

# 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  $\mathbf{f}_t = (f_{1t} \cdots f_{rt})'$ :

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

where the covariance matrix  $\boldsymbol{\Sigma}_0$  of the idiosyncratic errors  $\boldsymbol{\epsilon}_t$  is a diagonal matrix and  $\mathbf{\Lambda}$  is the  $m \times r$  matrix of factor loadings  $\Lambda_{ij}$  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](mailto: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](mailto: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, [hedibertf@insper.edu.br](mailto:hedibertf@insper.edu.br)

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:

$$\boldsymbol{\Omega} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Sigma}_0. \quad (1.2)$$

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  $\mathbf{u}_t = \mathbf{y}_t - \boldsymbol{\mu}$  exhibits a factor structure as in (1.2). In a factor-augmented model with conditional mean  $\boldsymbol{\mu}_t$ , the zero-mean innovations  $\mathbf{u}_t = \mathbf{y}_t - \boldsymbol{\mu}_t$  (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

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

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

Bayesian approaches with  $H = \infty$  a priori 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  $\beta_k$ , is random a priori 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  $\beta_H$  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  $\beta_{ij}$  of  $\beta_H$  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  $\beta_{ij}$  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  $\beta_H$  to obtain final estimates of the unknown factor dimension  $r$  and a unique rotation  $\mathbf{\Lambda}$  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  $\beta_H$  during sampling; however, leaving  $\beta_H$  unconstrained makes it difficult to recover the true number of factors and to estimate  $\mathbf{\Lambda}$ .

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  $\mathbf{\Lambda}$ . Econometric identification yields a unique decomposition of the covariance matrix  $\mathbf{\Omega}$  in (1.2) into the cross-covariance matrix  $\mathbf{\Lambda}\mathbf{\Lambda}'$  and the covariance matrix  $\mathbf{\Sigma}_0$  of the uncorrelated idiosyncratic errors and identifies a unique factor loading matrix  $\mathbf{\Lambda}$  from  $\mathbf{\Lambda}\mathbf{\Lambda}'$ . Even if the decomposition is unique (which need not be the case), it is well-known that  $\mathbf{\Lambda}$  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  $\mathbf{\Lambda}$  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  $\mathbf{\Omega}$  into  $\mathbf{\Lambda}\mathbf{\Lambda}'$  and  $\mathbf{\Sigma}_0$ ; 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  $\beta_H$  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  $\beta_H$  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  $\Sigma_0$  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  $\beta_{ij}$ , 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.

$$\mathbf{y}_t = \beta_H \mathbf{f}_t^H + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}_m(\mathbf{0}, \Sigma_H), \quad \mathbf{f}_t^H \sim \mathcal{N}_H(\mathbf{0}, \mathbf{I}_H). \quad (2.1)$$

We impose an exchangeable shrinkage process prior on the columns of  $\beta_H$  to achieve column sparsity with  $k < H$  non-zero columns, collected in the  $m \times k$  submatrix  $\beta_k$ , see Section 3.1 for details. We summarize sparsity by the so-called sparsity matrix  $\delta_H$  which is a binary indicator matrix of 0s and 1s of the same dimension as  $\beta_H$  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  $\delta_k$  be the sparsity matrix corresponding to the non-zero columns  $\beta_k$  of  $\beta_H$ . To achieve identification in a sparse EFA model, we assume that  $\delta_k$  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  $j$ th 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  $\beta_k \beta_k'$  are uniquely identified from the covariance matrix  $\boldsymbol{\Omega} = \beta_k \beta_k' + \boldsymbol{\Sigma}_k$  implied by the non-zero columns  $\beta_k$  of  $\boldsymbol{\beta}_H$  and by  $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}_H$ . Since variance identification implies that  $\boldsymbol{\Lambda} \boldsymbol{\Lambda}' = \beta_k \beta_k'$ , it follows that  $r = k$  and  $\beta_r = \boldsymbol{\Lambda} \mathbf{P}$  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  $\beta_k$  and  $\boldsymbol{\Lambda}$  leads to rotational identification up to signed permutations  $\beta_k \mathbf{P}_\pm \mathbf{P}_\rho$ , where the permutation matrix  $\mathbf{P}_\rho$  corresponds to one of  $k!$  possible column permutations in  $\beta_k$  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  $\beta_k$  is variance identified,  $r = k$  and  $\boldsymbol{\Lambda}$  is uniquely recovered by reordering the columns of  $\beta_r$  such that the pivots  $l_1 < \dots < l_r$  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 posterior 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  $\beta_k$  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  $\beta_k$ . 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

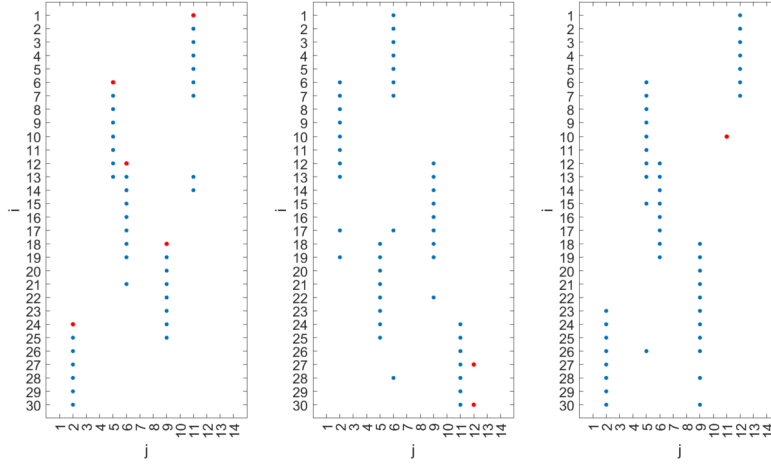


Figure 1: Posteriors draws of  $\delta_k$  from sparse BFA with  $m = 30$  and  $H = 14$  (zero loadings are left blank). Left:  $\delta_k$  with  $k = 5$  obeying the 3579 counting rule with the pivot rows  $(l_1, l_2, l_3, l_4, l_5) = (24, 6, 12, 18, 1)$  (marked red); center:  $\delta_k$  with  $k = 6$  violating the 3579 counting rule due to a column with only two non-zero elements (marked red); right:  $\delta_k$  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  $\mathbf{\Lambda}$ . More specifically, Frühwirth-Schnatter et al. (2023, Theorem 4) show as a third contribution that imposing a UGLT structure on the non-zero columns  $\beta_k$  of  $\beta_H$  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  $\beta_k$  and  $\Sigma_k$  take the following form:

$$\beta_k = (\mathbf{\Lambda} \quad \mathbf{\Xi}) \mathbf{P}_{\pm} \mathbf{P}_{\rho}, \quad \mathbf{\Xi} = \begin{pmatrix} \mathbf{0} \\ \Xi_{l_{\text{sp}}} \\ \mathbf{0} \end{pmatrix}, \quad \Sigma_k = \text{Diag}(\sigma_1^2, \dots, \sigma_{l_{\text{sp}}}^2 - \Xi_{l_{\text{sp}}}^2, \dots, \sigma_m^2), \quad (2.2)$$

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  $l_{\text{sp}}$  different from the pivot rows  $\mathbf{l}_r = (l_1, \dots, l_r)$  in  $\mathbf{\Lambda}$ . A similar representation holds for higher degrees  $k > r$  of overfitting, with  $\mathbf{\Xi}$  containing  $s = k - r$  spurious columns that obey a UGLT structure, i.e. the pivots  $\mathbf{l}_{\Xi}$  of  $\mathbf{\Xi}$  lie in different rows and are distinct from the pivots  $\mathbf{l}_r$  of  $\mathbf{\Lambda}$ .

Hence, if a posterior draw  $\beta_k$  from the EFA model (2.1) contains  $s$  spurious columns  $\mathbf{\Xi}$ , then they can be absorbed into the idiosyncratic errors by defining their covariance matrix as  $\Sigma_r = \Sigma_k + \mathbf{\Xi}\mathbf{\Xi}'$ . This leaves  $r = k - s$  active columns  $\beta_r$  (i.e. columns with at least two non-zero loadings) in  $\beta_H$ , which are extracted and postprocessed as above: if  $\beta_r$  obeys the 3579 counting rule, then  $\mathbf{\Lambda}$  is identified up to a signed permutation from  $\beta_r = \mathbf{\Lambda}\mathbf{P}_{\pm}\mathbf{P}_{\rho}$ , while  $\Sigma_0 = \Sigma_r$ , and the number of active columns  $r$  provides a posterior draw of the unknown factor dimension. Otherwise,  $\beta_r$  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  $\Xi$  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,

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

by absorbing the  $r_{sp}$  spurious columns  $\Xi$  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  $r_{sp} \in \{1, \dots, k-r\}$  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  $\Xi$  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  $\boldsymbol{\Sigma}_H = \boldsymbol{\Sigma}_r - \Xi \Xi'$ . The only non-zero loading  $\Xi_{l_{sp}}$  in  $\Xi$  can take any value such that the  $l_{sp}$ -th diagonal element of  $\boldsymbol{\Sigma}_H$  remains positive, i.e.  $\Sigma_{H, l_{sp}, l_{sp}} = \sigma_{l_{sp}}^2 - (\Xi_{l_{sp}})^2 > 0$ . This entire move only affects the  $l_{sp}$ -th row  $\boldsymbol{\beta}_{r, l_{sp}, \cdot}$  of  $\boldsymbol{\beta}_r$ . More specifically, for  $t = 1, \dots, T$ :

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

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

## 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  $\beta_{ij}$  in  $\beta_H$ ,

$$\beta_{ij}|\tau_j \sim (1 - \tau_j)\Delta_0 + \tau_j P_{\text{slab}}(\beta_{ij}), \quad (3.1)$$

where  $\Delta_0$  is a Dirac-spike at zero and  $P_{\text{slab}}$  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  $\tau_1, \dots, \tau_H$ :

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

The ESP prior turns model (2.1) into a *sparse* EFA model, where the number  $k$  of non-zero columns in  $\delta_H$  is random a priori, 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,

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

which converges to the 2PB prior (Ghahramani et al., 2007) for  $H \rightarrow \infty$ . For  $\gamma = 1$ , the finite one-parameter-beta (1PB) prior results which converges to the Indian buffet process prior (Teh et al., 2007) for  $H \rightarrow \infty$  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  $\alpha$  and  $\gamma$  are equipped with the hyperpriors  $\alpha \sim \mathcal{G}(a^\alpha, b^\alpha)$  and  $\gamma \sim \mathcal{G}(a^\gamma, b^\gamma)$ , since they are instrumental in controlling prior column sparsity. For the 1PB prior, for instance, the decreasing order statistics  $\tau_{(1)} > \dots > \tau_{(H)}$  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):

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

With the largest slab probability following  $\tau_{(1)} \sim \mathcal{B}(\alpha, 1)$ , 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  $\beta_H$ , 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  $\delta_k$  of the sparsity matrix  $\delta_H$  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  $l_j \in \mathcal{L}(\mathbf{l}_{r,-j})$ :

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



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  $\delta_{ij}$  below  $l_j$  are subject to variable selection,

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

with column-specific probability  $\tau_j$  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  $j$ th column  $\boldsymbol{\delta}_{:,j}^r$  of  $\boldsymbol{\delta}^r$  can be expressed both conditionally as well as marginalized w.r.t.  $\tau_j$ :

$$\Pr(\boldsymbol{\delta}_{:,j}^r | l_j, \tau_j) = \tau_j^{d_j-1} (1 - \tau_j)^{m-l_j-d_j+1}, \quad (3.7)$$

$$\Pr(\boldsymbol{\delta}_{:,j}^r | l_j) = \frac{B(a_H + d_j - 1, b_H + m - l_j - d_j + 1)}{B(a_H, b_H)}. \quad (3.8)$$

### 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  $\boldsymbol{\Sigma}_r = \text{Diag}(\sigma_1^2, \dots, \sigma_m^2)$  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:

$$\Xi_{l_{sp}}^2 | \sigma_{l_{sp}}^2 \sim \mathcal{U} \left[ 0, \sigma_{l_{sp}}^2 \right]. \quad (3.9)$$

This ensures for all  $l_{sp} \in \mathbf{l}_\Xi$  that  $\boldsymbol{\Sigma}_{k,l_{sp},l_{sp}} = \sigma_{l_{sp}}^2 - \Xi_{l_{sp}}^2 > 0$ . 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.<sup>1</sup>

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  $P_{\text{slab}}$  (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  $\boldsymbol{\beta}_i^\delta$  of unconstrained elements takes the form  $\boldsymbol{\beta}_i^\delta | \sigma_i^2 \sim \mathcal{N}_{q_i}(\mathbf{0}, \mathbf{B}_{i0}^\delta \sigma_i^2)$ , where  $\mathbf{B}_{i0}^\delta$  is a diagonal matrix. The variance of this prior is assumed to depend on the idiosyncratic variance  $\sigma_i^2$ , 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.

<sup>1</sup>Note that this is a major improvement compared to Frühwirth-Schnatter and Lopes (2018).

A common choice for  $P_{\text{slab}}$  is to introduce a global shrinkage parameter  $\kappa$ ,

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

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

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

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

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

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  $p(\beta_{i\cdot}^\delta|\sigma_i^2, b, \mathbf{f}_r) \propto p(\mathbf{y}_i^\delta|\mathbf{f}_r, \beta_{i\cdot}^\delta, \sigma_i^2)^b$  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  $\sigma_1^2, \dots, \sigma_m^2$  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  $\beta_{i1}, \dots, \beta_{ir}$  together with  $\sigma_i^2$  determine the *communalities*  $R_i^2$  as the proportion of variance explained by the common factors:

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

where  $\Omega_{ii} = \sum_{\ell=1}^r \beta_{i\ell}^2 + \sigma_i^2$  is the  $i$ th diagonal element of  $\mathbf{\Omega}$ . 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,  $\sigma_i^2 \sim \mathcal{G}^{-1}(c^\sigma, C_0)$ , 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  $\Omega_{11}, \dots, \Omega_{mm}$ . In this case, it is preferable to use an inverse gamma prior with heterogenous scales  $C_{0i}$ :

$$\sigma_i^2 \sim \mathcal{G}^{-1}(c^\sigma, C_{0i}). \quad (3.14)$$

We may assume that  $C_{0i} = b_i^\sigma$  are fixed hyperparameters. Alternatively, assuming random hyperparameters  $C_{0i} \sim \mathcal{G}(a^\sigma, a^\sigma/b_i^\sigma)$  with  $E(C_{0i}) = b_i^\sigma$  leads to a more general

prior which can be expressed as a rescaled F-distribution with the same prior expectation  $E(\sigma_i^2) = b_i^\sigma / (c^\sigma - 1)$  as (3.14), provided that  $c^\sigma > 1$ :

$$\sigma_i^2 \sim \frac{b_i^\sigma}{c^\sigma} F(2a^\sigma, 2c^\sigma). \quad (3.15)$$

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  $\sigma_i^2$ . Nevertheless, there exists a Bayesian analogue of the Heywood problem which takes the form of multi-modality of the posterior of  $\sigma_i^2$  with one mode lying at 0. Heywood problems typically occur if the constraint  $1/\sigma_i^2 \geq (\mathbf{\Omega}^{-1})_{ii}$  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  $\delta_k$  to satisfy the 3579 counting rule discussed in Section 2.1 is the following upper bound for  $k$ :

$$k \leq \lfloor (m-1)/2 \rfloor, \quad (3.16)$$

Prior distributions	Parameters	Values
Prior for $\tau_j, j = 1, \dots, H$		
$\tau_j   \alpha, \gamma, H \sim \mathcal{B}(\gamma \frac{\alpha}{H}, \gamma)$ ,	$a^\alpha, b^\alpha, H$	$a^\alpha = n_0, b^\alpha = a^\alpha(H - E_q)/H/E_q$
$\alpha \sim \mathcal{G}(a^\alpha, b^\alpha), \gamma \sim \mathcal{G}(a^\gamma, b^\gamma)$	$a^\gamma, b^\gamma$	$a^\gamma = b^\gamma = n_0$
Priors for $\sigma_i^2, i = 1, \dots, m$		
$\sigma_i^2 \sim \mathcal{G}^{-1}(c_0, C_0)$	$c_0, C_0$	$c_0 = 1, C_0 = 0.3$
$\sigma_i^2 \sim \mathcal{G}^{-1}(c^\sigma, b_i^\sigma)$	$c^\sigma, b_i^\sigma$	$b_i^\sigma = (c^\sigma - 1)(1 - E_R)\Omega_{ii}$
$\sigma_i^2 \sim (b_i^\sigma/c^\sigma) F(2a^\sigma, 2c^\sigma)$	$c^\sigma, b_i^\sigma, a^\sigma$	$a^\sigma = n_0$
Slab priors for $\beta_{ij}$		
Fractional prior	$b$	$b = 1/(mT)$
$\beta_{ij}   \sigma_i^2, \kappa \sim \mathcal{N}(0, \kappa \sigma_i^2)$ ,	$c^\kappa, b^\kappa$	$c^\kappa = n_0$
$\kappa \sim \mathcal{G}^{-1}(c^\kappa, b^\kappa)$		$b^\kappa = c^\kappa E_R / (1 - E_R) / E_q$
$\beta_{ij}   \theta_j, \sigma_i^2, \kappa \sim \mathcal{N}(0, \kappa \theta_j \sigma_i^2)$ ,	$c^\kappa, b^\kappa$	
$\kappa \sim \mathcal{G}^{-1}(c^\kappa, b^\kappa), \theta_j \sim F(2a^\theta, 2c^\theta)$	$a^\theta, c^\theta$	$a^\theta = n_0, c^\theta = 2.5$
$\beta_{ij}   \omega_{ij}, \theta_j, \sigma_i^2, \kappa \sim \mathcal{N}(0, \kappa \theta_j \sigma_i^2 \omega_{ij})$ ,	$c^\kappa, b^\kappa$	
$\kappa \sim \mathcal{G}^{-1}(c^\kappa, b^\kappa), \theta_j \sim F(2a^\theta, 2c^\theta)$ ,	$a^\theta, c^\theta$	$a^\omega = c^\omega = 0.5$ (horseshoe)
$\omega_{ij} \sim F(2a^\omega, 2c^\omega)$	$a^\omega, c^\omega$	$a^\omega = c^\omega = 0.2$ (triple gamma)

Table 1: Prior choices depending on five hyperparameters with default values  $H = \lfloor (m-1)/2 \rfloor$ ,  $E_q = 2$ ,  $E_R = 2/3$ ,  $c^\sigma = 2.5$  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  $\alpha \sim \mathcal{G}(a^\alpha, b^\alpha)$  accordingly. The choice of  $\alpha$  strongly impacts the expected row sparsity  $E_q = E(q_i|\alpha, H)$ , given by

$$E_q = \frac{\alpha}{1 + \alpha/H},$$

independently of  $\gamma$ . To match a prior guess of  $E_q$  with the prior expectation  $E_\alpha = E(\alpha|H) = H \cdot E_q / (H - E_q)$  of  $\alpha$ , we bind a given value of  $a^\alpha$  to the scale parameter  $b^\alpha = a^\alpha / E_\alpha$ . For large  $H$ , this yields  $E_\alpha \approx E_q$  a priori, whereas  $E_\alpha$  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  $E_q = 2$ . To center the 2PB prior at the 1PB prior (corresponding to  $\gamma = 1$ ), we choose  $b^\gamma = a^\gamma$  for a given value of  $a^\gamma$ .

For the exchangeable prior  $\sigma_i^2 \sim \mathcal{G}^{-1}(c^\sigma, C_0)$ , 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  $c^\sigma$  in prior (3.14) and (3.15) large enough to bound the prior of  $1/\sigma^2$  away from 0. Depending on the data,  $c^\sigma$  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  $b_i^\sigma$  both in (3.14) and (3.15) such that (3.13) holds on average, i.e.  $E(\sigma_i^2) = E(1 - R_i^2) E(\Omega_{ii})$ . Based on a prior guess  $E_R$  of the average amount of explained variance, this yields  $b_i^\sigma = (c^\sigma - 1)(1 - E_R)\bar{\Omega}_{ii}$ , 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  $\kappa$  for all hierarchical shrinkage priors and bind a given value of  $c^\kappa$  to the scale parameter  $b^\kappa = c^\kappa E_\kappa$ , where

$$E_\kappa = \frac{E_R}{(1 - E_R)E_q} \quad (3.17)$$

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  $\delta_{ij}$  and the standardized loadings  $\beta_{ij}^* = \beta_{ij} / \sqrt{\sigma_i^2 \kappa} \sim \mathcal{N}(0, 1)$ :

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

Using that the sum follows a  $\chi_{q_i}^2$ -distribution, and taking the expectation of both sides of the second equation yields (3.17). Various priors for the column specific shrinkage

parameters  $\theta_j$  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  $\theta_j$  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  $c^\sigma$  controlling Heywood problems, the prior expectation  $E_q$  of row sparsity, the prior expected fraction of explained variance  $E_R$  and the hyperparameters  $a^\alpha, a^\gamma, c^\kappa, a^\theta$  and, for the rescaled F-prior (3.15), also  $a^\sigma$ . 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  $\delta_H$  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  $\delta_H$ , and Kaufmann and Schuhmacher (2019), who sample  $\delta_H$  without imposing any constraint, we impose a UGLT structure on  $\delta_H$  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  $\delta_r$ . Due to the prior specification in Section 3, the number  $r_{sp}$  of spurious columns is a sufficient statistic for the remaining columns in  $\delta_H$  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  $\delta_r$  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  $\tau_j$  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  $\delta_{ij}$  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  $\tau_j$  (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  $\delta_r$ :
- (H) Update any unknown hyperparameters in the ESP prior (3.2) without conditioning on the slab probabilities  $\tau_r = (\tau_1, \dots, \tau_r)$ . 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  $\delta_r$  in a random order:
    - (a) Sample all indicators  $\delta_{ij}$  below the pivot  $l_j$  from  $p(\delta_{ij} | l_j, \delta_{r,-j}^r, \mathbf{f}_r, \tau_j, \mathbf{y})$  conditional on the remaining columns  $\delta_{r,-j}^r$ , the factors  $\mathbf{f}_r = (\mathbf{f}_1^r, \dots, \mathbf{f}_T^r)$  and  $\tau_j$ , without conditioning on  $\beta_r$  and  $\sigma_1^2, \dots, \sigma_m^2$ .
    - (b) If column  $\delta_{r,j}^r$  is spurious after this update, increase  $r_{sp}$  by one. Remove the  $j$ th column from  $\delta_r$ , the factors  $f_{jt}, t = 1, \dots, T$  from  $\mathbf{f}_r$  and  $\tau_j$  from  $\tau_r$  to define, respectively,  $\delta_{r-1}$ ,  $\mathbf{f}_{r-1}$  and  $\tau_{r-1}$  and decrease  $r$  by one.
  - (L) Loop over all columns  $j$  of  $\delta_r$  in a random order and sample a new pivot row  $l_j$  from  $p(l_j | \delta_{r,-j}^r, \mathbf{f}_r, \mathbf{y})$  without conditioning on  $\beta_r, \sigma_1^2, \dots, \sigma_m^2$  and the slab probabilities  $\tau_r$ . If column  $\delta_{r,j}^r$  is spurious after this update, proceed as in Step (D-b).
  - (P) Sample the model parameters  $\beta_r$  and  $\sigma_1^2, \dots, \sigma_m^2$  jointly conditional on the sparsity matrix  $\delta_r$  and the factors  $\mathbf{f}_r = (\mathbf{f}_1^r, \dots, \mathbf{f}_T^r)$  from  $p(\beta_r, \sigma_1^2, \dots, \sigma_m^2 | \delta_r, \mathbf{f}_r, \mathbf{y})$ .
  - (F) Sample the latent factors  $\mathbf{f}_r = (\mathbf{f}_1^r, \dots, \mathbf{f}_T^r)$  conditional on the model parameters  $\beta_r$  and  $\sigma_1^2, \dots, \sigma_m^2$  from  $p(\mathbf{f}_1^r, \dots, \mathbf{f}_T^r | \beta_r, \sigma_1^2, \dots, \sigma_m^2, \mathbf{y})$ .
  - (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  $\sigma_1^2, \dots, \sigma_m^2$ .
  - (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  $\delta_H$ .
  - (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  $\delta_r$ . Order the spurious columns such that  $l_{\Xi,1} < \dots < l_{\Xi,r_{sp}}$ .
  - (R-F) Loop over all spurious columns  $j_{sp}$  and sample the spurious factors  $\mathbf{f}_{j_{sp}} = (f_{j_{sp},1}, \dots, f_{j_{sp},T})$  independently for all  $t = 1, \dots, T$  from  $f_{j_{sp},t} | \mathbf{f}_t^r, \beta_r, \sigma_{l_{sp}}^2, y_{l_{sp},t} \sim \mathcal{N}(E_{j_{sp},t}, V_{j_{sp}})$ , where  $U_{j_{sp}}$  is a draw from a uniform distribution on  $[-1,1]$  and
 
$$V_{j_{sp}} = 1 - U_{j_{sp}}^2, \quad E_{j_{sp},t} = U_{j_{sp}}(y_{l_{sp},t} - \beta_{r,l_{sp},t} \mathbf{f}_t^r) / \sqrt{\sigma_{l_{sp}}^2}. \quad (4.1)$$
  - (R-H) Sample  $\tau_{j_{sp}} | l_{sp} \sim \mathcal{B}(a_H, b_H + m - l_{sp})$  for all spurious columns  $j_{sp}$ .
  - (R-D) Update all spurious columns from the last (with the largest pivot row) to the first (with the smallest pivot row): sample all  $(\delta_{i,j_{sp}}, i \in \{l_{sp} + 1, \dots, m\})$  below the pivot  $l_{sp}$  conditional on  $\tau_{j_{sp}}, \delta_r, \mathbf{f}_r$  and  $\mathbf{f}_{j_{sp}}$  without conditioning on  $\beta_r, \Xi$  and  $\sigma_1^2, \dots, \sigma_m^2$ . If a spurious column  $j_{sp}$  is turned into an active one, then decrease  $r_{sp}$  by 1, increase  $r$  by 1, add  $\delta_{\cdot,j_{sp}}$  to  $\delta_r$  and  $\mathbf{f}_{j_{sp}}$  to  $\mathbf{f}_r$ . Otherwise, remove  $\delta_{\cdot,j_{sp}}$  from  $\delta_{\Xi}$  and  $\mathbf{f}_{j_{sp}}$  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  $\beta_r$ , and  $\sigma_1^2, \dots, \sigma_m^2$ , see Appendix C.3. In Step (F), the conditional posterior  $p(\mathbf{f}_1^r, \dots, \mathbf{f}_T^r | \beta_r, \sigma_1^2, \dots, \sigma_m^2, \mathbf{y})$  factors into independent normal distributions given by:

$$\mathbf{f}_t^r | \mathbf{y}_t, \beta_r, \Sigma_r \sim \mathcal{N}_r \left( (\mathbf{I}_r + \beta_r' \Sigma_r^{-1} \beta_r)^{-1} \beta_r' \Sigma_r^{-1} \mathbf{y}_t, (\mathbf{I}_r + \beta_r' \Sigma_r^{-1} \beta_r)^{-1} \right), \quad (4.2)$$

where  $\Sigma_r = \text{Diag}(\sigma_1^2, \dots, \sigma_m^2)$ . For the hierarchical Gaussian priors (3.11) and (3.12), all unknown shrinkage parameters and, for the rescaled F-prior (3.15) on  $\sigma_1^2, \dots, \sigma_m^2$ , also the scaling parameters  $C_{01}, \dots, C_{0m}$  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  $\beta_H$ . Exploiting the results of Section 2.2, spurious columns in  $\delta_H$  are added and deleted in Step (R-S) by reversible jump MCMC (RJMCMC). Very conveniently, this step is independent of the pivots  $\mathbf{I}_\Xi$  and the loadings  $\Xi$  in the spurious columns, since the prior  $p(\delta_H, \beta_H, \Sigma_H | r_{sp})$  is invariant to the specific choice of  $\mathbf{I}_\Xi$  and  $\Xi$ , given  $r_{sp}$ . However, the prior odds that a zero column in  $\delta_H$  can be turned into an additional spurious column are equal to:

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

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  $\delta_H$  into a spurious column, with proposal density  $q_{\text{split}}(\delta_H^{\text{new}} | \delta_H) = p_s / (H - r - r_{sp})$ . A merge move turns one of the  $r_{sp} > 0$  spurious columns in  $\delta_H$  into a zero column, with proposal density  $q_{\text{merge}}(\delta_H^{\text{new}} | \delta_H) = p_s / r_{sp}$ . A split move is accepted with probability  $\min(1, A_{\text{split}}(r, r_{sp}))$ , where:

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

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

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

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  $\mathbb{I}_{\Xi}|\mathbf{I}_r$ , while the spurious factor loadings  $\Xi_{l_{sp}}$  are sampled from the prior (3.9). Given  $l_{sp}$ , the idiosyncratic variance  $\sigma_{l_{sp}}^2$  in the CFA model is split, with the help of a random variable  $U_{j_{sp}} \sim \mathcal{U}[-1, 1]$ , between  $\Xi_{l_{sp}}$  and an updated idiosyncratic variance  $\sigma_{l_{sp}}^{2,\text{new}}$ . More specifically:

$$\Xi_{l_{sp}} = U_{j_{sp}} \sqrt{\sigma_{l_{sp}}^2}, \quad \sigma_{l_{sp}}^{2,\text{new}} = (1 - U_{j_{sp}}^2) \sigma_{l_{sp}}^2. \quad (4.6)$$

Given  $\Xi_{l_{sp}}$  and  $\sigma_{l_{sp}}^{2,\text{new}}$ , factors  $f_{j_{sp},t}$  are proposed in Step (R-F) for each  $t = 1, \dots, T$  from the conditional density  $p(f_{j_{sp},t} | \mathbf{f}_t^r, \boldsymbol{\beta}_r, \sigma_{l_{sp}}^2, y_{l_{sp},t})$  given in (4.1). The slab probabilities  $\tau_{j_{sp}}$  are sampled in Step (R-H) as in Algorithm 1, Step (H), using that  $d_{j_{sp}} = 1$ . 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  $f_{j_{sp},t}$ . 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  $\mathbf{l}_r = (l_1, \dots, l_r)$  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  $l_j^{\text{new}}$  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  $d_j^{\text{new}}$  being the new number of non-zero elements in column  $j$ , the prior ratio  $R_{\text{move}}$  can be derived from (3.8):

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

Further details are provided in Appendix G.

## 4.3 Boosting MCMC

Step (F) and Step (P) in Algorithm 1 sample the factors  $(\mathbf{f}_1^r, \dots, \mathbf{f}_T^r)$  conditional on  $(\boldsymbol{\beta}_r, \boldsymbol{\Sigma}_r)$  and  $(\boldsymbol{\beta}_r, \boldsymbol{\Sigma}_r)$  conditional on  $(\mathbf{f}_1^r, \dots, \mathbf{f}_T^r)$ . Depending on the signal-to-noise ratio,



such full conditional Gibbs sampling tends to be poorly mixing. In a factor model where  $\mathbf{f}_t^r \sim \mathcal{N}_r(\mathbf{0}, \mathbf{I}_r)$ , the information in the data (the “signal”) can be quantified by the matrix  $\beta_r' \Sigma_r^{-1} \beta_r$  in comparison to the identity matrix  $\mathbf{I}_r$  (the “noise”) in the filter for  $\mathbf{f}_t^r | \mathbf{y}_t, \beta_r, \Sigma_r$ , 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  $\beta_r' \Sigma_r^{-1} \beta_r$  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-sufficiency interweaving strategy (ASIS) introduced by Yu and Meng (2011).

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

$$\mathbf{y}_t = \tilde{\beta}_r \tilde{\mathbf{f}}_t^r + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}_m(\mathbf{0}, \Sigma_r), \quad \tilde{\mathbf{f}}_t^r \sim \mathcal{N}_r(\mathbf{0}, \Psi), \quad (4.8)$$

where  $\Psi = \text{Diag}(\Psi_1, \dots, \Psi_r)$  is diagonal. The two systems are related by the transformations  $\tilde{\mathbf{f}}_t^r = (\Psi)^{1/2} \mathbf{f}_t^r$  and  $\tilde{\beta}_r = \beta_r (\Psi)^{-1/2}$ , where the nonzero elements in  $\tilde{\beta}_r$  have the same position as the nonzero elements in  $\beta_r$  and the sparsity matrix  $\delta_r$  is not affected by the transformation. The main difference between MDA and ASIS lies in the choice of  $\Psi$ . While  $\Psi_j$  is sampled from a working prior for MDA,  $\Psi_j$  is chosen in a deterministic fashion for ASIS. For illustration, Figure 2 shows posterior draws of  $\text{tr}(\beta_r' \Sigma_r^{-1} \beta_r)$  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  $\theta_j$  into the state equation by choosing  $\Psi_j = \theta_j$ . For the F-prior (3.15) on  $\sigma_1^2, \dots, \sigma_m^2$ , another useful strategy is *row boosting*, based on moving the random scales  $C_{0i}$  from the prior  $\sigma_i^2$  to the observation equation in all rows of the basic factor model. Full details for all boosting steps are provided in Appendix H.

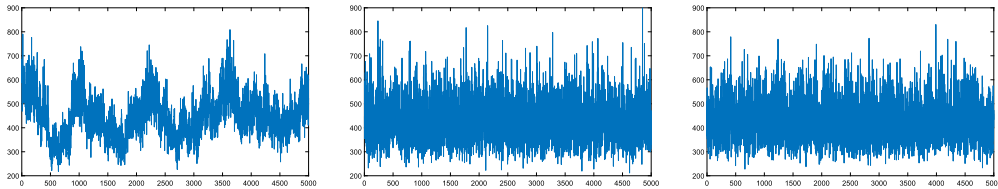


Figure 2: Exchange rate data (standardized, fractional prior and prior (3.14) for  $\sigma_i^2$ ); posterior draws of  $\text{tr}(\beta_r' \Sigma_r^{-1} \beta_r)$  without boosting (left-hand side), boosting through ASIS with  $\sqrt{\Psi_j}$  equal to the largest loading (in absolute values) (middle) and through MDA based on the working prior  $\Psi_j \sim \mathcal{G}^{-1}(1.5, 1.5)$  (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  $\sigma_1^2, \dots, \sigma_m^2$ , the modelsize  $d = \sum_{i,j} \delta_{ij}$ , and the communalities  $R_1^2, \dots, R_m^2$  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  $l_1, \dots, l_r$ . To estimate the factor loading matrix, further inference is performed conditionally on the posterior mode  $\tilde{r}$  and an estimator  $\hat{\mathbf{l}}_{\tilde{r}} = (\hat{l}_1, \dots, \hat{l}_{\tilde{r}})$  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  $E(\boldsymbol{\Lambda}|\mathbf{y}, \hat{\mathbf{l}}_{\tilde{r}})$ . 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  $\Lambda_{ij}$  and  $\delta_{ij}^\Lambda$  in  $\boldsymbol{\Lambda}$  and corresponding the sparsity matrix  $\boldsymbol{\delta}^\Lambda$ . Alternative estimators such as the sequence of pivots  $\mathbf{I}^*$  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.<sup>2</sup> We choose a 2PB prior to ensure column sparsity, combine various slab distributions for  $\beta_{ij}$  with various priors on  $\sigma_i^2$ , and use the default hyperparameters introduced in Table 1, namely  $H = \lfloor (m-1)/2 \rfloor$ ,  $c^\sigma = 2.5$ ,  $E_R = 2/3$ ,  $E_q = 2$  and  $a^\alpha = a^\gamma = c^\kappa = a^\theta = a^\sigma = n_0 = 6$ .

<sup>2</sup>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  $r_{\text{true}} = 5$  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  $\sigma_i^2$  is combined with the following slab distributions  $P_{\text{slab}}$  for  $\beta_{ij}$ : 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  $P_{\text{true}} = \Pr(\tilde{r} = r_{\text{true}}|\mathbf{y})$ , the posterior risk measures  $R_\Omega = E(L(\mathbf{\Omega}_r, \mathbf{\Omega}_0)|\mathbf{y})$ ,  $R_\Sigma = E(L(\mathbf{\Sigma}_r, \mathbf{\Sigma}_0)|\mathbf{y})$  and  $R_{\Omega^{-1}} = E(L(\mathbf{\Omega}_r^{-1}, \mathbf{\Omega}_0^{-1})|\mathbf{y})$  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  $\text{TP}_\Omega$  for non-zero and the false positive rate  $\text{FP}_\Omega$  for zero correlations in  $\mathbf{\Omega}_0$ , the bias  $B_d = E(d_r|\mathbf{y}) - d_{\text{true}}$  in model size, and the true positive rate  $\text{TP}_\delta$  and the false positive rate  $\text{FP}_\delta$  for the true sparsity pattern  $\delta^\Lambda$ . 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  $P_{\text{slab}}$ .

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  $P_{\text{slab}}$  has considerable impact on recovering the true sparsity pattern  $\delta^\Lambda$  in  $\mathbf{\Lambda}$  and  $\mathbf{\Omega}_0$ . The fractional prior has the smallest false positive rates  $\text{FP}_\delta$  and  $\text{FP}_\Omega$  both for  $\delta^\Lambda$  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  $\text{TP}_\delta$  and  $\text{TP}_\Omega$  both for  $\delta^\Lambda$  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.

	$\tilde{r}$	$P_{\text{true}}$	$R_{\Omega}$	$R_{\Omega^{-1}}$	$R_{\Sigma}$	$\text{TP}_{\Omega}$	$\text{FP}_{\Omega}$	$B_d$	$\text{TP}_{\delta}$	$\text{FP}_{\delta}$	
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.

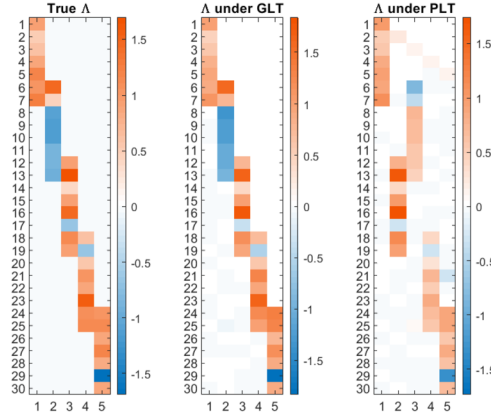


Figure 3: Comparing the true loading matrix  $\mathbf{\Lambda}$  (left) with the estimated loading matrix  $\hat{\mathbf{\Lambda}}$  under the UGLT condition with  $r$  unknown (middle) and under the PLT condition with  $r = 5$  known (right-hand side) for a randomly selected data set under the dedicated 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.<sup>3</sup> The data are demeaned and standardized. Note that  $H = 10$ . We combine the fractional prior (C.4) with the following priors on  $\sigma_i^2$ : prior (3.14) (HIG) and (3.15) (HF) with default settings, prior (3.14) with  $b_i^\sigma$  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.

<sup>3</sup>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

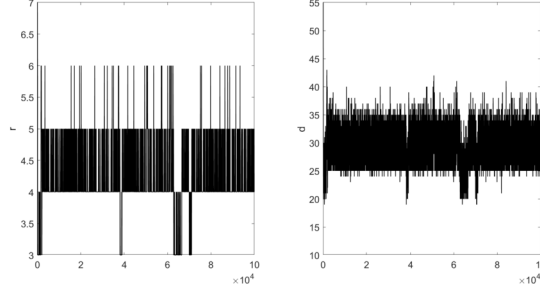


Figure 4: Exchange rate data (standardized); posterior draws of the factor dimension  $r$  (left) and model size  $d$  (right) including burn-in (fractional prior combined with the HIG prior).

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

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

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  $\alpha$  and  $\gamma$  under the fractional prior and various priors for  $\sigma_i^2$ .

Prior on $\sigma_i^2$	$\Pr(q_i = 0 \mathbf{y})$						
	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  $\delta_r$ , 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  $\delta_r$  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  $i$ th row sum of  $\delta_r$ . 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.

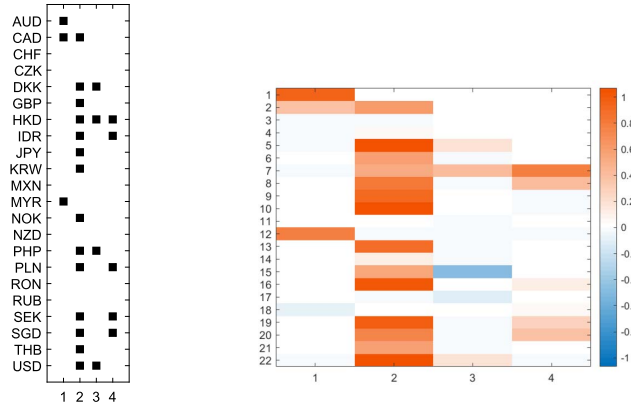


Figure 5: Exchange rate data (standardized); left hand side: sparsity matrix  $\delta_4$  corresponding to the median probability model (identical for all four priors); right hand side: estimated loading matrix  $E(\Lambda | \hat{l}_4, \mathbf{y})$  with  $\hat{l}_4 = (1, 2, 5, 7)$  for the HIG prior (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  $\beta_r$  are rotated into a GLT structure  $\Lambda$  by ordering the pivots such that  $l_1 < \dots < l_r$ . The sequence of pivots visited most often among all draws with  $r = \tilde{r} = 4$  is equal to  $\hat{l}_4 = (1, 2, 5, 7)$  for all priors and coincides with the sequence of pivots  $\mathbf{l}^*$  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  $\Lambda$ . All GLT draws where the pivots  $\mathbf{l}_4$  coincide with  $\hat{l}_4 = (1, 2, 5, 7)$  are used to identify the GLT representation of the factor loading matrix  $\Lambda$  and the marginal inclusion probabilities  $\Pr(\delta_{ij} = 1 | \mathbf{y}, \hat{l}_4)$ . 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.<sup>4</sup> Note that

<sup>4</sup> $T = 247$  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).

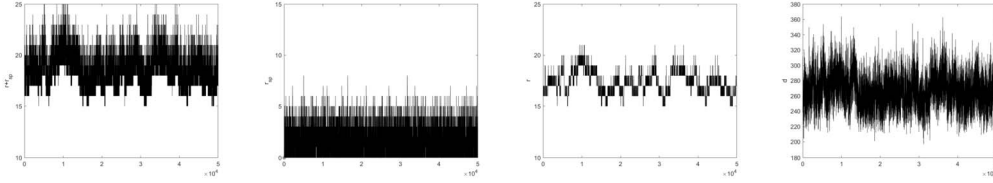


Figure 6: NYSE data; from left to right: posterior draws of the total number of non-zero columns  $r + r_{sp}$ , the number of spurious columns  $r_{sp}$ , the extracted number of factors  $r$  and the model dimension  $d$ .

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

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

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

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

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

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  $\sigma_i^2$ . 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  $\mathbf{A}$  by ordering the pivots such that  $l_1 < \dots < l_r$ . The sequence of pivots visited most often among all draws of varying dimension  $r$  is equal to  $\mathbf{I}^* = (1, 2, 3, 4, 5, 8, 9, 11, 13, 15, 18, 19, 20, 32, 46, 48)$  which implies that the estimator  $r^* = 16$  is identical with the posterior mode  $\tilde{r} = 16$ . Furthermore, the sequence of pivots  $\hat{\mathbf{I}}_{16}$  visited most often among all draws of dimension  $r = 16$  coincides with  $\mathbf{I}^*$ .



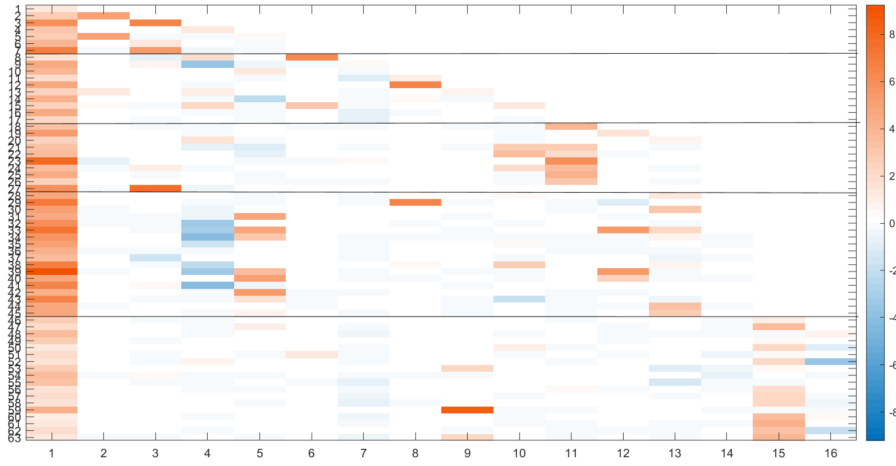


Figure 7: NYSE data; estimated GLT representation of the factor loading matrix  $\Lambda$ .

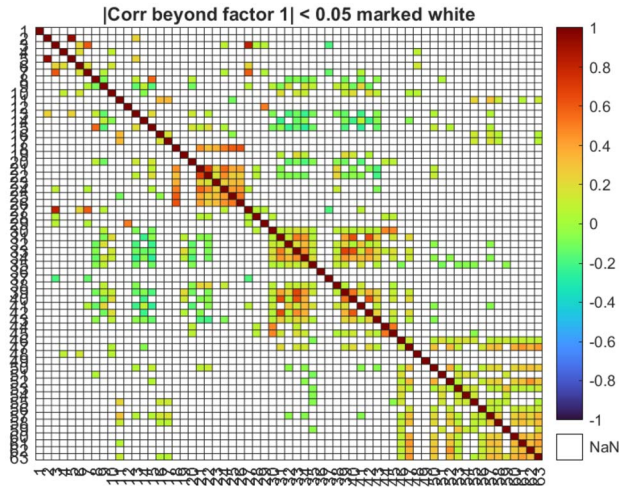


Figure 8: NYSE data; estimated marginal correlation matrix  $E(\Omega^*|\mathbf{y})$ , where  $\Omega_{i\ell}^* = \text{Corr}((y_{it} - \Lambda_{i1}f_{1t})(y_{\ell t} - \Lambda_{\ell 1}f_{1t}))$ .

All GLT draws where the pivots  $\mathbf{l}_r$  coincide with  $\mathbf{l}^* = \hat{\mathbf{l}}_{16}$  are used to identify the GLT representation of the factor loading matrix  $\Lambda$ , 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  $\Omega^*$  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  $\delta_{ij}$  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](https://doi.org/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-aos1456>. 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](https://arxiv.org/abs/2307.07045). 2
- Hosszejni, D. and Frühwirth-Schnatter, S. (2022). “Cover it up! Bipartite graphs uncover identifiability in sparse factor analysis.” [arXiv:2211.00671](https://arxiv.org/abs/2211.00671). 4, 5, 18
- Jöreskog, K. G. (1969). “A general approach to confirmatory maximum likelihood factor analysis.” *Psychometrika*, 34: 183–202. [MR0221659](https://doi.org/10.1007/BF02289658). 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](https://doi.org/10.1016/j.jeconom.2018.11.007). 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](https://doi.org/10.1080/10618600.2017.1322091). 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](https://doi.org/10.1002/jae.2566). 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](https://doi.org/10.1016/j.jeconom.2018.11.008). 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](https://doi.org/10.1214/23-ba1410). 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](https://doi.org/10.2333/bhmk.29.23). 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](https://doi.org/10.1093/biomet/asaa008). 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](https://doi.org/10.1007/BF0236762). 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](https://doi.org/10.1007/BF02291561). 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](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](https://arxiv.org/abs/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-sufficiency 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  $\mathbf{y}_t$ ), 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  $\beta_H = (\beta_{ij})_{ij}$  with  $1 \leq i \leq m$  and  $1 \leq j \leq H$ . The  $(i, j)$  component of the associated binary matrix  $\delta_H$  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  $\delta_H$ , 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](mailto: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,  $\sigma_i^2$ , 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 –  $\beta_{ij}$  – 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  $\beta_{ij}$  or  $\Lambda$ ), 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  $\delta_{ij}$  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  $\delta_H$  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,  $\tau_j$ , 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  $\delta_H$  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

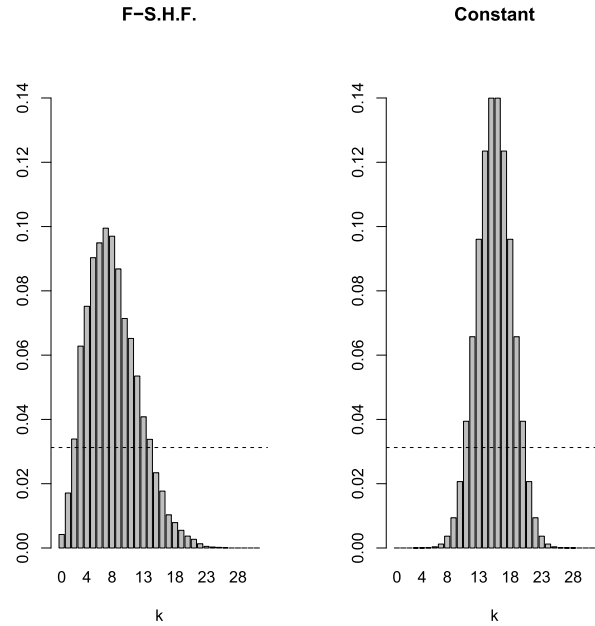


Figure 1: For  $m = 63$  and  $H = 31$ , the prior on  $k$  induced by the prior on  $\delta_H$  without constraints. On the left, the authors' proposal; on the right, the constant prior. The dashed line shows a prior that adjusts for multiplicity.

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 Ministry 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](mailto:niko.hauzenberger@strath.ac.uk)

†Department of Economics, University of Strathclyde, Glasgow, United Kingdom, [gary.koop@strath.ac.uk](mailto: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 hereafter, 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  $|\mathbf{\Lambda}_1| \rightarrow 0$  (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](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](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](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

## Contributed Discussion

Alejandra Avalos-Pacheco<sup>\*,†</sup>, Roberta De Vito<sup>‡</sup> and Gregor Zens<sup>§</sup>

**Introduction** We congratulate the authors on their significant contributions to Bayesian factor analysis, not only in the present paper, but also in a series of prior works including Conti et al. (2014), Kastner et al. (2017), Frühwirth-Schnatter et al. (2023), and Frühwirth-Schnatter (2023). In this discussion, we highlight several issues that arise when extending the proposed framework to a more general *multi-study setting*, where study-specific factors are considered in addition to ‘global’ factors and loadings.

Multi-study factor models have wide applications, especially in biostatistics and medical research, where integrating data from multiple studies is a common challenge (De Vito et al., 2019, 2021). Similar approaches are also used in multi-population demography (Li and Lee, 2005). Regardless of the field, the appropriate number of global and study-specific factors is often unknown, and model selection is typically based on ad hoc criteria. The proposed framework, therefore, has considerable potential to inform discussions in these research areas. However, when considering multi-study extensions, two key questions arise: (1) how do the authors’ identification strategies adapt to a multi-study setting? and (2) how does the computational implementation extend and scale within multi-study frameworks?

**Multi-Study Settings** Recently, several factor model extensions for multi-study settings have been developed, including multi-study factor models (De Vito et al., 2019, 2021), perturbed factor models (Roy et al., 2021), joint factor regression and batch effect correction models (Avalos-Pacheco et al., 2022), subspace factor models (Chandra et al., 2023), combinatorial factor models (Grabski et al., 2023) and multi-study factor regression models (De Vito and Avalos-Pacheco, 2023). Our focus here is on a basic multi-study factor model due to its flexibility, simplicity, and similarity to the authors’ framework. Following the authors’ notation, a sample  $\mathbf{y}_s = (\mathbf{y}_{1s}, \dots, \mathbf{y}_{Ts})'$  from study  $s$  ( $s = 1, \dots, S$ ) of  $T$  ( $t = 1, \dots, T$ ), multivariate observations  $\mathbf{y}_{ts}$  of dimension  $m$  is modeled as

$$\begin{aligned} \mathbf{y}_{ts} &= \boldsymbol{\beta}_H \mathbf{f}_{ts}^H + \boldsymbol{\Phi}_{H_s} \mathbf{l}_{ts}^{H_s} + \boldsymbol{\epsilon}_{ts}, \\ \boldsymbol{\epsilon}_{ts} &\sim N_m(0, \boldsymbol{\Sigma}_s) \quad \mathbf{f}_{ts}^H \sim N_H(0, \mathbf{I}_H) \quad \mathbf{l}_{ts}^{H_s} \sim N_{H_s}(0, \mathbf{I}_{H_s}), \end{aligned} \tag{1}$$

with a finite number of common and study-specific factors,  $H < \infty$  and  $H_s < \infty$ . Marginally, each study has a covariance matrix  $\text{Cov}(\mathbf{y}_s) = \boldsymbol{\Omega}_s = \boldsymbol{\beta}_k \boldsymbol{\beta}_k^\top + \boldsymbol{\Phi}_{j_s} \boldsymbol{\Phi}_{j_s}^\top + \boldsymbol{\Sigma}_s$ , where  $\boldsymbol{\beta}_k$  and  $\boldsymbol{\Phi}_{j_s}$  are formed by the non-zero columns of  $\boldsymbol{\beta}_H$  and  $\boldsymbol{\Phi}_{H_s}$  respectively, and

---

<sup>\*</sup>Institute of Applied Statistics, Johannes Kepler University Linz (JKU Linz), Linz, Austria, [alejandra.avalos\\_pacheco@jku.at](mailto:alejandra.avalos_pacheco@jku.at)

<sup>†</sup>Harvard-MIT Center for Regulatory Science, Harvard Medical School, Boston, MA, USA

<sup>‡</sup>Department of Biostatistics and Data Science Initiative, Brown University, Providence, USA, [roberta\\_devito@brown.edu](mailto:roberta_devito@brown.edu)

<sup>§</sup>International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria, [zens@iiasa.ac.at](mailto:zens@iiasa.ac.at)

$\Sigma_s$  is a diagonal matrix with strictly positive diagonal elements. The ‘global’ covariance component  $\beta_k \beta_k^\top$  encodes reproducible biological pathways that traditional factor analysis approaches may miss due to batch effects (De Vito et al., 2019, 2021; Hansen et al., 2024).

**Identification Issues** Factor analysis typically encounters several identifiability issues, including a potential indeterminacy of the covariance decomposition, rotational invariance, and the more trivial issues of invariance with respect to column and sign permutations of the factors and the loading matrices. For standard factor models, these identifiability issues have been elegantly resolved by the authors using unordered generalized lower triangular (UGLT) loading structures and the *3579 counting rule*, in combination with post-processing steps to reorder the posterior draws.

Extending these solutions to the multi-study setting is likely to present additional challenges. For example, obtaining a unique solution to the decomposition  $\Omega_s = \beta_k \beta_k^\top + \Phi_{j_s} \Phi_{j_s}^\top + \Sigma_s$  is more difficult due to the additive covariance decomposition into more than two terms (De Vito et al., 2019, 2021). While rotational invariance is expected to be resolved by independent UGLT structures for the global and study-specific loadings matrices, it is likely that an extension of the *3579 counting rule* or additional constraints and postprocessing steps will be needed for a unique variance decomposition. Hence, further investigation into extensions to the identification strategy proposed by the authors is needed to ensure a unique solution in multi-study factor models.

**Computational Considerations** In multi-study contexts, the number of loadings increases rapidly with each additional study, making model space exploration via MCMC more challenging and raising questions about computational scalability. A key issue is whether a conditional posterior sampling strategy would be feasible, in which ‘global’ quantities are updated conditionally on study-specific quantities, and so on. While this is arguably the simplest and most direct extension of the proposed MCMC scheme, it could potentially severely reduce sampling efficiency. This concern is particularly relevant given that the current MCMC scheme already involves updating global factors one at a time, conditional on all the others. An alternative approach might involve marginalized sampling strategies, where sets of factors and loadings are updated in model representations where all other factors are integrated out instead of conditioned on. While this could significantly improve MCMC efficiency, it likely also introduces significant computational overhead due to the need to sample from potentially high-dimensional posterior densities.

Three potential extensions of the proposed computational toolkit that could improve scalability include the use of 1) continuous shrinkage priors and post-processing methods to achieve sparsity (Hahn and Carvalho, 2015); 2) more efficient algorithms for model space exploration (Zanella and Roberts, 2019; Griffin et al., 2021) and 3) approximate Bayesian methods for factor models (Hansen et al., 2024). Of course, pursuing any of these avenues raises several new questions, particularly regarding theoretical guarantees and how to effectively combine them with the identification strategies proposed by the authors.



## References

- Avalos-Pacheco, A., Rossell, D., and Savage, R. S. (2022). “Heterogeneous large datasets integration using Bayesian factor regression.” *Bayesian Analysis*, 17(1): 33–66. MR4377136. doi: <https://doi.org/10.1214/20-ba1240>. 45
- Chandra, N. K., Dunson, D. B., and Xu, J. (2023). “Inferring covariance structure from multiple data sources via subspace factor analysis.” *arXiv preprint arXiv:2305.04113*. 45
- Conti, G., Frühwirth-Schnatter, S., Heckman, J. J., and Piatek, R. (2014). “Bayesian exploratory factor analysis.” *Journal of Econometrics*, 183(1): 31–57. MR3269916. doi: <https://doi.org/10.1016/j.jeconom.2014.06.008>. 45
- De Vito, R. and Avalos-Pacheco, A. (2023). “Multi-study factor regression model: an application in nutritional epidemiology.” *arXiv preprint arXiv:2304.13077*. MR3337656. 45
- De Vito, R., Bellio, R., Trippa, L., and Parmigiani, G. (2019). “Multi-study factor analysis.” *Biometrics*, 75(1): 337–346. MR3953734. doi: <https://doi.org/10.1111/biom.12974>. 45, 46
- De Vito, R., Bellio, R., Trippa, L., and Parmigiani, G. (2021). “Bayesian multistudy factor analysis for high-throughput biological data.” *The Annals of Applied Statistics*, 15(4): 1723–1741. MR4355073. doi: <https://doi.org/10.1214/21-aos1456>. 45, 46
- Frühwirth-Schnatter, S. (2023). “Generalized cumulative shrinkage process priors with applications to sparse Bayesian factor analysis.” *Philosophical Transactions of the Royal Society A*, 381(2247): 20220148. MR4590506. 45
- 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): 26. 45
- Grabski, I. N., Vito, R. D., Trippa, L., and Parmigiani, G. (2023). “Bayesian combinatorial multi-study factor analysis.” *The Annals of Applied Statistics* (in press). MR4637664. doi: <https://doi.org/10.1214/22-aos1715>. 45
- Griffin, J. E., Łatuszyński, K., and Steel, M. F. (2021). “In search of lost mixing time: adaptive Markov chain Monte Carlo schemes for Bayesian variable selection with very large  $p$ .” *Biometrika*, 108(1): 53–69. MR4226189. doi: <https://doi.org/10.1093/biomet/asaa055>. 46
- Hahn, P. R. and Carvalho, C. M. (2015). “Decoupling shrinkage and selection in Bayesian linear models: a posterior summary perspective.” *Journal of the American Statistical Association*, 110(509): 435–448. MR3338514. doi: <https://doi.org/10.1080/01621459.2014.993077>. 46
- Hansen, B., Avalos-Pacheco, A., Russo, M., and De Vito, R. (2024). “Fast variational inference for Bayesian factor analysis in single and multi-study settings.” *Journal*

- of Computational and Graphical Statistics*, 0(0): 1–13. [10.1080/10618600.2024.2356173](https://doi.org/10.1080/10618600.2024.2356173) 46
- 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(4): 905–917. [MR3765354](https://doi.org/10.1080/10618600.2017.1322091). doi: <https://doi.org/10.1080/10618600.2017.1322091>. 45
- Li, N. and Lee, R. (2005). “Coherent mortality forecasts for a group of populations: An extension of the Lee-Carter method.” *Demography*, 42: 575–594. 45
- Roy, A., Lavine, I., Herring, A. H., and Dunson, D. B. (2021). “Perturbed factor analysis: Accounting for group differences in exposure profiles.” *The Annals of Applied Statistics*, 15(3): 1386. [MR4316654](https://doi.org/10.1214/20-aas1435). doi: <https://doi.org/10.1214/20-aas1435>. 45
- Zanella, G. and Roberts, G. (2019). “Scalable importance tempering and Bayesian variable selection.” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(3): 489–517. [MR3961496](https://doi.org/10.1111/rssb.12496). 46



# Contributed Discussion

Francesco Denti\* and Stefano Rizzelli†

## 1 Introduction

We congratulate the authors for their valuable contributions to Bayesian factor analysis. Such a well-thought-out paper highlights that, while linear factor models may seem fairly simple, they require careful evaluation of numerous aspects to ensure model interpretability, effective signal recovery, and feasible posterior simulation (see also Bolfarine et al., 2024). We hereafter focus on methodological and computational aspects of signal-noise separation and sparsity induction, drawing connections between factor analysis and mixture modeling. Notably, techniques for ensuring parsimony in the Bayesian analysis of high-dimensional models have been the subject of a vast literature over the past two decades (see, e.g., Bhattacharya et al., 2015; Bhadra et al., 2019; Chandra et al., 2023). In this work, sparsity on the matrix of factor loadings is imposed by spike-and-slab priors, mixing a point mass at 0 and a diffuse distribution.

## 2 Handling Sparsity in Factor Analysis: Much to Learn from Mixtures?

Of particular interest is the incorporation of shrinkage through the slab distribution, devised as a continuous scale mixture of normal distributions (Bai et al., 2022). Specifically, the authors suggest employing *global*, *column-specific*, and *local* scale hyperparameters, assigning them inverse-gamma or F distributions. This choice reminisces continuous shrinkage priors used in Boss et al. (2024) for regression with grouped covariates, including global, group- and predictor-specific scale hyperparameters endowed with Cauchy, gamma, and inverse-gamma hyperpriors. Such an approach trades flexibility in the design of shrinkage profiles with a large number of hyperparameters to update in posterior computations.

A convenient, more parsimonious alternative could be to use a sparse discrete mixture of continuous mixtures as in Denti et al. (2023a), postulating that, a priori, the factor loadings have the following conditional distribution (cfr. Equation (3.12)):

$$(\beta_{ij} \mid \delta_{ij} = 1, \boldsymbol{\pi}_j, \{\omega_{lj}^*\}_{l=1}^L, \kappa, \theta_j, \sigma_i^2) \sim \sum_{l=1}^L \pi_{lj} \mathcal{N}(0, \kappa \cdot \theta_j \cdot \sigma_i^2 \cdot \omega_{lj}^*) \quad (1)$$

where  $\boldsymbol{\pi}_j$  is a vector of column-specific mixture weights modeled with a prior on a  $L$ -dimensional simplex. Moreover,  $\{\omega_{lj}^*\}_{l=1}^L \sim \nu$  denote the mixture component shrinkage parameters, for  $j = 1, \dots, r$ . The base measure  $\nu$  can be selected according to the

---

\*Department of Statistics, University of Padova, [francesco.denti@unipd.it](mailto:francesco.denti@unipd.it)

†Department of Statistics, University of Padova, [stefano.rizzelli@unipd.it](mailto:stefano.rizzelli@unipd.it)

problem at hand and the desired degree of sparsity; e.g.,  $\nu$  can be a half-Cauchy distribution to obtain a finite mixture of horseshoe priors (Carvalho et al., 2010). The model is augmented using column-specific categorical membership labels, denoted with  $z_{ij} \in \{1, \dots, L\}$ , obtaining:

$$(\beta_{ij} \mid \delta_{ij} = 1, z_{ij} = l, \omega_{lj}^*, \kappa, \theta_j, \sigma_i^2) \sim \mathcal{N}(0, \kappa \cdot \theta_j \cdot \sigma_i^2 \cdot \omega_{lj}^*), \quad (2)$$

such that  $\mathbb{P}[z_{ij} = l \mid \boldsymbol{\pi}_j] = \pi_{lj}$ . Adopting this prior structure in a factor model would (i) allow the sharing of information in posterior computation of shrinkage factors and (ii) segment the factor loadings in column-specific tiers of magnitude, which would help the interpretability of the model. For example, if in a generic column  $j$  a specific  $\omega_{lj}^*$  has posterior mode close to zero, then all the coordinates of  $\mathbf{y}_t$  assigned to the  $l$ -th cluster can be regarded as unaffected by the  $j$ -th factor. To achieve an even more parsimonious parameterization in terms of scale parameters, a common atoms structure could be introduced across the columns of  $\boldsymbol{\beta}$  following Denti et al. (2023b); D’Angelo and Denti (2024), setting  $\omega_{lj}^* = \omega_l^*$ , for each  $l = 1, \dots, L$ ,  $j = 1, \dots, r$ , and  $\{\omega_l^*\}_{l=1}^L \sim \nu$ .

In this work, factor loadings are further pulled towards zero by assigning to slab probability weights a beta distribution with shape hyperparameters  $\gamma\alpha/H$  and  $\gamma$ . When adopting a hierarchical Bayesian approach, the choice of the hyperprior of  $\alpha$  is delicate, given the important role this hyperparameter plays in controlling the degree of sparsity. In particular, conditionally on  $\gamma$  and  $\alpha$ , the number  $r$  of nonzero columns in the sparsity matrix  $\boldsymbol{\delta}_H$  (see Section 3.1 of the paper) or, in other words, the number of active factors, is a priori a binomial random variable of parameters  $H$  and  $\alpha/(\alpha + H)$ , respectively. A way to bypass hyperprior specification might be to fix hyperparameter  $\alpha$  in an empirical Bayes fashion, as done in the context of mixture models with Dirichlet prior (see, e.g., Liu, 1996; Rizzelli et al., 2024). This could be done by maximising with respect to  $\alpha$  the marginal likelihood arising from the unconstrained factor model and the priors in Equations (2.1), (3.1)–(3.3), and (3.14) of the paper. Assuming for simplicity that  $\gamma = 1$  and a generic slab density  $p_{\text{slab}}$ , such marginal likelihood boils down to

$$m(\mathbf{y}|\alpha) = \sum_{r=0}^H \binom{H}{r} \left(\frac{\alpha}{H + \alpha}\right)^r \left(\frac{H}{H + \alpha}\right)^{H-r} p(\mathbf{y}|r)$$

where, denoting  $\boldsymbol{\xi}_r = (\mathbf{f}, \beta_{1,1}, \dots, \beta_{r,m})$  and by  $\varphi$  and  $t_\nu$  the standard Gaussian and  $\nu$ -degrees of freedom Student-t probability densities, the term  $p(\mathbf{y}|r)$  is proportional to

$$\int \prod_{t=1}^T \left[ \prod_{l=1}^m t_{2c^\sigma} \left( \sqrt{\frac{c^\sigma}{C_{i0}}} \left( y_{tj} - \sum_{j=1}^r \beta_{ij} f_{tj} \right) \right) \prod_{j=1}^r \mathcal{N}(f_{tj}; 0, 1) \right] \prod_{i=1}^m \prod_{j=1}^r p_{\text{slab}}(\beta_{ij}) d\boldsymbol{\xi}_r.$$

It is easy to see that the first derivative  $(\partial/\partial\alpha)m(\mathbf{y}|\alpha)$  equals zero whenever the hyperparameter  $\alpha$  is such that the posterior mean of the number of active factors  $r$  (under the unconstrained model) equals its prior expectation  $H\alpha/(H + \alpha)$ . It would thus be interesting to investigate whether one could mutuate techniques in McAuliffe et al. (2006), Section 3, for calculating the maximiser. Data-dependent selections of  $\alpha$  of the form  $\hat{\alpha} = H\hat{r}/(H - \hat{r})$  may also be considered, where  $\hat{r}$  is a frequentist estimator of the number of active factors (see, e.g., Bai and Ng, 2002; Bai, 2003; Cragg and Donald, 1997; Barigozzi and Cho, 2020), as they yield an empirical Bayes prior on  $r$  with mean  $\hat{r}$  (provided that  $\hat{r} < H$ ).

## References

- Bai, J. (2003). “Inferential theory for factor models of large dimensions.” *Econometrica*, 71(1): 135–171. MR1956857. doi: <https://doi.org/10.1111/1468-0262.00392>. 50
- Bai, J. and Ng, S. (2002). “Determining the number of factors in approximate factor models.” *Econometrica*, 70(1): 191–221. MR1926259. doi: <https://doi.org/10.1111/1468-0262.00273>. 50
- Bai, R., Moran, G. E., Antonelli, J. L., Chen, Y., and Boland, M. R. (2022). “Spike-and-slab group lassos for grouped regression and sparse generalized additive models.” *Journal of the American Statistical Association*, 117(537): 184–197. MR4399078. doi: <https://doi.org/10.1080/01621459.2020.1765784>. 49
- Barigozzi, M. and Cho, H. (2020). “Consistent estimation of high-dimensional factor models when the factor number is over-estimated.” *Electronic Journal of Statistics*, 14(2): 2892–2921. MR4134347. doi: <https://doi.org/10.1214/20-EJS1741>. 50
- Bhadra, A., Datta, J., Polson, N. G., and Willard, B. (2019). “Lasso meets horseshoe: a survey.” *Statistical Science*, 34(3): 405–427. MR4017521. doi: <https://doi.org/10.1214/19-ST700>. 49
- Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2015). “Dirichlet–Laplace priors for optimal shrinkage.” *Journal of the American Statistical Association*, 110(512): 1479–1490. MR3449048. doi: <https://doi.org/10.1080/01621459.2014.960967>. 49
- 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>. 49
- Boss, J., Datta, J., Wang, X., Park, S. K., Kang, J., and Mukherjee, B. (2024). “Group inverse-gamma gamma shrinkage for sparse linear models with block-correlated regressors.” *Bayesian Analysis*, 19(3): 785–814. MR4770323. doi: <https://doi.org/10.1214/23-ba1371>. 49
- Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). “The horseshoe estimator for sparse signals.” *Biometrika*, 97: 465–480. MR2650751. doi: <https://doi.org/10.1093/biomet/asq017>. 50
- Chandra, N. K., Canale, A., and Dunson, D. B. (2023). “Escaping the curse of dimensionality in Bayesian model-based clustering.” *Journal of Machine Learning Research*, 24(144): 1–42. MR4596091. 49
- Cragg, J. G. and Donald, S. G. (1997). “Inferring the rank of a matrix.” *Journal of Econometrics*, 76(1): 223–250. MR1435888. doi: [https://doi.org/10.1016/0304-4076\(95\)01790-9](https://doi.org/10.1016/0304-4076(95)01790-9). 50
- D’Angelo, L. and Denti, F. (2024). “A finite-infinite shared atoms nested model for the Bayesian analysis of large grouped data.” *Bayesian Analysis* (to appear). doi: <https://doi.org/10.1214/24-BA1458>. 50

- Denti, F., Azevedo, R., Lo, C., Wheeler, D. G., Gandhi, S. P., Guindani, M., and Shahbaba, B. (2023a). “A horseshoe mixture model for Bayesian screening with an application to light sheet fluorescence microscopy in brain imaging.” *Annals of Applied Statistics*, 17: 2639–2658. MR4637684. doi: <https://doi.org/10.1214/23-aos1736>. 49
- Denti, F., Camerlenghi, F., Guindani, M., and Mira, A. (2023b). “A common atoms model for the Bayesian nonparametric analysis of nested data.” *Journal of the American Statistical Association*, 118(541): 405–416. MR4571130. doi: <https://doi.org/10.1080/01621459.2021.1933499>. 50
- Liu, J. S. (1996). “Nonparametric hierarchical Bayes via sequential imputations.” *The Annals of Statistics*, 24(3): 911–930. MR1401830. doi: <https://doi.org/10.1214/aos/1032526949>. 50
- McAuliffe, J., Blei, D., and Jordan, M. (2006). “Nonparametric empirical Bayes for the Dirichlet process mixture model.” *Statistics and Computing*, 16: 5–14. MR2224185. doi: <https://doi.org/10.1007/s11222-006-5196-2>. 50
- Rizzelli, S., Rousseau, J., and Petrone, S. (2024). “Empirical Bayes in Bayesian learning: understanding a common practice.” [arXiv:2402.19036](https://arxiv.org/abs/2402.19036). 50

## Contributed Discussion

Luke Vrotsos\* and Mike West†

We applaud the authors on innovative contributions to Bayesian factor modelling and computation. Their creative developments are of broad importance in model uncertainty analysis, particularly in the use of encompassing models within which data-supported submodels can be explored. New sparsity priors on factor loadings neatly extend prior approaches (going back at least to West, 2003), and their MCMC algorithms will enhance access by users. Our comments address three themes: goals, factor model assumptions, and connections between factor models and graphical models.

**A. Goals** When do we care about the number of factors as anything but a nuisance parameter? In what applications is the number of factors of primary interest with respect to inferential or decision goals? In financial examples, prediction and decisions are what matter. The traditional parameter-focussed approach scores models on purely statistical metrics, with no real regard for broader forecast or decision goals. In contrast, fully subjective Bayesian predictive decision synthesis (e.g. Tallman and West, 2023, 2024) stresses articulation of goals and model scoring modulo goals; its application to the setting of uncertain numbers of factors would seem to represent practical opportunities.

**B. Dependent Factor Models** Much of the literature continues to insist on models with uncorrelated factors. However, sparsity of factor loadings provides opportunity to exploit dependent factor models. This was recognised by Zhou et al. (2014) in time series, emphasising forecasting superiority of models with (often high and time-varying) dependencies among factors, and by Conti et al. (2014) in traditional random sampling models as noted in the current paper. It would be interesting to hear more from the authors on this, and on potential extensions of their machinery for dependent factors.

**C. Graphical Models and Factor Models** A sparse factor model defines a 0/1 sparsity pattern in the covariance matrix. In contrast, a sparse (normal) graphical model has a conditional independence graph underlying all dependencies, corresponding to a 0/1 sparsity pattern in the precision matrix (e.g. Jones et al., 2005; Jones and West, 2005; Cron and West, 2016). These are different approaches to parsimonious modelling while – in a very real sense – graphical models are more general. They “compete” as inverting a sparse covariance matrix typically yields a full precision matrix, and vice-versa. One existing approach to reconciliation (the only one to our knowledge) is that of Yoshida and West (2010); the authors’ methodology could apply to that framework.

However, our main interest here is in deeper theoretical connections between sparse graphical and factor models. New theory in Vrotsos and West (2024) reconciles some

---

\*PhD student, Department of Statistical Science, Duke University, [luke.vrotsos@duke.edu](mailto:luke.vrotsos@duke.edu)

†The Arts and Sciences Distinguished Professor Emeritus of Statistics and Decision Sciences, Duke University, [mike.west@duke.edu](mailto:mike.west@duke.edu)

classes of sparse factor models with inherently sparse graphical models, advancing understanding of how the number of factors and sparse factor structures relate to explicit assumptions about multivariate conditional independencies. Importantly, the sparse factor models are not primary, but are implied by specification of the more fundamental graphical structures. Further, these implied sparse factor models can have strong factor dependencies as well as dependence between factors and residuals.

These new theoretical connections arose from our interests in simultaneous graphical dynamic linear models (SGDLMs: Gruber and West, 2016, 2017), a rich, flexible and scalable class of multivariate time series models. In the simpler framework and notation of the current authors, the essential structure for a single  $m$ -dimensional vector observation  $\mathbf{y}$  involves the *simultaneous equation* specification  $\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\Gamma}\mathbf{y} + \boldsymbol{\nu}$  where:  $\boldsymbol{\mu}$  may involve exogenous predictors;  $\boldsymbol{\nu}$  is zero mean normal with diagonal precision matrix  $\boldsymbol{\Delta}$ ; and the  $m \times m$  *simultaneous coefficient* matrix  $\boldsymbol{\Gamma}$  is sparse with zero diagonal entries. Given a pattern of zeros in  $\boldsymbol{\Gamma}$ , each variable  $y_i$  is regressed on only those other  $y_j$  for which the  $(i,j)$  element of  $\boldsymbol{\Gamma}$  is non-zero. This is reflected in the implied directed (but usually *not* acyclic) graph where this set of the  $y_j$  are *simultaneous parents* of  $y_i$ ; correspondingly, this  $y_i$  is a *child variable* of each of its parents. The implied normal distribution of  $\mathbf{y}$  has mean vector  $(\mathbf{I} - \boldsymbol{\Gamma})^{-1}\boldsymbol{\mu}$  and precision matrix  $(\mathbf{I} - \boldsymbol{\Gamma})'\boldsymbol{\Delta}(\mathbf{I} - \boldsymbol{\Gamma})$ . With high levels of sparsity in  $\boldsymbol{\Gamma}$ , this precision matrix will be sparse with off-diagonal zeros representing the unique, underlying conditional independence graph.

The new theory in Vrotsos and West (2024) uses the sparse singular value decomposition  $\boldsymbol{\Gamma} = \boldsymbol{\Lambda}\mathbf{D}\mathbf{S}$  to give the equivalent model form  $\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{f} + \boldsymbol{\nu}$  with  $r$ -dimensional factor vector  $\mathbf{f} = \mathbf{D}\mathbf{S}\mathbf{y}$ , the  $r \times r$  diagonal matrix  $\mathbf{D}$  of positive singular values, and the  $m \times r$  loadings matrix  $\boldsymbol{\Lambda}$ . The factor vector  $\mathbf{f}$  is a direct linear combination of elements of  $\mathbf{y}$ ; in the time series setting, there are evident connections with so-called “linear scalar components” (e.g. Tiao and Tsay, 1989). Typically, both  $\boldsymbol{\Lambda}$  and  $\mathbf{S}$  are sparse, with sparsity patterns inherited from the sparsity pattern of  $\boldsymbol{\Gamma}$ . Hence each factor depends on a subset of the  $y_i$  (through  $\mathbf{S}$ ) while each variable  $y_i$  loads on only a subset of the factors (through  $\boldsymbol{\Lambda}$ ). The new results on sparsity and structure of this factor representation exploit linear algebra and graph theory to show the following.

The number of factors  $r \leq m$  cannot exceed – and typically equals –  $\text{rank}(\boldsymbol{\Gamma})$ . A very sparse  $\boldsymbol{\Gamma}$  will have low rank and small implied factor dimension. The theory is sharper in typical cases when  $\text{rank}(\boldsymbol{\Gamma})$  is the number of columns with at least one non-zero entry. A zero column  $j$  implies that  $y_j$  is not a parental predictor at all, so then the number of factors is the number of the  $y_j$  that are parents of at least one other variable. This is a property of the graph and sparsity pattern of  $\boldsymbol{\Gamma}$ , not at all dependent on the values of non-zero elements. Coupled with this are results about the precise patterns of non-zero entries in  $\boldsymbol{\Lambda}$  and  $\mathbf{S}$ , underpinning the definitions of factors. Non-zero elements in rows of  $\mathbf{S}$  – defining the factors – relate to subsets of variables with intersecting child sets; non-zero elements in columns of  $\boldsymbol{\Lambda}$  – defining the variables loaded on factors – are defined by subsets of variables with intersecting parental sets. A macroeconomic application in Vrotsos and West (2024) explores these concepts and the theory fully detailed there, with what we regard as illuminating examples with detailed contextual interpretation of inferred factor processes. There are technical details of factor identification within

subsets of factors defined on common parental sets, directly resolved by a version of the authors' "pivot" series that follows the earlier use of "founder" variables (Carvalho et al., 2008). These connections seem to offer opportunities to extend Bayesian sparse factor analysis in new ways, linking intimately to the main concern of the current authors in uncertainty about the number of factors, in particular.

## References

- Carvalho, C. M., Lucas, J. E., Wang, Q., Chang, J., Nevins, J. R., and West, M. (2008). "High-dimensional sparse factor modelling – applications in gene expression genomics." *Journal of the American Statistical Association*, 103: 1438–1456. MR2655722. doi: <https://doi.org/10.1198/016214508000000869>. 3
- 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>. 1
- Cron, A. J. and West, M. (2016). "Models of random sparse eigenmatrices matrices and Bayesian analysis of multivariate structure." In Frigessi, A., Bühlmann, P., Glad, I., Langaas, M., Richardson, S., and Vannucci, M. (eds.), *Statistical Analysis for High Dimensional Data*, 123–154. Springer. MR3616267. 1
- Gruber, L. F. and West, M. (2016). "GPU-accelerated Bayesian learning and forecasting in simultaneous graphical dynamic linear models." *Bayesian Analysis*, 11: 125–149. MR3447094. doi: <https://doi.org/10.1214/15-BA946>. 2
- Gruber, L. F. and West, M. (2017). "Bayesian forecasting and scalable multivariate volatility analysis using simultaneous graphical dynamic linear models." *Econometrics and Statistics*, 3: 3–22. MR3666239. doi: <https://doi.org/10.1016/j.ecosta.2017.03.003>. 2
- Jones, B., Dobra, A., Carvalho, C. M., Hans, C., Carter, C., and West, M. (2005). "Experiments in stochastic computation for high-dimensional graphical models." *Statistical Science*, 20: 388–400. MR2210226. doi: <https://doi.org/10.1214/088342305000000304>. 1
- Jones, B. and West, M. (2005). "Covariance decomposition in undirected Gaussian graphical models." *Biometrika*, 92: 779–786. MR2234185. doi: <https://doi.org/10.1093/biomet/92.4.779>. 1
- Tallman, E. and West, M. (2023). "Bayesian predictive decision synthesis." *Journal of the Royal Statistical Society (Ser. B)*, 86: 340–363. MR4754087. doi: <https://doi.org/10.1093/jrsss/qkad109>. 1
- Tallman, E. and West, M. (2024). "Predictive decision synthesis for portfolios: Betting on better models." In Mazur, S. and Österhol, P. (eds.), *Recent Developments in Bayesian Econometrics and Their Applications*. Springer. arXiv:2405.01598. 1
- Tiao, G. C. and Tsay, R. S. (1989). "Model specification in multivariate time se-

- ries.” *Journal of the Royal Statistical Society: Series B (Methodological)*, 51: 157–195. [MR1007452](#). 2
- Vrotsos, L. and West, M. (2024). “Dynamic graphical models: Theory, structure and counterfactual forecasting.” *Submitted for publication*. [arXiv:2410.06125](#). 1, 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., David, A. P., Heckerman, D., Smith, A. F. M., and West, M. (eds.), *Bayesian Statistics 7*, 723–732. Oxford. [MR2003537](#). 1
- Yoshida, R. and West, M. (2010). “Bayesian learning in sparse graphical factor models via annealed entropy.” *Journal of Machine Learning Research*, 11: 1771–1798. [MR2653356](#). 1
- Zhou, X., Nakajima, J., and West, M. (2014). “Bayesian forecasting and portfolio decisions using dynamic dependent sparse factor models.” *International Journal of Forecasting*, 30: 963–980. 1



## Contributed Discussion

Ihnwhi Heo<sup>\*</sup>, Julius M. Pfadt<sup>†</sup>, and Eric-Jan Wagenmakers<sup>†</sup>

Identifying the number of factors has remained a fundamental challenge in factor analysis research. We commend and congratulate Frühwirth-Schnatter et al. (2024) for their inspiring work in addressing this methodological issue. Their approach to achieving sparsity in overfitting exploratory factor analysis models is supported by two computational strategies. First, the authors introduce an unordered generalized lower triangular (UGLT) representation of the factor loading matrix to address rotational invariance. Second, they develop a customized reversible jump Markov chain Monte Carlo (MCMC) algorithm for efficient posterior inference of the number of factors and other model parameters. Their contributions are substantial and have been well summarized in invited discussions. The thorough theoretical exposition, supplemented by empirical examples, positions this work as a stepping stone for further exploration.

In this contributed discussion, we point out avenues for expanding upon the ideas of Frühwirth-Schnatter et al. (2024). The methods proposed by the authors implicitly assume the homogeneity of the population of interest. Put differently, the estimated factor structure is assumed to apply uniformly across the entire population. However, this assumption often needs to be relaxed due to the inherent heterogeneity in the population. One approach to account for population heterogeneity and consider differences between qualitatively distinct subpopulations—termed latent classes—is the mixture modeling framework (Frühwirth-Schnatter, 2010; McLachlan and Peel, 2000). With this consideration in mind, we suggest that an interesting direction is the incorporation of the mixture modeling framework to induce sparsity across different latent classes.

Within the mixture modeling framework, one important consideration is that the number of latent classes is typically predetermined using statistical criteria such as information criteria. Afterward, the factor structure for each latent class is estimated conditional on the estimated group membership. The next step will be to examine whether estimating separate factor structures for each latent class based on the UGLT representation is computationally feasible. Here, the label switching problem must be addressed. Common solutions include adding order constraints to parameters during estimation or post-processing the chains using some permutation techniques to reorder the MCMC output. Developing a scalable MCMC algorithm is essential in this context.

Another direction of research regards the estimation of the number of latent classes without resorting solely to information criteria. Instead of predetermining the dimensionality of factors based on information criteria, the dimension can be solely determined based on the data. A related idea can arise by borrowing the principles of Bayesian non-parametrics (Gershman and Blei, 2012; Ghahramani, 2013). Relatedly, Grushanina and

---

<sup>\*</sup>Department of Psychological Sciences, University of California, Merced, [ihnwhi.heo@gmail.com](mailto:ihnwhi.heo@gmail.com)

<sup>†</sup>Department of Psychology, University of Amsterdam, [julius.pfadt@gmail.com](mailto:julius.pfadt@gmail.com); [ej.wagenmakers@gmail.com](mailto:ej.wagenmakers@gmail.com)

Frühwirth-Schnatter (2023) developed an automatic inference process to assign a non-parametric prior for estimating the dimension of factors. The development of MCMC algorithms that integrate the data-driven estimation of factor dimensionality and induce sparsity warrants further research.




An important parameter in implementing mixture models is the mixing proportion. In the Bayesian framework, Dirichlet distributions are commonly used as prior distributions for this parameter, with hyperparameters representing the prior proportion of individuals in each class. Implementing this prior alongside the spike and slab priors for simultaneously imposing row and column sparsity while establishing factor structures could be an important contribution to future research.

We believe that expanding the ideas of Frühwirth-Schnatter et al. (2024) into the factor mixture modeling framework can reveal important and nuanced details about population heterogeneity, particularly with regard to the factor dimensions within each latent class. Looking ahead, the integration of their approach, along with its mixture extensions, into open-source statistical software such as JASP (JASP Team, 2024) could augment the accessibility and dissemination of these methods. We are confident that the work of Frühwirth-Schnatter et al. (2024) represents a substantial contribution that will inspire further advancements in both methodological and applied research. We once again commend the authors for their outstanding work.

## References

- Frühwirth-Schnatter, S. (2010). *Finite Mixture and Markov Switching Models*. Springer. MR2265601. 57
- Frühwirth-Schnatter, S., Hosszejni, D., and Lopes, H. F. (2024). “Sparse Bayesian factor analysis when the number of factors is unknown.” *Bayesian Analysis*, 1–31. 57, 58
- Gershman, S. J. and Blei, D. M. (2012). “A tutorial on Bayesian nonparametric models.” *Journal of Mathematical Psychology*, 56(1): 1–12. MR2903470. doi: <https://doi.org/10.1016/j.jmp.2011.08.004>. 57
- Ghahramani, Z. (2013). “Bayesian non-parametrics and the probabilistic approach to modelling.” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1984): 20110553. MR3005667. doi: <https://doi.org/10.1098/rsta.2011.0553>. 57
- 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. 57
- JASP Team (2024). “JASP (Version 0.19.0) [Computer software].” <https://jasp-stats.org/>. 58
- McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. John Wiley & Sons. MR1789474. doi: <https://doi.org/10.1002/0471721182>. 57

## Contributed Discussion

Szymon Urbas<sup>\*,§</sup>, Kate Finucane<sup>†</sup>,  
Isobel Claire Gormley<sup>†,‡</sup>, and Keefe Murphy<sup>\*,§</sup>

We congratulate the authors on their innovative contributions to the area of sparse factor analysis. Their work tackles the long-standing challenge of estimating the number of common factors that give rise to high-dimensional data, a challenge that has hampered the more widespread practical usage of factor-analytic models. The work advances the state of the art and provides an elegant, theoretically sound basis for inference on the number of common factors. We feel the proposed model has rich utility — in biological and chemical science applications especially — as its column and row sparsity features lend themselves to interpretable analyses of large datasets. Row sparsity is of particular interest in areas that use high-throughput data, such as spectral data, where methods that identify and remove extraneous variables (e.g., Casa et al., 2022; Urbas et al., 2024) are especially useful due to the typically high numbers of redundant variables encountered in high-dimensional settings.

We have a number of comments and queries related to computational efficiency, other aspects of practical utility, and potential model extensions. Firstly, we commend the authors on their well-formulated reversible-jump algorithm, particularly where ‘split’ proposals are concerned. Its self-stabilising behaviour is evident from the corresponding MH ratio (4.4). By substituting in the authors’ suggestions of  $a_H = \alpha/H$ ,  $b_H = 1$ , and  $H = \lfloor (m-1)/2 \rfloor$ , letting  $m$  be odd (for presentation clarity), and further assuming that  $r_{sp} = \lambda m$ , with  $\lambda \in [1/m, 1/2 - r/m)$ , we arrive at a (roughly) constant approximate proportion of column splits at each iteration:

$$(H - r - r_{sp}) \times \min\{1, A_{\text{split}}(r, r_{sp})\} = \min\left\{\left(\frac{1}{2} - \lambda\right)m - r - \frac{1}{2}, \frac{\alpha(1 - 2\lambda - \frac{2r+1}{m})^2}{2(\lambda - \frac{1-\lambda}{m} - \frac{1}{m^2})}\right\}.$$

This expression does not blow up or reduce to zero as  $m$  increases, suggesting desirable mixing properties even for challenging problems involving many variables. Interestingly, this all depends on the shrinkage prior hyperparameters used and possibly may not hold for other choices. While the authors have endeavoured to increase accessibility of their work to applied practitioners by requiring only five hyperparameters to be specified, it is noted in Section 3.4 that the recommended default setting for  $H$  is computationally inefficient for large  $m$  and advised that a smaller value below this upper limit should be chosen. As we also have concerns about the effect of  $H$  on  $E_q$  (the expected number of non-zero row elements) and whether the suggested default  $E_q = 2$  remains sensible when  $H$  is large, at which point  $E_q$  converges to  $\alpha$ , we would appreciate further guidance.

---

<sup>\*</sup>Department of Mathematics and Statistics, Maynooth University, [szymon.urbas@mu.ie](mailto:szymon.urbas@mu.ie)

<sup>†</sup>School of Mathematics and Statistics, University College Dublin, [kate.finucane@ucdconnect.ie](mailto:kate.finucane@ucdconnect.ie)

<sup>‡</sup>VistaMilk and Insight SFI Research Centres, University College Dublin, [claire.gormley@ucd.ie](mailto:claire.gormley@ucd.ie)

<sup>§</sup>Hamilton Institute, Maynooth University, [keefe.murphy@mu.ie](mailto:keefe.murphy@mu.ie)

Another area of interest is the discarding of post-burn-in samples. Table 3 reports that roughly 6–10% of samples are removed. We are curious to know whether or not this proportion remains constant as  $m$  increases. If so, this bodes well for high-dimensional applications. If not, we worry that chains may require many iterations in order to obtain reliable posterior inference. Additionally, were the method to be extended to the realm of mixture models, there is ambiguity in how the post-processing would be carried out. We are especially interested in how the method would behave in mixture settings since it has been shown that clustering performance tends to improve when the number of factors is allowed to vary across components (Murphy et al., 2020). Would the loading matrices of all mixture components each need to satisfy the 3579 rule for a sample to be accepted, or could the post-processing be conducted separately for each component? If the former case is true, we conjecture that mixtures with many components would require more discarded samples. Specifically, if we let  $p_g$  denote the component-specific proportion of valid posterior samples, then only  $\prod_{g=1}^G p_g$  samples would be retained in a mixture with  $G$  components. Our broader query about discarding samples is also pertinent to potential model extensions for which the per-iteration computational costs would be more expensive, such as the adoption of truncated distributional assumptions, under which rejection sampling or Gibbs steps involving auxiliary variables are common, albeit computationally demanding sampling approaches. These assumptions are required in settings where negative values are not meaningful, such as when modelling non-negative biological quantities (D’Angelo et al., 2021) or imputing positively constrained data (Finucane et al., 2024), where sampling from truncated multivariate distributions can be particularly costly. Overall, we suspect that requiring a large proportion of discarded samples could impact the method’s utility, particularly in large  $m$  settings, and hamper efforts to embed the approach within more complex modelling extensions.

Finally, we would like to stress that the issue of rotational invariance, which is so elegantly addressed via the imposition of UGLT structures on the loadings, is a direct consequence of the assumption that  $f_t \sim \mathcal{N}_r(\mathbf{0}, \mathbf{I}_r)$ . An alternative modelling strategy is to assume heteroscedastic latent factors by replacing  $\mathbf{I}_r$  with  $\Sigma_f$ , a diagonal matrix with non-identical entries. This yields a marginal covariance  $\Omega = \Lambda \Sigma_f \Lambda^\top + \Sigma_0$  which is no longer invariant under arbitrary rotations. It has been shown by Roy et al. (2021) that this simple modification can help to efficiently recover the true loading structure without requiring Procrustean post-processing to rotationally align the samples. It is well-known that  $(\Lambda, \mathbf{f})$  and  $(\Lambda \mathbf{P}, \mathbf{P}^\top \mathbf{f})$  have equivalent likelihood for any orthonormal matrix  $\mathbf{P}$  under a traditional factor model. This is not the case under a heteroscedastic factor model, however, unless  $\mathbf{P} = \mathbf{P}_\pm \mathbf{P}_\rho$  is specifically a signed permutation matrix. As this matches the identifiability conditions given in Section 2.1, we are eager to learn the implications (if any) of imposing UGLT structures in the presence of heteroscedastic factors. Are there additional benefits, such that imposing such structures would still be worthwhile? Would incorporating  $\Sigma_f$  simplify matters, or perhaps offset the need to discard samples which are not variance-identified? Relatedly, although steps to post-process  $\Lambda$  to address rotation, sign, and permutation invariance have been proposed (Papastamoulis and Ntzoufras, 2022), we wish to highlight that there is also a need to post-process  $\mathbf{f}$  if posterior summaries of the factor scores are of interest, as is often the case, for example, in psychology (Cattell, 1978). It is unclear if the proposed strategies for ensuring identifiability of the loadings also ensure validity of the draws of  $\mathbf{f}$ .

### Funding

This work is partially supported by Science Foundation Ireland (SFI) and the Department of Agriculture, Food and Marine, under grant numbers 18/CRT/6049, 12/RC/2289\_P2, and 16/RC/3835.

### References

- Casa, A., O’Callaghan, T. F., and Murphy, T. B. (2022). “Parsimonious Bayesian factor analysis for modelling latent structures in spectroscopy data.” *The Annals of Applied Statistics*, 16(4): 2417–2436. MR4489217. doi: <https://doi.org/10.1214/21-aos1597>. 59
- Cattell, R. B. (1978). *The Scientific Use of Factor Analysis in Behavioral and Life Sciences*. New York: Plenum. 60
- D’Angelo, S., Brennan, L., and Gormley, I. C. (2021). “Inferring food intake from multiple biomarkers using a latent variable model.” *The Annals of Applied Statistics*, 15(4): 2043–2060. MR4355088. doi: <https://doi.org/10.1214/21-aos1478>. 60
- Finucane, K., Brennan, L., and Gormley, I. C. (2024). “Missing data imputation using a truncated infinite factor model with application to metabolomics data.” [arXiv:2410.10633](https://arxiv.org/abs/2410.10633). 60
- Murphy, K., Viroli, C., and Gormley, I. C. (2020). “Infinite mixtures of infinite factor analysers.” *Bayesian Analysis*, 15(3): 937–963. MR4132655. doi: <https://doi.org/10.1214/19-BA1179>. 60
- Papastamoulis, P. and Ntzoufras, I. (2022). “On the identifiability of Bayesian factor analytic models.” *Statistics and Computing*, 32(2): 1–29. MR4394853. doi: <https://doi.org/10.1007/s11222-022-10084-4>. 60
- Roy, A., Lavine, I., Herring, A. H., and Dunson, D. B. (2021). “Perturbed factor analysis: accounting for group differences in exposure profiles.” *The Annals of Applied Statistics*, 15(3): 1386–1404. MR4316654. doi: <https://doi.org/10.1214/20-aos1435>. 60
- Urbas, S., Lovera, P., Daly, R., O’Riordan, A., Berry, D., and Gormley, I. C. (2024). “Predicting milk traits from spectral data using Bayesian probabilistic partial least squares regression.” *The Annals of Applied Statistics*, 18(4): 3486–3506. doi: <https://doi.org/10.1214/24-AOAS1947>. 59

## Contributed Discussion

Alessandro Casa<sup>\*</sup>, Michael Fop<sup>‡</sup>, and Silvia D’Angelo<sup>†</sup>

We would like to congratulate the authors on their valuable contribution to the Bayesian factor analysis framework. The paper combines different ideas, simultaneously addressing identifiability issues and performing estimation and inference on the number of factors  $r$ . Moreover, the proposed approach builds nicely on the sparse Bayesian factor analysis literature, with a clever elicitation of suitable shrinkage priors paired with a tailored MCMC estimation procedure. We wish to raise a few discussion points and hear the authors perspectives on them.

**Loading Matrix Structure** The proposal relies on an identification strategy based on unordered generalized lower triangular structures (UGLT, Frühwirth-Schnatter et al., 2023) which impose fewer restrictions compared to other common strategies. However, since the pivots must lie in different rows, this approach still requires that at least the first  $r - 1$  variables are represented as a combination of fewer than  $r$  factors. In fact, the  $j$ -th variable, for  $j = 1, \dots, r - 1$  is constrained to have at most  $j$  non-zero loadings. Therefore, UGLT structures potentially exacerbates one of the limitations of the PLT by introducing an ordering of the  $m$  original variables, with at least the first  $r - 1$  ones being described in terms of fewer common factors  $\mathbf{f}_t^H$ . Although this can be addressed by permuting the rows of  $\beta_H$ , in applied frameworks where the observed variables follow a natural ordering, this may conflict with the one imposed by the UGLT structure thus preventing row permutations. For example, Casa et al. (2022) consider Bayesian factor analysis to model high-dimensional spectroscopy data where the variables are often redundant. Here, alterations of their order would compromise the chemical interpretation of the analysis. Moreover, highly correlated variables are expected to share similar representations as functions of the latent factors. In their work, Casa et al. (2022) devise a strategy where rows of  $\beta_H$  are clustered and forced to be equal to cluster-specific values to account for redundancies. However, the UGLT structure makes it more difficult for the first  $r - 1$  rows (at least) of  $\beta_H$  to share a similar pattern with the remaining  $m - r + 1$  rows. This might hinder the detection of redundancies in the observed variables, a common and to some extent overlooked phenomenon in high-dimensional scenarios. It would be interesting to gather authors’ thoughts on the possibility of generalizing their procedure to scenarios where the nature of the data influences the structure of the loading matrix.

**Row Sparsity** The proposed approach cleverly adopt shrinkage priors imposing column sparsity, helping to identify the number of active factors. On the other hand, on

---

<sup>\*</sup>Faculty of Economics and Management, Free University of Bozen-Bolzano, Italy, [alessandro.casa@unibz.it](mailto:alessandro.casa@unibz.it)

<sup>†</sup>School of Computer Science and Statistics, Trinity College Dublin, Ireland, [dangelos@tcd.ie](mailto:dangelos@tcd.ie)

<sup>‡</sup>School of Mathematics and Statistics, University College Dublin, Ireland, [michael.fop@ucd.ie](mailto:michael.fop@ucd.ie)

the row level the induced sparsity seems to be less structured. Nonetheless, in high-dimensional settings factor analysis can be helpful also to identify irrelevant variables being uncorrelated with the other ones as the corresponding row of  $\beta_H$  is equal to 0. From a frequentist perspective, Hirose and Konishi (2012) resort to a penalized estimation procedure which enforces entire rows of the loading matrix to be equal to 0. Although this behavior seems to appear naturally in the real data applications in the paper, it may be worthwhile to reflect on the potential for a more structured approach to impose sparsity on the rows. Additionally, it is important to consider if and how this might conflict with the *3579 counting rule*, which necessitates a specific number of non-zero loadings for submatrices consisting of subsets of the columns in  $\beta_H$ .

**Prior on  $\gamma$**  The hierarchical prior structure defines a scheme that allows the generation of sparse  $\delta_k$  matrices respecting the UGLT structure. The slab probabilities  $\tau_j$  have an impact on the sparsity of  $\delta_k$  and are assumed to arise from a 2PB distribution,  $\tau_j \sim \mathcal{B}(\gamma \frac{\alpha}{H}, \gamma)$ . Ghahramani et al. (2007) show that in such context the expected number of active latent features is  $\bar{k} \sim \text{Pois}(\alpha \sum_{j=1}^m \frac{\gamma}{\gamma+j-1})$ , and that  $\bar{k}$  grows as  $\alpha\gamma \log m$  as  $m$  increases. The authors assume the prior  $\gamma \sim \mathcal{G}(\alpha^\gamma, \beta^\gamma)$ , setting  $\alpha^\gamma = \beta^\gamma = 6$ , implying  $\mathbb{E}[\gamma] = 1$ . Interestingly, this choice is independent of the dimension of the data  $m$ . Additionally, the authors specify a prior for  $\alpha$  such that  $\mathbb{E}[\alpha] \approx 2$  for large  $m$  and  $H = \lfloor (m-1)/2 \rfloor$ . Such specification implies that the expected number of latent features  $\bar{k}$  grows a priori  $\approx 2 \log m$ . While this setting proves valid in the applications discussed in the paper, it may induce situations where the number of latent features could be too small and grow at a slow rate in higher dimensional settings, where more latent factors might be needed to adequately represent the data.

Figure 1 illustrates this aspect of the prior specification. The solid orange line represents the  $\bar{k}$  values as a function of  $m \in [10, 500]$  for  $\gamma = 1$ , the expected value of  $\mathcal{G}(6, 6)$ . The value of  $\bar{k}$  grows very little as  $m$  increases; for example, for  $m = 200$ ,  $\bar{k} = 11.76$ , while for  $m = 500$ ,  $\bar{k} = 13.59$ . To allow  $\bar{k}$  to grow at faster rate, an option is the alternative prior specification  $\gamma \sim \mathcal{G}(m^{2c}, m^c)$ , in which the expected value is an explicit function of the dimension  $m$ :  $\mathbb{E}[\gamma] = m^c$ . Under this prior,  $\bar{k}$  will grow  $\approx 2m^c \log m$ , allowing more control on the prior expected number of active latent features. Given  $m$ , for appropriate choices  $0 < c < 1$ ,  $\bar{k} < H$ . The solid light blue and green lines show the  $\bar{k}$  values for  $\gamma = m^c$  with  $c = 0.3$  and  $c = 0.5$ , respectively. The value of  $\bar{k}$  increases at a faster rate in higher dimensional situations; for example, in the case  $c = 0.3$ ,  $\bar{k} = 37.60$  and  $\bar{k} = 57.31$  when  $m = 200$  and  $m = 500$ , respectively. This alternative prior specification can enable greater flexibility in relation to the number of active factors.

It would be valuable to hear the authors' perspectives on these considerations and the alternative prior, particularly regarding the application to higher-dimensional settings.

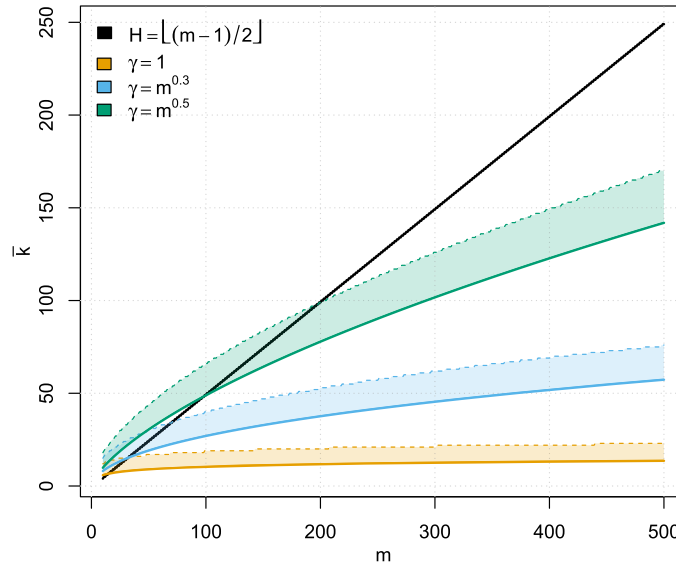


Figure 1: Expected number of active factors  $\bar{k}$  for varying data dimension  $m$  and prior specifications of  $\gamma$  prior. The dashed lines and shaded areas denote the interval between the mean and the 99<sup>th</sup> percentile of the corresponding Poisson distribution,  $\text{Pois}(2 \sum_{j=1}^m \frac{\gamma}{\gamma+j-1})$ .

## References

- Casa, A., O’Callaghan, T. F., and Murphy, T. B. (2022). “Parsimonious Bayesian factor analysis for modelling latent structures in spectroscopy data.” *The Annals of Applied Statistics*, 16(4): 2417–2436. MR4489217. doi: <https://doi.org/10.1214/21-aos1597>. 62
- 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): 26. 62
- Ghahramani, Z., Griffiths, T. L., and Sollich, P. (2007). “Bayesian nonparametric latent feature models.” In *Bayesian Statistics 8: Proceedings of the Eighth Valencia International Meeting*, June 2–6, 2006. Oxford University Press. MR2433194. 63
- Hirose, K. and Konishi, S. (2012). “Variable selection via the weighted group lasso for factor analysis models.” *Canadian Journal of Statistics*, 40(2): 345–361. MR2927750. doi: <https://doi.org/10.1002/cjs.11129>. 63



## Contributed Discussion

Louise Alamichel<sup>\*</sup>, Julyan Arbel<sup>\*</sup>, Daria Bystrova<sup>†</sup>,  
Guillaume Kon Kam King<sup>‡</sup>, and Alessandro Lanteri<sup>§</sup>

In this discussion, we wish to highlight the novel contributions of the paper to the field of Bayesian factor analysis and comment on aspects of asymptotics, and algorithmics. The authors present an efficient framework for modeling dependent multivariate observations, such as time series, using a sparse Bayesian factor model with an unknown number of factors. A carefully chosen prior on factor loadings aids in model identification. The paper contributes by seamlessly coupling model estimation with factor dimension selection and using reversible jump MCMC to explore the factor dimension space. Additionally, it identifies simple structures through row and column sparsity, offering a practical solution for high-dimensional datasets.

**Asymptotics** The approach is at the crossroads between several strands of Bayesian nonparametric research, employing a combination of overfitted mixtures, spike-and-slab variable selection and continuous shrinkage priors (e.g. the horseshoe prior). The asymptotic study of Bayesian nonparametric or related models, where the number of parameters may grow with the amount of data, is interesting as they do not automatically satisfy a Bernstein-von-Mises theorem. Prior choices may have a substantial impact on their asymptotic behaviour, and Bayesian asymptotics may reveal interesting mechanisms and provide intuitions or guidelines for prior elicitation. A typical example is the asymptotic study of spike-and-slab linear regression with  $n$  observations and  $p$  variables ( $p \gg n$ ), which reveals, for instance, that a constant prior inclusion probability for every covariate gives exponentially small mass to sparse models as  $p$  diverges (Scott and Berger, 2010), or that a prior complexity penalty exponential in the model size and Laplace slabs ensure consistency and good frequentist properties (Castillo et al., 2015). One may wish to work by analogy and see if some of these insights might shed light on possible asymptotic properties of the model described here. However, various asymptotic regimes may be considered. The simplest is  $t \rightarrow \infty$  for  $m$  and  $H$  fixed, i.e. infinitely long time series for a finite number of variables: in this case, provided  $\mathbf{\Lambda}$  is identifiable, it should be perfectly identified. The second is both  $t \rightarrow \infty$  and  $m \rightarrow \infty$ , e.g. when the number of time series considered grows. In this case, for  $H$  fixed, intuition from the analysis of overfitted mixture models (Alamichel et al., 2024) suggests that eventually, all columns will be filled because as more time series are considered, more factors will be needed to describe them. This suggests allowing  $H$  to increase with  $m$  and possibly

---

<sup>\*</sup>Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France, [louise.alamichel@inria.fr](mailto:louise.alamichel@inria.fr), [julyan.arbel@inria.fr](mailto:julyan.arbel@inria.fr)

<sup>†</sup>Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000, Grenoble, France, [daria.bystrova@univ-grenoble-alpes.fr](mailto:daria.bystrova@univ-grenoble-alpes.fr)

<sup>‡</sup>Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France, [guillaume.konkamking@inrae.fr](mailto:guillaume.konkamking@inrae.fr)

<sup>§</sup>University of Turin, Department of Economics and Statistics, Torino, Italy, [alessandro.lanteri@unito.it](mailto:alessandro.lanteri@unito.it)

setting  $H = \infty$  (unless novelty can be exhausted and new times series become redundant). This is the typical Bayesian nonparametric setting which allows model complexity to automatically grow with the amount of data. Several authors studied the posterior consistency of the number of factors in sparse Bayesian factor models. Ročková and George (2016) used spike and slab prior with an Indian Buffet Process (IBP,  $H = \infty$ ) on the factor loading matrix  $\mathbf{\Lambda}$  and provided posterior tail bound on the number of factors. Ohn and Kim (2022) build on the spike and slab prior introduced in Ročková and George (2016) and introduced a two-parameter IBP spike-and-slab. In addition, the authors proved the posterior consistency of the number of factors. However, the prior depends on the sparsity level. The recent work Ohn et al. (2024) introduces a new prior that is adaptive to the sparsity level in the data with the desired posterior consistency properties. As they mention, one of the crucial properties of the proposed prior, which is essential for good asymptotic properties, is the dependence between row and column sparsity in  $\mathbf{\Lambda}$ . As  $m \rightarrow \infty$ , the probability of slab inclusion inside a given column is constant (eq. (3.6)). Therefore, as mentioned above, the spike-and-slab prior gives exponentially small mass to sparse columns. The authors encourage sparsity inside columns by considering extra regularisation in the form of continuous shrinkage priors in the slab, as discussed in Section 3.2. It is not entirely obvious how these spike-and-slab and continuous shrinkage sparsity mechanisms interact, and it would be interesting to study whether row and column sparsity have the type of dependence required by Ohn et al. (2024). Furthermore, looking at the simulations in Table 2, we notice that when the sparsity pattern presents heavily overlapped factors, the model size bias  $B_d$  is negative. In these cases, the estimate of the number of columns is still reasonable, but sparsity constraints seem to overshrink the coefficients inside the columns.

**Algorithmics** The proposed framework ensures rigorous identification conditions enforcing a UGLT structure and the 3579 counting rule. However, these two conditions are enforced in two different ways. The UGLT structure is cunningly imposed by introducing a prior on the pivots. The 3579 counting rule is applied in the post-process, eliminating the MCMC draws that do not satisfy the rule. A natural extension to the proposed algorithm would be to impose the 3579 rule, incorporating it into the prior choice. Retaining only posterior samples satisfying the 3579 rule amounts to looking at a conditional posterior, and it is not clear how this posterior relates to one where the 3579 constraint would have been applied a priori. Also related to the algorithmic approach, it would be interesting if there was a way to avoid the reversible jump MCMC algorithm. Miller and Harrison (2018) show that for Mixture of Finite Mixtures (MFM), it is possible to build on the connection between MFM and classical Bayesian nonparametric priors such as the Dirichlet Process to use efficient samplers based on Neal’s algorithms (Neal, 2000) for instance. This requires the availability of the Exchangeable Partition Probability Function. There is a related concept for factor models, introduced in Broderick et al. (2013), which is that of Exchangeable Factors Probability Function. It would be an interesting research question to investigate whether this can be employed to construct algorithms bypassing reversible jump MCMC.



In conclusion, this paper offers an elegant sparse Bayesian factor analysis model overcoming identification and computational challenges associated with unknown factor

dimensions. Future exploration of the asymptotic properties and sensitivity to prior choices, especially in different application domains, could further solidify its practical utility.

## References

- Alamichel, L., Bystrova, D., Arbel, J., and Kon Kam King, G. (2024). “Bayesian mixture models (in)consistency for the number of clusters.” *Scandinavian Journal of Statistics*, in press. 65
- Broderick, T., Pitman, J., and Jordan, M. I. (2013). “Feature allocations, probability functions, and paintboxes.” *Bayesian Analysis*. MR3150470. doi: <https://doi.org/10.1214/13-BA823>. 66
- Castillo, I., Schmidt-Hieber, J., and van der Vaart, A. (2015). “Bayesian linear regression with sparse priors.” *The Annals of Statistics*, 43(5). MR3375874. doi: <https://doi.org/10.1214/15-AOS1334>. 65
- Miller, J. W. and Harrison, M. T. (2018). “Mixture models with a prior on the number of components.” *Journal of the American Statistical Association*, 113(521): 340–356. MR3803469. doi: <https://doi.org/10.1080/01621459.2016.1255636>. 66
- Neal, R. M. (2000). “Markov chain sampling methods for Dirichlet process mixture models.” *Journal of Computational and Graphical Statistics*, 9(2): 249–265. MR1823804. doi: <https://doi.org/10.2307/1390653>. 66
- Ohn, I. and Kim, Y. (2022). “Posterior consistency of factor dimensionality in high-dimensional sparse factor models.” *Bayesian Analysis*, 17(2): 491–514. MR4483228. doi: <https://doi.org/10.1214/21-ba1261>. 66
- Ohn, I., Lin, L., and Kim, Y. (2024). “A Bayesian sparse factor model with adaptive posterior concentration.” *Bayesian Analysis*, 19(4): 1277–1301. MR4802851. doi: <https://doi.org/10.1214/23-ba1392>. 66
- Ročková, V. and George, E. I. (2016). “Fast Bayesian factor analysis via automatic rotations to sparsity.” *Journal of the American Statistical Association*, 111(516): 1608–1622. MR3601721. doi: <https://doi.org/10.1080/01621459.2015.1100620>. 66
- Scott, J. G. and Berger, J. O. (2010). “Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem.” *The Annals of Statistics*, 38(5): 2587–2619. MR2722450. doi: <https://doi.org/10.1214/10-AOS792>. 65

## Contributed Discussion

Alessandro Zito\* and Jeffrey W. Miller†

We commend Frühwirth–Schnatter, Hosszejini, and Lopes on their valuable contribution, which offers an insightful solution to two long-standing problems of traditional Bayesian factor modeling, namely (i) identifiability of the factor loadings, and (ii) selection of the number of factors to use in the model. Solving both problems within a Bayesian framework is desirable for improving uncertainty quantification and interpretability in many applications.

The authors introduce a novel extension of the lower triangularity condition: letting  $l_1, \dots, l_k$  denote the indices of the first nonzero entries in rows  $1, \dots, k$  of the loadings matrix, the *unordered generalized lower triangular* (UGLT) condition is that  $l_1, \dots, l_k$  are all distinct. The authors enforce this UGLT constraint during Markov chain Monte Carlo (MCMC) sampling, and use the fact that UGLT provides identifiability up to a signed permutation in order to recover full identifiability in post-processing steps after sampling.

This approach suggests intriguing possibilities for further improvement and use in other latent factorization models. First, it would be interesting to extend the methodology to generalized bilinear models (GBMs). GBMs provide an extension of factor analysis to non-Gaussian data by using a generalized linear model for each entry of the data matrix and parametrizing the linear predictor with a low-rank latent factorization (Choulakian, 1996; Miller and Carter, 2020; Nicol and Miller, 2023). This class of models is especially useful when dealing with the sparse, high-dimensional count matrices that are routinely encountered in genome sequencing. We have found that enforcing identifiability can help improve estimation accuracy in GBMs (Miller and Carter, 2020; Nicol and Miller, 2023), and selecting an appropriate number of latent factors remains a challenging problem for GBMs. Thus, the authors’ approach to solving these problems may be of particular utility for this class of models.

Another direction would be to consider applying the methodology to non-negative matrix factorization (Lee and Seung, 2000). Non-negative matrix factorization (NMF) is used in cancer genomics to perform mutational signatures analysis, which deconvolves patterns of mutations in the DNA of tumor cells to uncover the corresponding oncogenic processes (Nik-Zainal et al., 2012; Alexandrov et al., 2013). Denoting as  $X_{ij}$  the number of mutations of type  $i = 1, \dots, I$  in tumor sample  $j = 1, \dots, J$ , it is standard to model  $X_{ij} \sim \text{Poisson}(\sum_{k=1}^K r_{ik}\theta_{kj})$  where  $r_k = (r_{1k}, \dots, r_{Ik})$  is a *mutational signature* capturing the rate at which each mutation type occurs under process  $k$ , while  $\theta_{k1}, \dots, \theta_{kJ}$  are loadings representing the activity of signature  $k$  on each sample. Identification of specific signatures in patients shows promise for improving patient outcomes with precision therapeutics (Aguirre et al., 2018). Correctly identifying which signatures are active in which patients is crucial for the effectiveness of this treatment approach.

---

\*Department of Biostatistics, Harvard University, [azito@hsph.harvard.edu](mailto:azito@hsph.harvard.edu)

†Department of Biostatistics, Harvard University, [jwmiller@hsph.harvard.edu](mailto:jwmiller@hsph.harvard.edu)

Sparsity-inducing priors have been used in Bayesian NMF models for mutational signatures analysis, including the use of overfitted models with compressive hyperpriors for estimating the number of factors (Zito and Miller, 2024) and multi-study NMF frameworks to detect both common and cancer-specific mutational signatures (Grabski et al., 2023). It would be interesting to adapt the authors’ UGLT approach to devise a sparsity-inducing NMF model to jointly quantify uncertainty in the loadings, signatures, and the number of active signatures. In this respect, zeros in the loading matrix of a UGLT structure would translate into a more transparent representation where the presence or absence of signature activity in a patient is more explicit. Furthermore, the presence of structural zeros has a connection with the identifiability of the NMF model itself, as first suggested by Donoho and Stodden (2003). See Gillis (2020) for a detailed account and discussion.

Finally, we would like to point out that while the UGLT structure relaxes the strict lower triangular condition that is often used, it can still depend strongly on the values of just a few entries of the loadings matrix. If there is high uncertainty in these entries, then it seems possible that the resulting inferences could be strongly affected. The authors’ approach of performing post-processing conditionally on the most probable factor dimension  $r$  and pivot sequence  $l_1, \dots, l_r$  alleviates this issue, but also reduces the number of posterior samples used to form posterior estimates. This suggests the possibility of using more information about the sparsity pattern of each column, rather than just index of the first nonzero entry.

In summary, the usefulness of having a UGLT prior structure in a Gaussian factor model is clearly demonstrated in the article of Frühwirth-Schnatter, Hosszejini, and Lopes, and it may also have applications beyond traditional factor analysis as well.

## References

- Aguirre, A. J., Nowak, J. A., Camarda, N. D., et al. (2018). “Real-time genomic characterization of advanced pancreatic cancer to enable precision medicine.” *Cancer Discovery*, 8(9): 1096–1111. doi: <https://doi.org/10.1158/2159-8290.CD-18-0275>. 68
- Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., et al. (2013). “Signatures of mutational processes in human cancer.” *Nature*, 500(7463): 415–421. doi: <https://doi.org/10.1038/nature12477>. 68
- Choulakian, V. (1996). “Generalized bilinear models.” *Psychometrika*, 61: 271–283. doi: <https://doi.org/10.1007/BF02294339>. 68
- Donoho, D. and Stodden, V. (2003). “When does non-negative matrix factorization give a correct decomposition into parts?” In: Thrun, S., Saul, L., and Schölkopf, B. (eds.), *Advances in Neural Information Processing Systems*, volume 16. MIT Press. [https://proceedings.neurips.cc/paper\\_files/paper/2003/file/1843e35d41ccf6e63273495ba42df3c1-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2003/file/1843e35d41ccf6e63273495ba42df3c1-Paper.pdf). 69
- Gillis, N. (2020). *Nonnegative Matrix Factorization*. Data Science Book Series. Society for Industrial and Applied Mathematics. MR4191210. doi: <https://doi.org/10.1137/1.9781611976410>. 69

- Grabski, I., Trippa, L., and Parmigiani, G. (2023). “Bayesian multi-study non-negative matrix factorization for mutational signatures.” 1–12. doi: <https://doi.org/10.1101/2023.03.28.534619>. 69
- Lee, D. and Seung, H. S. (2000). “Algorithms for non-negative matrix factorization.” In: Leen, T., Dietterich, T., and Tresp, V. (eds.), *Advances in Neural Information Processing Systems*, volume 13. MIT Press. [https://proceedings.neurips.cc/paper\\_files/paper/2000/file/f9d1152547c0bde01830b7e8bd60024c-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2000/file/f9d1152547c0bde01830b7e8bd60024c-Paper.pdf). 68
- Miller, J. W. and Carter, S. L. (2020). “Inference in generalized bilinear models.” [arXiv:2010.04896](https://arxiv.org/abs/2010.04896). 68
- Nicol, P. B. and Miller, J. W. (2023). “Model-based dimensionality reduction for single-cell RNA-seq using generalized bilinear models.” doi: <https://doi.org/10.1101/2023.04.21.537881>. 68
- Nik-Zainal, S., Alexandrov, L. B., Wedge, D. C., et al. (2012). “Mutational processes molding the genomes of 21 breast cancers.” *Cell*, 149(5): 979–993. doi: <https://doi.org/10.1016/j.cell.2012.04.024>. 68
- Zito, A. and Miller, J. W. (2024). “Compressive Bayesian non-negative matrix factorization for mutational signatures analysis.” [arXiv:2404.10974](https://arxiv.org/abs/2404.10974) 69

# Contributed Discussion

Elizabeth Bersson\* 

## Introduction

I would like to congratulate the authors on their important contributions to latent factor modeling, and, in particular, the insights provided in this manuscript. In this discussion, I describe motivations for sparsity in covariance estimation and propose consideration of an alternative view of parsimony with respect to covariance estimation.

Accurate high dimensional covariance estimation is a challenging task. A common approach to estimate a covariance matrix in such a setting is to utilize a factor model that represents a covariance matrix as a diagonal variance matrix plus a possibly low-rank covariance matrix. In contrast to estimating an unstructured covariance matrix, such an approach can reduce the number of unknown parameters to be estimated, thereby resulting in more stable covariance estimates. In this work, the authors present a prior framework for a latent factor model that simultaneously encourages within-column and within-row sparsity in the factor loadings matrix. In what follows, I primarily elaborate on this contribution that incorporates two interpretations of model parsimony, low-rank covariance estimation and zero-sparsity shrinkage, in latent factor modeling.

## Sparsity in Modeling

In statistical modeling, parsimony, whereby the smallest number of parameters are used such that the population is accurately represented (Box et al., 2015), is often desirable for a number of reasons. For example, in estimation of a mean regression coefficient vector, a natural manifestation of parsimony is to encourage sparsity of the coefficients. In such a setting, a sparse solution refers to one in which many coefficients are exactly equal to zero, as with a lasso penalization (Tibshirani, 1996), or very nearly equal to zero, as with a ridge penalization (Hoerl and Kennard, 1970). Such penalized regression coefficient estimation approaches are desirable as they can yield estimates with reduced error and greater predictive accuracy. Additionally, this definition of sparsity in mean regression implies a model selection framework whereby fewer covariates are used to predict the outcome. In this way, a sparse regression coefficient solution features streamlined interpretation of model parameters.

While mean estimation has received more substantial attention in the literature, these notions of parsimony and sparsity have been extended to covariance estimation as well, and many popular covariance estimation approaches that encourage sparsity such as diagonality (Daniels and Kass, 1999) and thresholding (Bickel and Levina, 2008) have been proposed. Favorable theoretical results for such approaches suggest superior

---

\*Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, [ebersson@mit.edu](mailto:ebersson@mit.edu)

performance in the case of sparse population covariances. For a detailed review of regularization methods for high dimensional covariance estimation, see Pourahmadi (2013).

Similarly, prior distributions used for the factor loadings matrix commonly favor a sparse matrix whereby the factor loadings matrix is shrunk towards a zero matrix, corresponding to a diagonal prior marginal covariance matrix. Recently, some non-exchangeable priors have been proposed that offer flexibility, but still strongly favor zero-sparsity. For example, the structured increasing shrinkage process proposed in Schiavon et al. (2022) shares information across the rows of the factor loadings matrix; this framework also allows for both column- and row-sparsity.

## General Parsimony for Covariances

By providing a framework for simultaneously encouraging column-sparsity and row-sparsity in a factor loadings matrix, the approach detailed in Frühwirth-Schnatter et al. (2024) couples dimension reduction with structured covariance shrinkage estimation. In this way, the proposed framework encourages two interpretations of parsimony. For one, by encouraging column-sparsity, the proposed prior allows for a simpler model with respect to lower rank of the factor covariance matrix. For another, by concurrently encouraging row-sparsity in the factor matrix, a simpler model is achieved by allowing for fewer non-zero parameters in the factor loadings matrix.

To elaborate, row-sparsity in this work is motivated by the notion of utilizing a simple structure for factor modeling as introduced in Thurstone (1947), which allows for practically meaningful factors. The concept of a simple structure, nicely summarized in Joereskog (1966), emphasizes sparsity among the active factors. That is, among the non-zero columns of the factor loadings matrix, each row of the factor loadings matrix should preferably have a large number of zero elements. While the interpretability of such simple structures may be appealing, it can be a restrictive structure. For example, the most parsimonious model under this objective of row-sparsity corresponds to a zero factor loadings matrix, and, correspondingly, a diagonal prior covariance matrix.

For diagonal or sparse population covariance matrices, these developed methods are appealing; however, the plausibility of such a covariance structure in practice is uncertain. While this approach, and others mentioned, may result in more stable estimates when compared to alternatives with many more unknown parameters, a compromise may be available whereby a different interpretation of parsimony is considered for high dimensional covariance estimation. In particular, a more flexible alternative that allows for non-zero shrinkage among active columns of the factor loadings matrix might yield more accurate estimates in regimes with dense population covariances. When coupled with the benefit of dimension reduction via column-sparsity in factor modeling, this additional layer of flexibility that allows for prior non-zero correlations among the variables can potentially yield more accurate covariance estimates than current approaches that favor shrinkage towards a diagonal covariance.



## References

- Bickel, P. J. and Levina, E. (2008). “Covariance regularization by thresholding.” *Annals of Statistics*, 36(6): 2577–2604. MR2485008. doi: <https://doi.org/10.1214/08-AOS600>. 71
- Box, G., Jenkins, G., Reinsel, G., and Ljung, G. (2015). *Time Series Analysis: Forecasting and Control*. Hoboken: Wiley, 5th edition. MR3379415. 71
- Daniels, M. J. and Kass, R. E. (1999). “Nonconjugate Bayesian estimation of covariance matrices and its use in hierarchical models.” *Journal of the American Statistical Association*, 94(448): 1254–1263. MR1731487. doi: <https://doi.org/10.2307/2669939>. 71
- Frühwirth-Schnatter, S., Hosszejni, D., and Lopes, H. F. (2024). “Sparse Bayesian factor analysis when the number of factors is unknown.” *Bayesian Analysis*. 72
- Hoerl, A. E. and Kennard, R. W. (1970). “Ridge regression: biased estimation for nonorthogonal problems.” *Technometrics*, 12(1): 55–67. MR0611894. doi: <https://doi.org/10.1080/01966324.1981.10737061>. 71
- Joereskog, K. G. (1966). “Testing a simple structure hypothesis in factor analysis.” *Psychometrika*, 31(2): 165–178. MR0198612. doi: <https://doi.org/10.1007/BF02289505>. 72
- Pourahmadi, M. (2013). *High-Dimensional Covariance Estimation*. Hoboken: John Wiley & Sons. MR3235948. doi: <https://doi.org/10.1002/9781118573617>. 72
- Schiavon, L., Canale, A., and Dunson, D. B. (2022). “Generalized infinite factorization models.” *Biometrika*, 109(3): 817–835. MR4472850. doi: <https://doi.org/10.1093/biomet/asab056>. 72
- Thurstone, L. L. (1947). *Multiple Factor Analysis*. Chicago: University of Chicago, 406–427. MR1526847. doi: <https://doi.org/10.2307/2304512>. 72
- Tibshirani, R. (1996). “Regression shrinkage and selection via the lasso.” *Journal of the Royal Statistical Society: Series B*, 58(1): 267–288. MR1379242. 71

# Invited Discussion

Antonio Canale<sup>\*</sup>, Lorenzo Schiavon<sup>†</sup>, and Federica Stolf<sup>‡</sup>

## 1 Introduction

We would like to congratulate the authors for their valuable contribution to sparse Bayesian factor analysis literature. Coupling identifiability through the unordered generalized lower triangular (UGLT) structures discussed in Frühwirth-Schnatter et al. (2023) with an efficient Markov chain Monte Carlo (MCMC) algorithm enables straightforward uncertainty quantification for all unknowns, including the factor loading’s sparse structure and dimension, thereby facilitating interpretation—often the ultimate goal of latent factor methods. In this discussion, we aim to highlight how the proposed approach can also be advantageous when exogenous information about the variables is available and integrated into the inferential process, specifically in inferring the sparsity structure of the factor loading matrix.

## 2 Order Non-invariant Priors for Factor Loadings

In imposing a UGLT structure on the factor loadings, Frühwirth-Schnatter et al. (2024) implicitly propose a prior which is not invariant with respect to the order in which the variables appear. In fact, under the proposed prior, the order of the variables influences the induced covariance structure, by increasing the probability of having sparse loadings for first variables. To see this, consider the model

$$\mathbf{y}_t = \boldsymbol{\beta} \mathbf{f}_t + \mathbf{e}_t, \quad \mathbf{e}_t \sim N(0, \boldsymbol{\Sigma}), \quad \mathbf{f}_t \sim N(0, \mathbf{I}),$$

and introduce  $\pi_{ij}$  to denote the conditional prior probability that  $\beta_{ij}$  is not zero given the location of the other pivots, i.e.  $\pi_{ij} = \text{pr}(\beta_{ij} \neq 0 \mid \mathbf{l}_{r,-j})$ . Then, the uniform prior distribution on  $l_j \mid \mathbf{l}_{r,-j}$ , reported in Equation (3.5) of the paper leads to

$$\begin{aligned} \pi_{ij} &= \text{pr}(\delta_{ij} = 1 \mid \mathbf{l}_{r,-j}) \\ &= \sum_{h \in \mathcal{L}(\mathbf{l}_{r,-j})} \text{pr}(\delta_{ij} = 1 \mid l_j = h, \mathbf{l}_{r,-j}) \text{pr}(l_j = h \mid \mathbf{l}_{r,-j}) \\ &= \frac{1}{m-r+1} \mathbb{1}\{i \in \mathcal{L}(\mathbf{l}_{r,-j})\} + \tau_j \frac{i - R_{i,-j} - 1}{m-r+1}, \end{aligned} \quad (1)$$

where  $R_{i,-j}$  is the number of pivots in  $\mathcal{L}(\mathbf{l}_{r,-j})$  smaller than  $i$ . The second term in Equation (1) is not decreasing in  $i$ , that implies an ordering of the  $\pi_{ij}$  for  $i > 1$  and  $i \in \mathcal{L}(\mathbf{l}_{r,-j})$ . In other terms, the prior probability of being sparse for an element of any column  $j$  of  $\boldsymbol{\beta}$  tends to be smaller as index  $i$  increases.

---

<sup>\*</sup>Department of Statistical Sciences, University of Padova, [antonio.canale@unipd.it](mailto:antonio.canale@unipd.it)

<sup>†</sup>Department of Economics, Ca’ Foscari University of Venice, [lorenzo.schiavon@unive.it](mailto:lorenzo.schiavon@unive.it)

<sup>‡</sup>Department of Statistical Science, Duke University, [federica.stolf@duke.edu](mailto:federica.stolf@duke.edu)

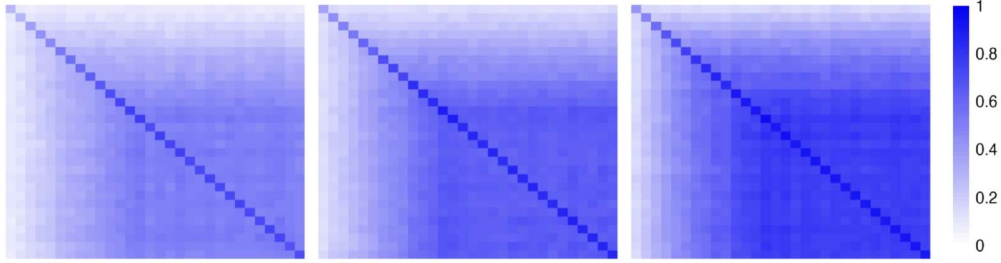


Figure 1: Monte Carlo mean across 500 simulations of  $\delta\delta^T$  with different values for the parameter of the beta prior:  $\alpha = 3$  (left panel),  $\alpha = 5$  (center panel) and  $\alpha = 7$  (right panel).

Figure 1 provides some numerical evidence of this behavior. We simulated 500 different sparsity matrices  $\delta$  from the prior proposed by the authors using as distribution on the slab probabilities the finite one-parameter-beta prior with three different values for the parameter  $\alpha$ , encoding different sparsity levels. Figure 1 shows the Monte Carlo average of  $\delta\delta^T$  as a measure of the prior probability of having a non null covariance in the global covariance matrix  $\Omega = \beta\beta^T + \Sigma$ . The expected decreasing sparsity for the elements of  $\beta$  induces a higher probability of having a null correlation between the first variables, *a priori*.

Although this behavior may seem undesirable when there is no obvious order between variables, in many applications, prior knowledge can guide the prior sparsity pattern. For example, in a previous work (Frühwirth-Schnatter et al., 2023), the authors modeled a dataset about the monthly log-returns from the New York Stock Exchange where firms were ordered by industrial sector. More generally, information about the relationships among variables may be linked to multiple observed traits of those variables. Consistently with this, Schiavon et al. (2022) proposed to induce a sparsity structure in the factor loading matrix by means of metacovariates, i.e. information associated to each marginal measured variable. Similarly to the approach presented by Frühwirth-Schnatter et al. (2024), this method is also not invariant with respect to the variables. However, the dependencies are informed by metacovariates that are associated with intrinsic similarities among the variables, rather than their order.

Following Schiavon et al. (2022) and adapting their notation to that of the paper by Frühwirth-Schnatter et al. (2024), the structured sparsity prior for the factor loadings is

$$\beta_{ij} \mid \gamma_j, \phi_{ij} \sim N(0, \gamma_j \phi_{ij}), \quad \text{pr}(\phi_{ij} = 1 \mid \theta_j) = c_m \text{logit}^{-1}(\mathbf{x}_i^\top \theta_j), \quad (2)$$

where  $\gamma_j$  is a column specific scale, and  $\phi_{ij}$  is a Bernoulli distributed local scale, with expectation depending on an offset  $c_m \in (0, 1)$  and on the metacovariates  $\mathbf{x}_i$  pertaining to the  $i$ -th marginal. These metacovariates can be categorical variables, such as the industrial sector in the previously mentioned stock exchange application, or continuous variables.

The family of factor models discussed by Schiavon et al. (2022), shares the usual identifiability issues, including the factor loadings rotational ambiguity. This issue is exacerbated in the context of structured sparsity, as obtaining a meaningful posterior summary that maintains a relation with the dependence from the metacovariates becomes a challenging task. In fact, standard MCMC post-processing techniques cannot be applied, as they would completely destroy the relationships across the various layers of the hierarchical model (2). As a result, the authors opted to use an approximation of the maximum a posteriori as the posterior point summary.

However, the solutions presented in the paper by Frühwirth-Schnatter et al. (2024) are likely to solve the identifiability issues also in the structured sparsity approach of Schiavon et al. (2022). We sketch this in the following pages.

### 3 Structured Sparsity UGLT Prior

Rather than deriving  $\pi_{ij}$  from the full conditional distribution of the pivot locations, we start specifying it in a way that is consistent with the structured sparsity approach. This allows us to obtain the prior distribution for the pivot locations as a consequence. Specifically, we aim at designing a prior process such that the prior probability that  $\beta_{ij}$  is not zero depends mainly on variable traits contained into metacovariates  $\mathbf{x}_i$ , rather than merely on variable ordering.

To achieve this, we specify a  $m \times r$  unconstrained sparsity matrix  $\Phi$  with binary entries  $\phi_{ij}$ . Similarly to the structured increasing shrinkage process in Schiavon et al. (2022), we propose to relate the probability  $\text{pr}(\phi_{ij} = 1)$  to the metacovariates  $\mathbf{x}_i$  through a logistic transformation of the linear predictor  $\mathbf{x}_i^\top \boldsymbol{\theta}_j$ . Then, indicating with  $\boldsymbol{\delta}$  the indicator matrix such that  $\delta_{ij} = \mathbb{1}(\beta_{ij} \neq 0)$ , we impose a UGLT structure on  $\boldsymbol{\delta}$  sampling uniformly a column index  $j$  and letting, for  $i = 1, \dots, m$ ,

$$\delta_{ij} = \phi_{ij} \mathbb{1}\{i \in \mathcal{L}(\mathbf{l}_{r,-j})\} + \phi_{ij} \mathbb{1}\{i \notin \mathcal{L}(\mathbf{l}_{r,-j})\} \mathbb{1}(i > l_j).$$

In other terms, given the pivots of the other columns  $\mathbf{l}_{r,-j}$ , the pivot  $l_j$  of column  $j$  corresponds to the first nonzero element of column  $j$  of the matrix  $\Phi$ , if  $i$  is not a pivot row in other columns. Notably, in doing this, we are creating a link between  $\pi_{ij} = \text{pr}(\delta_{ij} = 1 \mid \mathbf{l}_{r,-j})$  and the probability of  $\phi_{ij} = 1$ , thus having the metacovariates impacting each  $\pi_{ij}$ .

Under these settings the induced full conditional prior on the pivot location at column  $j$ , conditional to the locations of the other pivots, is

$$\text{pr}(l_j = i \mid \mathbf{l}_{r,-j}) = \mathbb{1}\{i \in \mathcal{L}(\mathbf{l}_{r,-j})\} \text{pr}(\phi_{ij} = 1) \prod_{\substack{m \in \mathcal{L}(\mathbf{l}_{r,-j}), \\ m < i}} \{1 - \text{pr}(\phi_{mj} = 1)\}.$$

Similarly,

$$\text{pr}(l_j < i \mid \mathbf{l}_{r,-j}) = \sum_{h=1}^{i-1} \text{pr}(l_j = h \mid \mathbf{l}_{r,-j}) = 1 - \prod_{\substack{m \in \mathcal{L}(\mathbf{l}_{r,-j}), \\ m < i}} \{1 - \text{pr}(\phi_{mj} = 1)\}.$$

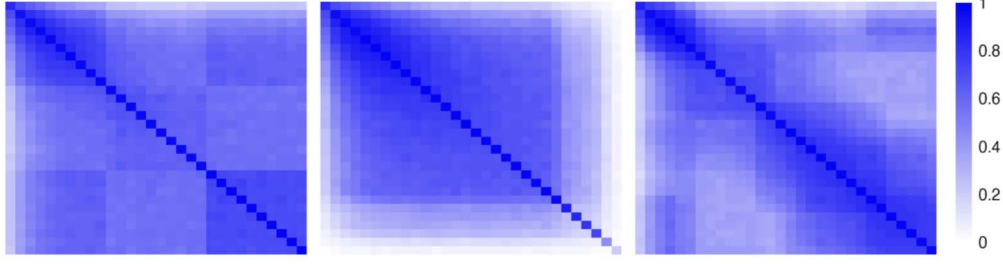


Figure 2: Monte Carlo mean across 500 simulations of  $\delta\delta^\top$ , under the structured sparsity prior with UGLT constraints and univariate metacovariates: categorical metacovariate with three levels (left); continuous metacovariate with linear relationship with respect to the variable order (center); continuous metacovariate with quadratic effect with respect to the variable order (right).

We are now able to derive the probability that  $\beta_{ij}$  is not sparse conditionally on the other pivots' locations, i.e.

$$\begin{aligned}\pi_{ij} &= \text{pr}(\delta_{ij} = 1 \mid \mathbf{l}_{r,-j}) \\ &= \text{pr}\{\phi_{ij} = 1 \mid i \in \mathcal{L}(\mathbf{l}_{r,-j})\} \mathbb{1}\{i \in \mathcal{L}(\mathbf{l}_{r,-j})\} \\ &\quad + \text{pr}\{\phi_{ij} = 1 \mid i \notin \mathcal{L}(\mathbf{l}_{r,-j}), l_j < i\} \text{pr}(l_j < i \mid \mathbf{l}_{r,-j}) \mathbb{1}\{i \notin \mathcal{L}(\mathbf{l}_{r,-j})\}.\end{aligned}$$

Similar in spirit to Schiavon et al. (2022), we set

$$\text{pr}\{\phi_{ij} \mid i \in \mathcal{L}(\mathbf{l}_{r,-j})\} = c_m \text{logit}^{-1}(\mathbf{x}_i^\top \boldsymbol{\theta}_j),$$

such that  $\pi_{ij} = c_m \text{logit}^{-1}(\mathbf{x}_i^\top \boldsymbol{\theta}_j)$  for  $i \in \mathcal{L}(\mathbf{l}_{r,-j})$ . In fact, the variable ordering still has an impact for the  $i \notin \mathcal{L}(\mathbf{l}_{r,-j})$  as those cannot be pivots. To mitigate this aspect, we specify

$$\text{pr}\{\phi_{ij} \mid i \notin \mathcal{L}(\mathbf{l}_{r,-j})\} = \min\{c_m \text{logit}^{-1}(\mathbf{x}_i^\top \boldsymbol{\theta}_j) / \text{pr}\{l_j < i \mid \mathbf{l}_{r,-j}\}, 1\},$$

resulting in

$$\begin{aligned}\pi_{ij} &= c_m \text{logit}^{-1}(\mathbf{x}_i^\top \boldsymbol{\theta}_j) \mathbb{1}\{i \in \mathcal{L}(\mathbf{l}_{r,-j})\} \\ &\quad + \min\{c_m \text{logit}^{-1}(\mathbf{x}_i^\top \boldsymbol{\theta}_j), \text{pr}(l_j < i \mid \mathbf{l}_{r,-j})\} \mathbb{1}\{i \notin \mathcal{L}(\mathbf{l}_{r,-j})\}.\end{aligned}\quad (3)$$

Figure 2 provides numerical examples illustrating the behavior of the proposed prior. Similarly to Figure 1, 500 different sparsity matrices  $\delta$  have been simulated independently using the structured sparsity structure prior in (3) under different scenarios for metacovariates. The left panel replicates a situation in which prior knowledge about grouping of variables is available. This setup mirrors the previously discussed applications to the New York Stock Exchange monthly log-returns, where the industrial sectors of the firms were known. In this case, we simulate using the proposed prior, with  $\mathbf{x}_i$

corresponding to a categorical metacovariate with three levels. In contrast, the second and third panel assume continuous metacovariates, with linear and quadratic effects, respectively, with respect to the variable order. The Monte Carlo averages of  $\delta\delta^T$  shown in the plots demonstrate different induced sparse correlation patterns, consistent with the information contained in the metacovariates, while the order-dependent decreasing sparsity behavior noticed under (1) almost vanishes. In the leftmost panel, the higher probability of non-null correlations between variables within the same group is clearly visible in blocks. In the middle panel, the first set of variables shows strong correlation *a priori*, while the third panel showcases that a simple reordering of the variables is not sufficient to control order-dependence when there is a more complex structure in prior knowledge.

## 4 Posterior Inference Under Structured UGLT Prior

We sketch the main steps for posterior inference under the structured UGLT approach presented. We integrate the UGLT structure update into the Gibbs sampler proposed by Schiavon et al. (2022) as follows.

Loop over the columns in random order and, for each column, perform the steps below.

1. Update the list  $\mathbf{l}_{r,-j}$ .
2. Update the parameter  $\boldsymbol{\theta}_j$  with an MH move. Following Schiavon et al. (2022), we sample a proposal for the parameter vector  $\boldsymbol{\theta}_j^*$  from the logit regression full conditional, assuming  $\text{pr}(\phi_{ij} = 1) = c_m \text{logit}^{-1}(\mathbf{x}_i^\top \boldsymbol{\theta}_j)$  for each  $i$ , and exploiting the Pólya–Gamma data augmentation proposed by Polson et al. (2013). We accept the proposal with probability  $\min\{p_{\boldsymbol{\theta}_j}(\boldsymbol{\theta}_j^* | -)/p_{\boldsymbol{\theta}_j}(\boldsymbol{\theta}_j | -), 1\}$ , where  $p_{\boldsymbol{\theta}_j}(\cdot | -)$  is proportional to the full conditional distribution of  $\boldsymbol{\theta}_j$ .
3. Set the prior probability of  $\phi_{ij}$  conditionally on  $\boldsymbol{\theta}_j$  and  $\mathbf{l}_{r,-j}$  as
 
$$\text{pr}\{\phi_{ij} | \boldsymbol{\theta}_j, \mathbf{l}_{r,-j}\} = c_m \text{logit}^{-1}(\mathbf{x}_i^\top \boldsymbol{\theta}_j) \mathbb{1}\{i \in \mathcal{L}(\mathbf{l}_{r,-j})\} \\ + \min\{c_m \text{logit}^{-1}(\mathbf{x}_i^\top \boldsymbol{\theta}_j) / \text{pr}(l_j < i | \mathbf{l}_{r,-j}), 1\} \mathbb{1}\{i \notin \mathcal{L}(\mathbf{l}_{r,-j})\}.$$
4. Update the  $m$  elements of the  $j$ -th column of  $\Phi$  in parallel from its Bernoulli full conditional, which is available in closed form. With the prior probability  $\text{pr}(\phi_{ij} | \boldsymbol{\theta}_j, \mathbf{l}_{r,-j})$  and the other columns of the loadings matrix known, we employ the strategy outlined in Schiavon et al. (2022).
5. Identify the pivot  $l_j$  as the first non-zero element in column  $j$  of  $\Phi$  that is not a pivot in other columns, i.e.  $i \in \mathcal{L}(\mathbf{l}_{r,-j})$  and such that  $\phi_{ij} = 1$ .
6. Update the  $j$ -th column of the sparsity matrix  $\boldsymbol{\delta}$  by setting  $\delta_{ij} = 0$  if  $i < l_j$ , and  $\delta_{ij} = \phi_{ij}$  otherwise.
7. Update the  $j$ -th column of  $\boldsymbol{\beta}$  conditionally on  $\boldsymbol{\delta}_{.,j}$ , and the factor model mean  $\boldsymbol{\beta}\mathbf{f}_t$ .

## References

- 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): 26. [74, 75](#)
- Frühwirth-Schnatter, S., Hosszejni, D., and Lopes, H. F. (2024). “Sparse Bayesian factor analysis when the number of factors is unknown.” *Bayesian Analysis*, 1–44. doi: <https://doi.org/10.1214/24-BA1423>. [74, 75, 76](#)
- Polson, N. G., Scott, J. G., and Windle, J. (2013). “Bayesian inference for logistic models using Pólya–Gamma latent variables.” *Journal of the American statistical Association*, 108(504): 1339–1349. [MR3174712](#). doi: <https://doi.org/10.1080/01621459.2013.829001>. [78](#)
- Schiavon, L., Canale, A., and Dunson, D. B. (2022). “Generalized infinite factorization models.” *Biometrika*, 109(3): 817–835. [MR4472850](#). doi: <https://doi.org/10.1093/biomet/asab056>. [75, 76, 77, 78](#)

# Invited Discussion

Christian Aßmann\*, Sylvia Kaufmann<sup>‡</sup>, and Markus Pape<sup>¶</sup>

## 1 Introduction

Frühwirth-Schnatter et al. (2024) investigate possibilities to induce sparsity in the loading matrix of a static factor model, while simultaneously estimating the number of latent factors. The suggested approach is based on the identification strategy imposing an unordered generalized lower triangular (UGLT) structure on the loading matrix, and the authors provide a comprehensive review of the properties of this identification approach. They thoroughly introduce the prior setup, which induces shrinkage in line with the desired UGLT structure. Different possibilities to steer the hyperparameters of the prior distribution are discussed and evidence from the literature is provided to highlight the different prior elicitation strategies. The paper also points at implications of prior specifications with regard to the Heywood problem, relating to multimodality of the posterior distribution with one mode at the boundary of the parameter space. Properly specified priors induce shrinkage away from the boundary of the parameter space. The prior specification gives rise to a special Markov Chain Monte Carlo (MCMC) setup. The innovative MCMC algorithm allows for sampling from the augmented posterior distribution implied by the high-dimensional parameter space when the number of factors is unknown, under an imposed UGLT structure on the loading matrix. The MCMC algorithm consists of two main blocks. In the first block, the parameters are sampled conditional on the number of latent factors, while in the second block, the number of factors is explored via a reversible jump (RJ) step. In addition to sampling from the full conditional distributions, boosting based on ancillarity-sufficiency interweaving is implemented to improve mixing. While the imposed UGLT structure ensures rotational identification, the draws are post-processed according to the 3579 counting rule to ensure variance identification as well. Applications are provided in form of a simulation study and two empirical illustrations using exchange rate data for 22 currencies against the euro and for returns of 63 firms listed at the NYSE.

In the following, we discuss several potential refinements and possible extensions that could be considered in future applications of the suggested approach.

## 2 Refinements

We discuss refinements related to the properties of the identification strategy and the induced shrinkage.

---

\*Leibniz Institute for Educational Trajectories, Bamberg, Germany, [christian.assmann@lifbi.de](mailto:christian.assmann@lifbi.de)

\*Chair of Survey Statistics and Data Analysis, Otto-Friedrich-Universität, Germany

<sup>‡</sup>Study Center Gerzensee, Switzerland, [sylvia.kaufmann@szgerzensee.ch](mailto:sylvia.kaufmann@szgerzensee.ch)

<sup>‡</sup>Faculty of Business and Economics, University of Basel, Switzerland

<sup>¶</sup>Department of Economics, Ruhr-Universität Bochum, Germany, [markus.pape@rub.de](mailto:markus.pape@rub.de)



## 2.1 Explore Rather Than Impose a (U)GLT Structure

Given that a UGLT structure may be transformed into a GLT structure by appropriate column permutations, we expose our arguments based on the GLT structure in the following. As discussed in Hauzenberger and Koop (2024), imposing a GLT structure when estimating a factor model is not order-invariant and needs a critical selection of leading or pivot units in the data set. Missing this consideration may induce the same issues brought forth against pre-imposing a PLT structure (Afmann et al., 2016; Chan et al., 2018; Frühwirth-Schnatter et al., 2023). In the present paper, Frühwirth-Schnatter et al. (2024) provide an MCMC scheme to identify so-called pivot units by posterior inference, without re-ordering units while sampling. The sampler thus provides an alternative to Carvalho et al. (2008), who detect pivot series by sequentially adding factors to the model as long as they occur to strongly load on series. The unit most strongly loaded by a factor is re-ordered into the group of leading units, which eventually leads to a PLT structure in the factor loading matrix. In contrast to that, imposing a GLT structure without re-ordering units introduces order non-invariance, which may eventually blur factor interpretation. For illustration, consider the sparse factor loading matrices

$$\Lambda = \begin{pmatrix} 0.7500 & 0 \\ 0 & 0 \\ -0.2800 & 0.6300 \\ 0 & 0.3000 \\ 0.7500 & -0.6000 \\ -0.4000 & -0.6000 \\ 0.2400 & 0 \end{pmatrix} \quad \text{and} \quad \tilde{\Lambda} = \begin{pmatrix} -0.4000 & -0.6000 \\ 0 & 0 \\ -0.2800 & 0.6300 \\ 0 & 0.3000 \\ 0.7500 & -0.6000 \\ 0.7500 & 0 \\ 0.2400 & 0 \end{pmatrix},$$

where underlying factors and factor interpretation remain unchanged when we exchange the first and sixth rows of  $\Lambda$ , to obtain  $\tilde{\Lambda}$ . While the units in  $\Lambda = (\lambda_1, \dots, \lambda_7)'$  with  $\lambda_i = (\lambda_{i1}, \lambda_{i2})'$ ,  $i = 1, \dots, 7$ , are ordered according to a GLT structure, they are not in  $\tilde{\Lambda}$ . The loadings are plotted as coordinates in Figure 1, where the solid lines represent the factor basis corresponding to the sparse representation. The label of units  $\lambda_i$  corresponds to the ordering in  $\tilde{\Lambda}$ . While this ordering does not conform to a sparse GLT representation, a GLT structure can be induced by rotation into (a dense) GLT form given as

$$\tilde{\Lambda}_R = \begin{pmatrix} 0.7211 & 0 \\ 0 & 0 \\ -0.3689 & 0.5824 \\ -0.2496 & 0.1664 \\ 0.0832 & -0.9569 \\ -0.4160 & -0.6240 \\ -0.1331 & 0.1997 \end{pmatrix}.$$

The rotation corresponds to a rotation in the factor basis, drawn as dashed lines in Figure 1. Imposing a GLT structure while estimating the model should eventually recover this representation. Although both representations are observationally equivalent, factor interpretation after rotation may be blurred or less obvious than under the sparse representation. We would argue that while both representations are variance and rotation

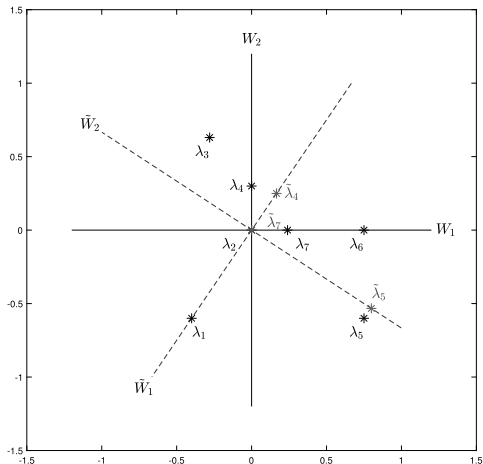


Figure 1: Coordinates corresponding to  $\tilde{\Lambda}$  (solid axes). Rotated basis corresponding to the rotation into the GLT structure  $\tilde{\Lambda}_R$  (dashed axes). Factor loadings corresponding to additional moderate shrinkage (gray), leading to  $\tilde{\Lambda}_{RS}$ . Axes are labeled towards the positive side..

identified, the one of interest is subject to the underlying research question. Interested in sparse factor analysis, we would prefer to recover the underlying sparse structure of the factor loading matrix, related to a potentially clear factor interpretation. The example illustrates that imposing a GLT structure without caring about the ordering of units may blur this inference.

The example suggests a potential refinement to the sampler proposed in Frühwirth-Schnatter et al. (2024). Rather than imposing a GLT structure, one could evaluate independently of unit ordering, whether a sparse loading matrix, sampled without imposing constraints, includes a GLT structure upon re-ordering the units. There may be various re-orderings leading to a GLT structure, but each one of them would be rotation identified. Such an online evaluation may be based on so-called set identification conditions, see Kaufmann and Pape (2024) who propose a geometric approach to factor model identification.

Another strategy to circumvent the order non-invariance of a GLT structure would be to re-order pivot units randomly to assess their stability, rather than exploring new pivot rows across a fixed ordering of units. Within the sampler, a re-ordering step would move a randomly chosen pivot unit into the last position, where none of the GLT constraints applies, i.e. the loading structure for this unit can be freely estimated. The unit retains its pivot status if the sampled loading structure remains unchanged when ranked last. Changes in the loading structure to the left of the pivot column are non-critical, as they do not affect the GLT structure.<sup>1</sup> Other changes are critical. First, the unit loses

<sup>1</sup>They are potentially critical for an UGLT structure, however. In a UGLT structure, zero loadings switching to non-zero may be critical for pivot units of higher rank order relative to the pivot unit re-ordered last.

its pivot status if there is a switch from a non-zero to a zero sampled pivot loading. Likewise, the unit also loses its pivot status if non-zero loadings are sampled to the right of the pivot element, i.e. elements that were initially restricted to zero by imposing a GLT structure. For instance, in the application to exchange rates (Subsection 5.2), we expect that the Australian dollar would also be loaded by the second factor if it were re-ordered last in the data set.

Overall, we conjecture that while a main part of the sampler proposed in Frühwirth-Schnatter et al. (2024), i.e. the RJMCMC step to detect the number of factors, would remain unchanged, relaxing the GLT constraint towards evaluating order-independent identification conditions or exploring pivot units by random re-ordering would largely simplify or increase the efficiency in sampling indicators and parameters. Eventually, the sampler would recover the true sparse, rotation identified structure underlying the data, independently of unit order.

## 2.2 When in, Don't Shrink

Frühwirth-Schnatter et al. (2024) induce shrinkage by specifying a Dirac spike and slab prior distribution on factor loadings. The probability of non-zero factor-specific loadings (the slab) follows an exchangeable shrinkage process (ESP) prior, see Legramanti et al. (2020), which increasingly penalizes the introduction of additional factors as the factor dimension increases. The prior provides the basis for designing the RJMCMC scheme to explore the factor dimension. Frühwirth-Schnatter et al. (2024) introduce additional column ( $\theta_j$ ) and row ( $\omega_{ij}$ ) shrinkage in the slab to induce sparsity. The combination is rather unconventional and extends the specification of Legramanti et al. (2020), who advise to formulate a rather diffuse slab to model the active loadings. It would seem more natural to induce row shrinkage by specifying an additional hierarchical layer in the ESP prior for the slab probability, in the spirit of Carvalho et al. (2008). Or, working only with a slab inducing global and local column and row shrinkage as suggested in Bhattacharya and Dunson (2011) or Kaufmann and Strachan (2024). Combining these two specifications raises the question which part, the ESP prior ( $\tau_j$ ) or factor-specific shrinkage ( $\theta_j$ ), dominates in determining model dimension and sparsity.

Returning to our previous example, we may illustrate that inducing too much sparsity when estimating the factor model under an imposed GLT structure may further blur factor interpretation, and additionally lead to biased factor-units associations (i.e. factor-driven correlations across units) and overemphasize the importance of some factors relative to others for some units. Inducing shrinkage in  $\tilde{\Lambda}_R$  (see Figure 1, black coordinates corresponding to the dashed factor basis), it is conceivable that some loadings may be shrunk towards some axis or even to zero, leading to

$$\tilde{\Lambda}_{RS} = \begin{pmatrix} 0.7211 & 0 \\ 0 & 0 \\ -0.3689 & 0.5824 \\ -0.3000 & 0 \\ 0 & -0.9605 \\ -0.4160 & -0.6240 \\ 0 & 0 \end{pmatrix}.$$

In Figure 1, the loadings (partially) shrunk to zero are indicated in gray. The non-zero loadings of Units 4 and 5 correspond to the length of the vector determined by the non-zero loadings in  $\tilde{\Lambda}_R$ , to keep the variance share explained by factors constant. Although the sparse solution is not very different from the dense one, some interpretations change fundamentally. The importance of Factors 1 and 2 is overemphasized for Units 4 and 5, respectively. The correlation between Units 6 and 7 is not captured by the factors any more, although they share a common factor component.

The bias introduced by shrinkage may be empirically relevant. For example, in the exchange rates application in Subsection 5.2, it happens that the Hong Kong dollar was pegged to the US dollar during the entire observation period, which reflects in a correlation coefficient of .999 between the two series. We would therefore expect these two series to be driven by the same factors, where even each factor would explain the same variance share in each series (i.e. equal factor loadings across units). Figure 5 in Frühwirth-Schnatter et al. (2024) conveys a very different interpretation. The Hong Kong dollar (Unit 7) serves as pivot for Factor 4, and is mainly loaded by Factor 4. The US dollar, ranked last, is mainly loaded by Factor 2, and not by Factor 4. Thus, the (interpretation of the) estimate is very difficult to reconcile with this empirical fact.<sup>2</sup>

The conclusion we draw from these considerations is that inducing shrinkage into a model estimated under an imposed GLT structure needs monitoring, to alleviate the issue of introducing a bias into the estimate of latent common or factor-driven correlation across units.

### 3 Extensions for Related Modeling Contexts

The suggested approach to handle sparsity in static factor models with a priori independent factors is an important archetypical model. Sparsity is prominent for data situations with the number of observational units tending to become small relative to the number of parameters involved. In order to become even more fruitful in general, the suggested approach could be extended to several related modeling contexts.

#### 3.1 Correlated or Orthogonal Factors

In particular time series usually share common persistent transitory or permanent dynamics that may be captured by latent dynamic processes, like a vector autoregression in factors leading to dynamic factor models. Usually, factors are assumed conditionally independent (Forni et al., 2000; Stock and Watson, 2002; Nakajima and West, 2013). However, induced sparsity in the factor loading matrix allows relaxing independence towards correlated factors. Combined with informative data, inducing sparsity into the factor loading matrix may recover a correlated factor basis (Beyeler and Kaufmann, 2021). Graphically, in Figure 1 a correlated factor basis would be reflected by non-orthogonal solid and dashed axes. This renders a sparse factor model a lot more flexible.

---

<sup>2</sup>Likewise, it is hardly conceivable that the Australian and New Zealand dollars do not share any common correlation.

Both applications in Frühwirth-Schnatter et al. (2024) fit a static factor model to exchange rate and stock price return series. Considering that common latent persistent dynamics are highly probable, modeling the unconditional distribution of factors, i.e. the static factor representation of a dynamic factor model, calls for correlated factors. We think that the approach would work well for a model specification with correlated factors.

Also factor models assuming orthogonal factors defined on the Stiefel manifold may be of interest, see Villani (2006), Koop et al. (2010), or Aßmann et al. (2024), as they imply orthogonal factor estimators and thus a strict variance decomposition serving as an interesting companion for sparse loading matrices. Combining the suggested approach with an orthogonal factor specification would require adjustments in the steps sampling the loadings and latent factors.

### 3.2 Data Types and Model Uncertainty

Related modeling contexts may also include factor models for binary, ordinal, and categorical dependent variables, and consideration of observed or fixed factor structures, see for example Edwards (2010) or Aßmann et al. (2023). While the latter seems straightforward to be incorporated, a change in scale of the dependent variable can be handled by data augmentation as suggested by Albert (1992) and Albert and Chib (1993). In line with consideration of categorical variables, a comparison with alternative model specifications as those considered in Conti et al. (2014) for continuous and binary dependent variables and dedicated factor structures may be of interest as well. The analysis of sparsity in these related model contexts is connected to the evaluation of whether the ESP prior performs similarly when considering categorical data, given the reduced amount of scaling information provided by such data constellations.

The benefits of the suggested approach may become most visible when embedded in ensembles of different factor models in empirical applications, thereby accounting for model uncertainty as well. While the Bayesian estimation paradigm allows for a conceptually straightforward handling of model uncertainty via model averaging, the ability to provide ensemble based estimation or forecasting is linked to the integration of the considered model specifications into one model space or the aggregation of the evidence from different model specifications. The approach suggested by Frühwirth-Schnatter et al. (2024) achieves this for the class of static factor models via augmenting the parameter space by the number of latent factors and providing a corresponding RJMCMC sampling scheme. Further analysis could investigate whether estimation of the number of latent factors based on rotational invariant quantities is as accurate as the estimator implied by the suggested approach. If not, this would strengthen the argument in favor of RJMCMC algorithms to explore model dimension, as numerically stable routines to calculate the marginal likelihood are currently lacking for some model specifications, like models with orthogonal factors. In cases where the number of latent factors can be accurately determined with other approaches, e.g. via prediction-based approaches or rotational invariant marginal data likelihood, there may be possibilities to construct weighting

schemes to aggregate or evaluate across different types of identifying restrictions, eventually enabling model averaging over confirmatory and shrinkage prior settings as well.

## 4 Conclusion

The paper of Frühwirth-Schnatter et al. (2024) is an excellent addition to the factor model toolbox in applied empirical research. The insights provided on sparsity round up the current literature. The suggested sampling scheme allows for exploring the model space with regard to the number of latent factors and has the potential to provide the basis for several fruitful extensions both in the direction suggested in the article by the authors and the ones discussed herein. Nevertheless, based on considerations exposed in our discussion, the order non-invariance of imposing a (U)GLT to induce rotation identification in the posterior estimate is not a subtle issue and needs to be addressed to recover the sparse structure underlying the factor loadings. Shrinkage as well needs to be monitored to obtain unbiased estimates of factor-determined correlations across the units in the data set.

### Funding

Christian Aßmann gratefully acknowledges funding received from the Deutsche Forschungsgesellschaft (DFG) within priority programme 2431, project number 539621548.

## References

- Albert, J. H. (1992). “Bayesian estimation of normal ogive item response curves using Gibbs sampling.” *Journal of Educational Statistics*, 17(3): 251–269. URL <http://www.jstor.org/stable/1165149?origin=crossref>. 85
- Albert, J. H. and Chib, S. (1993). “Bayesian analysis of binary and polychotomous response data.” *Journal of the American Statistical Association*, 88(422): 669–679. URL <http://www.jstor.org/stable/2290350>. MR1224394. 85
- Aßmann, C., Boysen-Hogrefe, J., and Pape, M. (2016). “Bayesian analysis of static and dynamic factor models: An ex-post approach towards the rotation problem.” *Journal of Econometrics*, 192(1): 190–206. MR3463672. doi: <https://doi.org/10.1016/j.jeconom.2015.10.010>. 81
- Aßmann, C., Boysen-Hogrefe, J., and Pape, M. (2024). “Post-processing for Bayesian analysis of reduced rank regression models with orthonormality restrictions.” *AStA Advances in Statistical Analysis*, 108(3): 577–609. MR4800687. doi: <https://doi.org/10.1007/s10182-023-00489-5>. 85
- Aßmann, C., Gaasch, J.-C., and Stingl, D. (2023). “A Bayesian approach towards missing covariate data in multilevel latent regression models.” *Psychometrika*, 88(4): 1495–1528. MR4668577. doi: <https://doi.org/10.1007/s11336-022-09888-0>. 85
- Beyeler, S. and Kaufmann, S. (2021). “Reduced-form factor augmented VAR – ex-

- exploiting sparsity to include meaningful factors.” *Journal of Applied Econometrics*, 36: 989–1012. MR4362604. doi: <https://doi.org/10.1002/jae.2852>. 84
- Bhattacharya, A. and Dunson, D. B. (2011). “Sparse Bayesian infinite factor models.” *Biometrika*, 98: 291–306. MR2806429. doi: <https://doi.org/10.1093/biomet/asr013>. 83
- Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q., and West, M. (2008). “High-dimensional sparse factor modeling: applications in gene expression genomics.” *Journal of the American Statistical Association*, 103(484): 1438–1456. MR2655722. doi: <https://doi.org/10.1198/016214508000000869>. 81, 83
- 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>. 81
- Conti, G., Frühwirth-Schnatter, S., Heckman, J. J., and Piatek, R. (2014). “Bayesian exploratory factor analysis.” *Journal of Econometrics*, 183(1): 31–57. MR3269916. doi: <https://doi.org/10.1016/j.jeconom.2014.06.008>. 85
- Edwards, M. C. (2010). “A Markov chain Monte Carlo approach to confirmatory item factor analysis.” *Psychometrika*, 75(3): 474–497. MR2719939. doi: <https://doi.org/10.1007/s11336-010-9161-9>. 85
- Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2000). “The generalized dynamic factor model: Identification and estimation.” *Review of Economics & Statistics*, 82: 540–554. MR1867540. doi: <https://doi.org/10.1017/S0266466601176048>. 84
- Frühwirth-Schnatter, S., Hosszejni, D., and Freitas Lopes, H. (2023). “When it counts – econometric identification of the basic factor model based on GLT structures.” *Econometrics*, 11: Article 26. 81
- Frühwirth-Schnatter, S., Hosszejni, D., and Lopes, H. F. (2024). “Sparse Bayesian factor analysis when the number of factors is unknown.” *Bayesian Analysis*, forthcoming. 80, 81, 82, 83, 84, 85, 86
- Hauzenberger, N. and Koop, G. (2024). “Invited Discussion: Sparse Bayesian factor analysis when the number of factors is unknown.” *Bayesian Analysis*, forthcoming. 81
- Kaufmann, S. and Pape, M. (2024). “A geometric approach to factor model identification.” Working Paper 24.06, Study Center Gerzensee. 82
- Kaufmann, S. and Strachan, R. W. (2024). “Dynamic factor models with common (drifting) stochastic trends.” Working Paper 24.02, Study Center Gerzensee. 83
- Koop, G., León-González, R., and Strachan, R. W. (2010). “Efficient posterior simulation for cointegrated models with priors on the cointegration space.” *Econometric Reviews*, 29(2): 224–242. MR2747499. doi: <https://doi.org/10.1080/07474930903382208>. 85

- Legramanti, S., Durante, D., and Dunson, D. (2020). “Bayesian cumulative shrinkage for infinite factorizations.” *Biometrika*, 107: 745–752. MR4138988. doi: <https://doi.org/10.1093/biomet/asaa008>. 83
- Nakajima, J. and West, M. (2013). “Bayesian dynamic factor models: latent threshold approach.” *Journal of Financial Econometrics*, 11: 116–153. 84
- Stock, J. H. and Watson, M. W. (2002). “Macroeconomic forecasting using diffusion indexes.” *Journal of Business & Economic Statistics*, 20: 147–162. MR1963257. doi: <https://doi.org/10.1198/073500102317351921>. 84
- Villani, M. (2006). “Bayesian point estimation of the cointegration space.” *Journal of Econometrics*, 134(2): 645–664. MR2328422. doi: <https://doi.org/10.1016/j.jeconom.2005.07.008>. 85



# Contributed Discussion

Margarita Grushanina\*

## 1 Overview

I congratulate the authors on their outstanding research which led to the development of new approaches to model estimation, selection, and identification in sparse Bayesian factor analysis. These significant contributions, their mathematical foundations and various applications are further illustrated in Hosszejni and Frühwirth-Schnatter (2022), Frühwirth-Schnatter (2023) and Frühwirth-Schnatter et al. (2023).

This approach is a novel and interesting addition to the literature which uses shrinkage process priors to learn factor dimensionality (Bhattacharya and Dunson, 2011; Legramanti et al., 2020; Frühwirth-Schnatter, 2023). The authors combine an exchangeable shrinkage process (ESP) prior (Frühwirth-Schnatter, 2023) with a Dirac-spike-and-slab prior on factor loadings to induce both row and column sparsity. The exploration of the factor dimensionality is performed via split and merge steps in the reversible jump MCMC. Finally, the paper offers an elegant and efficient solution to the problem of identification, which deals with both rotational invariance and variance identification in sparse factor models. The latter is often not given enough attention in the factor analytical literature, which is especially true for the literature involving automatic inference on factor dimensionality and infinite factorisations. In this discussion, I would like to explore how this approach could be applied to the framework of mixtures of factor analysers (MFA).

## 2 MFA framework

MFA represent a class of models where observations are divided into groups (clusters) and each cluster has its own cluster-specific factor model. Recently, MFA models have been developed which allow for automatic inference on varying factor dimensionality in each cluster (Murphy et al., 2020; Grushanina and Frühwirth-Schnatter, 2023). However, with unconstrained cluster-specific factor loading matrices, identification of factors and factor loadings has up to now not been sufficiently addressed in the case of MFA. One of the available options is to use a post-processing procedure, such as Aßmann et al. (2016), Poworoznek et al. (2021), or Papastamoulis and Ntzoufras (2022), among others. An example of this is Murphy et al. (2020) who use Procrustes rotation to achieve identified cluster-specific factor loadings, however, do not explicitly address variance identification in their sparse factor model setting.

If an MFA model assumes dense cluster-specific factor loading matrices and finite number of factors, variance identification can be achieved by imposing the upper limit on

---

\*Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, [m.grushanina@imperial.ac.uk](mailto:m.grushanina@imperial.ac.uk)

the number of factors  $r \leq (m-1)/2$  (Anderson and Rubin, 1956), assuming the notation as in Frühwirth-Schnatter et al. (2024). An example of such approach is Grushanina and Frühwirth-Schnatter (2023), where a finite ESP prior is employed to learn factor dimensionality in dense cluster-specific factor models and the upper limit on the number of factors in each cluster is set equal to  $H = (m-1)/2$ . However, in many applications of MFA models, sparse structure of cluster-specific loading matrices is of significant interest. With this in mind, I think an implementation of a similar algorithm as in Frühwirth-Schnatter et al. (2024) in an MFA setting is worth considering.

A straightforward approach would be to use the same prior on the cluster-specific factor loadings, which would yield for each cluster  $k$  the following prior:

$$\beta_{ij}^k | \tau_j^k \sim (1 - \tau_j^k) \Delta_0 + P_{slab}(\beta_{ij}^k), \quad \tau_j^k | H \sim \mathcal{B}(a_H, b_H), \quad j = 1, \dots, H.$$

However, several issues could potentially arise. One of them concerns computational feasibility. Although the authors propose an efficient MCMC algorithm, the costs of computation are likely to significantly increase if all the steps listed in Algorithm 1 will be performed  $k$  times at each of the  $M$  iterations, where  $k$  indicates the number of clusters. Also, the number of elements in  $\delta_H$ ,  $\mathbf{f}_H$ ,  $\beta_H$ ,  $\Sigma_H$  and  $\tau_H$  will also be multiplied by  $k$ , which may lead to memory problems in the case of high-dimensional data.

Another point to consider is that the partition of data into clusters in an MFA model algorithm is usually performed marginalised with respect to factors. This implies that in a cluster-specific factor model algorithm factors should be updated at the first step to ensure that in the subsequent steps factors derived from the observations in this particular cluster are used. More specifically, this means that step (F) will need to be placed before step (D) in the confirmatory factor analysis (CFA) block of the Algorithm 1 in Frühwirth-Schnatter et al. (2024) to ensure that the subsequent steps, which deal with imposing the unordered generalised lower triangular structure on the sparsity matrix  $\delta_r$ , are conditioned on the correct factors. This should work fine for the CFA block, however, considering that in the exploratory factor analysis block, step (R-F), spurious factors  $\mathbf{f}_{sp}$  are sampled conditional on the active factors  $\mathbf{f}_r$  updated based on the new loadings, I am not entirely sure if such change in structure will not affect the performance of the model.

It might also be worth looking into a simpler approach, for example, by assigning a normal prior to factor loadings and factorising the variance into column-specific and local shrinkage (Bhattacharya and Dunson, 2011; Schiavon et al., 2022). An ESP prior on the column-specific shrinkage parameter will push redundant columns of the factor loading matrix to zero. With a continuous shrinkage prior on the local shrinkage parameter some arbitrary thresholding will need to be applied during post-processing to obtain exact zeroes. However, this would leave identification to be dealt with entirely during post-processing.

## References

- Anderson, T. 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. 90
- 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.” *Journal of Econometrics*, 192: 190–206. MR3463672. doi: <https://doi.org/10.1016/j.jeconom.2015.10.010>. 89
- Bhattacharya, A. and Dunson, D. B. (2011). “Sparse Bayesian infinite factor models.” *Biometrika*, 98(2): 291–306. MR2806429. doi: <https://doi.org/10.1093/biomet/asr013>. 89, 90
- Frühwirth-Schnatter, S. (2023). “Generalised cumulative shrinkage process priors with application to sparse Bayesian factor analysis.” *Philosophical Transactions of the Royal Society A*, 381: 20220148. MR4590506. 89
- Frühwirth-Schnatter, S., Hosszejni, D., and Lopes, H. F. (2023). “When it counts – Econometric identification of factor models based on GLT structures.” *Econometrics*, 11(4): 26. doi: <https://doi.org/10.3390/econometrics11040026>. 89
- Frühwirth-Schnatter, S., Hosszejni, D., and Lopes, H. F. (2024). “Sparse Bayesian factor analysis when the number of factors is unknown.” *Bayesian Analysis*, forthcoming. 90
- 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. MR2265601. 89, 90
- Hosszejni, D. and Frühwirth-Schnatter, S. (2022). “Cover it up! Bipartite graphs uncover identifiability in sparse factor analysis.” arXiv:2211.00671. 89
- Legramanti, S., Durante, D., and Dunson, D. B. (2020). “Bayesian cumulative shrinkage for infinite factorizations.” *Biometrika*, 107(3): 745–752. MR4138988. doi: <https://doi.org/10.1093/biomet/asaa008>. 89
- Murphy, K., Viroli, C., and Gormley, I. (2020). “Infinite mixtures of infinite factor analysers.” *Bayesian Analysis*, 15(3): 937–963. MR4132655. doi: <https://doi.org/10.1214/19-BA1179>. 89
- Papastamoulis, P. and Ntzoufras, I. (2022). “On the identifiability of Bayesian factor analytic models.” *Statistics and Computing*, 32(2): 23. MR4394853. doi: <https://doi.org/10.1007/s11222-022-10084-4>. 89
- Poworoznek, E., Ferrari, F., and Dunson, D. (2021). “Efficiently resolving rotational ambiguity in Bayesian matrix sampling with matching.” arXiv:2107.13783. 89
- Schiavon, L., Canale, A., and Dunson, D. B. (2022). “Generalised infinite factorization models.” *Biometrika*, 109: 817–835. MR4472850. doi: <https://doi.org/10.1093/biomet/asab056>. 90

# Contributed Discussion

Roberto Casarin<sup>\*,†</sup>  and Antonio Peruzzi<sup>\*</sup> 

## 1 Introduction

The techniques suggested in Frühwirth-Schnatter et al. (2024), *FS-H-FL* hereafter, concern sparsity and factor selection and have enormous potential beyond standard factor analysis applications. We show how these techniques can be applied to Latent Space (LS) models for network data. These models suffer from well-known identification issues of the latent factors due to likelihood invariance to factor translation, reflection, and rotation (see Hoff et al., 2002). A set of observables can be instrumental in identifying the latent factors via auxiliary equations (see Liu et al., 2021). These, in turn, share many analogies with the equations used in factor modeling, and we argue that the factor loading restrictions may be beneficial for achieving identification.

## 2 Latent Space Models

Denote with  $W = \{w_{ij}, i, j = 1, \dots, n\}$  the adjacency matrix of a weighted network  $\mathcal{G}$ , where the weights are integer-valued,  $w_{ij} \in \mathbb{N}$ . We assume the following model:

$$w_{ij} \stackrel{\text{ind}}{\sim} \mathcal{Poi}(\theta_{ij}), \quad \theta_{ij} = g(\alpha - \|\mathbf{f}_i - \mathbf{f}_j\|^2),$$

where  $\mathcal{Poi}(\theta)$  denotes the Poisson distribution with intensity  $\theta$ ,  $g(\cdot) : \mathbb{R} \rightarrow \mathbb{R}^+$  is a link function,  $\alpha$  is an intercept parameter,  $\mathbf{f}_i$ ,  $i = 1, \dots, n$  is a collection of  $d$ -dimensional latent factors and  $\|\cdot\|$  denotes the Euclidean norm. To avoid translation issues, one can assume  $\sum_{i=1}^n f_{ik} = 0$  for  $k = 1, \dots, d$ .

The latent factors can be interpreted via a set of node-specific observables  $Y$  with the following interpretation factor model:

$$Y = \Lambda \mathbf{f} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{MN}_{p,n}(O, \Sigma_p, I_n),$$

where  $Y$  is an  $p \times n$  matrix of interpretation variables,  $\mathbf{f} = (\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_n)$  is a  $d \times n$  matrix obtained by stacking the factors,  $\Lambda = (\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \dots, \boldsymbol{\lambda}_d)$  is a  $p \times d$  matrix of loadings with  $\boldsymbol{\lambda}_k = (\lambda_{1k}, \lambda_{2k}, \dots, \lambda_{lk}, \dots, \lambda_{pk})'$  and  $\boldsymbol{\varepsilon}$  is a  $p \times n$  matrix of independent normal error terms with  $\Sigma_p = \text{Diag}(\sigma_1^2, \dots, \sigma_p^2)$ . We are interested in achieving *row sparsity* for  $\Lambda$ . Similarly to *FS-H-FL*, we assume the following prior distributions:

$$\begin{aligned} \alpha &\sim \mathcal{N}(0, \sigma_\alpha^2), \quad \mathbf{f}_i \sim \mathcal{N}_d(\mathbf{0}, (1 - 1/d)^{-1} I_d), \quad \sigma_i^2 \sim \mathcal{IG}(c_0, C_0), \\ \tau_l &\sim \mathcal{Be}(1, 1), \quad \sigma_k^2 \sim \mathcal{IG}(c_\sigma, b_\sigma), \quad \kappa \sim \mathcal{IG}(c_\kappa, b_\kappa), \\ \lambda_{lk} &| \kappa, \sigma_k^2, \tau_l \sim (1 - \tau_l) \delta_0 + \tau_l \mathcal{N}(0, \kappa \sigma_k^2). \end{aligned}$$

<sup>\*</sup>Department of Economics, Ca' Foscari University of Venice, [antonio.peruzzi@unive.it](mailto:antonio.peruzzi@unive.it)

<sup>†</sup>Venice Centre for Economic and Risk Analytics (VERA), [r.casarin@unive.it](mailto:r.casarin@unive.it)

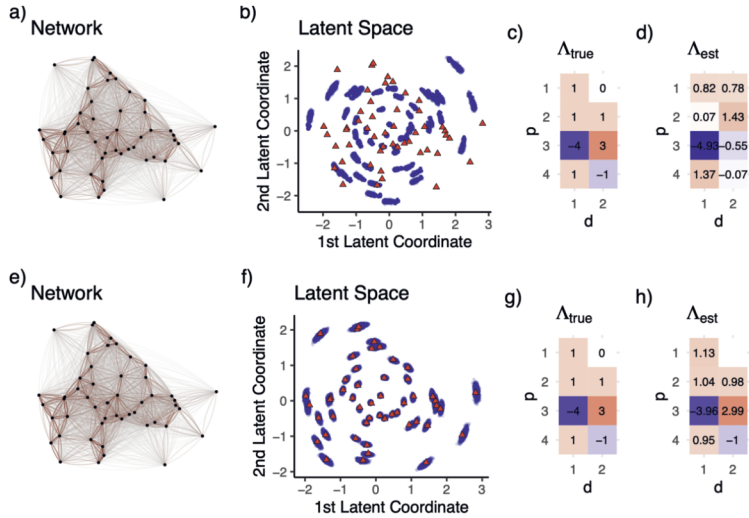


Figure 1: Results for an LS model without and with restrictions (top and bottom, respectively). Panel a) and e) report the observed network width edge gradient proportional to the absolute distance between the observed and predicted weight (darker edge colors). Panel b) and f) report the posterior draws (blue dots) against the true latent coordinates (red triangles). The true value of  $\Lambda$  is in Panel c) and g). Panel d) and h) report the posterior means of  $\Lambda$  without and with PLT restrictions, respectively.

Figure 1 presents the posterior results for an LS model with  $d = 2$  and  $p = 4$  for the unrestricted and restricted  $\Lambda$  (top and bottom panels, respectively). Panel b) shows the identification issue, and Panel f) the effectiveness of the restrictions on  $\Lambda$  to achieve identification of the set of latent factors  $\mathbf{f}$ . The factor identification is obtained via PLT restriction, i.e.  $\lambda_{kk} > 0$  and  $\lambda_{lk} = 0$  for  $k > l$ . As discussed in *FS-H-FL*, the PLT structure may be too restrictive. Therefore, we speculate on imposing an ordered or unordered GLT structure on  $\Lambda$ .

### 3 Conclusion

As further research, we suggest extending the authors' approach to nonlinear factor models. This is a stimulating work, and we are therefore very pleased to be able to propose the vote of thanks to the authors for their contribution.

#### Funding

This discussion was supported by the EU - NextGenerationEU, in the framework of the GRINS - Growing Resilient, INclusive and Sustainable project (GRINS PE00000018 - CUP H73C22000930001), National Recovery and Resilience Plan (NRRP) - PE9 - Mission 4, C2, Intervention 1.3. The views and opinions expressed are solely those of the authors and do not necessarily reflect those of the EU, nor can the EU be held responsible for them.

## References

- Frühwirth-Schnatter, S., Hosszejni, D., and Lopes, H. F. (2024). “Sparse Bayesian factor analysis when the number of factors is unknown.” *Bayesian Analysis*, 1–31. 92
- Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002). “Latent space approaches to social network analysis.” *Journal of the American Statistical Association*, 97(460): 1090–1098. MR1951262. doi: <https://doi.org/10.1198/016214502388618906>. 92
- Liu, H., Jin, I. H., Zhang, Z., and Yuan, Y. (2021). “Social network mediation analysis: a latent space approach.” *Psychometrika*, 86: 272–298. MR4242643. doi: <https://doi.org/10.1007/s11336-020-09736-z>. 92

## Contributed Discussion

Marta Catalano<sup>\*</sup>, Beatrice Franzolini<sup>†</sup>, Matteo Giordano<sup>‡</sup>, and Giovanni Rebaudo<sup>‡</sup>

We congratulate the authors for their contribution to the field of Bayesian factor analysis, which develops an appealing methodology for sparse recovery. We believe that their approach may become a useful tool in applications for statisticians and practitioners, thanks to its ability to simultaneously perform inference on identifiable sparse factor loadings and achieve data-driven model selection in terms of the number of factors.

This work also stimulates a wealth of interesting follow up questions, both on the specifics of the model at hand and also drawing from the broader contemporary Bayesian literature. In this discussion we expand upon some of these questions, with specific reference to multiple testing, dependence among slab probabilities, continuous alternatives to spike-and-slab priors, and the learning of the number of factors.

**Multiplicity and Multiple Testing** The sparse Bayesian factor model with UGLT structures estimates all parameters simultaneously. Nonetheless, the estimation process can be viewed as addressing two inferential challenges: first, a multiple testing problem, where each entry of the factor loading matrix is tested to determine whether it is active or not; and second, the estimation of the values of the active entries. In the context of multiple testing, the Bayesian framework naturally accommodates two desirable yet distinct types of penalties. The first is an Ockham’s razor penalty, typically arising from the use of marginal likelihoods, which penalizes models with more active loadings in factor analysis, thus *promoting sparsity*. The second is a multiplicity penalty, thanks to which the posterior inclusion probabilities for each factor loading decrease as the dimensions  $m$  or  $H$  increase, leading to a framework of *adaptive sparsity*. Multiplicity penalties relate to the frequentist challenge of multiple testing and generally provide better control of noisy signals and false discoveries by adopting a more cautious approach (Scott and Berger, 2010). It would be valuable to assess the presence and potential influence of the multiplicity penalty in the Bayesian model with UGLT structures. Specifically, while the prior on  $\tau_j$  seems to induce a multiplicity penalty in  $H$  (i.e., the number of columns), it would be worth exploring whether a similar penalty also applies to  $m$  (i.e., the number of rows) and whether such penalties are preserved after the post-processing procedure. Additionally, it would be interesting to investigate how these penalties affect the model’s recovery performance, for instance, by examining changes in the ROC curve.

**Individual Components in the Slab Probability** The prior for the factor loadings  $(\beta_{i,j})_{i,j}$  in (3.1) requires a multivariate spike and slab prior. The dependence between the marginals  $\beta_{i,j}$  is induced by column-specific slab probabilities  $\tau_j$  and a multivariate

---

<sup>\*</sup>Luis University, Italy, [mcatalano@luiss.it](mailto:mcatalano@luiss.it)

<sup>†</sup>Bocconi University, Italy, [beatrice.franzolini@unibocconi.it](mailto:beatrice.franzolini@unibocconi.it)

<sup>‡</sup>University of Torino, Italy, [matteo.giordano@unito.it](mailto:matteo.giordano@unito.it); [giovanni.rebaudo@unito.it](mailto:giovanni.rebaudo@unito.it)

hierarchical slab distribution with row-specific dependence, which shrinks each component towards zero. However, to achieve additional shrinkage for individual factor loadings, in (3.12) the authors propose to add an individual shrinking factor in the multivariate hierarchical slab distribution. Could a similar result be achieved by allowing for individual components in the slab probabilities? This would bring to the need for multivariate versions of the exchangeable or cumulative shrinkage process, which could possibly converge to a multivariate version of the Indian buffet process, e.g., the hierarchical Indian buffet process (Thibaux and Jordan, 2007; James et al., 2024).

### **A Role for Other Bayesian Approaches to Sparsity and Model Selection in Bayesian Factor Analysis?**

Implementing posterior-based inference with spike-and-slab priors is a notoriously challenging task, due to the underlying combinatorial problem of exploring the space containing all possible sparsity patterns. One of the key contributions of the present paper is the construction of an ad-hoc MCMC sampler (cf. Algorithm 1), which cleverly alternates between the two formulations, Exploratory and Confirmatory, of the Factor Analysis model. On the other hand, the construction of Bayesian models for sparse structures and variable selection is an issue of general interest in the broader Bayesian literature, where some continuous (and easier to work with) alternatives to the spike-and-slab prior have been shown, either theoretically or empirically, to be potential effective approaches. For example, Laplace priors are known to possess desirable sparsity-promoting properties at the level of the maximum-a-posterior estimators (Agapiou et al., 2018) and here could be deployed either directly onto the factor loadings  $\beta_{ij}$ , or column-wise for the matrix  $\beta_H$ . Excitingly, hierarchical Gaussian priors have recently been shown to possess variable selection properties when endowed with a horseshoe hyper-prior on the length-scale without the need for the additional spike-and-slab structure (Castillo and Randrianarisoa, 2024). Some further possibilities include the horseshoe priors themselves (and extensions thereof), Dirichlet–Laplace and the so-called R2-D2 priors, see Hirsh et al. (2022). These developments suggest the questions as to whether such sparse (or approximately sparse) continuous priors models could also be employed with success in Bayesian factor analysis.

**Learning the Number of Factors** Learning the number of factors simultaneously with estimating the factors is a challenging problem, which the authors address effectively in their proposal. Specifically, by imposing a UCLT structure, they facilitate the joint identification of the unknown number of factors  $r$  and the underlying factor model parameters  $\Lambda$  and  $\Sigma_0$  from the overfitting BFA model. Inference on the number of factors is validated through a simulation study, where the authors empirically demonstrate the model’s ability to recover the true number of factors,  $r_{\text{true}}$ . These findings give hope that the model can consistently estimate the true number of factors under a well-specified data-generating process. Establishing such consistency for the discrete parameter  $r_{\text{true}}$  under an identifiable and well-specified model might be achievable by leveraging Doob’s Theorem (Doob, 1949). In particular, demonstrating consistency for the true number of factors  $r_{\text{true}} (< H)$  in a subset of the possible values of  $r$  of prior probability one would imply consistency across all possible values of  $r$ , given that the prior distribution assigns positive probability to each possible value. This strategy has been successfully



employed, for instance, in proving the consistency of the number of mixture components in Bayesian finite mixture models with a prior on the number of components (Nobile, 1994; Miller, 2023).

## References

- Agapiou, S., Burger, M., Dashti, M., and Helin, T. (2018). “Sparsity-promoting and edge-preserving maximum a posteriori estimators in non-parametric Bayesian inverse problems.” *Inverse Problems*, 34: 1–37. MR3774703. doi: <https://doi.org/10.1088/1361-6420/aaacac>. 96
- Castillo, I. and Randrianarisoa, T. (2024). “Deep horseshoe Gaussian processes.” *Preprint at arXiv:2403.01737*. 96
- Doob, J. L. (1949). “Application of the theory of martingales.” In *Le Calcul des Probabilités et ses Applications*, volume 13, 23–27. MR0033460. 96
- Hirsh, S. M., Barajas-Solano, D. A., and Kutz, J. N. (2022). “Sparsifying priors for Bayesian uncertainty quantification in model discovery.” *Royal Society Open Science*, 9: 1–20. 96
- James, L. F., Lee, J., and Pandey, A. (2024). “Bayesian analysis of generalized hierarchical Indian buffet processes for within and across group sharing of latent features.” *Preprint at arXiv:2304.05244*. 96
- Miller, J. W. (2023). “Consistency of mixture models with a prior on the number of components.” *Dependence Modeling*, 11: 1–9. MR4557090. doi: <https://doi.org/10.1515/demo-2022-0150>. 97
- Nobile, A. (1994). “Bayesian analysis of finite mixture distributions.” Ph.D. thesis, Carnegie Mellon Univ. MR2692049. 97
- Scott, J. G. and Berger, J. O. (2010). “Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem.” *Annals of Statistics*, 38: 2587–2619. MR2722450. doi: <https://doi.org/10.1214/10-AOS792>. 95
- Thibaux, R. and Jordan, M. I. (2007). “Hierarchical beta processes and 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*, 564–571. San Juan, Puerto Rico: PMLR. 96

# Invited Discussion

Jamie L. Cross<sup>\*</sup>, Lennart Hoogerheide<sup>†,‡</sup>, and Herman K. van Dijk<sup>‡,§</sup>

## 1 Introduction

In their interesting and mathematically elegant paper Frühwirth-Schnatter et al. (2024) discuss the issue of determining a plausible number of latent components in a sparse factor model (technically: determining the rank of the factor space) which is a nontrivial issue in case of weak data, sparse model restrictions and diffuse prior information. In this context the authors focus on the connection between the theoretical issue of parametric identification restrictions including informative prior information and the operational issue of Bayesian Markov chain Monte Carlo estimation. Specifically, the authors achieve identification and inference which is independent from the ordering of the dependent variables by making use of the concept of Unordered Generalized Lower Triangular (UGLT) structure and for estimation they introduce a novel Markov chain Monte Carlo procedure which makes use of a reversible jump sampler. All this in order to learn about a plausible number of latent factors with substantial posterior probability.

We introduce two contributions to this research. We start to discuss the issue of identification restrictions within the authors' framework of a static factor model and present an operational alternative to the UGLT structure by introducing *orthogonal* parameter restrictions. Second, we propose the use of predictive likelihoods in combination with moving window estimation in order to determine a plausible time-varying number of factor model components. Our motivation stems from the observation that financial and economic relations vary over time. One of these time-varying relations is the increase in the correlations between equities during market downturns, that is, during equity market downturns fewer latent factors are assumed to be able to explain the same amount of variation in equity returns. We present empirical results on how a residual momentum strategy based on a time-varying latent factor model outperforms a standard momentum strategy using a portfolio of industrial stocks. This strategy has been popular among investors over a long time.

## 2 Identification Restrictions in Factor Models

For expository purposes we start with a basic multivariate regression model:

$$\mathbf{y}'_t = \mathbf{x}'_t \mathbf{B}' + \epsilon'_t, \quad \epsilon_t \sim \mathcal{NID}(\mathbf{0}, \Sigma), \quad (1)$$

where  $\mathbf{y}_t$  is an  $m$ -vector,  $\mathbf{x}_t$  is an  $r$ -vector and  $\mathbf{B}$  an  $m \times r$  matrix. It is well-known that a Bayesian analysis of this model using diffuse priors leads to a marginal posterior

---

<sup>\*</sup>Melbourne Business School, University of Melbourne, [j.cross@mbs.edu](mailto:j.cross@mbs.edu)

<sup>†</sup>Department of Econometrics, VU University Amsterdam, [l.f.hoogerheide@vu.nl](mailto:l.f.hoogerheide@vu.nl)

<sup>‡</sup>Tinbergen Institute

<sup>§</sup>Erasmus University Rotterdam and Norges Bank, [hkvandijk@ese.eur.nl](mailto:hkvandijk@ese.eur.nl)

of  $\mathbf{B}$  that is bell-shaped and belongs to the class of matrix Student-t distributions. Determining a plausible number of explanatory variables is a standard topic in an introductory Bayesian course. The connection between model structure and estimation is direct: analytical as well as simulation methods are used.

Next, consider a static factor model and adjust formula (1.1) of Frühwirth-Schnatter et al. (2024) as:

$$\mathbf{y}'_t = \mathbf{f}'_t \mathbf{\Lambda} + \boldsymbol{\epsilon}'_t, \quad \mathbf{f}_t \sim \mathcal{NID}(\mathbf{0}, \mathbf{I}_r), \quad \boldsymbol{\epsilon}_t \sim \mathcal{NID}(\mathbf{0}, \boldsymbol{\Sigma}_0), \quad \boldsymbol{\Sigma}_0 = \text{Diag}(\sigma_1^2, \dots, \sigma_m^2), \quad (2)$$

where  $\mathbf{f}_t$  is an  $r$ -vector and  $\mathbf{\Lambda}$  an  $r \times m$  matrix. The diagonal covariance matrix assumption with respect to the disturbances  $\boldsymbol{\epsilon}_t$  implies that all cross-sectional correlation is captured by the factors  $\mathbf{f}_t$ , in addition,  $\text{cor}(\mathbf{f}_t, \boldsymbol{\epsilon}_s) = 0$  for  $\forall s, t$ . The vector of observations  $\mathbf{x}_t$  is replaced by a vector of *unobserved* random factors  $\mathbf{f}_t$  and the matrix of coefficients  $\mathbf{B}'$  by a matrix  $\mathbf{\Lambda}$ , labeled factor loadings.

Let  $\mathbf{F}$  be the  $T \times r$  matrix of factors. The identification problem of  $\mathbf{F}$  and  $\mathbf{\Lambda}$  can be seen from the equality  $\mathbf{F}\mathbf{\Lambda} = \mathbf{F}\mathbf{R}\mathbf{R}^{-1}\mathbf{\Lambda}$  for an  $r \times r$  invertible (or invertible rotation) matrix  $\mathbf{R}$ , which has  $r^2$  free parameters. Hence, at least  $r^2$  parameter restrictions are needed for the model to be identified. The identity covariance matrix of the  $\mathbf{f}_t$  imposes  $\frac{r(r+1)}{2}$  restrictions, so an additional  $\frac{r(r-1)}{2}$  restrictions are required for identification. The transformed and/or rotated factors and loadings still provide the same likelihood value.

We note that the key feature of factor models is that the information in the  $m$  economic variables of interest  $\mathbf{y}_t$  can be compressed to a much lower number of  $r$  unobserved random factors  $\mathbf{f}_t$ . Given model and data, we intend to have this information dominate prior information. However, given the present likelihood of the model with a diffuse prior it is clear that there does not exist an operational estimation procedure for structural inference using Bayesian MCMC. Of course, structural identification is not a necessary condition for forecasting, see Geweke (2007).

Next, consider the static factor model with a triangular normalization on  $\mathbf{\Lambda}$ , given as:

$$\mathbf{\Lambda} = \left( \mathbf{\Lambda}_1^{(r \times r)} \quad \mathbf{\Lambda}_2^{(r \times (m-r))} \right), \quad \mathbf{\Lambda}_1 = \begin{pmatrix} \lambda_{11} & 0 & \cdots & 0 \\ \lambda_{21} & \lambda_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{r1} & \lambda_{r2} & \cdots & \lambda_{rr} \end{pmatrix}, \quad (3)$$

where  $\mathbf{\Lambda}_2$  is unrestricted. The triangular normalization on  $\mathbf{\Lambda}_1$  provides  $\frac{r(r-1)}{2}$  restrictions. Together with the restrictions on the covariance of the  $\mathbf{f}_t$  this gives parametric identification. Combining a diffuse prior with the likelihood yields a posterior which is unbounded (for  $\mathbf{F}$  tending to 0), but integrable.<sup>1</sup> Given the posterior structure and given an *a priori* fixed number of factors  $r$  the corresponding MCMC method is a basic Gibbs sampler. However, the important disadvantage is that inference depends on the ordering of the dependent variables.

<sup>1</sup>See Bastürk et al. (2017), Section 3.3 for proofs.

As mentioned, Frühwirth-Schnatter et al. (2024) achieve inference which is independent from the ordering of the dependent variables by making use of Unordered Generalized Lower Triangular structure. We propose to obtain this independence by making use of orthogonal normalization on the parameters of the model. The orthogonal normalization implies that in this case no preferred ordering of the variables is imposed and, conditionally upon a largest singular value, the region of integration of the factors and factor loadings is bounded. That is, the parametrization  $\mathbf{F}\mathbf{\Lambda}$  can be linked to the singular value decomposition  $\mathbf{F}\mathbf{\Lambda} = \mathbf{U}\mathbf{K}\mathbf{V}$ , where the rectangular  $T \times r$  matrix  $\mathbf{U}$  is an element of the Stiefel manifold  $\mathbf{U}'\mathbf{U} = \mathbf{I}_r$  and the  $r \times m$  matrix  $\mathbf{V}$  is an element of the Stiefel manifold  $\mathbf{V}'\mathbf{V} = \mathbf{I}_m$ .  $\mathbf{K}$  is a diagonal  $r \times r$  matrix with positive diagonal entries equal to the singular values of  $\mathbf{F}\mathbf{\Lambda}$ , denoted by  $\kappa = (\kappa_1, \dots, \kappa_r)'$ . The manifolds on which  $\mathbf{U}$  and  $\mathbf{V}$  are defined have finite volume conditionally upon a largest singular value and the region of integration of the factors and factor loadings is then bounded.

In order to achieve this we propose an approach that directly uses the structure of the singular value decomposition and makes use of a lasso type shrinkage prior for regularization, see Tibshirani (1996). As it is specified above, the singular value decomposition is not uniquely defined. Any simultaneous permutation of the columns of  $\mathbf{U}$ ,  $\mathbf{K}$  and  $\mathbf{V}$  also constitutes a singular value decomposition. A common way to avoid this ambiguity is by ordering the singular values that occur on the diagonal of  $\mathbf{K}$  as  $\kappa_1 \geq \kappa_2 \geq \dots \geq \kappa_r \geq 0$ , which is more straightforward than devising an ordering of the columns of  $\mathbf{F}$  and  $\mathbf{\Lambda}$ . Because of this ordering each element  $\kappa_{i+1}$  for  $i = 1, \dots, r - 1$  is bounded by  $\kappa_i$ . Only  $\kappa_1$  remains unbounded towards  $+\infty$ . Integrability is thus determined by the behavior of  $\kappa_1$ .<sup>2</sup>

A natural choice for  $\kappa_1$  that is consistent with the uniform prior on the simplex for  $\kappa_2, \dots, \kappa_r | \kappa_1$  is the exponential distribution. Conditional on  $\kappa_1$ , all model parameters (except the covariance matrix  $\Sigma$ ) are bounded to finite areas.

We conclude that given the standard form of a static factor model and using a lasso type shrinkage prior under orthogonal normalization on the parameters of the matrix with reduced rank, the marginal posteriors of these parameters are proper with finite first and higher order moments and inference is independent of the ordering of the dependent variables. For a survey on alternative identification restrictions with corresponding MCMC algorithms, see Koop et al. (2006).

### 3 Learning on a Plausible Number of Factor Model Components Using Predictive Likelihoods

We propose a ‘predictive likelihood’ approach to assess the number of factors in a factor model. The basic idea of this approach is to split the data into two parts: a training set and a ‘hold-out’ set of data. In the first part a possible weak prior is transformed to an

---

<sup>2</sup>If a singular value occurs more than once, then the columns of  $U$  and  $V$  corresponding to these singular values are not uniquely defined. Any other orthonormal basis that spans the same space will also do. Although the transformation between the matrix  $\mathbf{F}\mathbf{\Lambda}$  and its singular value decomposition  $(U, K, V)$  is still not invertible everywhere, this is an event with zero measure.

informative posterior which serves as a prior for the second part of the data. This also refrains from the Bartlett (1957) paradox occurring under diffuse priors.<sup>3</sup> The gain in the use of ‘prior data points’ is to obtain predictive likelihoods for the computation of reliable predictive model probabilities. It is important to note that predictive likelihoods evaluated at different times (using moving estimation windows) provide time-varying model probabilities. Any policy based on these model weights will therefore be time-varying as well.

Let  $M_r$  denote the factor model with  $r \ll m$  factors. A predictive likelihood for model  $M_r$  is computed by splitting the dataset as follows:

$$Y = \begin{pmatrix} Y_{t_0:t_1} \\ Y_{t_1+1:t_2} \end{pmatrix} = \begin{pmatrix} Y^* \\ \tilde{Y} \end{pmatrix}, \quad (4)$$

where observations from  $t_0$  to  $t_1$  are defined as the ‘training sample’ and observations from  $t_1 + 1$  to  $t_2$  are defined as the ‘hold-out sample’. The predictive likelihood for the hold-out sample is then defined as:

$$p(\tilde{Y}|Y^*, M_r) = \frac{p(\tilde{Y}, Y^*|M_r)}{p(Y^*|M_r)} = \frac{p(Y|M_r)}{p(Y^*|M_r)}. \quad (5)$$

Choosing the size of the hold-out samples is important. If the hold-out sample is very small, the qualities of the models may be hard to distinguish (with almost equal model probabilities), and the results may be sensitive to just a few hold-out observations. If the hold-out sample is very large, the results may be sensitive to the few observations in the small training sample, and the Bartlett (1957) paradox may imply that we choose a model with too small number of factors  $r$ . Naturally, a robustness check should be performed to see the effect of training and hold-out sample size selection.

A simple method to estimate model probabilities is the harmonic mean estimator (Newton and Raftery, 1994; Ardia et al., 2012), which has the advantage that it is easily estimated using a set of draws generated from the posterior distribution of parameters  $\theta_r$  of model  $M_r$ . The computational steps are as follows. Calculate two marginal likelihoods for each model  $M_r$ , a marginal likelihood for the whole sample and the second for the training sample. The full sample marginal likelihood is given as:

$$p(Y|M_r) = \int_{\theta_r} p(Y|\theta_r, M_r)p(\theta_r|M_r)d\theta_r \approx \left( \frac{1}{N} \sum_{i=1}^N p(Y|\theta_r^{f,i}, M_r)^{-1} \right)^{-1}$$

with posterior draws  $\theta_r^{f,i}$  ( $i = 1, \dots, N$ ) based on the full data sample. The training sample marginal likelihood is given as:

$$p(Y^*|M_r) = \int_{\theta_r} p(Y^*|\theta_r, M_r)p(\theta_r|M_r)d\theta_r \approx \left( \frac{1}{N} \sum_{i=1}^N p(Y^*|\theta_r^{*,i}, M_r)^{-1} \right)^{-1}$$

---

<sup>3</sup>Another approach is to construct a so-called imaginary sample by introducing a set of dummy observations. It yields a pragmatic class of priors proposed by Christopher Sims (Sims, 2005).

with posterior draws  $\theta_r^{*,i}$  ( $i = 1, \dots, N$ ) based on the training sample. Next calculate predictive likelihoods for each model  $M_r$  using (5) as:

$$p(\tilde{Y}|Y^*, M_r) = \frac{p(Y|M_r)}{p(Y^*|M_r)} \approx \frac{\sum_{i=1}^N p(Y^*|\theta_r^{*,i}, M_r)^{-1}}{\sum_{i=1}^N p(Y|\theta_r^{f,i}, M_r)^{-1}}. \quad (6)$$

From the predictive likelihoods for each model compute model probabilities for  $M_r$  for  $r \in 1, \dots, m-1$ :

$$p(M_r|Y) = \frac{p(\tilde{Y}|Y^*, M_r) \times p(M_r)}{\sum_{r'=1}^{m-1} p(\tilde{Y}|Y^*, M_{r'}) \times p(M_{r'})},$$

where  $p(M_r)$  is the prior model probability. An uninformative prior, such as  $p(M_r) = \frac{1}{m-1}$  is easy to use in this setting. Based on the predictive likelihood calculation from a rolling window of predictive likelihoods, an optimal model  $M^*$  can be chosen or Bayesian Model Averaging can be applied.

**Simulated Data Experiment** For illustrative purposes we apply the predictive likelihood approach for the factor model to simulated data. We consider simulated datasets with  $T = 100$  and  $T = 250$  observations. In order to see the effect of number of dependent variables  $m$  and number of factors  $r$  on the predictive likelihood methodology, we consider  $r = 1, 2$  common factors for  $m = 2, 4, 10, 20$  data series. For each simulation experiment, we apply the predictive likelihood approach with different sizes of training samples, consisting of 5%, 10%, 20% and 50% of observations. We replicate each simulation experiment 100 times.

Table 1 presents the posterior probabilities from all simulation experiments, where we report the posterior model probabilities for different number of factors averaged over 100 simulation experiments for each simulation setting. Posterior results are based on 4000 posterior draws, where the first 2000 draws are burn-in draws.

The results in Table 1 indicate that the highest probabilities (indicated by bold-face entries) for each simulation experiment, indicated in rows, correspond to the true number of factors. In most simulation studies, the posterior probability is very close to 1 for the correct model specification. Hence the predictive likelihoods provide a clear choice of models. Comparing the bottom panel of Table 1 with the other panels, we conclude that the predictive likelihood approach with a smaller training sample than 50% provides more clear indications of the correct number of factors, with posterior probabilities being closer to 1 compared to the same simulation setting but a larger training sample (50%). Thus, the length of the training sample should not be chosen too large compared to the total length of the sample and a sensitivity analysis with respect to the length of the training sample will give more confidence in the results.

Figure 1 presents the details of predictive likelihoods for two sets of simulated data and for each simulation. These data correspond to  $T = 100$ ,  $p = 2$ ,  $r = 1$  on the left panel of Figure 1 and  $T = 100$ ,  $p = 10$ ,  $r = 1$  on the right panel of Figure 1. For both simulation specifications, the correct number of factors  $r = 1$ , shown by the red lines in the figure, has the highest posterior probability in almost all simulation replications. We therefore conclude that the predictive likelihood approach accurately detects the number of factors, even with a small sample size.

5% training sample									
$T$	$m$	$r$	$\text{pr}(r = 1)$	$\text{pr}(r = 2)$	$\text{pr}(r = 3)$	$\text{pr}(r = 4)$	$\text{pr}(r = 5)$	$\text{pr}(r = 6)$	
100	2	1	<b>1.00</b>	0.00	–	–	–	–	
100	10	1	<b>0.96</b>	0.02	0.00	0.02	0.00	0.00	
250	4	2	0.00	<b>1.00</b>	0.00	0.00	–	–	
250	20	2	0.00	<b>1.00</b>	0.00	0.00	0.00	0.00	
10% training sample									
$T$	$m$	$r$	$\text{pr}(r = 1)$	$\text{pr}(r = 2)$	$\text{pr}(r = 3)$	$\text{pr}(r = 4)$	$\text{pr}(r = 5)$	$\text{pr}(r = 6)$	
100	2	1	<b>1.00</b>	0.00	–	–	–	–	
100	10	1	<b>0.89</b>	0.07	0.00	0.03	0.01	0.00	
250	4	2	0.00	<b>1.00</b>	0.00	0.00	–	–	
250	20	2	0.00	<b>1.00</b>	0.00	0.00	0.00	0.00	
20% training sample									
$T$	$m$	$r$	$\text{pr}(r = 1)$	$\text{pr}(r = 2)$	$\text{pr}(r = 3)$	$\text{pr}(r = 4)$	$\text{pr}(r = 5)$	$\text{pr}(r = 6)$	
100	2	1	<b>1.00</b>	0.00	–	–	–	–	
100	10	1	<b>0.77</b>	0.14	0.05	0.02	0.01	0.00	
250	4	2	0.01	<b>0.98</b>	0.01	0.00	–	–	
250	20	2	0.00	<b>0.95</b>	0.01	0.00	0.03	0.01	
50% training sample									
$T$	$m$	$r$	$\text{pr}(r = 1)$	$\text{pr}(r = 2)$	$\text{pr}(r = 3)$	$\text{pr}(r = 4)$	$\text{pr}(r = 5)$	$\text{pr}(r = 6)$	
100	2	1	<b>1.00</b>	0.00	–	–	–	–	
100	10	1	<b>0.34</b>	0.13	0.02	0.08	0.18	0.24	
250	4	2	0.00	<b>0.95</b>	0.05	0.00	–	–	
250	20	2	0.00	<b>0.88</b>	0.02	0.02	0.01	0.07	

Table 1: Average posterior probabilities from 100 simulation replications with  $T$  observations,  $m$  variables and  $r$  factors. Highest probabilities are indicated by **boldface** table entries.

**Equity Momentum at Work Using a Time-Varying Latent Factor Model** Financial momentum strategies are based on the expectation that past stock winners will continue to be winners and past stock losers will continue to be losers. Standard equity momentum strategy ranks stocks on their recent returns, skips a short period to overcome short term reversals and then buys stocks in the top of the ranking and short-sells stocks in the bottom of the ranking. We emphasize that standard momentum’s risk and return vary over time.

We compare the performance of residual momentum strategies based on the *residual returns* (returns in excess of what is to be expected based on the factors and factor loadings) with a standard momentum strategy. We use monthly return data on ten industry portfolios between 1960M7 and 2015M6, shown in Figure 2. The ten industries are labeled as ‘non-durables’, ‘durables’, ‘manufacturing’, ‘energy’, ‘hi-tech’, ‘telecom’, ‘shops’, ‘health’, ‘utilities’ and the final category ‘others’. This residual industry momentum strategy is a combination of residual momentum (Blitz et al., 2011) and industry momentum (Moskowitz and Grinblatt, 1999).

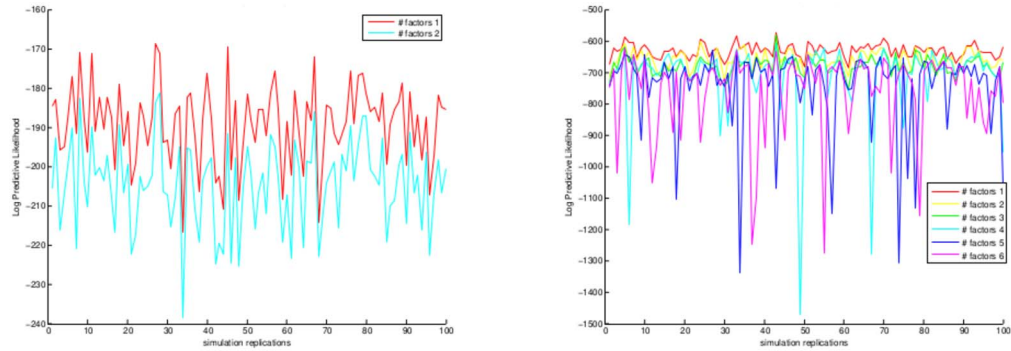


Figure 1: Log-predictive likelihoods for different number of factors for two sets of simulated data with  $T = 100$  observations. The left panel corresponds to  $m = 2$  series and  $r = 1$  common factor. The right panel corresponds to  $m = 10$  series and  $r = 1$  common factor. Each simulation experiment is repeated 100 times, as shown in the x-axes. Predictive likelihoods are calculated using 10 percent of the sample as the training sample.

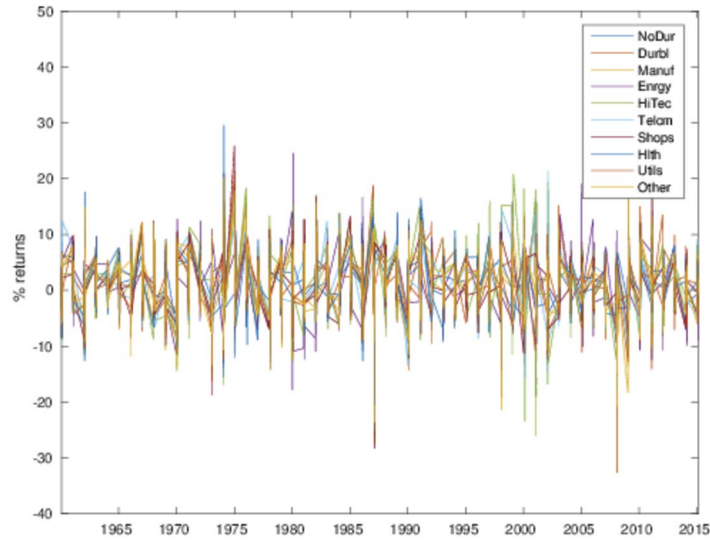


Figure 2: Monthly percentage returns for ten industry portfolios.

Results are presented in Table 2. In order to allow for changes in the number of factors, we apply the moving window estimation method as follows. Within one estimation period, the static factor model is estimated for  $r = 2, \dots, 5$  factors, and the optimal number of factors is chosen based on the predictive likelihoods. The moving window estimation is based on a sample of  $T = 240$  observations. We consider two



	Standard	Latent Factor Models	
		10%	20%
mean	-1.13	2.82	3.75
volatility	21.54	20.16	12.56
Sharpe ratio	-0.05	0.14	0.30
Largest loss	-68.52	-68.52	-20.95
Max. drawdown	135.3	68.52	35.19
Max. recovery	15	9	9

Table 2: Risk and return characteristics of standard momentum compared with momentum from time-varying factors.

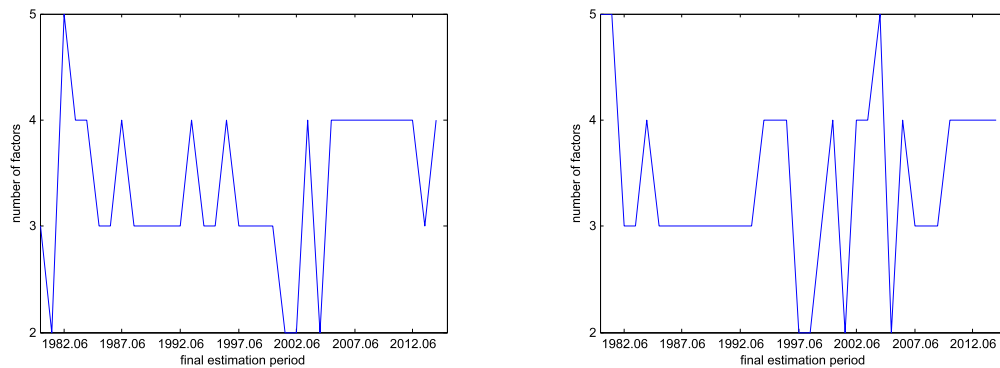


Figure 3: Optimal number of factors for 10 industry portfolios. The figure presents the number of factors with highest predictive likelihood at each estimation window for two training sample choices (10% or 20% of the estimation window). X-axes correspond to the final month in the estimation window.

cases for the predictive likelihood calculation with a ‘small training sample’ and a ‘large training sample’, consisting of 10% and 20% of the full moving window sample, respectively. We first analyze the evolution of the number of factors through the moving windows for the two training sample percentages. Figure 3 presents the number of factors with the highest predictive likelihood for each estimation sample under two training sample choices, where 10% and 20% of the total sample is used as training samples.

Figure 3 shows that the model with  $r = 3$  factors is the most frequently chosen model using both training samples. Despite this high frequency, the optimal number of factors according to predictive likelihoods changes substantially over time. The obtained number of factors in the left and right panels of Figure 3 are different, in particular, at the end of the sample period. On the one hand, this difference indicates that the training sample choice should be made with care in order to find the appropriate number of factors. On the other hand, this variation in the number of factors may influence the gains from a trading strategy, like momentum. For the latter reason, we next report the gains from trading strategies using both training samples.

One would expect that the number of optimal factors varies with the performance of equity markets, in particular fewer factors are present in the model during market declines. We have two major market declines in our sample: 2000–2002 and 2008 where equity markets lost 56 and 38 percent, respectively. We indeed find that during the equity market losses in 2000 to 2002 the optimal number of factors was 2. For the 2008 crash we do not find a smaller number of optimal factors. This needs to be explored in further research.

Next, we report the performance of a standard momentum strategy and compare it with a residual momentum strategy based on the factor model. The standard strategy is a benchmark and by definition does not depend on an underlying model. The ten industry portfolios are sorted on their mean (residual) returns in the last 12 months. The strategy is long in the best industry and short in the worst industry, where this position is held for 12 months. The first investment month is July 1980, as we require 240 months since 1960 for the first estimation of the model parameters.

In Table 2, we report the following risk and return measures for the returns of each strategy: mean return, volatility, Sharpe ratio, largest loss encountered, maximum drawdown, all measured in percentages, and maximum recovery period measured in months. These values are based on the realized returns of each investment strategy.

The standard industry momentum strategy does not yield positive average returns; its average annual return is minus 1.1 percent. The 20 percent Bayesian Factor Model scores better in all six criteria compared to the standard industry momentum strategy.

We conclude that a Bayesian latent factor model with a time-varying number of factors, moving window estimation and a training sample of 20 percent is able to outperform a standard momentum strategy for all criteria in this portfolio setting of ten industries. Apparently, the model adjusts quickly to big shocks and the number of optimal factors decreases when the equity market experiences large losses. However, more empirical work needs to be done to assess its properties adequately, which is outside the scope of the present paper.

## 4 Final Comments

Since the early nineteen-seventies there has been a strong tradition in Bayesian econometrics of studying the shape and integrability of posteriors of parameters of multivariate regression models with a reduced rank using different normalization restrictions and so-called *regularization priors*. Apart from the factor model, the other models are a time series model with an unknown number of non-stationary components and a structural instrumental variable regression model where number and strength of instrumental variables is not known. Research in this field was started in econometrics by Anderson and Rubin (1949) in simultaneous equations models and in 1956 by the same authors in factor models (Anderson and Rubin, 1956). Johansen (1991) treated reduced rank in a time series model with possibly non-stationary components. A survey of the extensive recent frequentist literature is beyond the scope of this paper. There exists also an emerging Bayesian literature about reduced rank estimation, see Bastürk et al. (2017), Section 3 and Appendix 3.2 for details.

We end with emphasizing that the paper by Frühwirth-Schnatter et al. (2024) gave much *food for thought*. We look forward to more theoretical and empirical work on the topic by the authors. In this context, the dynamic nature of many models in economics is relevant. It is very natural to allow parameters of such models to move through time. The well-known Normal or Kalman Filter is a fundamental tool for this and it helps to give identification in factor models. Although dynamic factor models make the mathematics of identification and possible MCMC algorithms more complex, yet, this is a promising research field. We note that the application of static factor models using moving estimation windows and predictive likelihoods for time-varying posterior probabilities of numbers of factors is also able to yield profitable residual momentum strategies that outperform benchmark strategies as the standard momentum strategy.

A second topic is to extend the work by the authors to the field of forecast combinations. Some recent work is given in Billio et al. (2013), Casarin et al. (2023), Aastveit et al. (2023), and Aastveit et al. (2024). The predictive probabilities introduced in the present paper can be used that framework.

#### Acknowledgments

This paper is an invited comment on Frühwirth-Schnatter et al. (2024), Sparse Bayesian Factor Analysis When the Number of Factors Is Unknown. It should not be reported as representing the views of Norges Bank. The views expressed are those of the authors and do not necessarily reflect those of Norges Bank.

## References

- Aastveit, K. A., Cross, J., Furlanetto, F., and Van Dijk, H. K. (2024). “Taylor rules with endogenous regimes.” Technical report, Tinbergen Institute Discussion Paper. [107](#)
- Aastveit, K. A., Cross, J. L., and van Dijk, H. K. (2023). “Quantifying time-varying forecast uncertainty and risk for the real price of oil.” *Journal of Business & Economic Statistics*, 41(2): 523–537. [MR4568040](#). doi: <https://doi.org/10.1080/07350015.2022.2039159>. [107](#)
- Anderson, T. W. and Rubin, H. (1949). “Estimation of the parameters of a single equation in a complete system of stochastic equations.” *The Annals of Mathematical Statistics*, 20(1): 46–63. [MR0028546](#). doi: <https://doi.org/10.1214/aoms/1177730090>. [106](#)
- Anderson, T. W. and Rubin, H. (1956). “Statistical inference in factor analysis.” In Neyman, J. (ed.), *Proceedings of the 3rd Berkeley Symposium on Mathematical Statistics and Probability, Vol. 5*, 111–150. Berkeley. [MR0084943](#). [106](#)
- Ardia, D., Basturk, N., Hoogerheide, L., and Van Dijk, H. K. (2012). “A comparative study of Monte Carlo methods for efficient evaluation of marginal likelihood.” *Computational Statistics and Data Analysis*, 56: 3398–3414. [MR2943902](#). doi: <https://doi.org/10.1016/j.csda.2010.09.001>. [101](#)

- Bartlett, M. S. (1957). “A comment on D. V. Lindley’s statistical paradox.” *Biometrika*, 44: 533–534. MR0207142. doi: <https://doi.org/10.1093/biomet/52.3-4.507>. 101
- Bastürk, N., Hoogerheide, L., and van Dijk, H. K. (2017). “Bayesian analysis of boundary and near-boundary evidence in econometric models with reduced rank.” *Bayesian Analysis*, 12(3): 879–917. MR3694006. doi: <https://doi.org/10.1214/17-BA1061>. 99, 106
- Billio, M., Casarin, R., Ravazzolo, F., and Van Dijk, H. K. (2013). “Time-varying combinations of predictive densities using nonlinear filtering.” *Journal of Econometrics*, 177(2): 213–232. MR3118557. doi: <https://doi.org/10.1016/j.jeconom.2013.04.009>. 107
- Blitz, D., Huij, J., and Martens, M. (2011). “Residual momentum.” *Journal of Empirical Finance*, 18: 506–521. 103
- Casarin, R., Grassi, S., Ravazzolo, F., and van Dijk, H. K. (2023). “A flexible predictive density combination for large financial data sets in regular and crisis periods.” *Journal of Econometrics*, 237(2): 105370. MR4664406. doi: <https://doi.org/10.1016/j.jeconom.2022.11.004>. 107
- Frühwirth-Schnatter, S., Hosszejni, D., and Lopes, H. F. (2024). “Sparse Bayesian factor analysis when the number of factors is unknown.” *Bayesian Analysis*. 98, 99, 100, 107
- Geweke, J. (2007). “Interpretation and inference in mixture models: Simple MCMC works.” *Computational Statistics & Data Analysis*, 51(7): 3529–3550. MR2367818. doi: <https://doi.org/10.1016/j.csda.2006.11.026>. 99
- Johansen, S. (1991). “Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models.” *Econometrica*, 59: 1551–1580. MR1135641. doi: <https://doi.org/10.2307/2938278>. 106
- Koop, G., Strachan, R., Van Dijk, H., and Villani, M. (2006). “Bayesian approaches to cointegration.” 100
- Moskowitz, T. J. and Grinblatt, M. (1999). “Do industries explain momentum.” *The Journal of Finance*, 54(4): 1249–1290. 103
- Newton, M. A. and Raftery, A. E. (1994). “Approximate Bayesian inference with the weighted likelihood bootstrap.” *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(1): 3–48. MR1257793. 101
- Sims, C. A. (2005). “Dummy observation priors revisited.” *Manuscript, Princeton University*. 101
- Tibshirani, R. (1996). “Regression shrinkage and selection via the lasso.” *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1): 267–288. MR1379242. 100

## Contributed Discussion

William R. P. Denault<sup>\*</sup>, Aliaksandr Hubin<sup>†</sup>, Valeria Vitelli<sup>‡</sup>, and Sylvia Richardson<sup>§</sup>

**Introduction** We congratulate Frühwirth-Schnatter et al. (2024) on their paper. Below, we discuss adaptations of the ideas presented in the paper to two problems: Empirical Bayes Matrix Factorization (EBMF), and latent binary Bayesian Neural Network (LBBNN).

**Improving EBMF** The work proposed by Frühwirth-Schnatter et al. (2024) offers a promising avenue for improving the Empirical Bayes Matrix Factorization (EBMF) framework (Wang and Stephens, 2021). In brief, EBMF fits a factor model using a fully factorized variational approximation of the posterior, and as many factor models, it suffers from identifiability issues. A potential improvement could involve modifying the coordinate ascent algorithm to enforce a Generalized Lower Triangular (GLT) structure on the loading matrix at each update, thereby ensuring identifiability. At present, the implementation of EBMF available through the `flashier` R package (Stevens, 2024) does not include priors that enforce GLT structures during the coordinate ascent procedure. As a preliminary investigation, we explored a naive version of GLT by permuting the columns of the loading matrix at each step to maximize the diagonal entries (we refer to this approach as *had hoc* permutation). This empirical approach demonstrated potential benefit in some cases by yielding sparser loading matrices and higher evidence lower bound (ELBO) scores when applied to the GTeX dataset (Lonsdale et al., 2013) (see Figure 1). Although our proposal is empirical and lacks the formal rigor of Frühwirth-Schnatter et al. (2024), we anticipate that refining the EBMF coordinate ascent algorithm by incorporating GLT structures could lead to improved approximations.

**Application to LBBNN** Bayesian neural networks (BNNs) with i.i.d. priors are not only over-parameterized but are also exposed to massive symmetries (Wiese et al., 2023) in the posterior. To resolve these problems, we propose to apply a triple Gamma shrinkage prior inspired by Frühwirth-Schnatter et al. (2024) for the slab component of the latent binary BNN (Hubin and Storvik, 2024). The network weights  $\beta_{ij}^{(l)}$  at each layer  $l$ , neuron  $i$ , and feature  $j$  follow a spike-and-slab distribution with slab:

$$p(\beta_{ij}^{(l)} | \gamma_{ij}^{(l)} = 1) = \iiint \mathcal{N}(\beta_{ij}^{(l)}; 0, \kappa \cdot \theta^{(l)} \cdot \zeta_i^{(l)} \cdot \sigma_0^2) dF(k) dF(\theta^{(l)}) dF(\zeta_i^{(l)}) \quad (1)$$

Here,  $\gamma_{ij}^{(l)} \in \{0, 1\}$  is a binary variable indicating whether the weight is drawn from the slab ( $\gamma_{ij}^{(l)} = 1$ ) or from the spike ( $\gamma_{ij}^{(l)} = 0$ ) at zero  $p(\beta_{ij}^{(l)} | \gamma_{ij}^{(l)} = 0) = \delta(\beta_{ij}^{(l)})$ .

---

<sup>\*</sup>Department of Human Genetics, University of Chicago, IL, USA, [wdenault@uchicago.edu](mailto:wdenault@uchicago.edu)

<sup>†</sup>BiAs, NMBU & University of Oslo & ØUC, Norway, [aliaksandr.hubin@nmbu.no](mailto:aliaksandr.hubin@nmbu.no)

<sup>‡</sup>Department of Biostatistics (OCBE), University of Oslo, Norway, [valeria.vitelli@medisin.uio.no](mailto:valeria.vitelli@medisin.uio.no)

<sup>§</sup>University of Cambridge, UK, & University of Oslo, Norway, [sylvia.richardson@mrc-bsu.cam.ac.uk](mailto:sylvia.richardson@mrc-bsu.cam.ac.uk)

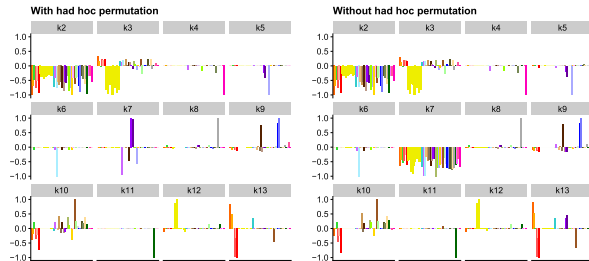


Figure 1: We fit EBMF using both the ad hoc permutation approach and the standard EBMF approach on GTeX data provided in the `flashier` package. The following parameters were used: a row-wise intercept, SVD for initialization,  $K_{\max} = 7$ , and an adaptive shrinkage prior (Stephens, 2017) for the loadings and factors. Each panel represents the posterior mean of the entries for each factor, as estimated by the `flashier` software (we omit the first intercept factor as all values are set to 1). The ELBO for the ad hoc permutation approach was  $-80131.77$ , and for the standard implementation it was  $-80456.5$ .

Further, in (1), the variance of the Gaussian slab is defined using three levels of parameters, where  $\kappa$  is a **global shrinkage** parameter shared across all layers,  $\theta^{(l)}$  is a **layer-specific shrinkage** parameter allowing to control (for logistic activations) the degree of linearity of a specific layer, and  $\zeta_i^{(l)}$  is a **neuron-specific shrinkage** parameter that breaks within layer symmetries. Lastly,  $\sigma_0^2$  is a fixed base variance. The priors on  $\kappa$ ,  $\theta^{(l)}$ , and  $\zeta_i^{(l)}$  are  $\kappa \sim F(2a_\kappa, 2b_\kappa)$ ,  $\theta^{(l)} \sim F(2a_{\theta^{(l)}}, 2b_{\theta^{(l)}})$ ,  $\zeta_i^{(l)} \sim F(2a_{\zeta_i^{(l)}}, 2b_{\zeta_i^{(l)}})$ .

Finally, the binary variable  $\gamma_{ij}^{(l)}$  follows  $p(\gamma_{ij}^{(l)}) = \text{BetaBinomial}(\gamma_{ij}^{(l)}; 1, \alpha, \beta)$ . We fix  $\sigma_0^2 = 1$ , and following Hubin and Storvik (2024), the other prior hyperparameters  $\alpha, \beta, a_\kappa, b_\kappa, a_{\theta^{(l)}}, b_{\theta^{(l)}}, a_{\zeta_i^{(l)}}, b_{\zeta_i^{(l)}}$  are found by an EB technique during the first 20 epochs. Integration in (1) is done by MC sampling at every iteration. Variational approximations and stochastic variational Bayes fully follow the mean-field variant from Hubin and Storvik (2024). The results obtained with the suggested approach are shown and discussed in Table 1. Triple Gamma shrinkage within slabs holds promise for LBBNNs, hence it is of interest to study it more thoroughly. MPM and pruning for LBBNN need further theoretical justification. Lastly, triple Gamma prior is of interest for standard BNN as the alternative to horseshoe (Ghosh et al., 2019).

### Funding

This work was supported by the Research Council of Norway, Integreat – Norwegian Centre for knowledge-driven machine learning, project number 332645. William R. P. Denault is supported by the Eric and Wendy Schmidt AI in Science Postdoctoral Fellowship at the University of Chicago, a Schmidt Sciences program.

Data	Metric	Median probability model				Dense model			
		min	median	max	density	min	median	max	density
MNIST	ACC	0.9814	0.9823	0.9835	0.2865	0.9846	0.9856	0.9865	1.0000
MNIST	ECE	0.0157	0.0162	0.0179	0.2865	0.0198	0.0210	0.0227	1.0000
FMNIST	ACC	0.8890	0.8920	0.8981	0.4201	0.8973	0.8988	0.9015	1.0000
FMNIST	ECE	0.0723	0.0762	0.0794	0.4201	0.0545	0.0567	0.0582	1.0000
KMNIST	ACC	0.9125	0.9138	0.9160	0.4639	0.9244	0.9256	0.9292	1.0000
KMNIST	ECE	0.0661	0.0677	0.0694	0.4639	0.0548	0.0574	0.0590	1.0000

Table 1: Predictive performance (accuracy, and ece) of the proposed LBBNN (summaries based on 10 runs). As compared to a recent comprehensive study (Anonymous, 2024), these results demonstrate marginal improvements to predictive performance and calibration in some of the settings. As expected, median probability model (MPM) sparsity is considerably lower as compared to the approaches without the shrinkage within the slab component. This is arguably due to some sparsity being moved to the slab components. Further sparsification is possible through slab pruning. Just as for other priors in LBBNN, MPM does not change significantly the performance of the full model, albeit providing sparsity.

## References

- Anonymous (2024). “Sparsifying Bayesian neural networks with latent binary variables and normalizing flows.” *Submitted to Transactions on Machine Learning Research*. Under review. URL <https://openreview.net/pdf?id=d6kqUKzG3V>. 111
- Frühwirth-Schnatter, S., Hosszejni, D., and Lopes, H. F. (2024). “Sparse Bayesian factor analysis when the number of factors is unknown.” *Bayesian Analysis*, 1–31. 109
- Ghosh, S., Yao, J., and Doshi-Velez, F. (2019). “Model selection in Bayesian neural networks via horseshoe priors.” *Journal of Machine Learning Research*, 20(182): 1–46. URL <http://jmlr.org/papers/v20/19-236.html>. MR4048993. 110
- Hubin, A. and Storvik, G. (2024). “Sparse Bayesian neural networks: Bridging model and parameter uncertainty through scalable variational inference.” *Mathematics*, 12(6): 788. URL <https://www.mdpi.com/2227-7390/12/6/788>. 109, 110
- Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., Foster, B., Moser, M., Karasik, E., Gillard, B., Ramsey, K., Sullivan, S., Bridge, J., Magazine, H., Syron, J., Fleming, J., Siminoff, L., Traino, H., Mosavel, M., Barker, L., Jewell, S., Rohrer, D., Maxim, D., Filkins, D., Harbach, P., Cortadillo, E., Berghuis, B., Turner, L., Hudson, E., Feenstra, K., Sobin, L., Robb, J., Branton, P., Korzeniewski, G., Shive, C., Tabor, D., Qi, L., Groch, K., Nampally, S., Buia, S., Zimmerman, A., Smith, A., Burges, R., Robinson, K., Valentino, K., Bradbury, D., Cosentino, M., Diaz-Mayoral, N., Kennedy, M., Engel, T., Williams, P., Erickson, K., Ardlie, K., Winckler, W., Getz, G., DeLuca, D., MacArthur, D., Kellis, M., Thomson, A., Young, T., Gelfand, E., Donovan, M., Meng, Y., Grant, G., Mash, D., Marcus, Y., Basile, M., Liu, J., Zhu, J., Tu, Z., Cox, N. J., Nicolae, D. L., Gamazon, E. R., Im, H. K., Konkashbaev, A., Pritchard, J., Stevens, M., Flutre, T.,

- Wen, X., Dermitzakis, E. T., Lappalainen, T., Guigo, R., Monlong, J., Sammeth, M., Koller, D., Battle, A., Mostafavi, S., McCarthy, M., Rivas, M., Maller, J., Rusyn, I., Nobel, A., Wright, F., Shabalin, A., Feolo, M., Sharopova, N., Sturcke, A., Paschal, J., Anderson, J. M., Wilder, E. L., Derr, L. K., Green, E. D., Struewing, J. P., Temple, G., Volpi, S., Boyer, J. T., Thomson, E. J., Guyer, M. S., Ng, C., Abdallah, A., Colantuoni, D., Insel, T. R., Koester, S. E., Little, A. R., Bender, P. K., Lehner, T., Yao, Y., Compton, C. C., Vaught, J. B., Sawyer, S., Lockhart, N. C., Demchok, J., and Moore, H. F. (2013). “The Genotype-Tissue Expression (GTEx) project.” *Nature Genetics*, 45(6): 580–585. Publisher: Nature Publishing Group. URL <https://www.nature.com/articles/ng.2653>. MR0446847. doi: <https://doi.org/10.1080/00033797500200181>. 109
- Stephens, M. (2017). “False discovery rates: a new deal.” *Biostatistics*, 18(2): 275–294. URL <https://academic.oup.com/biostatistics/article/18/2/275/2557030>. MR3824755. doi: <https://doi.org/10.1093/biostatistics/kxw041>. 110
- Stevens, J. R. (2024). “flashr: Creates flashcards of terms and definitions.” R package version 0.1.2. URL <https://cran.r-project.org/package=flashr>. 109
- Wang, W. and Stephens, M. (2021). “Empirical Bayes matrix factorization.” *Journal of Machine Learning Research*, 22(120): 1–40. URL <http://jmlr.org/papers/v22/20-589.html>. MR4279771. doi: <https://doi.org/10.1007/s00023-020-00971-9>. 109
- Wiese, J. G., Wimmer, L., Papamarkou, T., Bischl, B., Günemann, S., and Rügamer, D. (2023). “Towards efficient MCMC sampling in Bayesian neural networks by exploiting symmetry.” In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD): Research Track*, 459–474. doi: [https://doi.org/10.1007/978-3-031-43412-9\\_27](https://doi.org/10.1007/978-3-031-43412-9_27). 109