Insper

Parsimonious Bayesian Factor Analysis when the Number of Factors is Unknown

Sylvia Fruhwirth-Schnatter
Hedibert Freitas Lopes

Insper Working Paper
WPE: 345/2014

Parsimonious Bayesian Factor Analysis when the Number of Factors is Unknown

by

Sylvia Frühwirth-Schnatter

Institute for Statistics and Mathematics
Vienna University of Economics and Business
Gebäude D4, 4. Stock, Welthandelsplatz 1, A-1020 Wien
Email: sylvia.fruehwirth-schnatter@wu.ac.at

and

Hedibert Freitas Lopes

INSPER Institute of Education and Research
Rua Quatá 300, São Paulo, 04546-042, Brazil
E-mail: hedibertFL@insper.edu.br

Abstract

We introduce a new and general set of identifiability conditions for factor models which handles the ordering problem associated with current common practice. In addition, the new class of parsimonious Bayesian factor analysis leads to a factor loading matrix representation which is an intuitive and easy to implement factor selection scheme. We argue that the structuring the factor loadings matrix is in concordance with recent trends in applied factor analysis. Our MCMC scheme for posterior inference makes several improvements over the existing alternatives while outlining various strategies for conditional posterior inference in a factor selection scenario. Four applications, two based on synthetic data and two based on well known real data, are introduced to illustrate the applicability and generality of the new class of parsimonious factor models, as well as to highlight features of the proposed sampling schemes.

Key words: Identifiability; parsimony; Cholesky decomposition; rank defficiency; Heywood problem.

JEL codes: C01, C11,C31, C51, C52, C63.

1 Introduction

Selecting the number of common factors in factor analysis has been known since long to be a very difficult issue. Lopes & West (2004) are amongst the first ones to formally tackle this issue in standard normal linear factor analysis. They propose a reversible jump Markov chain Monte Carlo scheme whose independent proposal densities are based on multivariate normal and multivariate Student's t approximations to within model posterior distributions based on preliminary offline draws. Their main limitation is the high computational cost associated with these preliminary runs, particularly when either or both the number of observations and number of variables are large.

Our approach follows current trend in modern Bayesian factor analysis where more modeling structure is incorporated through the columns of the factor loadings matrix, through the common factors themselves or both. In Lopes, Salazar & Gamerman (2008), for example, a new class of spatial dynamic factor models is introduced where the temporal dependence is modeled by the latent dynamic factors while the spatial dependence is modeled by the columns of the factor loadings matrix. Other recent papers on structure factor models are, to name but a few, Lopes & Carvalho (2007) who model time-varying covariances via factor stochastic volatility models with time-varying loadings and Carvalho, Chang, Lucas, Nevins, & West (2008) who model high-dimensional gene expression data via sparse factor analysis.

We add to the discussion by showing how the true factor loading structure, including the number of factors, may be recovered when the fitted model is overfitting the number of factors. In this case, well-known identifiability problems arise, see Geweke & Singleton (1980) and Geweke & Zhou (1996). In the present paper, we formulate a new and general set of identifiability conditions that relax the usual lower triangular condition on the factor loading matrix. First of all, these conditions are able to handle the ordering problem present in most of the literature. Second, we show that under these conditions the factor loading matrix of an overfitting model takes a special form which allows easily to reconstruct the true number of factors and the true factor loading matrix.

Since this identification procedure relies on identifying zeros and non-zeros elements in the factor loading matrix, we perform parsimonious factor modeling and consider the selection of the elements of the factor loading matrix as a variable selection problem, as Frühwirth-Schnatter & Tüchler (2008) did for the related problem of parsimonious modelling of the Cholesky factors of a covariance matrix in a random-effects model. To this aim, we introduce a binary indicator matrix of the same dimension as the factor loading matrix and develop a computationally efficient Markov chain Monte Carlo (MCMC) scheme for posterior inference. Our approach not only leads to sparse factor loading structures as in the previous literature, but also allows to select the number of common factors by identifying overfitting factor models from the structure of the indicator matrix.

The paper is organized as follows. New identifiability and rank deficiency issues arise and

are properly dealt with in Section 2, along with other basic properties of factor analysis. Prior specification and posterior Bayesian inference are described in Section 3. Posterior inference is performed via a customized MCMC scheme described in Section 4. The section also outlines various strategies for conditional posterior inference in the factor selection scenario. Four applications, two based on synthetic data and two based on well known real data, are introduced in Section 5. They illustrate the applicability and generality of the new class of parsimonious factor models, as well as highlight features of the proposed sampling schemes. Section 6 concludes.

2 Factor Model Specification

2.1 The Basic Factor Model

Data on m variables are assumed to arise from a multivariate normal distribution $N_m(\mathbf{0}, \mathbf{\Omega})$ with zero mean and unknown covariance matrix $\mathbf{\Omega}$. A single observation is denoted by $\mathbf{y}_t = (y_{1t}, \ldots, y_{mt})'$ and a random sample is denoted by $\mathbf{y} = \{\mathbf{y}_t, t = 1, \ldots, T\}$. A factor model relates each observation $\mathbf{y}_t$ to a latent r-variate random variable $\mathbf{f}_t = (f_{1t} \cdots f_{rt})'$, the so-called common factors, through:

$$\mathbf{y}_t = \mathbf{\Lambda}\mathbf{f}_t + \boldsymbol{\epsilon}_t, \tag{1}$$

where $\mathbf{\Lambda}$ is the unknown $m \times r$ factor loading matrix with elements Λ_{ij}. The standard assumption is that $\mathbf{f}_t$, $\mathbf{f}_s$, $\boldsymbol{\epsilon}_t$, and $\boldsymbol{\epsilon}_s$ are pairwise independent for all t and s and that

$$\mathbf{f}_t \sim N_r(\mathbf{0}, \mathbf{I}_r), \tag{2}$$

$$\boldsymbol{\epsilon}_t \sim N_m(\mathbf{0}, \mathbf{\Sigma}_0), \qquad \mathbf{\Sigma}_0 = \text{Diag}\left(\sigma_1^2, \ldots, \sigma_m^2\right). \tag{3}$$

Assumption (3) implies that conditional on knowing $\mathbf{f}_t$ the m elements of $\mathbf{y}_t$ are independent and all dependence among these variables is explained by the common factors. This leads to the following constrained variance-covariance structure for $\mathbf{\Omega} = \text{V}(\mathbf{y}_t|\boldsymbol{\beta}, \mathbf{\Sigma}_0)$:

$$\mathbf{\Omega} = \mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Sigma}_0. \tag{4}$$

The idiosyncratic variances $\sigma_i^2 = \text{V}(y_{it}|\mathbf{f}_t, \boldsymbol{\beta}, \mathbf{\Sigma}_0)$ may be compared to the marginal variance $\text{V}(y_{it}|\boldsymbol{\beta}, \mathbf{\Sigma}_0)$ to determine the proportion of the variance of y_{it} that is explained by the common factors, also known as communalities (Bartholomew, 1987):

$$R_i^2 = 1 - \frac{\sigma_i^2}{\text{V}(y_{it}|\boldsymbol{\beta}, \mathbf{\Sigma}_0)} = \sum_{j=1}^{r} R_{ij}^2, \qquad R_{ij}^2 = \frac{\Lambda_{ij}^2}{\sum_{j=1}^{r} \Lambda_{ij}^2 + \sigma_i^2}. \tag{5}$$

2.2 New set of identifiability conditions

A well-known identification problem arises for the basic factor model. A rather trivial non-identifiability problem is *sign-switching*. Evidently, the signs of the elements of $\mathbf{f}_t$ and $\mathbf{\Lambda}$ are not identified: for each $j = 1, \ldots, r$ all elements $f_{j1}, \ldots, f_{jT}$ of the latent factors as well as the jth column of $\mathbf{\Lambda}$ may be changed by the same sign switch. A more serious problem is factor rotation. Without imposing further constraints on $\mathbf{\Lambda}$ the model is invariant under any transformation of the form $\mathbf{\Lambda}^\star = \mathbf{\Lambda}\mathbf{P}'$ and $\mathbf{f}_t^\star = \mathbf{P}\mathbf{f}_t$, where $\mathbf{P}$ is an arbitrary orthogonal matrix of dimension $k \times k$.

A common way of dealing with these problems is to constrain the upper triangular part of $\mathbf{\Lambda}$ to be zero and to assume that main diagonal elements of $\mathbf{\Lambda}$ are strictly positive, i.e. $\Lambda_{jj} > 0$ for all $j = 1, \ldots, k$, see e.g. Geweke & Zhou (1996). This constraint simultaneously prevents factor rotation and identifies the sign of the elements of $\mathbf{f}_t$ and $\mathbf{\Lambda}$, however it is generally too restrictive. It induces an order dependence among the responses and makes the appropriate choice of the first r response variables an important modeling decision (Carvalho et al., 2008). Well-known difficulties arise if one of the true factor loadings Λ_{jj} is equal to or close to 0, see e.g. Lopes & West (2004).

In the present paper we suggest a new and more general set of identifiability conditions for the basic factor model which handles the ordering problem in a more flexible way:

C1. $\mathbf{\Lambda}$ has full column-rank, i.e. $r = \text{rank}(\mathbf{\Lambda})$.

C2. $\mathbf{\Lambda}$ is a *generalized lower triangular* matrix, i.e. $l_1 < \ldots < l_r$, where l_j denotes for $j = 1, \ldots, r$ the row index of the top non-zero entry in the jth column of $\mathbf{\Lambda}$, i.e. $\Lambda_{l_j,j} \neq 0; \Lambda_{ij} = 0, \forall\, i < l_j$.

C3. $\mathbf{\Lambda}$ does not contain any column j where $\Lambda_{l_j,j}$ is the only non-zero element in column j.

Condition $\mathbf{C2}$ means that $\mathbf{\Lambda}$ is a generalized lower triangular matrix in the sense that the top non-zero entries in the r columns of $\mathbf{\Lambda}$ have increasing row indices $l_1, \ldots, l_r$ with $l_j \geq j$. These indices are well-defined, because $\mathbf{\Lambda}$ does not contain any zero column due to condition $\mathbf{C1}$. Condition $\mathbf{C2}$ covers the case where $\mathbf{\Lambda}$ is a lower triangular matrix with strictly non-zeros entries on the main diagonal ($l_j = j$ for $j = 1, \ldots, r$), but allows for more general forms of triangular matrices, if the ordering of the response variables is in conflict with this assumption. As we allow Λ_{jj} to be 0, response variables different from the first r ones are allowed to lead the factors. Indeed, for each factor j, the response variable l_j corresponding to the top non-zero element is the leading variable. Condition $\mathbf{C2}$ prevents factor rotation, but not sign switching. Sign switching is prevented by requiring additionally that $\Lambda_{l_j,j}$ is positive for each $j = 1, \ldots, r$.

Finally, condition $\mathbf{C3}$ is important to ensure that the true number of factors is identifiable from $\text{rank}(\mathbf{\Lambda})$ and will become clear in the light of our Theorem 1 below. The theorem proves that a factor model whose factor loading matrix contains a column with only a single non-zero element is observationally equivalent to a factor model with reduced rank.

2.3 The regression-type representation of a factor model

Assume that data $\mathbf{y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_T\}$ were generated by model (1) and that the number of factors r, as well as $\mathbf{\Lambda}$ and $\mathbf{\Sigma}_0$, should be estimated. The usual procedure is to fit a model with k factors,

$$\mathbf{y}_t = \boldsymbol{\beta}\mathbf{f}_t + \boldsymbol{\epsilon}_t, \qquad \boldsymbol{\epsilon}_t \sim N_m\left(\mathbf{0}, \mathbf{\Sigma}\right), \tag{6}$$

where $\boldsymbol{\beta}$ is a $m \times k$ coefficient matrix with elements β_{ij} and $\mathbf{\Sigma}$ is a diagonal matrix. We call (6) the regression-type representation of the true factor model (1) with k potential factors.

Identifying $\mathbf{\Lambda}$, $\mathbf{\Sigma}_0$, and r from (6) turns out to be surprisingly difficult, in particular, if the regression-type representation is overfitting the number of factors, i.e. $k > r$. In this case, the true model (1) is embedded in the regression-type representation (6) by adding $k - r$ zeros columns. The resulting coefficient matrix $\boldsymbol{\beta}_0$ is related to $\mathbf{\Lambda}$ through $\boldsymbol{\beta}_0\boldsymbol{\beta}_0^{'} = \mathbf{\Lambda}\mathbf{\Lambda}^{'}$. Hence $\text{rank}(\boldsymbol{\beta}_0) = \text{rank}(\mathbf{\Lambda}) = r$ and $\boldsymbol{\beta}_0$ is rank deficient, if $r < k$. This introduces a serious identifiability problem as shown by Geweke & Singleton (1980). More precisely, there exists a $k \times (k - r)$ matrix $\mathbf{Q}$ such that $\boldsymbol{\beta}_0\mathbf{Q} = \mathbf{O}_{m\times(k-r)}$ and $\mathbf{Q}^{'}\mathbf{Q} = \mathbf{I}_{k-r}$. For any $m \times (k - r)$ dimensional matrix $\mathbf{M}$ with mutually orthogonal rows another regression-type representation of the true factor model may be defined with parameters $\boldsymbol{\beta}$ and $\mathbf{\Sigma}$ given by:

$$\boldsymbol{\beta} = \boldsymbol{\beta}_0 + \mathbf{M}\mathbf{Q}^{'}, \qquad \mathbf{\Sigma} = \mathbf{\Sigma}_0 - \mathbf{M}\mathbf{M}^{'}. \tag{7}$$

The parameters $(\boldsymbol{\beta}, \mathbf{\Sigma})$ and $(\boldsymbol{\beta}_0, \mathbf{\Sigma}_0)$ define the same likelihood and are observationally equivalent since $\mathbf{\Omega} = \boldsymbol{\beta}_0\boldsymbol{\beta}_0^{'} + \mathbf{\Sigma}_0 = \boldsymbol{\beta}\boldsymbol{\beta}^{'} + \mathbf{\Sigma}$. Hence, $\boldsymbol{\beta}_0$ and $\mathbf{\Sigma}_0$ are not identifiable. It seems to be impossible to identity r and $\mathbf{\Lambda}$ from the coefficient matrix $\boldsymbol{\beta}$ in an overfitting regression-type representation, because the rank of $\boldsymbol{\beta}$ could be larger than r and it is not clear how the columns of $\boldsymbol{\beta}$ are related to $\mathbf{\Lambda}$ in this case. We solve this problem in Theorem 1.

Theorem 1. Assume that the data were generated by a basic factor model (1) obeying conditions $\mathbf{C1} - \mathbf{C3}$ and that a regression-type representation (6) with $k \geq r$ potential factors is fitted. Assume that the following condition holds for $\boldsymbol{\beta}$:

B1 The row indices of the top non-zero entry in each non-zero column of $\boldsymbol{\beta}$ are different.

Then r, $\mathbf{\Lambda}$ and $\mathbf{\Sigma}_0$ are related in the following way to the coefficient matrix $\boldsymbol{\beta}$ and the matrix $\mathbf{\Sigma}$ of the regression-type representation:

(a) r columns of $\boldsymbol{\beta}$ are, up to a column permutation, identical to the r columns of $\mathbf{\Lambda}$.

(b) If $\text{rank}(\boldsymbol{\beta}) = r$, then the remaining $k - r$ columns of $\boldsymbol{\beta}$ are zero columns. $\mathbf{\Lambda}$ is obtained by deleting the zero-column in $\boldsymbol{\beta}$ and permuting the remaining columns in such a way that condition **C2** is fullfilled. The number of factors r is equal to the number of non-zero columns in $\boldsymbol{\beta}$. Furthermore, $\mathbf{\Sigma}_0 = \mathbf{\Sigma}$.

(c) If rank($\boldsymbol{\beta}$) $> r$, then only $k -$ rank($\boldsymbol{\beta}$) of the remaining $k - r$ columns are zero columns, while $s =$ rank($\boldsymbol{\beta}$) $- r$ columns with column indices $j_1, \ldots, j_s$ differ from a zero column for a single element lying in s different rows $r_1, \ldots, r_s$. In this case $\boldsymbol{\Lambda}$ is obtained from $\boldsymbol{\beta}$ by deleting all zero columns and all columns where a single element differs from 0 and permuting the remaining columns in such a way that condition **C2** is fullfilled. The number of factors is equal to $r =$ rank($\boldsymbol{\beta}$) $- s$. Furthermore, $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma} + \mathbf{D}$, where $\mathbf{D}$ is a $m \times m$ diagonal matrix of rank s with non-zero diagonal elements $D_{r_l,r_l} = \beta^2_{r_l,j_l}$ for $l = 1, \ldots, s$.

A proof of this theorem is given in Appendix A.

Theorem 1 shows that it is not necessary to impose the usual assumption $\beta_{11} > 0, \ldots, \beta_{kk} > 0$ when fitting a regression-type model to achieve identifiability. Our more general condition $\mathbf{B1}$ allows $\boldsymbol{\beta}$ to be rank deficient and we do not force uniqueness of the position of the columns. Note that condition $\mathbf{B1}$ is also less stringent than the identifiability conditions $\mathbf{C1} - \mathbf{C3}$ on the underlying factor model.

Theorem 1 shows how easy it is to identify overfitting factor models under condition $\mathbf{B1}$. In particular, part (c) of Theorem 1 allows to identify coefficient matrices $\boldsymbol{\beta}$ in the regression-type representation which are spuriously overfitting the true number of factors. Condition $\mathbf{C3}$ guarantees that the underlying true factor loading matrix $\boldsymbol{\Lambda}$ it not of the overfitting type by itself. Overfitting is present whenever at least one column of $\boldsymbol{\beta}$ contains only a single non-zero element, i.e. whenever a certain factor is loading on exactly one of the variables. We found several examples of such factor loading matrices in the literature, see e.g. Lopes & West (2004, Figure 1). Because each of these columns increases the rank of the coefficient matrix by 1, an estimator of the number of factors based on the rank of the coefficient matrix is spuriously overfitting the true number of factors r. According to our theorem any such column may be replaced by a zero column, while the corresponding idiosyncratic variance is increased. This reduces the rank of the coefficient matrix by the number of such columns and the reduced rank is identical to the number of factors r.

Once, all zero column and all columns containing at most one non-zero element have been removed, the number of remaing columns defines r and $\boldsymbol{\Lambda}$ is obtained by reordering the remaining columns in such a way that condition **C2** is fullfilled. Condition **B1** guarantees that this reordering is possible and unique.

Theorem 1 is also useful, if the regression-type representation is not overfitting the number of factors, i.e. $k = r$. In this case, it follows that $\boldsymbol{\beta}$ and $\boldsymbol{\Lambda}$ coincide apart from a permutation of the columns and sign switching. Hence, condition **B1** allows the model to react flexibly, if the leading r responses were chosen poorly.

2.4 Parsimonious Bayesian Factor Analysis

Theorem 1 allows to select a possibly high number k of potential factors when fitting a regression-type model. Condition **B1** implies that $\boldsymbol{\beta}$ has at most $mk - k(k-1)/2$ free elements. Adding $\sigma_1^2, \ldots, \sigma_m^2$, gives a total number of parameters equal to $m(k+1)-k(k-1)/2$ which is bounded by the number of free elements in $\boldsymbol{\Omega}$, $m(m+1)/2$. This implies the following upper bound for k:

$$k \leq m + \frac{3}{2} - \sqrt{2m + \frac{9}{4}}. \tag{8}$$

Since the identification of r and $\boldsymbol{\Lambda}$ from $\boldsymbol{\beta}$ by Theorem 1 relies on identifying zero and non-zero elements in $\boldsymbol{\beta}$, we follow previous work on parsimonious factor modeling and consider the selection of the elements of $\boldsymbol{\beta}$ as a variable selection problem. To this aim, we introduce for each element β_{ij} a binary indicator δ_{ij} and define β_{ij} to be 0, if $\delta_{ij} = 0$, and leave β_{ij} unconstrained otherwise. In this way, we obtain an indicator matrix $\boldsymbol{\delta}$ of the same dimension as $\boldsymbol{\beta}$.

We use a Bayesian approach under priors to be described in Section 3 using MCMC methods to be described in Section 4. We search the space of all possible regression-type models obeying condition **B1** by sampling the indicator matrix $\boldsymbol{\delta}$ simultaneously with the remaining parameters. Note that condition **B1** which is a condition on $\boldsymbol{\beta}$ is fulfilled iff the corresponding indicator matrix $\boldsymbol{\delta}$ obeys Condition **B1**. Hence, Theorem 1 could be applied immediately to $\boldsymbol{\delta}$ to identify overfitting factor models. Furthermore, $\boldsymbol{\delta}$ is invariant to sign switching in the columns of $\boldsymbol{\beta}$.

As in previous work, our approach allows the identification of a parsimonious structure in the factor loading matrix in cases where this matrix is sparse. In addition to that, Theorem 1 allows the identification of the true number of factors r directly from the indicator matrix $\boldsymbol{\delta}$:

$$r = \sum_{j=1}^{k} I\{\sum_{i=1}^{m} \delta_{ij} > 1\}, \tag{9}$$

where $I\{\cdot\}$ is the indicator function, by taking spurious factors into account. Note that (9) is invariant to permuting the columns of $\boldsymbol{\delta}$ which is helpful for MCMC based inference with respect to r. Finally, our approach provides a principled way for inference on r, as opposed to previous work which are based on rather heuristic procedures to infer this quantity (Carvalho et al., 2008; Bhattacharya & Dunson, 2009).

3 Choosing Priors for Bayesian Inference

Since we perform Bayesian inference in the regression-type representation (6), we formulate a joint prior for the model indicator matrix $\boldsymbol{\delta}$, the variances $\sigma_1^2, \ldots, \sigma_m^2$, and the coefficient matrix $\boldsymbol{\beta}$ taking the form $p(\boldsymbol{\delta}, \sigma_1^2, \ldots, \sigma_m^2, \boldsymbol{\beta}) = p(\boldsymbol{\delta})p(\sigma_1^2, \ldots, \sigma_m^2)p(\boldsymbol{\beta}|\boldsymbol{\delta}, \sigma_1^2, \ldots, \sigma_m^2)$.

3.1 The Prior of the Indicator Matrix

To define $p(\boldsymbol{\delta})$, we use a hierarchical prior which allows different occurrence probabilities $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_k)$ of non-zero elements in the different columns of $\boldsymbol{\beta}$ and assume that all indicators are independent *a priori* given $\boldsymbol{\tau}$:

$$\Pr(\delta_{ij} = 1|\tau_j) = \tau_j, \qquad \tau_j \sim \mathcal{B}(a_0, b_0). \tag{10}$$

It is interesting to derive the corresponding prior $p(r|a_0, b_0, m, k)$ of the number factors r which depends not only on a_0 and b_0, but also on m and k. Due to column invariance in (9) we obtain that *a priori* r may be represented as $r = \sum_{j=1}^{k} I\{X_j > 1\}$, where $X_1, \ldots, X_k$ are independent random variables and X_j follows a Beta-binomial distribution with parameters $N_j = m - j + 1$ (i.e. the number of free elements in column j), a_0, and b_0. This result could be used to simulate $p(r|a_0, b_0, m, k)$, although it is possible, but tedious, to work out the prior probabilities analytically. Uniform priors for τ_j, i.e. $a_0 = b_0 = 1$, are not necessarily a good choice, because they lead to a more or less uniform prior of r over $\{1, \ldots, k\}$, in particular, if m is large. We recommend to tune for a given data set with known values m and k the hyperparameters a_0 and b_0 in such a way that $p(r|a_0, b_0, m, k)$ is in accordance with prior expectations concerning the number of factors, see Table 2.

Carvalho et al. (2008) introduced a seemingly more flexible prior in the context of sparse factor models where $\Pr(\delta_{ij} = 1|\tau_{ij}) = \tau_{ij}$, $\tau_{ij}|\rho_j \sim (1 - \rho_j)\mathcal{I}_0 + \rho_j \mathcal{B}(a_j m_j, a_j(1 - m_j))$, $\rho_j \sim \mathcal{B}(s\rho_0, s(1 - \rho_0))$, and $\mathcal{I}_0$ is a Dirac measure at 0. However, it is easy to verify that conditional on $\rho_1, \ldots, \rho_k$ all indicators are independent *a priori* with $\Pr(\delta_{ij} = 1|\rho_j) = \rho_j m_j$. Hence the only difference to prior (10) is that $\tau_j = \rho_j m_j$ is constrained *a priori* to lie in $[0, m_j]$ with m_j being a known hyperparameter rather in $[0, 1]$.

3.2 The Prior on the Idiosyncratic Variances

When estimating factor models using classical statistical methods, such as maximum likelihood (ML) estimation, it frequently happens that the optimal solution lies outside the admissible parameter space with one or more of the idiosyncratic variances σ_i^2s being negative, see e.g. Bartholomew (1987, Section 3.6). An empirical study in Jöreskog (1967) involving 11 data sets revealed that such improper solutions are quite frequent. This difficulty became known as Heywood problem and is likely to happen for samples with either T or m being small, for data where the true σ_i^2s are very unbalanced and for overfitting models where fitting more factors than present leads to an inflation of the communalities R_i^2 defined in (5), forcing σ_i^2 toward 0.

The introduction of a prior for each of the idiosyncratic variances σ_i^2s within a Bayesian framework naturally avoids negative values for σ_i^2, nevertheless there exists a Bayesian analogue of the Heywood problem which takes the form of multi-modality of the posterior of σ_i^2 with one mode lying at 0. Subsequently, the prior on the idiosyncratic variances $\sigma_1^2, \ldots, \sigma_m^2$ is selected

in such a way that Heywood problems are avoided. Heywood problems typically occur, if the constraint

$$\frac{1}{\sigma_i^2} \geq (\mathbf{\Omega}^{-1})_{ii} \quad \Leftrightarrow \quad \sigma_i^2 \leq \frac{1}{(\mathbf{\Omega}^{-1})_{ii}} \tag{11}$$

is violated, where the matrix $\mathbf{\Omega}$ is the covariance matrix of $\mathbf{y}_t$ defined in (4), see e.g. Bartholomew (1987, p. 54).

It is clear from inequality (11) that $1/\sigma_i^2$ has to be bounded away from 0. Therefore, improper priors on the idiosyncratic variances such as $p(\sigma_i^2) \propto 1/\sigma_i^2$, which have been used by several authors (Martin & McDonald, 1975; Akaike, 1987) are not able to prevent Heywood problems. We assume instead a proper inverted Gamma prior $\sigma_i^2 \sim \mathcal{G}^{-1}(c_0, C_{i0})$ for each of the idiosyncratic variances σ_i^2. First, we choose the number of degrees of freedom c_0 large enough to bound the prior away from 0, typically $c_0 = 2.5$. A prior with $c_0 = 1.1$ as in Lopes & West (2004) allows values too close to 0. Second, we reduce the occurrence probability of a Heywood problem which is equal to $\Pr(X \leq C_{i0}(\mathbf{\Omega}^{-1})_{ii})$ where $X \sim \mathcal{G}(c_0, 1)$ through the choice of C_{i0}. This probability decreases with C_{i0}, however, a downward bias may be introduced, if C_{i0} is too small, since $\mathrm{E}(\sigma_i^2) = C_{i0}/(c_0 - 1)$. We suggest to choose C_{i0} as the largest value for which the upper bound in (11) is fulfilled by the prior expectation $\mathrm{E}(\sigma_i^2)$:

$$C_{i0}/(c_0 - 1) \leq \frac{1}{(\mathbf{\Omega}^{-1})_{ii}}.$$

If $(\mathbf{\Omega}^{-1})_{ii}$ is estimated by the ith diagonal element of the inverse of the sample covariance matrix $\mathbf{S}_y$, this yields the following prior:

$$\sigma_i^2 \sim \mathcal{G}^{-1}\left(c_0, (c_0 - 1)/(\mathbf{S}_y^{-1})_{ii}\right). \tag{12}$$

Because $1 - R_i^2 = \sigma_i^2/\Omega_{ii}$, inequality (11) introduces an upper bound for $1 - R_i^2$ which is considerably smaller than 1 for small idiosyncratic variances σ_i^2. Hence, our prior is particularly sensible, if the communalities R_i^2 are rather unbalanced across variables and the variance of some observations is very well-explained by the common factors, while this is not the case for other variables.

If the idiosyncratic variances are not too unbalanced, we found it also useful to consider a hierarchical prior, where $C_{i0} \equiv C_0$ and C_0 is equipped with a $\mathcal{G}(g_0, G_0)$ prior with g_0 being a small integer, e.g. $g_0 = 5$. This prior allows for more shrinkage than the previous one. Once again we use the upper bound defined in (11) to choose G_0. We assume that the expected mode of the prior of σ_i^2 which is given by $\mathrm{E}(g_0/C_0) = g_0/c_0/G_0$ is smaller than the average of the upper bounds defined in (11), thus

$$\sigma_i^2 \sim \mathcal{G}^{-1}(c_0, C_0), \qquad C_0 \sim \mathcal{G}(g_0, G_0), \qquad G_0 = \frac{g_0}{c_0 \sum_{i=1}^{m} 1/(\mathbf{S}_y^{-1})_{ii}}. \tag{13}$$

Our case studies illustrate that these priors usually lead to unimodal posterior densities for the idiosyncratic variances.

3.3 The Prior on the Factor Loadings

We assume that the rows of the coefficient matrix $\boldsymbol{\beta}$ are independent *a priori* given the factors $\mathbf{f}_1, \ldots, \mathbf{f}_T$. Let $\boldsymbol{\beta}_{i\cdot}^{\boldsymbol{\delta}}$ be the vector of unconstrained elements in the ith row of $\boldsymbol{\beta}$ corresponding to $\boldsymbol{\delta}$. For each $i = 1, \ldots, m$, we assume that

$$\boldsymbol{\beta}_{i\cdot}^{\boldsymbol{\delta}} | \sigma_i^2 \sim N\left(\mathbf{b}_{i0}^{\boldsymbol{\delta}}, \mathbf{B}_{i0}^{\boldsymbol{\delta}} \sigma_i^2\right). \tag{14}$$

This choice allows the posterior of β_{ij} to be centered around 0, if the factor model is overfitting, while some authors (Lopes & West, 2004) assumed truncated normal priors for the diagonal elements of $\boldsymbol{\beta}$. The variance of the prior (14) depends on σ_i^2, because this allows joint drawing of $\boldsymbol{\beta}$ and $\sigma_1^2, \ldots, \sigma_m^2$ and, even more importantly, sampling the model indicators $\boldsymbol{\delta}$ without conditioning on the model parameters in the MCMC scheme to be discussed in Section 4.

The prior moments are either chosen as in Lopes & West (2004) and Ghosh & Dunson (2009), who considered a "unit scale" prior where $\mathbf{b}_{i0}^{\boldsymbol{\delta}} = \mathbf{0}$ and $\mathbf{B}_{i0}^{\boldsymbol{\delta}} = \mathbf{I}$. Alternatively, we use a fractional prior (O'Hagan, 1995) which was applied by several authors for variable selection in latent variable models (Smith & Kohn, 2002; Frühwirth-Schnatter & Tüchler, 2008; Tüchler, 2008).

The fractional prior can be interpreted as the posterior of a non-informative prior and a fraction b of the data. In the present context, we consider a conditionally fractional prior for the "regression model"

$$\tilde{\mathbf{y}}_i = \mathbf{X}_i^{\boldsymbol{\delta}} \boldsymbol{\beta}_{i\cdot}^{\boldsymbol{\delta}} + \tilde{\boldsymbol{\epsilon}}_i, \tag{15}$$

where $\tilde{\mathbf{y}}_i = (y_{i1} \cdots y_{iT})'$ and $\tilde{\boldsymbol{\epsilon}}_i = (\epsilon_{i1} \cdots \epsilon_{iT})'$. $\mathbf{X}_i^{\boldsymbol{\delta}}$ is a regressor matrix for $\boldsymbol{\beta}_{i\cdot}^{\boldsymbol{\delta}}$ constructed from the latent factor matrix $\mathbf{F} = (\mathbf{f}_1 \cdots \mathbf{f}_T)'$ in the following way. If no element in row i of $\boldsymbol{\beta}$ is restricted to 0, then $\mathbf{X}_i^{\boldsymbol{\delta}} = \mathbf{F}$. If some elements are restricted to 0, then $\mathbf{X}_i^{\boldsymbol{\delta}}$ is obtained from $\mathbf{F}$ by deleting all columns j where $\delta_{ij} = 0$, i.e. $\mathbf{X}_i^{\boldsymbol{\delta}} = \mathbf{F}\boldsymbol{\Pi}_i^{\boldsymbol{\delta}}$, where $\boldsymbol{\Pi}_i^{\boldsymbol{\delta}}$ is a $k \times \sum_{j=1}^{k} \delta_{ij}$ selection matrix, selecting those columns j of $\mathbf{F}$ where $\delta_{ij} \neq 0$.

Using $p(\boldsymbol{\beta}_{i\cdot}^{\boldsymbol{\delta}} | \sigma_i^2) \propto p(\tilde{\mathbf{y}}_i | \boldsymbol{\beta}_{i\cdot}^{\boldsymbol{\delta}}, \sigma_i^2)^b$ we obtain from regression model (15):

$$\boldsymbol{\beta}_{i\cdot}^{\boldsymbol{\delta}} | \sigma_i^2 \sim N\left(\mathbf{b}_{iT}, \mathbf{B}_{iT} \sigma_i^2 / b\right), \tag{16}$$

where $\mathbf{b}_{iT}$ and $\mathbf{B}_{iT}$ are the posterior moments under an non-informative prior:

$$\mathbf{B}_{iT} = \left((\mathbf{X}_i^{\boldsymbol{\delta}})' \mathbf{X}_i^{\boldsymbol{\delta}}\right)^{-1}, \qquad \mathbf{b}_{iT} = \mathbf{B}_{iT} (\mathbf{X}_i^{\boldsymbol{\delta}})' \tilde{\mathbf{y}}_i. \tag{17}$$

It is not entirely clear how to choose the fraction b for a factor model. If the regressors $\mathbf{f}_1, \ldots, \mathbf{f}_T$ were observed, then we would deal with m independent regression models for each of which T observations are available and the choice $b = 1/T$ would be appropriate. The factors, however, are latent and are estimated together with the other parameters. This ties the m regression models together. If we consider the multivariate regression model as a whole, then the total

number $N = mT$ of observations has to be taken into account which motivates choosing $b_N = 1/(Tm)$. In cases where the number of regressors d is of the same magnitude as the number of observations, Ley & Steel (2009) recommend to choose instead the risk inflation criterion $b_R = 1/d^2$ suggested by Foster & George (1994), because b_N implies a fairly small penalty for model size and may lead to overfitting models. In the present context this implies choosing $b_R = 1/d(k,m)^2$ where $d(k,m) = (km - k(k-1)/2)$ is the number of free elements in the coefficient matrix $\boldsymbol{\beta}$. This second criterion leads to a stronger penalty than b_N if $T < d(k,m)^2/m$.

4 MCMC Estimation

4.1 The MCMC Scheme

MCMC estimation including model specification search is implemented in the following way. We choose a lower triangular matrix with r non-zero columns as starting value for $\boldsymbol{\delta}$, and set each element, except the diagonal elements, to 0 with probability 50%. If $k \leq 10$, we sample the initial value r uniformly from $\{1, \ldots, k\}$, otherwise we sample r from $\min(\max(\mathcal{P}(r_0), 1), k)$, where $\mathcal{P}(r_0)$ is a Poisson distribution with mean r_0.

We choose starting values for the factors $\mathbf{f} = (\mathbf{f}_1, \ldots, \mathbf{f}_T)$ by sampling from the prior (2) and repeat the following sampling steps:

(a) Sample the indicator matrix $\boldsymbol{\delta}$ conditional on the factors $\mathbf{f}_1, \ldots, \mathbf{f}_T$ without conditioning on the model parameters $\boldsymbol{\beta}$ and $\sigma_1^2, \ldots, \sigma_m^2$ from $p(\boldsymbol{\delta}|\mathbf{f}_1, \ldots, \mathbf{f}_T, \boldsymbol{\tau}, \mathbf{y})$:

  (a-1) Try to turn all non-zero columns of $\boldsymbol{\delta}$ into zero columns.

  (a-2) Update the indicators jointly for each of the remaining non-zero columns j of $\boldsymbol{\delta}$:

    (a-2-1) try to move the top non-zero element l_j;

    (a-2-2) update the indicators in the rows $l_j + 1, \ldots, m$.

  (a-3) Try to turn all (or at least some) zero columns of $\boldsymbol{\delta}$ into a non-zero column.

(b) Sample the model parameters $\boldsymbol{\beta}$ and $\sigma_1^2, \ldots, \sigma_m^2$ jointly conditional on the indicator matrix $\boldsymbol{\delta}$ and the factors $\mathbf{f}_1, \ldots, \mathbf{f}_T$ from $p(\boldsymbol{\beta}, \sigma_1^2, \ldots, \sigma_m^2|\boldsymbol{\delta}, \mathbf{f}_1, \ldots, \mathbf{f}_T, \mathbf{y})$.

(c) Sample the latent factors $\mathbf{f}_1, \ldots, \mathbf{f}_T$ conditional on the model parameters $\boldsymbol{\beta}$ and $\sigma_1^2, \ldots, \sigma_m^2$ from $p(\mathbf{f}_1, \ldots, \mathbf{f}_T|\boldsymbol{\beta}, \sigma_1^2, \ldots, \sigma_m^2, \mathbf{y})$.

(d) Perform an acceleration step by moving temporarily to an expanded factor model.

(e) For each $j = 1, \ldots, k$, perform a random sign switch: substitute the draws of $\{f_{jt}\}_{t=1}^T$ and $\{\beta_{ij}\}_{i=j}^m$ with probability 0.5 by $\{-f_{jt}\}_{t=1}^T$ and $\{-\beta_{ij}\}_{i=j}^m$, otherwise leave these draws unchanged.

(f) Sample τ_j for $j = 1, \ldots, k$ from $\tau_j | \boldsymbol{\delta} \sim \mathcal{B}(a_0 + d_j, b_0 + p_j)$, where p_j is the number of free elements and $d_j = \sum_{i=1}^{m} \delta_{ij}$ is number of non-zero elements in column j.

To generate sensible values for the latent factors in the initial model specification, we found it useful to run the first few steps without variable selection.

Updating of the model indicators in Step (a) is done in a very efficient manner, by sampling all indicators in the same column simultaneously, see Subsection 4.1.1. Step (b) and (c) could be implemented as in Lopes & West (2004), however, we make several improvements. Using the full conditional posteriors developed in Appendix B.1, the parameters $\boldsymbol{\beta}_{i\cdot}$ and σ_i^2 could be sampled in Step (b) row by row, which might be slow if m is large. Instead, we show in Appendix B.2 that joint sampling of all idiosyncratic variances and all non-zero factor loadings is feasible.

In Step (c) a simplification is possible, as usually some columns of the coefficient matrix $\boldsymbol{\beta}$, say $l_1, \ldots, l_{k-r}$, are equal to zero. In this case, $f_{l_j,t}$ is sampled from $N(0,1)$ for $j = 1, \ldots, k-r$ and $t = 1, \ldots, T$, because the posterior of the latent factors $f_{l_1,t}, \ldots, f_{l_{k-r},t}$ is equal to the prior. The posterior of the remaining components $\mathbf{f}_t^1 = (f_{j_1,t}, \ldots, f_{j_r,t})$ is given by:

$$\mathbf{f}_t^1 | \mathbf{y}_t, \boldsymbol{\beta}, \boldsymbol{\Sigma} \sim N_r \left((\mathbf{I}_r + \boldsymbol{\beta}_1' \boldsymbol{\Sigma}^{-1} \boldsymbol{\beta}_1)^{-1} \boldsymbol{\beta}_1' \boldsymbol{\Sigma}^{-1} \mathbf{y}_t, (\mathbf{I}_r + \boldsymbol{\beta}_1' \boldsymbol{\Sigma}^{-1} \boldsymbol{\beta}_1)^{-1} \right), \tag{18}$$

where $\boldsymbol{\beta}_1$ is the $m \times r$ matrix containing the non-zero columns of $\boldsymbol{\beta}$. Step (d) is added to improve mixing, see Subsection 4.1.2 for details.

As described in Subsection 2.2 we loose identifiability of the signs of the elements of $\mathbf{f}_t$ and $\boldsymbol{\beta}$ with our approach. For each $j = 1, \ldots, k$ all elements $f_{j1}, \ldots, f_{jT}$ of the latent factor as well as the jth column of $\boldsymbol{\beta}$ may be changed by the same sign switch without changing the likelihood. In combination with the prior introduced in Subsection 3.3 this leads to multimodality of the posterior density. To make sure that our sampler explores all possible modes, a random sign switch is performed in Step (e) for each column..

4.1.1 Updating the Model Indicators

To sample δ_{ij} without conditioning on $\boldsymbol{\beta}$ and $(\sigma_1^2, \ldots, \sigma_m^2)$ one has to compute the log posterior odds O_{ij} of $\delta_{ij} = 1$ versus $\delta_{ij} = 0$ while holding $\boldsymbol{\delta}_{i,-j}$ and τ_j fixed:

$$O_{ij} = \log \frac{\Pr(\delta_{ij} = 1 | \boldsymbol{\delta}_{i,-j}, \tau_j, \tilde{\mathbf{y}}_i, \mathbf{f})}{\Pr(\delta_{ij} = 0 | \boldsymbol{\delta}_{i,-j}, \tau_j, \tilde{\mathbf{y}}_i, \mathbf{f})} = \log \frac{p(\tilde{\mathbf{y}}_i | \delta_{ij} = 1, \boldsymbol{\delta}_{i,-j}, \mathbf{f})}{p(\tilde{\mathbf{y}}_i | \delta_{ij} = 0, \boldsymbol{\delta}_{i,-j}, \mathbf{f})} + \log \frac{\tau_j}{1 - \tau_j}. \tag{19}$$

The marginal likelihoods $p(\tilde{\mathbf{y}}_i | \delta_{ij}, \boldsymbol{\delta}_{i,-j}, \mathbf{f})$ may be computed individually for each element (i, j) for $\delta_{ij} = 0$ and $\delta_{ij} = 1$ as is done e.g. in Carvalho et al. (2008), however, this is very inefficient, if km is large. We show in Appendix B.3 that it is possible to compute the log posterior odds O_{ij} for all indicators in column j jointly in a very efficient manner.

A Metropolis-Hastings step could be used to sample δ_{ij}: define $\delta_{ij}^{\text{new}} = 1 - \delta_{ij}^{\text{old}}$ and accept δ_{ij}^{new} with probability

$$\min\left\{1, \frac{\Pr(\delta_{ij}^{\text{new}}|\boldsymbol{\delta}_{i,-j}, \tau_j, \tilde{\mathbf{y}}_i, \mathbf{f})}{\Pr(\delta_{ij}^{\text{old}}|\boldsymbol{\delta}_{i,-j}, \tau_j, \tilde{\mathbf{y}}_i, \mathbf{f})}\right\}.$$

Based on the posterior odd O_{ij} and a uniform random number U this reads: if $\delta_{ij}^{\text{old}} = 0$, then accept $\delta_{ij}^{\text{new}} = 1$, iff $\log U \leq O_{ij}$; if $\delta_{ij}^{\text{old}} = 1$, then accept $\delta_{ij}^{\text{new}} = 0$, iff $\log U \leq -O_{ij}$; otherwise stay with δ_{ij}^{old}. Alternatively, a Gibbs step may be used, i.e. set $\delta_{ij} = 1$, iff $\log(U/(1-U)) \leq O_{ij}$, otherwise $\delta_{ij} = 0$. Since the indicators in column j are independent given τ_j, joint sampling of $\boldsymbol{\delta}_{\cdot,j}$ is easily vectorized for both methods, given the corresponding vector of the log posterior odds. We only have to guarantee that the resulting indicator matrix satisfies condition $\mathbf{B1}$.

Step (a) in our MCMC scheme contains dimension preserving as well as dimension changing moves. Step (a-1) reduces the number of factors by reducing the number of non-zero columns. We loop randomly over all non-zero column j and move to a zero column, if $\log U \leq -\sum_i O_{ij}$, where the summation is over all rows where $\delta_{ij}^{\text{old}} = 1$. Step (a-2) is a dimension preserving move and loops randomly over all remaining non-zero columns j. In a first step, we try to move the top non-zero element l_j to a randomly selected row i_n which, according to $\mathbf{B1}$, should not be occupied by the top non-zero elements of the remaining non-zero columns. If $i_n < l_j$, then $\delta_{i_n,j}^{\text{old}} = 0$ and we accept $\delta_{i_n,j}^{\text{new}} = 1$ according to a Metropolis-Hastings move, i.e. if $\log U \leq O_{i_n,j}$. If $i_n > l_j$, then we define $\delta_{l_j,j}^{\text{new}} = 0$, $\delta_{ij}^{\text{new}} = 0$ for all rows $I_1 = \{i : l_j < i < i_n, \delta_{ij}^{\text{old}} = 1\}$ and, if $\delta_{i_n,j}^{\text{old}} = 0$, $\delta_{i_n,j}^{\text{new}} = 1$, and accept this move, if $\log U \leq -O_{l_j,j} - \sum_{i \in I_1} O_{ij} + O_{i_n,j} I\{\delta_{i_n,j}^{\text{old}} = 0\}$. Given (a possibly new) leading element l_j the indicators δ_{ij} in the rows $\{i : l_j < i \leq m\}$ are updated simultaneously in a second step.

Finally, Step (a-3) increases the number of factors by turning a zero column into a non-zero column. If $k - r$ is not too large, say smaller than 10, then we loop over all zero columns, otherwise we pick the leftmost 10 zero columns, because the particular choice does not matter due to invariance to column permutations. Two moves are made to implement this step for a particular column j. First we propose a non-zero leading element by picking randomly a row l_j which is not occupied by the top non-zero elements of all remaining non-zero column of $\boldsymbol{\delta}$ and accept this move, if $\log U \leq O_{l_j,j}$. The column remains a zero column, if this move is not accepted. Otherwise, we update the remaining indicators δ_{ij} in the rows $\{i : l_j < i \leq m\}$ simultaneously in a second step.

For any move which involves the odds of more than one row, the vector of these odds is computed simultaneously using the result of Appendix B.3.

4.1.2 Marginal Data Augmentation for Factor Models

Recently, Ghosh & Dunson (2009) introduced parameter expansion to speed up the convergence of the sampler for a basic factor model which tends to be poorly mixing. They suggest to move

to an expanded model where the factors have a more general prior distribution than in (2):

$$\tilde{\mathbf{f}}_t \sim N_k\left(\mathbf{0}, \boldsymbol{\Psi}\right), \tag{20}$$

$$\mathbf{y}_t = \tilde{\boldsymbol{\beta}}\tilde{\mathbf{f}}_t + \boldsymbol{\epsilon}_t, \tag{21}$$

where $\boldsymbol{\Psi} = \text{Diag}(\Psi_1, \ldots, \Psi_k)$. They perform MCMC estimation in the expanded model by updating $\boldsymbol{\Psi}$ conditional on the remaining parameters at each sweep of the sampler. While their method successfully improves the efficiency of the sampler, it changes the prior of the parameters which is undesirable in a variable selection context where the prior strongly matters.

For this reason, we extend marginal data augmentation as discussed by van Dyk & Meng (2001) to the expanded factor model (20) and (21). To make sure that the prior remains unchanged, we assume that the working parameters $\Psi_1, \ldots, \Psi_k$ are independent *a priori* from the remaining parameters with prior $\Psi_j \sim \mathcal{G}\left(p_j, q_j\right)$.

Due to independence, we sample $\Psi_j^{\text{old}} \sim \mathcal{G}\left(p_j, q_j\right)$ to transform the standard basic factor model to the expanded version. We update Ψ_j^{new} in the expanded model by sampling from

$$\Psi_j^{\text{new}} \sim \mathcal{G}\left(p_j + \frac{T}{2}, q_j + \frac{1}{2\Psi_j^{\text{old}}}\sum_{t=1}^{T} f_{tj}^2\right), \tag{22}$$

and transform back to the original model. This leads to following acceleration in Step (d) which affects $\boldsymbol{\beta}$ and $\mathbf{f}_t$ for each $j = 1, \ldots, k$ in the following way:

$$\beta_{ij} = \beta_{ij}\sqrt{\frac{\Psi_j^{\text{old}}}{\Psi_j^{\text{new}}}}, \qquad i = 1, \ldots, m,$$

$$f_{tj} = f_{tj}\sqrt{\frac{\Psi_j^{\text{new}}}{\Psi_j^{\text{old}}}}, \qquad t = 1, \ldots, T.$$

4.2 Bayesian Posterior Inference Using the MCMC Draws

The posterior draws of $\boldsymbol{\delta}$, $\sigma_1^2, \ldots, \sigma_m^2$, and $\boldsymbol{\beta}$ obtained by the MCMC scheme described in Subsection 4.1 are draws for the regression-type representation (6) and have to be used in a careful way for posterior inference, because the position of the columns of $\boldsymbol{\delta}$ and $\boldsymbol{\beta}$ as well as the sign of the columns of $\boldsymbol{\beta}$ are not identified. Furthermore, the model may be overfitting in the sense of Theorem 1 which affects not only $\boldsymbol{\delta}$ and $\boldsymbol{\beta}$, but also some of the idiosyncratic variances.

Hence, for any particular functional of the posterior draws, it has to be investigated if this functional is affected by any of these identifiability problems. The number of factors r, for instance, which has been defined in (9) as a function of $\boldsymbol{\delta}$ is invariant to column permutations and sign switching and takes care of overfitting by definition, hence posterior draws for r are

obtained immediately from $\boldsymbol{\delta}$ by applying (9). To estimate the number r of factors for a particular data set, we may use the posterior mode $\tilde{r}$ of $p(r|\mathbf{y})$, where $p(r|\mathbf{y})$ is estimated by the frequency distribution obtained from the MCMC draws.

Also the model size d, defined as the number of non-zero elements in $\boldsymbol{\Lambda}$, i.e.

$$d = \sum_{j=1}^{k} d_j I\{d_j > 1\}, \qquad d_j = \sum_{j=1}^{m} \delta_{ij}. \tag{23}$$

is unaffected by any of these identifiability problems. Model size could be estimated by the posterior mode $\tilde{d}$ or the posterior mean $\mathrm{E}(d|\mathbf{y})$ of $p(d|\mathbf{y})$.

Posterior draws for the parameter $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Lambda}$ in the factor model (1) as well as for $\boldsymbol{\delta}^{\Lambda}$, where δ_{ij}^{Λ} is an indicator matrix for $\boldsymbol{\Lambda}$, are obtained from $\boldsymbol{\delta}$, $\sigma_1^2, \ldots, \sigma_m^2$, and $\boldsymbol{\beta}$ through various identifications steps. First, we use Theorem 1 to handle potential overfitting and define $\boldsymbol{\Sigma}_0 = \mathrm{Diag}(\sigma_1^2, \ldots, \sigma_m^2)$. For each column j of $\boldsymbol{\delta}$ with only a single non-zero element in row l_j, add $\beta_{l_j,j}^2$ to the idiosyncratic variance of feature l_j, i.e. $(\boldsymbol{\Sigma}_0)_{l_j,l_j} = (\boldsymbol{\Sigma}_0)_{l_j,l_j} + \beta_{l_s,j}^2$ and set the jth column of $\boldsymbol{\delta}$ and $\boldsymbol{\beta}$ equal to a zero column. $\boldsymbol{\Lambda}$ and $\boldsymbol{\delta}^{\Lambda}$ are then recovered from $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ by deleting all zero columns. Since the resulting matrices only obey condition **B1**, the position of the columns of $\boldsymbol{\Lambda}$ and $\boldsymbol{\delta}^{\Lambda}$ is not unique.

Nevertheless, certain functionals of $\boldsymbol{\delta}^{\Lambda}$ are invariant to column switching such as the number N_v of distinct models $\boldsymbol{\delta}^{\Lambda}$ visited by the search procedure and a ranking of these models in terms of their frequency. Model selection could be based on the highest probability model (HPM) which is the model defined by the indicator matrix $\boldsymbol{\delta}_H^{\Lambda}$ visited most often and its frequency p_H which may be regarded as an estimator of the posterior probability of this model. An alternative estimator of the number of factors r is given by the number r_H of non zero columns in the HPM and an alternative estimator for model size is given by the number d_H of non-zero elements in the HPM.

For further inference, we have to handle the column switching problem. To this aim, we sort the columns of $\boldsymbol{\Lambda}$ and $\boldsymbol{\delta}^{\Lambda}$ in such a way that the resulting row indices $\mathbf{l} = (l_1, \ldots, l_r)$ of the top non-zero elements are increasing and satisfy condition **C2**. However, column switching might still be present among the draws of $\boldsymbol{\Lambda}$ and $\boldsymbol{\delta}^{\Lambda}$. This happens, for instance, if the draws of $\mathbf{l}$ switch between $\mathbf{l}_1 = (1, 2, 3)$ and $\mathbf{l}_2 = (1, 3, 4)$. In this case, it is certain that 3 factors are present, however, only for two columns is the position of the top non-zero element certain, namely row 1 and row 3, while for the third column it is not certain whether the element in the second row is different from 0. Ordering the columns according to **C2** leads to column switching, because the third column for all draws where $\mathbf{l}_2$ holds corresponds to the second column for all draws where $\mathbf{l}_1$ holds.

A final step is needed to identify the position of the columns of $\boldsymbol{\Lambda}$ and $\boldsymbol{\delta}^{\Lambda}$, if $\mathbf{l}$ switches during MCMC estimation, which is typically the case, because the number of factors r changes as well between MCMC draws. To this aim, we determine the identifiability constraint $\mathbf{l}^{\star} = (l_1^{\star}, \ldots, l_{r_M}^{\star})$ visited most often. For all MCMC draws corresponding to $\mathbf{l}^{\star}$, i.e. the subsequence

$\boldsymbol{\Lambda}^\star$ and $(\boldsymbol{\delta}^\Lambda)^\star$ where $\mathbf{l}^{(m)} = \mathbf{l}^\star$ the positions of the columns is unique. Note that the MCMC draws $\boldsymbol{\Lambda}^\star$ and $(\boldsymbol{\delta}^\Lambda)^\star$ are draws from the conditional posterior of a r_M-factor model obeying Condition **C2** with the corresponding sequence $\mathbf{l}^\star$. We may then estimate for each indicator the marginal inclusion probability $\Pr(\delta_{ij}^\Lambda = 1|\mathbf{y}, \mathbf{l}^\star)$ under $\mathbf{l}^\star$ as the average over the elements of $(\boldsymbol{\delta}^\Lambda)^\star$. Note that $\Pr(\delta_{l_j,j}^\Lambda = 1|\mathbf{y}, \mathbf{l}^\star) = 1$ for $j = 1, \ldots, r_M$. Following Scott & Berger (2006), we determine the median probability model (MPM) by setting the indicators δ_{ij}^Λ in $\boldsymbol{\delta}_M^\Lambda$ to 1 iff $\Pr(\delta_{ij}^\Lambda = 1|\mathbf{y}, \mathbf{l}^\star) \geq 0.5$. The number of non-zero top elements r_M in the identifiability constraint $\mathbf{l}^\star$ is a third estimator of the number of factors, while the number d_M of non-zero elements in the MPM is yet another estimator of model size.

A discrepancy between the various estimators of the number of factors r is often a sign of a weakly informative likelihood and it might help to use a more informative prior for $p(r)$ by choosing the hyperparameters a_0 and b_0 accordingly. Also the structure of the indicator matrices $\boldsymbol{\delta}_H^\Lambda$ and $\boldsymbol{\delta}_M^\Lambda$ corresponding, respectively, to the HPM and the MPM may differ, in particular if the frequency p_H is small and some of the inclusion probabilities $\Pr(\delta_{ij}^\Lambda = 1|\mathbf{y}, \mathbf{l}^\star)$ are close to 0.5.

Finally, we have to address sign switching for $\boldsymbol{\Lambda}^\star$, because the non-zero factor loadings Λ_{ij} are identified only up to sign switching. Quantities which are invariant to sign switching are the communalities R_{ij}^2 defined in (5) and, of course, $(\boldsymbol{\delta}^\Lambda)^\star$. We handle sign switching in the posterior draws of $\boldsymbol{\Lambda}^\star$ in the following way. To obtain identification of the sign of the jth column, we impose the constraint $\Lambda_{l_j,j} > 0$ for all l_j. We used the top non-zero element in each column for identification, however any element $\Lambda_{r_j,j}$ with a high probability of being non-zero will do. The most efficient way to force these constraints is post-processing the posterior draws of $\boldsymbol{\Lambda}^\star$: for each draw, perform a sign switch similar to Step (e) in the MCMC scheme for any column j, where $\Lambda_{l_j,j} < 0$, and leave the signs unchanged, otherwise.

It may be of interest to perform posterior inference conditional on an arbitrary estimator $\hat{r}$ and an arbitrary constraint **C2** defined by $\mathbf{l} = (l_1, \ldots, l_{\hat{r}})$. To this aim, the MCMC scheme presented in Subsection 4.1 is started with $k = \hat{r}$ and an indicator matrix $\boldsymbol{\delta}$ obeying **C2**. During MCMC estimation, Step (a-1) and (a-3) are skipped to keep k fixed and Step (a-2-1) is skipped to hold $(l_1, \ldots, l_k)$ fixed. However, Step (a-2-2) is still performed to identify a parsimonious structure for the remaining elements of the factor loading matrix. The columns of the resulting posterior draws of $\boldsymbol{\Lambda}$ and $\boldsymbol{\delta}^\Lambda$ are identified up to sign switching. This minor identifiability problem is solved as above through the constraint $\Lambda_{l_j,j} > 0$, $j = 1, \ldots, k$. The posterior draws of $\boldsymbol{\delta}^\Lambda$ could be used to determine a k-factor HPM and a k-factor MPM obeying **C2**.

5 Illustrative Applications

We consider both simulated as well as real data to evaluate our procedure. We choose k equal to the maximum number of possible factor given by inequality (8), see Table 1, and tuned the hyperparameters a_0 and b_0 for each case study in order to match the prior $p(r|a_0, b_0, k, m)$ to

Table 1: Number of features m, number of observations T, maximum number of possible factors k, burn-in M_0, number of posterior draws M, and runtime in CPU minutes (averaged over priors) for the various case studies.

Data	m	k	T	M_0	M	CPU
Simulated	9	6	50/150/500	10,000	10,000	2.3/3.5/3.9
Exchange rate	6	3	143	10,000	50,000	5.7
Maxwell's – neurotic	10	6	148	10,000	60,000	9.2
Maxwell's – normal	10	6	810	30,000	300,000	74.2
Applicants	15	10	48	20,000	150,000	33.9

Table 2: Hyperparameters a_0 and b_0 and the corresponding prior distribution $p(r|a_0, b_0, k, m)$ (not reported for $r = 0$) for the various case studies

			r						
Data	a_0	b_0	1	2	3	4	5	6	7 - 10
Simulated	0.2	0.3	0.098	0.242	0.312	0.228	0.089	0.015	-
Exchange rate	1	1	0.228	0.448	0.286	-	-	-	-
Applicants	0.3	1.2	0.045	0.132	0.226	0.250	0.192	0.101	0.046
Maxwell's	0.3	0.7	0.137	0.280	0.304	0.185	0.060	0.008	-

prior expectations concerning the number of factors, see Table 2.

We study sensitivity of factor selection with respect to prior choices. We combine our new prior on the idiosyncratic variances where $c_0 = 2.5$ and C_{0i} is selected to avoid a Heywood problem with various fractional priors, namely $b = b_N$, and, if $T < d(k, m)^2/m$, $b = b_R$ as well as $b = 10^{-p}$ where p is a small integer. In addition, we study the conventional unit scale prior/inverted Gamma prior where $\mathbf{b}_{i0}^{\boldsymbol{\delta}} = \mathbf{0}$, $\mathbf{B}_{i0}^{\boldsymbol{\delta}} = \mathbf{I}$, and, respectively $c_0 = 1.1$ and $C_{i0} = 0.055$ (Lopes & West, 2004, LW) and $c_0 = 1$ and $C_{i0} = 0.2$ (Ghosh & Dunson, 2009, GD).

We run the MCMC scheme described in Section 4 for M iterations after a burn-in phase of M_0 draws using parameter expansion based on the Gamma prior $p_j = q_j = 0.1$. We use the inefficiency factor τ_d of the model size d defined in (23) to evaluate mixing of our MCMC scheme, see also Table 4, 7, and 11, and adjust M accordingly. The MCMC setting and the run times, averaged over priors, are summarized in Table 1. All implementation are carried out using MATLAB (Version 7.3.0) on a notebook with a 2.0 GHz processor. During the first 100 iterations no variable selection is performed.

5.1 Applications to Simulated Data

We reconsider a simulation study in Lopes & West (2004, Subsection 6.2), which was also reconsidered in Ghosh & Dunson (2009). The data are simulated from a three-factor model with 9 variables, where a considerable fraction of the factor loadings is zero and the idiosyncratic variances are rather unbalanced:

$$\mathbf{\Lambda}' = \begin{pmatrix} 0.99 & 0 & 0 & 0.99 & 0.99 & 0 & 0 & 0 & 0 \\ 0 & 0.95 & 0 & 0 & 0 & 0.95 & 0.95 & 0 & 0 \\ 0 & 0 & 0.9 & 0 & 0 & 0 & 0 & 0.9 & 0.9 \end{pmatrix},$$

$$(\sigma_1^2, \ldots, \sigma_9^2) = \begin{pmatrix} 0.02 & 0.19 & 0.36 & 0.02 & 0.02 & 0.19 & 0.19 & 0.36 & 0.36 \end{pmatrix}.$$

We consider samples of size $T = 50$, $T = 150$, and $T = 500$, and simulate 50 data sets for each T. Table 3 reports for each sample size T and for each prior the fraction of data set among the 50 simulated data set, for which the various estimators of the number of factors, namely $\tilde{r}$, r_H, and r_M were equal to the true value $r = 3$, and for which $\mathbf{l}^\star$, the highest posterior identifiability constraint, was equal to the true constraint **C2** with $l_1 = 1$, $l_2 = 2$, and $l_3 = 3$. In addition, the table summarizes the averages of the various estimators of model size, namely $\tilde{d}$, d_H and d_M in order to evaluate, if the true model size which is equal to $d = 9$ is over- or under estimated.

In most of the cases, the correct number of factors and the true identifiability constraint is identified for all simulated data set, however the prior is very influential on model size. The fixed scale priors used by Lopes & West (2004) and Ghosh & Dunson (2009) are overfitting the model size, even if T is large. For fractional priors, the fraction b strongly controls parsimony. Decreasing b, i.e. making the prior more vague, decreases model size and leads to more parsimonious solutions. However, if b is too small, like $b = 10^{-5}$, then the resulting model is underfitting, even if T is large, while choosing b too large, like $b = 10^{-2}$, leads to overfitting models. For large T, $b = b_N$ gives the highest hit rate for $\boldsymbol{\delta}$. For small T, b_R performs better than b_N, however, both for $T = 50$ and $T = 150$, a slightly smaller value than $\min(b_N, b_R)$, namely $b = 10^{-4}$, gives the best result.

To sum up, we find from this simulation study that our method not only selects the true number of factors, but is also able to identify the true constraint identifiability constraint and to reconstruct the finer structure of the factor loading matrix under factional priors. However, the fraction b has to be selected carefully. If b is too small, then the model will be underfitting, on the other hand, choosing b too large overfits model size. We identified $b = b_N$ and, if $T < d(k, m)^2/m$, $b = b_R$, as well as the adjacent value $b = 10^{-p}$, where p is the largest integer smaller than $-\log_{10} \min(b_N, b_R)$ as sensible values. The unit scale prior seems to be inferior to the fractional prior in connection with model selection in factor models and leads to overfitting models.

Table 3: Simulation Study ($m = 9$, true number of factors r equal to 3) – evaluating model and variable selection under different priors for 50 data sets simulated under different sample sizes T; hit rate (i.e. fraction of simulated time series where some estimator is equal to the true value) for r based on the posterior mode $\tilde{r}$, the non-zero columns r_H in the HPM and the non-zero columns r_M corresponding to the highest posterior identifiability constraint $\mathrm{l}^\star$; hit rate for the true constraint $\mathrm{l}^{\mathrm{true}} = (1, 2, 3)$ based on $\mathrm{l}^\star$; hit rate for the entire 0/1 pattern in $\boldsymbol{\delta}^\Lambda$ based on $\boldsymbol{\delta}^\Lambda_H$ corresponding to the HPM and $\boldsymbol{\delta}^\Lambda_M$ corresponding to the MPM; finally, average estimated model size d (the true model size being equal to 9) based on the posterior mode $\tilde{d}$, the HPM (d_H) and the MPM (d_M) and the average posterior frequency p_H for the HPM.

T	Prior	$\tilde{r}$	r_H	r_M	$\mathrm{l}^\star$	$\boldsymbol{\delta}^\Lambda_H$	$\boldsymbol{\delta}^\Lambda_M$	$\tilde{d}$	d_H	p_H	d_M
50	$b = 10^{-2}$	1	1	1	1	0.56	0.48	11.5	10.2	0.083	10
	$b_N = 2.2 \cdot 10^{-3}$	1	1	1	1	0.86	0.86	9.76	9.16	0.297	9.16
	$b = 10^{-3}$	1	1	1	1	0.88	0.86	9.38	9.14	0.436	9.16
	$b_R = 6.6 \cdot 10^{-4}$	1	1	1	1	0.88	0.9	9.24	9.12	0.479	9.1
	$b = 10^{-4}$	1	1	1	1	0.98	0.98	9.04	9.02	0.75	9.02
	$b = 10^{-5}$	0.96	0.96	0.96	0.98	0.96	0.96	8.88	8.88	0.885	8.88
	GD	1	1	1	1	0.66	0.54	12.2	10.7	0.059	10.3
	LW	1	1	1	1	0.52	0.4	12.7	11.4	0.053	10.9
150	$b = 10^{-2}$	1	1	1	1	0.74	0.6	11.4	10.1	0.093	9.94
	$b = 10^{-3}$	1	1	1	1	0.92	0.92	9.26	9.12	0.472	9.12
	$b_N = 7.4 \cdot 10^{-4}$	1	1	1	1	0.9	0.92	9.26	9.1	0.489	9.08
	$b_R = 6.6 \cdot 10^{-4}$	1	1	1	1	0.86	0.86	9.24	9.14	0.504	9.14
	$b = 10^{-4}$	1	1	1	1	0.94	0.96	9.08	9.06	0.742	9.04
	$b = 10^{-5}$	0.94	0.94	0.94	1	0.92	0.92	8.84	8.84	0.889	8.84
	GD	1	1	1	1	0.84	0.78	10.4	9.26	0.18	9.3
	LW	1	1	1	1	0.82	0.7	10.5	9.4	0.164	9.48
500	$b = 10^{-2}$	1	1	1	1	0.66	0.48	11.4	10	0.088	10
	$b = 10^{-3}$	1	1	1	1	0.9	0.9	9.18	9.1	0.475	9.1
	$b_R = 6.6 \cdot 10^{-4}$	1	1	1	1	0.96	0.96	9.18	9.04	0.507	9.04
	$b_N = 2.2 \cdot 10^{-4}$	1	1	1	1	0.96	0.96	9.06	9.04	0.683	9.04
	$b = 10^{-4}$	1	0.98	1	1	0.9	0.94	9.06	9.02	0.73	9.06
	$b = 10^{-5}$	0.86	0.86	0.86	0.9	0.84	0.84	8.6	8.6	0.902	8.6
	GD	1	1	1	1	0.88	0.88	9.52	9.12	0.361	9.12
	LW	1	1	1	1	0.86	0.86	9.46	9.14	0.374	9.14

Table 4: Exchange rate data; posterior distribution $p(r|\mathbf{y})$ of the number r of factors (bold number corresponding to the posterior mode $\tilde{r}$) and number of visited models N_v for various priors; frequency p_H, number of factors r_H, and model size d_H of the HPM; posterior probability $p(\mathbf{l}|\mathbf{y})$ of various identifiability constraints $\mathbf{l}$ (bold number corresponding to $\mathbf{l}^\star$); number of factors r_M and model size d_M of the MPM corresponding to $\mathbf{l}^\star$; inefficiency factor τ_d of the posterior draws of the model size d .

	$p(r\|\mathbf{y})$							$p(\mathbf{l}\|\mathbf{y})$				
Prior	1	2	3	N_v	p_H	r_H	d_H	(1,3)	(1,2)	r_M	d_M	τ_d
$b = 10^{-2}$	0	**0.799**	0.201	37	0.50	2	10	**0.496**	0.229	2	10	29.9
$b_N = 2.2 \cdot 10^{-3}$	0	**0.984**	0.016	24	0.85	2	10	**0.849**	0.122	2	10	11.9
$b = 10^{-3}$	0	**0.974**	0.026	15	0.89	2	10	**0.813**	0.139	2	10	23.0
GD	0	**0.966**	0.034	36	0.71	2	10	**0.706**	0.235	2	10	23.2
LW	0	**0.944**	0.056	37	0.69	2	10	**0.687**	0.227	2	10	17.8

5.2 Exchange Rate Data

We reanalyze the international exchange rate data studied in West & Harrison (1997, pp. 610-618) and Lopes & West (2004). The data are the changes in monthly exchange rates during the period 1/1975 to 12/1986, i.e. $T = 143$, for 6 currencies, namely US Dollar, Canadian Dollar, Japanese Yen, French Franc, Italian Lira, and Deutsche Mark. The data are standardized with respect to their mean and standard deviation. Selecting the number of factors for these data was carried out in Lopes & West (2004), using reversible jump MCMC (RJMCMC) and several approaches of computing the marginal likelihood, however, posterior inference turned out to be not very conclusive. For instance, Chib's estimator (Chib, 1995) selected a 3-factor model with probability one, while RJMCMC gave posterior probabilities of 0.88 and 0.12 for a two and a three-factor model, respectively.

Table 4 reports the posterior distribution $p(r|\mathbf{y})$ of the number r of factor derived for various priors using our new approach toward parsimonious factor modeling. Regardless of the prior, all estimators of r ($\tilde{r}$, r_H, r_M) choose a two-factor model. Depending on the prior, between 15 and 37 models were visited and the frequency p_H of the HPM varies between 50% and 89%. All priors select the constraint $\mathbf{C2}$ given by $\mathbf{l}^\star = (1, 3)$, however with differing posterior probabilities. The indicator matrix $\boldsymbol{\delta}_H^\Lambda$ corresponding to the HPM and the marginal inclusion probabilities $\Pr(\delta_{ij}^\Lambda = 1|\mathbf{y}, \mathbf{l}^\star)$ are reported for the fractional prior $b = b_N$ in Table 5.

Figure 1 shows that the posterior distributions of all idiosyncratic variances are unimodal due to the specific prior developed in Subsection 3.2. On the other hand, multimodality occurs under the improper prior $c_0 = 0$ and $C_{i0} = 0$.

To sum up, Table 4 leads to the conclusion to choose a 2-factor model and to force identifia-

Table 5: Exchange rate data; marginal inclusion posterior probabilities $\Pr(\delta_{ij}^{\Lambda} = 1|\mathbf{y}, \mathrm{l}^{\star})$ for a fractional prior with $b = b_N$ (bold elements correspond to the HPM).

	$\boldsymbol{\delta}_{\cdot 1}^{\Lambda}$	$\boldsymbol{\delta}_{\cdot 1}^{\Lambda}$
US	**1.**	0
Can	**1.**	0
Yen	**1.**	**1.**
FF	**1.**	**1.**
Lira	**1.**	**1.**
DM	**1.**	**1.**

Table 6: Exchange rate data; posterior means of the factor loading matrix, the idiosyncratic variances and the communalities for a two-factor model under the constraint **C2** with $l_1 = 1$ and $l_2 = 3$; bold numbers correspond to non-zero elements in the factor loading matrix of the 2-factor HPM.

Currency	$\mathrm{E}(\Lambda_{i1}\|\mathbf{y}, \mathrm{l}^{\star})$	$\mathrm{E}(\Lambda_{i2}\|\mathbf{y}, \mathrm{l}^{\star})$	$\mathrm{E}(\sigma_i^2\|\mathbf{y}, \mathrm{l}^{\star})$	$\mathrm{E}(R_{i1}\|\mathbf{y}, \mathrm{l}^{\star})$	$\mathrm{E}(R_{i2}\|\mathbf{y}, \mathrm{l}^{\star})$
US	**0.96**	0	0.081	91.8	0
Can	**0.951**	0	0.098	90.1	0
Yen	**0.449**	**0.418**	0.615	20.5	17.8
FF	**0.395**	**0.889**	0.053	16	78.7
Lira	**0.415**	**0.764**	0.241	17.5	58.3
DM	**0.408**	**0.765**	0.245	17	58.4

bility through the constraint $\mathrm{l}^{\star} = (1, 3)$. The constraint $\mathrm{l}^{\star}$ as well as the structure of $\boldsymbol{\delta}_H^{\Lambda}$ and $\boldsymbol{\delta}_M^{\Lambda}$ reported in Table 5 reveal that the ordering of various exchange rates is "poor" in the sense that the Canadian Dollar which is the second variable is not a suitable variable to lead the second factor, while any of the remaining exchange rates can take this place. Lopes & West (2004) came to a similar conclusion and interchanged the Canadian Dollar and the Japanese Yen in order to work with the conventional identifiability constraint on the main diagonal. Using our more general identifiability constraint **C2** reveals this information immediately and allows inference without the need to reorder the variable.

We estimate a two-factor model using conditional MCMC for the fractional prior $b = b_N$. Based on $\mathrm{l}^{\star}$, the sign of $\boldsymbol{\Lambda}$ is identified through the constraints $\Lambda_{11} > 0$ and $\Lambda_{32} > 0$. Table 6 shows the posterior means of the factor loading matrix, the idiosyncratic variances and the communalities. The bold cells in $\mathrm{E}(\boldsymbol{\Lambda}|\mathbf{y}, \mathrm{l}^{\star})$ indicate factor loadings that are non-zero according to the 2-factor HPM. As in Lopes & West (2004), we find that the first factor is a North American factor while the second factor is a European factor.

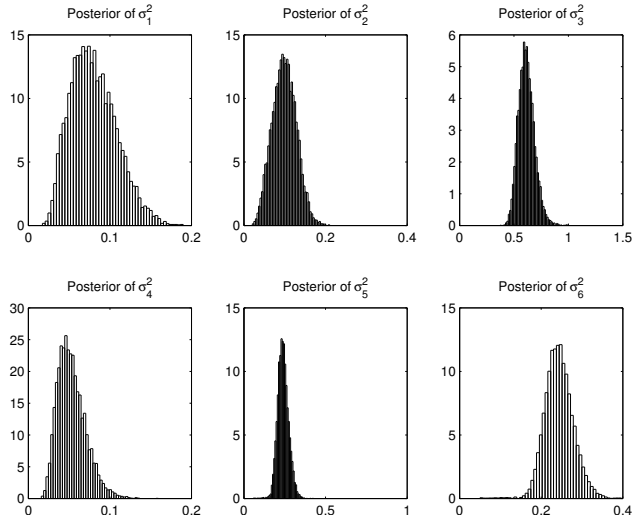

Figure 1: Exchange rate data; posterior densities $p(\sigma_i^2|\mathbf{y})$ of the idiosyncratic variances σ_i^2 for the fractional prior $b = b_N$.

Table 7: Maxwell's Children Data - neurotic children; posterior distribution $p(r|\mathbf{y})$ of the number r of factors (bold number corresponding to the posterior mode $\tilde{r}$) and highest posterior identifiability constraint $\mathbf{l}^\star$ with corresponding posterior probability $p(\mathbf{l}^\star|\mathbf{y})$ for various priors; number of visited models N_v; frequency p_H, number of factors r_H, and model size d_H of the HPM; number of factors r_M and model size d_M of the MPM corresponding to $\mathbf{l}^\star$; inefficiency factor τ_d of the posterior draws of the model size d.

	$p(r\|\mathbf{y})$					
Prior	2	3	4	5 - 6	$\mathbf{l}^\star$	$p(\mathbf{l}^\star\|\mathbf{y})$
$b = 10^{-3}$	**0.755**	0.231	0.014	0	(1,6)	0.532
$b_N = 6.8 \cdot 10^{-4}$	**0.828**	0.160	0.006	0	(1,6)	0.623
$b_R = 4.9 \cdot 10^{-4}$	**0.871**	0.127	0.001	0	(1,6)	0.589
$b = 10^{-4}$	**0.897**	0.098	0.005	0	(1,6)	0.802
GD	0.269	**0.482**	0.246	0.003	(1,2,3)	0.174
LW	0.027	0.199	**0.752**	0.023	(1,2,3,6)	0.249

Prior	N_v	p_H	r_H	d_H	r_M	d_M	τ_d
$b = 10^{-3}$	1472	0.20	2	12	2	12	30.9
$b_N = 6.8 \cdot 10^{-4}$	976	0.27	2	12	2	12	27.5
$b_R = 4.9 \cdot 10^{-4}$	768	0.34	2	12	2	12	22.6
$b = 10^{-4}$	421	0.45	2	12	2	12	18.1
GD	4694	0.06	2	15	3	19	40.6
LW	7253	0.01	4	24	4	24	32.5

Table 8: Maxwell's Children Data - normal children; posterior distribution $p(r|\mathbf{y})$ of the number r of factors (bold number corresponding to the posterior mode $\tilde{r}$) and highest posterior identifiability constraint $\mathbf{l}^\star$ with corresponding posterior probability $p(\mathbf{l}^\star|\mathbf{y})$ for various priors; number of visited models N_v; frequency p_H, number of factors r_H, and model size d_H of the HPM; number of factors r_M and model size d_M of the MPM corresponding to $\mathbf{l}^\star$; inefficiency factor τ_d of the posterior draws of the model size d.

	$p(r\|\mathbf{y})$					
Prior	3	4	5	6	$\mathbf{l}^\star$	$p(\mathbf{l}^\star\|\mathbf{y})$
$b = 10^{-3}$	0	0.391	**0.604**	0.005	(1,2,4,5,6)	0.254
$b_N = 1.2 \cdot 10^{-4}$	0	**0.884**	0.116	0	(1,2,4,5)	0.366
$b = 10^{-4}$	0	**0.891**	0.104	0.005	(1,2,4,5)	0.484
GD	0	0.396	**0.594**	0	(1,2,4,5,6)	0.229
LW	0	0.262	**0.727**	0.011	(1,2,4,5,6)	0.259

Prior	N_v	p_H	r_H	d_H	r_M	d_M	τ_d
$b = 10^{-3}$	4045	1.79	5	26	5	27	30.1
$b_N = 1.2 \cdot 10^{-4}$	1272	11.41	4	23	4	23	28.5
$b = 10^{-4}$	1296	12.17	4	23	4	23	29.1
GD	4568	1.46	5	29	5	28	32.7
LW	5387	1.56	5	28	5	28	30.7

5.3 Maxwell's Children Data

In our next example we reanalyze two data sets considered in Maxwell (1961) which are scores on 10 tests for a sample of $T = 148$ children attending a psychiatric clinic as well as a sample of $T = 810$ normal children. The first five tests are cognitive tests – (1) verbal ability, (2) spatial ability, (3) reasoning, (4) numerical ability and (5) verbal fluency. The resulting tests are inventories for assessing orectic tendencies, namely (6) neurotic questionaire, (7) ways to be different, (8) worries and anxieties, (9) interests and (10) annoyances. While psychological theory suggests that a 2-factor model is sufficient to account for the variation between the test scores, the significance test considered in Maxwell (1961) suggested to fit a 3-factor model to the first and a 4-factor model to the second data set. For the second data set, a Heywood problem is present for $k = 4$ and ML estimation leads to an improper solution with the idiosyncratic variance σ_8^2 being virtually 0. After eliminating variable $y_{8\cdot}$, Maxwell (1961) and Jöreskog (1967) fitted a three factor model to the remaining data. A reanalysis of these data in Ihara & Kano (1995) suggests that a three factor model should also be appropriate for the entire data set.

We compare these results with parsimonious factor modeling. The paper by Maxwell (1961) provides only the $m \times m$ correlation matrix for both groups of children, which is sufficient for

Table 9: Maxwell's Children Data; marginal inclusion posterior probabilities $\Pr(\delta_{ij}^{\Lambda} = 1|\mathbf{y}, \mathbf{l}^{\star})$ (bold elements correspond to the HPM); left hand side: neurotic children based on **C2** given by $l_1 = 1$, $l_2 = 6$ and the fractional prior $b = b_R$; right hand size: normal children, based on **C2** given by $l_1 = 1$, $l_2 = 2$, $l_3 = 4$, and $l_4 = 5$ and the fractional prior $b = b_N$.

	j	
Test i	1	2
1	**1**	0
2	**1**	0
3	**1**	0
4	**1**	0
5	**1**	0
6	0.313	**1**
7	**0.964**	**1**
8	0.178	**1**
9	**0.915**	**0.996**
10	0.157	**0.999**

	j			
Test i	1	2	3	4
1	**1**	0	0	0
2	**1**	**1**	0	0
3	**1**	**0.999**	0	0
4	**1**	**0.935**	**1**	0
5	**1**	0.186	0.018	**1**
6	**1**	**0.661**	0.016	**1**
7	**1**	0.308	**1**	**1**
8	**1**	0.044	0.207	**1**
9	**1**	0.197	0.036	**1**
10	**1**	0.072	**1**	**1**

Table 10: Maxwell's Children Data – neurotic children; posterior means of the factor loading matrix, the idiosyncratic variances and the communalities for a 2-factor model based on **C2** given by $l_1 = 1$, $l_2 = 6$ and the fractional prior $b = b_R$.

Test i	$\mathrm{E}(\Lambda_{i1}\|\mathbf{y})$	$\mathrm{E}(\Lambda_{i2}\|\mathbf{y})$	$\mathrm{E}(\sigma_i^2\|\mathbf{y})$	$\mathrm{E}(R_{i1}\|\mathbf{y})$	$\mathrm{E}(R_{i2}\|\mathbf{y})$
1	0.836	0	0.282	71.2	0
2	0.605	0	0.482	43.2	0
3	0.713	0	0.491	50.8	0
4	0.626	0	0.541	42	0
5	0.679	0	0.447	50.7	0
6	-0.117	0.615	0.526	2.7	40.9
7	-0.337	0.634	0.562	11.1	37.3
8	-0.064	0.671	0.428	1.3	50.6
9	-0.32	0.421	0.838	9.81	16.4
10	-0.057	0.411	0.789	1.1	18

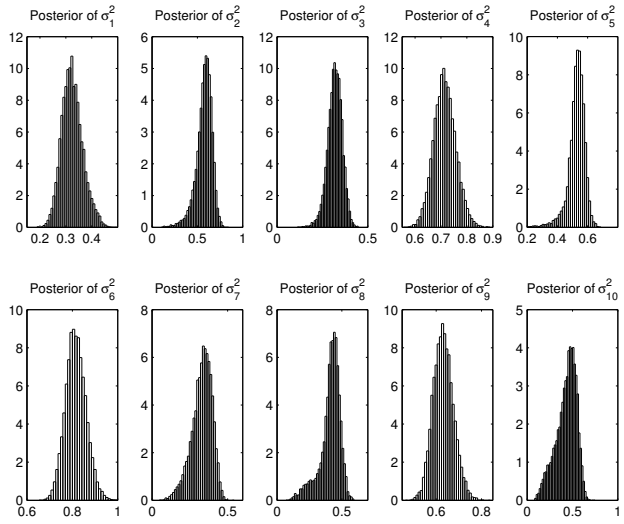

Figure 2: Maxwell's Children Data – normal children; posterior densities $p(\sigma_i^2|\mathbf{y}, \mathbf{l}^\star)$ of the idiosyncratic variances σ_i^2 for a 4-factor model based on **C2** given by $l_1 = 1$, $l_2 = 2$, $l_3 = 4$, and $l_4 = 5$ and the fractional prior $b = b_N$.

carrying out ML estimation. However, since we need a complete data set for our Bayesian analysis, we simulated two data sets of size $T = 148$ and $T = 810$, respectively, from a multivariate normal distribution having exactly the same correlation matrices as the ones published by Maxwell (1961) in Table 1 for neurotic and in Table 4 for normal children.

Table 7 and Table 8 report for both data sets the posterior distribution $p(r|\mathbf{y})$ for different priors. For the neurotic children, all estimators select a two factor model under the various fractional priors and the HPM and the MPM coincide, see Table 9 for $b = b_R$. The highest posterior identifiability constraint $\mathbf{l}^\star$ is given by $l_1 = 1$ and $l_2 = 6$. In comparison to that, the fixed scale priors are overfitting the number of factors and strongly overfit the number of non-zero factor loadings.

For the normal children we find that the prior is very influential on model size. The number of factors is equal to 4 for those priors that performed best for the simulated data considered in Subsection 5.1, namely $b = b_N$ and $b = 10^{-4}$, while the other priors choose a 5-factor model. None of the priors supports a three-factor model. Whenever a 4-factor model is selected, the highest posterior identifiability constraint $\mathbf{l}^\star$ is given by $l_1 = 1$, $l_2 = 2$, $l_3 = 4$, and $l_4 = 5$.

The structure of $\boldsymbol{\delta}^\Lambda$ reported in Table 9 reveals that for both data sets the ordering of the test scores would have been a poor choice under the conventional identifiability constraint on the main diagonal. Our procedure is robust in this respect and indicates that the scores are grouped and others than the first cognitive scores are leading the factors. To sum up, we decided to choose a 2-factor model with **C2** given by $l_1 = 1$ and $l_2 = 6$ for the neurotic children and a 4-factor model with **C2** given by $l_1 = 1$, $l_2 = 2$, $l_3 = 4$, and $l_4 = 5$ for the normal children.

Table 11: Kendall's Applicants Data; posterior distribution $p(r|\mathbf{y})$ of the number r of factors (bold number corresponding to the posterior mode $\tilde{r}$) and highest posterior identifiability constraint $\mathbf{l}^\star$ with corresponding posterior probability $p(\mathbf{l}^\star|\mathbf{y})$ for various priors; number of visited models N_v; frequency p_H (in percent), number of factors r_H, and model size d_H of the HPM; number of factors r_M and model size d_M of the MPM corresponding to $\mathbf{l}^\star$; inefficiency factor τ_d of the posterior draws of the model size d.

	$p(r\|\mathbf{y})$							
Prior	2	3	4	5	6	7 -8	$\mathbf{l}^\star$	$p(\mathbf{l}^\star\|\mathbf{y})$
$b = 10^{-2}$	0	0	0.004	**0.647**	0.291	0.058	(1,2,3,4,6)	0.605
$b_N = 1.4 \cdot 10^{-3}$	0	0.007	0.052	**0.876**	0.064	0.001	(1,2,3,4,6)	0.826
$b = 10^{-3}$	0	0	0.062	**0.901**	0.037	0	(1,2,3,4,6)	0.82
$b = 10^{-4}$	0	0.176	**0.524**	0.296	0.004	0	(1,2,4,6)	0.378
$b_R = 9.1 \cdot 10^{-5}$	0	0.155	**0.494**	0.346	0.006	0	(1,2,4,6)	0.329
$b = 10^{-5}$	0.084	**0.835**	0.081	0	0	0	(1,2,4)	0.721
GD	0	0.145	**0.832**	0.022	0	0	(1,2,3,4)	0.813
LW	0	0.01	**0.831**	0.157	0.002	0	(1,2,3,4)	0.781

Prior	N_v	$100p_H$	r_H	d_H	r_M	d_M	τ_δ
$b = 10^{-2}$	14855	0.0467	5	36	5	33	19.8
$b_N = 1.4 \cdot 10^{-3}$	11569	0.46	5	28	5	27	18.3
$b = 10^{-3}$	10647	0.55	5	25	5	27	16.8
$b = 10^{-4}$	3935	1.96	4	21	4	22	38.2
$b_R = 9.1 \cdot 10^{-5}$	5411	1.72	3	18	4	22	38.2
$b = 10^{-5}$	1632	10.87	3	17	3	17	28.
GD	14667	0.09	4	34	4	35	22.1
LW	14733	0.09	4	35	4	36	15.5

Table 10 shows the posterior means of the factor loading matrix $\mathbf{\Lambda}$, the idiosyncratic variances and the communalities for the neurotic children. The bold cells in the factor loading matrices indicate factor loadings that are non-zero according to the 2-factor HPM. The column signs are identified through the constraint $\Lambda_{11} > 0$ and $\Lambda_{62} > 0$. As in Maxwell (1961), we find that the first factor corresponds to cognitive abilities, while the second factor corresponds to orectic tendencies.

Finally, Figure 2 shows for the normal children that in contrast to Maxwell (1961) the posterior distributions of all idiosyncratic variances are unimodal and that also the posterior of σ_8^2 is bounded away from 0. This again demonstrates the usefulness of the specific prior developed in Subsection 3.2.

Table 12: Kendall's Applicants Data; inclusion probabilities for the indicator matrix $\boldsymbol{\delta}^{\Lambda}$ (fractional prior with $b = b_R$).

(i,j)		1	2	3	4
1	application letter	**1**	0	0	0
2	appearance	0.01	**1**	0	0
3	academic ability	0.015	0.024	0	0
4	likeability	0.083	0.263	**1**	0
5	self confidence	**0.974**	**1**	0.007	0
6	lucidity	0.516	**1**	0.011	**1**
7	honesty	0.292	0.092	**0.998**	0.002
8	salesmanship	0.017	**1**	0.005	0.003
9	experience	**1**	0.047	0.005	0.002
10	drive	0.217	**1**	0.004	0.007
11	ambition	0.047	**1**	0.004	0.007
12	grasp	0.26	**1**	0.008	**0.999**
13	potential	0.057	**1**	**0.529**	**0.57**
14	keenness to join	0.019	**0.902**	**0.986**	0.008
15	suitability	**1**	**0.956**	0.008	0.005

5.4 Kendall's Applicants Data

Our final example are data considered in Press & Shigemasu (1989, Table 1) and Rowe (2003, Table 9.2) which are scores on a ten-point scale on 15 characteristics of $T = 48$ applicants. Press & Shigemasu (1989) postulated that a 4 factor model is appropriate for these data and we want to compare this finding with parsimonious factor modeling. As in Press & Shigemasu (1989), the data are standardized. Table 11 reports the posterior distribution $p(r|\mathbf{y})$ of the number r of factors for different priors. The prior is very influential on model size for this data set and the number of estimated factors ranges from 3 to 5. Fractional priors based on $b = b_R$ and $b = 10^{-4}$ as well as the unit scale priors chooses a 4 factor model. The unit scale priors support the standard triangular identifiability constraints, however, these priors lead to larger models than the two fractional priors which support the identifiability constraint $l_1 = 1$, $l_2 = 2$, $l_3 = 4$, and $l_4 = 6$.

For all priors but the fractional prior with $b = 10^{-5}$ the frequency of the HPM is extremely small (often smaller than one percent) and the HPM is different from the MPM. The inclusion probabilities for various indicators in the factor loading matrix are close to 0.5, see Table 12 for the fractional prior b_R. A more detailed investigation (results not reported) revealed that the various priors exercise a strong influence on these probabilities some of which are bigger than 0.5 for some priors and smaller than 0.5 for others. Hence the decision which factor loadings

Table 13: Kendall's Applicants Data; posterior means of the factor loading matrix, the idiosyncratic variances and the communalities (in percent) for a 4-factor model with $b = b_R$ under the constraint **C2** given by $l_1 = 1$, $l_2 = 2$, $l_3 = 4$, and $l_4 = 6$.

		Parameter					Communality			
i	Item	Λ_{i1}	Λ_{i2}	Λ_{i3}	Λ_{i4}	σ_i^2	E(R_{i1})	E(R_{i2})	E(R_{i3})	E(R_{i4})
1	application letter	0.643	0	0	0	0.556	42.7	0	0	0
2	appearance	0.008	0.361	0	0	0.819	0.26	15.2	0	0
3	academic ability	0.009	-0.001	0	0	0.947	0.29	0.03	0	0
4	likeability	0.008	0.895	0.102	0	0.102	0.25	85.4	3.6	0
5	self confidence	-0.339	-0.001	0.976	0	0.12	10.3	0.01	79.4	0
6	lucidity	-0.126	0.004	0.867	0.082	0.105	2.75	0.07	71.4	15.5
7	honesty	-0.165	0.687	0.03	0.000	0.465	5.6	47.7	0.97	0.02
8	salesmanship	-0.000	-0.001	0.933	-0.000	0.126	0.1	0.01	87	0.01
9	experience	0.774	-0.001	0.012	0.000	0.379	60.7	0.02	0.4	0.02
10	drive	0.094	0.000	0.836	0.002	0.223	2.9	0.01	73.2	0.07
11	ambition	-0.011	-0.000	0.915	0.001	0.16	0.2	0.01	83.4	0.03
12	grasp	0.068	0.002	0.783	0.163	0.054	1.9	0.07	65	27.2
13	potential	0.026	0.205	0.737	0.191	0.178	0.82	8.1	62.9	7.3
14	keenness to join	0.004	0.525	0.397	-0.001	0.386	0.13	34.2	21.3	0.06
15	suitability	0.738	0.001	0.355	0.001	0.168	63.6	0.03	16.5	0.02

are zero or not is not easy for this data set and explains the extreme variability of the model sizes d_H and d_M in Table 11 across the various priors.

We decided to investigate a model with 4 factors under the constraint **C2** given by $l_1 = 1$, $l_2 = 2$, $l_3 = 4$, and $l_4 = 6$ in more detail and reran MCMC as described at the end of Subsection 4.2 for a fractional prior with $b = b_R$. Table 13 shows the factor loading matrix, the idiosyncratic variances and the communalities. Based on $\mathrm{I}^\star$, identifiability is achieved through the constraint $\Lambda_{11} > 0$, $\Lambda_{22} > 0$, $\Lambda_{43} > 0$, and $\Lambda_{64} > 0$. When analyzing which factors have high communalities for the different items we find a similar interpretation of the factors as Rowe (2003, Table 9.6), namely that factor 1 is a measure of position match (application letter, experience, suitability), factor 2 could be described as charisma (likeability, honesty, keenness to join), and factor 3 is measure of personality (self confidence, lucidity, salesmanship, drive, ambition, grasp, potential).

6 Conclusion

We introduced a new approach to Bayesian inference and Bayesian model search for the important class of Gaussian factor models. Our main contributions are two-fold. First, we lay down

a new and general set of identifiability conditions that handles the ordering problem present in most of the current literature. This leads directly to our second main contribution, i.e. a new strategy for searching the space of parsimonious/sparse factor loading matrices. To that end, we designed a highly computationally efficient MCMC scheme for posterior inference which makes several improvements over the existing alternatives while outlining various strategies for conditional posterior inference in a factor selection scenario. In additional, the prior specification for all model parameters is carefully studied with particular emphasis to *a)* analyzing the prior influence on our factor selection scheme, *b)* avoiding the Heywood problem (negative idiosyncratic variances), and *c)* studying alternative prior specifications for the components of the factor loading matrix that tackle underfitting and overfitting situations. Our applications offer strong empirical evidence that our method is able to select the true number of factors. It is also able to pinpoint the true identifiability constraint and reconstruct the structure of the factor loading matrix.

A Appendix A: Proof of Theorem 1

It is possible to embed the true model (1) in a regression-type representation with k potential factors by permuting the columns of $\mathbf{\Lambda}$ arbitrarily and adding $k-r$ zeros columns in between these columns at arbitrary positions. Denote the resulting coefficient matrix by $\boldsymbol{\beta}_0$. Subsequently we refer to zero columns of $\boldsymbol{\beta}_0$ as the "unidentified" columns, while the remaining columns of $\boldsymbol{\beta}_0$ are called "identified".

From Geweke & Singleton (1980) we known that, since $\boldsymbol{\beta}_0$ has rank $r < k$, there exists a $k \times (k-r)$ matrix $\mathbf{Q}$ such that

$$\boldsymbol{\beta}_0 \mathbf{Q} = \mathbf{O}_{m \times (k-r)}, \tag{24}$$

$\mathbf{Q}^{'}\mathbf{Q} = \mathbf{I}_{k-r}$, and, for any $m \times (k-r)$ dimensional matrix $\mathbf{M}$ with mutually orthogonal rows,

$$\boldsymbol{\beta} = \boldsymbol{\beta}_0 + \mathbf{M}\mathbf{Q}^{'}. \tag{25}$$

Split the columns of $\boldsymbol{\beta}$ in the following way: let $\boldsymbol{\beta}_1$ denote the $m \times r$ submatrix containing the identified columns and let $\boldsymbol{\beta}_2$ denote the $m \times (k-r)$ submatrix containing the remaining columns. Note that the corresponding submatrices of $\boldsymbol{\beta}_0$ are $\mathbf{\Lambda}^{\rho}$, where $\mathbf{\Lambda}^{\rho}$ is obtained from $\mathbf{\Lambda}$ by permuting the columns, and a zero matrix by construction. Split the rows of $\mathbf{Q}$ in a similar way. Let $\mathbf{Q}_1$ and $\mathbf{Q}_2$ denote those submatrices of size $r \times (k-r)$ and $(k-r) \times (k-r)$ which contain, respectively, the rows corresponding to the identified and the non-identified columns. Then we obtain from (25):

$$\boldsymbol{\beta}_1 = \mathbf{\Lambda}^{\rho} + \mathbf{M}\mathbf{Q}_1^{'}, \tag{26}$$

$$\boldsymbol{\beta}_2 = \mathbf{M}\mathbf{Q}_2^{'}. \tag{27}$$

Because $\boldsymbol{\beta}_0\mathbf{Q} = \boldsymbol{\Lambda}^\rho\mathbf{Q}_1$, we obtain from (24) that $\boldsymbol{\Lambda}^\rho\mathbf{Q}_1 = \mathbf{O}_{m\times(k-r)}$ and therefore $(\boldsymbol{\Lambda}^\rho)'\boldsymbol{\Lambda}^\rho\mathbf{Q}_1 = \mathbf{O}_{r\times(k-r)}$. Since $\text{rank}\big((\boldsymbol{\Lambda}^\rho)'\boldsymbol{\Lambda}\big) = \text{rank}(\boldsymbol{\Lambda}^\rho) = \text{rank}(\boldsymbol{\Lambda}) = r$ it follows that $(\boldsymbol{\Lambda}^\rho)'\boldsymbol{\Lambda}^\rho$ is invertible and therefore $\mathbf{Q}_1 = \mathbf{O}_{r\times(k-r)}$. Consequently, the matrix $\mathbf{M}\mathbf{Q}_1'$ appearing in (26) is equal to a zero matrix,

$$\mathbf{M}\mathbf{Q}_1' = \mathbf{O}_{m\times r}, \tag{28}$$

which proves the part (a) of our theorem, because $\boldsymbol{\beta}_1 = \boldsymbol{\Lambda}^\rho + \mathbf{M}\mathbf{Q}_1' = \boldsymbol{\Lambda}^\rho$. Furthermore, because $\mathbf{Q}'\mathbf{Q} = \mathbf{Q}_1'\mathbf{Q}_1 + \mathbf{Q}_2'\mathbf{Q}_2 = \mathbf{I}_{k-r}$ and $\mathbf{Q}_1'\mathbf{Q}_1 = \mathbf{O}_{(k-r)\times(k-r)}$ we obtain

$$\mathbf{Q}_2'\mathbf{Q}_2 = \mathbf{I}_{k-r}. \tag{29}$$

To prove part (b) and (c), we recall that $\boldsymbol{\beta}$ obeys condition **B1**. Hence, $\text{rank}(\boldsymbol{\beta}) = \text{rank}(\boldsymbol{\Lambda}) + s$, where s is equal to the number of non-zero columns among the non-identified columns of $\boldsymbol{\beta}$.

If $\text{rank}(\boldsymbol{\beta}) = \text{rank}(\boldsymbol{\Lambda})$, then $s = 0$ which means that all non-identified columns of $\boldsymbol{\beta}$ are zero columns and consequently $\boldsymbol{\beta}_2 = \mathbf{O}_{m\times(k-r)}$. From (27) we obtain that $\mathbf{M}\mathbf{Q}_2' = \mathbf{O}_{m\times(k-r)}$. Together with (28) we obtain: $\mathbf{M}\mathbf{M}' = \mathbf{M}\mathbf{Q}'\mathbf{Q}\mathbf{M}' = \mathbf{O}_{m\times m}$. This proves part (b) of our theorem.

If $s = \text{rank}(\boldsymbol{\beta}) - \text{rank}(\boldsymbol{\Lambda}) > 0$, then exactly $k - r - s$ among the $k - r$ columns of $\boldsymbol{\beta}_2$ are zero columns, while the remaining s columns are non-zero columns. Let $j_1, \ldots, j_s$ denote the index of these columns in $\boldsymbol{\beta}$. Let $\mathbf{A}$ denote the $m \times s$ matrix containing these columns and let $r_1, \ldots, r_s$ denote the row indices of the top non-zero element in each column of $\mathbf{A}$, i.e. $A_{r_l,l} = \beta_{r_l,j_l} \neq 0$ for $l = 1, \ldots, s$. Using (27) and (29) we obtain:

$$\mathbf{A}\mathbf{A}' = \boldsymbol{\beta}_2\boldsymbol{\beta}_2' = \mathbf{M}\mathbf{Q}_2'\mathbf{Q}_2\mathbf{M}' = \mathbf{M}\mathbf{M}' = \mathbf{D}, \tag{30}$$

where $\mathbf{D}$ is a $m \times m$ diagonal matrix with $\text{rank}(\mathbf{D}) = \text{rank}(\boldsymbol{\beta}_3) = s$.

Now let $\mathbf{L}$ denote the $s \times s$ matrix formed from the rows $r_1, \ldots, r_s$ of $\mathbf{A}$ and let $\mathbf{B}$ denote the $(m - s) \times s$ matrix formed from the remaining rows. Evidently, $\mathbf{L}$ is a lower triangular matrix with diagonal elements $L_{ll} = \beta_{r_l,j_l} \neq 0$ for $l = 1, \ldots, s$ and therefore of full rank. We obtain from (30):

$$\mathbf{L}\mathbf{L}' = \tilde{\mathbf{D}}, \tag{31}$$

$$\mathbf{L}\mathbf{B}' = \mathbf{O}_{(k-s)\times(m-s)}, \tag{32}$$

where $\tilde{\mathbf{D}}$ is a $s \times s$ diagonal matrix with $\text{rank}\Big(\tilde{\mathbf{D}}\Big) = \text{rank}(\mathbf{L}) = s$ formed from the elements $D_{j_1,j_1}, \ldots, D_{r_s,r_s}$ of $\mathbf{D}$. It follows from (31) that $\mathbf{L}$ is the Cholesky decomposition of a diagonal matrix of rank s and therefore a diagonal matrix of full rank. Multiplying (32) by $\mathbf{L}^{-1}$ yields $\mathbf{B} = \mathbf{O}_{(m-s)\times s}$. Therefore, the elements $A_{r_l,l} = \beta_{r_l,j_l}$, $l = 1, \ldots, s$ are the only non-zero elements of $\mathbf{A}$. Since the lth column of$\mathbf{A}$ is identical with the j_lth column of $\boldsymbol{\beta}$, we obtain that the j_lth column of $\boldsymbol{\beta}$ is equal to 0 apart from element β_{r_l,j_l}. Finally, combining (7) and (30), we obtain $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0 - \mathbf{D}$. This proves part (c) of our theorem.

B Appendix B: Details of MCMC Estimation

B.1 Full Conditional Posteriors

The parameters $(\boldsymbol{\beta}_{i\cdot}^{\boldsymbol{\delta}}, \sigma_i^2)$ are independent *a posteriori* across rows given $\mathbf{f}$. The posterior distribution $p(\boldsymbol{\beta}_{i\cdot}^{\boldsymbol{\delta}}, \sigma_i^2|\mathbf{y}, \mathbf{f}, \boldsymbol{\delta})$ is derived from the "regression model" (15). The precise form of the posterior moments depends on the number of non-zero elements in the ith row of $\boldsymbol{\beta}$, i.e. $q_i = \sum_{j=1}^{k} \delta_{ij}$.

For any zero row, i.e. if $q_i = 0$, we are dealing in (15) with a "null" model without regressors. Hence the posterior of σ_i^2 is given by

$$\sigma_i^2|\tilde{\mathbf{y}}_i, \mathbf{f} \sim \mathcal{G}^{-1}\left(c_T^n, C_{iT}^n\right), \tag{33}$$

$$c_T^n = c_0 + \frac{T}{2}, \qquad C_{iT}^n = C_{i0} + \frac{1}{2}\sum_{t=1}^{T} y_{it}^2. \tag{34}$$

For all remaining rows, i.e. if $q_i > 0$, the joint posterior of $(\boldsymbol{\beta}_{i\cdot}^{\boldsymbol{\delta}}, \sigma_i^2)$ is given by:

$$\sigma_i^2|\tilde{\mathbf{y}}_i, \mathbf{f} \sim \mathcal{G}^{-1}\left(c_T, C_{iT}^{\boldsymbol{\delta}}\right), \tag{35}$$

$$\boldsymbol{\beta}_{i\cdot}^{\boldsymbol{\delta}}|\sigma_i^2, \tilde{\mathbf{y}}_i, \mathbf{f} \sim N\left(\mathbf{B}_{iT}^{\boldsymbol{\delta}}\mathbf{m}_{iT}^{\boldsymbol{\delta}}, \mathbf{B}_{iT}^{\boldsymbol{\delta}}\sigma_i^2\right), \tag{36}$$

where the moments depend on the prior chosen for $\boldsymbol{\beta}_{i\cdot}^{\boldsymbol{\delta}}$. Under the fractional prior

$$(\mathbf{B}_{iT}^{\boldsymbol{\delta}})^{-1} = (\mathbf{X}_i^{\boldsymbol{\delta}})'\mathbf{X}_i^{\boldsymbol{\delta}}, \qquad \mathbf{m}_{iT}^{\boldsymbol{\delta}} = (\mathbf{X}_i^{\boldsymbol{\delta}})'\tilde{\mathbf{y}}_i, \qquad c_T = c_0 + \frac{(1-b)T}{2}, \tag{37}$$

$$C_{iT}^{\boldsymbol{\delta}} = C_{i0} + \frac{(1-b)}{2}\left(\tilde{\mathbf{y}}_i'\tilde{\mathbf{y}}_i - (\mathbf{m}_{iT}^{\boldsymbol{\delta}})'\mathbf{B}_{iT}^{\boldsymbol{\delta}}\mathbf{m}_{iT}^{\boldsymbol{\delta}}\right), \tag{38}$$

otherwise:

$$(\mathbf{B}_{iT}^{\boldsymbol{\delta}})^{-1} = (\mathbf{B}_{i0}^{\boldsymbol{\delta}})^{-1} + (\mathbf{X}_i^{\boldsymbol{\delta}})'\mathbf{X}_i^{\boldsymbol{\delta}}, \qquad \mathbf{m}_{iT}^{\boldsymbol{\delta}} = (\mathbf{B}_{i0}^{\boldsymbol{\delta}})^{-1}\mathbf{b}_{i0}^{\boldsymbol{\delta}} + (\mathbf{X}_i^{\boldsymbol{\delta}})'\tilde{\mathbf{y}}_i, \tag{39}$$

$$c_T = c_0 + \frac{T}{2}, \tag{40}$$

$$C_{iT}^{\boldsymbol{\delta}} = C_{i0} + \frac{1}{2}\left(\tilde{\mathbf{y}}_i'\tilde{\mathbf{y}}_i + (\mathbf{b}_{i0}^{\boldsymbol{\delta}})'(\mathbf{B}_{i0}^{\boldsymbol{\delta}})^{-1}\mathbf{b}_{i0}^{\boldsymbol{\delta}} - (\mathbf{m}_{iT}^{\boldsymbol{\delta}})'\mathbf{B}_{iT}^{\boldsymbol{\delta}}\mathbf{m}_{iT}^{\boldsymbol{\delta}}\right). \tag{41}$$

The marginal likelihood $p(\tilde{\mathbf{y}}_i|\boldsymbol{\delta}_{i,\cdot}, \mathbf{f})$ of each regression model (15) is relevant for sampling the indicators $\boldsymbol{\delta}_{i,\cdot}$ in row i. If all elements of $\boldsymbol{\delta}_{i,\cdot}$ in row i are 0, i.e $q_i = 0$, then the marginal likelihood simplifies to

$$p(\tilde{\mathbf{y}}_i|\boldsymbol{\delta}_{i,\cdot}, \mathbf{f}) = \frac{\Gamma(c_T^n)(C_{i0})^{c_0}}{(2\pi)^{T/2}\Gamma(c_0)(C_{iT}^n)^{c_T^n}}, \tag{42}$$

where c_T^n and C_{iT}^n are the posterior moments of σ_i^2 under a "null" model, given by (33). If at least one element of $\boldsymbol{\delta}_{i,\cdot}$ is different from 0, then the marginal likelihood reads under the fractional prior:

$$p(\tilde{\mathbf{y}}_i|\boldsymbol{\delta}_{i,\cdot}, \mathbf{f}) = \frac{b^{q_i/2}\Gamma(c_T)(C_{i0})^{c_0}}{(2\pi)^{T(1-b)/2}\Gamma(c_0)(C_{iT}^{\boldsymbol{\delta}})^{c_T}}, \tag{43}$$

while for any other prior the marginal likelihood is given by:

$$p(\tilde{\mathbf{y}}_i|\boldsymbol{\delta}_{i,\cdot},\mathbf{f}) = \frac{1}{(2\pi)^{T/2}}\frac{|\mathbf{B}_{iT}^{\boldsymbol{\delta}}|^{1/2}}{|\mathbf{B}_{i0}^{\boldsymbol{\delta}}|^{1/2}}\frac{\Gamma(c_T)(C_{i0})^{c_0}}{\Gamma(c_0)(C_{iT}^{\boldsymbol{\delta}})^{c_T}}, \tag{44}$$

where $\mathbf{B}_{iT}^{\boldsymbol{\delta}}$, c_T and $C_{iT}^{\boldsymbol{\delta}}$ are the posterior moments of $p(\boldsymbol{\beta}_{i\cdot}^{\boldsymbol{\delta}}, \sigma_i^2|\boldsymbol{\delta}_{i\cdot}, \tilde{\mathbf{y}}_i, \mathbf{f})$ given by (35) and (36).

B.2 Joint Sampling of idiosyncratic variances and factor loadings

Joint sampling all idiosyncratic variances and all factor loadings for a multi-factor model is quite challenging, but feasible. As in Appendix B.1, we distinguish between zero rows ($q_i = 0$) and non-zero rows ($q_i > 0$) in $\boldsymbol{\beta}$. Joint sampling of the idiosyncratic variances for all zero rows is easily vectorized using (33).

Let $i_1, \ldots, i_n$ be the indices of the non-zeros rows of $\boldsymbol{\beta}$, i.e. $q_{i_j} > 0$ for $j = 1, \ldots, n$ with n being the total number of non-zero rows. Let $\boldsymbol{\beta}^{\boldsymbol{\delta}} = (\boldsymbol{\beta}_{i_1\cdot}^{\boldsymbol{\delta}}, \ldots, \boldsymbol{\beta}_{i_n\cdot}^{\boldsymbol{\delta}})$ be a vector obtained by stacking row by row all non-zero elements in each row. Let $d_{\boldsymbol{\delta}} = \sum_i q_i$ be the total number of non-zero elements in $\boldsymbol{\beta}^{\boldsymbol{\delta}}$.

To sample the idiosyncratic variances $\sigma_{i_1}^2, \ldots, \sigma_{i_n}^2$ and the non-zero factor loadings $\boldsymbol{\beta}^{\boldsymbol{\delta}}$ jointly, we proceed in the following way:

1. Construct the information matrix $\mathbf{P}$ and the covector $\mathbf{m}$ of the joint posterior

$$\boldsymbol{\beta}^{\boldsymbol{\delta}}|\sigma_{i_1}^2, \ldots, \sigma_{i_n}^2, \mathbf{f}, \mathbf{y} \sim N_{d_{\boldsymbol{\delta}}}\left(\mathbf{P}^{-1}\mathbf{m}, \mathbf{P}^{-1}\mathbf{D}\right).$$

   $\mathbf{D} = \text{Diag}\left(\sigma_{i_1}^2\mathbf{1}_{1\times q_{i_1}} \cdots \sigma_{i_n}^2\mathbf{1}_{1\times q_{i_n}}\right)$, with $\mathbf{1}_{1\times l}$ being a $1\times l$ row vector of ones, is a $d_{\boldsymbol{\delta}}\times d_{\boldsymbol{\delta}}$ diagonal matrix containing the idiosyncratic variances, while the $d_{\boldsymbol{\delta}} \times d_{\boldsymbol{\delta}}$ matrix $\mathbf{P}$ and the $d_{\boldsymbol{\delta}} \times 1$ vector $\mathbf{m}$ are given by:

$$\mathbf{P} = \begin{pmatrix} (\mathbf{B}_{i_1,T}^{\boldsymbol{\delta}})^{-1} & \mathbf{O} & \cdots & \mathbf{O} \\ \mathbf{O} & (\mathbf{B}_{i_2,T}^{\boldsymbol{\delta}})^{-1} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{O} \\ \mathbf{O} & \cdots & \mathbf{O} & (\mathbf{B}_{i_n,T}^{\boldsymbol{\delta}})^{-1} \end{pmatrix}, \qquad \mathbf{m} = \begin{pmatrix} \mathbf{m}_{i_1,T}^{\boldsymbol{\delta}} \\ \vdots \\ \mathbf{m}_{i_n,T}^{\boldsymbol{\delta}} \end{pmatrix},$$

   where $(\mathbf{B}_{i_j,T}^{\boldsymbol{\delta}})^{-1}$ and $\mathbf{m}_{i_j,T}^{\boldsymbol{\delta}}$ are the information matrix and the covector appearing in the posterior (36) of the non-zero elements in row i_j. $\mathbf{P}$ is a sparse band matrix with maximal band width equal to $\max q_{i_j}$.

2. Compute the Cholesky decomposition $\mathbf{P} = \mathbf{L}\mathbf{L}'$, where $\mathbf{L}$ is lower triangular, using a special algorithm developed for band matrices. Next, solve $\mathbf{L}\mathbf{x} = \mathbf{m}$ for $\mathbf{x}$ using an algorithm specially designed for triangular matrices. Evidently, $\mathbf{x}$ is a $d_{\boldsymbol{\delta}} \times 1$ vector.

3. Sample $\sigma^2_{i_1}, \ldots, \sigma^2_{i_n}$ jointly from (35) where for each $j = 1, \ldots, n$ the posterior scale parameter $C^{\boldsymbol{\delta}}_{i_j,T}$ is obtained from $\mathbf{x}$ through:

$$(\mathbf{m}^{\boldsymbol{\delta}}_{i,T})' \mathbf{B}^{\boldsymbol{\delta}}_{iT} \mathbf{m}^{\boldsymbol{\delta}}_{i,T} = \mathbf{x}'_{i_j} \mathbf{x}_{i_j}, \tag{45}$$

with $\mathbf{x}_{i_j}$ being the q_{i_j} dimensional sub vector of $\mathbf{x}$ corresponding to $\boldsymbol{\beta}^{\boldsymbol{\delta}}_{i_j,\cdot}$.

4. Finally, define the diagonal matrix $\mathbf{D}$ from $\sigma^2_{i_1}, \ldots, \sigma^2_{i_n}$ as described above and draw $\mathbf{z} \sim N_{d_{\boldsymbol{\delta}}}(\mathbf{0}, \mathbf{D})$. Solving the system

$$\mathbf{L}' \boldsymbol{\beta}^{\boldsymbol{\delta}} = \mathbf{x} + \mathbf{z} \tag{46}$$

for $\boldsymbol{\beta}^{\boldsymbol{\delta}}$ leads to a draw from the joint posterior $\boldsymbol{\beta}^{\boldsymbol{\delta}} | \sigma^2_{i_1}, \ldots, \sigma^2_{i_n}, \mathbf{y}, \mathbf{f}$.

Proof. Let $\mathbf{L}_{i_j}$ be the $q_{i_j} \times q_{i_j}$ sub matrix of $\mathbf{L}$ corresponding to $\boldsymbol{\beta}^{\boldsymbol{\delta}}_{i_j,\cdot}$. Evidently, $\mathbf{L}_{i_j}$ is equal to the Cholesky decomposition of the individual information matrix $(\mathbf{B}^{\boldsymbol{\delta}}_{i_j,T})^{-1}$. Furthermore, the q_{i_j} dimensional sub vector $\mathbf{x}_{i_j}$ corresponding to $\boldsymbol{\beta}^{\boldsymbol{\delta}}_{i_j,\cdot}$ satisfies $\mathbf{L}_{i_j} \mathbf{x}_{i_j} = \mathbf{m}^{\boldsymbol{\delta}}_{i_j,T}$. Therefore:

$$\mathbf{x}'_{i_j} \mathbf{x}_{i_j} = (\mathbf{m}^{\boldsymbol{\delta}}_{i_j,T})' (\mathbf{L}'_{i_j})^{-1} \mathbf{L}^{-1}_{i_j} \mathbf{m}^{\boldsymbol{\delta}}_{i_j,T} = (\mathbf{m}^{\boldsymbol{\delta}}_{i_j,T})' (\mathbf{L}_{i_j} \mathbf{L}'_{i_j})^{-1} \mathbf{m}^{\boldsymbol{\delta}}_{i_j,T} = (\mathbf{m}^{\boldsymbol{\delta}}_{i_j,T})' \mathbf{B}^{\boldsymbol{\delta}}_{i_j,T} \mathbf{m}^{\boldsymbol{\delta}}_{i_j,T}.$$

This proves (45). Next, we prove that the solution $\boldsymbol{\beta}^{\boldsymbol{\delta}}$ of (46) is a posterior draw. Note that $\mathbf{L}\mathbf{L}' \boldsymbol{\beta}^{\boldsymbol{\delta}} = \mathbf{L}\mathbf{x} + \mathbf{L}\mathbf{z} = \mathbf{m} + \mathbf{L}\mathbf{z}$. Therefore

$$\boldsymbol{\beta}^{\boldsymbol{\delta}} = (\mathbf{L}\mathbf{L}')^{-1} \mathbf{m} + (\mathbf{L}\mathbf{L}')^{-1} \mathbf{L}\mathbf{z} = \mathbf{P}^{-1} \mathbf{m} + (\mathbf{L}')^{-1} \mathbf{z}.$$

It follows immediately that $\boldsymbol{\beta}^{\boldsymbol{\delta}} \sim N_{d_{\boldsymbol{\delta}}}(\mathbf{P}^{-1}\mathbf{m}, \mathbf{P}^{-1}\mathbf{D})$, because $\mathrm{E}(\boldsymbol{\beta}^{\boldsymbol{\delta}}) = \mathbf{P}^{-1}\mathbf{m}$ and $\mathrm{V}(\boldsymbol{\beta}^{\boldsymbol{\delta}}) = (\mathbf{L}')^{-1} \mathbf{D} \mathbf{L}^{-1} = (\mathbf{L}')^{-1} \mathbf{L}^{-1} \mathbf{D} = \mathbf{P}^{-1} \mathbf{D}$, because $\mathbf{D}\mathbf{L}^{-1} = \mathbf{L}^{-1}\mathbf{D}$.

B.3 Joint Sampling of All Indicators in a Column

To derive O_{ij}, the marginal likelihoods $p(\tilde{\mathbf{y}}_i | \boldsymbol{\delta}_{i,\cdot}, \mathbf{f})$ may be computed both for $\delta_{ij} = 0$ and $\delta_{ij} = 1$ individually for each row $i = 1, \ldots, m$ as in Appendix B.1, however, this is very inefficient, if m is large. Subsequently, we show that it is possible to compute the log posterior odd for all indicators in column j jointly in a very efficient manner.

In (19), the ratio of the marginal likelihoods where δ_{ij} switched between 1 and 0 is required. The precise form of this ratio depends on the remaining indicators $\boldsymbol{\delta}_{i,-j}$ in row i. It is easy to vectorize the computation of O_{ij} for all rows i , where $\sum_{l \neq j} \delta_{il} = 0$, i.e. if all elements of $\boldsymbol{\delta}_{i,-j}$ are zero. Because we are dealing with a "null"model under $\delta_{ij} = 0$ and a one-dimensional factor model under $\delta_{ij} = 1$ computation of the log odds O_{ij} is easy using (42) – (44):

$$\log \frac{p(\tilde{\mathbf{y}}_i | \delta_{ij} = 1, \boldsymbol{\delta}_{i,-j}, \mathbf{f})}{p(\tilde{\mathbf{y}}_i | \delta_{ij} = 0, \boldsymbol{\delta}_{i,-j}, \mathbf{f})} = \log \frac{\Gamma(c_T)(C^n_{iT})^{c^n_T}}{\Gamma(c^n_T)(C_{iT})^{c_T}} + B_i, \tag{47}$$

where $B_i = 0.5\log(b(2\pi)^{bT})$ for a fractional prior and $B_i = 0.5\log(B_{iT}/B_{i0,jj})$, where $B_{i0,jj}$ is jth diagonal element of $\mathbf{B}_{i0}$, for any other prior. c_T^n and C_{iT}^n are the posterior moments of the null model, see (33), while C_{iT} and B_{iT} are the posterior moments for $\delta_{ij} = 1$ and simplify for a fractional prior to

$$C_{iT} = C_{i0} + \frac{(1-b)}{2}\sum_{t=1}^{T}(y_{it} - f_{jt}b_{iT})^2, \quad b_{iT} = \left(\sum_{t=1}^{T} f_{jt}y_{it}\right) / \left(\sum_{t=1}^{T} f_{jt}^2\right),$$

and for any other prior to:

$$B_{iT} = B_{i0,jj}/(B_{i0,jj} + \sum_{t=1}^{T} f_{jt}^2), \qquad b_{iT} = B_{iT}(b_{i0,j}/B_{i0,jj} + \sum_{t=1}^{T} f_{jt}y_{it})$$

$$C_{iT} = C_{i0} + \frac{1}{2}\sum_{t=1}^{T}(y_{it} - f_{jt}b_{iT})^2 + \frac{1}{2B_{i0,jj}}(b_{iT} - b_{i0,j})^2.$$

It is also possible to vectorize the computation of the ratio of the marginal likelihoods for the remaining rows $i \in \{i_1, \ldots, i_n\}$, where at least one element of $\boldsymbol{\delta}_{i,-j}$ is different from zero. Using (43) and (44), the ratio reads:

$$\log\frac{p(\tilde{\mathbf{y}}_i|\delta_{ij}=1,\boldsymbol{\delta}_{i,-j},\mathbf{f})}{p(\tilde{\mathbf{y}}_i|\delta_{ij}=0,\boldsymbol{\delta}_{i,-j},\mathbf{f})} = c_T\log\frac{C_{iT}^0}{C_{iT}} + B_i,$$

where $B_i = 0.5\log b$ for a fractional prior and $B_i = 0.5\log(|\mathbf{B}_{iT}|/|\mathbf{B}_{iT}^0|) - 0.5\log(|\mathbf{B}_{i0}|/|\mathbf{B}_{i0}^0|)$ for any other prior. Here C_{iT}, $\mathbf{B}_{iT}$ and $\mathbf{B}_{i0}$ refer to a model where $\delta_{ij} = 1$, while C_{iT}^0, $\mathbf{B}_{iT}^0$ and $\mathbf{B}_{i0}^0$ refer to a model where $\delta_{ij} = 0$. To compute all relevant posterior moments simultaneously, we proceed in the following way:

1. Set in each row $i_1, \ldots, i_n$ the indicator $\delta_{i_l,j} = 1$. Reorder the columns of the factor loading matrix in such a way, that the jth column appears last. This is simply done by permuting the column of $\mathbf{F}$ appropriately before defining $\mathbf{X}_{i_l}^{\boldsymbol{\delta}}$. While the fractional prior is not affected by this, it might be necessary to reorder the prior mean and the prior covariance matrix for alternative priors.

2. Set up the information matrix $\mathbf{P}$ and the covector $\mathbf{m}$ of the corresponding joint posterior of all non-zero factor loadings in the rows $i_1, \ldots, i_n$ as described in Appendix B.2. Compute the Cholesky decomposition $\mathbf{L}$ of $\mathbf{P}$ and the corresponding vector $\mathbf{x}$ solving $\mathbf{L}\mathbf{x} = \mathbf{m}$.

3. Knowing $\mathbf{L}$ and $\mathbf{x}$, a vectorized computation of the log ratio (47) for all rows $i \in \{i_1, \ldots, i_n\}$ is possible. The posterior moments $C_{i_l,T}$ are directly available from the appropriate sub vectors $\mathbf{x}_{i_l}$ of $\mathbf{x}$, see (45). When we switch to a model where $\delta_{i_l,j} = 0$, then for the fractional prior

$$C_{i_l,T}^0 = C_{i_l,T} + \frac{1-b}{2}(x_{i_l}^\star)^2, \tag{48}$$

while for any other prior:

$$C^0_{i_l,T} = C_{i_l,T} + \frac{1}{2}(x^\star_{i_l})^2, \tag{49}$$

where $x^\star_{i_l} = (\mathbf{x}_{i_l})_{q_{i_l}}$ is the last element of $\mathbf{x}_{i_l}$. Furthermore,

$$0.5 \log(|\mathbf{B}_{i_l,T}|/|\mathbf{B}^0_{i_l,T}|) = -\log L^\star_{i_l}, \tag{50}$$

where $L^\star_{i_l} = (\mathbf{L}_i)_{q_{i_l},q_{i_l}}$ is the last diagonal element of the submatrix $\mathbf{L}_{i_l}$.

Proof of 3. When we switch to a model where the indicators are 0, then the information matrix $\mathbf{P}^0$ and the covector $\mathbf{m}^0$ of the joint posterior of the remaining non-zero factor loadings is obtained from $\mathbf{P}$ and $\mathbf{m}$ simply by deleting all rows (and columns) corresponding to $\delta_{i_1,j}, \ldots, \delta_{i_n,j}$, and the Cholesky decomposition $\mathbf{L}^0$ of $\mathbf{P}^0$ is obtained from $\mathbf{L}$ in the same way. Also the vector $\mathbf{x}^0$ solving $\mathbf{L}^0\mathbf{x}^0 = \mathbf{m}^0$ is obtained from $\mathbf{x}$ simply by deleting the rows corresponding to $\delta_{i_1,j}, \ldots, \delta_{i_n,j}$. This last result is easily seen by considering the subsystem $\mathbf{L}_{i_l}\mathbf{x}_{i_l} = \mathbf{m}^{\boldsymbol{\delta}}_{i_l,T}$ corresponding to the i_lth row. Because

$$\mathbf{L}_{i_l} = \begin{pmatrix} \mathbf{L}^0_{i_l} & \mathbf{O} \\ \mathbf{l}_{i_l} & (\mathbf{L}_i)_{q_{i_l},q_{i_l}} \end{pmatrix}, \tag{51}$$

we obtain $\mathbf{L}^0_{i_l}\mathbf{x}^0_{i_l} = \mathbf{m}^0_{i_l}$, where $\mathbf{x}^0_{i_l}$ is obtained from $\mathbf{x}_{i_l}$ by deleting the last element $x^\star_{i_l} = (\mathbf{x}_{i_l})_{q_{i_l}}$. Hence, $\mathbf{x}^0_{i_l}$ defines the desired subvector of $\mathbf{x}^0$ to compute $C^0_{i_l,T}$ as in (45). Since $(\mathbf{x}^0_{i_l})'\mathbf{x}^0_{i_l} = \mathbf{x}'_{i_l}\mathbf{x}_{i_l} - (\mathbf{x}_{i_l})^2_{q_{i_l}}$ we obtain from (35) that (48) and (49) hold. Note, however, that this simple relationship would not hold without reordering the columns as described above.

Finally, to compute the marginal likelihood for a standard prior, the ratio of the determinants $|\mathbf{B}_{i_l,T}|/|\mathbf{B}^0_{i_l,T}|$ is required. Since the lower triangular matrices $\mathbf{L}_{i_l}$ and $\mathbf{L}^0_{i_l}$ are, respectively, the Cholesky decomposition of $\mathbf{B}^{-1}_{i_l,T}$ and $(\mathbf{B}^0_{i_l,T})^{-1}$, we obtain:

$$1/|\mathbf{B}_{i_l,T}|^{1/2} = |(\mathbf{B}_{i_l,T})^{-1}|^{1/2} = |\mathbf{L}_{i_l}|, \tag{52}$$

where $|\mathbf{L}_{i_l}|$ is the product of the diagonal elements of $\mathbf{L}_{i_l}$. Computing $|\mathbf{B}^0_{i_l,T}|$ in the same ways and using (51) proves (50).

References

AKAIKE, H. (1987). Factor analysis and AIC. *Psychometrika* **52**, 317–332.

BARTHOLOMEW, D. J. (1987). *Latent Variable Models and Factor Analysis*. London: Charles Griffin.

BHATTACHARYA, A. & DUNSON, D. (2009). Sparse Bayesian infinite factor models. Duke Statistics Discussion Papers 2009-23, Duke University.

CARVALHO, C. M., CHANG, J., LUCAS, J. E., NEVINS, J. WANG, Q., & WEST, M. (2008). High-dimensional sparse factor modeling: Applications in gene expression genomics. *Journal of the American Statistical Association* **103**, 1438–1456.

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

FOSTER, D. P. & GEORGE, E. I. (1994). The risk inflation criterion for multiple regression. *The Annals of Statistics* **22**, 1947–1975.

FRÜHWIRTH-SCHNATTER, S. & TÜCHLER, R. (2008). Bayesian parsimonious covariance estimation for hierarchical linear mixed models. *Statistics and Computing* **18**, 1–13.

GEWEKE, J. F. & SINGLETON, K. J. (1980). Interpreting the likelihood ratio statistic in factor models when sample size is small. *Journal of the American Statistical Association* **75**, 133–137.

GEWEKE, J. F. & ZHOU, G. (1996). Measuring the pricing error of the arbitrage pricing theory. *Review of Financial Studies* **9**, 557–587.

GHOSH, J. & DUNSON, D. B. (2009). Default prior distributions and efficient posterior computation in Bayesian factor analysis. *Journal of Computational and Graphical Statistics* **18**, 306–320.

IHARA, M. & KANO, Y. (1995). Identifiability of full, marginal, and conditional factor analysis model. *Statistics and Probability Letters* **23**, 343–350.

JÖRESKOG, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika* **32**, 443–482.

LEY, E. & STEEL, M. F. J. (2009). On the effect of prior assumptions in Bayesian model averaging with applications to growth regression. *Journal of Applied Econometrics* **24**, 651–674.

LOPES, H. F. & CARVALHO, C. M. (2007). Factor stochastic volatility with time varying loadings and Markov switching regimes. *Journal of Statistical Planning and Inference* **137**, 3082–91.

LOPES, H. F., SALAZAR, E. & GAMERMAN, D. (2008). Spatial dynamic factor analysis. *Bayesian Analysis* **3**, 759–792.

LOPES, H. F. & WEST, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica* **14**, 41–67.

MARTIN, J. K. & MCDONALD, R. P. (1975). Bayesian estimation in unrestricted factor analysis: a treatment for Heywood cases. *Psychometrika* **40**, 505–517.

MAXWELL, A. E. (1961). Recent trends in factor analysis. *Journal of the Royal Statistical Society, Ser. A* **124**, 49–59.

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

PRESS, S. J. & SHIGEMASU, K. (1989). Bayesian inference in factor analysis. In *Contributions to Probability and Statistics. Essays in Honor of Ingram Olkin*, L. J. Gleser, M. D. Perlman, S. Press & A. Sampson, eds. New York: Springer, pp. 271–287.

ROWE, D. B. (2003). *Multivariate Bayesian Statistics*. London: Chapman & Hall.

SCOTT, J. G. & BERGER, J. O. (2006). An exploration of aspects of Bayesian multiple testing. *Journal of Statistical Planning and Inference* **136**, 2144–2162.

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

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

VAN DYK, D. & MENG, X.-L. (2001). The art of data augmentation. *Journal of Computational and Graphical Statistics* **10**, 1–50.

WEST, M. & HARRISON, P. J. (1997). *Bayesian Forecasting and Dynamic Models*. New York: Springer, 2nd ed.