Estimating Number of Factors by Adjusted Eigenvalues Thresholding

Jianqing Fan*
Department of Operations Research and Financial Engineering,
Bendheim Center for Finance, Princeton University
and
Jianhua Guo* Shurong Zheng*
School of Mathematics and Statistics and KLAS, Northeast Normal University

September 25, 2019

Abstract

Determining the number of common factors is an important and practical topic in high dimensional factor models. The existing literatures are mainly based on the eigenvalues of the covariance matrix. Due to the incomparability of the eigenvalues of the covariance matrix caused by heterogeneous scales of observed variables, it is very difficult to give an accurate relationship between these eigenvalues and the number of common factors. To overcome this limitation, we appeal to the correlation matrix and show surprisingly that the number of eigenvalues greater than 1 of population correlation matrix is the same as the number of common factors under some mild conditions. To utilize such a relationship, we study the random matrix theory based on the sample correlation matrix in order to correct the biases in estimating the top eigenvalues and to take into account of estimation errors in eigenvalue estimation. This leads us to propose adjusted correlation thresholding (ACT) for determining the

*Jianqing Fan is supported by NSF grants DMS-1712591 and DMS-1947097 and NIH grant R01-GM072611. Jianhua Guo and Shurong Zheng gratefully acknowledge National Natural Science Foundation of China (NNSFC) grant 11690012.

number of common factors in high dimensional factor models, taking into account the sampling variabilities and biases of top sample eigenvalues. We also establish the optimality of the proposed methods in terms of minimal signal strength and optimal threshold. Simulation studies lend further support to our proposed method and show that our estimator outperforms other competing methods in most of our testing cases.

Keywords: Factor models, number of factors, random matrices, adjusted eigenvalues, bias corrections.

1 Introduction

High-dimensional factor models find many applications in finance, economics, and genomics, or more generally high-dimensional data where the dependence of measurements can be attributed to a relatively small number of common factors (Fan et al., 2018). Determining the number of factors is an important issue in applications of factor models. The methods are typically based on eigenvalues or rank of the loading matrix. For example, Lewbel (1991) and Kong, Liu, and Zhou (Kong et al.) obtained the number of factors by testing the rank of the loading matrix.

There are rich literatures on eigenvalues based methods for selecting the number of common factors, which have been studied from three different perspectives. The first one is through model selection. Bai and Ng (2002) proposed three PC and three IC criteria by using the penalties to determine the number of common factors. Hallin and Liska (2007) developed an information criterion to determine the number of common factors in the general dynamic model. Li et al. (2017) proposed the information criteria similar to Bai and Ng (2002) to determine the number of common factors when the number of factors increases with the sample size. Su and Wang (2017) used the BIC information criterion to determine the number of common factors for time-varying factor models.

The second perspective is through hypothesis testing or confidence intervals. Onatski (2009) proposed a test statistic $T_{ON} = \max_{r_{\min}<i\leq r_{\max}} (\hat{\lambda}_i - \hat{\lambda}_{i+1})/(\hat{\lambda}_{i+1} - \hat{\lambda}_{i+2})$ to test the number of common factors where $\hat{\lambda}_i$ is the ith largest eigenvalue of the estimated covariance matrix, and $r_{\min}$ and $r_{\max}$ are pre-specified lower and upper bounds of the number of common factors and estimated the number of factors by

$$\hat{K}_{ON} = \arg \max_{r_{\min}<i\leq r_{\max}} (\hat{\lambda}_i - \hat{\lambda}_{i+1})/(\hat{\lambda}_{i+1} - \hat{\lambda}_{i+2}).$$

Kapetanios (2010) used the statistic $\tau_i(\hat{\lambda}_i - \hat{\lambda}_{r_{\max}+1})$ to test the number of common factors where τ_i is the normalized constant. Pan and Yao (2008) used the Ljung-Box-Pierce portmanteau test statistic to determine the number of common factors. Based on a confidence interval of the largest non-spiked eigenvalue of the estimated covariance matrix, Cai et al. (2017) proposed an algorithm to determine the number of common factors under the convergence regime that the dimension and the sample size tend to infinity proportionally.

The third perspective is through estimation. Onatski (2010) used the maximum eigengap to determine the number of common factors and proposed the eigenvalue difference criterion as follows:

$$\hat{K}_{ED} = \max\{i \leq r_{\max} : \hat{\lambda}_i - \hat{\lambda}_{i+1} \geq s\},$$

with s being a given threshold. Onatski (2010) stated that the difference between ED and Bai-Ng criteria is that the threshold of ED is sharp and the threshold of Bai-Ng criteria has more freedom. Wang (2012) and Lam and Yao (2012) proposed to use the ratios of two adjacent eigenvalues to determine the number of factors, which estimates K by

$$\hat{K}_{ER} = \arg \max_{1 \leq i \leq r_{\max}} \hat{\lambda}_i / \hat{\lambda}_{i+1}. \tag{1}$$

Ahn and Horenstein (2013) proposed also "ER" method independently, in addition to the "GR" method:

$$\hat{K}_{GR} = \arg \max_{1 \leq i \leq r_{\max}} \log(V_{i-1}/V_i) / \log(V_i/V_{i+1}),$$

with $V_i = \sum_{j=i+1}^{p} \hat{\lambda}_j$.

The aforementioned methods are all based on the eigenvalues of the covariance matrix, which is assumed to admit the sum of a low-rank matrix and a sparse matrix. Let $\mathbf{B} = (b_{ij})$ be a $p \times K$ dimensional matrix with $K < p$ that represents the factor loading matrix and diag$(\nu_1^2, \cdots, \nu_p^2)$ be the diagonal matrix that represents the variances of idiosyncratic noises.

Then, the covariance matrix of observed high-dimensional data is given by

$$\mathbf{\Sigma} = \mathbf{B}\mathbf{B}^T + \text{diag}(\nu_1^2, \cdots, \nu_p^2), \tag{2}$$

where T denotes the transpose of a vector or a matrix. A drawback of the covariance based methods is that it does not take into account the scales of the observed variables. For this reason, the existing methods can easily be inconsistent. For example, even for the simplest factor model (2) with the population covariance matrix $\mathbf{\Sigma} = \mathbf{B}\mathbf{B}^T + \text{diag}(\underbrace{1, \cdots, 1}_{K}, \nu_{K+1}^2, 1 \cdots, 1)$ where $\mathbf{B}^T = (\mathbf{B}_1^T, \mathbf{0}_{K\times(p-K)})$, $\mathbf{B}_1$ is of $K \times K$ dimension and $\text{rank}(\mathbf{B}_1) = K$, under some mild conditions of $\mathbf{B}_1$ and ν_{K+1}^2, we can show that

$$\begin{aligned} P(\hat{K}_{ON} \geq K+1) \to 1, \qquad & P(\hat{K}_{ED} \geq K+1) \to 1, \\ P(\hat{K}_{ER} \geq K+1) \to 1, \qquad & P(\hat{K}_{GR} \geq K+1) \to 1, \end{aligned} \tag{3}$$

but in fact, the true number of common factors is K. The proof of (3) will be given in Appendix B.

The correlation matrix clearly overcomes the scaling drawback of the covariance matrix. The $p \times p$ dimensional correlation matrix of $\mathbf{\Sigma}$ is given by

$$\mathbf{R} = [\text{diag}(\mathbf{\Sigma})]^{-1/2}\mathbf{\Sigma}[\text{diag}(\mathbf{\Sigma})]^{-1/2}, \tag{4}$$

where $\text{diag}(\mathbf{\Sigma})$ is the diagonal matrix by replacing the off-diagonal elements of $\mathbf{\Sigma}$ by zeros. Using the sample correlation matrix for factor analysis will overcome the aforementioned disadvantages of using sample covariance matrix. In fact, when the dimension is fixed and the sample size tends to infinity, Guttman (1954), Kaiser (1960, 1961) and Johnson and Wichern (2007) (page 491) have established a lower bound: the number of the eigenvalues satisfying $\max\{j : \hat{\lambda}_j > 1, j \in \{1, \cdots, p\}\}$ is smaller than or equal to the number of the

common factors. But the existing literatures haven't shown that they are indeed the same under certain conditions. Moreover, their estimation techniques

$$\hat{K}_u = \max\{j : \hat{\lambda}_j > 1, j \in [p]\} \quad \text{where} \quad [p] = \{1, \cdots, p\}$$

can not be consistent in the high dimensional setting since sample correlation matrices are inconsistent. Can such a simple, tuning parameter-free method be modified so that it is consistent for high dimensional factor models? This paper gives an affirmative answer via some high-dimensional adjustments of threshold parameters, leveraging on the random matrix theory.

The main contributions of this paper are as follows:

- Firstly, we establish the concise relationship between the eigenvalues of population correlation matrices and the number of common factors, that is, give the condition under which

$$K = \max\{j : \lambda_j(\mathbf{R}) > 1, j \in [p]\} \tag{5}$$

where $\lambda_1(\mathbf{R}) \geq \lambda_2(\mathbf{R}) \geq \cdots \geq \lambda_p(\mathbf{R})$ are the eigenvalues of correlation matrix $\mathbf{R}$ and K is the true number of common factors.

In factor analysis, the eigenvalues of correlation matrix are frequently used to evaluate the contributions of selected factors. Since $\sum_{j=1}^{p} \lambda_j(\mathbf{R}) = p$, some of eigenvalues of $\mathbf{R}$ are greater than 1 and the remaining eigenvalues of $\mathbf{R}$ are equal to or less than 1. It has been shown (Guttman, 1954; Kaiser, 1960, 1961) that the number of common factors is less than or equal to the number of $\mathbf{R}$'s eigenvalues greater than 1. One of contributions is to show that they are indeed the same. The results presented in Table 1 illustrate this point where $\{b_{\ell j}, \ell \in [p], j \in [K-1]\}$ are i.i.d. from the uniform

distribution U$(-1,1)$, $\nu_1^2 = \ldots = \nu_p^2 = \sigma^2$, $b_{1K}, \cdots, b_{pK}$ are i.i.d. from $U(-1,1)$ in Scenario 1 and $b_{1K} = \cdots = b_{pK} = 0$ in Scenario 2.

Table 1: Number of eigenvalues of $\mathbf{R}$ satisfying $\lambda_j(\mathbf{R}) > 1$

		Scenario 1			Scenario 2		
		$\sigma^2 = 1$	2	3	$\sigma^2 = 1$	2	3
K	p	rank($\mathbf{B}$)=K			rank($\mathbf{B}$)=K-1		
5	50	5	5	5	4	4	4
	100	5	5	5	4	4	4
10	50	10	10	10	9	9	9
	100	10	10	10	9	9	9

- Secondly, we propose a bias corrected estimator $\hat{\lambda}_i^C$ for $\lambda_i(R)$, which in general differs from the i^{th} largest eigenvalue $\hat{\lambda}_i$ of sample correlation matrix and develop a new estimator for the number of common factors as follows:

$$\hat{K}^C = \max\{j : \hat{\lambda}_j^C > s, \quad j \in [r_{\max}]\}, \qquad s = 1 + \sqrt{p/(n-1)} \tag{6}$$

under the regime $\rho_{n-1} = p/(n-1) \to \rho \in (0, \infty)$, where p is the dimension and n is the sample size. Our newly proposed method $\hat{K}^C$ does not depend on any tuning parameter and is even simpler than the eigenvalue ratio method $\hat{K}_{ER}$ or $\hat{K}_{ED}$ which involves the tuning parameter s. On the other hand, by (5), a naive method is

$$\hat{K}_u = \max\{j : \hat{\lambda}_j > 1, j \in [p]\}.$$

but this method overestimates K when n and p are of the same order.

Let $p/(n-1) \to \rho$ and

$$\begin{aligned}\mathcal{F}(v_0) \;=\; \{\mathbf{R}: \;& \mathbf{R} \text{ is the correlation matrix (8) of the observed vector} \\ & \text{in the factor model (7) and } \lambda_K(\mathbf{R}) > v_0\},\end{aligned}$$

where v_0 is a positive constant, representing signal strength. We will show that the optimal lower bounds for the signal strength v_0 and threshold s are

$$v_0 = 1 + \sqrt{\rho}, \quad s = 1 + \sqrt{p/(n-1)}$$

in the following sense.

Minimum signal strength v_0: We will show

(i). When $v_0 < 1 + \sqrt{\rho}$, there exists $\mathbf{R} \in \mathcal{F}(v_0)$ where no method based on the eigenvalues of the sample correlation matrix can give a consistent estimate of K.

(ii). When $v_0 = 1 + \sqrt{\rho}$, our method $\hat{K}^C$ can consistently estimate K.

Optimal threshold s. Let $\hat{K}^C(s)$ emphasize the dependence of $\hat{K}^C$ on a general s. We will prove

$$\begin{cases} P(\hat{K}^C(s) = K) \to 1, & \forall\ \mathbf{R} \in \mathcal{F}(\nu_0) \text{ if } s = 1 + \sqrt{p/(n-1)}, \\ P(\hat{K}^C(s) > K) \to 1, & \exists\ \mathbf{R} \in \mathcal{F}(\nu_0) \text{ if } s < 1 + \sqrt{p/(n-1)}, \\ P(\hat{K}^C(s) < K) \to 1, & \exists\ \mathbf{R} \in \mathcal{F}(\nu_0) \text{ if } s > 1 + \sqrt{p/(n-1)}, \end{cases}$$

In other words, the threshold $s = 1 + \sqrt{p/(n-1)}$ is optimal. We have conducted extensive simulations to compare our method with those in Bai and Ng (2002), Onatski (2005, 2010), Lam and Yao (2012), Ahn and Horenstein (2013). Simulation results

show that in most of our testing cases, our estimation method outperforms the competing ones. Even in the remaining cases considered in this paper, our estimation method has comparable performance to other competing methods.

- Thirdly, we derive the asymptotic properties of the largest K sample eigenvalues of the sample correlation matrix in high dimensional factor models. This is an important contribution to random matrix theories. The results may be used for other inference problems in high dimensional factor models.

The arrangement of this paper is as follows: Section 2 reviews the factor model, defines the common factors in detail and establishes the relationship between the number of common factors and the eigenvalues of the population correlation matrix. Section 3 proposes an estimation technique of the number of common factors based on a study on the random matrix theory of sample correlation matrix and demonstrates the convergence of the proposed estimator for the number of common factors. Section 4 investigates the optimality of the proposed estimator in high dimensional factor model. Section 5 presents extensive simulation results. Section 6 conducts two empirical studies. Section 7 concludes. Most of technical proofs are given in the appendix.

2 High Dimensional Factor Model

We now briefly review the factor model. In the factor model, the observable variable $\mathbf{y}$ can be decomposed as

$$\mathbf{y} = \boldsymbol{\alpha} + \mathbf{B}\mathbf{f} + \boldsymbol{\epsilon}, \tag{7}$$

where $\mathbf{y} = (y_1, \cdots, y_p)^T$ is the p-dimensional observable vector, $\mathbf{f} = (f_1, \cdots, f_K)^T$ is the K-dimensional latent factor vector, $\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_p)^T$ is the p-dimensional error vector,

$\boldsymbol{\alpha}$ is the p-dimensional intercept vector and $\mathbf{B}$ is the $p \times K$ dimensional loading matrix. Following Bai and Ng (2002), define the number of factors as rank($\mathbf{B}$). We impose the following conditions.

- **Condition C1:** The factors $f_1, \ldots, f_K$ are mutually independent; the factor vector $(f_1, \cdots, f_K)$ is independent of the error vector $(\epsilon_1, \ldots, \epsilon_p)$;
- **Condition C2:** $\mathrm{E}(\mathbf{f}) = \mathbf{0}_K, \mathrm{Cov}(\mathbf{f}) = \mathbf{I}_K$;
- **Condition C3:** $\mathrm{E}(\boldsymbol{\epsilon}) = \mathbf{0}_p, \mathrm{Cov}(\boldsymbol{\epsilon}) = \boldsymbol{\Psi} > \mathbf{0}_{p\times p}$ where $\boldsymbol{\Psi}$ may be not diagonal (but sparse);
- **Condition C4:** $p > K$ and the loading matrix $\mathbf{B}$ is of full column rank, i.e., $\mathrm{rank}(\mathbf{B}) = K$.

Write $\mathbf{B} = (\mathbf{b}_1, \cdots, \mathbf{b}_K)$ and $\mathbf{b}_j = (b_{1j}, \cdots, b_{pj})^T$ is a p-dimensional column vector for $j \in [K]$. If there is at most one coefficient $b_{\ell j} \neq 0$ with $\ell \in [p]$ for some $j \in [K]$, that is, f_j is only related to y_j and not related to $y_1, .., y_{j-1}, y_{j+1}, \cdots, y_p$, we can put f_j in ϵ_j. Thus, without loss of generality, we will define the common factor as follows:

Definition of Common Factors: If there are at least two coefficients $b_{\ell_1 j}, b_{\ell_2 j} \neq 0$ with $\ell_1, \ell_2 \in [p]$ for some $j \in [K]$, call the factor f_j as a common factor.

This paper focuses on determining the number of common factors under Conditions C1-C2-C3-C4-C5 where

Condition C5: For every $j \in [K]$, there are at least two coefficients $b_{\ell_1 j}, b_{\ell_2 j} \neq 0$ with $\ell_1, \ell_2 \in [p]$.

Conditions C1-C2-C3 have been frequently imposed (Bai and Ng, 2002; Johnson and Wichern, 2007). They are related to identifiability and moment conditions. Although they

are often used, Conditions C4-C5 are not explicitly written. Condition C4 shows that $\mathbf{B}$ is of full column rank, that is, $\text{rank}(\mathbf{B}) = K$. Condition C5 shows that every factor f_j has an impact on at least two observed variables.

By definition (7) of the factor model, we have

$$\mathbf{\Sigma} = \text{Cov}(\mathbf{y}) = \mathbf{B}\mathbf{B}^T + \mathbf{\Psi}.$$

Let the (j, j) entry of $\mathbf{\Sigma}$ be σ_{jj} for $j \in [p]$. Then by (4), the population correlation matrix of $\mathbf{y}$ in the factor model is

$$\mathbf{R} = \mathbf{Q}\mathbf{Q}^T = \mathbf{Q}_1\mathbf{Q}_1^T + \mathbf{Q}_2\mathbf{Q}_2^T, \tag{8}$$

where $\text{diag}(\mathbf{\Sigma}) = \text{diag}(\sigma_{11}, \cdots, \sigma_{pp})$ and

$$\begin{aligned} &\mathbf{Q} = [\text{diag}(\mathbf{\Sigma})]^{-1/2}(\mathbf{B}, \mathbf{\Psi}^{1/2}) = (\mathbf{Q}_1, \mathbf{Q}_2), \\ &\mathbf{Q}_1 = [\text{diag}(\mathbf{\Sigma})]^{-1/2}\mathbf{B}, \quad \mathbf{Q}_2 = [\text{diag}(\mathbf{\Sigma})]^{-1/2}\mathbf{\Psi}^{1/2}. \end{aligned} \tag{9}$$

In fact, $\mathbf{Q}_1\mathbf{Q}_1^T$ and $\mathbf{Q}_2\mathbf{Q}_2^T$ include the information of the factors $f_1, \cdots, f_K$ and errors $\epsilon_1, \cdots, \epsilon_p$ on $\mathbf{y}$, respectively. Let $\|\mathbf{M}\|_F$ denote the Frobenius norm of a matrix or a vector $\mathbf{M}$ and $\|\mathbf{M}\| = \sqrt{\lambda_1(\mathbf{M}\mathbf{M}^T)}$ the operator norm. The following theorem shows how to determine the number of factors from the population correlation matrix.

Theorem 1 *Under Conditions C1-C2-C3-C4-C5, if* $\|[\text{diag}(\mathbf{\Sigma})]^{-1}\mathbf{\Psi}\| \leq 1$, *we have*

$$\lambda_j(\mathbf{R}) \leq 1, \quad j = K+1, \ldots, p.$$

In addition, we have

$$K = \max\{j : \lambda_j(\mathbf{R}) > 1, j \in [p]\}, \tag{10}$$

when p is large enough and there exists three non-negative constants $\delta_1 > \delta_2 + \delta_3 \geq 0$, $\delta_3 < 0.5$ satisfying

$$\begin{aligned} &\|[\mathrm{diag}(\mathbf{\Sigma})]^{-1/2}\mathbf{B}\|_F^2 = O(p^{\delta_1}), \quad K = O(p^{\delta_3}), \\ &\|\mathbf{B}^T[\mathrm{diag}(\mathbf{\Sigma})]^{-1}\mathbf{B}\| \cdot \|\{\mathbf{B}^T[\mathrm{diag}(\mathbf{\Sigma})]^{-1}\mathbf{B}\}^{-1}\| = O(p^{\delta_2}), \\ &\|[\mathrm{diag}(\mathbf{\Sigma})]^{-1}\mathbf{\Psi}\| \leq 1. \end{aligned} \tag{11}$$

Theorem 1 gives a sufficient condition to ensure that the number of $\mathbf{R}$'s eigenvalues greater than 1 is equal to the number of common factors. Note that (11) imposes a restriction on the condition number of the matrix $\mathbf{B}^T[\mathrm{diag}(\mathbf{\Sigma})]^{-1}\mathbf{B}$. Without loss of generality, assume $\mathrm{diag}(\mathbf{\Sigma}) = \mathbf{I}_p$. Then $\mathbf{R} = \mathbf{B}\mathbf{B}^T + \mathbf{\Psi}$ and the conditions (11) become

$$K = O(p^{\delta_3}),\ \|\mathbf{\Psi}\| \leq 1,\ \|\mathbf{B}\|_F^2 = O(p^{\delta_1}),\ \|\mathbf{B}^T\mathbf{B}\| \cdot \|(\mathbf{B}^T\mathbf{B})^{-1}\| = O(p^{\delta_2}).$$

3 Properties of sample correlation matrix under factor model

Let $\mathbf{y}_1, \ldots, \mathbf{y}_n$ be an i.i.d. sample of size n from (7):

$$\mathbf{y}_i = \boldsymbol{\alpha} + \mathbf{B}\mathbf{f}_i + \boldsymbol{\epsilon}_i,\ i \in [n] = \{1, \cdots, n\}.$$

Then the sample covariance matrix and sample correlation matrix are

$$\begin{aligned} \hat{\mathbf{\Sigma}}_n &= n^{-1}\sum_{i=1}^{n}(\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})^T, \\ \hat{\mathbf{R}} &= [\mathrm{diag}(\hat{\mathbf{\Sigma}}_n)]^{-1/2}\hat{\mathbf{\Sigma}}_n[\mathrm{diag}(\hat{\mathbf{\Sigma}}_n)]^{-1/2}, \end{aligned} \tag{12}$$

where $\bar{\mathbf{y}} = n^{-1}\sum_{i=1}^{n}\mathbf{y}_i$ is the sample mean. Let the empirical spectral distributions (*ESD*) of $\hat{\mathbf{R}}$ and $\mathbf{R}$ be $F_n(t)$ and $H_{p-K}(t)$ as follows:

$$F_n(t) = \frac{1}{p-K}\sum_{j=K+1}^{p} 1(\lambda_j(\hat{\mathbf{R}}) \le t), \quad H_{p-K}(t) = \frac{1}{p-K}\sum_{j=K+1}^{p} 1(\lambda_j(\mathbf{R}) \le t), \tag{13}$$

for any real number t with $1(\cdot)$ being an indicator function.

3.1 Spectral properties of sample correlation matrix

In order to estimate the number of common factors, we first derive some fundamental results in random matrix theories: the Stieltjes equation of the limiting spectral distribution (*LSD*) $F(t)$ of the ESD $F_n(t)$ and the almost sure convergence of sample spiked eigenvalues $\lambda_1(\hat{\mathbf{R}}), \cdots, \lambda_K(\hat{\mathbf{R}})$ of $\hat{\mathbf{R}}$. There are some existing literatures on the spectral properties of $\hat{\mathbf{R}}$ when $\mathbf{R}$ is of special structures. Bao, Pan and Zhou (2011) derived the Tracy-Widom law of the maximum eigenvalue of $\hat{\mathbf{R}}$ as $\mathbf{R} = \mathbf{I}_p$ and $p/n \to \rho \in (0, \infty)$. El Karoui (2007) established the LSD of $\hat{\mathbf{R}}$ for the elliptical distribution as $p/n \to \rho \in (0, \infty)$ with the bounded spectral norm $\|\mathbf{R}\|$. Gao et al. (2017) obtained the central limit theorem of $\hat{\mathbf{R}}$ for the case $\mathbf{R} = \mathbf{I}_p$ and $p/n \to \rho \in (0, \infty)$. However, for the general factor model (7), the population correlation matrix is not $\mathbf{I}_p$. Theorem 2 below gives the Stieltjes equation of the LSD of $\hat{\mathbf{R}}$ for general case. For convergence of sample spiked eigenvalues $\lambda_1(\hat{\mathbf{R}}), \cdots, \lambda_K(\hat{\mathbf{R}})$, we have not found the related literatures and Theorem 3 below fills the void.

In order to derive Theorems 2-3, additional assumptions are needed.

Assumption (a). Letting $\mathbf{x}_i = (x_{1i}, \cdots, x_{p+K,i})^T = (f_{1i}, \cdots, f_{Ki}, e_{1i}, \cdots, e_{pi})^T$, $(e_{1i}, \cdots, e_{pi}) = (\epsilon_{1i}, \cdots, \epsilon_{pi})\mathbf{\Psi}^{-1/2}$, $\{x_{ji}, j \in [p+K], i \in [n]\}$ are independent random variables satisfying:

$$\frac{1}{n(p+K)\eta_n^4}\sum_{j=1}^{p+K}\sum_{i=1}^{n}\mathrm{E}|x_{ji}^4|1(|x_{ji}| > \eta_n\sqrt{n}) \to 0, \tag{14}$$

where $0 < K\eta_n \to 0$ and $K\eta_n \log n \to +\infty$.

Assumption (b). $\sup_{j\in[p+K]} \mathrm{E}(|x_{j1}|^{6+\delta_0})$ is bounded for all p, K for some $\delta_0 > 0$.

Assumption (c). The ratio of dimension to sample size $\rho_n = p/n \to \rho \in (0, \infty)$ as $n \to \infty$.

Assumption (d). The number of common factors satisfies $K = o(p^{1/6})$.

Assumption (e). $\|[\mathrm{diag}(\mathbf{\Sigma})]^{-1}\mathbf{\Psi}\| \leq 1$ and the limiting spectral distribution $H(t)$ of the ESD $H_{p-K}(t)$ from the eigenvalues $\lambda_{K+1}(\mathbf{R}), \cdots, \lambda_p(\mathbf{R})$ of $\mathbf{R}$ exists.

Remark 1 *The assumption (14) is the Lindeberg condition. By Theorem 1, it is known that $\lambda_j(\mathbf{R}) \leq 1$ for $j = K+1, \cdots, p$ if $\|[\mathrm{diag}(\mathbf{\Sigma})]^{-1}\mathbf{\Psi}\| \leq 1$. Thus, the support set of $H(t)$ is in $[0, 1]$.*

Lemma 1 *For the high dimensional factor model (7) satisfying Conditions C1-C2-C3-C4-C5, under Assumptions (a)-(b)-(c)-(d)-(e), we have*

$$\max_{j\in[p]} |\hat{\sigma}_{jj} - 1| = o_{a.s.}(1),$$

where $\hat{\sigma}_{jj} = n^{-1}\sum_{i=1}^n \mathbf{e}_j^T\mathbf{Q}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T\mathbf{Q}^T\mathbf{e}_j$ with $\bar{\mathbf{x}} = n^{-1}\sum_{i=1}^n \mathbf{x}_i$, $\mathbf{Q}$ being defined in (9) and $\mathbf{e}_j$ is the jth column of $\mathbf{I}_p$.

The proof of Lemma 1 is given in Appendix D. As we impose weak moment conditions, we need to use the truncation tricks and hence the proof is somewhat lengthy. For $z \in \mathcal{C}^+$, let the Stieltjes transform be

$$\begin{aligned} m_n(z) &= (p-K)^{-1}\sum_{j=K+1}^{p}(\lambda_j(\hat{\mathbf{R}}) - z)^{-1} = \int \frac{1}{t-z}dF_n(t), \\ \underline{m}_n(z) &= \int \frac{1}{t-z}d\underline{F}_n(t) = -(1-\rho_{K,n-1})z^{-1} + \rho_{K,n-1}m_n(z), \\ m(z) &= \int \frac{1}{t-z}dF(t), \quad \underline{m}(z) = \int \frac{1}{t-z}d\underline{F}(t) = -(1-\rho)z^{-1} + \rho m(z), \end{aligned} \tag{15}$$

where $F_n(z)$ is defined in (13), $\rho_{K,n-1} = (p-K)/(n-1)$, $\underline{F}_n(x) = (1-\rho_{K,n-1})1(x > 0) + \rho_{K,n-1}F_n(x)$, $\underline{F}(x) = (1-\rho)1(x>0) + \rho F(x)$ and $\mathcal{C}^+$ denotes the upper plane of the two-dimensional complex space. Then $\underline{m}(z)$ and $m(z)$ satisfy the equations (16) and (17).

Theorem 2 *For the high dimensional factor model (7) satisfying Conditions C1-C2-C3-C4-C5 and Assumptions (a)-(b)-(c)-(d)-(e), we have*

$$|m_n(z) - m(z)| = o_{a.s.}(1), \quad |\underline{m}_n(z) - \underline{m}(z)| = o_{a.s.}(1),$$

$$z = -\frac{1}{\underline{m}(z)} + \rho \int \frac{tdH(t)}{1+t\underline{m}(z)} = -\underline{m}^{-1}(z)\psi(-\underline{m}^{-1}(z)), \tag{16}$$

where $z \in \mathcal{C}^+$ *and*

$$\psi(x) = 1 + \rho \int \frac{t}{x-t} dH(t). \tag{17}$$

3.2 Bias correction of sample eigenvalues

Let $\hat{\lambda}_j = \lambda_j(\hat{\mathbf{R}})$ and $\lambda_j = \lambda_j(\mathbf{R})$ for $j \in [p]$. For any given j , define

$$\begin{aligned} m_{n,j}(z) &= (p-j)^{-1}[\sum_{\ell=j+1}^{p} (\hat{\lambda}_\ell - z)^{-1} + ((3\hat{\lambda}_j + \hat{\lambda}_{j+1})/4 - z)^{-1}], \\ \underline{m}_{n,j}(z) &= -(1-\rho_{j,n-1})z^{-1} + \rho_{j,n-1} m_{n,j}(z), \end{aligned}$$

with $\rho_{j,n-1} = (p-j)/(n-1)$. Let the corrected eigenvalue of $\hat{\lambda}_j$ be

$$\hat{\lambda}_j^C = -\frac{1}{\underline{m}_{n,j}(\hat{\lambda}_j)}, \ j \in [r_{\max}].$$

The following theorem, whose proof is given in Appendix I, shows that the corrected empirical eigenvalues are consistent.

Theorem 3 *For the high dimensional factor model (7) satisfying Conditions C1-C2-C3-C4-C5 and Assumptions (a)-(b)-(c)-(d)-(e), for $j \in [K]$, if $\lambda_j \geq \lambda_{K+1}(\mathbf{R})(1+\sqrt{\rho}) + \delta$ for some $\delta > 0$,*

$$\frac{\hat{\lambda}_j^C}{\lambda_j} = 1 + o_p(1) \quad \textit{and} \quad \frac{\hat{\lambda}_j}{\lambda_j} = \psi(\lambda_j) + o_p(1), \tag{18}$$

In particular, if in addition λ_j is bounded for $j \in [K]$, we have

$$\hat{\lambda}_j^C = \lambda_j + o_p(1) \quad \textit{and} \quad \hat{\lambda}_j = \lambda_j \psi(\lambda_j) + o_p(1).$$

Remark 2 *By Remark 1 and (10), we have $\lambda_j > 1$ for $j \in [K]$ under the conditions of Theorem 1 and the support of $H(t)$ being in $[0, 1]$. By (17) and (18), if $\lambda_j > \lambda_{K+1}(1+\sqrt{\rho})$ and is bounded, we have*

$$\hat{\lambda}_j - \lambda_j = \lambda_j \psi(\lambda_j) - \lambda_j = \rho \int \frac{\lambda_j t}{\lambda_j - t} dH(t) + o_p(1), \ j \in [K].$$

In other words, the sample eigenvalue $\hat{\lambda}_j$ is not a consistent estimator of λ_j for $j \in [K]$. This is due to the inconsistency of the high dimensional sample correlation matrix. On the other hand, from (18), we show that the corrected eigenvalue $\hat{\lambda}_j^C$ is consistent for $j \in [K]$.

4 Minimum signals and optimal threshold

We will adopt the notation and estimator defined in the introduction. Our aim is to find minimal signal strength v_0 for consistent estimation of the number of factors and to give the optimal threshold level for our estimator.

4.1 Minimal signal strength

The following theorem shows the minimal signal strength.

Theorem 4 (**Minimal signal strength** v_0). *For the high dimensional factor model (7) satisfying Conditions C1-C2-C3-C4-C5 and Assumptions (a)-(b)-(c)-(d)-(e), and for any estimation method $\hat{K}_{any}$ of the number of common factors by detecting the difference between $\{\lambda_j(\hat{\mathbf{R}}), j \in [K]\}$ and $\{\lambda_j(\hat{\mathbf{R}}), j = K+1, \cdots, p\}$, it holds that*

$$\limsup_{n\to\infty} \inf_{\mathbf{R}\in\mathcal{F}(v_0)} P(\hat{K}_{any} = K) < 1,$$

if $v_0 < 1 + \sqrt{\rho}$.

Proof. Let us take ϵ such that $(1-\epsilon)(1+\sqrt{\rho}) > \max\{1, v_0\}$ and $\mathbf{R} \in \mathcal{F}(v_0)$ such that $\mathbf{R} = \mathbf{\Sigma} = \mathrm{diag}(\mathbf{R})$ and

$$\lambda_p(\mathbf{R}) = \cdots = \lambda_{K+1}(\mathbf{R}) = 1 - \epsilon < \lambda_K(\mathbf{R}) = (1-\epsilon)(1+\sqrt{\rho}).$$

Then by (18), we have $\lambda_K(\hat{\mathbf{R}}) = (1-\epsilon)(1+\sqrt{\rho})^2 + o_p(1)$ because $H(t)$ is the limit of the empirical distribution function of $\{1-\epsilon, \cdots, 1-\epsilon\}$. By (S.46) in the supplementary material, we have $|\lambda_{K+1}(\hat{\mathbf{R}}) - \lambda_{K+1}(\mathbf{S}_n)| = o_{a.s.}(1)$. By Theorem 1.1 of Baik and Silverstein (2006), we have $(1-\epsilon)^{-1}\lambda_{K+1}(\mathbf{S}_n) = (1+\sqrt{\rho})^2 + o_{a.s.}(1)$. Thus we have

$$\lambda_{K+1}(\hat{\mathbf{R}}) = (1-\epsilon)(1+\sqrt{\rho})^2 + o_{a.s.}(1). \tag{19}$$

Hence, when n, p are large enough, $\lambda_K(\hat{\mathbf{R}})$ and $\lambda_{K+1}(\hat{\mathbf{R}})$ will be indifferentiable. That is, when n, p are large enough, the difference between $\lambda_K(\hat{\mathbf{R}})$ and $\lambda_{K+1}(\hat{\mathbf{R}})$ can't be detected by any method.

Remark 3 *The above theorem shows that $v_0 = 1 + \sqrt{\rho}$ is the minimal signal strength. Thus, throughout the rest of the paper, we will consider the estimation method in the set of the correlation matrix $\mathbf{R}$ as follows:*

$$\begin{aligned}\mathcal{F}(1+\sqrt{\rho}) \;=\; \{\mathbf{R}:\;\; & \mathbf{R} \textit{ is the correlation matrix (8) of the observed vector} \\ & \textit{in the factor model (7) and } \lambda_K(\mathbf{R}) > 1+\sqrt{\rho}\}.\end{aligned}$$

4.2 Optimal threshold

Recall our estimation method,

$$\hat{K}^C(s) = \max\{j : \hat{\lambda}_j^C > s, j \in [r_{\max}]\}, \tag{20}$$

where $r_{\max}$ is a pre-specified positive integer and the maximum of the empty set is defined as 0. The following theorem establishes the optimal bound of the threshold s.

Theorem 5 *For the high dimensional factor model (7) satisfying Conditions C1-C2-C3-C4-C5 and Assumptions (a)-(b)-(c)-(d)-(e), we have*

$$\limsup_{n\to\infty} \sup_{\mathbf{R}\in\mathcal{F}(1+\sqrt{\rho})} P(\hat{K}^C(s) > K) = 1, \quad \textit{if } s < 1+\sqrt{\rho}, \tag{21}$$

$$\limsup_{n\to\infty} \sup_{\mathbf{R}\in\mathcal{F}(1+\sqrt{\rho})} P(\hat{K}^C(s) < K) = 1, \quad \textit{if } s > 1+\sqrt{\rho}, \tag{22}$$

where s doesn't depend on n and p.

Proof. To (21), let $(1-\epsilon)(1+\sqrt{\rho}) > s$ and $\mathbf{R} \in \mathcal{F}(1+\sqrt{\rho})$ satisfy

$$\lambda_p(\mathbf{R}) = \ldots = \lambda_{K+1}(\mathbf{R}) = 1-\epsilon < 1+\sqrt{\rho} < \lambda_K(\mathbf{R}).$$

By (16) and (18), we have $\lambda_{K+1}(\hat{\mathbf{R}}) = \hat{\lambda}_{K+1}^C + \rho\frac{\hat{\lambda}_{K+1}^C(1-\epsilon)}{\hat{\lambda}_{K+1}^C-(1-\epsilon)} + o_p(1)$, because $H(t)$ is the limit of the empirical distribution function of $\{1-\epsilon, \cdots, 1-\epsilon\}$. It then follows from $\lambda_{K+1}(\hat{\mathbf{R}}) = (1-\epsilon)(1+\sqrt{\rho})^2 + o_p(1)$ that

$$\hat{\lambda}_{K+1}^C = (1-\epsilon)(1+\sqrt{\rho}) + o_p(1).$$

That is, $\hat{\lambda}_j^C \geq (1-\epsilon)(1+\sqrt{\rho}) + o_p(1),\ j \in [K+1]$. By using $(1-\epsilon)(1+\sqrt{\rho}) > s$, we conclude that

$$\limsup_{n\to\infty} \sup_{\mathbf{R}\in\mathcal{F}(1+\sqrt{\rho})} P(\hat{K}^C(s) > K) = 1,$$

when $s < 1 + \sqrt{\rho}$.

To prove (22), let $\mathbf{\Sigma} = \mathbf{R} \in \mathcal{F}(1 + \sqrt{\rho})$ satisfy

$$\lambda_p(\mathbf{R}) \leq \cdots \leq \lambda_{K+1}(\mathbf{R}) \leq 1 < 1 + \sqrt{\rho} < \lambda_K(\mathbf{R}) < s \leq \cdots \leq \lambda_1(\mathbf{R}).$$

By (18) and $\lambda_K(\mathbf{R}) < s$, we have $\hat{\lambda}_K^C = \lambda_K(\mathbf{R}) + o_p(1) < s + o_p(1)$ which means $\hat{\lambda}_j^C < s,\ j = p, \cdots, K$ in probability. Thus

$$\limsup_{n\to\infty} \sup_{\mathbf{R}\in\mathcal{F}(1+\sqrt{\rho})} P(\hat{K}^C(s) < K) = 1,$$

if $s > 1 + \sqrt{\rho}$. ■

Theorem 5 shows that the choices $s < 1 + \sqrt{\rho}$ and $s > 1 + \sqrt{\rho}$ are not optimal for the threshold parameter s in our estimation method. The following theorem will show that $s = 1 + \sqrt{\rho}$ is optimal.

Theorem 6 *For the high dimensional factor model (7) satisfying Conditions C1-C2-C3-C4-C5 and Assumptions (a)-(b)-(c)-(d)-(e), for* $\mathbf{R} \in \mathcal{F}(1+\sqrt{\rho})$*, we have when* $s = 1+\sqrt{\rho}$*,*

$$P(\hat{K}^C(s) = K) \to 1. \tag{23}$$

Proof. For any $\mathbf{R} \in \mathcal{F}(1 + \sqrt{\rho})$, we have

$$\lambda_p(\mathbf{R}) \leq \cdots \leq \lambda_{K+1}(\mathbf{R}) \leq 1 < 1 + \sqrt{\rho} + \epsilon_0 < \lambda_K(\mathbf{R}) \leq \cdots \leq \lambda_1(\mathbf{R}),$$

for a very small positive constant ϵ_0. By (16), we have $\lambda_{K+1}(\hat{\mathbf{R}}) = \hat{\lambda}_{K+1}^C + \rho \int \frac{\hat{\lambda}_{K+1}^C t}{\hat{\lambda}_{K+1}^C - t} dH(t) + o_p(1)$. Thus, we have

$$\hat{\lambda}_{K+1}^C \leq 1 + \sqrt{\rho} + o_p(1), \tag{24}$$

because of $\lambda_{K+1}(\hat{\mathbf{R}}) \leq (1+\sqrt{\rho})\psi(1+\sqrt{\rho}) + o_p(1)$ by Lemma S.6 in the supplementary material. By (18), we have

$$\hat{\lambda}_j^C \geq 1 + \sqrt{\rho} + \epsilon_0 + o_p(1),\ j \in [K]. \tag{25}$$

Thus, by (24) and (25), when $s = 1 + \sqrt{\rho}$, we have

$$\lim_{n\to\infty} P(\hat{K}^C(s) = K) = 1.$$

Summary of Method: We propose

$$\hat{K} = \max\{j : \hat{\lambda}_j^C > 1 + \sqrt{\rho_{n-1}}, j \in [r_{\max}]\},$$

where $\rho_{n-1} = p/(n-1)$. This is a simple and tuning free method.

5 Simulation studies

We evaluate the finite-sample performance of the proposed method by simulation studies. Because our proposed estimating method is based on the adjusted correlation thresholding, we will label the proposed estimating method as "ACT". We compare our ACT method with 13 existing methods: "PC_1", "PC_2","PC_3","IC_1", "IC_2" and "IC_3" in Bai and Ng (2002), "ON_1", "ON_2" and "ON_3" in Onasti (2005), "NON" in Onatski (2010), "ER" and "GR" in Ahn and Horenstein (2013). Due to the similarity of simulation results, we only present PC_3, IC_3, ON_2, ER, GR and ACT. The sample sizes are taken to be $n = 150, 300$ and the dimension is $p = 100, 300, 500, 1000$. Recall the factor model $\mathbf{y} = \mathbf{Bf} + \boldsymbol{\epsilon}$ in (7). For the Gaussian population, assume that ϵ_i's are iid from $N(0, \nu_i^2)$ and $f_1, \cdots, f_K$ are iid from $N(0,1)$. For the uniform population, assume that ϵ_i are iid from Unif$(0, 2\sqrt{3\nu_i^2})$ and

$f_1, \cdots, f_K$ are iid from Unif$(0, 2\sqrt{3})$. We set the true number of common factors $K = 5$. For every case, we conduct 1000 replications to summarize the empirical percentages of true estimation, overestimation, underestimation of the number of common factors, and the average number of common factors. We consider the following four cases for the factor loading matrix $\mathbf{B} = (b_{\ell j})_{\ell \in [p], j \in [K]}$. They can be verified to satisfy the imposed conditions, in particular, those in Theorem 1.

Case 1: Let $b_{\ell j} = \sqrt{3p^{-1/2}}$ for $\ell, j \in [K]$ and $b_{\ell j} = a_{\ell j}\sqrt{3(p-j)^{-1}}$ for $\ell \in \{K+1, \cdots, p\}, j \in [K]$ and $a_{\ell j} = -1$ if $\ell = Kj$ or $a_{\ell j} = 1$ if $\ell \neq Kj$. Assume that $\nu_1^2 = \cdots = \nu_p^2 = 0.55^2$. The model is from Harding (2013).

Case 2: Let $b_{\ell j}$ be iid from $N(0, 1)$ and $\nu_1^2, \cdots, \nu_p^2$ be iid from Unif$(0, 180)$.

Case 3: Let $b_{\ell j}$ be iid from $N(0, 1)$ and $\nu_1^2 = \cdots = \nu_p^2 = 36$. The model is used in Bai and Ng (2002) and Onatski (2010).

Case 4: Let $b_{jj} = 1$, $b_{\ell j}$ be iid from $N(0, 0.04)$ for $j \neq \ell$ and $\nu_1^2, \cdots, \nu_p^2$ be iid from Unif$(0, 5.5)$.

The simulation results for Cases 1–4 with $n = 300$ are presented respectively in Tables 2–5. The results for the cases with $n = 150$ are similar and are omitted. From these tables, we can see that except for very few settings, "ACT" behaves very well for almost all parameter setups. Even for these few settings, the percentiles of true estimation of "ACT" are also similar to those of "ER" and "GR".

Table 2: Percentages of the estimated number of common factors for Case 1 with $n = 300$ in 1000 simulations: "TRUE", "OVER" and "UNDER" truly estimates, overestimates and underestimates the number of common factors, respectively. "AVE" is the average of the estimated number of common factors.

p		PC_3	IC_3	ON_2	ER	GR	ACT
		Gaussian population					
100	TRUE	99.8	87.6	100	41.8	75.7	100
	OVER	0	0	0	0	0	0
	UNDER	0.2	12.4	0	58.2	24.3	0
	AVE	5	4.88	5	2.87	4.37	5
300	TRUE	92.0	55.9	99.9	4.2	8.7	100
	OVER	0	0	0.1	0	0	0
	UNDER	8,0	44.1	0	95.8	91.3	0
	AVE	4.92	4.56	5	2.09	2.56	5
500	TRUE	0	0	100	0	0.2	99.6
	OVER	0	0	0	0	0	0.4
	UNDER	100	100	0	100	99.8	0
	AVE	3.88	3.52	5	1.78	1.96	5
1000	TRUE	0	0	79.1	0	0	89.0
	OVER	0	0	0	0	0	1.9
	UNDER	100	100	20.9	100	100	9.1
	AVE	1.81	1.33	4.79	1.34	1.38	4.93
		Uniform population					
100	TRUE	99.9	90.3	99.9	44.7	81.2	100
	OVER	0	0	0.1	0	0	0
	UNDER	0.1	9.7	0	55.3	18.8	0
	AVE	5	4.9	5	2.97	4.52	5
300	TRUE	93.3	62.0	100	4.2	9.0	100
	OVER	0	0	0	0	0	0
	UNDER	6.7	38.0	0	95.8	91.0	0
	AVE	4.93	4.62	5	2.07	2.55	5
500	TRUE	0	0	99.9	0.1	0.3	99.2
	OVER	0	0	0.1	0	0	0.8
	UNDER	100	100	0	99.9	99.7	0
	AVE	3.92	3.55	5	1.75	1.97	5.01
1000	TRUE	0	0	83.3	0	0	89.8
	OVER	0	0	0.1	0	0	1.5
	UNDER	100	100	16.6	100	100	8.7
	AVE	1.82	1.29	4.84	1.32	1.37	4.93

Table 3: Percentages of the estimated number of common factors for Case 2 with $n = 300$ in 1000 simulations: "TRUE", "OVER" and "UNDER" truly estimates, overestimates and underestimates the number of common factors, respectively. "AVE" is the average of the estimated number of common factors.

p		PC_3	IC_3	ON_2	ER	GR	ACT
		Gaussian population					
100	TRUE	0	0	0.1	4.2	4.4	64.3
	OVER	0	0	0	6.6	7.3	0.10
	UNDER	100	100	99.9	89.2	88.3	35.6
	AVE	1.18	1	1.53	2.29	2.37	4.58
300	TRUE	47.0	1.7	31.2	27.0	28.2	98.9
	OVER	0	0	0.1	0.4	0.4	1.1
	UNDER	53.0	98.3	68.7	72.6	71.4	0
	AVE	4.42	2.81	4.17	3.01	3.07	5.01
500	TRUE	0	0	98.8	88.9	89.7	98.9
	OVER	0	0	0	0	0	1.1
	UNDER	100	100	1.2	11.1	10.3	0
	AVE	2.44	1.16	4.99	4.76	4.78	5.01
1000	TRUE	0	0	99.9	99.9	99.9	99.1
	OVER	0	0	0.1	0	0	0.9
	UNDER	100	100	0	0.1	0.1	0
	AVE	1.17	1	5	5	5	5.01
		Uniform population					
100	TRUE	0	0	0.1	5.0	5.4	60.7
	OVER	0	0	0.1	8.4	9.0	0.4
	UNDER	100	100	99.8	86.6	85.6	38.9
	AVE	1.17	1	1.57	2.38	2.45	4.54
300	TRUE	48.4	1.4	37.8	31.7	33.7	99.4
	OVER	0	0	0	0.3	0.4	0.6
	UNDER	51.6	98.6	62.2	68.0	65.9	0
	AVE	4.45	2.83	4.27	3.16	3.25	5.01
500	TRUE	0	0	99.4	91.0	91.6	99.1
	OVER	0	0	0.1	0	0	0.9
	UNDER	100	100	0.5	9.0	8.4	0
	AVE	2.44	1.13	5	4.81	4.83	5.01
1000	TRUE	0	0	99.9	100	100	99.0
	OVER	0	0	0.1	0	0	1.0
	UNDER	100	100	0	0	0	0
	AVE	1.12	1	5	5	5	5.01

Table 4: Percentages of the estimated number of common factors for Case 3 with $n = 300$ in 1000 simulations: "TRUE", "OVER" and "UNDER" truly estimates, overestimates and underestimates the number of common factors, respectively. "AVE" is the average of the estimated number of common factors.

p		PC_3	IC_3	ON_2	ER	GR	ACT
		Gaussian population					
100	TRUE	0	0	0.1	5.5	5.8	0
	OVER	0	0	0	9.6	9.7	0
	UNDER	100	100	99.9	84.9	84.5	100
	AVE	1	1	1.27	2.51	2.54	1.06
300	TRUE	0	0	1.1	4.2	4.6	5.4
	OVER	0	0	0	0.8	0.9	0
	UNDER	100	100	98.9	95	94.5	94.6
	AVE	1	1	2.85	2.1	2.14	2.91
500	TRUE	0	0	32.5	26.0	27.3	71.3
	OVER	0	0	0	0.2	0.2	2.8
	UNDER	100	100	67.5	73.8	72.5	25.9
	AVE	1	1	4.2	2.92	2.97	4.74
1000	TRUE	0	0	99.6	92.3	92.7	96.2
	OVER	0	0	0	0	0	3.8
	UNDER	100	100	0.4	7.7	7.3	0
	AVE	1	1	5	4.81	4.83	5.04
		Uniform population					
100	TRUE	0	0	0	5.0	5.1	0
	OVER	0	0	0	6.8	7.0	0
	UNDER	100	100	100	88.2	87.9	100
	AVE	1	1	1.27	2.33	2.35	1.08
300	TRUE	0	0	0.5	5.2	5.5	4.6
	OVER	0	0	0	1.2	1.3	0.10
	UNDER	100	100	99.5	93.6	93.2	95.30
	AVE	1	1	2.87	2.25	2.28	2.92
500	TRUE	0	0	37.3	31.5	32.6	76.10
	OVER	0	0	0.1	0.2	0.2	1.10
	UNDER	100	100	62.6	68.3	67.2	22.80
	AVE	1	1	4.26	3.08	3.13	4.76
1000	TRUE	0	0	99.8	94.5	94.7	96.8
	OVER	0	0	0.1	0	0	3.2
	UNDER	100	100	0.1	5.5	5.3	0
	AVE	1	1	5	4.88	4.88	5.03

Table 5: Percentages of the estimated number of common factors for Case 4 with $n = 300$ in 1000 simulations: "TRUE", "OVER" and "UNDER" truly estimates, overestimates and underestimates the number of common factors, respectively. "AVE" is the average of the estimated number of common factors.

p		PC_3	IC_3	ON_2	ER	GR	ACT
		Gaussian population					
100	TRUE	0.2	0	0.7	3.9	4.6	98.20
	OVER	0	0	0	1.9	2.4	0.20
	UNDER	99.8	100	99.3	94.2	93	1.60
	AVE	2.4	1	2.85	2.14	2.21	4.99
300	TRUE	99.5	81.7	97.8	81.6	83	99.3
	OVER	0.1	0	0.1	0	0	0.7
	UNDER	0.4	18.3	2.1	18.4	17.0	0
	AVE	5	4.81	4.98	4.55	4.6	5.01
500	TRUE	63.9	18.5	100	99.9	99.9	99.4
	OVER	0	0	0	0	0	0.6
	UNDER	36.1	81.5	0	0.1	0.1	0
	AVE	4.63	3.81	5	5	5	5.01
1000	TRUE	4.9	0.1	99.9	100	100	99.5
	OVER	0	0	0.1	0	0	0.5
	UNDER	95.1	99.9	0.0	0	0	0
	AVE	3.6	2.54	5	5	5	5
		Uniform population					
100	TRUE	0.3	0	1.3	4.7	5.0	96.0
	OVER	0	0	0	2.4	2.8	0.5
	UNDER	99.7	100	98.7	92.9	92.2	3.5
	AVE	2.32	1	2.87	2.21	2.28	4.97
300	TRUE	99.6	88.1	98.8	87.7	88.7	99.6
	OVER	0	0	0.1	0	0	0.4
	UNDER	0.4	11.9	1.1	12.3	11.3	0
	AVE	5	4.88	4.99	4.73	4.76	5
500	TRUE	67.1	18.1	99.8	99.8	99.8	99.7
	OVER	0	0	0.2	0	0	0.3
	UNDER	32.9	81.9	0	0.2	0.2	0
	AVE	4.66	3.85	5	5	5	5
1000	TRUE	6.4	0.2	99.9	100	100	99.3
	OVER	0	0	0.1	0	0	0.7
	UNDER	93.6	99.8	0	0	0	0
	AVE	3.71	2.54	5	5	5	5.01

6 Empirical Studies

This section analyzes two real data on economics and finance to demonstrate our proposed estimation method ACT.

Example 1 (Macroeconomic time series): We use the monthly macroeconomic datasets from March, 1960 to December, 2014 used by McCraken and Ng (2017). Series 64, 66, 101 and 130 are removed because of missing observations. Following McCraken and Ng (2017), outliers are removed where an outlier is defined as an observation that deviates from the sample mean by more than ten interquantile ranges. After the datasets are cleaned, the data dimension is $p = 123$ and the sample size is $n = 583$. McCraken and Ng (2017) used PC_2 to select nine factors by using the sample covariance matrix whose nine largest eigenvalues are 3.91×10^{11}, 1.20×10^{10}, 4.77×10^{9}, 3.06×10^{9}, 7.25×10^{8}, 3.60×10^{8}, 1.38×10^{8} and 2.83×10^{7} and 9.00×10^{6}. However, the marginal variances of these 123 time series vary widely from 2.80×10^{-4} to 1.80×10^{11}, which jeopardizes the fidelity of the covariance matrix based methods. Our estimation method ACT selects six factors by using the sample correlation matrix whose top nine eigenvalues are 73.10, 17.81, 10.22, 7.00, 4.80, 1.97, 1.53, 1.17, 1.06.

In terms percent of variance explained by the selected factors, the 9 selected factors explain 99.99% of total variation, whereas the 6 selected factors explain 99.95% of total variation. This is mainly due to the leading eigenvalue which is an order of magnitude larger than the rest. If we look at the standardized variables (the eigenvalues from the correlation matrix), the selected 9 factors explain 96.49% of total variations whereas the selected 6 factors explain 93.43%.

We now examine whether the number of factors that influences the equity market has changed before and after financial crisis. As an illustration, we use the stationarily

transformed macroeconomic time series (McCraken and Ng, 2017)

Transformed Macro Data Before the Financial Crisis:

We now use the stationarily transformed (McCraken and Ng, 2017) monthly macroeconomic datasets from January, 1960 to December, 2007 with sample size $n = 576$ and $p = 123$. Using PC_2 as in McCraken and Ng (2017), nine factors are selected. The nine largest eigenvalues for the sample covariance are 2.28×10^4, 13.06, 1.53, 0.88, 0.74, 0.40, 0.32, 0.24, 0.18. Again, the marginal sample variances for these 123 transformed series vary widely from 8.47×10^{-7} to 2.28×10^4. On the other hand, our estimation method ACT selects 10 factors by using the sample correlation matrix. The 10 largest eigenvalues of the sample correlation matrix are 18.37, 8.55, 7.66, 6.16, 5.86, 4.01, 3.76, 3.53, 2.89, 2.56. The variances explained by 9 selected and 10 selected factors are both around 99.99% due to a very spike top eigenvalue. In terms of percentage of variance explained by the standardized variables, ten factors explain 51.53% whereas nine factors explain 49.45%.

Transformed Macro Data After the Financial Crisis:

The period covers the data from January, 2010 to October, 2018 with the sample size is $n = 106$ and $p = 123$. Again, PC_2 selects 9 factors. The 9 largest eigenvalues are 5.68×10^4, 2.77, 0.86, 0.35, 0.17, 0.11, 0.08, 0.08, 0.03. Again, the marginal sample variances vary largely from 5.35×10^{-7} to 5.74×10^4. In contrast, our estimation method ACT chooses 7 factors. The 9 largest eigenvalues of correlation matrix are 16.48, 12.20, 9.23, 5.75, 5.68, 5.27, 4.22, 3.82 and 3.53. Moreover, nine selected factors explain 53.83% total variation in 123 series, whereas 7 factors explain 47.85% total variation in 123 series, which is similar to pre-crisis period by using the same method.

Example 2 (100 Fama-French portfolios): We now estimate the number of factors using the excess returns of Fama-French 100 portfolios. The data can be downloaded from

the data library of Professor Kenneth French's website. Again, we divide the data into two periods: before and after financial crisis.

Before the Financial Crisis:

Following Fan et al. (2012), we use the daily returns of 100 industrial portfolios formed on the basis of size and book-to-market ratio from January 2, 1998 to December 31, 2007. We note that the 71th and 100th portfolios have very large variances that possibly jeopardize the covariance matrix based methods. PC_3, IC_3, ON_2, ER and GR estimate the number of factors as 10, 10, 6, 3 and 3, respectively. The largest 10 eigenvalues of the sample covariance matrix are 1824.45, 885.13, 117.39, 9.74, 5.38, 3.17, 2.31, 2.14, 1.86, 1.59. Ten factors explain 98.86% total variation in the 100 portfolios; four factors (suggested by ACT) explain 98.29% total variation in 100 portfolios. On the other hand, ACT selects four factors. The largest 10 eigenvalues of sample correlation matrix are 65.81, 5.74, 2.57, 1.95, 1.10, 0.97, 0.90, 0.83, 0.72, 0.63. Ten factors explain 81.26% total variation, whereas four factors explain 76.09% total variation in 100 portfolios, in the standardized variables.

Table 6: Percent R^2 of well-known risk factors explained by PC-factors

	Rm-Rf	SMB	HML	Momentum
Before crisis (4 selected factors)	0.953	0.931	0.829	0.141
Before crisis (3 selected factors)	0.947	0.813	0.821	0.132
After crisis (3 selected factors)	0.982	0.891	0.917	0.155

The well-known risk factors for equity markets are Fama-French factors (Fama and French, 1993, 2015) and the momentum factor (Carhart, 1997). To examine how these known factors can be explained by the unsupervised learning method (PCA with number of

factors selected by ACT), we regress each known risk factor on the four principal component factors and report the coefficients of determination R^2 in Table 6. As comparison, we also regress these Fama-French factors on the 3 selected factors and report the result in the same table.

First of all, the well-known Fama-French factors (Rm-Rf is the market factor; SMB is the size factor; HML is the value factor; these three factors are nearly uncorrelated) are explained very well by the factors learned from the principal components. Regarding PC-factors as true factors (subject to learning or estimation errors), the results lend further support that the Fama-French factors are three most important factors, spanning essentially the same space as the first three PC-factors (regressing on first 3 PCs yields similar results). Such a confirmation of Fama-French factors appears new. On the other hand, the momentum factors can not be explained well by the first four principal components, which is a surprise. Figure 1 depicts how well these four well-known equity risk factors can be explained by the four principal components. As expected, the four principal components explain the Fama-French factors better than the first four principal components. On the other hand, the Carhart's momentum factor is not supported by the principal components.

We measure the difference between four given risk factors and four learned factors by using their projection matrix. Let A be a $n \times 4$ matrix formed by the time series of the four known factors and B be an $n \times 4$ matrix formed the four principal component factors. Define the projection matrix as $P_A = A(A^TA)^{-1}A^T$ and $P_B = B(B^TB)^{-1}B^T$. We then measure the difference between the space spanned by the four well-known factors and four learned PC-factors by using the Operator norm and Frobenius norm. For our data, they are

$$\|P_A - P_B\|_2 = 0.973, \quad \|P_A - P_B\|_F = 1.591.$$

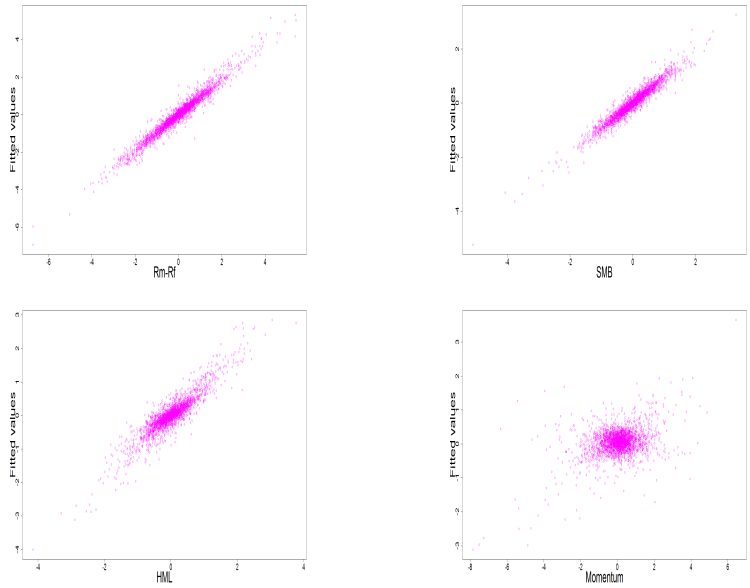

Figure 1: Graphical display for every observable factor v.s. its fitted values by regressing every observable factor on the four principal component factors before economic crisis.

In a similar vein, we measure the difference between the 3 Fama-French factors (A) and the three principal factors (B) by using the projection matrices. They are

$$\|P_A - P_B\|_2 = 0.481, \quad \|P_A - P_B\|_F = 0.949.$$

After the Financial Crisis:

We extrac the data from January 4, 2010 to April 30, 2019. The covariance matrix based methods PC_3, IC_3, ON_2, ER and GR estimate number of factors as 6, 6, 6, 1 and 1, respectively. On the other hand, ACT selects three factors which explain 85.90% total variation in 100 portfolios with three largest eigenvalues being 80.62, 3.22 and 2.06 based on the sample correlation matrix.

The R^2 of each the four well-known risk factors determined by three principal component factors is depicted in Table 6. Again, this confirms once more that the famous Fama-French factors aligned well with the first 3 principal components. Indeed, the differences between the two spaces are

$$\|P_A - P_B\|_2 = 0.406, \quad \|P_A - P_B\|_F = 0.708,$$

smaller than what it is before the financial crisis. On the other hand, the momentum factors are still not explained well by the PC factors.

7 Conclusions

Based on the sample correlation matrix, this paper discovers the equality between the number of eigenvalues exceeding one and the number of latent factors. To utilize such a relationship, we study the random matrix theory based on the sample correlation in order

to correct the biases in estimating the top eigenvalues and to take into account of estimation errors in eigenvalue estimation. This gives rise naturally to the adjusted correlation thresholding (ACT) for determining the number of common factors in high dimensional factor models. The estimation method overcomes the disadvantages of using the sample covariance matrix which allows observable variables incomparable in their scales. Simulation studies show that our proposed estimation method outperforms competing methods in the literature. This paper considers the iid samples from the static factor model. But in practice, people also care about the dynamic factor model. Our future work will establish the relationship between the population correlation matrix and the number of common factors in the dynamic factor models, and propose estimating the number of factors in the high dimensional dynamic factor model.

SUPPLEMENTARY MATERIAL

Title: Supplementary material for "Estimating Number of Factors by Adjusted Eigenvalues Thresholding". The material includes 8 lemmas and their proofs, and the proofs of Lemma 1 and Theorem 1, 2, 3. (SuppleFileFactor.pdf)

R codes for ACT: R codes are used for simulation studies in Section 5 and empirical studies in Section 6 (simuexam zipped file).

Data sets: Data sets are used in empirical studies in Section 6. (data zipped file)

References

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

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

Baik, J. and J. W. Silverstein (2006). Eigenvalues of large sample covariance matrices of spiked population models. *Journal of Multivariate Analysis 97*, 1382–1408.

Cai, T. T., X. Han, and G. M. Pan (2017). Limiting laws for divergence spiked eigenvalues and largest non-spiked eigenvalue of sample covariance matrices. *arXiv:1711.00217v2*.

Carhart, M. M. (1997). On persistence in mutual fund performance. *The Journal of finance 52*(1), 57–82.

El Karoui, N. (2007). On spectral properties of large dimensional correlation matrices and covariance matrices computed from elliptically distributed data. Technical report, Technical report from Department of Statistics, University of California, Berkeley.

Fama, E. F. and K. R. French (1993). Common risk factors in the returns on stocks and bonds. *Journal of financial economics 33*(1), 3–56.

Fama, E. F. and K. R. French (2015). A five-factor asset pricing model. *Journal of financial economics 116*(1), 1–22.

Fan, J., K. Wang, Y. Zhong, and Z. Zhu (2018). Robust high dimensional factor models with applications to statistical machine learning. *Statistical Science pending revision*.

Fan, J., J. Zhang, and K. Yu (2012). Vast portfolio selection with gross-exposure constrains. *Journal of the American Statistical Association 107*(498), 592–606.

Gao, J. T., X. Han, G. M. Pan, and Y. R. Yang (2017). High-dimensional correlation matrices: the central limit theorem and its application. *Journal of the Royal Statistical Society (Series B) 79*(3), 677–693.

Guttman, L. (1954). Some necessary conditions for common-factor analysis. *Psychometrika 19*(2), 149–161.

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

Harding, M. (2013). Estimating the number of factors in large dimensional factor models (manuscript).

Johnson, R. A. and D. W. Wichern (2007). *Applied Multivariate Statistical Analysis*. 6th Edition. Prentice Hall.

Kaiser, H. K. (1960). The application of electronic computers fot factor analysis. *Educational and Psychological Measurement XX*(1), 141–151.

Kaiser, H. K. (1961). A note on guttman's lower bound for the number of common factors. *The British Journal of Statistical Psychology XIV*(1), 1–2.

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

Kong, X. B., Z. Liu, and W. Zhou. A rank test for the number of factors with high-frequency data. *Journal of Econometrics 211*(2), 439–460.

Lam, C. and Q. Yao (2012). Factor modeling for high-dimensional time series: inference for the number of factors. *The Annals of Statistics 40*(40), 694–726.

Lewbel, A. (1991). The rank of demand systems: theory and nonparametric estimation. *Econometrica 59*, 711–730.

Li, H. J., Q. Li, and Y. T. Shi (2017). Determinging the number of factors when the number of factors can increase with sample size. *Journal of Econometrics 197*, 76–86.

McCraken, M. W. and S. Ng (2017). Fred-md: A monthly databased for macroeconomic research. *Journal of Business and Economic Statistics 79*(3), 677–693.

Onatski, A. (2005). A formal statistical test for the number of factors in the approximate factor models. *mimeo Columbia University[399, 400]*.

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

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

Pan, J. Z. and Q. W. Yao (2008). Modelling multiple time series via common factors. *Biometrika 95*(2), 365–379.

Su, L. J. and X. Wang (2017). On time-varying factor modesl: estimation and testing. *Journal of Econometrics 199*, 84–101.

Wang, H. (2012). Factor profiling for ultra high dimensional variable selection. *Biometrika 99*, 15–28.