Federal Reserve Bank of Philadelphia

Working Papers
Research Department

WP 19-23
April 2019
https://doi.org/10.21799/frbp.wp.2019.23

A Generalized Factor Model with Local Factors

Simon Freyaldenhoven
Federal Reserve Bank of Philadelphia

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

A Generalized Factor Model with Local Factors*

Simon Freyaldenhoven†

Federal Reserve Bank of Philadelphia

April 19, 2019

Abstract

I extend the theory on factor models by incorporating local factors into the model. Local factors only affect an unknown subset of the observed variables. This implies a continuum of eigenvalues of the covariance matrix, as is commonly observed in applications. I 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. I then introduce a new class of estimators to determine the number of those relevant factors. Unlike existing estimators, my estimators use not only the eigenvalues of the covariance matrix, but also its eigenvectors. I find strong evidence of local factors in a large panel of US macroeconomic indicators.

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 advice. Their insightful comments on earlier versions of this draft have been of great help. 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 and the 2017 NBER/NSF Time Series Conference for helpful comments and suggestions. An earlier draft of this paper was circulated under the title "Factor Models of Arbitrary Strength".

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

†simon.freyaldenhoven@phil.frb.org

1 Introduction

Factor models allow for a large number of economic variables to be distilled into a small number of reference variables, enabling the analysis of otherwise prohibitively complex datasets. This paper generalizes standard factor models by introducing a novel theoretical framework incorporating factors of varying strength. Here, the strength of a factor is defined by the number of outcomes it affects.[1] Instead of ruling out "local" factors that only affect a subset of the observed variables, as is commonly done in the literature, I explicitly allow for such factors. I derive which factors are strong enough to be economically important and which factors are estimable using the common principal component estimator. I 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 derived from the empirical distribution of the eigenvalues. I argue that in a setting with local factors there is additional information in the eigenvectors and propose to exploit this additional information by incorporating partial sums of the eigenvectors into the estimator.

While local factors have long been acknowledged, the current literature requires a clear distinction between "large" and "small" groups of affected variables, ruling out factors that drive a decreasing fraction of the observables. For example, the handbook chapter of Connor and Korajczyk (1995) distinguishes between factors affecting at most a fixed number of firms and factors affecting at least a constant proportion of all firms.[2] This paper proposes a more general model that allows for groups of intermediate sizes. This generalization provides a better approximation to the data under a given sample size.

Although the standard model implies a clearly visible separation of the eigenvalues of the covariance matrix into two groups (large eigenvalues representing factor-related variation and small eigenvalues representing idiosyncratic variation), such a visible separation is typically not found in practice. For example, a popular dataset in which factor models have been used is the "Stock & Watson" dataset (Stock and Watson (2002a), De Mol et al. (2008)), consisting of a large panel of US macroeconomic indicators. Figure 1 depicts the distribution of eigenvalues in an updated

[1] Note that this is different from the weak factor framework of Onatski (2012) and Kleibergen (2009).

[2] Specifically, Connor and Korajczyk (1995) state: "Suppose that there is a large number (n) of assets each representing the common shares of one firm. Each firm belongs to one of a large number (m) of industries each with a small number (h, with h approximately equal to n/m) of firms. Idiosyncratic returns are correlated within industries but uncorrelated across industries. [...] Holding h constant and letting n and m increase, this series of covariance matrices has bounded eigenvalues. [...] On the other hand, suppose that there is a small number, k, of sectors, each containing n/k firms. All firms within sector j are subject to sector shock f_j with unit betas (for simplicity). Firms in sector j are unaffected by the shocks of other sectors. Given these assumptions, the sector shocks constitute pervasive risk. Note the clear distinction between industries (a small proportion of the firms are in each industry) versus sectors (a substantial proportion of the firms are in each sector)."

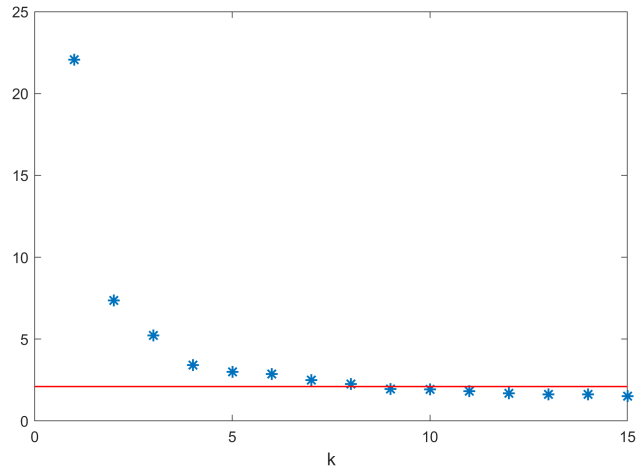


Figure 1: 20 largest eigenvalues of the covariance matrix for a dataset of 94 macroeconomic indicators in the US. Solid line indicates cutoff chosen according to Bai and Ng (2002) (with $r_{max} = 15$) to determine the number of factors. A more detailed discussion of this application can be found in Section 6.

vintage of this data.[3] The solid line indicates the cutoff between the two groups as chosen by a criterion of Bai and Ng (2002). A model implying a continuum of eigenvalues represents the data much better than such a classification into two groups. In finance, 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.

There are two kinds of weak factors that may induce a continuous decay in the distribution of eigenvalues. Such a factor can either have a weak effect on all observables, or it can affect only a subset of observables. This paper will largely focus on the latter scenario, although some results extend to the former. Recently a number of empirical studies have also 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. (2013)). The findings in De Mol et al. (2008) further provide empirical evidence for such a structure (see the discussion in Freyaldenhoven (2019)). 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. Han and Caner (2017) consider a model in which some factors may be less pervasive but treat all of those "local" factors as noise. To the best of my knowledge, the only theoretical papers in the direction of local factors in the sense of this paper are Wang (2008) and Choi et al. (2018). However, unlike those papers, I do not require the group structure and factor strengths to be known to the practitioner a priori.

Onatski (2009, 2010, 2012) proposes a 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

[3] I discuss the data in more detail in Section 6.

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, assume bounded eigenvalues of the covariance matrix, and remain largely agnostic about the factors themselves. By considering a diverging eigenvalue regime and explicitly modeling the factors, we are able to impose less restrictive assumptions on the error structure. There is also a related literature on sparse factor models under a Bayesian framework (e.g. 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), Antoine and Renault (2012)).

Before I 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)) .

Empirical evidence on whether weaker factors are priced appears somewhat mixed (e.g. Shukla and Trzcinka (1990)). In Section 3.2, I use the results in Green and Hollifield (1992) adapted to my framework to derive theoretical bounds on the strength of factors that will be priced. I find that the number of factors that are priced depends directly on the degree of diversification of the portfolios on the efficient frontier. The better diversified these portfolios are, the smaller the number of factors that have a non-zero factor premium.

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. I 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. I 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.

As a stylized model they consider macroeconomic panels with two factors. Some series are driven by two factors, some are only affected by one factor, and others are not associated with any factor. For example, the first n_1 series (only affected by the first factor) might be output and employment type series, the next n_2 series might be prices (affected by the second factor), the following n_3 series represent interest rates and are affected by both, and variations in the remaining series are purely idiosyncratic. If, for example, the cross section contains relatively few series representing prices and interest rates, this fits the framework of this paper.

Boivin and Ng (2006) use a Monte Carlo study to establish that the performance of the principal component estimator deteriorates as more "noisy" series are added, effectively making factors local in the sense of this paper. I provide an analytical framework, tying the convergence rate of a factor estimate to the factor's strength, that can help to explain their result.

2 A Model with Local Factors

To set up notation, define an n-dimensional process by X_t, $t = 1, 2, ..., T$. Let F_k, $k = 1, 2, ..., r$ denote the true factors. $\Lambda = [\lambda_{\cdot 1} \lambda_{\cdot 2} \cdots \lambda_{\cdot r}] = [\lambda_{1\cdot} \lambda_{2\cdot} \cdots \lambda_{n\cdot}]'$ denotes the matrix of factor loadings. Throughout, I 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. I assume that the data has a static factor structure[4]:

$$\underset{(T\times n)}{X^{(n)}} = \underset{(T\times r)}{F^{(n)}} \underset{(r\times n)}{\Lambda^{(n)\prime}} + \underset{(T\times n)}{e^{(n)}}. \tag{1}$$

I treat both Λ and F as parameters of the distribution of X. Throughout I 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 $\|B\|^2 = tr(B'B) = \sum_{ij} b_{ij}^2$. I further 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 $b_n = O(a_n)$. I 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.

[4] A more general setup would be the dynamic factor model of Forni et al. (2000), allowing for factor loadings that are represented by possibly infinite lag polynomials. However, whenever the order of such lag polynomials is bounded, the model can be rewritten in the static form with constant factor loadings, where the factors are augmented by a set of their own lags (see, e.g. Stock and Watson (2006)).

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* $|F_{tk}| < C\ \forall\ t$.

Part (a) of Assumption 1 requires n and T to be comparable even asymptotically. This assumption is common in the literature (e.g. Onatski (2012), Ahn and Horenstein (2013)) and plausible in many applications of factor models. Part (b) of Assumption 1 is considerably weaker than the standard assumptions in the literature (e.g. Stock and Watson (2002a), Bai and Ng (2002), Bai (2003), Ahn and Horenstein (2013)) in that the entries in $D^{(n)}$ are not assumed to diverge proportionally to n, thus allowing for weaker factors. All entries in $D^{(n)}$ can have different rates. Thus, rather than assuming pervasiveness of all factors, one can think of Assumptions 1(b)-(c) as identifying restrictions. I treat the factors in (1) as "primitive" exogenous forces. Because these forces are primitive, they do not have common causes and it is natural to treat them as approximately uncorrelated. The correlation between any two observables then of course arises because they are influenced by one or more of the same factors (Bernanke (1986)).

To simplify notation, I will omit the superscript (n) on matrices X, Λ, F, D and e in what follows.

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 allows for the loadings of any given factor k to be concentrated on an asymptotically vanishing fraction of variables. 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.

As a specific example, consider a cross section of n assets 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, $\sum_{i\in\mathcal{A}_l} \lambda_{ik}^2 = \sqrt{n}$ and $\sum_{i\notin\mathcal{A}_l} \lambda_{il}^2 = 0$, such that Assumption 2 holds for $\alpha_l = 0.5$. 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 a constant $d \in (0, 1]$ (which may depend on c), such that*

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

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

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

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

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

(b) $\frac{\Lambda_{.k}' e' F_{.l}}{n^{\frac{1}{2}\alpha_k} T^{\frac{1}{2}}} = O_p(1)$.

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)). 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).

Remark 1. *Let $r_1 + r_2 = r$, $\alpha_k > \tau$ for $k = 1, \ldots, r_1$ and $\alpha_k \leq \tau$ for $k = r_1 + 1, \ldots, r$ for some fixed value of $\tau \in [0, 1)$. In words: Let r_1 be the number of factors affecting proportionally more than n^τ variables, while the remaining factors are less pervasive. We can then rewrite the factor structure* (1) *as*

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

where the weakest r_2 factors are incorporated into the error term u. Effectively this is a factor model with r_1 factors, where $\psi_1(uu'/n)$ is no longer bounded. We can therefore think of Assumptions 1-4 as a generalization of standard factor models in two ways: they allow for the presence of weaker factors and they allow for stronger dependence in the error term.

By including more (weaker) factors, a practitioner can choose how much of the correlation among the observables she wishes to explicitly model. Note that one can generally always include additional factors, even if the corresponding eigenvalue is bounded. Throughout this paper, one can therefore think of r as an upper bound on the number of factors. However, this immediately raises the question of how many factors a practitioner should keep in the model. We model this choice of r_1 through the complexity parameter τ. A practitioner chooses a threshold $\tau \in [0,1]$ to indicate a lower bound on the strength of the factors she wishes to keep in the model. I discuss this choice in Sections 3.1-3.2.

Although I treat r as fixed, thus not allowing the number of factors to grow with the sample size, conceptually, my framework would allow for this. Allowing the number of factors to grow with the sample size is left as an interesting extension for future research.

All auxiliary lemmata for the proofs in the following sections are relegated to the Online Appendix.

3 Weak Asymptotics

I 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. I start with the following lemma:

Lemma 1. *Under Assumptions 1 and 2:*

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

Proof. If $k \leq r$:

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

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:*

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

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

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

where $\sigma_k(A)$ denotes the kth largest singular value of a matrix A. Therefore, with $A = F\Lambda'$, $B = e$ and $l = 1$, 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 I therefore conclude, using Lemma 1 and Assumption 3(e) respectively for the two eigenvalues on the RHS:

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

Similarly, again by Weyl's inequalities:

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

and I 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)). I 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.[5]

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 after the rth eigenvalue, 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 the tuning parameter τ (the lower bound on the pervasiveness of factors one wishes to keep in the model), I next consider the following two questions:

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

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

3.1 The Principal Component Estimator

I will begin 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 (2002a), Bai and Ng (2002), Bai (2003)). I obtain the following theorem:

Theorem 2. *Let $\hat{F}_k$ be defined as the standardized eigenvector corresponding to the kth largest eigenvalue of $\frac{XX'}{n}$. Then, 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. First define a matrix H as follows:

$$H = \Lambda'\Lambda \frac{F'\hat{F}}{T} \hat{D}_K^{-1},$$

where $\hat{D}_K$ is a diagonal matrix with the K largest eigenvalues of $\frac{X'X}{T}$ on the main diagonal. By Lemma 9: $H_{.k} = \iota_k + O_p(n^{\frac{1}{4}\alpha_1 - \frac{1}{2}\alpha_k}) + O_p(n^{\frac{1}{2}-\alpha_k})$.

[5]Note that we can replace Assumption 2 with the high level assumption $\sum_{i=1}^{n} \lambda_{ik}^2 \asymp n^{\alpha_k}$ and Theorem 1 still holds. 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.

Consequently, combining this with Lemma 10:

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

□

I note that this result may be of interest to a practitioner for two reasons. First, this establishes a lower bound in terms of factor strength for which we are able to prove consistency of the principal component estimator ($\alpha_k > \frac{1}{2}$). 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.[6] In cases in which factor estimates are used that correspond to weaker factors, Theorem 2 at least suggests to be cautious with respect to the standard errors of these estimates.

I also obtain a similar result for the factor loadings:

Theorem 3. *Let $\hat{\Lambda}' = \frac{\hat{F}'X}{T}$, with $\hat{F}$ defined as before. Then, 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.

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

where the last equality follows from Lemmata 7, 9 and 11 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, I obtain convergence of the principal component estimator for both the factors and the loadings as long as $\alpha_k > .5$. The following table provides an empirical test to assess the adequacy

[6] The exception is Onatski (2012), but his setup is quite distinct from the framework in this paper.

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 PC estimator for the kth factor (Bai (2003), Onatski (2012)). The exact DGP can be found in Online Appendix A.1. The numbers in Table 1 are in line with my 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, I observe that the estimation error seems to dominate the signal when the factor strength is below this threshold.

n	81	256	625	1296
$\lvert\mathcal{A}_2\rvert = n^{1/4}$	0.070	0.050	0.043	0.039
$\lvert\mathcal{A}_2\rvert = n^{1/2}$	0.114	0.098	0.086	0.078
$\lvert\mathcal{A}_2\rvert = n^{3/4}$	0.765	0.950	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.

3.2 Which Factors Matter?

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

To this end I next present two theoretical arguments why factors affecting proportionally more than $\sqrt{n}$ of the cross sections ($\tau = .5$) will be the natural target in many instances. One is derived from the Arbitrage Pricing Theory of Ross (1976) and a second argument builds on the works of Gabaix (2011) in the context of aggregate fluctuations in the economy. Note that for the two examples that follow I treat the factors as random in line with the literature in those fields.

3.2.1 Arbitrage Pricing Theory

I assume that the n-vector of demeaned asset returns $R_t - \mathbb{E}(R_t)$ for a given t follows a factor structure with potentially local factors as in the previous sections:

$$R_i - \mathbb{E}(R_i) = \underset{(1\times r)}{\lambda_i'} \underset{(r\times 1)}{F} + \underset{(1\times 1)}{e_i} = \underset{(1\times K)}{\lambda_i^{K\prime}} \underset{(K\times 1)}{F^K} + e_i^K, \tag{2}$$

treating the factors as random and the errors as uncorrelated with the factors. Equation (2) emphasizes again that, in the framework of this paper, we can always choose to move some of the weaker

factors into the error structure at the expense of more correlation in the error term. Denote the return of a portfolio by $R^p = \sum_i^n w_i R_i$, with $\sum_i^n w_i = 1$. I formalize the term "well-diversified" by imposing a bound on the sup-norm of the weights: $|w_i| \leq W_n \; \forall i$.[7] Following Green and Hollifield (1992), I say that exact APT pricing holds if the mean returns are in the span of the factor loadings and a constant vector:

$$E(R_j) = (1 - \sum_k^K \lambda_{jk}) E(R_0^*) + \sum_k^K \lambda_{jk} E(R_k^*),$$

where the portfolios $R_k^*, k = 0, \ldots, K^*$ are "factor-mimicking" portfolios. Their construction is detailed in Online Appendix C.3 and conditions for their existence are given in Huberman et al. (1987). Similarly, I define exact APT to hold in the limit, if, as n increases, there exist sequences of feasible factor-mimicking portfolios R_{nk}^*, such that for any fixed j

$$lim_{n\to\infty} E(R_j) - [(1 - \sum_k^K \lambda_{jk}) E(R_{n0}^*) + \sum_k^K \lambda_{jk} E(R_{nk}^*)] = 0.$$

Finally denote by ν_n the return on the global minimum variance portfolio when there are n assets and assume that the mean-variance frontier does not become vertical in the limit, such that there remains a meaningful trade-off between mean and variance.[8] I then obtain the following proposition:

Proposition 1. *Consider the sequence of efficient (minimum variance) portfolios for some mean return $\mu \neq lim_{n\to\infty} \nu_n$. If*

(i) $W_n = o(\frac{1}{n^\gamma})$, $\gamma > \frac{1}{2}$ *for every such portfolio, and*

(ii) $\lim_{n\to\infty} \sum_{i=1}^n \left| Cov(e_i, e_j) \right| = O(\sqrt{n})$,

then exact APT pricing holds in the limit with respect to the strongest K factors, where K is defined such that $\alpha_k > \gamma$ for $k = 1, 2, \ldots, K$ and $\alpha_k \leq \gamma$ for $k \geq K + 1$.

The proof can be found in Online Appendix C.3 and largely follows the proof of Theorem 3 in Green and Hollifield (1992).

Proposition 1 states that exact APT holds in the limit if the efficient portfolios are well diversified. Further, the number of factors that are priced depends directly on the degree of diversification

[7] While Chamberlain (1983) defines portfolio diversification through the ℓ_2-norm, the norm that proves tractable here is the sup-norm. Either definition formalizes the idea that the weights on individual assets get small as the universe of assets expands.

[8] This is the equivalent of the "absence of arbitrage" assumption in the Hilbert space setting of Chamberlain and Rothschild (1983).

of the portfolios on the efficient frontier. The better diversified these portfolios are, the smaller the number of factors that have a non-zero factor premium.

In particular, with $W_n = o(\frac{1}{\sqrt{n}})$, which yields diversification in the sense of Chamberlain and Rothschild (1983) and Chamberlain (1983), Proposition 1 establishes that exact APT pricing holds in the limit with respect to the r_1 factors affecting proportionally more than $\sqrt{n}$ of the assets (factors with $\alpha_k > .5$).

Proposition 1 holds under more general conditions than the approximate factor model of Chamberlain and Rothschild (1983). I do not require all eigenvalues of the error covariance matrix to be bounded, but explicitly allow for additional, weaker factors. Instead of ruling out the existence of such weaker factors, Proposition 1 establishes that they will not be priced.

3.2.2 Aggregate Fluctuations in the Economy

Consider a simple "Islands" economy with n firms as in Gabaix (2011). Firm i produces a quantity S_{it} of the consumption good. Instead of modeling firm-evel growth rates as unrelated, I model them as a combination of r mutually independent shocks that may affect several firms, on top of the idiosyncratic shocks. Firm i thus experiences a growth rate equal to

$$\frac{\Delta S_{i,t+1}}{S_{it}} = \frac{S_{i,t+1} - S_{it}}{S_{it}} = \lambda_i F_{t+1} + \sigma_i \varepsilon_{i,t+1}, \tag{3}$$

where σ_i is firm i's volatility, and the $\varepsilon_{i,t+1}$ are uncorrelated random variables with mean zero and variance 1. Firms' growth rates may be correlated through the presence of the first component. However, I do not impose the factors to be pervasive and likely $\lambda_{ik} = 0$ for most firm-factor combinations. Intuitively, these factors can correspond to economy wide shocks but also sector shocks or the introduction of policies only affecting a subset of firms, including shocks that affect as few as two firms. Thus (3) is quite general.

In this stylized model, GDP growth is given by:

$$\begin{aligned}\frac{\Delta Y_{t+1}}{Y_t} &= \frac{1}{Y_t}\sum_{i=1}^{n} \Delta S_{i,t+1} = \sum_{i=1}^{n} \frac{S_{it}}{Y_t}[\lambda_i F_{t+1} + \varepsilon_{i,t+1}] \\ &= \sum_{i=1}^{n} \frac{S_{it}}{Y_t}\lambda_i F_{t+1} + \sum_{i=1}^{n} \frac{S_{it}}{Y_t}\varepsilon_{i,t+1}.\end{aligned}$$

It follows that the variance of GDP growth at time $(t+1)$ conditional on time t information is

equal to

$$Var_t\left(\sum_{i=1}^{n}\frac{S_{it}}{Y_t}\lambda_i F_{t+1}+\sum_{i=1}^{n}\frac{S_{it}}{Y_t}\varepsilon_{i,t+1}\right)=Var_t\left(\sum_{i=1}^{n}\frac{S_{it}}{Y_t}\lambda_i F_{t+1}\right)+Var_t\left(\sum_{i=1}^{n}\frac{S_{it}}{Y_t}\varepsilon_{i,t+1}\right)$$
$$=Var_t\left(\sum_{i=1}^{n}\frac{S_{it}}{Y_t}\sum_{k=1}^{r}\lambda_{ik}F_{k,t+1}\right)+\sum_{i=1}^{n}\left(\frac{S_{it}}{Y_t}\right)^2\sigma_i^2.$$

For ease of notation, consider firms of equal size ($S_{it}=\frac{Y_t}{n}$) and identical standard deviation ($\sigma_i=\sigma$), and normalize the factors such that $Var(F_{kt})=1$. Further assume that, for a given k, the factor loadings are 1 on a subset of size $|\mathcal{A}_k|\asymp n^{\alpha_k}$ and zero everywhere else[9]. Then:

$$Var_t\left(\frac{\Delta Y_{t+1}}{Y_t}\right)=\sum_{k=1}^{r}\left(\sum_{i\in\mathcal{A}_k}\frac{1}{n}\right)^2+\sum_{i=1}^{n}\frac{1}{n^2}\sigma^2$$
$$\asymp\sum_{k=1}^{r}n^{2\alpha_k-2}+\frac{\sigma}{n}. \tag{4}$$

It immediately follows that, absent any factors ($r=0$), $\sigma_{GDP}=\sqrt{Var_t(\frac{\Delta Y_{t+1}}{Y_t})}=\frac{\sigma}{\sqrt{n}}$, which is the reason macroeconomists often appeal to aggregate shocks, since idiosyncratic fluctuations disappear in the aggregate at rate $\sqrt{n}$. Next, consider an economy with r shocks, where r_1 is the number of factors with $\alpha_k>.5$:

$$Var_t\left(\frac{\Delta Y_{t+1}}{Y_t}\right)\asymp\sum_{k=1}^{r_1}n^{2\alpha_k-2}+\sum_{k=r_1+1}^{r}n^{2\alpha_k-2}+\frac{\sigma}{n}$$
$$=\sum_{k=1}^{r_1}n^{2\alpha_k-2}+O_p(\frac{1}{n}).$$

Equation (4) establishes that the important shocks are those with $\alpha_k>\frac{1}{2}$ and that the standard rate of convergence breaks down whenever shocks exist that affect more than $\sqrt{n}$ firms.

This is in line with the granularity conditions derived in Gabaix (2011), who considers heterogeneous firm sizes that may grow with n. Intuitively, with the growth rate of the economy given by the sum of both the idiosyncratic and factor shocks in my context, we can think of the sector shocks as additional but larger firms. Then the economy consists of $n+r$ components (with $r<<n$). 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 exactly to the limit on sector size stated above.

[9]Defining the loadings instead in a more general way as in Assumption 2 does not alter any conclusions.

The key implication for the purposes of this paper is that, in order to understand the origins of fluctuations, the important shocks are precisely those that affect proportionally more than $\sqrt{n}$ firms.

4 Determining the Number of Factors

In many applications, the number of factors is of interest in itself, as illustrated in the last section. For example, we may be interested in the number of fundamental shocks in the economy that contribute to the surprisingly large standard deviation (more than 8 percentage points) of the Federal Reserve Board's Index of Industrial Production and Capacity Utilization(Foerster et al. (2011)). In finance, this number can be interpreted as the number of sources of nondiversifiable risk. In other cases the number of factors must be known to implement various estimation and forecasting procedures. For example, in factor-augmented VAR models, impulse responses based on an incorrect number of factors may be misleading and result in bad policy suggestions (Bernanke et al. (2005), Giannone et al. (2006)). Onatski (2015) discusses the consequences of a misspecified number of factors for the squared error of the estimated common component. The implications of the number of factors on the R^2 of the common component in explaining movements in individual series are discussed in Section 6.

The target of estimation in this section will be defined by a complexity parameter τ such that r_1 is the number of factors that affect proportionally more than n^τ cross sections. For the reasons outlined in the previous sections, the number of factors r_1 that a practitioner is usually interested in will be such that $\alpha_k > .5$ for $k = 1, \ldots, r_1$. This corresponds to complexity parameter $\tau = .5$.

Estimating the number of factors in factor models has been a subject of interest for some time now (e.g. Bai and Ng (2002), Onatski (2010), Ahn and Horenstein (2013)). To the best of my knowledge, all existing estimators are derived from the distribution of eigenvalues of the matrix $\frac{X'X}{T}$ (or equivalently the singular values of X). For example, the information criteria introduced in Bai and Ng (2002) effectively count the number of eigenvalues above a certain threshold, Ahn and Horenstein (2013) consider the ratio of subsequent eigenvalues, and Onatski (2010) uses the difference between subsequent eigenvalues to determine the number of factors.

While the first two methods explicitly require strong factors, "weak" factors are allowed for in Onatski (2010). In the framework of Onatski (2010) some of the "large" eigenvalues do not necessarily diverge to infinity. Essentially, Onatski's proposed estimator counts the number of eigenvalues that are too large to come from the idiosyncratic errors. While the work of Onatski provides an insightful and novel framework allowing for weak factors, the required assumptions on the error term are quite restrictive. Further, estimating the number of factors from the empirical distribution of eigenvalues still rests on a separability between the two groups of eigenvalues.

In conclusion, all existing methods to estimate the number of factors can be interpreted as formalizing the heuristic approach based on a visual inspection of the scree plot, which dates back to Cattell (1966). However, Theorem 1 established that, in the presence of local factors, the eigenvalues can no longer easily be separated into two groups (large eigenvalues representing factor-related variation and small eigenvalues representing idiosyncratic variation).

The novel insight here is that in scenarios with local factors, the eigenvectors of the matrix $\frac{X'X}{T}$ carry valuable information, which is discarded when solely considering the eigenvalue distribution. Intuitively, the hope is to exploit this additional information to "tilt the eigenvalues" in order to reintroduce a gap between those eigenvalues corresponding to factors with $\alpha_k > \tau$ and those below this threshold.

In order to incorporate the eigenvectors into the inference on the number of factors, consider the following quantity:

$$\hat{T}^u_{zk} \equiv \psi_k(\frac{X'X}{T})\hat{S}^u_{zk} \equiv \psi_k(\frac{X'X}{T})\left(\frac{1}{z}\sum_i^z \frac{\hat{\lambda}^2_{ik}}{\sqrt{\frac{1}{n}\sum_{i=1}^n \hat{\lambda}^2_{ik}}}\right)^u, \tag{5}$$

where, with slight abuse of notation, $\hat{\lambda}^2_{ik}$ are the squared entries of the kth eigenvector sorted in decreasing order, such that I take a partial sum over the z largest entries in the second part. One can think of $\hat{T}^u_{zk}$ as combining the kth eigenvalue of the matrix $\frac{X'X}{T}$ (the first component) with a measure of how concentrated the corresponding eigenvector is on a subset of the observables (the second component). A factor that is highly influential on a subset of covariates but unrelated to the majority of outcomes will be difficult to detect using solely the eigenvalue of the $\frac{X'X}{T}$. However, the second part of (5) will scale this eigenvalue up to enable a practitioner to detect its presence. The power u plays the role of a tuning parameter that governs the relative weight on the eigenvalue versus the eigenvector. With $u = 0$ the second part vanishes and $\hat{T}^u_{zk}$ reduces to just the eigenvalue. On the other hand, with $u = 2$, $\hat{T}^2_{zk}$ only depends on the largest z entries in the kth eigenvector.

Figure 2 illustrates the behavior of $\hat{T}^u_{zk}$ as a function of factor strength and the tuning parameter u. Moving from right to left in Figure 2, the number of covariates affected by the factor increases, with n^{α_k} outcomes influenced by the factor. On the right edge, only a fixed number of covariates is affected ($\alpha_k = 0$), while the left edge of the figure corresponds to a scenario in which the factor is relevant for all covariates ($\alpha_k = 1$). Moving from front to back, the value of the tuning parameter u varies from 0 to 2.

The front edge of the plane, with $u = 0$, simply corresponds to the corresponding eigenvalue $\psi_k(\frac{X'X}{T})$. With only a fixed number of covariates affected by F_k, this eigenvalue remains bounded (front right corner). As the factor affects more covariates, the eigenvalue begins to diverge at an increasing rate (see Theorem 1). One conventional estimator for the number of factors would count

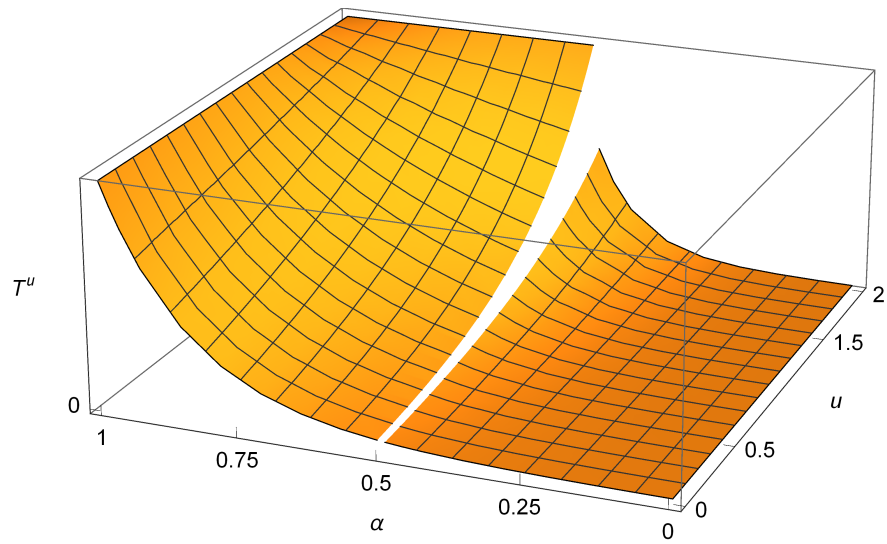


Figure 2: Theoretical divergence rate of $\hat{T}^u_{zk}$ as a function of both factor strength (α_k) and tuning parameter u for $z = \sqrt{n}\sqrt{loglog(n)}$ with $n = 500$. Note the steep region at $\alpha_k = \sqrt{n}$ in the back of the picture.

the number of eigenvalues above a chosen threshold K_n. As argued earlier, an often appealing choice for this threshold is $K_n \asymp \sqrt{n}$, such that we aim to keep all factors in the model with $\alpha_k > 0.5$. However, the relevant curve is rather flat around this cutoff. This suggests that such an estimator would be very sensitive to the choice of the threshold in finite samples. In fact, this is well known in the literature (e.g., Alessi et al. (2010)). Note that any estimator based on only the eigenvalues will share this problem.

In contrast, on the back edge of Figure 2 the behavior of $\hat{T}^2_{zk}$ is depicted.[10] Exploiting the information in the eigenvectors (by setting $u > 0$) induces a steep region in the statistic around the desired minimum factor strength, thereby helping to discriminate between factors above and below this threshold. The position of the steep increase can be chosen by a practitioner through the second tuning parameter z.

To derive this result formally, I start by defining the following class of (infeasible) quantities T^u_{zk}. For $u \in [0, 2]$:

$$T^u_{zk} = \psi_k\left(\frac{\Lambda F' F \Lambda'}{T}\right) S^u_{zk} = \psi_k\left(\frac{\Lambda F' F \Lambda'}{T}\right)\left(\frac{1}{z}\sum_i^z \frac{\lambda^2_{ik}}{\sqrt{\frac{1}{n}\sum_{i=1}^n \lambda^2_{ik}}}\right)^u,$$

where λ^2_{ik} are sorted in decreasing order. Note that for $u = 0$, $T^0_{zk} = \psi_k(\frac{\Lambda F' F \Lambda'}{T})$.

The behavior of T^u_{zk} is summarized in the following lemma:

[10]Setting $u > 2$ is possible and would result in a quantity that is even more peaked around the threshold parameter τ. The equivalent of Figure 2 extending up to $u = 3$ is depicted in Online Appendix B. With $u > 2$, T^u_{zk} is no longer monotonically increasing in α_k, the measure of factor strength. I will therefore restrict my analysis to $u \in [0, 2]$ in the remainder of this paper.

Lemma 2. *Under Assumptions 1-2, 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 $k \leq r$, with $u \in [0,2]$:*

(a) If $\alpha_k > \tau$: $T^u_{zk} \asymp n^{(1-\frac{1}{2}u)\alpha_k + \frac{1}{2}u}$

(b) If $\alpha_k \leq \tau$: $T^u_{zk} \asymp n^{(1+\frac{1}{2}u)\alpha_k + (\frac{1}{2}-\tau)u} g(n)^{-u}$.

Further, for $k = r+1, \ldots, r_{max}$: $T^u_{zk} = 0$.

Proof. Using Assumption 1 I can rewrite T^u_{zk} as follows:

$$
\begin{aligned}
T^u_{zk} &= \psi_k\left(\frac{\Lambda F' F \Lambda'}{T}\right) S^u_{zk} = \psi_k\left(\frac{\Lambda F' F \Lambda'}{T}\right)\left(\frac{1}{z}\sum_i^z \frac{\lambda^2_{ik}}{\sqrt{\frac{1}{n}\sum_{i=1}^n \lambda^2_{ik}}}\right)^u \\
&= \psi_k\big(\Lambda'\Lambda\big)\big(\sum_{i=1}^n \lambda^2_{ik}\big)^{-\frac{1}{2}u}\big(\frac{n^{\frac{1}{2}}}{z}\sum_i^z \lambda^2_{ik}\big)^u \\
&= \psi_k\big(\Lambda'\Lambda\big)^{1-\frac{1}{2}u} n^{\frac{1}{2}u}\big(\frac{1}{z}\sum_i^z \lambda^2_{ik}\big)^u. \qquad (6)
\end{aligned}
$$

First consider scenario (a). With $\alpha_k > \tau$, the last part of (6) is simply an average of the square of the z largest loadings. Combining Assumption 2(a) with the fact that $|\lambda_{ik}| < C \ \forall i$, we immediately have $T^u_{zk} \asymp n^{(1-\frac{1}{2}u)\alpha_k} n^{\frac{1}{2}u}$.

Next, for part (b), let $\alpha_k \leq \tau$: There are only $|\mathcal{A}_k| \asymp n^{\alpha_k}$ "large" loadings in the sum of equation (6) and Assumption 2 implies that

$$
\frac{1}{z}\sum_i^z \lambda^2_{ik} = \frac{1}{z}\sum_{i\in\mathcal{A}_k} \lambda^2_{ik} + \frac{1}{z}\sum_{i\notin\mathcal{A}_k} \lambda^2_{ik} \asymp \frac{n^{\alpha_k - \tau}}{g(n)},
$$

and it follows that $T^u_{zk} \asymp n^{(1-\frac{1}{2}u)\alpha_k} n^{\frac{1}{2}u} \frac{n^{(\alpha_k-\tau)u}}{g(n)^u}$.

For $k > r$, $\lambda_{ik} = 0 \ \forall i$, and this completes the proof. □

Since both Λ and F are unobserved, T^u_{zk} is infeasible to compute in practice. I will therefore use the feasible alternative to T^u_{zk}, introduced in (5) and repeated below:

$$
\hat{T}^u_{zk} \equiv \psi_k(\frac{X'X}{T})\hat{S}^u_{zk} \equiv \psi_k(\frac{X'X}{T})\left(\frac{1}{z}\sum_i^z \frac{\hat{\lambda}^2_{ik}}{\sqrt{\frac{1}{n}\sum_{i=1}^n \hat{\lambda}^2_{ik}}}\right)^u. \qquad (7)
$$

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 $\alpha_k > \tau$*:*

$$\hat{T}^u_{zk} \asymp n^{\frac{1}{2}u+(1-\frac{1}{2}u)\alpha_k}$$

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

$$\hat{T}^u_{zk} \asymp n^{(1+\frac{1}{2}u)\alpha_k+(\frac{1}{2}-\tau)u} g(n)^{-u}$$

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

$$\hat{T}^u_{zk} \asymp n^{\alpha_k} \qquad \textit{for } u = 0$$
$$\hat{T}^u_{zk} = O_p(n^{(1+\frac{1}{2}u)\alpha_k+(\frac{1}{2}-\tau)u} g(n)^{-u}) \qquad \textit{for } u > 0$$

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

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

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

As $\hat{T}^u_{zk}$ is the key quantity in what follows, I also present a corollary that simplifies the notation and covers most cases before I further discuss Theorem 4 and its implications. I argued in Section 3 that the important factors are usually those that affect proportionally more than $\sqrt{n}$ of the outcomes, so that $\tau = .5$ will often be the natural choice, and I will use this threshold going forward, omitting the corresponding subscript z and writing simply $\hat{T}^u_k$ 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 $\alpha_k > \frac{1}{2}$*:*

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

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

$$\hat{T}^u_k \asymp n^{\alpha_k} \qquad \textit{for } u = 0$$
$$\hat{T}^u_k = O_p\left(n^{(1+\frac{1}{2}u)\alpha_k} g(n)^{-u}\right) \qquad \textit{for } u > 0$$

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

$$\hat{T}_k^u = O_p(g(n)^{-u}).$$

The theoretical rates of Corollary 1 were illustrated graphically in Figure 2. To gain intuition, suppose $\alpha_k = 1$ for $k = 1, 2, ..., r$, which corresponds to the standard setup in the literature. Then $\hat{T}_{zk}^u \asymp n$ for $k = 1, 2, \ldots, r$, regardless of the choice of u (see the left edge of Figure 2). For $k > r$, $\hat{T}_{zk}^0 = O_p(1)$ and $\hat{T}_{zk}^u = O_p(g(n)^{-u})$ if $u > 0$. This means that under the standard setup with only strong factors, the behavior of $\hat{T}_{zk}^u$ is invariant to the choice of u (in terms of its rate of divergence) up to the very slowly diverging sequence $g(n)$. In contrast, for all local factors with $\alpha_k \in (0, 1)$ the divergence properties of $\hat{T}_{zk}^u$ depend on the power u. For example, let $\alpha_k \in (.5, 1]$. Then, $\hat{T}_k^0 \asymp n^{\alpha_k}$, $\hat{T}_k^1 \asymp n^{\frac{1}{2}+\frac{1}{2}\alpha_k}$ and $\hat{T}_k^2 \asymp n$. $\hat{T}_k^2$ has the appealing property that it does not depend on the factor strength for $\alpha_k \in (.5, 1]$. Thus it allows us to distinguish factors above the threshold $\tau = .5$ from idiosyncratic noise at the same rate as strong factors.

An ideal statistic would be discontinuous with a large jump at a user chosen threshold, thereby making it straightforward to identify the number of factors above this threshold. While $\hat{T}_k^2$ is discontinuous at τ, the discontinuity is small. However, we do obtain a steeper slope leading up to the threshold.[11]

4.1 Proposed Estimators

In this subsection, I derive consistency for several estimators for the number of factors r_1 analogous to those that have been proposed in the literature, but based on $\hat{T}_k^u$ instead of the eigenvalues $\hat{T}_k^0$. In particular, I focus on the case $u = 2$ and 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))

[11] The introduction of the slowly diverging sequence $g(n)$ in the definition of z in Lemma 2 and Theorem 4 will be useful in the construction of estimators based on $\hat{T}_k^u$. Specifically, this additional term is responsible for the gap at $\alpha = .5$ and allows us to perfectly separate any factor F_k with factor strength $\alpha_k = .5 + \epsilon$ from a factor F_l with $\alpha_l = .5$ for any $\epsilon > 0$. This relies on the fact that $g(n) = o(n^\epsilon)$ for any $\epsilon > 0$. For most empirically relevant sample sizes, this will only be a good approximation when $g(n)$ diverges extremely slowly. Without the $g(n)$ term, the results will still hold generically, except at the singular point $\alpha_k = .5$, where threshold and divergence rate coincide.

4.1.1 Thresholding Estimators

I start by considering the estimators introduced in Bai and Ng (2002). I will denote by PC the number k that minimizes the criterion function

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

where

$$V(k) = \min_{\Lambda,F_k}(NT)^{-1}\sum_{i=1}^{n}\sum_{t=1}^{T}(X_{ti} - \lambda_i^k F_t^k)^2 = \frac{1}{nT}\sum_{j=k+1}^{n}\psi_j(X'X), \tag{8}$$

and $\hat{\sigma}^2$ is an estimator of the unconditional variance of the idiosyncratic error. The second equality in (8) follows from the fact that $V(k)$ is the best approximation of X of rank k. We can alternatively represent $\hat{\sigma}^2$ as $V(r_{max}) = \frac{1}{n}\sum_{j=r_{max}}^{n}\psi_j\left(\frac{X'X}{T}\right)$.[12] Therefore, $BN(k)$ is a function of only the empirical distribution of the eigenvalues and will be equivalent to a thresholding procedure for the aforementioned. Unifying notation in terms of the eigenvalues and using $c = n/T$, this can be seen by rewriting their estimator as:

$$\begin{aligned} PC &= \arg\min_k V(k) + k\hat{\sigma}^2\left(\frac{n+T}{nT}\right)log\left(\frac{nT}{n+T}\right) \\ &= \arg\min_k \frac{1}{n}\sum_{l=k+1}^{n}\psi_l\left(\frac{X'X}{T}\right) + k\hat{\sigma}^2\left(\frac{c+1}{n}\right)log\left(\frac{n}{c+1}\right) \\ &= \max k \qquad \text{s.t. } \psi_k\left(\frac{X'X}{T}\right) > \hat{\sigma}^2(c+1)\log(\frac{n}{c+1}).^{13} \end{aligned} \tag{9}$$

Instead of deriving my estimator solely from the empirical distribution of the eigenvalues, I will consider the following criterion for a fixed constant Q:

$$TC = \max k \qquad \text{s.t. } \hat{T}_k^2 > Q\frac{n}{h(n)}, \tag{10}$$

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, TC is a consistent estimator for the number of factors r_1 such that $\alpha_k > 0.5$ for $k = 1, \ldots, r_1$ and $\alpha_k \leq 0.5$ for $k > r_1$.*

[12] Bai and Ng (2002) consider a total of 6 estimators that differ slightly in their penalty term that is added to $V(k)$ and include a version in logarithms. However, their performances are similar to the ones considered here, and the corresponding results are therefore omitted.

[13] Similarly Kapetanios (2004) suggests simply using a cutoff value $b = (1 + \sqrt{n/T})^2 + 1$ and estimating the number of factors as the number of empirical eigenvalues above this threshold.

Proof. I first show that $\lim_{n\to\infty} P(\hat{T}_k^2 > c\frac{n}{h(n)}) = 1$ for $k = 1, \ldots, r_1$. In this case, $\alpha_k > 0.5$. By Theorem 4, $\hat{T}_k^2 \asymp n$. Thus, $n = O_p(\hat{T}_k^2)$. Combining this with $\frac{1}{h(n)} = o_p(1)$ I obtain $\frac{n}{h(n)} = o_p(\hat{T}_k^2)$ and thus

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

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

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

Next, consider the case $\alpha_k \leq 0.5$. Then $\hat{T}_k^2 = O_p(\frac{n^{2\alpha_k}}{g(n)^2})$ by Theorem 4. But $O_p(n^{2\alpha_k}/g(n)^2) = O_p(n/g(n)^2) = o_p(n/h(n))$ by the definition of $h(n)$ and thus, for any $\varepsilon > 0$,

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

Letting $\varepsilon = 1$, this gives

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

completing the proof. □

In practice I propose to incorporate an estimator of the variance into the model, letting $h(n) = \frac{g(n)}{Q_2\hat{\sigma}^2}$ such that TC becomes

$$TC = \max k \qquad \text{s.t. } \hat{T}_k^2 > Q_1\hat{\sigma}^2\frac{n}{g(n)}$$

for some fixed constant Q_1. This is justified because, by Theorem 1

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{j=r_{max}}^{n} \psi_j\left(\frac{X'X}{T}\right) \leq \psi_{r_{max}}\left(\frac{X'X}{T}\right) \leq C$$

and, similarly

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{j=r_{max}}^{n} \psi_j\left(\frac{X'X}{T}\right) \geq \frac{1}{n}\sum_{j=r_{max}}^{[dn]} \psi_j\left(\frac{X'X}{T}\right) \geq c_1\psi_{[dn]}\left(\frac{X'X}{T}\right) \geq c,$$

where the last inequality uses Weyl's inequalities in a similar way to Theorem 1 to establish that $\psi_{[dn]}\left(\frac{X'X}{T}\right)$ is bounded away from zero, thereby guaranteeing that $\hat{\sigma}^2$ is bounded both above and below. Therefore $h(n) = Qg(n)$, where Q is a finite, positive constant and $h(n)$ fulfills the conditions stated below the definition of TC in (10).

While TC is therefore a consistent estimate for r_1, I next derive the properties of the existing estimator PC in a setting with local factors. The implicit target of estimation using the PC criterion of Bai and Ng (2002) will be different than the cutoff argued for in this paper. In fact, (9) suggests that the PC criterion will estimate the number of factors affecting more than $log(n)$ outcomes, which is confirmed in the following corollary:

Corollary 2. *Under Assumptions 1-4 PC 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 PC will not be a consistent estimator for r_1, unless there exists no factor k with $0 < \alpha_k \leq 0.5$, in which case r^* and r_1 coincide. However, we can also consider an analogous estimator to PC designed to estimate the number of factors with $\alpha_k > .5$:

$$PC_{\sqrt{n}} = \max k \qquad \text{s.t. } \psi_k\left(\frac{X'X}{T}\right) > \hat{\sigma}^2(c+1)\sqrt{\frac{n}{c+1}}g(n). \tag{11}$$

It is then easy to show that:

Corollary 3. *Under Assumptions 1-4 $PC_{\sqrt{n}}$ is a consistent estimator for the number of factors r_1 such that $\alpha_k > 0.5$ for $k = 1, \ldots, r_1$ and $\alpha_k \leq 0.5$ for $k > r_1$.*

Given the equivalence established in (9), 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 $PC_{\sqrt{n}}$ and TC are consistent estimators for r_1, recall the discussion surrounding Figure 2. Based on the steeper slope of $\hat{T}_k^2$ around the chosen threshold ($\sqrt{n}$), TC 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 $\hat{r} = ED$, the first number at which

this difference becomes larger than some constant threshold Q:[14]

$$ED = \max\{k \leq r_{max} : \psi_k(\frac{X'X}{T}) - \psi_{k+1}(\frac{X'X}{T}) \geq Q\} = \max\{k \leq r_{max} : \hat{T}_k^0 - \hat{T}_{k+1}^0 \geq Q\}.$$

Of course, this method also has an analogue when using $\hat{T}_k^u$ for $u > 0$. 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. I therefore focus on the case $\tau = .5$ as before and define

$$TD = \max\{k \leq r_{max} : \hat{T}_k^2 - \hat{T}_{k+1}^2 \geq \frac{n}{h(n)}\},$$

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, TD is a consistent estimator for the number of factors r_1 such that $\alpha_k > 0.5$ for $k = 1, \ldots, r_1$ and $\alpha_k \leq 0.5$ for $k > r_1$.*

Proof. First note that, because $\hat{T}_k^2 = O_p(n^{2\alpha_k}/g(n)^2) = o_p(\frac{n}{h(n)})$ for any k with $\alpha_k \leq 0.5$, $(\hat{T}_k^2 - \hat{T}_{k+1}^2) = o_p(\frac{n}{h(n)})$ for $k > r_1$.

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

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

□

4.1.3 Ratio Estimators

The most recent 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), defined as

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

[14] 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.

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

Because the ratio estimator explicitly relies on a large gap in the eigenvalue distribution, I 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 $k = 1, \ldots, r$.

On the other hand, the ratio estimator has the significant advantage that it is less dependent on any tuning parameter. It also tends to perform well in finite samples. In line with the ER estimator above I suggest a similar estimator based on the quantity $\hat{T}_k^2$:

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

Theorem 7. *Under Assumptions 1-5, TR is a consistent estimator for the number of factors r_1 such that $\alpha_k > 0.5$ for $k = 1, \ldots, r_1$ and $\alpha_k \leq 0.5$ for $k > r_1$.*

Proof. First consider $k = r_1 + 1, \ldots, r_{max}$. By Theorem 4, $\hat{T}_k^2 = O_p(\frac{1}{g(n)^2})$ and thus for every $c_1 > 0$, $lim_{n\to\infty} P\left(\hat{T}_k^2 \geq c_1\right) = 0$. Further, by Lemma 13 there exists a constant $c_2 > 0$, such that $lim_{n\to\infty} P\left(n\hat{T}_{k^*}^2 \geq c_2\right) = 1$ for $k^* = r_1 + 1, \ldots, r_{max}$. Then, for any $k = r_1 + 1, \ldots, r_{max}$, any finite $c > 0$, setting $c_1 = c * c_2$ yields

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

Next, consider $k = r_1$. By Assumption 5 $\alpha_k > 0.5$ and there exists a finite $q_1 > 0$ such that $lim_{n\to\infty} P\left(\hat{T}_{r_1}^2 > q_1 n\right) = 1$. Using Assumption 5 again, $\hat{T}_{r_1+1}^2 = O_p(\frac{1}{g(n)^2})$ and thus for every

$q_2 > 0$, $P\left(\hat{T}^2_{r_1+1} \geq q_2\right) = 0$. Then, for any $q > 0$ and setting $q_2 = q_1/q$:

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

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

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

□

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

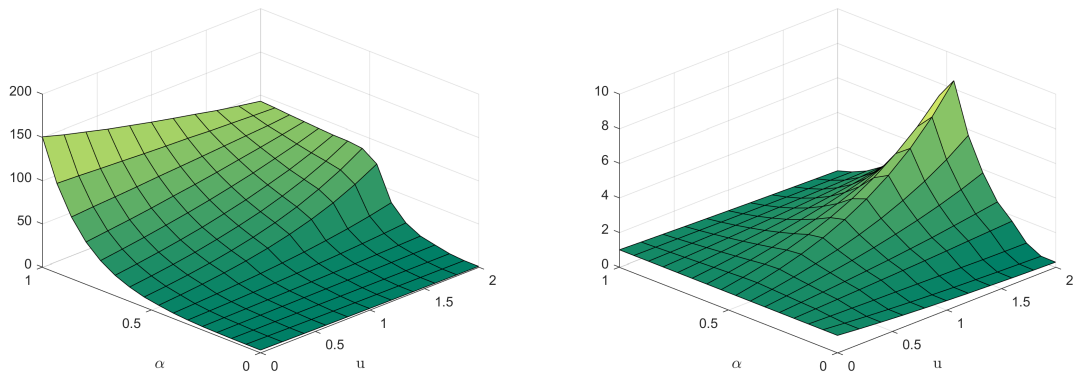
The proof largely follows the same arguments as those in the proof of Theorem 7 and is therefore relegated to Online Appendix C.2.

It is perhaps surprising that both estimators require an equal degree of separation. One might expect that the weak and strong factors need to be less well separated to obtain a consistent estimator for higher values of u. To ensure this theoretically, a lower bound on $\hat{T}^2_k$ for $k > r_1$, which in turn depends on the partial sum in S^2, would be needed. I leave this for future research. On an intuitive level however, the accompanied increase in slope around the targeted factor strength when using $\hat{T}^u_k$ for some $u > 0$ should again improve the performance of this estimator.

5 Simulations

I next present simulation evidence to assess the adequacy of the asymptotic approximations to the finite sample results. In what follows, fix $g(n) = 0.7\sqrt{loglog(n)}$.[15]

[15]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.



(a) Estimated $\hat{T}^u_{z1}$ **(b)** Estimated $\hat{S}^u_{z1}$

Figure 3: Empirical behavior of key quantities in simple DGP. Depicted are $\hat{T}^u_{z1}$ and $\hat{S}^u_{z1}$ as a function of both factor strength α and tuning parameter u. 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.

I start by depicting the empirical analog to Figure 2 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 2, I then vary α from 0 to 1 and u from 0 to 2. Figure 3a depicts the average value of $\hat{T}^u_1$ across 500 simulations. Note the close resemblance in shape to Figure 2. As u increases, a steep increase in $\hat{T}^u_1$ emerges around $\tau = .5$. I take Figure 3a as an encouraging sign that the finite sample behavior of $\hat{T}^u$ is well-approximated by the asymptotic theory of Section 3 at least in this simple setting.

I also emphasize that of the two quantities depicted on the horizontal axes α is unknown to a practitioner, while u is a tuning parameter that can be varied. Thus, for a given dataset, u can be varied as an exploratory tool. If F_k is a local factor in the sense of this paper, the divergence rate of $\hat{T}^u_k$ changes as u increases. Although not explicitly in my model, it is clear that the same does not hold if F_k is a weak but global factor with a small effect on all outcomes. The change in shape associated with an increase in the tuning parameter u is therefore indicative of the underlying structure and a practitioner might be interested in the behavior of $\hat{T}^u_z$ when u increases. Since $\hat{S}^u_k = \hat{T}^u_k/\hat{T}^0_k$, this amounts to looking at $\hat{S}^u_k$ (the "peakedness" of the eigenvector) directly. With $\tau = \frac{1}{2}$ and $u = 2$ for simplicity, $\hat{S}^2_k$ behaves as follows:

$$\hat{S}^2_k = \hat{T}^2_k/\hat{T}^0_k \asymp \begin{cases} n^{1-\alpha_k} & \text{for } \alpha_k \geq \frac{1}{2} \\ O_p(n^{\alpha_k}/g(n)^2) & \text{for } \alpha_k < \frac{1}{2}. \end{cases}$$

For the simple DGP introduced above, $\hat{S}_1^u$ is depicted in Figure 3b. It suggests that the finite sample behavior of $\hat{S}_k^2$ is also well-approximated by the asymptotic rates above. While the eigenvalue is monotonically increasing in factor strength, $\hat{S}_k^2$ takes its highest value at $\alpha_k = .5$.

I next consider more realistic settings as they might be observed in practice. I consider a 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[16] and will be varied later on. The variables exhibit a factor structure with 6 independent 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. There are three additional factors G_1, G_2, G_3 also drawn from a standard normal, which I consider too weak to be pervasive. 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. For the idiosyncratic part I allow for both cross-sectional and intertemporal correlation. I model the errors as

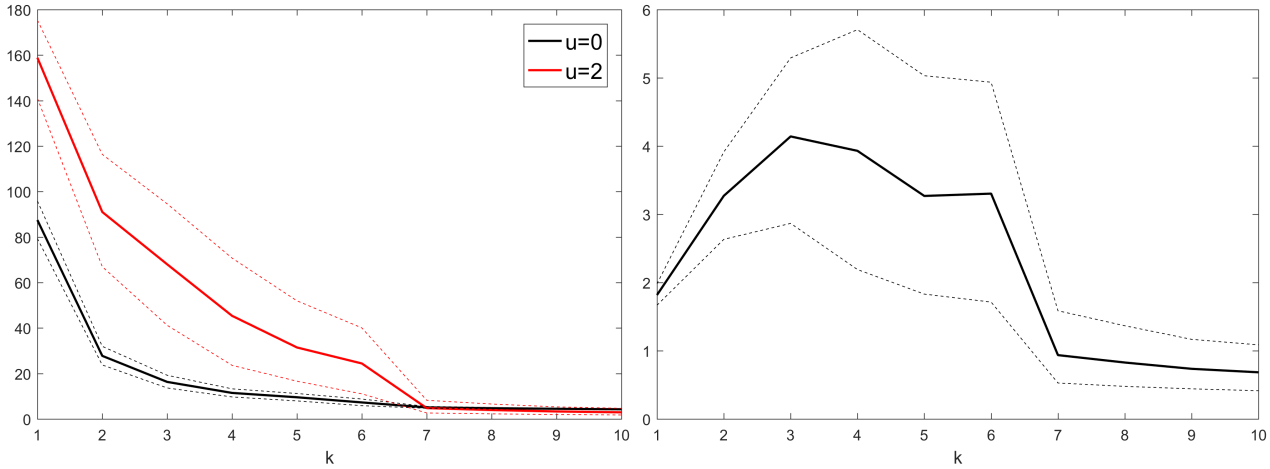
$$\begin{aligned} e_{ti} &= \rho e_{t-1,i} + (1-\rho^2)^{1/2} v_{ti} \\ v_{ti} &= \beta v_{t,i-1} + (1-\beta^2)^{1/2} u_{ti}, \quad u_{ti} \overset{i.i.d.}{\sim} N(0,1), \end{aligned}$$

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 I set $\theta = 1.5$ in my baseline model. The factor structure and signal-to-noise ratio of the 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 $\hat{T}_k^u, u \in \{0, 2\}$ and $\hat{S}_k^2$ in simulated data from the baseline DGP. It is constructed from 1000 realizations, with the bold line depicting the average and the dashed lines depicting the pointwise 5th and 95th percentile of the respective quantities. Note that $\hat{T}_k^0$ and $\hat{T}_k^2$ in Figure 4a correspond to the front and back edge of Figure 2 respectively. I observe an encouraging resemblance with a jump at $\hat{r} = 6$ when $u = 2$. This is due to the behavior of S_k^u, depicted in Figure 4b. The eigenvectors corresponding to more local factors are indeed more concentrated on a subset of its entries.

I next depict the ratios and differences of subsequent values of $\hat{T}_k^u$ in Figure 5. Consider an

[16] For example Bai and Ng (2002) consider sample sizes in both dimensions between 40 and 8000.



(a) Empirical distribution of $\hat{T}_k^u$.

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

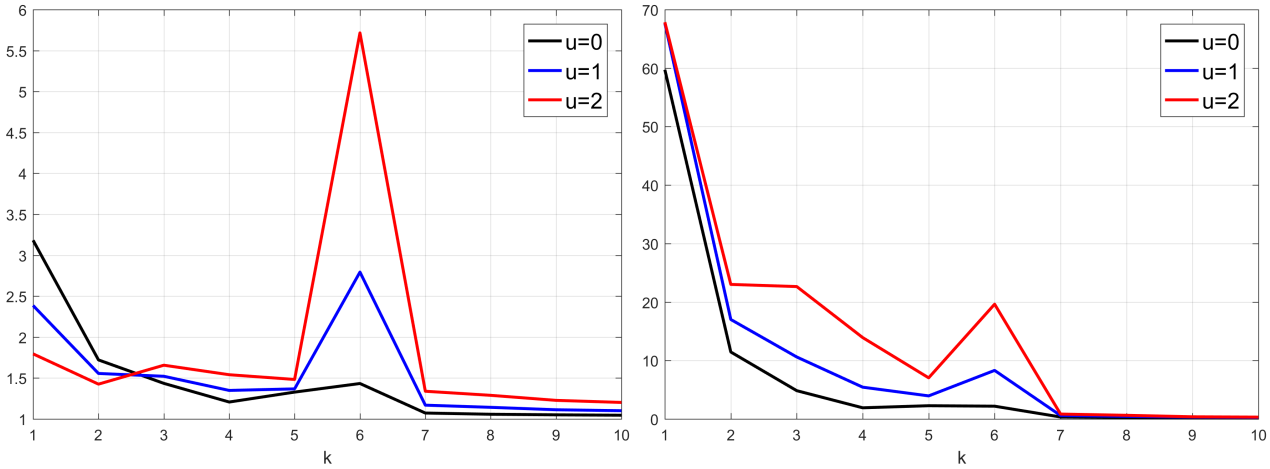
Figure 4: Illustration of key quantities in a simulated dataset. Solid line depicts average across 1000 realizations. Dashed lines illustrate 5th and 95th quantile. Data generated by baseline DGP, with $(n, T) = (300, 500)$, $(\rho, \beta) = (0.3, 0.1)$, $r_1 = 6$, $\theta = 1.5$. x-axis corresponds to first ten eigenvalues/eigenvectors of the covariance matrix.

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

I will next explore the performance of the various estimators from Section 4.1 for varying amounts of correlation in the error terms, various values of the signal-to-noise ratio, and varying sample sizes.

Figures 6 - 7 depict the performance of various estimators for the number of factors as the correlation in the idiosyncratic noise increases. In these figures, I 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. Where applicable, I let $h(n) = .1\hat{\sigma}^2\sqrt{loglog(n)}$.

Figure 6 depicts the percentage of simulations in which an estimator correctly estimates the number of 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 PC to estimate the number of factors. The second panel (b) uses the thresholding estimator TC. Panel (c) uses the maximum ratio of two subsequent values of $T_k^0 = \psi_k(\frac{X'X}{T})$ and panel (d) uses the maximum ratio of two subsequent values of T_k^2. 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



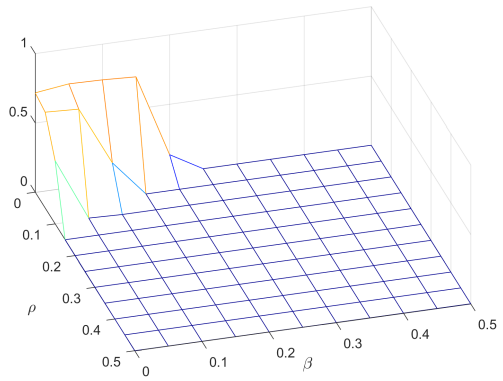
(a) Ratio $\hat{T}_k^u/\hat{T}_{k+1}^u$, $u \in \{0,1,2\}$ **(b)** Difference $\hat{T}_k^u - \hat{T}_{k+1}^u$, $u \in \{0,1,2\}$

Figure 5: Depicted are averages across 1000 replications in a simulated dataset. Data generated by baseline DGP, with $(n,T) = (300,500)$, $(\rho,\beta) = (0.3,0.1)$, $\theta = 1.5$, $r_1 = 6$. x-axis corresponds to first ten eigenvalues/eigenvectors of the covariance matrix.

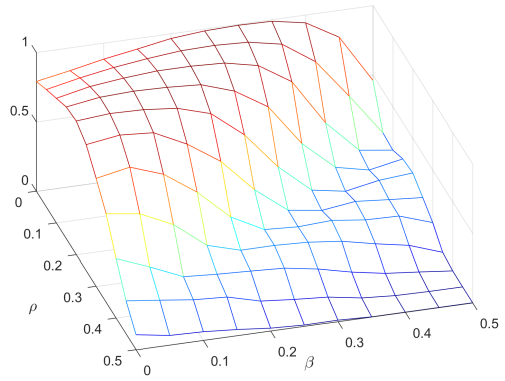
on the eigenvalues T_k^0, while panels (b) and (d) depict the corresponding estimators based on T_k^2. Panel (e) depicts the estimated number $\hat{r}$ using the ED estimator of Onatski (2010), while panel (f) depicts the alternative thresholding estimator based on the eigenvalues in $PC_{\sqrt{n}}$.

Figure 6 shows that, even under an exact factor structure, with $(\rho,\beta) = (0,0)$, only the two estimators incorporating the additional information in the eigenvectors (Figures 6b and 6d) reliably give an estimate of $\hat{r} = 6$. In fact, three of the six depicted estimators perform poorly across the entire parameter space: PC, ER or $PC_{\sqrt{n}}$ correctly identify $r = 6$ in less than 80% of simulations for all values of (ρ,β). The ED estimator selects a decreasing number of factors as the amount of correlation in the error term increases, and performs well in a region of the parameter space with modest correlation. The estimators based upon T^2 are 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, I conclude that the TR estimator, based on the ratio of subsequent values of $\hat{T}_k^2$, performs best across 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 simulations point to a significant gain in performance. Specifically, $\hat{r}^{TR} = 5.83$, and the best performing existing estimator based on the eigenvalues yields $\hat{r}^{ED} = 4.51$ on average.

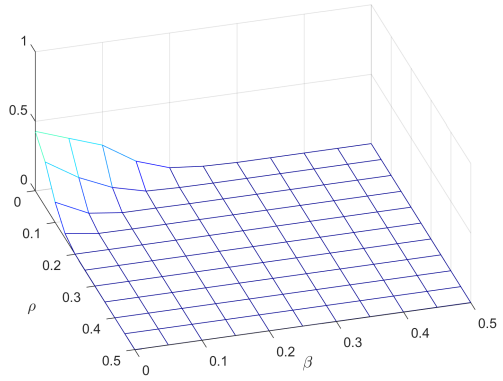
Fixing the correlation structure in the errors back to $(\rho,\beta) = (0.3,0.1)$, Figure 8 depicts the sensitivity of the estimators to the amount of noise in the data by varying θ. Estimators using



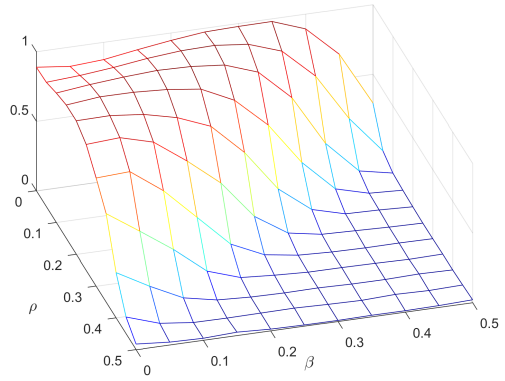
(a) Information criterion from Bai and Ng (2002) (PC)



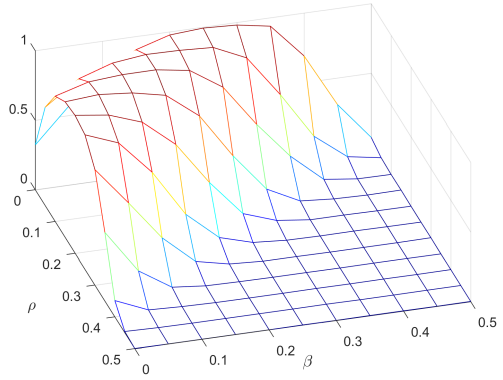
(b) Thresholding based on T^2 (TC)



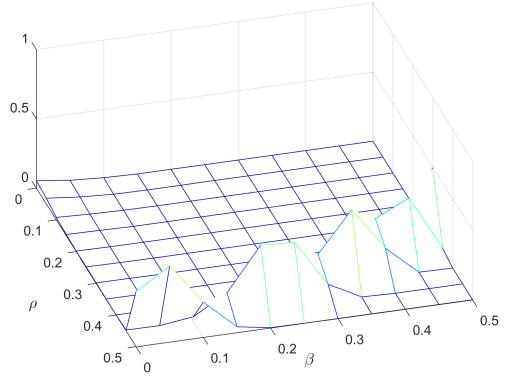
(c) Maximum ratio of two subsequent eigenvalues as in Ahn and Horenstein (2013) (ER)



(d) Maximum ratio of two subsequent values of T^2 (TR)

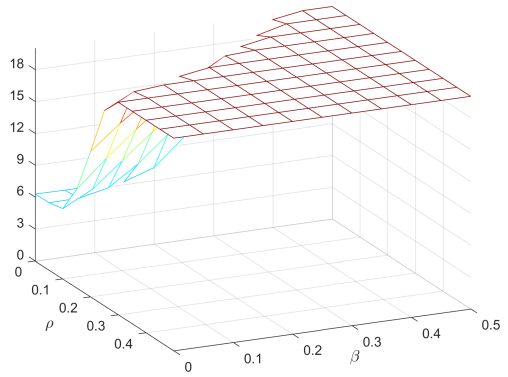


(e) Difference of two subsequent eigenvalues as in Onatski (2010) (ED)

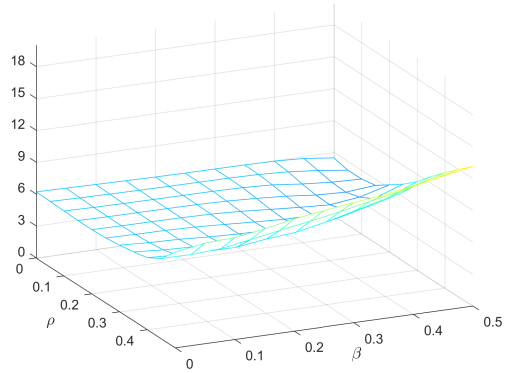


(f) Thresholding based on T^0 ($PC_{\sqrt{n}}$)

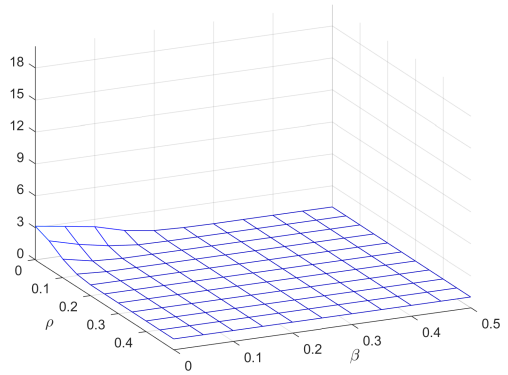
Figure 6: Percentage of simulations correctly estimating the number of "relevant" factors r_1 as both cross-sectional and intertemporal correlation is varied along a grid of (ρ, β). Data generated by baseline DGP, with $(n, T) = (300, 500)$, $\theta = 1.5$, and $r_1 = 6$. Figure based on 500 replications.



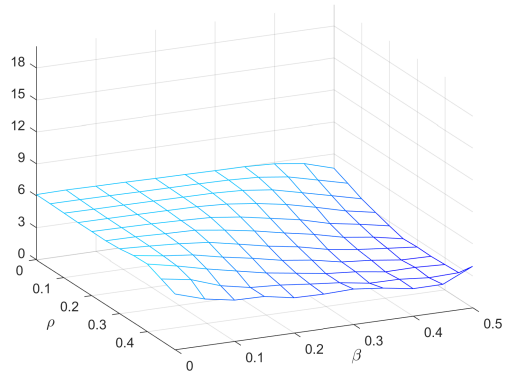
(a) Information criterion from Bai and Ng (2002) (PC)



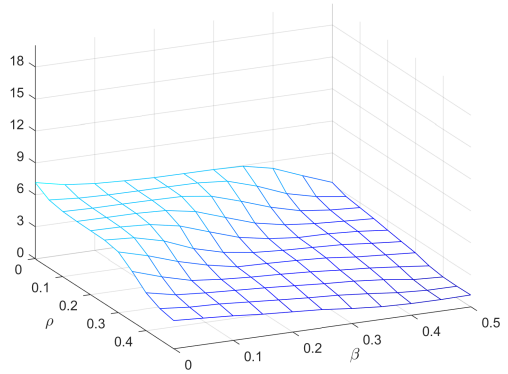
(b) Thresholding based on T^2 (TC)



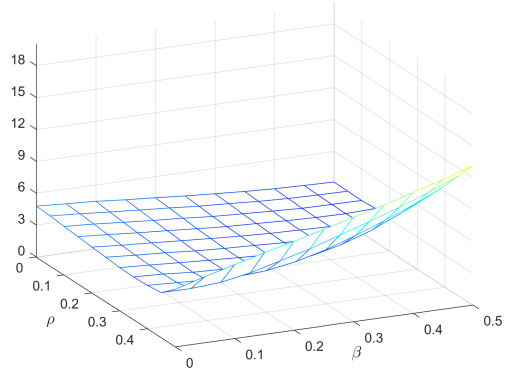
(c) Maximum ratio of two subsequent eigenvalues as in Ahn and Horenstein (2013) (ER)



(d) Maximum ratio of two subsequent values of T^2 (TR)

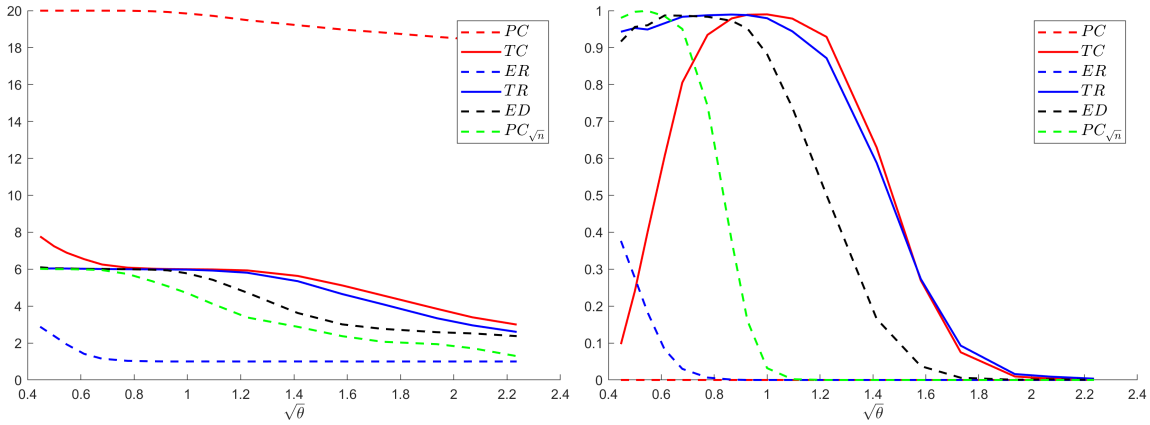


(e) Difference of two subsequent eigenvalues as in Onatski (2010) (ED)



(f) Thresholding based on T^0 ($PC_{\sqrt{n}}$)

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 PC, TC, ER, TR, ED and $PC_{\sqrt{n}}$

(b) % of simulations correctly estimating number of factors r_1 according to PC, TC, ER, TR, ED and $PC_{\sqrt{n}}$

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 information inherent in the eigenvectors remain correct for a considerably larger range of θ compared to their counterparts derived solely from the empirical eigenvalues. The TR estimator performs best in this dimension.

For the final set of simulations I vary both the cross-sectional dimension as the well as the time horizon of the data. Table 2 depicts the results. Each entry in Table 2 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 PC and ER directly with TC and TR, the previous pattern holds up: Setting $u > 0$ significantly improves the performance of the estimator. The TR estimator tends to perform best across the estimators considered.

In conclusion, I find that the TR estimator tends to perform best across most of the DGPs considered here.

In Online Appendix A.2, I repeat the analysis of this section with an alternative DGP that has a strong factor structure. In particular, I 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 $\alpha_k = 1$ for $k = 1, \ldots, 6$ as is usually the case in the literature. I find that, under a strong factor structure, estimators incorporating the partial sums in the eigenvector generally perform no worse than existing estimators, although the ED estimator of Onatski (2010) tends to perform particularly well in smaller samples. I therefore conclude that raising $\hat{T}_k^u$ to a power $u > 0$ has

n	T	ER	TR	PC	$PC_{\sqrt{n}}$	TC	ED
100	100	1.03 / 0.00	4.02 / **0.02**	20 / 0.00	14.7 / 0.00	14.3 / 0.00	1.53 / 0.00
100	150	1.02 / 0.00	3.52 / **0.03**	20 / 0.00	11.8 / 0.00	12.4 / 0.00	1.74 / 0.00
150	100	1.02 / 0.00	3.16 / **0.02**	20 / 0.00	13.7 / 0.00	14.6 / 0.00	1.7 / 0.00
150	250	1.01 / 0.00	3.86 / 0.09	20 / 0.00	5.94 / **0.61**	8.62 / 0.04	2.2 / 0.02
150	500	1.01 / 0.00	4.98 / 0.42	19.1 / 0.00	4.6 / 0.05	6.25 / **0.62**	3.28 / 0.20
300	250	1 / 0.00	4.39 / 0.19	20 / 0.00	4.09 / 0.00	6.5 / **0.45**	2.72 / 0.02
300	500	1 / 0.00	5.82 / **0.88**	19.5 / 0.00	3.38 / 0.00	5.93 / 0.93	4.62 / 0.49
300	750	1 / 0.00	5.96 / **0.97**	16.9 / 0.00	3.28 / 0.00	5.97 / 0.97	5.76 / 0.86
500	250	1 / 0.00	4.58 / 0.16	20 / 0.00	3.16 / 0.00	5.81 / **0.45**	3.02 / 0.01
500	500	1 / 0.00	5.92 / **0.94**	20 / 0.00	3.01 / 0.00	5.95 / 0.95	5.27 / 0.62
500	750	1 / 0.00	6 / **1.00**	17.8 / 0.00	3 / 0.00	5.99 / 0.99	5.95 / 0.95
1000	1000	1 / 0.00	6 / **1.00**	16.8 / 0.00	2.98 / 0.00	6 / 1.00	6 / 1.00

Table 2: 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.

little implications if all factors are strong, but yields significant performance gains if local factors are present in the data.

Based on these findings, my recommendation for estimating the number factors r_1 is therefore to use the TR estimator 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. Let $z = 0.7\sqrt{log(log(n))}\sqrt{n}$, rounded to the nearest integer[17], and compute

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

3. Set

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

[17] $g(n) = .7\sqrt{loglog(n)} \approx 1$ for most relevant sample sizes, and this recommendation is therefore generic.

6 A Factor Model with Local Factors of the US Economy

Two classic applications where factor models have proven particularly useful are macroeconomic monitoring and forecasting (see Stock and Watson (2016) for a good review). This section describes the factor model estimated from a large panel of US macroeconomic indicators under the weaker assumptions maintained in this paper and illustrates the implications of the presence of local factors.

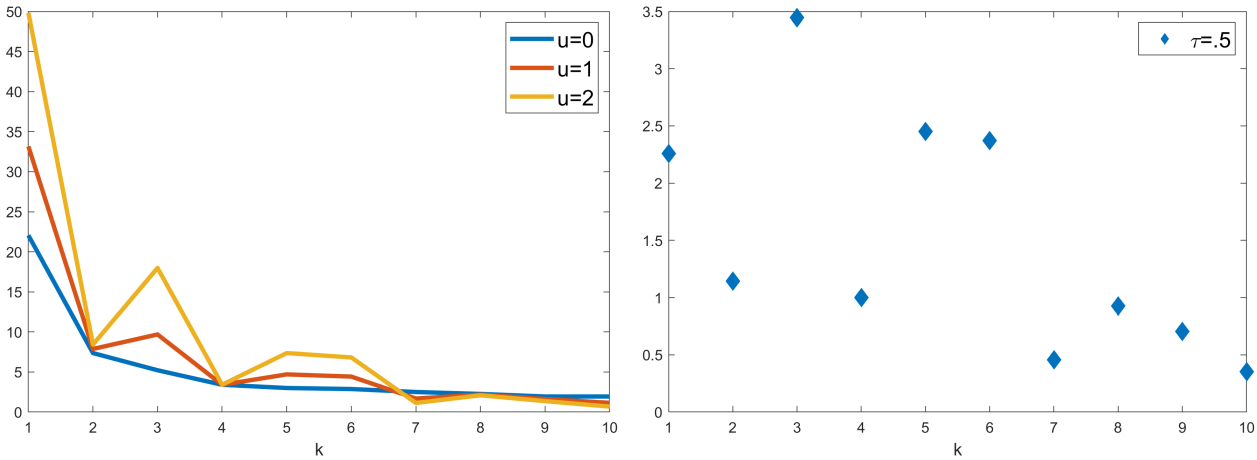
I employ one of the standard datasets in the factor model literature in macroeconomics (see, e.g. Stock and Watson (2005) and De Mol et al. (2008)). The data contains quarterly observations of 207 macroeconomic variables, primarily for the US economy. In particular, I use the vintage of the dataset used 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. All variables have been transformed to achieve approximate stationarity and a small number of outliers were removed. I 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)[18].

The dataset consists of series at multiple levels of aggregation. I only use the disaggregated time series in my estimation of the factor structure and disregard the aggregates (Boivin and Ng (2006), Stock and Watson (2016)). This elimination leaves 139 variables in the data. Only 94 of those series are available for the entire sample and I will restrict my analysis to those 94 time series. This allows for a straightforward implementation of the principal component estimator.[19]

I 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 $\hat{T}_k^u, u \in \{0, 1, 2\}$ and $\hat{S}_k^2$ in the data (setting $u = 0$ reproduces Figure 1 from the introduction). If $\hat{T}_k^u$ corresponds to a local factor with a strong effect on a subset of outcomes, it will be scaled up if $u > 0$. On the other hand, this does not hold for an eigenvalue corresponding to a factor that weakly affects all outcomes. This is illustrated in Figure 9b. In particular, with $\tau = 0.5$, I find that the 3rd, 5th, and 6th eigenvector are particularly concentrated on its largest loadings. 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.

[18] 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, I refer the reader to Stock and Watson (2016).

[19] Alternatively one could analyze the full sample of 139 disaggregated variables using the EM algorithm of Stock and Watson (2002b) to handle missing observations.



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

(b) Empirical distribution of $\hat{S}_k^2$

Figure 9: Illustration of key quantities for $\tau = 0.5$ in a dataset of US macroeconomic indicators

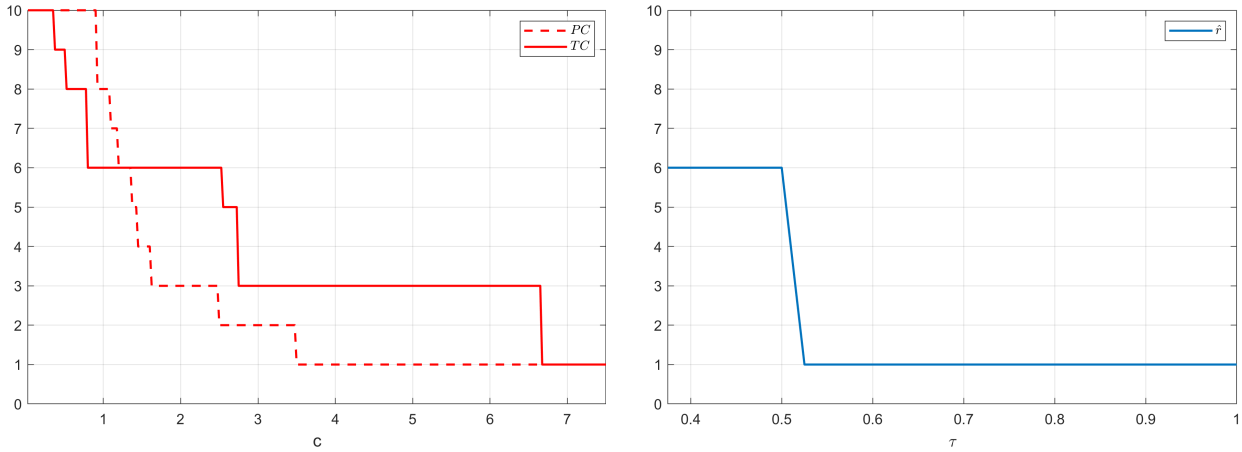
I next summarize the results of the 6 estimators considered throughout this paper in Table 3. While both estimators derived from $\hat{T}_k^2$ suggest the presence of 6 factors in the data, the three existing estimators from the literature (ER, PC, ED) find evidence for 1, 8, and 3 factors respectively.

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) suggests to vary this threshold and explore how the estimated number of factors changes. The result is depicted in Figure 10a. It depicts the estimated number of factors based on PC and TC 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. Using $\hat{T}_k^0$ yields significantly more ambiguous results than an equivalent procedure based on $\hat{T}_k^2$. Decreasing 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. Using a thresholding rule based on $\hat{T}_k^2$ instead, we observe two flat regions in the graph at $\hat{r} = 3$ and $\hat{r} = 6$ respectively.

Alternatively, a practitioner could vary τ and observe how the estimated number of factors changes. Recall that we may think of τ as a complexity parameter, because varying τ shifts the steep region of Figure 2. I demonstrate this in Figure 10b, which depicts the value of the TR estimator as a function of the complexity parameter τ. Figure 10b suggests that there is one "global"

Estimator	ER	TR	PC	$PC_{\sqrt{n}}$	TC	ED
Estimated number of factors	1	6	8	3	6	3

Table 3: Estimated number of factors in macroeconomic panel for the six estimators considered throughout this paper.



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

(b) TR, the maximum of $\hat{T}_k^2/\hat{T}_{k+1}^2$, $k \in \{2, 3, \ldots, 10\}$ for varying tuning parameter $\tau \in [.375, 1]$.

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

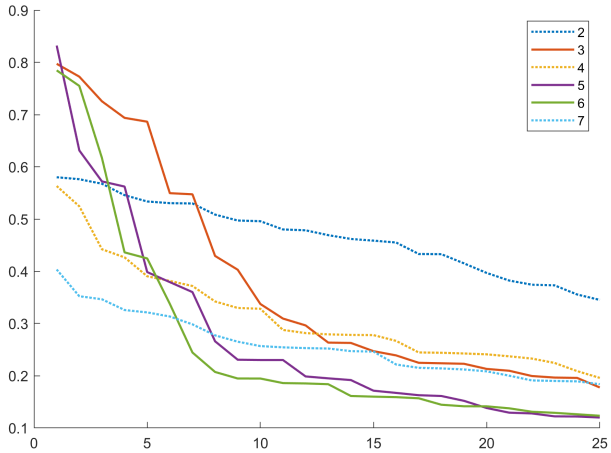
factor in the data and that the second most important factor is significantly weaker than the first one, as indicated by the first flat region of the graph. Next, the 7th factor appears to be significantly weaker than the 6th, as indicated by the second flat region in Figure 10b. Thus, Figure 10 suggests that the choice of $\hat{r} = 6$ is quite robust when incorporating the additional information in the eigenvectors.

Figure 9b also suggests a more nuanced interpretation of the factors, with factors 3, 5 and 6 appearing more "local". This is further illustrated in Figure 11a. For each factor, all associated loadings were ordered (in absolute value). Figure 11a then depicts how the largest 25 loadings decay for factors 2-7. The three factors identified as "local" above can easily be distinguished. They exhibit some large loadings, combined with a steeper subsequent drop-off.

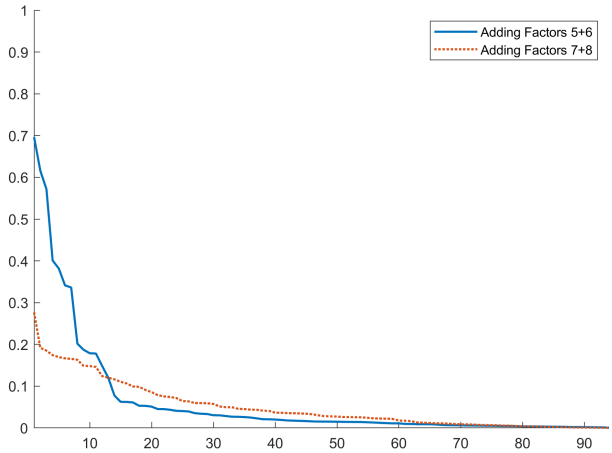
Related, 11b illustrates the importance of the factors for each series by considering the R^2 of the common component in explaining movements each series. For a given series in the panel, this measures the variation in the series due to contemporaneous variation in the factors.

Figure 11b depicts how the addition of weaker factors affects the individual R^2 of the 94 series in the panel. Sorted from highest to lowest, it illustrates how the local factors 5 and 6 are extremely important for a subset of the outcomes. For example, for the most impacted series of the panel, factors 5 and 6 explain around 70 percent of the variation in that series. While this subset of outcomes is very well explained by factors 5 and 6, associated with large jumps in the corresponding R^2, factors 7 and 8 do not have such a strong effect on a subset of outcomes. The 10 largest increments in the model R^2 are on average 0.39 and 0.18 for factors 5+6 and 7+8 respectively.[20]

[20]Note that the smooth nature of the scree plot implies that the 7th and 8th factors are not significantly weaker than



(a) Absolute value of 25 largest loadings for factors 2-7. Each line corresponds to the largest 25 loadings (in absolute value) for a specific factor. Solid lines correspond to factors identified earlier as "local."



(b) Incremental R^2 of the common component for each series, ordered from largest to smallest, for adding factors 5 and 6 (solid blue line), and factors 7 and 8 (dotted red line) to the model.

Figure 11: Both panels visualize how some factors in this dataset are more "local" than others.

Factor 3	Factor 5	Factor 6
PPI: Int. Material: Supplies & Components	**Nonfarm: Unit Nonlabor Payments**	**tb6m-tb3m**
PPI: Industrial Commodities	**Nonfarm: Unit Labor Cost**	**GS1-Tb3m**
PPI: Finished Consumer Goods	**Nonfarm: Real Compensation Per Hour**	**GS10-Tb3m**
PPI: Crude Petroleum Defl by PCE(LFE)	**BS: Real Compensation Per Hour**	**S&P'S STOCK PRICE INDEX**
Gasoline and other energy goods	PPI: Finished Consumer Foods	**DOW JONES IA**
BS: Real Compensation Per Hour	Food & beverages for off-premises consump	Consumer Loans, All Commercial Banks
Nonfarm: Real Compensation Per Hour	**Nonfarm: Output Per Hour of All Persons**	**BAA-GS10 Spread**
BS: Implicit Price Deflator	PPI: Finished Consumer Goods	
ISM Manufacturing: Prices Paid Index		

Table 4: Variables corresponding to largest loadings for actors 3, 5 and 6, the most local factors. Red coloring indicates a negative loading, while black indicates a positive loading.

Further, by treating factors as local, the resulting factors may be easier to interpret as they only correspond to a small subset of the observables, contrasting with conventional factors, which are often hard to interpret. Table 4 shows which economic indicators correspond to the largest loadings (in absolute value) associated with the three local factors. Variables with a negative loading are shown in red. For factor 3, I note that six of the nine variables, printed in bold, represent price indices as classified in the handbook chapter of Stock and Watson (2016). Additionally the fourth entry, while classified as an "Oil market variable," also represents a price index. The remaining two variables are both classified as "Productivity and Earnings" and it is worth noting that they have the opposite sign. Next, of the five series classified as "Productivity and Earnings" in the data, all five of these are associated with factor 5, emphasized in bold. Further, the remaining three entries are all price indices. I also note that the 6th factor is highly concentrated on a number of

factors 5 and 6. Adding factors 5 and 6 to the model increases the R^2 of the common component by 0.06 on average, while adding factors 7 and 8 to the model increases the R^2 of the common component by 0.05.

financial variables, specifically spreads and stock market indicators (again emphasized in bold). Further, this factor is associated with a negative return on the stock market and an increase in the interest rate spread.

This aids in the interpretation of the factors. For example, based on the discussion above, the 6th factor could be interpreted as indicating a flight from stocks into safe assets, such as bonds.

The previous discussion illustrates the advantage of taking the eigenvectors into account when selecting the number of factors as proposed in this paper. Without this additional information, factors 5 and 6 are missed by two of the three existing estimators in the literature (cf. Table 3). But these factors are highly influential on a subset of outcomes (e.g. Table 11). Failure to include them in the model would thus result in a model that does poorly in explaining this part of the economy.

7 Concluding Remarks

This paper develops a framework for factor models that allows for local factors that only affect an unknown subset of the observables. In many economic models I find that factors affecting proportionally more than $\sqrt{n}$ of the n observed variables are of economic interest. Under standard assumptions on the error terms, this coincides with the number of factors that can be estimated consistently using the principal component estimator. I further show that existing estimators for the number of factors in general do not yield a consistent estimate for this number of "relevant" factors. To estimate the number of economically important and estimable factors consistently, I argue that there is additional information in the eigenvectors that has not been exploited in the past. I demonstrate how one can easily incorporate this information into some of the prominent estimators commonly used. Monte Carlo evidence suggests significant finite sample gains over existing estimators.

In cases in which there is no clear gap in the distribution of eigenvalues, the theory developed in this paper provides a viable framework. It further provides a theoretical foundation that justifies the use of both factor models and the principal component estimator in datasets with no such clear gap.

In addition, the methods of this paper provide a novel insight into the structure of the data. There are two potential reasons subsequent factors may appear "weak" in a given dataset - either a weaker factor can have a weak effect on all observables, or it can have a strong impact on only a subset of observables (which I call a "local" factor in this paper). By using only the eigenvalues of $\frac{X'X}{T}$, these two kinds of factors will be treated equally. However, depending on the economic model or context, a researcher may be more interested in one or the other. By incorporating information from the eigenvectors, I allow a practitioner to distinguish between the two cases.

I implement my methods in one of the canonical datasets used in the factor model literature

and find strong evidence that there are indeed local factors present in the data.

The analysis in this paper suggests a number of promising topics for future research. Perhaps most interestingly, I conjecture that the principal component estimator considered in this paper can be substantially improved upon (at least in finite samples) using the sparsity assumptions of the model for the estimation of the factors. A regularized estimation approach suggests itself and is currently investigated in a separate project (Freyaldenhoven (2019)).

References

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.

Antoine, B. and E. Renault (2012). Efficient minimum distance estimation with multiple rates of convergence. *Journal of Econometrics 170*(2), 350–367.

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.

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

Bernanke, B. S., J. Boivin, and P. Eliasz (2005). Measuring the effects of monetary policy: A factor-augmented vector autoregressive (FAVAR) approach. *The Quarterly Journal of Economics 120*(1), 387–422.

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. *The 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.

Chamberlain, G. (1983). Funds, factors, and diversification in arbitrage pricing models. *Econometrica 51*(5), 1305–1323.

Chamberlain, G. and M. Rothschild (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica 51*(5), 1281–1304.

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. *The Econometrics Journal 14*(1), C45–C90.

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

Connor, G. and R. A. Korajczyk (1995). The arbitrage pricing theory and multifactor models of asset returns. In R. A. Jarrow, V. Maksimovic, and W. T. Ziemba (Eds.), *Handbooks in operations research and management science*, Volume 9, Chapter 4, pp. 87–144. Elsevier.

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.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2000). The generalized dynamic-factor model: Identification and estimation. *Review of Economics and Statistics 82*(4), 540–554.

Freyaldenhoven, S. (2019). Identification through sparsity in factor models. Working paper, Federal Reserve Bank of Philadelphia.

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

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*.

Giannone, D., L. Reichlin, and L. Sala (2006). VARs, common factors and the empirical validation of equilibrium business cycle models. *Journal of Econometrics 132*(1), 257–279.

Green, R. C. and B. Hollifield (1992). When will mean-variance efficient portfolios be well diversified? *The 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.

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

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. *The 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. *The 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. *The Journal of Finance 45*(5), 1541–1564.

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

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

Stock, J. H. and M. W. Watson (2006). Forecasting with many predictors. In G. Elliott, C. Granger, and A. Timmermann (Eds.), *Handbook of Economic Forecasting*, Volume 1, Chapter 10, pp. 515–554. Elsevier.

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.

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

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