Federal Reserve Bank of Philadelphia

Working Papers
Research Department

WP 20-25
June 2020
https://doi.org/10.21799/frbp.wp.2020.25

# Identification Through Sparsity in Factor Models

**Simon Freyaldenhoven**
Federal Reserve Bank of Philadelphia Research Department

**ISSN:** 1962-5361
**Disclaimer:** This Philadelphia Fed working paper represents preliminary research that is being circulated for discussion purposes. The views expressed in these papers are solely those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Philadelphia or the Federal Reserve System. Any errors or omissions are the responsibility of the authors. Philadelphia Fed working papers are free to download at: https://philadelphiafed.org/research-and-data/publications/working-papers.

# Identification Through Sparsity in Factor Models

Simon Freyaldenhoven

Federal Reserve Bank of Philadelphia*

June 16, 2020

**Abstract**

Factor models are generally subject to a rotational indeterminacy, meaning that individual factors are only identified up to a rotation. In the presence of local factors, which only affect a subset of the outcomes, we show that the implied sparsity of the loading matrix can be used to solve this rotational indeterminacy. We further prove that a rotation criterion based on the $\ell_1$-norm of the loading matrix can be used to achieve identification even under approximate sparsity in the loading matrix. This enables us to consistently estimate individual factors, and to interpret them as structural objects. Monte Carlo simulations suggest that our criterion performs better than widely used heuristics, and we find strong evidence for the presence of local factors in financial and macroeconomic datasets.

JEL codes: C38, C55

KEYWORDS: identification, factor models, sparsity, local factors

*Email: simon.freyaldenhoven@phil.frb.org. I thank Jushan Bai, Chris Hansen, Adam McCloskey, Jesse Shapiro, and seminar audiences at Brown University, the Bundesbank and the Federal Reserve Bank of Philadelphia, as well as participants of the 2019 Greater New York Area Econometrics Colloquium and the 2019 European Conference of the Econometrics Community for helpful comments and suggestions. Le Xu provided excellent research assistance.

**Disclaimer:** This Philadelphia Fed working paper represents preliminary research that is being circulated for discussion purposes. The views expressed in these papers are solely those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Philadelphia or the Federal Reserve System. Any errors or omissions are the responsibility of the authors. Philadelphia Fed working papers are free to download at https://philadelphiafed.org/research-and-data/publications/working-papers

# 1 Introduction

Factor models are subject to a rotational indeterminacy, meaning that the individual factors are only identified 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 [...]."

Similarly, Ludvigson and Ng (2009) state,

> "Because the factors are identifiable only up to an $r \times r$ matrix, a detailed interpretation of the individual factors would be inappropriate. Moreover, we caution that any labeling of the factors is imperfect, because each is influenced to some degree by all the variables in our large dataset and the orthogonalization means that no one of them will correspond exactly to a precise economic concept like output or unemployment, which are naturally correlated. Nonetheless, it is useful to show that the factors capture relevant macroeconomic information. We do so here by briefly characterizing the factors as they relate to the underlying variables in our panel dataset. Figure 1 shows that the first factor loads heavily on measures of employment and production [...]."

I demonstrate in this paper that a framework with local factors can solve this indeterminacy, allowing a researcher to recover individual factors, instead of only the factor space. Local factors are factors that only affect a subset of the observables (Ando and Bai 2017, Freyaldenhoven 2019). Economic examples are industry-specific shocks in a firm-level dataset. To achieve identification, we introduce a criterion based on the $\ell_1$-norm of the loading matrix, which is minimized when local factors are individually identified. This criterion allows us to consistently estimate the individual factors, enabling their economic interpretation as structural objects. The key insight is that, if the true loading matrix $\Lambda$ is sparse, any other rotation $\tilde{\Lambda} = \Lambda R$ will be less sparse.

[1] The two papers have a combined citation count of more than 3000 as of May 2019.

In some instances, it is sufficient to identify the factor space, and the rotational indeterminacy of the estimated factors is inconsequential, though even in those cases identifying the individual factors can facilitate a better understanding of the data. In other cases, there are important empirical implications of this rotational indeterminacy. For example, extracted factors are often used as right-hand side variables, among others in factor-augmented VARs (Ludvigson and Ng 2007). The rotational indeterminacy of the estimated factors is only inconsequential if the entire factor space is used in the conditioning set of a linear model. If only a subset of the extracted factors, or non-linear transformations of the extracted factors are used, this result clearly no longer holds.

There are both appealing economic reasons and statistical evidence for the presence of local factors and thus the presence of zeros in the loading matrix (see, for example, Freyaldenhoven (2019), De Mol et al. (2008); we expand on this discussion in Section 6).

Despite the large literature on both factor models and sparsity, little work has been done on the intersection of the two. Arguably one reason is that most sparsity patterns in the loading matrix are not invariant to rotations. Kristensen (2017) considers adding an $\ell_1$-penalty to the estimation of the loading matrix via principal components to produce "sparse" loadings. However, even if the true loadings are sparse, the chosen rotation (normalization) inherent to the principal component estimator means that the estimated loadings will generally no longer be sparse even in population. It is therefore crucial to first identify the "correct" rotation of the loading matrix before any regularized estimation of the loadings is conducted.

Other previous works aimed at incorporating variable selection into factor models include Bai (2008), Hirose and Konishi (2012), Caner and Han (2014), and Cheng et al. (2016). The method proposed by Bai (2008) involves two separate estimation steps, one to identify the relevant observables, and one to estimate the factor structure. Therefore, once the relevant observables are selected, all of those variables will be affected by all factors. In Hirose and Konishi (2012), Caner and Han (2014), and Cheng et al. (2016), the sparsity assumptions take a form closely related to the Group Lasso (Yuan and Lin 2006) by setting entire columns of the loading matrix to zero. These methods effectively reduce the number of estimated factors. In contrast, in our model, each factor may affect a (potentially different) subset of the observables.

Work on the estimation of sparse principal components in the statistics literature is naturally related to the estimation of sparse factors. Early work in this direction can be found in Jolliffe et al. (2003) and Zou et al. (2006). However, since Principal Component Analysis does not depend on any probability model, it is quite distinct from factor analysis. 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.

Finally, a number of papers have assumed a known group structure (e.g., Boivin and Ng 2006, Wang 2008, Dias et al. 2013, Moench et al. 2013, Choi et al. 2018). Unlike those papers, we do

not require the group structure and factors' strengths to be known a priori. Ando and Bai (2017) and Freyaldenhoven (2019) also do not require knowledge of the group structure a priori, but the focus of the first paper is on estimation of the factor space, and the focus of the second paper is on estimating the number of factors. Neither addresses identification of individual factors.

Bai and Ng (2013) provides an overview of the different normalizations commonly used in economics. Of course, a choice of normalization is equivalent to a choice of rotation for the estimated factors. In practice, the choice of rotation often appears somewhat arbitrary. We suspect that this is at least partially due to the limited availability of formal results. We are unaware of attempts to use the sparsity pattern (or the $\ell_1$-norm) of the loading matrix to either choose a normalization or to identify individual factors in the econometrics literature. However, there is a large strand of literature in psychology that considers rotation criteria aimed to simplify the loading matrix, going back to at least Carroll (1953) and Kaiser (1958) (also see Katz and Rohlf 1974, Rozeboom 1991, Jennrich 2006). The focus of this literature is usually on the algorithmic implementation of the optimization, and consistency results are only derived for special cases. Section 3 provides a more detailed comparison of our approach to these methods.

Throughout, we will consider rotations in which we keep the $\ell_2$-norm of $\Lambda$ constant. Thus, our work is also related to recent work in statistics that considers the difference between the $\ell_1$ and $\ell_2$ norm as a penalty term in regularized estimation (Yin et al. 2014, Yin et al. 2015).

The paper proceeds as follows. After setting up our model and fixing notation in Section 2, we first provide an intuitive illustration of our proposed criterion in Section 3. We then show in Section 4 that the true loading matrix $\Lambda^*$ is the unique minimum of the $\ell_0$-norm across rotations of the loading matrix under a sparsity assumption. However, minimizing the $\ell_0$-norm directly will generally be infeasible in practice for two reasons. First, this requires exact (rather than approximate) sparsity in the loadings. In fact, even under exact sparsity, any estimation error will generally mean that there are no exact zeros in the estimated loading matrix. Second, minimizing the $\ell_0$-norm directly will generally be computationally prohibitive. One can compare this to high-dimensional sparse linear regression models, where optimal subset selection is similarly infeasible. On the other hand, a vast body of literature exists documenting both the theoretical and practical appeal of using the $\ell^1$-norm instead as a regularization in linear regression models (Tibshirani 1996, Bühlmann and Van De Geer 2011, Belloni and Chernozhukov 2011). We similarly propose to minimize the $\ell^1$-norm of $\Lambda$ to make our approach feasible and establish the consistency of this approach. Despite the resemblance to regularized estimation methods, we emphasize that there is no "shrinkage" involved in our estimator. Instead, we use the $\ell_1$-norm as a criterion to select the most sparse loading matrix $\Lambda$ from among a set of rotations. Section 5 provides Monte Carlo evidence that supports our asymptotic results in finite sample. Finally, we apply our identification strategy to a panel of US macroeconomic indicators as well as a panel of individual stock returns

in Section 6, and identify and interpret the resulting local factors.

# 2 Preliminaries

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

$$\underset{(n\times 1)}{X_t} = \underset{(n\times r)}{\Lambda^*}\underset{(r\times 1)}{F_t} + \underset{(n\times 1)}{e_t}\ \forall t, \qquad \text{or more compactly,} \qquad \underset{(T\times n)}{X} = \underset{(T\times r)}{F}\underset{(r\times n)}{\Lambda^{*\prime}} + \underset{(T\times n)}{e}, \tag{1}$$

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. We use the running indices $s, t$ for the $T$ observations, $i, j$ for the $n$ variables, and $k, l$ for the $r$ factors throughout. To rule out pathological cases, we will assume throughout that $rank(\Lambda^*) = r$.

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, for instance, $\|A\|_0$ will count the non-zero entries of a matrix $A$. We use the term generalized permutation matrix for a matrix $P^*$ that can be expressed as the product of an invertible diagonal matrix $D$ and a permutation matrix $P$, with its dimension usually obvious from context. 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 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 rescaling 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. Then we can redefine $\Lambda^0 = \Lambda^*(H')^{-1}$ and $F^0 = FH$. This rotation may well be oblique since we do not restrict $H$ 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 or 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.[2]

Throughout the paper, we assume the data has been centered, such that $\mathbb{E}(X_i = 0)$. Auxiliary lemmata are relegated to the Appendix.

[2]See also Ahn and Horenstein (2013) and Onatski (2010) for alternative ways to determine the number of factors. Freyaldenhoven (2019) addresses the issue of estimating the number of factors under the presence of local factors, affecting only a subset of the observables.

# 3 Intuition and Connection to Existing Rotation Criteria

We start with a geometric illustration for the trivial case of only two outcomes (and two factors). This makes it easy to depict rotations graphically, and the intuition carries through to the results in the remainder of the paper.

To this end, consider rotating the unit vector $i_1 = (1, 0)$ in the x,y-plane, analogous to rotating a loading vector for $n = 2$ (one can think of $i_1$ as a hypothetical $\lambda^*_{\bullet 1}$). Keeping its Euclidean length equal to one, this rotation produces a circle, depicted by the small grey dots in Figure 1, where the distance from the origin represents the $\ell_2$-norm of each rotation (which is of course constant and equal to one).

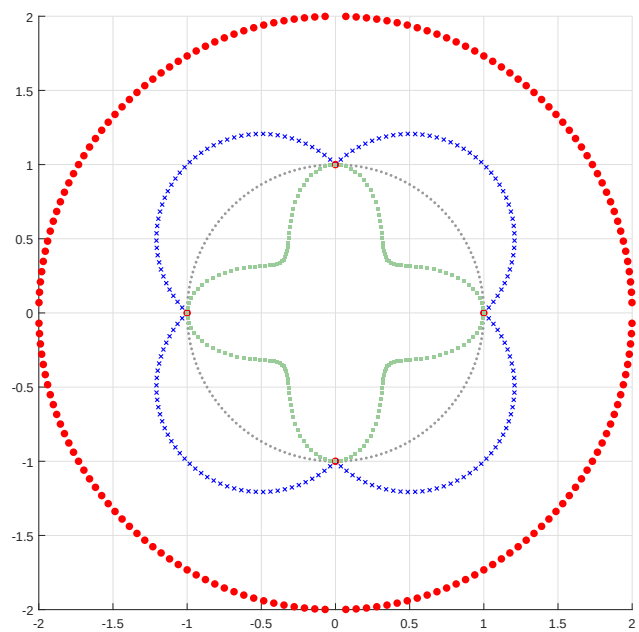



**Figure 1:** Different norms as a function of varying rotations of the unit vector. The size of each norm is illustrated by their distance from the origin. Depicted are $\ell_0$-norm (large, red circles), $\ell_1$-norm (blue crosses), $\ell_2$-norm (small, grey circles), and $\ell_4$-norm (green squares).

For each of these rotations, we next consider the corresponding $\ell_0$-(pseudo)norm. As we rotate the unit vector, we scale up each rotation such that its Euclidean distance from the origin equals the value of its $\ell_0$-(pseudo)norm. This corresponds to the outermost series, depicted by large red circles in Figure 1. We see that almost all rotations have two non-zero elements, such that $\|\cdot\|_0 = 2$, with the exceptions of the four vectors that align with either of the coordinate axes, where $\|\cdot\|_0 = 1$. Figure 1 also illustrates why minimizing the $\ell_0$-(pseudo)norm directly is infeasible. The objective function is flat across almost all rotations with discontinuities at the minima. Thus we would have to resort to grid search, and a fine grid is needed to obtain even approximate sparsity. While possible in an application with just two factors, such a grid search quickly becomes

computationally infeasible when searching over the surface of an $r$-dimensional sphere.

We therefore argue for the use of the $\ell_1$-norm instead. In Figure 1, the value of this norm across rotations is depicted by blue crosses. The $\ell_1$-norm falls in between the $\ell_0$- and $\ell_2$-norms, with its size again represented by the distance of each cross from the origin. Two things are worth noting. First, we note the kink points that occur whenever one of the entries is equal to exactly zero, which is reminiscent of the Lasso in high-dimensional linear regressions. We exploit the presence of these kink points in Section 4 to establish that the $\ell_1$-norm is indeed minimized at sparse rotations. Second, the $\ell_1$-norm is continuous and decreases toward a local minimum in its neighborhood, which makes it computationally appealing (we discuss the implementation in more detail in Online Appendix C).

Finally, the innermost green squares in Figure 1 represent $\|\cdot\|_4^4$, which, in slight abuse of language, we refer to as the $\ell_4$-norm. We added the $\ell_4$-norm to illustrate the connection of our method to the existing literature that considers rotation criteria aimed to simplify the loading matrix. Most of the existing rotation criteria in this literature are quartic functions of the loadings and maximize a variant of the following criterion function $Q(\cdot)$ over rotations of some initial estimate $\Lambda^0$:

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

If we only consider orthogonal rotations for now (which is equivalent to restricting $R$ to be orthonormal), (2) simplifies to

$$Q(\Lambda^0 R) = Q(\Lambda) = \sum_{k=1}^{r}\left[\sum_{i=1}^{n}\lambda_{ik}^4 - \frac{\theta}{n}\left(\sum_{i=1}^{n}\lambda_{ik}^2\right)^2\right]. \tag{3}$$

Setting $\theta$ 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 $\ell_2$-norm.

We see in Figure 1 that maximizing the $\ell_4$-norm across rotations will also produce four maxima, and with $n = 2$, these coincide with the minima for the $\ell_0$- and $\ell_1$-norms. In order to gain an intuition for the difference between the two approaches (maximizing $\ell_4$, minimizing $\ell_1$), it is instructive to also consider maximizing the $\ell_\infty$-norm. Using the $\ell_\infty$-norm would identify the rotation with the largest entry. In contrast, minimizing the $\ell_0$-norm essentially identifies the rotation with the smallest entries. While minimizing the $\ell_1$-norm is a relaxation of the $\ell_0$-norm, maximizing the $\ell_4$-norm can be thought of as a relaxation of the $\ell_\infty$-norm. Given the "resource-constraint" of a

constant $\ell_2$-norm, the solutions to the two optimization problems (maximizing $\ell_4$, minimizing $\ell_1$) will often look similar. However, our formal sparsity assumptions have direct implications for the behavior of the $\ell_0$- and $\ell_1$-norms, but not the $\ell_4$- or $\ell_\infty$-norms. We postulate that this is the reason why, under sparsity assumptions, formal results have been difficult to achieve using 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)), in more detail in Online Appendix B.

## 3.1 A Stylized Example

Before we present our theoretical results in Section 4, the following simple two-factor model allows us to demonstrate our proposed method in a concrete example. More comprehensive simulation results for a variety of data-generating processes can be found in Section 5.

We set $T = 224$ and $n = 207$ and consider the following baseline DGP[3]:

$$\underset{(224\times 207)}{X} = \underset{(224\times 2)}{F}\underset{(2\times 207)}{\Lambda^{*\prime}} + \underset{(224\times 207)}{e}, \tag{4}$$

where we model the errors as

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

Onatski (2010) argues that parameter values of $(\rho, \beta) = (0.3, 0.1)$ are good approximations to many financial datasets. 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 structure of the loading matrix $\Lambda^*$ is

$$\Lambda^* = \begin{bmatrix} \lambda^*_{1:m_1,1} & 0 \\ 0 & \lambda^*_{(n+1)-m_2:n,2} \end{bmatrix}, \tag{5}$$

with the non-zero entries $\lambda^*_{ik} \overset{i.i.d.}{\sim} U(0.1, 2.9)$. Fixing $m_1 = m_2 = 120$, the resulting loading matrix $\Lambda^*$ is depicted in Figure 2. Note that, with $m_1 = m_2 = 120$, some outcomes are affected by both factors. Our goal is to recover $\Lambda^*$.

By construction, the Principal Component estimator $\Lambda^0$ will estimate a rotation of the true loadings and factors that satisfies $\lambda^{0\prime}_{\bullet 1}\lambda^0_{\bullet 2} = 0$ and $F^{0\prime}_{\bullet 1}F^0_{\bullet 2} = 0$. Figure 3 depicts this estimate.[4] Two

[3]This mirrors the dimension of our macroeconomic dataset in Section 6.

[4]To compute the Principal Component estimator, we take the singular value decomposition $X = UDV'$. The

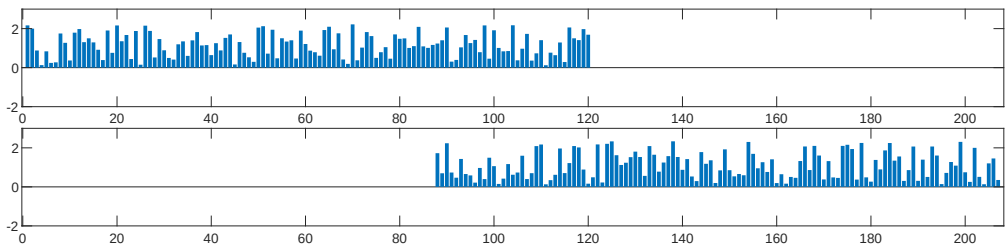



**Figure 2:** Illustration of true loading matrix $\Lambda^*$ for baseline DGP. Top panel depicts $\lambda^*_{\bullet 1}$, bottom panel $\lambda^*_{\bullet 2}$. For each factor, the loadings associated with all 207 outcomes are depicted.

things are worth noting. First, the rotation chosen by the normalization inherent to the Principal Component estimator results in an estimate of the loading matrix that has large loadings everywhere. Second, comparing Figures 2 and 3, we conclude that neither of the estimated loading vectors closely resembles $\lambda^*_{\bullet 1}$ or $\lambda^*_{\bullet 2}$.

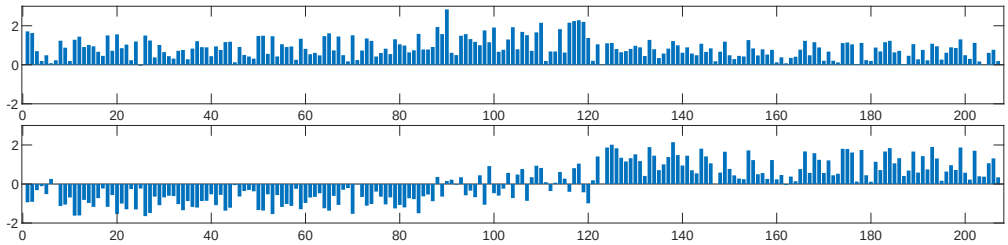



**Figure 3:** Illustration of Principal Component estimate $\Lambda^0$ for baseline DGP. Top panel depicts $\lambda^0_{\bullet 1}$, bottom panel $\lambda^0_{\bullet 2}$. For each factor, the loadings associated with all 207 outcomes are depicted.

Our proposed estimator takes $\Lambda^0$ as a starting point and is equal to the rotation of $\Lambda^0$ that minimizes the $\ell_1$-norm of the loading vectors.

Figure 4 depicts the value of $\|\lambda_{\bullet k}\|_1$ across rotations in the space spanned by the Principal Component estimator $\Lambda^0$. Specifically, it depicts how $\|\lambda_{\bullet k}\|_1 = \|w_1\lambda^0_{\bullet l1} + w_2\lambda^0_{\bullet l2}\|_1$ 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 $[w_1, w_2] = [sin(\theta), cos(\theta)]$, and depict $\|\lambda_{\bullet k}\|_1$ as a function of the angle $\theta$. This is depicted in Figure 4. It also allows a straightforward comparison to Figure 1, since now Figure 4 is essentially a flattened version of the $\ell_1$-norm in Figure 1.

The first minimum occurs at $\tilde{\theta}_1$. This minimum corresponds to weights of $[w_1\ w_2] = [-0.76\ -0.65]$, and consequently an estimated loading vector of $\tilde{\lambda}_{\bullet 1} = -0.76\lambda^0_{\bullet 1} - 0.65\lambda^0_{\bullet 2}$. The second minimum occurs at $\tilde{\theta}_2$. This minimum corresponds to weights of $[w_1, w_2] = [-0.78, 0.62]$, and consequently a second estimated loading vector of $\tilde{\lambda}_{\bullet 2} = -0.78\lambda^0_{\bullet 1} + 0.62\lambda^0_{\bullet 2}$. There are two more local minima in Figure 4 for positive values of $\theta$. These correspond to the estimates $-\tilde{\lambda}_{\bullet 1}$ and

leading $r$ columns of $V$ are used as $\lambda^0_{\bullet 1}, \ldots, \lambda^0_{\bullet r}$.

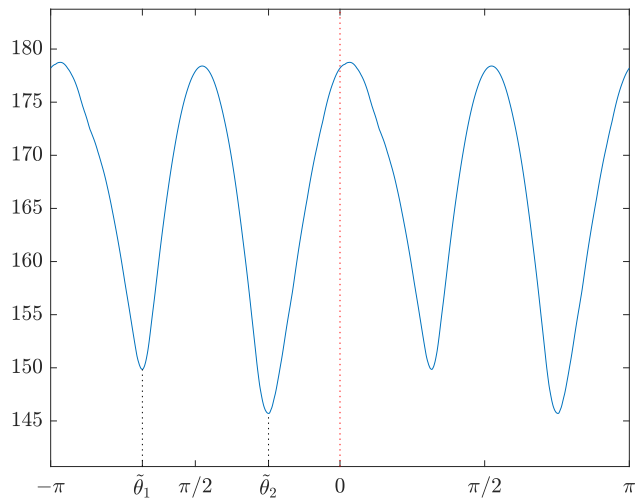



**Figure 4:** Objective function across all rotations in the space spanned by the initial estimate $\Lambda^0$. Depicted is $\|\lambda_{\bullet k}\|_1 = \|sin(\theta)\lambda^0_{\bullet l1} + cos(\theta)\lambda^0_{\bullet l2}\|_1$ as a function of $\theta$.

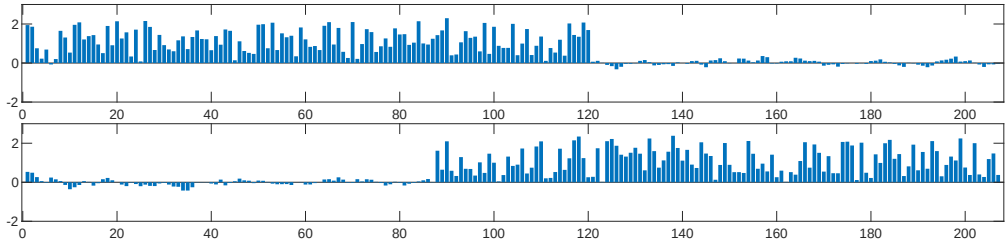



**Figure 5:** Illustration of $\tilde{\Lambda}$, the non-singular "rotated" matrix with the smallest $\ell_1$-norm $\|\Lambda\|_1 = \sum_{i,k} |\lambda_{ik}|$ for baseline DGP. Top panel depicts $\tilde{\lambda}_{\bullet 1}$, bottom panel $\tilde{\lambda}_{\bullet 2}$. For each factor, the loadings associated with all 207 outcomes are depicted.

$-\tilde{\lambda}_{\bullet 2}$ and are a consequence of the sign indeterminacy in the loading vectors, which is why Figure 4 looks identical for negative and positive values of $\theta$.

Combining $[\tilde{\lambda}_{\bullet 1}, \tilde{\lambda}_{\bullet 2}] = \tilde{\Lambda}$, we obtain our proposed estimator for $\Lambda^*$. $\tilde{\Lambda}$ is depicted in Figure 5. Comparing Figures 2 and 5, we conclude that $\tilde{\Lambda}$ is close to $\Lambda^*$, and that we are able to identify $\Lambda^*$ using our proposed criterion.

*Remark* 1. Even though $\tilde{\Lambda}$ is close to $\Lambda^*$, we note that $\tilde{\lambda}_{ik} \neq 0$ for all $i, k$. This is mainly due to the estimation error inherent to the preliminary estimate $\Lambda^0$. Having identified the correct rotation of $\Lambda^*$, 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.

# 4 Identification

## 4.1 Under Exact Sparsity

We start by assuming an exact sparsity pattern in the loading matrix. We will relax this assumption in Section 4.2.

**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* $\lambda_{ik}^* < C \; \forall i \in \mathcal{A}_k$ *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: it collects the indices of all outcomes affected by that factor. Some later results additionally require Assumption 1(c), which allows us to distinguish active loadings from inactive ones in terms of their $\ell_1$-norms.

**Assumption 2.** *Define $\Lambda_{\bullet,-m}^*$ as the $n$ by $(r-1)$ submatrix of $\Lambda^*$ obtained by deleting the $m$th column in $\Lambda^*$, and $\mathcal{A}_{z,-m}$ as the support of a linear combination $\Lambda_{\bullet,-m}^* z$ for a given $(r-1)$ vector of finite weights $z$ and let $b_k(z) = \max|\mathcal{B}|$, such that*

$$\left[\Lambda_{\bullet,-k}^{*\mathcal{A}_k}\right]^{\mathcal{B}} z = \left[\lambda_{\bullet,k}^{*\mathcal{A}_k}\right]^{\mathcal{B}}. \tag{6}$$

*Then, there exists a set of factors $\mathcal{F}^b$, such that $\forall F_k \in \mathcal{F}^b$, $\left|\mathcal{A}_k^c \cap \mathcal{A}_{z,-k}\right| > b_k(z)$ for all $z$.*

Different versions of the set $\mathcal{F}^b$, which approximately defines a group of factors whose associated active sets are not supersets of another factor's active set (see the discussion below), will appear throughout the paper. We will generally be able to show identification for factors in $\mathcal{F}^b$. For instance, in a two-factor model, with a global factor affecting all outcomes and a local factor affecting only a subset of the outcomes, only the latter will be in $\mathcal{F}^b$. In such a model, the local factor is identified, whereas the global one is not. However, these "local" factors, only affecting a subset of the outcomes, are often precisely those that are economically interesting because we may find them interpretable. To ease notation we note that Assumption 2 reduces to the following simple condition if $r = 2$ before further discussion.

**Assumption 2'.** *Let $b = \max|\mathcal{B}|$, such that $\mathcal{B} \subseteq (\mathcal{A}_1 \cap \mathcal{A}_2)$, and for all $i \in \mathcal{B}$*

$$c^* \lambda_{i,1}^* = \lambda_{i,2}^* \tag{7}$$

*for some constant $c^*$.*

*Then, there exists a set of factors $\mathcal{F}^b$, such that $\forall F_k \in \mathcal{F}^b$, $\left|\mathcal{A}_k^c \cap \mathcal{A}_l\right| > b$ for $l \neq k$.*

In order to gain intuition for Assumptions 2 and 2', we first consider three special cases under Assumption 2' (and thus with $r = 2$).

1. Suppose $\mathcal{A}_1 \cap \mathcal{A}_2 = \varnothing$ (The two factors affect different, non-overlapping groups of outcomes). Then, $b = 0$, while $\left|\mathcal{A}_k^c \cap \mathcal{A}_l\right| = |\mathcal{A}_l| > 0$. Therefore, $\mathcal{A}_1 \cap \mathcal{A}_2 = \varnothing$ implies that $F_1, F_2 \in \mathcal{F}^b$.

2. 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 it immediately follows that $F_1 \notin \mathcal{F}^b$. Thus, whenever $\mathcal{A}_k$ is a superset of another set, $F_k$ cannot be a member of the set $\mathcal{F}^b$.

3. Suppose $\frac{\lambda_{i1}}{\lambda_{i2}} \neq \frac{\lambda_{j1}}{\lambda_{j2}}$ for all $i, j \in \mathcal{A}_1 \cap \mathcal{A}_2$. Then $b = 1$. It follows that $F_1 \in \mathcal{F}^b$ if there are at least two outcomes affected by $F_2$, but not $F_1$.

Intuitively, a larger value of $b$ would mean that the two loading vectors are more similar on the intersection of their supports. Specifically, with two factors, $b$ is defined in Assumption 2' as the size of the largest set of non-zero entries in the loading vectors such that the two loading vectors are multiples of each other on that set. Now suppose $F_1 \in \mathcal{F}^b$. Then, the substantive part of Assumption 2', $|\mathcal{A}_1^c \cap \mathcal{A}_2| > b$, reduces to the statement that more than $b$ outcomes are affected by $F_2$, but not $F_1$. This is a more restrictive version of the assumption that $\mathcal{A}_1$ may not be a superset of $\mathcal{A}_2$. How much more restrictive will depend on the similarity between the loading vectors, embodied by $b$.

This intuition also holds for Assumption 2, where $b_k(z)$ is defined as the size of the largest set $\mathcal{B} \subseteq \mathcal{A}_k$ such that we can represent the loading vector $\lambda_{\bullet k}^*$ as an exact linear combination with weights $z$ of the remaining loading vectors on this set. Thus, $b_k = max_z b_k(z)$ is the size of the largest set such that we can represent the loading vector $\lambda_{\bullet k}^*$ as an exact linear combination of the remaining loading vectors on this set. A small value for $b_k$ (e.g., $b_k = r - 1$) means this set is small, and intuitively implies that the different loading vectors are more different.

The restriction $\left|\mathcal{A}_k^c \cap \mathcal{A}_{z,-k}\right| > b_k(z)$ is slightly more restrictive than the assumption that the active set for factor $k$ is not a superset of any other factor, and becomes more restrictive for larger values of $b_k(z)$.

Assumption 2 therefore implies a trade-off between the similarity in the supports of different loading vectors, and the similarity in the loadings on their joint support. The closer to collinearity the loading vectors are on their joint support (corresponding to a large value of $b_k$), the more distinct we require their active sets to achieve identification.

*Remark* 2. It will generally also be possible to identify the subspace spanned by a given subset of factors, 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" in the remainder of this paper.

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

$$\min_R \|\Lambda\|_0 = \min_R \|\Lambda^* R\|_0, \qquad \text{s.t. } R \text{ is nonsingular.} \tag{8}$$

We start with the case in which all factors are identifiable to keep notation simple:

**Corollary 1.** *Suppose Assumptions 1(a)-(b) and 2 hold, and $F_k \in \mathcal{F}^b$ for $k = 1, \ldots, r$. If $\tilde{\Lambda}$ is a solution to* (8)*, then $\tilde{\Lambda} = \Lambda^* P^*$ for some generalized permutation matrix $P^*$.*

Corollary 1 is a direct consequence of Theorem 1 below and its proof is therefore omitted. It 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 (and thereby factors), up to an arbitrary relabeling of the factors and arbitrary scaling constants.

In most settings of economic interest, there will be at least some factor $F_k$, such that $F_k \notin \mathcal{F}^b$ (for instance, a global factor). In such a case, suppose $F_k \in \mathcal{F}^b$ for $k = 1, \ldots, r^*$ and $F_k \notin \mathcal{F}^b$ for $k = r^* + 1, \ldots, r$, and partition $\Lambda^*$ accordingly: $\Lambda^* = [\Lambda^*_{\bullet,1:r^*} \Lambda^*_{\bullet,r^*+1:r}]$. The partitioning of the factors as above is without loss of generality, since it can always be achieved by a simple relabeling of the factors.

**Theorem 1.** *Suppose Assumptions 1(a)-(b) and 2 hold. If $\tilde{\Lambda}$ is a solution to* (8)*, then the corresponding $R$ is such that for every $l = 1, \ldots, r^*$, there exists an index $k$ (which depends on $l$), such that $R_{l,k} \neq 0$ and $R_{l',k} = 0 \; \forall l' \neq l$.*

*Proof.* Columns in $\tilde{\Lambda}$ are linear combinations of the columns in $\Lambda^*$ ($\tilde{\lambda}_{\bullet k} = \Lambda^* R_{\bullet k} = \sum_{l=1}^r \lambda^*_{\bullet l} R_{lk}$). Since $R$ is full rank, at least one entry $R_{lk}$ cannot be equal to zero for each $k$.

Suppose Theorem 1 did not hold, and for some $l^* \in 1, \ldots, r^*$, there exists an index $k$ and a set $\mathcal{L}_k \ni l^*, |\mathcal{L}_k| > 1$, such that not only $R_{l^*,k} \neq 0$, but also $R_{l',k} \neq 0$ for all $l' \in \mathcal{L}_k, l' \neq l^*$, and consider

$$\|\tilde{\lambda}_{\bullet k}\|_0 = \| \sum_{l \in \mathcal{L}_k} \lambda^*_{\bullet l} R_{lk} \|_0.$$

Defining $w^k = \frac{R_{\bullet k}}{-R_{l^*,k}}$, and using the fact that $R_{l^*,k} \neq 0$,

$$\|\sum_{l\in\mathcal{L}_k} \lambda^*_{\bullet l} R_{lk}\|_0 = \|\sum_{l\in\mathcal{L}_k} \lambda^*_{\bullet l} \frac{R_{lk}}{R_{l^*,k}}\|_0 = |\mathcal{A}_{w^k,-l^*} \cup \mathcal{A}_{l^*}| - |\mathcal{C}|$$
$$= |\mathcal{A}_{l^*}| + |\mathcal{A}_{w^k,-l^*} \cap \mathcal{A}^c_{l^*}| - |\mathcal{C}|\,,$$

where $\mathcal{C}$ is defined as the set of indices $i$, such that

$$\sum_{l\neq l^*, l\in\mathcal{L}_k} -\frac{R_{\bullet k}}{R_{l^*,k}} \lambda^*_{il} = \lambda^*_{i,l^*}.$$

By Assumption 2, $|\mathcal{A}_{w^k,-l^*} \cap \mathcal{A}^c_{l^*}| > |\mathcal{C}|$, and therefore

$$|\mathcal{A}_{l^*}| + |\mathcal{A}_{w^k,-l^*} \cap \mathcal{A}^c_{l^*}| - |\mathcal{C}| > |\mathcal{A}_{l^*}|.$$

Now consider setting $R_{l'k}$, $l' \in \mathcal{L}_k$ to zero, and denote the corresponding matrix by $R^*$, with $R^*_{l^*k}$ being the remaining non-zero entry in the $k$th column. Note that this will always be possible while preserving the non-singularity of $R^*$ until for every $l = 1, \ldots, r^*$, there exists an index $k$, such that $R_{l,k} \neq 0$ and $R_{l',k} = 0\ \forall l' \neq l$. It immediately follows that

$$\|\tilde{\lambda}_{\bullet k}\|_0 = \|\sum_{l\in\mathcal{L}_k} \lambda_{\bullet l} R_{lk}\|_0 > |\mathcal{A}_{l^*}| = \|\lambda_{\bullet l^*} R^*_{l^*,k}\|_0,$$

which contradicts that $R$ is a solution to the minimization problem (8).

We conclude that for every $l = 1, \ldots, r^*$, there exists an index $k$, such that $R_{l,k} \neq 0$ and $R_{l',k} = 0\ \forall l' \neq l$, and that the corresponding index $l$ is distinct for each $k$. This completes the proof. □

Theorem 1 establishes the following: If the true DGP includes local factors ($\mathcal{F}^b$ is non-empty), such local factors can be identified by maximizing the degree of sparsity in the loading matrix across feasible rotations. The intuition is that, of all possible rotations of local factors, none will be as sparse as the truth, $\Lambda^*$. Note that Theorem 1 does not say anything about factors that are not in $\mathcal{F}^b$. For instance, if there are global factors with a corresponding loading vector that is dense, identification of such factors based on a sparsity criterion will clearly be impossible.

*Remark* 3. Theorem 1 still holds if we did not observe $\Lambda^*$ directly, but instead had access to a rotation of $\Lambda^*$, $\Lambda^0 = \Lambda^* H$, where $H$ is a nonsingular matrix.

**Proposition 1.** *Suppose* $|\mathcal{A}_k| = n$ *for* $k = 1, \ldots, r$*, and define* $b^* = \max_k \max_z b_k(z)$*, where* $b_k(z)$ *is defined in Assumption 2. Then, for any nonsingular matrix* $R$*,* $|\tilde{\mathcal{A}}_k| \geq n - b^*$ *for* $k = 1, \ldots, r$*,*

*where* $\tilde{\Lambda} = \Lambda^* R$ *and* $|\tilde{\mathcal{A}}_k| = \|\tilde{\lambda}_{\bullet k}\|_0$.

*Proof.* Suppose Proposition 1 does not hold, and there exists a linear combination $R_{\bullet k}$, $\|R_{\bullet k}\|_0 > 0$, such that $\|\Lambda^* R_{\bullet k}\|_0 = \|\tilde{\lambda}_{\bullet k}\|_0 < n - b^*$. Then, there must exist a set $\mathcal{D}$, with $|\mathcal{D}| > b^*$ such that

$$\sum \lambda^*_{il} R_{lk} = 0 \; \forall i \in \mathcal{D}.$$

Since $|\mathcal{A}_k| = n$ for $k = 1, \ldots, r$, this also implies that, for some $l^* \in \{1, \ldots, r\}$,

$$-\sum_{l \neq l^*} \lambda^*_{il} R_{lk} = \lambda^*_{i,l^*} \; \forall i \in \mathcal{D}.$$

But since, by definition of $b^*$, $|\mathcal{D}| \leq b^*$, this leads to the desired contradiction. □

Proposition1 states that, if there are no local factors in the data and all factors affect all outcomes, there does not exist a rotation of the true factor loadings with a significant sparsity pattern. The implication is that, whenever a rotation is found with a high degree of sparsity, local factors must be present in the data. In line with our discussion following Assumption 2, how much sparsity can be achieved across rotations will depend on two things: the degree of sparsity in the true loading matrix, and how close to collinear the different loading vectors are.

Minimizing the $\ell_0$-norm directly is infeasible in practice (we will return to this discussion in Section 4.2). We therefore turn our attention to the $\ell_1$-norm of the loading matrix next.

**Assumption 3.** *Let* $V_k$ *denote the set of all linear combination* $v_{\bullet k}$ *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 define*

$$\beta^k(v_{\bullet k}) = \left| \sum_{i \in \mathcal{A}_k} |v_{ik}| \, \mathbf{1}\{\lambda^*_{ik} v_{ik} \geq 0\} - \sum_{i \in \mathcal{A}_k} |v_{ik}| \, \mathbf{1}\{\lambda^*_{ik} v_{ik} < 0\} \right|. \tag{9}$$

*Then, there exists a set of factors* $\mathcal{F}^{exact}$, *such that,* $\forall F_k \in \mathcal{F}^{exact}$,

$$\|v_{\bullet k}^{\mathcal{A}_k^c}\|_1 > \beta^k(v_{\bullet k}) \qquad \forall v_{\bullet k} \in V_k. \tag{10}$$

Assumption 3 is similar to Assumption 2 in the previous section, with $\mathcal{F}^{exact}$ approximately defining a group of factors whose associated active sets are not supersets of another factor's active set.

To see this, consider the case $r = 2$ again, in which case $v_{\bullet 1} = q_1 \lambda^*_{\bullet 1} + q_2 \lambda^*_{\bullet 2}$ for some constants $q_1, q_2$. Further, by definition of $\mathcal{A}_1$, $\|v_{\bullet 1}^{\mathcal{A}_1^c}\|_1 = q_2 \|\lambda_{\bullet 2}^{*\mathcal{A}_1^c}\|_1$, a constant times the sum of the absolute values of $\lambda^*_{2i}$ on $\mathcal{A}_1^c$. It follows that $F_1 \in \mathcal{F}^{exact}$ iff $\|\lambda_{\bullet 2}^{*\mathcal{A}_1^c}\|_1 > \frac{1}{q_2} \beta^k(v_{\bullet k})$. Consider the following three specific examples:

1. Suppose $\mathcal{A}_1 \cap \mathcal{A}_2 = \varnothing$. Then, $\lambda^*_{\bullet 1} \perp v_{\bullet 1}$ implies $v_{\bullet 1} = \lambda^*_{\bullet 2}$, and thus $\beta^1(v_{\bullet 1}) = \beta^1(\lambda^*_{\bullet 2}) = 0$. Further, note that $\|\lambda^{*\mathcal{A}_1^c}_{\bullet 2}\|_1 = \|\lambda^*_{\bullet 2}\|_1$. It follows from $rank(\Lambda^*) = r$ that $\|\lambda_{\bullet 2}\|_1 > 0$, and therefore $\mathcal{A}_1 \cap \mathcal{A}_2 = \varnothing$ implies that (10) holds and $F_1 \in \mathcal{F}^{exact}$. Clearly, the same reasoning can be applied to also show that $F_2 \in \mathcal{F}^{exact}$ if $\mathcal{A}_1 \cap \mathcal{A}_2 = \varnothing$.

2. Suppose $\mathcal{A}_2 \subseteq \mathcal{A}_1$. Then, $\lambda^{*\mathcal{A}_1^c}_{\bullet 2} = 0$, and it immediately follows that $F_1 \notin \mathcal{F}^{exact}$. Thus, whenever $\mathcal{A}_k$ is a superset of another set, $F_k$ cannot be a member of the set $\mathcal{F}^{exact}$.
   In general, Assumption 3 implies a lower bound on the degree of sparsity for each factor in $\mathcal{F}^{exact}$ and implies that $\mathcal{F}^{exact}$ will never include "global" factors. [5]

3. Suppose $|\mathcal{A}_k| \asymp n$ for $k = 1, 2$ as $n \to \infty$, and

$$\lambda^*_{ik} \begin{cases} \overset{i.i.d.}{\sim} N(0, \sigma) & \text{if } i \in \mathcal{A}_k \\ = 0 & \text{else.} \end{cases} \tag{11}$$

   Then, it can be shown (see Online Appendix A) that $\beta^k(v_{\bullet k}) = O_p(\sqrt{n})$ and that $F_1 \in \mathcal{F}^{exact}$ if $\sqrt{n} = o_p(\|\lambda^{*\mathcal{A}_1^c}_{\bullet 2}\|_1)$. Intuitively, this states that $F_1 \in \mathcal{F}^{exact}$ if more than $\sqrt{n}$ outcomes are affected by $F_2$, but not $F_1$.

We provide some further discussion of Assumption 3 and alternative assumptions sufficient for our results in Online Appendix A.

In what follows, we will work with an initial rotation of $\Lambda^*$, rather than with $\Lambda^*$ directly. We will denote this as $\Lambda^0 = \Lambda^* H$, where $H$ is nonsingular. $\Lambda^0$ has the property that its columns have equal length and are orthogonal, such that $\frac{\Lambda^{0\prime}\Lambda^0}{n} = I$. While $\Lambda^0$ is not unique, clearly such a rotation always exists. Intuitively, one can think of $\Lambda^0$ as the rotation of $\Lambda^*$ that is estimated by the Principal Component estimator, at this point ignoring any estimation error.

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 $\Upsilon$ with $\|\Upsilon\|_2 = 1$. When considering the $l_1$-norm of $\lambda_{\bullet k}$ for different linear combinations $\Upsilon$, we therefore hold the $l_2$-norm of $\lambda_{\bullet k}$ constant across those combinations. To this end, consider a $(r \times r)$ matrix R, and let $\Lambda = \Lambda^0 R$. We first note that

$$\min_R \|\Lambda\|_1 = \min_R \sum_{k=1}^r \|\lambda_{\bullet k}\|_1 = \min_R \sum_{k=1}^r \| \sum_{l=1}^r \lambda^0_{\bullet l} R_{lk}\|_1 = \sum_{k=1}^r \min_{R_{\bullet k}} \| \sum_{l=1}^r \lambda^0_{\bullet l} R_{lk}\|_1, \tag{12}$$

and consider a constrained version of the above:

$$\min_R \|\Lambda\|_1 = \min_R \|\Lambda^0 R\|_1 \qquad \text{such that } R'_{\bullet k} R_{\bullet k} = 1 \ \forall k. \tag{13}$$

[5] In fact, the number of loadings that are equal to zero for each factor in $\mathcal{F}^{exact}$ must be at least proportional to $\beta^k = \max_{v_{\bullet k} \in V_k} \beta^k(v_{\bullet k})$. To see this, note that $\lambda^*_{il} < C \ \forall i, l$ implies that $\|v^{\mathcal{A}^c_{\bullet k}}\|_1 = o(\beta^k)$ if $|\mathcal{A}^c_k| = o(\beta^k)$.

Because (12) is separable in $k$ and consists of $k$ identical parts, we will consider one part at a time:

$$\min_{R_{\bullet k}} \| \sum_{l=1}^{r} \lambda_{\bullet l}^{0} R_{lk} \|_1 \qquad \text{such that } R_{\bullet k}' R_{\bullet k} = 1. \tag{14}$$

Since $R_{\bullet k}' R_{\bullet k} = 1$ implies a constant $l_2$-norm, one can also loosely think of (14) as minimizing the difference between the $l_1$- and $l_2$-norms of $\lambda_{\bullet k} = \Lambda^* R_{\bullet k}$ across different linear combinations $R_{\bullet k}$. We obtain the following result.

**Theorem 2.** *Suppose Assumptions 1 and 3 hold 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 $F_k \in \mathcal{F}^{exact}$, $\lambda_k^*$ is a local minimum of* (14).

*Proof.* 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 $\Lambda^*$ without changing its length (in the standard $\ell_2$-sense). Next, note that $v_{\bullet k} = v_{\bullet k}^{\mathcal{A}_k} + v_{\bullet k}^{\mathcal{A}_k^c}$, and therefore

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

where the inequality above follows from Lemma 2 when we consider a small neighborhood around $\lambda_{\bullet k}^*$. To make this explicit, set $w_1 = \sqrt{1 - w_2^2}$ and $|w_2| = \epsilon$. We need 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 1, the first part can be made arbitrarily small for small enough values of $\epsilon$. On the other hand, the second part is larger than zero by Assumption 3 and does not depend on $\epsilon$. 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 $\Lambda^*$ and hence $\|\lambda_{\bullet k}^*\|_1$ is a local minimum of (14). □

It immediately follows that any set of local minima of (14) for $k = 1, \ldots, r$ is also a local minimum of (13). 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 (13) spans the same space that is spanned by $\Lambda^0$.

## 4.2 Under Approximate Sparsity

Theorems 1 and 2 require exact sparsity. However, such a pattern may only provide an approximation to the true loading matrix, and we may think of $\Lambda^*$ as approximately sparse instead. We therefore next redefine the sets $\mathcal{A}_k$ to relax our sparsity assumption and to include models with approximate sparsity in the loading matrix.

**Assumption 4.** *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)* $\sum_{i \in \mathcal{A}_k} \left|\lambda^*_{ik}\right| > c_0 n$ *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* $c, C$.

Assumption 4(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 4(a) restricting how much. Assumption 4(b) can be thought of as a pervasiveness assumption. Together with Assumption 4(c), it states that each factor affects a constant fraction of all outcomes, which is commonly maintained in the literature. For our main result (Theorem 4), we require access to a $\sqrt{n}$ consistent estimate of $\Lambda^*$. The ability to obtain such a $\sqrt{n}$ consistent estimate generally implies that factors must be pervasive (Freyaldenhoven 2019).

**Assumption 5.** *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 define*

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

*Then, there exists a set of factors $\mathcal{F}$, such that, for any $F_k \in \mathcal{F}$, and some $c_{min} > 0$ and $N < \infty$, whenever $n > N$:*

$$\|v_{\bullet k}^{\mathcal{A}_k^c}\|_1 - \beta^k(v_{\bullet k}) > c_{min} n^{\frac{3}{4}} \qquad \forall v_{\bullet k} \in V_k.$$

Assumption 5 slightly strengthens Assumption 3 in order to accommodate non-zero entries of $\lambda^*_{\bullet k}$ on $\mathcal{A}_k^c$, with a trade-off similar to the one we observed in Assumptions 2 and 3.

**Theorem 3.** *Suppose $n \to \infty$, Assumptions 4 and 5 hold 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 $F_k \in \mathcal{F}$, there exists a local minimum $\bar{\lambda}_{\bullet k}$ of* (14)*, such that*

$$\frac{1}{n}\|\bar{\lambda}_{\bullet k} - \lambda^*_{\bullet k}\|^2 = O(\frac{1}{\sqrt{n}}). \tag{16}$$

*Proof.* 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 $\Lambda^*$ without changing its length (in the standard $\ell_2$-sense). Then,

$$\begin{aligned}
\|\lambda_{\bullet k}\|_1 &= \|w_1\lambda^*_{\bullet k} + w_2 v_{\bullet k}\|_1 = \|w_1\lambda^{*\mathcal{A}_k}_{\bullet k} + w_2 v^{\mathcal{A}_k}_{\bullet k}\|_1 + \|w_1\lambda^{*\mathcal{A}^c_k}_{\bullet k} + w_2 v^{\mathcal{A}^c_k}_{\bullet k}\|_1 \\
&\geq \|w_1\lambda^{*\mathcal{A}_k}_{\bullet k} + w_2 v^{\mathcal{A}_k}_{\bullet k}\|_1 + \|w_2 v^{\mathcal{A}^c_k}_{\bullet k}\|_1 - |w_1|\,\|\lambda^{*\mathcal{A}^c_k}_{\bullet k}\|_1 \\
&\geq |w_1|\,\|\lambda^{*\mathcal{A}_k}_{\bullet k}\|_1 - |w_2|\,\beta^k(v_{\bullet k}) + w_2\|v^{\mathcal{A}^c_k}_{\bullet k}\|_1 - |w_1|\,\|\lambda^{*\mathcal{A}^c_k}_{\bullet k}\|_1 \\
&= |w_1|\,\|\lambda^{*\mathcal{A}_k}_{\bullet k}\|_1 + |w_2|\left(\|v^{\mathcal{A}^c_k}_{\bullet k}\|_1 - \beta^k(v_{\bullet k})\right) - |w_1|\,\|\lambda^{*\mathcal{A}^c_k}_{\bullet k}\|_1,
\end{aligned}$$

where the second inequality above follows from Lemma 2 when we consider a small neighborhood around $\lambda^*_k$. To make this explicit, set $w_1 = \sqrt{1 - w_2^2}$ and $|w_2| = \epsilon$. We need to show that

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

Note that

$$\begin{aligned}
& \left(\sqrt{1-\epsilon^2} - 1\right)\|\lambda^{*\mathcal{A}_k}_{\bullet k}\|_1 + \epsilon\left(\|v^{\mathcal{A}^c_k}_{\bullet k}\|_1 - \beta^k(v_{\bullet k})\right) - \left(\sqrt{1-\epsilon^2} + 1\right)\|\lambda^{*\mathcal{A}^c_k}_{\bullet k}\|_1 \\
& \qquad > \underbrace{(\sqrt{1-\epsilon^2} - 1)\|\lambda^{*\mathcal{A}_k}_{\bullet k}\|_1}_{\equiv \mathbf{I}} + \underbrace{\epsilon(\|v^{\mathcal{A}^c_k}_{\bullet k}\|_1 - \beta^k(v_{\bullet k}))}_{\equiv \mathbf{II}} - \underbrace{2\|\lambda^{*\mathcal{A}^c_k}_{\bullet k}\|_1}_{\equiv \mathbf{III}}
\end{aligned}$$

and choose $\epsilon = \frac{\bar{C}}{c_{min}} n^{-\frac{1}{4}}$. Then, we have the following bounds on the three terms above:

$$\begin{aligned}
\mathbf{I} &= \left(\sqrt{1-\epsilon^2} - 1\right) \|\lambda_{\bullet k}^{*\mathcal{A}_k}\|_1 = O(\frac{1}{\sqrt{n}})O(n) = O(\sqrt{n}), \\
\mathbf{II} &= \epsilon(\|v_{\bullet k}^{\mathcal{A}_k^c}\|_1 - \beta^k(v_{\bullet k})) > n^{-\frac{1}{4}}\bar{C}n^{\frac{3}{4}} = \bar{C}\sqrt{n} \quad \text{for any } \bar{C} < \infty, \\
\mathbf{III} &= 2\|\lambda_{\bullet k}^{*\mathcal{A}_k^c}\|_1 = O(\sqrt{n}),
\end{aligned}$$

where the rate in **I** follows from the fact that for $0 < \epsilon < 1$, $\left|\sqrt{1-\epsilon^2} - 1\right| < \epsilon^2 = O(\frac{1}{\sqrt{n}})$. The inequality for **II** follows from Assumption 5, and the stated rate in **III** follows from Assumption 4 (a). Thus the second term dominates and is positive.

This guarantees that $\|\lambda_{\bullet k}\|_1 > \|\lambda_{\bullet k}^*\|_1$. Since $\Lambda^*$ is a feasible solution to (14), and by continuity of the $\ell_1$-norm, there must therefore exist a $\bar{\lambda}_{\bullet k} = w_1\lambda_{\bullet k}^* + w_2 v_{\bullet k}$, $w_2 = O(n^{-\frac{1}{4}})$, with $w_1^2 + w_2^2 = 1$, that is a local minimum of (14).

It further follows that

$$\begin{aligned}
\|\lambda_{\bullet k}^* - \bar{\lambda}_{\bullet k}\| &= \|\lambda_{\bullet k}^* - (w_1\lambda_{\bullet k}^* + w_2 v_{\bullet k})\| = \|\lambda_{\bullet k}^* - \sqrt{1-w_2^2}\lambda_{\bullet k}^* - w_2 v_{\bullet k}\| \\
&= w_2\|\frac{1-\sqrt{1-w_2^2}}{w_2}\lambda_{\bullet k}^* - v_{\bullet k}\| < w_2\|\lambda_{\bullet k}^* - v_{\bullet k}\|,
\end{aligned}$$

where the inequality follows from the fact that $\frac{1-\sqrt{1-w_2^2}}{w_2} < 1$ for $0 < w_2 < 1$. Finally, by orthogonality of $\lambda_{\bullet k}^*$ and $v_{\bullet k}$, and since $\|\lambda_{\bullet k}^*\| = \|v_{\bullet k}\| = \sqrt{n}$,

$$w_2\|\lambda_{\bullet k}^* - v_{\bullet k}\| = w_2\sqrt{2n}.$$

We therefore conclude that, with $w_2 = O(n^{-\frac{1}{4}})$,

$$\|\lambda_k^* - \bar{\lambda}_{\bullet k}\| < w_2\|\lambda_{\bullet k}^* - v_{\bullet k}\| = O(n^{-\frac{1}{4}})\sqrt{2n} = O(n^{\frac{1}{4}}),$$

and the desired result follows. □

Theorems 1-3 all assumed access to an initial rotation of $\Lambda^*$, $\Lambda^0 = \Lambda^* H$. In practice, we will only have access to an estimate of such a rotation. That is, we will have access to an initial $\Lambda^0$, such that $\|\lambda_{ik}^0 - \lambda_{i\bullet}^* H_{\bullet k}\| = O_p(\frac{1}{\sqrt{n}})$. The main result of the paper follows and establishes that any $\sqrt{n}$ consistent estimate of the space spanned by the loadings is sufficient to recover the individual loading vectors corresponding to factors in $\mathcal{F}$.

*Remark* 4. We remain agnostic about where such an initial estimate may come from. An obvious candidate that fulfills the $\sqrt{n}$ consistency requirement above under some regularity conditions

would be the Principal Component estimator (Bai and Ng 2002, Bai 2003). We also use this estimator in our simulations and applications.

**Theorem 4.** *Suppose $n \to \infty$, Assumptions 4 and 5 hold, and $F_k \in \mathcal{F}$. Further suppose we have access to an initial estimate $\Lambda^0$, such that $\|\lambda^0_{ik} - \lambda^*_{i\bullet} H_{\bullet k}\| = O_p(\frac{1}{\sqrt{n}})$, where $H$ is nonsingular, and $\frac{\Lambda^{0\prime}\Lambda^0}{n} = I$. Then, there exists a local minimum $\bar{\lambda}_{\bullet k}$ of* (14)*, such that*

$$\frac{1}{n}\|\bar{\lambda}_{\bullet k} - \lambda^*_{\bullet k}\|^2 = O_p(\frac{1}{\sqrt{n}}). \tag{17}$$

*Proof.* Since $(\lambda^0_{ik} - \sum_{l=1}^r \lambda^*_{il} H_{lk}) = O(\frac{1}{\sqrt{n}})$ and H is nonsingular, there exists a $(r \times 1)$ vector $\psi^i$, such that

$$\lambda^0_{ik} - \sum_{l=1}^r \lambda^*_{il} H_{lk} = \sum_{l=1}^r \psi^i_l H_{lk},$$

where $\psi^i_l = O(\frac{1}{\sqrt{n}})$ for $l = 1, \ldots, r$. Thus,

$$\begin{aligned} & \lambda^0_{ik} - (\lambda^*_{i\bullet} - \psi^i_\bullet) H_{\bullet k} = 0 \\ \Leftrightarrow \quad & \lambda^0_{ik} - \dot{\lambda}_{i\bullet} H_{\bullet k} = 0, \end{aligned}$$

by defining $\dot{\lambda}_{i\bullet} \equiv (\lambda^*_{i\bullet} - \psi^i_\bullet)$. It follows that $\Lambda^0$ is an exact rotation of $\dot{\Lambda}$, where $\dot{\lambda}_{ik} - \lambda^*_{ik} = O(\frac{1}{\sqrt{n}})$ for all $i, k$.

Also note that, under Assumption 4(a), $\sum_{i \notin \mathcal{A}_k} \left|\dot{\lambda}_{ik}\right| = O(\sqrt{n})$ for all $F_k \in \mathcal{F}$, since $\sum_{i \notin \mathcal{A}_k} \left|\dot{\lambda}_{ik}\right| = \sum_{i \notin \mathcal{A}_k} \left|(\lambda^*_{ik} + (\dot{\lambda}_{ik} - \lambda^*_{ik})\right| \leq \sum_{i \notin \mathcal{A}_k} \left|\lambda^*_{ik}\right| + \sum_{i \notin \mathcal{A}_k} \left|\dot{\lambda}_{ik} - \lambda^*_{ik}\right| = O(\sqrt{n})$. Similarly, Assumptions 4(b)-4(c) hold for $\dot{\Lambda}$ if they hold for $\Lambda^*$ and thus the conditions in Assumption 4 hold for $\dot{\Lambda}$ if they are satisfied for $\Lambda^*$.

We next show that the conditions in Assumption 5 also hold for $\dot{\Lambda}$ if they are satisfied for $\Lambda^*$. Suppose $F_k \in \mathcal{F}$ and let $\dot{v}_{\bullet k}$ be an arbitrary linear combination of $\dot{\lambda}_{\bullet l}$, $l = 1, \ldots, r$, such that $\|\dot{v}_{\bullet k}\|_2^2 = n$ and $\dot{\lambda}_{\bullet k} \perp \dot{v}_{\bullet k}$. Then, there exists a $v_{\bullet k} \in V_k$, such that

$$\begin{aligned} \beta^k(\dot{v}_{\bullet k}) &= \left| \sum_{i \in \mathcal{A}_k} |\dot{v}_{ik}| \, \mathbf{1}\{\dot{\lambda}_{ik}\dot{v}_{ik} \geq 0\} - \sum_{i \in \mathcal{A}_k} |\dot{v}_{ik}| \, \mathbf{1}\{\dot{\lambda}_{ik}\dot{v}_{ik} < 0\} \right| \\ &= \left| \sum_{i \in \mathcal{A}_k} |\dot{v}_{ik}| \, \mathbf{1}\{\lambda^*_{ik}\dot{v}_{ik} \geq 0\} - \sum_{i \in \mathcal{A}_k} |\dot{v}_{ik}| \, \mathbf{1}\{\lambda^*_{ik}\dot{v}_{ik} < 0\} \right| \\ &= \left| \left( \sum_{i \in \mathcal{A}_k} |v_{ik}| \, \mathbf{1}\{\lambda^*_{ik}\dot{v}_{ik} \geq 0\} - \sum_{i \in \mathcal{A}_k} |v_{ik}| \, \mathbf{1}\{\lambda^*_{ik}\dot{v}_{ik} < 0\} \right) \right. \end{aligned}$$

$$
\begin{aligned}
&\quad + \left( \sum_{i \in \mathcal{A}_k} (|\dot{v}_{ik}| - |v_{ik}|) \mathbf{1}\{\lambda^*_{ik} \dot{v}_{ik} \geq 0\} - \sum_{i \in \mathcal{A}_k} (|\dot{v}_{ik}| - |v_{ik}|) \mathbf{1}\{\lambda^*_{ik} \dot{v}_{ik} < 0\} \right) \Bigg| \\
&= \left| \sum_{i \in \mathcal{A}_k} |v_{ik}| \, \mathbf{1}\{\lambda^*_{ik} \dot{v}_{ik} \geq 0\} - \sum_{i \in \mathcal{A}_k} |v_{ik}| \, \mathbf{1}\{\lambda^*_{ik} \dot{v}_{ik} < 0\} + O(\sqrt{n}) \right| \\
&= \left| \sum_{i \in \mathcal{A}_k} |v_{ik}| \, \mathbf{1}\{\lambda^*_{ik} v_{ik} \geq 0\} - \sum_{i \in \mathcal{A}_k} |v_{ik}| \, \mathbf{1}\{\lambda^*_{ik} v_{ik} < 0\} \right| + O(\sqrt{n}) \\
&\leq \beta^k(v_{\bullet k}) + O(\sqrt{n}),
\end{aligned}
$$

where the first equality follows from the fact that the signs of $\dot{\lambda}_{ik}$ and $\lambda^*_{ik}$ will be identical on $\mathcal{A}_k$, because of the lower bound on $\lambda^*_{ik}$ from Assumption 4(c). The third equality follows from Lemma 3 and the fourth equality from Lemma 4.

Thus, if $F_k \in \mathcal{F}$, and for some $c_{min} > 0$ and $N < \infty$, whenever $n > N$,

$$
\|v_{\bullet k}^{\mathcal{A}_k^c}\|_1 - \beta^k(v_{\bullet k}) > c_{min} n^{\frac{3}{4}} \qquad \forall v_{\bullet k} \in V_k,
$$

the following must also be true:

$$
\|\dot{v}_{\bullet k}^{\mathcal{A}_k^c}\|_1 - \beta^k(\dot{v}_{\bullet k}) > c_{min} n^{\frac{3}{4}} \qquad \forall v_{\bullet k} \in V_k.
$$

Therefore, Assumption 5 is satisfied for $\dot{\Lambda}$ if it is satisfied for $\Lambda^*$ and we can invoke Theorem 3 to yield:

$$
\frac{1}{n} \|\bar{\lambda}_{\bullet k} - \dot{\lambda}_{\bullet k}\|^2 = O(\frac{1}{\sqrt{n}}).
$$

Using the triangle inequality, it then immediately also follows that

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

□

The number of elements in $\Lambda^*$ increases with $n$. Thus Theorem 4 uses an averaged norm to establish consistency of the estimated loadings, similar to Ando and Bai (2017).

Theorem 4 suggests the following simple two-step algorithm to consistently estimate any local factors.[6]

1. Obtain a $\sqrt{n}$ consistent estimate $\Lambda^0$ that forms an orthonormal basis of factor space.

[6] We discuss our algorithmic implementation in more detail in Online Appendix C.

2. Find the rotation $\tilde{R}$ that minimizes the $l_1$-norm of the loadings:

$$\min_R \| \sum_{l=1}^{r} \lambda^0_{\bullet l} R_{lk} \|_1 \qquad \text{such that } R'_{\bullet k} R_{\bullet k} = 1, R \text{ nonsingular.} \tag{18}$$

By Theorem 4, if there are local factors present in the data, their true loading vectors will be (close to) an argmin of (18).

# 5 Simulations

This section presents results from Monte Carlo simulations in order to evaluate the performance of our proposed criterion in practice.

We start by revisiting the baseline DGP from our stylized example in Section 3.1 but move beyond a single realization. In order to summarize the performance of an estimator numerically, a simple measure is the cosine similarity between the columns in $\Lambda^*$ and $\hat{\Lambda}$. Because the factors can always be reordered, we take the maximum cosine similarity between any estimated vector and each vector of the true loadings. Formally, define the maximum cosine similarity $MC_l(\hat{\Lambda})$ between the true loading vector $\lambda^*_{\bullet l}$ and an estimate $\hat{\Lambda}$ as

$$MC_l(\hat{\Lambda}) = \max_k \frac{\hat{\lambda}'_{\bullet k} \lambda^*_{\bullet l}}{\|\hat{\lambda}_{\bullet k}\| \|\lambda^*_{\bullet l}\|} \qquad \text{for } l = 1, \ldots, r. \tag{19}$$

One can think of $MC_l$, $l = 1, \ldots, r$ as a measure of how well we estimate each of the $r$ loading vectors.

The maximum correlation corresponding to Figures 2-5 in Section 3.1 is depicted in Table 1, averaged across 1000 realizations. While $\lambda^*_{ik} \overset{i.i.d.}{\sim} U(0.1, 2.9)$ for $i \in \mathcal{A}_k$ was chosen to satisfy the upper and lower bounds assumed on $\Lambda^*$ in the previous section, we also consider $\lambda^*_{ik} = 1 + \eta_{ik}$, $\eta_{ik} \overset{i.i.d.}{\sim} N(0, 1)$ for $i \in \mathcal{A}_k$. Modeling the loadings as normally distributed is common in the literature (Bai 2003, Freyaldenhoven 2019).

The DGP underlying the first two columns is identical to that underlying Figures 2-5 and has uniformly distributed loadings on $\mathcal{A}_k$. We simulate new loadings in $\Lambda^*$ for each realization. The first column confirms that the Principal Component estimator does not successfully recover either of the two loading vectors. On the other hand, the second column provides evidence that our proposed estimator can identify the true loading matrix $\Lambda^*$. Because both factors affect the same number of outcomes in our baseline DGP, the two rows look identical. Columns 3-4 demonstrate that changing the distribution of the loadings $\lambda_{ik}$ on $\mathcal{A}_k$ has no effect on our results.

| | $\lambda_{ik}^* \sim U(0.1, 1.9)$ | | $\lambda_{ik}^* = 1 + \eta_{ik}, \eta_{ik} \sim N(0,1)$ | |
|---|---|---|---|---|
| Estimator $\hat{\Lambda}$ | $\Lambda^0$ | $\tilde{\Lambda}$ | $\Lambda^0$ | $\tilde{\Lambda}$ |
| $MC_1$ | 0.778 | 0.990 | 0.778 | 0.994 |
| $MC_2$ | 0.778 | 0.990 | 0.775 | 0.994 |

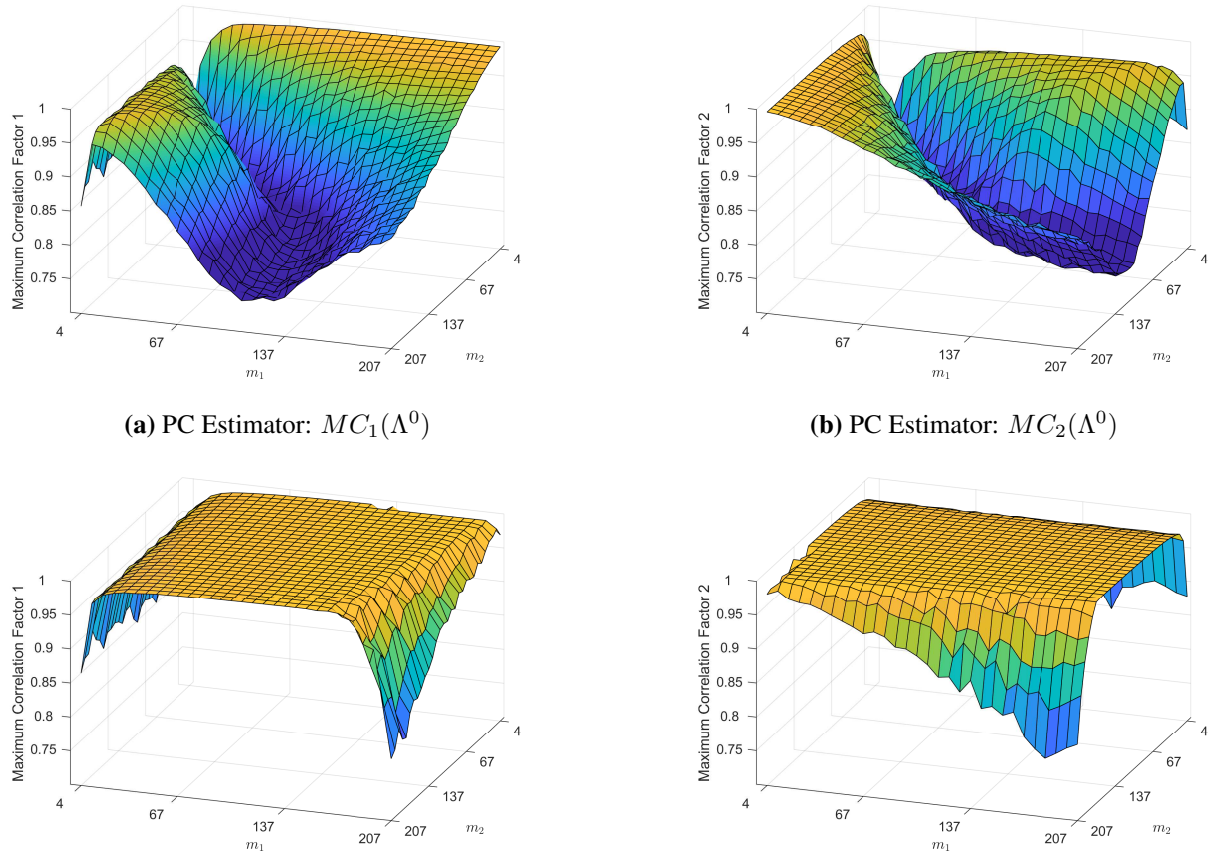
**Table 1:** Maximum cosine similarity $MC_l(\hat{\Lambda})$ across DGPs and estimators. $\Lambda^0$ refers to the Principal Component estimator, while $\tilde{\Lambda}$ represents our proposed rotation that minimizes the $\ell_1$-norm across all rotations. All numbers are based on 1000 realizations.

The previous results confirm that our proposed estimator can indeed recover the correct sparsity pattern in the loading matrix, at least under the relatively simple DGP of this section. We next consider a variety of data-generating processes to approximate a range of situations a practitioner might encounter in practice.

## 5.1 Results for a Variety of Data-Generating Processes

We start by varying the degree of sparsity in the loading matrix. To this end, we vary the values of $m_1$ and $m_2$ in the previous DGP and maintain that $\lambda_{ik}^* = 1 + \eta_{ik}$, $\eta_{ik} \overset{i.i.d.}{\sim} N(0,1)$, for $i \in \mathcal{A}_k$. The corresponding result is depicted in Figure 6. Using the maximum cosine similarity introduced above, Figure 6 depicts how well we are able to estimate the true factor loadings $\lambda_1^*$ and $\lambda_2^*$ as a function of $m_1$ and $m_2$. Panels 6a and 6b depict the performance of the Principal Component estimator. Unsurprisingly, the numbers reported in panels 6a and 6b are generally significantly below one. The exception to this are cases in which one factor is extremely weak. Effectively, the data follows a factor structure with a single factor in such cases, and consequently there is no rotational indeterminacy and the sole 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 factors throughout most of the parameter space using the rotation that minimizes the $\ell_1$-norm of the loadings. The exception occurs in the regions of weak identification 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 factors in $\mathcal{F}$ are identified, and clearly $F_1 \notin \mathcal{F}$ in this region, this is not surprising. On the opposite side of Figure 6c only a handful of outcomes are affected by $F_1$. $F_1$ is therefore only weakly identified, and our initial estimate of the factor space is poor, resulting in a maximum cosine similarity less than one. Of course, the Principal Component estimator fares no better in this region of the parameter space (cf. Figure 6a). The same reasoning applies to Figure 6d. We further conclude from panels 6c-6d that an identification failure for one of the factors does not imply identification failure for the other.




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

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

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

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

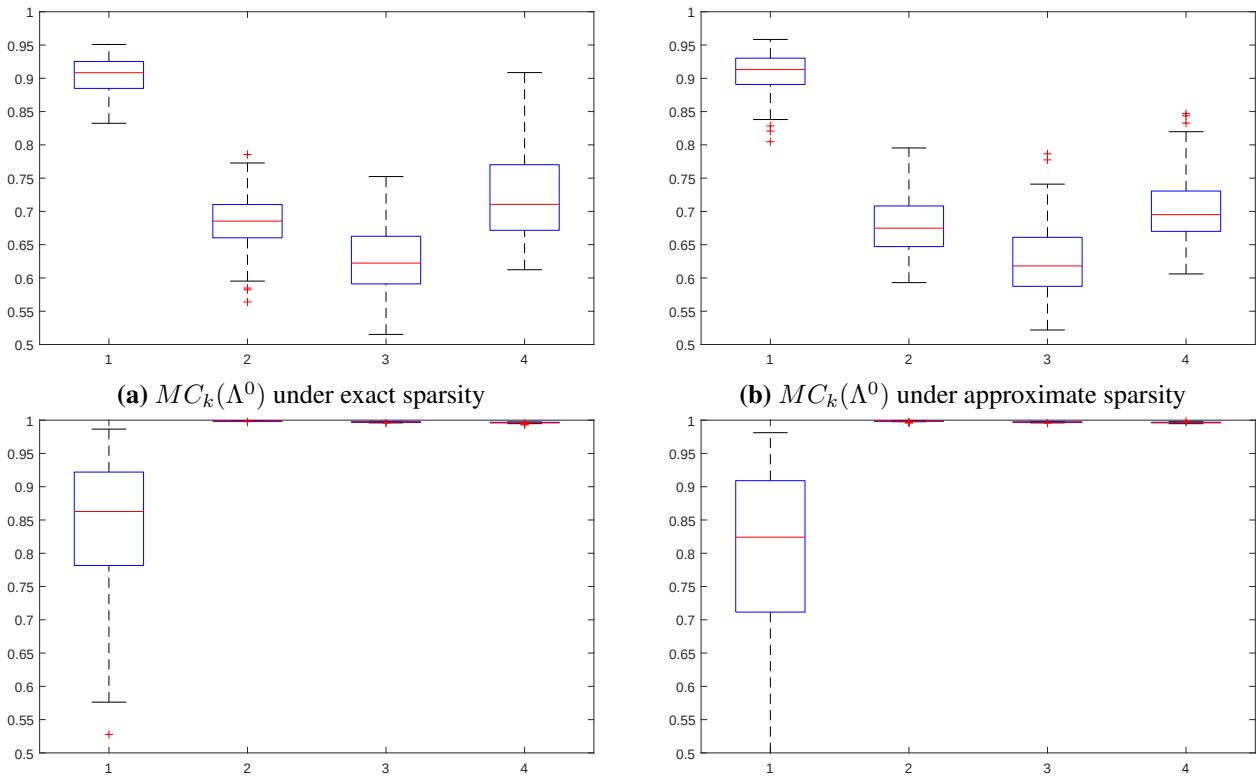
**Figure 6:** Maximum cosine similarity of estimators with each of the true loading vectors $\lambda^*_{\bullet k}$ as a function of the degree of sparsity in the loading matrix. $m_k$ refers to the number of non-zero entries in the $k$th column of $\Lambda^*$. Depicted are averages over 100 realizations.

We next increase the size of the model and consider a DGP with $(T, n) = (500, 300)$[7] 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

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

We consider the first factor "global," while the remaining three are local to varying degrees, meaning they do not affect all outcomes. Specifically, the 500-by-4 loading matrix $\Lambda^*$ has entries

[7] $(T, n) = (500, 300)$ falls roughly in the middle of the range of dimensions usually considered in the literature. For example, Bai and Ng (2002) consider sample sizes in both dimensions between 40 and 8000.




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

**(b)** $MC_k(\Lambda^0)$ under approximate sparsity

**(c)** $MC_k(\tilde{\Lambda})$ under exact 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 $\lambda^*_{\bullet k}$. $\Lambda^0$ 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 100 realizations.

$\lambda^*_{ik} = (1 + \eta_{ik})$ with $\eta_{ik} \overset{i.i.d.}{\sim} N(0, 1)$ 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 $\{|\mathcal{A}_k|\}_{k=1}^4 = \{500, 269, 144, 106\}$ for the four factors[8]. 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, $\lambda^*_{ik} \overset{i.i.d.}{\sim} N(0, \sigma^2)$, and $\sigma^2 = \frac{1}{n}$ for all $i \in \mathcal{A}_k^c$.

Figure 7 uses a boxplot to visualize the performance of $\Lambda^0$ and $\tilde{\Lambda}$. It depicts the maximum cosine similarity for each factor across 100 realizations of the DGP. The data underlying Figures 7a and 7c has an exact sparsity pattern. As expected, we do not consistently recover the true loadings using the Principal Component estimator, as illustrated in Figure 7a.

On the other hand, Figure 7c depicts the similarity between the truth and the estimated factor loadings that have been rotated to minimize the $\ell_1$-norm of $\Lambda$. Since the first factor does not exhibit any sparsity, there is no information in the $\ell_1$-norm that could help identify this factor. As a consequence, the similarity is below one, and identification fails for this "global" factor.

[8]This corresponds to group sizes $\{|\mathcal{A}_k|\}_{k=1}^4 = \{n, n^{0.9}, n^{0.8}, n^{0.75}\}$, rounded to the nearest integer.

On the other hand, the three local factors exhibit maximum cosine similarities that are visually indistinguishable from one in all realizations.

Underlying Figures 7b and 7d is the variant of our DGP with only an approximate sparsity pattern in the loading matrix. Based on Figures 7b 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.

In Online Appendix B, 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. We find that our criterion outperforms these alternative methods.

# 6 Applications

We next apply our proposed method to two economic applications in which factor models have been widely used, and that capture two important scenarios a practitioner might encounter. In the first scenario, we expect the majority of the factors to be local, while in the second scenario, local factors may or may not be present.

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 whether our method can detect these local factors and uncover 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.

## 6.1 Identifying Common Shocks in International Asset Returns

The idea of common risk factors across assets, implied by a factor structure in their returns, underlies the Arbitrage Pricing Theory of Ross (1976), such that, for the excess return of asset $i$ at time $t$, $x_{it}$, we obtain

$$x_{it} = R_{it} - \mathbb{E}(R_{it}) = \lambda_i F_t + e_{it}, \tag{20}$$

in the notation of this paper. As common factors are not observed, we need to replace $F_t$ and $\lambda_i$ by their estimates $\hat{F}_t$ and $\hat{\lambda}_i$. In financial economics, these estimates are commonly obtained by Principal Component Analysis (Connor and Korajczyk 1986, Ludvigson and Ng 2007).

We might be interested in $\mathbb{E}_t\, x_{t+1}$, the conditional mean of $x_{t+1}$, using the information up to

time $t$, as in Ludvigson and Ng (2007), who suggest a model akin to

$$x_{it+1} = \alpha' \hat{G}_t + \varepsilon_{it}, \tag{21}$$

where $\hat{G}_t$ is formed from a subset of $\hat{F}_t$ and/or functions of its individual components $\hat{F}_{kt}$ (e.g., $\hat{F}_{1t}^2$). However, since the Principal Component estimator will estimate a rotation of the true factors and their loadings, selecting a subset and/or taking non-linear transformations of $\hat{F}_t$ will not be invariant to the chosen rotation. We propose to avoid this issue by using our proposed rotation criterion to identify the individual factors and their loadings.

We consider daily returns for a large number of stocks from different parts of the world. In particular, our data consists of individual stock returns for companies that were either part of the *DAX30* (Germany), the *FTSE100* (UK), the *S&P100* (US), the *CAC40* (France), or the *TA100* (Middle East) on April 23, 2015.[9] 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, and will accordingly assume $r = 8$ in what follows. Figures 8 and 9 illustrate the loading vectors in $\Lambda^0$ and $\tilde{\Lambda}$. The thin dashed lines separate the geographical groups as described above, in the order of Frankfurt, London, New York, Paris, and Tel Aviv.

We see in Figure 8 that, as expected, the normalization implied by the Principal Component estimator results in a dense loading matrix, meaning each factor is associated with a large number of significant loadings. While there are some discernible regional patterns (e.g., $\lambda_{\bullet 2}^0$ has positive entries for all stocks listed in the US, and negative entries for all others), it appears unlikely we can think of these estimated loadings (and the implied factors) as structural objects.[10]

Figure 9 depicts our proposed estimator $\tilde{\Lambda}$. In contrast to Figure 8, we see that its loading vectors are highly concentrated on a subset of outcomes.[11] In particular, 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 Middle East, $\tilde{\lambda}_{\bullet 3}$ and $\tilde{\lambda}_{\bullet 7}$ are concentrated on stocks in the US, and $\tilde{\lambda}_{\bullet 8}$ is concentrated on stocks in the UK. The exception is $\tilde{\lambda}_{\bullet 2}$, whose large entries are dispersed geographically. However, of the 24 entries with a loading larger than 2, all 24 stocks belong to either

[9] We further restrict the stocks in the TA100 to those with a weight by market capitalization in the TA100 of at least .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 E.1.

[10] Instead, we conjecture that this pattern is largely an artificial result of the chosen normalization. For example, suppose the first loading vector happens to capture a structural object, and suppose there is a second factor that positively affects all US stocks. Intuitively, the negative loadings associated with all international stocks in $\lambda_{\bullet 2}^0$ will then be necessary to guarantee orthogonality between $\lambda_{\bullet 2}^0$ and $\lambda_{\bullet 1}^0$.

[11] We again stress that there is no "shrinkage" involved in our estimator, and therefore 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 (2019) for a potential approach to such regularization.

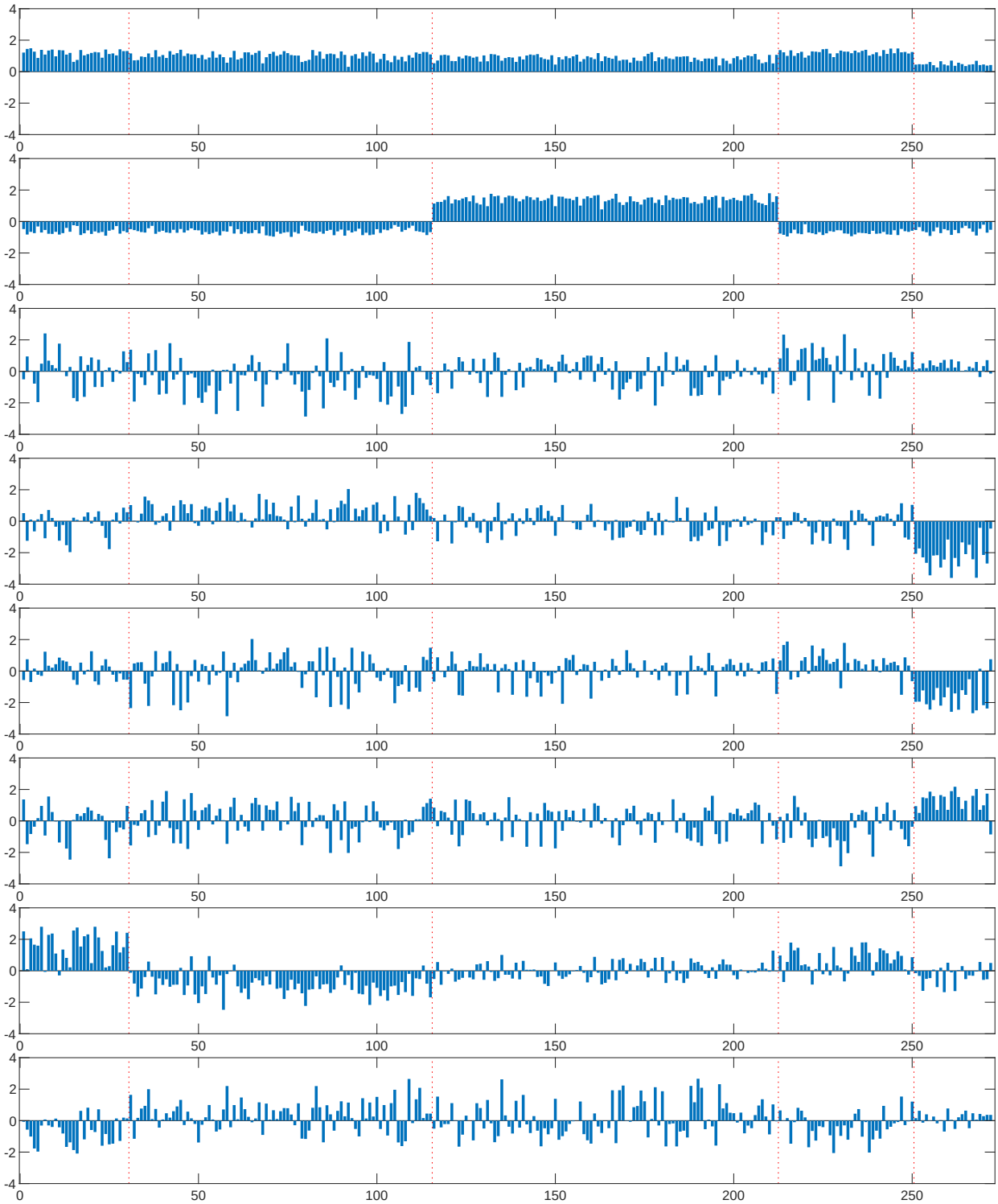



**Figure 8:** Illustration of PCA loadings $\lambda^0_{\bullet k}$ for $k = 1, \ldots, 8$ in panel of international asset returns. Bars correspond to the loadings of the 272 individual stocks. Geographical groups are Germany, UK, US, France, and Middle East, separated by dashed lines.

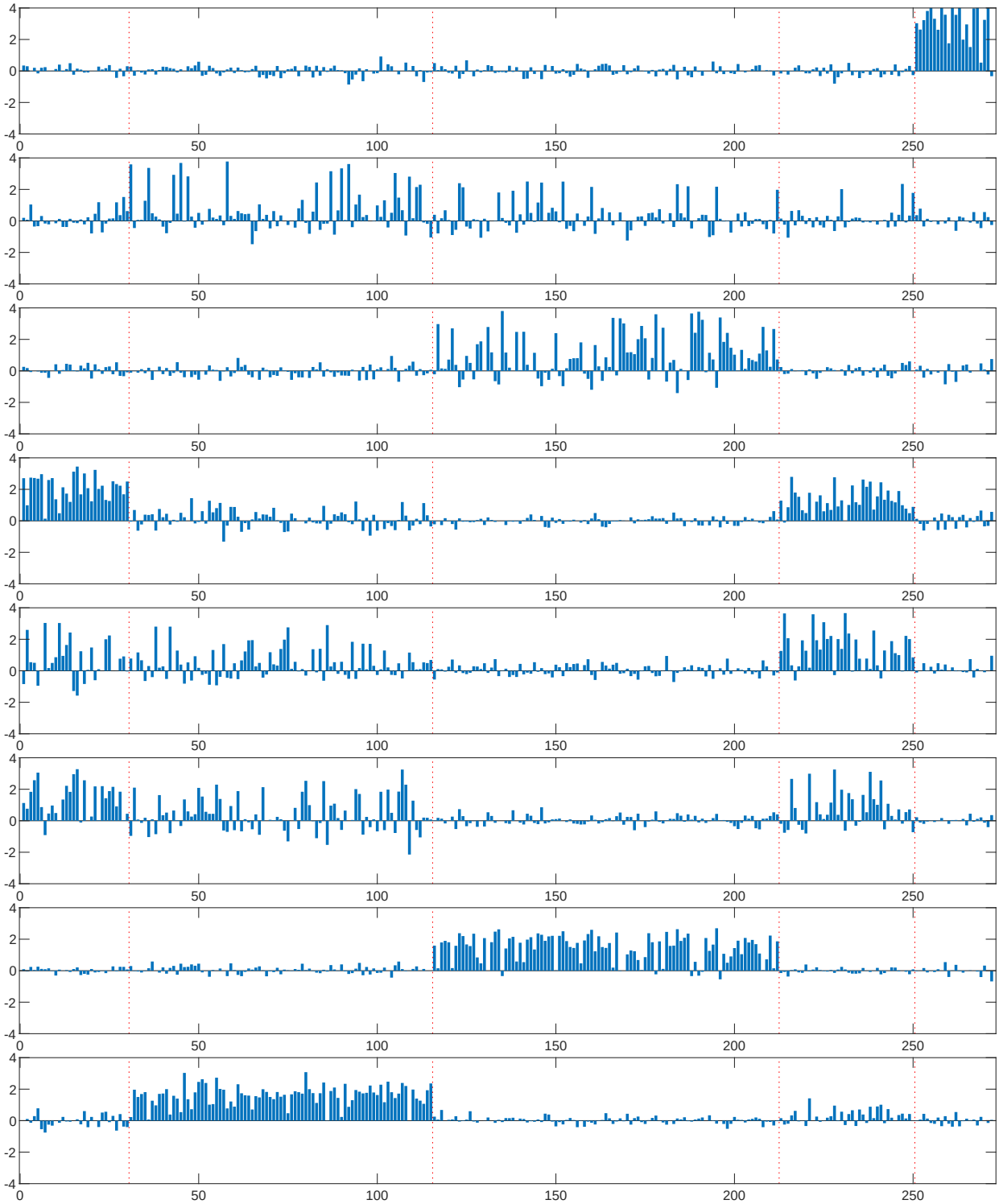



**Figure 9:** Illustration of the rotated loading vectors $\tilde{\lambda}_{\bullet k}$ for $k = 1, \ldots, 8$ in panel of international asset returns. Bars correspond to the loadings of the 272 individual stocks for the $k$th estimated loading vector. Geographical groups are Germany, UK, US, France, and Middle East, separated by dashed lines.

the *Oil & Gas* or the *Mining* sectors, enabling us to clearly label this a sectoral shock.

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

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

Figure 10 illustrates how, in this dataset, all columns of $\tilde{\Lambda}$ have significantly smaller $\ell_1$-norms than those in $\Lambda^0$. We take this as a sign that all eight estimated factors are identified using our proposed rotation criterion.

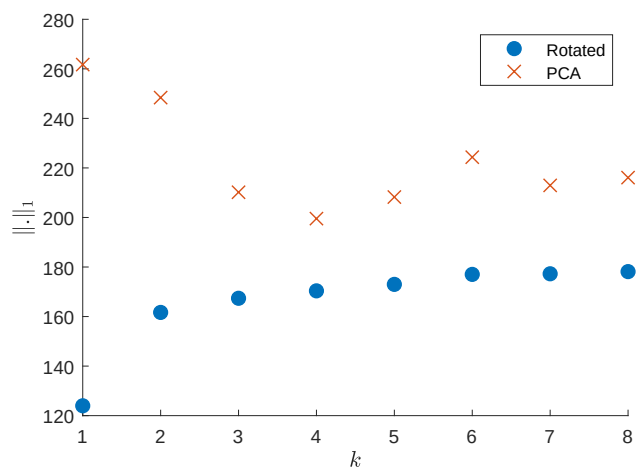



**Figure 10:** For each $k = 1, \ldots, 8$, the points above represent $\|\lambda_{\bullet k}\|_1$, the $\ell_1$-norm of the corresponding column in $\Lambda$ for daily returns of an international sample of stock returns. Blue dots correspond to columns in $\tilde{\Lambda}$, while red crosses correspond to $\Lambda^0$. Smaller values tend to be associated with a greater degree of sparsity in the loading vector.

*Remark* 5. $\tilde{\Lambda}$ is also appealing for the construction of factor-mimicking portfolios.[12] Not only do our results allow the construction of factor-mimicking portfolios for individual factors but, in finance, trading costs are usually modeled as either a proportional transaction cost or as a fixed brokerage fee per transaction (e.g., Liu 2004, Brodie et al. 2009). For the construction of factor-mimicking portfolios, a proportional transaction cost implies that the cost is directly related to the $\ell_1$-norm of the loadings, and with a fixed fee per transaction, factor-mimicking portfolios with fewer non-zero positions will imply lower transaction costs.

[12] See Huberman et al. (1987) and Green and Hollifield (1992) for more details.

## 6.2 Identifying Shocks to the US Economy

We next apply our identification strategy to a large panel of US macroeconomic indicators. Versions of this dataset have been used extensively in the macroeconomic literature (Stock and Watson 2005, De Mol et al. 2008, Stock and Watson 2016). In particular, we use the FRED-QD data collected and maintained by Michael W. McCracken.[13] Among others, it includes real activity variables, prices, productivity and earnings, interest rates and spreads, money and credit, asset and wealth variables, oil market variables, and indicators representing international activities. We use data from 1967Q1-2019Q1 and follow the initial transformations of the raw data as outlined in McCracken and Ng (2016) to achieve approximate stationarity and remove a small number of outliers.

The dataset consists of series at multiple levels of aggregation. We only use the disaggregated time series in my estimation of the factor structure and disregard the aggregates (Boivin and Ng 2006, Stock and Watson 2016). We also drop a small number of series with missing observations, so that the final data contains 206 quarterly observations of 166 macroeconomic variables, primarily for the US economy.

Using an earlier vintage of a closely related dataset, De Mol et al. (2008) focuses on macroeconomic forecasting, with the monthly industrial production index and the consumer price index as their primary prediction targets. When comparing forecast models on this data, they make 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 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."

One explanation for these findings is that there are local factors present in the data, with each factor only affecting a subset of the observed indicators (which will generally belong to the same economic category). In such a setting, if the overlap between the active sets $\mathcal{A}_k$ is small, the Lasso will tend to select a single variable from each group $\mathcal{A}_k$ as a noisy proxy for $F_k$, and the selected set of regressors will approximately span the space of the common factors. However, selection within groups will be unstable and sensitive to minor perturbations of the data, which would explain the

[13] Data are available at `https://research.stlouisfed.org/econ/mccracken/fred-databases`. For a full description of the data, as well as a more detailed description of the transformations of the raw data, we refer the reader to McCracken (2019).

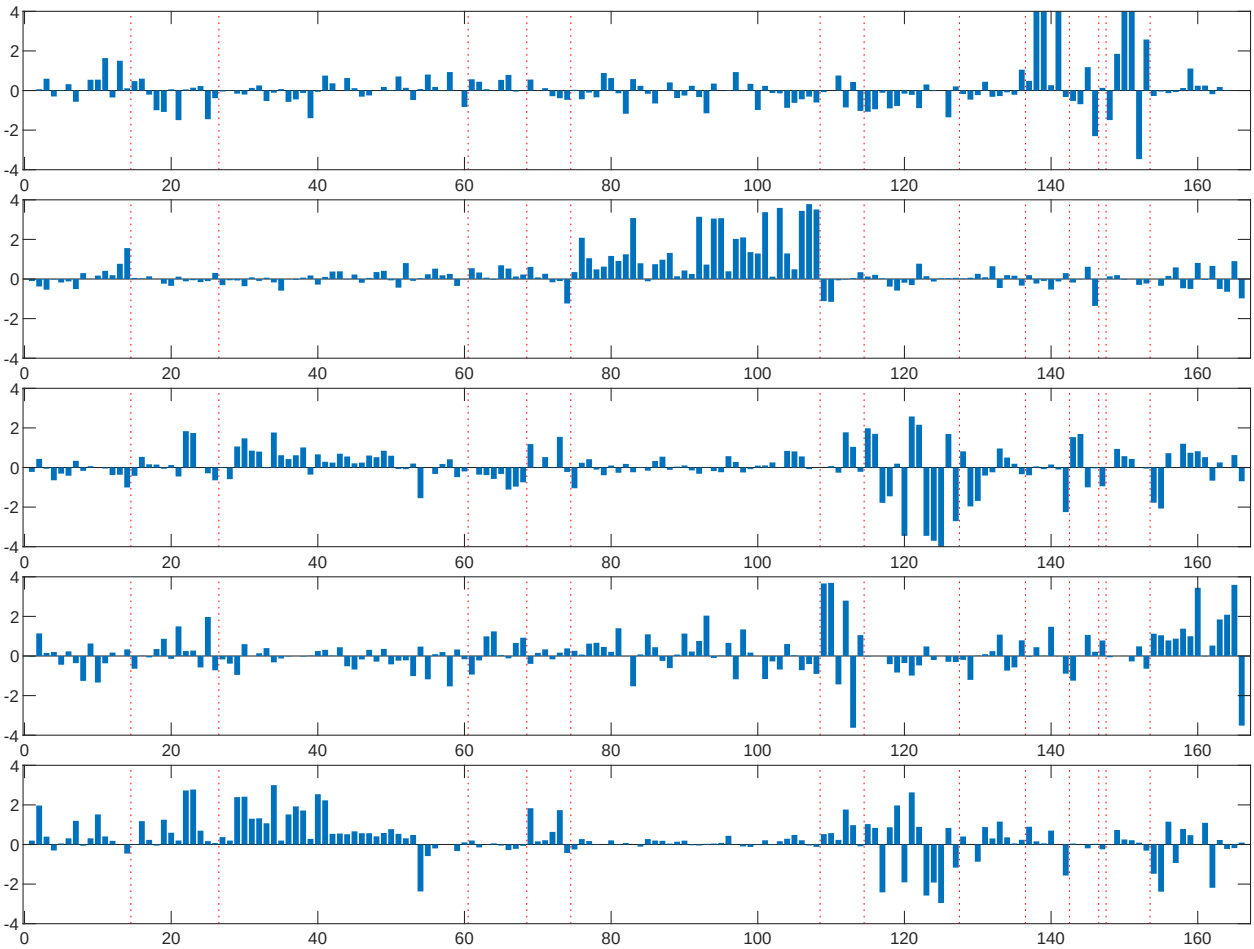



**Figure 11:** Illustration of the rotated loading vectors $\tilde{\lambda}_{\bullet k}$ for $k = 1, \ldots, 5$. Bars correspond to the 166 individual indicators for the $k$th estimated loading vector.

varying choices of variables from the same groups across subsamples. Under such a DGP, our proposed method will rotate the factors to uncover the underlying group structure in a data-driven way.

For our exercise, we first determine the number of factors to be eight, using the Information Criterion of Bai and Ng (2002), and accordingly assume $r = 8$ in what follows. Unsurprisingly, we find that all Principal Components generally load on most of the 166 observed outcomes. The estimated loadings using the Principal Component estimator $\Lambda^0$ can be found in Online Appendix Figure 5.

In contrast, Figure 11 depicts the five columns of our proposed estimator $\tilde{\Lambda}$ with the lowest $\ell_1$-norm. 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 first factor is mainly associated with household balance sheets and stock markets (groups 10 and 13). This captures the intuitive notion that an increase in asset prices will be associated

| 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.

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 second rotated factor almost exclusively drives all price variables (group 6), allowing an easy interpretation. The remaining loading vectors become harder to interpret as their loadings are less concentrated. This is illustrated in Figure 12, which depicts the $\ell_1$-norm of each of the columns in $\tilde{\Lambda}$. Based on Figure 12, we conclude that two factors are identified. Lacking a formal test for identification, we are unsure about factors 3-5, and provide a brief discussion of them next.

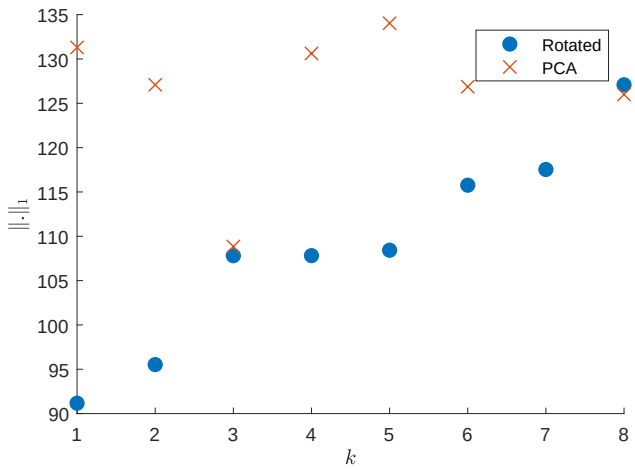



**Figure 12:** For each $k = 1, \ldots, 8$, the points above represent $\|\lambda_{\bullet k}\|_1$, the $\ell_1$-norm of the corresponding column in $\Lambda$ for a panel of US macroeconomic indicators. Blue dots correspond to columns in $\tilde{\Lambda}$, while red crosses correspond to $\Lambda^0$. Smaller values tend to be associated with a greater degree of sparsity in the loading vector.

The third factor affects interest rates and employment indices the most. It is also associated

with a decrease in federal debt and nonfinancial noncorporate business sector liabilities, as well as an increase in capacity utilization (groups 22 and 23). The fourth factor captures the co-movement of Earnings and Productivity indicators (group 7) with Non-Household Balance Sheet variables. The fifth is concentrated on interest rate movements, but also has some effect on employment and industrial production.

Based on the $\ell_1$-norms of their respective loadings, as depicted in Figure 12, we caution against the interpretation of the remaining three factors as structural objects. Their loadings are depicted in Online Appendix Figure 6.

We conclude that we find at least two loading vectors with large enough degrees of sparsity to consider them identified.

# 7 Conclusion

We establish conditions for identification of individual factors in factor models. Specifically, we show that under (approximate) sparsity in the loading matrix a rotation criterion based on the $\ell_1$-norm of the loading matrix can be used to achieve such identification. We validate our proposed criterion in simulations and find that it outperforms existing and widely used heuristics aimed at simplifying the loading matrix.

# References


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

Ando, T. and J. Bai (2017). 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.

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

Bai, J. (2008). Factor models. In S. N. Durlauf and L. E. Blume (Eds.), *The New Palgrave Dictionary of Economics*. Palgrave Macmillan.

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

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

Belloni, A. and V. Chernozhukov (2011). High dimensional sparse econometric models: An introduction. In P. Alquier, E. Gautier, and G. Stoltz (Eds.), *Inverse Problems and High-Dimensional Estimation*. Springer, Berlin, Heidelberg.

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

Brodie, J., I. Daubechies, C. De Mol, D. Giannone, and I. Loris (2009). Sparse and stable Markowitz portfolios. *Proceedings of the National Academy of Sciences 106*(30), 12267–12272.

Bühlmann, P. and S. Van De Geer (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Caner, M. and X. Han (2014). Selecting the correct number of factors in approximate factor models: The large panel case with group bridge estimators. *Journal of Business & Economic Statistics 32*(3), 359–374.

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

Cheng, X., Z. Liao, and F. Schorfheide (2016). Shrinkage estimation of high-dimensional factor models with structural instabilities. *The Review of Economic Studies 83*(4), 1511–1543.

Choi, I., D. Kim, Y. J. Kim, and N.-S. Kwark (2018). A multilevel factor model: Identification, asymptotic theory and applications. *Journal of Applied Econometrics 33*(3), 355–377.

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

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

Dias, F., M. Pinheiro, and A. Rua (2013). Determining the number of global and country-specific factors in the euro area. *Studies in Nonlinear Dynamics and Econometrics 17*(5), 573–617.

Freyaldenhoven, S. (2019). A generalized factor model with local factors. Working paper, Federal Reserve Bank of Philadelphia.

Green, R. C. and B. Hollifield (1992). When will mean-variance efficient portfolios be well diversified? *The Journal of Finance 47*(5), 1785–1809.

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

Hirose, K. and S. Konishi (2012). Variable selection via the weighted group lasso for factor analysis models. *The Canadian Journal of Statistics 40*(2), 345–361.

Huberman, G., S. Kandel, and R. F. Stambaugh (1987). Mimicking portfolios and exact arbitrage pricing. *The Journal of Finance 42*(1), 1–9.

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

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

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika 23*, 187–200.

Katz, J. O. and F. J. Rohlf (1974). Functionplane – a new approach to simple structure rotation. *Psychometrika 39*, 37–51.

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

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

Liu, H. (2004). Optimal consumption and investment with transaction costs and multiple risky assets. *The Journal of Finance 59*(1), 289–338.

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

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

McCracken, M. W. (2019). Fred-qd updated appendix. Working paper, Federal Reserve Bank of St. Louis.

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

Moench, E., S. Ng, and S. Potter (2013). Dynamic hierarchical factor models. *Review of Economics and Statistics 95*(5), 1811–1817.

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

Pelger, M. and R. Xiong (2019). Interpretable sparse proximate factors for large dimensions. Working paper.

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

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

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

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

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

Stock, J. H. and M. W. Watson (2005). Implications of dynamic factor models for VAR analysis. *National Bureau of Economic Research Working Paper 11467*.

Stock, J. H. and M. W. Watson (2012). Disentangling the channels of the 2007—09 recession. *Brookings Papers on Economic Activity* (1), 81–135.

Stock, J. H. and M. W. Watson (2016). Dynamic factor models, factor-augmented vector autoregressions, and structural vector autoregressions in macroeconomics. *Handbook of Macroeconomics 2*, 415–525.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological) 58*(1), 267–288.

Wang, P. (2008). Large dimensional factor models with a multi-level factor structure. Working paper, New York University.

Yin, P., E. Esser, and J. Xin (2014). Ratio and difference of l1 and l2 norms and sparse representation with coherent dictionaries. *Communications in Information and Systems 14*(2), 87–109.

Yin, P., Y. Lou, Q. He, and J. Xin (2015). Minimization of l1-2 for compressed sensing. *SIAM Journal of Scientific Computing 37*(1), A536–563.

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B 68*(1), 49–67.

Zou, H., T. Hastie, and R. Tibshirani (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics 15*(2), 265–286.

# A Auxiliary Lemmata

**Lemma 1.**

$$\lim_{\epsilon \to 0} \frac{1 - \sqrt{1-\epsilon^2}}{\epsilon} = 0.$$

*Proof.* This immediately follows from L'Hospital's rule:

$$\lim_{\epsilon \to 0} \frac{1 - \sqrt{1-\epsilon^2}}{\epsilon} = \lim_{\epsilon \to 0} \frac{2\epsilon}{2\sqrt{1-\epsilon^2}} = 0.$$

□

**Lemma 2.** *Under Assumptions 3 and 4, there exists a $c_0 < \infty$, such that for all $\epsilon \in (0, c_0)$:*

$$\|\sqrt{1-\epsilon^2}\lambda_{\bullet k}^{*\mathcal{A}_k} + \epsilon v_{\bullet k}^{\mathcal{A}_k}\|_1 \geq \sqrt{1-\epsilon^2}\|\lambda_{\bullet k}^{*\mathcal{A}_k}\|_1 - \epsilon\, \beta^k(v_{\bullet k}).$$

*Proof.* Since $\left|\lambda_{ik}^*\right| > c\ \forall i \in \mathcal{A}_k$, for small enough $\epsilon > 0$:

$$\begin{aligned}
&\|\sqrt{1-\epsilon^2}\lambda_{\bullet k}^{*\mathcal{A}_k} + \epsilon v_{\bullet k}^{\mathcal{A}_k}\|_1 \\
&= \sum_{i \in \mathcal{A}_k} \left(\sqrt{1-\epsilon^2}|\lambda_{ik}^*| + \epsilon\left|v_{ik}^{\mathcal{A}_k}\right|\right) \mathbf{1}\{\lambda_{ik}^* v_{ik}^{\mathcal{A}_k} \geq 0\} + \sum_{i \in \mathcal{A}_k} \left(\sqrt{1-\epsilon^2}|\lambda_{ik}^*| - \epsilon\left|v_{ik}^{\mathcal{A}_k}\right|\right) \mathbf{1}\{\lambda_{ik}^* v_{ik}^{\mathcal{A}_k} < 0\} \\
&= \sqrt{1-\epsilon^2}\|\lambda_{\bullet k}^{*\mathcal{A}_k}\|_1 + \epsilon \left(\sum_{i \in \mathcal{A}_k} \left|v_{ik}^{\mathcal{A}_k}\right| \mathbf{1}\{\lambda_{ik}^* v_{ik}^{\mathcal{A}_k} \geq 0\} - \sum_{i \in \mathcal{A}_k} \left|v_{ik}^{\mathcal{A}_k}\right| \mathbf{1}\{\lambda_{ik}^* v_{ik}^{\mathcal{A}_k} < 0\}\right) \\
&= \sqrt{1-\epsilon^2}\|\lambda_{\bullet k}^{*\mathcal{A}_k}\|_1 + \epsilon \left(\sum_{i \in \mathcal{A}_k} |v_{ik}|\, \mathbf{1}\{\lambda_{ik}^* v_{ik} \geq 0\} - \sum_{i \in \mathcal{A}_k} |v_{ik}|\, \mathbf{1}\{\lambda_{ik}^* v_{ik} < 0\}\right) \\
&\geq \sqrt{1-\epsilon^2}\|\lambda_{\bullet k}^{*\mathcal{A}_k}\|_1 - \epsilon\, \beta^k(v_{\bullet k}).
\end{aligned}$$

□

**Lemma 3.** *Suppose Assumption 4 holds and let $\dot{\lambda}_{\bullet k}$ be a vector with $\|\dot{\lambda}_{ik} - \lambda_{ik}^*\| = O_p(\frac{1}{\sqrt{n}})\ \forall i$ as $n \to \infty$ and let $\dot{v}_{\bullet k}$ be a linear combination of $\dot{\lambda}_{\bullet l}$, $l = 1, \ldots, r$, such that $\|\dot{v}_{\bullet k}\|_2^2 = n$ and $\dot{\lambda}_{\bullet k} \perp \dot{v}_{\bullet k}$.*

*Then, there exists a vector $v_{\bullet k}$ with*

1. $\lambda_{\bullet k}^{*\prime} v_{\bullet k} = 0$,
2. $(\dot{v}_{ik} - v_{ik}) = O_p(\frac{1}{\sqrt{n}})\ \forall i$.

*Proof.* Define $S_k = \sum_i^{\mathcal{A}_k^c} \big(\lambda_{ik}^* - \dot{\lambda}_{ik}\big)\dot{v}_{ik}$ and note that for some constant $V < \infty$, $S_k \leq V \sum_i^{\mathcal{A}_k^c} \big(\lambda_{ik}^* - \dot{\lambda}_{ik}\big) = O(\sqrt{n})$.

Let $s_{ik} = -\frac{S_k}{\lambda_{ik}^* |\mathcal{A}_k|}$ and

$$v_{ik} = \begin{cases} \dot{v}_{ik} - \frac{\dot{v}_{ik}}{\lambda_{ik}^*}(\lambda_{ik}^* - \dot{\lambda}_{ik}) + s_{ik} & \text{if } i \in \mathcal{A}_k \\ \dot{v}_{ik} & \text{otherwise.} \end{cases} \tag{22}$$

Then,

$$\begin{aligned}
\sum_{i=1}^n \lambda_{ik}^* v_{ik} &= \sum_i^{\mathcal{A}_k} \lambda_{ik}^* v_{ik} + \sum_i^{\mathcal{A}_k^c} \lambda_{ik}^* v_{ik} \\
&= \sum_i^{\mathcal{A}_k} \lambda_{ik}^* \Big( \dot{v}_{ik} - \frac{\dot{v}_{ik}}{\lambda_{ik}^*}(\lambda_{ik}^* - \dot{\lambda}_{ik}) + s_{ik} \Big) + \sum_i^{\mathcal{A}_k^c} \lambda_{ik}^* \dot{v}_{ik} \\
&= \sum_i^{\mathcal{A}_k} \Big( \lambda_{ik}^* \dot{v}_{ik} - \dot{v}_{ik}(\lambda_{ik}^* - \dot{\lambda}_{ik}) + \lambda_{ik}^* s_{ik} \Big) + \sum_i^{\mathcal{A}_k^c} \lambda_{ik}^* \dot{v}_{ik} \\
&= \sum_i^{\mathcal{A}_k} \Big( \dot{\lambda}_{ik} \dot{v}_{ik} + \lambda_{ik}^* s_{ik} \Big) + \sum_i^{\mathcal{A}_k^c} \lambda_{ik}^* \dot{v}_{ik} \\
&= \sum_i^n \dot{\lambda}_{ik} \dot{v}_{ik} - \sum_i^{\mathcal{A}_k^c} \dot{\lambda}_{ik} \dot{v}_{ik} + \sum_i^{\mathcal{A}_k} \lambda_{ik}^* s_{ik} + \sum_i^{\mathcal{A}_k^c} \lambda_{ik}^* \dot{v}_{ik} \\
&= -\sum_i^{\mathcal{A}_k^c} \dot{\lambda}_{ik} \dot{v}_{ik} + \sum_i^{\mathcal{A}_k} \lambda_{ik}^* s_{ik} + \sum_i^{\mathcal{A}_k^c} \lambda_{ik}^* \dot{v}_{ik} \\
&= \sum_i^{\mathcal{A}_k} \lambda_{ik}^* s_{ik} + \sum_i^{\mathcal{A}_k^c} \big(\lambda_{ik}^* - \dot{\lambda}_{ik}\big) \dot{v}_{ik} \\
&= -\sum_i^{\mathcal{A}_k} \lambda_{ik}^* \frac{S_k}{\lambda_{ik}^* |\mathcal{A}_k|} + S_k \\
&= 0.
\end{aligned}$$

Further, since by Assumption 4, $|\mathcal{A}_k| > c_0 n$ for some $c_0 > 0$, $s_{ik} = O_p(\frac{1}{\sqrt{n}})$. Combined with the fact that $\frac{v_{ik}}{\dot{\lambda}_{ik}}(\dot{\lambda}_{ik} - \lambda_{ik}^*) = O_p(\frac{1}{\sqrt{n}})$, we therefore conclude that $\dot{v}_{ik} - v_{ik} = O_p(\frac{1}{\sqrt{n}})$, which completes the proof. □

Intuitively, Lemma 3 states that, if two vectors are "close," their orthogonal complements are also close.

**Lemma 4.** *Suppose Assumptions 4-5 hold and $F_k \in \mathcal{F}$. Assume $\dot{\lambda}_{ik} - \lambda^*_{ik} = O(\frac{1}{\sqrt{n}})\ \forall i$ and let $\dot{v}_{\bullet k}$ be a linear combination of $\dot{\lambda}_{\bullet l}$, $l = 1, \ldots, r$, such that $\|\dot{v}_{\bullet k}\|_2^2 = n$ and $\dot{\lambda}_{\bullet k} \perp \dot{v}_{\bullet k}$. Then,*

$$\Bigg(\sum_{i\in\mathcal{A}_k} |v_{ik}|\, \mathbf{1}\{\lambda^*_{ik}\dot{v}_{ik} \geq 0\} - \sum_{i\in\mathcal{A}_k} |v_{ik}|\, \mathbf{1}\{\lambda^*_{ki}\dot{v}_{ik} < 0\}\Bigg)$$
$$= \Bigg(\sum_{i\in\mathcal{A}_k} |v_{ik}|\, \mathbf{1}\{\lambda^*_{ik}v_{ik} \geq 0\} - \sum_{i\in\mathcal{A}_k} |v_{ik}|\, \mathbf{1}\{\lambda^*_{ik}v_{ik} < 0\}\Bigg) + O(\sqrt{n}).$$

*Proof.* Let $\mathcal{S} = \{i : v_{ik} = O_p(\frac{1}{\sqrt{n}})\}$. Then,

$$\begin{aligned}
&\Bigg(\sum_{i\in\mathcal{A}_k} |v_{ik}|\, \mathbf{1}\{\lambda^*_{ik}\dot{v}_{ik} \geq 0\} - \sum_{i\in\mathcal{A}_k} |v_{ik}|\, \mathbf{1}\{\lambda^*_{ik}\dot{v}_{ik} < 0\}\Bigg) \\
&= \Bigg(\Big(\sum_{i\in\mathcal{A}_k\cap\mathcal{S}} |v_{ik}|\, \mathbf{1}\{\lambda^*_{ik}\dot{v}_{ik} \geq 0\} + \sum_{i\in\mathcal{A}_k\cap\mathcal{S}^c} |v_{ik}|\, \mathbf{1}\{\lambda^*_{ik}\dot{v}_{ik} \geq 0\}\Big) \\
&\qquad - \Big(\sum_{i\in\mathcal{A}_k\cap\mathcal{S}} |v_{ik}|\, \mathbf{1}\{\lambda^*_{ik}\dot{v}_{ik} < 0\} + \sum_{i\in\mathcal{A}_k\cap\mathcal{S}^c} |v_{ik}|\, \mathbf{1}\{\lambda^*_{ik}\dot{v}_{ik} < 0\}\Big)\Bigg) \\
&= \Bigg(\Big(\sum_{i\in\mathcal{A}_k\cap\mathcal{S}} |v_{ik}|\, \mathbf{1}\{\lambda^*_{ik}\dot{v}_{ik} \geq 0\} + \sum_{i\in\mathcal{A}_k\cap\mathcal{S}^c} |v_{ik}|\, \mathbf{1}\{\lambda^*_{ik}v_{ik} \geq 0\}\Big) \\
&\qquad - \Big(\sum_{i\in\mathcal{A}_k\cap\mathcal{S}} |v_{ik}|\, \mathbf{1}\{\lambda^*_{ik}\dot{v}_{ik} < 0\} + \sum_{i\in\mathcal{A}_k\cap\mathcal{S}^c} |v_{ik}|\, \mathbf{1}\{\lambda^*_{ik}v_{ik} < 0\}\Big)\Bigg) \\
&= \Bigg(\Big(O_p(\frac{1}{\sqrt{n}}) + \sum_{i\in\mathcal{A}_k\cap\mathcal{S}^c} |v_{ik}|\, \mathbf{1}\{\lambda^*_{ik}v_{ik} \geq 0\}\Big) \\
&\qquad - \Big(O_p(\frac{1}{\sqrt{n}}) + \sum_{i\in\mathcal{A}_k\cap\mathcal{S}^c} |v_{ik}|\, \mathbf{1}\{\lambda^*_{ik}v_{ik} < 0\}\Big)\Bigg) \\
&= \Bigg(\sum_{i\in\mathcal{A}_k\cap\mathcal{S}^c} |v_{ik}|\, \mathbf{1}\{\lambda^*_{ik}v_{ik} \geq 0\} - \sum_{i\in\mathcal{A}_k\cap\mathcal{S}^c} |v_{ik}|\, \mathbf{1}\{\lambda^*_{ik}v_{ik} < 0\}\Bigg) + O_p(\frac{1}{\sqrt{n}}) \\
&= \Bigg(\sum_{i\in\mathcal{A}_k} |v_{ik}|\, \mathbf{1}\{\lambda^*_{ik}v_{ik} \geq 0\} - \sum_{i\in\mathcal{A}_k} |v_{ik}|\, \mathbf{1}\{\lambda^*_{ik}v_{ik} < 0\}\Bigg) + O_p(\frac{1}{\sqrt{n}}),
\end{aligned}$$

where we used Lemma 3 in the second and third equality above. □