Non-Robustness Issue for Estimating the Number of Factors in High Dimensional Data

Zhenhao Gong

University of Connecticut

September 18, 2021

Overview

1. Introduction
2. Literature Review
3. Basic Factor Model and PCA
4. Strong and Weak Factors Estimation
5. Simulation Design for Non-Robustness Issue
6. Numerical Results and Conclusion

Introduction

- Factor analysis is one of the cores method for handling large data matrices and high dimensional data, with wide applications in macroeconomics, finance and other areas.
- A big challenge in factor analysis is how to estimate the number of factors. Most methods for estimating the number of factors are based on the results from random matrix theory (RMT), which require i.i.d and gaussian assumption on the error terms.
- The theme of my third year paper is to show that whether those methods for estimating the number strong and weak factors are robust by simulation when the error terms are serially and cross-sectionally correlated or have non-gaussian distributions.

Overview

1 Introduction

2 Literature Review

3 Basic Factor Model and PCA

4 Strong and Weak Factors Estimation

5 Simulation Design for Non-Robustness Issue

6 Numerical Results and Conclusion

Classical methods

1. Under the asymptotic regime that N is fixed and $T \to \infty$

- Sequential hypotheses test: Lawley, 1956; Bartelett, 1950; Anderson and Rubin, 1956;
- Scree test: Cattell, 1966; Cattell and Vogelmann, 1977;
- Kaiser's rule: (Kaiser, 1960);
- Parallel analysis (PA): Horn, 1965; Buja and Eyuboglu, 1992;
- Information criteria based methods such as Bayesian Information Criteria (BIC) and Akaike Information Criteria (AIC): Wax and Kailath, 1985; Fishler et al., 2002;

The fundamental problem the for those classical factor estimation methods is that they does not apply when both N and $T \to \infty$.

Methods for Large Matrices

2. Under the asymptotic regime that both $N, T \to \infty$, methods for estimating the number of strong factors

- Information criteria based methods: Bai and Ng (2002), Bai (2003); later improved by L. Alessi et al (2010);
- Eigenvalues difference based method: Onatski (2010);
- Eigenvalues ratio based method: Ahn and Horenstein (2013);

3. Methods for estimating the number of weak factors

- Information criteria based method: Nadakuditi and Edelman (2008);
- Bi-cross-validation based method: Owen and Wang (2015);

Overview

1. Introduction
2. Literature Review
3. Basic Factor Model and PCA
4. Strong and Weak Factors Estimation
5. Simulation Design for Non-Robustness Issue
6. Numerical Results and Conclusion

Basic Factor Model

- Let Y_{it} be the observed data for the ith cross-section unit at time t, for $i = 1, 2, \cdots, N$ and $t = 1, \cdots, T$. The factor model for Y_{it} is given by

$$Y_{it} = X_{it} + e_{it} = L_i^{'} F_t + e_{it}, \tag{1}$$

  where r_0 is the true number of factors, F_t is a $(r_0 \times 1)$ vector of common factors, L_i is a $(r_0 \times 1)$ vector of loadings associated with F_t, and e_{it} is the idiosyncratic errors of Y_{it}. Note that the factors, their loadings, as well as the idiosyncratic errors are not observable.

- The factor model (1) can be put in a matrix form as:

$$Y_t = LF_t + e_t, \tag{2}$$

  with $L = (L_1, L_2, \cdots, L_N)^{'} \in \mathbb{R}^{N \times r_0}$.

PCA and SVD in Factor Analysis

- As a common statistical method for dimension reduction of the data, principle component analysis (PCA) is closely related to factor analysis. Basically, PCA tries to find linear combinations of the observed variables to maximize the sample variance.
- Let $\widehat{\Sigma}_Y = YY^{'}/T$ be the sample covariance matrix corresponding to the population covariance matrix $\Sigma_Y = E(Y_t Y_t')$ assuming $E(Y_i) = 0$ for $i = 1, 2, \cdots, N$. Let the eigenvalue decomposition of $\widehat{\Sigma}_Y$ be $\widehat{\Sigma}_Y = P\Lambda P^{'}$, where $\Lambda = \text{diag}(\lambda_1, \lambda_2, \cdots, \lambda_N)$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N \geq 0$ and P is an orthogonal $N \times N$ matrix.
- Then the columns of P are eigenvectors of $\widehat{\Sigma}_Y$, and the rows of $P'Y$ are called principle components (PCs).

PCA and SVD in Factor Analysis

- Similar with PCA, one can use the singular value decomposition (SVD) of Y to derive eigenvectors and PCs.
- Let $Y = \sqrt{T}\hat{U}\hat{D}\hat{V}'$ where $\hat{U} \in \mathbb{R}^{N \times \min(N,T)}$, $\hat{V} \in \mathbb{R}^{T \times \min(N,T)}$, and $\hat{D} = \text{diag}(\hat{d}_1, \hat{d}_2, \cdots, \hat{d}_{\min(N,T)})$, with $\hat{U}'\hat{U} = \hat{V}'\hat{V} = I_{\min(N,T)}$, and $\hat{d}_1 \geq \hat{d}_2 \geq, \cdots, \geq \hat{d}_{\min(N,T)}$.
- Note that

$$\begin{aligned}\hat{\Sigma}_Y = YY'/T &= \hat{U}\hat{D}\hat{V}'\hat{V}\hat{D}\hat{U}' \\ &= \hat{U}(\hat{D})^2\hat{U}',\end{aligned}$$

so it is clear that the first r ($r \leq \min(N, T)$) eigenvectors of $\hat{\Sigma}_Y$ are columns of $\hat{U}$ and the first r principle components are rows of $\hat{U}'Y = \sqrt{T}\hat{D}\hat{V}'$ and $\hat{d}_r^2 = \lambda_r$ for $r = 1, 2, 3, \cdots, \min(N, T)$.

PCA and SVD in Factor Analysis

- To use PCA in factor analysis, we essentially use the eigenvectors of sample covariance matrix to estimate the linear space of factors and then estimate the linear space of loadings by least square method.
- Let r be an arbitrary number such that $0 < r < \min\{N, T\}$, the least square method seeks $L = (L_1, L_2, \cdots, L_N)' \in \mathbb{R}^{r \times N}$ and $F = (F_1, F_2, \cdots, F_T) \in \mathbb{R}^{r \times T}$ such that

$$V(\hat{L}, \hat{F}) = \min_{F,L} \left\| Y - L'F \right\|_F^2, \tag{3}$$

subject to $FF'/T = I_r$, and $L'L$ is diagonal for identification.

Note: $\|A\|_F$ is denoted as Frobenius norm for matrix A, defined by $\|A\|_F = \mathrm{tr}^{\frac{1}{2}}(AA')$.

PCA and SVD in Factor Analysis

- The matrix form for factor model can be expressed as

$$Y = L'F + e,$$

with $Y \in \mathbb{R}^{N\times T}$, $L \in \mathbb{R}^{r\times N}$, $F \in \mathbb{R}^{r\times T}$, and $e_t \in \mathbb{R}^{N\times T}$.

- Hence, for each given F, the least squares estimator of L' is $\hat{L}' = YF'/T$. Substituting this into equation (3), the objective function now becomes

$$\begin{aligned}\|Y - YF'F/T\|_F^2 &= \text{tr}(Y - YF'F/T)(Y - YF'F/T)' \\ &= \text{tr}\{Y(I_N - F'F/T)Y'\},\end{aligned}$$

which is identical to maximizing $\text{tr}\{F(Y'Y/T)F'\}$. Now it is clear that the estimated factor matrix $\hat{F}'_{\text{pc}}$, is $\sqrt{T}$ times the eigenvectors corresponding to the r largest eigenvalues of $T \times T$ matrix $Y'Y$. Given $\hat{F}'_{\text{pc}}$, $\hat{L}'_{\text{pc}} = Y\hat{F}'_{\text{pc}}/T$ is the corresponding loading matrix.

Overview

1. Introduction
2. Literature Review
3. Basic Factor Model and PCA
4. Strong and Weak Factors Estimation
5. Simulation Design for Non-Robustness Issue
6. Numerical Results and Conclusion

Strong and Weak Factor Assumptions

Assuming factors F_t and noise e_t are uncorrelated and have zero mean, and normalization $E(F_t^{'}F_t) = I_r$, then the covariance matrix of the factor model is given by

$$\Sigma_Y = LL^{'} + \Sigma_u,$$

where Σ_Y and Σ_u are the $N \times N$ covariance matrix of Y_t and e_t.

- For strong factors, it is assumed that $L^{'}L/N \to \Sigma_L$ for some $r_0 \times r_0$ positive definite matrices Σ_L as $N \to \infty$. In this case, the top r_0 eigenvalues of Σ_Y are diverge at the rate $O(N)$ while the rest of its eigenvalues are bounded as $N, T \to \infty$.
- For weak factors, it is assumed that $L^{'}L \to \Sigma_L$ as $N \to \infty$ for some positive definite matrix Σ_L. In this case, all the eigenvalues of Σ_Y are bounded as $N, T \to \infty$.

Consistency of PCA Estimators

- Under the weak factor assumption, PCA or MLE estimators of loadings and factors are inconsistent as $N, T \to \infty$.
- We can illustrate this by a simple example. Define

$$Y_{it} = L_i F_t + \epsilon_{it}, \quad \epsilon_{it} \overset{\text{iid}}{\sim} \mathcal{N}(0,1),$$

and let $r = 1$ for simplicity. If L_i is known, then the OLS estimator for F_t is

$$\hat{F}_t = \frac{\sum_{i=1}^N L_i Y_{it}}{\sum_{i=1}^N L_i^2} = F_t + \frac{\sum_{i=1}^N L_i \epsilon_{it}}{\sum_{i=1}^N L_i^2},$$

so we have $E(\hat{F}_t) = F_t$ and $Var(\hat{F}_t) = 1/\sum_{i=1}^N L_i^2$. Hence, for $\hat{F}_t$ to be consistent, we need $\sum_{i=1}^N L_i^2 \to \infty$ (strong factor assumption) such that $\mathbb{P}(|\hat{F}_t - F_t| > \delta) \leq Var(\hat{F}_t)/\delta^2 \to 0$ as $N \to \infty$.

Why weak factors is important?

There are several reasons why we need to estimate the number of weak factors except for strong ones.

- In many real finance and macroeconomics data sets where both N and T are large, the empirical observations show that the eigenvalues of the sample covariance matrices do not obviously separated into large and small eigenvalues groups.
- Including useful weak factors in the number of estimated factors can decrease the estimation error for recovering the underlying signal matrix.
- In many studies, the strong factors are obvious and uninteresting while the weak factors have useful insights.

Strong Factor Estimation

- There are many methods to consistently estimate the number of factors under strong factor assumption as $N, T \to \infty$.
- Let's take the information criteria based methods developed by Bai and Ng (2002) as a example and show why the strong factor assumption is critical. Define

$$V(r) = \frac{1}{NT} \left\| Y - \hat{L}_r^{\mathrm{pc}} \hat{F}_r^{\mathrm{pc}} \right\|_F^2,$$

where $\hat{L}_r^{\mathrm{pc}}$ and $\hat{F}_r^{\mathrm{pc}}$ be the principal component estimators as we discussed before,and r is an arbitrary number such that $0 < r < \min\{N, T\}$. Note this loss function is decreasing in r.

Strong Factor Estimation

- Based on $V(r)$, Bai and Ng have shown that if $g(N,T) \to 0$, and $\min\{\sqrt{N}, \sqrt{T}\}g(N,T) \to \infty$ as $N, T \to \infty$, then the estimator $\hat{r}$ defined by

$$\hat{r}_{\mathrm{IC}} = \mathrm{argmin}_{0 \le r \le r_{max}}\{V(r) + rg(N,T)\},$$

  is a consistent estimator: $lim_{N,T\to\infty}\mathbb{P}(\hat{r}_{IC} = r) = 1$.
- Note that the penalty function $rg(N,T)$ is increasing in r, so the penalized loss function is minimized by balancing these two functions at the true factor number r_0 asymptotically.
- The strong factor assumption is critical for this method since it ensures that the principal component estimators $\hat{L}_r^{\mathrm{pc}}$ and $\hat{F}_r^{\mathrm{pc}}$ are consistent as $N, T \to \infty$.

Strong Factor Estimation

- Onatski (2010) developed an estimator based on the difference of two adjacent eigenvalues (ED) of sample covariance matrix. The estimator he proposed is

$$\hat{r}_{\mathrm{ED}} = \max\left\{r \leq r_{max} : \lambda_r - \lambda_{r+1} \geq \delta\right\},$$

  where δ is some fixed number, λ_i is the i-th largest eigenvalue of $\hat{\Sigma}_Y$.
- Ahn and Horenstein (2013) proposed an estimator by simply maximizing the ratio of two adjacent eigenvalues of the sample covariance matrix. The estimator is defined as

$$\hat{r}_{\mathrm{ER}} = \mathrm{argmin}_{0 \leq r \leq r_{max}} \lambda_r / \lambda_{r+1},$$

  with $\lambda_0 = \sum_{r=1}^{\min(N,T)} \lambda_r / \log\min(N, T)$.

Weak Factor Estimation

- The results from random matrix theory (RMT) show that, even for white noise case $\Sigma_u = \sigma^2 I_N$, PCA estimators of the loadings and factors are inconsistent as $N, T \to \infty$. Specifically, there exists a phase transition phenomenon in the limit: let ξ_k denotes the k-th largest eigenvalue of population covariance matrix Σ_Y, if

$$\xi_k < (\sqrt{N/T} + 1)\sigma^2,$$

  there is a little chance to detect of the k-th factor using PCA or MLE as $T, N \to \infty$. (Kritchman and Nadler, 2009)
- Define the number of detectable factors as

$$\#\{k \leq N : \xi_k > (\sqrt{N/T} + 1)\sigma^2\},$$

  then one goal is to estimate the number of detectable factors.

Weak Factor Estimation

- Nadakuditi and Edelman (2008) developed an method (NE) based on the distribution of the sample eigenvalues to estimate the number of detectable factors (signals) in white noise and high dimensional data using the results from random matrix theory (RMT).
- The reason why we need to use the results of RMT spreading of the sample eigenvalues in the high-dimensional regime can be precisely described by new analytical results from RMT.
- Instead of estimating the number of detectable factors, one may prefer estimating the number of useful factors. The number of useful factors recover an underlying signal matrix more precisely than using the true number of factors or detectable factors.

Factor Taxonomy

From the results of RMT, there are two thresholds of signal strength: a detection threshold (μ_F) and an estimation threshold (μ_F^*). Based on these two asymptotic thresholds, each factor can be roughly placed into four categories:

1. Undetectable factor, $d_i^2 < \mu_F$, the factor is asymptotically undetectable by PCA or MLE based methods.
2. Harmful weak factor, $\mu_F < d_i^2 < \mu_F^*$, including the factor in the model will make the loss for recovering signal matrix larger.
3. Useful weak factor, $\mu_F^* < d_i^2 = O(1)$, including the factor will reduce the loss for recovering signal matrix.
4. Strong factor, d_i^2 grows proportionally to N.

Note: d_i is the i-th largest singular value of Y. We have the identity for $d_i^2 = \lambda_i$, where λ_i is the i-th largest eigenvalue of $\hat{\Sigma}_Y$.

Weak Factor Estimation

- Owen and Wang (2015) proposed a bi-cross-validation (BCV) based method to estimate the number of useful factors, using randomly held-out submatrices of the data matrix. The model is:

$$Y = X + \Sigma^{\frac{1}{2}} E = LF + \Sigma^{\frac{1}{2}} E, \tag{4}$$

  where $X \in \mathbb{R}^{N \times T}$ (signal matrix) is a product of loading matrix $L \in \mathbb{R}^{N \times r_0}$ and factor matrix $F \in \mathbb{R}^{r_0 \times T}$. The noise matrix $E \in \mathbb{R}^{N \times T}$ with entries $e_{it} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ with $\Sigma = \text{diag}(\sigma_1^2, \sigma_2^2, \cdots \sigma_N^2)$.
- The goal is to recover the signal matrix X despite the heteroscedastic noise using the criteria function:

$$\text{Err}_X(\hat{X}) = \mathbb{E}\left(\|\hat{X} - X\|_F^2 \right), \tag{5}$$

  where $\|A\|_F = \text{tr}^{\frac{1}{2}}(AA')$.

Weak Factor Estimation

- For each number of factor $r \geq 0$, let M be a method that gives a estimator $\hat{X}^M(r)$ of X using Y and r. The oracle number of factors for M is defined as

$$r_M^* = \underset{r}{\operatorname{argmin}} \left(\left\| \hat{X}^M(r) - X \right\|_F^2 \right),$$

  and the corresponding oracle estimate of X is $\hat{X}_{\text{opt}}^M = \hat{X}^M\left(r_M^*\right)$.
- For a good method M, when all the factors are strong enough, we should have $r_M^* = r_0$; while $r_M^* < r_0$ with weak enough factors.
- The algorithm for recovering the signal matrix X has two steps:
  1. Devise a method M to estimate X given oracle number of factors r_M^*.
  2. With such a method in hand, we need a means to estimate r_M^*.

Early Stopping Alternation (ESA)

- The method they devised for estimating X at a given r is early stopping alternation (ESA). Based on the sample variance, an initial estimator of Σ in noise matrix $\Sigma^{\frac{1}{2}}E$ is given by

$$\hat{\Sigma} = \text{diag}\left(\left(Y - \frac{1}{T}Y\mathbf{1}_{T\times T}\right)\left(Y - \frac{1}{T}Y\mathbf{1}_{T\times T}\right)'\right). \tag{6}$$

- Based on $\hat{\Sigma}$, the estimator $\hat{X}(r)$ for each r is defined as

$$\hat{X}(r) = \hat{\Sigma}^{\frac{1}{2}}\tilde{Y}(r), \tag{7}$$

with reweighted matrix $\tilde{Y}(r) = \hat{\Sigma}^{-\frac{1}{2}}Y(r)$, where $\hat{\Sigma}^{-\frac{1}{2}}Y(r) = U(r)D(r)V(r)'$ is the truncated singular value decomposition (SVD) of the matrix Y after normalization.

Early Stopping Alternation (ESA)

- Given an estimator $\hat{X}(r)$, the variance estimator $\hat{\Sigma}$ can be updated by:

$$\hat{\Sigma} = \frac{1}{n} \operatorname{diag}\left[(Y - \hat{X}(r))(Y - \hat{X}(r))'\right]. \tag{8}$$

- The ESA method simply start at (6) and iterate the (7) and (8) for some number m of times and then stop. By this process, they find that taking $m = 3$ works almost as well as whichever m is used to give the smallest estimation error in (5).
- Note that SVD after normalization of each variable is equivalent to ESA starting from (6) with $m = 1$, so ESA with m $= 3$ can be understood as applying truncated SVD on a more properly reweighted data ($\hat{\Sigma}^{-\frac{1}{2}} Y$) than one gets with m $= 1$.

BCV Method

- The means they developed for estimating r^*_{ESA} is bi-cross-vaildation (BCV) based method. In this method, the data matrix Y into four blocks by randomly select N_0 rows and T_0 columns as the held-out block as below

$$Y = \begin{pmatrix} Y_{00} & Y_{01} \\ Y_{10} & Y_{11} \end{pmatrix},$$

  where Y_{00} is the selected $N_0 \times T_0$ held-out block, and Y_{01}, Y_{10}, and Y_{11} are the other three held-in blocks.

- Correspondingly, X and Σ can be partitioned as

$$X = \begin{pmatrix} X_{00} & X_{01} \\ X_{10} & X_{11} \end{pmatrix}, \quad \text{and} \quad \Sigma = \begin{pmatrix} \Sigma_0 & 0 \\ 0 & \Sigma_1 \end{pmatrix}.$$

  The idea of BCV method is that, for each candidate r, we first use the three held-in blocks to estimate X_{00} and then select the optimal r^* based on the BCV estimated prediction error.

BCV Method

- More specifically, model (4) can be rewritten in terms of the four blocks:

$$\begin{pmatrix} Y_{00} & Y_{01} \\ Y_{10} & Y_{11} \end{pmatrix} = \begin{pmatrix} X_{00} & X_{01} \\ X_{10} & X_{11} \end{pmatrix} + \begin{pmatrix} \Sigma_0 & 0 \\ 0 & \Sigma_1 \end{pmatrix}^{\frac{1}{2}} \begin{pmatrix} E_{00} & E_{01} \\ E_{10} & E_{11} \end{pmatrix}$$
$$= \begin{pmatrix} L_0 F_0 & L_0 F_1 \\ L_1 F_0 & L_1 F_1 \end{pmatrix} + \begin{pmatrix} \Sigma_0^{\frac{1}{2}} E_{00} & \Sigma_0^{\frac{1}{2}} E_{01} \\ \Sigma_1^{\frac{1}{2}} E_{10} & \Sigma_1^{\frac{1}{2}} E_{11} \end{pmatrix}.$$

- Note that the held-in block

$$Y_{11} = X_{11} + \Sigma_1^{\frac{1}{2}} E_{11} = L_1 F_1 + \Sigma_1^{\frac{1}{2}} E_{11}$$

also has the factor structure as (4), so we can get estimators $\hat{X}_{11}(r)$ and $\hat{\Sigma}_1$ by using ESA for each r.

BCV Method

- The estimator $\hat{X}_{11}(r)$ can be decomposed as $\hat{X}_{11}(r) = \hat{L}_1 \hat{F}_1$ for $r < \text{rank}(Y_{11})$.
- Then L_0 and F_0 can be estimated by solving the linear regression model $Y_{01} = L_0 \hat{F}_1 + \Sigma_0^{\frac{1}{2}} E_{01}$ and $Y_{10} = \hat{L}_1 F_0 + \hat{\Sigma}_1^{\frac{1}{2}} E_{10}$. These least square solutions are

$$\hat{L}_0 = Y_{01} \hat{F}_1' \left(\hat{F}_1 \hat{F}_1' \right)^{-1} \text{ and } \quad \hat{F}_0 = \left(\hat{L}_1' \hat{\Sigma}_1^{-1} \hat{L}_1 \right)^{-1} \hat{L}_1' \hat{\Sigma}_1^{-1} Y_{10},$$

  which do not depend on the unknown Σ_0.
- Then the estimator of X_{00} is $\hat{X}_{00}(r) = \hat{L}_0 \hat{F}_0$ given r. They proved that the estimate $\hat{X}_{00}(r)$ is unique, though the decomposition of $\hat{X}_{11}(r)$ is not unique.

BCV Method

- With estimated $\hat{X}_{00}(r)$, the cross-validation prediction average squared error for block Y_{00} then can be defined as

$$\begin{aligned}\mathbb{E}\left[\widehat{\mathrm{PE}}_r\left(Y_{00}\right)\right] =& \mathbb{E}\left[\frac{1}{N_0 T_0}\left\|Y_{00}-\hat{X}_{00}(r)\right\|_F^2\right] \\ =& \mathbb{E}\left[\frac{1}{N_0 T_0}\mathrm{Err}_{X_{00}}\left(\hat{X}_{00}(r)\right)\right]+\frac{1}{N_0}\sum_{i=1}^{N_0}\sigma_i^2.\end{aligned}$$

- By repeating the above random partitioning step K times independently, the average BCV mean squared prediction error for Y is

$$\widehat{\mathrm{PE}}(r)=\frac{1}{K}\sum_{k=1}^{K}\widehat{\mathrm{PE}}_r(Y_{00}^{(k)}),$$

where $Y_{00}^{(k)}$ means the k-th time of random partition.

BCV Method

- The BCV estimate of the number of useful factors r^*_{ESA} is then

$$\hat{r}^*_{ESA} = \arg\min_r \widehat{\mathrm{PE}}(r), \quad 0 \leq r \leq r_{\max}.$$

- Lastly, for choosing the size of the holdout Y_{00}, we can define the true prediction error for ESA as:

$$\mathrm{PE}(r) = \frac{1}{TN}\|X - \hat{X}(r)\|_F^2 + \frac{1}{N}\sum_i \sigma_i^2, \tag{9}$$

  then the optimal number of factors for ESA is $\hat{r}^*_0 = \arg\min_r \mathrm{PE}(r)$.
- Perry (2009) proved that $\hat{r}^*_{ESA}$ and $\hat{r}^*_0$ track each other asymptotically if the relative size of the held-out matrix Y_{00} satisfies the following theorem.

BCV method

Theorem

For factor model (4), if r_0 is fixed and $N/T \to \gamma \in (0, \infty)$, then $\hat{r}^*_{ESA}$ and $\hat{r}^*_0$ converge to the same value if

$$\sqrt{\rho} = \frac{\sqrt{2}}{\sqrt{\bar{\gamma}} + \sqrt{\bar{\gamma} + 3}}$$

holds, where

$$\bar{\gamma} = \left(\frac{\gamma^{1/2} + \gamma^{-1/2}}{2} \right)^2, \quad \text{and} \quad \rho = \frac{T - T_0}{T} \cdot \frac{N - N_0}{N}.$$

Here ρ is the fraction of entries from Y in the held-in block Y_{11}. For example, $\rho \approx 22\%$ if Y is square with $c = 1$.

Overview

1. Introduction
2. Literature Review
3. Basic Factor Model and PCA
4. Strong and Weak Factors Estimation
5. Simulation Design for Non-Robustness Issue
6. Numerical Results and Conclusion

Issue of Non-Robustness

- Most methods for estimating the number of factors are based on the results from random matrix theory (RMT), which require i.i.d and Gaussian assumption on the error terms. These restrictions may not appropriate when we want to apply them in practice.
- The purpose of this simulation design is to show that whether all the methods we have introduced before are robust for estimating the number of factors (strong, weak and mixed), when the error terms are cross-sectional and high serial correlated or have non-guassian distributions.

Strong factors only

- First, we consider there are only strong factors in the generated data and apply all of the methods in the approximate factor model with serial, cross-sectional or non-guassian error terms.
- This simulation design follows the design of Baltagi, Kao, and Peng (2014) and Onatski (2012). Specifically, we consider the following data-generating process (DGP):

$$
\begin{aligned}
&Y_{it} = \sum_{j=1}^{r} \lambda_{ij} F_{tj} + e_{it}, \quad \text{where} \\
&\lambda_{ij}, F_{tj} \overset{\text{iid}}{\sim} \mathcal{N}(0,1), \quad i = 1, \cdots, N, t = 1, 2, \cdots, T, \\
&e_{it} = \rho_1 e_{it-1} + (1-\rho_1^2)^{1/2} \xi_{it}, \\
&\xi_{it} = \rho_2 \xi_{i-1,t} + (1-\rho_2^2)^{1/2} \epsilon_{it}, \quad \epsilon_{it} \overset{\text{iid}}{\sim} \mathcal{N}(0,1).
\end{aligned}
$$

Strong factor only

- We let $r = 2$, and consider the three cases for e_{it} below:

  1. high serial correlation only, $\rho_1 = 0.9$ and $\rho_2 = 0$;
  2. mild cross-sectional correlation only, $\rho_1 = 0$ and $\rho_2 = 0.5$;
  3. non-guassion distributions only, $\rho_1 = \rho_2 = 0$ with four types of distributions for e_{it}: normal, gamma, lognormal and chi-square with mean zero and variance 0.5.

  We use guassian and i.i.d error terms as our benchmark.

- The methods we use to apply for estimating the number of factors in the generated data are IC2, ER, ED, NE and BCV methods, where IC2 is one of the six criterions proposed by Bai and Ng (2002). We choose this criterion simply because it has the largest penalty term so that the probability of overestimation is the smallest.

Mixed with strong and weak factors

- Next, we consider there are weak factors or mixed with strong and weak factors in the generated data and apply all of the methods in the approximate factor model with serial, cross-sectional or non-guassian error terms.
- This simulation design follows the design of Owen and Wang (2015) and Onatski (2012). Consider the model for generating data as:

$$\begin{aligned} Y &= X + \Sigma^{\frac{1}{2}} E \\ &= \Sigma^{\frac{1}{2}}(\Sigma^{-\frac{1}{2}} X + E) = \Sigma^{\frac{1}{2}}(\sqrt{T}\hat{U}\hat{D}\hat{V}' + E) \end{aligned}$$

where $\sqrt{T}\hat{U}\hat{D}\hat{V}'$ is the singular value decomposition (SVD) for $\Sigma^{-\frac{1}{2}} X$ as we have introduced before. The singular value matrix $\hat{D} = \text{diag}(\hat{d}_1, \hat{d}_2, \cdots, \hat{d}_{\min(N,T)})$ defines the strength of each factor.

Generating the signal matrix

Based on those four categories of factor strength, we can generate the weighted signal matrix $\Sigma^{-\frac{1}{2}}X = \sqrt{T}UDV'$ with strong and weak factors in two steps as follows:

- Six testing scenarios are described in the table below:

	Scenario					
	1	2	3	4	5	6
# Undetectable	1	1	1	1	1	1
# Harmful	1	1	1	3	3	6
# Useful	6	4	3	1	3	1
# Strong	0	2	3	3	1	0

- U and V: they are generated uniformly from the Stiefel manifold $V_k(\mathbb{R}^N)$ and $V_k(\mathbb{R}^T)$ by certain process following (Perry 2009).

Generating the noise

We consider three cases below for constructing the noise term $\Sigma^{\frac{1}{2}} E$ and use $E = (e_{it})_{N\times T} : e_{it} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ and $\Sigma = I_N$ as our benchmark.

- Case 1: Error term with serial correlation,
  1. $E = (e_{it})_{N\times T} : e_{it} = \rho_1 e_{it-1} + (1-\rho_1^2)^{1/2}\epsilon_{it}, \quad \epsilon_{it} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$
  2. $\Sigma = I_N$: assume homoskedasiticity.

  We let $\rho_1 = 0.9$ for a high degree of the serial correlation.
- Case 2: Error term with cross-section correlation,
  1. $E = (e_{it})_{N\times T} : e_{it} = \rho_2 e_{i-1,t} + (1-\rho_2^2)^{1/2}\epsilon_{it}, \ \epsilon_{it} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$
  2. $\Sigma = I_N$: assume homoskedasiticity.

  We let $\rho_1 = 0.5$ for a mild degree of the cross-section correlation.

Generating the noise

- Case 3: Error term with non-guassion distributions,
  1. $E = (e_{itj})_{N \times T}$: here we consider three types of non-guassion distributions for e_{it}: gamma, log-normal and chi-square with mean zero and variance 0.5.
  2. $\Sigma = I_N$: assume homoskedasiticity.
- Data dimensions:
  1. There are 5 different (N, T) pairs are considered in simulations, 5 types of error terms, 5 factor estimation methods, and 6 types of factor strengths.
  2. In total there are $5 \times 5 \times 5 = 125$ scenarios for strong factor estimation and $5 \times 5 \times 5 \times 6 = 750$ scenarios for strong and weak factor estimation. For strong factor cases, each was simulated 1000 times. For strong and weak factor cases, each was simulated 100 times.

Overview

1. Introduction
2. Literature Review
3. Basic Factor Model and PCA
4. Strong and Weak Factors Estimation
5. Simulation Design for Non-Robustness Issue
6. Numerical Results and Conclusion

Results for strong factors only

- All of those five methods can precisely choose the number of factors when factors are "strong" as long as the noise term is white.
- When the data are serially correlated, all of those five methods have poor performances to estimate the number of factors except ER method when T is large.
- When the data are cross-sectionally correlated, ED and ER are robust for estimating the number factors. IC2, NE, and BCV have relatively good performances when N is large.
- All of those five method are not robust when the noise term in the data have non-gaussian distributions such as lognormal, gamma, and chi-square distribution.
- Overall, for serially and cross-sectionally correlated data with guassion distribution noise and large T, our simulation results suggest to use ER method to estimate the number of factors in practice.

Results for strong and weak factors

- In all the those five methods, NE is designed to estimate the number of strong and weak factors in white noise and BCV is designed to estimate the number of strong and weak factors in heteroscedastic noise. Other three methods are designed to estimate the number of strong factors only.
- For BCV and NE method, our results show that they perform quite well when we have cross-sectionally correlated data and the error terms in the factor model have guassian and chi-square distributions. They are not robust to other cases of error terms however.

Results for strong and weak factors

- For ED method, our results show that it is quite robust to all types of error terms in type 4 and type 6 factor strengths. That is, it is robust when the ratio of strong to useful weak factors is large. Also, for type2 - type5 factor strengths, the ED method performs quite well for large N when the error terms in the factor model are highly serial correlated.
- For ER method, it is fail to estimate the number of weak factors in all cases as we expected, since it designs for estimating the number of strong factors only.
- For IC2 method, our simulation results show that it performs quite well in all types of mixed factor strengths when the error term in factor model is white. Besides, it also robust to cross-sectional correlated error terms in all types of mixed factor strengths when N is large.

The End, thanks!