Factor Modelling for High-Dimensional Time Series: a selective survey on dimension reduction approach

Qiwei Yao

Department of Statistics

London School of Economics

q.yao@lse.ac.uk

- Basic setting: ideas, insights

  Pan & Y (2008), Lam, Y & Bathia (2011), Lam & Y (2012)

- Extension

  - Nonstationarity
  - Endogeneity
  - Nonlinearity

  Joint work with Jinyuan Chang (U Melbourne) & Bin Guo (Peking U)

- Driven by hidden Markov chains

  alternative way to model change-point and nonstationarity

  Work in progress with Rong Chen & Xiolu Liu (Rutgers U)

Let $\{\mathbf{y}_t\}$ be a $p \times 1$ time series defined by

$$\mathbf{y}_t = \mathbf{A}\mathbf{x}_t + \boldsymbol{\varepsilon}_t,$$

Let $\{\mathbf{y}_t\}$ be a $p \times 1$ time series defined by

$$\mathbf{y}_t = \mathbf{A}\mathbf{x}_t + \boldsymbol{\varepsilon}_t,$$

$\mathbf{x}_t$: $r \times 1$ unobservable factors, r $(< p)$ unknown

$\mathbf{A}$: $p \times r$ unknown constant factor loading matrix

$\{\boldsymbol{\varepsilon}_t\}$: vector WN($\boldsymbol{\mu}_\varepsilon, \boldsymbol{\Sigma}_\varepsilon$)

no linear combinations of $\mathbf{x}_t$ are WN.

Let $\{\mathbf{y}_t\}$ be a $p \times 1$ time series defined by

$$\mathbf{y}_t = \mathbf{A}\mathbf{x}_t + \boldsymbol{\varepsilon}_t,$$

$\mathbf{x}_t$: $r \times 1$ unobservable factors, r $(< p)$ unknown

$\mathbf{A}$: $p \times r$ unknown constant factor loading matrix

$\{\boldsymbol{\varepsilon}_t\}$: vector WN($\boldsymbol{\mu}_\varepsilon, \boldsymbol{\Sigma}_\varepsilon$)

no linear combinations of $\mathbf{x}_t$ are WN.

Lack of identification: $(\mathbf{A}, \mathbf{x}_t)$ may be replaced by $(\mathbf{A}\mathbf{H}, \mathbf{H}^{-1}\mathbf{x}_t)$ for any invertible $\mathbf{H}$.

Therefore, we assume $\mathbf{A}'\mathbf{A} = \mathbf{I}_r$

But factor loading space $\mathcal{M}(\mathbf{A})$ is uniquely defined

Let $\{\mathbf{y}_t\}$ be a $p \times 1$ time series defined by

$$\mathbf{y}_t = \mathbf{A}\mathbf{x}_t + \boldsymbol{\varepsilon}_t,$$

$\mathbf{x}_t$: $r \times 1$ unobservable factors, r $(< p)$ unknown

$\mathbf{A}$: $p \times r$ unknown constant factor loading matrix

$\{\boldsymbol{\varepsilon}_t\}$: vector WN($\boldsymbol{\mu}_\varepsilon, \boldsymbol{\Sigma}_\varepsilon$)

no linear combinations of $\mathbf{x}_t$ are WN.

Lack of identification: $(\mathbf{A}, \mathbf{x}_t)$ may be replaced by $(\mathbf{A}\mathbf{H}, \mathbf{H}^{-1}\mathbf{x}_t)$ for any invertible $\mathbf{H}$.

Therefore, we assume $\mathbf{A}'\mathbf{A} = \mathbf{I}_r$

But factor loading space $\mathcal{M}(\mathbf{A})$ is uniquely defined

The model is not new: Peña & Box (1987).

What is new?

- No distributional assumption on ε_t. More significantly, allow correlation between ε_t and $\mathbf{x}_{t+k}$ ($k \geq 1$).

What is new?

- No distributional assumption on ε_t. More significantly, allow correlation between ε_t and $\mathbf{x}_{t+k}$ ($k \geq 1$).
- Factor $\mathbf{x}_t$, and therefore also $\mathbf{y}_t$, may be nonstationary, not necessarily driven by unit roots.

What is new?

- No distributional assumption on $\boldsymbol{\varepsilon}_t$. More significantly, allow correlation between $\boldsymbol{\varepsilon}_t$ and $\mathbf{x}_{t+k}$ ($k \geq 1$).
- Factor $\mathbf{x}_t$, and therefore also $\mathbf{y}_t$, may be nonstationary, not necessarily driven by unit roots.
- A new estimation method for stationary cases: an eigenanalysis, applicable when $p >> n$

What is new?

- No distributional assumption on ε_t. More significantly, allow correlation between ε_t and $\mathbf{x}_{t+k}$ ($k \geq 1$).
- Factor $\mathbf{x}_t$, and therefore also $\mathbf{y}_t$, may be nonstationary, not necessarily driven by unit roots.
- A new estimation method for stationary cases: an eigenanalysis, applicable when $p >> n$

Key: estimate $\mathbf{A}$ (and r), or more precisely, $\mathcal{M}(\mathbf{A})$.

With available an estimator $\widehat{\mathbf{A}}$, a natural estimator for factor and the associated residuals are

$$\widehat{\mathbf{x}}_t = \widehat{\mathbf{A}}'\mathbf{y}_t, \qquad \widehat{\varepsilon}_t = (\mathbf{I}_p - \widehat{\mathbf{A}}\widehat{\mathbf{A}}')\mathbf{y}_t.$$

By modelling the lower-dimensional $\widehat{\mathbf{x}}_t$, we obtain the dynamical model for $\mathbf{y}_t$:

$$\widehat{\mathbf{y}}_t = \widehat{\mathbf{A}}\widehat{\mathbf{x}}_t.$$

Nonstationary factors

C1. $\boldsymbol{\varepsilon}_t \sim \mathrm{WN}(\boldsymbol{\mu}_{\boldsymbol{\varepsilon}}, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}})$, $\mathbf{c}'\mathbf{x}_t$ is not white noise for any constant $\mathbf{c} \in \mathcal{R}^p$. Furthermore $\mathbf{A}'\mathbf{A} = \mathbf{I}_r$.

Let $\mathbf{B} = (\mathbf{b}_1, \cdots, \mathbf{b}_{p-r})$ be a $p \times (p-r)$ matrix such that

$(\mathbf{A}, \mathbf{B})$ is a $p \times p$ orthogonal matrix, i.e.

$$\mathbf{B}'\mathbf{A} = 0, \qquad \mathbf{B}'\mathbf{B} = \mathbf{I}_{p-r}.$$

Since $\mathbf{y}_t = \mathbf{A}\mathbf{x}_t + \boldsymbol{\varepsilon}_t$,

$$\mathbf{B}'\mathbf{y}_t = \mathbf{B}'\boldsymbol{\varepsilon}_t$$

i.e. $\{\mathbf{B}'\mathbf{y}_t,\ t = 0, \pm 1, \cdots\}$ is WN.

Therefore

$$\mathrm{Corr}(\mathbf{b}_i'\mathbf{y}_t, \mathbf{b}_j'\mathbf{y}_{t-k}) = 0 \quad \forall\ 1 \le i, j \le p-r \ \text{ and } \ k \ge 1.$$

Search for mutually orthogonal directions $\mathbf{b}_1, \mathbf{b}_2, \cdots$ one by one such that the projection of $\mathbf{y}_t$ on each of those directions is a white noise.

Stop the search when such a direction is no longer available, and take $p - k$ as the estimated value of r, where k is the number of directions obtained in the search.

Search for mutually orthogonal directions $\mathbf{b}_1, \mathbf{b}_2, \cdots$ one by one such that the projection of $\mathbf{y}_t$ on each of those directions is a white noise.

Stop the search when such a direction is no longer available, and take $p - k$ as the estimated value of r, where k is the number of directions obtained in the search.

Latent factor process $\mathbf{x}_t$ stays latent in searching for $\mathcal{M}(\mathbf{A})$ or, equivalently, $\mathcal{M}(\mathbf{B})$.

See Pan and Yao (2008) for further details.

Stationary models

C2. $\mathbf{x}_t$ is stationary, and Cov$(\mathbf{x}_t, \boldsymbol{\varepsilon}_{t+k}) = 0$ for any $k \geq 0$.

Put $\boldsymbol{\Sigma}_y(k) = \mathrm{Cov}(\mathbf{y}_{t+k}, \mathbf{y}_t)$, $\boldsymbol{\Sigma}_x(k) = \mathrm{Cov}(\mathbf{x}_{t+k}, \mathbf{x}_t)$, $\boldsymbol{\Sigma}_{x\varepsilon}(k) = \mathrm{Cov}(\mathbf{x}_{t+k}, \boldsymbol{\varepsilon}_t)$. By $\mathbf{y}_t = \mathbf{A}\mathbf{x}_t + \boldsymbol{\varepsilon}_t$,

$$\boldsymbol{\Sigma}_y(k) = \mathbf{A}\boldsymbol{\Sigma}_x(k)\mathbf{A}' + \mathbf{A}\boldsymbol{\Sigma}_{x\varepsilon}(k), \qquad k \geq 1.$$

For a prescribed integer $k_0 \geq 1$, define

$$\mathbf{M} = \sum_{k=1}^{k_0} \boldsymbol{\Sigma}_y(k)\boldsymbol{\Sigma}_y(k)'.$$

Then $\mathbf{MB} = 0$, i.e. the columns of $\mathbf{B}$ are the eigenvectors of $\mathbf{M}$ corresponding to zero-eigenvalues.

Hence under C1 and C2,

Hence under C1 and C2,

The factor loading space $\mathcal{M}(\mathbf{A})$ are spanned by the eigenvectors of $\mathbf{M}$ corresponding to its non-zero eigenvalues, and the number of the non-zero eigenvalues is r.

Hence under C1 and C2,

> The factor loading space $\mathcal{M}(\mathbf{A})$ are spanned by the eigenvectors of $\mathbf{M}$ corresponding to its non-zero eigenvalues, and the number of the non-zero eigenvalues is r.

Let $\widehat{\mathbf{M}} = \sum_{k=1}^{k_0} \widehat{\mathbf{\Sigma}}_y(k)\widehat{\mathbf{\Sigma}}_y(k)'$, where $\widehat{\mathbf{\Sigma}}_y(k)$ denotes the sample covariance matrix of $\mathbf{y}_t$ at lag k.

Hence under C1 and C2,

The factor loading space $\mathcal{M}(\mathbf{A})$ are spanned by the eigenvectors of $\mathbf{M}$ corresponding to its non-zero eigenvalues, and the number of the non-zero eigenvalues is r.

Let $\widehat{\mathbf{M}} = \sum_{k=1}^{k_0} \widehat{\mathbf{\Sigma}}_y(k)\widehat{\mathbf{\Sigma}}_y(k)'$, where $\widehat{\mathbf{\Sigma}}_y(k)$ denotes the sample covariance matrix of $\mathbf{y}_t$ at lag k.

$\widehat{r}$: No. of non-zero eigenvalues of $\widehat{\mathbf{M}}$,

$\widehat{\mathbf{A}}$: its columns are the $\widehat{r}$ orthonormal eigenvectors of $\widehat{\mathbf{M}}$ corresponding to its $\widehat{r}$ largest eigenvalues.

Estimation for r

Let $\widehat{\lambda}_1 \geq \cdots \geq \widehat{\lambda}_p$ be the eigenvalues of $\widehat{\mathrm{M}}$.

Define

$$\widehat{r} = \arg \min_{1 \leq i \leq R} \widehat{\lambda}_{i+1} / \widehat{\lambda}_i,$$

where $1 \leq R < p$ is an integer.

In practice, we may let R be a (large) integer, e.g. $R = p/2$ or $p/3$.

This is an enhanced eyeball test, see Wang (2010).

Remark. The estimator $\widehat{r}$ benefits from the *fast convergence rates* of the estimators for the zero-eigenvalues.

Asymptotics I: $n \to \infty$ and p fixed

(i) $\mathbf{y}_t$ is strictly stationary, $E(||\mathbf{y}_t||^4) < \infty$.

(ii) $\mathbf{y}_t$ is ψ-mixing satisfying $\sum_j j\psi(j)^{\frac{1}{2}} < \infty$.

(iii) $\mathbf{M}$ has r non-zero eigenvalues $\lambda_1 > \cdots > \lambda_r > 0$.

Theorem 1. Under conditions C1 – C2 and (i) – (iii),

(i) $\widehat{\lambda}_j - \lambda_j = O_P(n^{-1/2})$ for $1 \le j \le r$,

(ii) $\widehat{\lambda}_{r+k} = O_P(n^{-1})$ for $1 \le k \le p - r$,

(iii) $D\{\mathcal{M}(\widehat{\mathbf{A}}),\ \mathcal{M}(\mathbf{A})\} = O_P(n^{-1/2})$ provided $\widehat{r} = r$ a.s.,

where

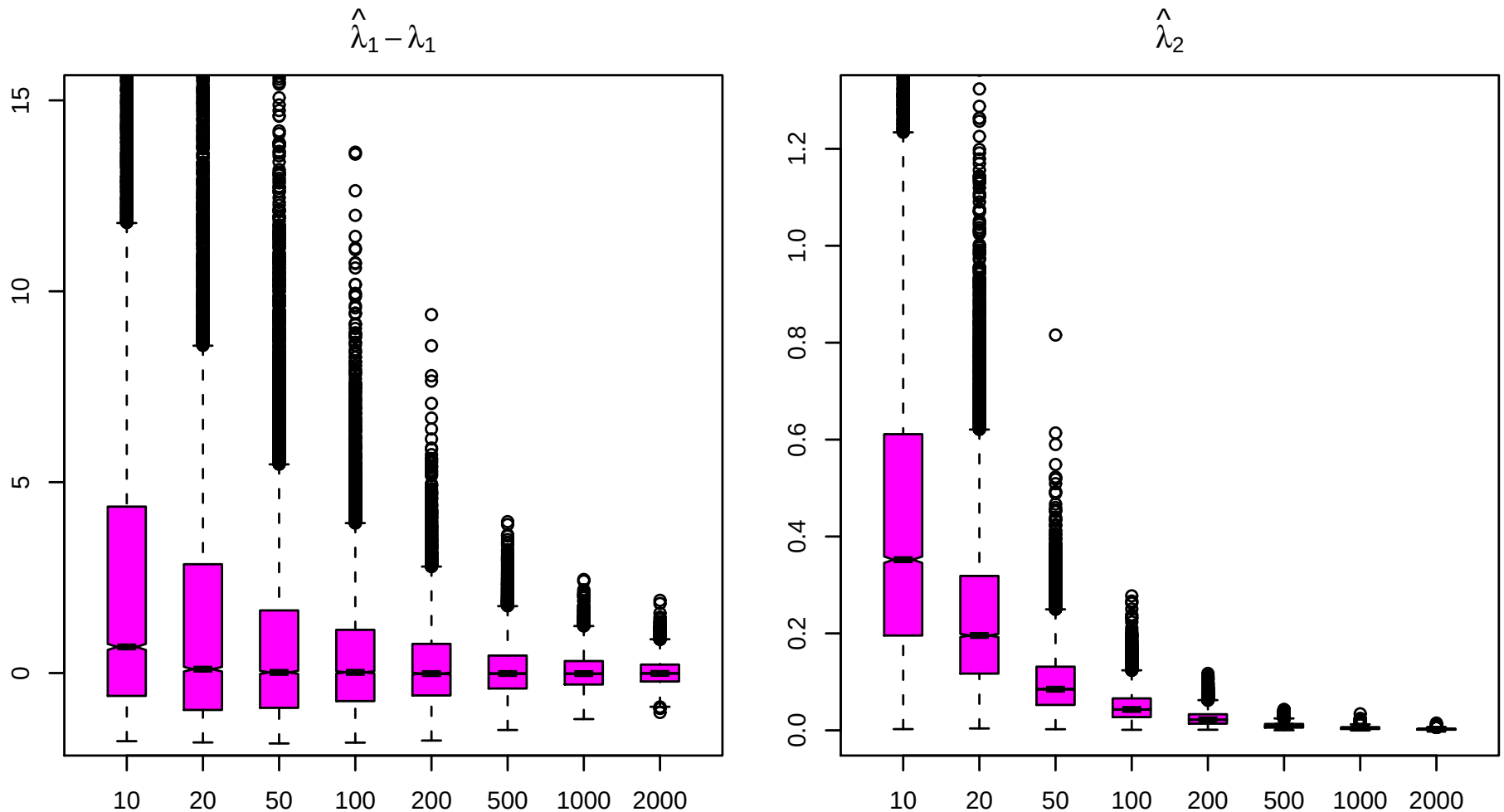
$$D\{\mathcal{M}(\widehat{\mathbf{A}}),\ \mathcal{M}(\mathbf{A})\} = 1 - \frac{1}{r}\,\mathrm{tr}(\mathbf{A}\mathbf{A}'\widehat{\mathbf{A}}\widehat{\mathbf{A}}').$$

(For the proof, see Bathia, Yao & Ziegelmann (2010).)

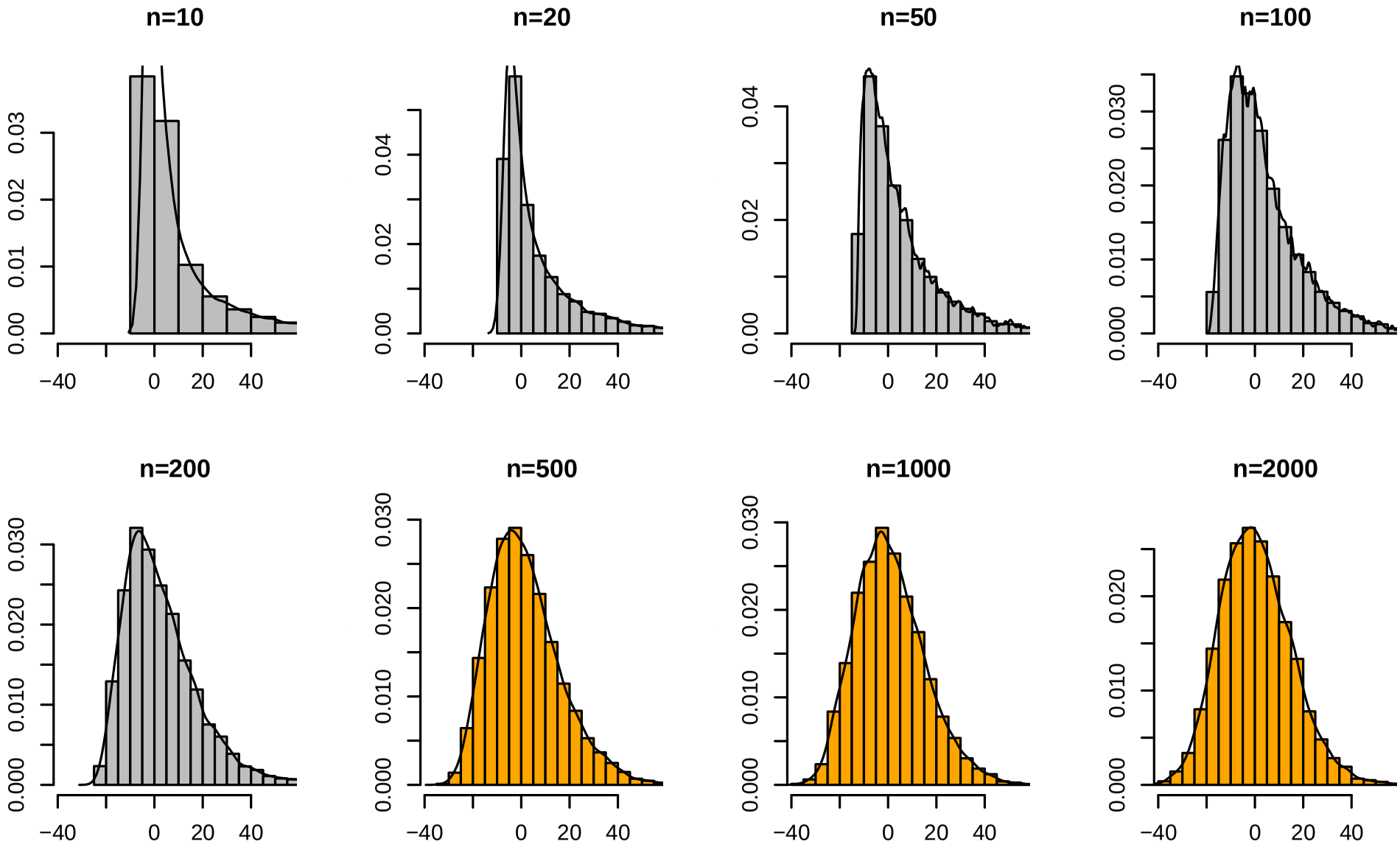
Numerical illustration: $\lambda_1 = 1.884,\ \lambda_2 = \lambda_3 = \lambda_4 = 0$ $(p = 4, r = 1)$

(Simulation replications: 10,000 times)

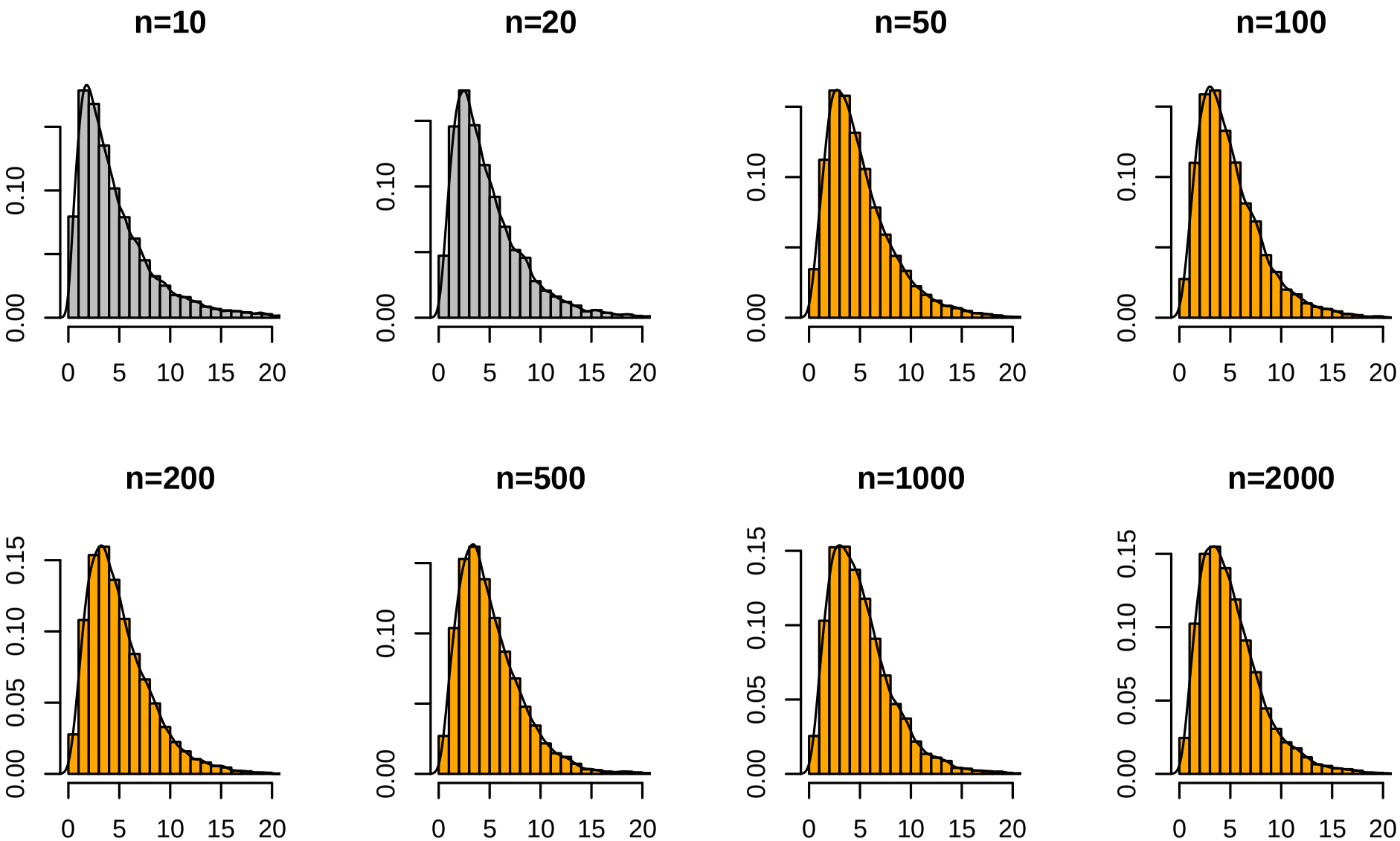
Boxplots of estimation errors against n



Histogram of $\sqrt{n}(\widehat{\lambda}_1 - \lambda_1)$



Histogram of $n\widehat{\lambda}_2$



Why $\widehat{\lambda}_{r+k} = O_P(n^{-1})$**?**

$\mathbf{M}$ is defined as a quadratic functions of the $\mathbf{\Sigma}_y(k)$:

$$\mathbf{M} = \sum_{k=1}^{k_0} \mathbf{\Sigma}_y(k)\mathbf{\Sigma}_y(k)'$$

In this sense, we are estimating λ^2 instead of λ itself.

Note

$$\widehat{\lambda}^2 - \lambda^2 = (\widehat{\lambda} + \lambda)(\widehat{\lambda} - \lambda) = \begin{cases} O_P(|\widehat{\lambda} - \lambda|) & \text{if } \lambda \neq 0 \\ \widehat{\lambda}^2 & \text{if } \lambda = 0 \end{cases}$$

$$= \begin{cases} O_P(n^{-1/2}) & \text{if } \lambda \neq 0 \\ O_P(n^{-1}) & \text{if } \lambda = 0. \end{cases}$$

When p is as large as n?

Repeat the numerical illustration: $p = n/2, \ r = 1$, i.e.

$$\lambda_1 = 1.884, \ \lambda_k = 0 \text{ for } k \geq 2.$$

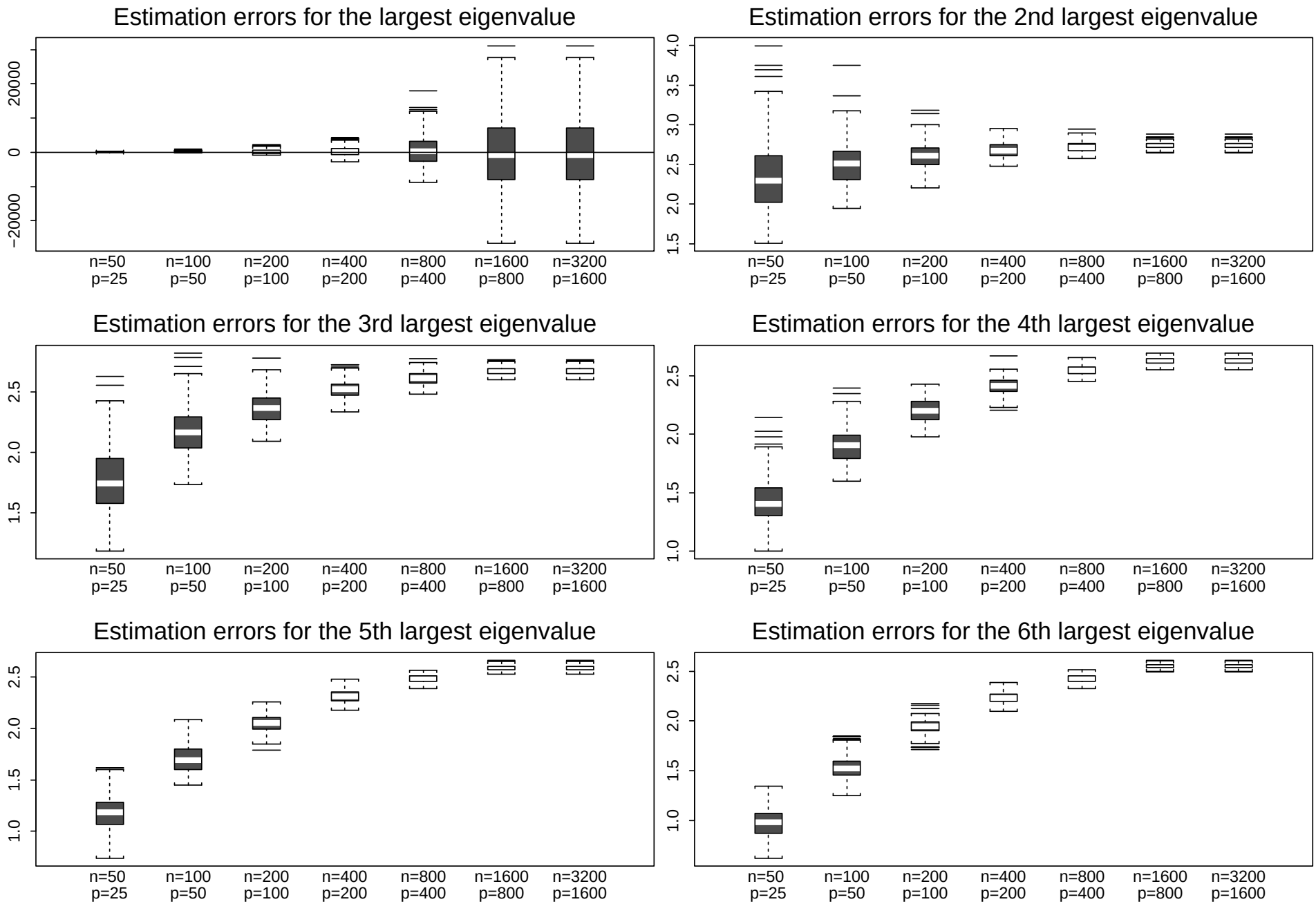
(Simulation replications: 200 times)

Take-away messages:

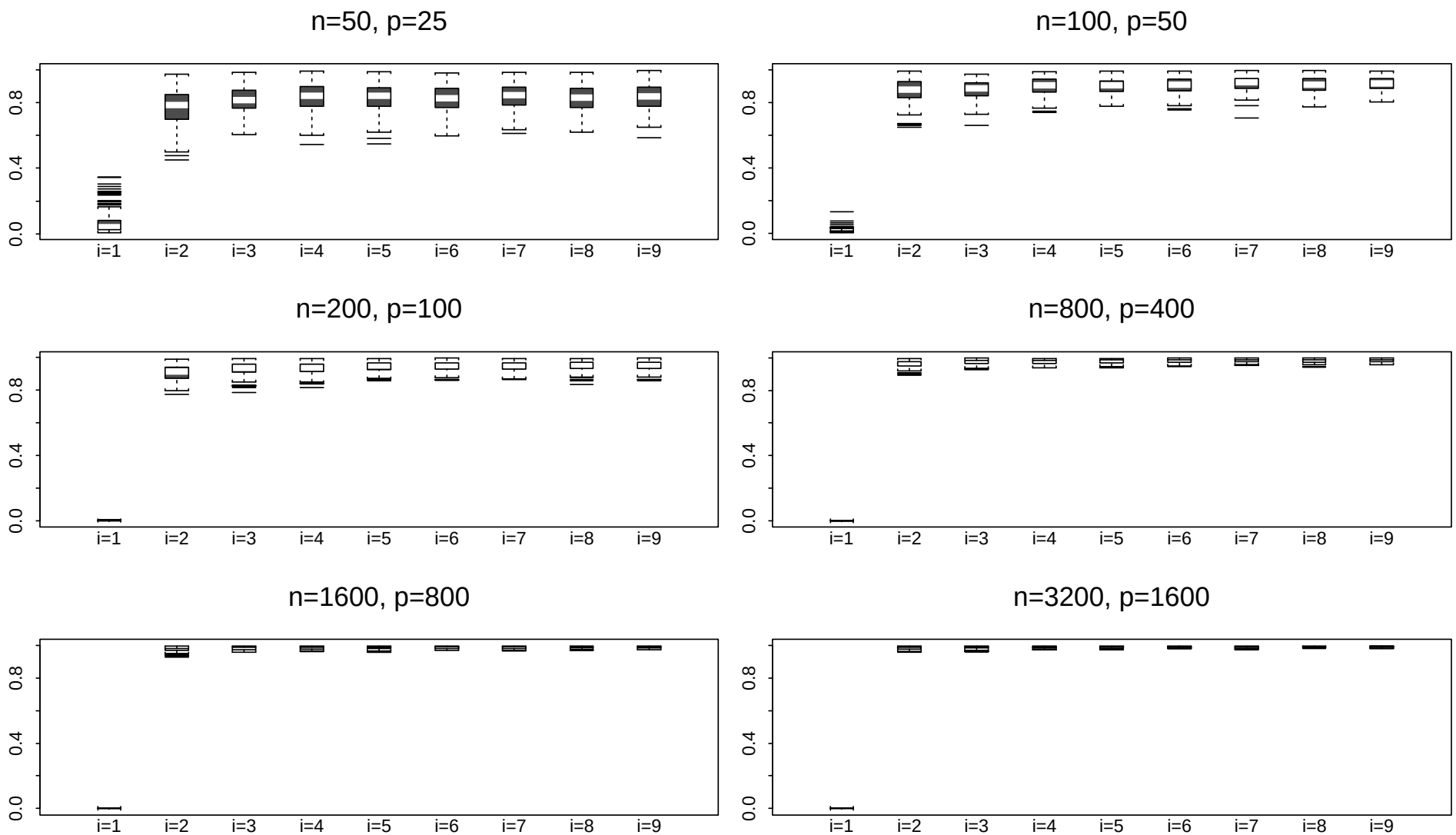
(i) $\widehat{\lambda}_j$ are no longer consistent.

(ii) The ratio-based estimator $\widehat{r}$ still works well.

(iii) $\widehat{\lambda}_{j+1}/\widehat{\lambda}_j \to 1$ for $j \geq 2$.


Estimation errors for the largest eigenvalue
20000
0
−20000
Estimation errors for the 2nd largest eigenvalue
4.0
3.5
3.0
2.5
2.0
1.5
Estimation errors for the 3rd largest eigenvalue
2.5
2.0
1.5
Estimation errors for the 4th largest eigenvalue
2.5
2.0
1.5
1.0
Estimation errors for the 5th largest eigenvalue
2.5
2.0
1.5
1.0
Estimation errors for the 6th largest eigenvalue
2.5
2.0
1.5
1.0
n=50 p=25
n=100 p=50
n=200 p=100
n=400 p=200
n=800 p=400
n=1600 p=800
n=3200 p=1600

Boxplots of ratios $\widehat{\lambda}_{i+1}/\widehat{\lambda}_i$



Asymptotics II: $n \to \infty$, $p \to \infty$ and r fixed

Recall model: $\mathbf{y}_t = \mathbf{A}\mathbf{x}_t + \boldsymbol{\varepsilon}_t$, and $\mathbf{A}$ is $p \times r$

(a) Assumptions on Strength of factors: (i) $\mathbf{A} = (\mathbf{a}_1, \cdots, \mathbf{a}_r)$, $||\mathbf{a}_i||^2 \asymp p^{1-\delta}, \quad i = 1, \cdots, r, \quad 0 \le \delta \le 1.$

(ii) For $k = 0, 1, \cdots, k_0$, $\boldsymbol{\Sigma}_x(k) \equiv \mathrm{Cov}(\mathbf{x}_{t+k}, \mathbf{x}_t)$ is full-ranked, and $\boldsymbol{\Sigma}_{x,\epsilon}(k) \equiv \mathrm{Cov}(\mathbf{x}_{t+k}, \boldsymbol{\varepsilon}_t) = O(1)$ elementwisely.

We call

- factors are strong if $\delta = 0$,
- factors are weak if $\delta > 0$.

Standardization '$\mathbf{A}'\mathbf{A} = \mathbf{I}_r$' + (i, ii) imply:

$$\|\boldsymbol{\Sigma}_x(k)\| \asymp p^{1-\delta} \asymp \|\boldsymbol{\Sigma}_x(k)\|_{\min}, \quad \|\boldsymbol{\Sigma}_{x,\epsilon}(k)\| = O(p^{1-\delta/2}),$$

where $a \asymp b$ represents $a = O(b)$ & $b = O(a)$, $\|\mathbf{A}\|^2 = \lambda_{\max}(\mathbf{A}\mathbf{A}')$ and $\|\mathbf{A}\|^2_{\min} = \max\{\lambda_{\min}(\mathbf{A}\mathbf{A}'),\ \lambda_{\min}(\mathbf{A}'\mathbf{A})\}$.

(b) For $k = 0, 1, \cdots, k_0$, $\|\mathbf{\Sigma}_{x,\epsilon}(k)\| = o(p^{1-\delta})$.

(c) The conditions used when p is fixed:

(i) $\mathbf{y}_t$ is strictly stationary, $E(||\mathbf{y}_t||^4) < \infty$.

(ii) $\mathbf{y}_t$ is ψ-mixing satisfying $\sum_j j\psi(j)^{\frac{1}{2}} < \infty$.

(iii) $\mathbf{M}$ has r non-zero eigenvalues $\lambda_1 > \cdots > \lambda_r > 0$.

Theorem 2. Under conditions C1 – C2 and (a) – (c),

$$\|\widehat{\mathbf{A}} - \mathbf{A}\| = O_P(p^{\delta} n^{-1/2})$$

provided $p^{\delta} n^{-1/2} \to 0$.

Remark. (i) When all factors are strong (i.e. $\delta = 0$), the convergence rate is $n^{-1/2}$, independent of the dimension p.

(ii) The similar result holds for <u>the factor based estimator</u> of $\{\text{Var}(\mathbf{y}_t)\}^{-1}$, but not for the estimator of $\text{Var}(\mathbf{y}_t)$.

Numerical illustration

A simple model with one strong factor: $\mathbf{y}_t = \mathbf{A}x_t + \boldsymbol{\varepsilon}_t$, where

$\mathbf{A}$ is $p \times 1$, with i-th element $2\cos(2\pi i/p)$,

$x_t = 0.9x_{t-1} + \eta_t$,

$\varepsilon_{tj},\ \eta_t$ are independent $N(0, 2^2)$.

Simulation replications: 50

Means and Standard Errors of $\|\widehat{\mathbf{A}} - \mathbf{A}\|$

$\|\widehat{\mathbf{A}} - \mathbf{A}\|$	$n = 200$	$n = 500$
$p = 20$	.022 (.005)	.014 (.003)
$p = 180$	.023 (.004)	.014 (.002)
$p = 400$	.022 (.004)	.014 (.002)
$p = 1000$	.023 (.004)	.014 (.002)

Theorem 3. Under conditions C1 - C2 & (a) - (c), and $\frac{p^\delta}{\sqrt{n}} \to 0$,

(i) $|\widehat{\lambda}_i - \lambda_i| = O_P(p^{2-\delta}/n^{1/2})$ for $i = 1, \cdots, r$,

(ii) $\widehat{\lambda}_j = O_P(p^2/n)$ for $j = r+1, \cdots, p$,

(iii) $\frac{\widehat{\lambda}_{j+1}}{\widehat{\lambda}_j} \asymp 1$ for $j = 1, \cdots, r-1$, and $\frac{\widehat{\lambda}_{r+1}}{\widehat{\lambda}_r} = O_P(\frac{p^{2\delta}}{n}) \xrightarrow{P} 0$.

Conjecture: $\frac{\widehat{\lambda}_{j+1}}{\widehat{\lambda}_j} \xrightarrow{P} 1$ for $j = (k_0+1)r+1, \cdots, (k_0+1)r+K$, where $K \geq 1$ is a constant.

A refined result: Under additional conditions such as ε_{tj} being *sub-Gaussian* and $\delta = 0$ (i.e. strong factors), it holds that

$$\widehat{\lambda}_{r+1}/\widehat{\lambda}_r = O_P(p^{-1/2}n^{-1}).$$

Numerical illustration

Simulation model with <u>3 factors</u>: $\mathbf{y}_t = \mathbf{A}\mathbf{x}_t + \boldsymbol{\varepsilon}_t$, where

$\mathbf{A} = (a_{ij})$ is $p \times 3$, $a_{ij} p^{\frac{\delta}{2}} \sim U(-1,\, 1)$, i.e. $||a_j||^2 = O(p^{1-\delta})$,

$x_{tj} = \alpha_j x_{t-1,j} + e_{tj}$, with $\alpha_1 = .6,\ \alpha_2 = -.5,\ \alpha_3 = .3$,

$\varepsilon_{tj},\ e_{tj}$ are independent $N(0, 1)$.

Simulation replications: 200 times.

Report the relative frequency estimates for $P(\widehat{r} = 3)$, where

$$\widehat{r} = \arg \min_{1 \le j \le 10} \widehat{\lambda}_{j+1} / \widehat{\lambda}_j,$$

and $\widehat{\lambda}_1 \ge \widehat{\lambda}_2 \ge \cdots \ge \widehat{\lambda}_p \ge 0$ are the eigenvalues of $\widehat{\mathbf{M}}$.

Estimates for $P(\widehat{r} = 3)$

n	50	100	200	400	800	1600	3200
$p\ (= .2n)$	10	20	40	80	260	320	640
$\delta = 0$	.165	.680	.940	.995	1	1	1
$\delta = .5$	.075	.155	.270	.570	.980	1	1
$p\ (= .5n)$	25	50	100	200	400	800	1600
$\delta = 0$	.410	.800	.980	1	1	1	1
$\delta = .5$	.090	.285	.285	.820	.960	1	1
$p\ (= .8n)$	40	80	160	320	640	1280	2560
$\delta = 0$	.560	.815	.990	1	1	1	1
$\delta = .5$	.060	.180	.490	.745	.970	1	1

(i) With p/n fixed, estimation improves as $n \nearrow$

(ii) Blessing of dimensionality: With n fixed, estimation improves as $p \nearrow$, when factors are strong.

Summary

- Latent factor process $\mathbf{x}_t$ is not featured in estimating the factor loading space

- Faster convergence rates for zero-eigenvalues are helpful in determining the number of factors

- Blessing of dimensionality when factors are strong.

Extension

Consider

$$\mathbf{y}_t = \mathbf{D}\mathbf{z}_t + \mathbf{A}\mathbf{x}_t + \boldsymbol{\varepsilon}_t,$$

$\mathbf{z}_t$ is a observable $m \times 1$ regressor, $\mathbf{D}$ is unknown regression coefficient matrix

Let $\boldsymbol{\varepsilon}_t$ be uncorrelated with $(\mathbf{z}_t, \mathbf{x}_t)$.

If we know $\mathbf{D}$, $\mathbf{y}_t - \mathbf{D}\mathbf{z}_t = \mathbf{A}\mathbf{x}_t + \boldsymbol{\varepsilon}_t$. The method applies.

Let

$$\widehat{\mathbf{D}} = (\widehat{\mathbf{d}}_1, \ldots, \widehat{\mathbf{d}}_p)', \qquad \widehat{\mathbf{d}}_i = \Big(\sum_{t=1}^{n} \mathbf{z}_t \mathbf{z}_t'\Big)^{-1} \sum_{t=1}^{n} y_{i,t}\, \mathbf{z}_t$$

Proceed with

$$\widehat{\boldsymbol{\eta}}_t \equiv \mathbf{y}_t - \widehat{\mathbf{D}}\mathbf{z}_t \approx \mathbf{A}\mathbf{x}_t + \boldsymbol{\varepsilon}_t$$

(Nonstationary) conditions

C0. $\mathrm{Cov}(\mathbf{z}_t, \mathbf{x}_t) = 0$.

C1. $(\mathbf{y}_t, \mathbf{z}_t, \mathbf{x}_t)$ is α-mixing with $\sum_k \alpha(k)^{1-2/\gamma} < \infty$ for some $\gamma > 0$.

C2. All components of $\mathbf{z}_t$, $\mathbf{A}\mathbf{x}_t$, $\boldsymbol{\varepsilon}_t$ have finite 2γ-th moments.

C3. $\lambda_{\min}\{E(\mathbf{z}_t\mathbf{z}_t')\} > C > 0$.

Under C0 – C3, $\|\widehat{\mathbf{D}} - \mathbf{D}\|_F = O_p\big(p^{1/2}T^{-1/2}\big)$ as both $n, p \to \infty$.

C4. $C_1 p^{1-\delta} \le \|\boldsymbol{\Sigma}_x(k)\|_{\min} \le \|\boldsymbol{\Sigma}_x(k)\|_2 \le C_2 p^{1-\delta}$ for all $k = 1, \ldots, k_0$, where $C_2 > C_1 > 0$ are constants, and δ is the factor strength index, where $\boldsymbol{\Sigma}_x(k) = \frac{1}{n-k}\sum_{t=1}^{n-k} \mathrm{Cov}(\mathbf{x}_{t+k}, \mathbf{x}_t)$.

C5. The r non-zero eigenvalues of $\mathbf{M}$ are distinct.

Under C0 – C5, the convergence rates for $\widehat{\mathbf{A}}$ can be derived as before.

Endogeneity: $\mathrm{Cov}(\mathbf{z}_t, \mathbf{x}_t) \neq 0$

$\mathbf{D}$ is no longer identifiable, or it is replaced by $\mathbf{D}^\star$:

$$\begin{aligned}
\mathbf{y}_t &= [\mathbf{D} + \mathbf{A}E(\mathbf{x}_t\mathbf{z}_t')\{E(\mathbf{z}_t\mathbf{z}_t')\}^{-1}]\mathbf{z}_t \\
&+ \mathbf{A}[\mathbf{x}_t - E(\mathbf{x}_t\mathbf{z}_t')\{E(\mathbf{z}_t\mathbf{z}_t')\}^{-1}\mathbf{z}_t] + \boldsymbol{\varepsilon}_t \\
&\equiv \mathbf{D}^\star\mathbf{z}_t + \mathbf{A}\mathbf{x}_t^\star + \boldsymbol{\varepsilon}_t,
\end{aligned}$$

while the factor loading unchanged.

If the original '$\mathbf{D}$' is the target, employ an instrumental variable $\mathbf{w}_t$: Cov$(\mathbf{w}_t, \mathbf{z}_t) \neq 0$, and $\mathbf{w}_t$ is uncorrelated with $(\mathbf{x}_t, \boldsymbol{\varepsilon}_t)$

$$\mathbf{y}_t\mathbf{w}_t' = \mathbf{D}\mathbf{z}_t\mathbf{w}_t' + \boldsymbol{\varepsilon}_t^\star, \quad \boldsymbol{\varepsilon}_t^\star = \mathbf{A}\mathbf{x}_t\mathbf{w}_t' + \boldsymbol{\varepsilon}_t\mathbf{w}_t'$$

$$\widehat{\mathbf{D}} = \Big(\frac{1}{n}\sum_{t=1}^{n}\mathbf{y}_t\mathbf{w}_t'\mathbf{R}'\Big)\Big(\frac{1}{n}\sum_{t=1}^{n}\mathbf{z}_t\mathbf{w}_t'\mathbf{R}'\Big)^{-1}, \quad \mathrm{rank}(\mathbf{R}) = m.$$

Nonlinear regression function

Consider model

$$\mathbf{y}_t = \mathbf{g}(\mathbf{u}_t) + \mathbf{A}\mathbf{x}_t + \boldsymbol{\varepsilon}_t,$$

where $\mathbf{u}_t$ is an observed process with fixed dimension, $\mathbf{g} = (g_1, \ldots, g_p)'$ is an unknown nonlinear function.

Let

$$g_i(\mathbf{u}) = \sum_{j=1}^{\infty} d_{i,j} l_j(\mathbf{u}), \qquad i = 1, 2, \ldots,$$

where $\{l_j(\cdot)\}$ is a set of base functions. Suppose only use the first m terms:

$$\mathbf{z}_t = (l_1(\mathbf{u}_t), \ldots, l_m(\mathbf{u}_t))'$$

Then

$$\mathbf{y}_t = \mathbf{D}\mathbf{z}_t + \mathbf{A}\mathbf{x}_t + \boldsymbol{\varepsilon}_t + \mathbf{e}_t$$

Now $m \to \infty$ together with n and p.

Model driven by a hidden Markov chain

Let z_t be a hidden Markov chain such that given $z_t = k$,

$$\mathbf{y}_t = \boldsymbol{\mu}_k + \mathbf{A}_k \mathbf{x}_t + \boldsymbol{\varepsilon}_t^{(k)} \quad \text{and} \quad \boldsymbol{\varepsilon}_t^{(k)} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_k),$$

where $\mathbf{A}_k' \mathbf{A}_k = \mathbf{I}_r$, $E(\mathbf{y}_t | z_t = k) = \boldsymbol{\mu}_k$, $E(\mathbf{x}_t | z_t) = 0$ and

$$\pi_{jk} = P(z_{t+1} = k | z_t = j) = P(z_{t+1} = k | z_t = j, \mathbf{y}_t), \quad j, k = 1, \cdots, m.$$

Typically m is small.

Let $(\mathbf{A}_k, \mathbf{B}_k)$ be an orthogonal matrix.

Two-step iterative algorithm based on $\{\mathbf{B}_k \mathbf{y}_t\}$ (instead of $\{\mathbf{y}_t\}$):

1. Given $\pi_{ij}, \boldsymbol{\mu}_k, \mathbf{B}_k, \boldsymbol{\Sigma}_{k,B}$, estimate $\{z_t\}$ by maximizing

$$\sum_{t=1}^{n-1} \log \left(\sum_{i,j=1}^{m} I(z_t = i, z_{t+1} = j) \pi_{ij} f(\mathbf{B}_j' \mathbf{y}_{t+1} | z_{t+1} = j) \right)$$

   - Use the Viterbi algorithm.

2. Given $\{z_t\}$,

$$\widetilde{\pi}_{ij} = \frac{\sum_{t=1}^{n-1} I(z_t = i,\ z_{t+1} = j)}{\sum_{t=1}^{n-1} I(z_t = i)}, \quad \widetilde{\boldsymbol{\mu}}_k = \frac{\sum_{t=1}^{n} \mathbf{y}_t I(z_t = k)}{\sum_{t=1}^{n} I(z_t = k)}$$

$$\widetilde{\mathbf{U}}_{j,k} = \frac{\sum_{t=1}^{n-j} (\mathbf{y}_{t+j} - \widetilde{\boldsymbol{\mu}}_k)(\mathbf{y}_t - \widetilde{\boldsymbol{\mu}}_k)' I(z_{t+j} = k)}{\sum_{t=1}^{n-j} I(z_{t+j} = k)}$$

$$\widetilde{\mathbf{M}}_k = \sum_{j=1}^{k_0} \widetilde{\mathbf{U}}_{j,k} \widetilde{\mathbf{U}}'_{j,k}$$

Let $\lambda_{1,k} \geq \cdots \geq \lambda_{p,k} \geq 0$ be the p eigenvalues of $\widetilde{\mathbf{M}}_k$, and $\boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_p$ be the corresponding orthonormal eigenvectors,

$$\widetilde{\mathbf{A}}_k = (\boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_d) \quad \text{and} \quad \widetilde{\mathbf{B}}_k = (\boldsymbol{\gamma}_{d+1}, \cdots, \boldsymbol{\gamma}_d).$$

Prediction

Let $\mathbf{y}_t = (\mathbf{y}_1, \cdots, \mathbf{y}_t)$,

$$\pi_k(t|t) = P(z_t = k|\mathbf{y}_t) \quad \text{and} \quad \pi_k(t+1|t) = P(z_{t+1} = k|\mathbf{y}_t).$$

Then $\pi_k(t+1|t) = \sum_{j=1}^m \pi_j(t|t)\,\pi_{jk}$, and

$$\pi_k(t|t) = \frac{\pi_k(t|t-1) f(\mathbf{y}_t|z_t = k, \mathbf{y}_{t-1})}{\sum_{j=1}^m \pi_j(t|t-1) f(\mathbf{y}_t|z_t = j, \mathbf{y}_{t-1})} \approx \frac{\pi_k(t|t-1) f(\mathbf{B}_k'\mathbf{y}_t|z_t = k)}{\sum_{j=1}^m \pi_j(t|t-1) f(\mathbf{B}_j'\mathbf{y}_t|z_t = j)}.$$

One-step-ahead prediction:

$$\mathbf{y}_n(1) = \widehat{\boldsymbol{\mu}}_{\widehat{k}} + \widehat{\mathbf{A}}_{\widehat{k}}\,\mathbf{x}_n(1), \quad \text{where} \quad \widehat{k} = \arg\max_{1 \le k \le m} \widehat{\pi}_k(n+1|n), \quad \text{or}$$

$$\mathbf{y}_n(1) = \sum_{k=1}^m \widehat{\pi}_k(n+1|n)\{\widehat{\boldsymbol{\mu}}_k + \widehat{\mathbf{A}}_k\,\mathbf{x}_n(1)\}.$$