Factor analysis of panel data

Application of LS estimation method

Panel data

Data

Different types of data

- time series
- cross-sectional
- panel data

Data

Different types of data

- time series
- cross-sectional
- panel data

Panel data

Examples of panel data:

- collection of time series of a number of variables (time-series–cross-section):
  - OECD data set
  - Penn World Tables: national income of 188 countries over 50 years
  - ...
- collection of cross sectional data at several points in time (longitudinal data)
  - National Longitudinal Survey of Labor Market experience (NLS)
  - Michigan Panle Study of Income Dynamics (PSID)
  - German Socioeconomic Panel (GSOEP)
  - ...

Panel data

Examples of panel data:

- collection of time series of a number of variables (time-series–cross-section):
    - OECD data set
    - Penn World Tables: national income of 188 countries over 50 years
    - ...
- collection of cross sectional data at several points in time (longitudinal data)
    - National Longitudinal Survey of Labor Market experience (NLS)
    - Michigan Panle Study of Income Dynamics (PSID)
    - German Socioeconomic Panel (GSOEP)
    - ...

Panel data

Examples of panel data:

- collection of time series of a number of variables (time-series–cross-section):
  - OECD data set
  - Penn World Tables: national income of 188 countries over 50 years
  - ...
- collection of cross sectional data at several points in time (longitudinal data)
  - National Longitudinal Survey of Labor Market experience (NLS)
  - Michigan Panle Study of Income Dynamics (PSID)
  - German Socioeconomic Panel (GSOEP)
  - ...

Panel data

Examples of panel data:

- collection of time series of a number of variables (time-series–cross-section):
  - OECD data set
  - Penn World Tables: national income of 188 countries over 50 years
  - ...
- collection of cross sectional data at several points in time (longitudinal data)
  - National Longitudinal Survey of Labor Market experience (NLS)
  - Michigan Panle Study of Income Dynamics (PSID)
  - German Socioeconomic Panel (GSOEP)
  - ...

Panel data

Examples of panel data:

- collection of time series of a number of variables (time-series–cross-section):
  - OECD data set
  - Penn World Tables: national income of 188 countries over 50 years
  - ...
- collection of cross sectional data at several points in time (longitudinal data)
  - National Longitudinal Survey of Labor Market experience (NLS)
  - Michigan Panle Study of Income Dynamics (PSID)
  - German Socioeconomic Panel (GSOEP)
  - ...

Time-series/cross-section

Electricity prices:

- hourly prices for every day
- day-ahead pricing

Day/Hour	1	2	...	24
01.01.2014				
02.01.2014				
⋮				
31.12.2016				

Time-series/cross-section

Electricity prices:

- hourly prices for every day
- day-ahead pricing

Day/Hour	1	2	...	24
01.01.2014				
02.01.2014				
⋮				
31.12.2016				

Electricity prices

Features of the electricity prices:

- daily, weekly and yearly seasonality
- co-movement of prices within one day
- high volatility: spikes

Question: should hourly prices be modeled jointly or separately?

Electricity prices

Features of the electricity prices:

- daily, weekly and yearly seasonality
- co-movement of prices within one day
- high volatility: spikes

Question: should hourly prices be modeled jointly or separately?

Panel data

Factor model

Factor model

Assumption: the co-movement of variables Y_{ti}, $i = 1, ..., N$ is govern by K common factors F_{tk}, where $K \ll N$.

Model:

$$Y_{ti} = F_t \lambda_i + e_{ti},$$

where

- $F_t = [F_{t1}, ..., F_{tK}]$ is a $(1 \times K)$ vector of common factors,
- $\lambda_i = [\lambda_{i1}, ..., \lambda_{iK}]'$ is a $(K \times 1)$ vector of factor loadings,
- e_{ti} is an idiosyncratic component (remark: e_{it} are only weakly correlated across sections i).

Factor model

Assumption: the co-movement of variables Y_{ti}, $i = 1, ..., N$ is govern by K common factors F_{tk}, where $K \ll N$.

Model:

$$Y_{ti} = F_t \lambda_i + e_{ti},$$

where

- $F_t = [F_{t1}, ..., F_{tK}]$ is a $(1 \times K)$ vector of common factors,
- $\lambda_i = [\lambda_{i1}, ..., \lambda_{iK}]'$ is a $(K \times 1)$ vector of factor loadings,
- e_{ti} is an idiosyncratic component (remark: e_{it} are only weakly correlated across sections i).

Factor model

Assumption: the co-movement of variables Y_{ti}, $i = 1, ..., N$ is govern by K common factors F_{tk}, where $K \ll N$.

Model:

$$Y_{ti} = F_t \lambda_i + e_{ti},$$

where

- $F_t = [F_{t1}, ..., F_{tK}]$ is a $(1 \times K)$ vector of common factors,
- $\lambda_i = [\lambda_{i1}, ..., \lambda_{iK}]'$ is a $(K \times 1)$ vector of factor loadings,
- e_{ti} is an idiosyncratic component (remark: e_{it} are only weakly correlated across sections i).

Factor model

Assumption: the co-movement of variables Y_{ti}, $i = 1, ..., N$ is govern by K common factors F_{tk}, where $K \ll N$.

Model:

$$Y_{ti} = F_t \lambda_i + e_{ti},$$

where

- $F_t = [F_{t1}, ..., F_{tK}]$ is a $(1 \times K)$ vector of common factors,
- $\lambda_i = [\lambda_{i1}, ..., \lambda_{iK}]'$ is a $(K \times 1)$ vector of factor loadings,
- e_{ti} is an idiosyncratic component (remark: e_{it} are only weakly correlated across sections i).

Factor model

Assumption: the co-movement of variables Y_{ti}, $i = 1, ..., N$ is govern by K common factors F_{tk}, where $K \ll N$.

Model:

$$Y_{ti} = F_t \lambda_i + e_{ti},$$

where

- $F_t = [F_{t1}, ..., F_{tK}]$ is a $(1 \times K)$ vector of common factors,
- $\lambda_i = [\lambda_{i1}, ..., \lambda_{iK}]'$ is a $(K \times 1)$ vector of factor loadings,
- e_{ti} is an idiosyncratic component (remark: e_{it} are only weakly correlated across sections i).

Factor model

Assumption: the co-movement of variables Y_{ti}, $i = 1, ..., N$ is govern by K common factors F_{tk}, where $K \ll N$.

Model:

$$Y_{ti} = F_t \lambda_i + e_{ti},$$

where

- $F_t = [F_{t1}, ..., F_{tK}]$ is a $(1 \times K)$ vector of common factors,
- $\lambda_i = [\lambda_{i1}, ..., \lambda_{iK}]'$ is a $(K \times 1)$ vector of factor loadings,
- e_{ti} is an idiosyncratic component (remark: e_{it} are only weakly correlated across sections i).

Factor model

Assumption: the co-movement of variables Y_{ti}, $i = 1, ..., N$ is govern by K common factors F_{tk}, where $K \ll N$.

Model:

$$Y_{ti} = F_t \lambda_i + e_{ti},$$

where

- $F_t = [F_{t1}, ..., F_{tK}]$ is a $(1 \times K)$ vector of common factors,
- $\lambda_i = [\lambda_{i1}, ..., \lambda_{iK}]'$ is a $(K \times 1)$ vector of factor loadings,
- e_{ti} is an idiosyncratic component (remark: e_{it} are only weakly correlated across sections i).

Factor model

Assumption: the co-movement of variables Y_{ti}, $i = 1, ..., N$ is govern by K common factors F_{tk}, where $K \ll N$.

Model:

$$Y_{ti} = F_t \lambda_i + e_{ti},$$

where

- $F_t = [F_{t1}, ..., F_{tK}]$ is a $(1 \times K)$ vector of common factors,
- $\lambda_i = [\lambda_{i1}, ..., \lambda_{iK}]'$ is a $(K \times 1)$ vector of factor loadings,
- e_{ti} is an idiosyncratic component (remark: e_{it} are only weakly correlated across sections i).

Factor model

Notation

- Y - a $(T \times N)$ panel of observations,
- F - a $(T \times K)$ matrix of common factors,
- λ - a $(K \times N)$ matrix of loadings, $\lambda = [\lambda_1, ..., \lambda_K]$
- e - a $(T \times N)$ panel of idiosyncratic components,

Model:

$$Y = F\lambda + e$$

Factor model

Notation

- Y - a $(T \times N)$ panel of observations,
- F - a $(T \times K)$ matrix of common factors,
- λ - a $(K \times N)$ matrix of loadings, $\lambda = [\lambda_1, ..., \lambda_K]$
- e - a $(T \times N)$ panel of idiosyncratic components,

Model:

$$Y = F\lambda + e$$

Factor model

Notation

- Y - a $(T \times N)$ panel of observations,
- F - a $(T \times K)$ matrix of common factors,
- λ - a $(K \times N)$ matrix of loadings, $\lambda = [\lambda_1, ..., \lambda_K]$
- e - a $(T \times N)$ panel of idiosyncratic components,

Model:

$$Y = F\lambda + e$$

Factor model

Notation

- Y - a $(T \times N)$ panel of observations,
- F - a $(T \times K)$ matrix of common factors,
- λ - a $(K \times N)$ matrix of loadings, $\lambda = [\lambda_1, ..., \lambda_K]$
- e - a $(T \times N)$ panel of idiosyncratic components,

Model:

$$Y = F\lambda + e$$

Factor model

Notation

- Y - a $(T \times N)$ panel of observations,
- F - a $(T \times K)$ matrix of common factors,
- λ - a $(K \times N)$ matrix of loadings, $\lambda = [\lambda_1, ..., \lambda_K]$
- e - a $(T \times N)$ panel of idiosyncratic components,

Model:

$$Y = F\lambda + e$$

LS estimation of a Factor model

Model

$$Y = F\lambda + e$$

Problems:

- Factors are not known,
- If factors are known, λ could be estimated with LS...
- Identification: when F is a factor then $G = FA$ is a factor, if $rk(A) = K$. What are loadings then?
- How to identify and estimate F?

LS estimation of a Factor model

Model

$$Y = F\lambda + e$$

Problems:

- Factors are not known,
- If factors are known, λ could be estimated with LS...
- Identification: when F is a factor then $G = FA$ is a factor, if $rk(A) = K$. What are loadings then?
- How to identify and estimate F?

LS estimation of a Factor model

Model

$$Y = F\lambda + e$$

Problems:

- Factors are not known,
- If factors are known, λ could be estimated with LS...
- Identification: when F is a factor then $G = FA$ is a factor, if $rk(A) = K$. What are loadings then?
- How to identify and estimate F?

LS estimation of a Factor model

Model

$$Y = F\lambda + e$$

Problems:

- Factors are not known,
- If factors are known, λ could be estimated with LS...
- Identification: when F is a factor then $G = FA$ is a factor, if $rk(A) = K$. What are loadings then?
- How to identify and estimate F?

LS estimation of a Factor model

Model

$$Y = F\lambda + e$$

Problems:

- Factors are not known,
- If factors are known, λ could be estimated with LS...
- Identification: when F is a factor then $G = FA$ is a factor, if $rk(A) = K$. What are loadings then?
- How to identify and estimate F?

Identification and estimation

Identification:

$$\frac{F'F}{T} = I_K$$

Estimation: minimize the sum of squares of idiosyncratic components

$$L = \sum_{i=1}^{N} \sum_{t=1}^{T} e_{it}^2$$

Identification and estimation

Identification:

$$\frac{F'F}{T} = I_K$$

Estimation: minimize the sum of squares of idiosyncratic components

$$L = \sum_{i=1}^{N} \sum_{t=1}^{T} e_{it}^2$$

Identification and estimation

Identification:

$$\frac{F'F}{T} = I_K$$

Estimation: minimize the sum of squares of idiosyncratic components

$$L = \sum_{i=1}^{N} \sum_{t=1}^{T} e_{it}^2$$

Identification and estimation

Identification:

$$\frac{F'F}{T} = I_K$$

Estimation: minimize the sum of squares of idiosyncratic components

$$L = \sum_{i=1}^{N} \sum_{t=1}^{T} e_{it}^2$$

Loss function

Loss funciton

$$L = tr(e'e)$$

since $e = F\lambda$

$$L = tr((Y - F\lambda)'(Y - F\lambda))$$

Loss function

Loss funciton

$$L = tr(e'e)$$

since $e = F\lambda$

$$L = tr((Y - F\lambda)'(Y - F\lambda))$$

Loss function

Suppose F is known then

$$\hat{\lambda} = (F'F)^{-1}F'Y$$

Hence

$$e = M_F Y,$$

where

$$M_F = I - F(F'F)^{-1}F'$$

Loss function

Suppose F is known then

$$\hat{\lambda} = (F'F)^{-1}F'Y$$

Hence

$$e = M_F Y,$$

where

$$M_F = I - F(F'F)^{-1}F'$$

Loss function

Suppose F is known then

$$\hat{\lambda} = (F'F)^{-1}F'Y$$

Hence

$$e = M_F Y,$$

where

$$M_F = I - F(F'F)^{-1}F'$$

Loss function

It can be shown that

$$L = tr(YY') - tr(F(F'F)^{-1}F'YY')$$

Under identification condition

$$L = tr(Y'Y) - \frac{1}{T}tr(F'YY'F)$$

If F minimizes L then it maximizes

$$\frac{1}{T}tr(F'YY'F)$$

Loss function

It can be shown that

$$L = tr(YY') - tr(F(F'F)^{-1}F'YY')$$

Under identification condition

$$L = tr(Y'Y) - \frac{1}{T} tr(F'YY'F)$$

If F minimizes L then it maximizes

$$\frac{1}{T} tr(F'YY'F)$$

Loss function

It can be shown that

$$L = tr(YY') - tr(F(F'F)^{-1}F'YY')$$

Under identification condition

$$L = tr(Y'Y) - \frac{1}{T}tr(F'YY'F)$$

If F minimizes L then it maximizes

$$\frac{1}{T}tr(F'YY'F)$$

Optimization problem

Let

$$\tilde{F} = F/\sqrt{T}$$

then

$$\frac{1}{T} tr(F'YY'F) = tr(\tilde{F}'YY'\tilde{F})$$

and

$$\tilde{F}'\tilde{F} = I_K$$

Optimization problem

Let

$$\tilde{F} = F/\sqrt{T}$$

then

$$\frac{1}{T} tr(F'YY'F) = tr(\tilde{F}'YY'\tilde{F})$$

and

$$\tilde{F}'\tilde{F} = I_K$$

Optimization problem

Let

$$\tilde{F} = F/\sqrt{T}$$

then

$$\frac{1}{T} tr(F'YY'F) = tr(\tilde{F}'YY'\tilde{F})$$

and

$$\tilde{F}'\tilde{F} = I_K$$

Optimization problem

Let

$$\tilde{F} = F/\sqrt{T}$$

then

$$\frac{1}{T} tr(F'YY'F) = tr(\tilde{F}'YY'\tilde{F})$$

and

$$\tilde{F}'\tilde{F} = I_K$$

LS estimator

It can be shown that

$$\hat{\hat{F}} = V_{1:K},$$

where $V_{1:K}$ are the eigenvectors of YY' corresponding with the K largest eigenvalues $(\gamma_1, ..., \gamma_K)$.

$$\hat{F} = \sqrt{T} V_{1:K},$$

$$\hat{\lambda} = \hat{F}' Y / T$$

LS estimator

It can be shown that

$$\hat{\tilde{F}} = V_{1:K},$$

where $V_{1:K}$ are the eigenvectors of YY' corresponding with the K largest eigenvalues $(\gamma_1, ..., \gamma_K)$.

$$\hat{F} = \sqrt{T} V_{1:K},$$

$$\hat{\lambda} = \hat{F}' Y / T$$

Variance of Y explained by Factors

Total variance

$$tr(Y'Y) = \sum_{i=1}^{T} \gamma_i$$

Variance of residuals

$$tr(e'e) = tr(Y'Y) - tr(F'YY'F)/T$$

Explained variance

$$tr(F'YY'F)/T = \sum_{i=1}^{K} \gamma_i$$

Variance of Y explained by Factors

Total variance

$$tr(Y'Y) = \sum_{i=1}^{T} \gamma_i$$

Variance of residuals

$$tr(e'e) = tr(Y'Y) - tr(F'YY'F)/T$$

Explained variance

$$tr(F'YY'F)/T = \sum_{i=1}^{K} \gamma_i$$

Variance of Y explained by Factors

Total variance

$$tr(Y'Y) = \sum_{i=1}^{T} \gamma_i$$

Variance of residuals

$$tr(e'e) = tr(Y'Y) - tr(F'YY'F)/T$$

Explained variance

$$tr(F'YY'F)/T = \sum_{i=1}^{K} \gamma_i$$

Variance of Y explained by Factors

Total variance

$$tr(Y'Y) = \sum_{i=1}^{T} \gamma_i$$

Variance of residuals

$$tr(e'e) = tr(Y'Y) - tr(F'YY'F)/T$$

Explained variance

$$tr(F'YY'F)/T = \sum_{i=1}^{K} \gamma_i$$

Variance of Y explained by Factors

Total variance

$$tr(Y'Y) = \sum_{i=1}^{T} \gamma_i$$

Variance of residuals

$$tr(e'e) = tr(Y'Y) - tr(F'YY'F)/T$$

Explained variance

$$tr(F'YY'F)/T = \sum_{i=1}^{K} \gamma_i$$

Variance of Y explained by Factors

Total variance

$$tr(Y'Y) = \sum_{i=1}^{T} \gamma_i$$

Variance of residuals

$$tr(e'e) = tr(Y'Y) - tr(F'YY'F)/T$$

Explained variance

$$tr(F'YY'F)/T = \sum_{i=1}^{K} \gamma_i$$

Share of explained variance

Share of explained variance

$$\frac{\sum_{i=1}^{K} \gamma_i}{\sum_{i=1}^{T} \gamma_i}$$

Can be used to choose the number of factors.

Number of factors

Notation

- $k = 1, ..., k_{max}$ - considered number of factors
- $e^{(k)}$ - idiosyncratic component for k factors
- $V(k) = \frac{1}{NT} \sum_{t=1}^{T} \sum_{i=1}^{N} (e_{ti}^{(k)})^2$
- $\hat{\sigma}^2 = V(k_{max})$ - a consistent estimator of the variance

Number of factors - IC

Bai, Ng (2002), Determining the number of factors in approximate Factor model, Econometrica, Vol. 70, No. 1, pp. 191-221

$$PC_1(k) = V(k) + k\hat{\sigma}^2\left(\frac{N+T}{NT}\right)\ln\left(\frac{NT}{N+T}\right)$$

$$IPC_1(k) = \ln V(k) + k\left(\frac{N+T}{NT}\right)\ln\left(\frac{NT}{N+T}\right)$$

Number of factors - IC

Procedure

- Set the maximum considered number of factors k_{max}
- Compute IC for $k = 1, 2, ..., k_{max}$
- Choose the number of factors minimizing IC.

Number of factors - IC

Procedure

- Set the maximum considered number of factors k_{max}
- Compute IC for $k = 1, 2, ..., k_{max}$
- Choose the number of factors minimizing IC.

Number of factors - IC

Procedure

- Set the maximum considered number of factors k_{max}
- Compute IC for $k = 1, 2, ..., k_{max}$
- Choose the number of factors minimizing IC.

Number of factors - IC

Procedure

- Set the maximum considered number of factors k_{max}
- Compute IC for $k = 1, 2, ..., k_{max}$
- Choose the number of factors minimizing IC.

Factors

Notice that

- H is a $(K \times K)$ matrix with ortonormal columns ($H'H = I_K$)
- $\hat{G} = \hat{F}H$ and $\hat{\lambda}_G = H^{-1}\hat{\lambda}$
- Do $\hat{G}$ and $\hat{\lambda}_G$ maximize the loss function?
- Do $\hat{G}$ satisfies the identification condition?

Factors

Notice that

- H is a $(K \times K)$ matrix with ortonormal columns $(H'H = I_K)$
- $\hat{G} = \hat{F}H$ and $\hat{\lambda}_G = H^{-1}\hat{\lambda}$
- Do $\hat{G}$ and $\hat{\lambda}_G$ maximize the loss function?
- Do $\hat{G}$ satisfies the identification condition?

Factors

Notice that

- H is a $(K \times K)$ matrix with ortonormal columns ($H'H = I_K$)
- $\hat{G} = \hat{F}H$ and $\hat{\lambda}_G = H^{-1}\hat{\lambda}$
- Do $\hat{G}$ and $\hat{\lambda}_G$ maximize the loss function?
- Do $\hat{G}$ satisfies the identification condition?

Factors

Notice that

- H is a $(K \times K)$ matrix with ortonormal columns ($H'H = I_K$)
- $\hat{G} = \hat{F}H$ and $\hat{\lambda}_G = H^{-1}\hat{\lambda}$
- Do $\hat{G}$ and $\hat{\lambda}_G$ maximize the loss function?
- Do $\hat{G}$ satisfies the identification condition?

Factors

Notice that

- H is a $(K \times K)$ matrix with ortonormal columns $(H'H = I_K)$
- $\hat{G} = \hat{F}H$ and $\hat{\lambda}_G = H^{-1}\hat{\lambda}$
- Do $\hat{G}$ and $\hat{\lambda}_G$ maximize the loss function?
- Do $\hat{G}$ satisfies the identification condition?

Factors

Question: how to relate the true factors with the estimator $\hat{F}$

- no direct link - possible rotation
- interpretation - based on loading matrix

Factors

Question: how to relate the true factors with the estimator $\hat{F}$

- no direct link - possible rotation
- interpretation - based on loading matrix

Factors

Question: how to relate the true factors with the estimator $\hat{F}$

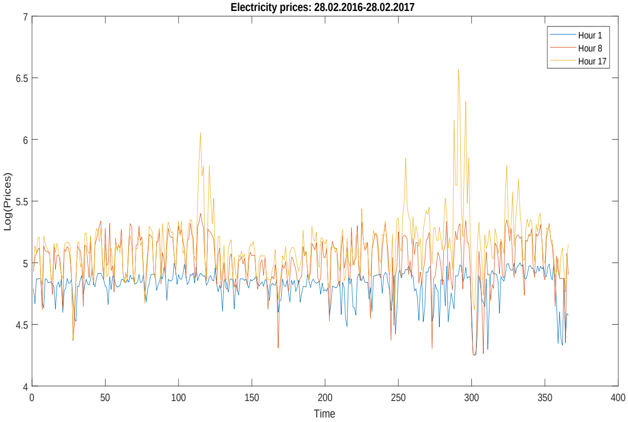
- no direct link - possible rotation
- interpretation - based on loading matrix

Electricity prices

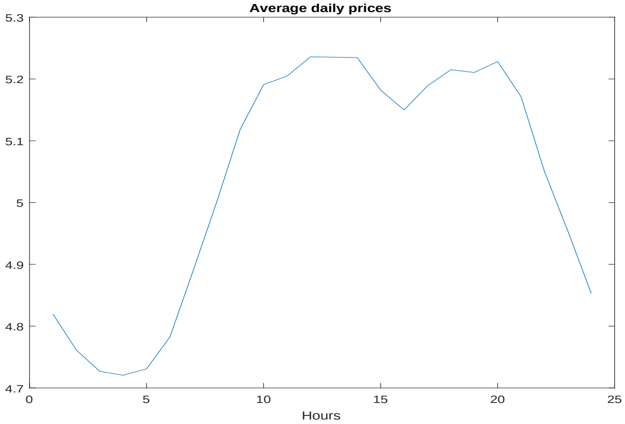
Data:

- Polish power exchange - day ahead market
- Hourly data
- Time span: 01.01.2015 - 28.02.2017
- Transformed into logarithms
- Panel of dimension (790×24)

Prices


Electricity prices: 28.02.2016-28.02.2017
Hour 1
Hour 8
Hour 17
Log(Prices)
Time
7
6.5
6
5.5
5
4.5
4
0
50
100
150
200
250
300
350
400

Average prices



Model specification

- Maximum number of factors is $k_{max} = 4$
- Hourly averages are removed
- Question: should the data be standardized?

Model specification

- Maximum number of factors is $k_{max} = 4$
- Hourly averages are removed
- Question: should the data be standardized?

Model specification

- Maximum number of factors is $k_{max} = 4$
- Hourly averages are removed
- Question: should the data be standardized?

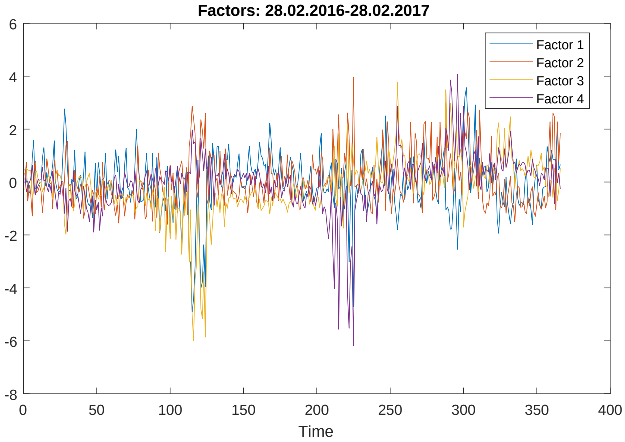
Model specification

- Maximum number of factors is $k_{max} = 4$
- Hourly averages are removed
- Question: should the data be standardized?

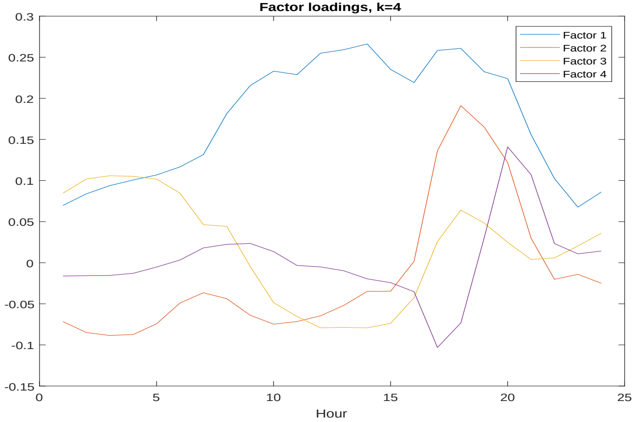
Information criteria and explained variability

k	Variability	PC	IPC
1	0.6822	0.0141	-4.1605
2	0.7922	0.0099	-4.4501
3	0.8801	0.0066	-4.8647
4	0.9168	0.0055	-5.0952

Factors


Factors: 28.02.2016-28.02.2017
Factor 1
Factor 2
Factor 3
Factor 4
6
4
2
0
-2
-4
-6
-8
0
50
100
150
200
250
300
350
400
Time

Loadings


Factor loadings, k=4
0.3
0.25
0.2
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
0
5
10
15
20
25
Hour
Factor 1
Factor 2
Factor 3
Factor 4

Conclusions

- Factor models could be useful in reducing the dimension of the data set.
- Factors could be estimated with the Least Square method
- In a case of an Electricity market
  - Four factors explain 91.7% of variability of log prices
  - First factor is the average level of prices
  - Second factor describes the evening peak
  - Third factor represents the morning peak