Estimation theory – Report 3

Marta Frankowska, 208581
Agnieszka Szkutek, 208619

December 1, 2017

Contents

1 Model

In both exercises we will be using the Factor model

$$Y_{T\times N} = F_{T\times K} \cdot \lambda_{K\times N} + e_{T\times N},$$

where

- $Y_{T\times N}$ panel of observations
- $F_{T\times K}$ matrix of common (latent) factors
- $\lambda_{K\times N}$ matrix of loadings
- $e_{T\times N}$ panel of specific components

To calculate F and λ we use the following formulas:

$$\hat{F} = \sqrt{T} V_{1:K} \quad \text{and} \quad \hat{\lambda} = \frac{\hat{F}'Y}{T},$$

where $V_{1:K}$ are eigenvectors of YY' corresponding to the K largest eigenvalues.

1.1 Selecting optimal number of factors

Notation:

- $K = 1, 2, \ldots, K_{\max}$ - the number of factors,
- $e^{(K)}$ - the individual components for K factors,
- $V(K) = \frac{1}{NT} \sum_{t=1}^{T} \sum_{i=1}^{N} \left(e_{ti}^{(K)}\right)^2$,
- $\hat{\sigma}^2 = V(K_{\max})$ - consistent estimator of variance.

Information criteria:

- $PC_1(K) = V(K) + K\hat{\sigma}^2 \frac{N+T}{NT} \ln \frac{NT}{N+T}$,
- $IPC_1(K) = \log V(K) + K \frac{N+T}{NT} \ln \frac{NT}{N+T}$.

Algorithm:

1. Set $K_{\max}$;
2. Compute $IC(K)$ for $K = 1, \ldots, K_{\max}$;
3. Choose $\hat{K}$ such that $IC(\hat{K}) = \min_{1 \leq K \leq K_{\max}} IC(K)$.

```
factor.model.est <-
  function(Y, K_max, draw) #function returning which K we should choose
  {
    T <- nrow(Y)
    N <- ncol(Y)

    eigen.decomp <- eigen(Y %*% t(Y))
    eigen.values <- eigen.decomp$values
    eigen.vectors <- eigen.decomp$vector
    # we calculate F, lambda and e for K_max
    F <- sqrt(T) * eigen.vectors[, 1:K_max]
    lambda <- t(F) %*% Y / T
    e <- Y - F %*% lambda
    sigma2.hat <- sum(e ^ 2) / (N * T)

    PC1 <- 1:K_max
    IPC1 <- 1:K_max
    for (K in 1:K_max)
    {
      # we calculate F, lambda and e for K
      F <- sqrt(T) * eigen.vectors[, 1:K]
      lambda <- t(F) %*% Y / T
      e <- Y - F %*% lambda
      V <- sum(e ^ 2) / (N * T)
      # we calculate PC1 and IPC1 for K
```

```
    PC1[K] <-
      V + K * sigma2.hat * ((N + T) / (N * T)) * log(N * T / (N + T))
    IPC1[K] <-
      log(V) + K * ((N + T) / (N * T)) * log(N * T / (N + T))
  }
  # choose K which gives the minimal value
  PC1_K <- which.min(PC1)
  IPC1_K <- which.min(IPC1)

  if (draw) {
    max.y <- max(max(IPC1), max(PC1)) + 2
    par(mfrow = c(1,1), mar=c(4,4,1,2))
    matplot(1:K_max, cbind(IPC1, PC1), pch=1, col=c("blue", "red"),
            xlab="K", ylab="IC",
            ylim = c(min(PC1[PC1_K],IPC1[IPC1_K]),
                     max(max(IPC1),max(PC1))+2))
    legend(1, max.y, c("IPC1", "PC1"), col = c("blue", "red"), pch=1)
  }
  return (list(PC1_K, IPC1_K))
}
```

2 Exercise 1

We will be using data from file *dataLab3.xlsx*, where Y size is $T \times N = 100 \times 100$. To calculate the number of factors K we will use function *factor.model.est*.

2.1 Part 1

First, we calculate the number of factors for the whole sample.

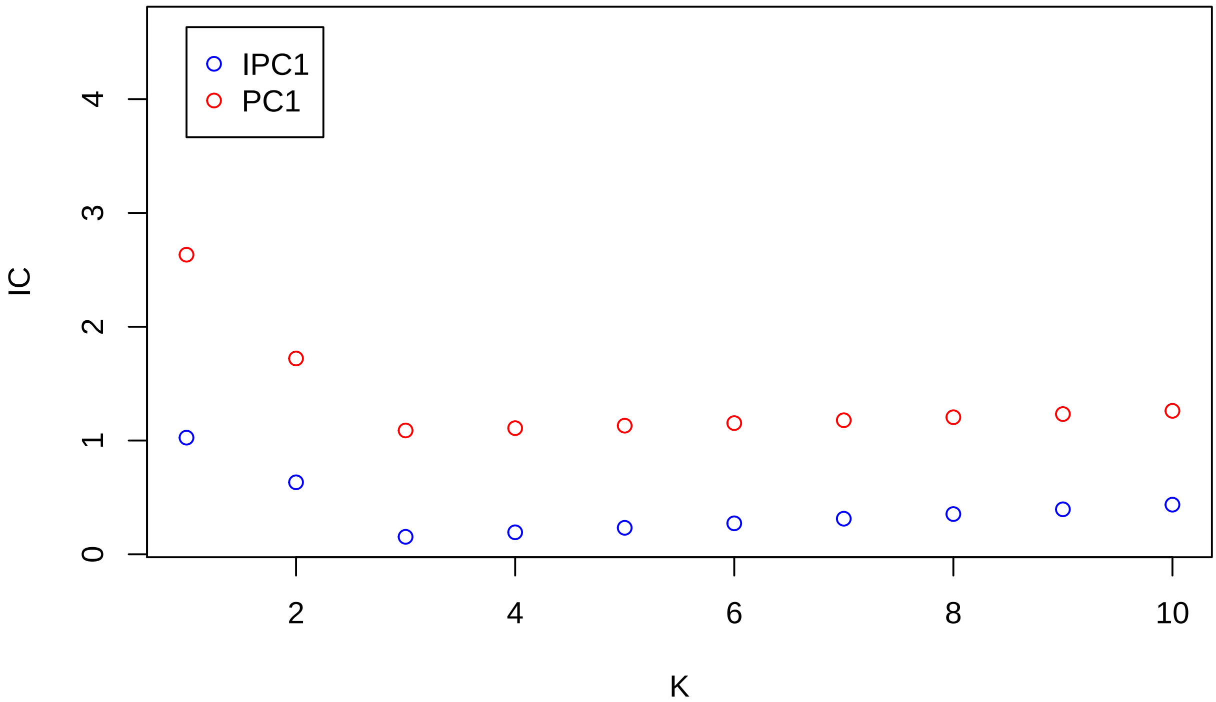


Figure 1: Information criteria

The function returns the same $\hat{K}$ for both PC_1 and IPC_1, it is equal to

	PC1	IPC1
K	3.00	3.00

Table 1: K returned for both Information criteria

Share of explained variance

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

where γ_i are the eigenvalues of YY' can also be used to choose the number of factors.

The plot shows the variability of the factors

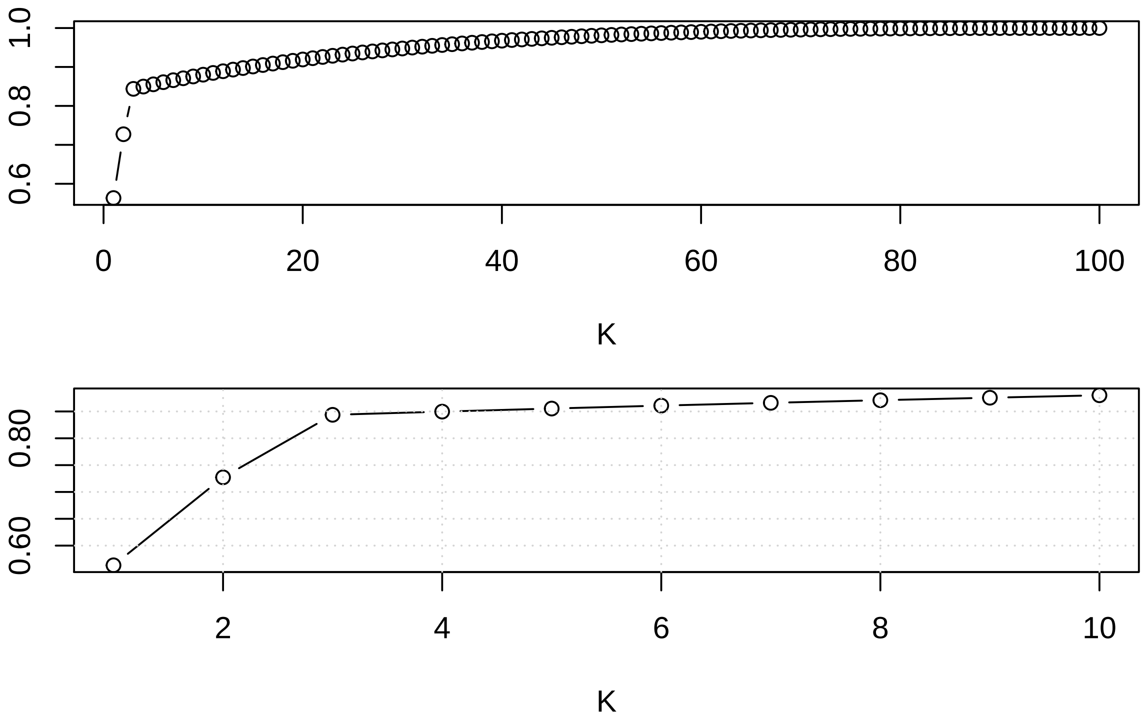


Figure 2: Share of explained variance depending on K

We can observe that for K we calculated from Information Criteria, the share of explained variance is more than 0.8, and to be exact we can calculate it from the formula (1):

```
## [1] 0.8437638
```

We can also say that the results are correct by looking at the eigenvalues

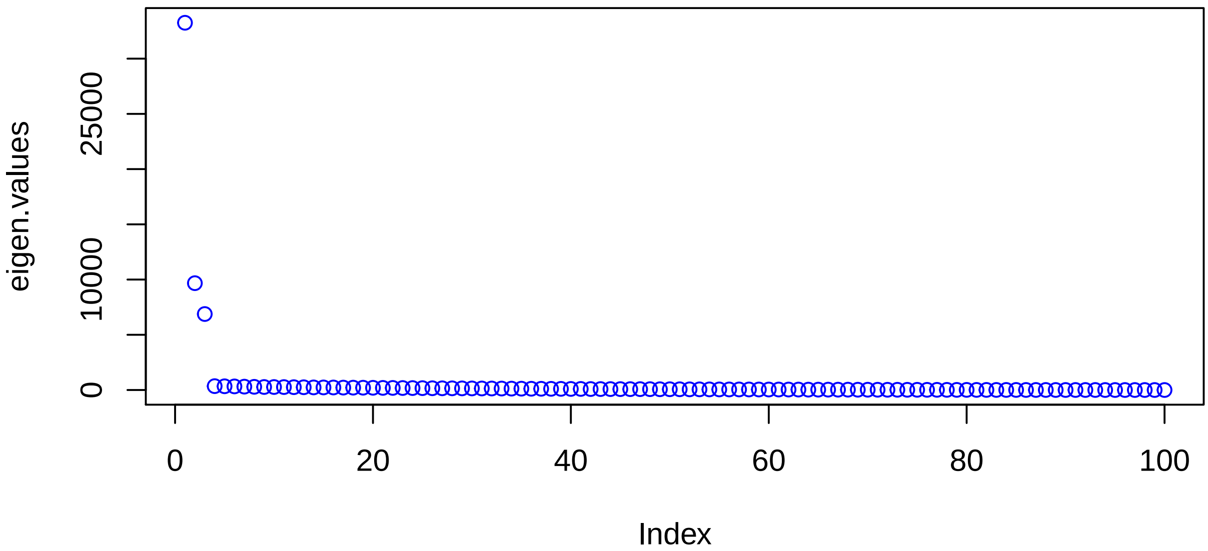


Figure 3: Eigenvalues of YY'

Now we can calculate $\hat{F}$ and $\hat{\lambda}$

```
source("functions.R")
Y <- as.matrix(read_excel('dataLab3.xlsx', col_names = FALSE))
N <- ncol(Y); T <- nrow(Y)
K <- as.numeric(factor.model.est(Y, 10, FALSE)[1])

eigen.decomp <- eigen(Y %*% t(Y))
eigen.vectors <- eigen.decomp$vectors
F <- sqrt(T) * eigen.vectors[, 1:K]
lambda <- t(F) %*% Y / T
```

To check if the results are correct, we can check the following condition

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

We calculate $\frac{F'F}{T}$ and the result is as follows

```
##              [,1]         [,2]          [,3]
## [1,]  1.000000e+00 3.774758e-17 -3.474998e-16
## [2,]  3.774758e-17 1.000000e+00  2.642331e-16
## [3,] -3.474998e-16 2.642331e-16  1.000000e+00
```

Taking into account the numerical errors we can assume that the above matrix is an identity matrix.

2.2 Part 2 and 3

Now we will compare estimated number of factors for the whole sample, the first 20 columns and the first 20 rows.

	PC1	IPC1
whole sample	3.00	3.00
first 20 columns	9.00	3.00
first 20 rows	9.00	3.00

Table 2: Comparison of results

We can observe that the results from IPC_1 and PC_1 differ depending on the size of the data. In particular for the sample of the first 20 columns the Information criteria look as follows

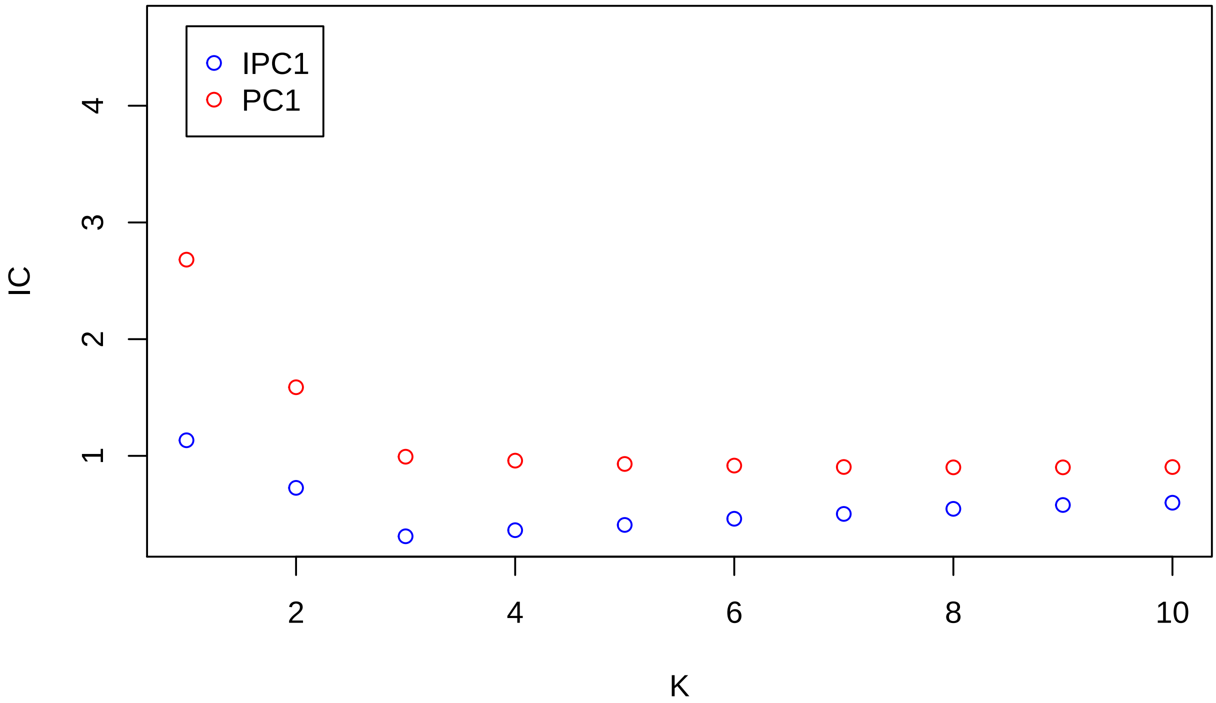


Figure 4: Information criteria for the first 20 columns

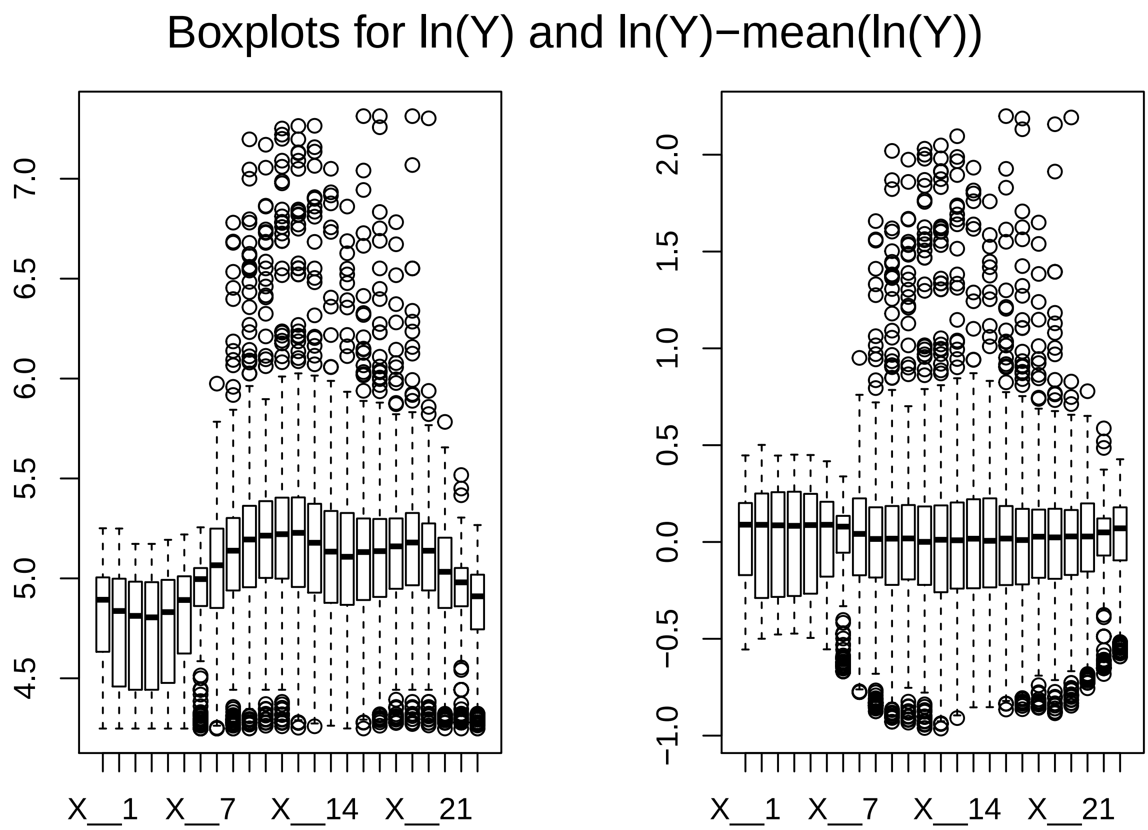
3 Exercise 2

In this exercise we will be working with data representing electricity prices from the balancing market. Each row describes the day, whereas the column describes the hour.

3.1 Part 1

We transform the data into logarithms and calculate mean for each column. Then we subtract the mean from each column.

```
Y <- as.matrix(read_excel('RB.xlsx', col_names = FALSE))
N <- ncol(Y); T <- nrow(Y)
Y <- scale(log(Y), center = TRUE, scale = FALSE)
```



3.2 Part 2

The plot shows the variability of the factors

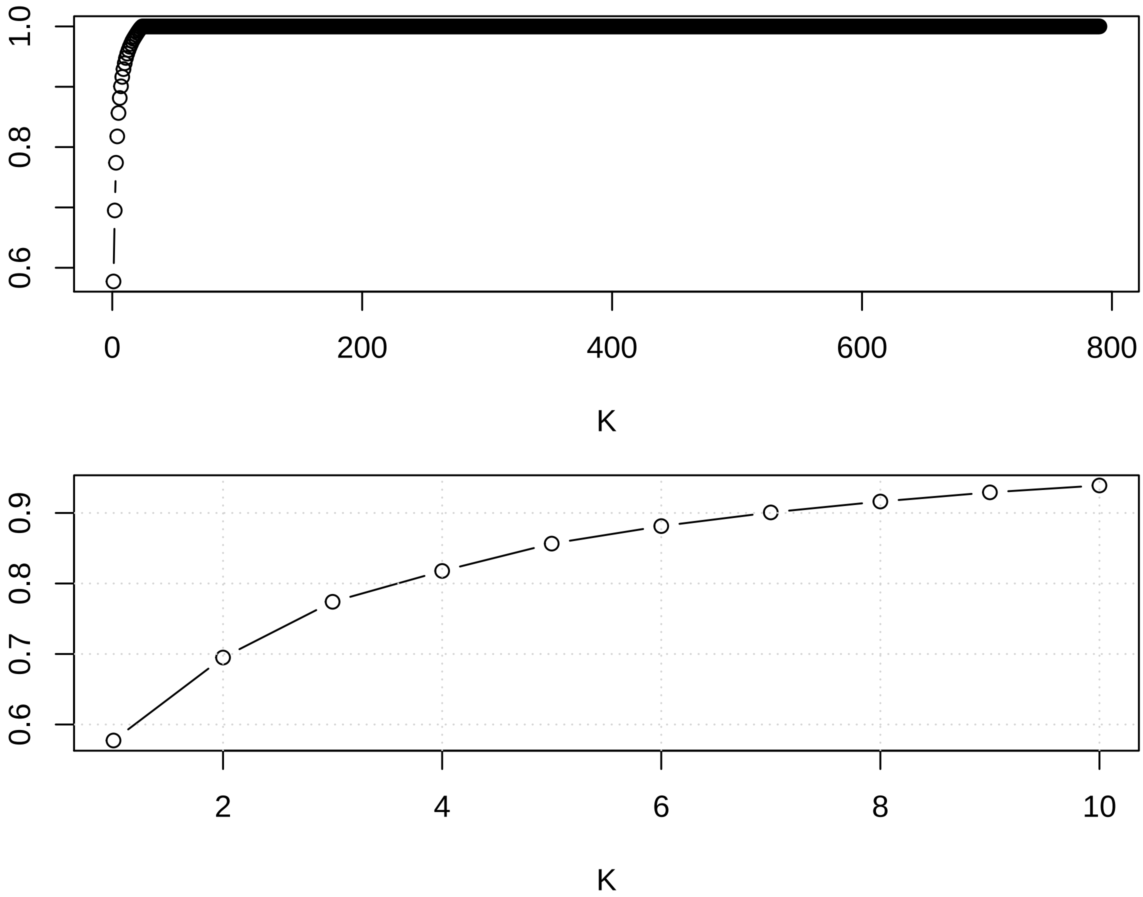


Figure 5: Share of explained variance

We can observe that if we want to have the Factor model which explains at least 80% of panel variability, we have to choose K equal to

```
## [1] 4
```

Now we can calculate $\hat{F}$ and $\hat{\lambda}$ and check the following condition

$$\frac{F'F}{T} = I.$$

We calculate $\frac{F'F}{T}$ and the result is as follows

```
##               [,1]          [,2]          [,3]          [,4]
## [1,]  1.000000e+00 -2.754477e-17  8.460181e-17  1.048388e-16
## [2,] -2.754477e-17  1.000000e+00 -3.147974e-17 -1.433453e-16
## [3,]  8.460181e-17 -3.147974e-17  1.000000e+00 -4.216037e-17
## [4,]  1.048388e-16 -1.433453e-16 -4.216037e-17  1.000000e+00
```

Taking into account the numerical errors we can assume that the above matrix is an identity matrix.

3.3 Part 3

We want to compute the information criteria with $K_{\max} = 8$. They suggest the following number of factors:

	Suggested no. of factors
PC1	8.00
IPC1	8.00

Table 3: Suggested number of factors

What is more, if we increase $K_{\max}$ the Information criteria will return the following results

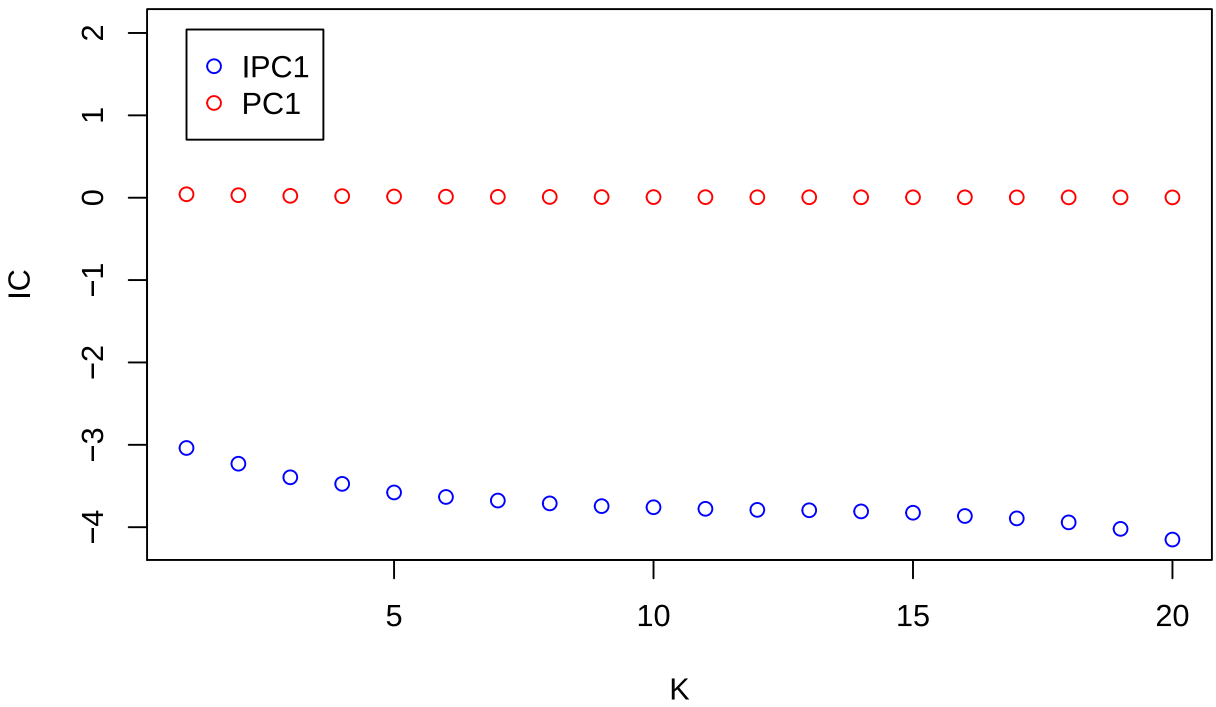


Figure 6: Information criteria for increased K_{max}

The bigger $K_{\max}$ we take then the bigger K is returned from Information criteria. So we concluded that Information criteria don't work in this case.

3.4 Part 4

Since Information criteria don't work for this data, we take K calculated from share of explained variance. We are going to plot loadings of the first two factors. We change signs of values in λ and F, so that values in λ in 17th column are non-negative.

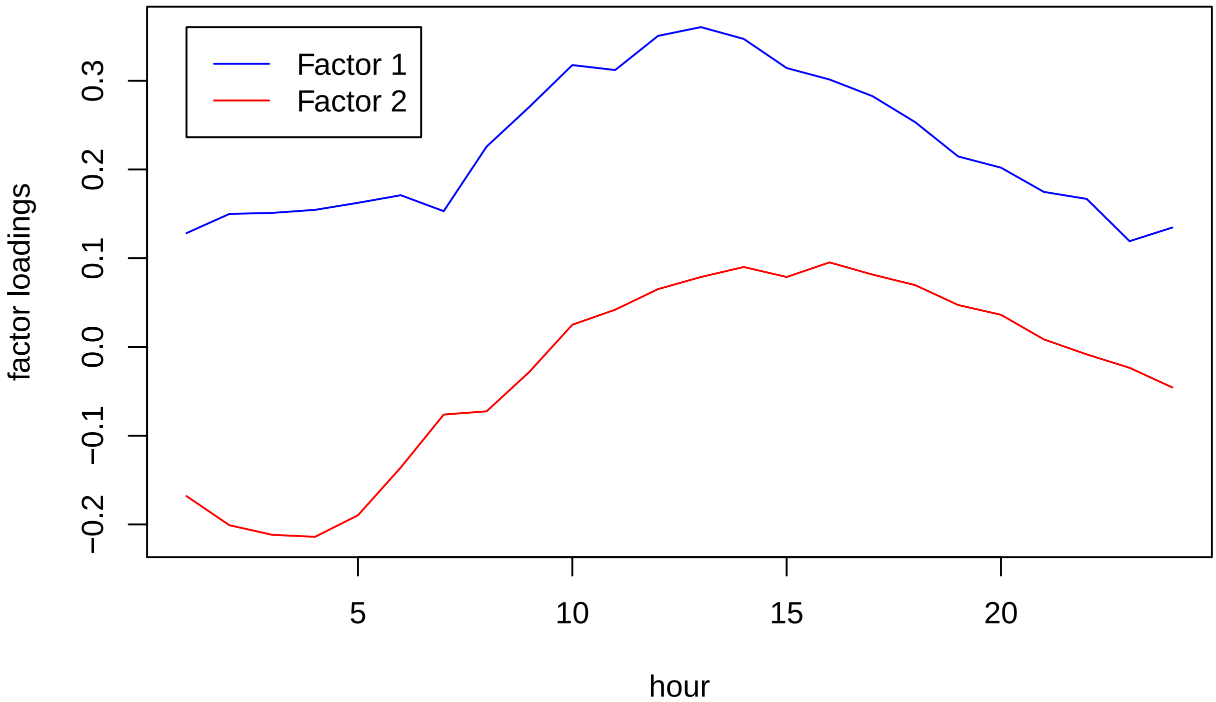


Figure 7: Factor loadings for K=4

Factor 1 describes the noon peak and Factor 2 describes lower prices of electricity at night.