Kiel

Working Papers

Kiel Institute for the World Economy

Bayesian Analysis of Dynamic Factor Models: An Ex-Post Approach towards the Rotation Problem

Christian Aßmann, Jens Boysen-Hogrefe and Markus Pape

No. 1902 | February 2014

Web: www.ifw-kiel.de

Kiel Institute for the World Economy, Hindenburgufer 66, 24105 Kiel, Germany

Kiel Working Paper No. 1902 | February 2014

Bayesian Analysis of Dynamic Factor Models: An Ex-Post Approach towards the Rotation Problem*

Christian Aßmann, Jens Boysen-Hogrefe and Markus Pape

Abstract:

Due to their indeterminacies, static and dynamic factor models require identifying assumptions to guarantee uniqueness of the parameter estimates. The indeterminacy of the parameter estimates with respect to orthogonal transformations is known as the rotation problem. The typical strategy in Bayesian factor analysis to solve the rotation problem is to introduce ex-ante constraints on certain model parameters via degenerate and truncated prior distributions. This strategy, however, results in posterior distributions whose shapes depend on the ordering of the variables in the data set. We propose an alternative approach where the rotation problem is solved ex-post using Procrustean postprocessing. The resulting order invariance of the posterior estimates is illustrated in a simulation study and an empirical application using a well-known data set containing 120 macroeconomic time series. Favorable properties of the ex-post approach with respect to convergence, statistical and numerical accuracy are revealed..

Keywords: Bayesian Estimation; Factor Models; Multimodality; Rotation Problem; Ordering Problem; Orthogonal Transformation.

JEL classification: C11; C31; C38; C51; C52.

Christian Aßmann
University of Bamberg
96045 Bamberg, Germany
Telephone: +49 951 863 3927
E-mail: christian.assmann@uni-bamberg.de

Jens Boysen-Hogrefe
Kiel Institute for the World Economy
24105 Kiel, Germany
Telephone: +49 431 8814 210
E-mail: jens.hogrefe@ifw-kiel.de

Markus Pape
University of Cologne
Universitätsstr. 22a
50937 Köln
Telephone: +49 221 470 2283
E-mail: markus.pape(@statistik.uni-koeln.de

1 Introduction

A latent factor model describes the influence of unobservable factors on observable data through factor loadings. Identifying assumptions are required to obtain unique estimates for the model quantities of interest, such as factors and loadings. The seminal paper of Anderson and Rubin (1956) deals with the question of model identification and shows that after restricting the covariance of factor innovations to unity the model is still invariant under orthogonal transformations of loadings and factors. Anderson and Rubin (1956) call this the *rotation problem* and provide different solutions, one of which is to constrain the loadings matrix to a lower triangular matrix.

Following the setup of Anderson and Rubin (1956), Geweke and Zhou (1996) discuss the Bayesian analysis of a factor model. To deal with the *rotation problem*, they suggest an identification scheme that has been widely used in many applications, sometimes in the slightly modified form of Aguilar and West (2000). This scheme constrains the loadings matrix to a positive lower triangular matrix. We will refer to this as positive lower triangular (PLT) identification scheme and to the accordingly constrained estimation approach as the PLT approach. Bai and Wang (2012) show that the PLT scheme solves the *rotation problem* also for the dynamic model. The PLT scheme guarantees a unique mode of the likelihood underlying the posterior distribution. It does not, however, guarantee the non-existence of local modes. The impact of constraints on the shape of the likelihood is discussed e.g. by Loken (2005). The constraints influence the shape of the likelihood and thus the shape of the posterior distribution. This is problematic since local modes can negatively affect the convergence behavior of Markov Chain Monte Carlo (MCMC) sampling schemes used for estimation purposes, see e.g. Celeux et al. (2000). As the constraints are imposed on particular elements of the loadings matrix, inference results depend on the ordering of the data. This is likewise observed by Carvalho et al. (2008). They call the variables whose loadings are constrained for identification purposes *factor founders* and develop an evolutionary search algorithm to choose the most appropriate subset of variables as *factor founders*. Similarly, Frühwirth-Schnatter and Lopes (2010) suggest a flexible approach that imposes a *generalized lower triangular* structure on the loadings matrix. Altogether, use of ex-ante identification via constraints on the parameter space may influence inference results with respect to the model parameters and functions of these parameters.[1]

The use of parameter constraints for identification and their consequences on inference are also discussed in the econometric literature for finite mixture models. Similar to factor models, finite mixture models are typically not identified, as labels of the mixture components can be changed by permutation. Thus, given symmetric priors, the posterior distribution has multiple symmetric modes. Identification can be achieved by fixing the ordering of the labels with respect to at least one of the parameters that are subject to label switching. However, if this identifying assumption is

[1] Accordingly, Lopes and West (2004) find that model selection criteria used to choose the number of factors are influenced by the way the variables are ordered and thus by the position of the restrictions on the parameter space.

introduced by prior distributions, the choice of the constraint may have a substantial impact on the shape of the posterior distribution and estimates derived therefrom, see Stephens (2000). Moreover, the posterior distribution may have multiple local modes, which has severe consequences for the mixing behavior of the Gibbs sampler. To cure this problem Celeux (1998) and Stephens (2000) suggest to achieve identification via post processing the output of the unconstrained sampler using relabeling algorithms.

Following the literature on finite mixture models, we propose an ex-post approach to fix the *rotation problem* that is suitable for the Bayesian analysis of both static and dynamic factor models. Ex-post identification, also recommended by Chan et al. (2013), can be framed as a decision theoretic approach and provides a unique posterior estimator, compare Celeux (1998), Celeux et al. (2000), and Stephens (2000). The suggested approach does not constrain the parameter space to solve the *rotation problem*, but obtains model identification by means of re-transformations of the output of a Gibbs sampler using orthogonal matrices. These orthogonal matrices are determined using a loss function adequately defined for the static and dynamic factor model. The minimization of the loss function in static factor models is based on the orthogonal Procrustes transformation proposed by Kristof (1964) and Schönemann (1966). Additionally, we use a weighting scheme as discussed by Lissitz et al. (1976), hence we refer it as the weighted orthogonal Procrustes (WOP) approach. For the dynamic factor model we use a parametrization of the orthogonal matrix allowing for numerical optimization of the defined loss function.[2] Inference results from the post-processed Gibbs output obtained under different orderings of the variables differ only by a single orthogonal transformation. Similarly, rotations of the solution with respect to criteria like Varimax or Quartimax can be applied to facilitate interpretability of the results. Our approach is thus purely exploratory.

To illustrate the properties of our ex-post approach towards the *rotation problem*, we provide a simulation study with static and dynamic factor models. We compare our inference results from the WOP ex-post approach with those from the PLT ex-ante approach by Geweke and Zhou (1996). We check both corresponding samplers for their convergence properties, as well as statistical and numerical accuracy. Convergence is generally obtained faster for the WOP approach. Statistical accuracy is similar to that of the PLT approach for parameters invariant under orthogonal transformations and if the PLT approach does not produce pathological posterior distributions. Such pathological cases do not occur under the WOP approach, which also shows much higher numerical accuracy than the PLT approach.

In an empirical application, we analyze the panel of 120 macroeconomic time series from Bernanke et al. (2005) using both the PLT and the WOP approach. As a first exercise, we choose series as fac-

[2]In an approach for sparse factor models, Kaufmann and Schumacher (2013) perform temporary orthogonal transformations of the model parameters to satisfy an alternative identification scheme suggested by Anderson and Rubin (1956), such that the outer product of the loadings matrix with itself is diagonal. Under this identification, the latent factors are sampled and afterwards transformed back into the original parametrization. This approach works well for sparse factor models, but seems to be inappropriate for the exactly identified exploratory factor models discussed here.

tor founders that are particularly fit for this purpose and estimate the model repeatedly. Afterwards, we perform repeated estimations of the model under randomly chosen orderings of the series. The WOP approach is found to be numerically more stable than the PLT approach in both exercises.

The paper proceeds as follows. Section 2 provides the dynamic factor model and discusses briefly the identification of the model. Section 3 introduces the novel identification approach for static and dynamic factor models. Section 4 illustrates the differences between the WOP and the PLT approach by means of a simple example. Section 5 presents a simulation study that compares both approaches. Section 6 provides an empirical illustration using the data set of Bernanke et al. (2005). Section 7 concludes.

2 Model setup and identification problem

In a dynamic factor model the comovements in a data panel with N variables and time dimension T are represented by K factors that relate to the data via loadings. The dynamic factor model takes the form

$$y_t \quad = \quad \Lambda_0 f_t + \Lambda_1 f_{t-1} + ... + \Lambda_S f_{t-S} + e_t, \quad t = 1, \ldots, T, \tag{1}$$

where y_t is an $N \times 1$ demeaned and stationary vector of observed data, f_t is a $K \times 1$ vector of K latent factors, Λ_s, $s = 0, \ldots, S$ representing $N \times K$ matrices of loadings, and e_t denotes a $N \times 1$ vector of errors with $e_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \Sigma)$ and $\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_N^2)$.[3] Further, the K factors follow a vector autoregressive process of order P given as

$$f_t \quad = \quad \Phi_1 f_{t-1} + \Phi_2 f_{t-2} + ... + \Phi_P f_{t-P} + \epsilon_t, \tag{2}$$

where Φ_p, $p = 1, \ldots, P$ are $K \times K$ persistence matrices, and $\epsilon_t \overset{\text{iid}}{\sim} \mathcal{N}(0, I_K)$. Setting the covariance of ϵ_t to the identity matrix solves the identification problem up to the *rotation problem* as discussed by Anderson and Rubin (1956) for static factor models. Bai and Wang (2012) show that this also holds for the dynamic factor model described here. Further, the value of S has no impact of the identification scheme. To illustrate the point we consider the likelihood with

$$\Theta = (\text{vec}(\Lambda_0), \ldots, \text{vec}(\Lambda_S), \text{vec}(\Phi_1), \ldots, \text{vec}(\Phi_P), \text{diag}(\Sigma)) \tag{3}$$

[3]Note that the model could be further enhanced by an autoregressive process of order Q for errors e_t as discussed by Kaufmann and Schumacher (2013).

summarizing all model parameters, $Y = (y_1, \ldots, y_T)$ and $f_0 = \ldots = f_{-\max\{S-1,P-1\}} = 0$ given as

$$\mathcal{L}(Y|\Theta) = \int_{f_T} \cdots \int_{f_1} \prod_{t=1}^{T} p(y_t|\Theta, f_t, \ldots, f_{t-S}) p(f_t|\Theta, f_{t-1}, \ldots, f_{t-P}) df_1 \ldots df_T \tag{4}$$

$$= \int_{f_T} \cdots \int_{f_1} (2\pi)^{-\frac{TN}{2}} |\Sigma|^{-\frac{T}{2}} \exp\left\{ -\frac{1}{2} \sum_{t=1}^{T} \left((y_t - \sum_{s=0}^{S} \Lambda_s f_{t-s})' \Sigma^{-1} (y_t - \sum_{s=0}^{S} \Lambda_s f_{t-s}) \right) \right\}$$

$$(2\pi)^{-\frac{TK}{2}} |\Omega_\epsilon|^{-\frac{T}{2}} \exp\left\{ -\frac{1}{2} \sum_{t=1}^{T} (f_t - \sum_{p=1}^{P} \Phi_p f_{t-p})' (f_t - \sum_{p=1}^{P} \Phi_p f_{t-p}) \right\} df_1 \ldots df_T.$$

The likelihood is invariant under the following transformation.[4] Define for any orthogonal $K \times K$ matrix D the transformation

$$\begin{aligned} \tilde{\Theta} &= (\text{vec}(\tilde{\Lambda}_0), \ldots, \text{vec}(\tilde{\Lambda}_S), \text{vec}(\tilde{\Phi}_1), \ldots, \text{vec}(\tilde{\Phi}_P), \text{diag}(\tilde{\Sigma})) \\ &= (\text{vec}(\Lambda_0 D), \ldots, \text{vec}(\Lambda_S D), \text{vec}(D^{-1}\Phi_1 D), \ldots, \text{vec}(D^{-1}\Phi_P D), \text{diag}(\Sigma)) = H(D)\Theta, \end{aligned} \tag{5}$$

with

$$H(D) = \begin{pmatrix} (D' \otimes I_{N(S+1)}) & 0 & 0 \\ 0 & I_P \otimes (D' \otimes D^{-1}) & 0 \\ 0 & 0 & I_N \end{pmatrix}, \tag{6}$$

where $|\det(H^{-1}(D))| = |\det(D)^{-(N(S+1)+K+1)}| = 1$. For completion, considering $\tilde{f}_t = D^{-1} f_t$, $t = 1, \ldots, T$, and $d\tilde{f}_t = |\det(D)| df_t = df_t$ and taking into account that the transformation has no impact on the range of parameters yields $\mathcal{L}(Y|\Theta) = \mathcal{L}(Y|\tilde{\Theta})$, i.e. the likelihood remains the same under the transformation in Equation (5). We will refer to this invariance of the likelihood as the *rotation problem*.

The invariance of the likelihood kernel is transferred to the posterior distribution and thus posterior estimators, when the chosen a priori distribution are as well invariant under the transformation described in Equation (5). As the *rotation problem* does not involve Σ, we choose the commonly used conjugate prior given as

$$\pi(\Sigma) = \prod_{i=1}^{N} \frac{\beta_{0i}^{\alpha_{0i}}}{\Gamma(\alpha_{0i})} \sigma_i^{-2(\alpha_{0i}+1)} \exp\left\{ -\frac{\beta_{0i}}{\sigma_i^2} \right\}. \tag{7}$$

[4] The static case arising for $S = P = 0$ corresponds to the closed form likelihood given as

$$(2\pi)^{-\frac{TN}{2}} |\Lambda_0 \Lambda_0' + \Sigma|^{-\frac{T}{2}} \exp\left\{ -\frac{1}{2} \sum_{t=1}^{T} y_t' (\Lambda_0 \Lambda_0' + \Sigma)^{-1} y_t \right\}.$$

With regard to invariance of the likelihood, the same caveats as in the dynamic case apply.

The priors for Λ_s, $s = 0, \ldots, S$ and Φ_p, $p = 1, \ldots, P$ are chosen as

$$\pi(\Phi_1, \ldots, \Phi_P) \propto c, \quad c > 0, \tag{8}$$

and

$$\pi(\Lambda_0, \ldots, \Lambda_S) = \prod_{s=0}^{S} (2\pi)^{-\frac{KN}{2}} |\Omega_{\Lambda_s}|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} (\mathrm{vec}(\Lambda_s) - \mu_{\Lambda_s})' \Omega_{\Lambda_s}^{-1} (\mathrm{vec}(\Lambda_s) - \mu_{\Lambda_s}) \right\} \tag{9}$$

respectively. The normal prior for $\{\Lambda_s\}_{s=0}^{S}$ is in line with the specification of Bai and Wang (2012), but does not impose constraints. The constant prior for $\{\Phi_p\}_{p=1}^{P}$ likewise follows the specification of Bai and Wang (2012), or, more generally, the specifications for Bayesian vector autoregressive modeling by Ni and Sun (2005). Additional stationarity constraints can be imposed by demanding that the eigenvalues of the companion matrix of $\{\Phi_p\}_{p=1}^{P}$ are all less than 1 in absolute value, see e.g. Hamilton (1994, ch.10). Note that the eigenvalues of the companion matrix are unaffected by the transformation in Equation (5).[5] We require that all mean vectors are set to zero, i.e. $\mu_{\Lambda_s} = 0$, $s = 0, \ldots, S$ and $\Omega_{\Lambda_s} = \Upsilon_s \otimes I_K$, $s = 0, \ldots, S$ with Υ_s a positive diagonal $N \times N$ matrix. The so far stated posterior distribution

$$p(\Theta|Y) \propto \mathcal{L}(Y|\Theta)\pi(\Sigma)\pi(\Phi_1, \ldots, \Phi_P)\pi(\Lambda_0, \ldots, \Lambda_S) \tag{10}$$

is then invariant under the transformation in Equation (5).

The model setup is directly accessible in state-space form. This allows for sampling using the methodology presented in Carter and Kohn (1994). Appendix A gives a detailed description of the corresponding Gibbs sampler, which we will call the *unconstrained Gibbs sampler* in the following, because it does not impose any constraints on the loadings matrix in order to solve the *rotation problem*. The unconstrained Gibbs sampler can access all orthogonal transformations of the parameter space described by Equation (5).[6]

Conversely, there exist several constrained sampling approaches, which solve the *rotation problem* by means of imposing constraints on the loadings matrix. In a model with K factors, the scheme of Geweke and Zhou (1996) constrains the first K rows and columns to form a positive lower triangular matrix, hence we call it the positive lower triangular (PLT) identification scheme. This is obtained by using a Dirac Delta prior at zero for all elements above the diagonal. The positivity constraints are imposed by either accepting only such draws with exclusively positive diagonal elements or,

[5]The constant prior for $\{\Phi_p\}_{p=1}^{P}$ can also be replaced by normal priors with zero mean and a covariance matrix that equals the unity matrix times a constant, since this distribution is also not affected by the transformation in Equation (5).

[6]Analogous to the random relabeling approach of Frühwirth-Schnatter (2006), mixing can be sped up by adding random orthogonal transformations of the parameter space in each draw.

alternatively, by using prior distributions truncated below at zero for the diagonal elements.[7] In Chapter 4, some properties of the PLT approach are discussed in more detail.

Leaving the model unidentified with respect to the corresponding $K(K-1)/2$ parameters and the signs of the factors and the corresponding column vectors of the loadings matrices, can also be understood as a parameter expansion of the model, introducing as additional parameter the orthogonal matrix D that varies in each iteration of the sampler. As D is only identified conditional on a subset of the model parameters, it qualifies as a *working parameter* in the sense of Meng and van Dyk (1999) and van Dyk and Meng (2001), being redundant, but useful in the estimation process. Gelman et al. (2008) argue in favor of expanding a model by redundant parameters in order to improve a sampler's mixing behavior. A similar parameter expansion approach has been suggested for static factor models by Chan et al. (2013). Whereas usually, parameter expansion approaches involve deliberately sampling the additional parameters, for the unconstrained sampler discussed here, leaving the model partially unidentified automatically results in the evolvement of an additional parameter governing the state of the sampler with respect to the directed parameters. This state parameter D can be expressed in terms of an orthogonal matrix that affects the directed parameters via $H(D)$. The task of the ex-post identification approach discussed in the following section is therefore to remove the effect of the state parameter D in such a way that all the draws from the sampler can be assumed to have originated from the sampler in a unique state $\overline{D}$.

3 An ex-post approach towards the rotation problem

The *rotation problem* is solved when the uniqueness of the estimator derived from the posterior distribution is ensured. The uniqueness is ensured when the invariance of the posterior distribution under the transformation in Equation (5) is inhibited. This is possible via ex-ante restrictions on the parameter space hindering the mapping of any points within the admissible parameter space by orthogonal matrices. While ex-ante restrictions are routinely applied in many econometric frameworks, ex-post identification is prominent for finite mixture models, see Celeux et al. (2000), Stephens (2000), Frühwirth-Schnatter (2001, 2006) and Grün and Leisch (2009).[8] To address the *rotation problem* hin-

[7]In a similar approach, Aguilar and West (2000) use a Dirac Delta prior at one for the diagonal elements, which also solves the scaling indeterminacy, so the variances of the factor innovations can be freely estimated. Yet another approach follows the scheme by Jöreskog (1979), where the top $K \times K$ section of the loadings matrix is constrained to the identity matrix. In turn, all elements of the covariance matrix of the factor innovations can be freely estimated. This approach is discussed in more detail in Bai and Wang (2012).

[8]In the context of finite mixture models ex-post identification is used as the posterior is invariant under permutation of mixing components, i.e. when label switching (Redner and Walker, 1984) occurs in the output of an unconstrained sampler. Richardson and Green (1997) advise to use different identifiability constraints when postprocessing the MCMC output. For finite mixtures Stephens (2000) and Frühwirth-Schnatter (2001) propose the use of relabeling algorithms that screen the output of the unconstrained sampler and sort the labels to minimize some divergence measures, e.g. Kullback-Leibler distances. The main idea behind the relabeling approach in finite mixtures is that the output of the unconstrained sampler in fact stems from a mixture distribution. The mixing is discrete and occurs via

dering inference, we propose an ex-post approach for Bayesian analysis of dynamic factor models, which can also be motivated as a decision-theoretic approach, see e.g. Stephens (2000).

The suggested ex-post identification approach is based on the observation that the unconstrained sampler provides a realized sample $\{\Theta^{(r)}\}_{r=1}^{R}$ from the posterior distribution which can equivalently be interpreted as a sample taking the form $\{H(D^{(r)})\Theta^{(r)}\}_{r=1}^{R}$, i.e. a sample given as a transformation of the realized sample by an arbitrary sequence of orthogonal matrices $\{D^{(r)}\}_{r=1}^{R}$. All samples taking the form $\{H(D^{(r)})\Theta^{(r)}\}_{r=1}^{R}$ are assigned the same posterior probability. Due to this indeterminacy, we refer to the unconstrained sample as *orthogonally mixing*. Each of the interpretations would result in a different estimate of Θ. To distinguish between the different interpretations of the form $\{H(D^{(r)})\Theta^{(r)}\}_{r=1}^{R}$, and correspondingly ensure uniqueness of the estimate, we advocate the use of a loss function approach allowing for discrimination of the loss invoked under different orthogonal transformations of the realized sample with respect to parameter estimation.

If for any point the minimal loss can be uniquely determined, *orthogonal mixing* is immaterial for parameter estimation, the *rotation problem* is fixed, and the corresponding estimator is uniquely identified. Following Jasra et al. (2005), we define a loss function as a mapping of the set of possible estimators $\{\Theta^*\}$ and each of the parameter values Θ within the parameter space on the real line, i.e. $L : \{\Theta^*\} \times \Theta \to [0, \infty)$ such that

$$L(\Theta^*, \Theta) = \min_{D}\{L_D(\Theta^*, \Theta(D))\}, \quad \text{s.t. } D'D = I, \tag{11}$$

with $L_D(\Theta^*, \Theta(D))$ denoting for given Θ^* the loss invoked for any transformation of Θ as described in Equation (5). The optimal estimator, corresponding to the optimal action in a decision-theoretic framework, is then defined as

$$\Theta^* = \arg \min_{\Theta^* \in \{\Theta^*\}} \int_{\Theta} L(\Theta^*, \Theta) p(\Theta|Y) d\Theta. \tag{12}$$

For computational reasons, an Monte Carlo (MC) approximation is used for the integral involved in Equation (12), thus we obtain

$$\Theta^* = \arg \min_{\Theta^* \in \{\Theta^*\}} \frac{1}{R} \sum_{r=1}^{R} L(\Theta^*, \Theta^{(r)}), \tag{13}$$

where $\Theta^{(r)}$, $r = 1, \ldots, R$ denotes a sample from the unconstrained posterior distribution. Based on the defined loss function the following algorithm is implemented. For a given initialization of the estimator Θ_0^*, solve for each point Equation (11). Then for given a sequence $\{D^{(r)}\}_{r=1}^{R}$ find a corresponding estimator Θ^* and iterate until convergence. The choice of the loss function is restricted with

permutations of the labels. The relabeling algorithm fixes the invariance of the likelihood with respect to a specific permutation based on a decision criterion and reverses thus the mixing.

regard to solvability and uniqueness of the solution to the particular minimization problem. Since the relative entropy, also known as the Kullback-Leibler distance, between the posterior distribution and the distribution of interest, vanishes as the number of draws from the posterior distribution goes to infinity, see Clarke et al. (1990), this measure is obviously an appropriate choice for the loss function. Moreover, a quadratic loss function produces estimators that are first-order equivalent to the estimators obtained under a Kullback-Leibler loss function in many cases, see Cheng et al. (1999). Thus we suggest the following quadratic loss function for the considered dynamic factor model denoted as

$$\begin{aligned} L_D(\Theta^*, \Theta^{(r)}(D^{(r)})) &= L_D(\Theta^*, H(D^{(r)})\Theta^{(r)}) = \mathrm{tr}\left[(H(D^{(r)})\Theta^{(r)} - \Theta^*)'(H(D^{(r)})\Theta^{(r)} - \Theta^*)\right] \quad (14) \\ &= L_{D,1}(\Theta^*, \Theta^{(r)}(D^{(r)})) + L_{D,2}(\Theta^*, \Theta^{(r)}(D^{(r)})) + L_{D,3}(\Theta^*, \Theta^{(r)}(D^{(r)})), \end{aligned}$$

with

$$\begin{aligned} L_{D,1}(\Theta^*, \Theta^{(r)}(D^{(r)})) &= \sum_{s=0}^{S} \mathrm{tr}\left[(\Lambda_s^{(r)}(D^{(r)}) - \Lambda_s^*)'(\Lambda_s^{(r)}(D^{(r)}) - \Lambda_s^*)\right] \quad (15) \\ &= \mathrm{tr}\left[(\bar{\Lambda}^{(r)}D^{(r)} - \bar{\Lambda}^*)'(\bar{\Lambda}^{(r)}D^{(r)} - \bar{\Lambda}^*)\right], \\ L_{D,2}(\Theta^*, \Theta^{(r)}(D^{(r)})) &= \sum_{p=1}^{P} \mathrm{tr}\left[(\Phi_p^{(r)}(D^{(r)}) - \Phi_p^*)'(\Phi_p^{(r)}(D^{(r)}) - \Phi_p^*)\right] \quad (16) \\ &= \sum_{p=1}^{P} \mathrm{tr}\left[(D^{(r)\prime}\Phi_p^{(r)}D^{(r)} - \Phi_p^*)'(D^{(r)\prime}\Phi_p^{(r)}D^{(r)} - \Phi_p^*)\right], \\ L_{D,3}(\Theta^*, \Theta^{(r)}(D^{(r)})) &= \mathrm{tr}[(\mathrm{diag}(\Sigma) - \mathrm{diag}(\Sigma^*))'(\mathrm{diag}(\Sigma) - \mathrm{diag}(\Sigma^*))], \quad (17) \end{aligned}$$

where $\bar{\Lambda}^{(r)} = (\Lambda_0^{(r)\prime}, \ldots, \Lambda_S^{(r)\prime})'$, $\tilde{\Phi} = (\Phi_1^{(r)\prime}, \ldots, \Phi_P^{(r)\prime})'$ and $\bar{\Lambda}^* = (\Lambda_0^{*\prime}, \ldots, \Lambda_S^{*\prime})'$, respectively. The dependence on the orthogonal matrix D is thus operationalized as $\Lambda_s(D) = \Lambda_s D$, $s = 0, \ldots, S$ and $\Phi_p(D) = D'\Phi_p D$, $p = 1, \ldots, P$. Using the MC version of the expected posterior loss results in the following minimization problem

$$\{\{D^{(r)}\}_{r=1}^R, \Theta^*\} = \arg\min \sum_{r=1}^{R} L_D(\Theta^*, \Theta^{(r)}(D^{(r)})) \quad \text{s.t.} \quad D^{(r)}D^{(r)\prime} = I, \; r = 1, \ldots, R. \quad (18)$$

Note that the standard static factor model is nested in the above for $S = P = 0$. For illustration, the solution of the optimization problem is discussed for the static as well as the dynamic factor model, as the solution principle for the static case applies for the dynamic case as well.

Static Factor Model A solution to the optimization problem stated in Equation (18) applied to the static factor model is obtained iteratively via a two-step optimization. The algorithm needs initialization with regard to $\Theta^* = \{\mathrm{vec}(\bar{\Lambda}^*), \mathrm{diag}(\Sigma^*)\}$, where we choose the last draw of the unconstrained sampler for convenience.

Step 1 For given Θ^* the following minimization problem for $D^{(r)}$ has to be solved for each $r = 1, \ldots, R$, i.e.

$$D^{(r)} = \arg\min L_{D,1}(\Theta^*, \Theta^{(r)}(D^{(r)})) \quad \text{s.t.} \quad D^{(r)\prime}D^{(r)} = I. \tag{19}$$

The solution of this orthogonal Procrustes (OP) problem is provided by Kristof (1964) and Schönemann (1966), see also Golub and van Loan (2013). It involves the following calculations:

1.1 Define $S_r = \bar{\Lambda}^{(r)\prime}\bar{\Lambda}^*$.

1.2 Do the singular value decomposition $S_r = U_r M_r V_r'$, where U_r and V_r denote the matrix of eigenvectors of $S_r S_r'$ and $S_r' S_r$, respectively, and M_r denotes a diagonal matrix of singular values, which are the square roots of the eigenvalues of $S_r S_r'$ and $S_r' S_r$. Note that the eigenvalues of $S_r S_r'$ and $S_r' S_r$ are identical.

1.3 Obtain the orthogonal transformation matrix $D^{(r)} = U_r V_r'$.

For further details on the derivation of this solution, see Schönemann (1966). Note that if the dispersion between the cross sections is rather large, the solution may be improved via consideration of weights, turning the problem to be solved into a weighted orthogonal Procrustes (WOP) problem, see e.g. Lissitz et al. (1976) and Koschat and Swayne (1991). Thus Step 1.1 above is altered into

1.1a Define $S_r = \bar{\Lambda}^{(r)\prime} W \bar{\Lambda}^*$,

where the weighting matrix W has to be diagonal with strictly positive diagonal elements and is initialized as the inverses of the estimated lengths of the loading vectors, i.e.

$$W = R\left(\sum_{r=1}^{R} \sqrt{(\bar{\Lambda}^{(r)}\bar{\Lambda}^{(r)\prime}) \odot I_{(S+1)N}}\right)^{-1}. \tag{20}$$

Consecutively, we use as weights a function of the number of factors and the determinants of the estimated covariance matrices, which are a measure invariant to orthogonal transformations, i.e. $W = \operatorname{diag}(w_1, \ldots, w_{(S+1)N})$, where

$$w_i = \det\left(\frac{1}{R}\sum_{r=1}^{R}(\bar{\lambda}_i^{(r)} - \lambda_i^*)(\bar{\lambda}_i^{(r)} - \lambda_i^*)'\right)^{-\frac{1}{K}}, \quad i = 1, \ldots, (S+1)N. \tag{21}$$

The weighting scheme scales the loadings in such a way that the estimated covariance matrix has determinant 1 for each variable.

Step 2 Choose $\bar{\Lambda}^*$ and Σ^* as

$$\bar{\Lambda}^* = \frac{1}{R}\sum_{r=1}^{R} \bar{\Lambda}^{(r)} D^{(r)} \quad \text{and} \quad \Sigma^* = \frac{1}{R}\sum_{r=1}^{R} \Sigma^{(r)}. \tag{22}$$

As Step 1 minimizes the (weighted) distance between the transformed observations and the given $\bar{\Lambda}^*$, it provides an unique orientation to each sampled $\bar{\Lambda}^{(r)}$. In Step 2, the estimator is determined based on an orientated sample. For arbitrary initial choices of $\bar{\Lambda}^*$ taken from the unconstrained sampler output, less than ten iterations usually suffice to achieve convergence to a fixed point $\bar{\Lambda}^*$. Convergence is assumed if the sum of squared deviations between two successive matrices $\bar{\Lambda}^*$ does not exceed a predefined threshold value, say 10^{-9}.

The following proposition summarizes the suggested ex-post approach for the static factor model.

Proposition 3.1. *The ex-post approach solves the rotation problem for the static factor model.*

Proof. The orthogonal matrix $D^{(r)}$ that minimizes the loss function in Equation (19) representing the orthogonal Procrustes problem is unique conditional on almost every $\Theta^{(r)}$ and Θ^*, where the elements in $\Theta^{(r)}$ are random variables following a nondegenerate probability distribution as implied by the chosen prior distributions. The availability of a unique solution to the orthogonal Procrustes problem providing a minimum is shown by Kristof (1964), Schönemann (1966) and Golub and van Loan (2013) and for the weighted orthogonal Procrustes problem by Lissitz et al. (1976). Following Golub and van Loan (2013) the minimization problem stated in Equation (19) is equivalent to the maximization of $\text{tr}(D^{(r)\prime}\bar{\Lambda}^{(r)\prime}\bar{\Lambda}^*)$, where the maximizing $D^{(r)}$ can be found by calculation of the singular value decomposition of $\bar{\Lambda}^{(r)\prime}\bar{\Lambda}^*$. If $U_r(\bar{\Lambda}^{(r)\prime}\bar{\Lambda}^*)V_r' = M_r = \text{diag}(m_r^{(1)}, \ldots, m_r^{(K)})$ is the singular value decomposition of this matrix and we define the orthogonal matrix $Z_r = V_r' D^{(r)\prime} U_r$, then

$$\text{tr}(D^{(r)\prime}\bar{\Lambda}^{(r)\prime}\bar{\Lambda}^*) = \text{tr}(D^{(r)\prime}U_r M_r V_r') = \text{tr}(Z_r M_r) = \sum_{k=1}^{K} z_r^{(kk)} m_r^{(k)} \leq \sum_{k=1}^{K} m_{(r)}^{k}.$$

The upper bound is then attained by setting $D^{(r)} = U_r V_r'$, which implies $Z_r = I_K$. Note that there exist points, however, where at least one singular value of $\bar{\Lambda}^{(r)\prime}\bar{\Lambda}^*$ is zero. In these cases, the left and right eigenvectors related to these singular values are not uniquely determined and thus no unique solution to the orthogonal Procrustes problem exists. However, these points occur with probability zero.

The unsolved *rotation problem* implies that within the parameter space pairs of points can be defined, with two points being pairwise orthogonal transformations of each other according to Equation (5). Denote such a pair as $\Theta^{(1)}$ and $\Theta^{(2)}$ with $\Theta^{(2)} = H(D_0)\Theta^{(1)}$, where D_0 is an orthogonal matrix. To show that the *rotation problem* is solved by the suggested ex-post approach, one has to show that no such pairs can be defined after postprocessing. After postprocessing, $\Theta^{(1)}$ and $\Theta^{(2)}$ take the form $H(D_1)\Theta^{(1)}$ and $H(D_2)\Theta^{(2)}$ respectively, where D_i, $i = 1, 2$ implies minimal loss with regard to Θ^*. Since D_1 and D_2 are uniquely defined as shown above and $H(D_2)\Theta^{(2)} = H(D_0 D_2)\Theta^{(1)}$ we have consequently $D_0 D_2 = D_1$, where we use the fact that the product of two orthogonal matrices is itself an orthogonal matrix, and orthogonal matrices commute. Assuming without loss of generality that $D_1 = I_K$, we have $D_2 = D_0'$. This implies that after postprocessing all points that can be represented as orthogonal transformations of $\Theta^{(1)}$ are collapsed into $\Theta^{(1)}$ as the point invoking minimal loss and thus enter the parameter estimation as $\Theta^{(1)}$. □

Next, we consider the case of the dynamic factor model. The corresponding ex-post approach is based on an extended loss function considering the dynamic factor structure as well.

Dynamic Factor Model The algorithm for the dynamic factor model differs with regard to Step 1 from the algorithm presented for the static factor model.

Step 1 For given Θ^* the following minimization problem for $D^{(r)}$ has to be solved for each $r = 1, \ldots, R$, i.e.

$$D^{(r)} = \arg\min(L_{D,1}(\Theta^*, \Theta^{(r)}(D^{(r)})) + L_{D,2}(\Theta^*, \Theta^{(r)}(D^{(r)}))) \quad \text{s.t.} \quad D^{(r)\prime} D^{(r)} = I. \tag{23}$$

The solution is based on numerical optimization using a parametrization of $D^{(r)}$ ensuring orthogonality. Since every orthogonal matrix D can be decomposed into a reflection matrix F with $\det(F) = \det(D) = \pm 1$ and a corresponding rotation matrix which can be factorized into $K^* = |\{(i,j) : i,j \in \{1,\ldots,K\}, j > i\}|$ Givens rotation matrices, we can parameterize any orthogonal matrix as

$$D = \begin{cases} D_+ = F_+ \prod_{(i,j):i,j\in\{1,\ldots,K\},j>i} G_{i,j,K} & \text{if } \det(D) = 1, \\ D_- = F_- \prod_{(i,j):i,j\in\{1,\ldots,K\},j>i} G_{i,j,K} & \text{if } \det(D) = -1, \end{cases} \tag{24}$$

where

$$F_+ = \begin{pmatrix} I_{K-1} & 0 \\ 0 & 1 \end{pmatrix}, \quad F_- = \begin{pmatrix} I_{K-1} & 0 \\ 0 & -1 \end{pmatrix}$$

and

$$G_{i,j,K} = \begin{pmatrix} g_{11} & \cdots & g_{1K} \\ \vdots & & \vdots \\ g_{K1} & \cdots & g_{KK} \end{pmatrix}, \text{ with } g_{rs} = \begin{cases} 1, & \text{for } i \neq r = s \neq j \\ \cos(\gamma_{(i,j)}), & \text{for } r = s = i \text{ and } r = s = j, \\ -\sin(\gamma_{(i,j)}), & \text{for } r = j, s = i, \\ \sin(\gamma_{(i,j)}), & \text{for } r = i, s = j, \\ 0, & \text{else,} \end{cases}$$

and $\gamma_{(i,j)} \in [-\pi, \pi)$ for all $\{(i,j) : i,j \in \{1,\ldots,K\}, j > i\}$.[9] This parametrization allows for a numerical optimization providing two matrices $D_-^{(r)}$ and $D_+^{(r)}$, where $D^{(r)}$ is then chosen as

$$D^{(r)} = \arg \min_{D_-^{(r)}, D_+^{(r)}} \{L_{D_-}(\Theta^*, \Theta(D_-^{(r)})), L_{D_+}(\Theta^*, \Theta(D_+^{(r)}))\}. \tag{25}$$

As the starting value for the numerical optimization we choose the solution defined by the WOP algorithm applied to $L_{D,1}(\Theta^*, \Theta^{(r)}(D^{(r)}))$ only. Convergence is quickly achieved and the overall improvement of the target value is generally very small, lying below 3% for all considered data scenarios.[10]

Step 2 Choose Θ^* as

$$\Theta^* = \frac{1}{R} \sum_{r=1}^{R} H(D^{(r)}) \Theta^{(r)}. \tag{26}$$

Following the line of arguments presented in case of the static factor model, we state the properties of the proposed ex-post approach towards the rotation problem for the dynamic factor model in form of two propositions.

Proposition 3.2. *The orthogonal matrix $D^{(r)}$ that minimizes the loss function in Equation (23) is unique conditional on almost every $\Theta^{(r)}$ and Θ^*, where the elements in $\Theta^{(r)}$ are random variables following a non-degenerate probability distribution as implied by prior distributions.*

Proof. The proof of Proposition 3.2 is given in Appendix B. □

Proposition 3.3. *The ex-post identification approach solves the rotation problem for the dynamic factor model.*

[9] This parametrization resembles the one used by Anderson et al. (1987), which is different with respect to the domain of the angular parameters, which is $\gamma_{(i,j)} \in [-\frac{\pi}{2}, \frac{\pi}{2})$, whereas in our decomposition, it is $\gamma_{(i,j)} \in [-\pi, \pi)$. Extending the domain accordingly allows to reduce the number of reflection parameters from K to 1, hence, our approach is more parsimonious with respect to the number of parameters, and, having all but one of the parameters living in the continuous space, is more easy to handle in optimizations.

[10] The accuracy of the numerical optimization procedure has been assessed via comparison of the numerical with the analytical solution in the static case.

Proof. The unsolved *rotation problem* implies that within the parameter space pairs of points can be defined, with two points being pairwise orthogonal transformations of each other according to Equation (5). Denote such a pair as $\Theta^{(1)}$ and $\Theta^{(2)}$ with $\Theta^{(2)} = H(D_0)\Theta^{(1)}$, where D_0 is an orthogonal matrix. To show that the *rotation problem* is solved by the suggested ex-post approach, one has to show that no such pairs can be defined after postprocessing. After postprocessing, $\Theta^{(1)}$ and $\Theta^{(2)}$ take the form $H(D_1)\Theta^{(1)}$ and $H(D_2)\Theta^{(2)}$ respectively, where D_i, $i = 1, 2$ implies minimal loss with regard to Θ^*. Since D_1 and D_2 are uniquely defined as shown in Proposition 3.2 and $H(D_2)\Theta^{(2)} = H(D_0 D_2)\Theta^{(1)}$ we have consequently $D_0 D_2 = D_1$, where we use the fact that the product of two orthogonal matrices is itself an orthogonal matrix, and orthogonal matrices commute. Assuming without loss of generality that $D_1 = I_K$, we have $D_2 = D_0'$. This implies that after postprocessing all points that can be represented as orthogonal transformations of $\Theta^{(1)}$ are collapsed into $\Theta^{(1)}$ as the point invoking minimal loss and thus enter the parameter estimation as $\Theta^{(1)}$. □

As the minimum of the loss function in Equation (14) is unique for almost every Θ^* and $\Theta^{(r)}$, there exists no orthogonal transformation of the postprocessed $\Theta^{(r)}$ that can further reduce the value of the loss function. Note that the ex-post identification approach yields a unique estimate for almost every input $\{\Theta^{(r)}\}_{r=1}^R$ with finite R conditional on the initial choice of Θ^*. A single round of the *WOP* approach thus generates a unique sequence $\{D^{(r)}\}_{r=1}^R$ for almost every Gibbs output with $R < \infty$ and a unique updated Θ^*. This argument applies iteratively until convergence providing a sequence $\{D_*^{(r)}\}_{r=1}^R$.

At convergence, we have that all orthogonal transformations of the sample based on a single orthogonal matrix, say D_{**}, imply the same loss at convergence and the alter the estimator accordingly as

$$\frac{1}{R}\sum_{r=1}^R \operatorname{tr}\left[\left(H(D_*^{(r)})\Theta^{(r)} - \frac{1}{R}\sum_{r=1}^R H(D_*^{(r)})\Theta^{(r)}\right)' H(D_{**})'H(D_{**})\left(H(D_*^{(r)})\Theta^{(r)} - \frac{1}{R}\sum_{r=1}^R H(D_*^{(r)})\Theta^{(r)}\right)\right]$$
$$= \frac{1}{R}\sum_{r=1}^R \operatorname{tr}\left[\left(H(D_{**}D_*^{(r)})\Theta^{(r)} - \frac{1}{R}\sum_{r=1}^R H(D_{**}D_*^{(r)})\Theta^{(r)}\right)' \left(H(D_{**}D_*^{(r)})\Theta^{(r)} - \frac{1}{R}\sum_{r=1}^R H(D_{**}D_*^{(r)})\Theta^{(r)}\right)\right].$$

Note that after the first iteration of the proposed ex post approach the difference of two distinct initializations with regard regard to the two implied sequences of orthogonal matrices, say $\{D_1^{(r)}\}_{r=1}^R$ and $\{D_2^{(r)}\}_{r=1}^R$ is captured within the sequence $\{D_1^{(r)} D_2^{(r)\prime}\}_{r=1}^R$, where we observe in all applications that this series converges to a unique orthogonal matrix, say $\bar{D}_{12}$, for each $r = 1. \ldots, R$. This implies that the effect of different initializations are mitigated at convergence up to an transformation based on a single orthogonal matrix.

We consider it an important advantage that the estimator implied by the defined loss function invokes the same loss and likelihood value as any orthogonal transformation of the parameter estimator, which reflects the ex-ante invariance concerning a specific choice for the orthogonal matrix involved in $H(D)$. To obtain a well-interpretable result, the user is free to apply an orthogonal transformation D_{**} to the resulting ex-post identified posterior distribution afterwards. Methods like Varimax or Quartimax can be applied to the postprocessed Bayes estimates, compare Kaiser (1958).

4 Comparison of ex-ante and ex-post solutions of the rotation problem

To illustrate the proposed ex-post approach towards the rotation problem we consider a static factor model, i.e. $S = P = 0$, and compare it to the ex-ante approach proposed by Geweke and Zhou (1996) for Bayesian factor analysis. In the ex-ante scheme the parameter space of the loadings is constrained to a positive lower triangular (PLT) matrix, i.e.

$$\Lambda = \begin{pmatrix} \lambda_{11} & \lambda_{21} & \dots & & & \lambda_{N1} \\ 0 & \lambda_{22} & \lambda_{32} & \dots & & \lambda_{N2} \\ \vdots & & \ddots & \ddots & & \vdots \\ 0 & \dots & 0 & \lambda_{KK} & \dots & \lambda_{NK} \end{pmatrix}', \quad \text{with} \quad \lambda_{ii} > 0, i = 1, \dots, K. \tag{27}$$

This is achieved in Maximum Likelihood factor analysis by constraining the upper triangular elements of the loadings matrix and finding the maximum of the accordingly constrained likelihood. In Bayesian factor analysis, the constraints are introduced by accordingly defined prior distributions, which are highly informative. In particular, a Dirac Delta prior is used for every λ_{ik} with $i < k$, and the prior on $(\lambda_{ii}, \dots, \lambda_{iK})'$ for $i \leq K$ is $K - i + 1$-variate Normal truncated below at zero along the first dimension. These constraints could equivalently be imposed ex-post based on a loss function defined in correspondence with the one in Equation (14). Following Stephens (2000), the corresponding risk criterion can be formulated as

$$-\sum_{i=2}^{K}\sum_{k=1}^{i-1} \log\left[I(\bar{\lambda}_{ik}(D) = 0)\right] - \sum_{i=1}^{K} \log\left[I(\bar{\lambda}_{ii}(D) > 0)\right] + \left(\sum_{i=1}^{N}\sum_{k=1}^{K}(\bar{\lambda}_{ik}(D) - \bar{\lambda}^{*}_{ik})^2\right), \tag{28}$$

where $\log(0)$ is defined by its limit value $-\infty$. The MC version of this loss function subject to $D'D = I$ can be minimized by performing a QR decomposition of Λ'. However, the resulting estimator and its corresponding orthogonal transformation do not invoke the same loss.

A well-known issue under the PLT scheme is that inference results depend on the ordering of the variables. This has been observed e.g. by Lopes and West (2004), Carvalho et al. (2008) or recently Chan et al. (2013). The reason for this ordering dependence can be motivated as follows. Consider the factor model given as $y_t = \Lambda f_t + e_t$. Then consider a $N \times N$ permutation matrix P that is premultiplied to y_t resulting in $Py_t = P\Lambda f_t + Pe_t$. Following Chan et al. (2013), when Λ has PLT form, then $P\Lambda$ almost surely does not have PLT form, since the set of matrices satisfying the PLT constraints under both orderings has Lebesgue measure zero. This implies that almost all admissible points under one set of PLT constraints are inadmissible under a different set of PLT constraints. Consider the transformation of a (posterior) distribution that satisfies the PLT constraints under a certain ordering of the data ($PLT|O_1$) into a distribution that satisfies the PLT constraints under a different ordering of the data ($PLT|O_2$), i.e.

$$PLT|O_1 \rightarrow PLT|O_2 = \{D_{\Lambda_{O_1}} : \Lambda_{O_1} D_{\Lambda_{O_1}} \in PLT|O_2\} \quad \text{for all} \quad \Lambda_{O_1} \in PLT|O_1. \tag{29}$$

This transformation involves an infinite number of orthogonal matrices D_Λ in order to let the transformed matrices satisfy the second set of constraints.[11] For the (posterior) distribution to remain unchanged except for an orthogonal transformation, there would have to be a unique orthogonal matrix performing this operation for every Λ admissible under the first set of constraints.

A further issue is multimodality. Imposing a lower triangularity constraint onto Λ without additionally demanding positive signs on the diagonal elements ensures *local* identification, see Anderson and Rubin (1956). Thus every reflection of a subset of columns of Λ yields the same likelihood value. Jennrich (1978) calls this phenomenon *transparent* multimodality. As Loken (2005) shows, introducing nonzero constraints leads to another type of multimodality, sometimes referred to *genuine* multimodality.[12] Whereas the constraints ensure that the parameter space contains only one global mode, they may induce multiple local modes that can no longer be mapped onto each other by axis reflections. Following the notion of Millsap (2001), imposing constraints may induce a solution from an equivalence class that is different from the equivalence class containing the true parameters, where an equivalence class corresponds to all points by the orthogonal transformation of the model parameters given in Equation (5).

To illustrate the issue of order dependence and multimodality, we discuss a small example that demonstrates the effect of the PLT restrictions on the shape of the likelihood. We start with a set of parameters for a model with $K = 2$ orthogonal static factors having unit variance and $N = 10$ variables, of which the first five are arranged in three different orderings, while the remaining five

[11] All matrices Λ already satisfying both sets of constraints are transformed by the identity matrix, whereas all matrices whose top $K \times K$ section is identical are transformed by the same orthogonal matrix. Finally, for those matrices whose top $K \times K$ section is singular, there exists no unique orthogonal matrix, see also Chan et al. (2013).

[12] Rubin and Thayer (1982) claim that this issue is also relevant in maximum likelihood confirmatory factor analysis, which is refuted by Bentler and Tanaka (1983), who attribute these findings to numerical problems. Such numerical problems, however, are potentially relevant for Bayesian factor analysis.

stay identical. This data set is simulated using as parameters

$$\Lambda_0 = \begin{pmatrix} 0.100 & -0.200 & 0.500 & 0.600 & 0.100 & 0.174 & -0.153 & -0.470 & 0.186 & -0.577 \\ 0.000 & 0.200 & -0.100 & 0.400 & -0.900 & 0.429 & -0.392 & 0.652 & 0.282 & -0.541 \end{pmatrix}' \tag{30}$$

and

$$\Sigma = \text{diag}(0.990, 0.920, 0.740, 0.480, 0.180, 0.786, 0.823, 0.354, 0.886, 0.374). \tag{31}$$

The three orderings we consider are the original one above, denoted $Y|O_1$, the second with ordering 2,3,1,4,5,6,7,8,9,10, denoted $Y|O_2$, and the third with ordering 5,2,1,3,4,6,7,8,9,10, denoted $Y|O_3$.

We first obtain the principal components estimate for Λ, which is afterwards transformed by $D_{\Lambda^*|O_1}$, $D_{\Lambda^*|O_2}$ and $D_{\Lambda^*|O_3}$, respectively, to satisfy the PLT constraints for each of the three orderings $Y|O_1$, $Y|O_2$ and $Y|O_3$ respectively. Of course, all three solutions attain the same log likelihood value. Next, we consider all possible orthogonal transformations of Λ^* under the three orderings. Those orthogonal transformations that are merely rotations can be expressed by an orthogonal matrix D_+ with $\det(D_+) = 1$. Those that involve a label switching between the factors or a reflection about a single axis require an orthogonal matrix D_- with $\det(D_-) = -1$. As can be seen from the decomposition of orthogonal matrices described in Equation (24), all orthogonal matrices in the $\mathbb{R}^2$ are expressible as a product of one rotation and one axis reflection. For illustration purposes, we replace the axis reflection by a permutation, which just adds a rotation by π to it, but leaves the determinant unchanged at -1. We first consider the rotations by applying a grid for the rotation angle between $-\pi$ and π. To incorporate the effects of the permutation, we also apply the grid of rotation angles to the loadings matrix with permuted columns. A permutation implies a label switching, accordingly the constraints on the loadings of the first factor are exchanged with constraints on the loadings of the second factor. To illustrate the effect of the identification constraints, all transformations are afterwards subject to the initial PLT constraints, i.e. all unconstrained parameters are transformed by the orthogonal matrix, the two parameters that are constrained to positivity are set to a small value $\epsilon > 0$ and the loading that is set to zero remains zero.[13]

Figure 1 shows the results of the exercise. The solid lines correspond to the transformations by means of rotations and the dashed lines to the transformations by means of a permutation and subsequent rotation. While under the first ordering, the likelihood is almost perfectly flat, hence the identification constraints have almost no effect at all, under the second ordering, the descent from the global mode is also quite flat in one direction, but considerably steeper in the other direction.

[13]Note that if there were no constraints, an orthogonal transformation would leave the likelihood value unchanged, and if there was merely the zero constraint, there likelihood profile on the circle would be bimodal under rotations with two identical modes exactly 180 degrees apart, see Loken (2005), and another two modes would evolve under the label switching and rotation from an identical likelihood profile, but shifted by 90 degrees.

Under the third ordering, the likelihood declines steeply in both directions. This result corroborates the finding in Carvalho et al. (2008) that some orderings of the data, corresponding to a particular choice of *factor founders*, provide more favorable inference results than others. Aside from the shape, the permutation and subsequent rotation induces a second mode, which is usually slightly lower than the first one. This mode is even present under the presumably well-behaved third ordering.

To illustrate the consequences for Bayesian analysis, we estimate the model parameters under all three orderings, incorporating the PLT constraints. We then repeat the estimation with the unconstrained sampler, and finally post-process the results of the unconstrained sampler using the WOP approach. The first row of Figure 2 shows the output of the unconstrained Gibbs sampler with respect to the loadings parameters of variable 8 on both factors, λ_8, under the three orderings, plotted as Gibbs sequences on the left and as bivariate contour plots on the right. The second row shows the according sampler output under the PLT constraint, and the third one the output post-processed with WOP. In all cases, 20,000 draws have been discarded as burn-in, another 20,000 have been kept. The posterior distribution of λ_8 from the unconstrained sampler clearly shows its invariance with respect to orthogonal transformations, whereas the shape of the bivariate posterior distribution of λ_8 under the three sets of PLT constraints varies in accordance with the likelihood profile shown in Figure 1. Eventually, the unconstrained Gibbs output postprocessed with WOP shows the same shape under all three orderings, hence being ordering-invariant, and is clearly unimodal. Unlike the PLT ex-ante approach towards the *rotation problem*, the WOP ex-post approach uses all information from the likelihood, including information on λ_{ik} for $i \leq k$, which is overridden in the PLT approach by the informative priors on the loadings of the factor founders. This does not only lead to ordering dependence of the resulting posterior estimates, but may also evoke multimodality of the posterior distribution.

5 Simulation study

To evaluate the properties of the proposed WOP ex-post approach, we now perform several simulation experiments, where we use the PLT ex-ante approach as a benchmark. The simulation experiments are designed to analyze the convergence, statistical and numerical properties of both approaches. We consider 24 different scenarios with the following features in common: First, at least 20% of the variation in the data is explained by the factors. Second, the simulated loadings matrices $\bar{\Lambda}$ all satisfy the PLT constraints, so the output from the PLT approach can be directly compared to the simulated parameters. Third, the diagonal elements of Λ_0 are chosen such that their explanatory power is the average of the explanatory power of all loadings on the same factor, which qualifies them as *factor founders* in the sense of Carvalho et al. (2008).

We simulate data sets with $T = 100$ and $T = 500$, respectively, using either $N = 30$ or $N = 100$

variables. Each of these scenarios is paired with $K = 2$ and $K = 4$ stationary factors, which are either static or follow $VAR(1)$ or $VAR(4)$ processes. The factor loadings are simulated according to the aforementioned conditions, where $\bar{\Lambda} = \Lambda_0$, i.e. $S = 0$. We investigate the corresponding 24 different scenarios. In all cases, the number of factors is assumed to be known. The priors are chosen as given in Chapter 3 with hyperparameters $\Upsilon_s = 100$, $s = 0, \dots, S$, $c = 1$, and $\alpha_{0i} = \beta_{0i} = 1$, $i = 1, \dots, N$.

Unlike the PLT approach, the WOP approach does not yield results where the estimated $\bar{\Lambda}$ is positive lower triangular. The actual values for all elements Θ^* used in the simulation are known, however, therefore it is possible to find the orthogonal transformation of the estimated Θ^* that minimizes the loss according to Equation (14), where no weighting scheme is used. This transformation treats the estimation errors of all entries of all elements of Θ^* alike, whereas transforming them such that $\bar{\Lambda}$ satisfies the PLT constraints reduces the errors of the constrained elements to zero at the cost of the inflating the errors of the remaining elements. Thus minimizing the loss for the estimate obtained from WOP might be considered giving WOP an unfair advantage. Therefore, we apply the same transformation on the estimates obtained under the PLT constraints, thus reducing the errors for the parameter estimates as well.

First, we analyze the convergence properties of the PLT and WOP approaches. Convergence is checked using the test by Geweke (1991), adjusting for autocorrelation in the draws by means of the heteroscedasticity and autocorrelation-robust covariance estimator by Newey and West (1987). We discard the initial 5,000 draws of the sampler and fix the length of the sample to be kept at 10,000 draws. While no convergence is found for the most recent 10,000 draws, the sequence is extended by another 1,000 draws. The burn-in sequence is not allowed to extend 100,000 draws, so we do not extend the sequence any further then and assume that it will not converge. Convergence statistics are evaluated for orthogonally invariant quantities only, i.e. the sum of squared loadings per variable, the idiosyncratic error variances, and the determinants of the persistence matrices of the factors at each lag. The total number of parameters monitored is therefore $2N + P$, and convergence is assumed if the Geweke statistic indicates convergence for 90% of the quantities, where the tests use $\alpha = 0.05$. Convergence for the WOP algorithm is assumed if the sum of squared differences between Θ^* in two subsequent iterations falls below 10^{-9}. We simulate 50 different samples for each scenario, which do not all converge for a burn-in limited to 100,000 draws. We therefore report both the number of cases where no convergence is attained and the convergence speed for 25 randomly selected converged sequences for each scenario. Table 1 shows the results. The scenarios with $K = 2$ factors do not experience any difficulties with respect to convergence, neither for the PLT approach nor for the WOP approach. The latter, however, requires on average about 600 additional draws on top of the initial burn-in of 5,000 draws, whereas the former requires about 1,500 additional draws. The scenarios with $K = 4$ factors require more iterations to converge and sometimes fail altogether. In particular, the scenario with $P = 4$, $N = 100$ and $T = 100$ stands out. It fails to converge for 4 out of 50 samples under the WOP approach and in 21 out of 50 samples under the PLT approach. Occasional non-convergence can be observed in some other scenarios for the PLT approach, while the WOP approach always converges. Beyond the initial burn-in, the WOP approach requires on average another 4,000 draws to converge, while whe PLT approach needs another 8,000. Note, however, that monitoring convergence of orthogonally invariant quantities that are functions of the parameters is not the same as monitoring convergence for the directed parameters in the case of the PLT sampler. If convergence is indicated for the orthogonally invariant quantities, estimates for the directed parameters may still perform poorly for the reasons discussed in Section . Yet, focusing on these quantities is the only feasible approach for comparing convergence behavior of the WOP and PLT approaches. Convergence of the directed parameters, however, always implies convergence of the orthogonally invariant quantities, hence the results can serve as a lower bound for the actual convergence in the PLT approach.

Second, we analyze the statistical properties of the posterior distributions obtained under both identification approaches. Table 2 reports the squared correlation coefficient between the simulated factors and the estimated factors. The results are overall similar for WOP and PLT in the scenarios with $K = 2$ factors. In scenarios with $K = 4$ factors, however, PLT sometimes performs considerably worse than WOP. The large standard deviations in these cases reveal that this is due to particular samples, indicating that under circumstances convenient for PLT, the results are similar to those obtained with WOP. Conversely, the WOP scheme provides at least as good results as the PLT approach for all model setups. Table 3 shows the root mean-squared errors (RMSE) for the loadings parameters. Since the models involve up to 2,000 loadings parameters, we only report the 5, 25, 50, 75 and 95 % quantiles of the RMSE for each parameter. With the single exception of the model with $N = 30$, $T = 500$, $K = 2$ and $P = 4$, all reported quantiles are lower for WOP than for PLT. In several models, particularly such models with static factors, the difference is negligible, whereas in other models, particularly those with high lag orders in the VAR process generating the factors, results are substantially in favor of WOP. Table 5 reports the RMSE for the persistence parameters. These parameters only exist in 16 out of the 24 models, and there are PK^2 of them in each model, hence their number is ranging between 4 and 64. Even though all parameters could be reported for the small models, due to the size of the large models, we again report the 5, 25, 50, 75 and 95 % quantiles of the RMSE. Judging by the median, the RMSE is often similar and tends to be better for WOP than for PLT. The upper quantiles, however, reveal that particularly in models with a more complex dynamic factor structure, WOP fares better than PLT. If we want to use large values for P, it thus seems advisable to use WOP rather than PLT for the estimation. Generally, estimates for quantities invariant to orthogonal transformations are very similar for PLT and WOP, as seen for the idiosyncratic variances. This consequently also holds for estimates for the product of factors and loadings, i.e. the systematic part of the model.[14] The posterior distributions obtained under the PLT constraints tend to be non-elliptical, as seen in Figure 2 and discussed in Section 5. This can also be highlighted by looking at the difference between $\widehat{\Lambda F'}$ and $\hat{\Lambda}\hat{F}'$, where $F = [f_1', \ldots, f_T']'$, e.g. measured as the Frobenius norm of the matrix of differences. This measure of divergence is generally much larger under PLT than under WOP, for scenarios with $K = 4$ dynamic factors, the divergence can be up to 1,000 times larger.

Eventually, we assess the numerical properties of both approaches, using 25 converged sequences. The directed parameter estimates are again transformed as described before. Table 6 shows the numerical standard errors for the loadings parameters. For all the models, the numerical standard errors are substantially larger for PLT than for WOP, particularly for models with a $K = 4$ factors and more complex persistence patterns in the factors. Looking at parameters invariant to orthogonal transformations, the verdict is different: Table 7 shows very similar results for the median numerical standard error for PLT and WOP, while the right tails reveal substantial differences for some models in favor of WOP. The same holds for the product of factors and loadings, which are not explicitly reported here. The persistence parameters in the factors, again a set of directed parameters, are evaluated in Table 8. Once again, the numerical standard errors are small for WOP, but large for PLT.

Altogether, the simulation study shows that both identification approaches yield very similar inference results for parameters invariant to orthogonal transformations. Modest improvements can be obtained by skipping ex-ante identification and using the WOP approach instead. Conversely, when inference on directed parameters is concerned, the WOP approach provides much better results than the PLT approach. These results hold for statistical as well as numerical properties. Since convergence is checked based on orthogonally invariant quantities only, the poorer performance of the estimates from the PLT approach is likely due to the non-elliptical shape of the posterior distributions and possible multimodality. These properties are induced by the ex-ante constraints and make the posterior distribution difficult to handle. Using the unconstrained sampler and postprocessing its output by the WOP approach prevents such problems.

6 Empirical example

To illustrate the WOP approach, we apply it to a data set of 120 macroeconomic time series from Bernanke et al. (2005). The time series are measured at monthly frequency over the period from January 1959 until August 2001, and undergo different types transformations to ensure stationarity. These transformations are described in detail by Bernanke et al. (2005). We replicate one of the model setups of Bernanke et al. (2005), with $S = 0$, $K = 4$ factors and $P = 7$ lags in the factor

[14] Results not reported here, but available upon request.

dynamics. Instead of considering the Fed Funds rate as an observable factor and the three remaining factors as latent, we assume that all factors are latent. The WOP approach allows for the estimation of factor-augmented vector-autoregressive models as well, however, this is beyond the scope of this paper.

First, we perform 20 repeated estimations of a specific ordering of the data using the PLT and the WOP approach. The four data series that are then affected by the PLT constraints, hence the factor founders, are the Fed Funds (FFYF), the industrial production (IP), the monetary base (FM2), and the consumer price index (PMCP). To make results under both approaches comparable, the posterior means under WOP are rotated such that they satisfy the PLT constraints. The required orthogonal matrix Q is found by first performing the QR decomposition $\bar{\Lambda}' = QR$, see e.g. Golub and van Loan (2013). Taking into account the signs of the diagonal elements of the corresponding R gives the additional column reflections to produce a loadings matrix satisfying the PLT constraints. The remaining model parameters can then be transformed accordingly. Figure 3 shows the 20 Bayes estimates for all four factors under both approaches. The correlation between the Fed Funds rate and the first factor is 0.9989 for PLT and 0.9987 for WOP. The numerical variation of the first two factors is slightly larger under the PLT approach compared to the WOP approach, while it is much larger for the last two factors. This result underlines that WOP provides Bayes estimates that have superior numerical properties.

Next, we consider 20 different orderings of the time series and estimate the accordingly specified factor model for each of them both by PLT and WOP.[15] Table 9 shows the average standard deviation over all 480 loadings parameters for each of the 20 different orderings of the variables. While the average standard deviations under the WOP approach vary only little, they deviate substantially from each other under PLT. Moreover, the smallest average standard deviation under PLT is almost as small as under WOP, while all other average standard deviations under the PLT approach are bigger. Figure 4 shows the results of the factor estimates, where the results under both approaches are orthogonally transformed such that the PLT structure holds with respect to the initial four factor founders. The results under the WOP approach are virtually identical compared to the first exercise, where the average correlation between the first factor and FYFF is again 0.9987. This illustrates that the estimation under the WOP approach is indeed invariant to the ordering of the variables. Results obtained under the PLT identification show clear variations, which are much bigger than under the conveniently ordered time series. The average correlation between the first factor and FYFF is now only 0.8418, with 12 out of the 20 orderings reaching a correlation of more than 0.99, but 6 out of them failing to exceed even 0.7. Orthogonally mapping pairs of parameter estimates obtained under the different orderings onto each other yields an average deviation, measured as the Frobenius norm of the matrix of differences, that is about 15 times larger for the estimates obtained

[15]The number of different orderings, or choice of factor founders, is prohibitively large, attaining 197 million, so we choose only a small random sample.

from the PLT approach, compared to those obtained from the WOP approach. This underlines the order dependency of the estimation under the PLT identification approach. The correlation results correspond to the result of the regression of the estimated onto the simulated factors in the simulation study in Section 5, shown in Table 2. It must be noted that while rather convenient orderings for PLT exist, they still do not outperform the results obtained under the WOP approach.

Figures 5 and 6 show the corresponding estimates for the factor loadings, giving the median as well as the 10% and 90% quantiles of the estimates. While the median under PLT with conveniently chosen factor founders resembles that under WOP, the range of the estimates is much wider. For randomly chosen orderings of the data series, the median under PLT is generally much closer to zero than the median under WOP. In summary, the results of the empirical example underline the results of the simulation study and highlight that the WOP identification approach has favorable numerical properties and provides ordering invariant inference.

7 Conclusion

We propose a novel ex-post approach to solve the rotation problem in Bayesian analysis of static and dynamic factor models. It has been common practice to impose constraints on the loadings matrix ex-ante by using truncated and degenerate prior distributions on its upper triangular elements, such that the sampler exclusively draws lower triangular loadings matrices with positive diagonal elements (PLT). If the PLT approach is used, however, inference results have been observed to be order-dependent. Thus we suggest to refrain from imposing ex-ante constraints via according prior distributions. Instead, we propose to use an orthogonally unconstrained sampler, which does not introduce constraints by the according prior distributions, but instead uses prior distributions for all model parameters that are invariant under orthogonal transformations. Using the orthogonally unconstrained sampler also avoids numerical problems that may occur when sampling from truncated distributions. The rotation problem is subsequently solved in a postprocessing step, where the distance between each draw of the unconstrained sampler and a fixed point is minimized. For static models the minimization problem for each draw of the sampler is the weighted orthogonal Procrustes problem, which has almost surely a unique analytic solution, hence the approach is likewise called weighted orthogonal Procrustes (WOP) approach. For dynamic models a unique solution exists almost surely, which can be found by using a numerical optimization routine.

The WOP approach has several desirable properties. The shape of the posterior distribution does not depend on the ordering of the data, hence the inference results are likewise no longer order-dependent. Furthermore, estimation and interpretation can be treated separately, as arbitrary rotation procedures, such like Varimax or Quartimax, can be applied to the posterior mean of the postprocessed Gibbs output. In a simulation study as well as in an application to a large macroeconomic data set, we compare the WOP approach with the commonly used PLT. Both exercises confirm the order-independence of the WOP approach, which also converges faster and yields lower MC errors.

Acknowledgements

For very helpful comments and thoughtful suggestions they provided on earlier versions of the paper, we thank Sylvia Frühwirth-Schnatter, Uwe Jensen, Sylvia Kaufmann, Roman Liesenfeld, Milan Stehlik, Helga Wagner, the participants of the Statistische Woche 2012, of the European Seminar on Bayesian Econometrics 2012, of the Vienna Workshop on High-Dimensional Time Series 2013, and of the ESEM 2013, and two anonymous referees.

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., Olkin, I., and Underhill, L. (1987). Generation of Random Orthogonal Matrices. *SIAM Journal on Scientific and Statistical Computing*, 8:625–629.

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.

Bai, J. and Wang, P. (2012). Identification and estimation of dynamic factor models. MPRA Paper 38434, University Library of Munich, Germany.

Bentler, P. and Tanaka, J. (1983). Problems with EM algorithms for ML Factor Analysis. *Psychometrika*, 48(2):247–251.

Bernanke, B., Boivin, J., and Eliasz, P. S. (2005). Measuring the Effects of Monetary Policy: A Factor-augmented Vector Autoregressive (FAVAR) Approach. *The Quarterly Journal of Economics*, 120(1):387–422.

Carter, C. K. and Kohn, R. (1994). On Gibbs sampling for state space models. *Biometrika*, 81(3):541–553.

Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q., and West, M. (2008). High-Dimensional Sparse Factor Modeling: Applications in Gene Expression Genomics. *Journal of the American Statistical Association*, 103(4):1438–1456.

Celeux, G. (1998). Bayesian inference for mixtures: The label-switching problem. In Payne, R. and Green, P. J., editors, *COMPSTAT 98—Proc. in Computational Statistics*, pages 227–233.

Celeux, G., Hurn, M., and Robert, C. (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, 95(451):957–970.

Chan, J., Leon-Gonzalez, R., and Strachan, R. W. (2013). Invariant Inference and Efficient Computation in the Static Factor Model. mimeo.

Cheng, M.-Y., Hall, P., and Turlach, B. A. (1999). High-derivative parametric enhancements of nonparametric curve estimators. *Biometrika*, 86(2):417–428.

Clarke, B., Andrew, and Barron, R. (1990). Information-theoretic Asymptotics of Bayes Methods. *IEEE Transactions on Information Theory*, 36:453–471.

Frühwirth-Schnatter, S. (2001). Fully Bayesian Analysis of Switching Gaussian State Space Models. *Annals of the Institute of Statistical Mathematics*, 53(1):31–49.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer, New York.

Frühwirth-Schnatter, S. and Lopes, H. F. (2010). Parsimonious Bayesian Factor Analysis when the Number of Factors is Unknown. Technical report, University of Chicago Booth School of Business.

Gelman, A., van Dyk, D. A., Huang, Z., and Boscardin, J. W. (2008). Using Redundant Parameterizations to Fit Hierarchical Models. *Journal of Computational and Graphical Statistics*, 17(1):95–122.

Geweke, J. (1991). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. Staff Report 148, Federal Reserve Bank of Minneapolis.

Geweke, J. and Zhou, G. (1996). Measuring the Pricing Error of the Arbitrage Pricing Theory. *Review of Financial Studies*, 9(2):557–587.

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

Grün, B. and Leisch, F. (2009). Dealing With Label Switching in Mixture Models Under Genuine Multimodality. *Journal of Multivariate Analysis*, 100(5):851–861.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press, Princeton NJ.

Jasra, A., Holmes, C., and Stephens, D. (2005). Markov chain monte carlo methods and the label switching problem in bayesian mixture modeling. *Statistical Science*, 20(1):50–67.

Jennrich, R. I. (1978). Rotational Equivalence of Factor Loading Matrices with Prespecified Values. *Psychometrika*, 43:421–426.

Jöreskog, K. G. (1979). Structural Equations Models in the Social Sciences: Specification, Estimation and Testing. In Magidson, J., editor, *Advances in Factor Analysis and Structural Equation Models*. Abt Books, Cambridge, MA.

Kaufmann, S. and Schumacher, C. (2013). Bayesian Estimation of Sparse Dynamic Factor Models With Order-Independent Identification. Working Paper 1304, Study Center Gerzensee.

Koschat, M. and Swayne, D. (1991). A weighted Procrustes criterion. *Psychometrika*, 56(2):229–239.

Kristof, W. (1964). Die beste orthogonale tranformation zur gegenseitigen überführung zweier factorenmatrizen. *Diagnostica*, 10:87–90.

Lissitz, R., Schönemann, P., and Lingoes, J. (1976). A solution to the weighted procrustes problem in which the transformation is in agreement with the loss function. *Psychometrika*, 41(4):547–550.

Loken, E. (2005). Identification Constraints and Inference in Factor Analysis Models. *Structural Equation Modeling*, 12:232–244.

Lopes, H. F. and West, M. (2004). Bayesian Model Assessment in Factor Analysis. *Statistica Sinica*, 14:41–67.

Meng, X.-L. and van Dyk, D. A. (1999). Seeking Efficient Data Augmentation Schemes via Conditional and Marginal Augmentation. *Biometrika*, 86:301–320.

Millsap, R. E. (2001). When Trivial Constraints Are Not Trivial: The Choice of Uniqueness Constraints in Confirmatory Factor Analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, 8(1):1–17.

Newey, W. K. and West, K. D. (1987). A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica*, 55(3):703–708.

Ni, S. and Sun, D. (2005). Bayesian estimates for vector autoregressive models. *Journal of Business & Economic Statistics*, 23:105–117.

Redner, R. A. and Walker, H. F. (1984). Mixture Densities, Maximum Likelihood and the EM Algorithm. *SIAM Review*, 26(2):195–239.

Richardson, S. and Green, P. J. (1997). On Bayesian Analysis of Mixtures With an Unknown Number of Components. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 59(4):731–792.

Rubin, D. and Thayer, D. (1982). EM algorithms for ML factor analysis. *Psychometrika*, 47(1):69–76.

Schönemann, P. H. (1966). A generalized solution to the orthogonal Procrustes Problem. *Psychometrika*, 31(1):1–10.

Shumway, R. and Stoffer, D. (2010). *Time Series Analysis and Its Applications: With R Examples*. Springer Texts in Statistics. Springer.

Smith, J. O. (2007). *Introduction to Digital Filters with Audio Applications*. W3K Publishing, http://www.w3k.org/books.

Stephens, M. (2000). Dealing with Label Switching in Mixture Models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 62(4):795–809.

Tippett, M. K., Anderson, J. L., Bishop, C. H., Hamill, T. M., and Whitaker, J. S. (2003). Ensemble square root filters. *Monthly Weather Review*, 131:1485–1490.

van Dyk, D. A. and Meng, X.-L. (2001). The Art of Data Augmentation. *Journal of Computational and Graphical Statistics*, 10:1–50.

Tables

			PLT						WOP					
N	T	P	$K=2$			$K=4$			$K=2$			$K=4$		
30	100	0	0	6160	(1028)	0	6600	(2041)	0	5280	(614)	0	5640	(1186)
30	100	1	0	5840	(1344)	0	11160	(6176)	0	5800	(1384)	0	5760	(1200)
30	100	4	0	6760	(3666)	0	13440	(13961)	0	5800	(1414)	0	17120	(13581)
30	500	0	0	5480	(714)	0	6680	(1973)	0	5320	(690)	0	5840	(1214)
30	500	1	0	6200	(3000)	0	8440	(2022)	0	5360	(638)	0	7320	(2155)
30	500	4	0	6160	(2544)	2	14200	(14428)	0	6760	(1665)	0	7240	(2891)
100	100	0	0	6000	(1414)	0	5640	(810)	0	5400	(645)	0	6160	(2014)
100	100	1	0	6200	(1633)	1	19320	(18538)	0	5560	(1044)	0	7440	(2501)
100	100	4	0	7480	(4094)	21	42160	(24535)	0	5800	(1291)	4	26720	(19661)
100	500	0	0	6240	(1234)	0	7920	(3651)	0	5160	(473)	0	6760	(2314)
100	500	1	0	8160	(2075)	0	8800	(4041)	0	5280	(614)	0	5760	(1012)
100	500	4	0	7080	(2971)	1	14280	(8299)	0	5640	(757)	0	7160	(2495)

Table 1: Number of sequences not converged after 100,000 iterations, average length of burn-in for 25 randomly chosen converged sequences per model. Standard errors in parentheses. Convergence is monitored for orthogonally invariant statistics of the parameters and assumed to hold if Geweke's (1991) test does not reject the Null hypothesis of convergence for at least 90% of the parameters with $\alpha = 5\%$. Heteroskedasticity and autocorrelation (HAC) robust standard errors have been used.

							PLT									
N	T	P	2 factors				4 factors									
30	100	0	0.9766	(0.0037)	0.9759	(0.0069)	0.9681	(0.0110)	0.9638	(0.0155)	0.9538	(0.0265)	0.9583	(0.0195)
30	100	1	0.9774	(0.0041)	0.9926	(0.0031)	0.8508	(0.1724)	0.9036	(0.1092)	0.8815	(0.1014)	0.9236	(0.0723)
30	100	4	0.9764	(0.0110)	0.9811	(0.0117)	0.5425	(0.2410)	0.6621	(0.1961)	0.6695	(0.1391)	0.8119	(0.1248)
30	500	0	0.9778	(0.0019)	0.9789	(0.0018)	0.9750	(0.0046)	0.9713	(0.0068)	0.9713	(0.0106)	0.9715	(0.0048)
30	500	1	0.9820	(0.0041)	0.9726	(0.0045)	0.9812	(0.0118)	0.9642	(0.0143)	0.9819	(0.0160)	0.9826	(0.0087)
30	500	4	0.9859	(0.0042)	0.9926	(0.0026)	0.8680	(0.1089)	0.6660	(0.2018)	0.8653	(0.0781)	0.5558	(0.2012)
100	100	0	0.9903	(0.0027)	0.9902	(0.0023)	0.9788	(0.0178)	0.9814	(0.0093)	0.9775	(0.0170)	0.9828	(0.0105)
100	100	1	0.9857	(0.0258)	0.9927	(0.0144)	0.8558	(0.1967)	0.8754	(0.1116)	0.9183	(0.1050)	0.7879	(0.2655)
100	100	4	0.9949	(0.0033)	0.9951	(0.0025)	0.6181	(0.2319)	0.7312	(0.1716)	0.8010	(0.2242)	0.6259	(0.2203)
100	500	0	0.9911	(0.0009)	0.9910	(0.0011)	0.9886	(0.0032)	0.9891	(0.0028)	0.9881	(0.0042)	0.9875	(0.0059)
100	500	1	0.9963	(0.0009)	0.9917	(0.0012)	0.9887	(0.0049)	0.9900	(0.0062)	0.9893	(0.0047)	0.9877	(0.0048)
100	500	4	0.9936	(0.0026)	0.9951	(0.0023)	0.7817	(0.1170)	0.8493	(0.1194)	0.7313	(0.1362)	0.4349	(0.1946)
							WOP									
N	T	P	2 factors				4 factors									
30	100	0	0.9766	(0.0038)	0.9761	(0.0071)	0.9738	(0.0060)	0.9701	(0.0064)	0.9724	(0.0060)	0.9681	(0.0097)
30	100	1	0.9775	(0.0044)	0.9927	(0.0026)	0.9724	(0.0076)	0.9865	(0.0072)	0.9753	(0.0071)	0.9878	(0.0043)
30	100	4	0.9788	(0.0081)	0.9828	(0.0082)	0.9792	(0.0084)	0.9722	(0.0109)	0.9879	(0.0049)	0.9829	(0.0064)
30	500	0	0.9903	(0.0025)	0.9901	(0.0023)	0.9843	(0.0060)	0.9862	(0.0047)	0.9863	(0.0045)	0.9859	(0.0049)
30	500	1	0.9915	(0.0026)	0.9961	(0.0018)	0.9967	(0.0019)	0.9897	(0.0040)	0.9934	(0.0028)	0.9861	(0.0051)
30	500	4	0.9962	(0.0017)	0.9959	(0.0016)	0.9691	(0.0241)	0.9656	(0.0133)	0.9770	(0.0091)	0.9654	(0.0143)
100	100	0	0.9778	(0.0019)	0.9790	(0.0019)	0.9771	(0.0021)	0.9745	(0.0023)	0.9762	(0.0026)	0.9728	(0.0025)
100	100	1	0.9843	(0.0033)	0.9749	(0.0037)	0.9874	(0.0035)	0.9725	(0.0028)	0.9897	(0.0022)	0.9841	(0.0033)
100	100	4	0.9803	(0.0049)	0.9861	(0.0042)	0.9809	(0.0039)	0.9894	(0.0028)	0.9811	(0.0021)	0.9851	(0.0029)
100	500	0	0.9913	(0.0008)	0.9911	(0.0009)	0.9899	(0.0014)	0.9902	(0.0015)	0.9896	(0.0016)	0.9903	(0.0012)
100	500	1	0.9964	(0.0007)	0.9919	(0.0010)	0.9900	(0.0016)	0.9919	(0.0017)	0.9906	(0.0018)	0.9890	(0.0025)
100	500	4	0.9933	(0.0022)	0.9953	(0.0012)	0.7887	(0.0714)	0.9571	(0.0243)	0.8946	(0.0580)	0.7824	(0.1170)

Table 2: Squared correlation between the estimated factors and the simulated factors from 25 randomly chosen converged sequences per model using PLT/WOP identification and optimal orientation of the estimated parameters, i.e. results are subsequently orthogonally transformed, such that the distance between the estimated and simulated parameters is minimized. Standard errors in parentheses.

				PLT					WOP				
N	T	K	P	q_{05}	q_{25}	q_{50}	q_{75}	q_{95}	q_{05}	q_{25}	q_{50}	q_{75}	q_{95}
30	100	2	0	0.0674	0.1066	0.1232	0.1441	0.1814	0.0658	0.0962	0.1137	0.1321	0.1612
30	100	2	1	0.1851	0.2375	0.2779	0.3100	0.4021	0.0583	0.0871	0.1002	0.1197	0.1569
30	100	2	4	0.2246	0.3090	0.3770	0.4226	0.5283	0.1104	0.1652	0.2428	0.3429	0.4748
30	100	4	0	0.1023	0.1782	0.2366	0.2927	0.3625	0.0871	0.1131	0.1304	0.1484	0.1851
30	100	4	1	0.2449	0.3920	0.4939	0.5811	0.7308	0.0752	0.1224	0.1548	0.1899	0.2554
30	100	4	4	0.3225	0.5027	0.6588	0.8169	1.0892	0.1051	0.2163	0.3103	0.3848	0.5060
30	500	2	0	0.0318	0.0523	0.0597	0.0713	0.0835	0.0317	0.0515	0.0586	0.0667	0.0796
30	500	2	1	0.0521	0.0908	0.1277	0.1588	0.2130	0.0403	0.0722	0.1188	0.1396	0.1743
30	500	2	4	0.1013	0.1648	0.2242	0.3163	0.4940	0.0572	0.1497	0.2135	0.3646	0.5385
30	500	4	0	0.0542	0.0886	0.1094	0.1378	0.1907	0.0456	0.0603	0.0679	0.0755	0.0941
30	500	4	1	0.0865	0.1389	0.1635	0.2123	0.2779	0.0463	0.0797	0.1101	0.1524	0.2432
30	500	4	4	0.1850	0.3829	0.5224	0.6727	1.0558	0.0549	0.1338	0.1881	0.2444	0.3656
100	100	2	0	0.0778	0.1075	0.1220	0.1373	0.1734	0.0756	0.1002	0.1135	0.1279	0.1616
100	100	2	1	0.0968	0.1260	0.1760	0.2128	0.2644	0.0616	0.0777	0.0961	0.1139	0.1574
100	100	2	4	0.1841	0.2895	0.3244	0.3617	0.4115	0.0763	0.1050	0.1269	0.1493	0.1841
100	100	4	0	0.1146	0.1807	0.2122	0.2515	0.3265	0.0876	0.1128	0.1297	0.1476	0.1831
100	100	4	1	0.2452	0.3696	0.4462	0.5426	0.7669	0.0783	0.1059	0.1216	0.1377	0.1801
100	100	4	4	0.4041	0.5659	0.6458	0.7598	0.9625	0.1349	0.2746	0.4840	0.6739	1.0730
100	500	2	0	0.0439	0.0547	0.0633	0.0712	0.0871	0.0387	0.0481	0.0541	0.0620	0.0729
100	500	2	1	0.0350	0.0479	0.0565	0.0656	0.0779	0.0302	0.0386	0.0455	0.0561	0.0720
100	500	2	4	0.0638	0.0782	0.0871	0.0985	0.1193	0.0423	0.0529	0.0671	0.1015	0.1334
100	500	4	0	0.0647	0.0887	0.1059	0.1325	0.2109	0.0451	0.0574	0.0656	0.0732	0.0899
100	500	4	1	0.0590	0.0772	0.0903	0.1100	0.1477	0.0479	0.0626	0.0757	0.0956	0.1608
100	500	4	4	0.4626	0.6600	0.8145	0.9796	1.4613	0.1417	0.2389	0.3435	0.4449	0.7268

Table 3: RMSEs for the loadings parameters from 25 randomly chosen converged sequences per model. Instead of reporting all NK parameters per model, we only report the 5%, 25%, 50%, 75% and 95% quantile. The left five columns show the results obtained under PLT identification and optimal orientation of the estimated parameters, i.e. results are subsequently orthogonally transformed, such that the distance between the estimated and simulated parameters is minimized, and the right five columns show the results obtained under WOP identification and optimal orientation of the estimated parameters, i.e. results are subsequently orthogonally transformed, such that the distance between the estimated and simulated parameters is minimized.

				PLT					WOP				
N	T	K	P	q_{05}	q_{25}	q_{50}	q_{75}	q_{95}	q_{05}	q_{25}	q_{50}	q_{75}	q_{95}
30	100	2	0	0.0656	0.1057	0.1357	0.1590	0.2571	0.0655	0.1063	0.1365	0.1588	0.2574
30	100	2	1	0.0606	0.1185	0.1374	0.1776	0.2638	0.0479	0.1138	0.1336	0.1720	0.2537
30	100	2	4	0.0667	0.1082	0.1416	0.1804	0.2542	0.0582	0.1013	0.1343	0.1747	0.2548
30	100	4	0	0.0877	0.1299	0.1497	0.1690	0.2726	0.0863	0.1296	0.1498	0.1690	0.2722
30	100	4	1	0.0740	0.1287	0.1635	0.1897	0.3468	0.0737	0.1257	0.1493	0.1864	0.2596
30	100	4	4	0.0828	0.1355	0.1564	0.2014	0.3884	0.0826	0.1210	0.1467	0.1840	0.3510
30	500	2	0	0.0260	0.0449	0.0606	0.0712	0.0991	0.0259	0.0450	0.0607	0.0712	0.0989
30	500	2	1	0.0264	0.0406	0.0601	0.0768	0.1068	0.0263	0.0406	0.0602	0.0769	0.1061
30	500	2	4	0.0320	0.0436	0.0584	0.0681	0.1299	0.0251	0.0391	0.0563	0.0662	0.1297
30	500	4	0	0.0367	0.0528	0.0609	0.0817	0.1179	0.0365	0.0526	0.0608	0.0822	0.1178
30	500	4	1	0.0369	0.0533	0.0676	0.0777	0.0963	0.0370	0.0485	0.0653	0.0755	0.0961
30	500	4	4	0.0359	0.0573	0.0676	0.0863	0.1923	0.0359	0.0551	0.0632	0.0793	0.1241
100	100	2	0	0.0613	0.1071	0.1372	0.1799	0.2561	0.0613	0.1068	0.1370	0.1786	0.2556
100	100	2	1	0.0616	0.0993	0.1426	0.1720	0.2626	0.0607	0.0970	0.1426	0.1720	0.2521
100	100	2	4	0.0612	0.1114	0.1408	0.1720	0.2866	0.0595	0.1074	0.1368	0.1697	0.2822
100	100	4	0	0.0691	0.1053	0.1372	0.1828	0.2930	0.0690	0.1058	0.1373	0.1829	0.2932
100	100	4	1	0.0678	0.1105	0.1405	0.1728	0.3355	0.0682	0.1079	0.1378	0.1723	0.2783
100	100	4	4	0.0693	0.1055	0.1450	0.1771	0.3532	0.0734	0.1104	0.1399	0.1703	0.2966
100	500	2	0	0.0275	0.0450	0.0584	0.0787	0.1222	0.0275	0.0450	0.0584	0.0786	0.1221
100	500	2	1	0.0276	0.0445	0.0589	0.0788	0.1217	0.0277	0.0446	0.0588	0.0786	0.1220
100	500	2	4	0.0273	0.0464	0.0614	0.0764	0.1170	0.0270	0.0462	0.0614	0.0764	0.1170
100	500	4	0	0.0291	0.0476	0.0647	0.0798	0.1283	0.0292	0.0478	0.0645	0.0798	0.1280
100	500	4	1	0.0288	0.0457	0.0638	0.0770	0.1205	0.0288	0.0454	0.0630	0.0771	0.1203
100	500	4	4	0.0333	0.0540	0.0773	0.0948	0.2345	0.0332	0.0546	0.0748	0.0989	0.1767

Table 4: RMSEs for the idiosyncratic variance parameters from 25 randomly chosen converged sequences per model. Instead of reporting all N parameters per model, we only report the 5%, 25%, 50%, 75% and 95% quantile. The left five columns show the results obtained under PLT identification, and the right five columns show the results obtained under WOP identification.

				PLT					WOP				
N	T	K	P	q_{05}	q_{25}	q_{50}	q_{75}	q_{95}	q_{05}	q_{25}	q_{50}	q_{75}	q_{95}
30	100	2	1	0.0679	0.0936	0.1160	0.1386	0.1644	0.0439	0.0525	0.0686	0.0900	0.1067
30	100	2	4	0.0709	0.0922	0.0986	0.1230	0.1781	0.0781	0.0907	0.0997	0.1068	0.1222
30	100	4	1	0.0545	0.0620	0.0774	0.1162	0.2533	0.0488	0.0603	0.0763	0.0936	0.1093
30	100	4	4	0.0594	0.0726	0.0958	0.1282	0.2215	0.0721	0.0913	0.1060	0.1190	0.1567
30	500	2	1	0.0392	0.0420	0.0456	0.0647	0.0907	0.0350	0.0385	0.0425	0.0605	0.0850
30	500	2	4	0.0366	0.0473	0.0500	0.0563	0.0764	0.0334	0.0446	0.0497	0.0561	0.0818
30	500	4	1	0.0240	0.0344	0.0403	0.0550	0.1019	0.0196	0.0281	0.0321	0.0418	0.0492
30	500	4	4	0.0330	0.0475	0.0647	0.0912	0.1509	0.0270	0.0374	0.0437	0.0470	0.0599
100	100	2	1	0.0558	0.0564	0.0746	0.0976	0.1052	0.0392	0.0518	0.0684	0.0830	0.0925
100	100	2	4	0.0732	0.1242	0.1447	0.1633	0.1888	0.0724	0.0934	0.1143	0.1307	0.1695
100	100	4	1	0.0582	0.0738	0.0773	0.1034	0.2500	0.0378	0.0568	0.0727	0.0847	0.1259
100	100	4	4	0.0640	0.0853	0.1055	0.1324	0.1705	0.0817	0.1100	0.1204	0.1378	0.1614
100	500	2	1	0.0259	0.0260	0.0298	0.0409	0.0519	0.0265	0.0271	0.0320	0.0423	0.0511
100	500	2	4	0.0346	0.0386	0.0425	0.0460	0.0507	0.0339	0.0396	0.0423	0.0462	0.0513
100	500	4	1	0.0267	0.0321	0.0377	0.0438	0.0491	0.0270	0.0337	0.0375	0.0442	0.0515
100	500	4	4	0.0463	0.0661	0.0914	0.1228	0.2093	0.0424	0.0585	0.0772	0.1047	0.2131

Table 5: RMSEs for the persistence parameters in the factors from 25 randomly chosen converged sequences. Instead of reporting all PK^2 parameters per model, we only report the 5%, 25%, 50%, 75% and 95% quantile. The left five columns show the results obtained under PLT identification, and the right five columns show the results obtained under WOP identification.

				PLT					WOP				
N	T	K	P	q_{05}	q_{25}	q_{50}	q_{75}	q_{95}	q_{05}	q_{25}	q_{50}	q_{75}	q_{95}
30	100	2	0	0.0072	0.0143	0.0168	0.0231	0.0364	0.0021	0.0046	0.0057	0.0072	0.0096
30	100	2	1	0.0364	0.0835	0.1196	0.1653	0.3328	0.0029	0.0047	0.0058	0.0079	0.0101
30	100	2	4	0.0095	0.0178	0.0238	0.0292	0.0372	0.0028	0.0048	0.0059	0.0075	0.0096
30	100	4	0	0.0377	0.0605	0.0814	0.1023	0.1557	0.0014	0.0046	0.0053	0.0061	0.0085
30	100	4	1	0.0476	0.0978	0.1743	0.2194	0.3184	0.0038	0.0068	0.0091	0.0133	0.0199
30	100	4	4	0.0622	0.0879	0.1090	0.1297	0.2060	0.0031	0.0067	0.0082	0.0105	0.0140
30	500	2	0	0.0044	0.0076	0.0093	0.0110	0.0153	0.0011	0.0020	0.0024	0.0028	0.0037
30	500	2	1	0.0054	0.0087	0.0119	0.0182	0.0270	0.0015	0.0029	0.0036	0.0041	0.0052
30	500	2	4	0.0081	0.0133	0.0166	0.0203	0.0246	0.0019	0.0044	0.0051	0.0065	0.0097
30	500	4	0	0.0189	0.0341	0.0586	0.0855	0.1335	0.0010	0.0029	0.0033	0.0036	0.0045
30	500	4	1	0.0339	0.0572	0.1512	0.2194	0.3231	0.0013	0.0037	0.0048	0.0063	0.0087
30	500	4	4	0.0451	0.1073	0.1643	0.2141	0.3405	0.0015	0.0036	0.0046	0.0053	0.0066
100	100	2	0	0.0289	0.0441	0.0550	0.0674	0.0939	0.0026	0.0034	0.0040	0.0047	0.0059
100	100	2	1	0.0145	0.0308	0.0712	0.0993	0.1271	0.0039	0.0053	0.0061	0.0071	0.0085
100	100	2	4	0.0473	0.0872	0.1249	0.1699	0.2145	0.0033	0.0050	0.0060	0.0089	0.0110
100	100	4	0	0.0307	0.0448	0.0555	0.0714	0.0917	0.0035	0.0047	0.0055	0.0063	0.0082
100	100	4	1	0.1009	0.1446	0.1891	0.2523	0.3768	0.0045	0.0063	0.0071	0.0082	0.0099
100	100	4	4	0.0560	0.0979	0.1255	0.1559	0.2894	0.0108	0.0153	0.0202	0.0271	0.0375
100	500	2	0	0.0054	0.0124	0.0263	0.0373	0.0488	0.0027	0.0033	0.0036	0.0040	0.0050
100	500	2	1	0.0131	0.0226	0.0286	0.0348	0.0430	0.0030	0.0042	0.0048	0.0052	0.0062
100	500	2	4	0.0212	0.0446	0.0590	0.0672	0.0806	0.0030	0.0040	0.0047	0.0058	0.0074
100	500	4	0	0.0308	0.0445	0.0581	0.0814	0.1637	0.0032	0.0040	0.0045	0.0048	0.0059
100	500	4	1	0.0297	0.0446	0.0588	0.0749	0.1719	0.0035	0.0046	0.0051	0.0057	0.0069
100	500	4	4	0.1023	0.1663	0.2556	0.3656	0.5931	0.0037	0.0058	0.0074	0.0096	0.0150

Table 6: Numerical standard errors for the loadings parameters from 25 randomly chosen converged sequences. Instead of reporting all NK parameters per model, we only report the 5%, 25%, 50%, 75% and 95% quantile. The left five columns show the results obtained under PLT identification, and the right five columns show the results obtained under WOP identification.

				PLT					WOP				
N	T	K	P	q_{05}	q_{25}	q_{50}	q_{75}	q_{95}	q_{05}	q_{25}	q_{50}	q_{75}	q_{95}
30	100	2	0	0.0007	0.0010	0.0012	0.0015	0.0025	0.0007	0.0011	0.0015	0.0019	0.0028
30	100	2	1	0.0007	0.0010	0.0013	0.0019	0.0029	0.0006	0.0010	0.0013	0.0018	0.0029
30	100	2	4	0.0006	0.0010	0.0013	0.0018	0.0028	0.0004	0.0011	0.0014	0.0019	0.0030
30	100	4	0	0.0010	0.0016	0.0019	0.0023	0.0114	0.0009	0.0015	0.0019	0.0023	0.0030
30	100	4	1	0.0011	0.0015	0.0021	0.0024	0.3164	0.0011	0.0016	0.0018	0.0023	0.0030
30	100	4	4	0.0009	0.0016	0.0021	0.0025	0.0533	0.0010	0.0017	0.0018	0.0022	0.0037
30	500	2	0	0.0002	0.0004	0.0005	0.0007	0.0013	0.0003	0.0005	0.0006	0.0009	0.0012
30	500	2	1	0.0002	0.0004	0.0006	0.0007	0.0016	0.0003	0.0005	0.0006	0.0009	0.0011
30	500	2	4	0.0003	0.0004	0.0006	0.0007	0.0013	0.0003	0.0005	0.0006	0.0008	0.0012
30	500	4	0	0.0004	0.0006	0.0007	0.0010	0.0031	0.0004	0.0006	0.0008	0.0012	0.0015
30	500	4	1	0.0004	0.0006	0.0007	0.0009	0.0174	0.0004	0.0007	0.0008	0.0011	0.0015
30	500	4	4	0.0004	0.0006	0.0007	0.0011	0.0243	0.0004	0.0006	0.0008	0.0010	0.0016
100	100	2	0	0.0005	0.0009	0.0012	0.0016	0.0028	0.0006	0.0010	0.0015	0.0019	0.0031
100	100	2	1	0.0006	0.0009	0.0011	0.0015	0.0024	0.0006	0.0011	0.0014	0.0019	0.0029
100	100	2	4	0.0006	0.0010	0.0014	0.0017	0.0029	0.0007	0.0011	0.0014	0.0018	0.0030
100	100	4	0	0.0007	0.0010	0.0013	0.0017	0.0032	0.0008	0.0012	0.0015	0.0021	0.0034
100	100	4	1	0.0007	0.0010	0.0014	0.0017	0.0031	0.0008	0.0012	0.0015	0.0019	0.0032
100	100	4	4	0.0007	0.0012	0.0015	0.0021	0.0055	0.0008	0.0012	0.0016	0.0020	0.0031
100	500	2	0	0.0002	0.0004	0.0005	0.0006	0.0011	0.0003	0.0005	0.0006	0.0008	0.0012
100	500	2	1	0.0002	0.0004	0.0005	0.0006	0.0012	0.0003	0.0005	0.0006	0.0007	0.0013
100	500	2	4	0.0003	0.0004	0.0005	0.0007	0.0011	0.0003	0.0005	0.0006	0.0008	0.0013
100	500	4	0	0.0003	0.0005	0.0006	0.0007	0.0012	0.0003	0.0005	0.0007	0.0009	0.0013
100	500	4	1	0.0003	0.0004	0.0005	0.0007	0.0013	0.0004	0.0005	0.0007	0.0008	0.0012
100	500	4	4	0.0006	0.0013	0.0027	0.0053	0.0278	0.0003	0.0006	0.0008	0.0010	0.0017

Table 7: Numerical standard errors for the idiosyncratic variances from 25 randomly chosen converged sequences. Instead of reporting all N parameters per model, we only report the 5%, 25%, 50%, 75% and 95% quantile. The left five columns show the results obtained under PLT identification, and the right five columns show the results obtained under WOP identification.

				PLT					WOP				
N	T	K	P	q_{05}	q_{25}	q_{50}	q_{75}	q_{95}	q_{05}	q_{25}	q_{50}	q_{75}	q_{95}
30	100	2	1	0.0487	0.0497	0.0601	0.0699	0.0701	0.0005	0.0007	0.0010	0.0018	0.0027
30	100	2	4	0.0023	0.0030	0.0035	0.0075	0.0109	0.0013	0.0014	0.0017	0.0020	0.0023
30	100	4	1	0.0187	0.0315	0.0363	0.0420	0.0460	0.0009	0.0010	0.0012	0.0014	0.0020
30	100	4	4	0.0145	0.0185	0.0226	0.0284	0.0454	0.0014	0.0017	0.0020	0.0021	0.0029
30	500	2	1	0.0021	0.0030	0.0048	0.0059	0.0059	0.0006	0.0006	0.0007	0.0008	0.0010
30	500	2	4	0.0005	0.0007	0.0021	0.0026	0.0036	0.0006	0.0006	0.0008	0.0009	0.0012
30	500	4	1	0.0090	0.0231	0.0342	0.0465	0.0941	0.0004	0.0006	0.0006	0.0008	0.0011
30	500	4	4	0.0070	0.0126	0.0168	0.0236	0.0377	0.0004	0.0006	0.0006	0.0007	0.0009
100	100	2	1	0.0053	0.0087	0.0141	0.0173	0.0174	0.0007	0.0008	0.0011	0.0015	0.0019
100	100	2	4	0.0234	0.0343	0.0576	0.0815	0.1044	0.0010	0.0012	0.0013	0.0020	0.0027
100	100	4	1	0.0252	0.0330	0.0402	0.0568	0.0934	0.0006	0.0012	0.0013	0.0016	0.0020
100	100	4	4	0.0108	0.0143	0.0191	0.0255	0.0339	0.0015	0.0025	0.0029	0.0037	0.0053
100	500	2	1	0.0082	0.0086	0.0130	0.0173	0.0174	0.0003	0.0004	0.0006	0.0010	0.0013
100	500	2	4	0.0015	0.0028	0.0051	0.0096	0.0119	0.0004	0.0005	0.0007	0.0007	0.0009
100	500	4	1	0.0021	0.0048	0.0072	0.0095	0.0202	0.0004	0.0006	0.0007	0.0008	0.0009
100	500	4	4	0.0151	0.0281	0.0316	0.0399	0.0959	0.0005	0.0008	0.0011	0.0015	0.0035

Table 8: Numerical standard errors for the persistence parameters in the factors from 25 randomly chosen converged sequences. Instead of reporting all PK^2 parameters per model, we only report the 5%, 25%, 50%, 75% and 95% quantile. The left five columns show the results obtained under PLT identification, and the right five columns show the results obtained under WOP identification.

	WOP		PLT	
1	0.0343	(0.0101)	0.2114	(0.1350)
2	0.0350	(0.0110)	0.0978	(0.0973)
3	0.0347	(0.0103)	0.1394	(0.1410)
4	0.0368	(0.0109)	0.1260	(0.1384)
5	0.0373	(0.0113)	0.1384	(0.1364)
6	0.0344	(0.0108)	0.0349	(0.0105)
7	0.0336	(0.0105)	0.1351	(0.1359)
8	0.0341	(0.0105)	0.0410	(0.0129)
9	0.0340	(0.0108)	0.1400	(0.0942)
10	0.0362	(0.0111)	0.1147	(0.0871)
11	0.0366	(0.0111)	0.1572	(0.1432)
12	0.0349	(0.0103)	0.0479	(0.0165)
13	0.0348	(0.0108)	0.1426	(0.1447)
14	0.0343	(0.0106)	0.1662	(0.1312)
15	0.0376	(0.0095)	0.1559	(0.1501)
16	0.0352	(0.0107)	0.1535	(0.1663)
17	0.0372	(0.0113)	0.1955	(0.1027)
18	0.0343	(0.0104)	0.1368	(0.1433)
19	0.0351	(0.0111)	0.0409	(0.0195)
20	0.0352	(0.0108)	0.1324	(0.1125)

Table 9: Standard deviations of posterior densities of λ_{ik}. Data have been permuted to obtain 20 different orderings. Standard deviations over the 480 entries of Λ are given in parentheses.

Figures

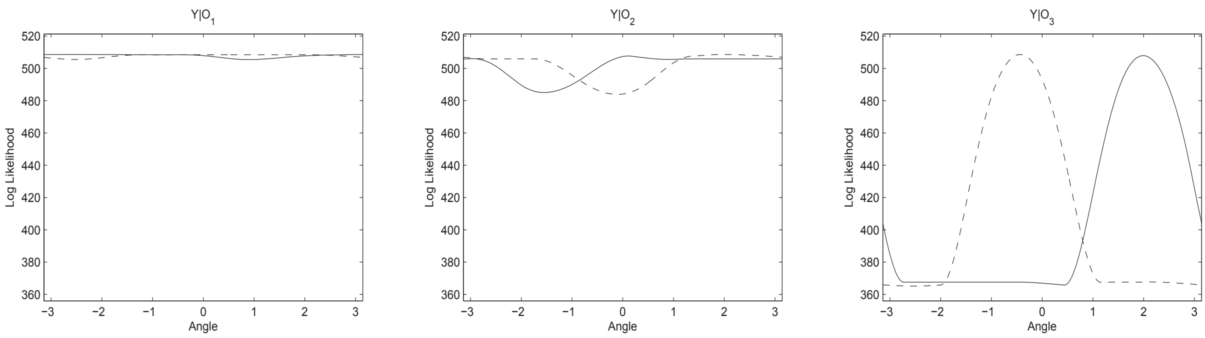

Figure 1: Log likelihood values of the principal components estimates, rotated along the circle, with constraints imposed.

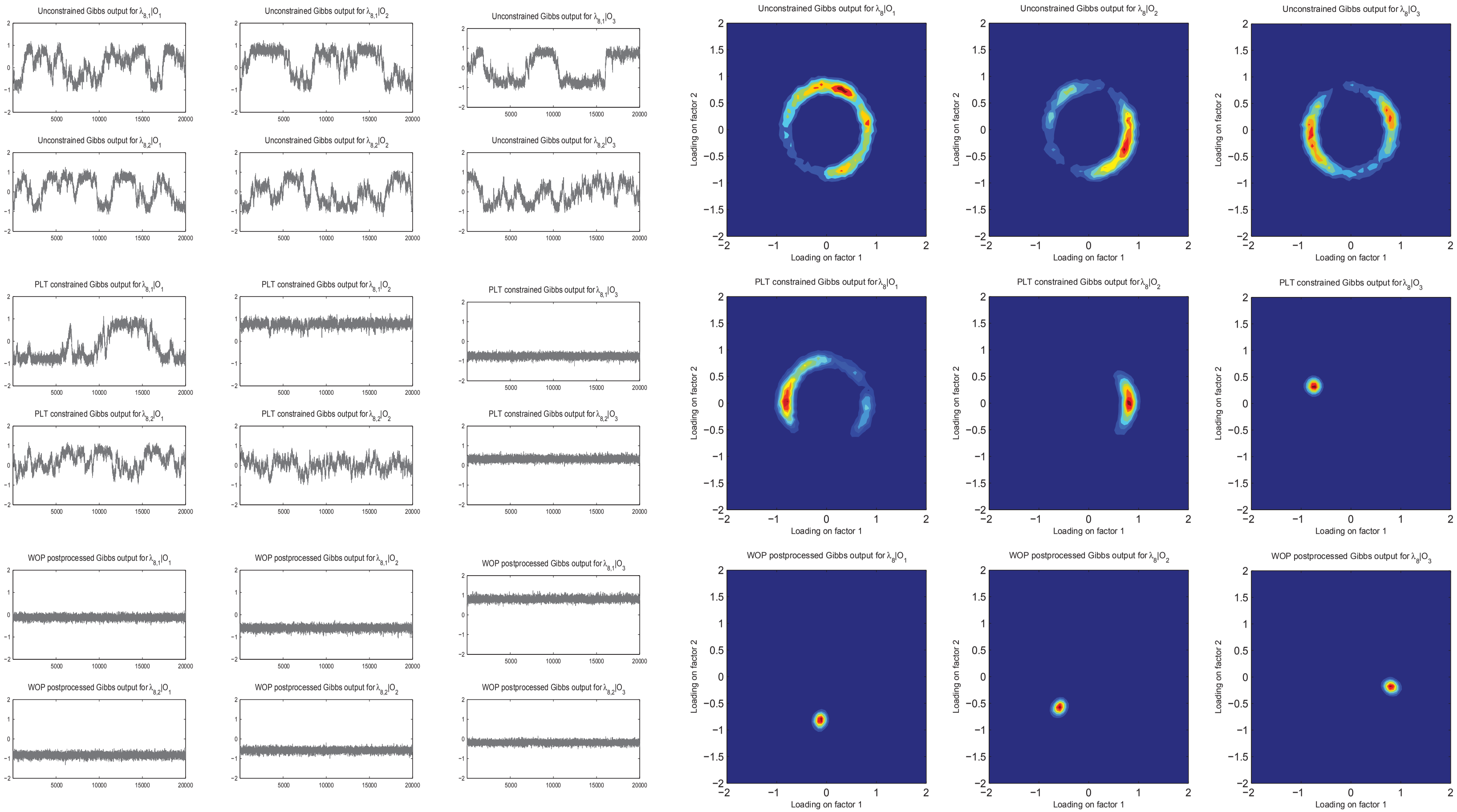

Figure 2: Gibbs sequences and contour plots for bivariate posterior distributions from the unconstrained sampler (top), the PLT constrained sampler (middle), and the unconstrained sampler after five WOP iterations and orientation (bottom). All plots show loadings on cross-section eight, for the three different orderings. Sample size is 20,000 draws, with the preceding 20,000 draws discarded as burn-in.

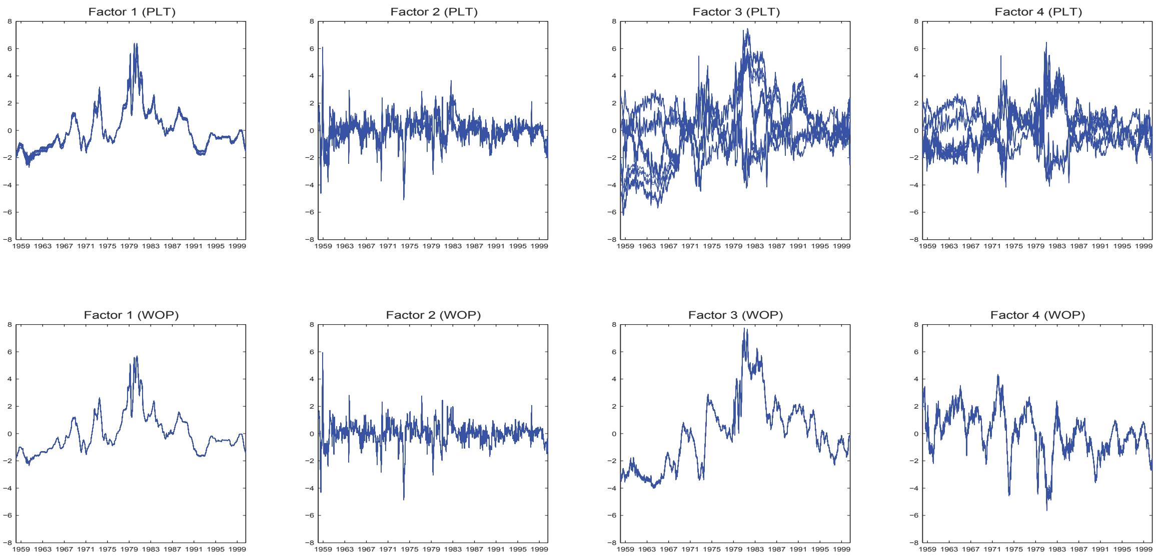

Figure 3: Estimated factors from 120 macroeconomic time series, displaying the results 20 randomly chosen converged sequences. Variables chosen as factor founders are Fed Funds Rate (FYFF), Industrial Production (IP), Monetary Base (FM2), and Consumer Price Index (PMCP). The first row shows the results from the PLT approach, the second row shows the results from the WOP approach, which have been orthogonally transformed to obtain the same tridiagonal loadings structure as PLT.

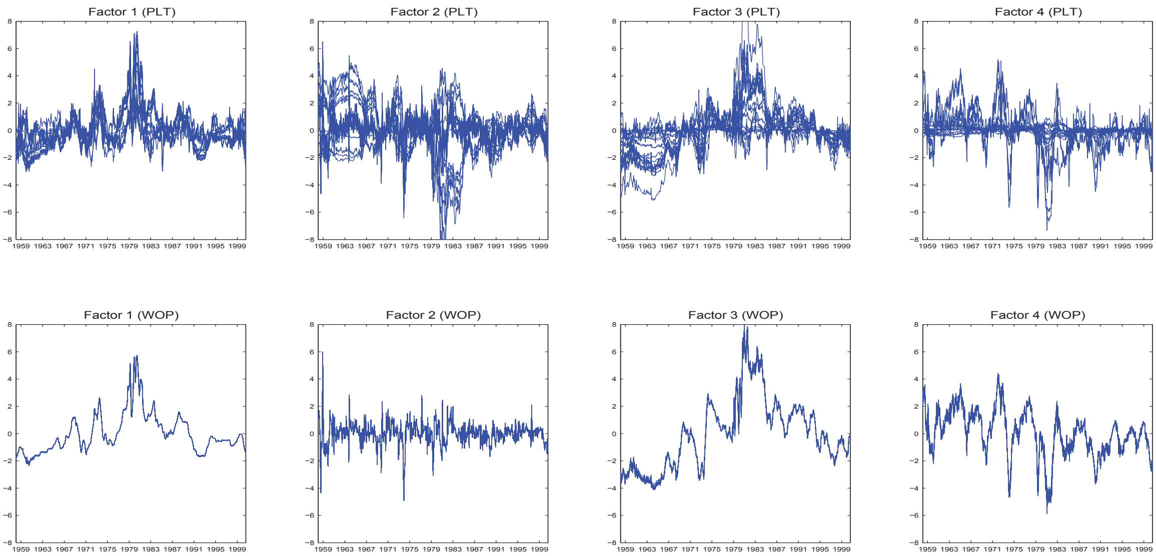

Figure 4: Estimated factors from 120 macroeconomic time series, displaying the results 20 randomly chosen converged sequences. Factor founders have been set randomly, results have afterwards been orthogonally transformed to create a positive lower triangular loadings matrix on the same four variables used as factor founders before, i.e. Fed Funds Rate (FYFF), Industrial Production (IP), Monetary Base (FM2), and Consumer Price Index (PMCP). The first row shows the results from the PLT approach, the second row shows the results from the WOP approach.

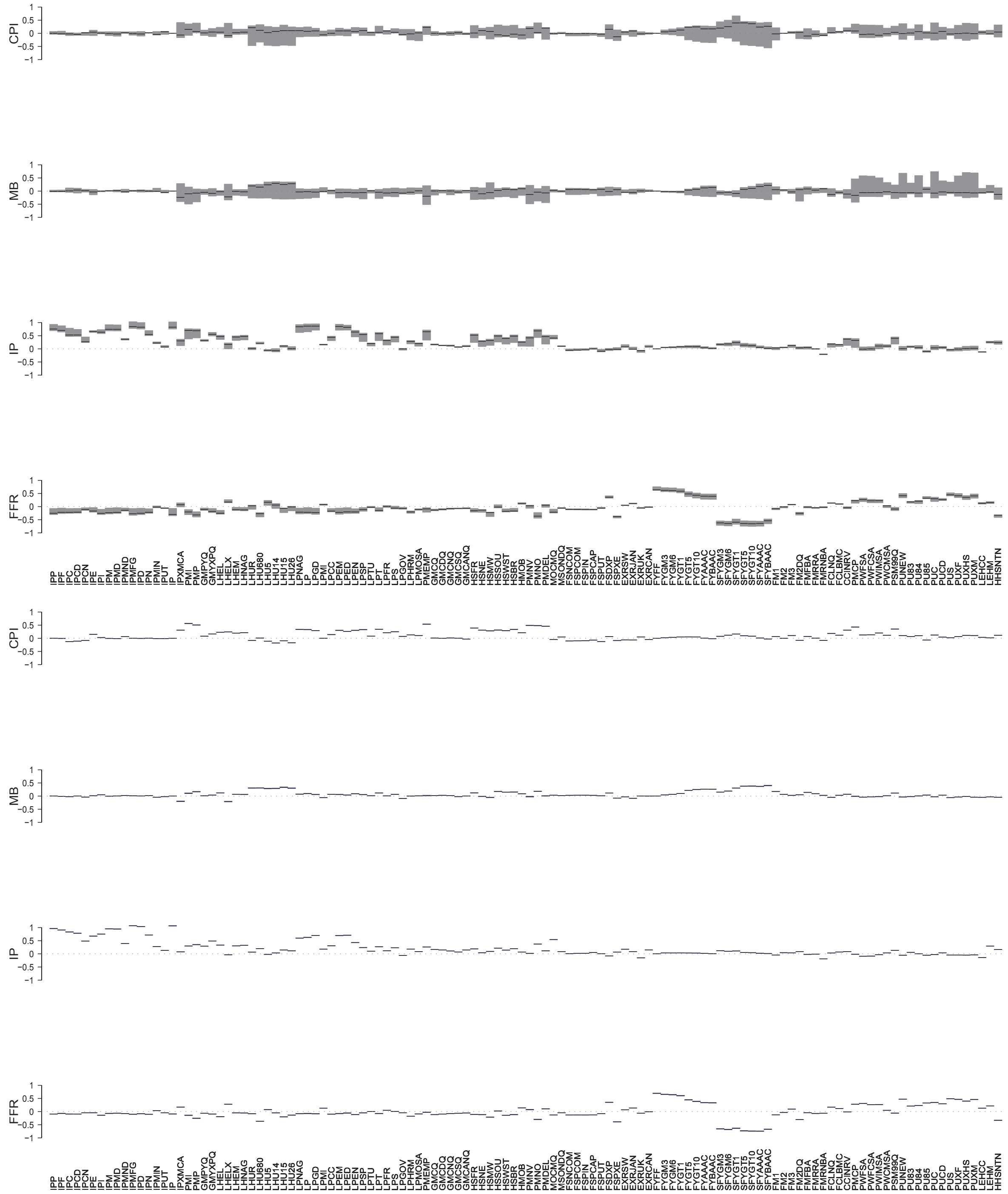

Figure 5: 10, 50 and 90 percent quantiles of the posterior estimates for the factor loadings obtained from 20 randomly chosen converged sequences. Variables chosen as factor founders are Fed Funds Rate (FYFF), Industrial Production (IP), Monetary Base (FM2), and Consumer Price Index (PMCP). The first plot shows the results from the PLT approach, the second plot shows the results from the WOP approach.

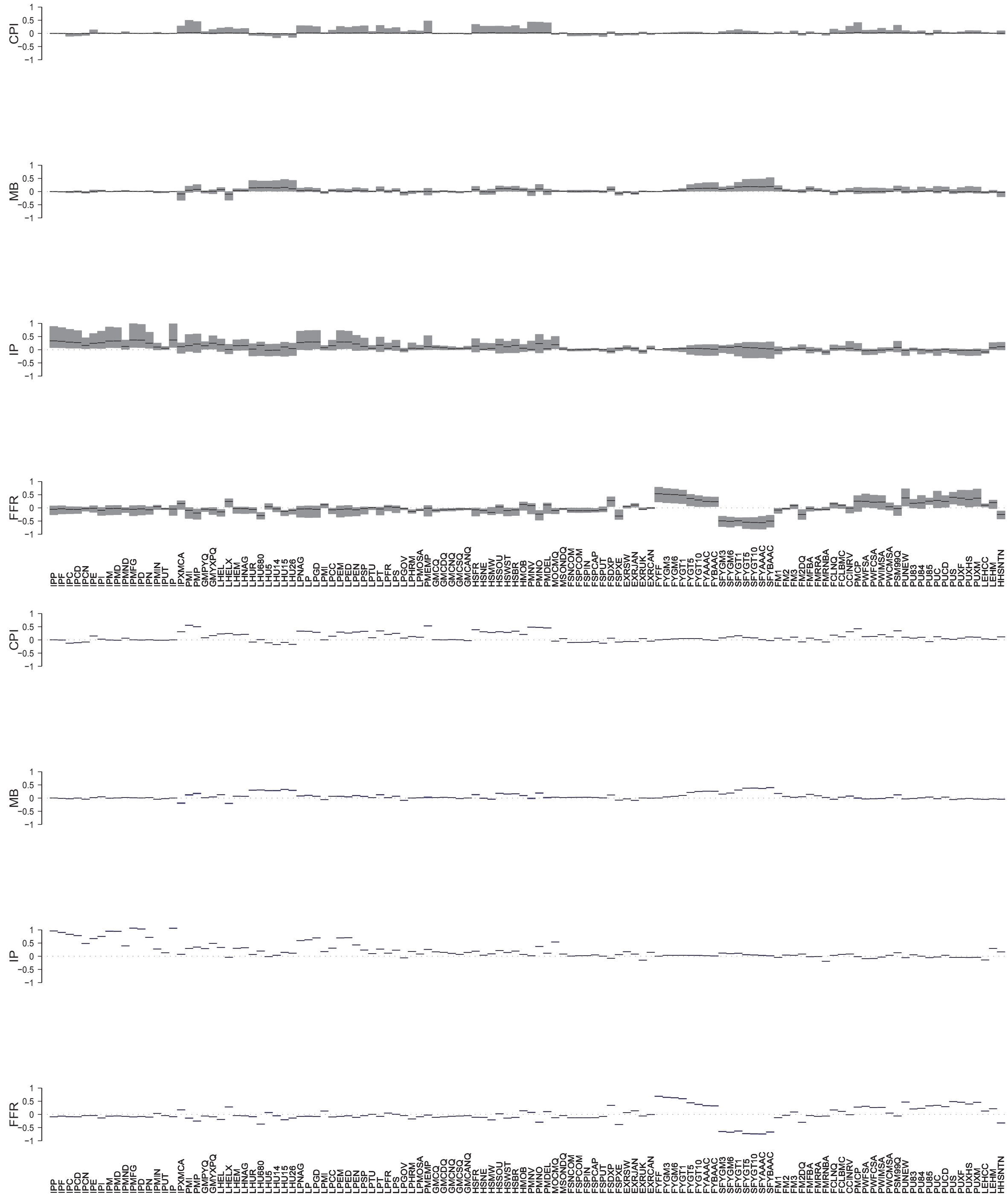

Figure 6: 10, 50 and 90 percent quantiles of the posterior estimates for the factor loadings obtained from 20 randomly chosen converged sequences. Factor founders have been set randomly, results have afterwards been orthogonally transformed to create a positive lower triangular loadings matrix on the same four variables used as factor founders before, i.e. Fed Funds Rate (FYFF), Industrial Production (IP), Monetary Base (FM2), and Consumer Price Index (PMCP). The first plot shows the results from the PLT approach, the second plot shows the results from the WOP approach.

A The unconstrained Gibbs sampler

For the model described in Equations (1) and (2) and prior distributions given in Equations (7) to (9), the unconstrained sampler proceeds by iterative sampling from the corresponding full conditional distributions, see also Bai and Wang (2012).

A.1 Sampling the latent factors by forward-filtering backward-sampling using a square-root Kalman filter

The latent dynamic factors are obtained via forward-filtering backward-sampling, using the ensemble-transform Kalman square-root filter in order to improve the performance of the sampling approach, see Tippett et al. (2003). Let

$$C = \max\{P, S+1\} \tag{32}$$

and define

$$G = \begin{pmatrix} \Phi_1 & \dots & \Phi_{C-1} & \Phi_C \\ I_K & & 0_K & 0_K \\ & \ddots & & \vdots \\ 0_K & & I_K & 0_K \end{pmatrix} \tag{33}$$

as the $CK \times CK$ extended block companion matrix of the latent dynamic factors, where $\Phi_c = 0_K$ for $c > P$,

$$E_t = [\epsilon_t' \ \ 0_{1\times(C-1)K}]' \tag{34}$$

as the vector of error terms in the state equation,

$$Q = \begin{pmatrix} \Psi_\epsilon^{-1} & 0_{K\times(C-1)K} \\ 0_{(C-1)K\times K} & 0_K \end{pmatrix} \tag{35}$$

as the corresponding covariance matrix, and

$$F_t = [f_t', \dots, f_{t-C}']' \tag{36}$$

as a vector of stacked latent factors containing the contemporary factors and C lags.[16] The state equation of the model then is obtained as

$$F_t = GF_{t-1} + E_t. \tag{37}$$

Accordingly, the observation equation is

$$y_t = HF_t + e_t, \tag{38}$$

where

$$H = [\Lambda_0, \dots, \Lambda_C], \tag{39}$$

with $\Lambda_c = 0_{N\times K}$ for $c > S$. With the state estimate at time $t-1$ being $\hat{F}_{t-1|t-1}$, where $\hat{F}_{0|0} = 0_{N\times 1}$, the predicted state at time t is

$$\hat{F}_{t|t-1} = G\hat{F}_{t-1|t-1}, \tag{40}$$

and the prediction covariance is

$$\hat{S}_{t|t-1} = G\hat{S}_{t-1|t-1}G' + Q, \tag{41}$$

where $\hat{S}_{0|0} = I_K$. Taking the observed value y_t into account, we obtain the prediction error

$$u_{t|t-1} = y_t - H\hat{F}_{t|t-1}. \tag{42}$$

The Kalman gain is obtained as

$$K_t = \hat{S}_{t|t-1}H'(H\hat{S}_{t|t-1}H + \Sigma)^{-1}, \tag{43}$$

hence the updating of the covariance matrix can be written as

$$\hat{S}_{t|t} = (I - K_tH)\hat{S}_{t|t-1}. \tag{44}$$

For the according updating step of the ETKF, we first perform a singular-value decomposition of $\hat{S}_{t|t-1}$ as

$$\hat{S}_{t|t-1} = A_{t|t-1}Z_{t|t-1}A'_{t|t-1}, \tag{45}$$

[16]Assume that $f_t = 0_{K\times 1}$ for $t \leq 0$ throughout.

and define the square root of the prediction covariance as

$$Z_t^f = A_{t|t-1} Z_{t|t-1}^{\frac{1}{2}}. \tag{46}$$

Considering the according singular-value decomposition of the innovation covariance matrix as

$$\hat{S}_{t|t} = A_{t|t} Z_{t|t} A'_{t|t}, \tag{47}$$

the corresponding square root can be defined as

$$Z_t^a = A_{t|t} Z_{t|t}^{\frac{1}{2}}. \tag{48}$$

The result from Equation (46) can be inserted into Equation (44) to obtain

$$\hat{S}_{t|t} = Z_t^f (I - Z_t^{f\prime} H' (H Z_t^f Z_t^{f\prime} H' + \Sigma)^{-1} H Z_t^f) Z_t^{f\prime}, \tag{49}$$

hence obtaining a square root of the term in parentheses by an according singular value decomposition of an equivalent expression by the Sherman-Morrison Woodbury identity,

$$(I + Z_t^{f\prime} H' \Sigma^{-1} H Z_t^f)^{-1} = B_t \Gamma_t B'_t, \tag{50}$$

or, equivalently,

$$I + Z_t^{f\prime} H' \Sigma^{-1} H Z_t^f = B_t \Gamma_t^{-1} B'_t. \tag{51}$$

The required square root is then

$$M_t = B_t \Gamma_t^{-\frac{1}{2}}, \tag{52}$$

allowing for the square-root updating as

$$Z_t^a = Z_t^f M_t. \tag{53}$$

Then the innovation covariance matrix can be rebuilt as

$$\hat{S}_{t|t} = Z_t^a Z_t^{a\prime}, \tag{54}$$

and the updated mean is

$$\hat{F}_{t|t} = \hat{F}_{t|t-1} + \hat{S}_{t|t} H \Sigma^{-1} u_{t|t-1}. \tag{55}$$

The factors are then obtained by backward-sampling from the resulting $\hat{F}_{t|T}$ and $\hat{S}_{t|T}$.

A.2 The remaining parameters

Throughout the paper, we assume diagonality for Σ resulting in

$$f(\Sigma|Y, \{\Lambda_s\}_{s=0}^S, \{\Phi_p\}_{p=1}^P, \{f_t\}_{t=1}^T) = \prod_{i=1}^N \frac{\beta_i^{\alpha_i}}{\Gamma(\alpha_i)} \left(\frac{1}{\sigma_i^2}\right)^{\alpha_i - 1} \exp\left\{-\frac{1}{\sigma_i^2}\beta_i\right\}, \tag{56}$$

where $\alpha_i = \frac{1}{2}T + \alpha_{0i}$ and $\beta_i = \frac{1}{2}\sum_{t=1}^T (y_{it} - \sum_{s=0}^S \lambda'_{s,i} f_{t-s})^2 + \beta_{0i}$ and $\alpha_{0i} = \beta_{0i} = 1$ for all $i = 1\ldots, N$. Due to diagonality of Σ, the full conditional distribution of the loadings can be factorized over the $S+1$ Λ_s matrices, and row-wise within these matrices, taking the N individual rows $\lambda_{s,i}$ per matrix into account. This yields the following full conditional distribution:

$$f(\{\Lambda_s\}_{s=0}^S|Y, \Sigma, \{\Phi_p\}_{p=1}^P, \{f_t\}_{t=1}^T) = \prod_{s=0}^S \prod_{i=1}^N (2\pi)^{-\frac{K}{2}} |\Omega_{\lambda_{s,i}}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\lambda_{s,i} - \mu_{\lambda_{s,i}})' \Omega_{\lambda_{s,i}}^{-1} (\lambda_{s,i} - \mu_{\lambda_{s,i}})\right\}, \tag{57}$$

where $\Omega_{\lambda_{s,i}} = (\frac{1}{\sigma_i^2}\sum_{t=1}^T f_{t-s} f'_{t-s} + (\Upsilon_s)_{i,i} I_K)^{-1}$ and $\mu_{\lambda_{s,i}} = \Omega_{\lambda_{s,i}}(\frac{1}{\sigma_i^2}\sum_{t=1}^T y_{it} f'_{t-s})$.
Finally, consider a stacked version of the persistence parameters for the factors,

$$\tilde{\Phi} = [\Phi'_1, \ \ldots, \ \Phi'_P]' \tag{58}$$

and denote a shortened $T - P \times K$ factor matrix starting at time point t as

$$\tilde{F}_t = [f_t, \ \ldots, \ f_{T-P+(t-1)}]', \tag{59}$$

and

$$\tilde{F} = [\tilde{F}_1, \ \ldots, \ \tilde{F}_P] \tag{60}$$

containing P such matrices. Then the full conditional distribution of $\tilde{\Phi}$ for normally distributed innovations in the factors and with an uninformative prior distribution obtains as

$$f(\tilde{\Phi}|Y, \Sigma, \{\Lambda_s\}_{s=0}^S, \{f_t\}_{t=1}^T) = (2\pi)^{-\frac{KP}{2}} |\Omega_{\tilde{\Phi}}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\text{vec}(\tilde{\Phi}) - \mu_{\tilde{\Phi}})' \Omega_{\tilde{\Phi}}{}^{-1} (\text{vec}(\tilde{\Phi}) - \mu_{\tilde{\Phi}})\right\}, \tag{61}$$

where $\Omega_{\tilde{\Phi}} = \Psi_\epsilon^{-1} \otimes (\tilde{F}'\tilde{F})^{-1}$ and $\mu_{\tilde{\Phi}} = \text{vec}((\tilde{F}'\tilde{F})^{-1}\tilde{F}'\tilde{F}_{P+1})$, see e.g. Ni and Sun (2005). It is easy to see that a parameter expansion of $\tilde{F}_t$ in terms of the orthogonal parameter D as described at the end of section 2 results in

$$\tilde{F}_t(D) = \tilde{F}_t D, \tag{62}$$

and, consequently,

$$\tilde{F}(D) = \tilde{F}\left(I_P \otimes D\right), \tag{63}$$

such that

$$\Omega_{\tilde{\Phi}}(D) = \Psi_\epsilon^{-1} \otimes (\tilde{F}\left(I_P \otimes D\right))'(\tilde{F}\left(I_P \otimes D\right))^{-1} = \Omega_{\tilde{\Phi}}, \tag{64}$$

and

$$\begin{aligned}\mu_{\tilde{\Phi}}(D) &= ((\tilde{F}\left(I_P \otimes D\right))'(\tilde{F}\left(I_P \otimes D\right)))^{-1}(\tilde{F}\left(I_P \otimes D\right))'(\tilde{F}_{P+1} D) \\ &= \left(I_P \otimes D\right)' \mu_{\tilde{\Phi}} D. \end{aligned} \tag{65}$$

B Proof of Proposition 3.2

Proof. Given a parametrization of D ensuring orthogonality the minimization problem in Equation (23) can be restated as

$$\begin{aligned} D &= \arg\min(L_1(\Theta^*, \Theta^{(r)}) + L_2(\Theta^*, \Theta^{(r)})) \\ &= \arg\max \operatorname{tr}(D'\bar{\Lambda}^{(r)\prime}\bar{\Lambda}^*) + \operatorname{tr}\left(\sum_{p=1}^{P} D'\Phi_p^{(r)\prime} D \Phi_p^*\right). \end{aligned}$$

To start with, let $K = 2$ and look at $\operatorname{tr}(D'\bar{\Lambda}^{(r)\prime}\bar{\Lambda}^*)$ first. Define

$$M = \bar{\Lambda}^{(r)\prime}\bar{\Lambda}^*, \tag{66}$$

and assume the parametrization $D = D_+$, i.e.

$$D_+ = \begin{pmatrix} \cos(\gamma_+) & -\sin(\gamma_+) \\ \sin(\gamma_+) & \cos(\gamma_+) \end{pmatrix}. \tag{67}$$

Then D_+ can be expressed in terms of an angle $\gamma_+ \in [-\pi, \pi)$ resulting in

$$\operatorname{tr}(D_+'M) = \operatorname{tr}\begin{pmatrix} m_{11}\cos(\gamma_+) + m_{21}\sin(\gamma_+) & m_{12}\cos(\gamma_+) + m_{22}\sin(\gamma_+) \\ -m_{11}\sin(\gamma_+) + m_{21}\cos(\gamma_+) & -m_{12}\sin(\gamma_+) + m_{22}\cos(\gamma_+) \end{pmatrix} \tag{68}$$

$$= (m_{11} + m_{22}) \cos(\gamma_+) + (m_{21} - m_{12}) \sin(\gamma_+)$$
$$= \sqrt{(m_{11} + m_{22})^2 + (m_{21} - m_{12})^2} \cos(\gamma_+ - \text{atan2}(m_{11} + m_{22}, m_{21} - m_{12}))$$
$$= A_+ \cos(\gamma_+ + \varphi_+), \tag{69}$$

which is a sinusoid, see e.g. Shumway and Stoffer (2010, Chapter 4.2.), with amplitude $A_+ = \sqrt{(m_{11} + m_{22})^2 + (m_{21} - m_{12})^2}$, phase $\varphi_+ = -\text{atan2}(m_{11}+m_{22}, m_{21}-m_{12})$, and frequency $\omega = \frac{1}{2\pi}$, i.e. there is exactly one maximum in the domain of γ for any choice of D_+.[17] Note that (69) uses the important equality

$$A \cos(\omega t + \varphi) = \sum_{i=1}^{n} A_i \cos(\omega t + \varphi_i), \tag{71}$$

where $A = \sqrt{\left(\sum_{i=1}^n A_i \cos(\varphi_i)\right)^2 + \left(\sum_{i=1}^n A_i \sin(\varphi_i)\right)^2}$ and $\varphi = \text{atan2}\left(\sum_{i=1}^n A_i \cos(\varphi_i), \sum_{i=1}^n A_i \sin(\varphi_i)\right)$, for which a proof can be found e.g. in Smith (2007). Next, consider the parametrization $D = D_-$. The resulting matrix for $K = 2$ is hence

$$D_- = \begin{pmatrix} \cos(\gamma_-) & \sin(\gamma_-) \\ \sin(\gamma_-) & -\cos(\gamma_-) \end{pmatrix}. \tag{72}$$

Then D_- can again be expressed in terms of an angle $\gamma_- \in [-\pi, \pi)$ resulting in

$$\text{tr}(D'_- M) = \text{tr} \begin{pmatrix} m_{11} \cos(\gamma_-) + m_{21} \sin(\gamma_-) & m_{12} \cos(\gamma_-) + m_{22} \sin(\gamma_-) \\ m_{11} \sin(\gamma_-) - m_{21} \cos(\gamma_-) & m_{12} \sin(\gamma_-) - m_{22} \cos(\gamma_-) \end{pmatrix} \tag{73}$$
$$= (m_{11} - m_{22}) \cos(\gamma_-) + (m_{12} + m_{21}) \sin(\gamma_-)$$
$$= \sqrt{(m_{11} - m_{22})^2 + (m_{12} + m_{21})^2} \cos(\gamma_- - \text{atan2}(m_{11} - m_{22}, m_{12} + m_{21}))$$
$$= A_- \cos(\gamma_- + \varphi_-), \tag{74}$$

which is also a sinusoid, but with amplitude $A_- = \sqrt{(m_{11} - m_{22})^2 + (m_{12} + m_{21})^2}$, phase $\varphi_- = -\text{atan2}(m_{11} - m_{22}, m_{12} + m_{21})$, and frequency $\omega = \frac{1}{2\pi}$. Thus, γ_- and γ_+ are uniquely identified if A_- and φ_- or A_+ and φ_+ respectively are all distinct from zero. Note that the events $m_{11} - m_{22} = 0$ and $m_{21} + m_{12} = 0$ or $m_{11} + m_{22} = 0$ and $m_{21} - m_{12} = 0$ corresponding to A_- and φ_- being zero or A_+ and φ_+ being zero respectively occur with probability zero since the corresponding restrictions on $\Theta^{(r)}$ and Θ^* denote a subspace of the parameter space. Further, the two maxima implied by γ_-

[17] The two-argument arctangent function atan2(y, x), defined on the interval $[-\pi, \pi)$ and based on a half-angle identity for the tangent, is given as

$$\text{atan2}(y, x) = 2\text{arctan} \frac{\sqrt{x^2 + y^2} - x}{y}. \tag{70}$$

and γ_+ are distinct with probability one since the event $A_- = A_+$ occurs as well with probability zero.

Now look at $\mathrm{tr}\left(\sum_{p=1}^{P} D'\Phi_p^{(r)'} D\Phi_p^*\right)$ and let $P = 1$. Assuming $D = D_+$ yields

$$\begin{aligned}\mathrm{tr}(D'\Phi^{(r)'}D\Phi^*) &= \phi_{12}^*(k_1(\gamma_+) + k_2(\gamma_+)) + \phi_{11}^*(k_3(\gamma_+) + k_4(\gamma_+)) \\ &\quad + \phi_{22}^*(k_5(\gamma_+) + k_6(\gamma_+)) + \phi_{21}^*(k_7(\gamma_+) + k_8(\gamma_+)),\end{aligned} \tag{75}$$

with

$$k_1(\gamma_+) = -\cos(\gamma_+)(\phi_{12}^{(r)}\cos(\gamma_+) + \phi_{11}^{(r)}\sin(\gamma_+)) = -\phi_{12}^{(r)}\cos^2(\gamma_+) + \phi_{11}^{(r)}\sin(\gamma_+)\cos(\gamma_+) \tag{76}$$
$$k_2(\gamma_+) = -\sin(\gamma_+)(\phi_{22}^{(r)}\cos(\gamma_+) + \phi_{21}^{(r)}\sin(\gamma_+)) = -\phi_{21}^{(r)}\sin^2(\gamma_+) - \phi_{22}^{(r)}\sin(\gamma_+)\cos(\gamma_+) \tag{77}$$
$$k_3(\gamma_+) = -\cos(\gamma_+)(\phi_{11}^{(r)}\cos(\gamma_+) - \phi_{12}^{(r)}\sin(\gamma_+)) = -\phi_{11}^{(r)}\cos^2(\gamma_+) - \phi_{12}^{(r)}\sin(\gamma_+)\cos(\gamma_+) \tag{78}$$
$$k_4(\gamma_+) = -\sin(\gamma_+)(\phi_{21}^{(r)}\cos(\gamma_+) - \phi_{22}^{(r)}\sin(\gamma_+)) = -\phi_{22}^{(r)}\sin^2(\gamma_+) - \phi_{21}^{(r)}\sin(\gamma_+)\cos(\gamma_+) \tag{79}$$
$$k_5(\gamma_+) = -\cos(\gamma_+)(\phi_{22}^{(r)}\cos(\gamma_+) + \phi_{21}^{(r)}\sin(\gamma_+)) = -\phi_{22}^{(r)}\cos^2(\gamma_+) + \phi_{21}^{(r)}\sin(\gamma_+)\cos(\gamma_+) \tag{80}$$
$$k_6(\gamma_+) = -\sin(\gamma_+)(\phi_{12}^{(r)}\cos(\gamma_+) + \phi_{11}^{(r)}\sin(\gamma_+)) = -\phi_{11}^{(r)}\sin^2(\gamma_+) + \phi_{12}^{(r)}\sin(\gamma_+)\cos(\gamma_+) \tag{81}$$
$$k_7(\gamma_+) = -\cos(\gamma_+)(\phi_{21}^{(r)}\cos(\gamma_+) - \phi_{22}^{(r)}\sin(\gamma_+)) = -\phi_{21}^{(r)}\cos^2(\gamma_+) - \phi_{22}^{(r)}\sin(\gamma_+)\cos(\gamma_+) \tag{82}$$
$$k_8(\gamma_+) = -\sin(\gamma_+)(\phi_{11}^{(r)}\cos(\gamma_+) - \phi_{12}^{(r)}\sin(\gamma_+)) = -\phi_{12}^{(r)}\sin^2(\gamma_+) + \phi_{11}^{(r)}\sin(\gamma_+)\cos(\gamma_+) \tag{83}$$

Consider Equations (77), (79), (81) and (83) and obtain

$$\begin{aligned} k_2 &= -\phi_{21}^{(r)}\sin^2(\gamma_+) - \phi_{22}^{(r)}\sin(\gamma_+)\cos(\gamma_+) \\ &= -\phi_{21}^{(r)}(1-\cos^2(\gamma_+)) - \phi_{22}^{(r)}\sin(\gamma_+)\cos(\gamma_+) = k_7 - \phi_{21}^{(r)} \end{aligned} \tag{84}$$
$$\begin{aligned} k_4 &= -\phi_{22}^{(r)}\sin^2(\gamma_+) - \phi_{21}^{(r)}\sin(\gamma_+)\cos(\gamma_+) \\ &= -\phi_{22}^{(r)}(1-\cos^2(\gamma_+)) - \phi_{21}^{(r)}\sin(\gamma_+)\cos(\gamma_+) = \phi_{22}^{(r)} - k_5 \end{aligned} \tag{85}$$
$$\begin{aligned} k_6 &= -\phi_{11}^{(r)}\sin^2(\gamma_+) + \phi_{12}^{(r)}\sin(\gamma_+)\cos(\gamma_+) \\ &= -\phi_{11}^{(r)}(1-\cos^2(\gamma_+)) + \phi_{12}^{(r)}\sin(\gamma_+)\cos(\gamma_+) = \phi_{11}^{(r)} - k_3 \end{aligned} \tag{86}$$
$$\begin{aligned} k_8 &= -\phi_{12}^{(r)}\sin^2(\gamma_+) + \phi_{11}^{(r)}\sin(\gamma_+)\cos(\gamma_+) \\ &= -\phi_{12}^{(r)}(1-\cos^2(\gamma_+)) + \phi_{11}^{(r)}\sin(\gamma_+)\cos(\gamma_+) = k_1 - \phi_{12}^{(r)} \end{aligned} \tag{87}$$

Inserting (76), (78), (80), (82), (84), (85), (86) and (87) into (75) yields

$$\begin{aligned}\mathrm{tr}(D'\Phi^{(r)'}D\Phi^*) &= \phi_{12}^*(k_1 + k_7 - \phi_{21}^{(r)}) + \phi_{11}^*(k_3 - k_5 + \phi_{22}^{(r)}) \\ &\quad + \phi_{22}^*(k_5 - k_3 + \phi_{11}^{(r)}) + \phi_{21}^*(k_7 + k_1 - \phi_{12}^{(r)}) \\ &= \underbrace{-\phi_{12}^*\phi_{21}^{(r)} - \phi_{11}^*\phi_{22}^{(r)} + \phi_{22}^*\phi_{11}^{(r)} - \phi_{21}^*\phi_{12}^{(r)}}_{=c_0} + \end{aligned}$$

$$
\begin{aligned}
&(\phi^*_{12}+\phi^*_{21})(k_1+k_7)+(\phi^*_{11}-\phi^*_{22})(k_3-k_5)\\
&=c_0+(\phi^*_{12}+\phi^*_{21})((\phi^{(r)}_{12}+\phi^{(r)}_{21})\cos^2(\gamma_+)+(\phi^{(r)}_{11}-\phi^{(r)}_{22})\sin(\gamma_+)\cos(\gamma_+))\\
&\quad\;+(\phi^*_{11}-\phi^*_{22})((\phi^{(r)}_{11}-\phi^{(r)}_{22})\cos^2(\gamma_+)-(\phi^{(r)}_{12}+\phi^{(r)}_{21})\sin(\gamma_+)\cos(\gamma_+))\\
&=c_0+\underbrace{((\phi^*_{12}+\phi^*_{21})(\phi^{(r)}_{12}+\phi^{(r)}_{21})+(\phi^*_{11}-\phi^*_{22})(\phi^{(r)}_{11}-\phi^{(r)}_{22}))}_{=c_1}\cos^2(\gamma_+)\\
&\quad\;+\underbrace{((\phi^*_{12}+\phi^*_{21})(\phi^{(r)}_{11}-\phi^{(r)}_{22})-(\phi^*_{11}-\phi^*_{22})(\phi^{(r)}_{12}+\phi^{(r)}_{21}))}_{=c_2}\sin(\gamma_+)\cos(\gamma_+)\\
&=c_0+(c_1\cos(\gamma_+)+c_2\sin(\gamma_+))\cos(\gamma_+)\\
&=c_0+(\underbrace{\sqrt{c_1^2+c_2^2}}_{=c_3}\cos(\gamma_++\underbrace{\operatorname{atan2}(c_1,c_2)}_{=c_4}))\cos(\gamma_+)\\
&=c_0+\frac{1}{2}c_3\cos(c_4)+\frac{1}{2}c_3\cos(2\gamma_+-c_4)\\
&=V_++A_+\cos(2\gamma_++\varphi_+),
\end{aligned}
\tag{88}
$$

where c_0 through c_4 are constant terms, and the second-last equality uses the fact that

$$
\cos(\gamma_1)\cos(\gamma_2)=\frac{1}{2}\left(\cos(\gamma_1-\gamma_2)+\cos(\gamma_1+\gamma_2)\right). \tag{89}
$$

The result of (88) is sinusoid with vertical shift $V_+=c_0+\frac{1}{2}c_3\cos(c_4)$, amplitude $A_+=\frac{1}{2}c_3$, phase $\varphi_+=-c_4$, and frequency $\omega=\frac{1}{\pi}$, i.e. there are exactly two maxima in the domain of γ for any choice of D_+. The equivalent result for $D=D_-$ obtains analogously. For $P>1$, reversing the order of summation and trace operator in $\operatorname{tr}\left(\sum_{p=1}^P D'\Phi_p^{(r)'}D\Phi_p^*\right)$, we obtain P such sinusoids, which all depend on the same γ_+, thus we can apply Equation (71) to the demeaned sinusoids and afterwards add the sum of the means again, another vertically shifted sinusoid with frequency $\frac{1}{\pi}$ and thus two maxima in the domain of γ_+. Note that the choice of $D=D_+$ yields the superposition of P sinusoids, whereas the choice of $D=D_-$ yields another superposition of P sinusoids, however with according changes in the phase, amplitude and vertical shift parameters. Although for D_+ as well as for D_- we find two maxima each, the maxima under the two parametrization are distinct with probability one as the restrictions on the parameter space causing coincidence of the two sets of maxima under the two parametrizations of the orthogonal matrix refer to a subspace of the parameter space having thus probability zero. Further using the same line of argument as above, for each of the parametrizations there exist two maxima with probability one.

To show the uniqueness of the maximum of $\operatorname{tr}(D'\bar{\Lambda}^{(r)'}\bar{\Lambda}^*)+\operatorname{tr}\left(\sum_{p=1}^P D'\Phi_p^{(r)'}D\Phi_p^*\right)$ we must consequently consider for both cases, $D=D_+$ and $D=D_-$, a superposition of two sinusoids with frequency $\frac{1}{2\pi}$ and $\frac{1}{\pi}$, respectively. The first of them has one peak and one trough on the interval $[-\pi,\pi)$, while the second has two of each. The sum over these two sinusoids has two peaks of identical height if and only if the peak of the first coincides with one of the two troughs of the second. Denoting the phase of $\text{tr}(D'\bar{\Lambda}^{(r)\prime}\bar{\Lambda}^*)$ as φ_Λ and the phase of tr $\left(\sum_{p=1}^{P} D'\Phi_p^{(r)\prime} D\Phi_p^*\right)$ as φ_Φ this implies the strict equality $\varphi_\Phi = \pi + 2\varphi_\Lambda$ corresponding to a restriction of the parameter space having probability zero.

Now consider the general case for $K > 2$. To derive the structure of the expression $\text{tr}(D'\bar{\Lambda}^{(r)\prime}\bar{\Lambda}^*)$, look at the matrix D first. D can be expressed as the product over $K(K-1)/2$ Givens rotation matrices and a reflection about the K^{th} axis. For the time being, the reflection is not considered. The Givens rotation matrices are functions in the angles $\underline{\gamma} = (\gamma_1, \ldots, \gamma_{\frac{K(K-1)}{2}})$. Thus, defining the constituent set of elements

$$\mathcal{CS} \quad = \quad \left\{\cos(\gamma_{k^*}), \sin(\gamma_{k^*}) = \cos\left(\gamma_{k^*} - \frac{\pi}{2}\right)\right\}_{k^*=1}^{K(K-1)/2}, \tag{90}$$

each entry of D can be characterized as

$$d_{ij} \quad = \quad \sum_{j^*=1}^{T_{ij}} a_{j^*}^{ij} \prod_{k^*=1}^{K(K-1)/2} \cos(\gamma_{k^*})^{b_{j^*k^*}^{ij}} \cos\left(\gamma_{k^*} - \frac{\pi}{2}\right)^{c_{j^*k^*}^{ij}}, \tag{91}$$

with T_{ij} denoting the number of subsets involved in d_{ij}, $a_{j^*}^{ij} \in \{-1, 1\}$, $b_{j^*k^*}^{ij}$ and $c_{j^*k^*}^{ij}$ taking either values 0 or 1, and $b_{j^*k^*}^{ij} + c_{j^*k^*}^{ij} \leq 1$. Then

$$\text{tr}(D'\bar{\Lambda}^{(r)\prime}\bar{\Lambda}^*) \quad = \quad \sum_{j=1}^{K} d'_{\cdot j}(\bar{\Lambda}^{(r)\prime}\bar{\Lambda}^*)_{\cdot j}, \tag{92}$$

where $D_{\cdot j}$ denotes the j^{th} column of D and $(\bar{\Lambda}^{(r)\prime}\bar{\Lambda}^*)_{\cdot j}$ denotes the j^{th} column of $\bar{\Lambda}^{(r)\prime}\bar{\Lambda}^*$. The same expression can also be stated in the structural form from Equation (91), hence

$$\text{tr}(D'\bar{\Lambda}^{(r)\prime}\bar{\Lambda}^*) \quad = \quad \sum_{j^*=1}^{T_{\text{tr}\bar{\Lambda}}} q_{j^*}^{\text{tr}\bar{\Lambda}} \prod_{k^*=1}^{K(K-1)/2} \cos(\gamma_{k^*})^{b_{j^*k^*}^{\text{tr}\bar{\Lambda}}} \cos\left(\gamma_{k^*} - \frac{\pi}{2}\right)^{c_{j^*k^*}^{\text{tr}\bar{\Lambda}}}, \tag{93}$$

where $T_{\text{tr}\bar{\Lambda}}$ denotes the number of subsets entering $\text{tr}(D'\bar{\Lambda}^{(r)\prime}\bar{\Lambda}^*)$, $q_{j^*}^{\text{tr}\bar{\Lambda}}$ is a function of the $a_{j^*}^{ij}$ and the elements in the matrix $\bar{\Lambda}^{(r)\prime}\bar{\Lambda}^*$, and $b_{j^*k^*}^{\text{tr}\bar{\Lambda}}$ and $c_{j^*k^*}^{\text{tr}\bar{\Lambda}}$ taking either values 0 or 1, and $b_{j^*k^*}^{\text{tr}\bar{\Lambda}} + c_{j^*k^*}^{\text{tr}\bar{\Lambda}} \leq 1$. It can be seen from Equation (92) that each subset involved in any d_{ij} enters Equation (93), which is hence a weighted sum over the union of products of all subsets of $\mathcal{CS}$ involved in D.

A weighted one-element subset of $\mathcal{CS}$ is a sinusoid with frequency $\frac{1}{2\pi}$, as discussed for $K = 2$. Since the γ_{k^*} are all mutually independent, the multiple-element subsets of $\mathcal{CS}$ are therefore sinusoids with the same frequency along each dimension and dimensionality not larger than $K(K-1)/2$. $\text{tr}(D'\bar{\Lambda}^{(r)\prime}\bar{\Lambda}^*)$ is then the superposition of $T_{\text{tr}\bar{\Lambda}}$ such sinusoids. In fact, all sinusoids can be treated as $K(K-1)/2$-variate sinusoids, which are constant along the dimensions whose angles they do not depend on. The superposition yields a unique maximum if each γ_{k^*} enters at least one of the sinusoids, otherwise the value of the respective γ_{k^*} is irrelevant for the maximization. Further, it must be ensured that all sinusoids have a unique maximum in the subset of $\underline{\gamma}$ they depend on, or, if this is not the case, the different parametrizations in $\underline{\gamma}$ all imply the same D.

Unlike the univariate sinusoids with frequency $\frac{1}{2\pi}$, however, multivariate ones have multiple maxima, because joint replacements of pairs of elements of $\underline{\gamma}$ can exploit the trigonometric identity

$$\cos(\gamma) \quad = \quad -\cos(\gamma+\pi) = -\cos(\gamma-\pi) = \cos(-\gamma). \tag{94}$$

Without loss of generality, consider the bivariate sinusoid $\cos(\gamma_1)\cos(\gamma_2)$, which has a maximum in $\Gamma = (0,0)$. By Equation (94), the exists a second maximum in $\underline{\gamma} = (\pi,\pi)$. The case of the bivariate sinusoid is also shown in the left panel of Figure 7. Accordingly, a trivariate sinusoid allows for $\binom{3}{2} = 3$ pairwise replacements, and a 4-variate sinusoid allows for $\binom{4}{2} + \binom{4}{4} = 6 + 1 = 7$ replacements, where the second term denotes the replacement of two pairs of angles by their counterparts at the same time. The number of additional redundant parametrizations for a $K(K-1)/2$-variate sinusoid is thus $\sum_{i^*=1}^{\lfloor\frac{K(K-1)}{2}/2\rfloor}\binom{\frac{K(K-1)}{2}}{2i^*}$, implying a total number of modes of $2^{\frac{K(K-1)}{2}-1}$. Note, however, that in order to obtain a redundant parametrization of D, *all* involved sinusoids must allow for the according pairwise replacements. The actual number of modes is therefore usually much smaller than $2^{\frac{K(K-1)}{2}-1}$. Consider e.g. $K = 3$, where the only admissible replacement for $\underline{\gamma} = (\gamma_1, \gamma_2, \gamma_3)$ is $\underline{\tilde{\gamma}} = (\gamma \pm \pi, \pm\pi - \gamma_2, \gamma_3 \pm \pi)$, where the sign of $\pm\pi$ must be chosen such that the angle is in the admissible range for $\underline{\gamma}$. Taking the redundant parametrizations into account, there exists thus a unique orthogonal matrix D providing a maximum for the involved sinusoids and thus a unique D maximizing $\mathrm{tr}(D'\bar{\Lambda}^{(r)\prime}\bar{\Lambda}^*)$.

An expression analogous to the one in Equation (93) can also be found for $\sum_{p=1}^{P}\mathrm{tr}(D'\Phi_p^{(r)\prime}D\Phi_p^*)$. Note that here, it is possible that the $\cos(\gamma_{k^*})$ enter in quadratic form, hence, the resultant sinusoids have frequency $\frac{1}{\pi}$. $\sum_{p=1}^{P}\mathrm{tr}(D'\Phi_p^{(r)\prime}D\Phi_p^*)$ then has the structural form

$$\sum_{j^*=1}^{T_{\mathrm{tr}\Phi}} p_{j^*}^{\mathrm{tr}\Phi} \prod_{k^*=1}^{K(K-1)} \cos(\gamma_{k^*})^{b_{j^*k^*}^{\mathrm{tr}\Phi}} \cos(\gamma_{k^*} - \frac{\pi}{2})^{c_{j^*k^*}^{\mathrm{tr}\Phi}}, \tag{95}$$

with $b_{j^*k^*}^{\mathrm{tr}\Phi}$ and $c_{j^*k^*}^{\mathrm{tr}\Phi}$ taking values $\{0,1,2\}$, $b_{j^*k^*}^{\mathrm{tr}\Phi} + c_{j^*k^*}^{\mathrm{tr}\Phi} \leq 2$, where $p_{j^*}^{\mathrm{tr}\Phi}$ is a function of the elements involved in the matrices $\Phi_p^{(r)}$ and Φ_p^*, $p = 1,\ldots,P$. Hence $\sum_{p=1}^{P}\mathrm{tr}(D'\Phi_p^{(r)\prime}D\Phi_p^*)$ is the sum of sinusoids having frequency $\frac{1}{\pi}$ or $\frac{1}{2\pi}$ along each dimension. Consequently, assuming that all γ_{k^*} enter the expression in Equation (95) at least once, the result is a superposition of $K(K-1)/2$-variate sinusoids, which do not exceed the frequency $\frac{1}{\pi}$ in any dimension. A bivariate sinusoid with frequency $\frac{1}{\pi}$ in each dimension is shown in the right panel of Figure 7. Each dimension where the frequency is doubled necessarily has twice as many maxima. Nonetheless, the number of maxima cannot exceed $2 \cdot 4^{\frac{K(K-1)}{2}-1}$ and is thus finite. Superimposing the sinusoids in $\sum_{j^*=1}^{T_{\mathrm{tr}\Phi}} p_{j^*}^{ij}$ with those in $\mathrm{tr}(D'\bar{\Lambda}^{(r)\prime}\bar{\Lambda}^*)$ thus results in a unique maximum almost surely, where the event that two maxima of the superimposed sinusoids are equally qualified by the sinusoids with lower frequency corresponds to a restriction on the parameter space and hence occurs with probability zero. The same maximization over $\underline{\gamma}$, but involving a reflection over the K^{th} axis, yields a lower or higher value with probability one. In the latter case, the corresponding matrix D with $\det(D) = -1$ yields the unique maximum, in the former case, the matrix D with $\det(D) = 1$, not involving the axis reflection, yields the unique maximum.

$\square$

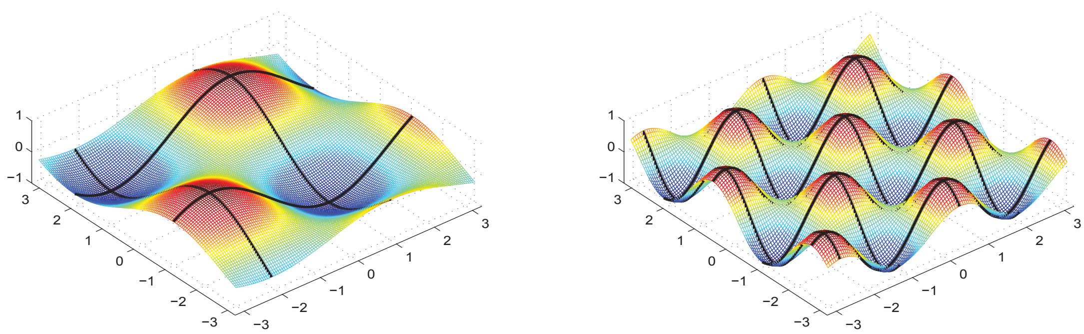

Figure 7: Bivariate sinusoids with frequency $\frac{1}{2\pi}$ along each dimension (left) and frequency $\frac{1}{\pi}$ along each dimension (right).