Bayesian Analysis (2024) **TBA**, Number TBA, pp. 1–44

Sparse Bayesian Factor Analysis When the Number of Factors Is Unknown

Sylvia Frühwirth-Schnatter[*], Darjus Hosszejni[†] and Hedibert Freitas Lopes[‡]

Abstract. There has been increased research interest in the subfield of sparse Bayesian factor analysis with shrinkage priors, which achieve additional sparsity beyond the natural parsimony of factor models. In this spirit, we estimate the number of common factors in the widely applied sparse latent factor model with spike-and-slab priors on the factor loadings matrix. Our framework leads to a natural, efficient and simultaneous coupling of model estimation and selection on one hand and model identification and rank estimation (number of factors) on the other hand. More precisely, by embedding the unordered generalised lower triangular loadings representation into overfitting sparse factor modelling, we obtain posterior summaries regarding factor loadings, common factors as well as the factor dimension via postprocessing draws from our efficient and customized Markov chain Monte Carlo scheme.

Keywords: hierarchical model, identifiability, point-mass mixture priors, marginal data augmentation, reversible jump MCMC, prior distribution, sparsity, Heywood problem, rotational invariance, ancillarity-sufficiency interweaving strategy, fractional priors.

MSC2020 subject classifications: Primary 62H25; secondary 62F15.

1 Introduction

Factor analysis aims at identifying common variation in multivariate observations and relating it to hidden causes, the so-called common factors, see Thurstone (1947) and, more recently, Anderson (2003). The common setup consists of a sample $\mathbf{y} = \{\mathbf{y}_1, \dots, \mathbf{y}_T\}$ of T multivariate observations $\mathbf{y}_t = (y_{1t}, \dots, y_{mt})'$ of dimension m. For a given factor dimension r, the basic factor model is defined as a latent variable model, involving the common factors $\mathbf{f}_t = (f_{1t} \cdots f_{rt})'$:

$$\mathbf{f}_t \sim \mathcal{N}_r\left(\mathbf{0}, \mathbf{I}_r\right), \quad \mathbf{y}_t = \mathbf{\Lambda}\mathbf{f}_t + \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \sim \mathcal{N}_m\left(\mathbf{0}, \mathbf{\Sigma}_0\right), \quad \mathbf{\Sigma}_0 = \mathrm{Diag}\left(\sigma_1^2, \dots, \sigma_m^2\right), \quad (1.1)$$

where the covariance matrix $\mathbf{\Sigma}_0$ of the idiosyncratic errors $\boldsymbol{\epsilon}_t$ is a diagonal matrix and $\mathbf{\Lambda}$ is the $m \times r$ matrix of factor loadings Λ_{ij} with a specific structure that facilitates econometric identification of this model; details follow. Model (1.1) implies that conditional on $\mathbf{f}_t$ the m elements of $\mathbf{y}_t$ are independent and all dependence among these variables

[*] Institute for Statistics and Mathematics, Vienna University of Economics and Business, Welthandelsplatz 1, 1020 Wien, Austria, sylvia.fruehwirth-schnatter@wu.ac.at

[†] Institute for Statistics and Mathematics, Vienna University of Economics and Business, Welthandelsplatz 1, 1020 Wien, Austria, darjus.hosszejni@wu.ac.at

[‡] School of Mathematical and Statistical Sciences, Arizona State University, Tempe, USA & Insper Institute of Education and Research, São Paulo, Brazil, hedibertfl@insper.edu.br

https://doi.org/10.1214/24-BA1423

is explained through the common factors. Assuming independence of $\mathbf{f}_t$ and $\boldsymbol{\epsilon}_t$ implies that, marginally, $\mathbf{y}_t$ arises from a multivariate normal distribution, $\mathbf{y}_t \sim \mathcal{N}_m(\mathbf{0}, \boldsymbol{\Omega})$, with zero mean and a covariance matrix $\boldsymbol{\Omega}$ with the following structure:

$$\boldsymbol{\Omega} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Sigma}_0. \tag{1.2}$$

Since r typically is (much) smaller than m, factor models yield a parsimonious representation of $\boldsymbol{\Omega}$ with (at most) $m(r+1)$ instead of the $m(m+1)/2$ parameters of an unconstrained covariance matrix. Hence, factor models proved to be very useful for covariance estimation, especially if m is large; see Fan et al. (2008), Forni et al. (2009), Bhattacharya and Dunson (2011) and Kastner (2019), among others.

The zero-mean assumption in model (1.1) can be alleviated. For data with a non-zero mean $\boldsymbol{\mu}$, the covariance matrix of $\boldsymbol{u}_t = \mathbf{y}_t - \boldsymbol{\mu}$ exhibits a factor structure as in (1.2). In a factor-augmented model with conditional mean $\boldsymbol{\mu}_t$, the zero-mean innovations $\boldsymbol{u}_t = \mathbf{y}_t - \boldsymbol{\mu}_t$ (rather than $\mathbf{y}_t$) follow model (1.1), while $\boldsymbol{\mu}_t$ is modelled separately. Examples include factor augmented mixed-outcome regression analysis (Conti et al., 2014), factor-augmented treatment effect models (Wagner et al., 2023), and mixtures of factor analyser models (Grushanina and Frühwirth-Schnatter, 2023), among others.

The recent years have seen many contributions in the field of sparse Bayesian factor analysis (BFA) which achieve additional sparsity beyond the natural parsimony of factor models. Shrinkage priors are employed that resolve two major challenges in factor analysis: First, by introducing *column sparsity* in an overfitting factor model, they lead to an automatic selection of the number of factors in situations where the true factor dimension r is unknown. Second, by introducing *row sparsity* they allow us to identify “simple structures” in the sense specified by Thurstone (1947) where in each row only a few non-zero loadings are present.

Choosing the factor dimension is in general a challenging problem, see Owen and Wang (2016) for a review. Often, the information criteria introduced by Bai and Ng (2002) are used also in a Bayesian context (Aßmann et al., 2016; Chan et al., 2018), other authors employ marginal likelihoods (Lee and Song, 2002; Lopes and West, 2004). Learning about the factor dimension is intrinsic in sparse BFA under priors that impose column sparsity in the overfitting model

$$\mathbf{y}_t = \boldsymbol{\beta}_H \mathbf{f}_t^H + \boldsymbol{\epsilon}_t, \qquad \boldsymbol{\epsilon}_t \sim \mathcal{N}_m(\mathbf{0}, \boldsymbol{\Sigma}_H), \quad \mathbf{f}_t^H \sim \mathcal{N}_H(\mathbf{0}, \mathbf{I}_H), \tag{1.3}$$

where $\boldsymbol{\beta}_H$ is an $m \times H$ loading matrix with elements β_{ij} and $\boldsymbol{\Sigma}_H$ is a diagonal matrix with strictly positive diagonal elements.

Bayesian approaches with $H = \infty$ apriori allow infinitely many columns in $\boldsymbol{\beta}_H$ which are increasingly pulled toward zero as the column index increases using priors such as the Indian buffet process prior (Griffiths and Ghahramani, 2006; Ročková and George, 2017), the multiplicative gamma process prior (Bhattacharya and Dunson, 2011; Durante, 2017; De Vito et al., 2021), or cumulative shrinkage process priors (Legramanti et al., 2020; Kowal and Canale, 2023). These prior choices ensure that the number k of non-zero columns in model (1.3), denoted by $\boldsymbol{\beta}_k$, is random apriori and takes finite values smaller than H with probability one.

Other authors allow H to be a finite number, assumed to be larger than the true number of factors r (Frühwirth-Schnatter and Lopes, 2010; Conti et al., 2014; Kaufmann and Schuhmacher, 2019) and we use such an *overfitting* BFA model in the present paper. To achieve column sparsity, we exploit a finite version of the two-parameter Beta prior to define a shrinkage process prior on $\boldsymbol{\beta}_H$ that induces increasing shrinkage of the factor loadings toward zero as the column index increases (Frühwirth-Schnatter, 2023). We employ spike-and-slab priors, where the elements β_{ij} of $\boldsymbol{\beta}_H$ are allowed to be exactly zero. Many authors considered spike-and-slab priors, where the identification of the non-zero factor loadings is treated as a variable selection problem, not only for basic factor models (West, 2003; Carvalho et al., 2008; Frühwirth-Schnatter and Lopes, 2010) but also for dedicated factor models with correlated factors (Conti et al., 2014) and dynamic factor models (Kaufmann and Schuhmacher, 2019). As opposed to continuous shrinkage priors on β_{ij} that are applied often in sparse BFA, spike-and-slab priors allow an explicit assessment of row sparsity in the loading matrix and identification of irrelevant variables y_{it} which are uncorrelated with the remaining variables in $\mathbf{y}_t$, since the entire row of the factor loading matrix is zero for these variables (Kaufmann and Schuhmacher, 2017).

A further challenge in sparse BFA is post-processing the posterior draws of $\boldsymbol{\beta}_H$ to obtain final estimates of the unknown factor dimension r and a unique rotation $\boldsymbol{\Lambda}$ of the unknown loading matrix. There is a growing literature in machine learning, statistics, and applied econometrics where more or less heuristic post-processing procedures are applied for this purpose (Aßmann et al., 2016; Kaufmann and Schuhmacher, 2019; Poworoznek et al., 2021; Papastamoulis and Ntzoufras, 2022). Often no constraints are imposed on $\boldsymbol{\beta}_H$ during sampling; however, leaving $\boldsymbol{\beta}_H$ unconstrained makes it difficult to recover the true number of factors and to estimate $\boldsymbol{\Lambda}$.

In the present paper, we pursue a more mathematical approach which relies on rigorous econometric identification in sparse BFA and also allows uncertainty quantification by deriving posterior distributions both for r and $\boldsymbol{\Lambda}$. Econometric identification yields a unique decomposition of the covariance matrix $\boldsymbol{\Omega}$ in (1.2) into the cross-covariance matrix $\boldsymbol{\Lambda}\boldsymbol{\Lambda}'$ and the covariance matrix $\boldsymbol{\Sigma}_0$ of the uncorrelated idiosyncratic errors and identifies a unique factor loading matrix $\boldsymbol{\Lambda}$ from $\boldsymbol{\Lambda}\boldsymbol{\Lambda}'$. Even if the decomposition is unique (which need not be the case), it is well-known that $\boldsymbol{\Lambda}$ is identified only up to a rotation. Following the pioneering work of Anderson and Rubin (1956), identification is achieved by imposing additional conditions (Reiersøl, 1950; Neudecker, 1990; Geweke and Zhou, 1996; Bai and Ng, 2013). The most popular condition requires $\boldsymbol{\Lambda}$ to be a lower triangular matrix with positive diagonal elements; however, such a *PLT structure* is rather restrictive (Jöreskog, 1969; Carvalho et al., 2008).

Recently, a new identification strategy based on unordered generalized lower triangular (UGLT) structures (Frühwirth-Schnatter and Lopes, 2018; Frühwirth-Schnatter et al., 2023) was introduced that addresses not only rotational invariance but also variance identification to ensure a unique decomposition of $\boldsymbol{\Omega}$ into $\boldsymbol{\Lambda}\boldsymbol{\Lambda}'$ and $\boldsymbol{\Sigma}_0$; a problem of which the literature is still less aware. By imposing such a UGLT structure on the non-zero columns of the loading matrix $\boldsymbol{\beta}_H$ in model (1.3), we achieve identification in the present paper. The UGLT structure only requires the top non-zero elements in each non-zero column of $\boldsymbol{\beta}_H$ to lie in arbitrary but distinct rows, and is a much weaker condition than a PLT structure. As shown in Frühwirth-Schnatter et al. (2023), on the one hand it is weak enough to ensure that any loading matrix can be rotated into a UGLT representation, on the other hand it is strong enough to ensure "controlled unidentifiability" up to column and sign switching which can be easily resolved.

For practical Bayesian inference, we develop a new and efficient Markov chain Monte Carlo (MCMC) procedure that delivers posterior draws from model (1.3) under point mass mixture priors, which is known to be particularly challenging (Pati et al., 2014). As part of our algorithm, we design a (simple) reversible jump MCMC sampler to navigate through the space of UGLT loading matrices of varying factor dimension. We achieve mathematically rigorous identification through post-processing the posterior draws and ensuring variance identification through the algorithm of Hosszejni and Frühwirth-Schnatter (2022). In this way, we recover the factor dimension r, the idiosyncratic variances $\mathbf{\Sigma}_0$ and an ordered GLT representation $\mathbf{\Lambda}$ of the loading matrix from the posterior draws. Our sampling as well as our identification strategy works under arbitrary choices for the slab distribution of β_{ij}, including fractional priors (Frühwirth-Schnatter and Lopes, 2010), the horseshoe prior (Zhao et al., 2016) and the Lasso prior (Ročková and George, 2017). In high-dimensional models, we work with structured priors with column-specific shrinkage (Legramanti et al., 2020) and employ the triple gamma prior (Cadonna et al., 2020) to achieve local separation of signal and noise.

The rest of the paper is organized as follows. Section 2 introduces sparse Bayesian exploratory factor analysis models with UGLT structures, while prior choices are discussed in Section 3. Section 4 introduces our innovative MCMC sampler for this model class and discusses post-processing to achieve identification. Section 5 illustrates the usefulness of the proposed methodology in various simulation settings and considers applications to exchange rate data and NYSE100 returns. Section 6 concludes.

2 Sparse Bayesian EFA models with UGLT structures

2.1 Model definition

Throughout the paper, we work with the exploratory factor analysis (EFA) model (1.3) with finite ($H < \infty$) potential common factors, i.e.

$$\mathbf{y}_t = \boldsymbol{\beta}_H \mathbf{f}_t^H + \boldsymbol{\epsilon}_t, \qquad \boldsymbol{\epsilon}_t \sim \mathcal{N}_m\left(\mathbf{0}, \mathbf{\Sigma}_H\right), \quad \mathbf{f}_t^H \sim \mathcal{N}_H\left(\mathbf{0}, \mathbf{I}_H\right). \tag{2.1}$$

We impose an exchangeable shrinkage process prior on the columns of $\boldsymbol{\beta}_H$ to achieve column sparsity with $k < H$ non-zero columns, collected in the $m \times k$ submatrix $\boldsymbol{\beta}_k$, see Section 3.1 for details. We summarize sparsity by the so-called sparsity matrix $\boldsymbol{\delta}_H$ which is a binary indicator matrix of 0s and 1s of the same dimension as $\boldsymbol{\beta}_H$ and contains the information which elements of a factor loading matrix are equal to zero and which elements are unconstrained, i.e. if $\delta_{ij} = 0$, then $\beta_{ij} = 0$, while $\beta_{ij} \in \mathbb{R}$ if $\delta_{ij} = 1$.

Let $\boldsymbol{\delta}_k$ be the sparsity matrix corresponding to the non-zero columns $\boldsymbol{\beta}_k$ of $\boldsymbol{\beta}_H$. To achieve identification in a sparse EFA model, we assume that $\boldsymbol{\delta}_k$ exhibits a UGLT structure (Frühwirth-Schnatter et al., 2023). Compared to the common literature, where all elements of $\boldsymbol{\delta}_H$ are left unspecified, this imposes the constraint on $\boldsymbol{\delta}_H$ that the top non-zero element in all non-zero columns $\boldsymbol{\delta}_k$ lie in different rows, see Figure 1 for examples of such matrices. More formally, let l_j denote the row index (also called pivot) of the top non-zero entry in the jth column of $\boldsymbol{\delta}_k$ (i.e. $\delta_{ij} = 0, \forall\, i < l_j$). $\boldsymbol{\delta}_k$ is said to be a UGLT structure, if the pivot elements $\mathbf{l}_k = (l_1, \dots, l_k)$ lie in different rows. As discussed in Frühwirth-Schnatter et al. (2023), this rather weak condition on $\boldsymbol{\delta}_H$ is sufficient for a mathematically rigorous identification of the parameters $(r, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}_0)$ in the underlying basic factor model (1.1) from the overfitting BFA model (2.1).

First, Frühwirth-Schnatter et al. (2023) prove that the so-called *3579 counting rule* is sufficient for variance identification which is easily violated for sparse Bayesian factor models. A sparsity matrix $\boldsymbol{\delta}_k$ satisfies the 3579 counting rule if the following condition is satisfied: for each $q = 1, \dots, k$ and for each submatrix consisting of q columns of $\boldsymbol{\delta}_k$, the number of nonzero rows in this sub-matrix is at least equal to $2q+1$. The 3579 counting rule states that every column of $\boldsymbol{\delta}_k$ should have at least 3, every pair of columns at least 5, every subset of 3 columns at least 7 elements and so forth. Hosszejni and Frühwirth-Schnatter (2022) provide an efficient algorithm to verify this rule. If the sparsity matrix $\boldsymbol{\delta}_k$ obeys the 3579 counting rule, then this implies that $\boldsymbol{\Sigma}_k$ and $\boldsymbol{\beta}_k\boldsymbol{\beta}_k'$ are uniquely identified from the covariance matrix $\boldsymbol{\Omega} = \boldsymbol{\beta}_k\boldsymbol{\beta}_k' + \boldsymbol{\Sigma}_k$ implied by the non-zero columns $\boldsymbol{\beta}_k$ of $\boldsymbol{\beta}_H$ and by $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}_H$. Since variance identification implies that $\boldsymbol{\Lambda}\boldsymbol{\Lambda}' = \boldsymbol{\beta}_k\boldsymbol{\beta}_k'$, it follows that $r = k$ and $\boldsymbol{\beta}_r = \boldsymbol{\Lambda}\mathbf{P}$ for some orthogonal matrix $\mathbf{P}$ (Anderson and Rubin, 1956, Lemma 5.1).

Second, Frühwirth-Schnatter et al. (2023) show that imposing a UGLT structure on $\boldsymbol{\beta}_k$ and $\boldsymbol{\Lambda}$ leads to rotational identification up to signed permutations $\boldsymbol{\beta}_k\mathbf{P}_\pm\mathbf{P}_\rho$, where the permutation matrix $\mathbf{P}_\rho$ corresponds to one of $k!$ possible column permutations in $\boldsymbol{\beta}_k$ and the reflection matrix $\mathbf{P}_\pm = \text{Diag}\,(\pm 1, \dots, \pm 1)$ to one of the 2^k possibilities to reverse the signs in a subset of columns. Provided that $\boldsymbol{\beta}_k$ is variance identified, $r = k$ and $\boldsymbol{\Lambda}$ is uniquely recovered by reordering the columns of $\boldsymbol{\beta}_r$ such that the pivots $l_1 < \ldots < l_r$ are increasing, while $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_r$. These insights are exploited in Section 4.4, where the posterior draws from a sparse EFA model with UGLT structure are screened in a post-processing manner to ensure full identification and to learn about the unknown factor dimension r, the loading matrix $\boldsymbol{\Lambda}$ as well as $\boldsymbol{\Sigma}_0$ from the data.

For illustration, we show in Figure 1 three posteriors draws of $\boldsymbol{\delta}_k$ for a sparse EFA factor analysis with $H = 14$ for artificial data with $m = 30$ that are part of an extensive simulation study in Section 5.1. All posterior draws exhibit $k < H$ non-zero columns as a result of imposing prior column sparsity. For the posterior draw $\boldsymbol{\delta}_k$ on the left, the number of non-zero columns $k = 5$ can be considered a posterior draw of the factor dimension r, since $\boldsymbol{\delta}_k$ obeys the 3579 counting rule. The pivots $(l_1, l_2, l_3, l_4, l_5) = (24, 6, 12, 18, 1)$ can be used to obtain a uniquely rotated posterior draw of $\boldsymbol{\Lambda}$, by reordering the columns of $\boldsymbol{\beta}_k$ such that the pivots (1,6,12,18,24) are increasing. The posterior draw $\boldsymbol{\delta}_k$ in the middle contains six non-zero columns which violate the 3579 counting rule, since the 12th column has only two non-zero elements. Such posterior draws are rejected during post-processing, as they do not allow unique identification of $\boldsymbol{\Lambda}$ from $\boldsymbol{\beta}_k$. The posterior draw $\boldsymbol{\delta}_k$ on the right also contains six non-zero columns with the 11th column being a so-called *spurious column* with a single non-zero factor loading. Such posterior

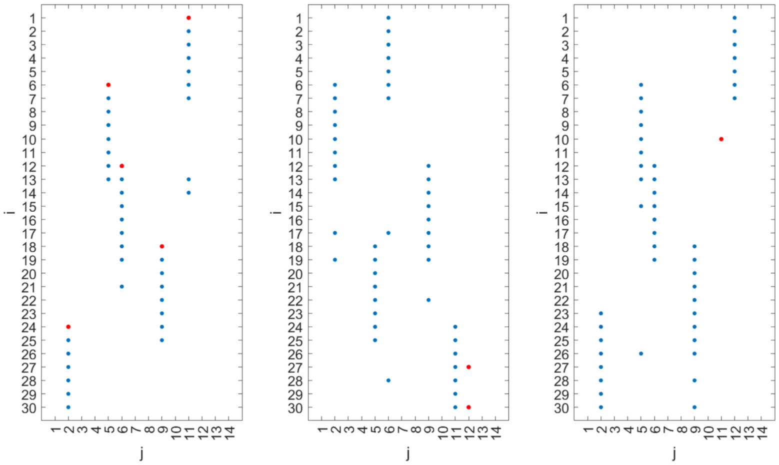

Figure 1: Posteriors draws of $\boldsymbol{\delta}_k$ from sparse BFA with $m = 30$ and $H = 14$ (zero loadings are left blank). Left: $\boldsymbol{\delta}_k$ with $k = 5$ obeying the 3579 counting rule with the pivot rows $(l_1, l_2, l_3, l_4, l_5) = (24, 6, 12, 18, 1)$ (marked red); center: $\boldsymbol{\delta}_k$ with $k = 6$ violating the 3579 counting rule due to a column with only two non-zero elements (marked red); right: $\boldsymbol{\delta}_k$ with $k = 6$ containing a spurious column (marked red).

draws obviously violate the 3579 counting rule, nevertheless they carry useful information about the factor dimension r and $\boldsymbol{\Lambda}$. More specifically, Frühwirth-Schnatter et al. (2023, Theorem 4) show as a third contribution that imposing a UGLT structure on the non-zero columns $\boldsymbol{\beta}_k$ of $\boldsymbol{\beta}_H$ in an EFA model favors posterior draws with such spurious columns, if the number of non-zero columns k overfits the true factor dimension r. For instance, if $k = r + 1$, then mathematically $\boldsymbol{\beta}_k$ and $\boldsymbol{\Sigma}_k$ take the following form:

$$\boldsymbol{\beta}_k = \begin{pmatrix} \boldsymbol{\Lambda} & \boldsymbol{\Xi} \end{pmatrix} \mathbf{P}_{\pm}\mathbf{P}_{\rho}, \quad \boldsymbol{\Xi} = \begin{pmatrix} \mathbf{0} \\ \Xi_{l_{\text{sp}}} \\ \mathbf{0} \end{pmatrix}, \quad \boldsymbol{\Sigma}_k = \text{Diag}\left(\sigma_1^2, \ldots, \sigma_{l_{\text{sp}}}^2 - \Xi_{l_{\text{sp}}}^2, \ldots, \sigma_m^2\right), \tag{2.2}$$

with a single non-zero factor loading $\Xi_{l_{\text{sp}}}$ satisfying $0 < \Xi_{l_{\text{sp}}}^2 < \sigma_{l_{\text{sp}}}^2$ which lies in a pivot row l_{sp} different from the pivot rows $\mathbf{l}_r = (l_1, \ldots, l_r)$ in $\boldsymbol{\Lambda}$. A similar representation holds for higher degrees $k > r$ of overfitting, with $\boldsymbol{\Xi}$ containing $s = k - r$ spurious columns that obey a UGLT structure, i.e. the pivots $\mathbf{l}_{\Xi}$ of $\boldsymbol{\Xi}$ lie in different rows and are distinct from the pivots $\mathbf{l}_r$ of $\boldsymbol{\Lambda}$.

Hence, if a posterior draw $\boldsymbol{\beta}_k$ from the EFA model (2.1) contains s spurious columns $\boldsymbol{\Xi}$, then they can be absorbed into the idiosyncratic errors by defining their covariance matrix as $\boldsymbol{\Sigma}_r = \boldsymbol{\Sigma}_k + \boldsymbol{\Xi}\boldsymbol{\Xi}'$. This leaves $r = k - s$ active columns $\boldsymbol{\beta}_r$ (i.e. columns with at least two non-zero loadings) in $\boldsymbol{\beta}_H$, which are extracted and postprocessed as above: if $\boldsymbol{\beta}_r$ obeys the 3579 counting rule, then $\boldsymbol{\Lambda}$ is identified up to a signed permutation from $\boldsymbol{\beta}_r = \boldsymbol{\Lambda}\mathbf{P}_{\pm}\mathbf{P}_{\rho}$, while $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_r$, and the number of active columns r provides a posterior draw of the unknown factor dimension. Otherwise, $\boldsymbol{\beta}_r$ is rejected.

2.2 Relating exploratory to confirmatory Bayesian factor analysis

The sparsity matrix $\boldsymbol{\delta}_H$ of the loading matrix $\boldsymbol{\beta}_H$ in the EFA model (2.1) allows us to classify factors into active (the corresponding column of $\boldsymbol{\delta}_H$ has at least two non-zero loadings), spurious (the corresponding column of $\boldsymbol{\delta}_H$ has a single non-zero loading) and inactive ones (the corresponding column of $\boldsymbol{\delta}_H$ is zero). This allows us to split $\boldsymbol{\delta}_H$ and $\boldsymbol{\beta}_H$ into $m \times r$ submatrices $\boldsymbol{\delta}_r$ and $\boldsymbol{\beta}_r$ with r active columns, $m \times r_{sp}$ submatrices $\boldsymbol{\delta}_\Xi$ and $\boldsymbol{\Xi}$ with r_{sp} spurious columns, and submatrices with $j_0 = H - r - r_{sp}$ zero columns, while the factors $\mathbf{f}_t^H$ are split into $\mathbf{f}_t^r$, $\mathbf{f}_t^\Xi$ and $\mathbf{f}_t^0$.

Exploiting representation (2.2), we extract the following model of factor dimension r which is embedded in any EFA model with UGLT structure,

$$\mathbf{f}_t^r \sim \mathcal{N}_r\left(\mathbf{0}, \mathbf{I}_r\right), \quad \mathbf{y}_t = \boldsymbol{\beta}_r \mathbf{f}_t^r + \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \sim \mathcal{N}_m\left(\mathbf{0}, \boldsymbol{\Sigma}_r\right), \quad \boldsymbol{\Sigma}_r = \boldsymbol{\Sigma}_H + \boldsymbol{\Xi}\boldsymbol{\Xi}', \tag{2.3}$$

by absorbing the r_{sp} spurious columns $\boldsymbol{\Xi}$ into the idiosyncratic error term. We call (2.3) the confirmatory factor analysis (CFA) model induced by the active columns $\boldsymbol{\beta}_r$ in the EFA model. The likelihood function is invariant to moving from the EFA model (2.1) to the CFA model (2.3), since the implied covariance matrix $\boldsymbol{\Omega} = \boldsymbol{\beta}_H \boldsymbol{\beta}'_H + \boldsymbol{\Sigma}_H = \boldsymbol{\beta}_r \boldsymbol{\beta}'_r + \boldsymbol{\Sigma}_r$ remains the same. On the other hand, we can move from the CFA model (2.3) to the EFA model (2.1) without changing the likelihood function by adding $r_{sp} \in \{1, \ldots, k-r\}$ spurious columns $\boldsymbol{\delta}_\Xi$ to $\boldsymbol{\delta}_r$. Moving forth and back between the EFA model (2.1) and the CFA model (2.3) is the cornerstone of an efficient MCMC algorithm developed in Section 4. In Section 3, priors are defined that are (largely) invariant to these moves.

For $r_{sp} = 1$, for instance, a single spurious column $\boldsymbol{\delta}_\Xi$ and $H - r - 1$ zero columns are added to $\boldsymbol{\delta}_r$ to define an EFA model with H columns. The only non-zero indicator in $\boldsymbol{\delta}_\Xi$ can lie in any row l_{sp} that is different from the pivots $\mathbf{l}_r$ in $\boldsymbol{\delta}_r$. A spurious column $\boldsymbol{\Xi}$ is added to $\boldsymbol{\beta}_r$ to define $\boldsymbol{\beta}_H$, while the covariance matrix of the idiosyncratic errors in the EFA model is defined as $\boldsymbol{\Sigma}_H = \boldsymbol{\Sigma}_r - \boldsymbol{\Xi}\boldsymbol{\Xi}'$. The only non-zero loading $\Xi_{l_{\text{sp}}}$ in $\boldsymbol{\Xi}$ can take any value such that the l_{sp}-th diagonal element of $\boldsymbol{\Sigma}_H$ remains positive, i.e. $\boldsymbol{\Sigma}_{H,l_{\text{sp}},l_{\text{sp}}} = \sigma^2_{l_{\text{sp}}} - (\Xi_{l_{\text{sp}}})^2 > 0$. This entire move only affects the l_{sp}-th row $\boldsymbol{\beta}_{r,l_{\text{sp}},\cdot}$ of $\boldsymbol{\beta}_r$. More specifically, for $t = 1, \ldots, T$:

$$\begin{aligned} y_{l_{\text{sp}},t} &= \boldsymbol{\beta}_{r,l_{\text{sp}},\cdot} \mathbf{f}_t^r + \epsilon_{l_{\text{sp}},t}, \quad \epsilon_{l_{\text{sp}},t} \sim \mathcal{N}\left(0, \sigma^2_{l_{\text{sp}}}\right), \\ y_{l_{\text{sp}},t} &= \boldsymbol{\beta}_{r,l_{\text{sp}},\cdot} \mathbf{f}_t^r + \Xi_{l_{\text{sp}}} f_t^\Xi + \tilde{\epsilon}_{l_{\text{sp}},t}, \quad \tilde{\epsilon}_{l_{\text{sp}},t} \sim \mathcal{N}\left(0, \sigma^2_{l_{\text{sp}}} - (\Xi_{l_{\text{sp}}})^2\right). \end{aligned} \tag{2.4}$$

By integrating model (2.4) with respect to the spurious factor f_t^Ξ, it can be verified that both models imply the same distribution $p(y_{l_{\text{sp}},t} | \boldsymbol{\beta}_{r,l_{\text{sp}},\cdot}, \mathbf{f}_t^r, \sigma^2_{l_{\text{sp}}})$, independently of $\Xi_{l_{\text{sp}}}$.

3 Prior specifications

3.1 Column sparsity through exchangeable shrinkage process priors

Bayesian inference is performed in the EFA model (2.1) with a finite number H of potential factors. We start with the description of an unconstrained model and below we introduce the UGLT structure as a constraint. Our starting point is the following Dirac-spike-and-slab prior for the factor loadings β_{ij} in $\boldsymbol{\beta}_H$,

$$\beta_{ij}|\tau_j \sim (1-\tau_j)\Delta_0 + \tau_j \mathrm{P}_{\mathrm{slab}}(\beta_{ij}), \tag{3.1}$$

where Δ_0 is a Dirac-spike at zero and $\mathrm{P}_{\mathrm{slab}}$ is a continuous slab distribution. Cumulative shrinkage where the columns of the loading matrix are increasingly pulled toward zero can be achieved in a factor model with $H < \infty$ by placing an exchangeable shrinkage process (ESP) prior on the slab probabilities $\tau_1, \dots, \tau_H$:

$$\tau_j|H \sim \mathcal{B}\left(a_H, b_H\right), \quad j = 1, \dots, H. \tag{3.2}$$

The ESP prior turns model (2.1) into a *sparse* EFA model, where the number k of non-zero columns in $\boldsymbol{\delta}_H$ is random apriori, taking values smaller than H with high probability. As shown by Frühwirth-Schnatter (2023), prior (3.2) has a representation as a finite cumulative shrinkage process (CUSP) prior (Legramanti et al., 2020). A prominent example of such an ESP prior is the finite two-parameter-beta (2PB) prior,

$$\tau_j|H \sim \mathcal{B}\left(\gamma\frac{\alpha}{H}, \gamma\right), \quad j = 1, \dots, H, \tag{3.3}$$

which converges to the 2PB prior (Ghahramani et al., 2007) for $H \to \infty$. For $\gamma = 1$, the finite one-parameter-beta (1PB) prior results which converges to the Indian buffet process prior (Teh et al., 2007) for $H \to \infty$ and has been employed by Ročková and George (2017) in sparse Bayesian factor analysis.

To adapt the ESP prior to the data at hand, the hyperparameters α and γ are equipped with the hyperpriors $\alpha \sim \mathcal{G}\left(a^{\alpha}, b^{\alpha}\right)$ and $\gamma \sim \mathcal{G}\left(a^{\gamma}, b^{\gamma}\right)$, since they are instrumental in controlling prior column sparsity. For the 1PB prior, for instance, the decreasing order statistics $\tau_{(1)} > \ldots > \tau_{(H)}$ of the slab probabilities can be expressed by the following stick-breaking representation in terms of independent beta random variables for $j = 1, \dots, H$ (Frühwirth-Schnatter, 2023):

$$\tau_{(j)} = \prod_{\ell=1}^{j} \nu_\ell, \quad \nu_\ell \sim \mathcal{B}\left(\alpha\frac{H-\ell+1}{H}, 1\right), \ell = 1, \dots, H. \tag{3.4}$$

With the largest slab probability following $\tau_{(1)} \sim \mathcal{B}\left(\alpha, 1\right)$, subsequent slab probabilities $\tau_{(j)} = \tau_{(j-1)}\nu_j$ are increasingly pulled toward zero as j increases and the 1PB prior induces considerable column sparsity, especially if $\alpha < H$.

Imposing a UGLT structure For given numbers r and r_{sp} of, respectively, active and spurious columns in $\boldsymbol{\beta}_H$, we define a prior $p(\mathbf{l}_\Xi|\mathbf{l}_r, r_{sp})p(\mathbf{l}_r|r)$ on the pivots $\mathbf{l}_r = (l_1, \dots, l_r)$ and $\mathbf{l}_\Xi = (l_{\Xi,1}, \dots, l_{\Xi,r_{sp}})$ such that the non-zero columns $\boldsymbol{\delta}_k$ of the sparsity matrix $\boldsymbol{\delta}_H$ exhibit a UGLT structure. The prior $p(\mathbf{l}_r|r)$ is defined as follows. Let $\mathcal{L}(\mathbf{l}) = \{i \in \{1, 2, \dots, m\} : i \notin \mathbf{l}\}$ be the set of all rows that are not used as pivots. Condition UGLT implies that each l_j has to be different from the pivots $\mathbf{l}_{r,-j}$ outside of column j and we assume a uniform prior distribution over all admissible pivots $l_j \in \mathcal{L}(\mathbf{l}_{r,-j})$:

$$p(l_j|\mathbf{l}_{r,-j}) = \frac{1}{|\mathcal{L}(\mathbf{l}_{r,-j})|} = \frac{1}{m-r+1}. \tag{3.5}$$

The conditional prior $p(\mathbf{l}_\Xi|\mathbf{l}_r, r_{sp})$ is uniform over all admissible values, i.e. $l_{\Xi,1}|\mathbf{l}_r$ is uniform over $\mathcal{L}(\mathbf{l}_r)$; $l_{\Xi,2}|l_{\Xi,1}, \mathbf{l}_r$ is uniform over $\mathcal{L}(\mathbf{l}_r \cup \{l_{\Xi,1}\})$, and so forth. Given the pivots l_j in all active columns $\boldsymbol{\delta}_r$, by definition $\delta_{l_j,j} = 1$ and $\delta_{ij} = 0$ for $i < l_j$, while the $m - l_j$ indicators δ_{ij} below l_j are subject to variable selection,

$$\Pr(\delta_{ij} = 1|l_j, \tau_j) = \begin{cases} 0, & i < l_j, \\ 1, & i = l_j, \\ \tau_j, & i = l_j + 1, \dots, m, \end{cases} \tag{3.6}$$

with column-specific probability τ_j following the ESP prior (3.2). With $d_j - 1$ successes and $m - l_j - d_j + 1$ failures in the experiment defined in (3.6), where $d_j = \sum_{i=1}^m \delta_{ij}$ is the number of non-zero indicators in columns j, the prior for the jth column $\boldsymbol{\delta}^r_{\cdot,j}$ of $\boldsymbol{\delta}^r$ can be expressed both conditionally as well as marginalized w.r.t. τ_j:

$$\Pr(\boldsymbol{\delta}^r_{\cdot,j}|l_j, \tau_j) = \tau_j^{d_j-1}(1 - \tau_j)^{m-l_j-d_j+1}, \tag{3.7}$$

$$\Pr(\boldsymbol{\delta}^r_{\cdot,j}|l_j) = \frac{B(a_H + d_j - 1, b_H + m - l_j - d_j + 1)}{B(a_H, b_H)}. \tag{3.8}$$

3.2 Choosing the slab distribution

To define a prior on the loading matrix $\boldsymbol{\beta}_H$ given $\boldsymbol{\delta}_H$, we first define a prior $p(\boldsymbol{\beta}_r|\boldsymbol{\Sigma}_r, \boldsymbol{\delta}_r)$ on the loading matrix $\boldsymbol{\beta}_r$ in the CFA model (2.3) containing the active columns of $\boldsymbol{\beta}_H$, conditional on $\boldsymbol{\Sigma}_r = \text{Diag}\left(\sigma_1^2, \dots, \sigma_m^2\right)$ and $\boldsymbol{\delta}_r$. When expanding the CFA model to an EFA model with r_{sp} columns, we define a prior $p(\boldsymbol{\Xi}|\boldsymbol{\beta}_r, \boldsymbol{\Sigma}_r, \mathbf{l}_\Xi)$ on the spurious loadings conditional on $\boldsymbol{\beta}_r$, $\boldsymbol{\Sigma}_r$, and $\mathbf{l}_\Xi$. The spurious factor loadings are assigned a uniform prior over all values that lead to a positive definite matrix $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}_r - \boldsymbol{\Xi}\boldsymbol{\Xi}'$ in the EFA model:

$$\Xi^2_{l_{\text{sp}}}|\sigma^2_{l_{\text{sp}}} \sim \mathcal{U}\left[0, \sigma^2_{l_{\text{sp}}}\right]. \tag{3.9}$$

This ensures for all $l_{\text{sp}} \in \mathbf{l}_\Xi$ that $\boldsymbol{\Sigma}_{k,l_{\text{sp}},l_{\text{sp}}} = \sigma^2_{l_{\text{sp}}} - \Xi^2_{l_{\text{sp}}} > 0$. By this definition, both the likelihood and the prior are invariant to moving between the EFA and the CFA model for a given number of spurious columns r_{sp}, regardless of the chosen slab distribution.[1]

The Dirac-spike-and-slab prior (3.1) is formulated for the factor loading matrix $\boldsymbol{\beta}_r$ in the CFA model. A broad range of slab distributions P_{slab} (which are briefly reviewed below) has been considered for sparse Bayesian factor analysis and can be combined with the reversible jump MCMC sampler we introduce in Section 4. Since the conditional likelihood function factors into a product over all rows of $\boldsymbol{\beta}_r$, prior independence of all rows i with $q_i = \sum_j \delta_{ij} > 0$ nonzero elements is assumed. A hierarchical Gaussian prior for the vector $\boldsymbol{\beta}^{\boldsymbol{\delta}}_{i\cdot}$ of unconstrained elements takes the form $\boldsymbol{\beta}^{\boldsymbol{\delta}}_{i\cdot}|\sigma_i^2 \sim \mathcal{N}_{q_i}\left(\mathbf{0}, \mathbf{B}^{\boldsymbol{\delta}}_{i0}\sigma_i^2\right)$, where $\mathbf{B}^{\boldsymbol{\delta}}_{i0}$ is a diagonal matrix. The variance of this prior is assumed to depend on the idiosyncratic variance σ_i^2, because this allows joint drawing of $\boldsymbol{\beta}_r$ and $\sigma_1^2, \dots, \sigma_m^2$ and, even more importantly, sampling the sparsity matrix $\boldsymbol{\delta}_r$ without conditioning on the model parameters during MCMC estimation, see Algorithm 1 in Section 4.

[1] Note that this is a major improvement compared to Frühwirth-Schnatter and Lopes (2018).

A common choice for P_{slab} is to introduce a global shrinkage parameter κ,

$$\beta_{ij}|\delta_{ij} = 1, \kappa \sim \mathcal{N}\left(0, \kappa\sigma_i^2\right), \tag{3.10}$$

which is either fixed or random with hyperprior $\kappa \sim \mathcal{G}^{-1}\left(c^\kappa, b^\kappa\right)$ or $\kappa \sim \mathrm{F}\left(2a^\kappa, 2c^\kappa\right)$. A popular extension are slab distributions with a column specific shrinkage parameter θ_j,

$$\beta_{ij}|\delta_{ij} = 1, \theta_j, \sigma_i^2, \kappa \sim \mathcal{N}\left(0, \kappa\theta_j\sigma_i^2\right), \tag{3.11}$$

where $\theta_j \sim \mathcal{G}^{-1}\left(c^\theta, b^\theta\right)$ either follows an inverse gamma prior (Legramanti et al., 2020) or a triple gamma prior, $\theta_j \sim \mathrm{F}\left(2a^\theta, 2c^\theta\right)$ (Cadonna et al., 2020; Frühwirth-Schnatter, 2023). This prior acts as a variance selection prior which pulls all factors $f_{jt}, t = 1, \ldots, T$, toward 0 for small values of θ_j. To achieve additional shrinkage for individual factor loadings, local shrinkage parameters ω_{ij} arising from an F-distribution can be introduced:

$$\beta_{ij}|\delta_{ij} = 1, \omega_{ij}, \theta_j, \sigma_i^2, \kappa \sim \mathcal{N}\left(0, \kappa\theta_j\sigma_i^2\omega_{ij}\right), \quad \omega_{ij} \sim \mathrm{F}\left(2a^\omega, 2c^\omega\right). \tag{3.12}$$

Related structured priors are employed in (Zhao et al., 2016; Schiavon et al., 2022), among others. As an alternative shrinkage prior, Frühwirth-Schnatter and Lopes (2010) introduced a conditionally conjugate fractional prior $p(\boldsymbol{\beta}_{i\cdot}^{\boldsymbol{\delta}}|\sigma_i^2, b, \mathbf{f}_r) \propto p(\tilde{\mathbf{y}}_i|\mathbf{f}_r, \boldsymbol{\beta}_{i\cdot}^{\boldsymbol{\delta}}, \sigma_i^2)^b$ in the spirit of O'Hagan (1995), see Appendix C.1 for details (Frühwirth-Schnatter et al. (2024)).

3.3 The prior on the idiosyncratic variances

Finally, we define a prior on the idiosyncratic variances $\sigma_1^2, \ldots, \sigma_m^2$ in the CFA model (2.3), taking two aspects into considerations. The first aspect in choosing this prior is whether the data are standardized, as often is recommended (Schiavon and Canale, 2020). For each variable y_{it}, the loadings $\beta_{i1}, \ldots, \beta_{ir}$ together with σ_i^2 determine the *communalities* R_i^2 as the proportion of variance explained by the common factors:

$$R_i^2 = \frac{\sum_{\ell=1}^r \beta_{i\ell}^2}{\Omega_{ii}} \quad \Leftrightarrow \quad \sigma_i^2 = (1 - R_i^2)\,\Omega_{ii}, \tag{3.13}$$

where $\Omega_{ii} = \sum_{\ell=1}^r \beta_{i\ell}^2 + \sigma_i^2$ is the ith diagonal element of $\boldsymbol{\Omega}$. For standardized data, where $\Omega_{ii} = 1$, $\sigma_i^2 = 1 - R_i^2$ is a scale-free parameter and the popular exchangeable inverse gamma prior, $\sigma_i^2 \sim \mathcal{G}^{-1}\left(c^\sigma, C_0\right)$, with constant scale C_0 is a sensible choice. However, for data that are not standardized, scale dependence of $\sigma_i^2 = (1 - R_i^2)\,\Omega_{ii}$ is to be expected, in particular in the presence of strong heterogeneity in the variances $\Omega_{11}, \ldots, \Omega_{mm}$. In this case, it is preferable to use an inverse gamma prior with heterogenous scales C_{0i}:

$$\sigma_i^2 \sim \mathcal{G}^{-1}\left(c^\sigma, C_{i0}\right). \tag{3.14}$$

We may assume that $C_{i0} = b_i^\sigma$ are fixed hyperparameters. Alternatively, assuming random hyperparameters $C_{i0} \sim \mathcal{G}\left(a^\sigma, a^\sigma/b_i^\sigma\right)$ with $\mathrm{E}(C_{i0}) = b_i^\sigma$ leads to a more general prior which can be expressed as a rescaled F-distribution with the same prior expectation $\mathrm{E}(\sigma_i^2) = b_i^\sigma/(c^\sigma - 1)$ as (3.14), provided that $c^\sigma > 1$:

$$\sigma_i^2 \sim \frac{b_i^\sigma}{c^\sigma} \, \mathrm{F}\left(2a^\sigma, 2c^\sigma\right). \tag{3.15}$$

Second, a difficulty known as Heywood problem should be considered when choosing this prior. This problem frequently occurs in ML estimation, with one or more estimators $\hat{\sigma}_i^2$s of the idiosyncratic variances being negative, see e.g. (Bartholomew, 1987). Putting a prior on the idiosyncratic variances 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. Heywood problems typically occur if the constraint $1/\sigma_i^2 \geq (\boldsymbol{\Omega}^{-1})_{ii}$ is violated for the covariance matrix of $\mathbf{y}_t$ (Bartholomew, 1987, p. 54). It is clear from this inequality that the prior of $1/\sigma_i^2$ has to be bounded away from 0. For this reason, Heywood problems might be an issue under improper priors such as $p(\sigma_i^2) \propto 1/\sigma_i^2$ (Martin and McDonald, 1975; Akaike, 1987) and proper priors with $c^\sigma > 0$ are preferable.

3.4 Choice of hyperparameters

For applications, we reduce the complex structure of the above priors to five hyperparameters. We summarize our choices in Table 1 and provide details in this section.

A necessary condition for $\boldsymbol{\delta}_k$ to satisfy the 3579 counting rule discussed in Section 2.1 is the following upper bound for k:

$$k \leq \lfloor (m-1)/2 \rfloor, \tag{3.16}$$

Prior distributions	Parameters	Values
Prior for $\tau_j, j = 1, \ldots, H$		
$\tau_j \mid \alpha, \gamma, H \sim \mathcal{B}\left(\gamma \frac{\alpha}{H}, \gamma\right)$,	a^α, b^α, H	$a^\alpha = n_0$, $b^\alpha = a^\alpha (H - \mathrm{E}_q)/H/\mathrm{E}_q$
$\alpha \sim \mathcal{G}\left(a^\alpha, b^\alpha\right), \gamma \sim \mathcal{G}\left(a^\gamma, b^\gamma\right)$	a^γ, b^γ	$a^\gamma = b^\gamma = n_0$
Priors for $\sigma_i^2, i = 1, \ldots, m$		
$\sigma_i^2 \sim \mathcal{G}^{-1}\left(c_0, C_0\right)$	c_0, C_0	$c_0 = 1, C_0 = 0.3$
$\sigma_i^2 \sim \mathcal{G}^{-1}\left(c^\sigma, b_i^\sigma\right)$	c^σ, b_i^σ	$b_i^\sigma = (c^\sigma - 1)(1 - \mathrm{E}_R)\Omega_{ii}$
$\sigma_i^2 \sim (b_i^\sigma / c^\sigma)\, \mathrm{F}\left(2a^\sigma, 2c^\sigma\right)$	$c^\sigma, b_i^\sigma, a^\sigma$	$a^\sigma = n_0$
Slab priors for β_{ij}		
Fractional prior	b	$b = 1/(mT)$
$\beta_{ij} \mid \sigma_i^2, \kappa \sim \mathcal{N}\left(0, \kappa\sigma_i^2\right)$,	c^κ, b^κ	$c^\kappa = n_0$
$\kappa \sim \mathcal{G}^{-1}\left(c^\kappa, b^\kappa\right)$		$b^\kappa = c^\kappa \mathrm{E}_R/(1 - \mathrm{E}_R)/\mathrm{E}_q$
$\beta_{ij} \mid \theta_j, \sigma_i^2, \kappa \sim \mathcal{N}\left(0, \kappa\theta_j\sigma_i^2\right)$,	c^κ, b^κ	
$\kappa \sim \mathcal{G}^{-1}\left(c^\kappa, b^\kappa\right)$, $\theta_j \sim \mathrm{F}\left(2a^\theta, 2c^\theta\right)$	a^θ, c^θ	$a^\theta = n_0$, $c^\theta = 2.5$
$\beta_{ij} \mid \omega_{ij}, \theta_j, \sigma_i^2, \kappa \sim \mathcal{N}\left(0, \kappa\theta_j\sigma_i^2\omega_{ij}\right)$,	c^κ, b^κ	
$\kappa \sim \mathcal{G}^{-1}\left(c^\kappa, b^\kappa\right)$, $\theta_j \sim \mathrm{F}\left(2a^\theta, 2c^\theta\right)$,	a^θ, c^θ	$a^\omega = c^\omega = 0.5$ (horseshoe)
$\omega_{ij} \sim \mathrm{F}\left(2a^\omega, 2c^\omega\right)$	a^ω, c^ω	$a^\omega = c^\omega = 0.2$ (triple gamma)

Table 1: Prior choices depending on five hyperparameters with default values $H = \lfloor (m-1)/2 \rfloor$, $\mathrm{E}_q = 2$, $\mathrm{E}_R = 2/3$, $c^\sigma = 2.5$ and $n_0 = 6$.

which we use as default for H. As discussed, this choice encourages spurious and zero columns in $\boldsymbol{\delta}_H$ which are essential for our strategy of recovering the factor dimension from the EFA model (2.1). If m is large, than choosing H below the upper bound (3.16) is sensible from a computational viewpoint.

In the vein of Thurstone (1947), we impose a simple structure on $\boldsymbol{\beta}_H$ by assuming that in each row the number of non-zero loadings $q_i = \sum_{j=1}^{H} \delta_{ij}$ is much smaller than H and choosing the hyperparameters in $\alpha \sim \mathcal{G}\left(a^{\alpha}, b^{\alpha}\right)$ accordingly. The choice of α strongly impacts the expected row sparsity $\mathrm{E}_q = \mathrm{E}(q_i|\alpha, H)$, given by

$$\mathrm{E}_q = \frac{\alpha}{1+\alpha/H},$$

independently of γ. To match a prior guess of E_q with the prior expectation $\mathrm{E}_\alpha = \mathrm{E}(\alpha|H) = H \cdot \mathrm{E}_q/(H - \mathrm{E}_q)$ of α, we bind a given value of a^{α} to the scale parameter $b^{\alpha} = a^{\alpha}/\mathrm{E}_\alpha$. For large H, this yields $\mathrm{E}_\alpha \approx \mathrm{E}_q$ apriori, whereas E_α is larger than E_q to achieve the same level of row sparsity for smaller values of H. A sensible choice in the spirit of Thurstone (1947) is $\mathrm{E}_q = 2$. To center the 2PB prior at the 1PB prior (corresponding to $\gamma = 1$), we choose $b^{\gamma} = a^{\gamma}$ for a given value of a^{γ}.

For the exchangeable prior $\sigma_i^2 \sim \mathcal{G}^{-1}\left(c^{\sigma}, C_0\right)$, a popular choice is $c^{\sigma} = 1$ and $C_0 = 0.3$ (Bhattacharya and Dunson, 2011). Following Frühwirth-Schnatter and Lopes (2010, 2018), we select c^{σ} in prior (3.14) and (3.15) large enough to bound the prior of $1/\sigma^2$ away from 0. Depending on the data, c^{σ} can be increased if any of the posteriors $p(\sigma_i^2|\mathbf{y})$ has a second mode at 0. For a given $c^{\sigma} > 1$, Frühwirth-Schnatter and Lopes (2018) select the scale parameter in (3.14) as $b_i^{\sigma} = (c^{\sigma} - 1)/(\widehat{\boldsymbol{\Omega}^{-1}})_{ii}$. Alternatively, we choose b_i^{σ} both in (3.14) and (3.15) such that (3.13) holds on average, i.e. $\mathrm{E}(\sigma_i^2) = \mathrm{E}(1 - R_i^2)\,\mathrm{E}(\Omega_{ii})$. Based on a prior guess E_R of the average amount of explained variance, this yields $b_i^{\sigma} = (c^{\sigma} - 1)(1 - \mathrm{E}_R)\bar{\Omega}_{ii}$, where $\bar{\Omega}_{ii} = 1$ for standardized data and otherwise $\bar{\Omega}_{ii} = \widehat{\boldsymbol{\Omega}}_{ii}$. See Appendix C.1 for details on estimating $\widehat{\boldsymbol{\Omega}^{-1}}$ and $\widehat{\boldsymbol{\Omega}}$.

Regarding the hyperparameters used for the prior $\beta_{ij}|\delta_{ij} = 1$ in the slab, we choose $b = 1/(mT)$ for the fractional prior (C.4) in the spirit of Foster and George (1994). We use the same prior on the global shrinkage parameter κ for all hierarchical shrinkage priors and bind a given value of c^{κ} to the scale parameter $b^{\kappa} = c^{\kappa}\mathrm{E}_\kappa$, where

$$\mathrm{E}_\kappa = \frac{\mathrm{E}_R}{(1 - \mathrm{E}_R)\mathrm{E}_q} \tag{3.17}$$

takes the prior information E_q and E_R used in the previous two priors into account. This choice is motivated for prior (3.10) by rewriting the coefficient of determination R_i^2 given in (3.13) in terms of δ_{ij} and the standardized loadings $\beta_{ij}^{\star} = \beta_{ij}/\sqrt{\sigma_i^2\kappa} \sim \mathcal{N}(0, 1)$:

$$R_i^2 = \frac{\kappa \sum_{j=1}^{r} (\beta_{ij}^{\star})^2 \delta_{ij}}{\kappa \sum_{j=1}^{r} (\beta_{ij}^{\star})^2 \delta_{ij} + 1} \quad \Rightarrow \quad R_i^2 = \kappa(1 - R_i^2)\chi^2_{q_i}.$$

Using that the sum follows a $\chi^2_{q_i}$-distribution, and taking the expectation of both sides of the second equation yields (3.17). Various priors for the column specific shrinkage parameters θ_j have been suggested, such as $a^{\theta} = c^{\theta} = 0.5$ (Zhao et al., 2016) or $(a^{\theta} = 2.5, c^{\theta} = 0.5)$ (Kowal and Canale, 2023). Following Frühwirth-Schnatter (2023), we choose $c^{\theta} = 2.5$ to fix the prior expectation of θ_j at around 1 and to impose a finite prior variance. Finally, regarding local shrinkage, for $a^{\omega} = c^{\omega} = 0.5$ the horseshoe prior employed by Zhao et al. (2016) results; choosing $a^{\omega} = c^{\omega} < 0.5$ yields a triple gamma (Cadonna et al., 2020) which imposes more aggressive shrinkage than the horseshoe.

This reduces the choice of hyperparameters to c^{σ} controlling Heywood problems, the prior expectation E_q of row sparsity, the prior expected fraction of explained variance E_R and the hyperparameters a^{α}, a^{γ}, c^{κ}, a^{θ} and, for the rescaled F-prior (3.15), also a^{σ}. Increasing these latter hyperparameters increases prior concentration around the chosen prior expectations. In our simulations and applications, we assume the same amount of prior information n_0 for any of these priors, i.e. $a^{\alpha} = a^{\gamma} = c^{\kappa} = a^{\theta} = a^{\sigma} = n_0$. We analyze prior sensitivity in Section 5 by comparing multiple priors.

4 MCMC estimation

MCMC estimation for sparse Bayesian factor models is notoriously difficult, since sampling the sparsity matrix $\boldsymbol{\delta}_H$ corresponds to navigating through an extremely high dimensional model space. In the present paper, we develop an innovative MCMC scheme for sparse Bayesian factor models where the factor dimension is unknown, summarized in Algorithm 1. To learn the number of factors, we sample from the posterior distribution of the EFA model (2.1), given the priors introduced in Section 3. As opposed to Carvalho et al. (2008), who operate under a PLT condition on the sparsity matrix $\boldsymbol{\delta}_H$, and Kaufmann and Schuhmacher (2019), who sample $\boldsymbol{\delta}_H$ without imposing any constraint, we impose a UGLT structure on $\boldsymbol{\delta}_H$ during MCMC sampling. As discussed in Section 2.1, this allows us to address identification of the factor model in a post-processing manner, see Section 4.4. Based on appropriate initial values (see Appendix A for details), we iterate M times through the various steps of Algorithm 1 and discard the first M_0 draws as burn-in.

Algorithm 1 consists of two main blocks. Block (CFA) operates in the confirmatory factor analysis model (2.3) corresponding to $\boldsymbol{\delta}_r$. Due to the prior specification in Section 3, the number r_{sp} of spurious columns is a sufficient statistic for the remaining columns in $\boldsymbol{\delta}_H$ and no further information is needed to update the parameters in the CFA model. To ensure that the loading matrix exhibits a UGLT structure, Step (L) performs MH steps that navigate through the space of all admissible $\boldsymbol{\delta}_r$ where the pivots $\mathbf{l}_r = (l_1, \dots, l_r)$ lie in different rows, see Section 4.2. Given $\mathbf{l}_r$, the hyperparameters a_H and b_H in the ESP prior (3.2) are updated in Step (H) using an MH step, see Appendix B. Both Step (L) and (H) are performed marginalized w.r.t. the slab probabilities $\boldsymbol{\tau}_r = (\tau_1, \dots, \tau_r)$. To sample τ_j for all columns j, the ESP prior (3.2) is combined with the likelihood (3.7). In Step (D), variable selection is performed in each column j for all indicators δ_{ij} below the pivot row l_j. This step potentially turns an active factor into a spurious one and in this way decreases the number of active factors r, while increasing r_{sp}. All moves in Step (D) are implemented conditionally on τ_j (and all shrinkage parameters for hierarchical Gaussian priors), as this allows efficient multimove sampling of all indicators

Algorithm 1 MCMC for sparse Bayesian factor models with UGLT structures.

```
(CFA) Update all unknowns in the CFA model (2.3) corresponding to $\boldsymbol{\delta}_r$:
    (H) Update any unknown hyperparameters in the ESP prior (3.2) without conditioning on the
        slab probabilities $\boldsymbol{\tau}_r = (\tau_1, \dots, \tau_r)$. For $j = 1, \dots, r$, sample $\tau_j|l_j, d_j \sim \mathcal{B}(a_H + d_j - 1, b_H + m - l_j - d_j + 1)$, where $d_j = \sum_{i=1}^m \delta_{ij}$.
    (D) Loop over all columns of the sparsity matrix $\boldsymbol{\delta}_r$ in a random order:
        (a) Sample all indicators $\delta_{ij}$ below the pivot $l_j$ from $p(\delta_{ij}|l_j, \boldsymbol{\delta}^r_{\cdot,-j}, \mathbf{f}_r, \tau_j, \mathbf{y})$ conditional
            on the remaining columns $\boldsymbol{\delta}^r_{\cdot,-j}$, the factors $\mathbf{f}_r = (\mathbf{f}^r_1, \dots, \mathbf{f}^r_T)$ and $\tau_j$, without conditioning on $\boldsymbol{\beta}_r$ and $\sigma^2_1, \dots, \sigma^2_m$.
        (b) If column $\boldsymbol{\delta}^r_{\cdot,j}$ is spurious after this update, increase $r_{sp}$ by one. Remove the $j$th
            column from $\boldsymbol{\delta}_r$, the factors $f_{jt}, t = 1, \dots, T$ from $\mathbf{f}_r$ and $\tau_j$ from $\boldsymbol{\tau}_r$ to define,
            respectively, $\boldsymbol{\delta}_{r-1}$, $\mathbf{f}_{r-1}$ and $\boldsymbol{\tau}_{r-1}$ and decrease $r$ by one.
    (L) Loop over all columns $j$ of $\boldsymbol{\delta}_r$ in a random order and sample a new pivot row $l_j$ from
        $p(l_j|\boldsymbol{\delta}^r_{\cdot,-j}, \mathbf{f}_r, \mathbf{y})$ without conditioning on $\boldsymbol{\beta}_r$, $\sigma^2_1, \dots, \sigma^2_m$ and the slab probabilities $\boldsymbol{\tau}_r$. If
        column $\boldsymbol{\delta}^r_{\cdot,j}$ is spurious after this update, proceed as in Step (D-b).
    (P) Sample the model parameters $\boldsymbol{\beta}_r$ and $\sigma^2_1, \dots, \sigma^2_m$ jointly conditional on the sparsity matrix
        $\boldsymbol{\delta}_r$ and the factors $\mathbf{f}_r = (\mathbf{f}^r_1, \dots, \mathbf{f}^r_T)$ from $p(\boldsymbol{\beta}_r, \sigma^2_1, \dots, \sigma^2_m|\boldsymbol{\delta}_r, \mathbf{f}_r, \mathbf{y})$.
    (F) Sample the latent factors $\mathbf{f}_r = (\mathbf{f}^r_1, \dots, \mathbf{f}^r_T)$ conditional on the model parameters $\boldsymbol{\beta}_r$ and
        $\sigma^2_1, \dots, \sigma^2_m$ from $p(\mathbf{f}^r_1, \dots, \mathbf{f}^r_T|\boldsymbol{\beta}_r, \sigma^2_1, \dots, \sigma^2_m, \mathbf{y})$.
    (S) For hierarchical Gaussian priors, update the global shrinkage parameter $\kappa$, the column-
        specific shrinkage parameters $\theta_1, \dots, \theta_r$ and all local shrinkage parameters $\omega_{ij}$ (if any) and
        recover $C_{01}, \dots, C_{0m}$ for the F-prior (3.15) on $\sigma^2_1, \dots, \sigma^2_m$.
    (A) Perform a boosting step to enhance mixing.
(EFA) Move from the current CFA model to an EFA model with $r_{sp}$ spurious columns and try to
      change $r_{sp}$, while holding the number of active factors $r$ fixed:
    (R-S) Perform an RJMCMC step to change the number $r_{sp}$ of spurious columns through a split
          move on a zero column or a merge move on a spurious column in $\boldsymbol{\delta}_H$.
    (R-L) Given $r_{sp}$, sample the pivot rows $\mathbf{l}_\Xi|\mathbf{l}_r$ of all $r_{sp}$ spurious columns sequentially from the
          set $\mathcal{L}(\mathbf{l}_r)$, where $\mathbf{l}_r$ are the pivot rows of the active factors $\boldsymbol{\delta}_r$. Order the spurious columns
          such that $l_{\Xi,1} < \dots < l_{\Xi,r_{sp}}$.
    (R-F) Loop over all spurious columns $j_{\mathrm{sp}}$ and sample the spurious factors $\mathbf{f}_{j_{\mathrm{sp}}} =$
          $(f_{j_{\mathrm{sp}},1}, \dots, f_{j_{\mathrm{sp}},T})$ independently for all $t = 1, \dots, T$ from $f_{j_{\mathrm{sp}},t}|\mathbf{f}^r_t, \boldsymbol{\beta}_r, \sigma^2_{l_{\mathrm{sp}}}, y_{l_{\mathrm{sp}},t} \sim$
          $\mathcal{N}\left(E_{j_{\mathrm{sp}},t}, V_{j_{\mathrm{sp}}}\right)$, where $U_{j_{\mathrm{sp}}}$ is a draw from a uniform distribution on [−1,1] and
              $V_{j_{\mathrm{sp}}} = 1 - U^2_{j_{\mathrm{sp}}}, \qquad E_{j_{\mathrm{sp}},t} = U_{j_{\mathrm{sp}}}(y_{l_{\mathrm{sp}},t} - \boldsymbol{\beta}_{r,l_{\mathrm{sp}},\cdot}\mathbf{f}^r_t)/\sqrt{\sigma^2_{l_{\mathrm{sp}}}}.$    (4.1)
    (R-H) Sample $\tau_{j_{\mathrm{sp}}}|l_{\mathrm{sp}} \sim \mathcal{B}\left(a_H, b_H + m - l_{\mathrm{sp}}\right)$ for all spurious columns $j_{\mathrm{sp}}$.
    (R-D) Update all spurious columns from the last (with the largest pivot row) to the first (with the
          smallest pivot row): sample all $(\delta_{i,j_{\mathrm{sp}}}, i \in \{l_{\mathrm{sp}} + 1, \dots, m\})$ below the pivot $l_{\mathrm{sp}}$ conditional
          on $\tau_{j_{\mathrm{sp}}}$, $\boldsymbol{\delta}_r$, $\mathbf{f}_r$ and $\mathbf{f}_{j_{\mathrm{sp}}}$ without conditioning on $\boldsymbol{\beta}_r$, $\boldsymbol{\Xi}$ and $\sigma^2_1, \dots, \sigma^2_m$. If a spurious column
          $j_{\mathrm{sp}}$ is turned into an active one, then decrease $r_{sp}$ by 1, increase $r$ by 1, add $\boldsymbol{\delta}_{\cdot,j_{\mathrm{sp}}}$ to $\boldsymbol{\delta}_r$
          and $\mathbf{f}_{j_{\mathrm{sp}}}$ to $\mathbf{f}_r$. Otherwise, remove $\boldsymbol{\delta}_{\cdot,j_{\mathrm{sp}}}$ from $\boldsymbol{\delta}_\Xi$ and $\mathbf{f}_{j_{\mathrm{sp}}}$ from $\mathbf{f}_\Xi$.
      Move from the current EFA model back to the CFA model and preserve $r_{sp}$.
```

$\{\delta_{ij}, i \in \{l_j + 1, \dots, m\}\}$, using Algorithm 2 in Appendix D.2. The remaining steps are quite standard in Bayesian factor analysis (Geweke and Singleton, 1980; Lopes and West, 2004). In Step (P), we use an efficient algorithm for multi-move sampling of all unknown model parameters $\boldsymbol{\beta}_r$, and $\sigma^2_1, \ldots, \sigma^2_m$, see Appendix C.3. In Step (F), the conditional posterior $p(\mathbf{f}^r_1, \ldots, \mathbf{f}^r_T | \boldsymbol{\beta}_r, \sigma^2_1, \ldots, \sigma^2_m, \mathbf{y})$ factors into independent normal distributions given by:

$$\mathbf{f}^r_t | \mathbf{y}_t, \boldsymbol{\beta}_r, \boldsymbol{\Sigma}_r \sim \mathcal{N}_r \left((\mathbf{I}_r + \boldsymbol{\beta}'_r \boldsymbol{\Sigma}_r^{-1} \boldsymbol{\beta}_r)^{-1} \boldsymbol{\beta}'_r \boldsymbol{\Sigma}_r^{-1} \mathbf{y}_t, (\mathbf{I}_r + \boldsymbol{\beta}'_r \boldsymbol{\Sigma}_r^{-1} \boldsymbol{\beta}_r)^{-1} \right), \tag{4.2}$$

where $\boldsymbol{\Sigma}_r = \text{Diag}\,(\sigma^2_1, \ldots, \sigma^2_m)$. For the hierarchical Gaussian priors (3.11) and (3.12), all unknown shrinkage parameters and, for the rescaled F-prior (3.15) on $\sigma^2_1, \ldots, \sigma^2_m$, also the scaling parameters $C_{01}, \ldots, C_{0m}$ are updated in Step (S) (see Appendix E), since Step (P) is performed conditional on these values. Finally, the boosting Step (A) is added to improve the mixing of the MCMC scheme, see Section 4.3.

In Block (EFA), the sampler moves from the current CFA model to an EFA model with r_{sp} spurious columns and performs dimension changing moves in the much larger space underlying this model. The sampler finally returns to a CFA model with a potentially larger number of active factors r, see Section 4.1 for more details.

4.1 Split and merge moves for overfitting models

Step (EFA) in Algorithm 1 is based on moving from the CFA model (2.3) to an EFA model (2.1) with r_{sp} spurious factors in $\boldsymbol{\beta}_H$. Exploiting the results of Section 2.2, spurious columns in $\boldsymbol{\delta}_H$ are added and deleted in Step (R-S) by reversible jump MCMC (RJMCMC). Very conveniently, this step is independent of the pivots $\mathbf{l}_\Xi$ and the loadings $\boldsymbol{\Xi}$ in the spurious columns, since the prior $p(\boldsymbol{\delta}_H, \boldsymbol{\beta}_H, \boldsymbol{\Sigma}_H | r_{sp})$ is invariant to the specific choice of $\mathbf{l}_\Xi$ and $\boldsymbol{\Xi}$, given r_{sp}. However, the prior odds that a zero column in $\boldsymbol{\delta}_H$ can be turned into an additional spurious column are equal to:

$$O^{\text{sp}}(r, r_{sp}) = \frac{a_H(m - r - r_{sp})}{b_H - 1 + m - r - r_{sp}}. \tag{4.3}$$

For $b_H = 1$, the prior odds (4.3) depend only on a_H, independently of the current number of active and spurious columns. But even in this case, simply adding or deleting spurious columns would lead to an invalid MCMC procedure and an RJMCMC step that incorporates $O^{\text{sp}}(r, r_{sp})$ is performed in Step (R-S). As opposed to other applications of RJMCMC, the acceptance rate is extremely easy to compute, see (4.4) and (4.5).

At each sweep of the sampler, a split or a merge move is performed with, respectively, probability $p_{\text{split}}(r, r_{sp})$ or $p_{\text{merge}}(r, r_{sp})$. A symmetric proposal is applied for all $0 \leq r_{sp} < H - r$ with $p_{\text{split}}(r, r_{sp}) = p_{\text{merge}}(r, r_{sp} + 1) = p_s$, where $p_s \leq 0.5$ is a tuning parameter, while $p_{\text{merge}}(r, r_{sp}) = 0$ for $r_{sp} = 0$ and $p_{\text{split}}(r, r_{sp}) = 0$ for $r_{sp} = H - r$. A split move turns one of the $H-(r+r_{sp})$ zero columns in $\boldsymbol{\delta}_H$ into a spurious column, with proposal density $q_{\text{split}}(\boldsymbol{\delta}^{\text{new}}_H | \boldsymbol{\delta}_H) = p_s/(H-r-r_{sp})$. A merge move turns one of the $r_{sp} > 0$ spurious columns in $\boldsymbol{\delta}_H$ into a zero column, with proposal density $q_{\text{merge}}(\boldsymbol{\delta}^{\text{new}}_H | \boldsymbol{\delta}_H) = p_s/r_{sp}$. A split move is accepted with probability $\min(1, A_{\text{split}}(r, r_{sp}))$, where:

$$A_{\text{split}}(r, r_{sp}) = \frac{q_{\text{merge}}(\boldsymbol{\delta}_H | \boldsymbol{\delta}^{\text{new}}_H)}{q_{\text{split}}(\boldsymbol{\delta}^{\text{new}}_H | \boldsymbol{\delta}_H)} O^{\text{sp}}(r, r_{sp}) = \frac{a_H(m - r - r_{sp})(H - r - r_{sp})}{(r_{sp} + 1)(b_H + m - r - r_{sp} - 1)}, \tag{4.4}$$

whereas a merge move is accepted with probability $\min(1, A_{\text{merge}}(r, r_{sp}))$, where

$$A_{\text{merge}}(r, r_{sp}) = \frac{1}{A_{\text{split}}(r, r_{sp} - 1)} = \frac{r_{sp}(b_H + m - r - r_{sp})}{a_H(m - r - r_{sp} + 1)(H - r - r_{sp} + 1)}. \tag{4.5}$$

There is a dynamic feature underlying this RJMCMC algorithm, with acceptance depending on the number of spurious columns r_{sp}. For $b_H = 1$, for instance, $A_{\text{split}}(r, r_{sp})$ is monotonically decreasing and $A_{\text{merge}}(r, r_{sp})$ is monotonically increasing in r_{sp}.

Once r_{sp} has been updated, Step (R-L) is trying to turn each spurious column into an active one. Since the likelihood is non-informative about spurious columns, pivots l_{sp} are sampled uniformly from the prior $\mathbf{l}_\Xi|\mathbf{l}_r$, while the spurious factor loadings $\Xi_{l_{\text{sp}}}$ are sampled from the prior (3.9). Given l_{sp}, the idiosyncratic variance $\sigma^2_{l_{\text{sp}}}$ in the CFA model is split, with the help of a random variable $U_{j_{\text{sp}}} \sim \mathcal{U}[-1, 1]$, between $\Xi_{l_{\text{sp}}}$ and an updated idiosyncratic variance $\sigma^{2,\text{new}}_{l_{\text{sp}}}$. More specifically:

$$\Xi_{l_{\text{sp}}} = U_{j_{\text{sp}}} \sqrt{\sigma^2_{l_{\text{sp}}}}, \qquad \sigma^{2,\text{new}}_{l_{\text{sp}}} = (1 - U^2_{j_{\text{sp}}})\sigma^2_{l_{\text{sp}}}. \tag{4.6}$$

Given $\Xi_{l_{\text{sp}}}$ and $\sigma^{2,\text{new}}_{l_{\text{sp}}}$, factors $f_{j_{\text{sp}},t}$ are proposed in Step (R-F) for each $t = 1, \ldots, T$ from the conditional density $p(f_{j_{\text{sp}},t}|\mathbf{f}^r_t, \boldsymbol{\beta}_r, \sigma^2_{l_{\text{sp}}}, y_{l_{\text{sp}},t})$ given in (4.1). The slab probabilities $\tau_{j_{\text{sp}}}$ are sampled in Step (R-H) as in Algorithm 1, Step (H), using that $d_{j_{\text{sp}}} = 1$. Finally, in Step (R-D) variable selection is performed in each spurious column on all indicators below l_{sp} as in Step (D) of Algorithm 1, conditional on $f_{j_{\text{sp}},t}$. Any spurious column that is turned into an active one is integrated into the CFA model, increasing in this way the number of active columns r. Further details and proofs are provided in Appendix F.

4.2 Special MCMC moves for unordered GLT structures

Step (L) in Algorithm 1 implements MH-moves to change the current position of the pivot rows $\mathbf{l}_r = (l_1, \ldots, l_r)$ in the r columns of the UGLT indicator matrix $\boldsymbol{\delta}_r$. To change $l_j|\mathbf{l}_{r,-j}$ given the remaining pivot rows $\mathbf{l}_{r,-j}$, we use several moves, namely shifting the pivot, adding a new pivot, deleting a pivot and switching the pivots (and additional indicators) between column j and a randomly selected column j'; see Figure G.1 for illustration. All moves are performed marginalized w.r.t. $\boldsymbol{\tau}_r$. Changing the pivot from l_j to l_j^{new} changes the number of unconstrained indicators, whereas the prior ratio $p(l_j^{\text{new}}|\mathbf{l}_{r,-j})/p(l_j|\mathbf{l}_{r,-j}) = 1$. With d_j^{new} being the new number of non-zero elements in column j, the prior ratio R_{move} can be derived from (3.8):

$$R_{\text{move}} = \frac{\Pr(\boldsymbol{\delta}^{\text{new}}_{\cdot,j}|l_j^{\text{new}})}{\Pr(\boldsymbol{\delta}_{\cdot,j}|l_j)} = \frac{B(a_H + d_j^{\text{new}} - 1, b_H + m - l_j^{\text{new}} - d_j^{\text{new}} + 1)}{B(a_H + d_j - 1, b_H + m - l_j - d_j + 1)}. \tag{4.7}$$

Further details are provided in Appendix G.

4.3 Boosting MCMC

Step (F) and Step (P) in Algorithm 1 sample the factors $(\mathbf{f}^r_1, \ldots, \mathbf{f}^r_T)$ conditional on $(\boldsymbol{\beta}_r, \boldsymbol{\Sigma}_r)$ and $(\boldsymbol{\beta}_r, \boldsymbol{\Sigma}_r)$ conditional on $(\mathbf{f}^r_1, \ldots, \mathbf{f}^r_T)$. Depending on the signal-to-noise ratio, such full conditional Gibbs sampling tends to be poorly mixing. In a factor model where $\mathbf{f}_t^r \sim \mathcal{N}_r\left(\mathbf{0}, \mathbf{I}_r\right)$, the information in the data (the "signal") can be quantified by the matrix $\boldsymbol{\beta}_r' \boldsymbol{\Sigma}_r^{-1} \boldsymbol{\beta}_r$ in comparison to the identity matrix $\mathbf{I}_r$ (the "noise") in the filter for $\mathbf{f}_t^r | \mathbf{y}_t, \boldsymbol{\beta}_r, \boldsymbol{\Sigma}_r$, see (4.2). One would expect that factor models with many measurements contain ample information to estimate the factors, however, this is true only if the information matrix $\boldsymbol{\beta}_r' \boldsymbol{\Sigma}_r^{-1} \boldsymbol{\beta}_r$ increases with m and most of the factor loadings are nonzero. Sparse factor models contain many columns with only a few non-zero loadings, leading to a low signal-to-noise ratio and, consequently, to a poorly mixing sampler. For such models, boosting steps are essential to obtain efficient MCMC schemes. Several papers (Ghosh and Dunson, 2009; Frühwirth-Schnatter and Lopes, 2010; Conti et al., 2014) apply marginal data augmentation (MDA) in the spirit of van Dyk and Meng (2001); others (Kastner et al., 2017; Frühwirth-Schnatter and Lopes, 2018) exploit the ancillarity-suffiency interweaving strategy (ASIS) introduced by Yu and Meng (2011).

Boosting is based on moving from the CFA model (2.3) where $\mathbf{f}_t^r \sim \mathcal{N}_r\left(\mathbf{0}, \mathbf{I}_r\right)$ to an expanded model with a more general prior:

$$\mathbf{y}_t = \tilde{\boldsymbol{\beta}}_r \tilde{\mathbf{f}}_t^r + \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \sim \mathcal{N}_m\left(\mathbf{0}, \boldsymbol{\Sigma}_r\right), \qquad \tilde{\mathbf{f}}_t^r \sim \mathcal{N}_r\left(\mathbf{0}, \boldsymbol{\Psi}\right), \tag{4.8}$$

where $\boldsymbol{\Psi} = \text{Diag}\left(\Psi_1, \ldots, \Psi_r\right)$ is diagonal. The two systems are related by the transformations $\tilde{\mathbf{f}}_t^r = (\boldsymbol{\Psi})^{1/2} \mathbf{f}_t^r$ and $\tilde{\boldsymbol{\beta}}_r = \boldsymbol{\beta}_r (\boldsymbol{\Psi})^{-1/2}$, where the nonzero elements in $\tilde{\boldsymbol{\beta}}_r$ have the same position as the nonzero elements in $\boldsymbol{\beta}_r$ and the sparsity matrix $\boldsymbol{\delta}_r$ is not affected by the transformation. The main difference between MDA and ASIS lies in the choice of $\boldsymbol{\Psi}$. While Ψ_j is sampled from a working prior for MDA, Ψ_j is chosen in a deterministic fashion for ASIS. For illustration, Figure 2 shows posterior draws of $\text{tr}(\boldsymbol{\beta}_r' \boldsymbol{\Sigma}_r^{-1} \boldsymbol{\beta}_r)$ for the exchange rate data to be discussed in Section 5.2 without boosting (left-hand panel) and illustrates the considerable efficiency gain when a boosting strategy such as ASIS (middle panel) or MDA (right-hand panel) is applied in Step (A).

For the hierarchical priors (3.11) and (3.12) we found it particularly useful to apply *column boosting* and interweave the column specific shrinkage parameter θ_j into the state equation by choosing $\Psi_j = \theta_j$. For the F-prior (3.15) on $\sigma_1^2, \ldots, \sigma_m^2$, another useful strategy is *row boosting*, based on moving the random scales C_{0i} from the prior σ_i^2 to the observation equation in all rows of the basic factor model. Full details for all boosting steps are provided in Appendix H.

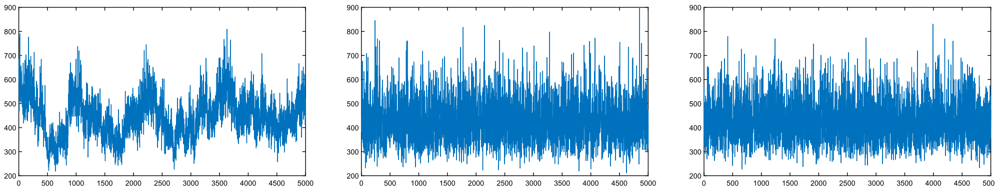

Figure 2: Exchange rate data (standardized, fractional prior and prior (3.14) for σ_i^2); posterior draws of $\text{tr}(\boldsymbol{\beta}_r' \boldsymbol{\Sigma}_r^{-1} \boldsymbol{\beta}_r)$ without boosting (left-hand side), boosting through ASIS with $\sqrt{\Psi_j}$ equal to the largest loading (in absolute values) (middle) and through MDA based on the working prior $\Psi_j \sim \mathcal{G}^{-1}\left(1.5, 1.5\right)$ (right-hand side).

4.4 Post-processing posterior draws

Algorithm 1 delivers posterior draws $(\boldsymbol{\delta}_r, \boldsymbol{\beta}_r, \boldsymbol{\Sigma}_r)$ in a CFA model with a varying number r of active columns. Our sampler imposes the (mild) condition that the pivots (the first non-zero loading in each column) lie in different rows, ensuring that all posterior draws of the loading matrix exhibit a UGLT structure. As discussed in Section 2.1, this allows identification during post-processing.

We use the 3579 counting rule and the algorithm of Hosszejni and Frühwirth-Schnatter (2022) to check for each draw $\boldsymbol{\delta}_r$ whether the variance decomposition is unique, and remove all draws that are not variance identified. Quantities that can be inferred from variance identified posteriors draws with varying factor dimension r include the covariance matrix $\boldsymbol{\Omega} = \boldsymbol{\beta}_r\boldsymbol{\beta}_r^T + \boldsymbol{\Sigma}_r$, the idiosyncratic variances $\sigma_1^2, \ldots, \sigma_m^2$, the modelsize $d = \sum_{i,j} \delta_{ij}$, and the communalities $R_1^2, \ldots, R_m^2$ defined in (3.13). Most importantly, variance identified posterior draws are instrumental for identifying the number of factors and the factor loading matrix. The number of nonzero columns of all variance identified draws $\boldsymbol{\delta}_r$ can be regarded as posterior draws of the unknown factor dimension r. The posterior distribution $p(r|\mathbf{y})$ derived from these draws yields uncertainty quantification and the posterior mode $\tilde{r}$ serves as an estimator of r.

Due to the UGLT structure imposed on $\boldsymbol{\beta}_r$, rotational invariance reduces to sign and column switching. $\boldsymbol{\beta}_r$ is rotated into a loading matrix $\boldsymbol{\Lambda}$ with GLT structure by ordering the columns such that the pivots are increasing, and the sign is reversed in all columns with a negative leading factor loading. The GLT draws $\boldsymbol{\Lambda}$ still exhibit a varying factor dimension r and posterior variation in $l_1, \ldots, l_r$. To estimate the factor loading matrix, further inference is performed conditionally on the posterior mode $\tilde{r}$ and an estimator $\hat{\mathbf{l}}_{\tilde{r}} = (\hat{l}_1, \ldots, \hat{l}_{\tilde{r}})$ of the pivots given by the sequence visited most often across all draws with factor dimension $r = \tilde{r}$. Bayesian model averaging over all GLT draws $\boldsymbol{\Lambda}$ with pivot $\hat{\mathbf{l}}_{\tilde{r}}$ yields the posterior mean $\mathrm{E}(\boldsymbol{\Lambda}|\mathbf{y}, \hat{\mathbf{l}}_{\tilde{r}})$. The marginal posterior $p(\Lambda_{ij}|\mathbf{y}, \hat{\mathbf{l}}_{\tilde{r}})$ and the marginal inclusion probability $\Pr(\delta_{ij}^{\Lambda} = 1|\mathbf{y}, \hat{\mathbf{l}}_{\tilde{r}})$ allow uncertainty quantification for individual elements Λ_{ij} and δ_{ij}^{Λ} in $\boldsymbol{\Lambda}$ and corresponding the sparsity matrix $\boldsymbol{\delta}^{\Lambda}$. Alternative estimators such as the sequence of pivots $\mathbf{l}^{\star}$ visited most often among all variance identified draws and more details are provided in Appendix I.

5 Applications

We discuss applications both to simulated as well as real data sets. For each data set, whether simulated or real, Algorithm 1 is used to generate and post-process M posterior draws after a burn-in of M_0 draws.[2] We choose a 2PB prior to ensure column sparsity, combine various slab distributions for β_{ij} with various priors on σ_i^2, and use the default hyperparameters introduced in Table 1, namely $H = \lfloor (m-1)/2 \rfloor$, $c^{\sigma} = 2.5$, $\mathrm{E}_R = 2/3$, $\mathrm{E}_q = 2$ and $a^{\alpha} = a^{\gamma} = c^{\kappa} = a^{\theta} = a^{\sigma} = n_0 = 6$.

[2] Tuning in Step (R) and Step (L) relies on $p_s = 0.5$, $p_{\text{shift}} = p_{\text{switch}} = 1/3$ and $p_a = 0.5$. Boosting in Step (A) relies on ASIS with $\sqrt{\Psi_j}$ being the largest loading (in absolute value) in column j.

5.1 Simulation study

We perform an extensive simulation study and summarize the main findings in this section. Full details on the simulation settings, the performance measures and additional results are provided in Appendix K. We assume $m = 30$, $T = 100$, and $r_{\mathrm{true}} = 5$ factors and consider six sparsity patterns $\mathbf{\Lambda}$, namely a dedicated factor model, a dedicated factor model with overlap, a two-block factor model, a sparse factor model with 50% overall sparsity, a model with a market factor that loads on all measurements and exhibits 60% sparsity in the remaining columns and a dense factor model with no zero loadings. 50 data sets are generated for each scenario from the basic factor model (1.1) under $\mathbf{\Sigma}_0 = \mathbf{I}$. Note that $H = 14$. Prior (3.14) for σ_i^2 is combined with the following slab distributions $\mathrm{P}_{\mathrm{slab}}$ for β_{ij}: a fractional prior (F), prior (3.10) with global shrinkage (G), prior (3.11) with column shrinkage (C), and prior (3.12), where local shrinkage with $a^\omega = c^\omega = 0.5$ relies on the horseshoe (H) and on a triple gamma with $a^\omega = c^\omega = 0.2$ (T). MCMC is performed with $M_0 = M = 4{,}000$ for all 300 data sets under each of these five priors, starting either with $r = 3$ or $r = 8$ active and $r_{sp} = 2$ spurious factors.

For each simulated data set, the variance identified posterior draws under a specific prior yield estimates of the posterior mode $\tilde{r}$, the posterior ordinate $P_{\mathrm{true}} = \Pr(\tilde{r} = r_{\mathrm{true}}|\mathbf{y})$, the posterior risk measures $R_\Omega = \mathrm{E}(L(\mathbf{\Omega}_r, \mathbf{\Omega}_0)|\mathbf{y})$, $R_\Sigma = \mathrm{E}(L(\mathbf{\Sigma}_r, \mathbf{\Sigma}_0)|\mathbf{y})$ and $R_{\Omega^{-1}} = \mathrm{E}(L(\mathbf{\Omega}_r^{-1}, \mathbf{\Omega}_0^{-1})|\mathbf{y})$ in recovering the true matrices $\mathbf{\Omega}_0$, $\mathbf{\Sigma}_0$, and $\mathbf{\Omega}_0^{-1}$, where L is the entropy (or Stein) loss (Yang and Berger, 1994), the true positive rate TP_Ω for non-zero and the false positive rate FP_Ω for zero correlations in $\mathbf{\Omega}_0$, the bias $\mathrm{B}_d = \mathrm{E}(d_r|\mathbf{y}) - d_{\mathrm{true}}$ in model size, and the true positive rate TP_δ and the false positive rate FP_δ for the true sparsity pattern $\boldsymbol{\delta}^\Lambda$. Tables 2 and K.1 report the average and Figures K.2 to K.7 the entire sampling distribution for these performance measures for all sparsity patterns and slab distributions $\mathrm{P}_{\mathrm{slab}}$.

In general, sparse UGLT Bayesian factor analysis has a high hit rate and correctly recovers the true number of factors through the posterior mode for most of the 1500 runs of our sampler, with 76 and, respectively, 14 under- and overfittings occurring mainly for the block and the market sparsity patterns. The choice of $\mathrm{P}_{\mathrm{slab}}$ has considerable impact on recovering the true sparsity pattern $\boldsymbol{\delta}^\Lambda$ in $\mathbf{\Lambda}$ and $\mathbf{\Omega}_0$. The fractional prior has the smallest false positive rates FP_δ and FP_Ω both for $\boldsymbol{\delta}^\Lambda$ and $\mathbf{\Omega}_0$ which is considerably smaller than for hierarchical shrinkage priors for all sparsity patterns. At the risk of higher false positive rates, the true positive rates TP_δ and TP_Ω both for $\boldsymbol{\delta}^\Lambda$ and $\mathbf{\Omega}_0$ are larger for hierarchical shrinkage priors than for the fractional prior, for which they are still high with a few exceptions. Overall, the fractional prior leads to the sparsest solutions with the smallest model size d, resulting in strong underfitting of d for the dense pattern but also in the smallest bias in d for other sparsity patterns. Regarding the estimates for $\mathbf{\Omega}_0$, $\mathbf{\Omega}_0^{-1}$ and $\mathbf{\Sigma}_0$, hierarchical shrinkage priors have a smaller average loss L than the fractional prior, even if the differences are significant only for the dense sparsity pattern.

For comparison, we perform for each of the sparsity patterns under each slab distribution sparse BFA under the PLT condition, assuming that the number of factors $r = 5$ is known and equal to the true value. Priors are the same as under UGLT with $H = 5$. MCMC is implemented by a simplification of Algorithm 1, see Appendix J.

		$\tilde{r}$	P_{true}	R_Ω	$R_{\Omega^{-1}}$	R_Σ	TP$_\Omega$	FP$_\Omega$	B$_d$	TP$_\delta$	FP$_\delta$
Dedicated											
UGLT	F	5	0.998	1.41	1.47	0.98	94.3	9.56	0.5	96.1	5.3
	G	5.04(0,2)	0.933	1.49	1.62	0.94	97.3	37.9	7.48	96.2	21.6
	T	5	0.962	1.48	1.58	0.94	97.1	59.4	17.3	95.4	38.3
PLT	F	–	–	2.66	2.98	1.67	85.3	34.5	3.25	28.2	75.2
	G	–	–	2.35	2.17	1.18	91.2	58.1	13.8	37.2	74.7
	T	–	–	2.29	2.24	1.26	93.1	72.9	28	44.4	77.1
Overlap											
UGLT	F	5.02(0,1)	0.985	1.61	1.64	0.95	96.3	8.31	1.02	97.7	4.6
	G	5	0.939	1.68	1.75	0.88	97.4	33.7	9.33	96.5	21.5
	T	5	0.972	1.72	1.83	0.92	98.8	53.7	22.2	98.2	37.2
PLT	F	–	–	2.56	2.49	1.29	88.8	28.6	3.83	37.5	66.1
	G	–	–	2.55	2.43	1.12	94.4	50.7	16.3	42.6	70.2
	T	–	–	2.83	2.62	1.4	95.9	64.9	33.4	53.6	71.5
Block											
UGLT	F	4.60(18,0)	0.636	3.00	2.78	1.34	88.7	7.25	−25.0	64.0	5.63
	G	4.90(6,1)	0.848	2.53	2.56	1.00	95.5	24.8	−8.17	78.5	12.9
	T	4.78(11,0)	0.764	2.64	2.61	1.13	97.5	42.2	3.49	82.5	23.6
PLT	F	–	–	4.05	4.19	1.69	83.5	19.2	−23.0	42	39.6
	G	–	–	3.04	3.85	1.18	92.7	37.2	−6.31	59.4	35.2
	T	–	–	4.01	4.59	1.79	95.1	49.1	11.8	61.6	47.5
Sparse											
UGLT	F	4.84(4,0)	0.92	2.64	2.43	1	93	7.85	0.06	90.6	9.9
	G	5.02(0,1)	0.94	2.32	2.44	0.88	98.6	27.5	17	94.5	26.7
	T	4.8(2,0)	0.94	2.31	2.4	0.94	95.6	38.1	35.4	96.5	41.0
PLT	F	–	–	3.67	3.59	1.33	91	18.4	3.7	56.6	47.0
	G	–	–	2.66	2.87	0.95	98.3	32.6	20.8	74.0	45.4
	T	–	–	2.83	2.74	1.13	98.9	42.4	41.2	82.0	52.6
Market											
UGLT	F	4.86(4,0)	0.919	2.78	2.43	1.03	96	0	−1.04	93.0	5.89
	G	5.02(0,1)	0.951	2.28	2.39	0.87	99.6	0	10.6	95.1	17.7
	T	4.84(6,1)	0.86	2.54	2.59	0.98	99.7	0	24	96.5	30.2
PLT	F	–	–	4.5	4.48	1.38	89.3	0	0.79	61.2	40.3
	G	–	–	2.55	2.62	0.92	99	0	14.4	76.4	38.1
	T	–	–	3.13	2.93	1.08	98.7	0	30	78.2	46.6
Dense											
UGLT	F	4.98(1,0)	0.976	5.94	4.45	1.07	94.1	0	−60.8	56.7	0
	G	5	0.989	3.79	3.72	0.85	98.9	0	−26.0	81.4	0
	T	5	0.99	4.26	3.99	0.99	99.4	0	−14.8	89.4	0
PLT	F	–	–	6.35	4.98	1.16	94.5	0	−58.1	58.6	0
	G	–	–	3.84	3.71	0.87	98.9	0	−27.4	80.4	0
	T	–	–	4.50	4.07	1.06	99.3	0	−17.1	87.8	0

For each performance measure, the average across 50 simulated data sets is reported. If $\tilde{r} \neq 5$, (a,b) report, respectively, cases of under- and overfitting.

Table 2: Performance of sparse Bayesian factor analysis under a UGLT condition with r unknown in comparison to a PLT condition with $r = r_{\text{true}} = 5$ known for all sparsity pattern.

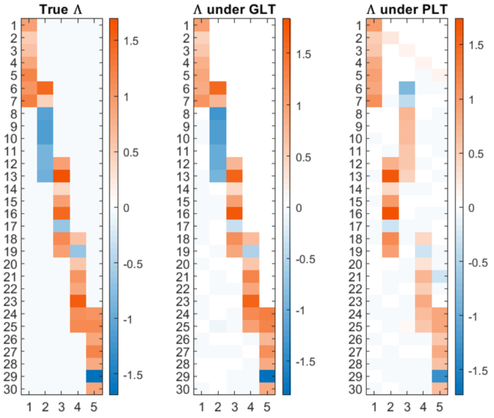

Figure 3: Comparing the true loading matrix $\mathbf{\Lambda}$ (left) with the estimated loading matrix $\hat{\mathbf{\Lambda}}$ under the UGLT condition with r unknown (middle) and under the PLT condition with $r = 5$ known (right-hand side) for a randomly selected data set under the dedicated with overlap scenario (fractional prior in the slab).

Eight performance measures are determined from these draws and compared to sparse BFA under the UGLT condition in Tables 2 and K.1 and Figures K.2 to K.7. Despite assuming the true number of factors, PLT shows worse performance with respect to recovering the true sparsity pattern in $\mathbf{\Lambda}$ and $\mathbf{\Omega}_0$, but also exhibits a higher loss L in estimating $\mathbf{\Omega}_0$, $\mathbf{\Omega}_0^{-1}$ and $\mathbf{\Sigma}_0$ with the exception of dense factor models which is the only sparsity pattern where the PLT pivots $(l_1, \dots, l_5) = (1, \dots, 5)$ coincide with the true pivots. For all other sparsity patterns, the PLT condition imposed on $\mathbf{\Lambda}$ does not really solve rotational invariance, but imposes an ordering on the columns of $\mathbf{\Lambda}$ that is in conflict with the GLT ordering, see Figure 3 for illustration.

5.2 Sparse Bayesian factor analysis for exchange rate data

As a first exercise on real data, we analyze log returns spanning $T = 96$ months from $m = 22$ exchange rates against the Euro.[3] The data are demeaned and standardized. Note that $H = 10$. We combine the fractional prior (C.4) with the following priors on σ_i^2: prior (3.14) (HIG) and (3.15) (HF) with default settings, prior (3.14) with b_i^σ chosen as in Frühwirth-Schnatter and Lopes (2018) (FSL) and $\sigma_i^2 \sim \mathcal{G}^{-1}(1, 0.3)$ (Bhattacharya and Dunson, 2011) (BD). Algorithm 1 is run for each prior for $M =$ 50,000 iterations, after a burn-in of 50,000. To verify convergence, independent MCMC chains are started with $r = 7$ active and $r_{sp} = 3$ spurious columns. The sampler shows good mixing across models of different dimension, with the inefficiency factor for model size d ranging from 7 (FSL) to 22 (HF). For illustration, Figure 4 shows all posterior draws of r and d including burn-in for the HIG prior.

[3] The data were obtained from the European Central Bank's Statistical Data Warehouse and range from January 3, 2000, to December 3, 2007. Table L.2 in Appendix L lists the 22 currencies. We derived

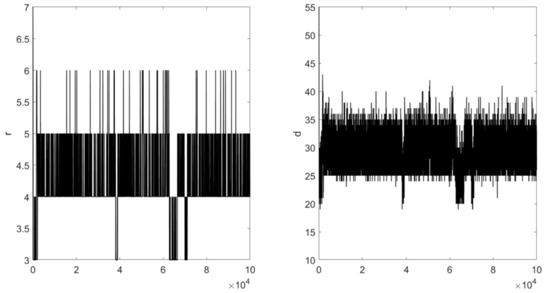

Figure 4: Exchange rate data (standardized); posterior draws of the factor dimension r (left) and model size d (right) including burn-in (fractional prior combined with the HIG prior).

	$p(r\|\mathbf{y})$									
Prior	0-2	3	4	5	6	7-10	$100p_V$	$\mathrm{E}(d\|\mathbf{y})$	$\mathrm{E}(\alpha\|\mathbf{y})$	$\mathrm{E}(\gamma\|\mathbf{y})$
HF	0	0.137	**0.834**	0.029	≈ 0	0	90.3	28	2.3	1.1
HIG	0	0.110	**0.874**	0.016	≈ 0	0	92.5	28	2.3	1.1
FSL	0	0.033	**0.954**	0.013	0	0	93.5	28	2.3	1.1
BD	0	0.033	**0.954**	0.013	0	0	93.6	28	2.2	1.1

Note: non-zero probabilities smaller than 10^{-3} are indicated by ≈ 0.

Table 3: Exchange rate data (standardized); posterior distribution $p(r|\mathbf{y})$ of the number of factors, fraction $100p_V$ of variance identified draws, posterior means of model size d and the hyperparameters α and γ under the fractional prior and various priors for σ_i^2.

	$\Pr(q_i = 0\|\mathbf{y})$						
Prior on σ_i^2	CHF	CZK	MXN	NZD	RON	RUB	remaining
HF	0.88	0.74	0.81	0.47	0.61	0.61	0
HIG	0.88	0.73	0.82	0.46	0.55	0.61	0
FSL	0.88	0.75	0.81	0.49	0.61	0.61	0
BD	0.87	0.72	0.82	0.48	0.62	0.64	0

Table 4: Exchange rate data (standardized); posterior probability of the event $\Pr(q_i = 0|\mathbf{y})$, where q_i is the row sum of $\boldsymbol{\delta}_r$, for various currencies.

Posterior inference as summarized in Table 3 is robust to the chosen prior. The fraction p_V of variance identified draws is in general very high and the posterior distribution $p(r|\mathbf{y})$ is highly concentrated at four factors. The indicator matrix $\boldsymbol{\delta}_r$ is sparse, with an average posterior model size of 28. The variance identified draws are used to explore if some measurements are uncorrelated with the remaining measurements. This is investigated in Table 4 through the posterior probability $\Pr(q_i = 0|\mathbf{y})$, where q_i is the ith row sum of $\boldsymbol{\delta}_r$. The Swiss franc (CHF), the Mexican peso (MXN) and the Czech koruna (CZK) have considerable probability to be uncorrelated with the rest, while the

the returns based on the first trading day in a month.

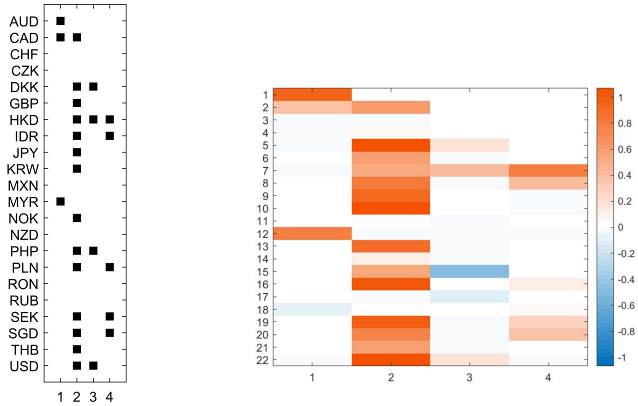

Figure 5: Exchange rate data (standardized); left hand side: sparsity matrix $\boldsymbol{\delta}_4$ corresponding to the median probability model (identical for all four priors); right hand side: estimated loading matrix $\mathrm{E}(\boldsymbol{\Lambda}|\hat{\mathbf{l}}_4, \mathbf{y})$ with $\hat{\mathbf{l}}_4 = (1, 2, 5, 7)$ for the HIG prior (nearly identical for all four priors).

situation is less clear for the New Zealand dollar (NZD), the Romania fourth leu (RON), and the Russian ruble (RUB). The remaining currencies are clearly correlated.

All posterior draws $\boldsymbol{\beta}_r$ are rotated into a GLT structure $\boldsymbol{\Lambda}$ by ordering the pivots such that $l_1 < \ldots < l_r$. The sequence of pivots visited most often among all draws with $r = \tilde{r} = 4$ is equal to $\hat{\mathbf{l}}_4 = (1, 2, 5, 7)$ for all priors and coincides with the sequence of pivots $\mathbf{l}^\star$ visited most often among all variance identified draws. Sign switching is resolved by imposing the constraint $\Lambda_{11} > 0$, $\Lambda_{22} > 0$, $\Lambda_{53} > 0$, and $\Lambda_{74} > 0$ on $\boldsymbol{\Lambda}$. All GLT draws where the pivots $\mathbf{l}_4$ coincide with $\hat{\mathbf{l}}_4 = (1, 2, 5, 7)$ are used to identify the GLT representation of the factor loading matrix $\boldsymbol{\Lambda}$ and the marginal inclusion probabilities $\Pr(\delta_{ij} = 1|\mathbf{y}, \hat{\mathbf{l}}_4)$. The analysis reveals a factor model with considerable sparsity, with many factor loadings being shrunk toward zero, see Figure 5 for illustration. Factor 2 is a common factor among the correlated currencies, while the remaining factors are three group specific, for the most part dedicated factors. Further results are reported in Appendix L.

5.3 Sparse factor analysis for NYSE stock returns

As a second application, we consider monthly log returns from $m = 63$ firms from the NYSE observed for $T = 247$ months from February 1999 till August 2019.[4] Note that

[4] $T = 247$ monthly returns (determined on the last trading day in each month) starting from February, 1999, of the largest 150 companies listed on the NYSE were downloaded from Bloomberg on September 13, 2019. After removing all companies with missing data, 103 firms remained. For our study, we consider the 63 firms belonging to the following five sectors: basic industries (1-7), non-durable consumer goods (8-17), energy (18-27), finance (28-45) and health care (46-63).

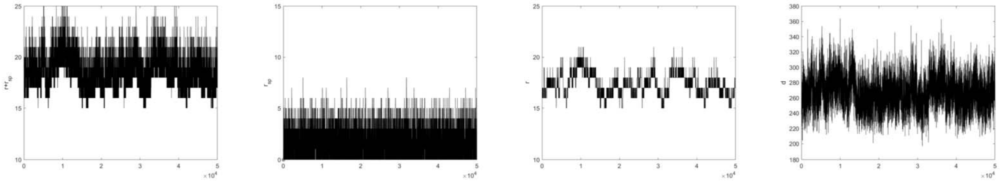

Figure 6: NYSE data; from left to right: posterior draws of the total number of non-zero columns $r + r_{sp}$, the number of spurious columns r_{sp}, the extracted number of factors r and the model dimension d.

$p(r\mid\mathbf{y})$										
0-14	15	16	17	18	19	20	21-31	E$(d\mid\mathbf{y})$	E$(\alpha\mid\mathbf{y})$	E$(\gamma\mid\mathbf{y})$
0	0.066	0.363	0.317	0.212	0.038	≈ 0	0	269	4.4	1.1

Note: non-zero probabilities smaller than 10^{-2} are indicated by ≈ 0.

Table 5: NYSE data; posterior distribution $p(r|\mathbf{y})$ of the number of factors; posterior means of model size d and the hyperparameters α and γ under prior (3.12) with $a^\omega = c^\omega = 0.2$ and the hierarchical F-prior (3.15) on σ_i^2.

$H = 31$. Since the data are not standardized, we fit an extended EFA model

$$\mathbf{y}_t = \boldsymbol{\mu} + \boldsymbol{\beta}_H \mathbf{f}_t^H + \boldsymbol{\epsilon}_t, \qquad \boldsymbol{\epsilon}_t \sim \mathcal{N}_m\left(\mathbf{0}, \boldsymbol{\Sigma}_H\right), \quad \mathbf{f}_t^H \sim \mathcal{N}_H\left(\mathbf{0}, \mathbf{I}_H\right),$$

with unknown mean $\boldsymbol{\mu}$, see Appendix M for details. As in the previous sections, we tried to apply a fractional prior as slab distribution, however, the fraction of variance identified posterior draws was extremely low (less than 1%). Instead, a hierarchically structured Gaussian shrinkage prior (3.12) is chosen, with local scaling parameters following a triple gamma prior with $a^\omega = c^\omega = 0.2$, and combined with the hierarchical F-prior (3.15) on σ_i^2. The fraction p_V of variance identified MCMC draws under this prior is roughly 32%. Algorithm 1 was applied to obtain $M = 50{,}000$ posterior draws after a burn-in of 50,000 draws, starting with $r = 20$ factors and $r_{sp} = 3$ spurious columns. The MCMC scheme shows relatively good mixing, despite the high dimensionality, as illustrated by Figure 6 showing posterior draws of the total number of non-zero columns, $r + r_{sp}$, the number of spurious columns r_{sp}, the extracted number of factors r, and the model dimension d.

As shown in Table 5, the posterior distribution $p(r|\mathbf{y})$ derived from the variance identified draws yields a posterior mode of $\tilde{r} = 16$, but also 17 or 18 factors receive considerable posterior evidence. For further inference, all posterior draws $\boldsymbol{\beta}_r$ are rotated into a GLT structure $\boldsymbol{\Lambda}$ by ordering the pivots such that $l_1 < \ldots < l_r$. The sequence of pivots visited most often among all draws of varying dimension r is equal to $\mathbf{l}^\star = (1, 2, 3, 4, 5, 8, 9, 11, 13, 15, 18, 19, 20, 32, 46, 48)$ which implies that the estimator $r^\star = 16$ is identical with the posterior mode $\tilde{r} = 16$. Furthermore, the sequence of pivots $\hat{\mathbf{l}}_{16}$ visited most often among all draws of dimension $r = 16$ coincides with $\mathbf{l}^\star$.

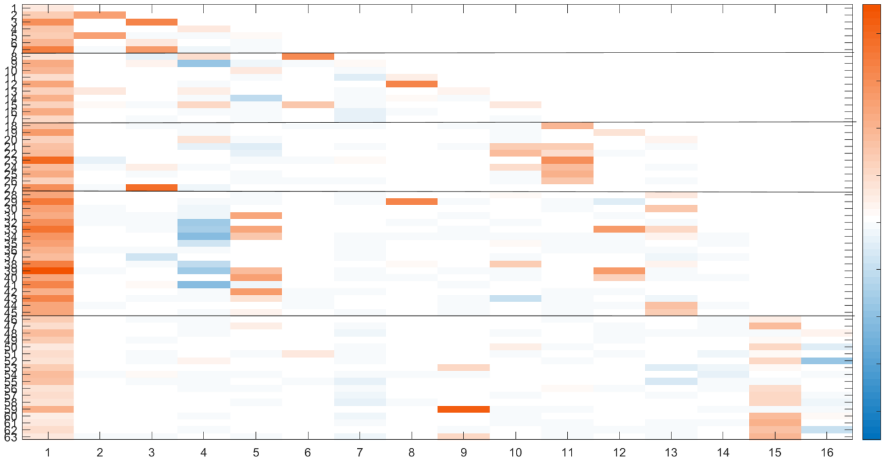

Figure 7: NYSE data; estimated GLT representation of the factor loading matrix $\boldsymbol{\Lambda}$.

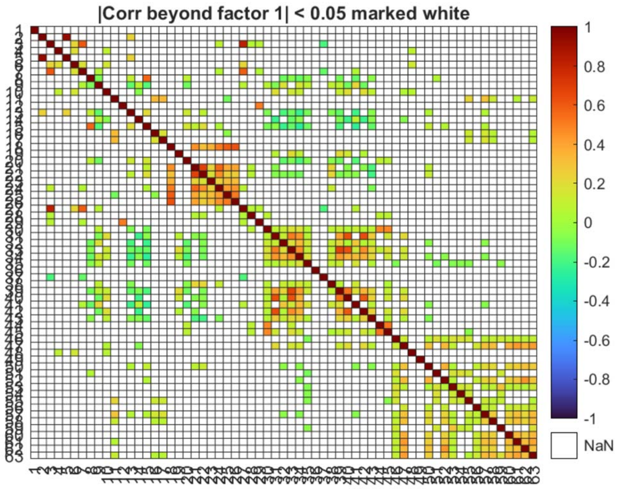

Figure 8: NYSE data; estimated marginal correlation matrix $\mathrm{E}(\boldsymbol{\Omega}^{\star}|\mathbf{y})$, where $\Omega^{\star}_{i\ell} = \mathrm{Corr}((y_{it} - \Lambda_{i1}f_{1t})(y_{\ell t} - \Lambda_{\ell 1}f_{1t}))$.

All GLT draws where the pivots $\mathbf{l}_r$ coincide with $\mathbf{l}^{\star} = \hat{\mathbf{l}}_{16}$ are used to identify the GLT representation of the factor loading matrix $\boldsymbol{\Lambda}$, see Figure 7. The analysis reveals a factor model with extreme sparsity. The first factor is a market factor that loads on all 63 firms. Several sector-specific factors emerge and capture industry specific correlations. Other factors capture cross-sectional correlations between specific firms. The remaining factors are weak factors with very sparse loadings; see also the estimated marginal correlation matrix $\boldsymbol{\Omega}^{\star}$ that remains after extracting the first factor in Figure 8.

6 Concluding remarks

We have estimated a fairly important and highly implemented class of sparse factor models when the number of common factors is unknown. Our framework leads to a natural, efficient and simultaneous coupling of model estimation and selection on one hand and model identification and rank estimation (number of factors) on the other hand. More precisely, by combining point-mass mixture priors with overfitting sparse factor modelling in an unordered generalised lower triangular loadings representation (Frühwirth-Schnatter et al., 2023), we obtain posterior summaries regarding factor loadings, common factors as well as the factor dimension via post-processing draws from our highly efficient and customised MCMC scheme. The new framework is readily available for some straightforward extensions. The reversible jump MCMC algorithm, for instance, can be applied to other factor models with minor modifications, in particular, to structures where all elements δ_{ij} in the sparsity matrix $\boldsymbol{\delta}_H$ are left unconstrained, see the studies in Frühwirth-Schnatter et al. (2023). The assumptions underlying the basic factor model can be substituted by idiosyncratic errors from Student-t distributions, by factors following Laplace (Grushanina and Frühwirth-Schnatter, 2021) or more general Gaussian mixtures priors (Piatek and Papaspiliopoulos, 2018) or by considering dynamic sparse factor models with stationary common factors (Kaufmann and Schuhmacher, 2019). A further interesting extension which is not built into the current analysis is to design a prior on the sparsity matrix that a priori distinguishes between pervasive factors that load on most measurements, group specific factors that load on selected measurements and factors that capture weak cross-sectional heterogeneity. Such approximate factor models are very popular in frequentist factor analysis (Chamberlain and Rothschild, 1983; Bai and Ng, 2002) and would deserve more attention from the Bayesian community. However, we leave this interesting idea for future research.

Supplementary Material

Supplementary material for: "Sparse Bayesian factor analysis when the number of factors is unknown" (DOI: 10.1214/24-BA1423SUPP; .pdf).

References

Akaike, H. (1987). "Factor analysis and AIC." *Psychometrika*, 52: 317–332. MR0914459. doi: https://doi.org/10.1007/BF02294359. 11

Anderson, T. W. (2003). *An introduction to multivariate statistical analysis*. Chichester: Wiley, 3rd edition. MR1990662. 1

Anderson, T. W. and Rubin, H. (1956). "Statistical inference in factor analysis." In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume V, 111–150. MR0084943. 3, 5

Aßmann, C., Boysen-Hogrefe, J., and Pape, M. (2016). "Bayesian analysis of static and dynamic factor models: An ex-post approach toward the rotation problem." *Jour-*

nal of Econometrics*, 192: 190–206. MR3463672. doi: https://doi.org/10.1016/j.jeconom.2015.10.010. 2, 3

Bai, J. and Ng, S. (2002). "Determining the number of factors in approximate factor models." *Econometrica*, 70: 191–221. MR1926259. doi: https://doi.org/10.1111/1468-0262.00273. 2, 26

Bai, J. and Ng, S. (2013). "Principal components estimation and identification of static factors." *Journal of Econometrics*, 176: 18–29. MR3067022. doi: https://doi.org/10.1016/j.jeconom.2013.03.007. 3

Bartholomew, D. J. (1987). *Latent variable models and factor analysis*. London: Charles Griffin. 11

Bhattacharya, A. and Dunson, D. (2011). "Sparse Bayesian infinite factor models." *Biometrika*, 98: 291–306. MR2806429. doi: https://doi.org/10.1093/biomet/asr013. 2, 12, 21

Cadonna, A., Frühwirth-Schnatter, S., and Knaus, P. (2020). "Triple the gamma – A unifying shrinkage prior for variance and variable selection in sparse state space and TVP models." *Econometrics*, 8: 20. 4, 10, 12

Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J., Wang, Q., and West, M. (2008). "High-dimensional sparse factor modeling: Applications in gene expression genomics." *Journal of the American Statistical Association*, 103: 1438–1456. MR2655722. doi: https://doi.org/10.1198/016214508000000869. 3, 13

Chamberlain, G. and Rothschild, M. (1983). "Arbitrage, factor structure, and mean-variance analysis on large asset markets." *Econometrica*, 51: 1281–1304. MR0736050. doi: https://doi.org/10.2307/1912275. 26

Chan, J., Leon-Gonzalez, R., and Strachan, R. W. (2018). "Invariant inference and efficient computation in the static factor model." *Journal of the American Statistical Association*, 113: 819–828. MR3832229. doi: https://doi.org/10.1080/01621459.2017.1287080. 2

Conti, G., Frühwirth-Schnatter, S., Heckman, J. J., and Piatek, R. (2014). "Bayesian exploratory factor analysis." *Journal of Econometrics*, 183: 31–57. MR3269916. doi: https://doi.org/10.1016/j.jeconom.2014.06.008. 2, 3, 17

De Vito, R., Bellio, R., Trippa, L., and Parmigiani, G. (2021). "Bayesian Multi-study factor analysis for high-throughput biological data." *The Annals of Applied Statistics*, 15: 1723 – 1741. MR4355073. doi: https://doi.org/10.1214/21-aoas1456. 2

Durante, D. (2017). "A note on the multiplicative gamma process." *Statistics and Probability Letters*, 122: 198–204. MR3584158. doi: https://doi.org/10.1016/j.spl.2016.11.014. 2

Fan, J., Fan, Y., and Lv, J. (2008). "High dimensional covariance matrix estimation using a factor model." *Journal of Econometrics*, 147: 186–197. MR2472991. doi: https://doi.org/10.1016/j.jeconom.2008.09.017. 2

Forni, M., Giannone, D., Lippi, M., and Reichlin, L. (2009). "Opening the black box: Structural factor models with large cross sections." *Econometric Theory*, 25: 1319–1347. MR2540502. doi: https://doi.org/10.1017/S026646660809052X. 2

Foster, D. P. and George, E. I. (1994). "The risk inflation criterion for multiple regression." *The Annals of Statistics*, 22: 1947–1975. MR1329177. doi: https://doi.org/10.1214/aos/1176325766. 12

Frühwirth-Schnatter, S. (2023). "Generalized cumulative shrinkage process priors with applications to sparse Bayesian factor analysis." *Philosophical Transactions of the Royal Society A*, 381: 20220148. MR4590506. 3, 8, 10, 12

Frühwirth-Schnatter, S., Hosszejni, D., and Lopes, H. (2023). "When it counts–Econometric identification of factor models based on GLT structures." *Econometrics*, 11(4): 26. doi: https://doi.org/10.3390/econometrics11040026. 3, 4, 5, 26

Frühwirth-Schnatter, S., Hosszejni, D., and Lopes, H. F. (2024). "Supplementary material for: "Sparse Bayesian factor analysis when the number of factors is unknown"." *Bayesian Analysis*. doi: https://doi.org/10.1214/24-BA1423SUPP. 10

Frühwirth-Schnatter, S. and Lopes, H. (2010). "Parsimonious Bayesian factor analysis when the number of factors is unknown." Research report, Booth School of Business, University of Chicago. 3, 4, 10, 12, 17

Frühwirth-Schnatter, S. and Lopes, H. (2018). "Sparse Bayesian factor analysis when the number of factors is unknown." arXiv:1804.04231. 3, 9, 12, 17, 21

Geweke, J. F. and 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. 13

Geweke, J. F. and Zhou, G. (1996). "Measuring the pricing error of the arbitrage pricing theory." *Review of Financial Studies*, 9: 557–587. 3

Ghahramani, Z., Griffiths, T. L., and Sollich, P. (2007). "Bayesian nonparametric latent feature models (with discussion and rejoinder)." In Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., and West, M. (eds.), *Bayesian statistics 8*. Oxford: Oxford University Press. MR2433194. 8

Ghosh, J. and Dunson, D. B. (2009). "Default prior distributions and efficient posterior computation in Bayesian factor analysis." *Journal of Computational and Graphical Statistics*, 18: 306–320. MR2749834. doi: https://doi.org/10.1198/jcgs.2009.07145. 17

Griffiths, T. L. and Ghahramani, Z. (2006). "Infinite latent feature models and the Indian buffet process." In Weiss, Y., Schölkopf, B., and Platt, J. (eds.), *Advances in neural information processing systems*, volume 18, 475–482. Cambridge, MA: MIT Press. MR2441315. 2

Grushanina, M. and Frühwirth-Schnatter, S. (2021). "Bayesian infinite factor models with non-Gaussian factors." In *JSM Proceedings, International Society of Bayesian*

Analysis (ISBA) Section, 396–415. Alexandria, VA: American Statistical Association. 26

Grushanina, M. and Frühwirth-Schnatter, S. (2023). "Dynamic mixture of finite mixtures of factor analysers with automatic inference on the number of clusters and factors." arXiv:2307.07045. 2

Hosszejni, D. and Frühwirth-Schnatter, S. (2022). "Cover it up! Bipartite graphs uncover identifiability in sparse factor analysis." arXiv:2211.00671. 4, 5, 18

Jöreskog, K. G. (1969). "A general approach to confirmatory maximum likelihood factor analysis." *Psychometrika*, 34: 183–202. MR0221659. doi: https://doi.org/10.1007/BF02289658. 3

Kastner, G. (2019). "Sparse Bayesian time-varying covariance estimation in many dimensions." *Journal of Econometrics*, 210: 98–115. MR3944765. doi: https://doi.org/10.1016/j.jeconom.2018.11.007. 2

Kastner, G., Frühwirth-Schnatter, S., and Lopes, H. F. (2017). "Efficient Bayesian inference for multivariate factor stochastic volatility models." *Journal of Computational and Graphical Statistics*, 26: 905–917. MR3765354. doi: https://doi.org/10.1080/10618600.2017.1322091. 17

Kaufmann, S. and Schuhmacher, C. (2017). "Identifying relevant and irrelevant variables in sparse factor models." *Journal of Applied Econometrics*, 32: 1123–1144. MR3714397. doi: https://doi.org/10.1002/jae.2566. 3

Kaufmann, S. and Schuhmacher, C. (2019). "Bayesian estimation of sparse dynamic factor models with order-independent and ex-post identification." *Journal of Econometrics*, 210: 116–134. MR3944766. doi: https://doi.org/10.1016/j.jeconom.2018.11.008. 3, 13, 26

Kowal, D. R. and Canale, A. (2023). "Semiparametric functional factor models with Bayesian rank selection." *Bayesian Analysis*, 18: 1161–1189. MR4675036. doi: https://doi.org/10.1214/23-ba1410. 2, 12

Lee, S.-Y. and Song, X.-Y. (2002). "Bayesian selection on the number of factors in a factor analysis model." *Behaviormetrika*, 29: 23–39. MR1894459. doi: https://doi.org/10.2333/bhmk.29.23. 2

Legramanti, S., Durante, D., and Dunson, D. B. (2020). "Bayesian cumulative shrinkage for infinite factorizations." *Biometrika*, 107: 745–752. MR4138988. doi: https://doi.org/10.1093/biomet/asaa008. 2, 4, 8, 10

Lopes, H. F. and West, M. (2004). "Bayesian model assessment in factor analysis." *Statistica Sinica*, 14: 41–67. MR2036762. 2, 13

Martin, J. K. and McDonald, R. P. (1975). "Bayesian estimation in unrestricted factor analysis: A treatment for Heywood cases." *Psychometrika*, 40: 505–517. MR0488503. doi: https://doi.org/10.1007/BF02291561. 11

Neudecker, H. (1990). "On the identification of restricted factor loading matrices: An alternative condition." *Journal of Mathematical Psychology*, 34: 237–241. MR1057287. doi: https://doi.org/10.1016/0022-2496(90)90004-S. 3

O'Hagan, A. (1995). "Fractional Bayes factors for model comparison." *Journal of the Royal Statistical Society, Ser. B*, 57: 99–138. MR1325379. 10

Owen, A. B. and Wang, J. (2016). "Bi-cross-validation for factor analysis." *Statistical Science*, 31: 119–139. MR3458596. doi: https://doi.org/10.1214/15-STS539. 2

Papastamoulis, P. and Ntzoufras, I. (2022). "On the identifiability of Bayesian factor analytic models." *Statistics and Computing*, 32: 23. MR4394853. doi: https://doi.org/10.1007/s11222-022-10084-4. 3

Pati, D., Bhattacharya, A., Pillai, N. S., and Dunson, D. B. (2014). "Posterior contraction in sparse Bayesian factor models for massive covariance matrices." *Annals of Statistics*, 42: 1102–1130. MR3210997. doi: https://doi.org/10.1214/14-AOS1215. 4

Piatek, R. and Papaspiliopoulos, O. (2018). "A Bayesian nonparametric approach to factor analysis." *Submitted*. 26

Poworoznek, E., Ferrari, F., and Dunson, D. (2021). "Efficiently resolving rotational ambiguity in Bayesian matrix sampling with matching." arXiv:2107.13783. 3

Reiersøl, O. (1950). "On the identifiability of parameters in Thurstone's multiple factor analysis." *Psychometrika*, 15: 121–149. MR0035966. doi: https://doi.org/10.1007/BF02289197. 3

Ročková, V. and George, E. I. (2017). "Fast Bayesian factor analysis via automatic rotation to sparsity." *Journal of the American Statistical Association*, 111: 1608–1622. MR3601721. doi: https://doi.org/10.1080/01621459.2015.1100620. 2, 4, 8

Schiavon, L. and Canale, A. (2020). "On the truncation criteria in infinite factor models." *Stat*, 9: e298. MR4156478. doi: https://doi.org/10.1007/s40065-018-0218-4. 10

Schiavon, L., Canale, A., and Dunson, D. B. (2022). "Generalized infinite factorization models." *Biometrika*, 109: 817–835. MR4472850. doi: https://doi.org/10.1093/biomet/asab056. 10

Teh, Y. W., Görür, D., and Ghahramani, Z. (2007). "Stick-breaking construction for the Indian buffet process." In Meila, M. and Shen, X. (eds.), *Proceedings of the eleventh international conference on artificial intelligence and statistics*, volume 2 of *Proceedings of Machine Learning Research*, 556–563. San Juan, Puerto Rico: PMLR. 8

Thurstone, L. L. (1947). *Multiple factor analysis*. Chicago: University of Chicago. MR1526847. doi: https://doi.org/10.2307/2304512. 1, 2, 11, 12

van Dyk, D. and Meng, X.-L. (2001). "The art of data augmentation." *Journal of Computational and Graphical Statistics*, 10: 1–50. MR1936358. doi: https://doi.org/10.1198/10618600152418584. 17

Wagner, H., Frühwirth-Schnatter, S., and Jacobi, L. (2023). "Factor-augmented

Bayesian treatment effects models for panel outcomes." *Econometrics and Statistics*, 28: 63–80. MR4644292. doi: https://doi.org/10.1016/j.ecosta.2022.04.003. 2

West, M. (2003). "Bayesian factor regression models in the "large p, small n" paradigm." In Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., and West, M. (eds.), *Bayesian statistics 7*, 733–742. Oxford: Oxford University Press. MR2003537. 3

Yang, R. and Berger, J. O. (1994). "Estimation of a covariance matrix using the reference prior." *The Annals of Statistics*, 22: 1195–1211. MR1311972. doi: https://doi.org/10.1214/aos/1176325625. 19

Yu, Y. and Meng, X.-L. (2011). "To center or not to center: That is not the question - An ancillarity-suffiency interweaving strategy (ASIS) for boosting MCMC efficiency." *Journal of Computational and Graphical Statistics*, 20: 531–615. MR2878987. doi: https://doi.org/10.1198/jcgs.2011.203main. 17

Zhao, S., Gao, C., Mukherjee, S., and Engelhardt, B. E. (2016). "Bayesian group factor analysis with structured sparsity." *Journal of Machine Learning Research*, 17: 1–47. MR3580349. 4, 10, 12

Invited Discussion

Gonzalo García-Donato[*]

1 Introduction

A common and challenging feature of factorial models is ignorance of the number k of latent variables. This parameter has a structural nature, with a huge effect on the final likelihood assumed. For example, if $k = 0$, the number of parameters in the factor model is m (the dimension of $\boldsymbol{y}_t$), while if $k = 1$, the underlying (unconstrained) factor model doubles its complexity with $2m$ parameters. The scenario is that of a *model selection* problem as opposed to an *estimation* problem where a single model (k in this context) is treated as known.

In this paper, the uncertainty about k is treated explicitly by assuming that k is unknown but lies in a pre-specified interval $0 \leq k \leq H$. Once this is properly subsumed within a Bayesian framework, the need to fix this parameter has been circumvented and we will be able to infer *a posteriori* about k and any other quantity of interest. The task is far from straightforward and poses extraordinary challenges that the authors address with skill. The result is a thorough addition to the literature on factor models, greatly expanding our understanding of these very popular tools in econometric applications.

A crucial aspect of this paper is the use of (Dirac) spike and slab priors (hereafter DSS) for the factor coefficients $\boldsymbol{\beta}_H = (\beta_{ij})_{ij}$ with $1 \leq i \leq m$ and $1 \leq j \leq H$. The (i, j) component of the associated binary matrix $\boldsymbol{\delta}_H$ is zero if $\beta_{ij} = 0$ (a possible event because of the positive – the spike – mass at zero). These special priors are the key ingredient to substantiate the desired uncertainty about the number of factors and subsequent relational aspects over their components. For example, k becomes the number of non-zero columns in $\boldsymbol{\delta}_H$, and so on.

DSSs are one of the many priors that have emerged from the model selection literature. An unambiguous feature that reveals their model selection nature is that the slab component (usually a Gaussian density) is proper. If such a component were improper or would be a vague density, the results would be essentially arbitrary (Berger, 2006). Among the alternatives, DSS have the main distinction of assuming independence (perhaps conditional on hyperparameters) among their components. For variable selection, this leads to suboptimal priors (Bayarri et al., 2012), but their usefulness in solving complex problems like the one in this paper is unquestionable. Because of this independence, DSS obscures the existing differences between model selection and estimation. This is because each model prior is implicitly defined by integration, but this is not generally true for model selection priors (and the prior for a model nested in a particular model does not coincide with the marginal of the larger model). The surprising consequence is that the progress made in Bayesian model selection has had

[*] Department of Economy and Finance, University of Castilla-La Mancha, gonzalo.garciadonato@uclm.es

little impact on the progress made in DSS priors (and vice versa!). In a sense, the two lines of research have evolved in isolation from each other over the past decades. In this regard, the effort in this paper to incorporate more sophisticated model selection priors, such as the fractional priors, is a solid step towards reconciliation.

My discussion aims to revisit aspects of this work from a model selection perspective, trying to stimulate the possible benefits of such interactions.

2 Reconciling terms

Model selection (also called model choice or model uncertainty) is a branch of statistics that explicitly assumes that the model generating the data is unknown. Usually, this flexibility is limited to the assumption that the true model belongs to a fixed set of possibilities known as model space ($\mathcal{M}$); the so-called $\mathcal{M}$-closed perspective. Within the Bayesian paradigm, and given its intrinsic ability to handle all kinds of uncertainty, many important problems in statistics have been approached through the lens of model selection. This was the route taken by H. Jeffreys (Jeffreys, 1961) for the paradigmatic case of testing, where each hypothesis entertained is made equivalent to a competing model. Another very popular example is variable selection, where each subset of the originally considered variables defines a possible model, and where we find one of the origins of DSS priors (Mitchell and Beauchamp, 1988).

Almost automatically, Bayesian model selection procedures are parsimonious, in accordance with Occam's razor postulate (Berger and Pericchi, 2001). In modern language, we say that it induces sparsity, a desirable property exploited in the present work. Sparsity is a consequence of i) explicitly considering all models as plausible alternatives, and ii) a proper prior over the additional parameters, which has the effect of penalizing complexity.

The simplest model (say M_0) in $\mathcal{M}$ occupies a relevant place in model selection – in this work, M_0 is the model with only idiosyncratic variances, σ_i^2, and $\boldsymbol{\mu}$ –. M_0 allows us to distinguish between common parameters and new parameters. For common parameters the literature suggests (see e.g. Bayarri et al., 2012) that, under convenient reparameterizations, we can use objective (perhaps improper) priors, justifying limiting distributions of Eq. 3.14. This way avoids the need to manage additional hyperparameters; is completely objective and would likely counteract the reported Heywood problem. Of course, the devil is in the details and finding a reparameterization that makes all models invariant under the same group (Berger et al., 1998) is challenging. For new parameters – β_{ij} – the prior distribution must be a proper prior. The fact that this prior is usually centered on zero (cf. Eqs. 3.10–3.12) is also related to the importance of M_0 and leads to the second observation about its important role. The simplest model must be a sensible model, usually requiring an intercept ($\boldsymbol{\mu}$) that can be replaced by standardization (as is done in this paper).

Inference under model uncertainty is a complex problem called *model averaging* (MA) (which recognizes the fact that reports are the result of weighting inferences from different models). A highly recommended recent review of the topic with an emphasis on economics is Steel (2020). For prediction, MA is safe, but for estimation (e.g., to infer about β_{ij} or $\mathbf{\Lambda}$), we must be convinced that the parameters being weighted have a compatible meaning across models. Further, we must be prepared to aggregate posterior distributions that mix discrete and continuous distributions. For this reason, the Bayesian model selection software (García-Donato and Forte, 2018) returns MA in a way that takes into account the idiosyncratic nature of these parameters.

In this paper, because H is fixed, we are in an $\mathcal{M}$-closed problem (allowing $H = \infty$ has similarities to the $\mathcal{M}$-open perspective). The cardinality of the model space without restrictions is 2^{mH}, a number that grows easily with m and H. For example, for the application in Section 5.2, $\mathcal{M}$ has 2^{220}, a number of the order of 10^{66}. In the present paper, a very promising MCMC scheme is proposed to study such challenging $\mathcal{M}$, which has a reversible jump engine. The design of specific algorithms able to handle very large model spaces has been a fruitful area of research in model selection in recent years (Zanella, 2020; Zhou et al., 2022, see for example). Broadly speaking, the idea is to sample δ_{ij} in a way that prioritizes the best models and preserves the essential properties of an MCMC.

3 Sparsity vs. multiplicity

In model selection settings, the prior on δ_H usually has a large impact on the results, especially when $\mathcal{M}$ has a large cardinality. In the case of factor models with an unknown number of factors, such potential sensitivity is perhaps more worrisome given the dependence on H, a parameter that is fixed with some degree of arbitrariness.

Without constraints, the prior adopted here assumes that $\delta_{ij} \sim \text{Ber}(\tau_j)$ and the probability of success, τ_j, follows a beta distribution that depends on two hyperparameters that have independent gamma densities (Table 1). This prior induces both column and row sparsity. For the NYSE example with $H = 31$ and $m = 63$, this choice would lead to the prior on dimensionality k shown in Figure 1 (left) in this discussion. In the right side, I have plotted the distribution on k obtained with the constant prior $\delta_{ij} \sim \text{Ber}(.5)$.

There is no consensus in the literature on the exact role of the prior over model space. As in the present paper, a large majority of authors have used this prior to incorporate an additional sparsity effect (but recall that Bayesian model selection is already parsimonious). Castillo et al. (2015) is a prominent example in variable selection. Other authors have argued that such a prior should be responsible for controlling for multiplicity: the fact that more populated dimensions artificially increase their influence for purely combinatorial reasons (Scott and Berger, 2010). A clear message in Scott and Berger (2010) is that the constant prior does not provide control in this sense and should be avoided. This last hypothesis is the one assumed in García-Donato and Paulo (2022) for the closely related case of variable selection with qualitative variables (factors). There, the prior for $\boldsymbol{\delta}_H$ is assigned in such a way that it adjusts for column multiplicity – as opposed to column sparsity –, since all column dimensions receive the same probability (it is inversely proportional to $\binom{H}{k}$). This prior has the attractive additional property of being completely objective, independent of any parameter. The

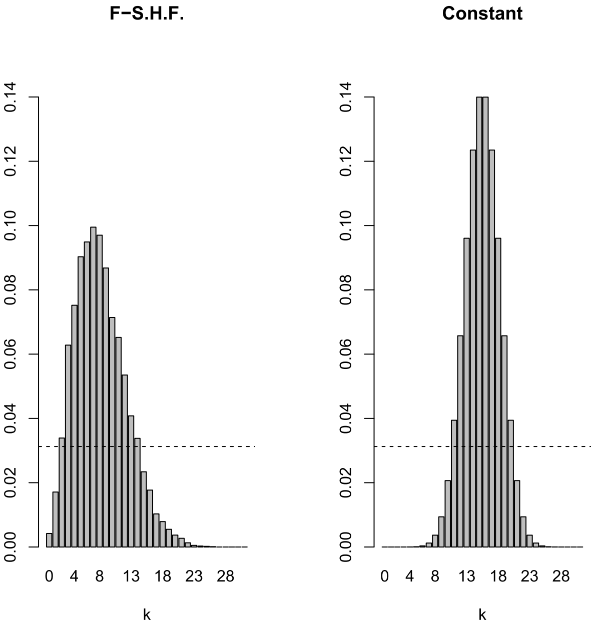

Figure 1: For $m = 63$ and $H = 31$, the prior on k induced by the prior on $\boldsymbol{\delta}_H$ without constraints. On the left, the authors' proposal; on the right, the constant prior. The dashed line shows a prior that adjusts for multiplicity.

dashed line in Figure 1 corresponds to this prior. Note that it is constant over the dimensions.

For each of the above possibilities, it is difficult to assess what the final prior would be once the relevant constraints in the present problem are incorporated. In the examples in the paper, the posterior distribution of k seems to be concentrated near $\frac{H}{2}$ (the most populated dimensions), which does not seem a strong sparse response. It also makes me think about the issue of multiplicity (this would be a revealing symptom in variable selection) and whether the proposed prior behaves similarly to the constant prior (as the similarities shown in Figure 1 seem to indicate).

Funding

This work has been partially supported by the Spanish Ministery of Science and Innovation under grant number PID2022-138201NB-I00 and by the Junta de Comunidades de Castilla-La Mancha under grant SBPLY/21/180501/000241.

References

Bayarri, M. J., Berger, J. O., Forte, A., and García-Donato, G. (2012). "Criteria for Bayesian model choice with application to variable selection." *The Annals of Statistics*, 40: 1550–1577. MR3015035. doi: https://doi.org/10.1214/12-AOS1013. 32, 33

Berger, J. O. (2006). "The case for objective Bayesian analysis." *Bayesian Analysis*, 1(3): 385–402. MR2221271. doi: https://doi.org/10.1214/06-BA115. 32

Berger, J. O., Pericchi, L. R., and Varshavsky, J. (1998). "Bayes factors and marginal distributions in invariant situations." *Sankhya: The Indian Journal of Statistics, Series A*, 60: 307–321. MR1718789. 33

Berger, J. O. and Pericchi, R. L. (2001). "Objective Bayesian methods for model selection: introduction and comparison (with discussion)." In Lahiri, P. (ed.), *Model Selection*, 135–207. Institute of Mathematical Statistics Lecture Notes- Monograph Series, volume 38. MR2000753. doi: https://doi.org/10.1214/lnms/1215540968. 33

Castillo, I., Schmidt-Hieber, J., and van der Vaart, A. (2015). "Bayesian linear regression with sparse priors." *Annals of Statistics*, 43: 1986–2018. MR3375874. doi: https://doi.org/10.1214/15-AOS1334. 34

García-Donato, G. and Forte, A. (2018). "Bayesian testing, variable selection and model averaging in linear models using R with BayesVarSel." *The R Journal*, 10(1): 155–174. 34

García-Donato, G. and Paulo, R. (2022). "Variable selection in the presence of factors: a model selection perspective." *Journal of the American Statistical Association*, 117(540): 1847–1857. MR4528475. doi: https://doi.org/10.1080/01621459.2021.1889565. 34

Jeffreys, H. (1961). *Theory of Probability*. London: Oxford University Press. MR0187257. 33

Mitchell, T. and Beauchamp, J. (1988). "Bayesian variable selection in linear regression." *Journal of the American Statistical Association*, 83: 1023–1032. MR0997578. 33

Scott, J. and Berger, J. (2010). "Bayes and Empirical-Bayes multiplicity adjustment in the variable-selection problem." *The Annals of Statistics*, 38: 2587–2619. MR2722450. doi: https://doi.org/10.1214/10-AOS792. 34

Steel, M. F. J. (2020). "Model averaging and its use in economics." *Journal of Economic Literature*, 58(3): 644–719. 34

Zanella, G. (2020). "Informed proposals for local MCMC in discrete spaces." *Journal of the American Statistical Association*, 115(530): 852–865. MR4107684. doi: https://doi.org/10.1080/01621459.2019.1585255. 34

Zhou, Q., Yang, J., Vats, D., Roberts, G. O., and Rosenthal, J. S. (2022). "Dimension-free mixing for high-dimensional Bayesian variable selection." *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(5): 1751–1784. MR4515557. doi: https://doi.org/10.1111/rssb.12546. 34

Invited Discussion

Niko Hauzenberger[*] and Gary Koop[†]

Introduction

Conditional on knowing the number of factors, r, analysis in static and dynamic factor models is straightforward for the Bayesian. However, inference on r is challenging. A Bayesian could use marginal likelihoods to select the number of factors (see Geweke, 1996). But in the standard big data setups nowadays (which involve a large number of variables/measurements m), this is computationally cumbersome, requiring the estimation of a large set of models that vary in r ($\leq m$).

Frühwirth-Schnatter et al. (2024) address this issue using an elegant combination of an *identified* factor model and a shrinkage prior which can *select* the number of factors (column-wise shrinkage) and shrink the factor loadings on active factors (row-wise shrinkage). Their strategy involves starting with an overfitting model — as is often done in the literature on mixture models, see e.g., Malsiner-Walli et al. (2016) and Grushanina and Frühwirth-Schnatter (2023) — then eliminating spurious factors (columns) and introducing additional sparsity in loadings (rows) of active factors. These additional exact zero factor loadings (achieved through row sparsification) not only make a parsimonious factor model even more parsimonious but also facilitate identification of the remaining active factors. In terms of computation, Frühwirth-Schnatter et al. (2024) use a novel and efficient reversible jump Markov chain Monte Carlo (MCMC) sampler that allows for the number of (active) factors r to vary during sampling. All in all, this paper is a valuable addition to the Bayesian factor literature.

There are three directions that paper differs from conventional approaches to Bayesian factor analysis: identification, prior choice and computation. We will organize our discussion around these three aspects. We will conclude with some thoughts on potential extensions of this model.

Identification

It is well known that without further restrictions, the static factor model is unidentified. The restrictions selected by the unordered generalized lower triangular (UGLT) structure used by the authors are estimated agnostically from the data, rather than imposed ad hoc or a priori. As explained in Frühwirth-Schnatter et al. (2024) and also in earlier work by the authors, Frühwirth-Schnatter et al. (2023), UGLT identification has advantages over the conventional identification scheme which involves assuming the

[*] Department of Economics, University of Strathclyde, Glasgow, United Kingdom, niko.hauzenberger@strath.ac.uk

[†] Department of Economics, University of Strathclyde, Glasgow, United Kingdom, gary.koop@strath.ac.uk

factor loading matrix to be lower triangular with positive numbers on the diagonal (they refer to this identification scheme as PLT). UGLT is much more flexible and is likely to be as good an identification restriction that is possible in the class of schemes that achieve identification through zero restrictions on the factor loadings. However, UGLT — similar to other more conventional zero restriction schemes such as PLT — can have the drawback that it does not necessarily always guarantee order invariance. Although, in terms of order invariance, freeing up the exact positions of the zero factor loadings constitutes a substantial improvement upon PLT and other ad hoc zero restriction schemes, UGLT still requires a minimal number of zero restrictions to ensure identifications of the r factors through r linearly independent rows in the factor loading matrix. This can make UGLT prone to a lack of order invariance as well.

But why is order invariance relevant for a Bayesian in the first place? Order invariance implies that posterior and predictive results depend on the way the variables are ordered. In the large Bayesian Vector Autoregression (VAR) literature there is a growing recognition that standard approaches are not order invariant and that the empirical effect of a lack of order invariance can be substantial. For instance, two different orderings of the variables might lead to almost identical point forecasts, but substantially different predictive variances and thus substantially different log predictive likelihoods, see Arias et al. (2023) and Chan et al. (2024).

As noted by Frühwirth-Schnatter et al. (2024), there are other identification schemes used in factor models. Chan et al. (2018), referred to as CLS herafter, achieve identification without using zero restrictions on the factor loadings. CLS consider the static factor model directly as a reduced-rank regression and develop a fully invariant specification of that regression model. The details of their identification scheme are not germane to the present discussion other than to note that it leads to order invariance. However, their empirical work suggests ordering issues are potentially important in factor models. In an empirical illustration involving six variables, CLS show how two different orderings can lead to log marginal likelihoods that differ by about 142 when using the PLT identification scheme. Of course, the log marginal likelihood is the same for every possible ordering using the identification scheme they suggest. CLS therefore strongly recommend using an order-invariant specification. Alternatively, researchers could also estimate the variable ordering from the data, similar to Wu and Koop (2023) in the VAR context, or average over all possible orderings. However, the latter strategy is feasible only when working with small m.

CLS provide theoretical/formal derivations and discussion about issues that arise when using zero restrictions on factor loadings. Let $\mathbf{\Lambda}_1$ be the $r \times r$ matrix containing the r rows of the factor loading matrix that are restricted to ensure identification. CLS show that the lack of order invariance arises with any identification scheme, such as the UGLT one, which restricts $\mathbf{\Lambda}_1$ to be non-singular, imposing r linearly independent rows in the factor loading matrix. This non-singularity rules out points where $|\mathbf{\Lambda}_1| = 0$ and this leads to a discontinuity which plays a key role in the transformation between different orderings. CLS show how their approach, which does not rule out $|\mathbf{\Lambda}_1| = 0$, allows for straightforward evaluation of marginal likelihoods for different choices of r using the Savage-Dickey density ratio. Thus, choosing the number of factors using marginal likelihoods is easy to do unlike in conventional approaches such as PLT and UGLT. Of course, Frühwirth-Schnatter et al. (2024) have an alternative method of choosing the number of factors using a clever hierarchical prior. But it is worth noting that the identification scheme of CLS has one good property that UGLT may lack when $|\mathbf{\Lambda}_1| \to 0$ (i.e., order invariance). And therefore it might be worth comparing UGLT with the CLS approach for extreme cases where $|\mathbf{\Lambda}_1| \approx 0$, and to investigate how UGLT behaves in the presence of discontinuities when the ordering of variables is most influential (as discussed in Section 3 of CLS).

Some Bayesians are happy working with unidentified models (at least when forecasting) since combining a proper prior with an unidentified likelihood will typically lead to a proper posterior and predictive. This allows us to speculate that, even without the UGLT identification restrictions, the model developed in Frühwirth-Schnatter et al. (2024) could be a very interesting one. Furthermore, in the recent VAR literature, identification can be achieved through relaxing the homoskedasticity and Normality assumptions for the VAR errors. This can be done, e.g., by allowing for stochastic volatility, regime-switching, or fat-tailed errors (see Rigobon, 2003; Lewis, 2022; Bertsche and Braun, 2022) instead of imposing exact zero restrictions on the error covariance matrix. Relaxing some assumptions in the Normal and homoskedastic static factor model of Frühwirth-Schnatter et al. (2024) might be one way (of many ways) forward, thereby combining UGLT with the identification through heteroskedasticity approach proposed by Sentana and Fiorentini (2001) for the static factor model.

In summary, the UGLT identifying structure of Frühwirth-Schnatter et al. (2024) does have some very nice properties as outlined in their paper. This makes it a useful addition to the Bayesian factor literature. However, other approaches exist with different properties which may have different advantages, especially as related to order invariance. When choosing identifying restrictions, the Bayesian must weigh the pros and cons of each. And it may not even be necessary to make a choice of identifying restrictions on the factor loadings if working with an unidentified model suffices or if identification is achieved in other ways (e.g., via heteroskedasticity).

Prior

Frühwirth-Schnatter et al. (2024) propose an exchangeable shrinkage process prior that does have many attractive properties. Specifically, this prior allows a researcher to effectively let the data decide on the number of (active) factors r in an almost tuning-free, automatic manner. In addition, it achieves row sparsity in the relevant block of the factor loading matrix associated with the active factors. Shrinkage along these two dimensions naturally leads to the quest of the desired/optimal level of column sparsity (which relates to the overall parsimony of the factor model) and row sparsity (which relates to the simplicity of the remaining structures according to terminology of Frühwirth-Schnatter et al. (2024)).

When working with exploratory factor models, researchers would most likely agree that it is desirable to obtain a column sparse specification with a small number of active factors only (i.e., where r is rather small) and where this small set of factors may even be sensible to interpret. Frühwirth-Schnatter et al. (2024) achieve column sparsity by combining a Dirac spike and slab prior on each factor loading with an exchangeable shrinkage process on column-specific inclusion probabilities, which increasingly pushes columns towards zero and thus automatically eliminates superfluous factors.

However, it is less obvious whether row sparsity is a generally desirable feature. This depends of course on the specific time series data at hand, but in macroeconomics the *illusion of sparsity* has recently received considerable attention (Giannone et al., 2021; Fava and Lopes, 2021; Gruber and Kastner, 2022). Giannone et al. (2021) list factor models as a typical dense statistical technique. But what about *sparse* factor models that aggressively induce row sparsity? In some applications, it may be desirable to have all these few factors load on many time series and thus be able to explain most of the variation in the measurements. This would be associated with a row-dense factor loading matrix and — according to Giannone et al. (2021) — such a row-dense but column-sparse factor model may be indeed considered dense overall, since most measurements load on at least one (common) factor. Frühwirth-Schnatter et al. (2024) consider two applications: one application uses monthly exchange rate data and the other uses monthly stock market returns. What both applications have in common is that the factors that tend to load on many time series are easier to interpret, while the more idiosyncratic factors (with only a very few associated non-zero factor loadings) tend to be more difficult to interpret. For example, in the financial application using stock market returns, the market factor (which loads (equally) on almost every single firm return and acts like a cross-sectional average or first principal component) and the industry-specific factors (which load on almost every firm within a given industry) can be labelled and interpreted relatively straightforwardly.

In the VAR context, Gruber and Kastner (2022) discuss the sparsity-inducing properties of various popular shrinkage priors using a sparsity measure proposed in Hoyer (2004). In the context of a static factor model and given a specific factor, this measure defines the sparsest possible estimate as having only one non-zero loading, while the densest estimate is defined as having all measurements load equally on this factor. It could be worthwhile to use such a sparsity measure to assess the a priori imposed overall degree of sparsity of the shrinkage prior.

Sampling and Computation

Frühwirth-Schnatter et al. (2024) propose an efficient reversible jump MCMC sampler. To substantially improve the sampling efficiency of the sparse factor model, they use MCMC boosting by considering either ancillarity-sufficiency interweaving strategy (ASIS) or marginal data augmentation (MDA) steps. In applications of Frühwirth-Schnatter et al. (2024), m is of moderate size ($m = 22$ in the exchange rate application and $m = 63$ in the stock return application) and a static factor model is assumed. In the static case, the proposed algorithm likely scales well even in higher dimensions using hundreds of variables. But what if a researcher wishes to use a dynamic factor model, where the state equation of the factors evolves according to a VAR. For example, in the case of $m = 201$, this would amount for an upper bound for the number of factor

$r^* = \frac{m-1}{2} = 100$ in their overfitting model. Is the proposed reversible jump MCMC computationally efficient in such a case? Probably yes, but only if the true number of dynamic factors is low.

Furthermore, Frühwirth-Schnatter et al. (2024) highlight the fact that working with an unidentified model and leaving the factor loading matrix fully free and unrestricted may harm posterior inference and sampling efficiency. Even in the unidentified case, post-processing might still be a valid option, particularly relying on the methods proposed in Kaufmann and Schumacher (2019), Chakraborty et al. (2020) or Bolfarine et al. (2024). For example, Bolfarine et al. (2024) represents a straightforward yet effective approach for ex-post sparsification of the factor loading matrix. This method aims to obtain a sparse posterior representation of posterior estimates and to decide on the number of factors r based on a loss measure. As argued by Bolfarine et al. (2024), it is not necessarily a competing approach but rather a complementary device and could be used for any overfitting model equipped with hierarchical shrinkage priors, as it just needs the posterior as input.

Potential Extensions from a Practitioner's View

In this section, we will discuss potential extensions from a practitioner's view, working in the field of macroeconomics or finance. Frühwirth-Schnatter et al. (2024) is about the Normal, homoskedastic, static factor model. Any of these assumptions could be relaxed or changed. Our discussion will mainly center on the question of what desirable features a factor model — used off the shelf for analyzing macroeconomic and financial time series data — should have.

The empirical macroeconomist would probably find the dynamic factor model the most interesting extension of the model of Frühwirth-Schnatter et al. (2024) since most macroeconomic data exhibits dependence over time. This would be straightforward to do although, as noted above, it could potentially cause problems for computation unless the number of (active) factors is small.

A second extension, commonly done with both macroeconomic and financial times series data, would involve adding stochastic volatility. This, too, would be straightforward to add. However, the recent Covid-19 pandemic, geopolitical tensions and earlier financial and Eurozone crises, raise the issue as to whether simply adding stochastic volatility is enough. These events caused severe economic shocks and turbulence on the financial markets. It is possible that the fundamental relationships between the r latent factors and the m observed measurements may have changed in response to these events. Accounting for this would require a very flexible model that not only allows the variance of the factors and/or idiosyncratic shocks to vary over time but also allows for time-varying factor loadings. Relating this discussion to that on sparsity, such a model would imply dynamic row sparsity and dynamic column sparsity (i.e., time-varying dimensions of the factor loading matrix). Such extensions would not be difficult to add and may be necessary when working with macroeconomic or financial data sets which include crisis periods.

Summary and Conclusions

Frühwirth-Schnatter et al. (2024) is an exceptionally fine paper and the methods described therein should belong in any practitioner's toolbox. In this discussion, we have offered some thoughts about identification in their model, highlighting the issue of order invariance. We have also discussed the prior and computational issues. The prior of Frühwirth-Schnatter et al. (2024) has attractive properties and, as their title emphasizes, their approach is about "sparse" Bayesian factor models. But the title of another paper we cite, Gruber and Kastner (2022), ends with the "Sparse or dense? It depends!" and we offer some thoughts on their prior in light of the sparse versus dense debate. The methods of Frühwirth-Schnatter et al. (2024) could be adapted to allow for row density instead of sparsity.

On computation, our comments relate to computational efficiency with larger m or r. We speculate that their methods would work well in the static factor model of any dimension, and in the dynamic factor model if r is small. But there may be worries with large r or in more complicated models. Furthermore, we offer some additional thoughts on the use of post-processing methods.

There are a myriad of interesting extensions of the static factor model and we discuss a few of them likely to be of most interest to the practitioners and argue that extending the methods of Frühwirth-Schnatter et al. (2024) to handle them would be straightforward.

Funding

N. Hauzenberger gratefully acknowledges financial support from the Oesterreichische Nationalbank (OeNB, Anniversary Fund, project no. 18763) and from the Austrian Science Fund (FWF), ZK-35.

References

Arias, J. E., Rubio-Ramirez, J. F., and Shin, M. (2023). "Macroeconomic forecasting and variable ordering in multivariate stochastic volatility models." *Journal of Econometrics*, 235(2): 1054–1086. MR4602902. doi: https://doi.org/10.1016/j.jeconom.2022.04.013. 38

Bertsche, D. and Braun, R. (2022). "Identification of structural vector autoregressions by stochastic volatility." *Journal of Business & Economic Statistics*, 40(1): 328–341. MR4356576. doi: https://doi.org/10.1080/07350015.2020.1813588. 39

Bolfarine, H., Carvalho, C. M., Lopes, H. F., and Murray, J. S. (2024). "Decoupling shrinkage and selection in Gaussian linear factor analysis." *Bayesian Analysis*, 19(1): 181–203. MR4692547. doi: https://doi.org/10.1214/22-ba1349. 41

Chakraborty, A., Bhattacharya, A., and Mallick, B. K. (2020). "Bayesian sparse multiple regression for simultaneous rank reduction and variable selection." *Biometrika*, 107(1): 205–221. MR4064149. doi: https://doi.org/10.1093/biomet/asz056. 41

Chan, J., Leon-Gonzalez, R., and Strachan, R. W. (2018). “Invariant inference and efficient computation in the static factor model.” *Journal of the American Statistical Association*, 113(522): 819–828. MR3832229. doi: https://doi.org/10.1080/01621459.2017.1287080. 38

Chan, J. C., Koop, G., and Yu, X. (2024). “Large order-invariant Bayesian VARs with stochastic volatility.” *Journal of Business & Economic Statistics*, 42(2): 825–837. MR4729010. doi: https://doi.org/10.1080/07350015.2023.2252039. 38

Fava, B. and Lopes, H. F. (2021). “The illusion of the illusion of sparsity: An exercise in prior sensitivity.” *Brazilian Journal of Probability and Statistics*, 35(4): 699–720. MR4350956. doi: https://doi.org/10.1214/21-bjps503. 40

Frühwirth-Schnatter, S., Hosszejni, D., and Lopes, H. F. (2024). “Sparse Bayesian factor analysis When the number of factors is unknown.” *Bayesian Analysis*, forthcoming. 37, 38, 39, 40, 41, 42

Frühwirth-Schnatter, S., Hosszejni, D., and Lopes, H. F. (2023). “When it counts – Econometric identification of the basic factor model based on GLT structures.” *Econometrics*, 11(4). 37

Geweke, J. (1996). “Bayesian reduced rank regression in econometrics.” *Journal of Econometrics*, 75(1): 121–146. MR1414507. doi: https://doi.org/10.1016/0304-4076(95)01773-9. 37

Giannone, D., Lenza, M., and Primiceri, G. E. (2021). “Economic predictions with big data: The illusion of sparsity.” *Econometrica*, 89(5): 2409–2437. 40

Gruber, L. and Kastner, G. (2022). “Forecasting macroeconomic data with Bayesian VARs: Sparse or dense? It depends!” *arXiv preprint arXiv:2206.04902*. MR4300614. doi: https://doi.org/10.1007/978-3-030-31150-6_3. 40, 42

Grushanina, M. and Frühwirth-Schnatter, S. (2023). “Dynamic mixture of finite mixtures of factor analysers with automatic inference on the number of clusters and factors.” *arXiv preprint arXiv:2307.07045*. MR2265601. 37

Hoyer, P. O. (2004). “Non-negative matrix factorization with sparseness constraints.” *Journal of Machine Learning Research*, 5(9). MR2248024. 40

Kaufmann, S. and Schumacher, C. (2019). “Bayesian estimation of sparse dynamic factor models with order-independent and ex-post mode identification.” *Journal of Econometrics*, 210(1): 116–134. MR3944766. doi: https://doi.org/10.1016/j.jeconom.2018.11.008. 41

Lewis, D. J. (2022). “Robust inference in models identified via heteroskedasticity.” *Review of Economics and Statistics*, 104(3): 510–524. 39

Malsiner-Walli, G., Frühwirth-Schnatter, S., and Grün, B. (2016). “Model-based clustering based on sparse finite Gaussian mixtures.” *Statistics and Computing*, 26(1): 303–324. MR3439375. doi: https://doi.org/10.1007/s11222-014-9500-2. 37

Rigobon, R. (2003). “Identification through heteroskedasticity.” *Review of Economics and Statistics*, 85(4): 777–792. 39

Sentana, E. and Fiorentini, G. (2001). "Identification, estimation and testing of conditionally heteroskedastic factor models." *Journal of Econometrics*, 102(2): 143–164. MR1842239. doi: `https://doi.org/10.1016/S0304-4076(01)00051-3`. 39

Wu, P. and Koop, G. (2023). "Estimating the ordering of variables in a VAR using a Plackett–Luce prior." *Economics Letters*, 230: 111247. MR4618824. doi: `https://doi.org/10.1016/j.econlet.2023.111247`. 38