Spiked Eigenvalues of High Dimensional Separable Sample Covariance Matrices

Guangming Pan,

Nanyang Technological University, Singapore

November 19, 2019

Outline

1. **Motivation**
   - Misleading PCA on Simulated Data
2. **High Dimensional Separable Covariance Model**
3. **Asymptotic Performance of Largest Eigenvalues**
4. **Inference on High Dimensional Time Series**
   - Implementing Factor Analysis on Our Model
   - Unit Root Models Satisfying Assumption 4
   - A New Test for Unit Root against Factor Model
   - More Thoughts about Panel Data Structures
5. **Simulations**
   - The Simulation about Proposition 3
   - The Simulation about Proposition 4
6. **Reference**

Outline

1. **Motivation**
   - Misleading PCA on Simulated Data
2. **High Dimensional Separable Covariance Model**
3. **Asymptotic Performance of Largest Eigenvalues**
4. **Inference on High Dimensional Time Series**
   - Implementing Factor Analysis on Our Model
   - Unit Root Models Satisfying Assumption 4
   - A New Test for Unit Root against Factor Model
   - More Thoughts about Panel Data Structures
5. **Simulations**
   - The Simulation about Proposition 3
   - The Simulation about Proposition 4
6. **Reference**

The model

$$y_{it} = \ell_{i1} f_{1t} + \ell_{i2} f_{2t} + \varepsilon_{it} = \boldsymbol{\ell}_i^* \mathbf{f}_t + \varepsilon_{it}, \quad i = 1, 2, \ldots, n; \; t = 1, 2, \ldots, T, \quad (1)$$

where $\mathbf{f}_t = (f_{1t}, f_{2t})^*$ are two common factors, $\boldsymbol{\ell}_i = (\ell_{i1}, \ell_{i2})^*$ are the corresponding factor loadings, and ε_{it} is the error component, in which the symbol "*" denotes the conventional conjugate transpose.

Scenario : No true common factors

Under this case, the factor loadings are generated as $\boldsymbol{\ell}_i = (0, 0)^*$. When the original data follow AR(1) model ($\gamma = 0.2$), Figures 1 and 2 provide all eigenvalues of the sample covariance matrix as $(T, n) = (20, 40)$ and $(T, n) = (40, 20)$, respectively. There are no spiked eigenvalues in view of these graphs, which correctly reflect the fact that there are no common factors in the original data.

Figures

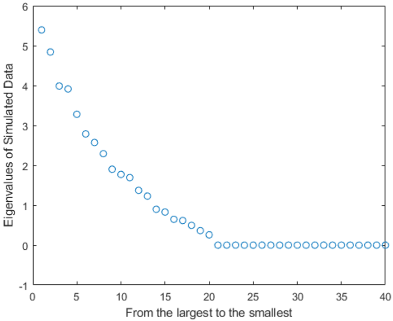

Figure: 1 $T = 20, n = 40, \gamma = 0.2$

Figures

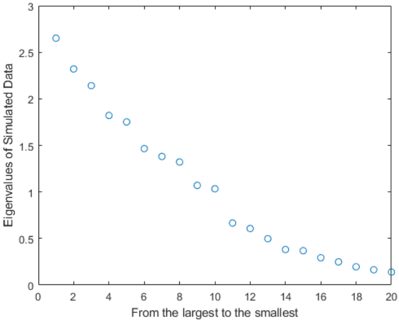

Figure: 2 $T = 40, n = 20, \gamma = 0.2$

Figures

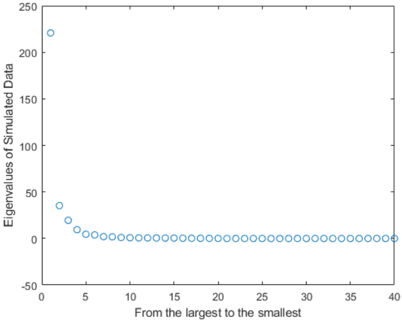

Figure: 3 $T = 20, n = 40, \gamma = 1$

Figures

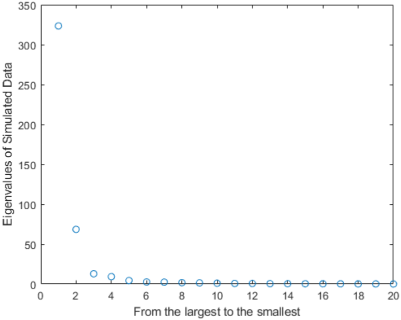

Figure: 4 $T = 40, n = 20, \gamma = 1$

Scenario : No true common factors

However, as the data observations are nonstationary ($\gamma = 1$), Figures 3 and 4 show that there is one spiked eigenvalue from the sample covariance matrix, while the true number of common factors is 0.
This example demonstrates that PCA may not be informative accurately on high dimensional data with dependent sample observations.

Outline

1. Motivation
   - Misleading PCA on Simulated Data
2. **High Dimensional Separable Covariance Model**
3. Asymptotic Performance of Largest Eigenvalues
4. Inference on High Dimensional Time Series
   - Implementing Factor Analysis on Our Model
   - Unit Root Models Satisfying Assumption 4
   - A New Test for Unit Root against Factor Model
   - More Thoughts about Panel Data Structures
5. Simulations
   - The Simulation about Proposition 3
   - The Simulation about Proposition 4
6. Reference

High Dimensional Separable Covariance Model

Consider an n-dimensional random vector $\mathbf{y}$ with observations $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T$. Pool all observations together into a $T \times n$ matrix $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T)^*$. The data matrix $\mathbf{Y}$ has the structure

$$\mathbf{Y} = \mathbf{\Gamma} \mathbf{X} \mathbf{\Omega}^{1/2}, \tag{2}$$

where $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n) = (x_{ij})_{(T+L)\times n}$ is a $(T+L) \times n$ random matrix with i.i.d. elements; $\mathbf{\Sigma} = \mathbf{\Gamma}\mathbf{\Gamma}^*$ and $\mathbf{\Omega}$ are $T \times T$ and $n \times n$ deterministic non-negative definite matrices, respectively. Here $\mathbf{\Gamma}$ is a $T \times (T+L)$ deterministic matrix.

Separable covariance matrix

- Actually the matrix $\mathbf{\Gamma}$ describes dependence among sample observations.
- The matrix $\mathbf{\Omega}$ measures cross-sectional dependence for $\mathbf{y}$ under study.
- Under this setting, the sample covariance matrix of $\mathbf{y}$ can be expressed as $\mathbf{\Gamma X \Omega X^* \Gamma^*}$. It is also called separable covariance matrix.

Largest spiked eigenvalues

- We are interested in the largest spiked eigenvalues of matrix $\mathbf{\Omega}$, which describes the cross-sectional dependence.
- In the classical procedure of using PCA, spiked empirical eigenvalues from sample covariance matrix $\mathbf{\Gamma X \Omega X^* \Gamma^*}$ are utilized to approximate those of the matrix $\mathbf{\Omega}$.
- In this paper, we investigate the spiked empirical eigenvalues from an innovative view: how the the spiked eigenvalues of the matrix $\mathbf{\Sigma}$ (due to the dependent sample) affect the spiked sample eigenvalues ?
- To this end, we do not impose any spiked structures on the matrix $\mathbf{\Omega}$.

spikiness of the matrix Σ

We assume spikiness of the matrix $\mathbf{\Sigma}$ through the following decomposition. Let the spectral decomposition of $\mathbf{\Gamma}$ be $\mathbf{V}\Lambda^{1/2}\mathbf{U}$, where $\mathbf{V}$ and $\mathbf{U}$ are $T \times T$ and $T \times (T+L)$ orthogonal matrices respectively ($\mathbf{V}\mathbf{V}^* = \mathbf{U}\mathbf{U}^* = \mathbf{I}$), Λ is a diagonal matrix composed by the descent ordered eigenvalues of $\mathbf{\Sigma} = \mathbf{\Gamma}\mathbf{\Gamma}^*$. Moreover, we write $\Lambda = \begin{pmatrix} \Lambda_S & 0 \\ 0 & \Lambda_P \end{pmatrix}$, where $\Lambda_S = diag(\mu_1, ..., \mu_K)$, $\Lambda_P = diag(\mu_{K+1}, ..., \mu_T)$, and $\mu_1, ..., \mu_K$ are referred to the spiked eigenvalues that are significantly bigger than the rest. In addition, we write $\mathbf{U} = \begin{pmatrix} \mathbf{U}_1 \\ \mathbf{U}_2 \end{pmatrix}$ and $\mathbf{\Sigma}_2 = \mathbf{U}_2^* \Lambda_P \mathbf{U}_2$.

Outline

1. Motivation
   - Misleading PCA on Simulated Data
2. High Dimensional Separable Covariance Model
3. **Asymptotic Performance of Largest Eigenvalues**
4. Inference on High Dimensional Time Series
   - Implementing Factor Analysis on Our Model
   - Unit Root Models Satisfying Assumption 4
   - A New Test for Unit Root against Factor Model
   - More Thoughts about Panel Data Structures
5. Simulations
   - The Simulation about Proposition 3
   - The Simulation about Proposition 4
6. Reference

Asymptotic Performance of Largest Eigenvalues

This section is to establish the asymptotic distribution of the largest spiked empirical eigenvalues. First, we make the following assumptions.

Assumption (Moment Conditions)

$\{x_{ij}$: $i = 1, ..., T + L$, $j = 1, ..., n\}$ are i.i.d random variables such that $\mathbb{E}x_{ij} = 0$. $\mathbb{E}|\sqrt{n}x_{ij}|^2 = 1$ and $\mathbb{E}|\sqrt{n}x_{ij}|^4 = \gamma_4 < \infty$.

Assumption 2

Assumption (Dependent Sample Structure)

$\alpha_{\mathcal{L}} = \mu_K = ... = \mu_{K-n_{\mathcal{L}}} < \alpha_{\mathcal{L}-1} = \mu_{K-n_{\mathcal{L}}+1} ... < \alpha_1 = \mu_{n_1} = ... = \mu_1$, where n_1,..., $n_{\mathcal{L}}$ are finite. Moreover, there exists a small constant $c > 0$ such that $\alpha_{i-1} - \alpha_i \geq c\alpha_i$ for $i = 1, 2, ..., \mathcal{L}$ and $\mu_K - \mu_{K+1} \geq c\mu_K$.

Assumption 3

Assumption (Cross-sectional Structure)

The matrix $\mathbf{\Omega}$ is nonnegative definite and its effective rank $r^*(\mathbf{\Omega}) = \frac{tr(\mathbf{\Omega})}{\|\mathbf{\Omega}\|_2} \to \infty$, where $\|\mathbf{\Omega}\|_2$ means the spectral norm.

Assumption 4

Assumption (Spiked Dependent Sample Structure)

The spiked eigenvalues of the population covariance matrix are much bigger than the rest of the eigenvalues. Precisely speaking, for $\forall \varepsilon > 0$, there is K_ε, independent of n and T, such that when n and T are big enough,

$$\frac{\sum_{i=K_\varepsilon}^{T} \mu_i}{\mu_K} < \frac{\varepsilon}{2}. \tag{3}$$

Remarks about Assumptions

Note that Assumptions 2 and 4 impose a spiked structure on $\mathbf{\Sigma}$ while Assumption 3 could endure either spiked or non-spiked structure on $\mathbf{\Omega}$. This is consistent with the aim of this paper to investigate the effect caused by dependent sample observations on the spiked sample eigenvalues.

Remarks about Assumptions

When $\mu_i = i^{-1-\sigma}$ and $\sigma > 0$ one can find that Assumption 4 holds. Moreover, Section 4.2 below shows that Assumption 4 holds in the unit root setting. In addition, define a near unit root model of the form:

$$y_{it} = \rho y_{i,t-1} + \sum_{h=0}^{L} b_h z_{i,t-h}, \tag{4}$$

where $T(1-\rho)$ is bounded as T goes to infinity. It can also be verified that Assumption 4 holds in such models. Also, heterogeneous high–dimensional time series models can also be covered if the corresponding variances satisfy Assumption 4.

the asymptotic joint distribution of the largest spiked eigenvalues

Denote the i-th largest sample eigenvalue of $\mathbf{\Gamma X \Omega X^* \Gamma^*}$ by λ_i. Set $m_i = \sum_{j=1}^{i-1} n_j$, for all $i = 1, 2, ..., \mathcal{L}$. The following theorem establishes the asymptotic joint distribution of the largest spiked eigenvalues.

the asymptotic joint distribution of the largest spiked eigenvalues

Theorem

Suppose that Assumptions 1-4 hold,

$$\frac{n}{\mu_i\sqrt{tr(\mathbf{\Omega}^2)}}\left(\lambda_{m_i+1}-\mu_i\frac{tr\mathbf{\Omega}}{n},\lambda_{m_i+2}-\mu_i\frac{tr\mathbf{\Omega}}{n},...,\lambda_{m_i+n_i}-\mu_i\frac{tr\mathbf{\Omega}}{n}\right)\xrightarrow{d}\mathcal{R}_i,$$

where $\mathcal{R}_i$ are the eigenvalues of an $n_i\times n_i$ matrix $\mathbf{R}_i$ with the Gaussian elements, $\mathbb{E}\mathbf{R}_i=0$, the covariance of the $(\mathbf{R}_i)_{k_1,l_1}$ and $(\mathbf{R}_i)_{k_2,l_2}$ is

$$\lim_{n\to\infty}\frac{n^2}{tr(\mathbf{\Omega}^2)}\times Cov(\mathbf{u}^*_{m_i+k_1}\mathbf{X}\mathbf{\Omega}\mathbf{X}^*\mathbf{u}_{m_i+l_1},\mathbf{u}^*_{m_i+k_2}\mathbf{X}\mathbf{\Omega}\mathbf{X}^*\mathbf{u}_{m_i+l_2})(6)$$

Here the limit of (6) is bounded.

the asymptotic joint distribution of the largest spiked eigenvalues

Theorem

Moreover, if $\frac{\mu_i}{\mu_j} \geq c > 1$, λ_{m_i+f} and λ_{m_j+g} are asymptotically independent, where $1 \leq f \leq n_i$ and $1 \leq g \leq n_j$. Particularly when $n_i = 1$ for all $i = 1, \ldots, K$ we have

$$\left(\frac{\lambda_1 - \mu_1 \frac{tr\boldsymbol{\Omega}}{n}}{\mu_1 \sqrt{var(\mathbf{u}_1^* \mathbf{X}\boldsymbol{\Omega}\mathbf{X}^* \mathbf{u}_1)}}, \frac{\lambda_2 - \mu_2 \frac{tr\boldsymbol{\Omega}}{n}}{\mu_2 \sqrt{var(\mathbf{u}_2^* \mathbf{X}\boldsymbol{\Omega}\mathbf{X}^* \mathbf{u}_2)}}, \ldots, \frac{\lambda_K - \mu_K \frac{tr\boldsymbol{\Omega}}{n}}{\mu_K \sqrt{var(\mathbf{u}_K^* \mathbf{X}\boldsymbol{\Omega}\mathbf{X}^* \mathbf{u}_K)}}\right) \xrightarrow{d} N(\mathbf{0}, \mathbf{I_K}).$$

Remarks about the Theorem

Note that

$$\sqrt{\frac{tr(\mathbf{\Omega})}{\|\mathbf{\Omega}\|_2}} = \frac{tr(\mathbf{\Omega})}{\sqrt{\|\mathbf{\Omega}\|_2 tr(\mathbf{\Omega})}} \leq \frac{tr(\mathbf{\Omega})}{\sqrt{tr(\mathbf{\Omega^2})}} \leq \frac{tr(\mathbf{\Omega})}{\|\mathbf{\Omega}\|_2}. \tag{7}$$

From this and Assumption 3, we can find that the standard deviation of λ_i has the smaller order than the mean of λ_i. So the sample eigenvalues $\{\lambda_i, i \leq K\}$ have the same order as $\mu_i \frac{tr\mathbf{\Omega}}{n}$.

Remarks about the Theorem

It indicates that the sample eigenvalues $\lambda_1, \cdots, \lambda_K$ are spiked under this case no matter whether $\mathbf{\Omega}$ has spiked eigenvalues or not. This phenomenon suggests that PCA may reflect inaccurate information of the cross-sectional structure $\mathbf{\Omega}$ due to the dependent sample observations. This is in contrast to the results Baik and Silverstein (2006) for the independent sample observations which establish one to one correspondence between the sample spiked eigenvalues and the population spiked eigenvalues due to the cross-sectional structure $\mathbf{\Omega}$.

Two Propositions

The following two propositions further investigate the relations between the leading sample eigenvalues and the eigenvalues of $\mathbf{\Sigma}$ due to the dependent sample observations.

Proposition

Under the conditions of Theorem 2, there exists a positive constant c such that $\liminf_{T\to\infty} \frac{\mu_1}{tr(\mathbf{\Gamma}\mathbf{\Gamma}^*)} > c$ and

$$\lim_{n,T\to\infty} P\left(\frac{\lambda_1}{tr(\mathbf{\Gamma}\mathbf{X}\mathbf{\Omega}\mathbf{X}^*\mathbf{\Gamma}^*)} > c\right) = 1. \tag{8}$$

Moreover, when $1 \le i < K$,

$$\frac{\lambda_i}{\lambda_{i+1}} - \frac{\mu_i}{\mu_{i+1}} \to 0 \text{ in probability as } n, T \to \infty. \tag{9}$$

Two Propositions

Proposition

Let the conditions of Theorem 2 hold. For $1 \le i < K - 1$, if $\min\left\{\frac{\mu_i}{\mu_{i+1}}, \frac{\mu_{i+1}}{\mu_{i+2}}\right\} \ge c > 1$, then $\frac{\frac{\lambda_i}{\mu_i} - \frac{\lambda_{i+1}}{\mu_{i+1}}}{\frac{\lambda_{i+2}}{\mu_{i+2}} - \frac{\lambda_{i+1}}{\mu_{i+1}}} = \frac{\lambda_i \frac{\mu_{i+1}}{\mu_i} - \lambda_{i+1}}{\lambda_{i+2} \frac{\mu_{i+1}}{\mu_{i+2}} - \lambda_{i+1}}$ has the same asymptotic distribution as $\frac{v_1 - v_2}{v_3 - v_2}$, where $\{v_i : 1 \le i \le 3\}$ are i.i.d standard normal random variables.

Remark about Propositions

Remark

Proposition 1 shows that the ratio of the neighboring spiked empirical eigenvalues approximate that of the spiked eigenvalues from the dependent sample structure. A central limit theorem is provided for the ratio statistic constructed from the spiked empirical eigenvalues in Proposition 2.

Outline

1. Motivation
   - Misleading PCA on Simulated Data
2. High Dimensional Separable Covariance Model
3. Asymptotic Performance of Largest Eigenvalues
4. **Inference on High Dimensional Time Series**
   - Implementing Factor Analysis on Our Model
   - Unit Root Models Satisfying Assumption 4
   - A New Test for Unit Root against Factor Model
   - More Thoughts about Panel Data Structures
5. Simulations
   - The Simulation about Proposition 3
   - The Simulation about Proposition 4
6. Reference

Outline

1. Motivation
   - Misleading PCA on Simulated Data
2. High Dimensional Separable Covariance Model
3. Asymptotic Performance of Largest Eigenvalues
4. **Inference on High Dimensional Time Series**
   - **Implementing Factor Analysis on Our Model**
   - Unit Root Models Satisfying Assumption 4
   - A New Test for Unit Root against Factor Model
   - More Thoughts about Panel Data Structures
5. Simulations
   - The Simulation about Proposition 3
   - The Simulation about Proposition 4
6. Reference

Implementing Factor Analysis on Our Model

- Proposition 1 implies that the largest sample eigenvalue has the same order as the sum of all eigenvalues.
- The largest eigenvalue is so large that the methods in Ahn and Horensten (2013) and Bai (2004) would both estimate the number of factors to be the one bigger than zero even though there is no factor in our model.
- Similarly, the relation between λ_i and λ_{i+1} leads to that Onatski (2010) would estimate the number of factors to be the one bigger than zero even though there is no factor in our model. We examine this one by one below.

The method in Onatski (2010)

- The method in Onatski (2010) is based on the difference between the i–th largest eigenvalue and the $i+1$–th one. In a few words, the idea of Onatski (2010) is that if there is no factor, for any $i \geq 1$, the difference between the i–th largest eigenvalue and the $i+1$–th one should be very small.
- Recalling (9), we can find that the difference between the i-th largest eigenvalue and the $i+1$th one in our model can be large when $\frac{\mu_i}{\mu_{i+1}} > 1$.
- In other words, the method in Onatski (2010) would get a non-zero estimate for the number of factors in our model when $\frac{\mu_i}{\mu_{i+1}} > 1$.

The method in Ahn and Horensten (2013)

- The method in Ahn and Horensten (2013) is based on the ratio between two successive largest eigenvalues. It defines a mock eigenvalue $\lambda_0 = \frac{\sum_{i=1}^{\min\{n,T\}} \lambda_i}{\ln(\min\{n,T\})}$. Then the estimator is

$$\tilde{k}_{ER} = \max_{0 \le k \le k_{max}} \frac{\lambda_k}{\lambda_{k+1}}. \tag{10}$$

- Note that λ_0 has a smaller order than the trace of the sample covariance matrix. Then the method of Ahn and Horensten (2013) implies that if there is no factor, the largest eigenvalue should have a smaller order than the trace of the sample covariance matrix.
- (8) shows that the largest eigenvalue in our model has the same order as the trace of the sample covariance matrix.
- In other words, the method in Ahn and Horensten (2013) would get a non-zero estimate for the number of factors in our model.

The method in Bai (2004)

- The methods in Bai (2004) is based on penalty functions. Briefly speaking, the idea of Bai (2004) is that if there is no factor, the largest eigenvalue should be smaller than the penalty function. The criterion has the form:

$$PC(k) = \min_{\Lambda} \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} (X_{it} - \lambda_i^{k'} \hat{F}_t^k)^2 + kg(n,T), \qquad (11)$$

  where $g(n,T)$ satisfies some properties in Bai (2004).
- However, we can find that all the penalty functions have a smaller order than the trace of the sample covariance matrix.
- Recalling (8), the methods in Bai (2004) would get a non-zero estimate for the number of factors in our model.

Outline

1. Motivation
   - Misleading PCA on Simulated Data
2. High Dimensional Separable Covariance Model
3. Asymptotic Performance of Largest Eigenvalues
4. **Inference on High Dimensional Time Series**
   - Implementing Factor Analysis on Our Model
   - **Unit Root Models Satisfying Assumption 4**
   - A New Test for Unit Root against Factor Model
   - More Thoughts about Panel Data Structures
5. Simulations
   - The Simulation about Proposition 3
   - The Simulation about Proposition 4
6. Reference

The unit root model

$$y_{it} = y_{i,t-1} + \sum_{h=0}^{L} b_h z_{i,t-h}, \tag{12}$$

where $1 \le i \le n$, $1 \le t \le T$ and L can be finite or infinite. Here

$$z_{it} = \sum_{s=1}^{n} \Upsilon_{is} x_{st}, \tag{13}$$

where $1 \le i \le n$ and $1 - L \le t \le T$.

The unit root model

Let $\mathbf{Y} = (y_{it})$ be an $n \times T$ matrix and $\bar{\mathbf{Y}}$ be an $n \times T$ matrix with all entries of the ith row being $\frac{\sum_{t=1}^{T} y_{it}}{T}$. We next specify some conditions so that Theorem 2 can be applied to the sample covariance matrix: $(\mathbf{Y} - \bar{\mathbf{Y}})^*(\mathbf{Y} - \bar{\mathbf{Y}})$.

Assumption 5

Assumption (Moment Conditions)

$\{x_{it}\colon\ i = 1, ..., n,\ t = 1 - L, ..., T\}$ are independent random variables such that $\mathbb{E}x_{it} = 0$. $\mathbb{E}|\sqrt{n}x_{it}|^2 = 1$ and $\mathbb{E}|\sqrt{n}x_{it}|^4 = \gamma_4 < \infty$. Write $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)$.

Assumption 6

Assumption (Cross-sectional Structure)

$\mathbf{\Omega} = \mathbf{\Upsilon}^*\mathbf{\Upsilon}$ satisfies Assumption 3 with the $n \times n$ matrix $\mathbf{\Upsilon} = (\Upsilon_{is})$.

Assumption 7

Assumption (Serial Correlation)

The coefficients $\{b_i\}_{i=0}^{L}$ in (12) satisfy $\sum_{i=0}^{L} i|b_i| < \infty$ and $\sum_{i=0}^{L} b_i \neq 0$.

The form of the unit root

Write $\mathbf{H} = \mathbf{I} - \frac{\mathbf{1}\mathbf{1}^*}{T}$, where the $T \times 1$ vector $\mathbf{1}$ consists of all one. Let $\mathbf{\Gamma} = \mathbf{HCW}$ and $\mu_1 \geq \mu_2 \geq ... \geq \mu_T$ be the ordered eigenvalues of $\mathbf{\Gamma\Gamma}^*$. With a simple calculation, the sample covariance matrix can be expressed as

$$(\mathbf{Y} - \bar{\mathbf{Y}})^*(\mathbf{Y} - \bar{\mathbf{Y}}) = \mathbf{H}\mathbf{Y}^*\mathbf{Y}\mathbf{H}^* = \mathbf{\Gamma}\mathbf{X}\mathbf{\Omega}\mathbf{X}^*\mathbf{\Gamma}^*. \tag{14}$$

Theorem of the unit root

Theorem

Let Assumptions 5-7 hold. Denoting the k-th largest eigenvalue of (14) by λ_k, for any fixed k, we have

$$\frac{n}{\sqrt{2tr(\mathbf{\Omega}^2)}}\left(\frac{\lambda_1-\mu_1\frac{tr\mathbf{\Omega}}{n}}{\mu_1},\frac{\lambda_2-\mu_2\frac{tr\mathbf{\Omega}}{n}}{\mu_2},\dots,\frac{\lambda_k-\mu_k\frac{tr\mathbf{\Omega}}{n}}{\mu_k}\right)\xrightarrow{d}N(\mathbf{0},\mathbf{I_k}).\ (15)$$

Propositions 1-2 hold as well for model (12).

Outline

1. Motivation
   - Misleading PCA on Simulated Data
2. High Dimensional Separable Covariance Model
3. Asymptotic Performance of Largest Eigenvalues
4. **Inference on High Dimensional Time Series**
   - Implementing Factor Analysis on Our Model
   - Unit Root Models Satisfying Assumption 4
   - **A New Test for Unit Root against Factor Model**
   - More Thoughts about Panel Data Structures
5. Simulations
   - The Simulation about Proposition 3
   - The Simulation about Proposition 4
6. Reference

Two models

We below focus on a nonstationary factor model of the form:

$$M1: \ y_{it} = \boldsymbol{\ell}_i^* \mathbf{f}_t + u_{it}, \tag{16}$$

where $\mathbf{f}_t$ is a r-dimensional(r is fixed) vector, u_{it} is a stationary term and $\{\mathbf{f}_t\}_{t=1,\cdots,T}$ are independent of $\{u_{it}\}_{i=1,\cdots,n,t=1,\cdots,T}$.
We then recall the unit root model discussed in Theorem 3 as follows:

$$M2: \ y_{it} = y_{i,t-1} + \sum_{h=0}^{L} b_h z_{i,t-h} \tag{17}$$

for $1 \leq i \leq n$ and $1 \leq t \leq T$, where L can be finite or infinite, $z_{it} = \sum_{s=1}^{n} \Upsilon_{is} x_{st}$ and Assumptions 5-7 hold.

some remarks

Note that Model M1 is equivalent to the following form:

$$M3: \ y_{it} = y_{i,t-1} + \boldsymbol{\ell}_i^*(\mathbf{f}_t - \mathbf{f}_{t-1}) + u_{it} - u_{i,t-1} \triangleq y_{i,t-1} + v_{it}, \tag{18}$$

which seems to be similar to model M2. However M3 is different from M2 in two aspects.

- At first, if $\mathbf{f}_t - \mathbf{f}_{t-1} \neq 0$, $\boldsymbol{\ell}_i^*(\mathbf{f}_t - \mathbf{f}_{t-1})$ could lead to a strong cross-sectional dependence (strong factor) such that Assumption 6 is violated.
- Furthermore, even if $\mathbf{f}_t - \mathbf{f}_{t-1} = 0$, then $v_{it} = u_{it} - u_{i,t-1}$ doesn't necessarily satisfy $\sum_{i=0}^{L} b_i \neq 0$ in Assumption 7.
- Here one should note that $(\sum_{i=0}^{L} b_i)^2$ contributes to the limit of the first few largest eigenvalues of the corresponding sample covariance matrix.

Differences between M1 and M2

As a result, the eigenvalues of these two models behavior differently.

- When the number of factors in M1 is r, there is a significant drop from the r-th largest eigenvalue of M1 to its $(r+1)$-th largest eigenvalue.
- In contrast the ratio of the i-th largest eigenvalue of M2 to its $(i+1)$-th largest eigenvalue is asymptotically equal to $(i+1)^2/i^2$ so that there is no significant drop between them.

The new test statistics

We below propose a new statistic for M1 and M2. Define

$$\bar{\mu}_i = \frac{1}{2\left(1 + \cos\left(\frac{(T-i)\pi}{T}\right)\right)} \tag{19}$$

and

$$T_{uf} = \frac{\lambda_1 \frac{\bar{\mu}_2}{\bar{\mu}_1} - \lambda_2}{\lambda_2 - \lambda_3 \frac{\bar{\mu}_2}{\bar{\mu}_3}}. \tag{20}$$

The new test statistics

Proposition

Under the conditions of Theorem 3, for any fixed k,

$$\lim_{n,T\to\infty} P\left(\tilde{k}_{ER} = \max_{0\leq i\leq k}\left\{\frac{\lambda_i}{\lambda_{i+1}}\right\} = 1\right) = 1. \tag{21}$$

Furthermore, when

$$\lim_{n,T\to\infty} \frac{\mathrm{tr}(\mathbf{\Omega})}{T\sqrt{\mathrm{tr}(\mathbf{\Omega^2})}} = 0, \tag{22}$$

the statistic T_{uf} has the same asymptotic distribution as $\frac{v_1-v_2}{v_2-v_3}$, where $\{v_i : 1 \leq i \leq 3\}$ are i.i.d standard normal variables.

The new test statistics

Remark

Note that $\frac{tr(\mathbf{\Omega})}{\sqrt{tr(\mathbf{\Omega^2})}} \leq \sqrt{n}$. So when $\frac{\sqrt{n}}{T} \to 0$, (22) holds.

Equation (21) implies that using $\tilde{k}_{ER}$ in Ahn and Horensten (2013) may mistakenly think a unit root model as a single factor model.

The new test statistics

However using the statistic T_{uf} could distinguish between them because for single factor models (see Assumption 1 in Onatski (2010), Assumptions A and B in Bai (2004) and Assumption A in Ahn and Horensten (2013)),

$$\frac{\lambda_1}{\lambda_2} \to \infty \text{ in probability as } n, T \to \infty. \tag{23}$$

This ensures the power of the statistic T_{uf} specified below.

Proposition

In single factor models(under the assumptions of Onatski (2010), Bai (2004) or Ahn and Horensten (2013)), the following holds

$$T_{uf} \to \infty \text{ in probability as } n, T \to \infty. \tag{24}$$

Outline

1. Motivation
   - Misleading PCA on Simulated Data
2. High Dimensional Separable Covariance Model
3. Asymptotic Performance of Largest Eigenvalues
4. **Inference on High Dimensional Time Series**
   - Implementing Factor Analysis on Our Model
   - Unit Root Models Satisfying Assumption 4
   - A New Test for Unit Root against Factor Model
   - More Thoughts about Panel Data Structures
5. Simulations
   - The Simulation about Proposition 3
   - The Simulation about Proposition 4
6. Reference

four kinds of panel data structures

We have more thoughts about distinguishing four kinds of panel data structures:

(1) stationary and weak cross-sectional dependence;

(2) stationary and strong cross–sectional dependence;

(3) unit root and weak cross–sectional dependence;

(4) unit root and strong cross–sectional dependence.

Here by strong cross–sectional dependence we mean that its effective rank $r^*(\mathbf{\Omega}) = \frac{tr(\mathbf{\Omega})}{\|\mathbf{\Omega}\|_2} \to c > 0$ while weak cross–sectional dependence implies that its effective rank $r^*(\mathbf{\Omega}) \to +\infty$.

stationary and weak cross-sectional dependence

- Theorem 2.3 of Zhang, Pan and Gao (2018) shows that when the data belongs to the first kind, the largest eigenvalue of sample covariance matrix has smaller order than the trace.
- On the other hand, the largest eigenvalues from three other types of data have the same order as the trace.
- So we can distinguish the first type, stationary and weak cross–sectional dependence, from others.

the remaining three cases

For the remaining three cases, we consider using PCA for them.

- Since PCA is a linear combination of data on the cross section we believe it has the same time-dependence as the initial data.
- In other words, from the first PC we may tell the difference between the second type stationary structure and the remaining two nonstationary structures since there are plenty of methods available for the univariate variable.
- So we can distinguish the second case from two others.
- Finally, we can use T_{uf} to distinguish the third case from the fourth case.

Outline

1. Motivation
   - Misleading PCA on Simulated Data
2. High Dimensional Separable Covariance Model
3. Asymptotic Performance of Largest Eigenvalues
4. Inference on High Dimensional Time Series
   - Implementing Factor Analysis on Our Model
   - Unit Root Models Satisfying Assumption 4
   - A New Test for Unit Root against Factor Model
   - More Thoughts about Panel Data Structures
5. **Simulations**
   - The Simulation about Proposition 3
   - The Simulation about Proposition 4
6. Reference

Outline

1. Motivation
   - Misleading PCA on Simulated Data
2. High Dimensional Separable Covariance Model
3. Asymptotic Performance of Largest Eigenvalues
4. Inference on High Dimensional Time Series
   - Implementing Factor Analysis on Our Model
   - Unit Root Models Satisfying Assumption 4
   - A New Test for Unit Root against Factor Model
   - More Thoughts about Panel Data Structures
5. **Simulations**
   - The Simulation about Proposition 3
   - The Simulation about Proposition 4
6. Reference

the critical value

At first we compute the critical value by simulating $\frac{v_1 - v_2}{v_2 - v_3}$ where $\{v_i : 1 \leq i \leq 3\}$ are i.i.d standard Gaussian random variables based on 500000 replications. The quantiles of $\frac{v_1 - v_2}{v_2 - v_3}$ are reported in Table 1.

Table: The quantiles of $\frac{v_1 - v_2}{v_2 - v_3}$ based on 500000 replications

2.5% quantile	5% quantile	95% quantile	97.5% quantile
-11.6549	-6.0392	4.9932	10.4598

Then we can use a two-tailed test with the critical values -11.6549 and 10.4598. We can also use one-side test with the critical values $C_{5L} = -6.0392$ or $C_{95R} = 4.9932$.

the setting and the size

We consider the following setting:

$$y_{it} = y_{i,t-1} + \psi z_{i,t-1} + z_{it},$$

where $\psi = 0.5$ and $\mathbf{\Omega} = \left(\Omega_{i,j}\right) = \left(0.3^{|i-j|}\right)$. The estimated sizes for the test statistic T_{uf} based on 1000 replications, different critical values and different values of n and T are reported in Tables 1-3. Tables 1-3 show that T_{uf} has stable sizes with different critical values. We can choose -11.6549 and 10.4598 as the critical values of a two-tailed test.

the size

Table: 1 The size results on T_{uf} based on 1000 replications

n	T	two-side	C_{5L}	C_{95R}
20	20	0.069	0.064	0.052
20	40	0.065	0.068	0.051
20	60	0.067	0.059	0.055
20	80	0.074	0.062	0.063
20	100	0.064	0.060	0.060
20	200	0.055	0.064	0.054
40	20	0.062	0.069	0.063
40	40	0.070	0.061	0.062
40	60	0.055	0.063	0.057
40	80	0.052	0.068	0.047
40	100	0.052	0.054	0.056
40	200	0.060	0.059	0.046

the size

Table: 2 The size results on T_{uf} based on 1000 replications

n	T	two-side	C_{5L}	C_{95R}
60	20	0.051	0.045	0.051
60	40	0.048	0.051	0.048
60	60	0.047	0.050	0.047
60	80	0.055	0.057	0.056
60	100	0.050	0.043	0.055
60	200	0.053	0.050	0.057
80	20	0.062	0.052	0.064
80	40	0.059	0.047	0.051
80	60	0.051	0.051	0.057
80	80	0.054	0.059	0.048
80	100	0.047	0.058	0.047
80	200	0.054	0.044	0.055

the size

Table: 3 The size results on T_{uf} based on 1000 replications

n	T	two-side	C_{5L}	C_{95R}
100	20	0.058	0.048	0.050
100	40	0.057	0.061	0.042
100	60	0.035	0.046	0.044
100	80	0.051	0.058	0.044
100	100	0.051	0.040	0.054
100	200	0.053	0.059	0.039
200	20	0.048	0.040	0.059
200	40	0.058	0.055	0.045
200	60	0.046	0.043	0.048
200	80	0.046	0.050	0.044
200	100	0.055	0.048	0.057
200	200	0.066	0.046	0.063

$\tilde{k}_{ER}$

We also calculate the proportion of $\tilde{k}_{ER} = 1$ with different values of the prescribed upper bound k in (21). Tables 4-6 show that (21) also works well, since the calculated proportion is approaching 1 as the dimension n and T both increase.

$\tilde{k}_{ER}$

Table: 4 The proportion of $\tilde{k}_{ER} = 1$ based on 1000 replications

n	T	$k = 5$	$k = 10$	$k = 15$
20	20	0.720	0.712	0.701
20	40	0.733	0.732	0.732
20	60	0.749	0.749	0.749
20	80	0.770	0.770	0.770
20	100	0.754	0.754	0.754
20	200	0.766	0.766	0.766
40	20	0.816	0.815	0.815
40	40	0.841	0.841	0.841
40	60	0.835	0.835	0.835
40	80	0.835	0.835	0.835
40	100	0.824	0.824	0.824
40	200	0.838	0.838	0.838

$\tilde{k}_{ER}$

Table: 5 The proportion of $\tilde{k}_{ER} = 1$ based on 1000 replications

n	T	$k = 5$	$k = 10$	$k = 15$
60	20	0.874	0.874	0.874
60	40	0.870	0.870	0.870
60	60	0.902	0.902	0.902
60	80	0.897	0.897	0.897
60	100	0.877	0.877	0.877
60	200	0.897	0.897	0.897
80	20	0.920	0.920	0.920
80	40	0.933	0.933	0.933
80	60	0.909	0.909	0.909
80	80	0.914	0.914	0.914
80	100	0.925	0.925	0.925
80	200	0.914	0.914	0.914

$\tilde{k}_{ER}$

Table: 6 The proportion of $\tilde{k}_{ER} = 1$ based on 1000 replications

n	T	$k = 5$	$k = 10$	$k = 15$
100	20	0.936	0.936	0.936
100	40	0.950	0.950	0.950
100	60	0.931	0.931	0.931
100	80	0.937	0.937	0.937
100	100	0.946	0.946	0.946
100	200	0.945	0.945	0.945
200	20	0.981	0.981	0.981
200	40	0.986	0.986	0.986
200	60	0.989	0.989	0.989
200	80	0.986	0.986	0.986
200	100	0.988	0.988	0.988
200	200	0.982	0.982	0.982

Outline

1. Motivation
   - Misleading PCA on Simulated Data
2. High Dimensional Separable Covariance Model
3. Asymptotic Performance of Largest Eigenvalues
4. Inference on High Dimensional Time Series
   - Implementing Factor Analysis on Our Model
   - Unit Root Models Satisfying Assumption 4
   - A New Test for Unit Root against Factor Model
   - More Thoughts about Panel Data Structures
5. **Simulations**
   - The Simulation about Proposition 3
   - The Simulation about Proposition 4
6. Reference

Simulations

Now we consider the single factor model:

$$y_{it} = l_i f_t + \sqrt{\theta} e_{it}. \tag{25}$$

We use the same error term e_{it} as in Ahn and Horensten (2013):$e_{it} = \sqrt{\frac{1-\rho^2}{1+2J\beta}}\tilde{e}_{it}$

$$\tilde{e}_{it} = \rho \tilde{e}_{i,t-1} + \epsilon_{it} + \sum_{h=\max(i-J,1)}^{i-1} \beta \epsilon_{ht} + \sum_{i+1}^{h=\min(i+J,n)} \beta \epsilon_{ht}, \tag{26}$$

where ϵ_{it} and l_i are all drawn independently from $N(0,1)$. We also use the most complicated setting of Ahn and Horensten (2013) which has both serially and cross-sectionally correlated errors: $\rho = 0.5, \beta = 0.2$ and $J = \max(10, n/20)$.

Simulations

Since $\rho < 1$, we can find that the error term is stationary. If f_t is also stationary, y_{it} is stationary. Then it will be very different from the unit root model and there are too many methods to test stationary and unit root.
So we focus on the case where f_t is nonstationary. We set $f_t = f_{t-1} + \tilde{f}_t$, where $\tilde{f}_t$ is drawn independently from $N(0,1)$.
Then the power of T_{uf} based on 1000 replications, different θ, the critical values of the two-sided test and different values of n and T, are reported in Tables 7-9. The power results given in Tables demonstrate that the proposed test works well numerically.

The power

Table: 7 The power results on T_{uf} based on 1000 replications and two-sided test

n	T	$\theta = 3$	$\theta = 1$	$\theta = 1/3$
20	20	0.087	0.232	0.582
20	40	0.114	0.380	0.725
20	60	0.147	0.485	0.843
20	80	0.197	0.605	0.896
20	100	0.249	0.634	0.953
20	200	0.455	0.868	0.995
40	20	0.116	0.295	0.679
40	40	0.173	0.474	0.848
40	60	0.260	0.637	0.946
40	80	0.293	0.715	0.974
40	100	0.389	0.768	0.980
40	200	0.620	0.932	0.999

The power

Table: 8 The power results on T_{uf} based on 1000 replications and two-sided test

n	T	$\theta = 3$	$\theta = 1$	$\theta = 1/3$
60	20	0.082	0.289	0.710
60	40	0.122	0.454	0.848
60	60	0.151	0.554	0.925
60	80	0.223	0.653	0.958
60	100	0.282	0.764	0.979
60	200	0.509	0.934	0.999
80	20	0.077	0.267	0.678
80	40	0.129	0.448	0.854
80	60	0.170	0.544	0.939
80	80	0.206	0.687	0.961
80	100	0.285	0.762	0.982
80	200	0.537	0.925	1

The power

Table: 9 The power results on T_{uf} based on 1000 replications and two-sided test

n	T	$\theta = 3$	$\theta = 1$	$\theta = 1/3$
100	20	0.061	0.288	0.689
100	40	0.095	0.437	0.866
100	60	0.157	0.603	0.968
100	80	0.208	0.709	0.980
100	100	0.287	0.806	0.985
100	200	0.599	0.969	1
200	20	0.041	0.238	0.674
200	40	0.085	0.492	0.894
200	60	0.192	0.680	0.978
200	80	0.327	0.794	0.995
200	100	0.442	0.883	0.999
200	200	0.761	0.991	1

Outline

1. Motivation
   - Misleading PCA on Simulated Data
2. High Dimensional Separable Covariance Model
3. Asymptotic Performance of Largest Eigenvalues
4. Inference on High Dimensional Time Series
   - Implementing Factor Analysis on Our Model
   - Unit Root Models Satisfying Assumption 4
   - A New Test for Unit Root against Factor Model
   - More Thoughts about Panel Data Structures
5. Simulations
   - The Simulation about Proposition 3
   - The Simulation about Proposition 4
6. **Reference**

AHN, S. C. and HORENSTEN, A. R. (2013). Eigenvalue ratio test for the number of factors. *Econometrica* **81**, 1203-1227.

BAI, J. (2004). Estimating cross-section common stochastic trends in nonstationary panel data. *Journal of Econometrics* **122**, 137-183.

BAIK, J. and SILVERSTEIN, J. W. (2006). Eigenvalues of large sample covariance matrices of spiked population models. *J. Multivariate Anal.* **97**, 1382–1408.

ONATSKI, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *Review of Economic and Statistics* **92**, 1004-1016.

ZHANG, B., PAN, G. M and GAO, J. (2018). CLT for largest eigenvalues and unit root tests for high-dimensional nonstationary time series. *Ann. Statist.* **46**, 2186-2215.

Thank You Very Much !