Structural-Factor Modeling of Big Dependent Data

Ruey S. Tsay

Booth School of Business, University of Chicago

January 11, 2019

Joint with: Zhaoxing Gao

Table of Contents

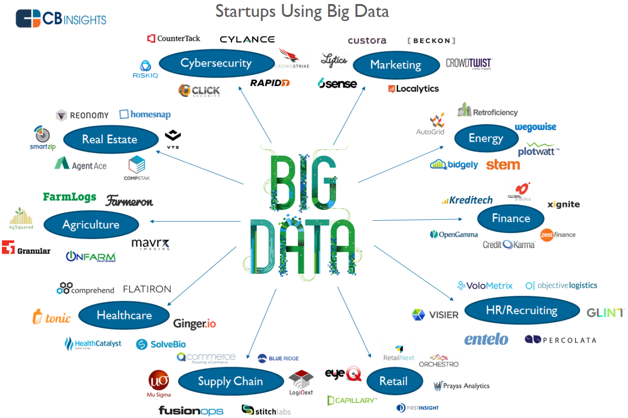
- Some Big Data Examples
- Available Statistical Methods
- Approximate Factor Model and Its limitations
- The Proposed Model and Methodology
- Theoretical Properties
- Numerical Results
- Conclusions

Data tsunami

Information and technology have revolutionized data collection.

- Millions of surveillance video cameras and billions of Internet searches and social media chats and tweets produce massive data that contain vital information about security, public health, consumer preference, business sentiments, economic health, among others.
- Billions of prescriptions and enormous amount of genetic and genomic information provide critical data on health and precision medicine.

Big data are ubiquitous



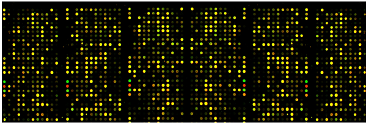
'There were 5 exabytes of information created between the dawn of civilization through 2003, but that much information is now created every 2 days' – Eric Schmidt, Former CEO of Google

What is Big Data?

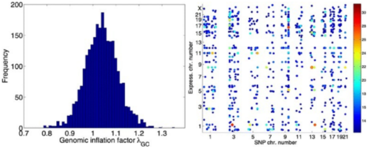
- Large and Complex data: *Structured data* (n and p are both large); *Unstructured data* (email, text, web, videos)
    - Biological Sci.: Genomics, Medicine, Genetics, Neuroscience
    - Engineering: Machine Learning, Computer Vision, Networks
    - Social Sci.: Economics, Business and Digital Humanities
    - Natural Sci.: Meteorology, Earth Science, Astronomy
- Characterizes contemporary scientific and decision problems

Examples: Biological Sciences

- Bioinformatics: disease classification/predicting clinical outcomes using microarray data or proteomic data



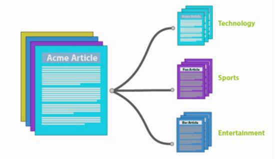
- Association studies between phenotypes and SNPs (eQTL)

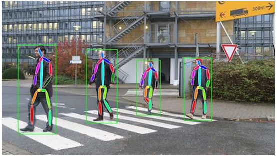
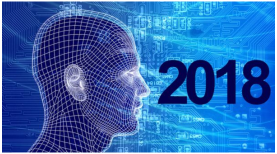


- Detecting activated voxels after stimuli in Neuroscience

Example: Machine Learning

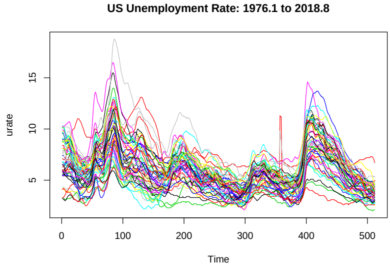
- Document or text classification: E-mail spam
- Computer vision, object classification (images, curves)
- Social media and Internet
- Online learning and recommendation
- Surveillance videos and network security

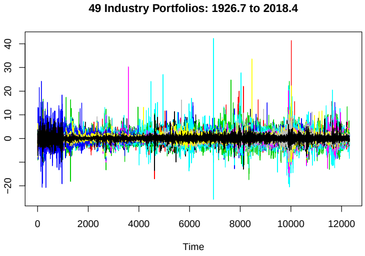


Example: Finance, Economics and Business

- Data: Stock, currency, derivative, commodity, high-frequency trades, macroeconomic, unstructured news and texts, consumers' confidence and business sentiments in social media and Internet

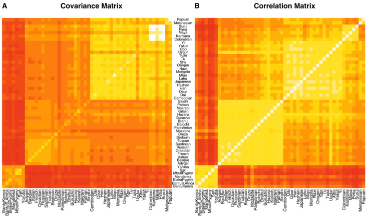




- Social media contains useful information on economic health, consumer confidence and preference, suppliers and demands
- Retail sales provide useful information on public health, economic health, consumer confidence and preference, etc.

Example: Finance, Economics and Business

- Risk and portfolio management: Managing 2K stocks involves 2m elements in covariance
- Credit: Default probability depends on firm specific attributes, market conditions, macroeconomic variables, feedback effects of firms, etc.



- Predicting Housing prices: 1000 neighborhoods require 1m parameters using, e.g. VAR(1),

$$\mathbf{X}_t = \mathbf{A}\mathbf{X}_{t-1} + \varepsilon_t.$$

What can Big Data do?

Hold great promises for understanding

- Heterogeneity: personalized medicine or services
- Commonality: in presence of large variations (noises)
- Dependence: financial data series

from large pools of variables, factors, genes, environments, and their interactions as well as latent factors.

Available statistical methods (TS)

1. Focus on *sparsity*
   - LASSO: Tibshirani (1996)
   - Group Lasso: Yuan and Lin (2006)
   - Elastic net: Zou and Hastie (2005)
   - SCAD: Fan and Li (2001)
   - Fused LASSO: Tibshirani et al. (2005)
2. Focus on *dimension reduction*
   - PCA: Pearson (1901)
   - CCA: Box and Tiao (1977)
   - SCM: Tiao and Tsay (1989)
   - Factor model: Peña and Box (1987), Bai and Ng (2002), Stock and Watson (2005), Lam and Yao (2011, 2012) etc.

Approximate factor model (Econ. & Finance)

The model:

$$\mathbf{y}_t = \mathbf{A}\mathbf{x}_t + \varepsilon_t, \tag{1}$$

where

- $\{\mathbf{y}_1, ..., \mathbf{y}_n\}$ with $\mathbf{y}_t = (y_{1t}, ..., y_{pt})' \in \mathbb{R}^p$ which are observable
- $\mathbf{A} \in \mathbb{R}^{p \times r}$, $\mathbf{x}_t \in \mathbb{R}^r$ are unknown
- ε_t is the idiosyncratic component

The goal

- Estimate the loading matrix $\mathbf{A}$
- Recover the factor process $\mathbf{x}_t$
- Estimate the number of common factors r

Available methods

1. Principal Component Analysis (PCA): Bai and Ng (2002, *Econometrica*), Bai (2003, *Econometrica*)...
   - ε_t is not necessarily white noise
   - $\widehat{\boldsymbol{\Sigma}}_y = n^{-1}\sum_{t=1}^{n} \mathbf{y}_t\mathbf{y}_t' = \widehat{\mathbf{P}}\widehat{\mathbf{D}}\widehat{\mathbf{P}}'$, $\widehat{\mathbf{A}} = \widehat{\mathbf{P}}_r\widehat{\mathbf{D}}_r^{1/2}$, $\widehat{\mathbf{x}}_t = \widehat{\mathbf{D}}_r^{-1/2}\widehat{\mathbf{P}}_r'\mathbf{y}_t$
   - $\widehat{\varepsilon}_t = \mathbf{y}_t - \widehat{\mathbf{A}}\widehat{\mathbf{x}}_t = (\mathbf{I}_p - \widehat{\mathbf{P}}_r\widehat{\mathbf{P}}_r')\mathbf{y}_t$
2. Eigen-analysis on Auto-covariances: Lam, Yao and Bathia (2011, *Biometrika*), Lam and Yao (2012, *AoS*)
   - Assume ε_t is vector white noise
   - $\widehat{\mathbf{M}} = \sum_{k=1}^{k_0} \widehat{\boldsymbol{\Gamma}}_k\widehat{\boldsymbol{\Gamma}}_k'$ with $\widehat{\boldsymbol{\Gamma}}_k = n^{-1}\sum_{t=k+1}^{n} \mathbf{y}_t\mathbf{y}_{t-k}'$, k_0 is fixed
   - $\widehat{\mathbf{A}}$ contains the eigenvectors of $\widehat{\mathbf{M}}$ corresponding to top r eigenvalues
   - $\widehat{\mathbf{x}}_t = \widehat{\mathbf{A}}'\mathbf{y}_t$, $\widehat{\varepsilon}_t = \mathbf{y}_t - \widehat{\mathbf{A}}\widehat{\mathbf{x}}_t = (\mathbf{I}_p - \widehat{\mathbf{A}}\widehat{\mathbf{A}}')\mathbf{y}_t$

Some fundamental issues

- PCA may fail if signal-to-noise ratio is low. For the analysis of high-dimensional financial series, as the market and economic information accumulates, the noise is often increasing faster than the signal;
- PCA cannot distinguish signal and noises in some sense. For example, some components in $\widehat{\mathbf{x}}_t$ might be white noises;
- $\widehat{\mathbf{x}}_t$ in Lam and Yao (2011) includes the noise components. When the largest eigenvalues of the noise covariance are diverging, the resulting estimators would deteriorate.;
- The information criterion in Bai and Ng (2002) and the ratio-based method in Lam and Yao (2011) may also fail if the largest eigenvalues of the covariance matrix of the noise are diverging.
- The sample covariance matrix of the estimated noises is singular if $r > 0$.

Contributions of the proposed method

1. Address the aforementioned issues from a different perspective
2. Provide a new model to understand the mechanism of factor models
3. Propose a Projected PCA to eliminate the (diverging) effect of the idiosyncratic terms
4. A new way to identify the number of factors, which is more reliable than the information criterion and ratio-based method.

Setting

Assume $\mathbf{y}_t$ with $E\mathbf{y}_t = \mathbf{0}$ and admits a latent structure:

$$\mathbf{y}_t = \mathbf{L}\begin{bmatrix} \mathbf{f}_t \\ \boldsymbol{\varepsilon}_t \end{bmatrix} = [\mathbf{L}_1, \mathbf{L}_2]\begin{bmatrix} \mathbf{f}_t \\ \boldsymbol{\varepsilon}_t \end{bmatrix} = \mathbf{L}_1\mathbf{f}_t + \mathbf{L}_2\boldsymbol{\varepsilon}_t, \qquad (2)$$

- $\mathbf{L} \in \mathbb{R}^{p\times p}$ is a full rank loading matrix, implying $\mathbf{L}^{-1}\mathbf{y}_t = \begin{bmatrix} \mathbf{f}_t \\ \boldsymbol{\varepsilon}_t \end{bmatrix}$,
- $\mathbf{f}_t = (f_{1t}, \ldots, f_{rt})'$ is a r-dimensional factor process,
- $\boldsymbol{\varepsilon}_t = (\varepsilon_{1t}, \ldots, \varepsilon_{vt})'$ is a v-dimensional white noise vector,
- r is a small and fixed nonnegative integer. Cov$(\mathbf{f}_t) = \mathbf{I}_r$, Cov$(\boldsymbol{\varepsilon}_t) = \mathbf{I}_v$, Cov$(\mathbf{f}_t, \boldsymbol{\varepsilon}_t) = 0$, and no linear combination of $\mathbf{f}_t$ is serially uncorrelated.

Where does this come from?

Canonical Correlation Analysis (CCA): Let $\boldsymbol{\eta}_t = (\mathbf{y}'_{t-1}, ..., \mathbf{y}'_{t-m})'$ for sufficient large m,

- $\boldsymbol{\Sigma}_y = \text{Cov}(\mathbf{y}_t)$, $\boldsymbol{\Sigma}_\eta = \text{Cov}(\boldsymbol{\eta}_t)$ and $\boldsymbol{\Sigma}_{y\eta} = \text{Cov}(\mathbf{y}_t, \boldsymbol{\eta}_t)$
- $\mathbf{L}^{-1}$ contains the eigenvectors of $\boldsymbol{\Sigma}_y^{-1}\boldsymbol{\Sigma}_{y\eta}\boldsymbol{\Sigma}_\eta^{-1}\boldsymbol{\Sigma}'_{y\eta}$ associated with its descending eigenvalues,
- $\mathbf{L}^{-1}\mathbf{y}_t$ has uncorrelated components and their correlation with the past lagged variables are in decreasing order
- Assume the top r eigenvalues are non-zero
- $\mathbf{L}^{-1}\mathbf{y}_t = (\mathbf{f}'_t, \boldsymbol{\varepsilon}'_t)'$

See Tiao and Tsay (1989, JRSSB): Include all finite-order VARMA models.

Why CCA does not always work in practice?

A natural method is to adopt the CCA at the sample level to recover the latent structure.

But,

- the sample covariance matrix is not consistent to the population one in high dimension;
- the sample precision matrix is not consistent to the population one in high dimension.

For instance, when the dimension is greater than the sample size.

Our method

SVD or QR

Let $\mathbf{L}_1 = \mathbf{A}_1\mathbf{Q}_1$ and $\mathbf{L}_2 = \mathbf{A}_2\mathbf{Q}_2$ where $\mathbf{A}_1'\mathbf{A}_1 = \mathbf{I}_r$ and $\mathbf{A}_2'\mathbf{A}_2 = \mathbf{I}_v$,

$$\mathbf{y}_t = \mathbf{L}_1\mathbf{f}_t + \mathbf{L}_2\boldsymbol{\varepsilon}_t = \mathbf{A}_1\mathbf{x}_t + \mathbf{A}_2\mathbf{e}_t, \quad (3)$$

where

- $\mathbf{A}_1$ is not orthogonal to $\mathbf{A}_2$ in general
- $\mathbf{A}_1$ and $\mathbf{x}_t$ are not uniquely identified since we can replace $(\mathbf{A}_1, \mathbf{x}_t)$ by $(\mathbf{A}_1\mathbf{H}, \mathbf{H}'\mathbf{x}_t)$
- The linear space spanned by the columns of $\mathbf{A}_1$, $\mathcal{M}(\mathbf{A}_1)$ is uniquely defined (*all equal to $\mathcal{M}(\mathbf{L}_1)$*)
- Denote $\mathbf{B}_1$ and $\mathbf{B}_2$ as orthonormal complements of $\mathbf{A}_1$ and $\mathbf{A}_2$, resp.

Orthonormal projections: SCM(0,0)

Let the past lagged vector $\boldsymbol{\eta}_t = (\mathbf{y}'_{t-1}, ..., \mathbf{y}'_{t-k_0})'$, we seek a direction $\mathbf{a} \in \mathbb{R}^p$ that solves the following

$$\max_{\mathbf{a}\in\mathbb{R}^p} \|\text{Cov}(\mathbf{a}'\mathbf{y}_t, \boldsymbol{\eta}_t)\|_2^2, \quad \text{subject to} \quad \mathbf{a}'\mathbf{a} = 1. \tag{4}$$

Equivalently, we solve

$$\boldsymbol{\Sigma}_{y\eta}\boldsymbol{\Sigma}'_{y\eta}\mathbf{a} = \lambda\mathbf{a}. \tag{5}$$

Since

$$\mathbf{M} := \boldsymbol{\Sigma}_{y\eta}\boldsymbol{\Sigma}'_{y\eta} = \sum_{k=1}^{k_0} \boldsymbol{\Sigma}_y(k)\boldsymbol{\Sigma}_y(k)' \tag{6}$$

where $\boldsymbol{\Sigma}_y(k) = \text{Cov}(\mathbf{y}_t, \mathbf{y}_{t-k})$, and $\mathbf{MB}_1 = \mathbf{0}$. Then $\mathbf{A}_1$ consists of r columns associated with the r nonzero eigenvalues of $\mathbf{M}$.

Projected PCA

Note that

$$\mathbf{y}_t = \mathbf{A}_1\mathbf{x}_t + \mathbf{A}_2\mathbf{e}_t, \tag{7}$$

and let $\mathbf{B}_1$ and $\mathbf{B}_2$ be the orthogonal complements of $\mathbf{A}_1$ and $\mathbf{A}_2$, respectively. Then

$$\mathbf{B}_1'\mathbf{y}_t = \mathbf{B}_1'\mathbf{A}_2\mathbf{e}_t, \tag{8}$$

$$\mathbf{B}_2'\mathbf{y}_t = \mathbf{B}_2'\mathbf{A}_1\mathbf{x}_t. \tag{9}$$

Observe that $\mathbf{B}_1'\mathbf{y}_t$ and $\mathbf{B}_2'\mathbf{y}_t$ are uncorrelated, then

$$\mathbf{B}_2'\mathbf{\Sigma}_y\mathbf{B}_1\mathbf{B}_1'\mathbf{\Sigma}_y\mathbf{B}_2 = \mathbf{0}, \tag{10}$$

which implies that $\mathbf{B}_2$ consists the eigenvectors corresponding to the zero eigenvalues of $\mathbf{S} := \mathbf{\Sigma}_y\mathbf{B}_1'\mathbf{B}_1'\mathbf{\Sigma}_y$. Once $\mathbf{A}_1$, $\mathbf{B}_1$ and $\mathbf{B}_2$ are given, then

$$\mathbf{x}_t = (\mathbf{B}_2'\mathbf{A}_1)^{-1}\mathbf{B}_2'\mathbf{y}_t. \tag{11}$$

Estimation: r is known

- Given the data $\{\mathbf{y}_t | t = 1, ..., n\}$, the first step is to perform an eigen-analysis on

$$\widehat{\mathbf{M}} = \sum_{k=1}^{k_0} \widehat{\boldsymbol{\Sigma}}_y(k) \widehat{\boldsymbol{\Sigma}}_y(k)', \tag{12}$$

  where $\widehat{\boldsymbol{\Sigma}}_y(k)$ is the lag-k sample auto-covariance matrix of $\mathbf{y}_t$.
- Let $\widehat{\mathbf{A}}_1 = (\widehat{\mathbf{a}}_1, ..., \widehat{\mathbf{a}}_r)$ and $\widehat{\mathbf{B}}_1 = (\widehat{\mathbf{b}}_1, ...\widehat{\mathbf{b}}_v)$. The second step is a projected PCA based on

$$\widehat{\mathbf{S}} = \widehat{\boldsymbol{\Sigma}}_y \widehat{\mathbf{B}}_1 \widehat{\mathbf{B}}_1' \widehat{\boldsymbol{\Sigma}}_y. \tag{13}$$

  That is, we project the data $\mathbf{y}_t$ onto the direction of $\widehat{\mathbf{B}}_1$, then perform PCA between the original data $\mathbf{y}_t$ and its projected coordinates.

Selection of $\widehat{\mathbf{B}}_2$

- p is low: $\widehat{\mathbf{b}}_2 = (\widehat{\mathbf{b}}_{v+1}, ..., \widehat{\mathbf{b}}_p)$, where $\widehat{\mathbf{b}}_{v+1}, ..., \widehat{\mathbf{b}}_p$ are the eigenvectors corresponding to the smallest r eigenvalues of $\widehat{\mathbf{S}}$.
- p is large:
  - Assume the largest K eigenvalues of $\mathbf{\Sigma}_e$ are diverging, which is a reasonable condition in the high-dimensional case;
  - Write $\mathbf{A}_2 = (\mathbf{A}_{21}, \mathbf{A}_{22})$ with $\mathbf{A}_{21} \in \mathbb{R}^{p \times K}$ and $\mathbf{A}_{22} \in \mathbb{R}^{p \times (v-K)}$, consider $\mathbf{B}_2^* = (\mathbf{A}_{22}, \mathbf{B}_2) \in \mathbb{R}^{p \times (p-K)}$ and $\mathbf{B}_2^*$ consists of $p - K$ eigenvectors corresponding to the $p - K$ smallest eigenvalues of $\mathbf{S} = \mathbf{\Sigma}_y \mathbf{B}_1 \mathbf{B}_1' \mathbf{\Sigma}_y$.
  - Let $\widehat{\mathbf{B}}_2^*$ be an estimator of $\mathbf{B}_2^*$ consisting of $p - K$ eigenvectors associated with the $p - K$ smallest eigenvalues of $\widehat{\mathbf{S}}$. We then estimate $\widehat{\mathbf{B}}_2$ by $\widehat{\mathbf{B}}_2 = \widehat{\mathbf{B}}_2^* \widehat{\mathbf{R}}$, where $\widehat{\mathbf{R}} = (\widehat{\mathbf{r}}_1, \ldots, \widehat{\mathbf{r}}_r) \in \mathbb{R}^{(p-K) \times r}$ with $\widehat{\mathbf{r}}_i$ being the eigenvector associated with the i-th largest eigenvalues of $\widehat{\mathbf{B}}_2^{*\prime} \widehat{\mathbf{A}}_1 \widehat{\mathbf{A}}_1' \widehat{\mathbf{B}}_2^*$.
- Recovered factors: $\widehat{\mathbf{x}}_t = (\widehat{\mathbf{B}}_2' \widehat{\mathbf{A}}_1)^{-1} \widehat{\mathbf{B}}_2' \mathbf{y}_t$.

Determination of the number of common factors

Note that

$$\mathbf{B}_1'\mathbf{y}_t = \mathbf{B}_1'\mathbf{A}_2\mathbf{e}_t, \qquad (14)$$

which is a vector white noise process. Let $\widehat{\mathbf{G}}$ be the matrix of eigenvectors (in the decreasing order of eigenvalues) of $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{u}}_t = \widehat{\mathbf{G}}'\mathbf{y}_t = (\widehat{u}_{1t}, ..., \widehat{u}_{pt})'$.

- p is small: using Ljung-Box statistic $Q(m)$ to test the null hypothesis that $\widehat{u}_{it}$ is a white noise starting with $i = p$. If the null hypothesis is rejected, then $\widehat{r} = i$; otherwise, reduce i by one and repeat the testing process.
- p is large: high-dimensional white noise tests, Chang, Yao and Zhou (2017) and Tsay (2018+).

Theoretical properties: p is fixed and $n \to \infty$

Some Assumptions:

A1. The process $\{(\mathbf{y}_t, \mathbf{f}_t)\}$ is α-mixing with the mixing coefficient satisfying the condition $\sum_{k=1}^{\infty} \alpha_p(k)^{1-2/\gamma} < \infty$ for some $\gamma > 2$, where

$$\alpha_p(k) = \sup_i \sup_{A \in \mathcal{F}_{-\infty}^i, B \in \mathcal{F}_{i+k}^{\infty}} |P(A \cap B) - P(A)P(B)|,$$

and $\mathcal{F}_i^j$ is the σ-field generated by $\{(\mathbf{y}_t, \mathbf{f}_t) : i \leq t \leq j\}$.

A2. $E|f_{it}|^{2\gamma} < C_1$ and $E|\varepsilon_{jt}|^{2\gamma} < C_2$ for $1 \leq i \leq r$ and $1 \leq j \leq v$, where $C_1, C_2 > 0$ are some constants and γ is given in Assumption 1.

A3. $\lambda_1 > ... > \lambda_r > \lambda_{r+1} = ... = \lambda_p = 0$, where λ_i is the i-th largest eigenvalue of $\mathbf{M}$.

Theorem 1: p is fixed and $n \to \infty$

Theorem

Suppose Assumptions 1-3 hold and r is known and fixed. Then, for fixed p,

$$\|\widehat{\mathbf{A}}_1 - \mathbf{A}_1\|_2 = O_p(n^{-1/2}), \|\widehat{\mathbf{B}}_1 - \mathbf{B}_1\|_2 = O_p(n^{-1/2}), \|\widehat{\mathbf{B}}_2 - \mathbf{B}_2\|_2 = O_p(n^{-1/2})$$

as $n \to \infty$. Therefore,

$$\|\widehat{\mathbf{A}}_1 \widehat{\mathbf{x}}_t - \mathbf{A}_1 \mathbf{x}_t\|_2 = O_p(n^{-1/2}).$$

- The convergence rates of all estimates are standard at $\sqrt{n}$.

Theorem 2: p is fixed and $n \to \infty$

For two $p \times r$ half orthogonal matrices $\mathbf{H}_1$ and $\mathbf{H}_2$, define

$$D(\mathcal{M}(\mathbf{H}_1), \mathcal{M}(\mathbf{H}_2)) = \sqrt{1 - \frac{1}{r}\text{tr}(\mathbf{H}_1\mathbf{H}_1'\mathbf{H}_2\mathbf{H}_2')}. \quad (15)$$

Note that $D(\mathcal{M}(\mathbf{H}_1), \mathcal{M}(\mathbf{H}_2)) \in [0, 1]$. It is equal to 0 if and only if $\mathcal{M}(\mathbf{H}_1) = \mathcal{M}(\mathbf{H}_2)$, and to 1 if and only if $\mathcal{M}(\mathbf{H}_1) \perp \mathcal{M}(\mathbf{H}_2)$.
Theorem 2. Suppose Assumptions 1-2 hold and r is known and fixed. Then, for fixed p,

$$D(\mathcal{M}(\widehat{\mathbf{A}}_1), \mathcal{M}(\mathbf{A}_1)) = O_p(n^{-1/2}), \quad D(\mathcal{M}(\widehat{\mathbf{B}}_1), \mathcal{M}(\mathbf{B}_1)) = O_p(n^{-1/2})$$

and

$$D(\mathcal{M}(\widehat{\mathbf{B}}_2), \mathcal{M}(\mathbf{B}_2)) = O_p(n^{-1/2}),$$

as $n \to \infty$. The convergence rate of the extracted factors $\widehat{\mathbf{A}}_1\widehat{\mathbf{x}}_t$ is the same as that in Theorem 1.

Theoretical properties: $n \to \infty$ and $p \to \infty$

A4. (i) $\mathbf{L}_1 = (\mathbf{c}_1, ..., \mathbf{c}_r)$ such that $\|\mathbf{c}_j\|_2^2 \asymp p^{1-\delta_1}$, $j = 1, ..., r$ and $\delta_1 \in [0, 1)$; (ii) For each $j = 1, ..., r$ and δ_1 given in (i),

$$\min_{\theta_i \in \mathbb{R}, i \neq j} \|\mathbf{c}_j - \sum_{1 \leq i \leq r, i \neq j} \theta_i \mathbf{c}_i\|_2^2 \asymp p^{1-\delta_1}.$$

A5. (i) $\mathbf{L}_2$ admits a singular value decomposition $\mathbf{L}_2 = \mathbf{A}_2 \mathbf{D}_2 \mathbf{V}_2'$, where $\mathbf{A}_2 \in \mathbb{R}^{p \times v}$ is given before, $\mathbf{D}_2 = \text{diag}(d_1, ..., d_v)$ and $\mathbf{V}_2 \in \mathbb{R}^{v \times v}$ satisfying $\mathbf{V}_2' \mathbf{V}_2 = \mathbf{I}_v$; (ii) There exists a finite integer $0 < K < v$ such that $d_1 \asymp ... \asymp d_K \asymp p^{(1-\delta_2)/2}$ for some $\delta_2 \in [0, 1)$ and $d_{K+1} \asymp ... \asymp d_v \asymp 1$.

A6. $0 \leq \kappa_{\min} \leq \|\boldsymbol{\Sigma}_{f\varepsilon}(k)\|_2 \leq \kappa_{\max}$ for $1 \leq k \leq k_0$, where $\kappa_{\min}$ and $\kappa_{\max}$ can be either finite constants or diverging rates in relation to p and n.

A7. (i) For any $\mathbf{h} \in R^v$ with $\|\mathbf{h}\|_2 = 1$, $E|\mathbf{h}'\boldsymbol{\varepsilon}_t|^{2\gamma} < \infty$; (ii) $\sigma_{\min}(\mathbf{R}'\mathbf{B}_2^{*\prime}\mathbf{A}_1) \geq C_3$ for some constant $C_3 > 0$ and some half orthogonal matrix $\mathbf{R} \in \mathbb{R}^{(p-K) \times r}$ satisfying $\mathbf{R}'\mathbf{R} = \mathbf{I}_r$, where $\sigma_{\min}$ denotes the minimum non-zero singular value of a matrix.

Theorem 3: $p \to \infty$ and $n \to \infty$

Theorem

Suppose Assumptions 1-7 hold and r is known and fixed. As $n \to \infty$, if $p^{\delta_1} n^{-1/2} = o(1)$ or $\kappa_{\max}^{-1} p^{\delta_1/2+\delta_2/2} n^{-1/2} = o(1)$, then

$\|\widehat{\mathbf{A}}_1 - \mathbf{A}_1\|_2 = O_p(p^{\delta_1} n^{-1/2})$ *if* $\kappa_{\max} p^{\delta_1/2-\delta_2/2} = o(1)$,

$\|\widehat{\mathbf{A}}_1 - \mathbf{A}_1\|_2 = O_p(\kappa_{\min}^{-2} p^{\delta_2} n^{-1/2} + \kappa_{\min}^{-2} \kappa_{\max} p^{\delta_1/2+\delta_2/2} n^{-1/2})$ *if* $r \le K, \kappa_{\min}^{-1} p^{\delta_2/2-\delta_1/2} = o(1)$,

$\|\widehat{\mathbf{A}}_1 - \mathbf{A}_1\|_2 = O_p(\kappa_{\min}^{-2} p n^{-1/2} + \kappa_{\min}^{-2} \kappa_{\max} p^{1+\delta_1/2-\delta_2/2} n^{-1/2})$ *if* $r > K, \kappa_{\min}^{-1} p^{(1-\delta_1)/2} = o(1)$,

and the above results also hold for $\|\widehat{\mathbf{B}}_1 - \mathbf{B}_1\|_2$. *Furthermore,*

$$\|\widehat{\mathbf{B}}_2^* - \mathbf{B}_2^*\|_2 = O_p\left(p^{2\delta_2-\delta_1} n^{-1/2} + p^{\delta_2} n^{-1/2} + (1 + p^{2\delta_2-2\delta_1})\|\widehat{\mathbf{B}}_1 - \mathbf{B}_1\|_2\right).$$

Remarks

• If $\kappa_{max} = \kappa_{\min} = 0$, i.e., $\mathbf{f}_t$ and $\boldsymbol{\varepsilon}_s$ are independent for all t and s, we have

$$\|\widehat{\mathbf{A}}_1 - \mathbf{A}_1\|_2 = O_p(p^{\delta_1} n^{-1/2})$$

$$\|\widehat{\mathbf{B}}_2^* - \mathbf{B}_2^*\|_2 = O_p(p^{2\delta_2 - \delta_1} n^{-1/2} + p^{\delta_2} n^{-1/2} + p^{\delta_1} n^{-1/2}).$$

To guarantee they are consistent, we require $p^{\delta_1} n^{-1/2} = o(1)$, $p^{\delta_2} n^{-1/2} = o(1)$ and $p^{2\delta_2 - \delta_1} n^{-1/2} = o(1)$. When $p \asymp n^{1/2}$, it implies that $0 \leq \delta_1 < 1$, $0 \leq \delta_2 < 1$ and $\delta_2 < (1 + \delta_1)/2$, i.e., the ranges of δ_1 and δ_2 are pretty wide. On the other hand, if $p \asymp n$, we see that $0 \leq \delta_1 < 1/2$, $0 \leq \delta_2 < 1/2$ and $2\delta_2 - \delta_1 < 1/2$, these ranges become narrower if p is large.

• if $\delta_1 = \delta_2 = \delta$, we require $p^{\delta} n^{-1/2} = o(1)$.

Improving the rates

A8. For any $\mathbf{h} \in R^v$ with $\|\mathbf{h}\|_2 = 1$, there exists a constant $C_4 > 0$ such that

$$P(|\mathbf{h}'\boldsymbol{\varepsilon}_t| > x) \leq 2\exp(-C_4 x^2) \quad \text{for any } x > 0.$$

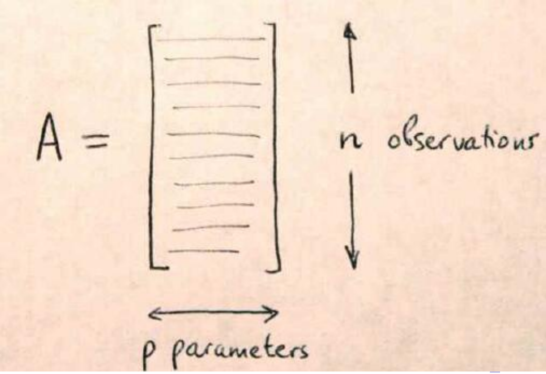
Assumption 8 implies that $\boldsymbol{\varepsilon}_t$ are sub-Gaussian. Examples of sub-Gaussian distributions include the standard normal distribution in R^v, the uniform distribution on the cube $[-1, 1]^v$, among others. See, for example, Vershynin (2018).

Why Sub-Gaussian?

For general ε_t with $E\varepsilon_t = \mathbf{0}$ and $\text{Cov}(\varepsilon_t) = \mathbf{I}_p$,

$$\|\frac{1}{n}\sum_{t=1}^{n} \varepsilon_t \varepsilon_t' - \mathbf{I}_p\|_2 = O_p(pn^{-1/2}).$$

Some famous results in Random Matrix Theory (RMT):



Random Matrix Theory

- Compute the $p \times p$ Wishart matrix $W = A'A$. The eigenvalues of $\sqrt{W}$ are called the singular values of A. For the largest singular values, the eigenvectors of W are the principal components.
- Bai-Yin law (1993): $s_{\min}(A) = \sqrt{n} - \sqrt{p} + o(\sqrt{p})$ and $s_{\max}(A) = \sqrt{n} + \sqrt{p} + o(\sqrt{p})$, under the assumptions that the entries of A are independent copies of a random variable with zero mean, unit variance, and finite fourth moment.
- This easily translates into the statement that the sample covariance matrix $\Sigma_n = \frac{1}{n} A'A$ nicely approximates the actual covariance matrix I_p:

$$\|\Sigma_n - I_p\|_2 \approx 2\sqrt{\frac{p}{n}} + \frac{p}{n}.$$

- Sub-Gaussian also holds: Vershynin (2018).

Improving the rates

Theorem:
Let Assumptions 1-8 hold and r is known and fixed, and $p^{\delta_1/2}n^{-1/2} = o(1)$, $p^{\delta_2/2}n^{-1/2} = o(1)$.
(i) Under the condition that $\delta_1 \le \delta_2$,

$$\|\widehat{\mathbf{A}}_1 - \mathbf{A}_1\|_2 = O_p(\downarrow)$$

$$\begin{cases} p^{\delta_1/2}n^{-1/2}, & \kappa_{\max} p^{\delta_1/2-\delta_2/2} = o(1), \\ \kappa_{\min}^{-2} p^{\delta_2-\delta_1/2} n^{-1/2} + \kappa_{\min}^{-2}\kappa_{\max} p^{\delta_2/2} n^{-1/2}, & r \le K, \kappa_{\min}^{-1} p^{\delta_2/2-\delta_1/2} = o(1 \\ \kappa_{\min}^{-2} p^{1-\delta_1/2} n^{-1/2} + \kappa_{\min}^{-2}\kappa_{\max} p^{1-\delta_2/2} n^{-1/2}, & r > K, \kappa_{\min}^{-1} p^{(1-\delta_1)/2} = o(1) \end{cases}$$

and the above results also hold for $\|\widehat{\mathbf{B}}_1 - \mathbf{B}_1\|_2$, and

$$\|\widehat{\mathbf{B}}_2^* - \mathbf{B}_2^*\|_2 = O_p(p^{2\delta_2-3\delta_1/2}n^{-1/2} + p^{2\delta_2-2\delta_1}\|\widehat{\mathbf{B}}_1 - \mathbf{B}_1\|_2).$$

Improving the rates

(ii) Under the condition that $\delta_1 > \delta_2$, if $\kappa_{\max} = 0$ and $p^{\delta_1 - \delta_2/2} n^{-1/2} = o(1)$, then

$$\|\widehat{\mathbf{A}}_1 - \mathbf{A}_1\|_2 = O_p(p^{\delta_1 - \delta_2/2} n^{-1/2}).$$

If $\kappa_{\max} >> 0$, then

$$\|\widehat{\mathbf{A}}_1 - \mathbf{A}_1\|_2 =$$

$$\begin{cases} O_p(\kappa_{\min}^{-2} \kappa_{\max} p^{\delta_1/2} n^{-1/2}), & r \le K, \kappa_{\min}^{-1} p^{\delta_2/2 - \delta_1/2} = o(1), \\ O_p(\kappa_{\min}^{-2} \kappa_{\max} p^{1+\delta_1/2-\delta_2} n^{-1/2}), & r > K, \kappa_{\min}^{-1} p^{(1-\delta_1)/2} = o(1), \end{cases}$$

and the above results also hold for $\|\widehat{\mathbf{B}}_1 - \mathbf{B}_1\|_2$, and

$$\|\widehat{\mathbf{B}}_2^* - \mathbf{B}_2^*\|_2 = O_p(p^{\delta_2/2} n^{-1/2} + \|\widehat{\mathbf{B}}_1 - \mathbf{B}_1\|_2).$$

- When $\kappa_{\min} = \kappa_{\max} = 0$ and $\delta_1 = \delta_2 = \delta$, we require $p^{\delta/2} n^{-1/2} = o(1)$.

Estimation error: extracted factors

Under the conditions in Theorem 3 or 4, we have

$$p^{-1/2}\|\widehat{\mathbf{A}}_1\widehat{\mathbf{x}}_t - \mathbf{A}_1\mathbf{x}_t\|_2 = O_p(p^{-1/2} + p^{-\delta_1/2}\|\widehat{\mathbf{A}}_1 - \mathbf{A}_1\|_2 + p^{-\delta_2/2}\|\widehat{\mathbf{B}}_2^* - \mathbf{B}_2^*\|_2).$$

• When $\delta_1 = \delta_2 = 0$, i.e. the factors and the noise terms are all strong, the convergence rate in Theorem 5 is $O_p(p^{-1/2} + n^{-1/2})$, which is the optimal rate specified in Theorem 3 of Bai (2003) when dealing with the traditional approximate factor models.

• In practice, Let $\widehat{\mu}_1 \geq ... \geq \widehat{\mu}_p$ be the sample eigenvalues of $\widehat{\mathbf{S}}$ and define $\widehat{K}_L$ as

$$\widehat{K}_L = \arg \min_{1 \leq j \leq \widehat{K}_U} \{\widehat{\mu}_{j+1}/\widehat{\mu}_j\}, \tag{16}$$

we suggest $\widehat{K}_U = \min\{\sqrt{p}, \sqrt{n}, p - \widehat{r}, 10\}$. Then the estimator $\widehat{K}$ for K can assume some value between $\widehat{K}_L$ and $\widehat{K}_U$.

• The consistency of $\widehat{r}$ follows from the consistencies of the corresponding tests.

Numerical results: simulation–small p

Setting: Consider Model (2) with common factors satisfying

$$\mathbf{f}_t = \mathbf{\Phi}\mathbf{f}_{t-1} + \mathbf{z}_t,$$

where $\mathbf{z}_t$ is a white noise process.

- $r = 3$, $p = 5, 10, 15, 20$, $n = 200, 500, 1000, 1500, 3000$
- the elements of $\mathbf{L}$ are drawn independently from $U(-2, 2)$, and the elements of $\mathbf{L}_2$ are then divided by $\sqrt{p}$ to balance the accumulated variances of f_{it} and ε_{it} for each component of $\mathbf{y}_t$. $\mathbf{\Phi}$ is a diagonal matrix with its diagonal elements being drawn independently from $U(0.5, 0.9)$,
- $\boldsymbol{\varepsilon}_t \sim N(0, \mathbf{I}_v)$ and $\mathbf{z}_t \sim N(0, \mathbf{I}_r)$
- We use 1000 replications for each (p, n)

$$\text{RMSE} = \left(\frac{1}{n}\sum_{t=1}^{n} \|\widehat{\mathbf{A}}_1\widehat{\mathbf{x}}_t - \mathbf{L}_1\mathbf{f}_t\|_2^2\right)^{1/2}.$$

Estimation of r

Table: Empirical probabilities $P(\widehat{r} = r)$ of various (p, n) configurations for the model of Example 1 with $r = 3$, where p and n are the dimension and the sample size, respectively. 1000 iterations are used.

		n				
	p	200	500	1000	1500	3000
$r = 3$	5	0.861	0.889	0.890	0.912	0.926
	10	0.683	0.718	0.723	0.735	0.748
	15	0.506	0.555	0.561	0.599	0.601
	20	0.395	0.425	0.441	0.447	0.453

Estimation of loadings and the RMSE

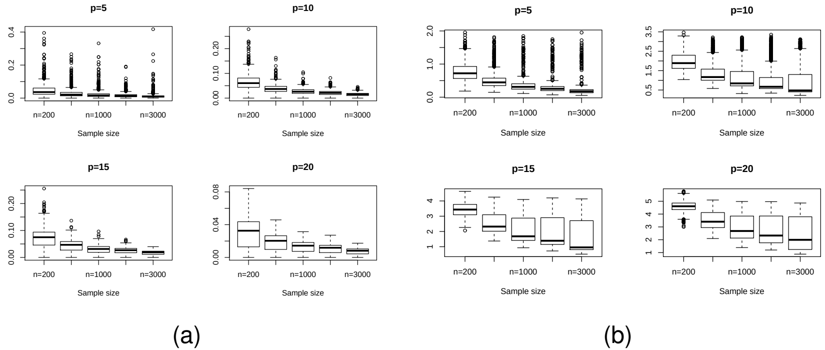


Figure: (a) Boxplots of $\bar{D}(\mathcal{M}(\widehat{\mathbf{A}}_1), \mathcal{M}(\mathbf{L}_1))$ when $r = 3$ under the scenario that p is relatively small in Example 1. (b) Boxplots of the RMSE when $r = 3$ under the scenario that p is relatively small in Example 1. The sample sizes are $200, 500, 1000, 1500, 3000$, respectively.

Simulation: p is large

- In this example, we consider Model (2) with $\mathbf{f}_t$ being the same as that in Example 1.
- $r = 5$; $K = 3, 7$; $p = 50, 100, 300, 500$; $n = 300, 500, 1000, 1500, 3000$;
- $(\delta_1, \delta_2) = (0, 0), (0.4, 0.5)$ and $(0.5, 0.4)$;
- For each setting, the elements of $\mathbf{L}$ are drawn independently from $U(-2, 2)$, and then we divide $\mathbf{L}_1$ by $p^{\delta_1/2}$, the first K columns of $\mathbf{L}_2$ by $p^{\delta_2/2}$ and the rest $v - K$ columns by p to satisfy Assumptions 4-5. $\mathbf{\Phi}$, $\boldsymbol{\varepsilon}_t$ and $\boldsymbol{\eta}_t$ are drawn similarly as those of Example 1. We use 1000 replications in each experiment.

Estimation of r

Table: Empirical probabilities $P(\widehat{r} = r)$ for Example 2 with $r = 5$ and $K = 3$, where p and n are the dimension and the sample size, respectively. δ_1 and δ_2 are the strength parameters corresponding to the factors and the errors, respectively. 1000 iterations are used.

		n				
(δ_1, δ_2)	p	300	500	1000	1500	3000
(0,0)	50	0.510	0.833	0.906	0.917	0.926
	100	0.538	0.799	0.910	0.916	0.922
	300	0.582	0.907	0.916	0.924	0.932
	500	0.560	0.888	0.918	0.928	0.932
(0.4,0.5)	50	0.717	0.903	0.928	0.929	0.935
	100	0.800	0.924	0.938	0.940	0.944
	300	0.858	0.904	0.928	0.932	0.952
	500	0.834	0.922	0.932	0.933	0.948
(0.5,0.4)	50	0.420	0.890	0.910	0.916	0.920
	100	0.508	0.868	0.912	0.928	0.936
	300	0.581	0.910	0.926	0.929	0.932
	500	0.678	0.928	0.936	0.938	0.934

Estimation of r

Table: Empirical probabilities $P(\widehat{r} = r)$ of Example 2 with $r = 5$ and $K = 7$, where p and n are the dimension and the sample size, respectively. δ_1 and δ_2 are the strength parameters corresponding to the factors and the errors, respectively. 1000 iterations are used.

(δ_1, δ_2)	p	n = 300	500	1000	1500	3000
(0,0)	50	0.418	0.688	0.904	0.908	0.910
	100	0.426	0.754	0.910	0.916	0.918
	300	0.406	0.686	0.914	0.925	0.926
	500	0.614	0.778	0.912	0.918	0.920
(0.4,0.5)	50	0.806	0.820	0.892	0.912	0.926
	100	0.800	0.914	0.922	0.904	0.922
	300	0.939	0.935	0.935	0.929	0.930
	500	0.898	0.904	0.926	0.930	0.933
(0.5,0.4)	50	0.332	0.856	0.900	0.928	0.938
	100	0.356	0.716	0.920	0.922	0.928
	300	0.384	0.688	0.924	0.936	0.945
	500	0.421	0.778	0.924	0.930	0.931

Comparisons

- Bai and Ng (2002):

$$\widehat{r} = \arg \min_{1 \le k \le \widetilde{k}} \left\{ \log(\frac{1}{np} \sum_{t=1}^{n} \|\widehat{\varepsilon}_t\|_2^2) + k \left(\frac{p+n}{np} \log(\frac{np}{p+n}) \right) \right\}, \quad (17)$$

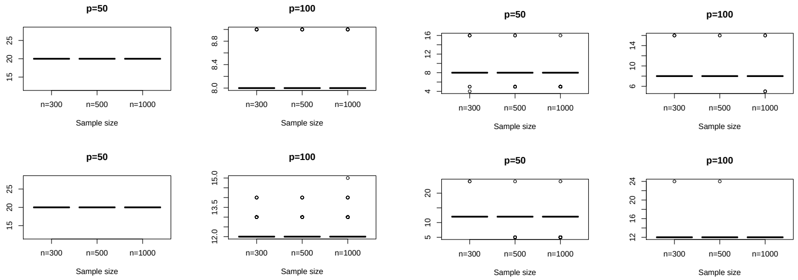
where we choose $\widetilde{k} = 20$ and $\widehat{\varepsilon}_t$ is the p-dimensional residuals obtained by the principal component analysis.

- Lam and Yao (2011):

$$\widehat{r} = \arg \min_{1 \le j \le R} \{ \frac{\widehat{\lambda}_{j+1}}{\widehat{\lambda}_j} \}, \quad (18)$$

where $\widehat{\lambda}_1, ..., \widehat{\lambda}_p$ are the eigenvalues of $\widehat{\mathbf{M}}$.

PCA and ratio



(a) (b)

Figure: (a) Boxplots of $\widehat{r}$ obtained by the information criterion method in (17) corresponding to BN when $r = 5$, $K = 3$ for the upper panel, and $K = 7$ for the lower panel of Example 2; (b) Boxplots of $\widehat{r}$ obtained by the ratio-based method in (18) corresponding to LYB when the true $r = 5$, $K = 3$ for the upper panel, and $K = 7$ for the lower panel of Example 2.

Estimation of loadings

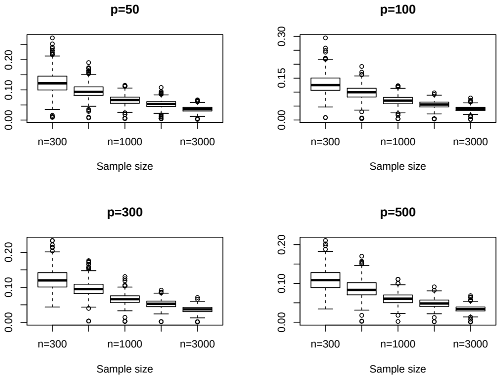


Figure: Boxplots of $\bar{D}(\mathcal{M}(\widehat{\mathbf{A}}_1), \mathcal{M}(\mathbf{L}_1))$ when $r = 3$ and $K = 5$ under the scenario that p is relatively large in Examle 2. $n = 300, 500, 1000, 1500, 3000$, respectively. 1000 iterations are used.

RMSE: comparisons

Table: The RMSE defined in when $r = 5$ and $K = 7$ in Example 2. $n = 300, 500, 1000, 1500, 3000$, respectively. Standard errors are given in the parentheses and 1000 iterations are used. GT denotes the proposed method, BN denotes the principal component analysis and LYB is the ratio one.

Method	p	n = 300	500	1000	1500	3000
GT		**1.510(0.233)**	**1.124(0.235)**	**0.770(0.235)**	**0.627(0.224)**	**0.488(0.273)**
LYB	50	3.056(0.085)	3.051(0.081)	3.056(0.075)	3.053(0.122)	2.976(0.400)
BN		3.058(0.086)	3.053(0.082)	3.058(0.075)	3.059(0.077)	3.055(0.074)
GT		**1.490(0.179)**	**1.148(0.188)**	**0.817(0.141)**	**0.677(0.126)**	**0.519(0.191)**
LYB	100	3.050(0.074)	3.056(0.065)	3.053(0.055)	3.046(0.159)	3.024(0.257)
BN		3.051(0.075)6	3.057(0.065)	3.054(0.055)	3.057(0.055)	3.052(0.052)
GT		**1.729(0.118)**	**1.463(0.107)**	**1.149(0.094)**	**1.107(0.079)**	**0.769(0.077)**
LYB	300	3.052(0.047)	3.055(0.047)	3.053(0.040)	3.056(0.037)	3.056(0.034)
BN		3.053(0.055)	3.056(0.047)	3.054(0.040)	3.056(0.037)	3.057(0.034)
GT		**1.753(0.089)**	**1.547(0.081)**	**1.285(0.052)**	**1.044(0.070)**	**0.861(0.047)**
LYB	500	3.057(0.053)	3.050(0.042)	3.054(0.035)	3.055(0.034)	3.055(0.027)
BN		3.058(0.053)	3.050(0.042)	3.054(0.035)	3.056(0.034)	3.055(0.027)

Real data

• In this example, we consider the daily returns of 49 Industry Portfolios which can be downloaded from `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`. There are many missing values in the data so we only apply the proposed method to the period from July 13, 1988 to November 23, 1990 for a total of 600 observations.

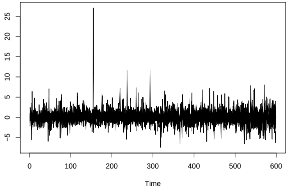


Figure: Time plots of daily returns of 49 Industry Portfolios with 600 observations from July 13, 1988 to November 23, 1990 of Example 3.

Eigenvalues of $\widehat{\mathbf{S}}$

• In the testing, we use $k_0 = 5$ in Equation (12), $m = 10$ in the test statistic $T(m)$, and the upper 95%-quantile 2.97 of the Gumbel distribution as the critical value of the test. We find that $\widehat{r} = 6$.

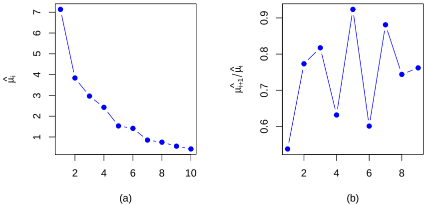


Figure: (a) The first 10 eigenvalues of $\widehat{\mathbf{S}}$ in Example 3; (b) The plots of the ratios for the eigenvalues $\widehat{\mu}_i$ of $\widehat{\mathbf{S}}$. In this example, the largest eigenvalue of $\widehat{\mathbf{x}}_t$ is 10.74, which is almost at the same level as $\widehat{\mu}_1 = 7.14$ of $\widehat{\mathbf{S}}$ with $p = 49$. This empirical phenomenon supports the assumption that the largest eigenvalue of the covariance matrix of the idiosyncratic terms tends to diverge for large p.

Recovered factors

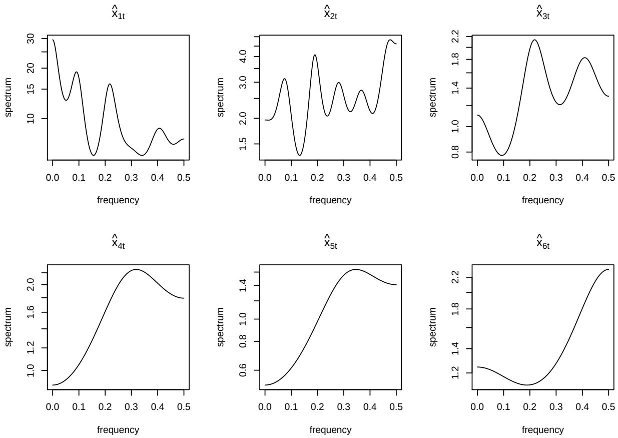


Figure: The spectral densities of 6 estimated common factors using the proposed methodology with $\widehat{K} = 7$ of Example 3.

PCA

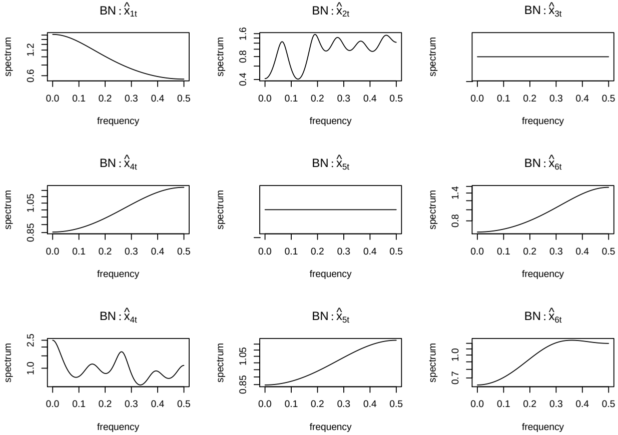


Figure: The spectral densities of the first 9 estimated factors using the principal component analysis in Bai and Ng (2002) of Example 3.

Spectrum of 6 transformed series

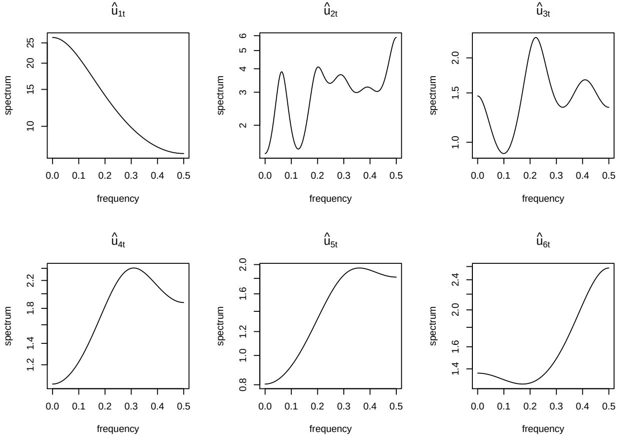


Figure: The spectral densities of first 6 transformed series using the eigen-analysis in Example 3.

Forecasting

Finally, we compare the forecasting performance of the proposed method with those of other methods. For the h-step ahead forecasts, we compare the actual and predicted values of the model estimated using data in the time span $[1, \tau]$ for $\tau = 500, ..., 600 - h$, and the associated h-step ahead forecast error is defined as

$$\text{FE}_h = \frac{1}{100 - h + 1} \sum_{\tau=500}^{600-h} \left(\frac{1}{\sqrt{p}} \|\widehat{\mathbf{y}}_{\tau+h} - \mathbf{y}_{\tau+h}\|_2 \right), \qquad (19)$$

where $p = 49$ in this example.

Forecasting

Table: The 1-step, 2-step and 3-step ahead forecast errors. Standard errors are given in the parentheses. GT denotes our method, BN denotes the principal component analysis in Bai and Ng (2002) and LYB is the one in Lam, Yao and Bathia (2011).

		GT							BN	LYB
		$\widehat{K}=1$	$\widehat{K}=2$	$\widehat{K}=3$	$\widehat{K}=4$	$\widehat{K}=5$	$\widehat{K}=6$	$\widehat{K}=7$		
1-step	AR(1)	1.152	1.161	1.159	1.162	1.158	1.158	1.159	**1.142**	1.157
		(0.469)	(0.484)	(0.482)	(0.489)	(0.487)	(0.483)	(0.487)	(0.442)	(0.465)
	AR(2)	1.164	1.165	1.166	1.168	1.164	1.165	1.164	1.156	1.162
		(0.474)	(0.480)	(0.482)	(0.493)	(0.486)	(0.483)	(0.485)	(0.446)	(0.470)
	AR(3)	1.170	1.172	1.172	1.174	1.169	1.170	1.168	1.168	1.162
		(0.477)	(0.485)	(0.489)	(0.498)	(0.493)	(0.493)	(0.496)	(0.441)	(0.470)
2-step	AR(1)	1.179	1.180	1.180	1.180	1.179	**1.178**	**1.178**	1.182	1.180
		(0.512)	(0.512)	(0.512)	(0.513)	(0.512)	(0.510)	(0.510)	(0.513)	(0.514)
	AR(2)	1.190	1.190	1.190	1.188	1.188	1.187	1.185	1.197	1.185
		(0.519)	(0.514)	(0.514)	(0.513)	(0.514)	(0.512)	(0.512)	(0.520)	(0.519)
	AR(3)	1.194	1.193	1.194	1.191	1.191	1.191	1.189	1.204	1.185
		(0.520)	(0.519)	(0.520)	(0.519)	(0.520)	(0.520)	(0.523)	(0.510)	(0.520)
3-step	AR(1)	1.181	**1.180**	**1.180**	**1.180**	**1.180**	**1.180**	**1.180**	1.184	1.184
		(0.511)	(0.511)	(0.511)	(0.510)	(0.511)	(0.510)	(0.510)	(0.514)	(0.513)
	AR(2)	1.185	1.183	1.183	1.183	1.183	1.182	1.182	1.190	1.187
		(0.510)	(0.510)	(0.508)	(0.508)	(0.508)	(0.507)	(0.508)	(0.514)	(0.512)
	AR(3)	1.187	1.184	1.184	1.184	1.184	1.184	1.184	1.198	1.188
		(0.517)	(0.513)	(0.513)	(0.512)	(0.514)	(0.518)	(0.520)	(0.510)	(0.514)

1-step ahead

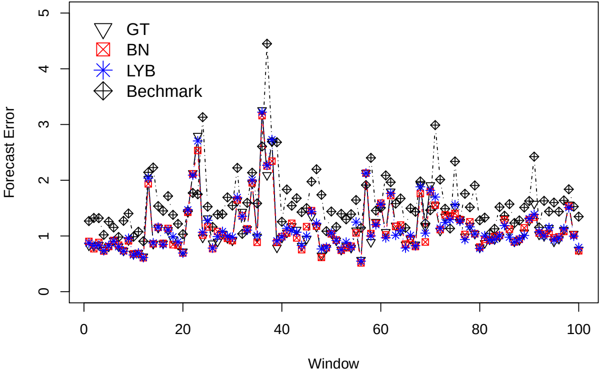


Figure: Time plots of the 1-step ahead point-wise forecast errors using AR(1) and VAR(1) models with $\widehat{K} = 1$ for various methods used in Example 3.

Conclusion

- This article introduced a new structured factor model for high-dimensional time series analysis.
- We allow the largest eigenvalues of the covariance matrix of the idiosyncratic components to diverge to infinity by imposing some structure on the noise terms.
- We propose a Project PCA to mitigate the diverging effect of the noises.
- A new way to identify the number of common factors based on white noise tests.

References

Thank You!

- Gao Z. and Tsay, R.S. (2018+). A Structural-Factor Approach to Modeling High-Dimensional Time Series. , *Revised and Resubmitted.* Available at *arXiv:1808.06518*.
- Gao Z. and Tsay, R.S. (2018+). Structural-Factor Modeling of High-Dimensional Time Series: Another Look at Approximate Factor Models with Diverging Eigenvalues. *Submitted*. Available at *arXiv:1808.07932*.
- Gao Z. and Tsay, R.S. (2018+). Structured Dynamic Matrix-Variate Factor Models. *Manuscript*.