Bayesian (non-)unique sparse factor modelling

Sylvia Kaufmann and Markus Pape

Working Paper 23.04

Bayesian (non-)unique sparse factor modelling

Sylvia Kaufmann* and Markus Pape†

December 31, 2023

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 associated with them, eventually determining factor interpretation. We 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. We propose two MCMC approaches for Bayesian inference and corresponding post-processing algorithms to uncover multiple sparse representations of the factor loadings matrix. We investigate both approaches in a simulation study. Applied to data on country-specific gross domestic product and U.S. price components series, we retrieve multiple sparse factor representations for each data set. Both approaches prove useful to discriminate between pervasive and weaker factors.

JEL classification: C11; C33; C55

Keywords: Multimodality; Sparsity; Pervasive and weak factors

*Study Center Gerzensee, Dorfstrasse 2, 3115 Gerzensee, and University of Basel, Switzerland

†Department of Economics, Ruhr-University Bochum, Universitaetsstrasse 150, 44801 Bochum, Germany

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, 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, potntially uncovering multiple sparse structures in Λ. Applications to large panels of country-specific gross domestic product (GDP) and U.S. price components 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 posterior

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, B and C contain details about posterior derivations, posterior processing, and the simulation study, respectively.

2 (Non-)Unique sparse factor representation

2.1 Specification

Consider a vector of observable data $Y = (y_1', \ldots, y_T')'$. Each y_t, $t = 1, \ldots, T$, denotes an $N \times 1$ vector of variables y_{it}, $i, \ldots, N$, 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, $\Lambda = \{\lambda_{ij} | i = 1, \ldots, N, j = 1, \ldots, K\}$ is the $N \times K$ factor loading matrix and ϵ_t is an $N \times 1$ vector of idiosyncratic components.[1] 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 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 first and second (unconditional) moments are stationary, which means that observed data in (2) is non-trending.

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

[1] 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: $E\left(\tilde{f}_t \tilde{f}_t'\right) = \Sigma_{\tilde{f}} = LL'$; $L^{-1}\Sigma_{\tilde{f}} L^{-1'} = I_K$. When post-multiplying $\tilde{\Lambda}$ with L, the factor model with correlated factors is observationally equivalent to system (2): $y_t = \tilde{\Lambda} L L^{-1} \tilde{f}_t + \epsilon_t = \Lambda f_t + \epsilon_t$.

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. Although the common components of various sparse representations may account for a similar share in data variation, solutions may nevertheless lead to differing 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 Σ_f spans a possibly correlated factor basis[2] 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 the $\mathbb{R}^N$ with $K < N$, or, as an $N \times K$ matrix with full column rank. The set of all K-frames in the $\mathbb{R}^N$ is then denoted as the (real) non-compact Stiefel manifold $V(K, N)$.[3] 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. Eventually, the set of all orthogonal $K \times K$ matrices is denoted as the (real) orthogonal group $O(K)$, corresponding to an orthogonal factor basis.

For example, Figure 1 plots following loading matrix as coordinates:

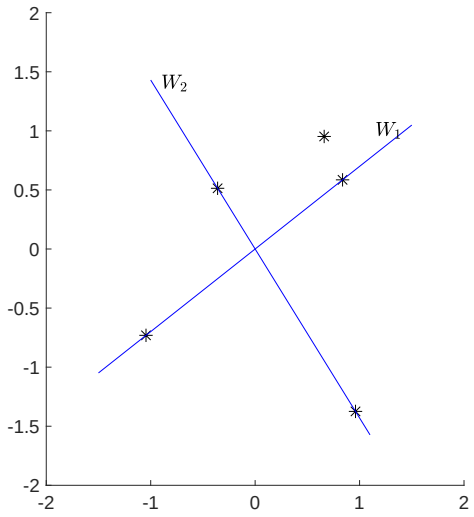
$$\Lambda = \begin{pmatrix} 0.66 & 0.95 \\ -1.05 & -0.73 \\ 0.96 & -1.37 \\ 0.84 & 0.59 \\ -0.36 & 0.51 \end{pmatrix}, \; \tilde{\Lambda} = \begin{pmatrix} 1.09 & 0.40 \\ -1.28 & 0.00 \\ 0.00 & -1.68 \\ 1.02 & 0.00 \\ 0.00 & 0.63 \end{pmatrix} \tag{5}$$

where coordinates for Λ are specified in terms of the graph's 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 loadings 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

[2] In (4), $\Sigma_f = I_K$ corresponds to an orthonormal factor basis.

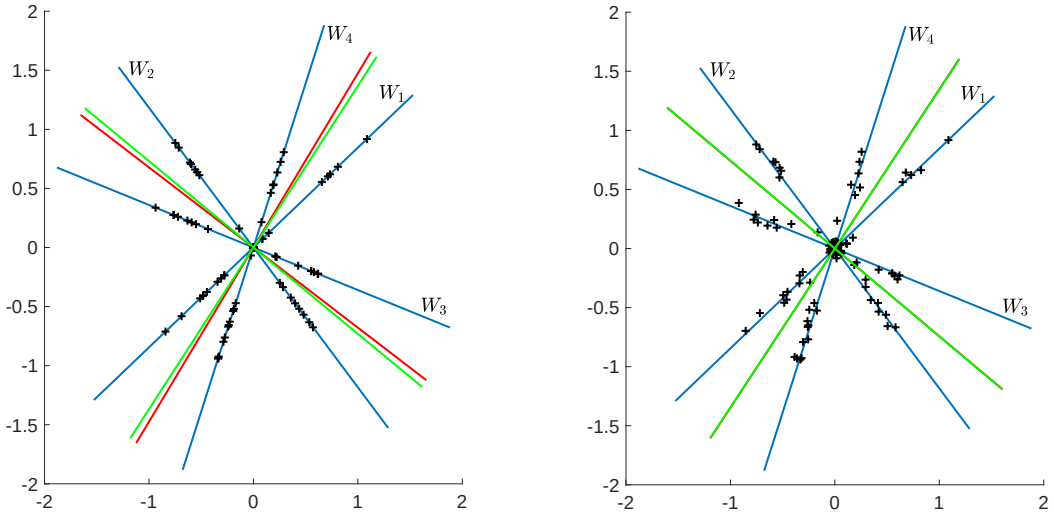
[3] Note that the Stiefel manifold is sometimes defined as the set of all orthogonal K-frames in the $\mathbb{R}^N$ and sometimes defined as the set of all independent K-frames in the $\mathbb{R}^N$. We use the latter - and wider - concept here.

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



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 to identify either second factor. This motivates to base inference on order-invariant estimation and identify factors, including their position 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.



Sometimes, the data may allow for multiple sparse representations. Figure (2) provides an illustration. Consider the left panel, showing 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}, $k_i = 1, \ldots, 4$, may be selected to span a factor basis, $W_{k_i} \in Gr(1, 2)$. Two combinations define an orthogonal factor basis, say $W_1 \perp W_2$, and $W_3 \perp W_4$, such that

either $\lambda_{i\cdot} \in W_1$ and $\lambda_{i\cdot} \in W_2$ for a first subset of sparse loadings, or $\lambda_{j\cdot} \in W_3$ and $\lambda_{j\cdot} \in W_4$ for a 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 of the two sparse representations. 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 representing the data. Such structures may be attributable to measurement errors and an unrestricted estimation may infer all factor loadings different from zero. However, 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 such a "noisy" bimodal representation. Loading vectors that were previously part of the zero space $W_{\{\}}$ are now located near the origin, but not exactly at the origin, whereas the loading vectors that previously fell into the one-dimensional subspaces W_{k_i} are now located near them. An approach designed to discover a sparse representation may end up with either set of orthogonal factors plotted in blue in the picture. As in the left panel of Figure 2, we show the result of a Varimax optimization and least squares minimization, which again span 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, but also including 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 practically at 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 proper factor founder series on which to pre-impose rotational identification 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* - see Aßmann et al. (2016), or Aßmann et al. (2023) for static factor settings with strictly orthogonal factors - 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 the 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. This is a general issue with spike and slab priors that has been discussed in machine learning e.g. by Titsias and Lázaro-Gredilla (2011) and 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.

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

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

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

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 precision u. 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), $\vartheta = \{\beta_{ij}, \rho_j, \tau_j | i = 1, \ldots, N,\ j = 1, \ldots, K\}$.

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

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

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. $\pi(\Lambda|\boldsymbol{y}, \boldsymbol{f}, \Sigma_\epsilon)$. 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 $\pi(\vartheta|\Lambda)$. See Appendix A for the derivation of posterior moments.

2. $\pi(\boldsymbol{f}|\boldsymbol{y}, \theta) = N(\mathrm{f}, \mathrm{F})$ 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 $\boldsymbol{\Lambda} = I_T \otimes \Lambda$.

To explore the full unconstrained posterior distribution, depending on the sampler each iteration is terminated by either random rotation or random permutation: Step 3. of, respectively, the unconstrained rotation or the sparse permutation sampler consists in

3.**U.** (**Unconstrained rotation**) Random rotation of the factor loadings and factors: The output of the unconstrained rotation sampler will display substantial autocorrelation with respect to its orientation. Mixing can be increased using an orthogonal matrix $D \in \mathbb{R}^{K\times K}$, which is drawn such that it is distributed with Haar measure, i.e., uniformly on the K-dimensional hypersphere.[4] The factors and factor-specific parameters are then transformed as

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

   Since the sampler is unconstrained, the entire posterior distribution is explored even if this step is omitted. In more complex setups, however, several hundred thousands iterations may be required to achieve this.

[4] To obtain the desired matrix D, the approach proposed by Mezzadri (2007) is used, sampling a matrix $M \in \mathbb{R}^{K\times K}$ with independent entries from a uniform distribution and applying the QR decomposition as $M = QR'$. Then the desired orthogonal matrix obtains as $D = QS$, where $S = \mathrm{diag}(\mathrm{sgn}(r_{1,1}), \ldots, \mathrm{sgn}(r_{K,K}))$.

3.**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, the k^{th} factor and corresponding loadings incur a sign switch. By 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.

4 Posterior processing: Multiple mode identification

Next, we describe the mode identification techniques using the output of the unconstrained rotation and sparse permutation sampler, based on the geometric representation motivated in Section 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). The posterior distribution is identified up to a final orthogonal transformation by an arbitrary orthogonal

matrix H_*. When appropriately chosen, a rotation H_* can uncover a sparse representation of Λ. 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 target will be to orient the factor basis such that subspaces spanned by as few factor basis axes as possible will intersect with each of the hyperellipsoids.

All parameters characterising the K-dimensional HPD hyperellipsoid for $\lambda_{i\cdot}$ can be inferred from the posterior sample:

center	$c_i \in \mathbb{R}^K$, $c_i = 1/M \sum_{m=1}^{M} \lambda_{i\cdot}^{(m)'}$
rotation (in matrix form)	$H_i \in \mathbf{O}(K)$,
rotation (in angle form)	$\gamma_i \in \mathbb{R}^P$, where $P = \binom{K}{2}$
radii/half-diameters	$r_i \in \mathbb{R}^K$

To obtain H_i, we first compute an estimate of the covariance matrix of $\lambda_{i\cdot}$, which is $\overline{\Psi}_i = 1/M \sum_{m=1}^{M} \lambda_{i\cdot}^{(m)'} \lambda_{i\cdot}^{(m)} - c_i c_i'$. The spectral decomposition $\overline{\Psi}_i = H_i W_i H_i'$ yields H_i, an orthogonal matrix, and W_i, a diagonal matrix with eigenvalues $w_1, \ldots, w_K$ on the diagonal. The Givens decomposition of H_i yields the Givens rotation angles $\gamma_i = (\gamma_{i,1}, \ldots, \gamma_{i,P})'$, where P is the number of axis pairs involved.[5] To obtain $r_i = (r_{i,1}, \ldots, r_{i,K})'$, we work with demeaned and decorrelated draws. We demean the draws $\lambda_{i\cdot}^{(m)}$ to obtain $\lambda_{i\cdot}^{(m),dem} = \lambda_{i\cdot}^{(m)} - c_i'$. Next, we decorrelate the demeaned draws to obtain $\lambda_{i\cdot}^{(m),dec} = \lambda_{i\cdot}^{(m),dem} H_i$. Finally, we standardize the demeaned and decorrelated draws to obtain $\lambda_{i\cdot}^{(m),stand} = \lambda_{i\cdot}^{(m),dem} H_i W_i^{-\frac{1}{2}}$. Denote the empirical $1-\alpha$ quantile of $\|\lambda_{i\cdot}^{stand}\|_2$ by $q_{1-\alpha}$, and determine the radii of the i^{th} hyperellipsoid as $r_{i,k} = q_{1-\alpha}\sqrt{w_k}$.

Figure 3 shows how a two-dimensional ellipsoid $\lambda_{i\cdot}$ may be re-constructed by reverting the steps just described. The first panel in the first row shows the unit circle that consists of the set of points $\{x_i | x_i' x_i = 1\}$. The second panel in the first row shows how the unit circle is expanded to an ellipse by scaling each of its points along the k^{th} dimension with radius $r_{i,k}$, such that the points become $R_i x_i$, with $R_i = \text{diag}(r_{i,1}, r_{i,2})$. The third panel in the first row shows how H_i is used to rotate the ellipse, and the points become $H_i R_i x_i$. Eventually, as shown in the first panel in the second row, the ellipse is shifted, translating all of its points as $H_i R_i x_i + c_i$. The procedure is generically applicable to higher-dimensional ellipsoids, see Appendix B.2 for a $K = 3$-dimensional example.

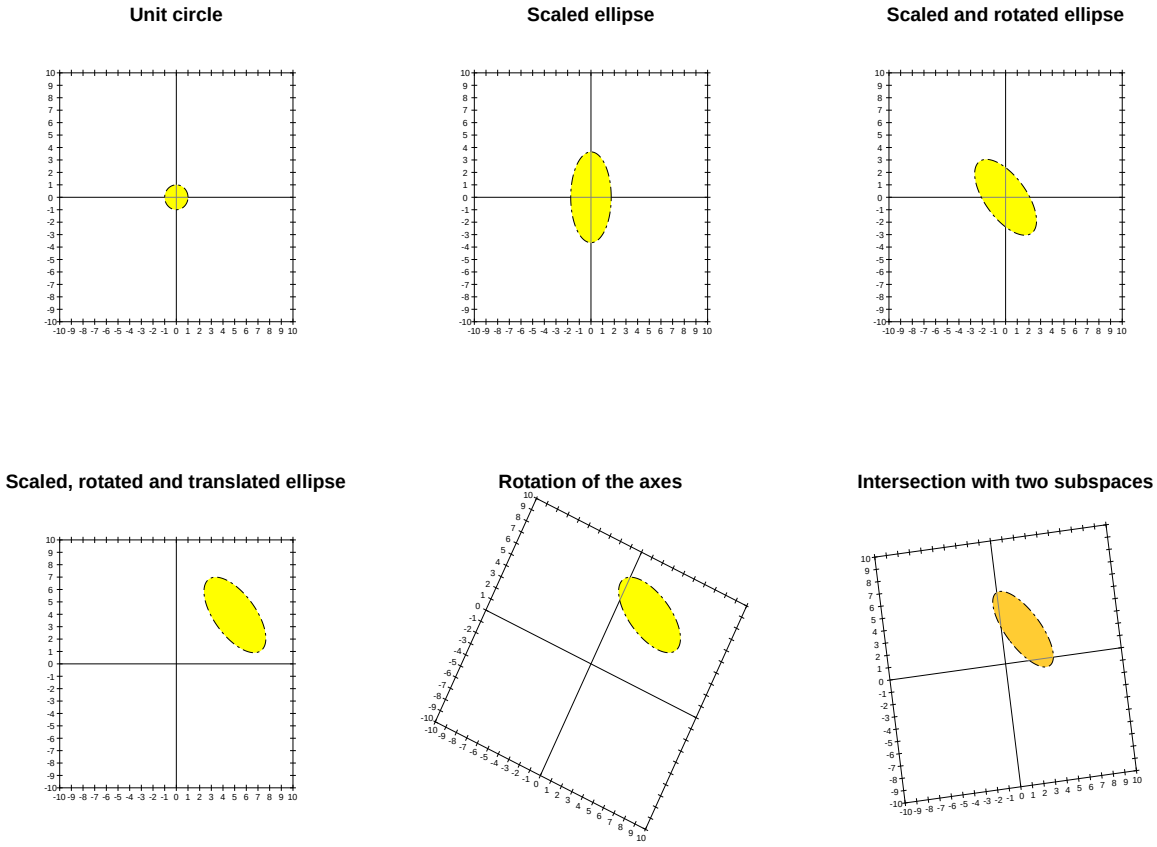
For general $K \in \mathbb{N}$, it holds that any point x_i lies inside the ellipsoid associated with $\lambda_{i\cdot}$ if and only if

$$\|(x_i - c_i)' H_i R_i^{-1}\|_2 < 1, \quad \text{where} \quad R_i = \text{diag}(r_{i,1}, \ldots, r_{i,K}). \tag{14}$$

To identify a sparse representation, we now have to find a rotation matrix H_*, such that as many ellipsoids as possible would intersect with low-dimensional subspaces of the space

[5] The Givens decomposition of a matrix H is described in Appendix B.1.

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.



spanned by H_*. Consider the second panel in the second row of Figure 3, which shows how a rotation of the coordinate axes to the right results in an intersection of the hyperellipsoid with the new second coordinate axis. Both axes are one-dimensional subspaces of the $\mathbb{R}^2$, and the intersection implies that a nonzero loading is only required for the second factor here, whereas the loading on the first factor can be set to zero.

We introduce an indicator matrix Δ to describe the sparse pattern in Λ, a matrix indicating non-zero coordinates of the subspaces that ellipsoid i intersects with, i.e. $\delta_{i,k} = 1$ if $\lambda_{i,k} \neq 0$, and zero otherwise. Note that if the origin is located within the i^{th} hyperellipsoid, the hyperellipsoid intersects with the zero-dimensional space, and hence, all loadings on variable i can be set to zero. That is, $\delta_{i,k} = 0$ for all $k \in \{1, \ldots, K\}$. This continues to hold for arbitrary rotations H_* of the coordinate system. Accordingly, if the i^{th} hyperellipsoid intersects with the k^{th} axis only, loadings on variable i can be set to zero for all factors except the k^{th} one, i.e., $\delta_{i,k} = 1$ and $\delta_{i,j} = 0$ for all $j \in \{1, \ldots, k-1, k+1, \ldots, K\}$.

The i^{th} hyperellipsoid intersects with a subspace of the rotated space if and only if there exists at least one point x_i from that subspace which is located within the hyperellipsoid, i.e., that satisfies the inequality in Equation (14). A natural candidate for x_i is the point

within the subspace at the minimal Mahalanobis distance to the hyperellipsoid's center c_i. We represent x_i by $x_i = H_*(S_{(k_i,j_i)}s_{(k_i,j_i),i})$, where $S_{(k_i,j_i)}$ is a $K \times K$ matrix including the standard vectors spanning the j_i^{th} subspace of dimension k_i, with $0 \leq k_i \leq K$ and $j_i \in \left\{1, \ldots, \binom{K}{k_i}\right\}$, and $s_{(k_i,j_i),i}$ is a $K \times 1$ vector of scaling factors. For instance, for $K = 3$, the matrix that spans the first subspace of dimension $k_i = 2$ corresponds to $S_{(2,1)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$, with $HS_{(2,j_i)} \in Gr(2,3)$ for arbitrary orthogonal matrices $H \in \mathbf{O}(3)$.[6] There are two dimensions with unique subspaces, namely the zero space, spanned by $S_{(0,1)}$, and the $\mathbb{R}^K$, spanned by $S_{(K,1)} = I_K$. To ensure that the point x_i does not fall into a subspace of lower dimension than $S_{(k_i,j_i)}$, we require that each element h of $s_{(k_i,j_i)}$ that scales the h^{th} diagonal element of $S_{(k_i,j_i)}$ is non-zero if the corresponding diagonal element is equal to 1, i.e., $s_{(k_i,j_i)h} \neq 0$ for every $h \in \{1, \ldots, K\}$ with $S_{(k_i,j_i),h,h} = 1$. This implies that every dimension contained in the subspace spanned by $S_{(k_i,j_i)}$ has a non-zero scaling factor.[7] Accordingly, $x_i = H_*(S_{(k_i,j_i)}s_{(k_i,j_i)})$ is a point located in the rotated subspace spanned by $S_{(k_i,j_i)}$, but not within any rotated lower-dimensional subspace.

Conditional on a rotation H_* and $S_{(k_i,j_i)}$, we determine an optimal $s_{(k_i,j_i),i}$, minimizing

$$\ell_{k_i,j_i,i}(S_{(k_i,j_i)}, s_{(k_i,j_i),i}, H_*) = \|(H_*(S_{(k_i,j_i)}s_{(k_i,j_i),i}) - c_i)'H_iR_i^{-1}\|_2,$$

which corresponds to the expression in Equation (14). Eventually, if the inequality in Equation (14) holds for $x_i = H_*(S_{(k_i,j_i)}s_{(k_i,j_i),i})$, the i^{th} hyperellipsoid intersects with the rotated subspace spanned by $S_{(k_i,j_i)}$. In that case, $\delta_{i\cdot}$ corresponds to the diagonal elements of $S_{(k_i,j_i)}$, $\delta_{i\cdot} = \text{diag}(S_{(k_i,j_i)})$. For instance, for $K = 3$, we may find a rotation H_* such that there exists a nonzero vector $s_{(2,1)}$ for $S_{(2,1)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$, so that the inequality in Equation (14) holds for $x_i = H_*\left(S_{(2,1)}s_{(2,1),i}\right)$, i.e., the i^{th} hyperellipsoid intersects with the rotated first two-dimensional subspace of the $\mathbb{R}^3$. Then we have $\delta_{i,\cdot} = \begin{pmatrix} 1 & 1 & 0 \end{pmatrix}$.

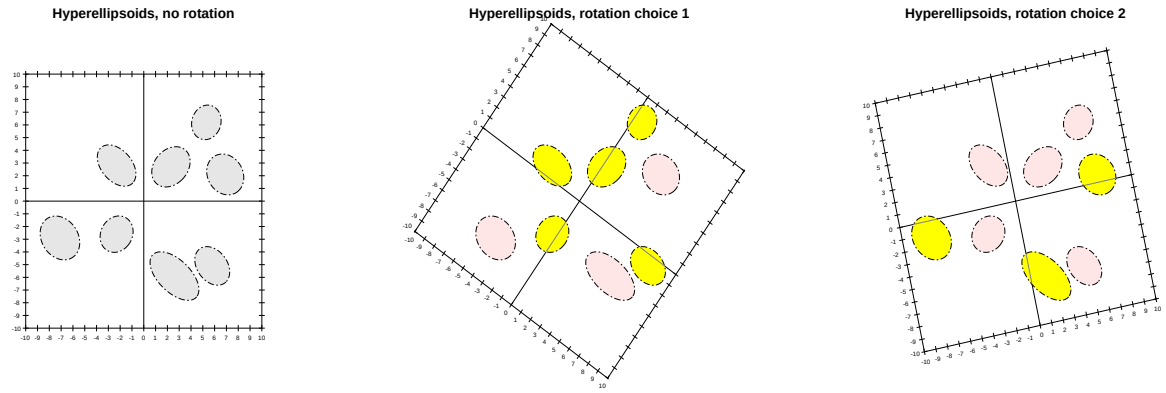
The task is now to find the optimal rotation matrix H_* for the entire coordinate system such that for each i, the point $x_i = H_*\left(S_{(k_i,j_i)}s_{(k_i,j_i),i}\right)$ in the rotated subspace spanned by $S_{(k_i,j_i)}$ falls inside the i^{th} hyperellipsoid, with k_i as small as possible and optimal scaling vector $s_{(k_i,j_i),i}$. When checking for sparsity under some rotation H_*, we therefore start with low-dimensional subspaces, i.e., $k = 1, \ldots, K-1$.

For example, consider the eight hyperellipsoids shown in the first panel of Figure 4. A rotation of the axes to the right by 35 degrees results in an intersection with the rotated axes for five of the hyperellipsoids. This is shown in the second panel, with the intersecting hyperellipsoids highlighted in yellow. A rotation of the axes to the left by 12 degrees results in an intersection with the rotated axes for the remaining three of the

[6] Note that for $2 \leq k \leq K-1$, this representation is redundant insofar as changes in H can accommodate for a different choice of $S_{(k_i,j_i)}$, however, as we consider multiple row vectors in our loading matrices, it is vital to distinguish between subspaces of equal dimension, as the orthogonal matrix H rotates each row vector towards one of these subspaces, albeit not necessarily the same subspace.

[7] For simplicity, we may also require that all $s_{(k_i,j_i),h} \neq 0$, since a scaling factor applied to a zero element of $S_{(k_i,j_i)}$ has no effect anyway.

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).



hyperellipsoids. This is shown in the third panel, with the intersecting hyperellipsoids again highlighted in yellow. Note that since the axes are one-dimensional subspaces of the $\mathbb{R}^2$, intersections of hyperellipsoids with an axis imply that a nonzero loading is only required for the factor corresponding to this axis. The first proposed rotation therefore results in two variables with nonzero loadings on the first factor only, three variables with nonzero loadings on the second factor only, and three variables with nonzero loadings on both factors. Accordingly, the second proposed rotation results in two variables with nonzero loadings on the first factor only, one variable with nonzero loadings on the second factor only, and five variables with nonzero loadings on both factors. This example also illustrates that reflections (corresponding to a sign change in factors) and permutations (change in factor order) do not induce changes in the number of zero or non-zero elements in each row of Δ. It hence suffices to run an optimization over the $\binom{K}{2}$ rotation angles required for a Givens decomposition of H_*.

In this context, it must be noted that the lowest-dimensional subspace k_i, $0 < k_i < K$, intersecting with the i^{th} hyperellipsoid may not be unique. Hence, if the same rotation H_* is applied to two distinct subspaces of the same dimension, say $S_{(k_i,j_{i,1})}$ and $S_{(k_i,j_{i,2})}$, where $j_{i,1} \neq j_{i,2}$, then both $x_{i,1} = H_*(S_{(k_i,j_{i,1})}s_{(k-i,j_{i,1}),i})$ and $x_{i,2} = H_*(S_{(k_i,j_{i,2})}s_{(k_i,j_{i,2}),i})$ may satisfy the inequality in Equation (14), but $x_i = H_*(S_{(l_i,j_i)}s_{(l_i,j_i),i})$ with $S_{(l_i,j_i)} = S_{(k_i,j_{i,1})}S_{(k_i,j_{i,2})}$ may not, where $S_{(l_i,j_i)}$, $l_i < k_i$, spans the l_i-dimensional subspace that combines the zero dimensions of both k_i-dimensional subspaces $S_{(k_i,j_{i,1})}$ and $S_{(k_i,j_{i,2})}$. The third panel in the second row of Figure 3 shows the most simple example of this situation. Under a rotation to the coordinate system by H_*, the ellipse in the $\mathbb{R}^2$ overlaps with both axes, i.e., distinct one-dimensional subspaces, but does not contain the origin, i.e., the unique zero-dimensional subspace. This means that either $\delta_{i,1} = 0$ or $\delta_{i,2} = 0$, but not both. Technically, this constitutes the possibility of multiple different sparse representations of Λ for the same choice of H_*.

With these topological considerations at hand, we now propose an algorithm in order to determine Δ. The algorithm proceeds as follows:

1. Find all hyperellipsoids containing the origin of the coordinate system, i.e. those variables i for which

$$\left\|(0_K - c_i)' H_i R_i^{-1}\right\|_2 < 1 \tag{15}$$

holds. Set $\delta_{i,k} = 0$ for all $k \in \{1, \ldots, K\}$. These hyperellipsoids do not need to be taken into account in the subsequent optimization.

2. Find the optimal rotation of the coordinate system

$$H_*^{opt} = \arg\min_{H_*} \left\{ \sum_{i=1}^{N} \ell_i^*(H_*) + \zeta \sum_{i=1}^{N} \sum_{k=1}^{K} \delta_{i,k}(H_*) \right\}, \tag{16}$$

where

$$\ell_i^*(H_*) = \ell_{k_i,j_i,i}(S_i^*(H_*), s_i^*(H_*), H_*) \quad \text{for every} \quad i \in \{1, \ldots, N\}, \tag{17}$$

and the second term in Equation (16) counts the number of nonzero elements in Δ and scales the result with a penalty term ζ. In simulations and applications, we used $\zeta = 10$. H_* is conveniently expressed as a function of the Givens rotation angles.

In the following, we explain in detail how the expression in Equation (17) is obtained for each hyperellipsoid and how $\delta_{i,k}(H_*)$ in Equation (16) is determined in cases not covered by step 1 of the algorithm.

First, consider the loss function

$$\ell_{k_i,j_i,i}(S_{(k_i,j_i)}, s_{(k_i,j_i),i}, H_*) = \|(H_*(S_{(k_i,j_i)} s_{(k_i,j_i),i}) - c_i)' H_i R_i^{-1}\|_2. \tag{18}$$

(a) For given S_{k_i,j_i} and H_*, the optimal vector of scaling factors is obtained as

$$s^*_{(k_i,j_i),i}(H_*) = \arg\min_{s_{(k_i,j_i),i}} \ell_{k_i,j_i,i}(S_{(k_i,j_i)}, s_{(k_i,j_i),i}, H_*). \tag{19}$$

(b) For given H_*, we attempt to find $S_{(k_i,j_i)}(H_*)$ for every hyperellipsoid by evaluating Equation (18) for the k_i-dimensional subspaces of the $\mathbb{R}^K$ spanned by $S_{(k_i,j_i)}(H_*)$, starting with $k_i = 1$. If for some $k_i \leq K - 1$,

$$\ell_{k_i,j_i,i}(S_{(k_i,j_i)(H_*)}, s^*_{(k_i,j_i),i}(H_*), H_*) < 1 \tag{20}$$

holds, the i^{th} hyperellipsoid intersects with the subspace spanned by $S_{(k_i,j_i)}(H_*)$ and rotated by H_*, and we can set $\delta_{i\cdot} = \text{diag}(S_{(k_i,j_i)}(H_*))$. Subspaces with $k > k_i$ do not have to be investigated in this case, since if a lower-dimensional subspace intersects with a hyperellipsoid, then all spaces of higher dimension containing the lower-dimensional subspace also intersect with the hyperellipsoid. Conversely, if Equation (20) holds for no $S_{(k_i,j_i)}(H_*)$ with $k_i \leq K - 1$, we set $\ell_i^*(H_*) = 0$ in Equation 17 and $\delta_{i\cdot} = 1'_K$.[8]

[8]Note that $\ell_i^*(H_*) = 0$ implies that c_i falls into the $\mathbb{R}^K$, which is indeed true. Note further that the penalty for nonzero loadings in this case is ζK, so the contribution of this hyperellipsoid to the total value of the loss function in Equation (16) is large.

If multiple subspaces of equal dimension k_i intersect with the hyperellipsoid, corresponding to the situation depicted in the third panel in the second row of Figure 3, we may choose among different strategies, such as always selecting the subspace for which Equation (20) yields the smallest loss (which is the approach we choose in the following), or selecting from the candidate subspaces randomly, or selecting a subspace that excludes the axes where we are particularly interested in having zero loadings for variable i.

(c) If there exists an $S_{(k_i,j_i)}(H_*)$ satisfying Equation (20), and, in case of nonuniqueness, has been chosen according to one of the strategies described in (b), we introduce the shorter notation $S_i^*(H_*) = S_{(k_i,j_i)}(H_*)$, and the corresponding optimal vector of scaling factors from Equation (19) is denoted as $s_i^*(H_*)$. This yields the expression in Equation (17).

Note that due to the aforementioned Givens decomposition, the algorithm may proceed through the axes in a pairwise fashion, always addressing two dimensions of the hyperellipsoids at a time. The process of finding the optimal overall rotation matrix H_*, described in step 2 of the algorithm, is thus replaced by $P = \binom{K}{2}$ pairwise optimizations. Each pairwise optimization is then concerned with one specific pair of axes at a time, implying $K = 2$ in step 2 of the algorithm, reducing the number of subspaces to be considered to merely two.[9] A few additional changes are implied accordingly. Denote the axes involved in one pairwise optimization as k_1 and k_2. Then the optimization can be simplified by first finding those hyperellipsoids for which the inequality in Equation (15) holds if only elements k_1 and k_2 of the resulting vector on the left-hand side are considered. These hyperellipsoids do not have to be taken into account in the current pairwise optimization, as $\delta_{i,k_1} = \delta_{i,k_2} = 0$ already holds, which cannot change as a result of a pairwise rotation. Furthermore, in step 2 of the algorithm, we replace the initial full rotation matrix H_* by a Givens rotation matrix around the axes k_1 and k_2. And eventually, every pairwise optimization needs to take the outcomes of previous pairwise optimizations into account. To achieve this, recall our earlier observation that rotating the coordinate system by a matrix H is equivalent to rotating all hyperellipsoids by its transpose H'. Say that in the first pairwise optimization, we determine the optimal rotation between the first two axes, denoted $H_{*(1,2)}^{opt}$. To incorporate the effect of this first pairwise optimization, the hyperellipsoids are afterwards rotated by ${H_{*(1,2)}^{opt}}'$. In the next pairwise optimization, the optimal rotation between the first and third axes $H_{*(1,3)}^{opt}$, conditional on $H_{*(1,2)}^{opt}$, is determined. After this pairwise optimization, the hyperellipsoids undergo the corresponding rotation by ${H_{*(1,3)}^{opt}}'$, and so forth. Each of the P pairwise optimizations involves a single angle parameter γ_p at a time that determines the corresponding $H_{*(k_1,k_2)}^{opt}$, rather than one optimization involving P angle parameters at once, and the full optimal rotation matrix can be rebuilt as

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

[9] The null space and the $\mathbb{R}^2$ are unaffected by all possible choices for the pairwise rotation here, so they do not need to be considered, and we only need the two one-dimensional subspaces of the $\mathbb{R}^2$.

where $H_{*,p}^{opt}$ denotes the optimal rotation about the p^{th} pair of axes. Note that to prevent the algorithm from getting trapped in local optima, the optimizations over the axis pairs may be performed repeatedly and in random order.

Independently of whether we use a joint or a pairwise optimization, we may want the algorithm to find more than a single mode. To achieve this, we can adjust the loss function that determines H_*^{opt}, such that it takes a set $\mathcal{H} = \{H_*^{prev,1}, \ldots, H_*^{prev,L}\}$ of L previously found rotation matrices into account, i.e.,

$$H_*^{opt} = \arg\min_{H_*} \left\{ \sum_{i=1}^{N} \ell_i^*(H_*) + \zeta \sum_{i=1}^{N} \sum_{k=1}^{K} \delta_{i,k}(H_*) + \psi \sum_{l=1}^{L} \|H_*' H_*^{prev,l} - P_s^*\|_2 \right\},$$

where $\psi > 0$ is a suitably chosen term to penalize solutions for H_*^{opt} that are too similar to previous solutions contained in $\mathcal{H}$, and

$$P_s^* = \arg\min_{P_s} \left(\|H_*' H_*^{prev,l} - P_s\|_2 \right)$$

is the K-dimensional signed permutation matrix with minimal distance to $H_*' H_*^{prev,l}$, in order to reflect that two different choices for H_*^{opt} may indeed be equivalent, save for rearranging columns and switching their signs.

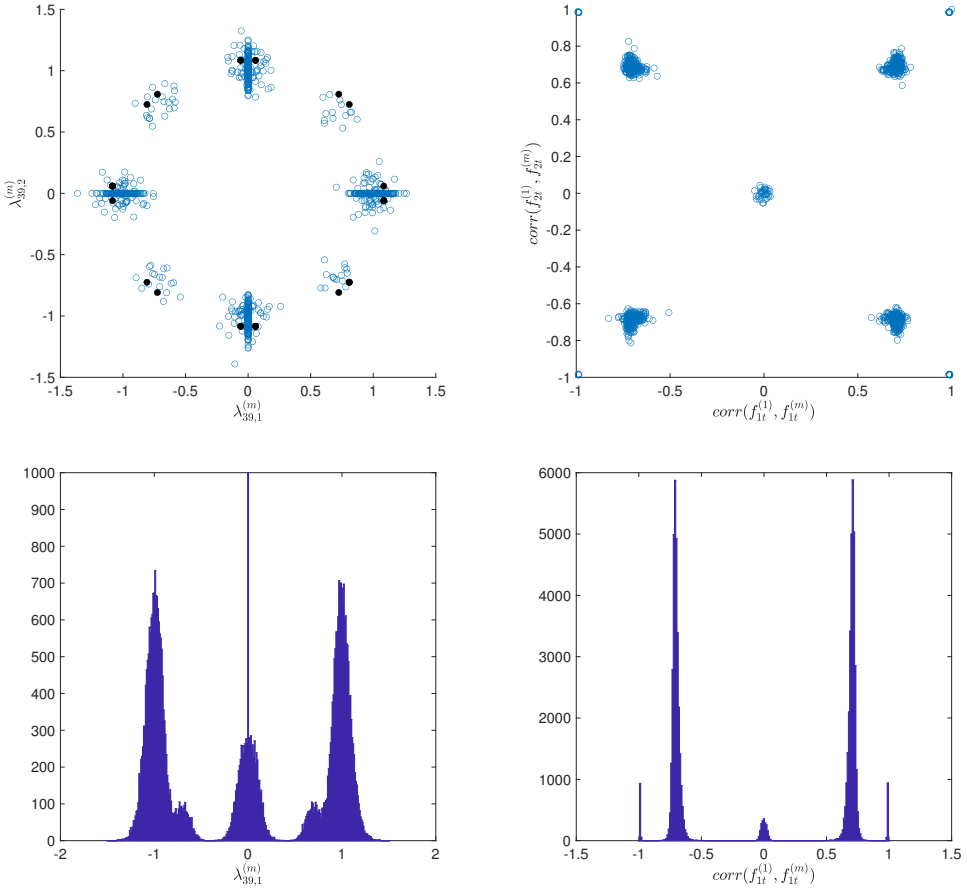
4.2 Mode identification using the output of the sparse permutation sampler

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 series, where the black dots represent all permutations of the true loadings for each sparse modes.[10] 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 a 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 specific series's factor loading only. When the number of factors is larger or sparse patterns are more complex, it may become difficult to determine a restriction discriminating between sparse modes in the first stage. 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.

[10] See Subsection 5.1 for a description of the simulation settings. We plot the output for the scenario data50ex_ln, where loadings of series 39 in the first and second modes are, respectively, $\lambda_{39,\cdot} = [0.81, -0.72]$ and $\lambda_{39,\cdot} = [.06, -1.08]$.

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 (b)).

(a) Unsorted draws



(b) Sorted draws

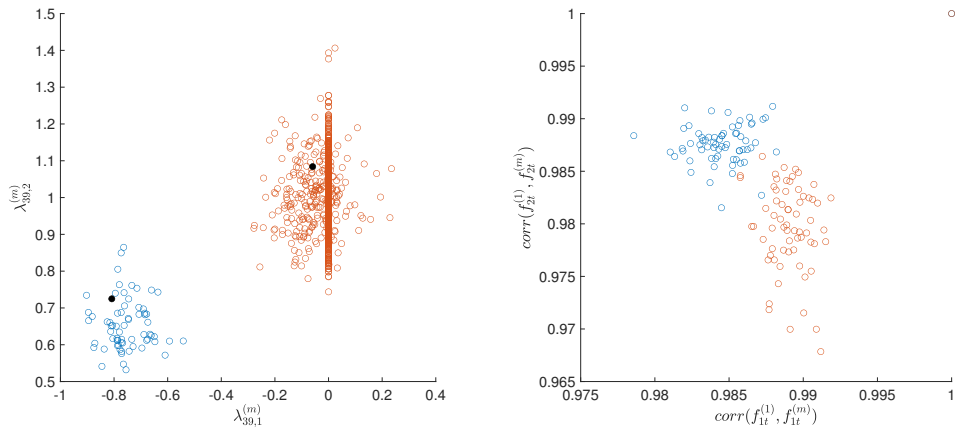
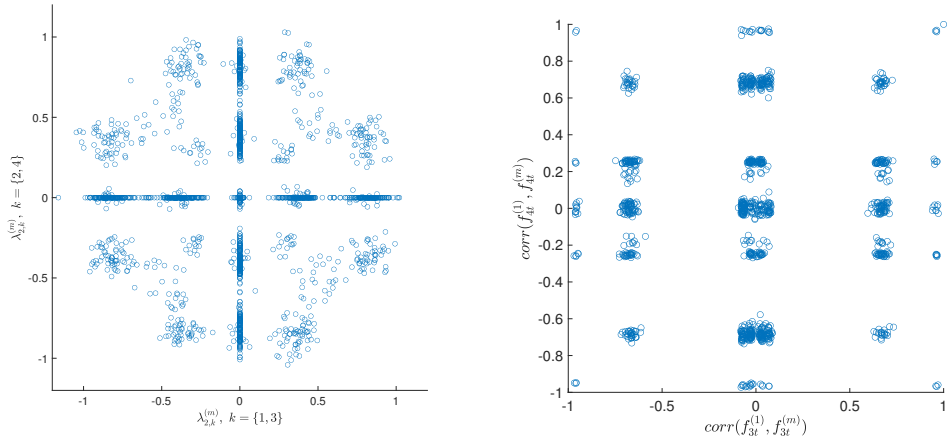
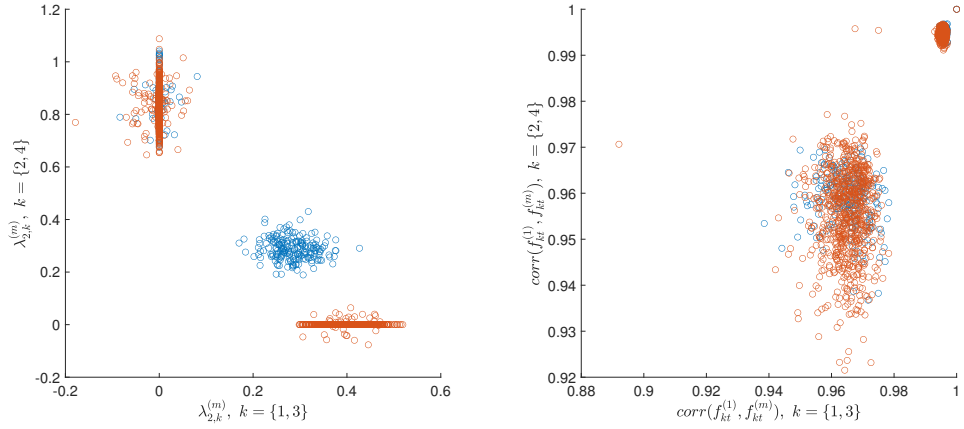


Figure 6: MCMC output for the scenario $K = 4$ factors with two pervasive factors and two sparse modes, K4m2_2pf_ln in Table 4. Left panels: Scatter plot of factor loadings for the second series; right panels: Scatter plots of correlations across draws of the first (third) factor against correlations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



However, factor draws 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 combining those K factors. The number of filled factor combinations 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.3 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 draws 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 combination $\mathcal{I}_Z = \{Z_1, \ldots, Z_K\} \subset \{1, \ldots, G\}$, $Z = 1, \ldots \binom{G}{K}$ has been drawn.

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.

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.

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 gives detailed explanations how such minimally correlated subspaces can be constructed.

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

The unconstrained rotation approach from Section 4.1 is applied to sequences of length 200,000 each, and the algorithm attempts to find two distinct modes, applying a penalty to the first mode when looking for the second mode, as described in step 5 of the algorithm.

Throughout, the highest posterior density 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.[11]

Table 1: $K = 2$, unconstrained rotation, $\alpha = 0.05$: The second column dislays which mode was found 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
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 shows the low-noise scenarios. Both simulated modes are almost always perfectly recovered for the exact sparsity scenarios. In the low-noise scenarios with approximate sparsity, in almost all cases one of the modes is perfectly recovered, with only 1 or 2 false zeros or non-zeros. One exception (data80ap_ln) produces a larger number of false non-zeros. However, the magnitude of false zero and non-zero loadings is small.

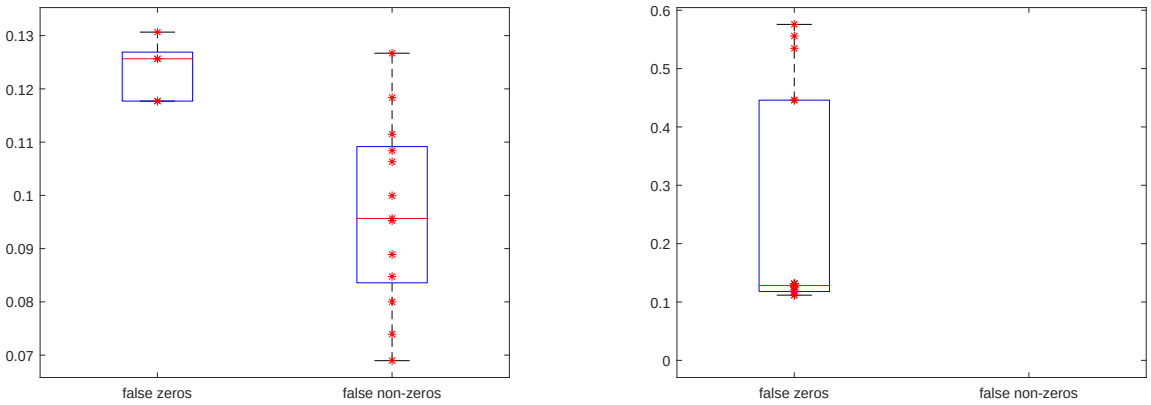
The lower part of the table displays the high-noise scenarios. In none of these, any false non-zeros are found. In three cases, both modes are recovered perfectly, and one mode

[11] 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 loading pattern.

is perfectly recovered in one case (data80ex). The number of false zeros is small, ranging from 1 to 3, but the average 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 false zeros and non-zeros across scenarios.

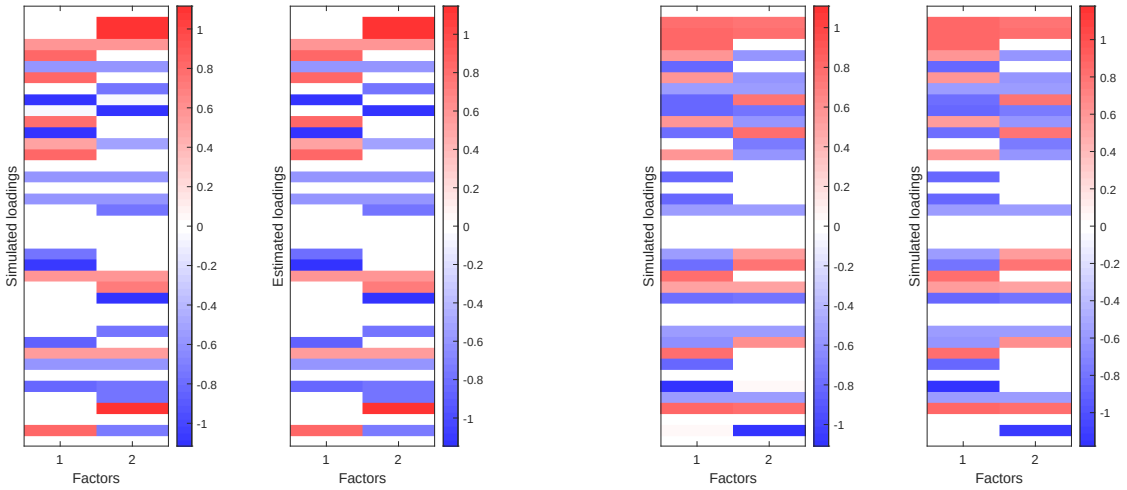
Figure 8 displays heat plots of simulated (left) and estimated (right) loadings for both modes of the scenario data50ex_thur_ln. They confirm that sparse patterns are well recovered by the procedure described in Subsection 4.1.

Figure 7: $K = 2$, unconstrained rotation: Boxplot of average absolute factor loadings (absolute values), 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 higher noise scenarios, hence no boxplot to display.



Lower noise ratio | Higher noise ratio

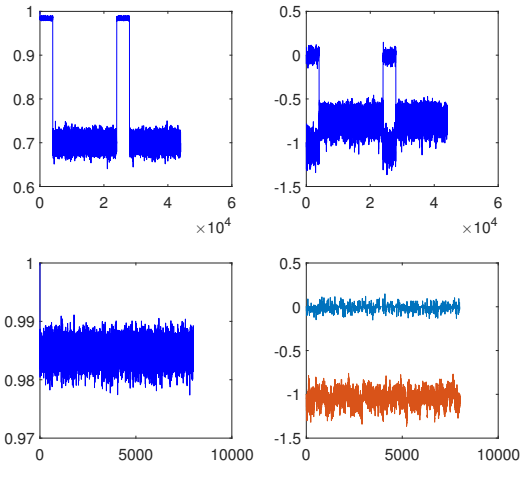
Figure 8: Heat plot of factor loadings, $K = 2$, scenario data50ex_thur_ln, estimated based on the unconstrained rotation approach.



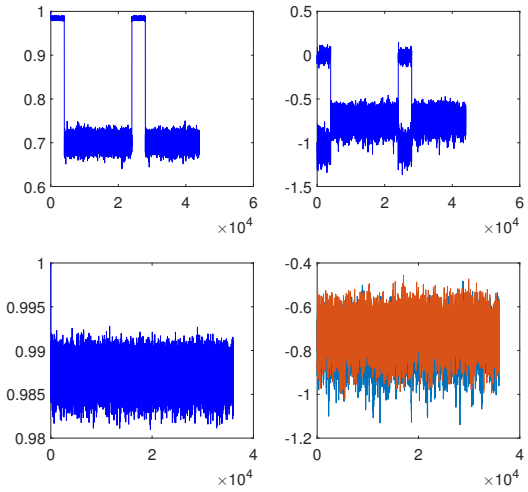
Mode 1 | Mode 2

Applying the sparse permutation sampler from Section 4.2 to the simulated data, 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: Posterior draws, unsorted and sorted, $K = 2$, scenario data50ex_thur_ln. 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 false zeros and non-zeros across low- and high-noise scenarios.

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

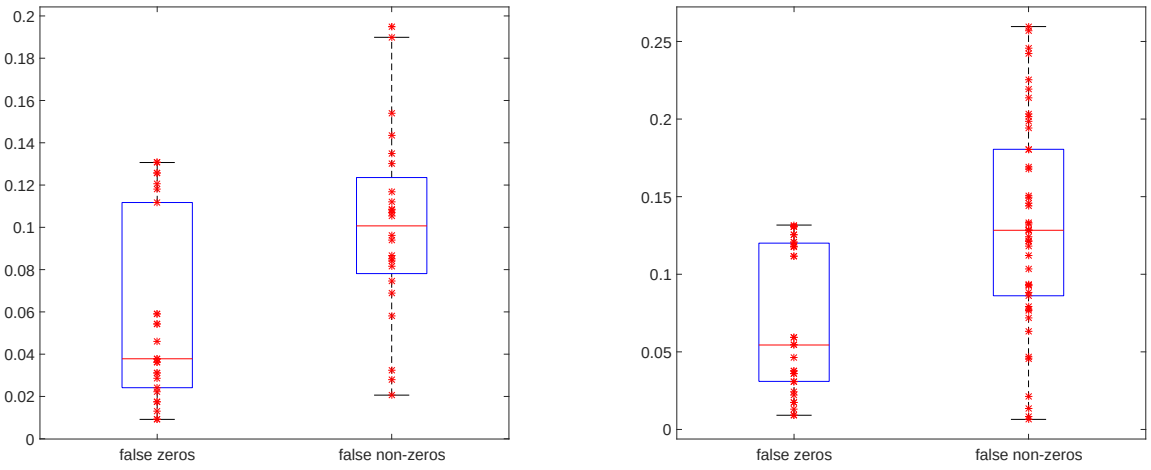
Scenario	Draws	False zeros		False non-zeros		Matching indices	
		Number	Average	Number	Average	Jaccard	Simple score
data50ex_thur_ln	8,000	8	0.03	0	-	0.83	0.90
	36,000	2	0.05	0	-	0.96	0.97
data50ex_ln	4,000	8	0.03	0	-	0.83	0.90
	40,000	2	0.05	2	0.08	0.92	0.95
data50ap_thur_ln	4,000	0	-	1	0.08	0.98	0.99
	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		

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. Also this procedure is able to recover well both underlying simulated sparse structures.

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

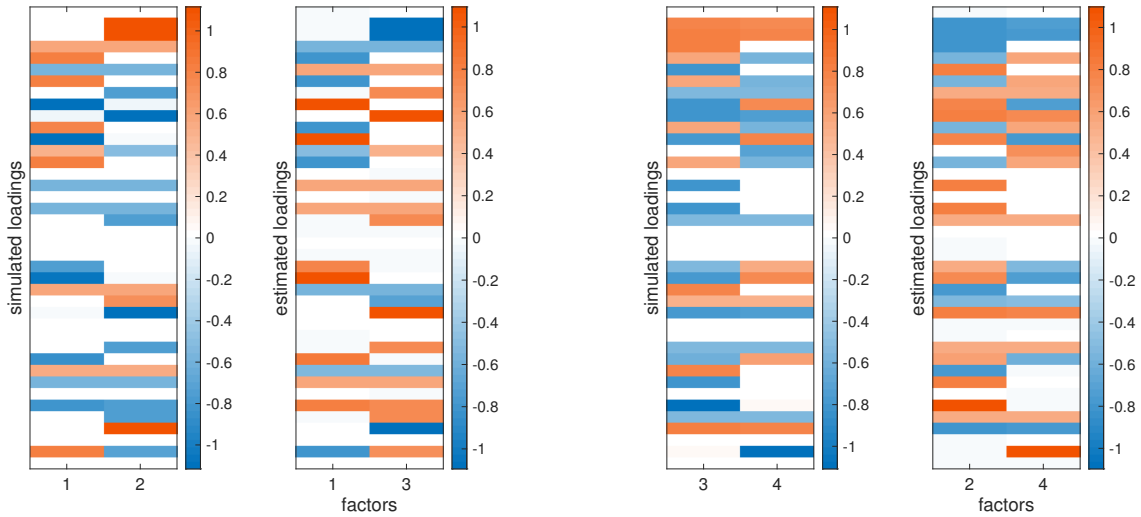
For $K = 3$ we simulate a pervasive factor, denoted as *1pf*, i.e., a strong factor driving all variables, and two weaker, i.e., local or group-specific, factors, which can be represented by two underlying sparse loading structures. For $K = 4$, we simulate one or two pervasive factors, denoted as *1pf* or *2pf*, complemented with, respectively, three or two weaker

Figure 10: Boxplot of factor loading (absolute values), $K = 2$, estimated with the sparse permutation sampler, pooling over scenarios and factors. 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.



Lower noise ratio Higher noise ratio

Figure 11: Heat plot of factor loadings, $K = 2$, scenario data50ex_thur_ln, estimated with the sparse permutation sampler.



Mode 1 Mode 2

factors. Again, we simulate data with a high and low signal-to-noise ratios, where the former is denoted as *ln*. Combining these features yields six settings, from which we simulate $N = 60$ series of length $T = 100$ each.

The ex-post approach from Section 4.1 is applied to sequences of length 200,000 each, as in the case of $K = 2$ factors, and the algorithm again attempts to find two distinct modes, applying a penalty to the first mode when looking for the second mode.Throughout, the

highest posterior density intervals are constructed with $\alpha = 0.05$.

Table 3 shows for the each scenario the comparison of each estimated mode with the closest simulated mode. It reports the number of false zeros and non-zeros per mode, and the Jaccard and simple matching coefficients between simulated and estimated factor loading structure for each mode. 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 scenario with higher noise, 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 reaches up to 39, and up to 12 false non-zeros are estimated. Especially in the scenario K4m2_2pf, the average across the 12 false non-zeros is 0.4 or 0.5, which indicates that the estimated sparse representations are somewhat different from the simulated ones. For the scenarios with two pervasive factors, the number and the average absolute value of false zeros seem very large. There is a good explanation for this feature, however, discussed below. The left panel in Figure 12 visualizes some details of Table 3 by displaying box-plots of average true values of false zeros and estimated values of the false non-zeros, pooled across scenarios with two pervasive factors.

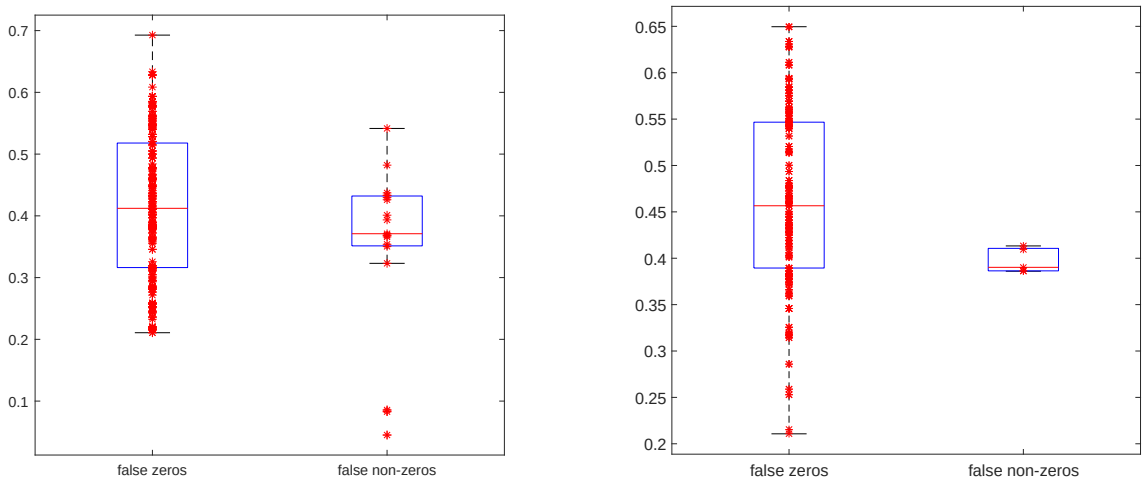
Table 3: $K = 3$, $K = 4$, unconstrained rotation, $\alpha = 0.05$: The second column shows which mode was found first and which second. Absolute true and estimated average are reported for, respectively, false zeros and non-zeros.

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, of each panel. Note that the factor orderings have been adjusted for better fit. In both top panels we see that the estimated loading structure for the pervasive factors is more sparse than for simulated factors, which reflects the large number reported false zeros in Table 3. However, the posterior rotation approach precisely induces sparsity. The bottom panels display a Varimax rotated version of the simulated pervasive factors in each left-hand heat plot. We see that the posterior rotation solution identifies a sparse structure closely resembling the Varimax rotation of simulated loadings. For the non-pervasive factors, there are several deviations, which reflects 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

Figure 12: Scenario K4m2_2pf and K4m2_2pf_ln, pooling over modes. Boxplot of factor loadings (absolute values). 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 output (b) Sparse permutation output



(a) (b)

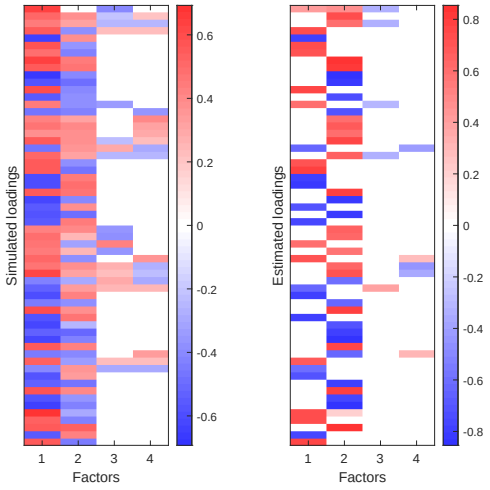
of a factor loading draw taken from the initial chain. We again retain the last 4,000 of each chain to obtain 64,0000 draws for posterior inference.

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 one of simulated modes is substantially larger in the remaining scenarios.[12] 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 throughout, and the number of false zeros is much smaller than in the analysis based on the unconstrained rotation approach. The true loadings in the case 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 the scenarios with two pervasive factors, we also identify a large number and average true value of false zeros. The explanation is the same as for results obtained by posterior rotation.

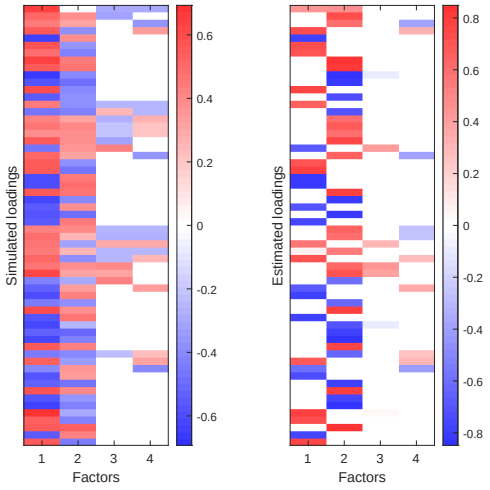
Figure 14 displays heatplots of simulated and estimated factor loadings for the scenario K4m2_2pf. As for posterior rotation, the sparse structure identified for the two pervasive factors comes close to a Varimax rotation of the simulated loadings (bottom panels). Overall, the loading structure of the two non-pervasive factors is recovered quite well for both modes, which clearly outperforms posterior rotation.

[12]Figure 28 in Appendix C.2 displays an example for sorted and unsorted posterior draws for the scenario K4m2_2pf, similar to Figure 9.

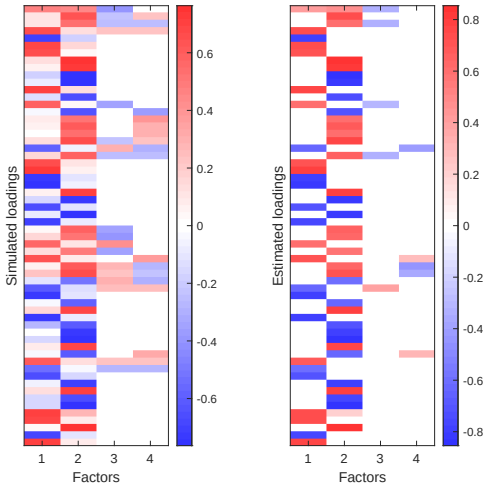
Figure 13: Heat plot of factor loadings, $K = 4$, scenario K4m2_2pf, unconstrained rotation. Second line: Varimax rotation of simulated loadings for the two pervasive factors.



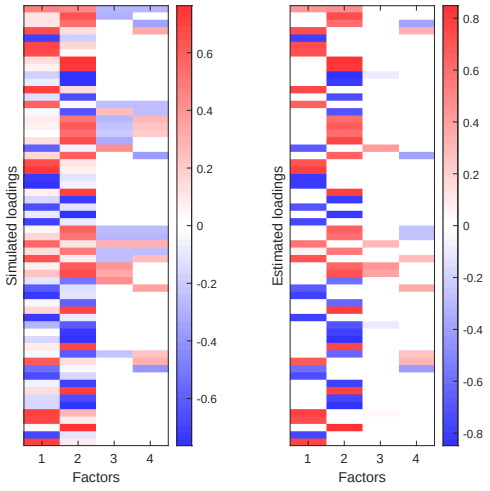
Mode 1



Mode 2



Mode 1

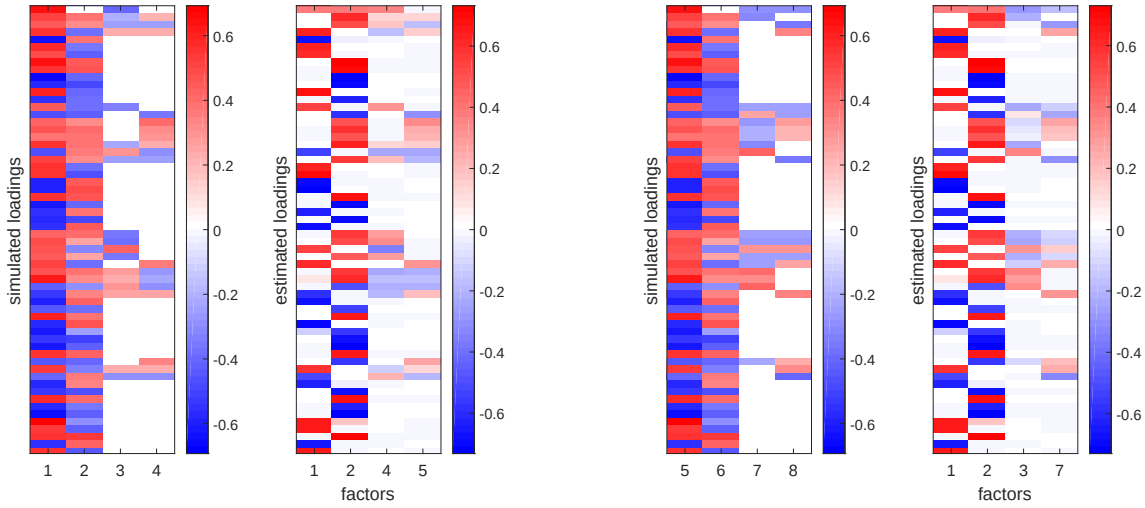


Mode 2

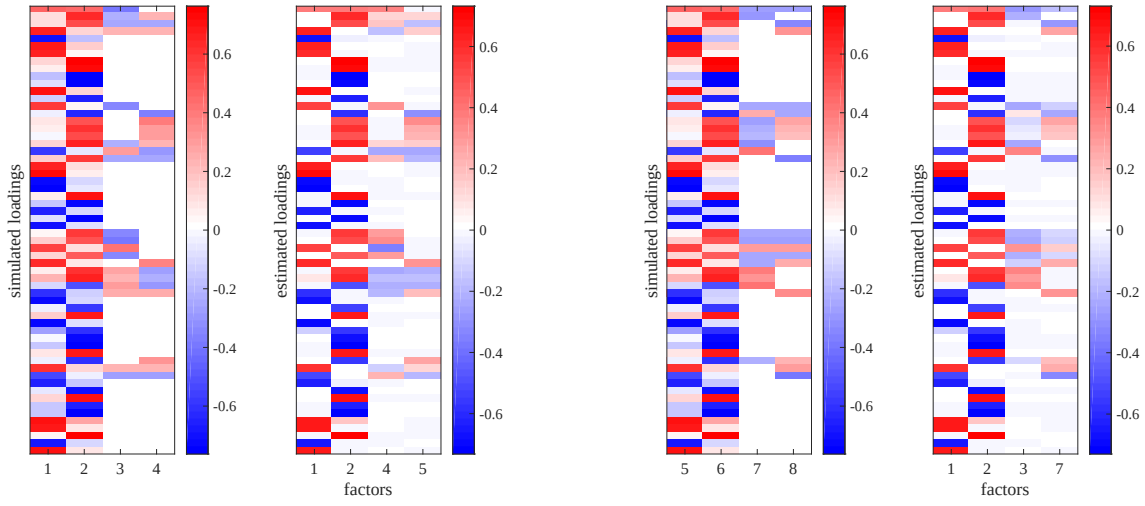
Table 4: Evaluations of the results obtained from the sparse permutation sampler: The first (second) line evaluates the first (second) mode. The column labelled Draws reports the number of posterior draws assigned to the simulated mode.

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: Heat plot of factor loadings, $K = 4$, scenario K4m2_2pf, estimated based on the sparse permutation sampler. Second line: Varimax rotation of simulated loadings of the two pervasive factors.



Mode 1 Mode 2



Mode 1 Mode 2

6 Applications

We revisit the datasets used in Kaufmann and Schumacher (2017): Monthly inflation 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.

6.1 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: Estimated factor loadings, based on the unconstrained rotation approach.

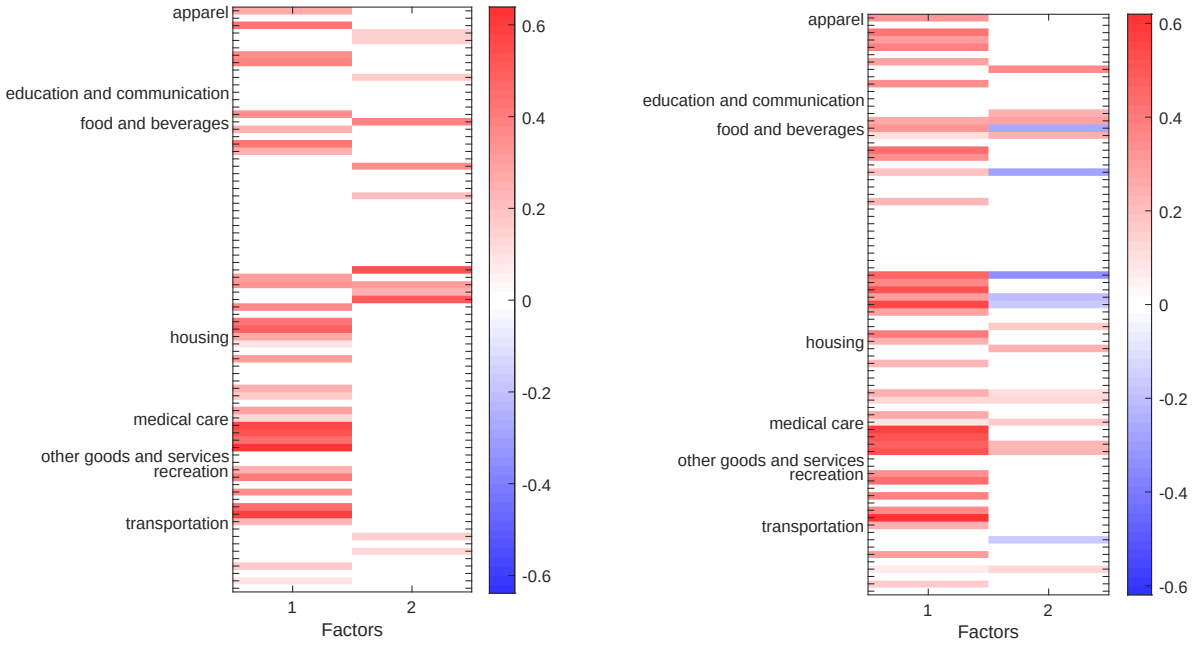
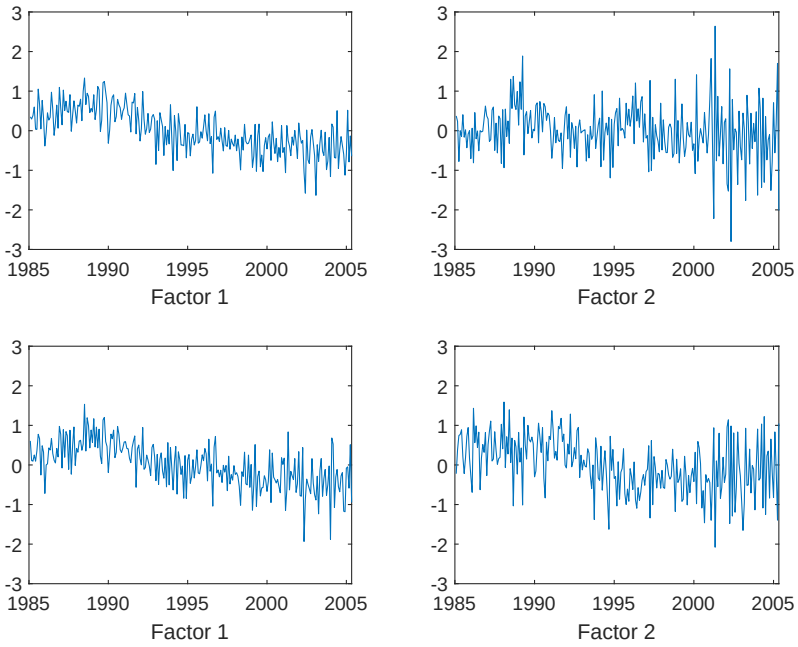


Figure 15 shows the loadings patterns identified by the unconstrained rotation approach. For the first mode, shown in the left panel, there are 33 non-zero loadings on the first factor and 12 non-zero loadings on the second factor. All loadings are nonnegative. For the second mode, shown in the right panel, there are 39 non-zero loadings on the first

factor and 18 non-zero loadings on the second factor. In this case, there are several negative loadings on the second factor.

Were the first factor pervasive, also factor loadings should be similar across modes. This is not quite the case shown in Figure 15 and hard to assess from mean factor plots displayed in Figure 16. However, the correlation across Factors 1 in 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. We conclude that there is evidence for a pervasive factor, despite the fact that the algorithm penalizes similar sparse solutions and optimizes to find differences in all factors.

Figure 16: US CPI: Mean factors, based on the unconstrained rotation approach.



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.3 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. See in Table 5 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.[13]. 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

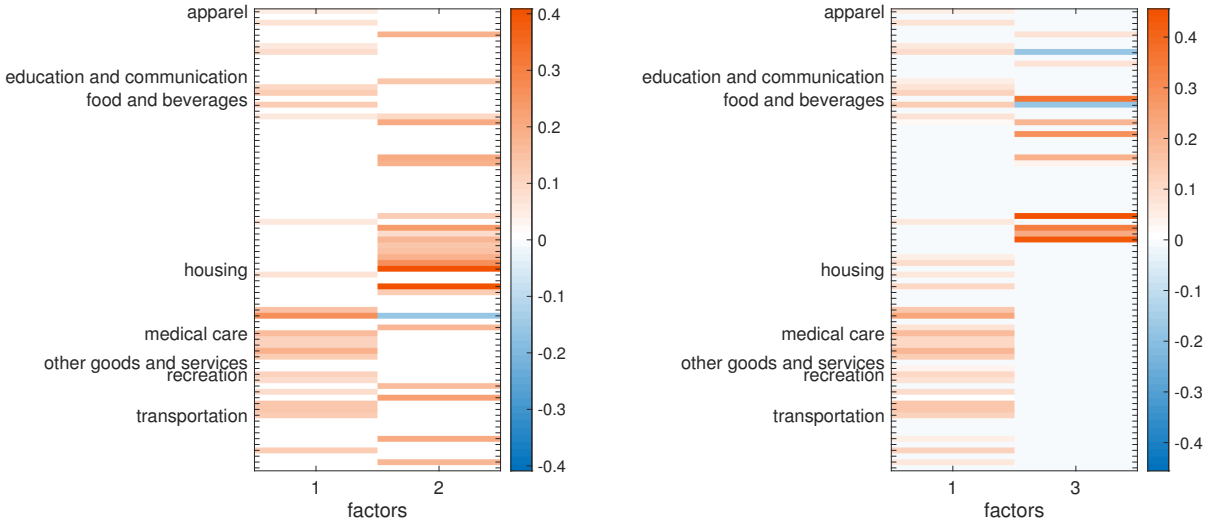
[13]Non-zero loadings are determined by loadings for which the median posterior probability of a non-zero factor loading is larger than 0.5, $\hat{\beta}_{ij} > 0.5$, with $\hat{\beta}_{ij} = median(\beta_{ij}^{(m)})$.

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

Table 5: US CPI: Sorted output

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: Median factor loadings, based on the sparse permutation sampler.



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.

Figure 18: US CPI: Factor correlations, across factor combinations

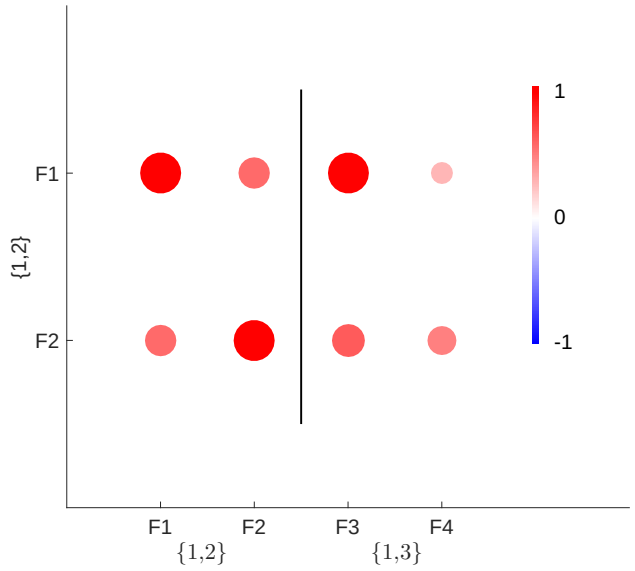
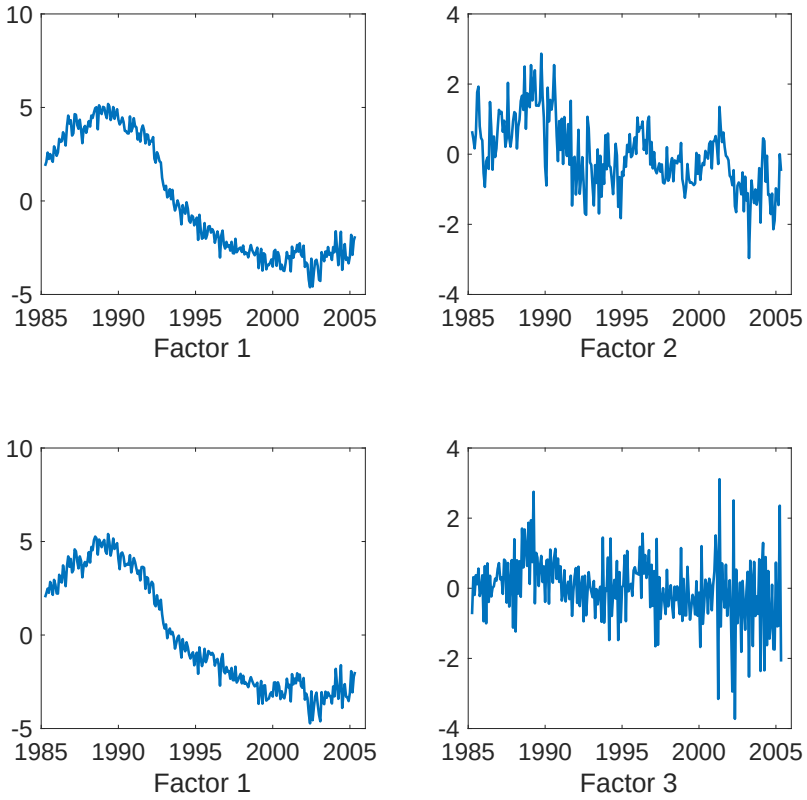


Figure 19: US CPI: Mean factors, based on the sparse permutation sampler.



6.2 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 20: GDP growth, unconstrained rotation: Mean factor loadings

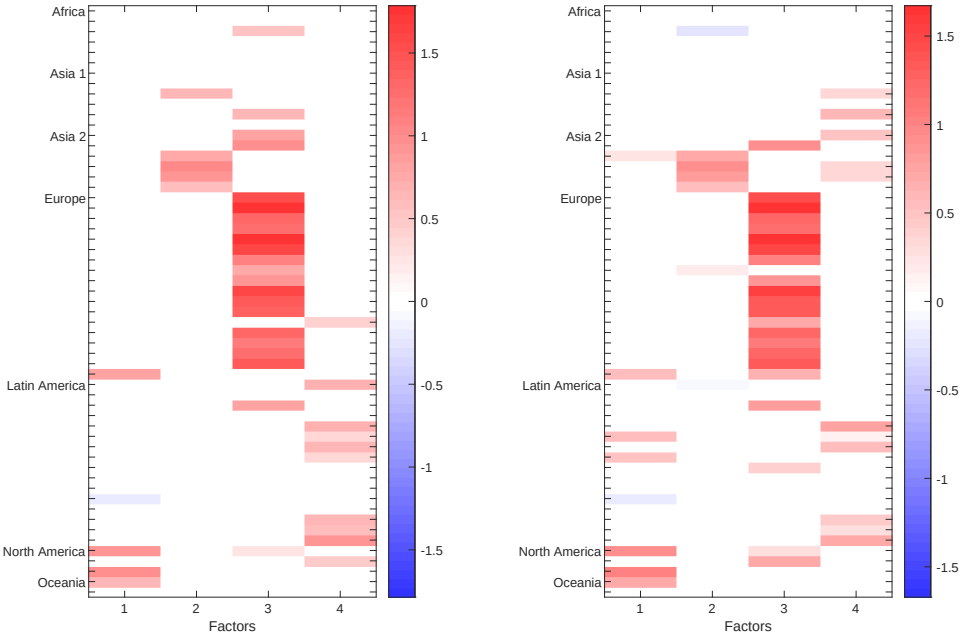
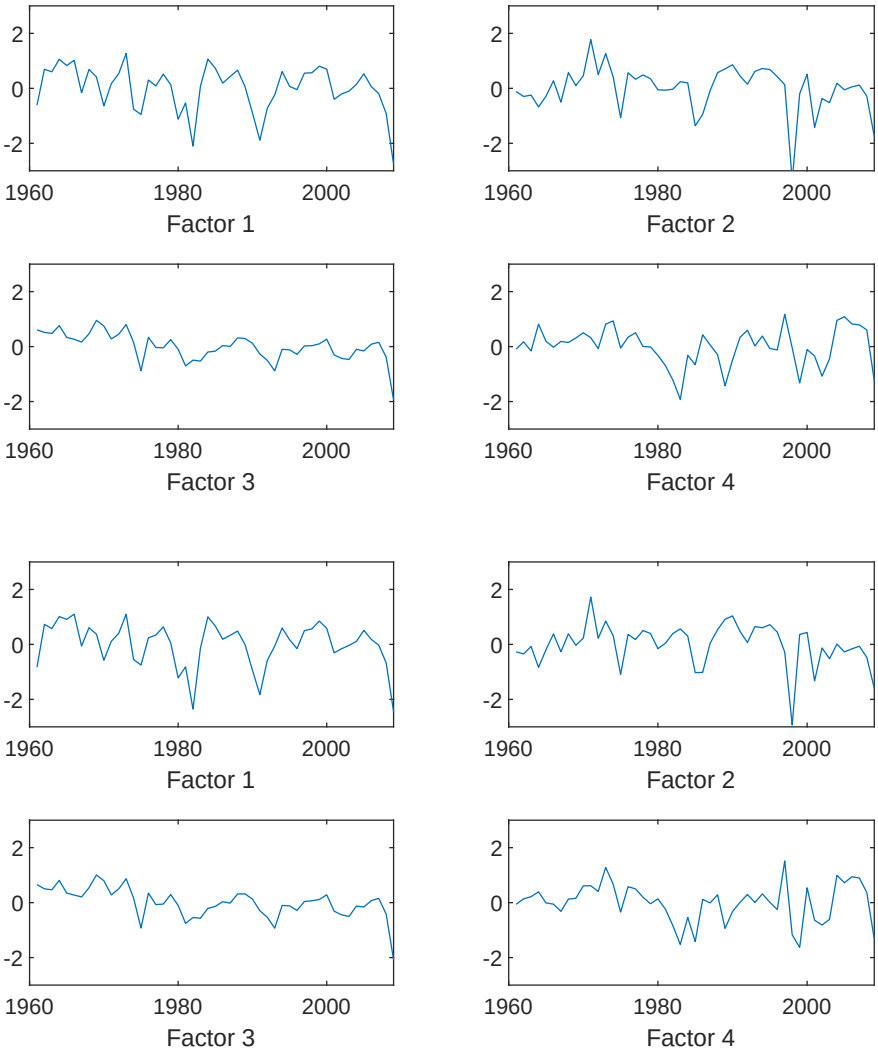


Figure 20 shows the loadings patterns identified by the unconstrained rotation approach. For the first mode, shown in the left panel, there are 5, 5, 22 and 12 non-zero loadings, respectively, on the four factors. There are two negative loadings, one on the first and one on the fourth factor, both close to zero. For the second mode, shown in the 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. We conclude that there are three pervasive factors, whereas the fourth factor is a local or weak one in each mode.

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. Clus-

Figure 21: GDP growth: Mean factors, obtained from the unconstrained rotation approach.



tering 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.3, 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.

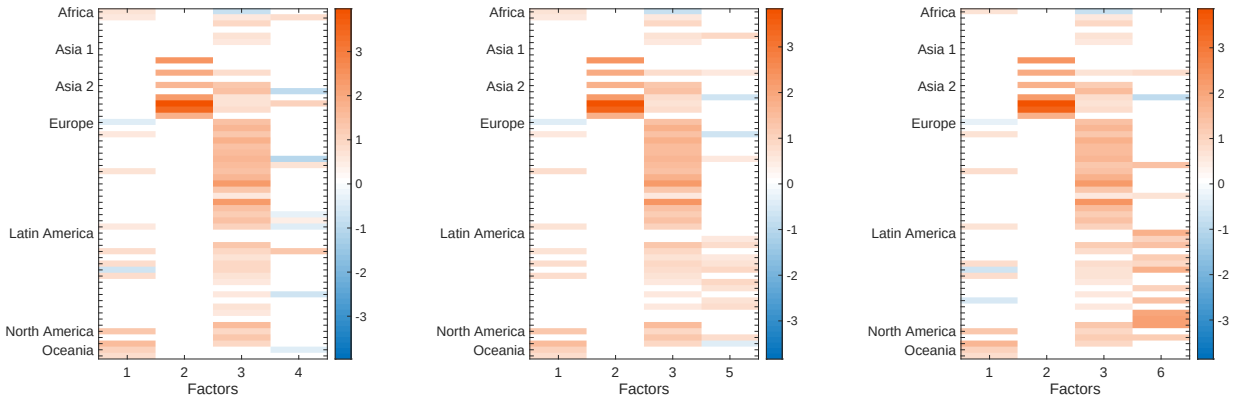
Sorting out the draws, only three factor combinations are visited. Table 6 reports again that almost all of the 65,000 draws can be assigned to either of the three modes. We identify 3 pervasive factors and 3 weaker ones. Figure 22 displays the loadings patterns. The number of non-zero loadings on the 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 factor, there are between 11 and 16 non-zero loadings, and the location of these vary substantially across factors.

Table 6: GDP growth: 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 22: GDP growth, sparse permutation: Mean factor loadings, averaged over draws with a non-zero probability larger than .5.



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.

Figure 23: GDP growth: Factor correlations, across factor combinations

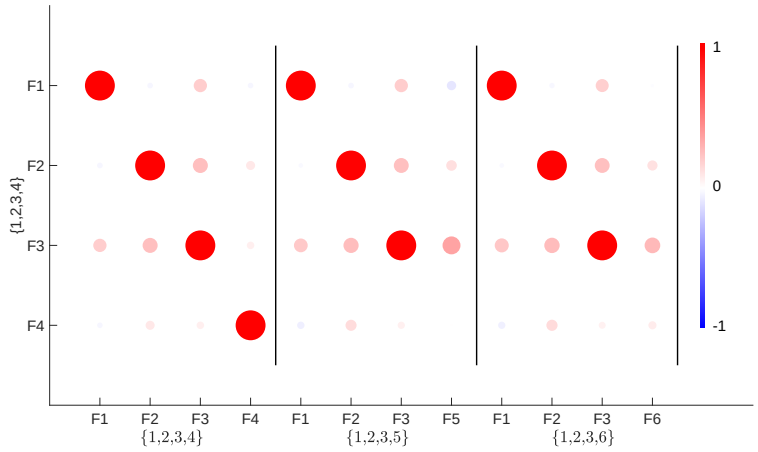
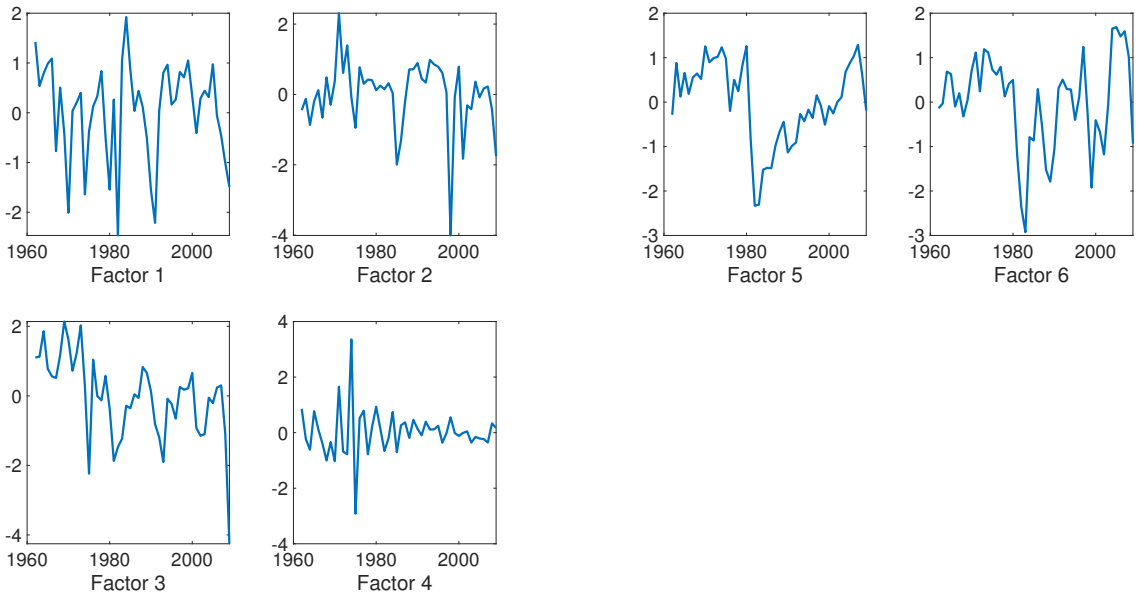


Figure 24: GDP growth: Mean factors, obtained from the sparse permutation sampler.



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 signs 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.

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.

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

A 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 $E(\beta_{ij}) = \rho_j b$ across variables. The marginal prior becomes

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

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

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

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 (21) are given by:

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

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

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

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

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

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

The hyperparameters τ_j and ρ_j are sampled from, respectively, an inverse Gamma $\pi(\tau_j|\cdot) \sim IG(g_j, G_j)$ and a Beta distribution, $\pi(\rho_j|\cdot) \sim B(r_{1j}, r_{2j})$, with

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

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

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

B Post-processing

B.1 Givens decomposition of an orthogonal matrix

The Givens decomposition of an orthogonal matrix H with $H'H = HH' = I_K$ can be performed as follows. First, it is necessary to find all pairs of axes k_1, $k_2 \in \{1, \ldots, K\}$ with $k_1 \neq k_2$. Note that there are $P = \binom{K}{2}$ such pairs, numbered $p = \{1, \ldots, P\}$ in the following. Then the following steps must be applied, 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
$$\gamma_p = \arctan2(g_{p,2,1}, g_{p,1,1}).$$

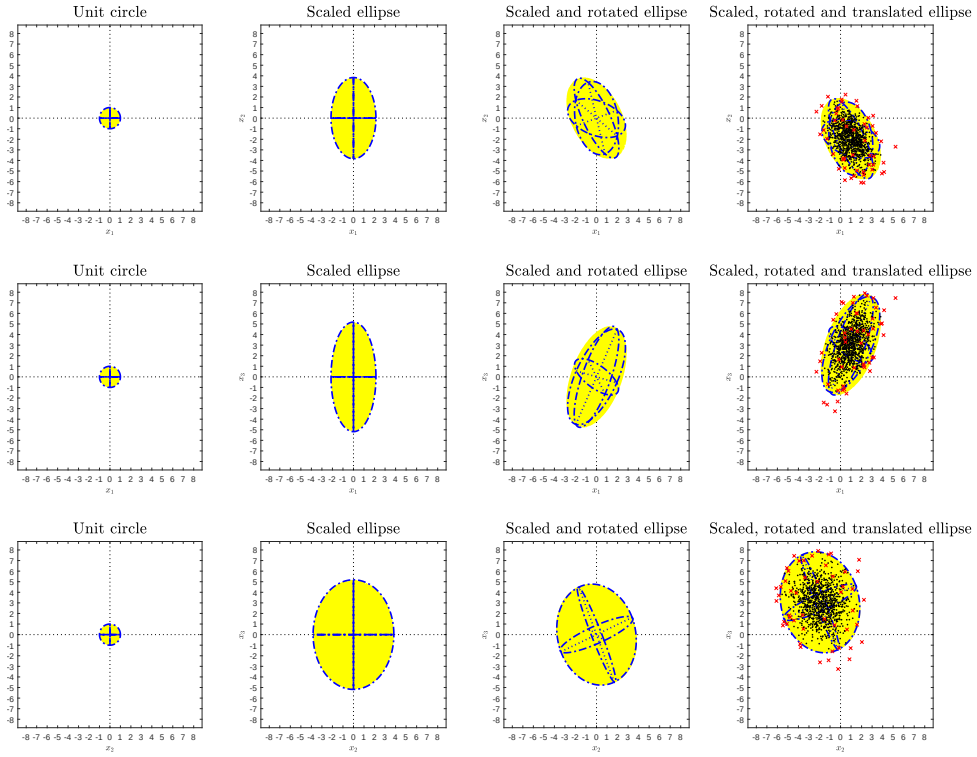
3. Replace the ${k_1}^{\text{th}}$ and ${k_2}^{\text{th}}$ row of matrix H, denoted as the submatrix $H_{\{k_1,k_2\},\cdot}$, by its rotated version $G_p H_{\{k_1,k_2\},\cdot}$.

4. If $p < P$, increment p and proceed with step 1, otherwise the decomposition is complete, in which case $H = I_K$ must hold.

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

Figure 25 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 actually a unit ball when all three dimensions are considered. The three ellipses in the second panels of each row represent the views on the expanded ellipsoid from three different angles. In the third panels of each row, the rotation has been applied, and the original

Figure 25: 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 shows axes 1 and 3, and bottom row shows 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.



axes are indicated within the resulting rotated ellipsoids shown. Again, the last panel of each row shows the data points inside and outside the ellipsoid as black dots and red marks. Note that the red marks apparently within the ellipsoid are in fact located in front of it or behind it. The share of points inside the ellipsoid is again guaranteed to be exactly $1 - \alpha$.

B.3 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

$$P(z_k^{(m)} = g) = \eta_g, \ g = 1, \ldots, G \tag{21}$$
$$\eta = (\eta_1, \ldots, \eta_G) \sim D(e_0, \ldots, e_0) \text{ with } e_0 = (G-1)/2 \tag{22}$$
$$\pi(f_k^{(m)} | z_k^{(m)} = g) \sim N\left(\mathrm{f}_g, \mathrm{F}_g\right), \text{ where } \mathrm{F}_g = \mathrm{diag}(\mathrm{F}_{g1}, \ldots, \mathrm{F}_{gT})$$
$$\pi(\mathrm{f}_g) \sim N\left(0_{T\times 1}, I_T\right), \ \mathrm{F}_{gt} \sim IG(\mathrm{s}_0, \mathrm{S}_0)$$

The prior for the mixture indicator (21)-(22) 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

$$P\left(z_k^{(m)} = g | f_k^{(m)}, \mathrm{f}_g, \mathrm{F}_g, \eta\right) \propto |\mathrm{F}_g|^{-1/2} \exp\left\{-0.5 \sum_{t=1}^{T} \frac{(\mathrm{sad}(f_{kt}^{(m)}) - \mathrm{f}_{gt})^2}{\mathrm{F}_{gt}}\right\} \eta_g \tag{23}$$

   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 model estimation.

   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 (23).

2. Update the cluster association probabilities: $\pi(\eta|\boldsymbol{z}) \sim D(e_1, \ldots, e_G)$ 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 $g = 1, \ldots, G$: $\pi\left(\mathrm{f}_g | \boldsymbol{z}, \boldsymbol{f}\right) \sim N\left(\bar{\mathrm{f}}_g, \bar{\mathrm{F}}_g\right)$, with moments

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

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

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

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.

C Additional material for the simulation study

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 $\mathbf{D} \in \mathcal{O}(K)$ 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 likewise 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 26, where $\mathbf{D}_2^{(2)'} = \mathbf{D}_1^{(2)}$.

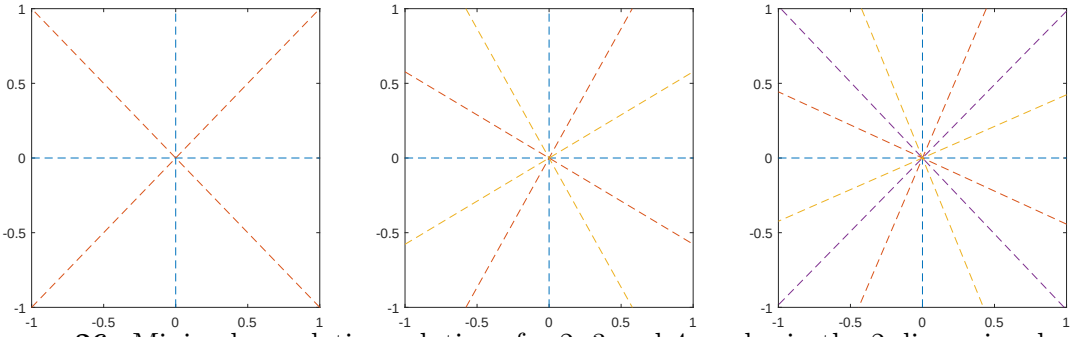


Figure 26: Minimal correlation solutions for 2, 3 and 4 modes in the 2-dimensional case.

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 27 shows three solutions for two modes in the 3-dimensional case.

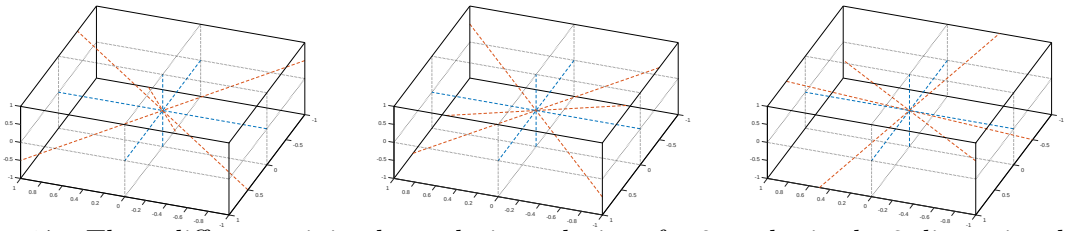
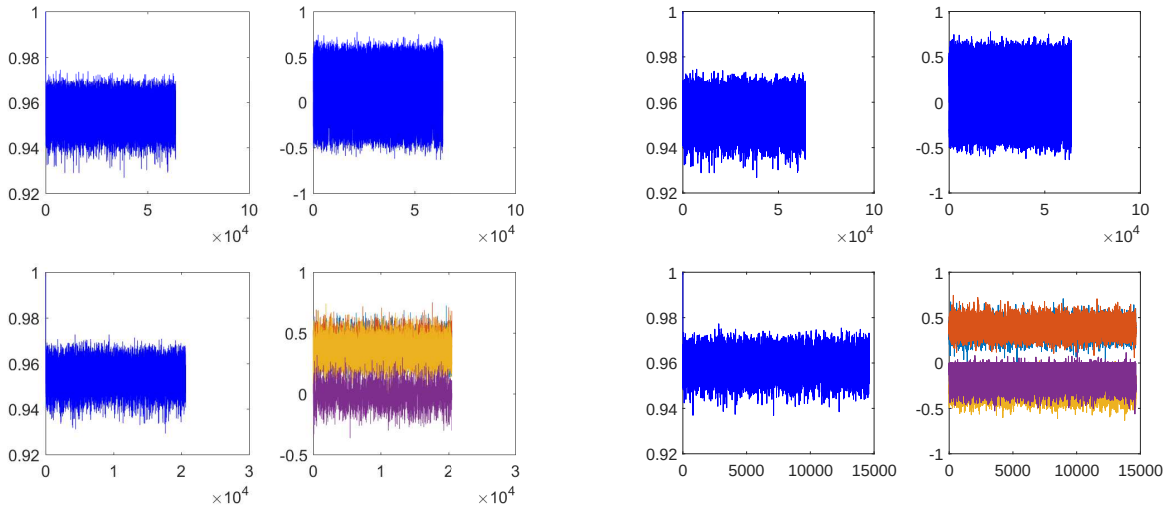


Figure 27: Three different minimal correlation solutions for 2 modes in the 3-dimensional case.

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 $\mathbf{D}_1^{(2)} = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}$ 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 $\mathbf{D}_1^{(4)} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \end{pmatrix}$, $\mathbf{D}_2^{(4)} = \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} \end{pmatrix}$, $\mathbf{D}_3^{(4)} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \end{pmatrix}$, 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 $\mathbf{D}_3^{(4)\prime} = \mathbf{D}_2^{(4)}$.

C.2 Additional figures from the simulation study

Figure 28: Posterior draws, unsorted and sorted, $K = 4$, two pervasive factors, scenario K4m2_2pf. 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

(b) Mode 2