1)/2$ elements of C are unconstrained and we are actually dealing with an arbitrary, positive definite variance-covariance matrix Q.

Our approach of choosing parsimonious variance-covariance matrices for a hierarchical model reduces the problem of variance-covariance selection to the more common problem of Bayesian variable selection in multiple regression models, as reviewed in George and McCulloch (1997). This relation becomes more evident by rewriting the observation equation (1) as follows. Depending on the indicators vector γ, various elements of C will be restricted to 0, whereas the remaining, non-zero elements of C are treated as an unknown parameter, denoted by C^γ. The parameter vector C^γ is constructed from C by stacking the non-zero

elements of C column by column. For known random effects $\tilde{z}_i$, observation equation (1) may be regarded as following regression model in C^γ:

$$y_i = Z_i^1\alpha + Z_i^2\beta^G + W_i^\gamma C^\gamma + \varepsilon_i, \tag{6}$$

where the predictor matrix W_i^γ depends on the design matrix Z_i^2, and on the latent random effects $\tilde{z}_i$. We will provide details of how W_i^γ is constructed at the end of this subsection. Like in standard Bayesian variable selection, elements in the predictor matrix W_i^γ will be included or deleted, depending on γ. As a notable difference, however, variable selection in (6) is with respect to predictors that are latent, rather than directly observed.

For a fixed value of γ, W_i^γ is constructed from the design matrix Z_i^2 and the latent random effects $\tilde{z}_i$ in the following way. For each column $C_{\cdot m}$ of C, the predictor matrix W_i^γ in (6) contains a sub-matrix $W_i^{\gamma_{\cdot m}}$, which corresponds to all non-zero elements of the column $C_{\cdot m}$:

$$W_i^\gamma = \left(\begin{array}{ccc} W_i^{\gamma_{\cdot 1}}\tilde{z}_{i1} & \cdots & W_i^{\gamma_{\cdot d}}\tilde{z}_{id} \end{array} \right).$$

$\tilde{z}_{im}, m = 1, \ldots, d$ refers to the m-th element of the latent variable $\tilde{z}_i$. For each column $C_{\cdot m}$ of C, the sub-matrix $W_i^{\gamma_{\cdot m}}$ is constructed as follows. If all elements in column $C_{\cdot m}$ were unrestricted, then $W_i^{\gamma_{\cdot m}}$ would be equal to the matrix Z_i^2. To account for the zero elements in column $C_{\cdot m}$, the following columns of Z_i^2 have to be deleted in order to obtain $W_i^{\gamma_{\cdot m}}$: the first $m-1$ columns (remember that C is lower triangular by definition) as well as those columns l, where C_{lm} is restricted to 0 ($\gamma_{lm} = 0$).

2.2.2. 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 Q, our approach is substantially different from Smith and Kohn (2002), who use the Cholesky decomposition $Q^{-1} = LSL'$ of the inverse of Q where L is a lower triangular matrix with ones in the diagonals and S 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 L, whereas the elements of S are assumed to be positive. If all indicators are equal to 1, all $d(d-1)/2$ elements of L are unconstrained, leading to the estimation of an arbitrary positive definite variance-covariance matrix Q as in our approach. If all indicator are equal to 0, however, Q is shrunk toward the diagonal matrix S^{-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 to reduce any of the random effects to a fixed one.

Our own approach is more flexible in this respect. As we work with the Cholesky decomposition of Q rather than Q^{-1}, it is possible to reduce some or all random effects to fixed ones, by choosing the indicators γ_{lm} appropriately. From

$$\beta_{il} = \beta_l^G + \sum_{m=1}^{l} C_{lm}\gamma_{lm}\tilde{z}_{im}, \tag{7}$$

where $\tilde{z}_{il} \sim \text{Normal}\,(0,1)$ are $l = 1, \ldots, d$ independent standard normals, we find that the l-th random effect β_{il} is shrunk toward a fixed effect with coefficient β_l^G, if all elements in

l-th line of C are equal to 0. In this case, the rank of Q is reduced by one. If all indicators were 0, then Q is equal to a zero matrix, and all random effects are shrunk toward an effect with fixed coefficient.

Our approach is related to Chen and Dunson (2003), who apply a similar but more specific approach to the Cholesky decomposition $Q = SLL'S$, where L is a lower triangular matrix with ones in the diagonal and S is a diagonal matrix. In order to reduce random effects to fixed ones, they allow the diagonal elements of S to have a positive probability of being zero, whereas no variable selection is performed for the elements of 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 Q especially in higher dimensional problems.

2.3. Prior Distributions

2.3.1. Defining the prior for the indicators γ

For Bayesian estimation one has to select the prior of the indicator variables γ_{lm}. Conditional on a known value $\tau \in [0,1]$, the indicator variables γ_{lm} are assumed to be apriori independent with

$$\Pr\{\gamma_{lm} = 1|\tau\} = \tau.$$

This implies that for fixed τ the number of non-zero elements in C follows the binomial distribution Bino (d_s, τ), where $d_s = d(d+1)/2$ is the total number of free parameters in C. For variance-covariance matrices Q of moderate size this density is fairly non-informative on the number of non-zeros 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 apriori probable number of non-zero elements will 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:

$$p(\gamma) = \int p(\gamma|\tau)p(\tau)d\tau = \text{Beta}(q_\gamma, d_s - q_\gamma + 1). \tag{8}$$

Here, Beta$(\cdot,\cdot)$ is the beta function, q_γ is the number of non-zero elements in C. Note that the marginal prior (8) implies apriori dependence between the elements of γ.

2.3.2. Selecting the Prior of the Variance-Covariance Matrix of the Random Effects

A convenient starting point for prior selection of the variance-covariance matrix of the random effects under the Cholesky decomposition is model (6) which is a linear normal regression model in C^γ. Conditional on knowing the indicator variable γ and the standardized random effects $\tilde{z}^N$, we choose as prior for the non-zero elements C^γ of the Cholesky decomposition C of Q, the conditionally conjugate normal prior

$$p(C^\gamma|\sigma_\varepsilon^2) \sim \text{Normal}\left(a_0, \sigma_\varepsilon^2 A_0\right). \tag{9}$$

This conditionally conjugate normal prior leads to a normal posterior distribution

$$p(C^\gamma|\tilde{z}^N, \sigma_\varepsilon^2, \alpha, \beta^G, y) \sim \text{Normal}\left(a_N, \sigma_\varepsilon^2 A_N\right), \tag{10}$$

where a_N and A_N are given by:

$$a_N = A_N \left(\sum_{i=1}^{N} (W_i^\gamma)'(y_i - Z_i^1\alpha - Z_i^2\beta^G) + A_0^{-1}a_0 \right),$$
$$A_N^{-1} = \sum_{i=1}^{N} (W_i^\gamma)'W_i^\gamma + A_0^{-1}.$$

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 Q, the normal distribution is a natural conjugate conditional prior for the free elements of the lower triangular matrix L in the Cholesky decomposition $Q^{-1} = LSL'$. In the context of random-effects models, a conditionally conjugate normal prior for the Cholesky factors of the variance-covariance matrix 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 of Chen and Dunson (2003), who considered the Cholesky decomposition $Q = SLL'S$, in various aspects. Chen and Dunson (2003) use a conditionally normal prior on the free elements of the lower triangular matrix L, and consider a zero inflated half normal distribution for the free elements in the diagonal matrix S, consisting of a mass point at zero (with probability $1 - \tau$) and a normal density with mean a_0 and variance A_0 truncated below zero. Their prior may be formulated in terms of d variable indicators $\gamma_l, l = 1, \ldots, d$, for the d free elements of S, 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 Q, whereas our prior is more flexible. Second, we include the diagonal into the Cholesky decomposition, which allows to define a normal prior on all non-zero elements of C, not only on the lower triangular matrix L.

2.3.3. Remaining Priors

It remains to choose a prior for the mean parameters (α, β^G) and the observation error variance σ_ε^2. For the mean parameters (α, β^G) we assume a joint Normal (b_0, B_0) prior distribution, whereas the observation error variance σ_ε^2 is a priori InvGamma $(s_0/2, S_0/2)$.

2.3.4. Prior Selection Without Variable Selection

The conditionally conjugate normal prior on the free elements of the Cholesky factors, introduced in Subsection 2.3.2 in the context of variance-covariance selection, is also of interest for the standard normal random-effects model, without doing variable selection on the elements of C. The conditionally conjugate normal prior on the $d(d+1)/2$ free elements of C, together with the non-centered parameterization (1) and (2), provides a convenient alternative to the inverted Wishart prior applied together with the centered parameterization (3) and (4). We will demonstrate in Section 3, that a straightforward Gibbs sampling scheme is available for this new prior, whereas the simulation study in Section 4 demonstrates, that this prior is less influential on posterior inference

3. MCMC Estimation

We introduce an MCMC scheme which simultaneously carries out model selection and estimation of all unknown parameters. MCMC estimation of random effects model without variable selection was considered by numerous authors. The parameterization of the random-effects model turns out to be of enormous importance for the convergence behavior of the MCMC chains. The influence of the parameterization of the mean on the convergence behavior of the straightforward Gibbs sampler was analyzed by Gelfand et al. (1995) and Papaspiliopoulos et al. (2003) for normal hierarchical linear models. Non-centering both of the mean and the variance-covariance matrix is investigated in Meng and van Dyk (1998) and van Dyk and Meng (2001) for random-effects models and in Frühwirth-Schnatter (2004) for more general state space models. In these articles, a criterion depending on the amount of heterogeneity captured by the random effects in comparison to the model error was established to choose the optimal parameterization when applying a full conditional Gibbs sampler. An algorithm that is insensitive towards the parameterization of the mean was introduced for mixtures of random effects models in Frühwirth-Schnatter et al. (2004). In the present article, we make use of the findings of Frühwirth-Schnatter et al. (2004), and samples the fixed effects and the mean parameters efficiently without conditioning on the random effects.

The non-centered parameterization based on the Cholesky decomposition, together with the priors defined in Section 2, give way to the following convenient Gibbs sampling scheme involving standard densities, only:

(i) Sample $\gamma_{lm}|\gamma_{\backslash lm}, \alpha, \beta^G, \sigma_\varepsilon^2, y$, where $\gamma_{\backslash lm}$ denotes the indicator vector γ without the element γ_{lm}, from a discrete density with two realizations.

(ii) Sample $C^\gamma|\alpha, \beta^G, \tilde{z}^N, \sigma_\varepsilon^2, y$ from a normal distribution.

(iii) Sample $\alpha, \beta^G|C^\gamma, \sigma_\varepsilon^2, y$ from a normal distribution.

(iv) Sample $\tilde{z}^N|\alpha, \beta^G, C^\gamma, \sigma_\varepsilon^2, y$ from a normal distribution.

(v) Sample $\sigma_\varepsilon^2|\alpha, \beta^G, \tilde{z}^N, y$ from an inverted Gamma distribution.

Subsequently, we will discuss each step in more detail.

3.1. 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 C of the variance-covariance matrix 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 $p(\gamma_{lm}|\gamma_{\backslash lm}, \alpha, \beta^G, \sigma_\varepsilon^2, y)$, where $\gamma_{\backslash lm}$ denotes the indicator vector γ without the element γ_{lm}. Then conditional on knowing γ, all non-zero elements C^γ of C are sampled from the appropriate distribution.

Note that the density $p(\gamma_{lm}|\gamma_{\backslash lm}, \alpha, \beta^G, \sigma_\varepsilon^2, y)$ is marginalized over the Cholesky factors in order to avoid the computational problems discussed e.g. in George and McCulloch (1997). To implement this step, the marginal likelihood $p(y|\gamma, \alpha, \beta^G, \tilde{z}^N, \sigma_\varepsilon^2)$ where C^γ is integrated out is required. As will be shown below, this quantity is readily available in closed form under an informative prior on C^γ, whereas further considerations are necessary under non-informative priors.

3.1.1. The marginal likelihood function under informative priors
The marginal likelihood $p(y|\gamma,\alpha,\beta^G,\tilde{z}^N,\sigma_\varepsilon^2)$ where C^γ is integrated out, is given by:

$$p(y|\gamma,\alpha,\beta^G,\tilde{z}^N,\sigma_\varepsilon^2)=\int p(y|\gamma,\alpha,\beta^G,\tilde{z}^N,\sigma_\varepsilon^2,C^\gamma)p(C^\gamma|\sigma_\varepsilon^2)dC^\gamma, \tag{11}$$

where $p(y|\gamma,\alpha,\beta^G,\tilde{z}^N,\sigma_\varepsilon^2,C^\gamma)$ is obtained as the following quadratic form:

$$\begin{aligned}&p(y|\gamma,\alpha,\beta^G,\tilde{z}^N,\sigma_\varepsilon^2,C^\gamma)= \\ &\left(\frac{1}{2\pi\sigma_\varepsilon^2}\right)^{NT/2}\exp\left(-\frac{1}{2\sigma_\varepsilon^2}\sum_{i=1}^{N}||y_i-W_i^\gamma C^\gamma-Z_i^1\alpha-Z_i^2\beta^G||_2\right).\end{aligned} \tag{12}$$

$||x||_2=\sum_j x_j^2$ is the L_2-norm of a vector $x=(x_1\cdots x_p)'$.

For a proper normal prior $p(C^\gamma|\sigma_\varepsilon^2)$, where in (9) $|A_0|>0$, the marginal likelihood (11) is a well-defined quantity:

$$\begin{aligned}&p(y|\gamma,\alpha,\beta^G,\tilde{z}^N,\sigma_\varepsilon^2)= \\ &\frac{|A_N|^{-1/2}}{|A_0|^{-1/2}}\left(\frac{1}{2\pi\sigma_\varepsilon^2}\right)^{NT/2}\exp\left(-\frac{1}{2\sigma_\varepsilon^2}(S^\gamma+(a_N-a_0)'A_0^{-1}(a_N-a_0))\right),\end{aligned} \tag{13}$$

where

$$S^\gamma=\sum_{i=1}^{N}||y_i-W_i^\gamma a_N-Z_i^1\alpha-Z_i^2\beta^G||_2. \tag{14}$$

a_N and A_N are the moments of the posterior $p(C^\gamma|\tilde{z}^N,\sigma_\varepsilon^2,\alpha,\beta^G,y)$ given in (10).

3.1.2. The marginal likelihood function under improper priors
Like in variable selection problems for the standard regression model, the specific choice of the prior moments a_0 and A_0 is likely to be rather influential on the posterior of the model indicator γ, see O'Hagan (1995) and George and McCulloch (1997). Furthermore, the marginal likelihood (13) is not well-defined under the improper prior $p(C^\gamma|\sigma_\varepsilon^2)\propto c$, which corresponds to choosing $a_0=0, A_0^{-1}=0$ in (9).

To obtain a meaningful marginal likelihood also under the improper prior $p(C^\gamma|\sigma_\varepsilon^2)\propto c$, 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 L in the Cholesky decomposition $Q^{-1}=LSL'$.

The basic idea of the fractional prior is to use part of the likelihood to construct a proper prior for model selection under the improper prior $p(C^\gamma|\sigma_\varepsilon^2)\propto c$:

$$\begin{aligned}&p(y|\gamma,\tilde{z}^N,\sigma_\varepsilon^2,\alpha,\beta^G,C^\gamma)^{1-b}p(y|\gamma,\tilde{z}^N,\sigma_\varepsilon^2,\alpha,\beta^G,C^\gamma)^b \\ &\propto p(y|\gamma,\tilde{z}^N,\sigma_\varepsilon^2,\alpha,\beta^G,C^\gamma)^{1-b}p(C^\gamma|\tilde{z}^N,\sigma_\varepsilon^2,\alpha,\beta^G,y^{TN\times b}),\end{aligned} \tag{15}$$

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

$$\begin{aligned}p(C^\gamma|\tilde{z}^N,\sigma_\varepsilon^2,\alpha,\beta^G,y^{TN\times b})&=p(y|\gamma,\tilde{z}^N,\sigma_\varepsilon^2,\alpha,\beta^G,C^\gamma)^b/p(y^{TN\times b}),\\ p(y^{TN\times b})&=\int p(y|\gamma,\tilde{z}^N,\sigma_\varepsilon^2,\alpha,\beta^G,C^\gamma)^b dC^\gamma.\end{aligned}$$

The fractional prior is easily shown to be the density of a multivariate normal distribution,

$$p(C^\gamma|\tilde{z}^N,\sigma_\varepsilon^2,\alpha,\beta^G,y^{TN\times b}) \sim \text{Normal}\left(a_N^I,\sigma_\varepsilon^2 A_N^I/b\right), \tag{16}$$

where a_N^I and A_N^I are equivalent the moments of the conditional posterior $p(C^\gamma|\tilde{z}^N,\sigma_\varepsilon^2,\alpha,\beta^G,y)$, based on the improper prior $a_0=0, A_0^{-1}=0$. Thus the fractional prior is centered in the posterior mean, obtained under an improper prior, however with the posterior variance-covariance matrix being multiplied by the factor $1/b$.

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 $p(y|\gamma,\alpha,\beta^G,\tilde{z}^N,\sigma_\varepsilon^2,C^\gamma)$. 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 purse 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 $p(y|\gamma,\alpha,\beta^G,\tilde{z}^N,\sigma_\varepsilon^2,C^\gamma)^{1-b}$:

$$\begin{aligned} &p(y|\gamma,\alpha,\beta^G,\tilde{z}^N,\sigma_\varepsilon^2) = \\ &\int p(y|\gamma,\alpha,\beta^G,\tilde{z}^N,\sigma_\varepsilon^2,C^\gamma)^{(1-b)}p(C^\gamma|\tilde{z}^N,\sigma_\varepsilon^2,\alpha,\beta^G,y^{TN\times b})dC^\gamma, \end{aligned} \tag{17}$$

where $p(C^\gamma|\tilde{z}^N,\sigma_\varepsilon^2,\alpha,\beta^G,y^{TN\times b})$ is equal to the fractional prior (16). As only quadratic forms in C^γ are involved both in the fractional prior as well as in the conditional likelihood, it is possible to carry out integration with respect to C^γ explicitly in (17):

$$p(y|\gamma,\alpha,\beta^G,\tilde{z}^N,\sigma_\varepsilon^2) = b^{p_\gamma/2}\left(\frac{1}{2\pi\sigma_\varepsilon^2}\right)^{NT(1-b)/2}\exp\left(-\frac{(1-b)}{2\sigma_\varepsilon^2}S^\gamma\right), \tag{18}$$

where $p_\gamma=\dim(C^\gamma)$ and S^γ is given by (14).

Following Berger and Pericchi (1996) we choose the fraction b for the fractional prior equal to $\frac{m}{N\cdot T}$, where m is the dimension of C^γ for the larger of the two compared models plus 1.

3.1.3. Sampling the indicators

To sample the indicators γ_{lm}, we use exactly the same algorithm as in Smith and Kohn (2002). Generate u from a uniform distribution on $[0,1]$. Let γ_{lj}^{old} denote the current value of γ_{lm}. Then,

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

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

(i-3) if $\gamma_{lm}^{old}=1$ and $u\le p(\gamma_{lm}=0)$, generate $v\sim U[0,1]$ and set $\gamma_{lm}^{new}=0$, if $v\le l(\gamma_{lm}=0)/(l(\gamma_{lm}=0)+l(\gamma_{lm}=1))$;

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

Here $p(\gamma_{lm} = i) = \Pr\{\gamma_{lm} = i|\gamma_{\setminus lm}\}, i = 0, 1$ is the conditional prior of γ_{lm}. $l(\gamma_{lm} = i)$ is equal the marginal likelihood $p(y|\gamma, \alpha, \beta^G, \tilde{z}^N, \sigma^2_\varepsilon)$ defined in (18) where γ_{lm} either takes the value $i = 0$ or $i = 1$. As $p(\gamma_{lm} = 0) \approx \hat{\tau}_\gamma$, the fraction of positive elements of C, we find the following: step (i-1) will occur most often, if this fraction is small; step (i-2) will occur most often, if this fraction is large; the other step 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.

3.1.4. The conditional prior of the indicators

To generate from $\gamma_{lm}|\gamma_{\setminus lm}, \alpha, \beta^G, C, \sigma^2_\varepsilon, y$, we need the conditional prior of γ_{lm} given the remaining elements. Let q_γ be the number of elements of C that are non-zero (before sampling γ^{new}_{lm}). If $\gamma^{old}_{lm} = 1$, then

$$p(\gamma_{lm} = 0) = h_1/(h_1 + 1), \quad p(\gamma_{lm} = 1) = 1/(h_1 + 1),$$

where

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

Note that $1/(h_1 + 1) \approx \hat{\tau}$, where $\hat{\tau} = q_\gamma/(d_s)$ is the estimated fraction of positive elements in C. If $\gamma^{old}_{lm} = 0$, then

$$p(\gamma_{lm} = 0) = h_0/(h_0 + 1), \quad p(\gamma_{lm} = 1) = 1/(h_0 + 1),$$

where

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

3.1.5. Sampling C^γ

We generate $C^\gamma|\gamma, \delta, \tilde{z}^N, \sigma^2_\varepsilon, y$ from the following normal posterior distribution:

$$C^\gamma|\gamma, \tilde{z}^N, \alpha, \beta^G, \sigma^2_\varepsilon, y \sim \text{Normal}\left(a_N, \sigma^2_\varepsilon A_N\right),$$

where a_N and A_N are given by the moments of the posterior $p(C^\gamma|\tilde{z}^N, \sigma^2_\varepsilon, \alpha, \beta^G, y)$ given in (10).

3.2. Sampling the remaining parameters

Conditional on knowing γ and C^γ we are dealing with a random-effects model with known variance-covariance matrix $Q = CC'$, where C is the Cholesky factor corresponding to γ and C^γ. Consequently, one could use any of the MCMC schemes in order to sample α, β^G, σ^2_ε, and the non-centered random effects $\tilde{z}^N$. Here we use the partially marginalized sampler of Frühwirth-Schnatter et al. (2004) which is slightly modified in order to deal with the non-centered parameterization.

3.2.1. Sampling α, β^G

From model (3) and (4) we derive the marginal heteroscedastic model:

$$y_i \sim \text{Normal}\left(Z_i^1\alpha + Z_i^2\beta^G, Z_i^2 Q Z_i^2 + \sigma_\varepsilon^2 I\right) \tag{19}$$

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

We sample the fixed effects α and the mean parameter β^G together in one block from model (19) with the random effects being integrated out. This yields the following posterior distribution:

$$p(\alpha, \beta^G | \gamma, C^\gamma, \sigma_\varepsilon^2, y) \sim \text{Normal}\left(B_N b_N, B_N\right),$$

where

$$B_N^{-1} = \sum_{i=1}^N [Z_i^1\ Z_i^2]'(Z_i^2 Q Z_i^2 + \sigma_\varepsilon^2 I)^{-1}[Z_i^1\ Z_i^2] + B_0^{-1},$$
$$b_N = B_N(\sum_{i=1}^N [Z_i^1\ Z_i^2]'(Z_i^2 Q Z_i^2 + \sigma_\varepsilon^2 I)^{-1} y_i + B_0^{-1} b_0).$$

3.2.2. Sampling $\tilde{z}^N$

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

$$\tilde{z}_i | \gamma, \alpha, \beta^G, C^\gamma, \sigma_\varepsilon^2, y \sim \text{Normal}\left(P_i p_i, P_i\right), \tag{20}$$

where

$$p_i = P_i(\sigma_\varepsilon^{-2}(Z_i^2 C)'(y_i - Z_i^1\alpha - Z_i^2\beta^G)),$$
$$P_i^{-1} = (\sigma_\varepsilon^{-2}(Z_i^2 C)' \cdot (Z_i^2 C) + I).$$

3.2.3. Sampling σ_ε^2

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

$$\sigma_\varepsilon^2 | \gamma, \alpha, \beta^G, \tilde{z}^N, y \sim \text{InvGamma}\left(s_N/2, S_N/2\right),$$

with $s_N = TN + s_0$ and

$$S_N = S_0 + S^\gamma + (a_N - a_0)' A_0^{-1}(a_N - a_0),$$

with S^γ being the sum of squared errors defined in (14).

3.3. Sampling Under Alternative Priors on the Variance-Covariance Matrix

The conditionally conjugate normal prior on the non-zero elements C^γ in the Cholesky factor C was chosen primarily for computational convenience, because it allows running a MCMC scheme involving standard densities, only. The MCMC scheme introduced above, however, is easily extended to deal with non-conjugate priors on the Cholesky factors. In this case the likelihood $p(y|\gamma, \tilde{z}^N, \sigma^2_\varepsilon, \alpha, \beta^G, C^\gamma)$ can be used to construct a Gaussian proposal for C^γ. The Metropolis-Hastings algorithm can then be applied to correct for the non-conjugate prior.

4. Simulation Study

The traditional Gibbs sampling algorithm that is based on the centered parameterization samples the variance-covariance matrix from an inverted Wishart distribution and is known to bias the estimated variance-covariance matrix (Natarajan and Kass (2000)). In the first simulation study we are going to examine whether the new algorithm that is based on the non-centered parameterization leads to an improvement in this respect. We simulate data with full variance-covariance matrices Q and dimensions: $d = 5, T_i = 10, N = 50$. The detailed parameter values used for simulation are given the Appendix.

We compare the results of five algorithms, two of which are based on the standard centered parameterization, whereas three of them use the methods introduced in this article. The two algorithms based on the centered parameterization use the conditional conjugate inverted Wishart distribution prior on the variance-covariance matrix Q, and are applied with two different prior scale matrices. The first prior scale matrix is chosen such that the prior expected variance-covariance matrix equals the identity matrix: $E(Q) = I$. This is the usual default choice if no additional prior information is available. In the second run we select the prior expected variance-covariance matrix to equal the true values $E(Q) = Q$, which typically will not be known in real applications and may be used as a benchmark for these kind of centered algorithms. In both cases the degrees of freedom are set to the minimal value.† The three algorithms based on the non-centered parameterization use the conditional conjugate normal prior on the Cholesky factors of the Cholesky decompostion $Q = CC'$ with different priors. In a first run, estimation based on the non-centered parameterization is carried out for a non-informative normal prior distribution for the free elements in the Cholesky factor C, and for the second run we choose a flat normal prior, with the mean being equal to the lower triangular of the identity matrix. For these two runs we do not include the variable selection step (i), but fix the indicators as $\gamma_{lm} = 1$ in order to obtain an arbitrary variance-covariance matrices. Finally we examine the performance of the new algorithm and carry out all steps (i)-(v), including variable selection.

We base our analysis on 100 data sets. Each algorithm was carried out for 25000 iterations and the first 15000 iterations were skipped for burn-in. We estimate the variance-covariance matrix for two different loss functions: The first Bayes estimate equals the posterior mean and corresponds to the squared error loss function:

$$L = \frac{1}{d^2}\sqrt{\sum_{l=1}^{d}\sum_{m=1}^{d}(\hat{Q}_{lm} - Q_{lm})^2}.$$

†The degrees of freedom have to fulfill $\alpha_0^Q > \frac{1}{2}(d+1)$. The prior scale matrix S_0 is derived from $S_0^Q = E(Q)(\alpha_0^Q - (d+1)/2)$.

Table 1. First simulation study: sample medians for the loss functions L and L_1 for algorithms based on the centered and non-centered parameterization

	L	eig_L^{max}	eig_L^{min}	L^{cond}	L_1	$eig_{L_1}^{max}$	$eig_{L_1}^{min}$	L_1^{cond}
$d = 5, T_i = 10, N = 50$								
true values	-	22.11	2.74	8.06	-	22.11	2.74	8.06
centered, prior $E(Q) = I$	0.32	21.24	1.55	13.75	0.94	18.29	1.18	15.49
centered, prior $E(Q) = Q$	0.28	22.22	2.05	11.49	0.44	19.48	1.74	11.96
non-cent., noninf. prior	0.37	26.69	2.50	10.70	0.41	22.89	2.02	10.97
non-cent., prior $E(C) = I$	0.38	27.10	2.67	10.79	0.36	23.15	2.14	11.06
non-cent., step (i) incl.	0.39	26.60	2.53	10.51	-	-	-	-

We give the sample median of the squared error loss (L), the biggest (eig_L^{max}) and the smallest (eig_L^{min}) eigenvalue, and the condition number (L^{cond}) of the posterior mean estimator of the variance-covariance matrix Q in the first four columns of Table 1. In Figure 1, 2, and 3 we make boxplots of these measures for the five algorithms.

Alternatively we choose the following estimate of the variance-covariance matrix

$$(E(Q^{-1}|y))^{-1},$$

which is the Bayes estimator with respect to the loss function

$$L_1(Q, \hat{Q}) = tr(\hat{Q}Q^{-1}) - log|\hat{Q}Q^{-1}| - d,$$

see Yang and Berger (1994) for details. In column 5-8 of Table 1 we give again the same measures for this estimator (L_1, $eig_{L_1}^{max}$, $eig_{L_1}^{min}$, L_1^{cond}) and in Figure 4, 5, and 6 we draw again the corresponding boxplots.

From the figures and from Table 1 we see that the non-centered algorithms yield better results for nearly all values. This is especially true for the L_1 metric, where the performance of the centered algorithm is worse for all measures. Interestingly the centered algorithm is outperformed by the non-centered algorithms even if we assume the true variance-covariance matrix as the prior scale matrix. For the squared error loss metric the results are not unique. The maximum eigenvalue is overestimated by the non-centered algorithms and the squared error loss is a little bit smaller for the centered parameterization. On the other hand the minimal eigenvalue as well as the condition number is estimated more accurately by the non-centered parameterization. The results for the new algorithm that additionally includes the variance-covariance selection step do not diverge from the results of the other two non-centered algorithms that do not include step (i). So in practice this variance-covariance selection step may be included into the non-centered algorithm without loss of quality of the estimated variance-covariance matrix, even when estimating a full variance-covariance matrix.

So far we demonstrated that the algorithms based on the non-centered parameterization yield improved estimates of the variance-covariance matrices in comparison to traditional algorithms which are based on the centered parameterization. In our second simulation study we are going to examine the ability of the new algorithm to find the true structure of the variance-covariance matrix. We use a variance-covariance matrix Q of dimension 15 times 15 so that we have 120 free elements in the matrix for which we have to carry out variance-covariance selection. The rank of Q is 12 and 52 off-diagonal elements are zero. We simulate data for $T_i = 20$ and $N = 150$. Details are given in the Appendix. We

Table 2. Second simulation study: Median of the percentage rates of correctly identified zero and non-zero elements in the lower triangular variance-covariance matrix

non-zero diagonal elements	zero diagonal elements	non-zero off-diagonal elements	zero off-diagonal elements
100	100	77.36	99.04

simulated 64 data sets and give the medians of percentage rates of correctly identified zero and non-zero values of Q in Table 2. The identification of the diagonal elements of the variance-covariance matrix is crucial for selecting fixed and random effects and in Table 2 we find that our algorithm identifies these effects perfectly. Concerning the off-diagonal elements, 99.04% of the zero off-diagonal elements are selected correctly and 77.36% of the non-zero off-diagonal elements are included into the model. Therefore the algorithm estimates a model that is more parsimonious than the model our simulated data were based on.

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. 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 Z_i^2 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 15000 iterations of our new procedure. The first 5000 iterations were discarded for burn-in. The probability for each element of the variance-covariance matrix Q to be non-zero may easily be derived from the simulations of the indicators γ and equation (5). In Table 3 we give these posterior probabilities for the elements of the variance-covariance matrix to be non-zero. Only the variance for the interaction of the third brand with the quadratic price effect (14th diagonal element in Table 3) has a low probability of 0.04 for being significantly different from zero and is estimated as a fixed effect here. All the other variances have posterior probabilities between 0.87 and 1 to be different from zero and are therefore determined as random effects. For the first nine design parameters, also all covariances are non-zero, as we can see from the corresponding posterior probabilities taking the value one in Table 3. For some of the other random effects the variance-covariance matrix is more sparse.

The power of the new variance-covariance selection method becomes obvious when comparing these results to results obtained by traditional methods. In the centered parameterization these data have been analyzed in Tüchler (2003) by means of a Gibbs sampling algorithm where the variance-covariance matrix was sampled from an inverted Wishart distribution. This work compares a model where all effects are included as random effects with

Table 3. Posterior probability for the elements of the variance-covariance matrix Q to be significantly different from zero (rounded).

1	1	1	1	1	1	1	1	1.00	0.01	0.44	0	0.00	0.00	0.00
-	1	1	1	1	1	1	1	1.00	0.01	0.44	1.00	0.98	0.02	0.00
-	-	1	1	1	1	1	1	1.00	0.01	0.44	1.00	0.98	0.02	0.00
-	-	-	1	1	1	1	1	1.00	1	1	0.05	0.03	0.00	0.01
-	-	-	-	1	1	1	1	1.00	0.01	1	0.04	0.01	0.00	0.01
-	-	-	-	-	1	1	1	1.00	0.63	0.46	0.91	0.90	0.02	0.01
-	-	-	-	-	-	1	1	1.00	0.07	1	0.31	0.25	0.00	0.02
-	-	-	-	-	-	-	1	1	1	0.46	1	0.05	0.00	0.87
-	-	-	-	-	-	-	-	1	1	0.44	0.20	0.05	0.00	0.01
-	-	-	-	-	-	-	-	-	1	0.01	0.17	0.02	0.00	0.01
-	-	-	-	-	-	-	-	-	-	1	0.07	0.00	0.00	0.02
-	-	-	-	-	-	-	-	-	-	-	1	0.98	0.02	0.87
-	-	-	-	-	-	-	-	-	-	-	-	0.98	0.02	0.04
-	-	-	-	-	-	-	-	-	-	-	-	-	0.04	0.00
-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.87

a model where the interaction effects of the brands with the quadratic price parameter are fixed by means of model likelihoods. Among these two random-effects models, the model likelihoods clearly favored the second one (logarithm of the model likelihood is -9222.36 for the model with fixed interaction effects and -9291.99 for the model with all effects specified as random). In contrast to that our new procedure selects just one single brand by quadratic price effect as fixed. The three others are selected as random effects although most of the corresponding covariance elements are set to zero. In Tüchler (2003) there are 54 additional unknown variance-covariance parameters in the full model in comparison to the model where interactions are fixed. In addition to that the new variable selection procedure suggests that most covariances of these interaction effects are not significantly different from zero (Table 3). Therefore the model likelihoods rather prefer the model with fixed interaction effects and fewer parameters. Our new procedure is more flexible and adds only 13 significant elements for the brand by quadratic price effects and enables us to make better use of the information in the data.

Our procedure is clearly more flexible than the alternative variable selection method for random effects models of Chen and Dunson (2003). For their method it is not possible to select non-zero covariances for effects with non-zero variances. All covariances are automatically included for those effects which are specified as random effects by the procedure.

In Table 4 we give the posterior probabilities for the elements of C to be different from zero. From comparison with Table 3 we find that the number of zero elements is much smaller in C than in Q. Interestingly estimation on the basis of the Cholesky decomposition proceeds with fewer parameters in C than in the resulting variance-covariance matrix and therefore offers a very parsimonious estimation tool.

6. Concluding Remarks

In this paper, we considered a non-centered parameterization of the standard random-effects model, which is based on the Cholesky decomposition of the variance-covariance matrix. The choice of this parameterization offers several advantages. First, posterior simulations using MCMC schemes are efficient and automatically deliver variance-covariance

Table 4. Posterior probability for the elements of the Cholesky factor matrix C to be significantly different from zero (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	1	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0.01	1	1	0	0	0	0	0	0	0	0	0	0	0
1	0	0	1	0.04	0	0	0	0	0	0	0	0	0	0
1	0.91	0.58	0.01	0.93	1	0	0	0	0	0	0	0	0	0
1	0.27	0.06	1	1	0.07	0.01	0	0	0	0	0	0	0	0
1	0	0.01	0	0	0.01	1	1	0	0	0	0	0	0	0
1	0.04	0	0	0	0	0.01	1	0.01	0	0	0	0	0	0
0.01	0	1	0	0	0.05	0	1	0.04	0.02	0	0	0	0	0
0.44	0	0	1	0.01	0.02	0.04	0	0.01	0.01	0.03	0	0	0	0
0	1	0	0.04	0.01	0	1	0.16	0	0	0	0	0	0	0
0	0.98	0.01	0	0	0	0.04	0	0.04	0.01	0.01	0.05	0.01	0	0
0	0.02	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0.01	0	0	0.87	0	0	0	0	0	0	0	0

matrices without the need to introduce any constraints, as the Cholesky factors of a variance-covariance matrix are unconstrained. This feature is rather desirable from a computational point of view.

Second, the regression type structure of the non-centered parameterization, where the elements of the Cholesky factors appear as a regression coefficient, allows to choose a simple, conditionally conjugate normal prior on the Cholesky factor. The first simulation study in Section 4 demonstrated that this prior is less influential on the estimated variance-covariance matrix than the inverted Wishart prior, which is the corresponding conditionally conjugate prior for the centered parameterization.

Finally, based on the non-centered parameterization, we were 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. In particular, with this method we are able to learn from the data for each effect, whether it is random or not. This result is potentially of great interest in many areas of applied statistics.

Acknowledgments

We would like to thank Mike Smith for many very helpful comments and suggestions. We thank Mena Stefan for competent and very helpful computational assistance with the simulation studies in Section 4. This work was supported by the Austrian Science Foundation (FWF) under grant SFB 010 ('Adaptive Information Systems and Modelling in Economics and Management Science').

A. Design of the simulation studies

For the first simulation study we simulated data from the random-effects model (3),(4) with design matrix Z_i^2 equal to

$$Z_i^2 = \begin{pmatrix} 1 & 1 & 0 & 0 & z_1 \\ 1 & 1 & 0 & 0 & z_2 \\ 1 & 1 & 0 & 0 & z_3 \\ 1 & 0 & 1 & 0 & z_1 \\ 1 & 0 & 1 & 0 & z_2 \\ 1 & 0 & 1 & 0 & z_3 \\ 1 & 0 & 0 & 1 & z_3 \\ 1 & 0 & 0 & 1 & z_4 \\ 1 & 0 & 0 & 0 & z_3 \\ 1 & 0 & 0 & 0 & z_4 \end{pmatrix},$$

where the values of z_1 vary between 0 and 0.2, those of z_3 vary between 4 and 4.2, those of z_4 vary between 6.4 and 7.2, and z_2 takes the value 2.1. We include no fixed effects ($\alpha = 0$), and the random effects have the mean parameter $\beta^G = [15\ 5\ 5\ 4.5\ -2]$ and variance-covariance matrix

$$Q = \begin{pmatrix} 12.4 & 0.6 & 2.9 & 3.9 & 4.4 \\ 0.6 & 14.5 & 4.0 & 2.9 & 2.2 \\ 2.9 & 4.0 & 10.0 & 3.3 & 2.6 \\ 3.9 & 2.9 & 3.3 & 7.3 & 2.7 \\ 4.4 & 2.2 & 2.6 & 2.7 & 5.2 \end{pmatrix}.$$

The model error variance σ_ε^2 equals 1.

For the second simulation study the design matrices Z_i^2 consist of 20 observations and 15 parameters and have a similar structure like the design matrices of the first simulation study. The mean parameter equals $\beta^G = [15\ 5\ 5\ 4.5\ -2\ -1.8\ -2.5\ 1\ 2\ .5\ -1\ 1\ 0.5\ -2\ -1]$ and the upper triangular part of the variance-covariance matrix Q writes

$$\begin{pmatrix} 99.6 & 3.3 & 0 & 0 & 0 & 0 & 0 & 94.8 & 23.1 & 60.6 & 48.5 & -8.6 & -33.2 & 2.5 & 5.7 \\ & 130.5 & 0 & 0 & 0 & 0 & 0 & 27.8 & 71.5 & 55.2 & 1.8 & -0.3 & -1.1 & 0.1 & 0.2 \\ & & 84.8 & 0 & 0 & 0 & 0 & 0 & 0 & 4.1 & 3.6 & 37.8 & 20.5 & 28.3 & 36.5 \\ & & & 63.5 & 0 & 0 & 0 & 0 & 0 & 0 & 3.4 & 2.9 & 0 & 0 & 0 \\ & & & & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & & & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & & & & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & & & & & 130 & 35.4 & 70.4 & 48.5 & -6.3 & -31.7 & 2.4 & 5.5 \\ & & & & & & & & 77.2 & 43.6 & 15.4 & 7.4 & -7.6 & 0.6 & 1.3 \\ & & & & & & & & & 82.3 & 30.2 & -0.4 & -16.8 & 6 & 9.3 \\ & & & & & & & & & & 47 & 6.2 & -10.2 & 6.3 & 9.1 \\ & & & & & & & & & & & 134.1 & 100 & 78.7 & 96.5 \\ & & & & & & & & & & & & 111.3 & 70.7 & 85.5 \\ & & & & & & & & & & & & & 68.3 & 71.1 \\ & & & & & & & & & & & & & & 98.5 \end{pmatrix}.$$

The model error variance σ_ε^2 equals 1.

References

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

Barnard, J., R. McCulloch, and X.-L. Meng (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica 10*, 1281–1311.

Berger, J. O. and L. R. Pericchi (1996). Objective bayesian methods for model selection: introduction and comparison. In P. Lahirini (Ed.), *Model Selection*, pp. 135–207. Beachwood.

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

Chiu, T. J., T. Leonard, and T. Kam-Wah (1996). The matrix-logarithmic covariance model. *Journal of the American Statistical Association 91*, 198–210.

Daniels, M. J. (1999). A prior for the variance in hierarchical models. *Canadian Journal of Statistics 27*, 567–578.

Daniels, M. J. and R. Kass (1999). Nonconjugate Bayesian estimation of covariance matrices and its use in hierarchical models. *Journal of the American Statistical Association 94*, 1254–1263.

Daniels, M. J. and R. Kass (2001). Shrinkage estimators for covariance matrices. *Biometrics 57*, 1173–1184.

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

Everson, P. J. and C. N. Morris (2000). Inference for multivariate normal hierarchal models. *Journal of Royal Statistical Society, Series B 62*, 399–412.

Frühwirth-Schnatter, S. (2004). Efficient Bayesian parameter estimation for state space models based on reparameterizations. In A. Harvey, S. J. Koopman, and N. Shephard (Eds.), *State Space and Unobserved Component Models: Theory and Applications. Proceedings of a Conference in Honour of James Durbin*, Cambridge, pp. 123–151. Cambridge University Press.

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

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

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

Hobert, J. P. and G. Casella (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association 91*, 1461–1473.

Leonard, T. and J. S. Hsu (1992). Bayesian inference for a covariance matrix. *The Annals of Statistics 20*, 1669–1696.

Liechty, J., M. Liechty, and P. Müller (2004). Bayesian correlation estimation. *Biometrika 5*, forthcoming.

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

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

Natarajan, R. and R. E. Kass (2000). Reference Bayesian methods for generalized linear mixed models. *Journal of the American Statistical Association 95*, 227–237.

Natarajan, R. and C. E. McCulloch (1998). Gibbs sampling with diffuse proper priors: A valid approach to data-driven inference? *Journal of Computational and Graphical Statistics 7*, 267–277.

O'Hagan, A. (1995). Fractional Bayes factors for model comparison (Disc: p118-138). *Journal of the Royal Statistical Society, Series B, Methodological 57*, 99–118.

Otter, T., R. Tüchler, and S. Frühwirth-Schnatter (2004). Capturing consumer heterogeneity in metric conjoint analysis using Bayesian mixture models. *International Journal of Marketing Research 21*, forthcoming.

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

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

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

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

Sun, D., R. K. Tsutakawa, and Z. He (2001). Propriety of posteriors with improper priors in hierarchical linear mixed models. *Statistica Sinica 11*, 77–95.

Tüchler, R. (2003). *Bayesian Modelling of Unobserved Heterogeneity - with an Application to Metric Conjoint Analysis*. Phd thesis, University of Technology, Vienna, Austria.

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

van Dyk, D. and X.-L. Meng (2001). The art of data augmentation. *Journal of Computational and Graphical Statistics 10*, 1–50.

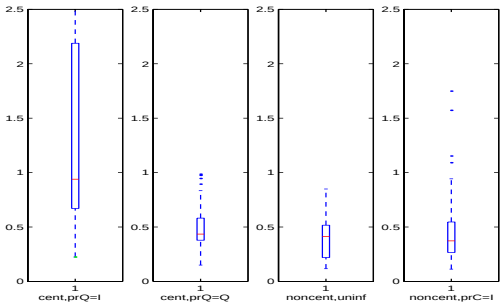


Fig. 1. First simulation study: boxplots for L_1 loss

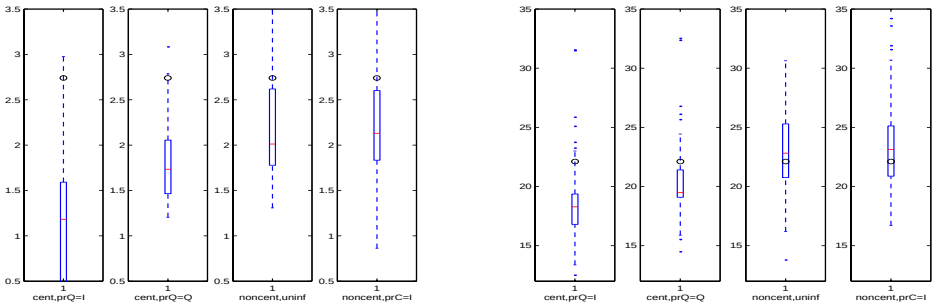


Fig. 2. First simulation study: boxplots for the smallest (left hand side) and the biggest (left hand side) eigenvalue of the Bayes estimator with resp. to L_1 loss, true values: 'o'

Wong, F., C. Carter, and R. Kohn (2003). Efficient estimation of covariance selection models. *Biometrika 90*, 809–830.

Yang, R. and J. O. Berger (1994). Estimation of a covariance matrix using the reference prior. *The Annals of Statistics 22*, 1195–1211.

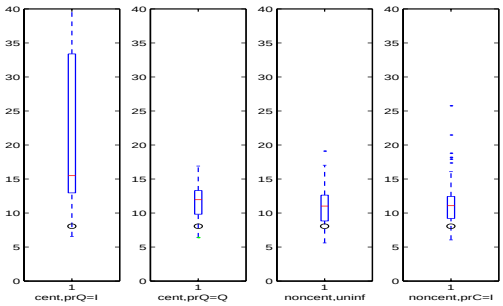


Fig. 3. First simulation study: boxplots for the condition number of the Bayes estimator with resp. to L_1 loss, true values: 'o'

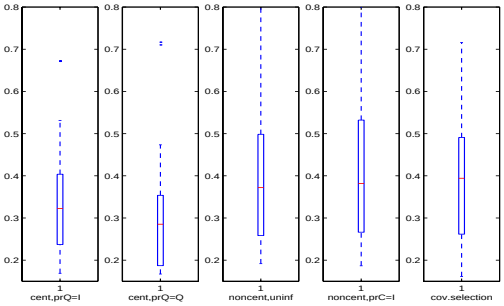


Fig. 4. First simulation study: boxplots for squared error loss

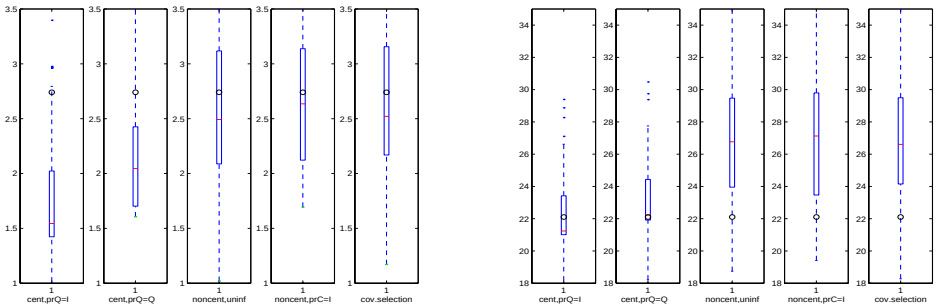


Fig. 5. First simulation study: boxplots for the smallest (left hand side) and the biggest (left hand side) eigenvalue of the Bayes estimator with resp. to squared error loss, true values: 'o'

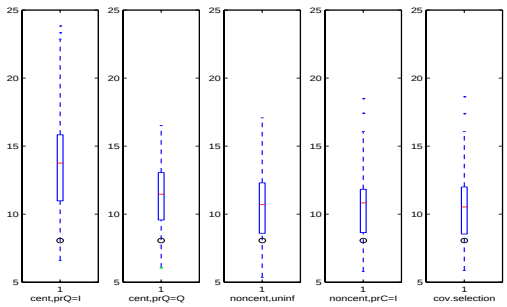


Fig. 6. First simulation study: boxplots for the condition number of the Bayes estimator with resp. to squared error loss, true values: 'o'