Federal Reserve Bank of Philadelphia

Working Papers
Research Department

WP 21-15
April 2021
https://doi.org/10.21799/frbp.wp.2021.15

Factor Models with Local Factors—Determining the Number of Relevant Factors

Simon Freyaldenhoven
Federal Reserve Bank of Philadelphia Research Department

ISSN: 1962-5361

Factor Models with Local Factors - Determining the Number of Relevant Factors*

Simon Freyaldenhoven†

Federal Reserve Bank of Philadelphia

April 9, 2021

Abstract

We extend the theory on factor models by incorporating "local" factors into the model. Local factors affect only an unknown subset of the observed variables. This implies a continuum of eigenvalues of the covariance matrix, as is commonly observed in applications. We derive which factors are pervasive enough to be economically important and which factors are pervasive enough to be estimable using the common principal component estimator. We then introduce a new class of estimators to determine the number of those relevant factors. Unlike existing estimators, our estimators use not only the eigenvalues of the covariance matrix, but also its eigenvectors. We find that incorporating partial sums of the eigenvectors into our estimators leads to significant gains in performance in simulations.

JEL-Classification: C38, C52, C55

KEYWORDS: high-dimensional data, factor models, weak factors, local factors, sparsity

*I am grateful to Frank Kleibergen, Adam McCloskey, Eric Renault and Jesse Shapiro for their insightful suggestions. I also thank Christian Hansen, Phillip Ketz, Susanne Schennach, and audiences at Brown University, the European School of Management and Technology, the Federal Reserve Bank of Philadelphia, the University of Amsterdam, the University of Konstanz, the University of Pennsylvania, the University of St. Gallen, the University of Virginia, as well as participants of the 2016 European Conference of the Econometrics Community, the 2016 Maastricht Workshop on Advances in Quantitative Economics, the 2017 CIREQ Econometrics Conference, the 2017 NBER/NSF Time Series Conference, and the 2019 Annual Conference of the International Association for Applied Econometrics for helpful comments. An earlier draft of this paper was circulated under the title "A Generalized Factor Model with Local Factors".

†simon.freyaldenhoven@phil.frb.org

1 Introduction

Standard factor models imply a clear partition of the eigenvalues of the observed covariance matrix into two sets – large eigenvalues representing factor-related variation and small eigenvalues representing idiosyncratic variation. However, a clear separation of the eigenvalues into one set of large eigenvalues and a second set of small eigenvalues is typically not found in practice. This is exemplified in Figure 1, which depicts the leading eigenvalues from a large panel of US macroeconomic indicators. Empirical studies on Arbitrage Pricing Theory (Ross 1976) similarly point to a continuum of factor strengths. For example, in a cross section of asset returns, Trzcinka (1986) finds that, while the first eigenvalue dominates, the first 6 eigenvalues diverge at differing rates.

This paper develops a model that can produce the kind of eigenvalue distributions typically found in practice by incorporating "local" factors into the model. Local factors affect only a subset of the observed variables. Their presence implies a continuum of eigenvalues, thereby describing data such as the one underlying Figure 1 better than a model that implies a large gap between two sets of eigenvalues.

In such a model, in which factors affect arbitrary subsets of the observed variables, we derive which factors are pervasive enough to be economically important and which factors are estimable using the common principal component estimator. We will refer to these as the "relevant" factors.

We then introduce a new class of estimators to determine the number of those relevant factors. While there exists a multitude of estimators for the number of factors (e.g. Bai and Ng 2002, Onatski 2010, Ahn and Horenstein 2013), existing estimators are generally derived from

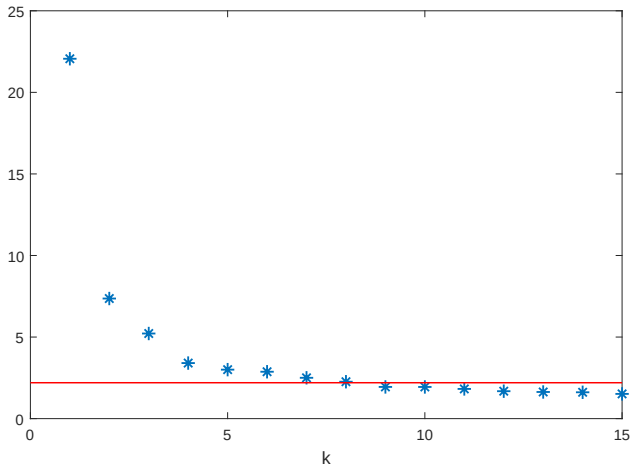


Figure 1: 20 largest eigenvalues of the covariance matrix for a dataset of 94 macroeconomic indicators in the US. The data is an updated vintage of the "Stock & Watson" dataset (Stock and Watson 2002, De Mol et al. 2008), a popular dataset in which factor models have been used in the literature, and has been transformed to achieve stationarity. Solid line indicates cutoff chosen according to Bai and Ng (2002) (with $r_{max} = 15$) to determine the number of factors.

the empirical distribution of the eigenvalues. Under the presence of local factors, these estimators are theoretically invalid and tend to perform poorly in practice. Instead, our proposed estimators combine the empirical eigenvalue distribution with information obtained from the corresponding eigenvectors. Exploiting this additional information by incorporating partial sums of the eigenvectors into our estimators leads to significant gains in performance.[1]

To fix ideas, consider the following single factor model:

$$x_{it} = \lambda_i F_t + e_{it}, \quad i = 1, \dots, n; \; t = 1, \dots, T, \tag{1}$$

where λ_i denotes the factor loading for outcome i, F_t the factor at time t, and e_{it} an idiosyncratic noise component. Standard factor models typically assume that $\sum_i^n \lambda_i^2$ diverges at rate n. Intuitively, this condition states that the factor is "strong" - on average, the loadings λ_i are bounded away from zero. In this paper, we relax this assumption and allow that $\sum_i^n \lambda_i^2$ diverges at rate n^α, $\alpha \in [0, 1]$. If $\alpha < 1$, we call the corresponding factor "weak". A factor can be weak for two reasons. One reason could be that the factor has a weak effect on all observables. Formally, in this case the individual loadings are modeled as drifting sequences (e.g. $\lambda_i^{(n)} = \frac{1}{\sqrt{n}}\lambda_i^*$, where λ_i^* are constants). Kleibergen (2009) and Onatski (2012) are examples of settings in which a weak factor affects all outcomes but its associated loadings are drifting towards zero. Alternatively, a factor can be weak because it affects only a subset of observables. This can similarly be modeled through drifting sequences. Instead of using drifting sequences for individual factor loadings, now the fraction of loadings that are non-zero are modeled as a drifting sequence, while the non-zero loadings themselves are fixed. This is the approach we pursue in this paper, and we refer to factors that are weak in this sense as "local". A similar framework is considered in Bailey et al. (2020) and Uematsu and Yamagata (2020).

Onatski (2009, 2010, 2012) proposes an alternative framework for weak factors through random matrix theory, and a similar model to the one used in this paper has been considered in the large body of literature on sparse PCA (e.g. Paul and Johnstone 2012, Cai et al. 2013). These papers typically build on stronger assumptions on the error terms, and remain largely agnostic about the factors themselves. By explicitly modeling the factors, we are able to impose less restrictive assumptions on the error structure. For a related literature on sparse factor models under a Bayesian framework see Carvalho et al. (2008), Gao et al. (2013) and Pati et al. (2014). Finally, by considering a continuum of factor strengths, this paper is similar in spirit to the extensive literature in econometrics on identification with varying convergence rates (e.g. Andrews and Cheng 2012).

A number of empirical studies have postulated a structure with group-specific factors, sometimes called hierarchical factor models (e.g. Boivin and Ng 2006, Moench et al. 2013, Dias et al.

[1] Strictly speaking, partial sums of the estimated factor loadings. Using the standard PCA estimator, this is equivalent to the eigenvectors up to a rescaling.

2013), that is naturally related to the setting in this paper. Ando and Bai (2017) and Han (2017) also consider group-specific (or regional) factors, but require all group sizes to be comparable to the overall cross-sectional dimension, in which case all factors are strong asymptotically. Han and Caner (2017) consider a model in which some factors may be less pervasive but treat all local factors as noise. Wang (2008) and Choi et al. (2018) allow for local factors in the sense of this paper, but require the group structure and factor strengths to be known to the practitioner a priori. Using a similar drifting sequence framework as this paper, Uematsu and Yamagata (2020) focus on alternatives to the PCA estimator, while Bailey et al. (2016) derive a measure of cross-sectional dependence in the data.

Alternative approaches to estimate the number of factors in factor models include Bai and Ng (2019), who introduce a rank restriction on the estimates, and the approximate minimum rank problem defined in Ten Berge and Kiers (1991), which effectively attempts to find the smallest number of factors r, such that the unexplained common variance is smaller than some threshold δ. Gagliardini et al. (2019) propose a diagnostic criterion to determine the number of factors in an economy with a continuum of assets, allowing for both missing observations and n much larger than T. Neither considers the drifting sequence framework with local factors of this paper.

Before we formally introduce the model, the following are some concrete examples of economic models to which this paper applies.

Example 1. Arbitrage Pricing Theory.
Consider an unobserved common shock that affects only a subset of the population at the company level, for example, a new law that affects only large firms. As the number of firms, n, increases, one reasonable assumption is that the number of large firms increases at a rate slower than n (Chudik et al. (2011)). Unlike traditional factor models, the framework of this paper allows for this and is in line with the empirical finding indicating that the largest eigenvalues of the sample covariance matrix of asset returns diverge at differing rates (e.g. Trzcinka 1986) .

The empirical evidence on whether weaker factors are priced is somewhat mixed (e.g. Shukla and Trzcinka 1990). In Section 3.2, we use the results in Green and Hollifield (1992) adapted to our framework to derive theoretical bounds on the strength of factors that will be priced. We find that expected returns of individual assets are determined by those factors affecting proportionally more than $\sqrt{n}$ assets.

Example 2. The origins of aggregate fluctuations.
There is an ongoing debate about the origins of fluctuations in the aggregate economy (see, e.g. Foerster et al. 2011). Long and Plosser (1983) suggest that sectoral shocks may account for GDP fluctuations. With a fixed number of sectors, these sectoral shocks affect a fixed proportion of firms and can be viewed as aggregate shocks themselves. In contrast, Horvath (1998) investigates conditions under which an economy with n sectors can have a volatility that does not decay according

to $\frac{1}{\sqrt{n}}$. By modeling sectoral shocks as local factors affecting the corresponding subset of firms, this can be mapped into the framework of this paper. We show in Section 3.2 that, in an economy with n firms, $\sqrt{n}$-convergence for the aggregate growth rate of the economy fails when there are sectoral shocks affecting proportionally more than $\sqrt{n}$ firms. We therefore find that aggregate fluctuations can be attributed to sectors proportionally larger than $\sqrt{n}$ firms.

Example 3. Macroeconomic forecasting.
In a widely cited paper Boivin and Ng (2006) investigate the properties of the principal component estimator in finite samples. Specifically, they document conditions under which adding more data can be undesirable for factor estimation.

Boivin and Ng (2006) use a Monte Carlo study to establish that the performance of the principal component estimator deteriorates as more series unrelated to the factors are added, effectively making factors local in the sense of this paper. In this paper, by tying the convergence rate of a factor estimate to the factor's strength, we provide a theoretical confirmation of their findings.

2 A Model with Local Factors

To set up notation, denote the pth largest eigenvalue of a matrix A by $\psi_p(A)$ and the Frobenius norm of a matrix B by $\|B\|$, such that $\|B\|^2 = tr(B'B) = \sum_{ij} b_{ij}^2$. We make extensive use of the notion that certain quantities diverge at particular rates and write $a_n \asymp b_n$ for two sequences a_n, b_n if $a_n = O(b_n)$ and $b_n = O(a_n)$. We write $Y_n = \bar{O}_p(n^\gamma)$ as shorthand for $Y_n = O_p(min\{1, n^\gamma\})$. Finally let ι_p denote a vector with a 1 at entry p and zeros everywhere else, with the dimension varying, but obvious from context.

We consider the n-dimensional process X_t, $t = 1, 2, ..., T$. Let F_k, $k = 1, 2, ..., r$ denote the factors, and $\Lambda = [\lambda_{\cdot 1} \lambda_{\cdot 2} \cdots \lambda_{\cdot r}] = [\lambda_{1\cdot} \lambda_{2\cdot} \cdots \lambda_{n\cdot}]'$ denote the matrix of factor loadings. Throughout, we use the running indices s and t for the T observations, indices i, j for the n variables, and k and l for the r factors.

We assume that there exists a (not necessarily unique) finite r, such that the data has a static factor representation,

$$\underset{(T\times n)}{X^{(n)}} = \underset{(T\times r)}{F^{(n)}} \underset{(r\times n)}{\Lambda^{(n)'}} + \underset{(T\times n)}{e^{(n)}}, \tag{2}$$

and the following assumptions are satisfied.

Assumption 1. *There exist positive constants c,C and a diagonal matrix $D_r^{(n)}$ with diagonal entries $d_1^{(n)}, d_2^{(n)}, \ldots, d_r^{(n)}$, such that as $n, T \to \infty$:*

(a) $n/T \to c$

(b) $\Lambda^{(n)\prime}\Lambda^{(n)} = D_r^{(n)},\ d_1^{(n)} > d_2^{(n)} > \ldots > d_r^{(n)}$ *and* $|\lambda_{ik}| < C\ \forall i$

(c) $\frac{1}{T}F^{(n)\prime}F^{(n)} = I_r$ *and* $|F_{tk}| < C\ \forall\ t.$

We treat both Λ and F as parameters of the distribution of X (as in, e.g. Onatski 2012). If F and Λ are random and independent of each other, all results should still hold provided $\mathbb{E}\,\|F_t\|^4$ and $\mathbb{E}\,\|\lambda_i\|^4$ are bounded.

Part (a) of Assumption 1 requires n and T to be comparable even asymptotically. This mainly simplifies exposition, is common in the literature (e.g. Onatski 2012, Ahn and Horenstein 2013), and plausible in many applications of factor models. Notably, we do not require the entries in $D^{(n)}$ to diverge proportionally to n, thus allowing for weaker factors. All entries in $D^{(n)}$ may diverge at different rates (or not diverge at all). Since they depend on the sample size, they should be thought of as drifting sequences. With that in mind, to simplify notation, we will omit the superscript (n) on matrices X, Λ, F, D and e in what follows.

We also maintain that the true factors and their loadings fulfill the orthogonality conditions (b)-(c). Intuitively, we think of the factors in (2) as "primitive" exogenous forces. Because these forces are primitive, they do not have common causes and it is natural to treat them as orthogonal (Bernanke 1986). With local factors, Assumption 1(b) will be approximately true if the groups of outcomes affected by various factors are sufficiently distinct. (For example, Assumption 1(b) will hold exactly if the loadings are block diagonal, also see the discussion in Bai and Ng (2013)). This is restrictive, but necessary to separately identify the factors (and their corresponding loadings). We show in Section 5 that in practice our proposed estimators are robust to violations of this assumption.

Assumption 2. *For each factor k, the entire set of indices $i = 1, 2, \ldots, n$ can be partitioned into a set of indices $\mathcal{A}_k$ with cardinality $|\mathcal{A}_k| \asymp n^{\alpha_k}$ for some $\alpha_k \in [0, 1]$ and its complement such that, as $n, T \to \infty$ for all k:*

(a) $\sum_{i\in\mathcal{A}_k} \lambda_{ik}^2 \asymp n^{\alpha_k}$

(b) $\sum_{i\notin\mathcal{A}_k} \lambda_{ik}^2 < C$ *for some* $C < \infty$.

Assumption 2 can be thought of as a sparsity assumption. It states that any given factor fulfills the conventional pervasiveness assumption only on an unknown subset of all outcomes ($\mathcal{A}_k$), while the remaining loadings are small in the sense that their squares are summable. For the local factors ($\alpha_k < 1$) the loadings of any given factor k are concentrated on an asymptotically vanishing fraction of variables. Assumption 2 implies that the diagonal elements in d_k from Assumption 1 diverge at rate α_k.

As a specific example, consider a cross section of n asset returns and an industry with a size proportional to $\sqrt{n}$ of the assets. Suppose there exists an industry-specific factor F_l that affects

only those assets: $\lambda_{il} = 1$ if $i \in \mathcal{A}_l$ and $\lambda_{il} = 0$ if $i \notin \mathcal{A}_l$. Then, $\sum_{i \in \mathcal{A}_l} \lambda_{il}^2 \asymp \sqrt{n}$ and $\sum_{i \notin \mathcal{A}_l} \lambda_{il}^2 = 0$, such that Assumption 2 holds with $\alpha_l = 1/2$.[2] The standard assumptions in the literature correspond to assuming $\alpha_k = 1$ for all factors, thus ruling out any such local factors.

Assumption 3. *There exist constants $c > 0$, $C < \infty$, and $d \in (0, 1]$, such that*

(a) $\mathbb{E}(e_{ti}) = 0$, $\mathbb{E}\,|e_{ti}|^4 \leq C$

(b) $\sum_{t=1}^{T} |\,\mathbb{E}\left(\frac{e_s' e_t}{n}\right)| \leq C \;\forall s \quad$ *and* $\quad \sum_{j=1}^{n} |\,\mathbb{E}\left(\frac{e_i' e_j}{T}\right)| \leq C \;\forall i$

(c) for every (t,s), $\mathbb{E}\,\|\frac{1}{\sqrt{n}}[e_s' e_t - \mathbb{E}(e_s' e_t)]\|^4 \leq C$

(d) $E\,\|\frac{1}{\sqrt{nT}} \sum_{s=1}^{T} F_s[e_s' e_t - \mathbb{E}(e_s' e_t)]\|^2 \leq C \;\forall t$

(e) $\psi_1\left(\frac{e'e}{T}\right) = O_p(1)$ *and* $P\left(\psi_{[dn]}\left(\frac{e'e}{T}\right) \geq c\right) = 1$ *for some* $d > 0$.

Assumption 4. *For any $k, l < r$:*

(a) $\frac{\lambda_{\cdot k}' e_t}{n^{\frac{1}{2}\alpha_k}} = O_p(1) \quad \forall t$

(b) $\frac{\lambda_{\cdot k}' e' F_l}{n^{\frac{1}{2}\alpha_k} T^{\frac{1}{2}}} = O_p(1)$.

Assumptions 3 and 4 concern the possibly correlated noise. Assumption 3 rules out that there is too much dependence in the error terms and is standard in the literature (Bai 2003, Bai and Ng 2006). More primitive conditions can be provided that imply part (e) (see Onatski 2015, Moon and Weidner 2017). The second part of Assumption 3 (e) ensures that at least a constant fraction of outcomes have non-zero idiosyncratic variances. Assumption 4 is weaker than one that requires a number of Central Limit Theorems to hold. With $\alpha_k = 1$ for $k = 1, ..., r$, it is implied by Assumptions F2 and F3 in Bai (2003).

2.1 Defining the number of relevant factors

In standard factor models, the number of common factors is readily defined. In a model with local factors, the definition of r is less clear. Since we allow for $\alpha_k \neq 1$ (and, in fact $\alpha_k = 0$), the number of factors is no longer uniquely defined.

For instance, suppose there exists a factor F_r with $\lambda_{1r}, \lambda_{2r} \neq 0$, and $\lambda_{ir} = 0$ otherwise. Thus, the factor F_r affects only two outcomes ($i = 1, 2$). Under such a DGP, a model with r factors, and a model with $(r-1)$ factors combined with some additional correlation between the error

[2] To explicitly write this down in terms of drifting sequences, denote the fraction of outcomes affected by factor F_l by $\pi_l^{(n)}$. Then, clearly $\pi_l^{(n)} = \frac{|\mathcal{A}_l|}{n} \asymp \frac{1}{\sqrt{n}}$.

terms e_1 and e_2, will both satisfy Assumptions 1-4. We therefore treat r as an upper bound on the number of "relevant" factors throughout this paper such that Assumptions 1-4 are satisfied, with the understanding that some of the r factors may be too weak to be relevant in practice.

To make this precise, let $r_1 + r_2 = r$, where $\alpha_k > \tau$ for $k = 1, \ldots, r_1$ and $\alpha_k \leq \tau$ for $k = r_1 + 1, \ldots, r$ for some fixed value of $\tau \in [0, 1)$. Thus, r_1 is the number of factors affecting proportionally more than n^τ variables, while any additional factors are less pervasive. We can rewrite the factor structure (2) as

$$\begin{aligned} \underset{(T\times n)}{X} &= \underset{(T\times r)}{F}\underset{(r\times n)}{\Lambda'} + \underset{(T\times n)}{e} \\ &= \underset{(T\times r_1)}{F^s}\underset{(r_1\times n)}{\Lambda^{s'}} + \underset{(T\times r_2)}{F^w}\underset{(r_2\times n)}{\Lambda^{w'}} + \underset{(T\times n)}{e} \\ &= \underset{(T\times r_1)}{F^s}\underset{(r_1\times n)}{\Lambda^{s'}} + \underset{(T\times n)}{\tilde{e}}, \end{aligned}$$

where the weakest r_2 factors are incorporated into the error term $\tilde{e}$. This is effectively a factor model with r_1 factors, where $\psi_1(\tilde{e}\tilde{e}'/n)$ may no longer be bounded.

This brings up the important question of which factors are relevant. We answer this question in Sections 3.1-3.2 based on both economic and statistical arguments. This will inform the choice of τ and define r_1, the number of relevant factors, which will be the target of estimation in Section 4.

3 Weak Asymptotics

We first show what the introduction of local factors implies for the empirical distribution of the eigenvalues of the matrix $\frac{X'X}{T}$. This is the quantity depicted in Figure 1 and often included in applications to justify the use of a factor model.

Lemma 1. *Under Assumptions 1 and 2:*

$$\psi_k\left(\frac{\Lambda F'F\Lambda'}{T}\right)\begin{cases} \asymp n^{\alpha_k}, & k = 1, 2, \ldots, r \\ = 0 & k > r. \end{cases}$$

Proof. If $k \leq r$:

$$\begin{aligned} \psi_k\left(\frac{\Lambda F'F\Lambda'}{T}\right) &= \psi_k(\Lambda\Lambda') = \psi_k(\Lambda'\Lambda) \\ &= \sum_{i=1}^{n}\lambda_{ik}^2 = \sum_{i\in\mathcal{A}_k}\lambda_{ik}^2 + \sum_{i\notin\mathcal{A}_k}\lambda_{ik}^2 \asymp n^{\alpha_k} + O(1) \\ &\asymp n^{\alpha_k}, \end{aligned}$$

where the equality in the second line follows from Assumption 2.

If $k > r$: the result immediately follows from the fact that $rank(\Lambda F' F \Lambda') = r$. □

The properties of the eigenvalues of the matrix $\frac{X'X}{T}$ then follow:

Theorem 1. *For any given factor k* $(k = 1, 2, \ldots, r)$*, under Assumptions 1-3:*

$$\psi_k\left(\frac{X'X}{T}\right) \begin{cases} \asymp n^{\alpha_k} \textit{ for } k = 1, 2...r \\ = O_p(1) \textit{ for } k = r+1, \ldots, n. \end{cases}$$

Proof. By the singular value version of Weyl's inequalities (Horn and Johnson 2012):

$$\sigma_{k+l-1}(A+B) \leq \sigma_k(A) + \sigma_l(B) \quad 1 \leq k, l \leq \min(n, T),\ k + l \leq \min(n, T),$$

where $\sigma_k(A)$ denotes the kth largest singular value of a matrix A. Therefore, for $k = 1, 2, \ldots, r_{max}$:

$$\sigma_k(X) \leq \sigma_k(F\Lambda') + \sigma_1(e).$$

Since $\sigma_k(A) = \sqrt{\psi_k(AA')}$ for any matrix A, it follows that

$$\sqrt{\psi_k(XX')} \leq \sqrt{\psi_k(F\Lambda'\Lambda F')} + \sqrt{\psi_1(ee')}.$$

And we therefore conclude, using Lemma 1 and Assumption 3(e) respectively for the two eigenvalues on the RHS:

$$\begin{aligned} \psi_k\left(\frac{XX'}{T}\right) &\leq \psi_k\left(\frac{F\Lambda'\Lambda F'}{T}\right) + \psi_1\left(\frac{ee'}{T}\right) + 2\sqrt{\psi_k\left(\frac{F\Lambda'\Lambda F'}{T}\right)}\sqrt{\psi_1\left(\frac{ee'}{T}\right)} \\ &\leq C_1 n^{\alpha_k} + O_p(1) + O_p(n^{\frac{1}{2}\alpha_k}) \leq C_2 n^{\alpha_k}. \end{aligned}$$

Similarly, again by Weyl's inequalities:

$$\begin{aligned} & \sigma_k(X - e) \leq \sigma_k(X) + \sigma_1(e) \\ \Rightarrow \quad & \sigma_k(F\Lambda') \leq \sigma_k(X) + \sigma_1(e) \\ \Rightarrow \quad & \sqrt{\psi_k\left(\frac{F\Lambda'\Lambda F'}{T}\right)} \leq \sqrt{\psi_k\left(\frac{XX'}{T}\right)} + \sqrt{\psi_1\left(\frac{ee'}{T}\right)} \\ \Rightarrow \quad & \sqrt{\psi_k\left(\frac{XX'}{T}\right)} \geq \sqrt{c_1 n^{\alpha_k}} - O_p(1) \end{aligned}$$

and we therefore also conclude that $\psi_k(\frac{XX'}{T}) \geq c_2 n^{\alpha_k}$. □

Under a scenario with r strong factors ($\alpha_k = 1$ for all $k = 1, 2, \ldots, r$), this reduces to the standard result in the literature: the first r eigenvalues diverge at rate n (Connor and Korajczyk 1993, Bai and Ng 2002, Hallin and Liska 2007). We extend this result to allow for weaker factors with the slower divergence rates of Theorem 1 for factors that affect only a subset of the observed variables.[3]

Theorem 1 provides a possible explanation for the continuum of eigenvalues often observed, as in Figure 1. While conventional factor models imply a large gap in the eigenvalue distribution, the eigenvalues corresponding to local factors will fall into this gap.

Recall the earlier distinction of factors into two groups: $F = [F_1, \ldots, F_{r_1}, F_{r_1+1}, \ldots, F_r] = [F^s, F^w]$, such that $r = r_1 + r_2$, $\alpha_k > \tau$ for $k = 1, 2.., r_1$ and $\alpha_k \leq \tau$ for $k = r_1 + 1, \ldots, r$ for some user specified threshold $\tau \in [0, 1]$. To provide guidance on how to choose τ (which factors are relevant), we next consider the following two questions:

1. When is a factor pervasive enough to be estimated consistently?

2. When is a factor pervasive enough to be of interest in some common economic models?

3.1 The Principal Component Estimator

We start with the first question and consider the standard estimator in the literature: estimation of both the factors and their loadings is achieved through the principal component estimator (see Stock and Watson 2002, Bai and Ng 2002, Bai 2003). Specifically, denoting the singular value decomposition of $\frac{1}{\sqrt{T}}X$ by $U\Sigma V'$, we obtain estimates $\hat{F}$ for K factors as $\hat{F} = \sqrt{T}U_{[1:K]}$, and $\hat{\Lambda} = V_{[1:K]}\Sigma_{[K,K]}$, where $M_{[1:K]}$ denotes the leading K columns and $M_{[K,K]}$ the $K \times K$ matrix consisting of the K leading columns and rows of a matrix M.

Define a $(r \times K)$ matrix $H = \Lambda'\Lambda\frac{F'\hat{F}}{T}\hat{D}_K^{-1}$, where $\hat{D}_K$ is a diagonal matrix with the K largest eigenvalues of $\frac{X'X}{T}$ on the main diagonal, and let $H_{.k}$ denote the kth column of H. We obtain the following Lemma.

Lemma 2. *Under Assumptions 1-4, with $\hat{F}$ and H defined as above:*

$$\hat{F}_{tk} - H'_{.k}F_t = O_p(n^{1-2\alpha_k}) + O_p(n^{\frac{1}{2}\alpha_1 - \alpha_k}).$$

The proof for Lemma 2 can be found in Online Appendix B.1). We achieve convergence as long as $\alpha_k > 1/2$. Further, in the case of r strong factors ($\alpha_r = \alpha_1 = 1$) and with $K = r$, the rates

[3]Theorem 1 still holds if we replace Assumption 2 with the high level assumption $\sum_{i=1}^{n} \lambda_{ik}^2 \asymp n^{\alpha_k}$. The result in Theorem 1 therefore extends to weak factors in general and does not need the sparsity pattern that is imposed by Assumption 2.

above reduce to

$$\hat{F}_t - H'F_t = O_p(\frac{1}{\sqrt{n}}).$$

This is in line with the literature (Bai 2003). Next, the orthogonality assumptions 1(b)-(c) imply the following for the rotation matrix H.

Lemma 3. *Under Assumptions 1-4,* $H_{.k} = \iota_k + O_p(n^{\frac{1}{4}\alpha_1 - \frac{1}{2}\alpha_k}) + O_p(n^{\frac{1}{2}-\alpha_k})$.

The proof for Lemma 3 can be found in Online Appendix B.1.[4] In the case of r strong factors ($\alpha_r = \alpha_1 = 1$), a similar result has been derived in Bai and Ng (2013). Combining Lemmata 2 and 3, we thus obtain the following:

Theorem 2. *Under Assumptions 1-4,*

$$\hat{F}_{tk} - F_{tk} = O_p(n^{1-2\alpha_k}) + O_p(n^{\frac{1}{4}\alpha_1 - \frac{1}{2}\alpha_k}) + O_p(n^{\frac{1}{2}-\alpha_k}).$$

Proof. By Lemma 3: $H_{.k} = \iota_k + O_p(n^{\frac{1}{4}\alpha_1 - \frac{1}{2}\alpha_k}) + O_p(n^{\frac{1}{2}-\alpha_k})$. Consequently, combining this with Lemma 2:

$$\begin{aligned}
\hat{F}_{tk} - F_{tk} &= (\hat{F}_{tk} - H'_{k.}F_t) + (H'_{k.} - \iota'_k)F_t \\
&= O_p(n^{1-2\alpha_k}) + O_p(n^{\frac{1}{2}\alpha_1 - \alpha_k}) + O_p(n^{\frac{1}{4}\alpha_1 - \frac{1}{2}\alpha_k}) + O_p(n^{\frac{1}{2}-\alpha_k}) \\
&= O_p(n^{\frac{1}{4}\alpha_1 - \frac{1}{2}\alpha_k}) + O_p(n^{\frac{1}{2}-\alpha_k}) + O_p(n^{1-2\alpha_k}).
\end{aligned}$$

□

This result may be of interest to a practitioner for two reasons. First, while PCA as a dimensionality reduction technique is always feasible, Theorem 2 establishes a lower bound in terms of factor strength for which we are able to prove consistency of the principal component estimator ($\alpha_k > \frac{1}{2}$) under a factor model. Further, even for factors that are estimated consistently, it suggests that the estimation of a factor becomes worse as its strength decreases (as documented in simulations in Boivin and Ng 2006). The intuition is clear: as fewer cross sections carry a signal about F_k, the precision of its estimate decreases. However, the fact that weaker factors tend to be estimated with less precision seems to be largely unaccounted for in the current literature. For cases in which factor estimates are used that correspond to weaker factors, Theorem 2 at least suggests being cautious with respect to the standard errors of these estimates.

We also note that a similar result holds for the factor loadings, which will be important for our results in Section 4:

[4]Here, and throughout the paper, we assume that the sign indeterminacy of $\hat{F}$ is resolved by adding the normalization that the diagonal elements of $\hat{F}'F$ are nonnegative.

Theorem 3. *Under Assumptions 1-4:*

$$\hat{\lambda}_{ik} - \lambda_{ik} = \bar{O}_p(n^{\frac{1}{4}\alpha_1 - \frac{1}{2}\alpha_k}) + \bar{O}_p(n^{\frac{1}{2}-\alpha_k}).$$

Proof. By definition, with $\frac{1}{\sqrt{T}}X = U\Sigma V'$,

$$\hat{\Lambda}' = \Sigma_{[K,K]}V'_{[1:K]} = U'_{[1:K]}U_{[1:K]}\Sigma_{[K,K]}V'_{[1:K]} = U'_{[1:K]}U\Sigma V' = \frac{1}{\sqrt{T}}U'_{[1:K]}X = \frac{\hat{F}'X}{T}.$$

Thus,

$$\begin{aligned}
\hat{\lambda}_{ik} &= \frac{\hat{F}'_k X_i}{T} = \frac{1}{T}\hat{F}'_k F\lambda_i + \frac{1}{T}\hat{F}'_k e_i \\
&= \lambda_{ik} + (\frac{\hat{F}'_k F}{T} - \iota'_k)\lambda_i + \frac{1}{T}(\hat{F}_k - F_k)'e_i + \frac{1}{T}F'e_i \\
&= \lambda_{ik} + (\frac{\hat{F}'_k F}{T} - \iota'_k)\lambda_i + \frac{1}{T}(\hat{F}_k - FH_{\cdot k})'e_i + \frac{1}{T}(H_{\cdot k} - \iota_k)'F'e_i + \frac{1}{T}F'e_i \\
&= \lambda_{ik} + O_p(n^{\frac{1}{4}\alpha_1 - \frac{1}{2}\alpha_k}) + O_p(n^{\frac{1}{2}-\alpha_k}) + O_p(n^{1-2\alpha_k}),
\end{aligned}$$

where the last equality follows from Lemma 8 and Lemmata 3 and 10 in the Online Appendix, as well as Assumption 4(b). Since also $\frac{\hat{F}'_k X_i}{T} \leq \frac{1}{T}\|\hat{F}\|\|X_i\| = O_p(1)$, this completes the proof. □

Thus, we obtain convergence of the principal component estimator for both the factors and the loadings as long as $\alpha_k > 1/2$. The following table provides an empirical test to assess the adequacy of these asymptotic results in approximating finite sample properties. Two factors were created, one strong ($\alpha_1 = 1$) and one weak ($\alpha_2 < 1$). The strength of the weak factor is varied with $\alpha_2 \in [0.25, 0.5, 0.75]$, and the correlation of the estimated factor $\hat{F}_2$ with its true counterpart is depicted in Table 1. This correlation can be thought of as a measure of consistency of the principal component estimator for the kth factor. The exact DGP can be found in Online Appendix A.1. The numbers in Table 1 are in line with our theoretical findings: Theorem 2 suggests that the correlation between F_2 and $\hat{F}_2$ will approach 1 when $\alpha_2 > .5$. This corresponds to the third row in Table 1. Additionally, we observe that the estimation error seems to dominate the signal when the factor strength is below this threshold.

3.2 Which Factors Matter?

Aside from practical issues in estimating factors that affect only a subset of the cross sections, it is also not clear which factors are of interest to a practitioner in the first place.

We next present two simplified economic models to illustrate that, in a model with local factors,

n	81	256	625	1296
$\lvert\mathcal{A}_2\rvert = n^{1/4}$	0.067	0.050	0.041	0.040
$\lvert\mathcal{A}_2\rvert = n^{1/2}$	0.109	0.099	0.089	0.079
$\lvert\mathcal{A}_2\rvert = n^{3/4}$	0.754	0.951	0.976	0.984

Table 1: Average correlation between estimate $\hat{F}_2$ and truth F_2 for differing factor strengths of local factor as sample size increases. Data has one global factor F_1 that affects all outcomes. Table based on 1000 repetitions. Detailed description of DGP in Online Appendix A.1.

the economically relevant factors are often also those that affect proportionally more than $\sqrt{n}$ of the cross sectional units ($\tau = 1/2$). The first model builds on the Arbitrage Pricing Theory (APT) of Ross (1976) and the second considers aggregate fluctuations in the economy.

3.2.1 Arbitrage Pricing Theory

Consider a vector of asset returns R_t, with the index j denoting an individual asset. We assume that for a large enough number of factors r the excess returns $R_t - \mathbb{E}(R_t)$ follow a factor structure that obeys Assumptions 1-4 with

$$R_i - \mathbb{E}(R_i) = \lambda_i F + e_i. \tag{3}$$

Notably, we continue to allow for local factors with $\alpha_k < 1$, in line with the empirical finding indicating that the largest eigenvalues of the sample covariance matrix of asset returns diverge at differing rates (e.g. Trzcinka 1986).

In a model with local factors, it is natural to ask which of these factors are strong enough to be priced. We are therefore interested in the number of factors APT dictates is needed such that the mean returns $E(R_i)$ are in the span of the factor loadings of those factors. Intuitively, this boils down to the question of which factor risk is diversifiable. Exposure to factors that affect so few outcomes (have a small enough α_k) that well-diversified portfolios have negligible variance stemming from those weaker factors should not be rewarded.

Assume that all efficient portfolios are well diversified.[5] Then, we show formally in the Online Appendix that APT pricing holds with respect to the factors affecting proportionally more than $\sqrt{n}$ of the assets (those factors with $\alpha_k > 1/2$). In other words, the mean returns are in the span of the factor loadings of the K factors with $\alpha_k > 1/2$.

Instead of ruling out the existence of non-pervasive factors with $\alpha_k < 1$, our model allows for the existence of local factors that affect only a small subset of assets (e.g. those in a particular

[5]Denoting portfolio weights by w_i, with $\sum_i^n w_i = 1$, we formalize the term "well-diversified" by imposing a bound on the sup-norm of the weights: $|w_i| \leq o(\frac{1}{\sqrt{n}})\ \forall i$.

industry). In such a model, Arbitrage Pricing Theory dictates that the relevant factors determining the mean returns of the assets are those with $\alpha_k > 1/2$. Weaker, less pervasive factors with $\alpha_k \leq 1/2$ should not be priced. A more detailed discussion with a formal Proposition following the arguments in Green and Hollifield (1992) can be found in Online Appendix B.3.

3.2.2 Aggregate Fluctuations in the Economy

There is an ongoing debate about the origins of fluctuations in the aggregate economy (e.g. Foerster et al. 2011), since economy-wide shocks are often insufficient in explaining aggregate fluctuations (Cochrane 1994). Gabaix (2011) provides conditions under which aggregate fluctuations can arise in a model with heterogeneous firm sizes. Acemoglu et al. (2012) provides conditions under which aggregate fluctuations can arise in a model with production networks with heterogeneity in centrality. Here, we are interested in the conditions under which aggregate fluctuations can arise in a model with sectoral shocks.

Consider a simple "Islands" economy with n firms. Firm i produces a quantity S_{it} of the consumption good and firm-level growth rates have a factor structure for large enough r that obeys Assumptions 1-4 with

$$\frac{\Delta S_{i,t+1}}{S_{it}} = \frac{S_{i,t+1} - S_{it}}{S_{it}} = \lambda_i F_{t+1} + \varepsilon_{i,t+1}. \tag{4}$$

Intuitively, a firm's growth rate is therefore a combination of shocks that may affect several firms (e.g. economy-wide or industry shocks) and idiosyncratic shocks. However, we do not impose the factors to be pervasive and likely $\lambda_{ik} = 0$ for most firm-factor combinations.

In this stylized model, GDP growth is given by

$$\frac{\Delta Y_{t+1}}{Y_t} = \frac{1}{Y_t}\sum_{i=1}^{n} \Delta S_{i,t+1} = \sum_{i=1}^{n} \frac{S_{it}}{Y_t}\lambda_i F_{t+1} + \sum_{i=1}^{n} \frac{S_{it}}{Y_t}\varepsilon_{i,t+1}.$$

With finite sized firms it is easy to show that the variance of GDP growth at time $(t+1)$ conditional on time t equals

$$Var_t\left(\frac{\Delta Y_{t+1}}{Y_t}\right) \asymp \sum_{k=1}^{r} n^{2\alpha_k - 2} + O_p(\frac{1}{n}). \tag{5}$$

By modeling sectoral shocks as local factors affecting the corresponding subset of firms, it immediately follows from (5) that $\sqrt{n}$-convergence for the aggregate growth rate of the economy fails if there are sectoral shocks affecting proportionally more than $\sqrt{n}$ firms. In other words, aggregate fluctuations can be attributed to sectors proportionally larger than $\sqrt{n}$ firms.

This is in line with the granularity conditions derived in Gabaix (2011), who considers an economy with heterogeneous firm sizes. Proposition 2 in Gabaix (2011) establishes that $\sigma_{GDP} \asymp \frac{1}{\sqrt{n}}$ only if the largest firm has a relative weight of at most $W_n = O(\frac{1}{\sqrt{n}})$. This corresponds to the limit on sector size stated above. This is also in line with the results in Acemoglu et al. (2012). There, an intersectoral network plays the same role as the firm size distribution in Gabaix (2011) and the distribution of sectoral sizes in our paper – shocks that take more central positions in the network have a disproportionate effect on aggregate output.

In order to understand the origins of aggregate fluctuations, the relevant sectors are therefore those that consist of more than $\sqrt{n}$ firms ($\alpha_k > 1/2$). A more detailed discussion of this model can be found in Online Appendix B.4.

4 Determining the Number of Relevant Factors

The results in the previous section suggest that the relevant factors will usually be those with $\alpha_k > \tau = 1/2$. We next consider estimation of this number of relevant factors.

Existing estimators for the number of factors are usually derived from the distribution of eigenvalues of the matrix $\frac{X'X}{T}$ (or equivalently the singular values of $\frac{1}{\sqrt{T}}X$), and can be interpreted as formalizing the heuristic approach based on a visual inspection of the scree plot, which dates back to Cattell (1966). Crucially, these estimators require separability between the two groups of eigenvalues (large eigenvalues representing factor-related variation and small eigenvalues representing idiosyncratic variation).

Theorem 1 established that, in the presence of local factors, the eigenvalues can no longer be easily separated into two groups. This is illustrated in Figure 2a, which depicts the theoretical rate of divergence for the kth eigenvalue if the corresponding factor has a pervasiveness α_k. With only a fixed number of outcomes affected by F_k ($\alpha_k = 0$), the corresponding eigenvalue remains bounded. As the factor affects more covariates, the eigenvalue begins to diverge at an increasing rate.

Rather than relying only on the eigenvalues, we propose that there is additional information in the eigenvectors of $\frac{X'X}{T}$. We show that these eigenvectors can be used to "tilt the eigenvalues" and thereby reintroduce a gap between those corresponding to factors with $\alpha_k > \tau$ and those below this threshold.

For intuition, consider the singular value decomposition of $\frac{1}{\sqrt{T}}X = U\Sigma V$ again. The right-singular vectors of $\frac{1}{\sqrt{T}}X$, V, are the eigenvectors of $\frac{X'X}{T}$. Denoting the first eigenvector (right-singular vector) V_1 by v, by construction $\sum_{i=1}^{n} v_i^2 = 1$. Now consider the partial sum $\sum_{i=1}^{n/2} v_i^2$. If the eigenvector is concentrated on its first $n/2$ entries, this sum is still equal to one. On the other hand, if the entries in the eigenvector are evenly spread out, $\sum_{i=1}^{n/2} v_i^2 \approx 0.5$.

Now consider a factor model with a single factor. The principal component estimator is constructed as $\hat{\lambda} = \sigma_1 v$, with σ_1 denoting the largest singular value, and is therefore just a multiple of the eigenvector v. Thus, $\sum_i^n \hat{\lambda}_i^2 = \sigma_1^2$. Further, if all estimated loadings $\hat{\lambda}_i$ are concentrated on a small subset $\mathcal{A}$, $\sum_{i\in\mathcal{A}} \hat{\lambda}_i^2 \approx \sigma_1^2$. On the other hand, if all n loadings are comparable in size, $\sum_{i\in\mathcal{A}} \hat{\lambda}_i^2 << \sigma_1^2$. Intuitively, measuring how concentrated the estimated loadings are on a subset of outcomes therefore carries additional information for the pervasiveness of the corresponding factor.

To derive this result formally, we introduce the following quantity:

$$\hat{\Upsilon}_{zk}^u \equiv \psi_k\left(\frac{X'X}{T}\right)\hat{S}_{zk}^u \equiv \psi_k\left(\frac{X'X}{T}\right)\left(\frac{1}{z}\sum_i^z \frac{\hat{\lambda}_{ik}^2}{\sqrt{\frac{1}{n}\sum_{i=1}^n \hat{\lambda}_{ik}^2}}\right)^u, \tag{6}$$

where, with some abuse of notation, the estimated squared loadings $\hat{\lambda}_{ik}^2$ are sorted in decreasing order, such that we take a partial sum over the z largest estimated loadings of the kth loading vector in the second part.

With $u = 0$, the second part vanishes and $\hat{\Upsilon}_{zk}^0 = \psi_k(\frac{X'X}{T})$, such that $\hat{\Upsilon}_{zk}^0$ is simply equal to the kth eigenvalue of $\frac{X'X}{T}$. On the other hand, for $u > 0$, $\hat{S}_{zk}^u \neq 1$, and the second part enters. Intuitively $\hat{S}_{zk}$ measures how concentrated the corresponding eigenvector is on its z largest entries. The power u governs the relative weight on the eigenvalue versus the eigenvector. The behavior of $\hat{\Upsilon}_{zk}^u$ is summarized in the following theorem.

Theorem 4. *Under Assumptions 1-4, choose a threshold $z = n^\tau g(n), \tau \in [0,1]$, such that (i) $g(n) \to \infty$ and (ii) $g(n)/n^\epsilon \to 0$ for any $\epsilon > 0$ as $n \to \infty$. Then, for any given factor $k \leq r_{max}$, with $u \in [0,2]$:*

(a) If $\alpha_k > \tau$:

$$\hat{\Upsilon}_{zk}^u \asymp n^{\frac{1}{2}u + (1-\frac{1}{2}u)\alpha_k} \tag{7}$$

(b) If $\max\{\frac{1+\tau}{3}, \frac{\alpha_1+4\tau}{6}\} < \alpha_k \leq \tau$:

$$\hat{\Upsilon}_{zk}^u \asymp n^{(1+\frac{1}{2}u)\alpha_k + (\frac{1}{2}-\tau)u} g(n)^{-u} \tag{8}$$

(c) If $0 < \alpha_k \leq \max\{\frac{1+\tau}{3}, \frac{\alpha_1+4\tau}{6}\}$:

$$\hat{\Upsilon}_{zk}^u \asymp n^{\alpha_k} \qquad \text{for } u = 0 \tag{9}$$

$$\hat{\Upsilon}_{zk}^u = O_p(n^{(1+\frac{1}{2}u)\alpha_k + (\frac{1}{2}-\tau)u} g(n)^{-u}) \qquad \text{for } u > 0 \tag{10}$$

(d) If $\alpha_k = 0$ *or* $k > r$*:*

$$\hat{\Upsilon}^u_{zk} = O_p(n^{(\frac{1}{2}-\tau)u} g(n)^{-u}). \tag{11}$$

The proof of Theorem 4 can be found in Online Appendix B.2.

As $\hat{\Upsilon}^u_{zk}$ is the key quantity in what follows, we also present a corollary that simplifies the notation and covers most cases before further discussion. We argued in Section 3 that the relevant factors are usually those that affect proportionally more than $\sqrt{n}$ of the outcomes. We will thus omit the subscript z and simply write $\hat{\Upsilon}^u_{zk} = \hat{\Upsilon}^u_k$ when $z = \sqrt{n}g(n)$, corresponding to $\tau = 1/2$, to obtain the following corollary:

Corollary 1. *Let* $z = \sqrt{n}g(n)$*, such that (i)* $g(n) \to \infty$ *and (ii)* $g(n)/n^\epsilon \to 0$ *for any* $\epsilon > 0$ *as* $n \to \infty$*. Then, under Assumptions 1-4, for any given factor* $k \leq r_{max}$ *and with* $u \in [0, 2]$*:*

(a) If $\alpha_k > \frac{1}{2}$*:*

$$\hat{\Upsilon}^u_k \asymp n^{\frac{1}{2}u + (1-\frac{1}{2}u)\alpha_k}$$

(b) If $\alpha_k \leq \frac{1}{2}$*:*

$$\begin{aligned} &\hat{\Upsilon}^u_k \asymp n^{\alpha_k} && \text{for } u = 0 \\ &\hat{\Upsilon}^u_k = O_p\left(n^{(1+\frac{1}{2}u)\alpha_k} g(n)^{-u}\right) && \text{for } u > 0 \end{aligned}$$

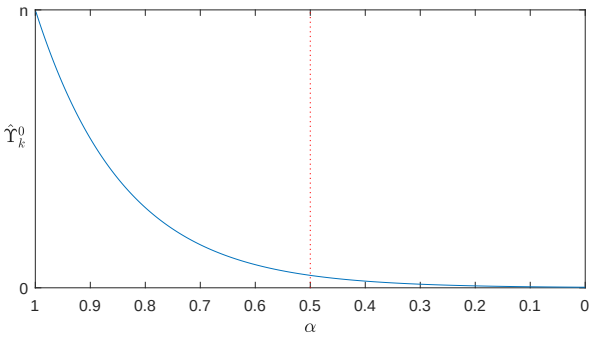
(c) If $\alpha_k = 0$ *or* $k > r$*:*

$$\hat{\Upsilon}^u_k = O_p(g(n)^{-u}).$$

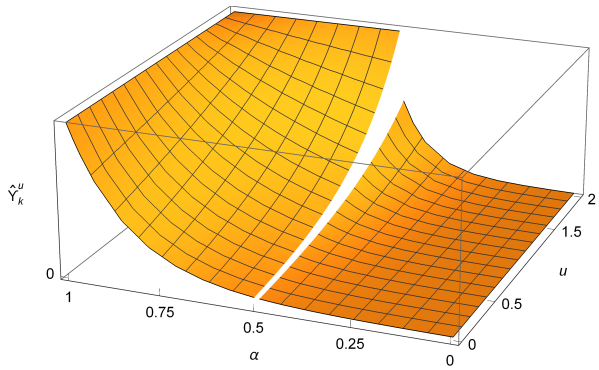
The theoretical rates of Corollary 1 are illustrated graphically in Figure 2b, which depicts the behavior of $\hat{\Upsilon}^u_k$ as a function of factor pervasiveness α_k and the parameter u.

Moving from right to left in Figure 2b, the number of covariates affected by the factor increases, with n^{α_k} outcomes influenced by the factor. On the right, only a fixed number of covariates are affected ($\alpha_k = 0$), while the left of the figure corresponds to a scenario in which the factor is relevant for all covariates ($\alpha_k = 1$). With $u = 0$, the front of Figure 2b is identical to Figure 2a (because $\hat{\Upsilon}^0_k = \psi_k(\frac{X'X}{T})$). Moving from front to back, u increases from 0 to 2 (thereby increasing the importance of the eigenvector).

The key insight is that setting $u > 0$ induces a steep region in the statistic $\hat{\Upsilon}^u_k$ around $\alpha_k = 1/2$, thereby helping to discriminate between factors above and below this threshold. While the flat



(a) Theoretical divergence rate of $\hat{\Upsilon}^0_{zk} = \psi_k(\frac{X'X}{T})$ as a function of factor strength (α_k).



(b) Theoretical divergence rate of $\hat{\Upsilon}^u_{zk}$ as a function of both factor strength (α_k) and tuning parameter u, which governs the relative weight of eigenvalue and eigenvector, for $z = \sqrt{n}\sqrt{loglog(n)}$

Figure 2: Illustration of theoretical divergence rates for key quantities. Figure created with $n = 500$.

slope around $\alpha_k = .5$ in Figure 2a suggests that existing estimators, using the eigenvalues of $\frac{X'X}{T}$ ($\hat{\Upsilon}^0_k$), will have low discriminatory power in distinguishing factors above and below this threshold, Figure 2b suggests that an estimator based on $\hat{\Upsilon}^2_k$ will perform well in distinguishing relevant factors with $\alpha_k > 1/2$ from those below this threshold.

For example, let $\alpha_k \in (0.5, 1]$. Then, $\hat{\Upsilon}^0_k \asymp n^{\alpha_k}$, while $\hat{\Upsilon}^2_k \asymp n$. Thus, $\hat{\Upsilon}^2_k$ does not depend on the factor strength for $\alpha_k \in (.5, 1]$ and can distinguish local factors with $\alpha_k > 1/2$ from the weakly correlated noise at the same rate as strong factors with $\alpha_k = 1$.

On the other hand, if $\alpha_k = 1$, $\hat{\Upsilon}^u_k \asymp n$, regardless of the choice of u, and if $\alpha_k = 0$ (or $k > r$), $\hat{\Upsilon}^0_k = O_p(1)$ and $\hat{\Upsilon}^u_k = O_p(g(n)^{-u})$ if $u > 0$. Thus, with only strong factors, the divergence rate of $\hat{\Upsilon}^u_k$ is invariant to the choice of u (up to the very slowly diverging sequence $g(n)$).[6]

4.1 Proposed Estimators for the Number of Relevant Factors

Analogous to existing estimators in the literature, but based on $\hat{\Upsilon}^2_k$ instead of $\hat{\Upsilon}^0_k$, we consider the following estimators:

1. An information criteria-like threshold (cf. Bai and Ng 2002, Kapetanios 2004)

2. The difference between two subsequent values (cf. Onatski 2010, Kapetanios 2010)

3. The ratio of two subsequent values (cf. Ahn and Horenstein 2013)

[6]In practice, we suggest using $g(n) = 0.7\sqrt{loglog(n)}$, which is approximately equal to one for most relevant sample sizes, and used throughout our simulations unless noted otherwise.

4.1.1 Thresholding Estimators

We start by considering one of the estimators introduced in Bai and Ng (2002). We will denote by EC the number k that minimizes the criterion function

$$\Gamma_{BN}(k) = V(k) + k\hat{\sigma}^2\left(\frac{n+T}{nT}\right)log\left(\frac{nT}{n+T}\right),$$

where

$$V(k) = \min_{\Lambda,F_k}(NT)^{-1}\sum_{i=1}^{n}\sum_{t=1}^{T}(X_{ti} - \lambda_i^k F_t^k)^2 = \frac{1}{nT}\sum_{j=k+1}^{n}\psi_j(X'X), \tag{12}$$

and $\hat{\sigma}^2$ is an estimator of the unconditional variance of the idiosyncratic error. The second equality in (12) follows from the fact that $V(k)$ is the best approximation of X of rank k. We can alternatively represent $\hat{\sigma}^2$ as $V(r_{max}) = \frac{1}{n}\sum_{j=r_{max}}^{n}\psi_j\left(\frac{X'X}{T}\right)$ for a fixed $r_{max} > r$. Unifying notation in terms of the eigenvalues and using $c = n/T$, we can rewrite their estimator as:

$$\begin{aligned} BN &= \underset{0\le k\le r_{max}}{\arg\min}\; V(k) + k\hat{\sigma}^2\left(\frac{n+T}{nT}\right)log\left(\frac{nT}{n+T}\right) \\ &= \underset{0\le k\le r_{max}}{\arg\min}\; \frac{1}{n}\sum_{l=k+1}^{n}\psi_l\left(\frac{X'X}{T}\right) + k\hat{\sigma}^2\left(\frac{c+1}{n}\right)log\left(\frac{n}{c+1}\right) \\ &= \max_{0\le k\le r_{max}} k \qquad \text{s.t. } \psi_k\left(\frac{X'X}{T}\right) > \hat{\sigma}^2(c+1)\log(\frac{n}{c+1}). \end{aligned} \tag{13}$$

Therefore, BN is equivalent to a thresholding on the eigenvalue distribution in our setting (also see the discussion in Bai and Ng (2019)).[7] Instead of deriving the estimator solely from the empirical distribution of the eigenvalues, we will consider the following estimator:

$$FC = \max_{0\le k\le r_{max}} k \qquad \text{s.t. } \hat{\Upsilon}_k^2 > \hat{\sigma}^2\frac{n}{h(n)}, \tag{14}$$

where the function $h(n)$ is such that (i) $h(n) \to \infty$ and (ii) $h(n)/g(n)^2 \to 0$ as $n \to \infty$, and $g(n)$ fulfills the conditions stated in Theorem 4. For example, $h(n) = g(n)$ is a valid choice.

Theorem 5. *Under Assumptions 1-4, FC is a consistent estimator for the number of factors r_1 such that $\alpha_k > 1/2$ for $k = 1, \dots, r_1$ and $\alpha_k \le 1/2$ for $k > r_1$.*

Proof. By Theorem 1 and Weyl's inequality, $\hat{\sigma}^2$ is bounded both above and below. Thus all rate conditions on $h(n)$ also hold for $\frac{1}{\hat{\sigma}^2}h(n)$. We first show that $\lim_{n\to\infty} P(\hat{\Upsilon}_k^2 > \hat{\sigma}^2\frac{n}{h(n)}) = 1$ for

[7] Throughout, we add a "mock eigenvalue" $\hat{\Upsilon}_0^0 = \hat{\Upsilon}_0^2 = \hat{\sigma}^2 n$ to allow for $r_1 = 0$. It is easy to show that our consistency results below are unaffected by its conclusion. Further, in line with Ahn and Horenstein (2013), our simulations suggest that our results are not sensitive to the exact choice of this mock eigenvalue.

$k = 1, \ldots, r_1$. In this case, $\alpha_k > 1/2$. By Theorem 4, $\hat{\Upsilon}_k^2 \asymp n$. Thus, $n = O_p(\hat{\Upsilon}_k^2)$. Combining this with $\frac{\hat{\sigma}^2}{h(n)} = o_p(1)$ we obtain $\hat{\sigma}^2 \frac{n}{h(n)} = o_p(\hat{\Upsilon}_k^2)$ and thus

$$lim_{n\to\infty} P\left(\left|\frac{\hat{\sigma}^2 \frac{n}{h(n)}}{\hat{\Upsilon}_k^2}\right| \geq \varepsilon\right) = 0$$

for any $\varepsilon > 0$. Letting $\varepsilon = 1$ and rearranging, it follows that

$$lim_{n\to\infty} P\left(\hat{\Upsilon}_k^2 \leq \hat{\sigma}^2 \frac{n}{h(n)}\right) = 0 \qquad \text{for } k = 1, \ldots, r_1.$$

Next, consider the case $\alpha_k \leq 1/2$. Then $\hat{\Upsilon}_k^2 = O_p(\frac{n^{2\alpha_k}}{g(n)^2})$ by Theorem 4. But $O_p(n^{2\alpha_k}/g(n)^2) = O_p(n/g(n)^2) = o_p(\hat{\sigma}^2 n/h(n))$ and thus for any $\varepsilon > 0$,

$$lim_{n\to\infty} P\left(\left|\frac{\hat{\Upsilon}_k^2}{\hat{\sigma}^2 \frac{n}{h(n)}}\right| \geq \varepsilon\right) = 0.$$

Letting $\varepsilon = 1$, this gives

$$lim_{n\to\infty} P\left(\hat{\Upsilon}_k^2 \geq \hat{\sigma}^2 \frac{n}{h(n)}\right) = 0 \qquad \text{for } k > r_1,$$

completing the proof. □

While FC is therefore a consistent estimate for r_1, we next derive the properties of the existing estimator BN in a setting with local factors. The implicit target of estimation using the BN criterion of Bai and Ng (2002) will be different than the cutoff argued for in this paper.

Corollary 2. *Under Assumptions 1-4 BN is a consistent estimator for the number of factors r^* such that $\alpha_k > 0$ for $k = 1, \ldots, r^*$ and $\alpha_k = 0$ for $k > r^*$.*

It is therefore clear that BN will not be a consistent estimator for r_1, and will tend to overestimate the number of relevant factors.

However, we can also consider an analogous estimator to BN, based on just the eigenvalues, that is designed to estimate the number of factors with $\alpha_k > .5$:

$$BN_{\sqrt{n}} = \max_{0\leq k\leq r_{max}} k \qquad \text{s.t. } \psi_k\left(\frac{X'X}{T}\right) > \hat{\sigma}^2(c+1)\sqrt{\frac{n}{c+1}}g(n). \tag{15}$$

It is then easy to show that:

Corollary 3. *Under Assumptions 1-4 $BN_{\sqrt{n}}$ is a consistent estimator for the number of factors r_1 such that $\alpha_k > 1/2$ for $k = 1, \ldots, r_1$ and $\alpha_k \leq 1/2$ for $k > r_1$.*

Given the equivalence established in (13), the proofs of Corollaries 2 and 3 follow the same arguments as the proof of Theorem 5 and are therefore omitted. While this section establishes that both $BN_{\sqrt{n}}$ and FC are consistent estimators for r_1, recall the discussion surrounding Figure 2b. Based on the steeper slope of $\hat{\Upsilon}_k^2$ around the chosen threshold ($\sqrt{n}$), FC is expected to perform better in finite samples.

4.1.2 Difference Estimators

Instead of choosing a cutoff value, Onatski (2010) establishes that the eigenvalues of the idiosyncratic errors cluster together, while the r eigenvalues corresponding to factors will remain separated. Based on this, one can construct an estimator based on the difference between two subsequent eigenvalues. Starting from r_{max} and successively looking at the difference between two subsequent eigenvalues in decreasing order, the estimator yields $\hat{r} = AO$, the first number at which this difference becomes larger than some threshold Q:[8]

$$AO = \max_{0 \leq k \leq r_{max}} k \quad \text{s.t. } \psi_k(\frac{X'X}{T}) - \psi_{k+1}(\frac{X'X}{T}) \geq Q = \max_{0 \leq k \leq r_{max}} k \quad \text{s.t. } \hat{\Upsilon}_k^0 - \hat{\Upsilon}_{k+1}^0 \geq Q.$$

Onatski (2010) considers any factors strong enough to be included in the model as soon as their cumulative effects grow with the sample size: the target of estimation r^* is defined as the number of factors with $lim_{n\to\infty}\|\mathcal{A}_k\| = \infty$. As discussed in previous sections, there are both theoretical and empirical reasons why a practitioner may conclude that some of these factors are too weak to be included in the model. Instead, an estimator for the number of relevant factors could take the form

$$FD = \max_{0 \leq k \leq r_{max}} k \quad \text{s.t. } \hat{\Upsilon}_k^2 - \hat{\Upsilon}_{k+1}^2 \geq \frac{n}{h(n)},$$

where $h(n)$ is a function such that (i) $h(n) \to \infty$ and (ii) $h(n)/g(n)^2 \to 0$ as $n \to \infty$, and $g(n)$ fulfills the conditions stated in Theorem 4.

Theorem 6. *Under Assumptions 1-4, FD is a consistent estimator for the number of factors r_1 such that $\alpha_k > 1/2$ for $k = 1, \ldots, r_1$ and $\alpha_k \leq 1/2$ for $k > r_1$.*

Proof. First note that, because $\hat{\Upsilon}_k^2 = O_p(n^{2\alpha_k}/g(n)^2) = o_p(\frac{n}{h(n)})$ for any k with $\alpha_k \leq 1/2$, $(\hat{\Upsilon}_k^2 - \hat{\Upsilon}_{k+1}^2) = o_p(\frac{n}{h(n)})$ for $k > r_1$.

[8] Under more restrictive assumptions on the error terms, Onatski (2010) proposes an appealing way to calibrate the tuning parameter Q, exploiting the shape of the eigenvalue distribution of the idiosyncratic noise at their edge, that unfortunately is no longer valid in the setup of this paper.

Next consider $k = r_1$. By Theorem 4, if $\alpha_k > .5$, $lim_{n\to\infty} P\left(\hat{\Upsilon}^2_{r_1} > Q_1 \frac{n}{h(n)}\right) = 1$ and, also by Theorem 4, $lim_{n\to\infty} P\left(\hat{\Upsilon}^2_{r_1+1} < Q_2 \frac{n}{h(n)}\right) = 1$, for any finite constants $Q_1, Q_2 > 0$. Choosing Q_1 and Q_2 such that $Q_1 - Q_2 = 1$ gives

$$lim_{n\to\infty} P\left((\hat{\Upsilon}^2_r - \hat{\Upsilon}^2_{r+1}) > \frac{n}{h(n)}\right) = 1.$$

□

4.1.3 Ratio Estimators

Another estimator that has been introduced to the literature and shown to perform well is based on the ratio of two subsequent eigenvalues following Ahn and Horenstein (2013):

$$AH = \underset{0\leq k\leq r_{max}}{\arg\max} \frac{\psi_k(\frac{X'X}{T})}{\psi_{k+1}(\frac{X'X}{T})} = \underset{1\leq k\leq r_{max}}{\arg\max} \frac{\hat{\Upsilon}^0_k}{\hat{\Upsilon}^0_{k+1}}.$$

Assumption 5. $\alpha_k > \frac{1}{2}$ *for* $k = 1, \ldots, r_1$ *and* $\alpha_k = 0$ *for* $k = r_1 + 1, \ldots, r_{max}$.

An advantage of the ratio estimator is that it is less dependent on any tuning parameter. Because the ratio estimator explicitly relies on a large gap in the eigenvalue distribution, we require an additional assumption of such a gap in Assumption 5 to establish consistency of ratio-based estimators below. Assumption 5 rules out any factors affecting an increasing number of covariates unless the number of affected covariates increases at a rate faster than $\sqrt{n}$. This assumption is somewhat restrictive, but still less restrictive than the setup of Ahn and Horenstein (2013), who impose $|\mathcal{A}_k| \asymp n$ for $k = 1, \ldots, r$. We expect that setting $u > 0$ should again improve the performance of this estimator in the presence of local factors and, in line with the ER estimator above, suggest a similar estimator based on the quantity $\hat{\Upsilon}^2_k$:

$$FR = \underset{0\leq k\leq r_{max}}{\arg\max} \frac{\hat{\Upsilon}^2_k}{\hat{\Upsilon}^2_{k+1}}.$$

Theorem 7. *Under Assumptions 1-5,* FR *is a consistent estimator for the number of factors* r_1 *such that* $\alpha_k > 1/2$ *for* $k = 1, \ldots, r_1$ *and* $\alpha_k \leq 1/2$ *for* $k > r_1$.

Proof. First consider $k = r_1 + 1, \ldots, r_{max}$. By Theorem 4, $\hat{\Upsilon}^2_k = O_p(\frac{1}{g(n)^2})$ and thus for every $c_1 > 0$, $lim_{n\to\infty} P\left(\hat{\Upsilon}^2_k \geq c_1\right) = 0$. Further, by Lemma 12 in the Online Appendix there exists a constant $c_2 > 0$, such that $lim_{n\to\infty} P\left(n\hat{\Upsilon}^2_{k^*} \geq c_2\right) = 1$ for $k^* = r_1 + 1, \ldots, r_{max}$. Then, for any

$k = r_1 + 1, \ldots, r_{max}$, any finite $c > 0$, setting $c_1 = cc_2$ yields

$$\begin{aligned} lim_{n\to\infty} P(\frac{\hat{\Upsilon}_k^2}{\hat{\Upsilon}_{k+1}^2} > cn) &= lim_{n\to\infty} \Bigg[P(\frac{\hat{\Upsilon}_k^2}{\hat{\Upsilon}_{k+1}^2} > cn | \hat{\Upsilon}_{k+1}^2 < \frac{c_2}{n}) P\left(\hat{\Upsilon}_{k+1}^2 < \frac{c_2}{n}\right) \\ &\qquad + P(\frac{\hat{\Upsilon}_k^2}{\hat{\Upsilon}_{k+1}^2} > cn | \hat{\Upsilon}_{k+1}^2 \geq \frac{c_2}{n}) P\left(\hat{\Upsilon}_{k+1}^2 \geq \frac{c_2}{n}\right) \Bigg] \\ &= lim_{n\to\infty} P(\frac{\hat{\Upsilon}_k^2}{\hat{\Upsilon}_{k+1}^2} > cn | \hat{\Upsilon}_{k+1}^2 \geq \frac{c_2}{n}) + 0 = lim_{n\to\infty} P(\hat{\Upsilon}_k^2 > cn\hat{\Upsilon}_{k+1}^2) \\ &\leq lim_{n\to\infty} P(\hat{\Upsilon}_k^2 > cc_2) = lim_{n\to\infty} P(\hat{\Upsilon}_k^2 > c_1) = 0. \end{aligned}$$

Next, consider $k = r_1$. By Assumption 5, $\alpha_k > 1/2$ and there exists a finite $q_1 > 0$ such that $lim_{n\to\infty} P\left(\hat{\Upsilon}_{r_1}^2 > q_1 n\right) = 1$. Using Assumption 5 again, $\hat{\Upsilon}_{r_1+1}^2 = O_p(\frac{1}{g(n)^2})$ and thus for every $q_2 > 0$, $P\left(\hat{\Upsilon}_{r_1+1}^2 \geq q_2\right) = 0$. Then, for any $q > 0$ and setting $q_2 = q_1/q$:

$$\begin{aligned} lim_{n\to\infty} P\left(\frac{\hat{\Upsilon}_{r_1}^2}{\hat{\Upsilon}_{r_1+1}^2} > qn\right) &= lim_{n\to\infty} \Bigg[P\left(\frac{\hat{\Upsilon}_{r_1}^2}{\hat{\Upsilon}_{r_1+1}^2} > qn | \hat{\Upsilon}_{r_1+1}^2 < q_2\right) P\left(\hat{\Upsilon}_{r_1+1}^2 < q_2\right) \\ &\qquad + P\left(\frac{\hat{\Upsilon}_{r_1}^2}{\hat{\Upsilon}_{r_1+1}^2} > qn | \hat{\Upsilon}_{r_1+1}^2 \geq q_2\right) P\left(\hat{\Upsilon}_{r_1+1}^2 \geq q_2\right) \Bigg] \\ &= lim_{n\to\infty} \left[P\left(\frac{\hat{\Upsilon}_{r_1}^2}{\hat{\Upsilon}_{r_1+1}^2} > qn | \hat{\Upsilon}_{r_1+1}^2 < q_2\right) + 0 \right] \\ &\geq lim_{n\to\infty} P(\hat{\Upsilon}_{r_1}^2 > q_2 qn) = lim_{n\to\infty} P(\hat{\Upsilon}_{r_1}^2 > q_1 n) = 1. \end{aligned}$$

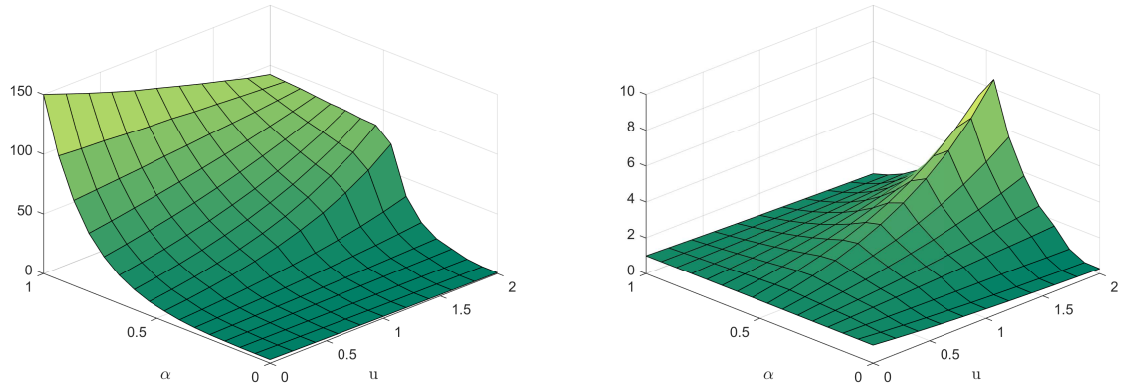
Finally, consider $k = 1, \ldots, r_1 - 1$. We note that in that case we already established that there exists a finite $q_1 > 0$ such that $lim_{n\to\infty} P\left(\hat{\Upsilon}_{k+1}^2 > q_1 n\right) = 1$. It then immediately follows that, for any $c_3 > 0$,

$$lim_{n\to\infty} P\left(\frac{\hat{\Upsilon}_k^2}{\hat{\Upsilon}_{k+1}^2} > c_3 n\right) = 0.$$

□

Corollary 4. *Under Assumptions 1-5, AH is a consistent estimator for the number of factors* r_1.

The proof largely follows the same arguments as those in the proof of Theorem 7 and is therefore relegated to Online Appendix B.2.



(a) Estimated $\hat{\Upsilon}_1^u$ **(b)** Estimated $\hat{S}_1^u$

Figure 3: Empirical behavior of key quantities in simple DGP. Depicted are $\hat{\Upsilon}_1^u = \psi_1(\frac{X'X}{T})\hat{S}_1^u$ and $\hat{S}_1^u$ as a function of both factor strength α and tuning parameter u, which governs the relative weight of eigenvalue and eigenvector. Data simulated with single factor of varying strength ($|\mathcal{A}| = n^\alpha$), $z = \sqrt{n}\sqrt{loglog(n)}$, and $n = 300$, $T = 500$. Figure depicts averages based on 500 simulations.

5 Simulations

We next present simulation evidence to assess the adequacy of our asymptotic approximations in finite sample. Throughout, we fix $g(n) = 0.7\sqrt{loglog(n)}$.[9]

We start by depicting the empirical analog to Figure 2b for a simple DGP in Figure 3. The simulated data has a single factor $F_1 \overset{i.i.d.}{\sim} N(0,1)$, with $T = 500$ and $n = 300$. All loadings are 1 on a random subset of covariates with cardinality $|\mathcal{A}| = n^\alpha$ and 0 everywhere else. Error terms u_{it} are $i.i.d.$ from a standard normal distribution, and each variable X_i is centered and divided by its standard deviation. In line with Figure 2b, we then vary α from 0 to 1 and u from 0 to 2. Figure 3a depicts the average value of $\hat{\Upsilon}_1^u$ across 500 simulations. Note the close resemblance in shape to Figure 2b. As u increases, a steep increase in $\hat{\Upsilon}_1^u$ emerges around $\tau = .5$. We take Figure 3a as an encouraging sign that the finite sample behavior of $\hat{\Upsilon}_1^u$ is well-approximated by the asymptotic theory of Section 3 at least in this simple setting.

To provide further intuition for the shape of Figure 3a, Figure 3b depicts $\hat{S}_1^u$ (the eigenvector component of $\hat{\Upsilon}_k^u$). We note that the measure of concentration for the eigenvector, $\hat{S}_1^2$, takes its highest value at $\alpha_1 = .5$.

We next consider more realistic settings as they might be observed in practice and consider a

[9]Note that $g(n)$ clearly fulfills the two criteria stated in Theorem 4: It grows with n, but at a very slow rate and is dominated by n^ε for any $\varepsilon > 0$. Further, in practice, $g(n) = 0.7\sqrt{loglog(n)} \approx 1$ for most relevant sample sizes. This recommendation is therefore generic. We show in Online Appendix A.3 that our results are robust across different values of a if we set $g(n) = a\sqrt{loglog(n)}$.

panel with

$$\underset{(500\times300)}{X} = \underset{(500\times6)}{F}\underset{(6\times300)}{\Lambda'} + \underset{(500\times3)}{G}\underset{(3\times300)}{\Lambda^{w'}} + \sqrt{\theta}\underset{(500\times300)}{e},$$

where $(T, n) = (500, 300)$ falls within the range of dimensions usually considered in the literature and will be varied later on. The variables exhibit a factor structure with 6 relevant factors $F_k, k = 1, 2...., 6$, drawn from a standard normal distribution. The 500×6 loading matrix Λ is created by filling random subsets of its columns with $(1 + \eta_{ik})$, where η_{ik} is drawn from a standard normal. These subsets will be of varying size and dictate which variables are affected by the corresponding factor, with the sequence of group sizes given by $\{|\mathcal{A}_k|\}_{k=1}^6 = \{n, n^{.85}, n^{.75}, n^{2/3}, n^{2/3}, n^{.6}\}$ rounded to the nearest integer for the 6 factors. All other entries in Λ are zero. Note that this implies that the orthogonality condition in Assumption 1(b) is in fact violated to better approximate a factor structure as it may occur in practice. The degree of this violation is depicted in Online Appendix Figure 1.

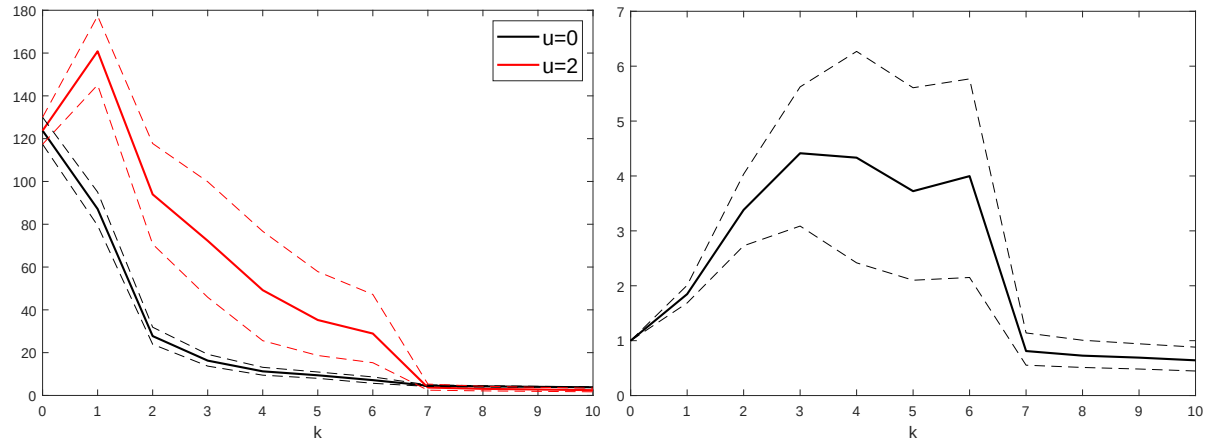
There are three additional factors – G_1, G_2, G_3 – also drawn from a standard normal, but which affect too few outcomes to be considered relevant. Their loading matrix Λ^w has entries $(1 + \eta_i)$, where η_i is drawn from a standard normal on random subsets of its columns with cardinalities $n^{1/3}$, $n^{1/4}$ and $log(n)$, again rounded to the nearest integer. All remaining entries are zero. Thus, in this DGP, $r_1 = 6$. For the idiosyncratic part, we allow for both cross-sectional and intertemporal correlation. We model the errors as

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

with baseline parameter values of $(\rho, \beta) = (0.3, 0.1)$ as in Onatski (2010). The parameter θ varies the signal-to-noise ratio and we set $\theta = 1.5$ in our baseline model. The factor structure and signal-to-noise ratio of our baseline DGP are designed to closely reproduce the scree plot in the macroeconomic application (see Figure 1 in the introduction).

Figure 4 depicts the behavior of both $\hat{\Upsilon}_k^u, u \in \{0, 2\}$ and $\hat{S}_k^2$ in simulated data from the baseline DGP. The bold line depicts the average and the dashed lines depict the pointwise 5th and 95th percentile of the respective quantities. As expected, the presence of local factors of varying pervasiveness means that the eigenvalues decrease gradually with no clear gap. On the other hand, we observe a jump at $k = 6$ when $u = 2$. The jump is due to the behavior of S_k^u, depicted in Figure 4b. According to this measure, the eigenvectors corresponding to the local factors F_2 to F_6 are indeed more concentrated on a subset of its entries.

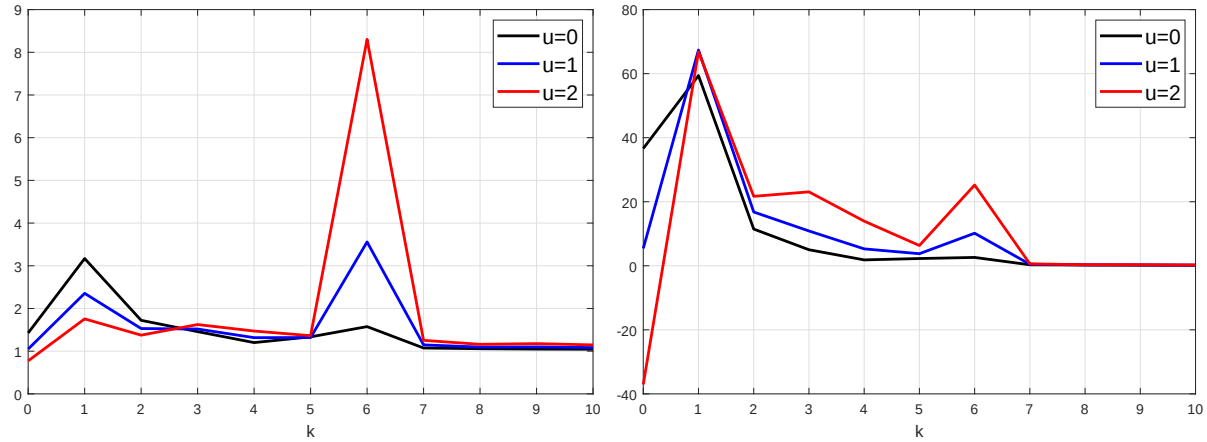
We next depict the ratios and differences of subsequent values of $\hat{\Upsilon}_k^u$ in Figure 5. Consider an estimator constructed as the maximum of subsequent ratios of $\hat{\Upsilon}_k^u$, which are depicted in Figure 5a.



(a) Empirical distribution of $\hat{\Upsilon}_k^u$. **(b)** Empirical distribution of $\hat{S}_k^2$.

Figure 4: Illustration of key quantities in a simulated dataset. Solid line depicts average across 1000 realizations. Dashed lines illustrate 5th and 95th quantile. Data generated by baseline DGP, with $(n, T) = (300, 500)$, $(\rho, \beta) = (0.3, 0.1)$, $r_1 = 6$, $\theta = 1.5$.

In contrast to an estimator derived solely from the eigenvalues of $\frac{X'X}{T}$ (AH), which suggests the presence of a single factor based on the average depicted here, incorporating the eigenvectors by setting $u = 2$ (FR) clearly yields an estimate of $\hat{r} = 6$. For an estimator based on the differences (Figure 5b) it is more difficult to gauge what the estimator would select from the picture, but we similarly observe a larger jump at $k = 6$ as the tuning parameter u increases.



(a) Ratio $\hat{\Upsilon}_k^u / \hat{\Upsilon}_{k+1}^u$, $u \in \{0, 1, 2\}$ **(b)** Difference $\hat{\Upsilon}_k^u - \hat{\Upsilon}_{k+1}^u$, $u \in \{0, 1, 2\}$

Figure 5: Depicted are averages across 1000 replications in a simulated dataset. Data generated by baseline DGP, with $(n, T) = (300, 500)$, $(\rho, \beta) = (0.3, 0.1)$, $\theta = 1.5$, $r_1 = 6$.

We will next explore the performance of the various estimators from Section 4.1 for varying amounts of correlation in the error terms, signal-to-noise ratios, and sample sizes. The estimators we consider are repeated in Table 2 for reference.

Figures 6 - 7 depict the performance of various estimators for the number of factors as the

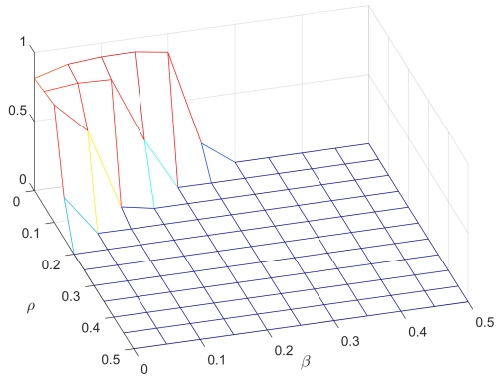
Estimator	Formula		Reference
AH	$\arg\max_{0\leq k\leq r_{max}}$	$\hat{\Upsilon}_k^0/\hat{\Upsilon}_{k+1}^0$	Ahn and Horenstein (2013)
FR	$\arg\max_{0\leq k\leq r_{max}}$	$\hat{\Upsilon}_k^2/\hat{\Upsilon}_{k+1}^2$	
BN	$\max_{0\leq k\leq r_{max}} k$	s.t. $\hat{\Upsilon}_k^0 > \hat{\sigma}^2(c+1)\log(\frac{n}{c+1})$	Bai and Ng (2002)
FC	$\max_{0\leq k\leq r_{max}} k$	s.t. $\hat{\Upsilon}_k^2 > \hat{\sigma}^2\frac{n}{0.1\sqrt{loglog(n)}}$	
$BN_{\sqrt{n}}$	$\max_{0\leq k\leq r_{max}} k$	s.t. $\hat{\Upsilon}_k^0 > \hat{\sigma}^2(c+1)\sqrt{\frac{n}{c+1}}g(n)$	
AO	$\max_{0\leq k\leq r_{max}} k$	s.t. $\hat{\Upsilon}_k^0 - \hat{\Upsilon}_{k+1}^0 \geq Q\}$	Onatski (2010)

Table 2: Definitions of the six estimators considered throughout. For all estimators we set $r_{max} = 15$ and $g(n) = 0.7\sqrt{loglog(n)}$ throughout our simulations. Note that $g(n) = 0.7\sqrt{loglog(n)} \approx 1$ for most relevant sample sizes. We show in Online Appendix A.3 that our results are robust across different values of a for $g(n) = a\sqrt{loglog(n)}$.

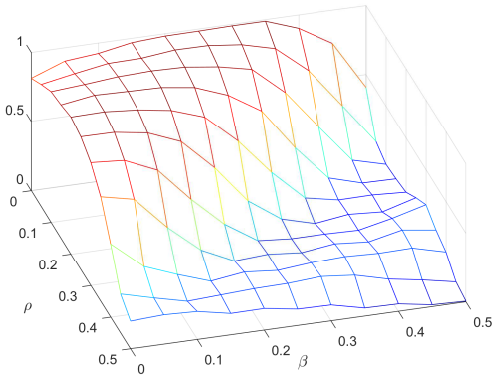
correlation in the idiosyncratic noise increases. In these figures, we vary the dependence structure of the error term along a two dimensional grid of (ρ, β), altering both the amount of autocorrelation and cross-sectional dependence. For the FC estimator, we need to prespecify a threshold function and use $h(n) = 0.1\sqrt{loglog(n)}$ throughout. However, we find that, in line with the literature (e.g. Alessi et al. (2010)), the estimated number of factors can be sensitive to the choice of $h(n)$.

Figure 6 depicts the percentage of simulations in which an estimator correctly estimates the number of relevant factors to be 6. Figure 7 depicts the average number of factors an estimator yields across simulations. In both figures, panel (a) uses the thresholding estimator based on $\hat{\Upsilon}_k^0$ (BN) to estimate the number of factors. Panel (b) uses the thresholding estimator based on $\hat{\Upsilon}_k^2$ (FC). Panel (c) uses the maximum ratio of two subsequent values of $\Upsilon_k^0 = \psi_k(\frac{X'X}{T})$ (AH) and panel (d) uses the maximum ratio of two subsequent values of Υ_k^2 (FR). Both panels (a) and (b), and (c) and (d) are therefore directly comparable to each other: panels (a) and (c) depict the results of the existing estimators based on the eigenvalues Υ_k^0, while panels (b) and (d) depict the corresponding estimators based on Υ_k^2. Panel (e) depicts the estimated number of factors using the AO estimator of Onatski (2010), while panel (f) depicts the alternative thresholding estimator based on the eigenvalues $BN_{\sqrt{n}}$.

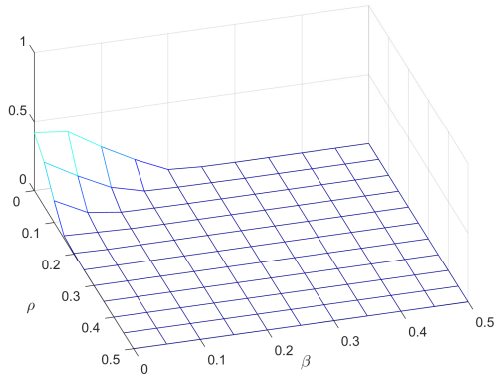
Under an exact factor structure, with $(\rho, \beta) = (0, 0)$, we see from Figure 6 that only the two estimators incorporating the additional information in the eigenvectors (Figures 6b and 6d) are able to identify the number of relevant factors reliably. They are also more robust to the introduction of dependence in the errors. They are on average correct in the simple case of no correlation in the error terms and remain correct for modest levels of cross-sectional and intertemporal correlation. In particular, when comparing panels (a) and (c) with panels (b) and (d), where the estimators are directly comparable, we observe a clear benefit from setting $u > 0$. Based on Figures 6 and 7, we conclude that the FR estimator, based on the ratio of subsequent values of $\hat{\Upsilon}_k^2$, performs best across



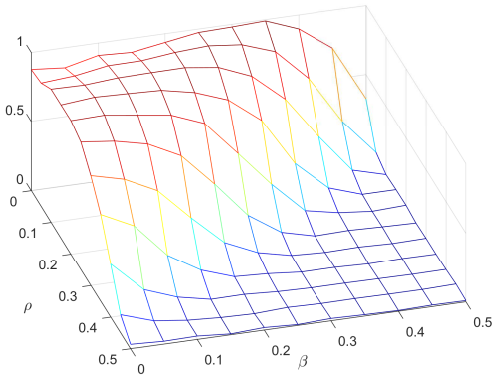
(a) Thresholding based on $\hat{\Upsilon}_k^0$ (BN)



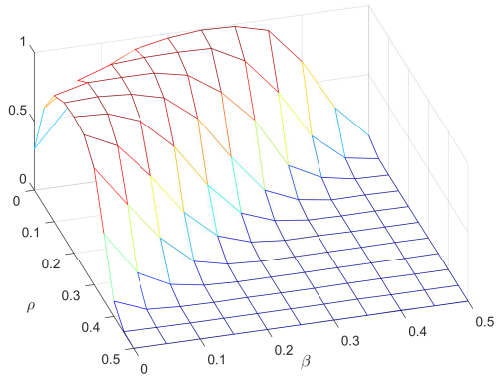
(b) Thresholding based on $\hat{\Upsilon}_k^2$ (FC)



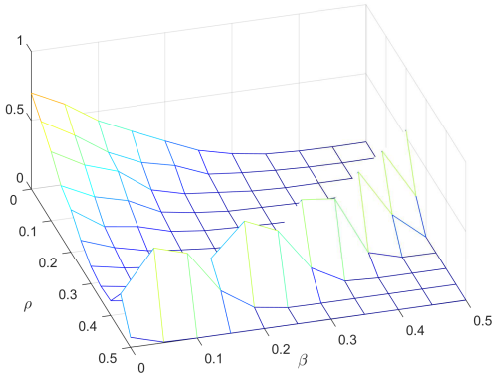
(c) Maximum ratio of two subsequent values of $\hat{\Upsilon}_k^0$ as in Ahn and Horenstein (2013) (AH)



(d) Maximum ratio of two subsequent values of $\hat{\Upsilon}_k^2$ (FR)

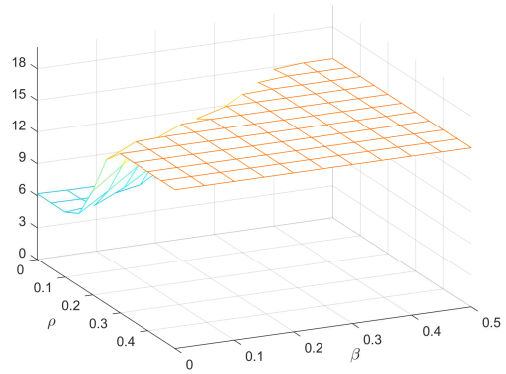


(e) Difference of two subsequent eigenvalues as in Onatski (2010) (AO)

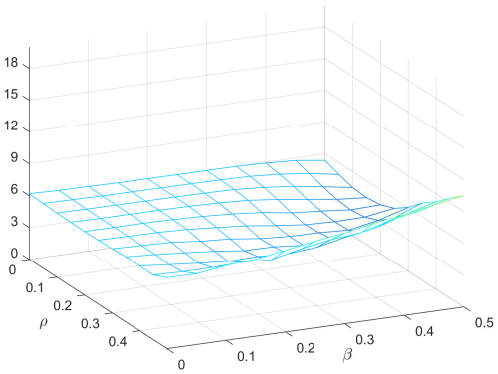


(f) Thresholding based on $\hat{\Upsilon}_k^0$ $(BN_{\sqrt{n}})$

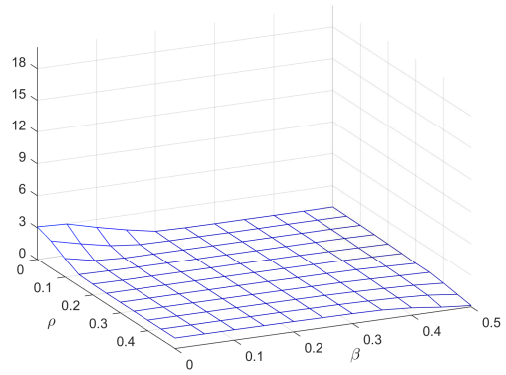
Figure 6: Percentage of simulations correctly estimating the number of "relevant" factors r_1 as both cross-sectional and intertemporal correlation is varied along a grid of (ρ, β). Data generated by baseline DGP, with $(n, T) = (300, 500)$, $\theta = 1.5$, and $r_1 = 6$. Figure based on 500 replications.



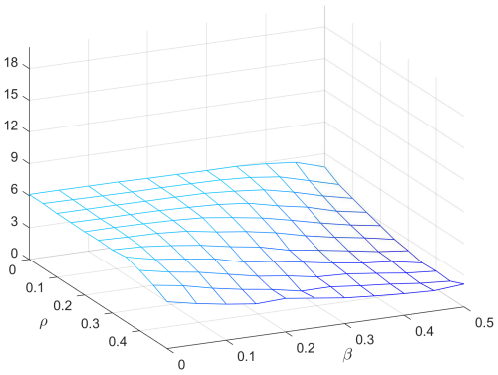
(a) Thresholding based on $\hat{\Upsilon}_k^0$ (BN)



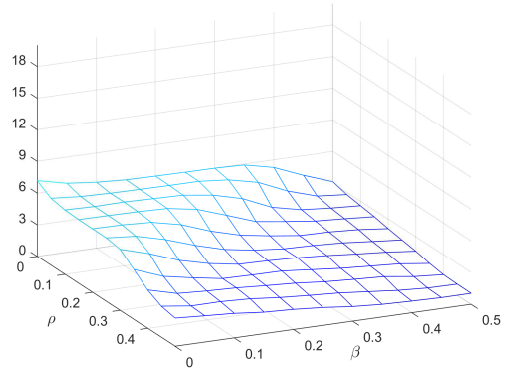
(b) Thresholding based on $\hat{\Upsilon}_k^2$ (FC)



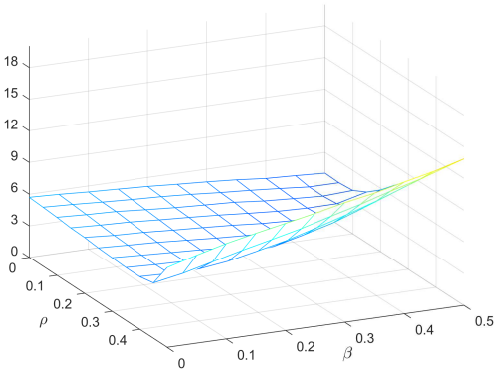
(c) Maximum ratio of two subsequent values of $\hat{\Upsilon}_k^0$ as in Ahn and Horenstein (2013) (AH)



(d) Maximum ratio of two subsequent values of $\hat{\Upsilon}_k^2$ (FR)

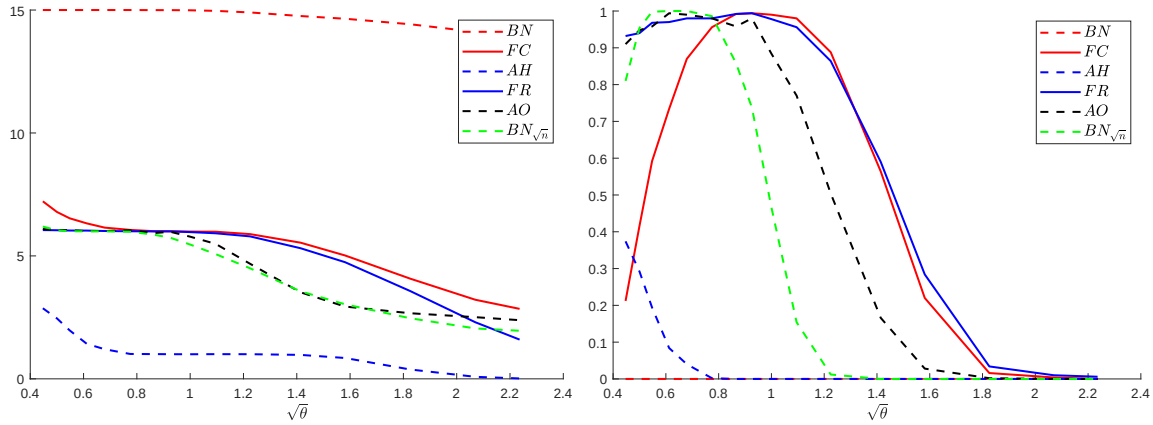


(e) Difference of two subsequent eigenvalues as in Onatski (2010) (AO)



(f) Thresholding based on $\hat{\Upsilon}_k^0$ $(BN_{\sqrt{n}})$

Figure 7: Average estimate for number of factors as both cross-sectional and intertemporal correlation is varied along a grid of (ρ, β). Data generated by baseline DGP, with $(n, T) = (300, 500)$, $\theta = 1.5$, and $r_1 = 6$. Figure based on 500 replications.



(a) Average estimate for number of factors according to BN, FC, AH, FR, AO and $BN_{\sqrt{n}}$

(b) % of simulations correctly estimating number of factors r_1 according to BN, FC, AH, FR, AO and $BN_{\sqrt{n}}$

Figure 8: Empirical behavior of estimators as the relative variance of idiosyncratic noise increases by varying θ. Data generated by baseline DGP, with $(n, T) = (300, 500)$, $(\rho, \beta) = (0.3, 0.1)$, and $r_1 = 6$. Figure based on 500 replications.

the parameter space. Further, Onatski (2010) argues that the parameter pair $(\rho, \beta) = (0.3, 0.1)$ describes the data well in many financial applications. At those parameter values, the FC and FR estimators clearly perform best. Specifically, on average the FR estimator yields an estimate of 5.74 factors, while the best performing existing estimator based on the eigenvalues (AO) yields 4.81 factors on average.

Fixing the correlation structure in the errors to $(\rho, \beta) = (0.3, 0.1)$, Figure 8 depicts the sensitivity of the estimators to the amount of noise in the data by varying θ. Estimators using the information inherent in the eigenvectors remain correct for a larger range of θ compared to their counterparts derived solely from the empirical eigenvalues. The FR estimator performs best in this dimension.

For the final set of simulations we vary both the cross-sectional dimension as the well as the time horizon of the data. Table 3 depicts the results. Each entry in Table 3 consists of two numbers $\hat{r}/\%$, where $\hat{r}$ is the average number of estimated factors, and $\%$ is the percentage correctly classifying $r = 6$. In small samples all estimators perform poorly. While the ratio- and difference-based estimators tend to underestimate the true number of factors, both estimators based on thresholding the eigenvalues tend to overestimate the number of factors in small samples. Again comparing BN and AH directly with FC and FR, the previous patterns are confirmed: Setting $u > 0$ significantly improves the performance of the estimator.

In Online Appendix A.4, we repeat the analysis of this section with an alternative DGP that has a strong factor structure. In particular, we use the same baseline DGP as in this section but set $\lambda_{ik} = 1 + \eta_{ik}$, $\eta_i \sim N(0, 1)$, for every entry in Λ and exclude the very weak factors G. Thus

n	T	ER	TR	PC	$PC_{\sqrt{n}}$	TC	ED
100	100	1.06 / 0.00	1.82 / 0.00	15 / 0.00	9.78 / 0.00	8.94 / 0.00	1.42 / 0.00
100	150	1.04 / 0.00	1.64 / 0.02	15 / 0.00	7.32 / 0.10	6.96 / **0.30**	1.72 / 0.00
150	100	1 / 0.00	1.92 / 0.00	15 / 0.00	8.66 / 0.00	10.2 / 0.00	1.62 / 0.00
150	250	1.02 / 0.00	2.4 / 0.04	15 / 0.00	4.04 / 0.02	5.66 / **0.54**	2.06 / 0.02
150	500	1.02 / 0.00	3.8 / 0.24	14.6 / 0.00	3.54 / 0.00	5.36 / **0.42**	3.1 / 0.12
300	250	1 / 0.00	4.78 / 0.24	15 / 0.00	3.08 / 0.00	5.68 / **0.46**	2.36 / 0.00
300	500	1 / 0.00	5.56 / 0.78	15 / 0.00	2.94 / 0.00	5.86 / **0.86**	4.66 / 0.50
300	750	1 / 0.00	5.96 / **0.96**	13 / 0.00	3.04 / 0.00	5.96 / **0.96**	5.74 / 0.86
500	250	1 / 0.00	4.5 / 0.16	15 / 0.00	2.88 / 0.00	5.9 / 0.54	2.86 / 0.02
500	500	1 / 0.00	5.98 / **0.98**	15 / 0.00	2.86 / 0.00	5.98 / **0.98**	5.22 / 0.66
500	750	1 / 0.00	5.98 / **0.98**	14.1 / 0.00	2.9 / 0.00	5.98 / **0.98**	5.94 / 0.94
1000	1000	1 / 0.00	6 / **1.00**	13.3 / 0.00	2.88 / 0.00	6 / **1.00**	6 / **1.00**

Table 3: Table depicts the performance of different estimators as the sample size is varied along a grid of (n, T). Data generated by baseline DGP, with $(\rho, \beta) = (0.3, 0.1)$, $\theta = 1.5$, and $r_1 = 6$. Each entry depicts a combination $\hat{r}/\%$, where $\hat{r}$ is the average number of estimated factors, and $\%$ is the percentage correctly classifying $r_1 = 6$. In each row, the highest percentage is highlighted. Table based on 500 replications.

$\alpha_k = 1$ for $k = 1, \ldots, 6$ as is usually the case in the literature. We find that, under a strong factor structure, estimators incorporating the partial sums in the eigenvector generally perform no worse than existing estimators, although the AO estimator of Onatski (2010) tends to perform particularly well in smaller samples. We therefore conclude that raising $\hat{\Upsilon}_k^u$ to a power $u > 0$ has little implication if all factors are strong, but yields significant performance gains if local factors are present in the data.

Based on these findings, our recommendation for estimating the number of relevant factors is to use the FR estimator (which is also less dependent on tuning parameters compared to FC, which require a researcher to prespecify a threshold function through the choice of $h(n)$) with its implementation outlined as follows:

1. Obtain preliminary estimates $\hat{F}$, $\hat{\Lambda}$ using the first r_{max} principal components, where r_{max} is large enough such that $\psi_k(\frac{X'X}{T})$ is bounded for $k > r_{max}$.

2. With $z = 0.7\sqrt{log(log(n))}\sqrt{n}$, rounded to the nearest integer, compute

$$\hat{\Upsilon}_{zk}^2 \equiv \psi_k(\frac{X'X}{T})\hat{S}_{zk}^u \equiv \psi_k(\frac{X'X}{T})\left(\frac{1}{z}\sum_i^z \frac{\hat{\lambda}_{ik}^2}{\sqrt{\frac{1}{n}\sum_{i=1}^n \hat{\lambda}_{ik}^2}}\right)^2.$$

3. Set

$$\hat{r} = FR = \underset{0 \leq k \leq r_{max}}{\arg\max} \frac{\hat{\Upsilon}_k^2}{\hat{\Upsilon}_{k+1}^2}.$$

6 The number of relevant factors for the US Economy

We next apply our estimators to one of the standard datasets in the factor model literature in macroeconomics (see, e.g. Stock and Watson 2005 and De Mol et al. 2008). The data contains quarterly observations of 207 macroeconomic variables, primarily for the US economy. In particular, we use the vintage of the dataset used in the handbook chapter of Stock and Watson (2016). 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 activity. The data ranges from 1959Q1-2014Q4.[10]

The dataset consists of series at multiple levels of aggregation. We only use the disaggregated time series in our estimation and, for simplicity, eliminate all series with missing observations. This leaves 94 time series.

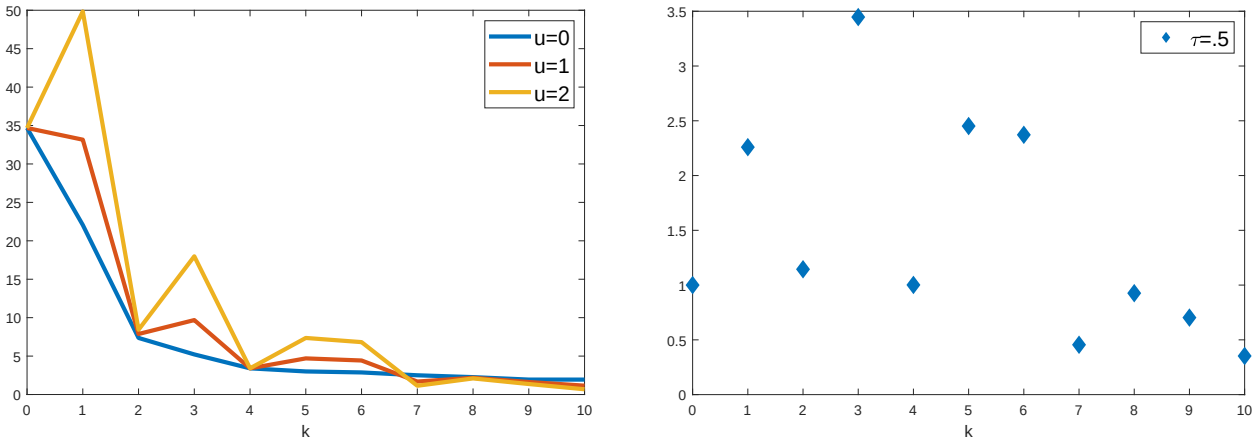
We start by depicting some of the key quantities introduced in this paper and provide some intuitive discussion of those. Figure 9 depicts the behavior of both $\hat{\Upsilon}_k^u, u \in \{0, 1, 2\}$ and $\hat{S}_k^2$ in the data (setting $u = 0$ reproduces Figure 1 from the introduction). If $\hat{\Upsilon}_k^u$ corresponds to a local factor with a strong effect on a subset of outcomes, it will be scaled up if $u > 0$. This is illustrated in Figure 9b, which suggests the 3rd, 5th, and 6th eigenvector are particularly concentrated on its largest loadings[11]. As a consequence, a visual inspection of Figure 9a indicates a drop-off at $k = 3$ and $k = 6$, suggesting the presence of either 3 or 6 factors, depending on the minimum strength of the factors a practitioner would like to keep in her model.

Table 4 summarizes the results of the 6 estimators considered throughout this paper. While both estimators derived from $\hat{\Upsilon}_k^2$ suggest the presence of 6 factors in the data, the three existing estimators from the literature (AH, BN, AO) find evidence for 1, 8, and 3 factors respectively.

To address the problem that the estimated number of factors is sensitive to the choice of cutoff under existing thresholding estimators, Alessi et al. (2010) suggest varying this threshold and ex-

[10] All variables have been transformed to achieve stationarity and a small number of outliers were removed. We follow the same transformations as Stock and Watson (2016) and also follow their practice in removing low-frequency trends in the data using a biweight low-pass filter, with a bandwidth of 100 quarters, as in Stock and Watson (2012). Data are available at `http://www.princeton.edu/~mwatson/publi.html`. For a full description of the data, as well as a more detailed description of the transformations to the raw data, we refer the reader to Stock and Watson (2016).

[11] This is confirmed in Online Appendix Figure 6, which depicts the distribution of the 25 largest loadings (in absolute value) for each factor. Online Appendix Table 2 further shows which economic indicators correspond to the largest loadings (in absolute value) associated with the three local factors.



(a) Empirical distribution of $\hat{\Upsilon}_k^u$ for $u \in \{0, 1, 2\}$ **(b)** Empirical distribution of $\hat{S}_k^2$

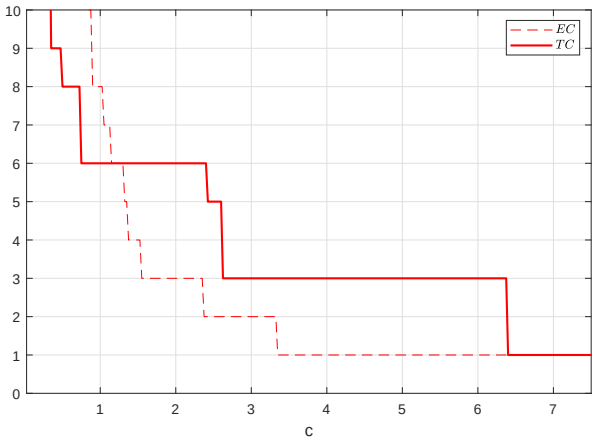
Figure 9: Illustration of key quantities for $\tau = 1/2$ in a dataset of US macroeconomic indicators

ploring how the estimated number of factors changes. We conduct a similar exercise in Figure 10a, which depicts the estimated number of factors based on BN and FC as a function of the tuning parameter c, which multiplies the thresholds in Theorem 5 and Corollary 2. Figure 10a confirms that incorporating the eigenvector makes the estimator less dependent on the choice of the tuning parameter. While varying the threshold for the eigenvalues leads to a gradual increase in the number of estimated factors, indicated by the absence of a prolonged flat region, varying the threshold for $\hat{\Upsilon}_k^2$, we observe two flat regions in the graph at $\hat{r} = 3$ and $\hat{r} = 6$ respectively.

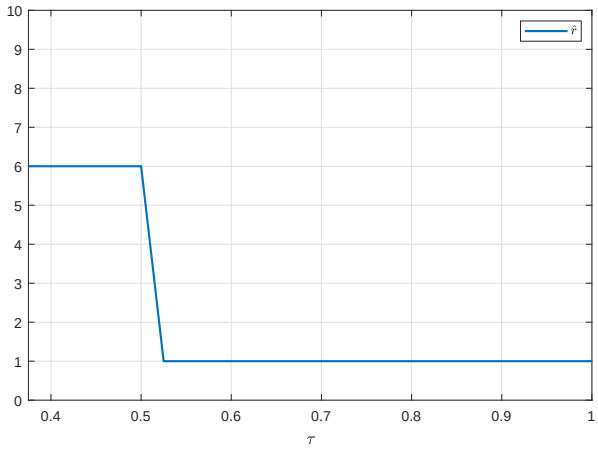
Using an estimator based on $\hat{\Upsilon}_k^2$, a practitioner could alternatively vary τ and observe how the estimated number of factors changes. We demonstrate this in Figure 10b, which depicts the value of the FR estimator as a function of the complexity parameter τ. Figure 10b suggests that there is one "global" factor in the data and that the second most important factor is significantly weaker than the first one, as indicated by the first flat region of the graph. Next, the 7th factor appears to be significantly weaker than the 6th, as indicated by the second flat region in Figure 10b. Thus,

Estimator	Formula		Estimated number of factors
AH	$\arg\max_{0\le k\le r_{max}}$	$\hat{\Upsilon}_k^0/\hat{\Upsilon}_{k+1}^0$	1
FR	$\arg\max_{0\le k\le r_{max}}$	$\hat{\Upsilon}_k^2/\hat{\Upsilon}_{k+1}^2$	6
BN	$\max_{0\le k\le r_{max}} k$	s.t. $\hat{\Upsilon}_k^0 > \hat{\sigma}^2(c+1)\log(\frac{n}{c+1})$	8
FC	$\max_{0\le k\le r_{max}} k$	s.t. $\hat{\Upsilon}_k^2 > \hat{\sigma}^2\frac{n}{g(n)}$	6
$BN_{\sqrt{n}}$	$\max_{0\le k\le r_{max}} k$	s.t. $\hat{\Upsilon}_k^0 > \hat{\sigma}^2(c+1)\sqrt{\frac{n}{c+1}}g(n)$	3
AO	$\max_{0\le k\le r_{max}} k$	s.t. $\hat{\Upsilon}_k^0 - \hat{\Upsilon}_{k+1}^0 \ge Q\}$	3

Table 4: Estimated number of factors in macroeconomic panel for the six estimators defined in Section 4 and considered throughout.



(a) Estimate $\hat{r}_1$ from thresholding based on $\hat{\Upsilon}_k^0$ (BN) and $\hat{\Upsilon}_k^2$ (FC) as a function of constant c, which multiplies cutoffs from Theorem 5 and Corollary 2.



(b) Estimate $\hat{r}_1$ from using the ratio estimator $FR = \hat{\Upsilon}_k^2/\hat{\Upsilon}_{k+1}^2$, $k \in \{0, \ldots, 10\}$ for varying tuning parameter $\tau \in [.375, 1]$.

Figure 10: Illustration of estimators in dataset of US macroeconomic indicators for varying tuning parameters

Figure 10 suggests that the choice of $\hat{r} = 6$ is quite robust.

7 Concluding Remarks

This paper develops factor models with local factors that affect only an unknown subset of the observables. It provides a theoretical foundation that justifies the use of both factor models and the principal component estimator in datasets with no clear gap in the eigenvalue distribution.

We argue that factors affecting proportionally more than $\sqrt{n}$ of the n observed variables are often the ones of economic interest. Under some regularity conditions, this coincides with the number of factors that can be estimated consistently using the principal component estimator.

We show that existing estimators in general do not yield a consistent estimate for this number of "relevant" factors. To estimate the number of economically important and estimable factors, we introduce novel estimators that exploit additional information in the eigenvectors of the covariance matrix. We validate our proposal in simulations and find significant finite sample gains over existing methods.

References

Acemoglu, D., V. M. Carvalho, A. Ozdaglar, and A. Tahbaz-Salehi (2012). The network origins of aggregate fluctuations. *Econometrica 80*(5), 1977–2016.

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

Alessi, L., M. Barigozzi, and M. Capasso (2010). Improved penalization for determining the number of factors in approximate factor models. *Statistics & Probability Letters 80*(23), 1806–1813.

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.

Andrews, D. W. and X. Cheng (2012). Estimation and inference with weak, semi-strong, and strong identification. *Econometrica 80*(5), 2153–2211.

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

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 (2006). Determining the number of factors in approximate factor models, errata. Technical report, Columbia University.

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

Bai, J. and S. Ng (2019). Rank regularized estimation of approximate factor models. *Journal of Econometrics 212*(1), 78–96.

Bailey, N., G. Kapetanios, and M. H. Pesaran (2016). Exponent of cross-sectional dependence: Estimation and inference. *Journal of Applied Econometrics 31*(6), 929–960.

Bailey, N., G. Kapetanios, and M. H. Pesaran (2020). Measurement of factor strength: Theory and practice. Technical report, CESifo Working Paper.

Bernanke, B. S. (1986). Alternative explanations of the money-income correlation. *Carnegie-Rochester Conference Series on Public Policy 25*, 49 – 99.

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

Cai, T. T., Z. Ma, and Y. Wu (2013). Sparse PCA: Optimal rates and adaptive estimation. *Annals of Statistics 41*(6), 3074–3110.

Carvalho, C. M., J. Chang, J. E. Lucas, J. R. Nevins, Q. Wang, and M. West (2008). High-dimensional sparse factor modeling: Applications in gene expression genomics. *Journal of the American Statistical Association 103*(484), 1438–1456.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research 1*(2), 245–276.

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.

Chudik, A., M. H. Pesaran, and E. Tosetti (2011). Weak and strong cross-section dependence and estimation of large panels. *Econometrics Journal 14*(1), C45–C90.

Cochrane, J. H. (1994). Shocks. In *Carnegie-Rochester Conference Series on Public Policy*, Volume 41, pp. 295–364. Elsevier.

Connor, G. and R. A. Korajczyk (1993). A test for the number of factors in an approximate factor model. *Journal of Finance 48*(4), 1263–1291.

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.

Foerster, A. T., P.-D. G. Sarte, and M. W. Watson (2011). Sectoral versus aggregate shocks: A structural factor analysis of industrial production. *Journal of Political Economy 119*(1), 1–38.

Gabaix, X. (2011). The granular origins of aggregate fluctuations. *Econometrica 79*(3), 733–772.

Gagliardini, P., E. Ossola, and O. Scaillet (2019). A diagnostic criterion for approximate factor structure. *Journal of Econometrics 212*(2), 503–521.

Gao, C., C. D. Brown, and B. E. Engelhardt (2013). A latent factor model with a mixture of sparse and dense factors to model gene expression data with confounding effects. *arXiv preprint arXiv:1310.4792*.

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

Hallin, M. and R. Liska (2007). The generalized dynamic factor model: Determining the number of factors. *Journal of the American Statistical Association 102*(478), 603–617.

Han, X. (2017). Shrinkage estimation of factor models with global and group-specific factors. Technical report, City University of Hong Kong.

Han, X. and M. Caner (2017). Determining the number of factors with potentially strong within-block correlations in error terms. *Econometric Reviews 36*(6-9), 946–969.

Horn, R. A. and C. R. Johnson (2012). *Matrix analysis*. Cambridge University Press.

Horvath, M. (1998). Cyclicality and sectoral linkages: Aggregate fluctuations from independent sectoral shocks. *Review of Economic Dynamics 1*(4), 781–808.

Kapetanios, G. (2004). A new method for determining the number of factors in factor models with large datasets. Technical report, Department of Economics, Queen Mary University of London.

Kapetanios, G. (2010). A testing procedure for determining the number of factors in approximate factor models with large datasets. *Journal of Business & Economic Statistics 28*(3), 397–409.

Kleibergen, F. (2009). Tests of risk premia in linear factor models. *Journal of Econometrics 149*(2), 149–173.

Long, J. B. and C. I. Plosser (1983). Real business cycles. *Journal of Political Economy 91*(1), 39–69.

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

Moon, H. R. and M. Weidner (2017). Dynamic linear panel regression models with interactive fixed effects. *Econometric Theory 33*, 158–195.

Onatski, A. (2009). Testing hypotheses about the number of factors in large factor models. *Econometrica 77*(5), 1447–1479.

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

Onatski, A. (2012). Asymptotics of the principal components estimator of large factor models with weakly influential factors. *Journal of Econometrics 168*(2), 244–258.

Onatski, A. (2015). Asymptotic analysis of the squared estimation error in misspecified factor models. *Journal of Econometrics 186*(2), 388–406.

Pati, D., A. Bhattacharya, N. S. Pillai, and D. Dunson (2014). Posterior contraction in sparse Bayesian factor models for massive covariance matrices. *Annals of Statistics 42*(3), 1102–1130.

Paul, D. and I. M. Johnstone (2012). Augmented sparse principal component analysis for high dimensional data. *arXiv preprint arXiv:1202.1242*.

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

Shukla, R. and C. Trzcinka (1990). Sequential tests of the arbitrage pricing theory: A comparison of principal components and maximum likelihood factors. *Journal of Finance 45*(5), 1541–1564.

Stock, J. H. and M. W. Watson (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association 97*(460), 1167–1179.

Stock, J. H. and M. W. Watson (2005). Implications of dynamic factor models for VAR analysis. Technical report, NBER Working Paper 114467.

Stock, J. H. and M. W. Watson (2012). Disentangling the channels of the 2007-09 recession. *Brookings Papers on Economic Activity 43*, 81–156.

Stock, J. H. and M. W. Watson (2016). Dynamic factor models, factor-augmented vector autoregressions, and structural vector autoregressions in macroeconomics. In J. B. Taylor and H. Uhlig (Eds.), *Handbook of Macroeconomics*, Volume 2, Chapter 8, pp. 415–525. Elsevier.

Ten Berge, J. M. and H. A. Kiers (1991). A numerical approach to the approximate and the exact minimum rank of a covariance matrix. *Psychometrika 56*(2), 309–315.

Trzcinka, C. (1986). On the number of factors in the arbitrage pricing model. *Journal of Finance 41*(2), 347–368.

Uematsu, Y. and T. Yamagata (2020). Estimation of weak factor models. Technical report, ISER Discussion paper 1053.

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