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


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 \( \{y_t\} \) be a \( p \times 1 \) time series defined by

\[
y_t = A x_t + \varepsilon_t,
\]
Let \( \{y_t\} \) be a \( p \times 1 \) time series defined by

\[
y_t = Ax_t + \varepsilon_t,
\]

\( x_t: r \times 1 \) unobservable factors, \( r (< p) \) unknown

\( A: p \times r \) unknown constant factor loading matrix

\( \{\varepsilon_t\}: \) vector \( WN(\mu_\varepsilon, \Sigma_\varepsilon) \)

no linear combinations of \( x_t \) are WN.
Let \( \{y_t\} \) be a \( p \times 1 \) time series defined by

\[
y_t = Ax_t + \varepsilon_t,
\]

\( x_t: r \times 1 \) unobservable factors, \( r (\leq p) \) unknown

\( A: p \times r \) unknown constant factor loading matrix

\( \{\varepsilon_t\}: \) vector \( \text{WN}(\mu_\varepsilon, \Sigma_\varepsilon) \)

no linear combinations of \( x_t \) are WN.

Lack of identification: \( (A, x_t) \) may be replaced by

\( (AH, H^{-1}x_t) \) for any invertible \( H \).

Therefore, we assume \( A' A = I_r \)

But factor loading space \( \mathcal{M}(A) \) is uniquely defined
Let $\{y_t\}$ be a $p \times 1$ time series defined by

$$y_t = Ax_t + \varepsilon_t,$$

$x_t$: $r \times 1$ unobservable factors, $r (< p)$ unknown
A: $p \times r$ unknown constant factor loading matrix
$\{\varepsilon_t\}$: vector $\text{WN}(\mu_\varepsilon, \Sigma_\varepsilon)$
no linear combinations of $x_t$ are WN.

Lack of identification: $(A, x_t)$ may be replaced by $(AH, H^{-1}x_t)$ for any invertible $H$.

Therefore, we assume $A'A = I_r$

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

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

- No distributional assumption on $\varepsilon_t$. More significantly, allow correlation between $\varepsilon_t$ and $x_{t+k}$ ($k \geq 1$).
What is new?

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

- Factor $x_t$, and therefore also $y_t$, may be nonstationary, not necessarily driven by unit roots.
What is new?

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

- Factor $x_t$, and therefore also $y_t$, 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 $\epsilon_t$. More significantly, allow correlation between $\epsilon_t$ and $x_{t+k}$ ($k \geq 1$).
- Factor $x_t$, and therefore also $y_t$, may be nonstationary, not necessarily driven by unit roots.
- A new estimation method for stationary cases: an eigenanalysis, applicable when $p >> n$

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

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

$$\hat{x}_t = \hat{A}'y_t, \quad \hat{\epsilon}_t = (I_p - \hat{A}\hat{A}')y_t.$$

By modelling the lower-dimensional $\hat{x}_t$, we obtain the dynamical model for $y_t$:

$$\hat{y}_t = \hat{A}\hat{x}_t.$$
Nonstationary factors

C1. \( \varepsilon_t \sim WN(\mu_\varepsilon, \Sigma_\varepsilon) \), \( c'x_t \) is not white noise for any constant \( c \in \mathbb{R}^p \). Furthermore \( A'A = I_r \).

Let \( B = (b_1, \cdots, b_{p-r}) \) be a \( p \times (p - r) \) matrix such that

\[
(A, B) \text{ is a } p \times p \text{ orthogonal matrix, i.e.}
\]

\[
B'A = 0, \quad B'B = I_{p-r}.
\]

Since \( y_t = Ax_t + \varepsilon_t \),

\[
B'y_t = B'\varepsilon_t
\]

i.e. \( \{B'y_t, \ t = 0, \pm 1, \cdots\} \) is WN.

Therefore

\[
\text{Corr}(b_i'y_t, b_j'y_{t-k}) = 0 \quad \forall \ 1 \leq i, j \leq p - r \text{ and } k \geq 1.
\]
Search for mutually orthogonal directions $b_1, b_2, \cdots$ one by one such that the projection of $y_t$ 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 $b_1, b_2, \cdots$ one by one such that the projection of $y_t$ 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 $x_t$ stays latent in searching for $\mathcal{M}(A)$ or, equivalently, $\mathcal{M}(B)$.

See Pan and Yao (2008) for further details.
Stationary models

C2. $x_t$ is stationary, and $\text{Cov}(x_t, \varepsilon_{t+k}) = 0$ for any $k \geq 0$.

Put $\Sigma_y(k) = \text{Cov}(y_{t+k}, y_t)$, $\Sigma_x(k) = \text{Cov}(x_{t+k}, x_t)$, $\Sigma_{x\varepsilon}(k) = \text{Cov}(x_{t+k}, \varepsilon_t)$. By $y_t = Ax_t + \varepsilon_t$,

$$\Sigma_y(k) = A \Sigma_x(k) A' + A \Sigma_{x\varepsilon}(k), \quad k \geq 1.$$  

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

$$M = \sum_{k=1}^{k_0} \Sigma_y(k) \Sigma_y(k)'.$$  

Then $MB = 0$, i.e. the columns of $B$ are the eigenvectors of $M$ corresponding to zero-eigenvalues.
Hence under C1 and C2,
Hence under C1 and C2,

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

Let $\hat{M} = \sum_{k=1}^{k_0} \hat{\Sigma}_y(k)\hat{\Sigma}_y(k)'$, where $\hat{\Sigma}_y(k)$ denotes the sample covariance matrix of $y_t$ at lag $k$. 
Hence under C1 and C2,

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

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

$\hat{r}$: No. of non-zero eigenvalues of $\hat{\mathcal{M}}$,

$\hat{A}$: its columns are the $\hat{r}$ orthonormal eigenvectors of $\hat{\mathcal{M}}$ corresponding to its $\hat{r}$ largest eigenvalues.
Estimation for $r$

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

Define

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

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 $\hat{r}$ benefits from the fast convergence rates of the estimators for the zero-eigenvalues.
Asymptotics I: $n \to \infty$ and $p$ fixed

(i) $y_t$ is strictly stationary, $E(||y_t||^4) < \infty$.

(ii) $y_t$ is $\psi$-mixing satisfying $\sum j \cdot \psi(j)^{1/2} < \infty$.

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

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

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

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

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

where

$$D\{M(\hat{A}), M(A)\} = 1 - \frac{1}{r} \text{tr}(AA'\hat{A}\hat{A}')$$

(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 \( \sqrt{n} (\hat{\lambda} - \lambda_1) \), for different values of \( n \):

- \( n = 10 \)
- \( n = 20 \)
- \( n = 50 \)
- \( n = 100 \)
- \( n = 200 \)
- \( n = 500 \)
- \( n = 1000 \)
- \( n = 2000 \)

(1) Histogram of \( n^\lambda \left( \lambda - \lambda_1 \right) \)
Histogram of $n\lambda^2$
Why $\hat{\lambda}_{r+k} = O_P(n^{-1})$?

$M$ is defined as a **quadratic** functions of the $\Sigma_y(k)$:

$$M = \sum_{k=1}^{k_0} \Sigma_y(k) \Sigma_y(k)'$$

In this sense, we are estimating $\lambda^2$ instead of $\lambda$ itself.

Note

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

$$= \begin{cases} O_P(n^{-1/2}) & \text{if } \lambda \neq 0 \\ O_P(n^{-1}) & \text{if } \lambda = 0. \end{cases}$$
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) $\hat{\lambda}_j$ are no longer consistent.

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

(iii) $\frac{\hat{\lambda}_{j+1}}{\hat{\lambda}_j} \to 1$ for $j \geq 2$. 
Boxplots of ratios $\hat{\lambda}_{i+1}/\hat{\lambda}_i$
Asymptotics II: \( n \to \infty, p \to \infty \) and \( r \) fixed

Recall model: \( y_t = A x_t + \varepsilon_t \), and \( A \) is \( p \times r \)

(a) Assumptions on Strength of factors: (i) \( A = (a_1, \ldots, a_r) \),

\[ \|a_i\|^2 \asymp p^{1-\delta}, \quad i = 1, \ldots, r, \quad 0 \leq \delta \leq 1. \]

(ii) For \( k = 0, 1, \ldots, k_0 \), \( \Sigma_x(k) \equiv \text{Cov}(x_{t+k}, x_t) \) is full-ranked, and \( \Sigma_{x,\varepsilon}(k) \equiv \text{Cov}(x_{t+k}, \varepsilon_t) = O(1) \) elementwisely.

We call

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

Standardization \( 'A' A = I_r' + (i, ii) \) imply:

\[ \|\Sigma_x(k)\| \asymp p^{1-\delta} \asymp \|\Sigma_x(k)\|_{\text{min}}, \quad \|\Sigma_{x,\varepsilon}(k)\| = O(p^{1-\delta/2}), \]

where \( a \asymp b \) represents \( a = O(b) \) & \( b = O(a) \), \( \|A\|^2 = \lambda_{\text{max}}(AA') \)

and \( \|A\|^2_{\text{min}} = \max\{\lambda_{\text{min}}(AA'), \lambda_{\text{min}}(A'A)\} \).
(b) For $k = 0, 1, \cdots, k_0$, $\|\Sigma_{x,\epsilon}(k)\| = o(p^{1-\delta})$.

(c) The conditions used when $p$ is fixed:
   (i) $y_t$ is strictly stationary, $E(||y_t||^4) < \infty$.
   (ii) $y_t$ is $\psi$-mixing satisfying $\sum_j j \psi(j)^{1/2} < \infty$.
   (iii) $M$ has $r$ non-zero eigenvalues $\lambda_1 > \cdots > \lambda_r > 0$.

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

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

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 the factor based estimator of $\{\text{Var}(y_t)\}^{-1}$, but not for the estimator of $\text{Var}(y_t)$.
**Numerical illustration**

A simple model with **one strong factor**: \( y_t = Ax_t + \varepsilon_t \), where

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

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

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

Simulation replications: 50

<table>
<thead>
<tr>
<th>( | \hat{A} - A | )</th>
<th>( n = 200 )</th>
<th>( n = 500 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p = 20 )</td>
<td>.022 (.005)</td>
<td>.014 (.003)</td>
</tr>
<tr>
<td>( p = 180 )</td>
<td>.023 (.004)</td>
<td>.014 (.002)</td>
</tr>
<tr>
<td>( p = 400 )</td>
<td>.022 (.004)</td>
<td>.014 (.002)</td>
</tr>
<tr>
<td>( p = 1000 )</td>
<td>.023 (.004)</td>
<td>.014 (.002)</td>
</tr>
</tbody>
</table>
Theorem 3. Under conditions C1 - C2 & (a) - (c), and $\frac{p^\delta}{\sqrt{n}} \to 0$,

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

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

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

Conjecture: $\frac{\hat{\lambda}_{j+1}}{\hat{\lambda}_j} \xrightarrow{P} 1$ 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 $\varepsilon_{tj}$ being sub-Gaussian and $\delta = 0$ (i.e. strong factors), it holds that

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

Simulation model with 3 factors: \( y_t = Ax_t + \varepsilon_t, \) where

\[ A = (a_{ij}) \text{ is } p \times 3, \quad a_{ij} \pi^2 \sim U(-1, 1), \quad \text{i.e. } ||a_j||^2 = O(p^{1-\delta}), \]

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

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

Simulation replications: 200 times.

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

\[ \hat{r} = \arg \min_{1 \leq j \leq 10} \frac{\hat{\lambda}_{j+1}}{\hat{\lambda}_j}, \]

and \( \hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p \geq 0 \) are the eigenvalues of \( \hat{M}. \)
Estimates for $P(\hat{r} = 3)$

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

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

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

- Latent factor process $x_t$ 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.
Consider

\[ y_t = Dz_t + Ax_t + \varepsilon_t, \]

\( z_t \) is a observable \( m \times 1 \) regressor, \( D \) is unknown regression coefficient matrix

Let \( \varepsilon_t \) be uncorrelated with \( (z_t, x_t) \).

If we know \( D \), \( y_t - Dz_t = Ax_t + \varepsilon_t \). The method applies.

Let

\[ \hat{D} = (\hat{d}_1, \ldots, \hat{d}_p)', \quad \hat{d}_i = \left( \sum_{t=1}^{n} z_t z_t' \right)^{-1} \sum_{t=1}^{n} y_{i,t} z_t \]

Proceed with

\[ \hat{\eta}_t \equiv y_t - \hat{D}z_t \approx Ax_t + \varepsilon_t \]
(Nonstationary) conditions

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

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

C2. All components of \( \mathbf{z}_t, \mathbf{A} \mathbf{x}_t, \varepsilon_t \) have finite \( 2\gamma \)-th moments.

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

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

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

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

Under C0 – C5, the convergence rates for \( \mathbf{\hat{A}} \) can be derived as before.
**Endogeneity:** $\text{Cov}(z_t, x_t) \neq 0$

$D$ is no longer identifiable, or it is replaced by $D^*$:

$$y_t = [D + AE(x_t z'_t)\{E(z_t z'_t)\}^{-1}]z_t$$

$$+ A[x_t - E(x_t z'_t)\{E(z_t z'_t)\}^{-1}z_t] + \varepsilon_t$$

$$\equiv D^*z_t + Ax^*_t + \varepsilon_t,$$

while the factor loading unchanged.

If the original ‘$D$’ is the target, employ an instrumental variable $w_t$: $\text{Cov}(w_t, z_t) \neq 0$, and $w_t$ is uncorrelated with $(x_t, \varepsilon_t)$

$$y_t w'_t = Dz_t w'_t + \varepsilon^*_t,$$

$$\hat{D} = \left(\frac{1}{n} \sum_{t=1}^{n} y_t w'_t R'\right)\left(\frac{1}{n} \sum_{t=1}^{n} z_t w'_t R'\right)^{-1}, \quad \text{rank}(R) = m.$$
Nonlinear regression function

Consider model

\[ y_t = g(u_t) + Ax_t + \varepsilon_t, \]

where \( u_t \) is an observed process with fixed dimension, \( g = (g_1, \ldots, g_p)' \) is an unknown nonlinear function.

Let

\[ g_i(u) = \sum_{j=1}^{\infty} d_{i,j} l_j(u), \quad i = 1, 2, \ldots, \]

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

\[ z_t = (l_1(u_t), \ldots, l_m(u_t))' \]

Then

\[ y_t = Dz_t + Ax_t + \varepsilon_t + e_t \]

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 $z_t = k$,

$$y_t = \mu_k + A_k x_t + \varepsilon_t^{(k)} \quad \text{and} \quad \varepsilon_t^{(k)} \sim N(0, \Sigma_k),$$

where $A_k' A_k = I_r$, $E(y_t | z_t = k) = \mu_k$, $E(x_t | z_t) = 0$ and

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

Typically $m$ is small.
Let \((A_k, B_k)\) be an orthogonal matrix.

Two-step iterative algorithm based on \(\{B_k y_t\}\) (instead of \(\{y_t\}\)):

1. Given \(\pi_{ij}, \mu_k, B_k, \Sigma_k, B\), estimate \(\{z_t\}\) by maximizing

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

- Use the Viterbi algorithm.
2. Given \( \{z_t\} \),

\[
\hat{\pi}_{ij} = \frac{\sum_{t=1}^{n-1} I(z_t = i, z_{t+1} = j)}{\sum_{t=1}^{n-1} I(z_t = i)}, \quad \hat{\mu}_k = \frac{\sum_{t=1}^{n} y_t I(z_t = k)}{\sum_{t=1}^{n} I(z_t = k)}
\]

\[
\hat{U}_{j,k} = \frac{\sum_{t=1}^{n-j} (y_t + j - \hat{\mu}_k)(y_t - \hat{\mu}_k)'I(z_{t+j} = k)}{\sum_{t=1}^{n-j} I(z_{t+j} = k)}
\]

\[
\hat{M}_k = \sum_{j=1}^{k_0} \hat{U}_{j,k} \hat{U}_{j,k}'
\]

Let \( \lambda_{1,k} \geq \cdots \geq \lambda_{p,k} \geq 0 \) be the \( p \) eigenvalues of \( \hat{M}_k \), and \( \gamma_1, \cdots, \gamma_p \) be the corresponding orthonormal eigenvectors,

\[
\hat{A}_k = (\gamma_1, \cdots, \gamma_d) \quad \text{and} \quad \hat{B}_k = (\gamma_{d+1}, \cdots, \gamma_d).
\]
Prediction

Let $y_t = (y_1, \ldots, y_t)$,

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

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

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

One-step-ahead prediction:

$$
y_{n}(1) = \hat{\mu}_{\hat{k}} + \hat{A}_{\hat{k}} x_{n}(1), \quad \text{where} \quad \hat{k} = \arg\max_{1 \leq k \leq m} \hat{\pi}_k(n + 1|n), \quad \text{or}
$$

$$
y_{n}(1) = \sum_{k=1}^{m} \hat{\pi}_k(n + 1|n)\{\hat{\mu}_k + \hat{A}_k x_{n}(1)\}.
$$