Bayesian (non-)unique sparse factor modelling

Sylvia Kaufmann and Markus Pape

Working Paper 23.04R

Bayesian (non-)unique sparse factor modelling

Sylvia Kaufmann[a,*], Markus Pape[b]

[a]*Study Center Gerzensee, Dorfstrasse 2, 3115, Gerzensee, Switzerland, and University of Basel, Switzerland*
[b]*Ruhr University Bochum, Universitaetsstrasse 150, 44801, Bochum, Germany*

Abstract

Factor modelling extracts common information from a high-dimensional data set into few common components, where the latent factors usually explain a large share of data variation. Exploratory factor estimation induces sparsity into the loading matrix to associate units or series with those factors most strongly loading on them, eventually determining factor interpretation. The authors motivate geometrically under which circumstances it may be necessary to consider the existence of multiple sparse factor loading matrices with similar degrees of sparsity for a given data set. They propose two MCMC approaches for Bayesian inference and corresponding post-processing algorithms to uncover multiple sparse representations of the factor loading matrix. They investigate both approaches in a simulation study. Applying the methods to data on U.S. sectoral inflation and country-specific gross domestic product growth series, they retrieve multiple sparse factor representations for each data set. Both approaches prove useful to discriminate between pervasive and weaker factors.

Keywords: Multimodality, Sparsity, Pervasive and weak factors

JEL: C11, C33, C55

1. Introduction

We deal with condensing and extracting common information from high-dimensional data, using a factor model

$$\underset{N\times 1}{y_t} = \underset{(N\times K)(K\times 1)}{\Lambda f_t} + \underset{N\times 1}{\epsilon_t},$$

where nowadays typically $K << N$, and a considerable share of data variation is explained by these latent factors or the common component Λf_t,

$$\Sigma_y = \Lambda \Sigma_f \Lambda' + \Sigma_\epsilon, \tag{1}$$

with $\Sigma_y = E(y_t y_t')$, $\Sigma_f = E(f_t f_t')$ and Σ_ϵ diagonal. Factor identification, ultimately determining factor interpretation, has been approached by setting over- or rotational identification restrictions before estimation (Geweke and Zhou,

*corresponding author
Email addresses: sylvia.kaufmann@szgerzensee.ch (Sylvia Kaufmann), markus.pape@rub.de (Markus Pape)

1996; Aguilar and West, 2000; Bernanke et al., 2005), typically on the loading matrix. Also interested in identifying factors, we will explore two ways of proceeding, which do not call for over-identifying restrictions. The first one extracts factors under a generic or just-identified specification (Λ unrestricted, $\Sigma_f = I_K$) and rotates ex-post towards a factor identifying specification (Aßmann et al., 2016; Chan et al., 2018; Aßmann et al., 2023). The second one induces or estimates an association of units or data series with those factors most strongly determining them (West, 2003; Lucas et al., 2006; Kaufmann and Schumacher, 2019). Under both approaches we seek to determine a sparse factor loading matrix, where the non-zero loadings ultimately yield a factor interpretation. The interesting issue arising here is whether the induced or estimated sparse structure is unique or whether there may be multiple sparse factor loading matrices, i.e. factor representations, where each explains approximately the same share of data variation and results in potentially different factor interpretations.

Generally, identification conditions developed in the literature do not rule out local non-uniqueness, i.e. multiple sparse loading matrices that represent different sparse factor models, fitting a given data set potentially similarly well. We motivate geometrically when different sparse loading matrices may arise and lead potentially to different interpretations of underlying factors. We contribute in various dimensions to exploratory, data-driven factor analysis. Both procedures we explore estimate factor models based on order-invariant, just-identified Bayesian posterior inference. Local or rotational identification is obtained by processing the posterior output with algorithms closely related to machine learning procedures, potentially uncovering multiple sparse structures in Λ. Applications to large panels of country-specific gross domestic product (GDP) and U.S. sectoral inflation rates reveal that multiple sparse structures can be uncovered when weak factors underly data variation, a feature discussed in psychometrics (Briggs and MacCallum, 2003) as well as in the econometrics literature, see Freyaldenhoven (2022) and references therein.

In Section 2 we present the model specification and introduce a geometric interpretation of factor models. We motivate why multiple sparse representations may arise. Section 3 outlines the Bayesian framework and the two approaches, based on different priors, to uncover multiple sparse representations. In Section 4, we describe in detail the post-processing algorithms, the first based on optimal rotation and the second on posterior clustering, sorting out factor draws into typical groups of joint factor draws. In Section 5, an extensive simulation study demonstrates the good properties of both approaches, based on scenarios also including pervasive factors, that is factors that load on most and the same units across various sparse representations. Section 6 reports the applications on U.S. monthly sectoral inflation rates and yearly GDP growth rates of countries listed in the Penn World Table. For both datasets, we are able to identify multiple sparse representations. We extract pervasive factors as well as some weaker factors, each identifiable jointly with the pervasive ones, but too weak to be jointly identifiable all together. Section 7 concludes. Appendices A to D contain additional details and results concerning posterior sampling, post-processing, the simulation study and the oracle property of the post-processing procedures when an overfitting number of factors is estimated, respectively.

2. (Non-)Unique sparse factor representation

2.1. Specification

Consider a vector of observable data $Y = (y_1', \dots, y_T')'$. Each y_t, $t = 1, \dots, T$, denotes an $N \times 1$ vector of variables y_{it}, $i, \dots, N$, and can be represented as

$$y_t = \Lambda f_t + \epsilon_t,\ \epsilon_t \sim i.i.d.\ N(0, \Sigma_\epsilon), \tag{2}$$

$$E(f_t f_t') = I_K,\ \Sigma_\epsilon \text{ diagonal with elements } \sigma_i^2, \tag{3}$$

with $K << N$ and where f_t is a $K \times 1$ vector of latent factors, $\Lambda = \{\lambda_{ij} | i = 1, \dots, N, j = 1, \dots, K\}$ is the $N \times K$ factor loading matrix and ϵ_t is an $N \times 1$ vector of idiosyncratic components. We assume without loss of generality an identity covariance matrix for factors, given that correlated factors $\tilde{f}_t$ can be de-correlated by using e.g. a Cholesky decomposition of the factor covariance: $E\left(\tilde{f}_t \tilde{f}_t'\right) = \Sigma_{\tilde{f}} = LL'$ and $L^{-1}\Sigma_{\tilde{f}} L^{-1'} = I_K$. When post-multiplying $\tilde{\Lambda}$ with L, the factor model with correlated factors is observationally equivalent to system (2): $y_t = \tilde{\Lambda} L L^{-1} \tilde{f}_t + \epsilon_t = \Lambda f_t + \epsilon_t$.

As common variation is captured by the factor component only, Σ_ϵ is diagonal and $E(f_t \epsilon_t') = 0$. Although we allow for extensions in the applications, we abstract from a dynamic representation of factors and idiosyncratic errors, as the variance of components in (2) can be interpreted in terms of unconditional variances. We assume that the first and second (unconditional) moments are, respectively, zero and constant, which implies y_t to follow a covariance-stationary process.

In (2), underlying factors are usually unobserved, and we rely on observed data variation, $\Sigma_y = E(y_t y_t')$, to extract the common component:

$$\Sigma_y = \Lambda\Lambda' + \Sigma_\epsilon. \tag{4}$$

Finding a solution to (4) does not only mean mathematically solving the system of $N(N+1)/2$ independent equations. A valid decomposition requires Σ_ϵ to be positive definite and $\Sigma_y - \Sigma_\epsilon$ positive semi-definite and of lower-rank K (Anderson and Rubin, 1956). Questions that arise are (1) does a solution exist and is it unique, which concerns *global* identification; (2) is Σ_ϵ unique, which concerns *local* identification, and (3) for an identified solution, how to determine the orientation of the factor basis and factor order, which concerns *rotational* or *mode* identification. In the following, we deal with local and rotational identification.

In particular we are interested in identifying factors or an orientation of the factor basis which induce a sparse factor loading matrix. This is achieved by specifying a sparse prior distribution on factor loadings or shrinking loadings to zero after an optimal rotation of the factor basis. Thus, we obtain a sparse factor loading matrix by statistical inference. Although the common components of various sparse representations may account for a similar share in data variation, solutions may lead to different elements in Σ_ϵ, which would entail local non-uniqueness. Finding different sparse representations by orthogonal rotation deals with rotational or mode identification. We do not provide an in-depth discussion of identification in the present paper. The interested reader may refer to Kaufmann and Pape (2023), where

we summarize the most important results and provide a geometric approach to identification, including an algorithm to assess the identification properties of a factor model.

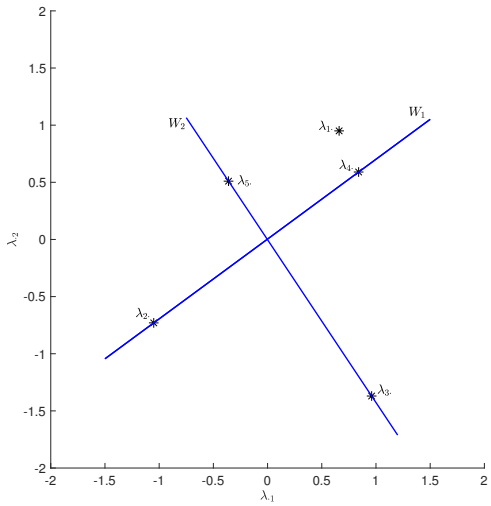
2.2. A geometric interpretation of factor models

To motivate the possibility of multiple sparse factor decompositions, we use the geometric representation of a factor model, where $\Sigma_f = I_K$ spans an orthonormal factor basis, and each row $\lambda_{i\cdot}$ in Λ represents weights attached to basis vectors and corresponds to cartesian coordinates in a K-dimensional space (Lawley and Maxwell, 1971).

For the following exposition it is useful to introduce some geometric and topological concepts. First, denote as a K-frame a set of K independent column vectors in $\mathbb{R}^N$ with $K < N$ or a $N \times K$ matrix with full column rank. The set of all K-frames in $\mathbb{R}^N$ is then denoted as the (real) non-compact Stiefel manifold $V(K, N)$. Note that the column vectors of the K-frame are not required to be orthogonal, as sometimes defined, like in Chan et al. (2018).

As the K independent column vectors in a K-frame span the K-dimensional (real) vector space $\mathbb{R}^K$, we may consider its k-dimensional subspaces for $k < K$. The set of all k-dimensional linear subspaces of $\mathbb{R}^K$ is then denoted as the (real) Grassmann manifold $Gr(k, K)$. For instance, $Gr(1, 2)$ is the set of all lines through the origin in a plane. Last, the set of all orthogonal $K \times K$ matrices is denoted as the (real) orthogonal group $O(K)$, corresponding to an orthogonal factor basis.

Figure 1: Five factor loadings, four of which are located in 1-dimensional subspaces.

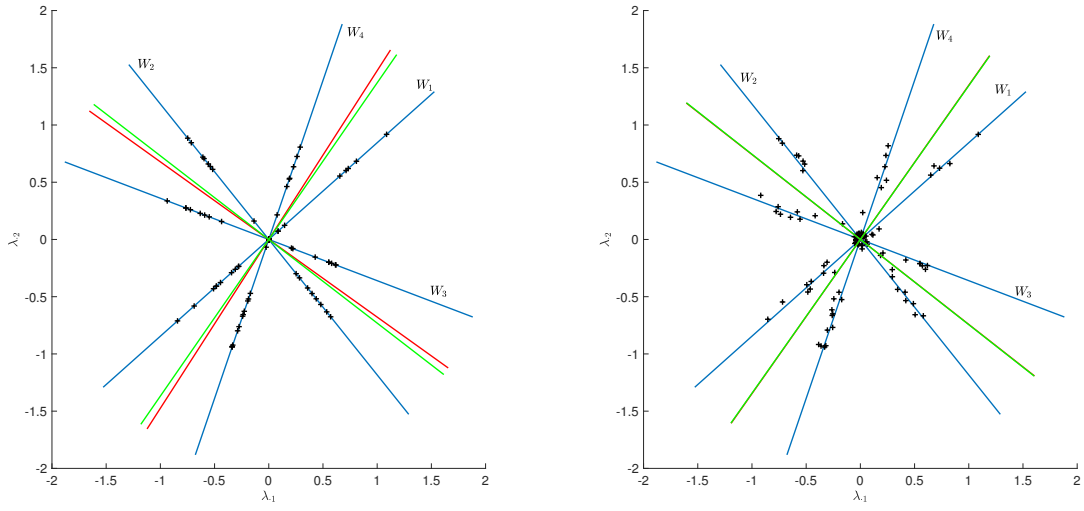


For example, Figure 1 plots the following loading matrices as coordinates:

$$\Lambda = \begin{pmatrix} 0.66 & 0.95 \\ -1.05 & -0.73 \\ 0.96 & -1.37 \\ 0.84 & 0.59 \\ -0.36 & 0.51 \end{pmatrix}, \ \tilde{\Lambda} = \begin{pmatrix} 1.09 & 0.40 \\ -1.28 & 0.00 \\ 0.00 & -1.68 \\ 1.02 & 0.00 \\ 0.00 & 0.63 \end{pmatrix}, \tag{5}$$

where coordinates for Λ are specified in terms of the x- and y-axis. We see that there are two pairs of row vectors in Λ, each located in a 1-dimensional subspace, $W_1 \in Gr(1,2)$ for $\lambda_{2\cdot}$ and $\lambda_{4\cdot}$ and $W_2 \in Gr(1,2)$ for $\lambda_{3\cdot}$ and $\lambda_{5\cdot}$. Both subspaces span an orthogonal factor basis $W_1 \perp W_2$, indicated with blue lines. The sparse loading matrix $\tilde{\Lambda}$ corresponds to the rotated factor basis. The example also illustrates the importance of choosing units when setting pre-defined identification restrictions onto the factor loading matrix. Choosing either $\lambda_{2\cdot}$ and $\lambda_{4\cdot}$ or $\lambda_{3\cdot}$ and $\lambda_{5\cdot}$ as leading units in Λ combined with identification restrictions such as lower diagonal or diagonal, would fail in identifying a second factor as each set of units is loaded by a single factor only. This motivates to base inference on order-invariant estimation and identify factors, including their order and sign, by processing the posterior output, as outlined in the next section.

Figure 2: Two exact sparse representations (left) and two "noisy" sparse representations (right) in a two-factor model. Rotation based on the Varimax criterion and based on least square minimization.



Multiple sparse representations may underlie the data, as Figure (2) illustrates. The left panel shows a model with $K = 2$ factors. It turns out that we can define multiple sparse representations of Λ. Each combination of two of the blue lines W_{k_i}, with $W_{k_i} \in Gr(1,2)$, $k_i = 1, \ldots, 4$, spans an orthogonal factor basis. The two combinations are either

$W_1 \perp W_2$ or $W_3 \perp W_4$, such that either $\lambda_{i\cdot} \in W_1(W_3)$ or $\lambda_{i\cdot} \in W_2(W_4)$ for a first (second) subset of loadings. We additionally show the solutions of rotations based on the Varimax criterion and least square minimization as green and red lines, respectively. Both fail to find any sparse representation. Instead, they result in slightly different orthogonal factor bases spanned in between the sparse representations.

A more realistic scenario is one of only approximate sparse structures underlying the data, where all factor loadings would be non-zero under unrestricted estimation. Nonetheless, multiple representations may underly data where a number of factor loadings may be large and non-zero and the remaining ones may be small and close to zero. The right panel of Figure 2 gives an illustration of a "noisy" bimodal representation. Loading vectors that were previously part of the zero space $W_{\{\}}$ are now located near, but not exactly at the origin, whereas the loading vectors that previously were elements of one of the one-dimensional subspaces W_{k_i} are now only closely located to them. An approach designed to uncover a sparse representation may end up with either set of orthogonal factors plotted in blue. The results of a Varimax optimization and least squares minimization span again a factor basis lying in-between the bases spanning a sparse representation.

In practical applications, this scenario may be relevant in particular for data driven by pervasive factors with nonzero loadings on almost all variables and local or group-specific factors, which load only on specific subsets of variables. Each mode or sparse representation would relate to a different set of weak factors, determining potentially different interpretations of weak factors. With many factor loadings at or near zero, Figure 2 may hence be understood as representing two pairs of weak or local factors.

3. Bayesian inference

As motivated in the previous subsection, multiple modes or sparse representations may arise in exploratory sparse factor analysis where informed by the data, elements of Λ are set endogenously to zero. We propose two Bayesian approaches, based on different priors, to obtain a posterior inference of the model. In view of the discussion in Subsection 2.2, where we illustrated the difficulty of selecting K leading units for pre-imposing factor- and rotation-identifying restrictions, both approaches are based on order-invariant, unconstrained Markov chain Monte Carlo (MCMC) samplers. Factor identification, including factor order and sign, then is obtained by processing the posterior MCMC output.

The approaches differ in terms of their computational involvement at each stage of posterior inference, either when sampling or post-processing. The first approach based on a normal prior for factor loadings and an *unconstrained rotation sampler* (Aßmann et al., 2016, 2023) needs a careful design of a posterior optimization algorithm to find multiple sparse representations of the factor loading matrix of (nearly) equal sparsity degree. The second approach builds on a spike and slab prior (Mitchell and Beauchamp, 1988; George and McCulloch, 1997; West, 2003) and uses a *sparse permutation sampler* to obtain a sample from the multimodal posterior distribution (Kaufmann and Schumacher, 2019). Although the sparse prior induces sparsity into the factor loading matrix, upon convergence to a mode the sampler looses entropy, making it very unlikely to visit other modes or sparse representations (Titsias and

Lázaro-Gredilla, 2011; Bengio et al., 2013). To circumvent the issue, we disturb the sampler after convergence by multiple random rotations and run multiple chains in parallel to detect different sparse modes.

3.1. Bayesian specification

The first building block of the Bayesian framework includes the specification of prior distributions, where in both approaches the prior specification for factor loadings is a standard one used in Bayesian (sparse) factor analysis. The first approach performs posterior inference based on an unconstrained normal prior distribution for the factor loadings

$$\pi\left(\lambda_{ij}\right) = N(0, \tau_0). \tag{6}$$

The second approach induces a sparse Λ by working with a hierarchical spike and slab prior (West, 2003; Carvalho et al., 2008):

$$\pi(\lambda_{ij}|\beta_{ij}, \tau_j) = (1 - \beta_{ij})\delta_0(\lambda_{ij}) + \beta_{ij}N(0, \tau_j), \tag{7}$$

$$\pi(\beta_{ij}|\rho_j) = (1 - \rho_j)\delta_0(\beta_{ij}) + \rho_j B(ab, a(1 - b)), \tag{8}$$

and

$$\pi(\rho_j) = B(r_0 s_0, r_0(1 - s_0)), \tag{9}$$

where δ_0 represents the Dirac Delta function assigning all probability mass to zero and $B(uv, u(1 - v))$ is the beta distribution with mean v and u ruling the precision. For τ_j, we assume an inverse Gamma prior distribution $IG(g_0, G_0)$. Note that both prior specifications are invariant with respect to factor and sign permutation, and the normal prior is also invariant with respect to factor rotation. This allows us to explore the unconstrained posterior distribution.

We introduce the following notation to lay out compactly the second building block, the likelihood, and the posterior inference. We stack all observations of variables y_t into $\boldsymbol{y} = (y_1', \ldots, y_T')'$ and all observations of unobserved factors into $\boldsymbol{f} = (f_1', \ldots, f_T')'$. Model parameters and hyperparameters are gathered in $\theta = \{\Lambda, \Sigma_\epsilon, \vartheta\}$, where ϑ collects all hyperparameters of the hierarchical prior (7)-(9), $\vartheta = \left\{\beta_{ij}, \rho_j, \tau_j | i = 1, \ldots, N,\ j = 1, \ldots, K\right\}$.

The complete data likelihood factorizes as

$$L(\boldsymbol{y}|\boldsymbol{f}, \theta) = \prod_{t=1}^{T} \pi(y_t|f_t, \theta), \tag{10}$$

with normal observation density

$$\pi(y_t|f_t, \theta) = \frac{1}{\sqrt{2\pi}|\Sigma_\epsilon|^{1/2}} \exp\left\{-\frac{1}{2}\left(y_t - \Lambda f_t\right)' \Sigma_\epsilon^{-1} \left(y_t - \Lambda f_t\right)\right\}.$$

To complete the prior specification, we assume a normal prior distribution for factors $\pi(\boldsymbol{f}) = N(0, \mathrm{F}_0)$, $\mathrm{F}_0 = I_{KT}$.

3.2. Posterior inference

Although the joint posterior distribution

$$\pi\left(\boldsymbol{f}, \theta|\boldsymbol{y}\right) = L(\boldsymbol{y}|\boldsymbol{f}, \theta)\pi\left(\theta|\vartheta\right)\pi\left(\vartheta\right) \tag{11}$$

is not available in closed form, we can derive full conditional distributions and rely on a Gibbs sampling scheme. To obtain draws from the posterior distribution, we sample repeatedly from

1. $\pi\left(\Lambda|\boldsymbol{y}, \boldsymbol{f}, \Sigma_\epsilon\right)$. Both the normal and the sparse prior are conditionally conjugate. Therefore, the posterior distributions will also be, respectively, normal and sparse. Under the sparse prior, we additionally update the hyperparameters and draw from $\pi\left(\vartheta|\Lambda\right)$. See Appendix A.1 for the derivation of posterior moments.
2. $\pi\left(\boldsymbol{f}|\boldsymbol{y}, \theta\right) = N(\mathrm{f}, \mathrm{F})$ with moments

$$\mathrm{F} = \left(\boldsymbol{\Lambda}'\left(I_T \otimes \Sigma_\epsilon^{-1}\right)\boldsymbol{\Lambda} + \mathrm{F}_0^{-1}\right)^{-1},\ \mathrm{f} = \mathrm{F}\left(\boldsymbol{\Lambda}'\left(I_T \otimes \Sigma_\epsilon^{-1}\right)\boldsymbol{y}\right),$$

   with $\boldsymbol{\Lambda} = I_T \otimes \Lambda$.
3. $\pi\left(\Sigma_\epsilon|\boldsymbol{y}, \boldsymbol{f}, \Lambda\right) = \prod_{i=1}^{N} IG(\mathrm{s}_i, \mathrm{S}_i)$ with moments

$$\mathrm{s}_i = \mathrm{s} + T/2,\ \mathrm{S}_i = \mathrm{S} + .5\sum_{t=1}^{T}(y_{it} - \lambda_{i\cdot} f_t)^2,\ \text{and s, S prior shape and scale, respectively.}$$

To explore the full unconstrained posterior distribution, depending on the sampler each iteration is terminated by either a random rotation or a random permutation of factors and factor-specific parameters. Step 4. in either scheme consists in

4.**U**. (**Unconstrained rotation**) Random rotation of the factor loadings and factors: Draw an orthogonal matrix $D \in \mathbb{R}^{K\times K}$, distributed with Haar measure, i.e., uniformly on the K-dimensional hypersphere (Mezzadri, 2007), and rotate factors and factor loadings:

$$\begin{aligned} \boldsymbol{f} &:= (I_T \otimes D)\boldsymbol{f}, \\ \Lambda &:= \Lambda D'. \end{aligned} \tag{12}$$

Enforcing rotation accelerates the exploration of the posterior distribution, especially in high-dimensional setups.

4.**S**. (**Sparse permutation**) Random permutation of factor position and sign: First, randomly draw a permutation $\varrho = (\varrho_1, \ldots, \varrho_K)$ of $\{1, \ldots K\}$ and apply it to factors, factor loadings and hyperparameters

$$\begin{aligned} \boldsymbol{f} &:= \varrho(\boldsymbol{f}) = \left\{f_{\varrho_j t}|j = 1, \ldots, K, t = 1, \ldots, T\right\}, \\ \{\Lambda, \vartheta\} &:= \varrho\left(\Lambda, \vartheta\right) = \left\{\lambda_{i\varrho_j}, \beta_{i\varrho_j}, \rho_{\varrho_j}, \tau_{\varrho_j}|i = 1, \ldots, N, j = 1, \ldots, K\right\}. \end{aligned} \tag{13}$$

Second, draw K independent Rademacher distributed random variables. If the k^{th} variable takes the value -1, switch the sign of the k^{th} factor and corresponding loadings. Implementing this step, the output of the sparse permutation sampler will display $2^K K!$ modes.

The unconstrained rotation sampler explores the unconstrained posterior distribution, and generally one MCMC chain or shorter parallel chains are run to obtain a sample from the posterior distribution. As mentioned earlier, the sparse permutation sampler may converge to a sparse representation and stay there, making it difficult for the sampler to visit other sparse representations. To enforce the sampler to visit additional potential sparse representations, we proceed as follows:

1. Simulate a first chain:
   Initialize the sampler, retain M_1 draws from the posterior after convergence.
2. Disturb and simulate $R-1$ chains in parallel:
   Initialize $R-1$ parallel MCMC chains, each by a random orthonormal rotation of a factor loading draw of the first chain, $\Lambda^{(0),r} = \Lambda^{(m)} D^{(r)}$, $m \in \{1, \ldots, M_1\}$. Retain M_r values after convergence.
3. Collect all $M = \sum_{r=1}^{R} M_r$ posterior draws.

Remark 1. To couple the output of parallel MCMC chains with a fixed number of burn-in and M_r of retained draws, we need to ensure that chains have converged, and we have retained enough draws to perform posterior inference. Assessing convergence in a high-dimensional, latent-variable model is not trivial, if we disregard the usual trace and autocorrelation plots or convergence diagnostics for single parameter values. It is even more involved if we (have to) take into consideration random factor and sign permutation applied at the end of each iteration. We suggest using a statistic to assess convergence of both the unidentified and identified posterior output, based on a sign-independent model parameter. We use the Jaccard matching index between the first draw of a (sub-)chain and all following draws, based on an indicator matrix for non-zero loadings, obtained by evaluating $\beta_{ij}^{(m)}$ for each draw m. The procedure is detailed in Appendix A.2 and illustrated for the MCMC output obtained for U.S. sectoral inflation rates (see Subsection 6.2). The alternative of implementing an exact procedure (Jacob et al., 2020) in this high-dimensional, latent-variable model is non-trivial if at all feasible. Random permutation of factors and factor sign at the end of each iteration make the problem even more difficult.

4. Posterior processing: Multiple mode identification

Next, we describe the mode identification procedures using the output of the unconstrained rotation and sparse permutation sampler, based on the geometric representation motivated in Subsection 2.2.

4.1. Mode identification using the output of the unconstrained rotation sampler

To obtain a sample from the posterior distribution of Λ, we first post-process the unconstrained sampler's output with the weighted orthogonal Procrustes (WOP) procedure to orient all draws towards a common factor basis, see Aßmann et al. (2016). After this step, the posterior distribution is identified up to a final orthogonal rotation H_*, $H_* H_*' = I_K$. When appropriately oriented, the matrix H_* will identify a sparse structure in Λ. In this subsection, we discuss an optimization approach to determine H_*.

Highest posterior density (HPD) K-dimensional hyperellipsoids, constructed for each $1 \times K$ row of factor loadings $\lambda_{i\cdot}$ in Λ, provide the basis for the optimization. The objective is to rotate the factor basis in such a way that sparse subspaces spanned by as few basis axes as possible will intersect with each of the hyperellipsoids, inducing a sparse representation for Λ. For non-elliptical posterior distributions, the corresponding K-dimensional HPD regions are represented by shells in K-dimensional space, see Hyndman (1996). In the following, we stick to the case of elliptical posterior distributions.

Definition 1. *Parameterization of a hyperellipsoid*

A K-dimensional hyperellipsoid is defined by three parameters: its center $c \in \mathbb{R}^K$, its orientation $H \in O(K)$ and its radii, or half-diameters $r \in \mathbb{R}^K$.

The orientation H can be replaced by a vector $\gamma \in \mathbb{R}^P$, where $P = \binom{K}{2}$, which contains the rotation angles from a Givens decomposition of H, see Appendix B.1. ◇

According to Definition 1, the parameters of the unit circle centered at the origin of the $\mathbb{R}^2$ space are $c = (0,\ 0)'$, $H = I_2$ or $\gamma = 0$ and $r = (1, 1)'$.

All parameters characterizing the K-dimensional HPD hyperellipsoid for $\lambda_{i\cdot}$ can be inferred from the posterior sample. The center is estimated as $\hat{c}_i = 1/M \sum_{m=1}^{M} \lambda_{i\cdot}^{(m)\prime}$. To obtain an estimate for the orientation H_i, we first compute an estimate of the covariance matrix of $\lambda_{i\cdot}$, $\widehat{\Psi}_i = 1/M \sum_{m=1}^{M} \lambda_{i\cdot}^{(m)\prime} \lambda_{i\cdot}^{(m)} - \hat{c}_i \hat{c}_i'$. The spectral decomposition $\widehat{\Psi}_i = \widehat{H}_i \widehat{W}_i \widehat{H}_i'$ yields $\widehat{H}_i$, an orthogonal matrix, and $\widehat{W}_i$, a diagonal matrix with eigenvalues $\hat{w}_{i,1}, \ldots, \hat{w}_{i,K}$ on the diagonal. The Givens decomposition of $\widehat{H}_i$ yields the Givens rotation angles $\hat{\gamma}_i = (\hat{\gamma}_{i,1}, \ldots, \hat{\gamma}_{i,P})'$, where P is the number of axis pairs involved. To obtain estimates for the radii $r_i = (r_{i,1}, \ldots, r_{i,K})'$, we work with demeaned and decorrelated draws. We demean the draws $\lambda_{i\cdot}^{(m)}$ to obtain $\lambda_{i\cdot}^{(m),dem} = \lambda_{i\cdot}^{(m)} - \hat{c}_i'$. Next, we decorrelate the demeaned draws to obtain $\lambda_{i\cdot}^{(m),dec} = \lambda_{i\cdot}^{(m),dem} \widehat{H}_i$. Finally, we standardize the demeaned and decorrelated draws to obtain $\lambda_{i\cdot}^{(m),stand} = \lambda_{i\cdot}^{(m),dem} \widehat{H}_i \widehat{W}_i^{-\frac{1}{2}}$. Denote the empirical $(1-\alpha)$ quantile of $\|\lambda_{i\cdot}^{(m),stand}\|_2$ by $q_{1-\alpha}$, where $\|\cdot\|_2$ denotes the Euclidean norm, and estimate the radii of the i^{th} hyperellipsoid as $\hat{r}_{i,k} = q_{1-\alpha} \sqrt{\hat{w}_{i,k}}$. Note that for convenience, we omit the hats on HPD parameter estimates in the following.

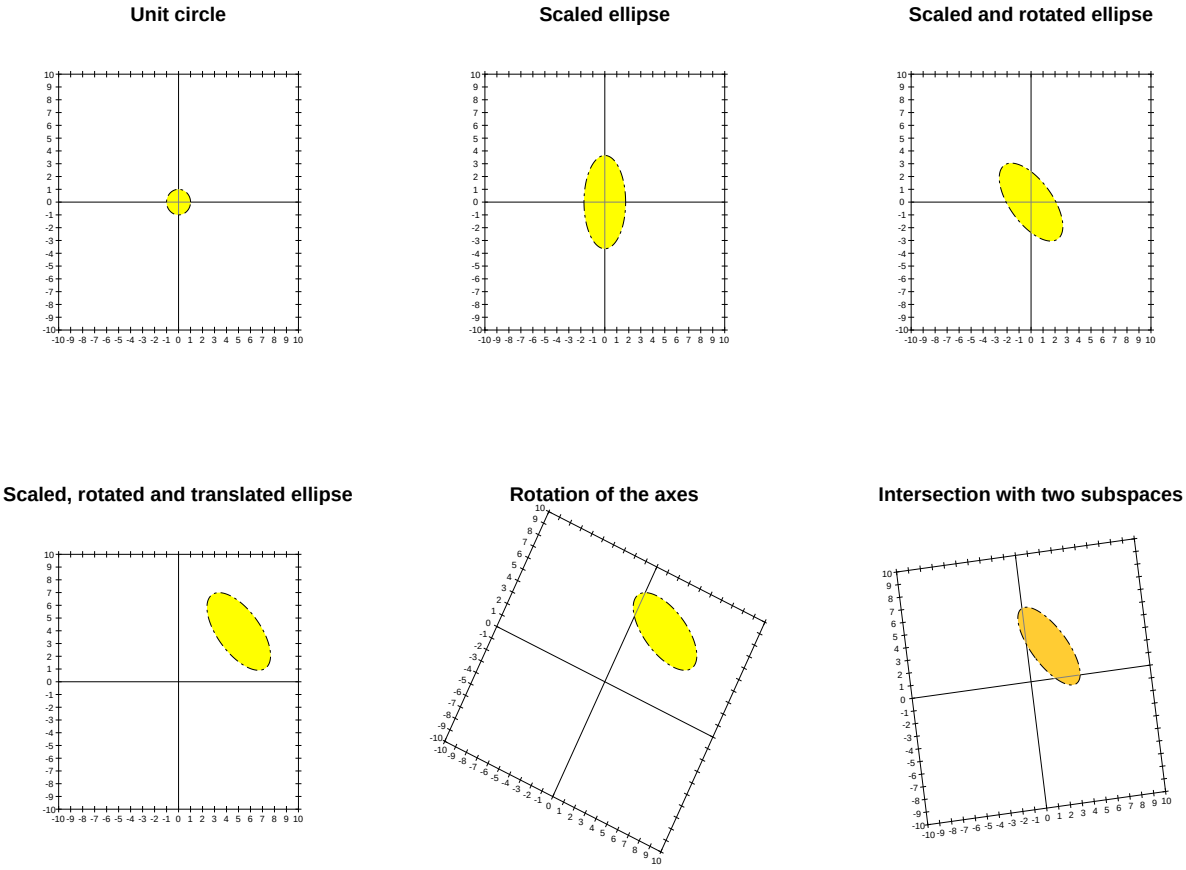
Definition 2. *Relation between a point and a hyperellipsoid in the $\mathbb{R}^K$ space*

For general $K \in \mathbb{N}$, it holds that any point $x \in \mathbb{R}^K$ lies inside the K-dimensional hyperellipsoid parameterized as in Definition 1, if and only if

$$\|(x - c)' H R^{-1}\|_2 < 1, \quad \text{where} \quad R = \text{diag}(r_1, \ldots, r_K). \qquad \diamond$$

Figure 3 illustrates how to re-construct a two-dimensional ellipsoid $\lambda_{i\cdot}$ with the three parameters. In the first row, the first panel shows the unit circle that consists of the set of points $\{x_i | x_i' x_i = 1\}$. In the second panel, the unit circle is expanded to an ellipse by scaling the points along the k^{th} dimension with radius $r_{i,k}$ to $R_i x_i$. In the third panel, the ellipse has been rotated by H_i to $H_i R_i x_i$. Eventually, the ellipse is shifted, translating all of its points to $H_i R_i x_i + c_i$ (First panel in the second row). The procedure is generically applicable to higher-dimensional ellipsoids, see Appendix B.2 for a $K = 3$-dimensional example.

Figure 3: 95% highest posterior density ellipsoid for $K = 2$, built from the unit circle (first row, first panel), which is first expanded (first row, second panel), then rotated (first row, third panel), and eventually translated (second row, first panel). A possible rotation of the coordinate system is shown in the second panel of the second row. The third panel of the second row shows a different hyperellipsoid, which intersects with both axes after rotating the coordinate system.



To identify a sparse representation, we look for a rotation matrix H_*, such that as many ellipsoids as possible will intersect with low-dimensional subspaces of the space spanned by H_*. The second panel in the second row of Figure 3 illustrates a rotation of the axes to the right by an angle that causes the hyperellipsoid to intersect with the y-axis. Both axes are one-dimensional subspaces of the $\mathbb{R}^2$ space. The intersection implies that, to represent the ellipse, a nonzero loading is necessary only for the second factor, while the loading for the first factor can be set to zero.

Definition 3. *Sparsity indicator matrix*
Let the matrix $\Delta \in \mathbb{R}^{N\times K}$ represent the sparse pattern in Λ, indicating non-zero coordinates of the subspaces ellipsoid i intersects with, i.e. $\delta_{ik} = 1$ if $\lambda_{ik} \neq 0$, and zero otherwise. Moreover, let $\delta_{i\cdot}$ denote the i^{th} row of matrix Δ. $\diamond$

Note that if the i^{th} hyperellipsoid includes the origin, the hyperellipsoid intersects with the zero-dimensional space, and hence, all loadings on variable i can be set to zero. That is, $\delta_{ik} = 0$ for all $k \in \{1, \ldots, K\}$. This also holds under arbitrary rotations H of the factor basis. Accordingly, if the i^{th} hyperellipsoid intersects with the k^{th} axis only, loadings of variable i can be set to zero except the k^{th} one, i.e., $\delta_{ik} = 1$ and $\delta_{ij} = 0$ for all $j \in \{1, \ldots, k-1, k+1, \ldots, K\}$ (see

Figure 3, second panel in the second row). If H_* causes the hyperellipsoid to overlap with different subsets, without including their intersection, the sparse representation is not unique. The third panel in the second row illustrates this situation. The ellipse intersects each axis, without including the origin. In this case, we can set either $\delta_{i\cdot} = (1,\ 0)$ or $\delta_{i\cdot} = (0,\ 1)$.

Definition 4. *Indexing k-elemental subsets of the set* $\{1, \ldots, K\}$

Consider the set $\mathcal{K} = \{1, \ldots, K\}$ with $K \geq 1$ and $0 \leq k \leq K$. Denote as the j^{th} k-elemental subset of $\mathcal{K}$ the set $\mathcal{K}_{k,j} \subseteq \{1, \ldots, K\}$ with $\left|\mathcal{K}_{k,j}\right| = k$ and $\psi_j = \sum_{i=1}^{K} 2^{K-i} \mathcal{I}_{\{i \in \mathcal{K}_{k,j}\}}$, such that $\psi_j > \psi_h$ for every $j < h$. ◇

For instance, consider $K = 3$, $\mathcal{K} = \{1, 2, 3\}$, and $k = 2$. The ordered 2-elemental subsets are $\mathcal{K}_{2,1} = \{1, 2\}$ with $\psi_1 = 6$, $\mathcal{K}_{2,2} = \{1, 3\}$ with $\psi_2 = 5$, and $\mathcal{K}_{2,3} = \{2, 3\}$ with $\psi_3 = 3$.

Definition 5. *Subspace and rotated subspace*

Let $K \geq 1$ and $0 \leq k \leq K$, and let the $K \times K$ matrix $S_{(k,j)}$ contain the standard vectors corresponding to the axes indicated by $\mathcal{K}_{k,j}$ and zero vectors elsewhere. Then $S_{(k,j)}$ spans the j^{th} k-dimensional subspace of the K-dimensional space.

Let $H \in O(K)$ be a rotation matrix. Then $HS_{(k,j)}$ spans the rotated j^{th} k-dimensional subspace of the K-dimensional space. ◇

For instance, for $K = 3$, the matrix that spans the first subspace of dimension $k = 2$ corresponds to $S_{(2,1)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$, with $HS_{(2,j)} \in Gr(2, 3)$, $H \in O(3)$. Note that for $2 \leq k \leq K - 1$, Definition 5 is redundant as the different kth-dimensional subspaces can be spanned by $S_{(k,1)}$ under appropriate rotation H. However, we need to consider all different $\binom{K}{k}$ elements as the diagonal elements of $S_{(k,j)}$ will determine $\delta_{i\cdot}$.

Definition 6. *Points in subspaces and rotated subspaces*

Let $s_{(k,j)}$ be a $K \times 1$ vector of nonzero scaling factors for $S_{(k,j)}$, such that $x = S_{(k,j)} s_{(k,j)}$ is a point within the subspace spanned by $S_{(k,j)}$.

Let $H \in O(K)$ be a rotation matrix. Then $x = H(S_{(k,j)} s_{(k,j)})$ is a point within the rotated subspace spanned by $HS_{(k,j)}$. ◇

Definition 7. *Optimal scaling vector*

Conditional on a rotation H and $S_{(k_i,j_i)}$, we minimize the Mahalanobis distance between $x_i = H(S_{(k_i,j_i)} s_{(k_i,j_i),i})$ and the center of the i^{th} hyperellipsoid c_i

$$\ell_{k_i,j_i,i}(S_{(k_i,j_i)}, s_{(k_i,j_i),i}, H) = \|(H(S_{(k_i,j_i)} s_{(k_i,j_i),i}) - c_i)' H_i R_i^{-1}\|_2,$$

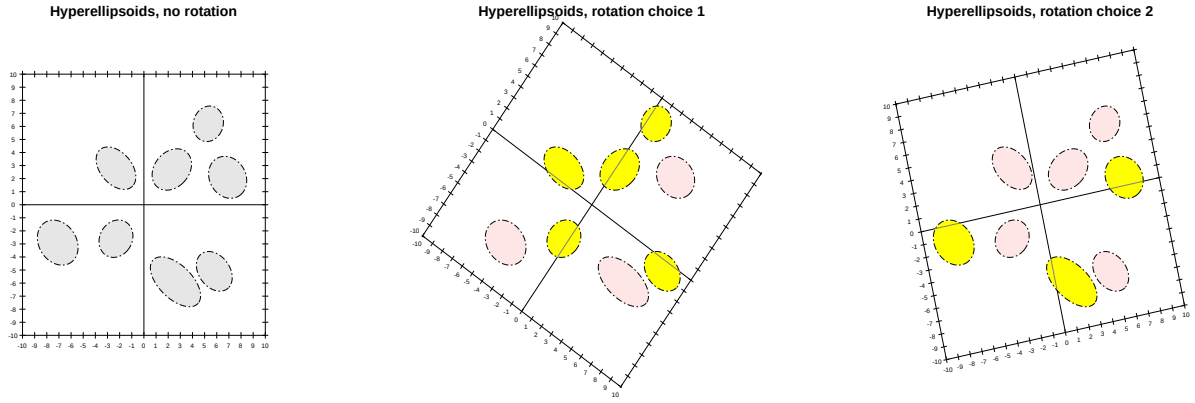
to obtain the optimal scaling vector

$$s^{opt}_{(k_i,j_i),i}(H) = \underset{s_{(k_i,j_i),i}}{\arg\min} \left\{ \ell_{k_i,j_i,i}(S_{(k_i,j_i)}, s_{(k_i,j_i),i}, H) \right\}. \quad \diamond$$

If $s^{opt}_{(k_i,j_i),i}(H)$ causes x_i to lie within the i^{th} hyperellipsoid (see Definition 2), the i^{th} hyperellipsoid intersects with the rotated subspace spanned by $HS_{(k_i,j_i)}$. When this is the case, $\lambda_{i\cdot}$ falls into the corresponding subspace $S_{(k_i,j_i)}$ and we set $\delta_{i\cdot}$ equal to the diagonal elements of $S_{(k_i,j_i)}$, $\delta_{i\cdot} = \text{diag}(S_{(k_i,j_i)})$.

We optimize H to obtain a minimum number of nonzero elements in Δ. This is achieved when a maximum number of hyperellipsoids intersect with low-dimensional subspaces, i.e. k_i should be as small as possible. When evaluating sparsity under some rotation H, we therefore start with low-dimensional subspaces.

Figure 4: 95% highest posterior density ellipsoids for eight row vectors $\lambda_{i\cdot}$ with $K = 2$ (first panel), first proposed axis rotation (second panel), and second proposed axis rotation (third panel).



For example, consider the eight hyperellipsoids shown in the first panel of Figure 4. A rotation of the axes to the right (left) by 35 (12) degrees causes five (three) hyperellipsoids to intersect with one of the rotated axes (highlighted in yellow). The first (second) rotation yields $\delta_{i\cdot} = (1,\ 0)$ for two (two) variables, $\delta_{i\cdot} = (0,\ 1)$ for three (one) variable(s), and $\delta_{i\cdot} = (1,\ 1)$ for three (five) variables. The first rotation induces the sparsest representation.

With these topological considerations at hand, we define the loss function to minimize:

$$\mathcal{L}(H) = \sum_{i \in \mathfrak{H}} \ell_i^*(H) + \zeta \sum_{i \in \mathfrak{H}} \sum_{k=1}^{K} \delta_{ik}(H). \tag{14}$$

The solution yields the optimal rotation

$$H_* = \arg\min_H \{\mathcal{L}(H)\}$$

Note that only a subset $\mathfrak{H} \subseteq \{1, \ldots, N\}$ of units contribute to the loss function. All loadings of those units with hyperellipsoids including the origin are set to zero, remain zero under any rotation H. Therefore $\delta_{i\cdot} = 0$ for all these units, see also the remarks after Definition 3. The first term in Equation (14) captures the contribution of the nonzero rows determined for Λ:

$$\ell_i^*(H) = \ell_{k_i, j_i, i}(S_i^*(H), s_i^*(H), H) \quad \text{for every} \quad i \in \mathfrak{H}, \tag{15}$$

where $S_i^*(H)$ spans a low-dimensional subspace, and where conditional on $S_i^*(H)$, $s_i^*(H) = s^{opt}_{(k_i, j_i), i}(H)$ (Definition 7).

The second term in Equation (14) acts as penalty by scaling the number of nonzero elements in Δ determined under rotation H. In simulations and applications, we used $\zeta = 10$.

For optimizing, it is convenient to express H as a function of the $\binom{K}{2}$ Givens rotation angles γ, and optimizing with respect to γ, see Appendix B.1. A starting value for H may be chosen randomly or set empirically by e.g. a Varimax rotation of $1/M \sum_{m=1}^{M} \Lambda^{(m)}$. We use the following algorithm to determine Δ and evaluate the loss function until convergence:

1. Set $i = \min\{\mathfrak{H}\}$.
2. While $i \leq \max\{\mathfrak{H}\}$
   (a) Set $k_i = 1$.
   (b) While $k_i \leq K - 1$
      (i) For every $j_i \in \left\{1, \ldots, \binom{K}{k_i}\right\}$, determine the optimal scaling vector $s^{opt}_{(k_i,j_i),i}(H)$ for the subspace spanned by S_{k_i,j_i}, see Definition 7.
      (ii) Determine $\mathfrak{J}_i$:

$$\mathfrak{J}_i = \left\{ j_i | \ell_{k_i,j_i,i}(S_{(k_i,j_i)}(H), s^{opt}_{(k_i,j_i),i}(H), H) < 1; j_i = 1, \ldots, \binom{K}{k_i} \right\} \tag{16}$$

      (aa) If $\mathfrak{J}_i = \varnothing$, increment k_i and proceed with (b).
      (bb) If $|\mathfrak{J}_i| = 1$ (c.f. Figure 3, second-row second panel), the rotation H allows for a sparse representation of the i^{th} row of Λ with k_i nonzero elements, with $S^*_i(H) = S_{k_i,j_i}$ and $s^*_i(H) = s^{opt}_{(k_i,j_i),i}(H)$.
      Calculate $\ell^*_i(H)$ (Equation (15)) and set $\delta_{i\cdot}(H) = \text{diag}(S^*_i(H))$.
      Set i to the next element of $\mathfrak{H}$ and go to 2.
      (cc) If $|\mathfrak{J}_i| > 1$ (c.f. Figure 3, second-row third panel), choose one of the following strategies:
         * Select the j_i contributing minimum loss $\ell^*_i(H)$. We always follow this strategy.
         * Select j_i randomly from $\mathfrak{J}_i$.
         * Select j_i corresponding to a subspace S_{k_i,j_i} which excludes those axes for which we are interested in setting loadings to zero for variable i.
      Set $S^*_i(H) = S_{k_i,j_i}$ and $s^*_i(H) = s^{opt}_{(k_i,j_i),i}(H)$.
      Calculate $\ell^*_i(H)$ (Equation (15)) and set $\delta_{i\cdot}(H) = \text{diag}(S^*_i(H))$.
      Set i to the next element of $\mathfrak{H}$ and go to 2.
   (c) If $\mathfrak{J}_i = \varnothing$ (and $k_i = K$), there is no sparse representation for the i^{th} hyperellipsoid, i.e. the i^{th} row of Λ.
   In this case, set

$$S^*_i(H) = \underset{S_{k_i-1,j_i}(H)}{\arg\min} \left\{ \ell_{k_i-1,j_i,i}(S_{(k_i-1,j_i)}, s^{opt}_{(k_i-1,j_i),i}(H), H) \right\}$$

   and determine $s^*_i(H)$ as in Definition 7 with $S_{k_i\cdot j_i} = S^*_i(H)$.
   Calculate $\ell^*_i(H)$ (Equation (15)) and set $\delta_{ik}(H) = 1$ for all $k \in \{1, \ldots, K\}$
   Set i to the next element of $\mathfrak{H}$ and go to 2.

3. Calculate $\mathcal{L}(H)$ (Equation (14)), adjust H.

   Return to 1. until $\mathcal{L}(H)$ converges. Upon convergence $H_* = H$.

Remark 2. In our simulations and applications, we optimize over the Givens rotation angles γ_* and use the Matlab® routine `fminunc` for optimization. The same routine is used to find the optimal scaling vectors $s^{opt}_{(k_i,j_i),i}(H)$.

Remark 3. The loss function may be optimized sequentially, using the algorithm to optimize with respect to Givens angles sequentially. Each angle γ_p involves the subspace spanned by two columns only, say k_1 and k_2. The set $\mathfrak{H}$ is thus determined for columns k_1 and k_2. This considerably accelerates the algorithm, as for $K = 2$, in step (b) we only need to evaluate the two one-dimensional subspaces $S_{(1,1)}$ and $S_{(1,2)}$. Each optimal pairwise rotation p, $H_{*,p}$ is incorporated into the hyperellipsoids by rotating them by $H'_{*,p}$ towards the optimized axes. The following optimization for another axes pair is undertaken conditional on all previous optimizations. The final optimal rotation H_* results then from

$$H_* = \prod_{p=1}^{P} H_{*,p}.$$

Appendix B.3 illustrates the convergence of the algorithm for the WOP-processed MCMC output of scenario K3m2_1pf in the simulation study (Subsection 5.2).

To determine further sparse modes, we can adjust the loss function and penalize solutions too close to previously determined solutions $\mathcal{H} = \{H^1_*, \ldots, H^L_*\}$:

$$H^{L+1}_* = \arg\min_H \left\{ \left(\sum_{i\in\mathfrak{H}} \ell^*_i(H) + \zeta \sum_{i\in\mathfrak{H}} \sum_{k=1}^{K} \delta_{i,k}(H) \right) / \left(\varkappa \min_l \left\{ \|H' H^l_* - P^*_{s,l}\|_2 + \epsilon \right\} \right) \right\}, \tag{17}$$

where $\varkappa > 0$ is a suitably chosen penalty and

$$P^*_{s,l} = \arg\min_{P_s} \left(\|H' H^l_* - P_s\|_2 \right)$$

is the K-dimensional signed permutation matrix P_s with minimal distance to $H' H^l_*$, to avoid solutions that only permute factor position and sign of previous solutions. A small value $\epsilon > 0$ ensures that the function is defined for every H.

4.2. Mode identification using the output of the sparse permutation sampler

As motivated in Subsection 3.2, we run multiple chains of the sparse permutation sampler to allow the sampler to converge to and stabilize at potentially more than one sparse mode. Each sparse representation will display $2^K K!$ modes due to the random permutations of factor positions and signs at the end of each iteration. In the presence of more than one sparse mode, visual tools like scatter plots or histograms that usually uncover label and sign switching within a mode may become inappropriate to discriminate between sparse modes in a first stage, or vice versa. For example, the upper-left scatter plot in Figure 5 visualizes the unsorted MCMC output for factor loadings of a unit, where the black dots represent all permutations of the true loadings for each sparse mode ($\lambda_{39,\cdot} = [0.81, -0.72]$ and $\lambda_{39,\cdot} = [.06, -1.08]$, are, respectively, the first and second mode for unit 39 of scenario data50ex_ln in Subsection 5.1.) Although the pattern discriminates well between the two sparse modes, one where both factor loadings are non-zero

and the other where one loading is nearly zero, it is difficult to find factor-identifying restrictions in the first mode, as factor loadings are very close to each other (in absolute terms). The histogram of draws below the scatter plot illustrates the difficulty in defining mode- and factor-identifying restrictions based on the marginal density of a unit-specific factor loading. When the number of factors is larger or sparse patterns are more complex, it becomes even more difficult to determine a restriction discriminating between sparse modes. The upper-left scatter plot of factor loadings in Figure 6 visualizes the situation for a simulated factor model with four factors and two sparse modes. Obviously, there is no way of separating draws into one of the modes, nor a way of identifying factors.

However, factor draws, $f_k^{(m)} = \left\{f_{kt}^{(m)} | t = 1, \ldots, T\right\}$, from the same posterior distribution will be highly correlated across each other. We visualize this in the upper right panels of Figures 5 and 6. These scatter plots suggest that groups of factor draws are well identified based on there cross-correlations. The right histogram in Figure 5 also shows a distinct group of highly correlated factor draws (the absolute correlation is nearly 1). Therefore, to identify potential multiple sparse modes and factors within modes we suggest to post-process the MCMC output based on correlations across factor draws (or correlations across factors and factor loadings draws). In a first step, we set up an overfitting mixture model for factors (or factors stacked with loadings), where the number of components G will be a multiple of the number of factors K. The number of filled components will indicate the number of distinct factors sampled. Each draw of (distinct) K factors is then assigned to the mode (the factor set) combining those K factors. The number of filled factor sets will indicate the number of sparse modes sampled. Within each mode, we re-order draws and switch sign accordingly to obtain factor identification.

We proceed in the following way:

1. Classify each factor draw $f_k^{(m)} = \left\{f_{kt}^{(m)} | t = 1, \ldots, T\right\}$, $k = 1, \ldots, K$, $m = 1, \ldots, M$ into one of $G \geq K$ clusters by estimating a mixture model with G components, where G is set to a multiple of K. The prior mixture probability η is assumed uniform Dirichlet and is specified in a way to allow for empty groups ex-post, $\pi(\eta) = D(e_0, \ldots, e_0)$, with $e_0 < G/2$ (Rousseau and Mengersen, 2011). Conditional on the component indicator $z_k^{(m)} \in \{1, \ldots, G\}$, $f_k^{(m)} | z_k^{(m)} = g \sim N(\mathrm{f}_g, \mathrm{F}_g)$, where the mean factor path $\mathrm{f}_g = \{\mathrm{f}_{gt} | t = 1, \ldots, T\}$ of component g is interpreted as *factor representative*.

   See Appendix B.4 for more details on the sampler.

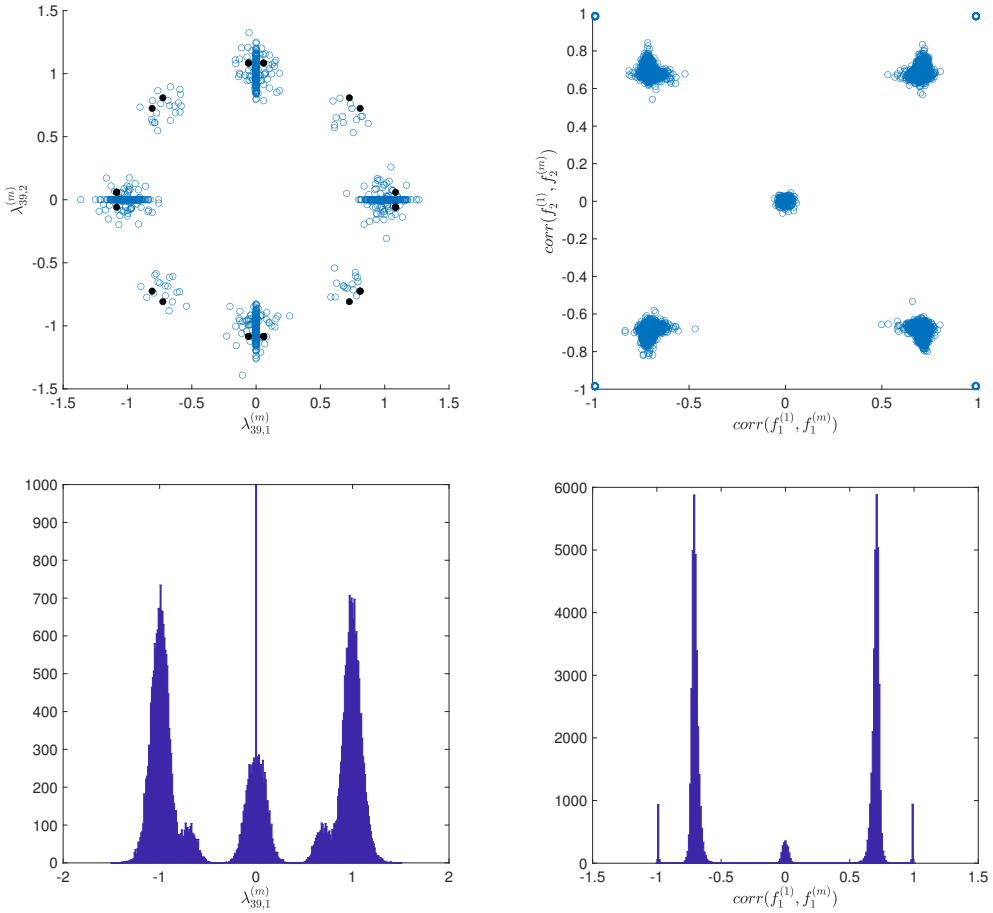
2. For posterior inference, retain those draws (m) of K factors, for which the association to components is unique, and re-order factors in ascending order of components $f^{(m)} = \left\{f_{z_k^{(m)}}^{(m)} | k = 1, \ldots, K; z_1^{(m)} < \cdots < z_K^{(m)}\right\}$. Change the sign of those draws negatively correlated with the factor representative $f_{z_k^{(m)}}^{(m)} := \mathrm{sign}(corr(f_{z_k^{(m)}}^{(m)}, \mathrm{f}_{z_k^{(m)}})) f_{z_k^{(m)}}^{(m)}$.

3. Finally, evaluate how many times (N_Z) a factor set $\mathcal{I}_Z = \{Z_1, \ldots, Z_K\} \subset \{1, \ldots, G\}$, $Z = 1, \ldots \binom{G}{K}$ has been drawn. Retain the most populated, e.g. the sets with more than 1,000 draws.

Remark 4. To obtain a sharper distinction between groups, we may stack factor and factor loading draws in the first step: Classify $(f_k^{(m)\prime} \lambda_k^{(m)\prime})' = \left\{f_{kt}^{(m)}, \lambda_{ik}^{(m)} | t = 1, \ldots, T; i = 1, \ldots, N\right\}$ into one of $G \geq K$ clusters by estimating a mixture model with G components.

Figure 5: MCMC output for the scenario $K = 2$ factors and two sparse modes, data50ex_ln in Table 2. Left panels: Scatter plots and histogram of factor loadings for a selected series; right panels: Scatter plots and histogram of correlations across draws for the first factor against correlations across draws for the second factor. Blue and red colors refer to the first and second identified mode, respectively. The black dots reflect all permutations of true factor loadings (Panel (a)) and mode-specific true factor loadings (Panel (c)).

(a) Unsorted draws



(b) Sorted draws

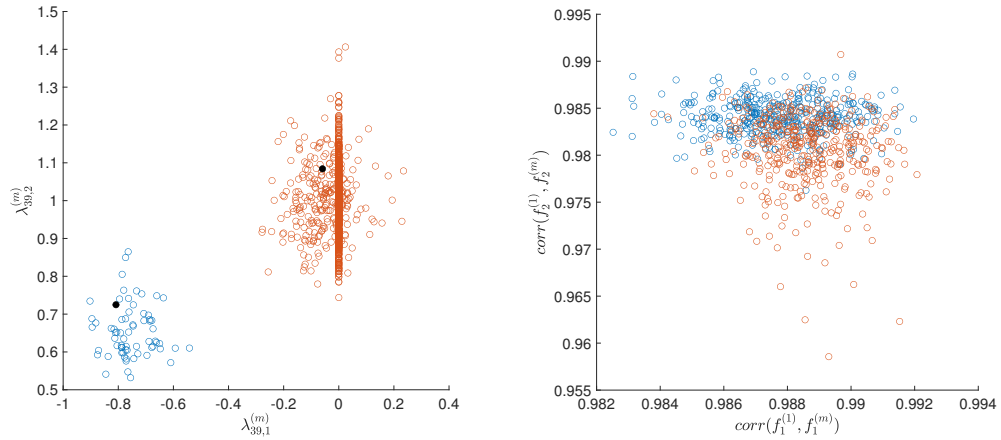
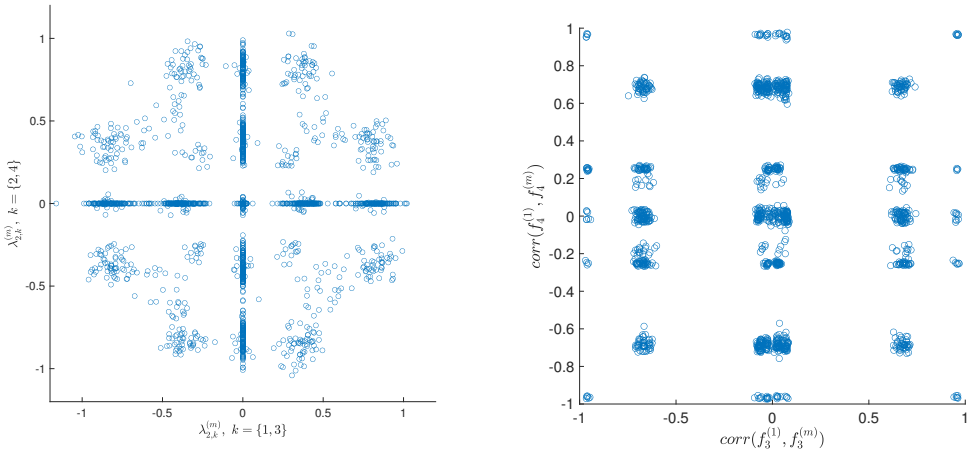
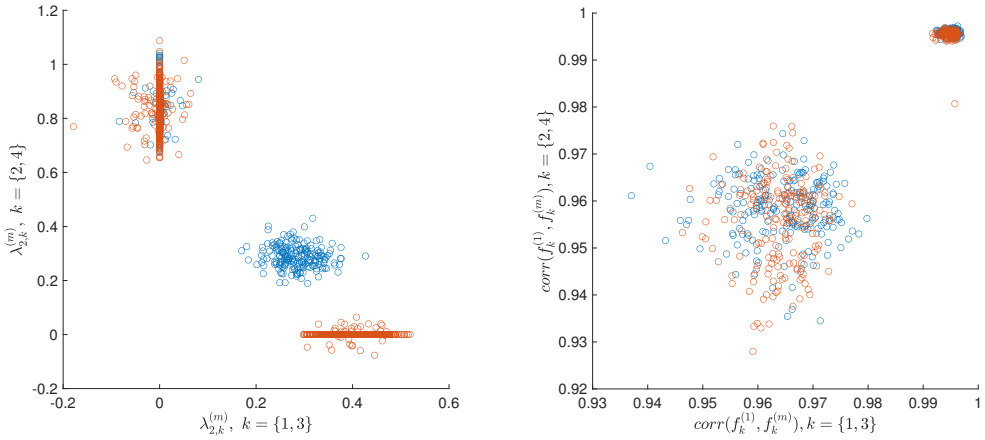


Figure 6: MCMC output for the scenario $K = 4$ factors with two pervasive factors and two sparse modes, K4m2_2pf_ln in Table 4. Left panels: Scatter plot of factor loadings for the second series; right panels: Scatter plots of correlations across draws of the first (third) factor against correlations across draws of the second (fourth) factor. Blue and red colors refer to the first and second identified mode, respectively.

(a) Unsorted draws



(b) Sorted draws



Remark 5. The sampler usually converges quite quickly. Nevertheless, an increasing dimension of $(f_k^{(m)\prime}\lambda_k^{(m)\prime})'$ and the posterior sample M may slow down considerably the clustering algorithm. Therefore, we may apply Step 1. only to a randomly chosen subset of posterior draws to determine the factor representatives. We then determine component association of each draw, $z_k^{(m)}$, based on the correlation with factor representatives, $z_k^{(m)} = g$ such that $|corr(f_k^{(m)}, \mathrm{f}_g)| = \max_c |corr(f_k^{(m)}, \mathrm{f}_c)|,\ c = 1, \ldots, G$.

The result of post-processing for the two examples is visualized in the bottom panels of Figures 5 and 6. The right scatter plots of factor correlations confirm that factor draws are well sorted out into both modes and the clustering allows for factor identification. The left scatter plots of factor loadings reflect two well identified modes for each setting, too. For $K = 2$, the two modes correspond to the ones we discerned from the scatter plot of the unsorted draws, one where both factor loadings are different from zero and the other one where one loading is shrunk towards zero. For $K = 4$, the scatter plot of sorted factor loadings reveals that the loading structure of two factors (and in fact these two factors) coincide across both modes, whereas the loading structure of the other two factors differ across modes. The characteristics plotted for one series carry over to loadings of all other series. We discuss these and further results in more details in Section 5.

5. Simulation study

We analyze two basic settings. In the first one, the data generating process (DGP) consists of two factors and two different underlying factor loading structures of about equal sparsity degree. In the second one, the DGP consists of three or four factors, where one or two of the factors are so-called *pervasive factors*. These are present in both underlying factor loading structures. The remaining factors are *local* or *unit-specific factors* with different loading structures of about equal sparsity degree. For each setting we simulate various scenarios.

We report results based on factor models estimated with the true number of factors K. Appendix D illustrates that the post-processing procedures recover the true number of factors when models are estimated for an overfitting number of factors. The recommendation for empirical analyses is to over- rather than underfit the number of factors in a first round.

5.1. $K = 2$ factors, two underlying sparse loading structures

We simulate data driven by two static factors and two underlying loading structures with overall 50% or 80% sparsity, denoted as *data50* or *data80*, respectively. The subspaces implied by the two different underlying sparse loading structures are minimally correlated with each other. Appendix C.1 details how to construct minimally correlated subspaces.

For each sparsity degree, we simulate loadings under an exact sparse pattern (*ex*), with exact zero loadings, or an approximate pattern (*ap*), with "noisy" zeros. Factors and idiosyncratic errors in some scenarios satisfy Thurstone's assumptions exactly (*thur*). The variance of the idiosyncratic errors is either large or low (*ln*), resulting in signal-to-noise ratios of approximately 0.8 to 1 and 4 to 5, respectively. These settings yield 16 scenarios from which we simulate $N = 40$ units of length $T = 100$ each.

The output of the unconstrained rotation sampler, a MCMC chain of length $M = 200,000$, is post-processed as described in Subsection 4.1. The algorithm is applied to find two distinct modes, penalizing the first mode when optimizing towards the second one (see (17)). Throughout, the HPD intervals are constructed with $\alpha = 0.05$. Table 1 displays a comparison of each estimated mode with the closest simulated mode. We report for each mode the number and the average absolute values of false zeros and false non-zeros, and the Jaccard and simple matching coefficients between simulated and estimated modes. To account for the effect of setting loadings to zero, reported average values in this table are those obtained from re-estimating the model conditional on the identified sparse loading pattern.

Table 1: $K = 2$, unconstrained rotation, $\alpha = 0.05$. The second column dislays which simulated mode was detected first and second. Absolute true and estimated average values are reported for, respectively, false zeros and non-zeros.

Scenario	Ordering	False zeros		False non-zeros		Matching indices	
		Number	Average	Number	Average	Jaccard	Simple score
data50ex_thur_ln	1	0	-	0	-	1.00	1.00
	2	0	-	0	-	1.00	1.00
data50ex_ln	1	0	-	0	-	1.00	1.00
	2	0	-	0	-	1.00	1.00
data50ap_thur_ln	1	1	0.12	1	0.10	0.95	0.98
	2	0	-	0	-	1.00	1.00
data50ap_ln	1	1	0.12	1	0.09	0.95	0.98
	2	0	-	0	-	1.00	1.00
data80ex_thur_ln	2	0	-	0	-	1.00	1.00
	1	0	-	0	-	1.00	1.00
data80ex_ln	2	0	-	0	-	1.00	1.00
	1	0	-	0	-	1.00	1.00
data80ap_thur_ln	2	2	0.13	1	0.08	0.91	0.96
	1	0	-	0	-	1.00	1.00
data80ap_ln	2	1	0.13	6	0.10	0.82	0.91
	1	0	-	4	0.11	0.87	0.95
overall average		0.31	0.12	0.81	0.10	0.97	0.99
data50ex_thur	1	0	-	0	-	1.00	1.00
	2	0	-	0	-	1.00	1.00
data50ex	1	0	-	0	-	1.00	1.00
	2	0	-	0	-	1.00	1.00
data50ap_thur	1	1	0.12	0	-	0.98	0.99
	2	1	0.13	0	-	0.98	0.99
data50ap	1	1	0.12	0	-	0.98	0.99
	2	1	0.13	0	-	0.98	0.99
data80ex_thur	2	0	-	0	-	1.00	1.00
	1	0	-	0	-	1.00	1.00
data80ex	2	0	-	0	-	1.00	1.00
	1	2	0.56	0	-	0.92	0.98
data80ap_thur	2	2	0.12	0	-	0.94	0.98
	1	3	0.23	0	-	0.89	0.96
data80ap	2	2	0.13	0	-	0.94	0.98
	1	3	0.37	0	-	0.89	0.96
overall average		1.00	0.24	0.00	-	0.97	0.99

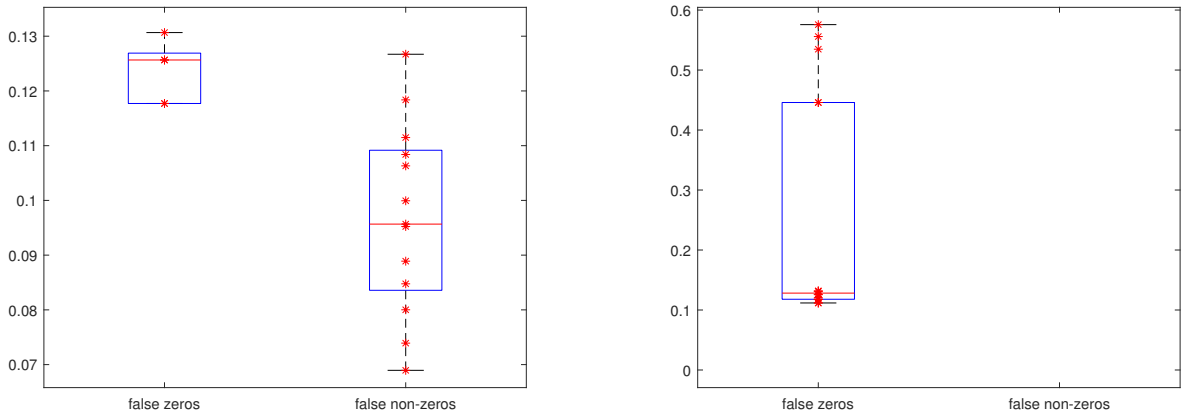
The upper part of the table displays the results for the low-noise scenarios. Both simulated modes are always perfectly recovered for the exact sparsity scenarios. In scenarios with approximate sparsity, one of the modes is perfectly recovered in almost all cases, with only 1 or 2 false zeros or non-zeros. One exception (data80ap_ln) produces a larger number of false non-zeros. However, their magnitude is small.

The lower part of the table displays the results for the high-noise scenarios. No false non-zeros are obtained in anyone of them. In three cases, both modes are recovered perfectly, and one mode is perfectly recovered in one case (data80ex). The number of false zeros is small, ranging from 1 to 3, but the average value of true loadings is

sometimes larger in magnitude, reaching up to 0.56. Apparently, a lower signal-to-noise ratio induces the procedure to occasionally identify slightly more sparsity than simulated. Figure 7 provides a graphical illustration of average absolute values of false zeros and non-zeros across scenarios.

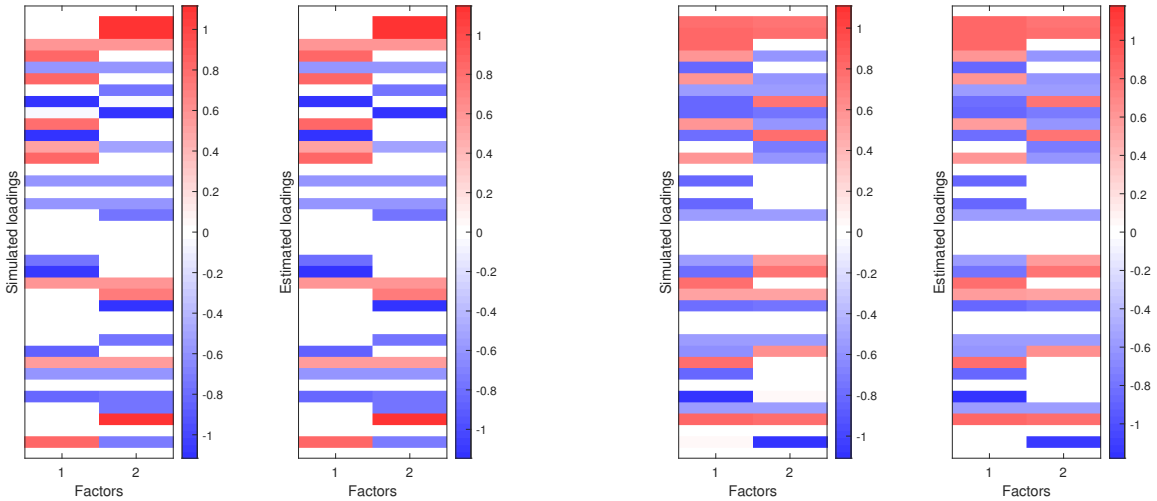
Figure 8 displays heat plots of simulated (left) and estimated (right) loadings for both modes of the scenario data50ex_thur_ln. They confirm that sparse patterns are well recovered by the optimization procedure described in Subsection 4.1.

Figure 7: $K = 2$, unconstrained rotation. Boxplot of average absolute values of factor loadings, pooled across scenarios. The centerline is the median, the edges correspond to the 25th and 75th percentiles (IQR), while the whiskers extend 1.5 times IQR beyond the edges. Note: No false non-zeros in the high-noise scenarios, hence no boxplot to display.



Low-noise ratio High-noise ratio

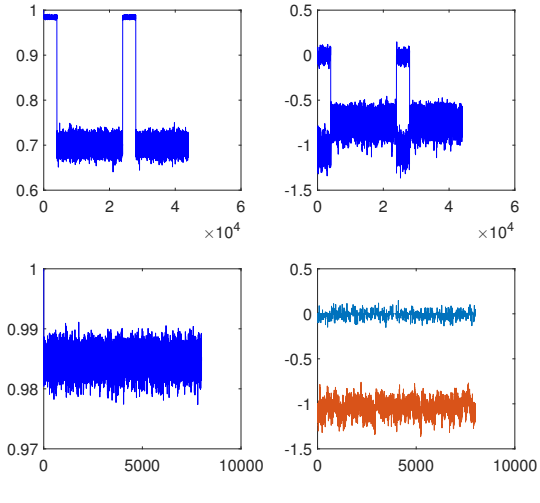
Figure 8: $K = 2$, unconstrained rotation, scenario data50ex_thur_ln. Heat plot of posterior mean factor loadings.



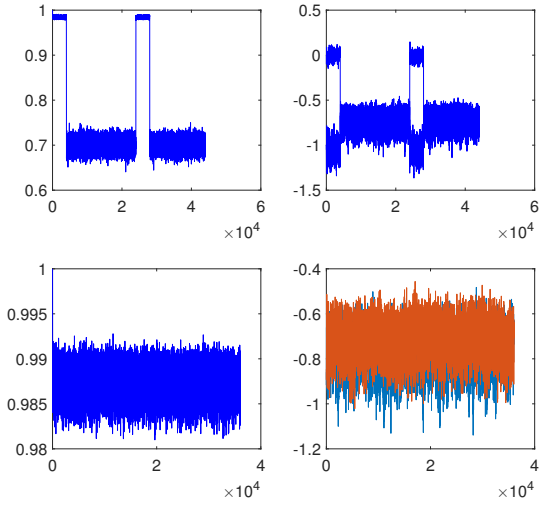
Mode 1 Mode 2

Applying the sparse permutation sampler described in Subsection 3.2, we estimate each scenario with 11 chains of 10,000 draws. After running an initial chain, starting values for 10 parallel chains are obtained by random orthonormal rotation of a draw for factor loadings of this initial chain. By retaining the last 4,000 of each chain, we obtain 44,000 draws for posterior inference.

Figure 9: $K = 2$, sparse permutation, scenario data50ex_thur_ln. Posterior draws, unsorted and sorted. From top left to bottom right: Correlation of the first with all other posterior draws of factor 1, posterior draws of a selected row of Λ, correlation of the first with all other sorted posterior draws of mode-specific factor 1, sorted posterior draws of a mode-specific row of Λ.



(a) Mode 1 (8,000 sorted draws)



(b) Mode 2 (36,000 sorted draws)

The first line in Figure 9 shows the unsorted draws from the sparse permutation sampler for the scenario data50ex_thur_ln. In each panel, the left figure plots the correlations of the first draw for the first factor with all remaining draws, while the right panel plots the unsorted draws for a selected row of Λ. After clustering and re-ordering the draws accordingly, 8,000 draws are allocated to the first mode (second line, left panel). The correlation of the first draw for the first factor with all remaining draws is close to 1, and the factor loadings for the selected row of Λ are all located near 0 and near -1, respectively. Accordingly, the remaining 36,000 draws are allocated to the second mode (second line, right panel). The correlation of the first draw for the first factor with all remaining draws is close to 1 here as well, and the factor loadings for the selected row of Λ are all located near -0.7 and -0.8, respectively. This implies that the first mode has a nonzero loading in the chosen row of Λ only for one of the factors, whereas the second mode has nonzero loadings in the chosen row of Λ for both factors.

Table 2 provides an overview of the estimation results obtained with the sparse permutation sampler. The number of false zero and non-zero loadings is somewhat higher than for the unconstrained rotation approach. However, the true loadings for false zeros are overall small in absolute value. The average absolute value of false non-zeros is in the same range as for the unconstrained rotation approach. Figure 10 provides a graphical illustration of the absolute values of false zeros and non-zeros across low- and high-noise scenarios. Overall, deviations of estimated from true

Table 2: $K = 2$, sparse permutation. The first (second) line evaluates the first (second) mode. The second column reports the number of posterior draws assigned to the respective mode. The posterior median of absolute true and estimated average values are reported for, respectively, false zeros and non-zeros.

Scenario	Draws	False zeros		False non-zeros		Matching indices	
		Number	Average	Number	Average	Jaccard	Simple score
data50ex_thur_ln	8,000	8	0.03	0	-	0.83	0.90
	36,000	2	0.05	0	-	0.96	0.97
data50ex_ln	4,000	8	0.03	0	-	0.83	0.90
	40,000	2	0.05	2	0.08	0.92	0.95
data50ap_thur_ln	4,000	0	-	1	0.08	0.98	0.99
	40,000	0	-	1	0.03	0.98	0.99
data50ap_ln	32,000	0	-	6	0.14	0.87	0.93
	12,000	0	-	4	0.08	0.92	0.95
data80ex_thur_ln	16,000	0	-	0	-	1.00	1.00
	28,000	2	0.04	0	-	0.93	0.97
data80ex_ln	40,000	0	-	2	0.08	0.94	0.97
	4,000	2	0.04	1	0.07	0.89	0.96
data80ap_thur_ln	28,000	2	0.13	1	0.15	0.91	0.96
	16,000	0	-	0	-	1.00	1.00
data80ap_ln	36,000	3	0.12	3	0.10	0.83	0.93
	8,000	2	0.12	3	0.12	0.83	0.94
overall average		1.9	0.05	1.5	0.10		
data50ex_thur	20,000	8	0.03	0	-	0.83	0.90
	24,000	2	0.05	0	-	0.96	0.97
data50ex	16,000	7	0.03	3	0.17	0.80	0.88
	28,000	2	0.05	6	0.07	0.86	0.90
data50ap_thur	32,000	1	0.12	0	-	0.97	0.99
	12,000	1	0.13	0	-	0.98	0.99
data50ap	44,000	1	0.12	5	0.12	0.87	0.93
	-	-	-	-	-	-	-
data80ex_thur	20,000	0	-	0	-	1.00	1.00
	24,000	2	0.04	0	-	0.93	0.97
data80ex	10,997	0	-	11	0.12	0.72	0.86
	33,001	2	0.04	8	0.16	0.71	0.88
data80ap_thur	28,000	3	0.12	0	-	0.91	0.96
	16,000	2	0.12	0	-	0.93	0.97
data80ap	8,939	2	0.13	4	0.16	0.83	0.93
	27,135	2	0.12	5	0.16	0.78	0.91
overall average		2.2	0.07	2.6	0.13		

values are small.

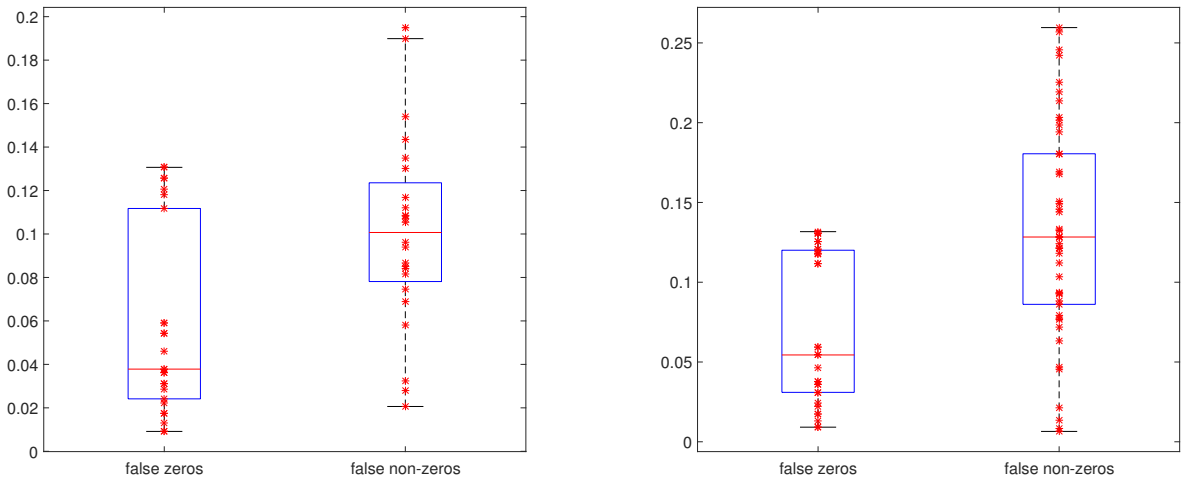
The heat plots for each mode of the factor loading matrices is shown in Figure 11, where simulated and estimated structures are displayed on, respectively, the left and right side. Note that the sign of estimated loadings has been adjusted such that the majority of loadings is positive for each factor. Therefore, the sign of estimated loadings is opposite to the simulated ones. The post-processing procedure recovers well both underlying simulated sparse structures.

5.2. Simulated $K = \{3, 4\}$ with pervasive and local factors

For $K = 3$ we simulate a pervasive factor (*1pf*), i.e. a strong factor driving all variables, and two weaker, i.e. local or group-specific, factors, which can be represented by two underlying sparse loading structures. For $K = 4$, we simulate one or two pervasive factors (*1pf* or *2pf*) complemented with, respectively, three or two weaker factors. Again, we simulate data with a high (*ln*) and low signal-to-noise ratios. Combining these features yields six settings, from which we simulate $N = 60$ series of length $T = 100$ each.

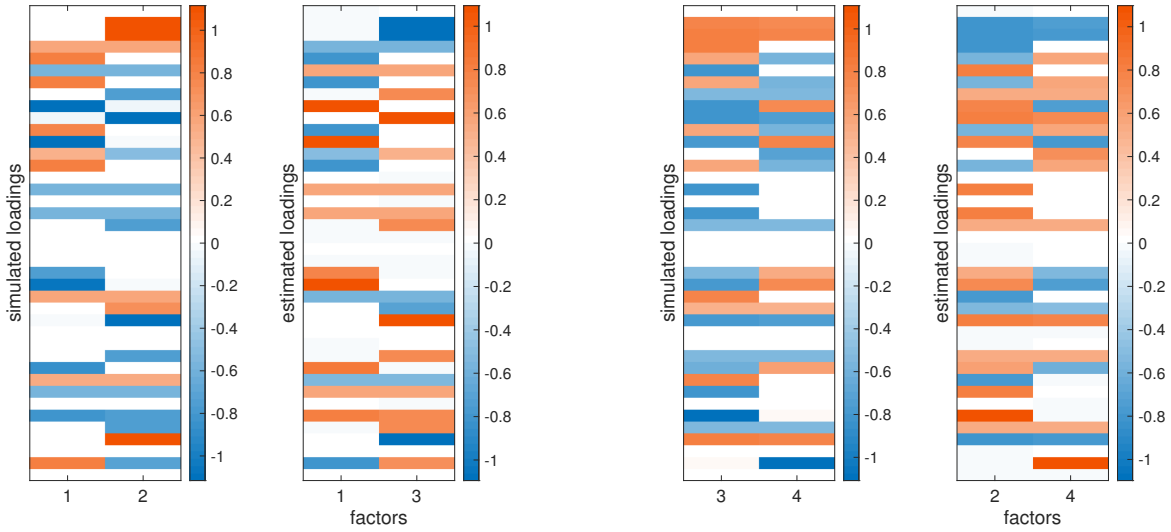
The post-processing approach described in Subsection 4.1 is applied to sequences of length 200,000 obtained from

Figure 10: $K = 2$, sparse permutation, pooling over scenarios and factors. Boxplot of absolute median factor loadings. The centerline is the median, the edges correspond to the 25th and 75th percentiles (IQR), while the whiskers extend 1.5 times IQR beyond the edges.



Low-noise ratio High-noise ratio

Figure 11: $K = 2$, sparse permutation, scenario data50ex_thur_ln. Heat plot of posterior median factor loadings.



Mode 1 Mode 2

the unconstrained rotation sampler. The algorithm is again guided to identify two distinct modes, penalizing the first mode when optimizing towards the second mode. Throughout, the HPD intervals are constructed with $\alpha = 0.05$.

Table 3 displays the comparison between estimated and the closest simulated modes. We report the number of false zeros and non-zeros for each mode, including the Jaccard and simple matching coefficients between simulated and

estimated factor loading structures. We also report the average absolute value across false zero and non-zero loadings. For the scenario K3m2_1pf_ln, both modes are perfectly recovered. For the corresponding high-noise scenario, there are 12 false zeros in both modes, with an average absolute true value of around 0.3. For the scenarios with $K = 4$ factors and one pervasive factor, the number of false zeros increases to 39, and to 12 for false non-zeros. Especially in the scenario K4m2_2pf, the average across absolute values of false non-zeros is around 0.4 or 0.5, which indicates that the estimated sparse representations differ somewhat from the simulated ones. For the scenarios with two pervasive factors, the number and the average absolute value of false zeros seem very large. We provide an explanation for this feature further below. The left panel in Figure 12 visualizes the results of Table 3. We display box-plots of average absolute true values of false zeros and estimated values of false non-zeros, pooled across scenarios with two pervasive factors.

Table 3: $K = 3$, $K = 4$, unconstrained rotation, $\alpha = 0.05$. The second column dislays which simulated mode was detected first and second. Absolute true and estimated average are reported for, respectively, false zeros and non-zeros.

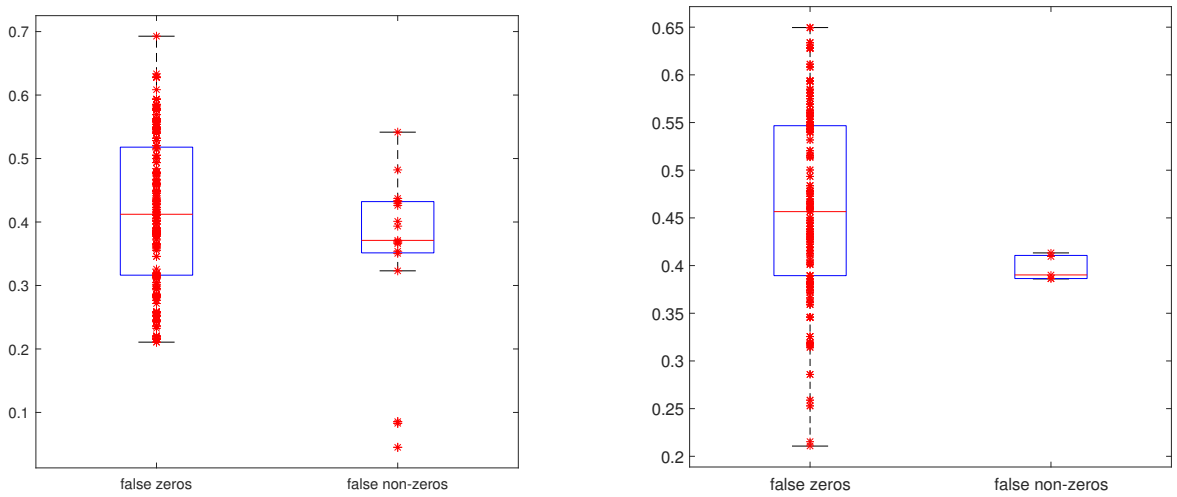
Scenario	Ordering	False zeros		False non-zeros		Matching indices	
		Number	Average	Number	Average	Jaccard	Simple score
K3m2_1pf	2	12	0.34	0	-	0.88	0.93
	1	12	0.30	0	-	0.88	0.93
K3m2_1pf_ln	2	0	-	0	-	1.00	1.00
	1	0	-	0	-	1.00	1.00
K4m2_1pf	2	39	0.28	9	0.24	0.66	0.80
	1	35	0.26	3	0.10	0.72	0.84
K4m2_1pf_ln	1	24	0.24	12	0.50	0.75	0.85
	2	24	0.22	12	0.44	0.75	0.85
K4m2_2pf	1	84	0.40	1	0.40	0.46	0.65
	2	80	0.41	4	0.16	0.47	0.65
K4m2_2pf_ln	2	53	0.46	13	0.39	0.65	0.78
	1	67	0.42	1	0.54	0.52	0.67

Figure 13 displays heat plots for factor loadings of the scenario K4m2_2pf, for simulated and estimated loadings on the left and right side, respectively, in each panel. In both top panels we see that the estimated loading structure for pervasive factors is sparser than the simulated structure, which reflects the large number of false zeros reported in Table 3. However, the post-processing procedure indeed optimizes the rotation to induce sparsity. The bottom panels display a Varimax rotated version of the simulated pervasive factors in each left heat plot. We see that the post-processing procedure identifies a sparse structure for pervasive factors that is very similar to a Varimax rotation of simulated loadings. The deviations between simulated and estimated structures for the weak factors reflect the results reported in Table 3.

The sparse permutation sampler was run with 16 chains of 10,000 draws. After an initial chain, starting values for factor loadings are obtained by random orthonormal rotations of a factor loading draw taken from the initial chain. We again retain the last 4,000 of each chain to obtain 64,0000 draws for posterior inference.

Table 4 provides an overview of the estimation results obtained with the sparse permutation sampler. Note that for all scenarios, the number of draws assigned to each mode do not sum up to 64,000. While the sum across modes is only slightly below 64,000 for the scenarios K3m2_1pf and K4m2_2pf_ln, the number of draws not assigned to

Figure 12: Scenario K4m2_2pf and K4m2_2pf_ln, pooling over modes. Boxplot of absolute factor loadings. The centerline is the median, the edges correspond to the 25th and 75th percentiles (IQR), while the whiskers extend 1.5 times IQR beyond the edges.

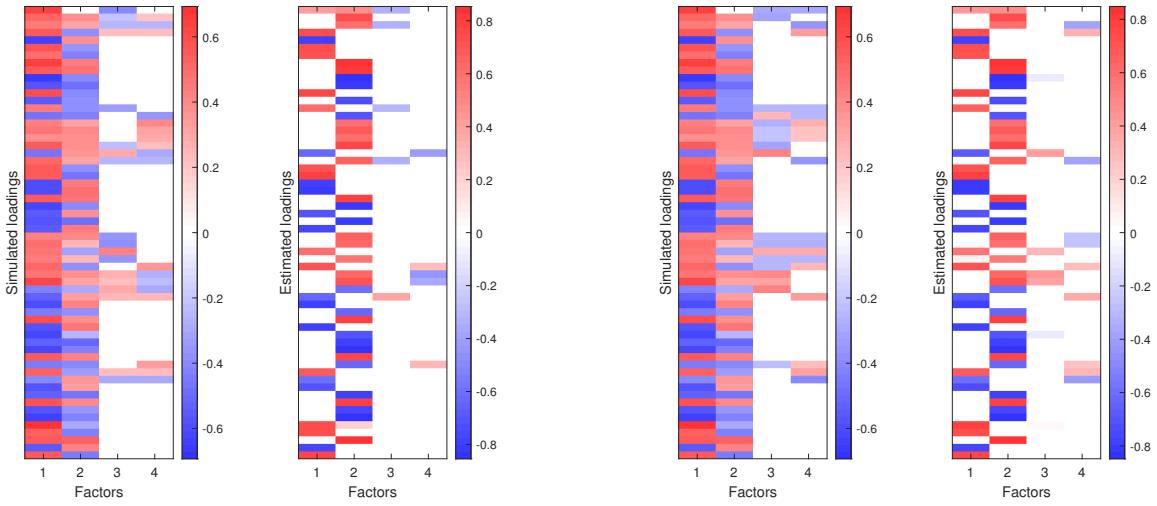


(a) Unconstrained rotation

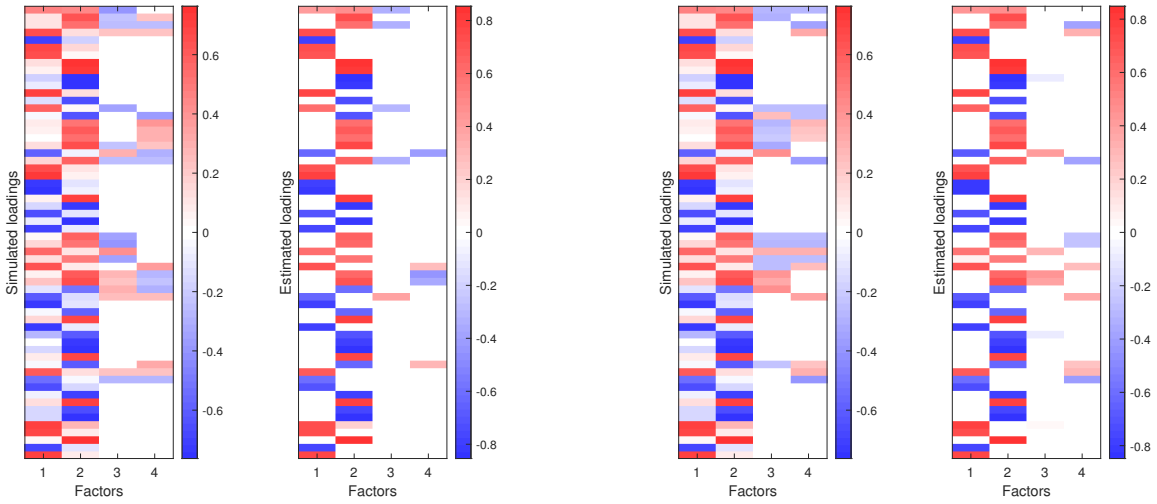
(b) Sparse permutation

one of simulated modes is substantially larger in the remaining scenarios. Nonetheless, for the scenarios with $K = 3$ factors, both modes are identified perfectly, and in the scenarios with $K = 4$ factors, there are only very few false non-zero loadings across all scenarios, and the number of false zeros is much lower than the one obtained by posterior rotation (see Table 3). When compared with the output obtained by posterior rotation, the true loadings of false zeros are smaller in magnitude for the scenarios with $K = 4$ and one pervasive factor, and similar for the scenarios with $K = 4$ and two pervasive factors (see also the right panel in Figure 12). Note that for these scenarios, the number of false zeros and average absolute values of true factor loadings are quit large. The sparsity induced by the sparse permutation sampler on loadings of the pervasive factors comes close to a Varimax rotation of the simulated loadings, see Figure 14 which displays heatplots of factor loadings. We also observe that the loading structure of the two weaker factors is recovered quite well for both modes, which clearly outperforms posterior rotation.

Figure 13: $K = 4$, unconstrained rotation, scenario K4m2_2pf. Heat plot of posterior mean factor loadings. Second line: Varimax rotation of simulated loadings for the two pervasive factors.



Mode 1 Mode 2

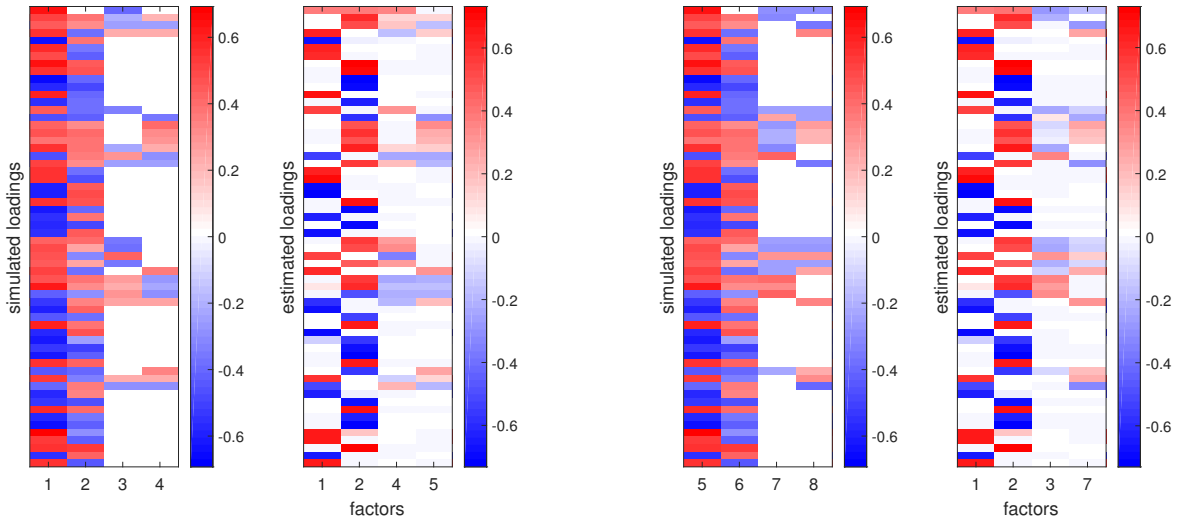


Mode 1 Mode 2

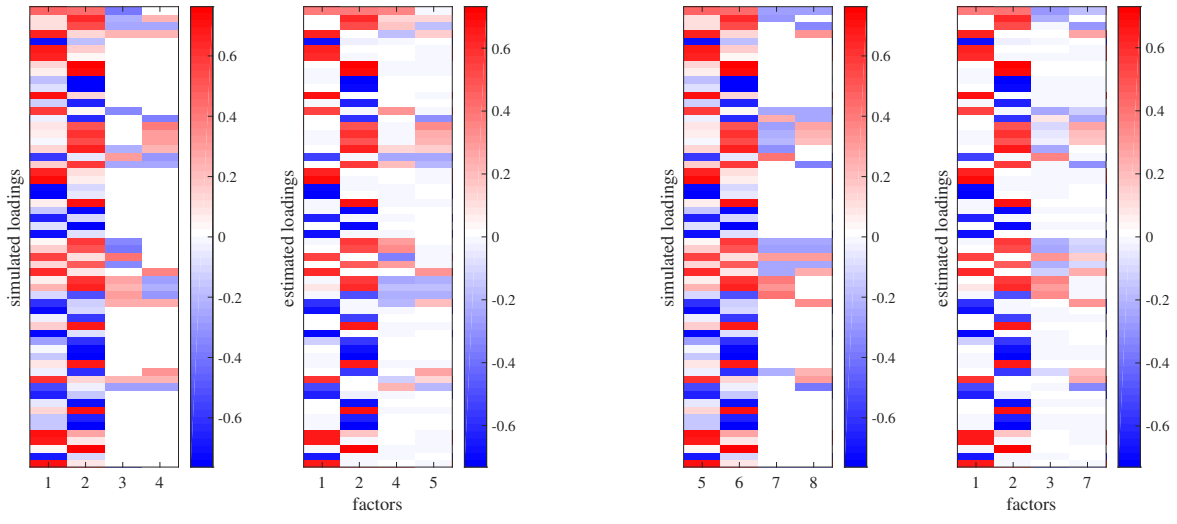
Table 4: $K = 3$, $K = 4$, sparse permutation. The first (second) line evaluates the first (second) mode. The second column reports the number of posterior draws assigned to the respective mode. The posterior median of absolute true and estimated average values are reported for, respectively, false zeros and non-zeros.

Scenario	Draws	False zeros		False non-zeros		Matching indices	
		Number	Mean	Number	Mean	Jaccard	Simple score
K3m2_1pf	27,999	0	-	0	-	1.00	1.00
	35,959	0	-	0	-	1.00	1.00
K3m2_1pf_ln	8,000	0	-	0	-	1.00	1.00
	43,813	0	-	0	-	1.00	1.00
K4m2_1pf	22,661	15	0.14	0	-	0.89	0.94
	13,712	10	0.14	0	-	0.92	0.96
K4m2_1pf_ln	4,000	0	-	0	-	1.00	1.00
	2,895	12	0.16	18	0.24	0.80	0.88
K4m2_2pf	20,670	57	0.48	1	0.39	0.63	0.76
	38,661	59	0.47	1	0.39	0.62	0.75
K4m2_2pf_ln	12,000	41	0.46	1	0.41	0.73	0.82
	51,898	40	0.46	1	0.41	0.74	0.83

Figure 14: $K = 4$, sparse permutation, scenario K4m2_2pf. Heat plot of posterior median factor loadings. Second line: Varimax rotation of simulated loadings of the two pervasive factors.



Mode 1 Mode 2



Mode 1 Mode 2

6. Applications

To illustrate the sampler and the post-processing procedures, we revisit the datasets used in Kaufmann and Schumacher (2017), namely monthly inflation series in US sectoral CPI components (Mackowiak et al., 2009) and yearly GDP growth rates of a multi-country panel used in Francis et al. (2017). To analyze the datasets, we extend the specification to include p autoregressive terms to capture factor dynamics and q terms to capture (independent) idiosyncratic dynamics. In the next subsection, we briefly expose the extended model specification and the conditional posterior distributions to sample the dynamic parameters. We justify that the post-processing algorithm described in Subsection 4.1, based on orthonormal rotations of an orthonormal factor basis, can be applied to the MCMC output of a dynamic factor model. Empirical results follow in the next two subsections.

6.1. Extending to dynamic factors and idiosyncratic components

To capture the observed persistence in time series, model (2)-(3) is extended:

$$y_t = \Lambda f_t + \varepsilon_t,\ \epsilon_t \sim i.i.d.\ N(0, \Sigma_\epsilon), \tag{18}$$

$$f_t = \Phi_1 f_{t-1} + \cdots + \Phi_p f_{t-p} + \nu_t,\ \nu_t \sim i.i.d.\ N(0, I_K) \tag{19}$$

$$\varepsilon_t = \Psi_1 \varepsilon_{t-1} + \cdots + \Psi_q \varepsilon_{t-q} + \epsilon_t,\ \epsilon_t \sim i.i.d.\ N(0, \Sigma_\epsilon). \tag{20}$$

The parameter vector θ includes also $\mathbf{\Phi} = \{\Phi_1, \ldots, \Phi_p\}$ and $\mathbf{\Psi} = \{\Psi_1, \ldots, \Psi_q\}$, and the likelihood is formulated in terms of filtered series $\tilde{y}_t = \Psi(L)y_t = y_t - \Psi_1 y_{t-1} - \cdots - \Psi_q y_{t-q}$ and $\tilde{f}_t = (\psi_1(L)f_t', \ldots, \psi_N(L)f_t')'$

$$L(\tilde{\mathbf{y}}|\tilde{\boldsymbol{f}}, \theta) = \prod_{t=1}^{T} \pi(\tilde{y}_t|\tilde{f}_t, \theta), \tag{21}$$

with normal observation density

$$\pi(\tilde{y}_t|\tilde{f}_t, \theta) = \frac{1}{\sqrt{2\pi}|\Sigma_\epsilon|^{1/2}} \exp\left\{-\frac{1}{2}\left(\tilde{y}_t - \mathbf{\Lambda}\tilde{f}_t\right)' \Sigma_\epsilon^{-1}\left(\tilde{y}_t - \mathbf{\Lambda}\tilde{f}_t\right)\right\},$$

with each row i of Λ as ith block-diagonal element of $\mathbf{\Lambda}$. Conditionally, factors are independent $f_t|f^{t-1}, \mathbf{\Phi} \sim N\left(\Phi_1 f_{t-1} + \cdots + \Phi_p f_{t-p}, I_K\right)$. The prior reflects this conditional independence and is specified in terms of filtered factors $\bar{f}_t = \Phi(L)f_t$, $\pi(\bar{\boldsymbol{f}}) = N(0, \mathrm{F}_0)$, with F_0 block-diagonal with elements $\mathrm{F}_0^{(i)}$ for initial conditions $\{f_0, \ldots, f_{-p+1}\}$ and I_K for f_t, $t = 1, \ldots, T$.

Posterior inference is obtained via the posterior sampler described in Subsection 3.2, applied to filtered data and including two additional sampling steps to update parameters $\{\mathbf{\Phi}, \mathbf{\Psi}\}$. In Step 3., in addition to sample from 3.i $\pi\left(\Sigma_\epsilon|\tilde{\mathbf{y}}, \tilde{\boldsymbol{f}}, \Lambda\right)$, we sample from

3.ii $\pi(\mathbf{\Phi}|\boldsymbol{f}) = N(\mathrm{p}, \mathrm{P})$, where moments P and p are derived based on the vector-autoregression (19).

3.iii $\pi(\mathbf{\Psi}|\boldsymbol{y}, \boldsymbol{f}, \Lambda, \Sigma_\epsilon) = \prod_{i=1}^{N} N(\mathrm{q}_i, \mathrm{Q}_i)$ where moments Q_i and q_i are derived independently for each series based on

(20), i.e.

$$\varepsilon_{it} = \psi_{i1}\varepsilon_{i,t-1} + \cdots + \psi_{iq}\varepsilon_{i,t-q} + \epsilon_{it},\ \epsilon_{it} \sim i.i.d.\ N(0, \sigma_i) \text{ and } \varepsilon_{it} = y_{it} - \lambda_{i\cdot} f_t.$$

At the end of each iteration, random rotation or permutation is applied to factor-specific parameters $\mathbf{\Phi}$, too. Both post-processing procedures can be applied as described in Section 4. In particular, given that factors are conditionally independent and prior and posterior distributions reflect this independence, we can post-process the MCMC sample of factors by orthonormal rotation to detect sparse modes.

6.2. *Monthly CPI sectoral inflation rates*

The dataset contains $N = 79$ sectoral inflation series covering the period February 1985 to May 2005, $T = 244$. We estimate a model with $K = 2$ factors, include $p = 4$ and $q = 2$ factor and idiosyncratic autoregressive terms, respectively, which reflects results documented in Mackowiak et al. (2009). Mackowiak et al. (2009) preferred a model with one over two factors, although results remain basically unchanged when including two factors. We revisit the dataset to evaluate whether the uncertainty about the number of factors may be due to underlying weak factors.

Figure 15: US CPI, uncontrained rotation. Mean estimated factor loadings

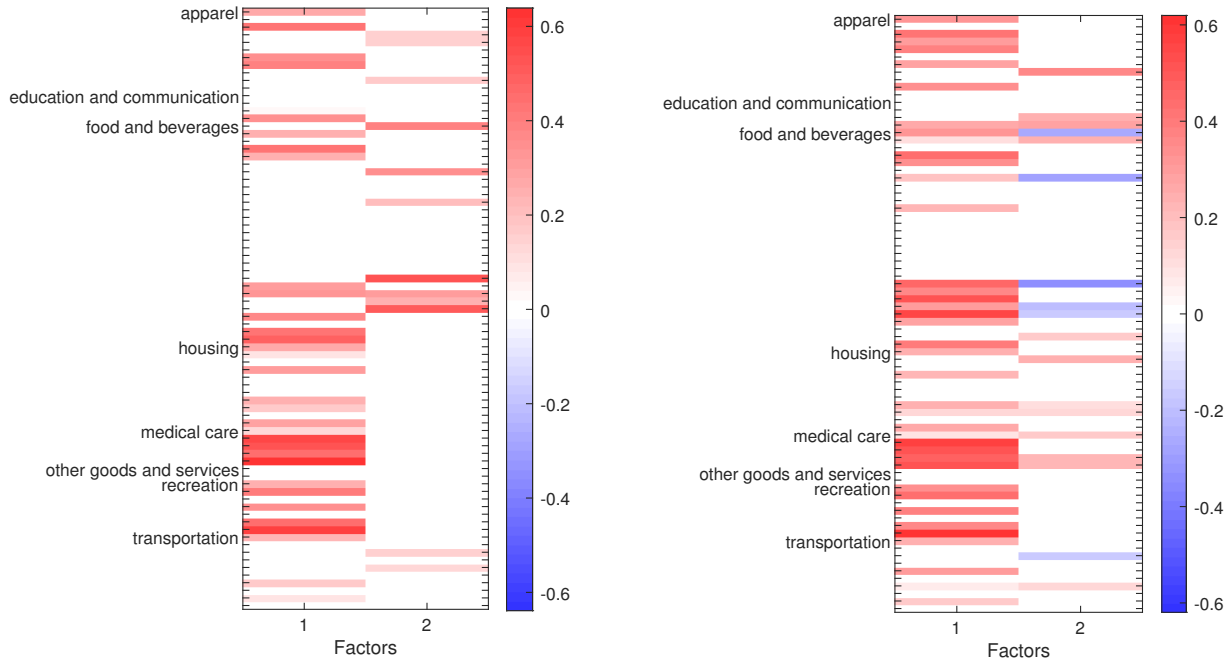
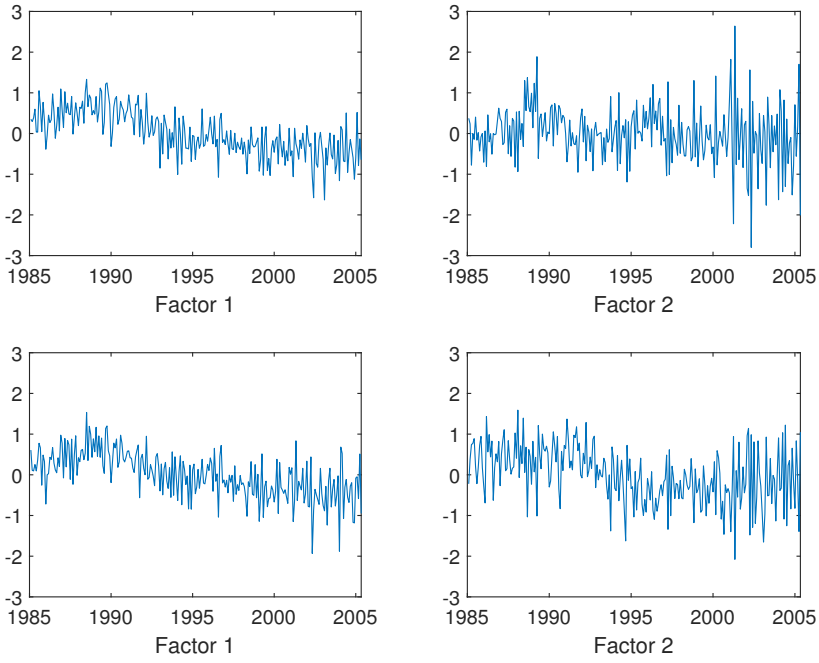


Figure 15 shows the loadings patterns identified by the unconstrained rotation approach. For the first mode (left panel) there are 33 non-zero loadings on the first factor and 12 non-zero loadings on the second factor. All loadings

are nonnegative. For the second mode (right panel), there are 39 non-zero loadings on the first factor and 18 non-zero, some of them negative, loadings on the second factor.

Were the first factor pervasive, factor loadings should be similar across modes. This is not quite the case, as shown in Figure 15 and hard to assess from mean factor plots displayed in Figure 16. However, the correlation across first factors of both modes is 0.86, while the correlation across factors in each mode is as low as 0.18 and 0.09 for Mode 1 and 2, respectively.

Figure 16: US CPI, unconstrained rotation. Mean factors of Mode 1 (first line) and Mode 2 (second line).



Using the sparse permutation sampler, the results are based on 13 chains of 11,000 draws, retaining the last 5,000, obtaining 65,000 draws for posterior inference. In a first round, clustering factor draws based on correlations we identify one pervasive factor. Therefore, we set $G = 3$ to post-process factor draws as described in Appendix B.4 setting $e_0 = .1(K/2-1)$. Each draw is assigned to one of the three components, potentially allowing for $\binom{3}{2} = 3$ factor combinations $\mathcal{I}_Z = \{Z_1, Z_2\} \subset \{1, 2, 3\}$, all Z_k different.

Sorting out the draws, only two factor combinations are visited. Table 5 reports that almost all of the 65,000 draws can be assigned to either of the two modes. We identify 1 pervasive factor and 2 weaker ones. Non-zero loadings are determined by loadings for which the median posterior probability of a non-zero factor loading is larger than 0.5, $q_{0.5}(\beta_{ij}^{(m)}) > 0.5$.

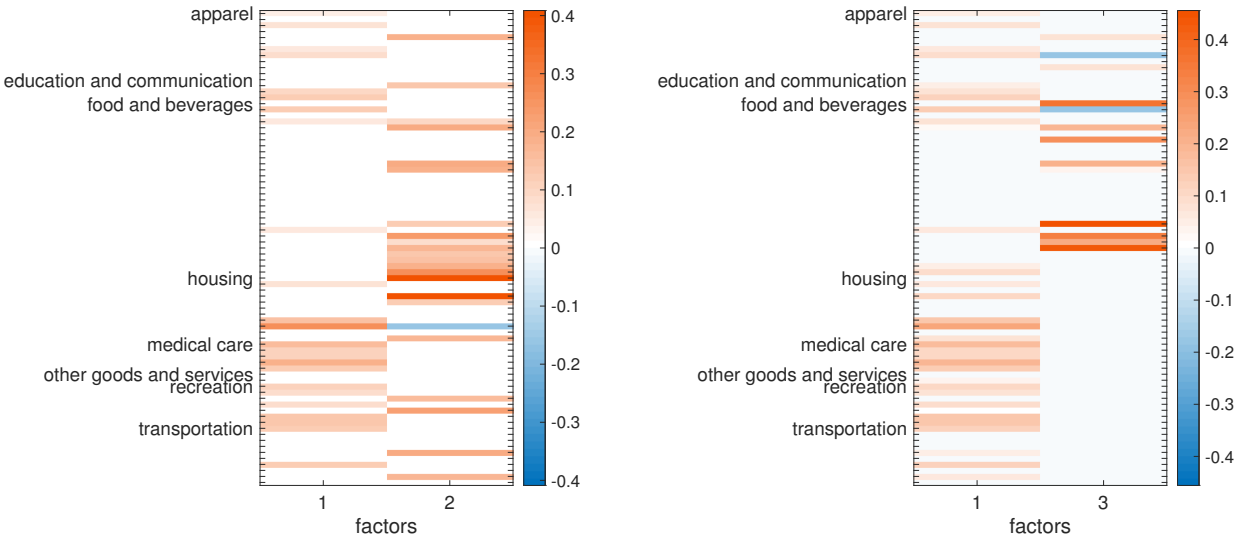
Figure 17 shows the loadings patterns. In the first mode, there are 25 non-zero loadings on the first factor and 23 on the second factor, while in the second mode, there are 34 non-zero loadings on the first factor and 13 on the second

factor. Negative loadings occur with the second factors, but are rare and overall close to zero.

Table 5: US CPI, sparse permutation. Sorted output, Jaccard matching indices computed across median loading matrices.

Factor combination	Draws	Non-zero loadings	Jaccard matching indices Compared to $\{1, 2\}$	
$\{1, 2\}$	53,086	25/23	-	-
$\{1, 3\}$	11,793	34/13	73.5	28.6

Figure 17: US CPI, sparse permutation. Median factor loadings



Mode 1 Mode 2

Looking at the correlations across mean factors (Figure 18), the correlation between the two first factors is virtually 1, while the two second factors show merely a correlation of about 0.5. The correlation between the two factors from both modes is 0.59 and 0.31, respectively, and hence somewhat larger than between the factors identified by the unconstrained rotation approach. The mean factors themselves are shown in Figure 19.

Overall, we conclude that both post-processing procedures yield evidence for a pervasive factor across two sparse modes and two weaker factors, each present in one mode.

6.3. Yearly GDP growth rates

The dataset contains $N = 57$ GDP growth series covering the years 1961 to 2009, $T = 49$. We estimate a model with $K = 4$ factors, include $p = 2$ and $q = 1$ factor and idiosyncratic autoregressive terms, respectively. We again revisit the data to uncover the number and characteristics of factors, i.e. whether a number of pervasive factors may be extracted with potentially differing local factors.

Figure 18: US CPI, sparse permutation. Factor correlations, across factor combinations.

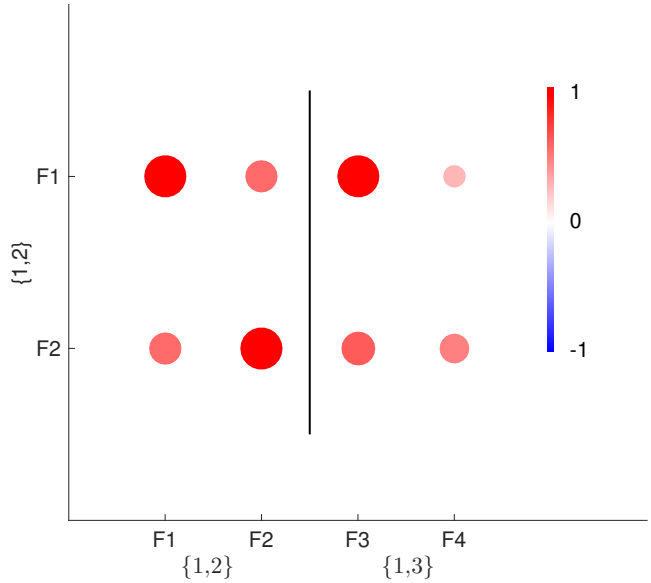


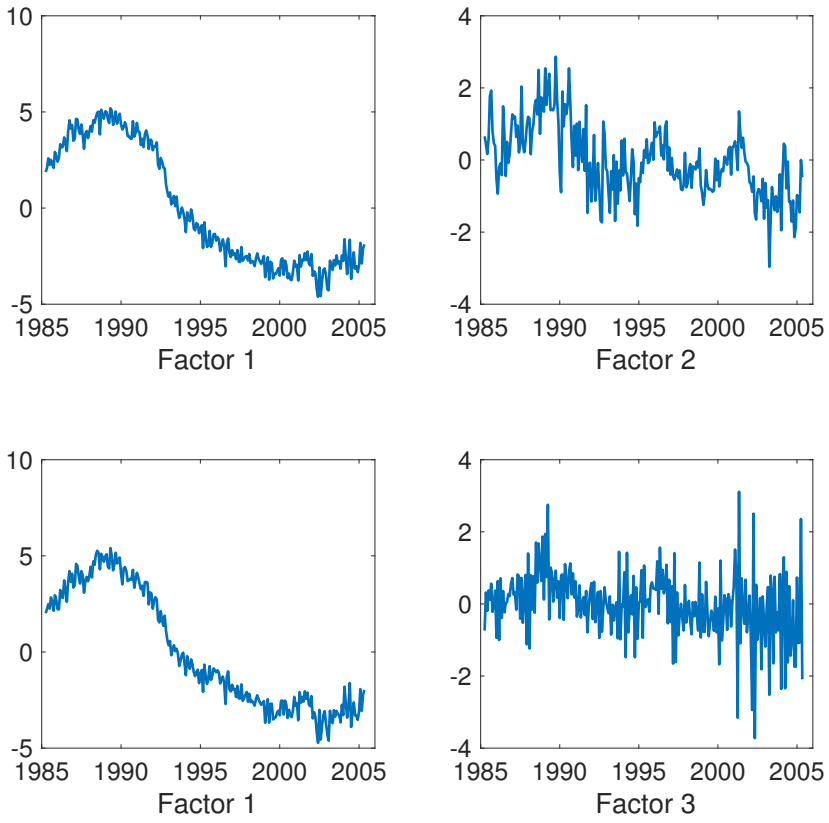
Figure 20 shows the loadings patterns identified by the unconstrained rotation approach. For the first mode (left panel), there are 5, 5, 22 and 12 non-zero loadings, respectively, for the four factors. There are two negative loadings, one on the first and one on the fourth factor, both close to zero. For the second mode (right panel), there are 9, 7, 22 and 11 non-zero loadings, respectively, on the four factors, with two negative loadings each on the first and second factors.

Figure 21 shows the factors, those corresponding to the first mode in the upper two rows, and those corresponding to the second mode in the lower two rows. Looking at the correlations between factors across modes, Factors 1, 2 and 3 correlate with, respectively, 0.99, 0.97 and 1 across modes, while the correlation between the fourth factor of each mode is somewhat lower.

For the sparse permutation sampler, the results are again based on 13 chains of 11,000 draws, retaining the last 5,000, obtaining 65,000 draws for posterior inference. Clustering factor draws in a first round based on correlations, we identify 3 pervasive factors. Therefore, we set $G = 7$ to post-cluster factor draws as described in Appendix B.4, setting $e_0 = 0.01(G/2 - 1)$ to allow for empty clusters. Each draw is assigned to one of seven components, potentially allowing for $\binom{7}{4} = 35$ factor combinations $\mathcal{I}_Z = \{Z_1, \ldots, Z_4\} \subset \{1, \ldots, 7\}$, all Z_k different.

Sorting out the draws, only three factor combinations are visited. Table 6 reports again that almost all of the 65,000 draws can be assigned to either of the three modes. We identify 3 pervasive factors and 3 weaker ones. Figure 22 displays the loadings patterns. The number of non-zero loadings on the three pervasive factors is virtually identical across the three modes, with 13 or 14 non-zero loadings on the first, 7 non-zero loadings on the second, and 41, 42 or 43 non-zero loadings on the third factor. Moreover, note that these non-zeros occur in the same places. For the fourth

Figure 19: US CPI, sparse permutation. Mean factors. Mode 1 (first line) and Mode 2 (second line).

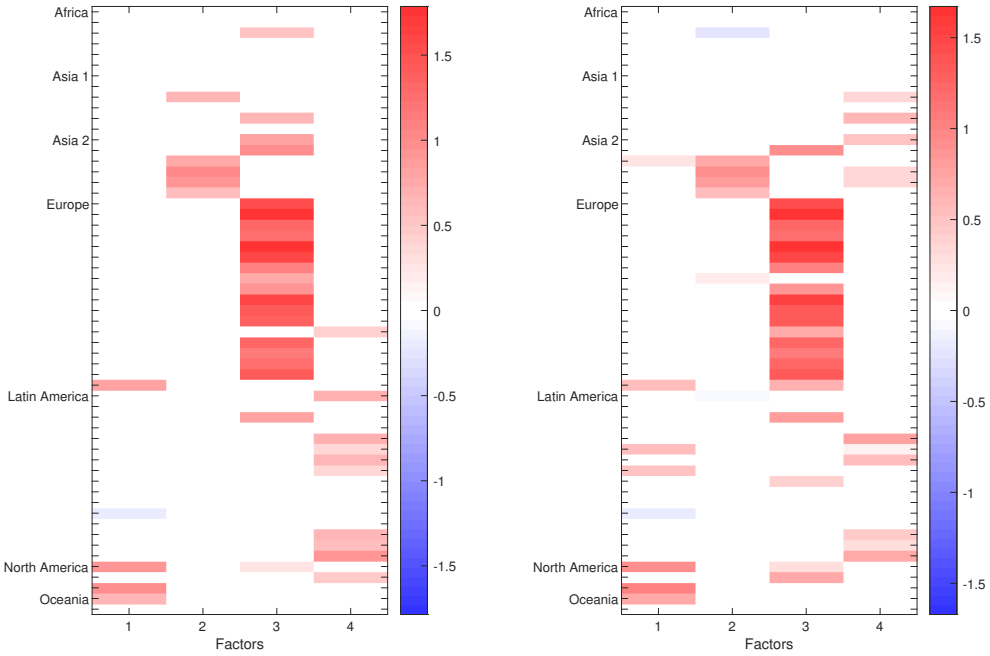


factor, there are between 11 and 16 non-zero loadings, and the location of these vary substantially across factors.

Looking at the correlations between factors across modes (Figure 23), the correlation between the first three pervasive factors across modes is virtually 1, while the correlations between the fourth factors across modes are close to zero. Correlations across factors of each mode are also low to moderate only. The mean factors themselves are shown in Figure 19.

We conclude that both post-processing procedures yield evidence for three pervasive factors. While post-processing the random rotation output yields two modes with two weaker factors, the post-processed output of the sparse sampler is able to identify three modes with one weak factor each. Sampling by random rotation is based on orthonormal rotations and likewise, when post-processing the MCMC output of the random rotation sampler, the optimization towards sparsity is based on orthonormal rotations. The sparse permutation sampler induces sparsity in the loading matrix while sampling, where the different sparse representations need not be nested within each other by orthonormal rotation. This may explain the ability of the sparse permutation sampler to uncover more weaker factors for this application.

Figure 20: GDP growth, unconstrained rotation: Mean factor loadings



Mode 1 Mode 2

Table 6: GDP growth, sparse permutation. Sorted output.

Factor combination	Draws	Non-zero loadings	Jaccard matching indices Compared to $\{1,2,3,4\}$			
$\{1,2,3,4\}$	16,181	14/7/43/11	-	-	-	-
$\{1,2,3,5\}$	18,037	13/7/42/16	92.9	1.0	97.7	3.9
$\{1,2,3,6\}$	30,473	13/7/41/16	80.0	1.0	95.4	3.9

Figure 21: GDP growth, unconstrained rotation. Mean factors. Mode 1 (first two row) and Mode 2 (bottom two rows).

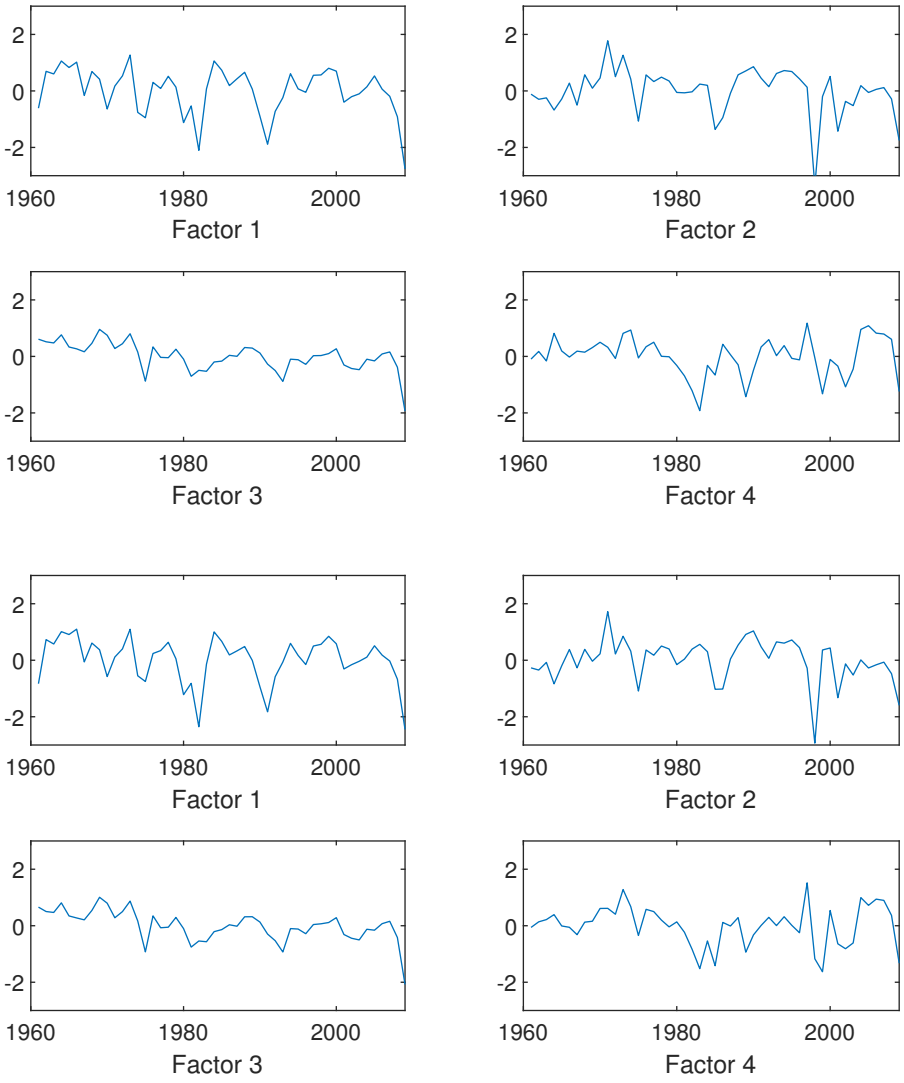


Figure 22: GDP growth, sparse permutation. Mean factor loadings, averaged over draws with a non-zero probability larger than 0.5.

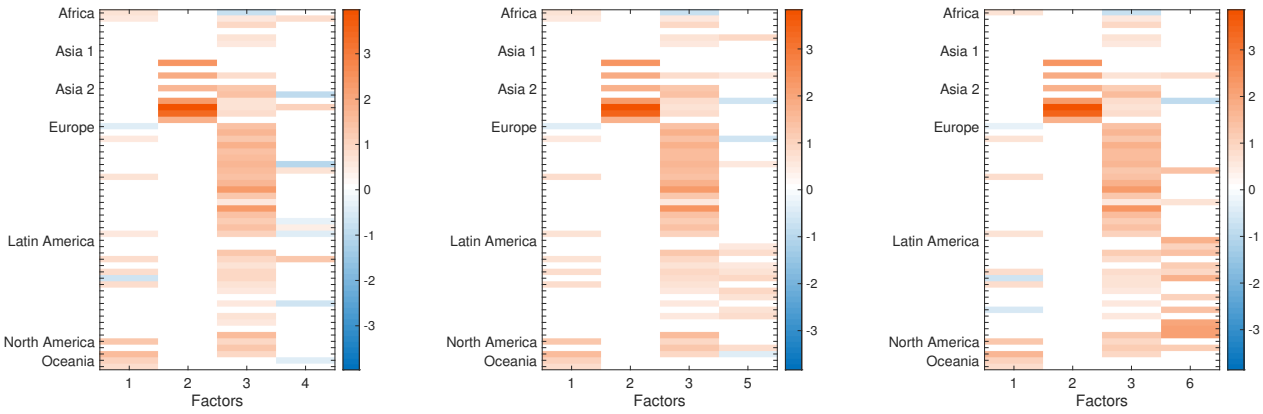


Figure 23: GDP growth, sparse permutation. Factor correlations, across factor combinations.

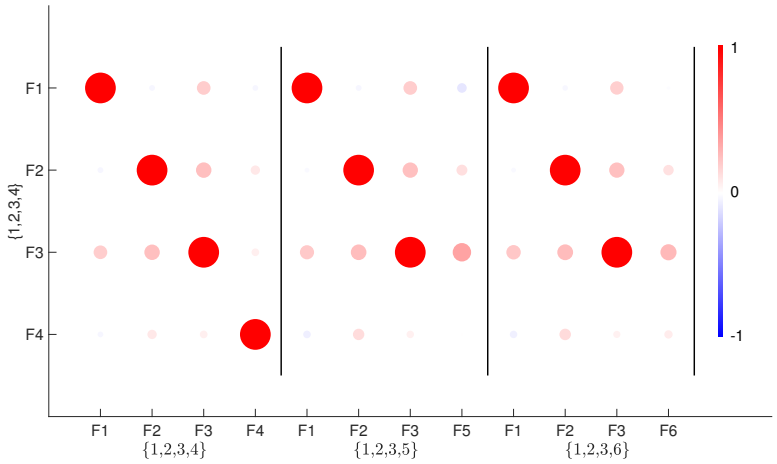
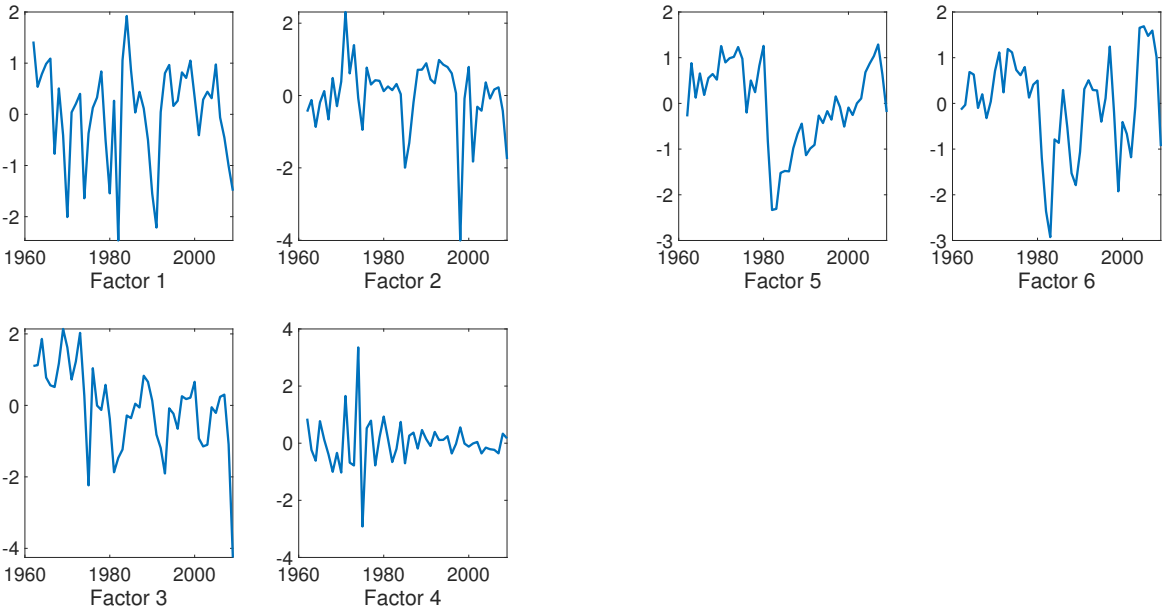


Figure 24: GDP growth, sparse permutation. Mean factors.



7. Conclusion

We present two approaches to uncover whether a sparse factor representation underlies high-dimensional data and whether the sparse representation is (locally) unique. Both approaches estimate the factor model within a Bayesian framework based on order-invariant, just-identified Markov chain Monte Carlo sampling. The first approach specifies a normal prior distribution for factor loadings and explores the unconstrained posterior distribution by implementing an unconstrained random rotation sampler. The second approach induces sparsity in the factor loading matrix by specifying a hierarchical point mass-normal mixture prior distribution on factor loadings. Random permutation of factor position and sign helps exploring the unconstrained posterior distribution. Given that the sampler may stabilize upon convergence to a sparse representation of the factor loading matrix, we run multiple chains in parallel to allow the sampler to converge to various sparse modes.

The posterior output of both samplers is post-processed to uncover potential multiple sparse representations of the factor model. The output of the unconstrained rotation sampler is optimally rotated towards sparse representations, i.e. towards different, most sparse representations displaying similar sparsity. The output of the sparse permutation sampler is post-processed to cluster factor and factor loading draws and group them into typical combinations of joint factor draws.

An extensive simulation exercise demonstrates that both approaches recover multiple underlying sparse representations, also in the presence of so-called pervasive factors, that is, factors affecting most and the same units in multiple sparse representations. We illustrate the importance of uncovering multiple sparse structures by applying the method to two datasets, for which the determination of the number of factors has been ambiguous in empirical applications. We show that pervasive factors underly each dataset, while some weaker factors are present, each identifiable jointly with the pervasive ones, but too weak to be jointly identifiable all together. The applications evidence that the sparse permutation sampler extracts pervasive factors of higher correlation across sparse representations than the rotated output of the unconstrained rotation sampler, and eventually identifies more weak factors.

Multiple sparse factor loading representations potentially lead to different factor and structural interpretations, which may be exploited in future research depending on the research question of interest.

References

Aguilar, O. and West, M. (2000). Bayesian Dynamic Factor Models and Portfolio Allocation. *Journal of Business & Economic Statistics*, 18(3):338–357.

Anderson, T. and Rubin, H. (1956). *Statistical Inference in Factor Models*, volume 5 of *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, pages 111–150. University of California Press.

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.

Aßmann, C., Boysen-Hogrefe, J., and Pape, M. (2023). Post-processing for Bayesian analysis of reduced rank regression models with orthonormality restrictions. *AStA Advances in Statistical Analysis, forthcoming*.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828.

Bernanke, B. S., Boivin, J., and Eliasz, P. (2005). Measuring the effects of monetary policy: A factor-augmented vector autoregressive (FAVAR) approach. *Quarterly Journal of Economics*, 120:387–422.

Briggs, N. E. and MacCallum, R. C. (2003). Recovery of weak common factors by maximum likelihood and ordinary least squares estimation. *Multivariate Behavioral Research*, 38:25–56.

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(4):1438–1456.

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.

Francis, N., Owyang, M. T., and Savascin, O. (2017). An endogenously clustered factor approach to international business cycles. *Journal of Applied Econometrics*, 32:1261–1276.

Freyaldenhoven, S. (2022). Factor models with local factors — Determining the number of relevant factors. *Journal of Econometrics*, 229:80–102.

George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Bayesian Statistics*, 5:609–620.

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

Golub, G. H. and van Loan, C. F. (2013). *Matrix Computations*. The Johns Hopkins University Press, 4th edition.

Hyndman, R. J. (1996). Computing and graphing highest density regions. *The American Statistician*, 50(2):120–126.

Jacob, P. E., O'Leary, J., and Atchadé, Y.F. (2020). Unbiased markov chain monte carlo methods with couplings. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82:543–600.

Kaufmann, S. and Pape, M. (2023). A geometric approach to factor model identification: Sato's $O(K)$ algorithm. mimeo.

Kaufmann, S. and Schumacher, C. (2017). Identifying relevant and irrelevant variables in sparse factor models. *Journal of Applied Econometrics*, 32:1123 – 1144.

Kaufmann, S. and Schumacher, C. (2019). Bayesian estimation of sparse dynamic factor models with order-independent and ex-post mode identification. *Journal of Econometrics*, 210:116–134.

Lawley, D. and Maxwell, A. (1971). *Factor Analysis as a Statistical Method*. Butterworths, London, 2nd edition.

Lucas, J., Carvalho, C., Wang, Q., Bild, A., Nevins, J., and West, M. (2006). Sparse Statistical Modelling in Gene Expression Genomics. In Do, K. A., Mueller, P., and Vannucci, M., editors, *Bayesian Inference for Gene Expression and Proteomics*. Cambridge University Press, Cambridge UK.

Mackowiak, B., Mönch, E., and Wiederholt, M. (2009). Sectoral price data and models of price setting. *Journal of Monetary Economics*, 56:78–99.

Mezzadri, F. (2007). How to generate random matrices from the classical compact groups. *Notices of the American Mathematical Society*, 54(5):592 – 604.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian Variable Selection in Linear Regression. *Journal of the American Statistical Association*, 83:1023–1032.

Rousseau, J. and Mengersen, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):689–710.

Titsias, M. and Lázaro-Gredilla, M. (2011). Spike and Slab Variational Inference for Multi-Task and Multiple Kernel Learning. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K., editors, *Advances in Neural Information Processing Systems 24*, pages 2339–2347. NIPS.

West, M. (2003). Bayesian Factor Regression Models in the "Large p, Small n" Paradigm. In *Bayesian Statistics 7*, pages 723–732. Oxford University Press.

Appendix A. Posterior sampling

Appendix A.1. Posterior distribution of factor loadings and hyperparameters

The prior (7)-(9) in Subection 3.1 implies a common base rate of a non-zero factor loading of $E\left(\beta_{ij}\right) = \rho_j b$ across variables. The marginal prior becomes

$$\pi\left(\lambda_{ij}|\rho_j\right) \sim (1-\rho_j b)\delta_0(\lambda_{ij}) + \rho_j b N\left(0, \tau_j\right).$$

For each factor j, transform the variables to

$$y_{it}^{(j)} = y_{it} - \sum_{l=1, l\neq j}^{k} \lambda_{il} f_{lt} = \lambda_{ij} f_{jt} + \epsilon_{it},$$

which isolates the effect of factor j on variable i. Combine the marginal prior with data information to sample independently across i from

$$\begin{aligned} \pi\left(\lambda_{ij}|\cdot\right) &= \prod_{t=1}^{T} \pi(y_{it}^{(j)}|\cdot)\left\{(1-\rho_j b)\delta_0(\lambda_{ij}) + \rho_j b N\left(0, \tau_j\right)\right\}, \\ &= P\left(\lambda_{ij}=0|\cdot\right)\delta_0(\lambda_{ij}) + P\left(\lambda_{ij}\neq 0|\cdot\right) N\left(m_{ij}, M_{ij}\right), \end{aligned}$$

with observation density $\pi(y_{it}^{(j)}|\cdot) = N\left(\lambda_{ij} f_{jt}, \sigma_i^2\right)$ and where

$$M_{ij} = \left(\frac{1}{\sigma_i^2}\sum_{t=1}^{T} f_{jt}^2 + \frac{1}{\tau_j}\right)^{-1}, \; m_{ij} = M_{ij}\left(\frac{1}{\sigma_i^2}\sum_{t=1}^{T} f_{jt} y_{it}^{(j)}\right).$$

The posterior odds of a non-zero factor loading in (A.1) are given by:

$$\frac{P\left(\lambda_{ij}\neq 0|\cdot\right)}{P\left(\lambda_{ij}=0|\cdot\right)} = \frac{\pi\left(\lambda_{ij}\right)|_{\lambda_{ij}=0}}{\pi\left(\lambda_{ij}|\cdot\right)|_{\lambda_{ij}=0}} \frac{\rho_j b}{1-\rho_j b} = \frac{N\left(0; 0, \tau_j\right)}{N\left(0; m_{ij}, M_{ij}\right)} \frac{\rho_j b}{1-\rho_j b}.$$

Conditional on λ_{ij} we update the variable specific probabilities β_{ij} and sample from $\pi(\beta_{ij}|\lambda_{ij}, \cdot)$. If $\lambda_{ij} = 0$

$$\begin{aligned} &\pi(\beta_{ij}|\lambda_{ij}=0, \cdot) \propto (1-\beta_{ij})\left[(1-\rho_j)\delta_0(\beta_{ij}) + \rho_j B\left(ab, a(1-b)\right)\right], \\ &P(\beta_{ij}=0|\lambda_{ij}=0, \cdot) \propto (1-\rho_j), \quad P(\beta_{ij}\neq 0|\lambda_{ij}=0, \cdot) \propto (1-b)\rho_j. \end{aligned}$$

That is, with posterior odds $(1-b)\rho_j/(1-\rho_j)$ we sample from $B\left(ab, a(1-b)+1\right)$ and set otherwise β_{ij} equal to zero. Conditional on $\lambda_{ij} \neq 0$ we obtain

$$\begin{aligned} &\pi(\beta_{ij}|\lambda_{ij}\neq 0, \cdot) \propto \beta_{ij} N\left(\lambda_{ij}; 0, \tau_j\right)\left[(1-\rho_j)\delta_0(\beta_{ij}) + \rho_j B\left(ab, a(1-b)\right)\right], \\ &P(\beta_{ij}=0|\lambda_{ij}\neq 0, \cdot) = 0, \quad P(\beta_{ij}\neq 0|\lambda_{ij}\neq 0, \cdot) = 1. \end{aligned}$$

In this case we sample β_{ij} from $B\left(ab+1, a(1-b)\right)$.

The hyperparameters τ_j and ρ_j are sampled from, respectively, an inverse Gamma $\pi\left(\tau_j|\cdot\right) \sim IG\left(g_j, G_j\right)$ and a Beta

distribution, $\pi\left(\rho_j|\cdot\right) \sim B\left(r_{1j}, r_{2j}\right)$, with

$$g_j = g_0 + \frac{1}{2}\sum_{i=1}^{N} I\left\{\lambda_{ij} \neq 0\right\}, \quad G_j = G_0 + \frac{1}{2}\sum_{i=1}^{N} {\lambda_{ij}}^2,$$

$$r_{1j} = r_0 s_0 + S_j, \quad r_{2j} = r_0(1 - s_0) + N - S_j, \text{ and } S_j = \sum_{i=1}^{N} I\left\{\beta_{ij} \neq 0\right\},$$

and $I\{\cdot\}$ is the indicator function.

Appendix A.2. Running parallel chains: Convergence

When executing parallel chains with a fixed burn-in and a fixed number of iterations, we need to assess whether both are sufficiently large for the sampler to converge for each chain, and whether we have produced enough iterations to perform posterior evaluations. Assessing convergence in a high-dimensional, latent-variable model is not trivial, if we disregard the usual trace and autocorrelation plots or convergence diagnostics for single parameter values. It is even more involved if we (have to) take into consideration random factor and sign permutation applied at the end of each iteration.

We suggest using a statistic to assess convergence of both the unidentified and identified posterior output, based on a sign-independent model parameter. We compute the Jaccard matching index between the first draw of a (sub-)chain and all following draws, based on an indicator matrix for non-zero loadings, obtained by evaluating $\beta_{ij}^{(m)} > .75$ for each draw m. To resolve factor permutation in the unidentified output, we re-order factor-specific columns to maximize the Jaccard matching index between the first draw $\tilde{m}$ of the (sub-)chain and each subsequent one.

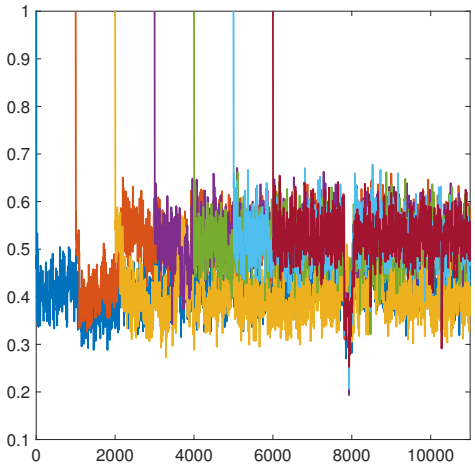
Figure A.25, Panels (a) and (b), plots the Jaccard matching index for the unidentified posterior output. Panel (a) displays various series of Jaccard matching index for the first chain of 11,000 draws, computed between starting draws $\tilde{m} = 1, 1001, 2001, \ldots, 6001$ and all subsequent draws. We observe that the index series for $\tilde{m} = 1$ yields very low values. As the sampler proceeds, the level of the index series increases, the overall value and the pattern stabilizes after $\tilde{m} = 5001$. The dip in the index around $m = 8000$ may represent draws from another mode present in the data. In Panel (b), we concatenate the Jaccard matching index series of all chains, each with first draw set to $\tilde{m} = 6001$. We observe that the level and the pattern of the series is similar across chains. From these plots, we conclude that a burn-in of 6000 is enough for each chain to reach convergence.

Panels (c) and (d) of Figure A.25 show the Jaccard matching index series for each mode, respectively, 1 and 2. Both series are unimodal, and as such reflect that the post-processing procedure identifies well both modes.

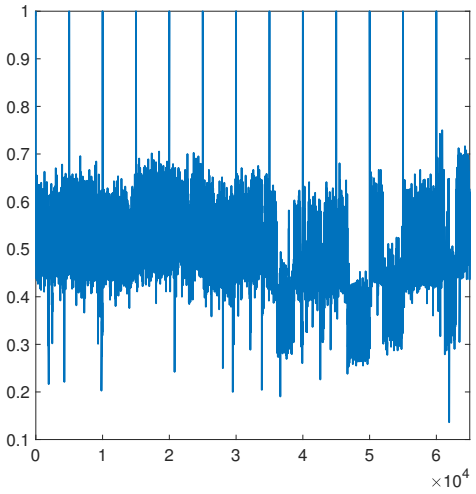
Figure A.26 plots the posterior output for both modes. The top-left plot in Panels (a) and (b) are the same, they plot the correlation of the first draw for Factor 2 with the all remaining ones. We clearly observe a bi-modal distribution, reflecting the two weaker factors. The top-right figures in each panel show the trace plots of factor loadings for a specific series in each mode, where the blue (red) trace corresponds to the loading of Factor 1 (2). We observe a bimodal trace plot for Factor 2 only. The bottom panels display similar graphs for each mode-identified posterior output. The correlation across factor draws and factor loadings as well show unimodal distributions, which again

reflects that posterior processing works well in identifying both modes. A strong factor (Factor 1) is the same across modes, whereas a weaker factor (Factor 2) characterizes each mode.

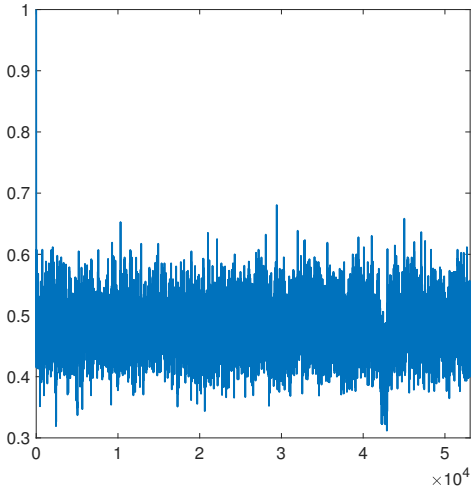
Figure A.25: US CPI: Jaccard index between the first draws (Draw $\tilde{m}$) of a (sub-)chain and the subsequent ones, based on indicator matrices obtained by evaluating $\beta_{ij}^{(m)} > .75$. To resolve factor permutation, we re-order factor-specific columns to maximize the Jaccard matching index between the first draw $\tilde{m}$ of the (sub-)chain and each subsequent one. To improve the visualization, the index starts at 1 (the Jaccard matching index for the 1st draw with itself). (a) Initial chain of 11,000 draws; (b) Concatenated, retained 6,000 draws of 13 chains; (c) Sorted output for the first mode (53,086 draws); (d) Sorted output for the second mode (11,793 draws).



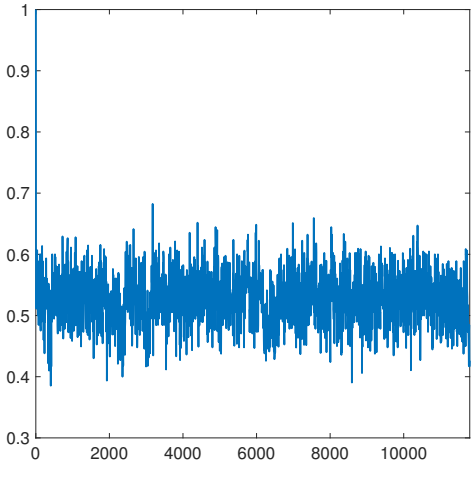
(a) $\tilde{m} = 1, 1001, \ldots, 6001$



(b) $\tilde{m} = 6001$ for all chains

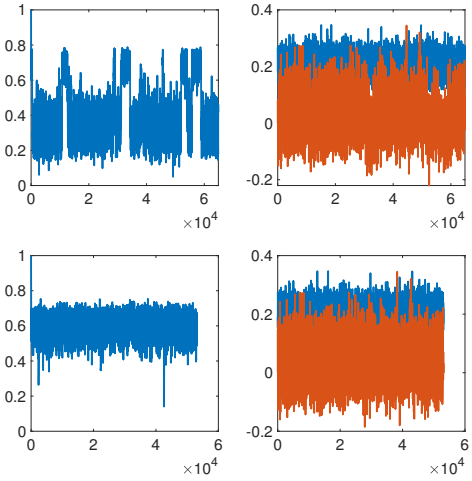


(c) $\tilde{m} = 1$, sorted output for mode 1

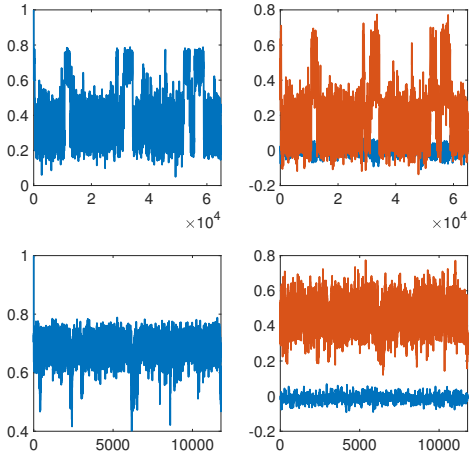


(d) $\tilde{m} = 1$ sorted output for mode 2

Figure A.26: US CPI: Posterior draws, unsorted and sorted, $K = 2$. From top left to bottom right: Correlation of the first with all other posterior draws of Factor 2, posterior draws of a selected row of Λ, correlation of the first with all other sorted posterior draws of mode-identified Factor 2, sorted posterior draws of a mode-identified row of Λ.).



(a) Mode 1 (53,086 draws)



(b) Mode 2 (11,793 draws)

Appendix B. Post-processing

Appendix B.1. Givens decomposition of an orthogonal matrix

An orthogonal matrix H with $\det(H) = 1$ is a rotation matrix. An orthogonal matrix H with $\det(H) = -1$ can be transformed into a rotation matrix by multiplying the last column of H by -1, inducing an axis reflection. Note that to identify a sparse pattern in the algorithm proposed in Subsection 4.1, axis reflections are ruled out, so it is sufficient to consider H with $\det(H) = 1$.

Rotation matrices of dimension $K > 2$ can be decomposed into Givens rotation matrices, see Golub and van Loan (2013), Section 5.1.

The Givens decomposition of an orthogonal matrix H with $H'H = HH' = I_K$ can be performed as follows. First, define all pairs of axes $k_1, k_2 \in \{1, \ldots, K\}$ with $k_1 \neq k_2$. There are $P = \binom{K}{2}$ such pairs, $p = \{1, \ldots, P\}$. Then apply following steps, starting with $p = 1$.

1. Determine the two-dimensional Givens rotation matrix

$$G_p = \frac{1}{\|(h_{k_1,k_1},\, h_{k_2,k_1})'\|_2} \cdot \begin{pmatrix} h_{k_1,k_1} & h_{k_2,k_1} \\ -h_{k_2,k_1} & h_{k_1,k_1} \end{pmatrix} = \begin{pmatrix} g_{p,1,1} & g_{p,1,2} \\ g_{p,2,1} & g_{p,2,2} \end{pmatrix}.$$

2. Calculate the Givens rotation angle of matrix G_p as

$$\gamma_p = \arctan2(g_{p,2,1}, g_{p,1,1}).$$

3. Replace the $k_1{}^{\text{th}}$ and $k_2{}^{\text{th}}$ row of matrix H, denoted as the submatrix $H_{\{k_1,k_2\},\cdot}$, by its rotated version $G_p H_{\{k_1,k_2\},\cdot}$.
4. If $p < P$, increment p and proceed with step 1, otherwise the decomposition is complete, in which case $H = I_K$.

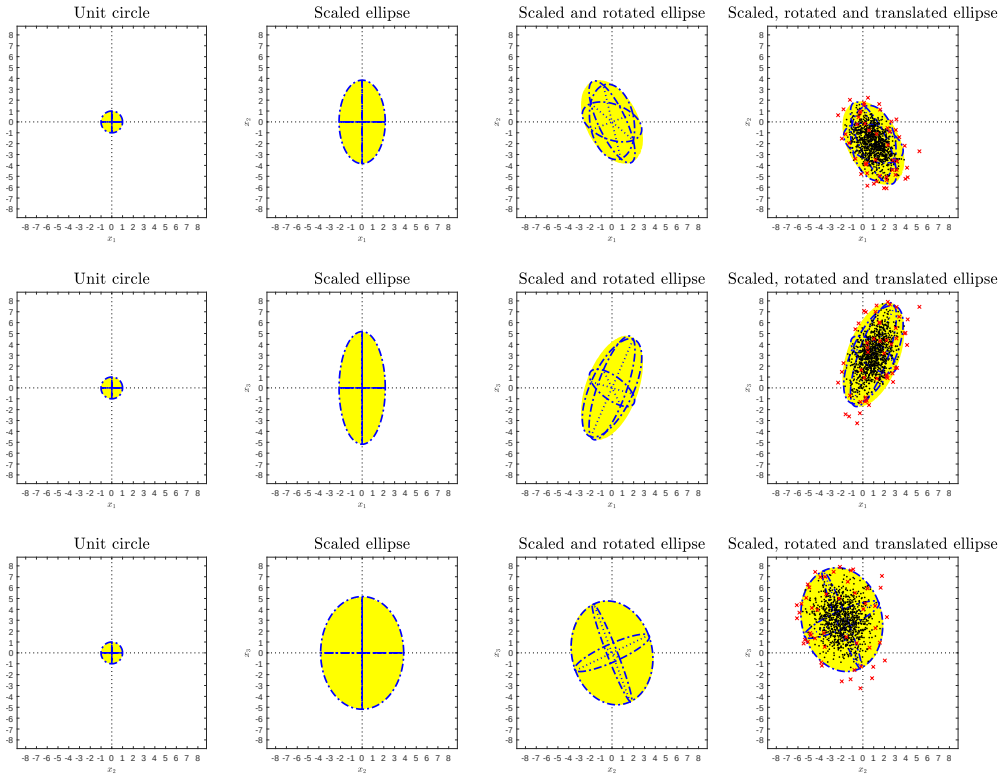
Appendix B.2. Constructing a $K = 3$-dimensional hyperellipsoids

Figure B.27 extends the construction exercise to three dimensions. Each row of the plot shows one pair of dimensions, with dimensions 1 and 2 in the first, 1 and 3 in the second, and 2 and 3 in the third row. The unit circle in the first panel of each row is hence a unit ball when all three dimensions are considered. The ellipses in the second panel of each row represent the expanded ellipsoid from three different angles. In the third panel of each row, the ellipse has been rotated, and the original axes are indicated within the rotated ellipsoid. Again, the last panel of each row shows the data points inside and outside the ellipsoid as black dots and red marks, respectively. Note that the red marks apparently within the ellipsoid are in fact located in front of or behind it. The share of points inside the ellipsoid is again exactly $1 - \alpha$.

Appendix B.3. Optimization towards sparsity

Figure B.28 shows an example of how the loss function in Equation (14) evolves as the algorithm described in Section 4.1 is applied to the output of the unconstrained sampler that has been postprocessed with the WOP procedure of Aßmann et al. (2016). The scenario that is analyzed is K3m2_1pf, as described in Section 5.2. In the left panel,

Figure B.27: 95% highest posterior density ellipsoid for $K = 3$, built by first expanding the unit ball, then applying a rotation and a translation. Top row shows axes 1 and 2, middle row axes 1 and 3, and bottom row axes 2 and 3. Original data points shown in black (inside the ellipsoid) and red (outside the ellipsoid) in the panels in the right column.



a single optimization is considered, where the algorithm proceeds through the axes in a pairwise fashion, addressing two dimensions of the hyperellipsoids at a time, as described at the end of Section 4.1. Each iteration therefore refers to a (Givens) rotation for one pair of axes. Convergence is reached after approximately 120 steps. In the right panel, the same optimization is repeated 50 times, with randomly chosen starting values for the rotation angles. The value of the loss function at convergence is very similar across repetitions. Sometimes it takes more than 500 iterations for the algorithm to converge. Usually, however, the value of the loss function is minimized after less than 50 iterations. Figure B.29 shows the Givens rotation angles for the same example, as implied by the choice of H_* in each iteration of the algorithm.

Throughout the simulation study and the empirical application, we used $\alpha = 0.05$ to construct the HPD hyperellipsoids that determine the sparse structures in Λ. Therefore, we provide an example to illustrate how the results change if α is varied. Reducing α makes the HPD hyperellipsoids wider, so we should see more shrinkage and hence, more sparsity. For the scenario K3m2_1pf, we start with $\alpha = 0.1$, for which the resulting estimate of Λ is shown in the

first panel of Figure B.30. Reducing α to 0.05 yields a pattern with five additional zero elements and two additional nonzero elements in mode 1, whereas mode 2 remains unchanged. The resulting estimate of Λ is shown in the second panel of Figure B.30. Eventually, we reduce α further to 0.01. The resulting pattern has 13 additional zero elements and one additional nonzero element in mode 1 and 17 additional zero elements and three additional nonzero elements in mode 2. The resulting estimate of Λ is shown in the third panel of Figure B.30. The degree of sparsity is increased when reducing α, though the identified sparse structure for lower values of α is not perfectly, but closely, nested in the structure obtained for higher values of α. Furthermore, we observe that as α is reduced, estimated nonzero loadings are larger in absolute terms.

Figure B.28: Loss function in Equation (14) shown for scenario K3m2_1pf. The left panel shows a single optimization with iterative pairwise axis (Givens) rotations. The right panel shows the same for 50 optimizations, using different randomly chosen starting points.

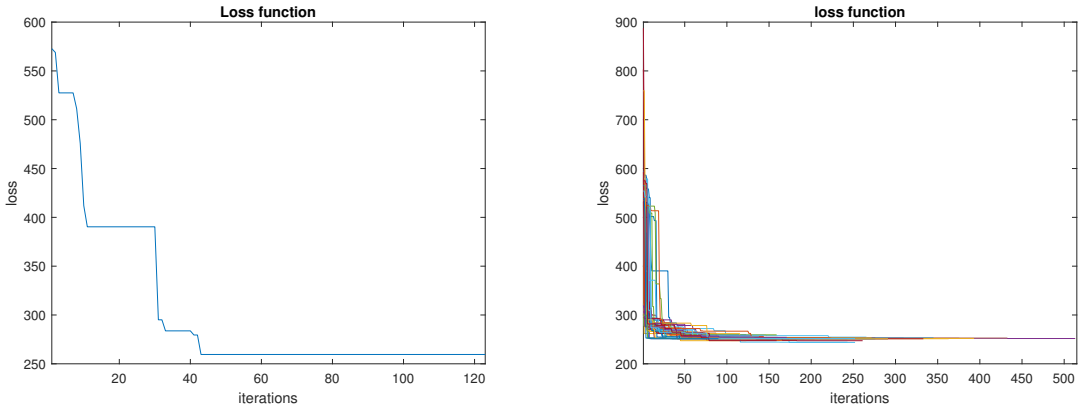


Figure B.29: Givens rotation angles implied by the optimal H_* for the pairwise axis rotations for scenario K3m2_1pf. The blue, red and yellow lines, respectively, denote the angles between axis pairs 1 and 2, 1 and 3, and 2 and 3.

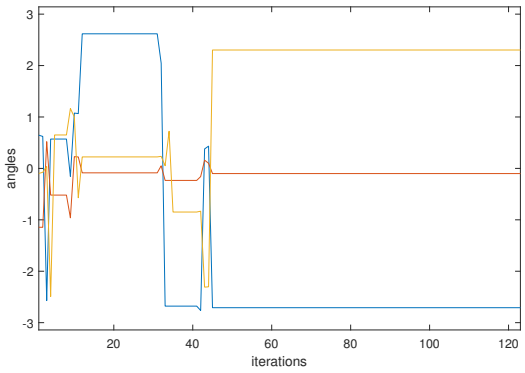
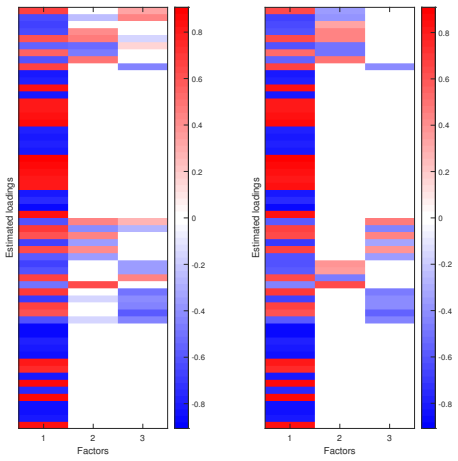
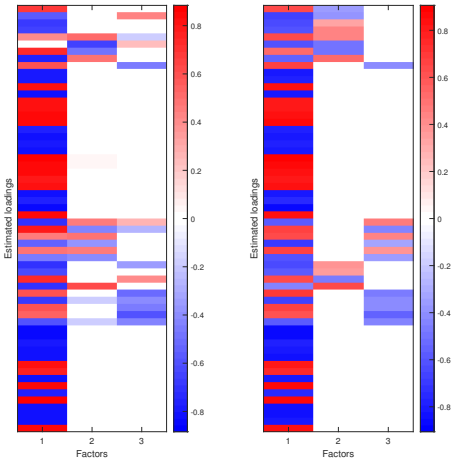


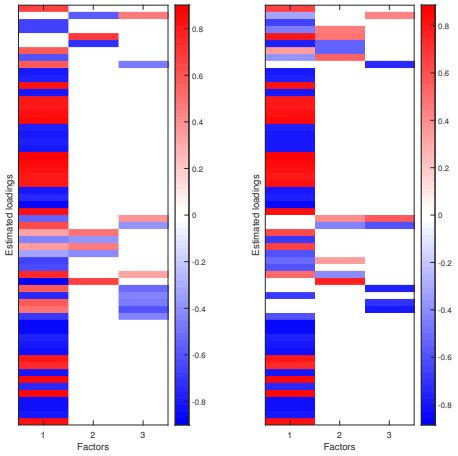
Figure B.30: Results for scenario K3m2_1pf for different choices of α, showing two modes for each choice of α.



$\alpha = 0.1$



$\alpha = 0.05$



$\alpha = 0.01$

Appendix B.4. Ex-post clustering of factor draws

The posterior output of the sparse permutation sampler has $2^K K!$ modes. Posterior mode identification will assign each factor draw $f_k^{(m)} = \{f_{kt}^{(m)} | t = 1, \ldots, T\}, k = 1, \ldots, K, m = 1, \ldots, M$ to one of K clusters, if we neglect the sign switch, i.e. if we sign-adjust appropriately the factor draws. If multiple sparse factor representations are possible, the posterior output will display a multiple of $2^K K!$ modes. In this case, the factor draws $f_k^{(m)}$ will group into $G \geq K$ clusters. To sort out the posterior output, we set up a mixture model with mixture indicator $z_k^{(m)} = \{1, \ldots, G\}$ which indicates the cluster $g = \{1, \ldots, G\}$ with which factor draw $f_k^{(m)}$ is associated. We define the following hierarchical prior model

$$P(z_k^{(m)} = g) = \eta_g, \; g = 1, \ldots, G, \tag{B.1}$$

$$\eta = (\eta_1, \ldots, \eta_G) \sim D(e_0, \ldots, e_0), \text{ with } e_0 = (G-1)/2, \tag{B.2}$$

$$\pi(f_k^{(m)} | z_k^{(m)} = g) \sim N\left(\mathrm{f}_g, \mathrm{F}_g\right), \text{ where } \mathrm{F}_g = \mathrm{diag}(\mathrm{F}_{g1}, \ldots, \mathrm{F}_{gT}),$$

and

$$\pi(\mathrm{f}_g) \sim N\left(0_{T\times 1}, I_T\right), \; \mathrm{F}_{gt} \sim IG(\mathrm{s}_0, \mathrm{S}_0).$$

The prior for the mixture indicator (B.1)-(B.2) is uniform discrete and the Dirichlet specification with $e_0 < (G-1)/2$ allows for empty clusters ex-post, Rousseau and Mengersen (2011).

An estimate of the clusters and cluster association for each draw is obtained by sampling iteratively over the following steps:

1. Update cluster association of each factor draw $f_k^{(m)}$, $k = 1, \ldots, K$, $m = 1, \ldots, M$: $\pi\left(z_k^{(m)} | f_k^{(m)}, \mathrm{f}_g, \mathrm{F}_g, \eta\right)$. The posterior probability of cluster association is proportional to

$$P\left(z_k^{(m)} = g | f_k^{(m)}, \mathrm{f}_g, \mathrm{F}_g, \eta\right) \propto |\mathrm{F}_g|^{-1/2} \exp\left\{-0.5 \sum_{t=1}^{T} \frac{(\mathrm{sad}(f_{kt}^{(m)}) - \mathrm{f}_{gt})^2}{\mathrm{F}_{gt}}\right\} \eta_g. \tag{B.3}$$

The expression $\mathrm{sad}(f_{kt}^{(m)})$ means sign adjustment according to

$$\mathrm{sad}(f_{kt}^{(m)}) = \begin{cases} f_{kt}^{(m)} & \text{if } \sum_{t=1}^{T}(f_{kt}^{(m)} - \mathrm{f}_{gt})^2 < \sum_{t=1}^{T}(-f_{kt}^{(m)} - \mathrm{f}_{gt})^2 \\ -f_{kt}^{(m)} & \text{if } \sum_{t=1}^{T}(f_{kt}^{(m)} - \mathrm{f}_{gt})^2 > \sum_{t=1}^{T}(-f_{gt}^{(m)} - \mathrm{f}_{gt})^2 \end{cases}.$$

This operation adjusts the sign of those draws which are negatively correlated to the factor mean due to random sign switching applied during sampling.

Simulate $U \sim (0, 1)$ and set $z_k^{(m)}$ equal to

$$g = \left(\sum_{l=1}^{G} I\left\{\left(\sum_{j=1}^{l} P(z_k^{(m)} = j | \cdot)\right) \leq U\right\}\right) + 1,$$

where $I\{\cdot\}$ represents the indicator function and $P(z_k^{(m)} = j|\cdot)$ are the normalized posterior cluster probabilities obtained from (B.3).

2. Update the cluster association probabilities: $\pi(\eta|z) \sim D(e_1, \ldots, e_G)$ with $e_g = e_0 + N_g$, $N_g = \sum_{k,m} I\{z_k^{(m)} = g\}$, $g = 1, \ldots, G$.

3. Update the factor representative f_g, i.e. the mean path of factors, in cluster $g = 1, \ldots, G$: $\pi\left(\mathrm{f}_g|z, \boldsymbol{f}\right) \sim N\left(\bar{\mathrm{f}}_g, \bar{\mathrm{F}}_g\right)$, with moments

$$\bar{\mathrm{F}}_g = (N_g \mathrm{F}_g^{-1} + I_T)^{-1} \text{ and } \bar{\mathrm{f}}_g = \bar{\mathrm{F}}_g \left(\mathrm{F}_g^{-1} \sum_{k,m} \mathrm{sad}(f_k^{(m)}) I\{z_k^{(m)} = g\} \right).$$

4. Update the time-specific variance of factors in cluster g: $\pi\left(\mathrm{F}_{gt}|z, \mathrm{f}_g, \boldsymbol{f}\right) \sim IG\left(\mathrm{s}_{gt}, \mathrm{S}_{gt}\right)$ with

$$\mathrm{s}_{gt} = \mathrm{s}_0 + 0.5 N_g \text{ and } \mathrm{S}_{gt} = \mathrm{S}_0 + 0.5 \sum_{k,m} (\mathrm{sad}(f_{kt}^{(m)}) - \mathrm{f}_{gt})^2 I\{z_k^{(m)} = g\}.$$

For factors, we set $\mathrm{s}_0 = 2$ and $\mathrm{S}_0 = 1$. When we set up a mixture model for factors stacked with factor loadings, we set $\mathrm{s}_0 = .3125$ and $\mathrm{S}_0 = 5$ for factor loadings.

Appendix C. Simulation

Appendix C.1. Orthogonal matrices for minimal correlation

In order to keep rotations of factors and loading matrices as far apart from each other as possible, consider that by assumption of static uncorrelated factors with identical unit variances, i.e. $F \sim (0, \mathbf{I}_K)$, we have $\mathbb{E}(FF') = \mathbf{I}_K$. Transforming the factors by an orthogonal matrix $\mathbf{D} \in O(K)$ yields $\tilde{F} = \mathbf{D}F$. Minimizing the variance between all members of the initial set of factors and all members of the rotated set of factors is therefore identical to minimizing the largest absolute element of the matrix $\mathbf{D}$. The covariance matrix of the initial and the rotated factors thus becomes

$$Cov\left(\begin{pmatrix} F & \tilde{F} \end{pmatrix}\right) = \begin{pmatrix} \mathbf{I}_K & \mathbf{D} \\ \mathbf{D}' & \mathbf{I}_K \end{pmatrix}.$$

If more than two modes are desired, use the result that for two orthogonal matrices $\mathbf{D}_1$ and $\mathbf{D}_2$, the matrix $\mathbf{D}_1\mathbf{D}_2$ is also orthogonal. Therefore, to obtain m modes that minimize the absolute correlation between any two factors, orthogonal matrices $\mathbf{D}_1$ to $\mathbf{D}_{m-1}$ are required, and, defining $\mathbf{D}_0 = \mathbf{I}$, it must hold that the largest absolute matrix elements of any product $\mathbf{D}_i'\mathbf{D}_j$ with $i \neq j$ and $i, j \in \{0, \ldots, m\}$ becomes as small as possible.

Two interesting results are explained in the following: First, for $m = 2$ and small values of K, minimizing the largest absolute element of the matrix $\mathbf{D}$ yields the same result as minimizing the variance of the absolute elements of $\mathbf{D}$. Moreover, if a solution for K_1 and K_2 has been found, say, $\mathbf{D}_{K_1}$ and $\mathbf{D}_{K_2}$, where $K_1 = K_2$ may hold, a solution for K_1K_2 is found as $\mathbf{D}_{K_1} \otimes \mathbf{D}_{K_2}$.

Consider e.g. the case $K = 2$. The rotation matrix that minimizes the angles between F and $F\mathbf{D}^{(2)}$ is either $\mathbf{D}_1^{(2)} = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}$ or $\mathbf{D}_2^{(2)} = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}$, as shown in Figure C.31, where $\mathbf{D}_2^{(2)\prime} = \mathbf{D}_1^{(2)}$.

Figure C.31: Minimal correlation solutions for 2, 3 and 4 modes in the 2-dimensional case.

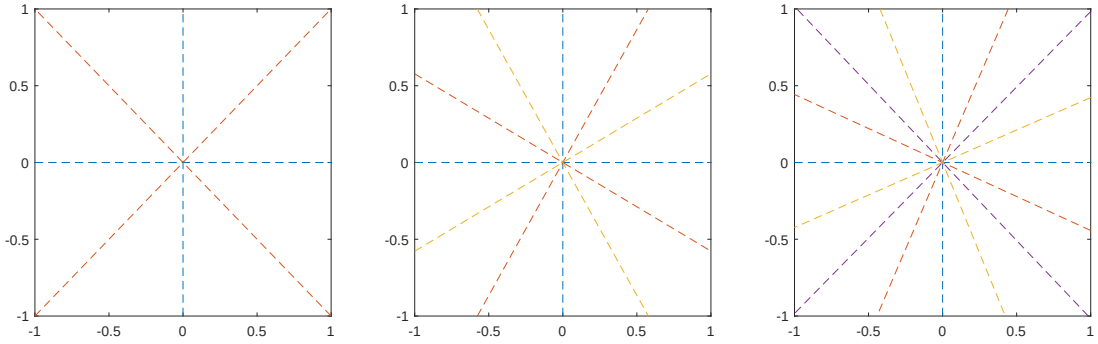
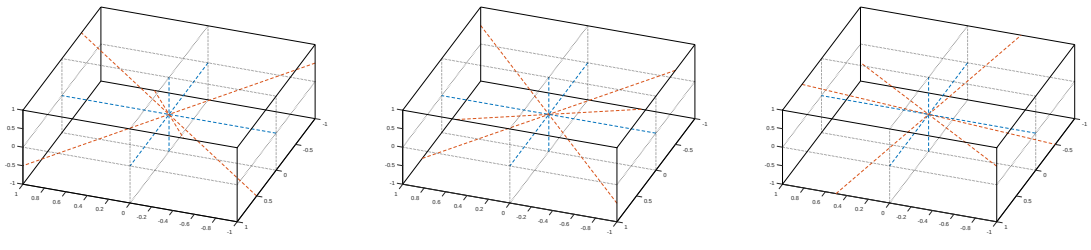


Figure C.32: Three different minimal correlation solutions for 2 modes in the 3-dimensional case.



Next, consider the case $K = 3$. The rotation matrix that minimizes the angles between F and $F\mathbf{D}$ can now take several different forms, one of which is $\mathbf{D}^{(3)} = \begin{pmatrix} -\frac{1}{3} & \frac{2}{3} & \frac{2}{3} \\ \frac{2}{3} & -\frac{1}{3} & \frac{2}{3} \\ \frac{2}{3} & \frac{2}{3} & -\frac{1}{3} \end{pmatrix}$. Figure C.32 shows three solutions for two modes in the 3-dimensional case.

Regarding $K = 4$, the above mentioned result can be used, i.e., solutions obtain as $\mathbf{D}_1^{(2)} \otimes \mathbf{D}_1^{(2)}$, $\mathbf{D}_1^{(2)} \otimes \mathbf{D}_2^{(2)}$, $\mathbf{D}_2^{(2)} \otimes \mathbf{D}_1^{(2)}$, and $\mathbf{D}_2^{(2)} \otimes \mathbf{D}_2^{(2)}$, with

$$\mathbf{D}_1^{(2)} = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \text{ and } \mathbf{D}_2^{(2)} = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}.$$

This yields

$$\mathbf{D}_1^{(4)} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \end{pmatrix}, \quad \mathbf{D}_2^{(4)} = \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} \end{pmatrix}, \quad \mathbf{D}_3^{(4)} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \end{pmatrix} \text{ and}$$

$$\mathbf{D}_4^{(4)} = \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \end{pmatrix},$$

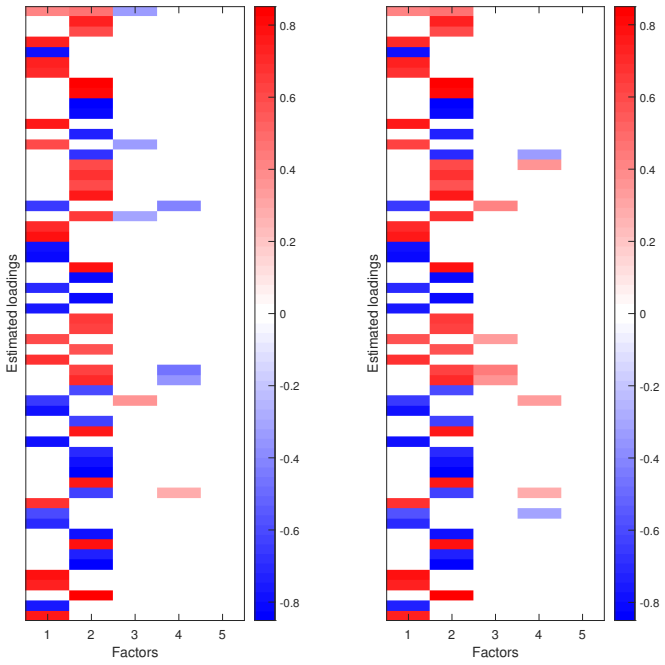
where $\mathbf{D}_4^{(4)\prime} = \mathbf{D}_1^{(4)}$ and $\mathbf{D}_3^{(4)\prime} = \mathbf{D}_2^{(4)}$.

Appendix D. Results for overfitted estimate

The following results were obtained for a factor model fitted to the simulated data set dataN60K4m2_2pf (see Subsection 5.2) with an over-fitted number of factors $K = 5$.

The unconstrained rotation approach consistently produces a sparsity indicator matrix Δ that contains an entire column of zeros. The corresponding loading matrices Λ estimated for both identified modes are shown in Figure D.33. The patterns for the first four factors, and particularly the pervasive factors, are similar to those in Figure 13, and accordingly, the redundant factor has zero loadings across all units.

Figure D.33: $K = 5$, unconstrained rotation, scenario K4m2_2pf, heat plot of posterior mean factor loadings.



For the sparse permutation sampler, we first run a chain of 10,000 draws, we initialize 15 parallel chains by random rotation of a factor loading draw. We discard the first 6,000 draws from each chain, leaving us with 64,000 draws for posterior evaluation. To post-process the MCMC output (see Subsection 4.2), we stack factor and factor loading draws, we set $G = 10$ and specify a Dirichlet prior on cluster probability that allows for empty groups, $e_0 = .01(G/2 - 1)$. If all factors were relevant in both modes, each cluster should be populated by virtually 32,000 draws. Table D.7 shows that all clusters are populated, although some of them with a very low number of draws (see the line labelled All). For example, less than 10,000 draws are assigned to clusters labelled Factors 3 and 6, respectively. On the other hand, 64,000 draws are assigned to clusters labelled Factors 2 and 4, respectively. These two clusters obviously correspond to the two simulated pervasive factors.

Before visualizing the results, we proceed and retain those draws (m) which represent a set of unique 5 factors. Table D.8 lists those factor sets that were drawn more than 1,000 times (For expositional convenience we omit $1,086$ draws assigned to the set $\{2, 4, 5, 7, 10\}$). We observe again that Factors 2 and 4 are elements of each set, while some other factors (like Factors 9 and 10) are elements of many but not all sets.

Table D.7: $K = 5$, sparse permutation, scenario K4m2_2pf. Clusters of factors and factor loadings draws. All: Total number of factor-specific MCMC draws; Retained: Number of MCMC draws accross factor combinations retained for posterior evaluation, i.e. those factor combinations with more than 1,000 assigned posterior draws (see also Table D.8).

Factor	1	2	3	4	5	6	7	8	9	10
All	13,725	64,000	9,551	64,000	27,809	9,859	12,251	18,589	57,913	42,304
Retained	13,722	62,049	8,582	62,049	25,882	8,917	12,247	18,581	57,867	40,349

Figures D.34 to D.36 visualize the processed posterior output. Figure D.34 displays the heatmap of the posterior median of factor-specific loadings and non-zero loading probabilities. Factors 2 and 4 load on many series, and the pattern of loadings coincides with the heatmap for loadings of Factors, respectively, 2 and 1 in Figure 14 (Estimate for $K = 4$). The heatmap reveals that we estimated an over-fitted number of factors, as only 3 of the remaining factors have non-zero loadings. Figures D.35 and D.36 plot the heatmap for, respectively, the median of factor loadings and the posterior mean of non-zero loading probabilities of each of the factor sets displayed in Table D.8. Each plot uncovers that we estimated an over-fitted number of factors. In each plot, the posterior median of all loadings of one of the factors is zero and the posterior mean of non-zero loading probabilities is virtually zero.

Facing such a posterior output, we would re-estimate the factor model for $K = 4$.

Table D.8: $K = 5$, sparse permutation, scenario K4m2_2pf. Sorted output for overfitted estimate. Factor combinations and number of draws. See Figures D.35 and D.36 for, respectively, the posterior median of factor loadings and the posterior mean of non-zero probabilities of each sorted factor combination.

Combination	Draws	Combination	Draws	Combination	Draws
$\{1, 2, 4, 5, 9\}$	4,634	$\{1, 2, 4, 5, 10\}$	1,323	$\{1, 2, 4, 9, 10\}$	7,765
$\{2, 3, 4, 5, 9\}$	3,110	$\{2, 3, 4, 9, 10\}$	5,472	$\{2, 4, 5, 6, 9\}$	2,934
$\{2, 4, 5, 8, 9\}$	4,558	$\{2, 4, 5, 8, 9\}$	6,464	$\{2, 4, 5, 8, 10\}$	1,773
$\{2, 4, 6, 9, 10\}$	5,983	$\{2, 4, 7, 9, 10\}$	6,603	$\{2, 4, 8, 9, 10\}$	10,344

Figure D.34: $K = 5$, sparse permutation, scenario K4m2_2pf. Clustered output for overfitted estimate. Factor-specific median factor loadings. Evaluated for factor-specific draws across factor combinations displayed in Table D.8.

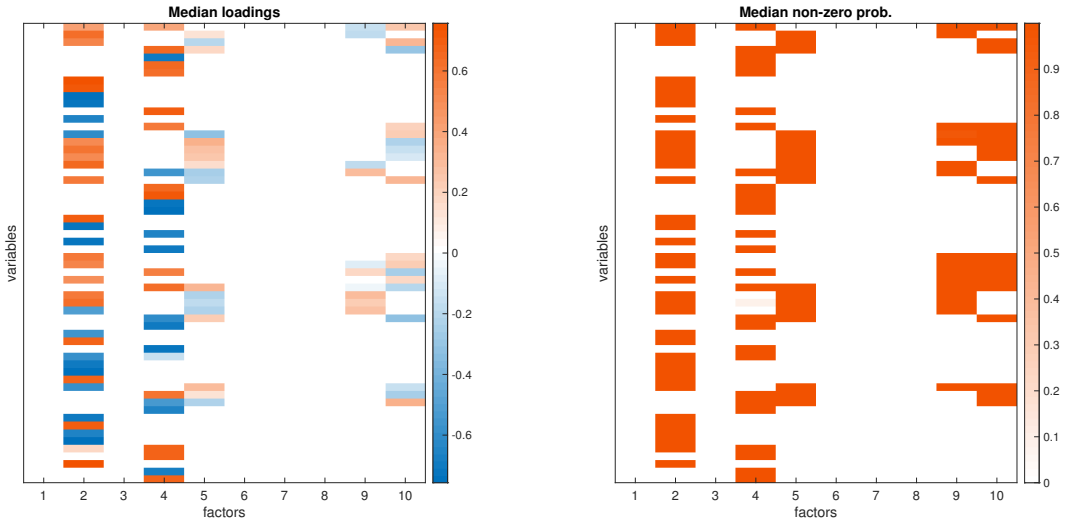


Figure D.35: $K = 5$, sparse permutation, scenario K4m2_2pf. Sorted output for overfitted estimate. Median factor loadings. See Table D.8 for the number of draws sorted into each factor combination.

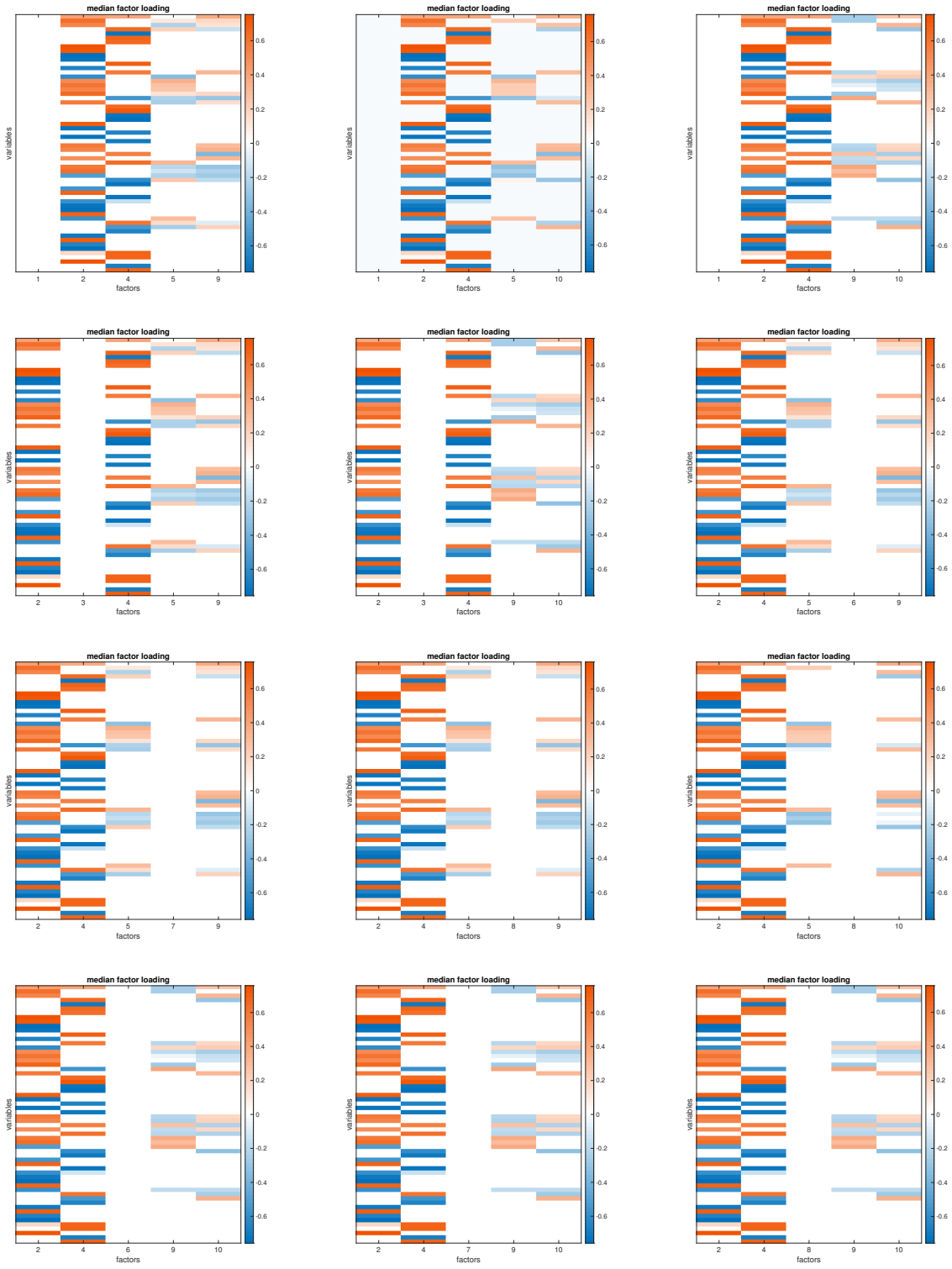


Figure D.36: $K = 5$, sparse permutation, scenario K4m2_2pf. Sorted output for overfitted estimate. Mean non-zero probability. See Table D.8 for the number of draws sorted into each factor combination.