FEDERAL RESERVE BANK
PHILADELPHIA

Working Papers
RESEARCH DEPARTMENT

Identification Through Sparsity in Factor Models

The ℓ_1-Rotation Criterion

Simon Freyaldenhoven
Federal Reserve Bank of Philadelphia

WP 20-25

PUBLISHED
June 2020

REVISED
June 2025

ISSN: 1962-5361

DOI: https://doi.org/10.21799/frbp.wp.2020.25

Identification Through Sparsity in Factor Models: The ℓ_1-Rotation Criterion

Simon Freyaldenhoven

Federal Reserve Bank of Philadelphia*

June 11, 2025

Abstract

Linear factor models are generally not identified. We provide sufficient conditions for identification: Under a sparsity assumption, we can estimate the individual loading vectors using a novel rotation criterion that minimizes the ℓ_1-norm of the loading matrix. This enables economic interpretation of the factors. Existing rotation criteria (e.g., Varimax, Kaiser 1958) are theoretically unjustified and perform worse in our simulations. We illustrate our method in two economic applications.

JEL codes: C38, C51, C55

KEYWORDS: identification, factor models, sparsity, local factors

*Email: simon.freyaldenhoven@phil.frb.org. I thank Jushan Bai, Richard Crump, Chris Hansen, Adam McCloskey, Jesse Shapiro, and participants at numerous seminars and conferences for their comments and suggestions. Joseph Huang, Ryan Kobler, Nathan Schor, and Le Xu provided excellent research assistance. Previous versions of this working paper were circulated as *Identification Through Sparsity in Factor Models*.

1 Introduction

Factor models are subject to a rotational indeterminacy, meaning that the individual factors and loading vectors are identified only up to a rotation. Although this rotational indeterminacy prohibits any economic interpretation of the estimated factors, even seminal papers in economics (e.g., Stock and Watson 2002, Ludvigson and Ng 2009[1]) often include a discussion on the economic interpretations of individual factors, usually preceded by the caveat that such an interpretation is theoretically unjustified. For example, Stock and Watson (2002) remark:

> "Because the factors are identified only up to a $k \times k$ matrix, detailed discussion of the individual factors is unwarranted. Nevertheless, [...] Figure 1 therefore displays the R^2 of the regression of the 215 individual time series against each of the six empirical factors [...] Broadly speaking, the first factor loads primarily on output and employment; the second on interest rate spreads, unemployment rates and capacity utilization rates [...]."

We show that the assumption of sparsity in the loading matrix can solve this indeterminacy, allowing a researcher to estimate how the individual factors affect the observed variables. Sparsity in the loading matrix is natural in many economic applications. It is implied by the presence of local factors – factors that affect only a subset of the observables. Economic examples include industry-specific shocks in a firm-level dataset.[2]

Formally, our first result is that the true loading matrix Λ^* achieves the minimum of the ℓ_0-norm across rotations of the loading matrix under a sparsity assumption. Intuitively this states that any rotation of a sparse loading vector will be less sparse. However, a rotation criterion based directly on the sparsity pattern (ℓ_0-norm) of the loading matrix will generally be infeasible. Our next result then establishes that the true loading matrix Λ^* also achieves a minimum of the ℓ_1-norm across rotations. Specifically, our proposed ℓ_1-rotation criterion enables a researcher to consistently estimate the individual loading vectors of any local factors. Our rotation criterion is easy to implement in practice, and simply requires a $\sqrt{n}$-consistent estimate of the loading space as a starting point. Despite the resemblance to regularized estimation methods with an ℓ_1-penalty, such as Sparse Principal Component Analysis, we emphasize that there is no "shrinkage" involved in our estimator. Instead, we use the ℓ_1-norm

[1] The two papers have a combined citation count of more than 4,500 as of October 2022.

[2] We formally introduce various definitions of "local" factors later on. Intuitively, the loading vector $\lambda^*_{\bullet k}$ of a local factor F_k satisfies a sparsity assumption (a subset of the loadings are equal to zero), and an additional assumption on the collinearity between $\lambda^*_{\bullet k}$ and remainder of the loading matrix.

as a criterion to select the most sparse loading matrix Λ from among a set of rotations. Applying our criterion to both an international panel of daily stock returns and a panel of US macroeconomic indicators enables us to identify individual loading vectors in both cases and to better understand the economic structure of the data.

A byproduct of our result is a diagnostic to determine whether local factors are present in a given dataset. This diagnostic effectively consists of counting the number of "small" loadings in the most sparse rotation of the loading matrix, and comparing it to the number of small loadings that could be expected if the true loading matrix was non-zero everywhere. We find strong evidence for the existence of local factors in both of our applications.

Despite the large literature on both factor models and sparsity, little work has been done on the intersection of the two. There are multiple papers on sparse principal components in the statistics literature (e.g., Jolliffe et al. 2003, Zou et al. 2006), but since principal component analysis (PCA) is a model-free dimensionality reduction technique, the object of interest is quite different and, as long as the eigenvalues of the covariance matrix are distinct, PCA gives a unique solution. For a Bayesian perspective, see Ročková and George (2016) and Kaufmann and Schumacher (2019), who use sparse priors to encourage sparsity in the loading matrix.

Kristensen (2017) considers sparse principal components to estimate a factor model, but makes no sparsity assumptions and instead assumes that both the factors and the loading vectors are orthogonal to one another (and have distinct entries on the diagonal). This means there is no rotation invariance in his setup: under this assumption, even the principal components estimator will identity the individual columns of Λ^* and F (Bai and Ng, 2013). We argue that sparsity in the loading matrix, which is both economically appealing and has statistically testable implications, is a more natural assumption in many settings than assuming that both the factors and the loading vectors are orthogonal and that the eigenvalues of the covariance matrix are distinct.

Another related literature considers hierarchical factor models with a known group structure (e.g., Boivin and Ng 2006, Moench et al. 2013, Choi et al. 2018). Unlike those papers, we neither require the group structure to be known a priori, nor require a hierarchical model in which each outcome belongs to only one group. Ando and Bai (2017), Uematsu and Yamagata (2022) and Freyaldenhoven (2022) also do not require knowledge of the group structure a priori, but the focus of the first two papers is on estimation of the factor space, and the focus of the third paper is on estimating the number of factors. Neither addresses identification of

individual factors.[3]

Perhaps most closely related to our work is a large and popular literature that considers rotation criteria aimed to simplify the loading matrix in factor models, going back to at least Carroll (1953) and Kaiser (1958) (also see Katz and Rohlf 1974, Rozeboom 1991, Jennrich 2006).[4] However, existing rotation criteria are generally missing formal consistency results. To the best of our knowledge, our ℓ_1-rotation is the first rotation criterion that comes with theoretical guarantees to recover the true loading vectors under a sparsity assumption. Remarkably, we also find that our criterion performs better than existing criteria across our simulations.

The paper proceeds as follows. After setting up our model and fixing notation in Section 2, we discuss a simple example and give an intuitive discussion of our results in Section 3. In Section 4, we show that the true loading matrix Λ^* is the unique minimum of the ℓ_0-norm across rotations under exact sparsity. In Section 5, we establish that Λ^* is also a minimum of the ℓ_1-norm across rotations and extend our results to allow for $\sqrt{n}$-consistent initial estimates of the loading space and approximate sparsity in the true loading vectors. Section 6 combines the results from Sections 4 and 5 for our main result. Section 7 provides Monte Carlo evidence that supports our asymptotic results in finite sample. In Section 8, we apply our results to a panel of individual stock returns as well as a panel of US macroeconomic indicators.

2 Preliminaries

We use standard notation in the literature on factor models and assume X follows a factor structure:

$$\underset{(n\times 1)}{X_t} = \underset{(n\times r)(r\times 1)}{\Lambda^*\ F_t} + \underset{(n\times 1)}{e_t}\ \forall t, \qquad \text{or more compactly,} \qquad \underset{(T\times n)}{X} = \underset{(T\times r)(r\times n)}{F\ \Lambda^{*'}} + \underset{(T\times n)}{e}, \quad (1)$$

where $\Lambda^* = [\lambda^*_{1\bullet}\lambda^*_{2\bullet}...\lambda^*_{n\bullet}]' = [\lambda^*_{\bullet 1}\lambda^*_{\bullet 2}...\lambda^*_{\bullet r}]$ denotes the matrix of true factor loadings, and F denotes the unobserved factors. We use the running indices i, j for the n variables, and k, l for the r factors throughout. To rule out pathological cases, we will assume throughout that $rank(\Lambda^*) = r$.

[3] An alternative approach to identify individual factors assumes factors are independent and non-Gaussian. Then higher order moments become available to address the lack of identification via, e.g., Independent Component Analysis (Hyvärinen and Oja, 2000). Also see Gouriéroux et al. (2017) or Drautzburg and Wright (2023) for this approach to identification in structural VARs.

[4] For example, the Varimax criterion (Kaiser 1958) is widely used across fields with more than 9,000 citations as of August 2021 and is included in many major statistical software applications (e.g., R, MATLAB and SAS).

Let $tr(A)$ denote the trace of a matrix A. We use the Frobenius norm for matrices, such that $\|A\|^2 = tr(A'A) = \sum_{i,j} a_{ij}^2$. Similarly, unless otherwise noted, $\|A\|_1$ and $\|A\|_0$ will be entrywise (pseudo-)norms, such that $\|A\|_1 = \sum_{i,j} \left|a_{ij}\right|$ and $\|A\|_0$ counts the non-zero entries of a matrix A. For two vectors a, b we write $a \perp b$ to denote that they are orthogonal. A set in a superscript of a vector x, always denoted by a script letter (e.g., $\mathcal{G}$), defines a vector $x^{\mathcal{G}}$ such that $x_i^{\mathcal{G}} = x_i$ whenever $i \in \mathcal{G}$ and $x_i^{\mathcal{G}} = 0$ otherwise. We write $a_n \asymp b_n$ for two sequences a_n, b_n if $a_n = O(b_n)$ and $b_n = O(a_n)$. We normalize the length of the true loading vectors throughout, and impose that $\sum_{i=1}^n \lambda_{il}^{*2} = n$ for $l = 1, \ldots, r$. Clearly, such a normalization of a loading vector $\lambda_{\bullet k}^*$ and its corresponding factor F_k is immaterial.

Equation (1) is observationally equivalent for different rotations of the loadings and factors. To see this, let H denote an arbitrary nonsingular matrix. We can redefine $\Lambda^0 = \Lambda^*(H')^{-1}$ and $F^0 = FH$. This rotation may well be oblique since H does not need to be unitary, and we make no assumption that either the factors or the loading vectors are orthogonal. In our view, there is no reason a priori to believe that the underlying factors, and in particular the loading vectors, are necessarily orthogonal.

Among others, Bai and Ng (2002) showed in their seminal paper that in factor models of large dimensions we can consistently estimate the number of factors under some regularity conditions. We will therefore assume the true number of factors r to be known in the remainder of this paper.[5] Throughout the paper, we assume the data has been centered, such that $\mathbb{E}(X_i) = 0$.[6] Auxiliary lemmata are relegated to the Online Appendix.

3 Intuition

We start with a stylized example and an intuitive discussion of our proposed criterion.

3.1 A Stylized Example

To fix ideas, consider the following simple factor model with two factors for a vector x_t of dimension $n = 207$:

$$x_t = \lambda_{\bullet 1}^* F_{1t} + \lambda_{\bullet 2}^* F_{2t} + e_t, \quad t = 1, \ldots, T, \tag{2}$$

[5] See also Ahn and Horenstein (2013) and Onatski (2010) for alternative ways to determine the number of factors. Freyaldenhoven (2022) addresses the issue of estimating the number of factors under the presence of local factors, affecting only a subset of the observables.

[6] While unnecessary from a theoretical standpoint, normalizing the scale of the observed variables may also be appealing in practice. Adding a normalization step will not affect any of the conclusions that follow under some mild conditions.

where $\lambda^*_{\bullet k}$ denotes the vector of loadings for factor k (denoted by F_{kt}), and e_t an idiosyncratic noise component. We discuss the data-generating process (DGP) in more detail in Section 7. Suppose both factors are local with the structure of the loading matrix Λ^* given by

$$\Lambda^* = \begin{bmatrix} \lambda^*_{1:m_1,1} & 0 \\ 0 & \lambda^*_{(n+1)-m_2:n,2} \end{bmatrix}, \tag{3}$$

where $m_1 = m_2 = 120$. Thus, 120 outcomes are affected by the first factor, and 120 outcomes are affected by the second factor. Note that, with $n = 207$, some outcomes are affected by both factors. For the non-zero entries, we set $\lambda^*_{ik} \overset{i.i.d.}{\sim} U(0.1, 2.9)$. Figure 1 visualizes the resulting loading matrix Λ^*. Our goal is to recover Λ^*.

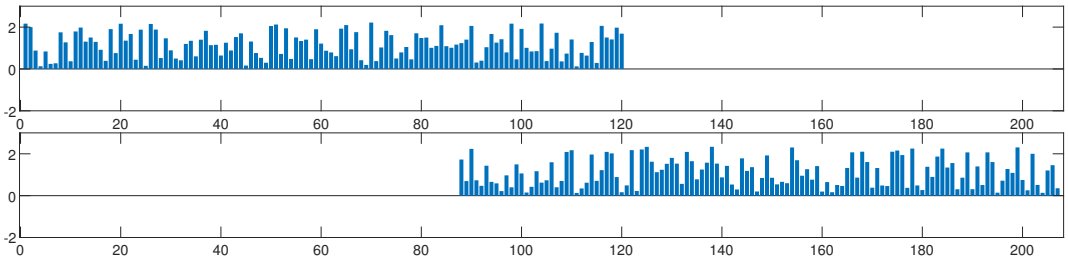


Figure 1: Illustration of true loading matrix Λ^* for stylized DGP. Top panel depicts $\lambda^*_{\bullet 1}$, bottom panel $\lambda^*_{\bullet 2}$. For each factor, the loadings associated with all 207 outcomes are depicted.

Under standard regularity conditions in the literature, it is well known that we can obtain estimates $\lambda^0_{\bullet 1}, \lambda^0_{\bullet 2}$, such that

$$\begin{aligned} \lambda^0_{i1} &= H_{11}\lambda^*_{i1} + H_{12}\lambda^*_{i2} + o_p(1) \\ \lambda^0_{i2} &= H_{21}\lambda^*_{i1} + H_{22}\lambda^*_{i2} + o_p(1), \end{aligned} \tag{4}$$

where H is an unknown nonsingular rotation matrix (e.g., Bai 2003).[7] Thus, the estimates $\lambda^0_{\bullet 1}$ and $\lambda^0_{\bullet 2}$ will in population be linear combinations of the true loading vectors $\lambda^*_{\bullet 1}$ and $\lambda^*_{\bullet 2}$. We make the following two observations (for now ignoring the $o_p(1)$ term in Equation (4)):

1. **Observation 1: Linear combinations of sparse loading vectors are generally dense.** For an arbitrary linear combination of the true loading vectors $\lambda^0_{\bullet 1} = H_{11}\lambda^*_{\bullet 1} + H_{12}\lambda^*_{\bullet 2}$ with $H_{11}, H_{12} \neq 0$ we will generally have $\lambda^0_{i1} \neq 0$ for $i = 1, \ldots, n$. Thus, even though the true loading vector $\lambda^*_{\bullet 1}$ is sparse (cf. Figure 1), a generic estimate $\lambda^0_{\bullet 1}$ will generally have non-zero entries everywhere.

[7] We state this result more formally in Assumption 3.

2. **Observation 2: There exists a linear combination of the estimated loading vectors that is sparse.**
   Since $\lambda^0_{\bullet 1}$ and $\lambda^0_{\bullet 2}$ are linear combinations of $\lambda^*_{\bullet 1}$ and $\lambda^*_{\bullet 2}$, it follows that $\lambda^*_{\bullet 1}$ and $\lambda^*_{\bullet 2}$ are also linear combinations of $\lambda^0_{\bullet 1}$ and $\lambda^0_{\bullet 2}$. In other words, there must exist weights w_1 and w_2, such that $\lambda^*_{\bullet 1} = w_1\lambda^0_{\bullet 1} + w_2\lambda^0_{\bullet 2}$. It then also follows that, if $\lambda^*_{\bullet 1}$ is sparse, there must exist a linear combination of $\lambda^0_{\bullet 1}$ and $\lambda^0_{\bullet 2}$ that is sparse.

Together, these two observations form the key insight of the paper: The sparsity pattern in the loading matrix is not invariant to rotations and can be used to achieve identification. We next illustrate our approach to identification in this stylized DGP. By construction, the Principal Component estimator Λ^0 will estimate a rotation H of the true loadings and factors that satisfies $\lambda^{0\prime}_{\bullet 1}\lambda^0_{\bullet 2} = 0$ and $F^{0\prime}_{\bullet 1}F^0_{\bullet 2} = 0$.[8] Figure 2 depicts this estimate. In line with

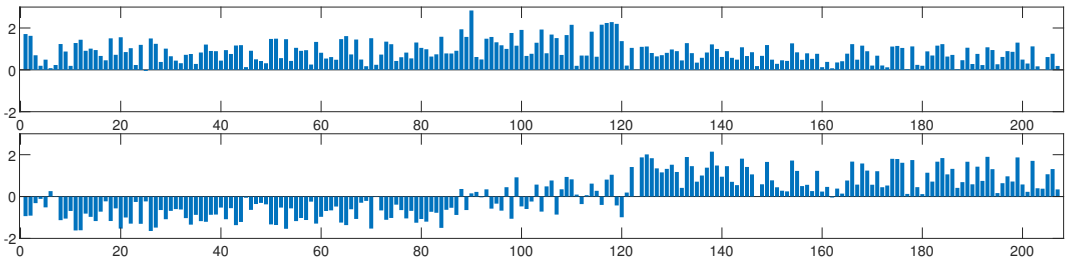


Figure 2: Illustration of Principal Component estimate Λ^0 for stylized DGP. Top panel depicts $\lambda^0_{\bullet 1}$, bottom panel $\lambda^0_{\bullet 2}$. For each factor, the loadings associated with all 207 outcomes are depicted.

Observation 1, the rotation matrix H inherent to the Principal Component estimator results in an estimate of the loading matrix with no discernible sparsity pattern. Further, comparing Figures 1 and 2, we conclude that neither of the estimated loading vectors closely resembles $\lambda^*_{\bullet 1}$ or $\lambda^*_{\bullet 2}$.

Following Observation 2, we are next interested in identifying a linear combination of $\lambda^0_{\bullet 1}$ and $\lambda^0_{\bullet 2}$ that is sparse. Because a rotation criterion that is directly based on the number of non-zero elements will generally be infeasible (we return to this later), our proposed estimator takes Λ^0 as a starting point and is equal to the rotation of Λ^0 that minimizes the ℓ_1-norm of the loading vectors. Figure 3 depicts the value of $\|\lambda_{\bullet k}\|_1$ across rotations in the space spanned by the Principal Component estimator Λ^0. Specifically, it depicts how $\|\lambda_{\bullet k}\|_1 = \|w_1\lambda^0_{\bullet 1} + w_2\lambda^0_{\bullet 2}\|_1$ changes as we vary the weights w_1, w_2, under the restriction that $w_1^2 + w_2^2 = 1$. A convenient way to enforce this restriction, and to depict the result graphically, is to let

[8] To compute the Principal Component estimator, we take the singular value decomposition $X = UDV'$. The leading r columns of V are used as $\lambda^0_{\bullet 1}, \ldots, \lambda^0_{\bullet r}$.

$[w_1, w_2] = [\sin(\theta), \cos(\theta)]$, and depict $\|\lambda_{\bullet k}\|_1$ as a function of the angle θ. This is depicted in Figure 3.

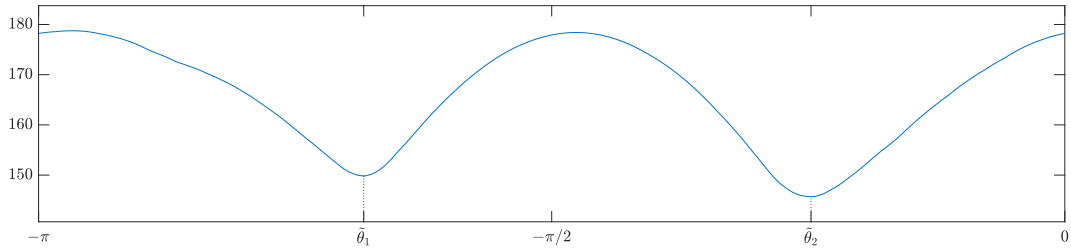


Figure 3: Objective function across rotations in the space spanned by the initial estimate Λ^0. Depicted is $\|\lambda_{\bullet k}\|_1 = \|\sin(\theta)\lambda^0_{\bullet 1} + \cos(\theta)\lambda^0_{\bullet 2}\|_1$ as a function of the angle θ.

We find two local minima at angles $\tilde{\theta}_1$ and $\tilde{\theta}_2$. The first minimum $\tilde{\theta}_1$ corresponds to weights of $[w_1 \ w_2] = [-0.70 \ 0.71]$, and consequently an estimated loading vector of $\tilde{\lambda}_{\bullet 1} = -0.70\lambda^0_{\bullet 1} + 0.71\lambda^0_{\bullet 2}$. The second minimum $\tilde{\theta}_2$ corresponds to weights of $[w_1, w_2] = [0.84, 0.54]$, and consequently a second estimated loading vector of $\tilde{\lambda}_{\bullet 2} = 0.84\lambda^0_{\bullet 1} + 0.54\lambda^0_{\bullet 2}$. Combining $[\tilde{\lambda}_{\bullet 1}, \tilde{\lambda}_{\bullet 2}] = \tilde{\Lambda}$, we obtain our proposed estimator for Λ^*. $\tilde{\Lambda}$ is depicted in Figure 4. Comparing Figures 1 and 4, we conclude that $\tilde{\Lambda}$ is close to Λ^*, and that we are able to identify the individual columns of Λ^* using our proposed criterion.

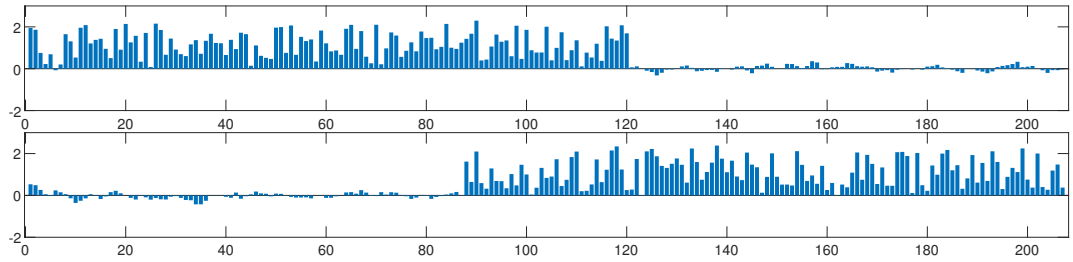


Figure 4: Illustration of $\tilde{\Lambda}$, the nonsingular "rotated" matrix with the smallest ℓ_1-norm $\|\Lambda\|_1 = \sum_{i,k} |\lambda_{ik}|$ for stylized DGP. Top panel depicts $\tilde{\lambda}_{\bullet 1}$, bottom panel $\tilde{\lambda}_{\bullet 2}$. For each factor, the loadings associated with all 207 outcomes are depicted.

Remark 1. Even though $\tilde{\Lambda}$ is close to Λ^*, we note that $\tilde{\lambda}_{ik} \neq 0$ for all i, k. This is expected because the preliminary estimate Λ^0 is subject to estimation error, and our method does not impose any regularization. Having identified the correct rotation of Λ^*, we conjecture that standard methods in regularized estimation, or even simple thresholding, can be used to further improve the estimate $\tilde{\Lambda}$ in practice. We leave this as an interesting avenue for future research.

It is also worth contrasting our proposal with existing regularized estimation procedures, such as variants of sparse PCA (cf. Kristensen, 2017). Regularizing the principal components

is akin to obtaining a sparse approximation of Λ^0 depicted in Figure 2. But this means inducing sparsity in an object (Λ^0) that will, in general, not be sparse. We instead propose to first identify a sparse rotation of the loading matrix ($\tilde{\Lambda}$, depicted in Figure 4). Any regularized estimation of Λ^* could then happen in a second step (to obtain a sparse approximation of $\tilde{\Lambda}$ rather than Λ^0) and is thus complementary to our approach.

Alternatively, we can approximate the number of loadings that are zero ($n - \|\lambda_{\bullet k}\|_0$) directly for each rotation by counting the number of "small" loadings. Figure 5 depicts the number of small loadings in $\lambda_{\bullet k}$ across rotations in the space spanned by the initial estimate Λ^0, again as a function of the angle θ . While in this case, with just two factors, it is feasible to find the rotation that minimizes the (approximate) ℓ_0-norm based on a visual inspection of Figure 5, the discontinuities and large number of local extrema of this function make this approach infeasible in higher dimensions (we expand on this in Online Appendix B). On the other hand, the angles $\tilde{\theta}_1$ and $\tilde{\theta}_2$ in Figure 5 are those found by minimizing the ℓ_1-norm of the loadings (which is continuous) and are identical to the local minima depicted in Figure 3. Importantly, the minima of the ℓ_1- and ℓ_0-norm are close to each other, and both are close to Λ^*.

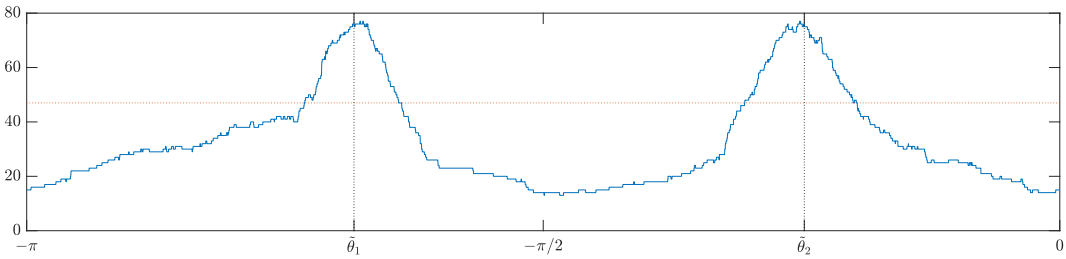


Figure 5: Depicted is the number of small loadings $Q_0 = \sum_{i=1}^{n} \mathbf{1}_{|\lambda_{ik}|<1/log(n)}$, where $\lambda_{\bullet k} = \sin(\theta)\lambda^0_{\bullet 1} + \cos(\theta)\lambda^0_{\bullet 2}$, as a function of the angle θ. Horizontal dashed red line represents critical value for assessing whether there are local factors in the data.

We also use Figure 5 to illustrate how one may be able to use the estimate $\tilde{\Lambda}$ to infer whether local factors exist in a given dataset. In Section 4.3, we suggest a diagnostic that amounts to counting the number of small loadings in the most sparse estimated loading vector $\tilde{\lambda}_{\bullet k}$. In Figure 5, this corresponds to checking whether, for either of the angles $\tilde{\theta}_1$ and $\tilde{\theta}_2$, the number of small coefficients, indicated by the blue line, is larger than a critical value, indicated by the horizontal red dashed line. Based on Figure 5, our diagnostic suggests the presence of local factors in this instance.[9]

[9]In Online Appendix A, we depict the equivalent of Figure 5 if both factors affect all outcomes. There, our diagnostic correctly suggests that no local factors are present in the data. Intuitively, if both $\lambda^*_{\bullet 1}$ and $\lambda^*_{\bullet 2}$ are non-zero everywhere, no linear combination of the two exists with a significant sparsity pattern.

3.2 Connection to Existing Rotation Criteria

A number of widely used rotation criteria exist aimed at simplifying the loading matrix, going back to at least Carroll (1953) and Kaiser (1958). These existing criteria usually use quartic functions of the loadings and maximize a variant of the following criterion function $Q(\cdot)$ over rotations of an initial estimate Λ^0:

$$Q(\Lambda^0 R) = Q(\Lambda) = \sum_{k=1}^{r}\sum_{l=1}^{k-1}(\sum_{i=1}^{n}\lambda_{ik}^2\lambda_{il}^2 - \frac{c}{n}\sum_{i=1}^{n}\lambda_{ik}^2\sum_{j=1}^{n}\lambda_{jl}^2). \tag{5}$$

If we consider only orthogonal rotations for now (which is equivalent to restricting R to be orthonormal), (5) simplifies to

$$Q(\Lambda^0 R) = Q(\Lambda) = \sum_{k=1}^{r}\left[\sum_{i=1}^{n}\lambda_{ik}^4 - \frac{c}{n}\left(\sum_{i=1}^{n}\lambda_{ik}^2\right)^2\right]. \tag{6}$$

For example, setting c to 0, 1, and $r/2$, respectively, results in the Quartimax (Carroll 1953), Varimax (Kaiser 1958), and Equamax (Saunders 1962) rotation criteria. Considering one loading vector at a time, it becomes clear that these are closely related to maximizing $\|\lambda_{\bullet k}\|_4^4 = \sum_{i=1}^{n}\lambda_{ik}^4$, subject to a constant ℓ_2-norm. In contrast, we propose to minimize $\|\lambda_{\bullet k}\|_1 = \sum_{i=1}^{n}|\lambda_{ik}|$, subject to a constant ℓ_2-norm.

To gain an intuition for the difference between the two approaches (maximizing ℓ_4, minimizing ℓ_1), it is instructive to first consider maximizing the ℓ_∞-norm and contrast this with minimizing the ℓ_0-norm across rotations. Intuitively, maximizing the ℓ_∞-norm identifies the rotation with the largest entry, while minimizing the ℓ_0-norm essentially identifies the rotation with the smallest entries. Minimizing the ℓ_1-norm is a relaxation of minimizing the ℓ_0-norm, while maximizing the ℓ_4-norm is a relaxation of maximizing the ℓ_∞-norm. Our sparsity assumptions have direct implications for the behavior of the ℓ_0- and ℓ_1-norms, but not the ℓ_4- or ℓ_∞-norms. We conjecture this is the reason why, under sparsity assumptions, formal results have been difficult to achieve using existing rotation criteria that are quartic functions of the loadings.

We discuss the connection between our proposed method and a variety of quartic criteria, including criteria that result in oblique factor rotations (e.g., Hendrickson and White 1964), further in Online Appendices B and E.

4 Minimizing the ℓ_0-norm

Let R be a $r \times r$ matrix and consider the following minimization problem:

$$\tilde{R} = \arg\min_{R} \|\Lambda^* R\|_0, \qquad \text{such that } R \text{ is nonsingular and } \|R_{\bullet k}\|_2 = 1 \ \forall k. \tag{7}$$

Define $\tilde{\Lambda} = \Lambda^* \tilde{R}$ as the rotation of Λ^* corresponding to this minimum.

Our first set of results is derived under an exact sparsity pattern in the loading matrix.

Assumption 1. *For each factor k, we can partition the set of indices $i = 1, 2, \ldots, n$ into a set of indices $\mathcal{A}_k$ with cardinality $|\mathcal{A}_k|$ and its complement $\mathcal{A}_k^c$, such that:*

(a) $\lambda_{ik}^* \neq 0$ *and* $\left|\lambda_{ik}^*\right| < C \ \forall i \in \mathcal{A}_k$ *and a constant* C.

(b) $\lambda_{ik}^* = 0 \ \forall i \notin \mathcal{A}_k$.

(c) $\exists c > 0$, *such that* $\left|\lambda_{ik}^*\right| > c \ \forall i \in \mathcal{A}_k$.

Parts (a)-(b) define $\mathcal{A}_k$ as the support of $\lambda_{\bullet k}^*$, and we may think of $\mathcal{A}_k$ as the "active set" for a given factor or loading vector: it collects the indices of all outcomes affected by that factor. On their own, parts (a)-(b) are thus merely a definition. Some results additionally require Assumption 1(c), which requires a gap between loadings on $\mathcal{A}_k$ and its complement.

4.1 Two Factors ($r = 2$)

To keep our notation simple, we start with a simplified version of our results for the case of two factors.

Definition 1'. *Let $b = \max |\mathcal{B}|$, such that $\mathcal{B} \subseteq (\mathcal{A}_1 \cap \mathcal{A}_2)$, and for all $i \in \mathcal{B}$*

$$c^* \lambda_{i1}^* = \lambda_{i2}^*, \tag{8}$$

for some constant c^*.

Define the set $\mathcal{I}_{\ell_0}^e = \left\{ k \in \{1, 2\} : \left|\mathcal{A}_k^c \cap \mathcal{A}_l\right| - b > 0 \text{ for all } l \neq k \right\}$.

Thus, b denotes the size of the largest set of non-zero entries in the loading vectors such that the two loading vectors are perfectly collinear on that set.[10] Intuitively, requiring a lower

[10] Since we generally treat the factor loadings as parameters rather than random variables, we do not specify b further in Definition 1'. More primitive conditions can be derived if we treat the loadings as random instead. For instance, suppose $|\mathcal{A}_k| \asymp n$ for $k = 1, 2$ as $n \to \infty$, and $\lambda_{ik}^* \overset{i.i.d.}{\sim} N(0, \sigma)$ if $i \in \mathcal{A}_k$, and $\lambda_{ik}^* = 0$ otherwise. Then, with probability 1, $\frac{\lambda_{i1}}{\lambda_{i2}} \neq \frac{\lambda_{j1}}{\lambda_{j2}}$ for all $i, j \in \mathcal{A}_1 \cap \mathcal{A}_2$, and it follows that $b = 1$.

bound on $\left|\mathcal{A}_k^c \cap \mathcal{A}_l\right|$ in the definition of $\mathcal{I}_{\ell_0}^e$ can be thought of as our sparsity assumption: $\left|\mathcal{A}_k^c \cap \mathcal{A}_l\right|$ will be larger when $\lambda_{\bullet k}^*$ has a significant sparsity pattern (since the set $\mathcal{A}_k^c$ is by construction larger when the set $\mathcal{A}_k$ is small). Thus, the more outcomes are unaffected by F_k, the more likely $k \in \mathcal{I}_{\ell_0}^e$.

To see that a naive definition of a local factor that simply requires sparsity in $\lambda_{\bullet k}^*$, and thus restricts the size of $\mathcal{A}_k$ (e.g. $k \in \mathcal{I}_{\ell_0}^e \Leftrightarrow |\mathcal{A}_k| < \alpha n$ for some $\alpha \in (0,1)$), cannot be sufficient for identification we give the following simple counterexample. Suppose $\mathcal{A}_1 = \mathcal{A}_2 = \{1, \ldots, n/2\}$, such that the first half of the observed outcomes are affected by both factors. Now consider a setting where only $i = \{1, \ldots, n/2\}$ is observed. In such a setting, we have a traditional factor model in which both factors affect all outcomes. Thus, the usual rotational indeterminacy (and thus lack of identification) applies. Further, simply adding an additional $n/2$ "noise variables" that are unrelated to both factors for $i = \{n/2+1, \ldots, n\}$ clearly cannot be be sufficient for identification. This illustrates why a simple sparsity assumption on $\lambda_{\bullet k}^*$ is not sufficient for identification.

It is also instructive to consider the following three specific examples:

1. Suppose that $\frac{\lambda_{i1}}{\lambda_{i2}} \neq \frac{\lambda_{j1}}{\lambda_{j2}}$ for all $i, j \in \mathcal{A}_1 \cap \mathcal{A}_2$. Then, $b = 1$. Further suppose that at least two distinct outcomes are unaffected by each factor (e.g., $\lambda_{11}^* = \lambda_{21}^* = \lambda_{32}^* = \lambda_{42}^* = 0$, and λ_{12}^*, λ_{22}^*, λ_{31}^*, and λ_{41}^* are non-zero).
   Then, $\left|\mathcal{A}_k^c \cap \mathcal{A}_l\right| > 1$ for $l \neq k$. Therefore $1, 2 \in \mathcal{I}_{\ell_0}^e$.

2. Suppose $\mathcal{A}_1 \cap \mathcal{A}_2 = \varnothing$ (The two factors affect different, non-overlapping groups of outcomes).
   Then, $b = 0$, while $\left|\mathcal{A}_k^c \cap \mathcal{A}_l\right| = |\mathcal{A}_l| > 0$ for $l \neq k$. Therefore $1, 2 \in \mathcal{I}_{\ell_0}^e$.

3. Suppose $\mathcal{A}_2 \subseteq \mathcal{A}_1$ (The second factor F_2 affects a subset of the outcomes affected by F_1).
   Then, $|\mathcal{A}_1^c \cap \mathcal{A}_2| = 0$, and $1 \notin \mathcal{I}_{\ell_0}^e$. Thus, whenever $\mathcal{A}_k$ is a superset of another active set $\mathcal{A}_l$, k cannot be a member of the set $\mathcal{I}_{\ell_0}^e$.

Remark 2. Four different versions of the set $\mathcal{I}$ ($\mathcal{I}_{\ell_0}^e$, $\mathcal{I}_{\ell_1}^e$, $\mathcal{I}_{\ell_1}$, $\mathcal{I}_{\ell_0}$) will appear throughout the paper. We will generally be able to show identification for the loading vectors corresponding to factors in these sets. They define sets of local factors (or, more precisely, the corresponding indices) whose loadings can be identified by sparsity under various conditions. In particular,

1. $\mathcal{I}_{\ell_0}^e$ will collect sparse loading vectors identifiable by rotating the ℓ_0-norm under exact sparsity and no estimation error (cf. Definitions 1' and 1),

2. $\mathcal{I}^e_{\ell_1}$ will collect collect sparse loading vectors identifiable by rotating the ℓ_1-norm under exact sparsity and no estimation error (cf. Definitions 2' and 2),

3. $\mathcal{I}_{\ell_1}$ will collect collect sparse loading vectors identifiable by rotating the ℓ_1-norm under approximate sparsity and allowing for estimation error (cf. Definition 3),

4. $\mathcal{I}_{\ell_0}$ will collect collect sparse loading vectors identifiable by rotating an approximate version of the ℓ_0-norm (which we introduce in Section 6) under approximate sparsity and allowing for estimation error (cf. Definition 4).

The sets $\mathcal{I}_{\ell_0}$ and $\mathcal{I}_{\ell_1}$ are not nested (and neither are $\mathcal{I}^e_{\ell_0}$ and $\mathcal{I}^e_{\ell_1}$), but as we argue throughout the paper both capture the idea of "local factors." That is, the more sparse $\lambda^*_{\bullet k}$ is, the more likely $k \in \mathcal{I}$ for any of its four versions ($\mathcal{I}^e_{\ell_0}, \mathcal{I}^e_{\ell_1}, \mathcal{I}_{\ell_1}, \mathcal{I}_{\ell_0}$). We thus expect them to be similar in practice. Further, perhaps unsurprisingly, the conditions for identification will be more stringent when allowing for estimation error, and thus $\mathcal{I}_{\ell_0} \subseteq \mathcal{I}^e_{\ell_0}$ and $\mathcal{I}_{\ell_1} \subseteq \mathcal{I}^e_{\ell_1}$ (this simply states that, if a loading vector can be identified without estimation error using a particular norm, then it can also be identified when allowing for such error). To keep the writing concise, we will call factors in these sets simply "local" throughout, with the understanding that these definitions are in fact slightly more restrictive and which factors fulfill the definition of a local factor will vary depending on a) which norm we use, and b) whether we allow for estimation error.

To gain further intuition for Definition 1', and how it will be used in our proofs, suppose $1 \in \mathcal{I}^e_{\ell_0}$ and consider rotating $\lambda^*_{\bullet 1}$ – i.e., consider adding a nonzero weight w_2 to $\lambda_{\bullet 1} = w_1\lambda^*_{\bullet 1} + w_2\lambda^*_{\bullet 2}$, such that $|w_1|, |w_2| \in (0, 1)$. Compared to $\lambda^*_{\bullet 1}$, $\lambda_{\bullet 1}$ now has additional non-zero entries on $\mathcal{A}_1^c \cap \mathcal{A}_2$. By definition, $\lambda_{\bullet 1}$ has more than b such additional non-zero entries. On the other hand, any entries that satisfy $\lambda^*_{i1} = -\frac{w_2}{w_1}\lambda^*_{i2}$ will be zero on $\lambda_{\bullet 1}$. But the number of entries that satisfy this last condition is bounded above by b. Definition 1' thus ensures that $\lambda_{\bullet 1}$ is less sparse than $\lambda^*_{\bullet 1}$ for any $w_2 \neq 0$ if $1 \in \mathcal{I}^e_{\ell_0}$.

If $\frac{\lambda_{i1}}{\lambda_{i2}} \neq \frac{\lambda_{j1}}{\lambda_{j2}}$ $\forall i, j \in \mathcal{A}_1 \cap \mathcal{A}_2$ ($b = 1$), this is guaranteed by having two outcomes affected by the second factor that are not affected by the first factor (cf. Example 1 above). If there is some collinearity between $\lambda^*_{\bullet 1}$ and $\lambda^*_{\bullet 2}$ (meaning $b > 1$), we intuitively require a larger amount of sparsity in $\lambda^*_{\bullet 1}$ such that $|\mathcal{A}_1^c \cap \mathcal{A}_2| > b$ is satisfied.

The general idea, which extends to the case of $r > 2$, is the following: There are two ways in which a linear combination of the true loading vectors can have a significant sparsity pattern: Either one of the true loading vectors is sparse, or the two loading vectors are collinear on a large subset of outcomes (this is the set $\mathcal{B}$). If $k \in \mathcal{I}^e_{\ell_0}$, this restricts the amount

of collinearity between the two loading vectors, or more precisely the size of the set $\mathcal{B}$, relative to the amount of sparsity in $\lambda^*_{\bullet k}$. To us, ruling out large subsets of outcomes on which loading vectors are perfectly collinear does not seem very restrictive. Then, if we find a linear combination of the true loading vectors with a significant sparsity pattern, this must stem from the fact that one of the true loading factors is indeed sparse.

Before we continue to the more general case and present a generalization of Definition 1' that allows for $r \geq 3$, we present our first formal result in the simple case in which we just have two local factors, similar to the stylized example in the previous section.

Proposition 1. *Suppose $r = 2$, Assumption 1(a)-(b) holds, and $k \in \mathcal{I}^e_{\ell_0}$ for $k = 1, 2$. Then $\tilde{\Lambda} = \Lambda^* P$ for some permutation matrix P.*

Proof. First note that, for any permutation matrix P, $\|\Lambda^* P\|_0 = \|\Lambda^*\|_0$, since reordering the columns clearly does not change the amount of sparsity.

Next, suppose that the solution to (7) is $\tilde{\Lambda} = \Lambda^* \tilde{R}$, where $\tilde{R}$ is not a permutation matrix. Then, at least one column $\tilde{\lambda}_{\bullet k}$ is a linear combination of both columns in Λ^*: $\tilde{\lambda}_{\bullet k} = \Lambda^* R_{\bullet k} = \lambda^*_{\bullet 1} R_{1k} + \lambda^*_{\bullet 2} R_{2k}$, with $R_{2k}, R_{1k} \neq 0$. It follows that

$$\begin{aligned} \|\tilde{\lambda}_{\bullet k}\|_0 = \|\lambda^*_{\bullet 1} R_{1k} + \lambda^*_{\bullet 2} R_{2k}\|_0 &= |\mathcal{A}_1 \cup \mathcal{A}_2| - |\mathcal{C}| \\ &= |\mathcal{A}_1| + |\mathcal{A}_2 \cap \mathcal{A}_1^c| - |\mathcal{C}| \\ &= |\mathcal{A}_2| + |\mathcal{A}_1 \cap \mathcal{A}_2^c| - |\mathcal{C}|\,, \end{aligned}$$

where $\mathcal{C}$ is defined as the set of indices i, such that $\frac{\lambda^*_{i1}}{\lambda^*_{i2}} = -\frac{R_{2k}}{R_{1k}}$.

Since, by Definition 1', $|\mathcal{A}_1^c \cap \mathcal{A}_2| > |\mathcal{C}|$ and $|\mathcal{A}_2^c \cap \mathcal{A}_1| > |\mathcal{C}|$, it follows that $\|\tilde{\lambda}_{\bullet k}\|_0 > \max(\|\lambda^*_{\bullet 1}\|_0, \|\lambda^*_{\bullet 2}\|_0)$, and thus $\|\Lambda^* \tilde{R}\|_0 > \|\Lambda^*\|_0$. □

Proposition 1 formalizes our previous claim that Λ^* is the most sparse representation among its rotations.

4.2 More Than Two Factors

Definition 1. *Let $\Lambda^*_{\bullet,-m}$ be the n by $(r-1)$ submatrix of Λ^* obtained by deleting the mth column in Λ^*, let $\mathcal{A}_{z,-m}$ be the support of a linear combination $\Lambda^*_{\bullet,-m} z$ for a given $(r-1)$ vector of finite weights z, and let $b_k(z) = \max |\mathcal{B}|$, $\mathcal{B} \subseteq \mathcal{A}_k$, such that*

$$\Lambda^{*\mathcal{A}_k}_{i,-k} z = \lambda^{*\mathcal{A}_k}_{ik} \; \forall i \in \mathcal{B}. \tag{9}$$

Define the set $\mathcal{I}^e_{\ell_0} = \Big\{ k \in \{1, \ldots, r\} : \big|\mathcal{A}_k^c \cap \mathcal{A}_{z,-k}\big| - b_k(z) > 0 \; \forall z \neq 0 \Big\}$.

In words, $b_k = max_z b_k(z)$ is the size of the largest set of non-zero loadings on $\lambda^*_{\bullet k}$ that can be replicated as an exact linear combination of the remaining loading vectors. A small value for b_k (e.g., $b_k = r - 1$) means this set is small: no large subset of outcomes exists for which the loading vector $\lambda^*_{\bullet k}$ is perfectly collinear with a linear combination of the remaining loading vectors.[11] The more collinearity there exists between $\lambda^*_{\bullet 1}$ and a linear combination of the remaining columns in Λ^* (meaning a larger b_k), the higher the amount of sparsity required in $\lambda^*_{\bullet 1}$ such that $\left|\mathcal{A}_k^c \cap \mathcal{A}_{z,-k}\right| - b_k(z) > 0$ is satisfied.

While we refer to factors corresponding to the various versions of the set $\mathcal{I}^e_{\ell_0}$ as local factors, we reiterate, following our discussion above, that simply requiring the loadings $\lambda_{\bullet k}$ to have a sparsity pattern is not sufficient to guarantee that $k \in \mathcal{I}^e_{\ell_0}$.

Remark 3. It will generally also be possible to identify the subspace spanned by a given subset of loading vectors, even if their active sets are closely related, as long as they are sufficiently distinct from the active sets of all other factors. In order not to further complicate our notation, we will ignore this and simply consider such factors "unidentifiable."

Next, we generalize the results of Proposition 1 to allow for more than two factors.

Corollary 1. *Suppose Assumption 1(a)-(b) holds, and $k \in \mathcal{I}^e_{\ell_0}$ for $k = 1, \ldots, r$. Then $\tilde{\Lambda} = \Lambda^* P$ for some permutation matrix P.*

Corollary 1 is a direct consequence of Theorem 1 below. It again states that, in a model with only local factors, any rotation of the true loading matrix will be less sparse than the truth. Thus, the rotation with the highest degree of sparsity identifies the individual loading vectors, up to an arbitrary relabeling of the factors. In most settings of economic interest, there will be at least some factor F_k, such that $k \notin \mathcal{I}^e_{\ell_0}$ (for instance, a global factor). In such a case, we obtain the following generalization of Corollary 1.

Theorem 1. *Suppose Assumption 1(a)-(b) holds. Then for every $l \in \mathcal{I}^e_{\ell_0}$, there exists an index k (which depends on l), such that $\tilde{R}_{lk} = 1$ and $\tilde{R}_{l',k} = 0 \; \forall l' \neq l$.*

The proof of Theorem 1 is similar to that of Proposition 1 and thus relegated to the Online Appendix. Theorem 1 establishes the following: If the true DGP includes local factors ($\mathcal{I}^e_{\ell_0}$ is non-empty with, say, $|\mathcal{I}^e_{\ell_0}| = r^*$), the corresponding r^* columns in Λ^* will also appear as columns in $\tilde{\Lambda}$. This means that the loading vectors for such local factors can be identified by maximizing the degree of sparsity in the loading matrix across rotations. Note that Theorem

[11] We again note that under additional assumptions, we can further specify b_k. For example, suppose we treat the loadings as random instead of fixed, and $\lambda^*_{ik} \overset{i.i.d.}{\sim} N(0, \sigma)$ if $i \in \mathcal{A}_k$, and $\lambda^*_{ik} = 0$ otherwise. Then, with probability 1, $b_k = r - 1$.

1 does not say anything about columns of Λ^* corresponding to indices that are not in $\mathcal{I}^e_{\ell_0}$, or equivalently, the remaining $r - r^*$ columns in $\tilde{\Lambda}$. These could be arbitrary linear combinations of the columns in Λ^*. This is intuitive: if there are global factors with a corresponding loading vector that has non-zero entries everywhere, identification of such loading vectors based on a sparsity criterion will clearly be impossible.

4.3 A Diagnostic for Sparsity in Λ^*

Intuitively, if all factors affect all outcomes ($|\mathcal{A}_k| = n$ for $k = 1, \ldots, r$) and loading vectors are not (close to) collinear, even the most sparse rotation in the space spanned by the true loading matrix will not have a significant sparsity pattern. On the other hand, if there exists a loading vector $\lambda^*_{\bullet k}$ with at least a constant fraction of its loadings equal to zero, then clearly a rotation with as many zeros exists.

We thus define the largest amount of sparsity across loading vectors for a given loading matrix Λ as $\mathcal{L}_0(\Lambda) = \max_k(n - \|\lambda_{\bullet k}\|_0)$ and suggest the following diagnostic to determine whether local factors are present:

$$LF = \mathbf{1}\{\hat{\mathcal{L}}_0(\tilde{\Lambda})) \geq \gamma n\}, \quad \text{where } \hat{\mathcal{L}}_0(\tilde{\Lambda}) = \max_k(\sum_{i=1}^{n} \mathbf{1}\{\left|\tilde{\lambda}_{ik}\right| < h_n\}), \tag{10}$$

where $h_n \to 0$ as $n \to \infty$. Thus, LF amounts to checking whether there exists a rotation such that the number of loadings smaller than the threshold h_n is larger than γn.[12] While deriving a formal test for the existence of local factors is beyond the scope of this paper, Online Appendix G provides some theoretical justification for our diagnostic.

5 Minimizing the ℓ_1-norm

Unfortunately, minimizing the ℓ_0-norm directly is infeasible in practice for two reasons.

First, we assumed thus far that a) Λ^* had an exact sparsity pattern, and b) we were able to consider rotations in the true loading space instead of rotations in the space spanned by an initial estimate of this space. In practice, if either Λ^* is only approximately sparse, or if we allow for estimation error in the initial estimate of the loading space, no rotation of the initial estimate will exist that has an exact sparsity pattern. This poses a challenge to minimizing the ℓ_0-norm directly.

Second, even if a rotation with exact zeros did exist, minimizing the ℓ_0-norm will gener-

[12] We reemphasize that counting the number of small loadings in an arbitrary rotation (e.g., using the Principal Component estimator Λ^0) would not work. In general, the number of small loadings in Λ^0 will be small even under sparsity in Λ^* (cf. Figures 1 and 2). It is therefore crucial to first find the most sparse rotation $\tilde{\Lambda}$.

ally be computationally prohibitive. One can compare this to high-dimensional sparse linear regression models, where optimal subset selection is generally infeasible. On the other hand, a vast body of literature exists documenting both the theoretical and practical appeal of using the ℓ_1-norm instead as a regularization in linear regression models (e.g., Bühlmann and Van De Geer 2011).[13] While both minimizing the ℓ_1-norm and maximizing any of the quartic rotation criteria discussed in Section 3.2 are still non-convex problems, finding a solution to these problems is much easier than minimizing the ℓ_0-norm directly. For a geometric illustration of this, see Online Appendix B (and recall our discussion surrounding Figures 3 and 5). We therefore turn our attention to the ℓ_1-norm of the loading matrix next.

In what follows, we will work with an initial linear transformation of Λ^*, rather than with Λ^* directly. We will denote this as $\Lambda^0 = \Lambda^* H$, where H is nonsingular. Λ^0 has the property that its columns have equal length and are orthogonal, such that $\frac{\Lambda^{0\prime}\Lambda^0}{n} = I$. Intuitively, one can think of Λ^0 as the rotation of Λ^* that is estimated by the Principal Component estimator, at this point still ignoring any estimation error.

Consider the following optimization problem:

$$\min_R \|\Lambda^0 R\|_1 \qquad \text{such that } R \text{ is nonsingular and } \|R_{\bullet k}\|_2 = 1 \; \forall k. \tag{11}$$

Noting that $\|\Lambda^0 R\|_1 = \sum_{k=1}^r \| \sum_{l=1}^r \lambda^0_{\bullet l} R_{lk}\|_1$, we see that (11) is separable in k and consists of k identical parts up to the nonsingularity constraint. We thus consider one part at a time:

$$\min_{R_{\bullet k}} \| \sum_{l=1}^r \lambda^0_{\bullet l} R_{lk}\|_1 \qquad \text{such that } \|R_{\bullet k}\|_2 = 1. \tag{12}$$

Importantly, $\frac{\Lambda^{0\prime}\Lambda^0}{n} = I$ implies $\|\lambda_{\bullet k}\|_2 = \|\Lambda^0 \Upsilon\|_2 = \sqrt{n}$ for any $(r \times 1)$ vector Υ with $\|\Upsilon\|_2 = 1$. When considering the ℓ_1-norm of $\lambda_{\bullet k}$ for different linear combinations Υ, we therefore hold the ℓ_2-norm of $\lambda_{\bullet k}$ constant across those combinations (and are thus looking at "rotations" rather than arbitrary linear combinations, also see Online Appendix B).

Remark 4. Implemented directly, the optimization problem in (12) is still computationally challenging as it involves finding minima of a non-convex function over the surface of an r-dimensional sphere. In practice, we translate the problem into spherical coordinates, such

[13] We note that in the context on linear regressions, there has been some progress in implementing an ℓ_0-based approach via mixed integer optimization in recent years (e.g., see Bertsimas et al. (2016), Huang et al. (2018), or Chen and Lee (2023).

that the constraint minimization problem in (12) simply becomes

$$\min_{\theta_k} \left\| \lambda^0_{\bullet 1} \cos(\theta_{1k}) + \lambda^0_{\bullet r} \Pi_{p=1}^{r-1} \sin(\theta_{pk}) + \sum_{l=2}^{r-1} \lambda^0_{\bullet l} \cos(\theta_l) \Pi_{p=2}^{l-1} \sin(\theta_{p-1,k}) \right\|_1 , \qquad (13)$$

an unconstrained optimization over the $(r-1)$ angles collected in the vector θ_k. We then solve (13) for a grid of starting points to find all local minima, similar to a standard implementation of existing quartic rotation criteria such as those based on (6). See Browne (2001) for a discussion and additional references. We discuss our transformation and the implementation of (13) in more detail in Online Appendix F.

5.1 Two Factors ($r = 2$)

We again start with a simplified version of our next definition for the case of two factors, and, to further simplify notation, start by further assuming that the two loading vectors are orthogonal.

Definition 2". *Suppose* $\frac{\Lambda^{*\prime}\Lambda^*}{n} = I_2$*, and, for* $l \neq k$*, let*

$$\beta^k = \left| \sum_{i \in \mathcal{A}_k} |\lambda^*_{il}| \, \mathbf{1}\{\lambda^*_{ik}\lambda^*_{il} > 0\} - \sum_{i \in \mathcal{A}_k} |\lambda^*_{il}| \, \mathbf{1}\{\lambda^*_{ik}\lambda^*_{il} < 0\} \right| . \qquad (14)$$

Define the set $\mathcal{I}^e_{\ell_1} = \left\{ k \in \{1,2\} : \|\lambda^{*\,\mathcal{A}_k^c}_{\bullet l}\|_1 > \beta^k . \right\}$.

Note the similarity in the definitions of $\mathcal{I}^e_{\ell_1}$ and $\mathcal{I}^e_{\ell_0}$ (cf. Definition 1'): If $1 \in \mathcal{I}^e_{\ell_0}$, the condition $|\mathcal{A}_1^c \cap \mathcal{A}_2| > b$ required a lower bound on the number of non-zero loadings in $\lambda_{\bullet 2}$ outside of the set $\mathcal{A}_1$. If 1 is in $\mathcal{I}^e_{\ell_1}$, Definition 2" requires a lower bound on the sum (of the absolute values) of the loadings $\lambda_{\bullet 2}$ outside of the set $\mathcal{A}_1$. Since $\|\lambda^{*\,\mathcal{A}_1^c}_{\bullet 2}\|_1$ will tend to be larger if $\lambda^*_{\bullet 1}$ has a significant sparsity pattern (because the set $\mathcal{A}_1^c$ is by construction larger when the set $\mathcal{A}_1$ is small), the more outcomes are unaffected by F_k, the likelier $k \in \mathcal{I}^e_{\ell_1}$.

To gain intuition for how the inequality inside the definition of $\mathcal{I}^e_{\ell_1}$ is used in the proofs, again consider rotating $\lambda^*_{\bullet 1}$ — i.e., adding a (small) non-zero weight w_2 to $\lambda_{\bullet 1} = w_1\lambda^*_{\bullet 1} + w_2\lambda^*_{\bullet 2}$. Compared to $\lambda^*_{\bullet 1}$, $\lambda_{\bullet 1}$ now has additional non-zero entries on $\mathcal{A}_1^c \cap \mathcal{A}_2$, increasing the ℓ_1-norm of $\lambda_{\bullet 1}$ by $|w_2| \, \|\lambda^{*\,\mathcal{A}_1^c}_{\bullet 2}\|_1$. On the other hand, the entries on $\mathcal{A}_1$ may either increase or decrease, depending on the sign of the product $\lambda^*_{i1}\lambda^*_{i2}$. If $\lambda^*_{i1}\lambda^*_{i2} > 0$ (cf. the first part of (14)), $|\lambda_{i1}| > |\lambda^*_{i1}|$. If $\lambda^*_{i1}\lambda^*_{i2} < 0$ (cf. the second part of (14)), $|\lambda_{i1}| < |\lambda^*_{i1}|$. The overall change in the ℓ_1-norm of $\lambda_{\bullet 1}$ on $\mathcal{A}_1$ is thus approximately $|w_2| \, \beta^k$. The inequality inside the definition

of $\mathcal{I}^e_{\ell_1}$ thus ensures that $\lambda_{\bullet 1}$ has a higher ℓ_1-norm than $\lambda^*_{\bullet 1}$.

Dropping the assumption that the two loading vectors are orthogonal, we update our definition of the set $\mathcal{I}^e_{\ell_1}$ as follows:

Definition 2'. *Let $v_{\bullet k} = q_1\lambda^*_{\bullet 1} + q_2\lambda^*_{\bullet 2}$ for constants q_1 and q_2, such that $\|v_{\bullet k}\|_2^2 = n$ and $\lambda^*_{\bullet k} \perp v_{\bullet k}$.*[14] *Let*

$$\beta^k(v_{\bullet k}) = \left| \sum_{i \in \mathcal{A}_k} |v_{ik}| \, \mathbf{1}\{\lambda^*_{ik} v_{ik} > 0\} - \sum_{i \in \mathcal{A}_k} |v_{ik}| \, \mathbf{1}\{\lambda^*_{ik} v_{ik} < 0\} \right|. \tag{15}$$

Define the set

$$\mathcal{I}^e_{\ell_1} = \left\{ k \in \{1, 2\} : \|v^{\mathcal{A}_k^c}_{\bullet k}\|_1 > \beta^k(v_{\bullet k}) \right\}. \tag{16}$$

Again consider F_1. By definition of $\mathcal{A}_1$, $\|v^{\mathcal{A}_1^c}_{\bullet 1}\|_1 = q_2 \|\lambda^{*\mathcal{A}_1^c}_{\bullet 2}\|_1$, a constant times the sum of the absolute values of $\lambda^*_{\bullet 2}$ on $\mathcal{A}_1^c$. Thus, $1 \in \mathcal{I}^e_{\ell_1}$ if $\|\lambda^{*\mathcal{A}_1^c}_{\bullet 2}\|_1 > \frac{1}{q_2}\beta^1(v_{\bullet 1})$. Compared to Definition 2", dropping the requirement that $\frac{\Lambda^{*\prime}\Lambda^*}{n} = I$ therefore does not change the main idea: We require a lower bound on the sum (of the absolute values) of the loadings $\lambda^*_{\bullet 2}$ outside of the set $\mathcal{A}_1$ for 1 to be in $\mathcal{I}^e_{\ell_1}$.

The reason we work with $v_{\bullet 1}$ instead of $\lambda^*_{\bullet 2}$ once we drop the assumption that $\frac{\Lambda^{*\prime}\Lambda^*}{n}$ is the following: We would like to express all rotations of $\lambda^*_{\bullet k}$ in the subspace spanned by Λ^*. We achieve this by using $\lambda_{\bullet k} = w_1\lambda^*_{\bullet k} + w_2 v_{\bullet k}$, such that $\|v_{\bullet k}\|_2^2 = n$, $\lambda^*_{\bullet k} \perp v_{\bullet k}$ and $w_1^2 + w_2^2 = 1$. Rotating $\lambda^*_{\bullet 1}$ then amounts to adding a (small) non-zero weight w_2 to $\lambda_{\bullet 1} = w_1\lambda^*_{\bullet 1} + w_2 v_{\bullet 1}$ and the discussion below Definition 2" still applies.

If $r > 2$, $v_{\bullet k}$ will still be a linear combination of $\lambda^*_{\bullet l}$, $l = 1, \ldots, r$ and the intuition above directly extends to the case of $r > 2$. Before we consider this case, it is again instructive to consider a few specific examples maintaining $r = 2$:[15]

1. Suppose that $\lambda^*_{ik} \in \{-c_k, c_k\}$ for all $i \in \mathcal{A}_k$ and $\frac{\Lambda^{*\prime}\Lambda^*}{n} = I_2$.
   Then $\beta^k = 0$ (see Lemma 9). If at least one distinct outcome is unaffected by each factor (e.g., $\lambda^*_{11} = \lambda^*_{22} = 0$, and λ^*_{12} and λ^*_{21} are non-zero), then $1, 2 \in \mathcal{I}^e_{\ell_1}$.

[14] Note that $v_{\bullet k}$ is unique up to a possible sign indeterminacy (which is immaterial since both (15) and (16) involve taking absolute values).

[15] We also again consider a scenario where we treat the loadings as random variables instead. Suppose that $|\mathcal{A}_k| \asymp n$ for $k = 1, 2$, and $\lambda^*_{ik} \overset{i.i.d.}{\sim} N(0, \sigma)$ if $i \in \mathcal{A}_k$, and $\lambda^*_{ik} = 0$ otherwise. Then, $\beta^k = O_p(\sqrt{n}) = o_p(n)$. For a longer discussion, see Online Appendix D.

2. Suppose $\mathcal{A}_1 \cap \mathcal{A}_2 = \varnothing$ (The two factors affect different, non-overlapping groups of outcomes).
   Then, $\beta^k = 0$, and $1, 2 \in \mathcal{I}^e_{\ell_1}$.

3. Suppose $\mathcal{A}_2 \subseteq \mathcal{A}_1$ (The second factor F_2 affects a subset of the outcomes affected by F_1).
   Then, $|\mathcal{A}_1^c \cap \mathcal{A}_2| = 0$ implies that $\|\lambda^{*\mathcal{A}_1^c}_{\bullet 2}\|_1 = 0$, and $1 \notin \mathcal{I}^e_{\ell_1}$. Thus, whenever $\mathcal{A}_k$ is a superset of another active set $\mathcal{A}_l$, k cannot be a member of the set $\mathcal{I}^e_{\ell_1}$.

5.2 More Than Two Factors

Definition 2. *Let V_k denote the set of all linear combinations $v_{\bullet k}$ of $\lambda^*_{\bullet l}$, $l = 1, \ldots, r$, such that $\|v_{\bullet k}\|_2^2 = n$ and $\lambda^*_{\bullet k} \perp v_{\bullet k}$ and let*

$$\beta^k(v_{\bullet k}) = \left| \sum_{i \in \mathcal{A}_k} |v_{ik}| \, \mathbf{1}\{\lambda^*_{ik} v_{ik} > 0\} - \sum_{i \in \mathcal{A}_k} |v_{ik}| \, \mathbf{1}\{\lambda^*_{ik} v_{ik} < 0\} \right|. \tag{17}$$

Define the set

$$\mathcal{I}^e_{\ell_1} = \left\{ k \in \{1, \ldots, r\} : \|v^{\mathcal{A}_k^c}_{\bullet k}\|_1 - \beta^k(v_{\bullet k}) > 0 \; \forall v_{\bullet k} \in V_k. \right\}. \tag{18}$$

The only difference with Definition 2' is that the vector $v_{\bullet k}$ is no longer unique, and we thus require the inequality inside (18) to hold for all vectors $v_{\bullet k}$ that are orthogonal to $\lambda^*_{\bullet k}$ and have unit length.

In Online Appendix D, we show that the inequality in the definition of $\mathcal{I}^e_{\ell_1}$ holds with high probability for a variety of designs of the loading matrix. There, we repeatedly simulate different loading matrices, and record the fraction of realizations with $k \in \mathcal{I}^e_{\ell_1}$ to provide some guidance for how likely $k \in \mathcal{I}^e_{\ell_1}$ in practice.

Theorem 2. *Suppose Assumption 1 holds and we have access to a rotation of the true loading matrix, $\Lambda^0 = \Lambda^* H$, where H is nonsingular and $\frac{\Lambda^{0\prime}\Lambda^0}{n} = I$. If $k \in \mathcal{I}^e_{\ell_1}$, the minimization in* (12) *has a local minimum at $R^*_{\bullet k}$, where $R^*_{\bullet k}$ satisfies $\lambda^*_{\bullet k} = \Lambda^0 R^*_{\bullet k}$.*

Proof. Suppose $k \in \mathcal{I}^e_{\ell_1}$ and consider $\lambda_{\bullet k} = w_1 \lambda^*_{\bullet k} + w_2 v_{\bullet k}$, where $v_{\bullet k}$ is an arbitrary linear combination of $\lambda^*_{\bullet l}$, $l = 1, \ldots, r$, such that $\|v_{\bullet k}\|_2^2 = n$, $\lambda^*_{\bullet k} \perp v_{\bullet k}$ and $w_1^2 + w_2^2 = 1$. This can be thought of as considering all rotations of $\lambda^*_{\bullet k}$ in the subspace spanned by Λ^* without changing its length (in the standard ℓ_2-sense). Next, note that $v_{\bullet k} = v^{\mathcal{A}_k}_{\bullet k} + v^{\mathcal{A}_k^c}_{\bullet k}$, and therefore

$$\|\lambda_{\bullet k}\|_1 = \|w_1 \lambda^*_{\bullet k} + w_2 v_{\bullet k}\|_1 = \|w_1 \lambda^{*\mathcal{A}_k}_{\bullet k} + w_2 v^{\mathcal{A}_k}_{\bullet k}\|_1 + |w_2| \, \|v^{\mathcal{A}_k^c}_{\bullet k}\|_1$$

$$
\begin{aligned}
&= |w_1| \, \|\lambda_{\bullet k}^{*\mathcal{A}_k}\|_1 - |w_2| \, \beta^k(v_{\bullet k}) + |w_2| \, \|v_{\bullet k}^{\mathcal{A}_k^c}\|_1 \qquad (19) \\
&= |w_1| \, \|\lambda_{\bullet k}^{*}\|_1 + |w_2| \left(\|v_{\bullet k}^{\mathcal{A}_k^c}\|_1 - \beta^k(v_{\bullet k}) \right),
\end{aligned}
$$

where (19) follows from Lemma 3 when we consider a small neighborhood around $\lambda_{\bullet k}^*$. To make this explicit and align notation with that of Lemma 3, set $w_1 = \sqrt{1 - w_2^2}$ and $|w_2| = \epsilon$. We want to show that

$$
\begin{aligned}
& \sqrt{1-\epsilon^2} \|\lambda_{\bullet k}^*\|_1 + \epsilon \left(\|v_{\bullet k}^{\mathcal{A}_k^c}\|_1 - \beta^k(v_{\bullet k}) \right) > \|\lambda_{\bullet k}^*\|_1 \\
\Leftrightarrow \quad & \left(\sqrt{1-\epsilon^2} - 1 \right) \|\lambda_{\bullet k}^*\|_1 + \epsilon \left(\|v_{\bullet k}^{\mathcal{A}_k^c}\|_1 - \beta^k(v_{\bullet k}) \right) > 0 \\
\Leftrightarrow \quad & \frac{\sqrt{1-\epsilon^2} - 1}{\epsilon} + \frac{\|v_{\bullet k}^{\mathcal{A}_k^c}\|_1 - \beta^k(v_{\bullet k})}{\|\lambda_{\bullet k}^*\|_1} > 0.
\end{aligned}
$$

By Lemma 2, the first part can be made arbitrarily small for small enough values of ϵ. On the other hand, the second part is strictly positive by Definition 2 and does not depend on ϵ. We therefore conclude that for a small enough neighborhood, $\|\lambda_{\bullet k}\|_1 > \|\lambda_{\bullet k}^*\|_1$ for all rotations of $\lambda_{\bullet k}^*$ in the subspace spanned by Λ^* and hence $\|\lambda_{\bullet k}^*\|_1$ is a local minimum of (12). □

Theorem 2 states that $R_{\bullet k}^*$ (and thus $\lambda_{\bullet k}^*$) is a minimum of (12) if F_k is a local factor ($k \in \mathcal{I}_{\ell_1}^e$). Note that any set of local minima of (12) for $k = 1, \ldots, r$ is also a local minimum of (11). By imposing the additional constraint that R is nonsingular, we rule out that multiple columns in R lead to the same $\lambda_{\bullet k}^*$ and ensure that any solution $\tilde{\Lambda} = \Lambda^0 \tilde{R}$ to (11) spans the same space as Λ^0.

Remark 5. It is worth noting that Theorem 2 does not rule out other local minima. Ruling out other minima is possible but requires stronger assumptions on Λ^* that may be unlikely to be broadly applicable in practice. We illustrate this further in Online Appendix C. Since, in general, other local minima may exist, we turn to the question of how to select the "correct" minima from these local minima next.

Algorithm 1. *Given a rotation of the true loading matrix, $\Lambda^0 = \Lambda^* H$, where H is nonsingular and $\frac{\Lambda^{0\prime}\Lambda^0}{n} = I$:*

1. *Find all local minima $R_{\bullet k}^p$, $p = 1, \ldots, P$, to* (12) *and collect them in a matrix $R^P = [R_{\bullet k}^1 \cdots R_{\bullet k}^P]$ to form the "candidate matrix" $\bar{\Lambda} = \Lambda^0 R^P$.*

2. *Append this matrix with Λ^0 to yield the $n \times (P + r)$ matrix $\ddot{\Lambda} = [\bar{\Lambda} \; \Lambda^0]$.*

3. *Pick r columns of $\ddot{\Lambda}$ forming a $(n \times r)$ matrix $\tilde{\Lambda}$ such that $\|\tilde{\Lambda}\|_0$ is minimized and $rank(\tilde{\Lambda}) = r$.*

Our next result establishes that Algorithm 1 provably selects the correct local minima, meaning that the solution to Algorithm 1 will include the loading vectors $\lambda^*_{\bullet k}$ for any $k \in \mathcal{I}^e_{\ell_1} \cap \mathcal{I}^e_{\ell_0}$.

Corollary 2. *Suppose Assumption 1 holds. Let $\tilde{\Lambda}$ denote the solution to Algorithm 1. If $k \in \mathcal{I}^e_{\ell_1} \cap \mathcal{I}^e_{\ell_0}$, then $\tilde{\lambda}_{\bullet l}$=$\lambda^*_{\bullet k}$ for some $l = 1, \ldots, r$.*

Proof. By Theorem 1, if $k \in \mathcal{I}^e_{\ell_0}$, $\lambda^*_{\bullet k}$ is part of the global minimum of (7) (it minimizes the ℓ_0-norm across all rotations).

By Theorem 2, if $k \in \mathcal{I}^e_{\ell_1}$, $\lambda^*_{\bullet k}$ is a local minimum of (12). Thus, there exists a column $\ddot{\lambda}_{\bullet p}$ in $\ddot{\Lambda}$ such that $\ddot{\lambda}_{\bullet p} = \lambda^*_{\bullet k}$.

The third step of Algorithm 1 is a restricted version of (7) (since it restricts the feasible set to a discrete set of vectors collected in R^P), but by Theorem 2, $\lambda^*_{\bullet k}$ is a feasible solution. Thus, since $\lambda^*_{\bullet k}$ is both part of the global minimum of the ℓ_0-norm, and a feasible solution to the third step of Algorithm 1, it also minimizes the ℓ_0-norm across the set of permissible values in the third step of Algorithm 1. □

Theorems 1 and 2, as well as Corollary 2, required exact sparsity and did not allow for any estimation error in Λ^0. We therefore next relax our assumptions accordingly.

Assumption 2. *For each factor F_k, we can partition the set of indices $i = 1, 2, \ldots, n$ into a set of indices $\mathcal{A}_k$ with cardinality $|\mathcal{A}_k|$ and its complement, such that as $n \to \infty$,*

(a) $\sum_{i \notin \mathcal{A}_k} \left|\lambda^*_{ik}\right| = O(\sqrt{n})$.

(b) $|\mathcal{A}_k| > c_0 n$ *for some* $c_0 > 0$.

(c) $\left|\lambda^*_{ik}\right| > c \; \forall i \in \mathcal{A}_k$ *and* $\left|\lambda^*_{ik}\right| < C \; \forall i$ *for constants* $0 < c, C < \infty$.

(d) $\sup_{i \notin \mathcal{A}_k} \lambda^*_{ik} = o(\frac{1}{\log n})$.

(e) There exists a $c > 0$, such that $B = \{i : |\lambda^*_{i\bullet} w| \in (\frac{1-c}{\log n}, \frac{1+c}{\log n})\} = \emptyset$ for any fixed $(r \times 1)$ weight vector w.

Assumption 2(a) relaxes the definition of $\mathcal{A}_k$ to allow for approximate sparsity. We may still think of $\mathcal{A}_k$ as the active (or important) set for a given factor F_k, but F_k may now also affect other outcomes, with Assumption 2(a) restricting how much. Assumption 2(b) can

be thought of as a pervasiveness assumption. Together with Assumption 2(c), it states that each factor affects a constant fraction of all outcomes, which is commonly maintained in the literature. For the results that follow, we require access to a $\sqrt{n}$-consistent estimate of the space spanned by Λ^* and the ability to obtain such a $\sqrt{n}$-consistent estimate generally implies that factors must be pervasive (Freyaldenhoven 2022).

Assumption 2(d)-(e) are only needed for Theorem 4 to accommodate our chosen approximation to the ℓ_0-norm. Assumption 2(d) effectively introduces a gap between loadings on $\mathcal{A}_k$ and its complement, while Assumption 2(e) rules out knife-edge cases, where linear combinations of loading vectors are almost, but not quite collinear.

Definition 3. *Define the set*

$$\mathcal{I}_{\ell_1} = \left\{ k \in \{1, \dots, r\} : \|v_{\bullet k}^{\mathcal{A}_k^c}\|_1 - \beta^k(v_{\bullet k}) > c_{min} n^{\frac{3}{4}} \; \forall v_{\bullet k} \in V_k \right\} \tag{20}$$

for some $c_{min} > 0$ and $N < \infty$, whenever $n > N$.

The only difference between Definitions 2 and 3 is that we require a larger term on the RHS of the inequality in (20) in order to accommodate non-zero entries of $\lambda^*_{\bullet k}$ on $\mathcal{A}_k^c$ (though we note that both lower bounds are of order $o(n)$). Note that this implies that $\mathcal{I}_{\ell_1} \subseteq \mathcal{I}^e_{\ell_1}$.

So far, we assumed access to an initial rotation of Λ^*, $\Lambda^0 = \Lambda^* H$. In practice, we will only have access to an estimate of such a rotation. We remain agnostic about where such an initial estimate may come from but simply require $\sqrt{n}$-consistency.

Assumption 3. *We have access to an initial estimate Λ^0 with $\frac{\Lambda^{0\prime}\Lambda^0}{n} = I$, such that $(\lambda^0_{ik} - \lambda^*_{i\bullet} H_{\bullet k}) = O_p(\frac{1}{\sqrt{n}})$, where H is nonsingular and the elements in H^{-1} are bounded above by some constant $C < \infty$.*

An obvious candidate that achieves $\sqrt{n}$ consistency under some regularity conditions (and, in particular, under the assumption that both n and T converge to infinity, and suitable restrictions on the relative size of n and T) would be the Principal Component estimator (Stock and Watson 2002, Bai and Ng 2002, Bai 2003)[16]. This is the estimator we use in our simulations and applications.

[16] A large literature exists detailing various conditions on the primitives of the model (e.g., among others, the amount of correlation in the error term e) that allows an estimate with this rate. We also note that convergence rates are typically derived under a simultaneous limit for both n and T. See Bai and Ng (2021) for a more detailed discussion of the rotation matrix H, and Uematsu and Yamagata (2022) or Ando and Bai (2017) for examples of alternative estimators under various sparsity assumptions.

Theorem 3. *Suppose $n \to \infty$, Assumptions 2(a)-(c) and 3 hold, and $k \in \mathcal{I}_{\ell_1}$. Then, there exists a local minimum of* (12) *at $\bar{R}_{\bullet k}$, with $\bar{\lambda}_{\bullet k} = \Lambda^0 \bar{R}_{\bullet k}$, such that*

$$\bar{\lambda}_{ik} = \lambda^*_{ik} + O_p(n^{-1/4}) \qquad \textit{for each } i, \tag{21}$$

and

$$\frac{1}{n}\|\lambda^*_{\bullet k} - \bar{\lambda}_{\bullet k}\|^2 = O_p(n^{-\frac{1}{2}}). \tag{22}$$

Theorem 3 establishes that the minimization problem in (12) yields local minima that consistently estimate the loadings (and the individual loading vectors) of local factors, even under approximate sparsity and allowing for estimation error in the initial estimate Λ^0.

While we give a formal proof in the Online Appendix, we briefly outline the arguments here. Theorem 2 established that, under exact sparsity, and given a rotation $\Lambda^0 = \Lambda^* H$, a local minimum of the ℓ_1-norm across rotations of Λ^0 exists at $\lambda^*_{\bullet k}$ if $k \in \mathcal{I}^e_{\ell_1}$. In Lemma 7, we show that an approximate version of this result holds under approximate sparsity (Assumption 2): given a rotation $\Lambda^0 = \Lambda^* H$, a local minimum of the ℓ_1-norm across rotations of Λ^0 exists that is close to $\lambda^*_{\bullet k}$ if $k \in \mathcal{I}_{\ell_1}$. Then, we show that, under approximation error (cf. Assumption 3), there exists a proxy $\dot{\Lambda}$ that is close to Λ^*, such that we can write Λ^0 as a rotation of this proxy and this proxy loading matrix $\dot{\Lambda}$ is also approximately sparse. We then use this to show that, if $k \in \mathcal{I}_{\ell_1}$, a local minimum of the ℓ_1-norm across rotations of Λ^0 exists close to $\dot{\lambda}_{\bullet k}$. Since $\dot{\lambda}_{\bullet k}$ is close to $\lambda^*_{\bullet k}$, this completes the proof.

6 Combining ℓ_0-norm and ℓ_1-norm

We note that Remark 5 still applies and the number of local minima may be larger than the number of factors in $\mathcal{I}_{\ell_1}$. To choose among the local minima when allowing for estimation error and approximate sparsity, we use an approximate version of the ℓ_0-norm, which we denote by $\|\cdot\|_a$. Specifically, let $\|\Lambda\|_a$ denote the number of entries in Λ with $\big|\lambda_{ij}\big| > \frac{1}{\log n}$.

Algorithm 2. *Given a $\sqrt{n}$-consistent estimate Λ^0 that forms an orthonormal basis of the loading space:*

1. *Find all local minima $R^p_{\bullet k}$, $p = 1, \ldots, P$, to* (12) *and collect them in a matrix $R^P = [R^1_{\bullet k} \cdots R^P_{\bullet k}]$ to form the "candidate matrix" $\bar{\Lambda} = \Lambda^0 R^P$.*

2. *Append this matrix with Λ^0 to yield the $n \times (P + r)$ matrix $\ddot{\Lambda} = [\bar{\Lambda}\ \Lambda^0]$.*

3. *Pick r columns of $\ddot{\Lambda}$ forming a $(n \times r)$ matrix $\tilde{\Lambda}$ such that $\|\tilde{\Lambda}\|_a$ is minimized and $rank(\tilde{\Lambda}) = r$.*

Definition 4. *Let $\Lambda^*_{\bullet,-m}$ be the n by $(r-1)$ submatrix of Λ^* obtained by deleting the mth column in Λ^*. Given a $(r-1)$ vector of finite weights z, let $\mathcal{A}^*_{z,-m}$ be the set of outcomes for a linear combination $x = \Lambda^*_{\bullet,-m} z$ with $x_i > \frac{1}{\log n}$. Let $b^*_k(z) = \max |\mathcal{B}|$, $\mathcal{B} \subseteq \mathcal{A}_k$, such that*

$$\Lambda^{*\mathcal{A}_k}_{i,-k} z \in \left(\lambda^{*\mathcal{A}_k}_{ik} - \frac{1}{\log n}, \lambda^{*\mathcal{A}_k}_{ik} + \frac{1}{\log n} \right) \forall i \in \mathcal{B}. \tag{23}$$

Define the set $\mathcal{I}_{\ell_0} = \left\{ k \in \{1, \ldots, r\} : \left| \mathcal{A}^c_k \cap \mathcal{A}^*_{z,-k} \right| - b^*_k(z) > 0 \; \forall z \neq 0 \right\}$.

Definition 4 is a generalization of Definition 1, which required an exact equality in lieu of (23): $b_k = max_z b_k(z)$ is the size of the largest set of non-zero loadings on $\lambda^*_{\bullet k}$ that can be approximately replicated as a linear combination of the remaining loading vectors. A small value for b_k (e.g., $b_k = r - 1$) means this set is small, and intuitively states that the loading vector $\lambda^*_{\bullet k}$ is not too similar to the remaining loading vectors. In Online Appendix J.2, we include a simplified version of Definition 4 for the case of two factors, analogous to Definition 1'.

Comparing Definitions 1 and 4, we note that $\mathcal{I}_{\ell_0} \subseteq \mathcal{I}^e_{\ell_0}$, since $b^*_k(z) > b_k(z)$, and $\mathcal{A}^*_{z,-m} \subseteq \mathcal{A}_{z,-m}$. This allows us to accommodate the use of $\|\cdot\|_a$ instead of $\|\cdot\|_0$ in the final step of Algorithm 2. In particular, under approximate sparsity and estimation error (Assumptions 2 and 3), the set $\mathcal{I}_{\ell_0}$ defines the set of loading vectors that can be identified by minimizing the ℓ_a-norm. Since this result and its corresponding proof closely resemble those of Proposition 1 and Theorem 1, we relegate a formal statement with proof to the Online Appendix (see Theorem OA1 therein). We then combine Theorems OA1 and 3 for our final result below (Theorem 4), which is similar to Corollary 2 but allows for estimation error. It establishes that the solution to Algorithm 2 will converge to the loading vectors of any local factor F_k, defined as $k \in \mathcal{I}_{\ell_0} \cap \mathcal{I}_{\ell_1}$.

Theorem 4. *Suppose Assumptions 2 and 3 hold. Let $\tilde{\Lambda}$ denote the solution to Algorithm 2. If $k \in \mathcal{I}_{\ell_0} \cap \mathcal{I}_{\ell_1}$,*

$$\tilde{\lambda}_{il} = \lambda^*_{ik} + O_p(n^{-1/4}) \qquad \textit{for each } i, \tag{24}$$

and

$$\frac{1}{n}\|\lambda^*_{\bullet k} - \tilde{\lambda}_{\bullet l}\|^2 = O_p(n^{-\frac{1}{2}}) \tag{25}$$

for some $l = 1, \ldots, r$.

Proof. By Theorem OA1, if $k \in \mathcal{I}_{\ell_0}$, there exists an $\tilde{\lambda}_{\bullet l} = \lambda^*_{\bullet k} + O_p(\frac{1}{\sqrt{n}})$ that is part of the global minimum that minimizes the ℓ_a-norm across all rotations of Λ^0.

By Theorem 3, if $k \in \mathcal{I}_{\ell_1}$, $\lambda^*_{\bullet k}$ is close to a local minimum of the ℓ_1-norm across all rotations of Λ^0: there exists a column $\ddot{\lambda}_{\bullet p}$ in $\ddot{\Lambda}$ such that $\ddot{\lambda}_{ip} = \lambda^*_{ik} + O_p(n^{-1/4})$.

Combining the two results, there exists a column $\ddot{\lambda}_{\bullet p}$ in $\ddot{\Lambda}$ such that $\|\ddot{\lambda}_{\bullet p}\|_a = \|\lambda^*_{\bullet k}\|_a = \|\tilde{\lambda}_{\bullet l}\|_a$. Since the third step of Algorithm 2 is a restricted version of minimizing the ℓ_a-norm across all rotations, but achieves the global minimum, the result follows. □

Remark 6. Theorem 4 establishes identification for the loading vectors $\lambda^*_{\bullet k}$, $k \in \mathcal{I}_{\ell_0} \cap \mathcal{I}_{\ell_1}$. One shortcoming of this result is that it may not be obvious in practice which loading vectors are in this set (and are thus identified and convey structural information), and which ones are not. In Section 8, we discuss this further in the context of our empirical applications and present some heuristics on how to determine which factors are in this set. Formally identifying which loading vectors are in this set would be an interesting avenue for future research.

Before we conclude this section, it is worth noting that all results throughout concerned the loadings Λ^*. A natural question is to what extent our consistency results for the individual loading vectors translate into consistency results for the corresponding individual factor realizations. Perhaps surprisingly, the realizations of individual factors will generally not be identified. In other words, knowing a loading vector $\lambda^*_{\bullet k}$, and thus how the corresponding factor F_k affects all outcomes, is *not* sufficient to identify the corresponding factor realizations F_{kt}, $t = 1, \ldots, T$, without further assumptions.

For intuition, suppose we were to form estimates for the factor realizations at each time period by a cross-sectional regression of the outcomes on the estimated factor loadings, such that

$$F_t = (\tilde{\Lambda}'\tilde{\Lambda})^{-1}\tilde{\Lambda}'X_t \qquad \text{for } t = 1, \ldots, T. \tag{26}$$

Intuitively, consistency of F_{kt} requires knowledge of *all* loading vectors $\lambda^*_{\bullet k}$, $k = 1, \ldots, r$. Thus, this usually rules out the existence of any global factors. However, a setting in which all factors are local (in which case the entire loading matrix Λ would be identified, such that

$\tilde{\Lambda} \approx \Lambda^*$) appears unlikely in most economic applications. It may be possible to achieve identification of the individual factors under additional restrictions (such as orthogonality of the factors or the loadings). We leave this for future research.

7 Simulations

This section presents results from Monte Carlo simulations to evaluate the performance of our proposals in finite sample. We start by revisiting the baseline DGP from our stylized example in Section 3.1 and provide some more details about this DGP. The factors $F_k, k = 1, 2$ are generated jointly normal with a correlation of 0.3, unit variances, and are $i.i.d.$ over time. The error terms have the following correlation structure:

$$\begin{aligned} e_{ti} &= \rho e_{t-1,i} + (1-\rho^2)^{1/2} v_{it}, \\ v_{ti} &= \beta v_{t,i-1} + (1-\beta^2)^{1/2} u_{it}, \quad u_{it} \overset{i.i.d.}{\sim} N(0,1), \end{aligned}$$

with $(\rho, \beta) = (0.3, 0.1)$, which Onatski (2010) argues are good approximations to many financial datasets. We simulate 2,000 realizations of our baseline DGP. For each realization, we simulate new loadings in Λ^*. Our goal is to recover Λ^*.

To summarize the performance of an estimator across simulation runs, we use the cosine similarity between the columns in Λ^* and an estimate $\hat{\Lambda}$. Because the factors can always be reordered, for each true loading vector $\lambda^*_{\bullet l}$, we use the maximum cosine similarity with any estimated loading vector to measure how closely we are able to recover $\lambda^*_{\bullet l}$. Formally, define the maximum cosine similarity $MC_l(\hat{\Lambda})$ between the true loading vector $\lambda^*_{\bullet l}$ and an estimate $\hat{\Lambda}$ as

$$MC_l(\hat{\Lambda}) = \max_k \frac{\hat{\lambda}'_{\bullet k} \lambda^*_{\bullet l}}{\|\hat{\lambda}_{\bullet k}\| \|\lambda^*_{\bullet l}\|} \quad \text{for } l = 1, \ldots, r. \tag{27}$$

Thus, a value of MC_l close to one means that one of the estimated loading vectors $\hat{\lambda}_{\bullet k}$, $k = 1, \ldots, r$, is close to $\lambda^*_{\bullet l}$.

The maximum cosine similarity corresponding to Figures 1-4 in Section 3.1 is depicted in the first two columns of Table 1. The first column confirms that the Principal Component estimator does not successfully recover either of the two loading vectors. On the other hand, consistent with Figure 4, our proposed estimator can successfully identify the true loading matrix Λ^*. Because both factors symmetrically affect the same number of outcomes in our baseline DGP, the two rows look similar. While $\lambda^*_{ik} \overset{i.i.d.}{\sim} U(0.1, 2.9)$ for $i \in \mathcal{A}_k$ was chosen to

For $i \in \mathcal{A}_k$:	$\lambda_{ik}^* \sim U(0.1, 1.9)$		$\lambda_{ik}^* \sim N(1,1)$	
Estimator $\hat{\Lambda}$	Λ^0	$\tilde{\Lambda}$	Λ^0	$\tilde{\Lambda}$
MC_1	0.779	0.990	0.776	0.994
MC_2	0.777	0.990	0.773	0.994

Table 1: Maximum cosine similarity $MC_l(\hat{\Lambda})$ across DGPs and estimators. Λ^0 refers to the Principal Component estimator, while $\tilde{\Lambda}$ represents our proposed rotation that minimizes the ℓ_1-norm across all rotations. Depicted are averages based on 2,000 realizations.

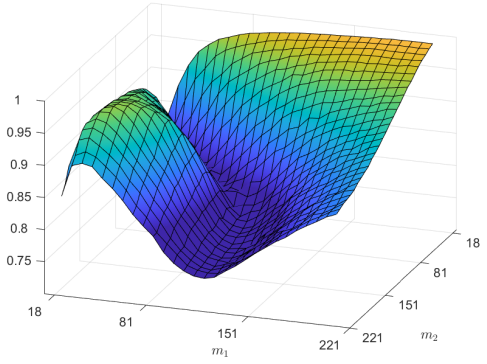
satisfy the upper and lower bounds assumed on Λ^* in the previous section, we also consider $\lambda_{ik}^* \overset{i.i.d.}{\sim} N(1,1)$ for $i \in \mathcal{A}_k$ in the third and fourth column of Table 1. Column 4 demonstrates that changing the distribution of the loadings λ_{ik}^* on $\mathcal{A}_k$ has no meaningful impact on our results. Finally, using $\mathbf{1}\{\hat{\mathcal{L}}_0(\tilde{\Lambda}) > \gamma n\}$ to determine whether local factors are present, we successfully detect the existence of local factors in all 2,000 simulation runs for this DGP.

The previous results confirm that our proposed ℓ_1-rotation and sparsity diagnostic work well in our baseline DGP: we can reliably detect the presence of local factors, and can correctly recover the sparsity pattern in the loading matrix, thereby identifying the individual loading vectors. We next consider a variety of data-generating processes to approximate a range of situations a practitioner might encounter in practice.

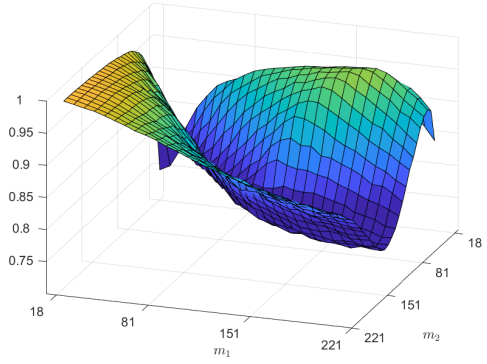
7.1 Results for a Variety of Data-Generating Processes

We next vary the degree of sparsity in the loading matrix by varying the values of $m_1 = |\mathcal{A}_1|$ and $m_2 = |\mathcal{A}_2|$. We maintain that $\lambda_{ik}^* \overset{i.i.d.}{\sim} N(1,1)$, for $i \in \mathcal{A}_k$. All other parameters remain unchanged from our baseline DGP. Figure 6 depicts how well we are able to estimate the true factor loadings $\lambda_{\bullet 1}^*$ and $\lambda_{\bullet 2}^*$ in this setting as a function of m_1 and m_2. Panels 6a and 6b depict the performance of the Principal Component estimator for $\lambda_{\bullet 1}^*$ and $\lambda_{\bullet 2}^*$. Unsurprisingly, the maximum cosine similarities are generally significantly below one. The exceptions to this are cases in which one factor is extremely weak. In such cases, the data effectively has a factor structure with a single factor, there is no rotational indeterminacy, and the sole strong factor is identified.

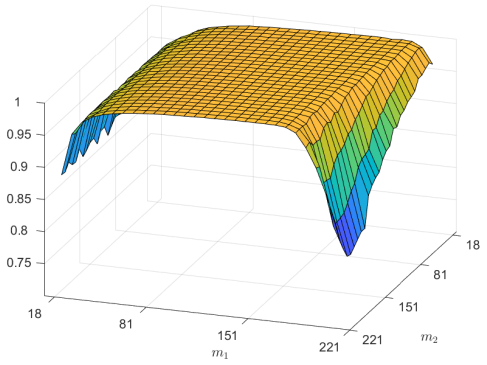
Panels 6c and 6d depict the maximum cosine similarity for our proposed estimate $\tilde{\Lambda}$. We are able to separately identify the two loading vectors throughout most of the parameter space using our ℓ_1-criterion. The exception occurs in the regions of the parameter space where a factor becomes either "global" or very weak. For example, along the right edge of Figure 6c, F_1 affects all observables. Since only the loading vectors corresponding to local factors are



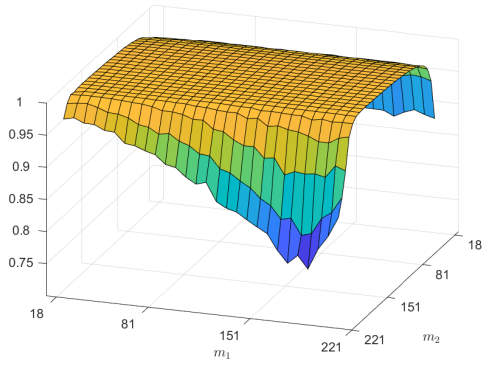
(a) PC Estimator: $MC_1(\Lambda^0)$



(b) PC Estimator: $MC_2(\Lambda^0)$



(c) Rotated Estimator: $MC_1(\tilde{\Lambda})$



(d) Rotated Estimator: $MC_2(\tilde{\Lambda})$

Figure 6: Maximum cosine similarity of estimators with each of the true loading vectors $\lambda^*_{\bullet k}$ as a function of the degree of sparsity in the loading matrix. m_k refers to the number of non-zero entries in the kth column of Λ^*. Figure based on our baseline DGP with $\lambda^*_{ik} \sim N(1,1)$ for $i \in \mathcal{A}_k$. Depicted are averages over 500 realizations.

identified, and clearly $1 \notin \mathcal{I}$ for any of the four defintions of local factors in this region, this is not surprising. On the opposite side of Figure 6c, only few outcomes are affected by F_1. $\lambda^*_{\bullet 1}$ is therefore only weakly identified, and our initial estimate of the loading space is poor, resulting in a maximum cosine similarity less than one. We further conclude from panels 6c-6d that an identification failure for one of the loading vectors does not imply identification failure for the other.

In Online Appendix G.1, we document the performance of our diagnostic to detect the presence of local factors. In summary, our method detects the existence of local factors in almost all realization whenever at least one loading vector has more than 25% of its entries equal to zero. On the other hand, under a dense DGP, our diagnostic suggests that no local factors are presented in at least 99% of all realizations.

We next increase the size of the model and consider a DGP with $(T, n) = (500, 300)$ and $r = 4$, with a small amount of correlation between the factors. Specifically, let $F_t \sim N(0, \Sigma_F)$, $i.i.d.$ over time, with

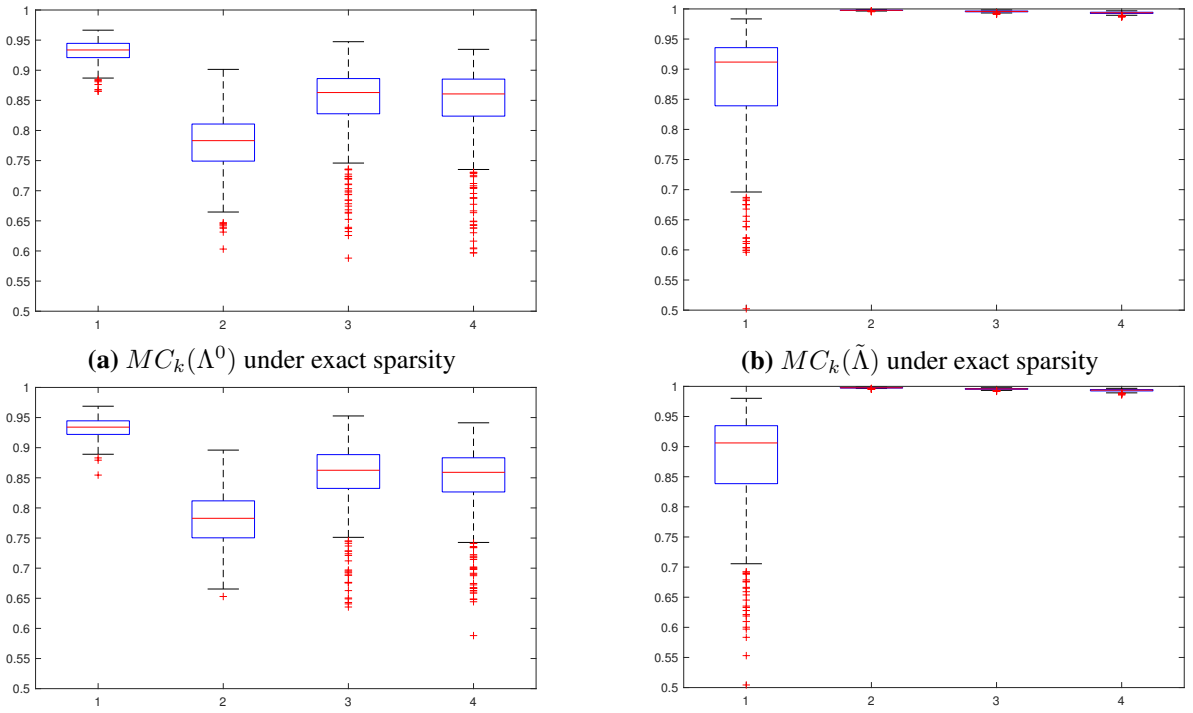
$$\Sigma_F = \begin{bmatrix} 1.0 & 0.3 & 0.0 & 0.0 \\ 0.3 & 1.0 & 0.3 & 0.0 \\ 0.0 & 0.3 & 1.0 & 0.3 \\ 0.0 & 0.0 & 0.3 & 1.0 \end{bmatrix}.$$

The first factor in this DGP is "global," while the remaining three are local to varying degrees. Specifically, the 300-by-4 loading matrix Λ^* has entries $\lambda^*_{ik} \overset{i.i.d.}{\sim} N(1, 1)$ if $i \in \mathcal{A}_k$, and $\lambda^*_{ik} = 0$ otherwise. The subsets $\mathcal{A}_k$ will be of varying size and dictate which variables are affected by each factor k, with the sequence of group sizes given by $\{|\mathcal{A}_k|\}_{k=1}^4 = \{300, 170, 96, 72\}$ for the four factors. The idiosyncratic component e_{it} is created the same way it was in our baseline DGP. Finally, we consider a variant of this DGP in which there is no exact sparsity, but rather an approximate version thereof. Here, $\lambda^*_{ik} \overset{i.i.d.}{\sim} N(0, \frac{1}{n})$, for all $i \in \mathcal{A}_k^c$.

Figure 7 uses a boxplot to visualize the performance of Λ^0 and $\tilde{\Lambda}$. It depicts the maximum cosine similarity for each factor across 500 realizations. The data underlying Figures 7a and 7b has an exact sparsity pattern ($\lambda^*_{ik} = 0$ if $i \in \mathcal{A}_k^c$). As expected, we do not consistently recover the true loadings using the Principal Component estimator Λ^0 (cf. Figure 7a). On the other hand, Figure 7b depicts the similarity between the truth, Λ^*, and our proposed estimate $\tilde{\Lambda}$. Since the first factor does not exhibit any sparsity, there is no information in the ℓ_1-norm that could help identify the corresponding loading vector. As a consequence, the similarity is below one, and identification fails for this loading vector. On the other hand, the loading vectors of the three local factors exhibit maximum cosine similarities that are visually indistinguishable from one in all realizations. Underlying Figures 7c and 7d is the variant of our DGP with approximate sparsity in the loading matrix. Based on Figures 7c and 7d, the above conclusions are unchanged. Our proposed estimator $\tilde{\Lambda}$ recovers the loading vectors associated with the three local factors in all realizations.[17]

In Online Appendix E, we compare the performance of our proposed estimator to a number of existing heuristics that are currently widely used to simplify the loading matrix, including some of the quartic criteria discussed in Section 3.2. We find that our ℓ_1-rotation performs better than these alternative methods.

[17] We also note that, for both DGPs (exact and approximate sparsity), our proposed diagnostic correctly detects the presence of local factors in all simulation runs.



(a) $MC_k(\Lambda^0)$ under exact sparsity

(b) $MC_k(\tilde{\Lambda})$ under exact sparsity

(c) $MC_k(\Lambda^0)$ under approximate sparsity

(d) $MC_k(\tilde{\Lambda})$ under approximate sparsity

Figure 7: Each panel depicts the maximum cosine similarity of an estimator with all four of the true loading vectors $\lambda^*_{\bullet k}$. Λ^0 denotes Principal Component estimator, while $\tilde{\Lambda}$ denotes estimate after proposed rotation. The first factor is global, factors 2-4 are local. Boxplots based on 500 realizations.

8 Applications

We next apply our rotation criterion to two economic applications in which factor models have been widely used, chosen to capture two scenarios a practitioner might encounter. First, we consider a dataset of international stock returns. Because of the global nature of this dataset, we expect the presence of region-specific factors in this dataset. We are therefore interested in whether our method can detect these local factors and recover the geographic structure of the data. Second, we consider a large panel of US macroeconomic indicators, where it is less clear a priori whether local factors are present.

8.1 International Asset Returns

Let R_{it} denote the return of asset i at time t. Following the Arbitrage Pricing Theory of Ross (1976) and Chamberlain and Rothschild (1983), we assume that unexpected returns $x_{it} = R_{it} - \mathbb{E}(R_i)$ follow a factor structure, such that

$$x_{it} = R_{it} - \mathbb{E}(R_i) = \lambda^*_{i\bullet} F_t + e_{it}. \tag{28}$$

We treat the common factors as unobserved, so we need to replace F_t and $\lambda_{i\bullet}^*$ by their estimates $\hat{F}_t$ and $\hat{\lambda}_{i\bullet}$. In financial economics, these estimates are commonly obtained by Principal Component analysis (Connor and Korajczyk 1986, Ludvigson and Ng 2007). We propose to identify the individual loading vectors using our ℓ_1-criterion.

Our dataset consists of daily returns for a large number of stocks from different parts of the world. In particular, it includes individual stock returns for companies that were part of the *DAX30* (Germany), the *FTSE100* (UK), the *S&P100* (US), the *CAC40* (France), or the *TA100* (Middle East) on April 23, 2015.[18] In total, the data covers 272 stocks spanning 687 observations from 01/01/2011 until 03/20/2015. We determine the number of factors to be eight using Bai and Ng (2002)'s Information Criterion with $r_{max} = 25$, and will accordingly use $r = 8$ in what follows.

To estimate the space spanned by these eight factors, we then estimate the leading eight principal components. Unsurprisingly, we find that each of the eight principal components loads on most of the 272 individual stocks. The estimated loadings corresponding to the Principal Component estimator Λ^0 can be found in Online Appendix Figure 11.

In contrast, Figure 8 depicts our proposed estimator $\tilde{\Lambda}$. The thin dashed lines separate the geographical groups as described above, in the order of Frankfurt, London, New York, Paris, and Tel Aviv. In contrast to the Principal Component estimate, we see that its loading vectors are highly concentrated on a subset of outcomes.[19] It reveals strong regional dependencies in asset returns as illustrated in Table 2. For example, $\tilde{\lambda}_{\bullet 1}$ is almost entirely concentrated on stocks in the US (similar to $\tilde{\lambda}_{\bullet 4}$), $\tilde{\lambda}_{\bullet 2}$ is concentrated on stocks in the Middle East, and $\tilde{\lambda}_{\bullet 3}$ is concentrated on stocks in the UK. $\tilde{\lambda}_{\bullet 5}$,$\tilde{\lambda}_{\bullet 6}$ and $\tilde{\lambda}_{\bullet 8}$ are European factors. The only loading vector without a strong geographical pattern is $\tilde{\lambda}_{\bullet 7}$. However, all 25 stocks with a loading larger than two for this factor belong to the *Oil & Gas* or the *Mining* sector, enabling us to clearly label this a sector-specific factor.

Revisiting Remark 6, we need to be cautious about which of the estimated loading vectors can be interpreted. Some loading vectors have both a significant sparsity pattern and distinct active sets (e.g., the second and fifth, concentrated on the Middle East and the Natural Resource sector, respectively). Thus, $k \in \mathcal{I}_{\ell_0} \cap \mathcal{I}_{\ell_1}$ seems plausible for $k \in \{2, 5\}$. In

[18] We further restrict the stocks in the TA100 to those with a weight by market capitalization in the TA100 of at least 0.5%. This makes the remaining stocks comparable in size to the rest of the sample. For a more detailed discussion of the data, see Online Appendix I.

[19] We again stress that there is no "shrinkage" involved in our estimator, such that our sparse representation of the factors fits the data exactly as well as a rotation with dense loadings. This also implies that none of the estimated loadings in $\tilde{\Lambda}$ will be exactly equal to zero. A further regularization step is beyond the scope of this paper. See Pelger and Xiong (2021) for a potential approach to such regularization.

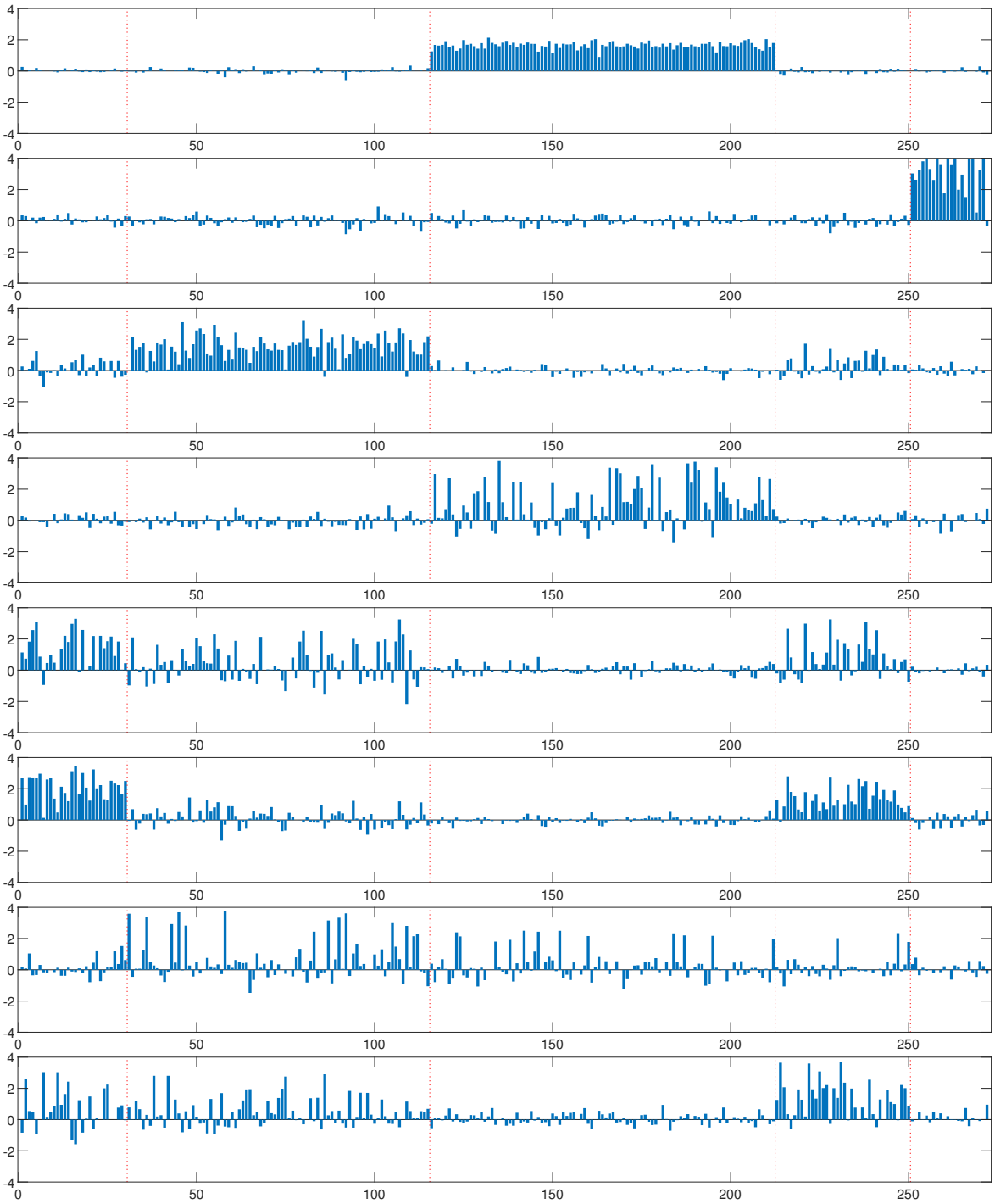


Figure 8: Illustration of the rotated loading vectors $\tilde{\lambda}_{\bullet k}$ for $k = 1, \ldots, 8$ in panel of international asset returns. Bars correspond to the loadings of the 272 individual stocks for the kth estimated loading vector. Geographical groups are Germany, UK, US, France, and Middle East, separated by dashed lines.

Factor	Region	Sector
1	US	
2	Middle East	
3	UK	
4	US	
5	Germany, UK, France	
6	Germany, France	
7	Global	Natural Resources (Oil and Mining)
8	Germany, UK, France	

Table 2: Interpretation of individual factors in panel of international asset returns, based on estimated loading matrix $\tilde{\Lambda}$.

other cases (e.g., the first and fourth, both concentrated in the US), loading vectors have very similar active sets (i.e. $\mathcal{A}_1 \approx \mathcal{A}_4$), suggesting that the corresponding loading vectors are still identified jointly, but not separately (cf. Remark 3). While this suggests that there are two US specific factors, a further labeling of these individual factors would be unwarranted.

Finally, because we find a rotation in which all columns of $\tilde{\Lambda}$ exhibit significant sparsity, our diagnostic suggests the existence of local factors in this dataset.

8.2 Macroeconomic Indicators

We next apply our identification strategy to a large panel of US macroeconomic indicators. In particular, we use the FRED-QD data collected and maintained by Michael W. McCracken.[20] Our final sample contains 206 quarterly observations of 166 macroeconomic variables.

Two papers that have looked into the nature of the optimal forecasting model in the context of a very similar dataset are De Mol et al. (2008) and Giannone et al. (2021). Both papers investigate how forecasts that use sparsity-inducing regularization compare to regularization methods that do not lead to a sparsity pattern in the predictors (such as ridge regressions or factor-augmented regressions). Specifically, De Mol et al. (2008) make the following two observations:

1. "The high correlation of the Lasso forecast with the PC forecast suggests that our data is highly collinear: Under collinearity, when appropriately selected, a few variables should capture the

[20]Data are available at `https://research.stlouisfed.org/econ/mccracken/fred-databases`. Versions of this dataset have been used extensively in the literature on macroeconomic forecasting (De Mol et al. 2008, Stock and Watson 2016). For a full description of the data, we refer the reader to McCracken (2019). We use data from 1967Q1-2019Q1 and follow the transformations of the raw data as outlined in McCracken and Ng (2016) to achieve stationarity and remove a small number of outliers. We use only the disaggregated time series in our estimation of the factor structure and disregard the aggregates (Boivin and Ng 2006, Stock and Watson 2016) and drop a small number of series with missing observations.

essence of the covariation of the data and, as principal components, span approximately the space of the common factors."

2. "The selection [of variables by the Lasso] is different at different points in the sample, although selected variables generally belong to the same economic category."

Similarly, Giannone et al. (2021) find a significant sparsity pattern for the predictors (more than 75% of their regressors have a coefficient of zero), but note that there is substantial uncertainty about the identity of the relevant predictors. These observations can be rationalized by the presence of local factors, with each factor affecting only a subset of the observed indicators (which will generally belong to the same economic category). In such a setting, a regularized estimator that induces sparsity in the individual components of X_t will tend to select a single variable from each group $\mathcal{A}_k$ as a noisy proxy for F_{kt}. The selected set of regressors will thus approximately span the space of the common factors. However, selection within groups will be unstable and sensitive to minor perturbations of the data, thereby leading to varying variable selection from the same groups across subsamples or posterior draws.

To determine whether a "group structure" with local factors is present in this application, we first determine the number of factors to be eight, using the Information Criterion of Bai and Ng (2002) with $r_{max} = 15$, and accordingly use $r = 8$ in what follows. To estimate the space spanned by these eight factors, we then estimate the leading eight principal components. Unsurprisingly, these load on most of the 166 observed outcomes. The estimated loadings using the Principal Component estimator Λ^0 can be found in Online Appendix Figure 10. In contrast, Figure 9 depicts our proposed estimator $\tilde{\Lambda}$. In order to gain an understanding of the factors, Table 3 reproduces the grouping of variables as suggested in McCracken (2019), which is in turn based on Stock and Watson (2012). The corresponding groups of variables are separated by dashed lines in Figure 9.

The first factor almost exclusively drives all price variables (group 6), allowing an easy interpretation as an aggregated price index of which we observe multiple measurements. The second factor mainly affects a combination of interest rates, employment indicators, and industrial production. The third factor is mainly associated with household balance sheets and stock markets (groups 10 and 13), capturing the intuitive notion that an increase in asset prices will be associated with an improvement in household balance sheets. Accordingly, almost all of those indicators are associated with positive loadings, with the exception of the dividend yield, which has a large negative loading. The picture is less clear for subsequent factors, and we therefore refrain from interpreting additional factors. However, we note that

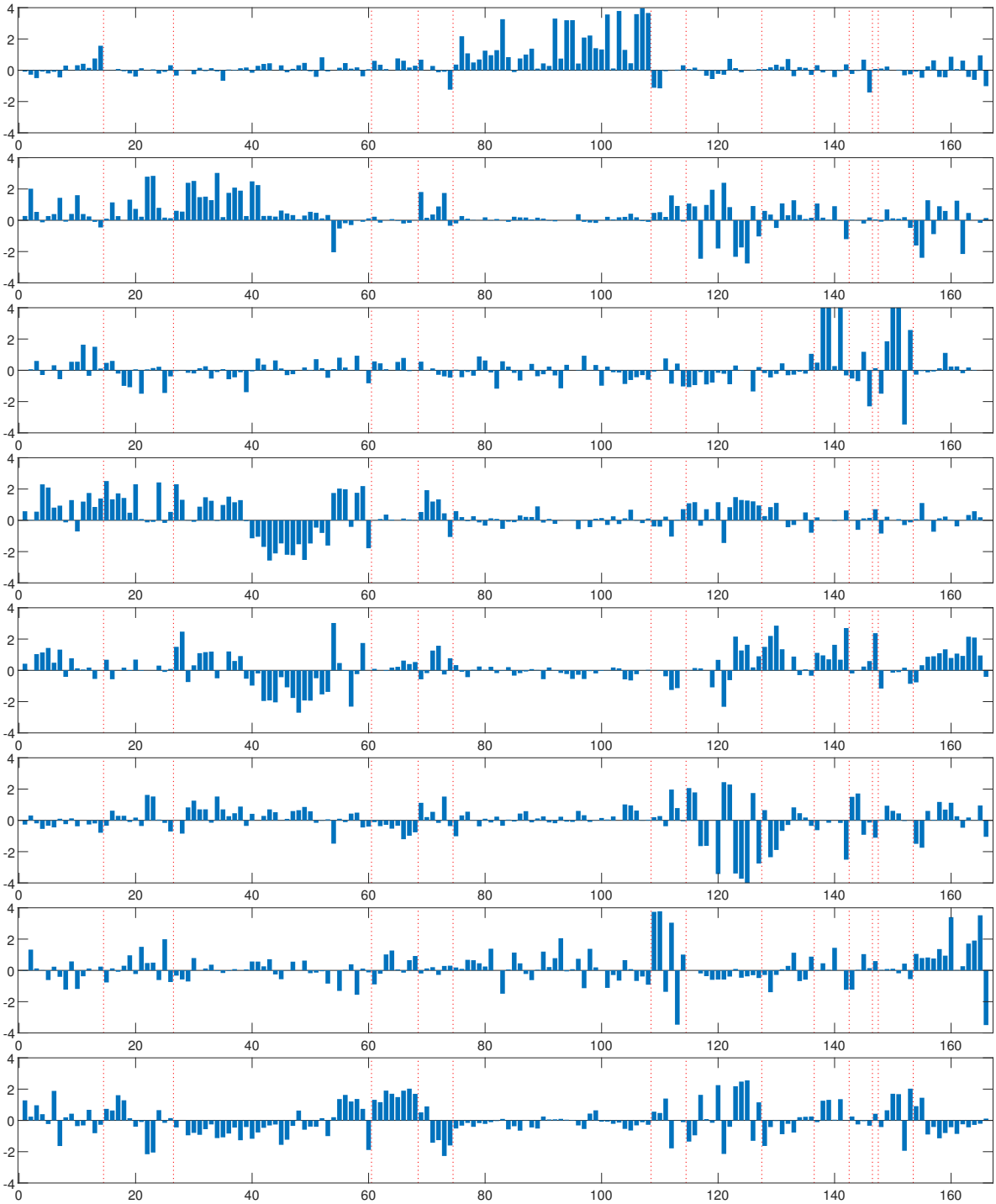


Figure 9: Illustration of the rotated loading vectors $\tilde{\lambda}_{\bullet k}$ for $k = 1, \ldots, 8$ in panel of macroeconomic indicators. Bars correspond to the 166 individual indicators for the kth estimated loading vector. Groups of variables are separated by dashed lines (see Table 3).

Group	Category	Associated Variables
1	National Income and Product Accounts (NIPA)	1-14
2	Industrial Production	15-26
3	Employment and Unemployment	27-60
4	Housing	61-68
5	Inventories, Orders, and Sales	69-74
6	Prices	75-108
7	Earnings and Productivity	109-114
8	Interest Rates	115-127
9	Money and Credit	128-136
10	Household Balance Sheets	137-142
11	Exchange Rates	143-146
12	Other	147
13	Stock Markets	148-153
14	Non-Household Balance Sheets	154-166

Table 3: Grouping of variables in panel of US macroeconomic indicators.

our testing criterion again suggests the existence of local factors in this dataset.

9 Conclusion

We introduce a new rotation criterion to simplify the loading matrix in factor models. Our rotation criterion minimizes the ℓ_1-norm of the loadings and is theoretically appealing. Unlike existing heuristics, such as the Varimax criterion (Kaiser 1958), we derive theoretical guarantees for our rotation criterion if the true loading matrix is sparse: Under (approximate) sparsity in the loading matrix, our ℓ_1-rotation can be used to identify the individual loading vectors.

Our ℓ_1-rotation criterion performs well across simulations and two economic applications. In our two applications, we find strong evidence that local factors are indeed present in the data in both cases. In both applications our method estimates sensible economic objects, which a researcher would not be able to recover otherwise.

References

Seung C Ahn and Alex R Horenstein. Eigenvalue ratio test for the number of factors. *Econometrica*, 81(3):1203–1227, 2013.

Tomohiro Ando and Jushan Bai. Clustering huge number of financial time series: A panel data approach with high-dimensional predictors and factor structures. *Journal of the American Statistical Association*, 112(519):1182–1198, 2017.

Jushan Bai. Inferential theory for factor models of large dimensions. *Econometrica*, 71(1): 135–171, 2003.

Jushan Bai and Serena Ng. Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221, 2002.

Jushan Bai and Serena Ng. Principal components estimation and identification of static factors. *Journal of Econometrics*, 176(1):18–29, 2013.

Jushan Bai and Serena Ng. Approximate factor models with weaker loadings. arXiv preprint, arxiv:2109.03773, 2021.

Dimitris Bertsimas, Angela King, and Rahul Mazumder. Best subset selection via a modern optimzation lens. *The Annals of Statistics*, 44(2):813–852, 2016.

Jean Boivin and Serena Ng. Are more data always better for factor analysis? *Journal of Econometrics*, 132(1):169–194, 2006.

Michael W Browne. An overview of analytic rotation in exploratory factor analysis. *Multivariate behavioral research*, 36(1):111–150, 2001.

Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: Methods, theory and applications*. Springer Science & Business Media, 2011.

John B Carroll. An analytical solution for approximating simple structure in factor analysis. *Psychometrika*, 18:23–38, 1953.

Gary Chamberlain and Michael Rothschild. Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica*, 51(5):1281–304, 1983.

Le-Yu Chen and Sokbae Lee. Sparse quantile regression. *Journal of Econometrics*, 235(2): 2195–2217, 2023.

In Choi, Dukpa Kim, Yun Jung Kim, and Noh-Sun Kwark. A multilevel factor model: Identification, asymptotic theory and applications. *Journal of Applied Econometrics*, 33(3): 355–377, 2018.

Gregory Connor and Robert A Korajczyk. Performance measurement with the arbitrage pricing theory: A new framework for analysis. *Journal of Financial Economics*, 15(3): 373–394, 1986.

Christine De Mol, Domenico Giannone, and Lucrezia Reichlin. Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components? *Journal of Econometrics*, 146(2):318–328, 2008.

Thorsten Drautzburg and Jonathan H Wright. Refining set-identification in vars through independence. *Journal of Econometrics*, 235(2):1827–1847, 2023.

Simon Freyaldenhoven. Factor models with local factors—determining the number of relevant factors. *Journal of Econometrics*, 229(1):80–102, 2022.

Domenico Giannone, Michele Lenza, and Giorgio E Primiceri. Economic predictions with big data: The illusion of sparsity. *Econometrica*, 89(5):2409–2437, 2021.

Christian Gouriéroux, Alain Monfort, and Jean-Paul Renne. Statistical inference for independent component analysis: Application to structural var models. *Journal of Econometrics*, 196(1):111–126, 2017.

Alan E Hendrickson and Paul Owen White. Promax: A quick method for rotation to oblique simple structure. *British Journal of Statistical Psychology*, 17(1):65–70, 1964.

Jian Huang, Yuling Jiao, Yanyan Liu, and Xiliang Lu. A constructive approach to l_0 penalized regression. *Journal of Machine Learning Research*, 19(10):1–37, 2018.

Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.

Robert I Jennrich. Rotation to simple loadings using component loss functions: The oblique case. *Psychometrika*, 71:173–191, 2006.

Ian T Jolliffe, Nickolay T Trendafilov, and Mudassir Uddin. A modified principal component technique based on the lasso. *Journal of Computational and Graphical Statistics*, 12(3): 531–547, 2003.

Henry F Kaiser. The Varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23:187–200, 1958.

Jeffrey Owen Katz and F James Rohlf. Functionplane – a new approach to simple structure rotation. *Psychometrika*, 39:37–51, 1974.

Sylvia Kaufmann and Christian Schumacher. Bayesian estimation of sparse dynamic factor models with order-independent and ex-post mode identification. *Journal of Econometrics*, 210(1):116–134, 2019.

Johannes T Kristensen. Diffusion indexes with sparse loadings. *Journal of Business & Economic Statistics*, 35(3):434–451, 2017.

Sydney C Ludvigson and Serena Ng. The empirical risk–return relation: A factor analysis approach. *Journal of Financial Economics*, 83(1):171–222, 2007.

Sydney C Ludvigson and Serena Ng. Macro factors in bond risk premia. *Review of Financial Studies*, 22(12):5027–5067, 2009.

Michael W McCracken. FRED-QD updated appendix. Working paper, Federal Reserve Bank of St. Louis, 2019.

Michael W McCracken and Serena Ng. FRED-MD: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, 34(4):574–589, 2016.

Emanuel Moench, Serena Ng, and Simon Potter. Dynamic hierarchical factor models. *The Review of Economics and Statistics*, 95(5):1811–1817, 2013.

Alexei Onatski. Determining the number of factors from empirical distribution of eigenvalues. *The Review of Economics and Statistics*, 92(4):1004–1016, 2010.

Markus Pelger and Ruoxuan Xiong. Interpretable sparse proximate factors for large dimensions. *Journal of Business & Economic Statistics*, 40(4), 2021.

Veronika Ročková and Edward I George. Fast Bayesian factor analysis via automatic rotations to sparsity. *Journal of the American Statistical Association*, 111(516):1608–1622, 2016.

Stephen A Ross. The arbitrage theory of capital asset pricing. *Journal of Economic Theory*, 13(3):341–360, 1976.

William W Rozeboom. Theory & practice of analytic hyperplane optimization. *Multivariate Behavioral Research*, 26(1):179–197, 1991.

David R Saunders. Trans-Varimax-some properties of the ratiomax and Equamax criteria for blind orthogonal rotation. *American Psychologist*, 17(6):395–396, 1962.

James H Stock and Mark W Watson. Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics*, 20(2):147–162, 2002.

James H Stock and Mark W Watson. Disentangling the channels of the 2007-09 recession. *Brookings Papers on Economic Activity*, (1):81–135, 2012.

James H Stock and Mark W Watson. Dynamic factor models, factor-augmented vector autoregressions, and structural vector autoregressions in macroeconomics. *Handbook of Macroeconomics*, 2:415–525, 2016.

Yoshimasa Uematsu and Takashi Yamagata. Estimation of sparsity-induced weak factor models. *Journal of Business & Economic Statistics*, 41(1), 2022.

Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286, 2006.