Clustering time series by dependence measures

Andrés M. Alonso[1,2] and Daniel Peña [1,3]

[1]Department of Statistics

[2]Institute Flores de Lemus

[3]UC3M-BS Institute of Financial Big Data

VI Jornada de Estadística - UAM
February 24, 2017, Cantoblanco

Outline

The problem

Time series clustering problems arise when we observe a sample of time series and we want to group them into different categories or clusters.

This a central problem in many application fields and hence time series clustering is nowadays an active research area in different disciplines including finance and economics, medicine, engineering, seismology and meteorology, among others.

Dissimilarity measures

Key point

The metric chosen to assess the dissimilarity between two data objects plays a crucial role in time series clustering.

Dissimilarity measures

Different dissimilarity criteria specifically designed to deal with time series have been proposed in the literature. Some examples are:

Maharaj, E.A. A significance test for classifying ARMA models, J. Statist. Comput. Simul. **54**, 305-331 (1996)

Kakizawa, Y., Shumway, R.H. and Taniguchi, M. Discrimination and clustering for multivariate time series, J. Am. Stat. Assoc. **93**, 328–340, (1998)

Vilar, J.A. and Pértega, S. Discriminant and cluster analysis for Gaussian stationary processes: Local linear fitting approach, J. Nonparametr. Stat. **16**, 443-462 (2004)

Caiado, J., Crato, N. and Peña, D. A periodogram-based metric for time series classification, Comput. Statist. Data Anal. **50**, 2668-2684 (2006)

Alonso, A.M., Berrendero, J.R., Hernández, A., Justel, A. Time series clustering based on forecast densities, Comput. Statist. Data Anal. **51**, 762–776 (2006)

Corduas, M., Piccolo, D. Time series clustering and classification by the autoregressive metric, Comput. Statist. Data Anal. **52**, 1860–1872 (2008)

Scotto, M., Alonso, A.M., Barbosa, S. Clustering time series of sea levels: an extreme value approach, J. Waterway, Port, Coastal, and Ocean Engineering **136**, 215–225 (2010)

Dissimilarity measures

Conceptually most of the dissimilarity criteria proposed for time series clustering lead to a notion of similarity relying on two possible criteria:

- Proximity between raw series data
- Proximity between underlying generating processes

In both cases, the classification task becomes inherently univariate since similarity searching is governed only by the behavior of each series but doesn't take into account the cross-dependency among the series.

Dissimilarity measures

Conceptually most of the dissimilarity criteria proposed for time series clustering lead to a notion of similarity relying on two possible criteria:

- Proximity between raw series data
- Proximity between underlying generating processes

In both cases, the classification task becomes inherently univariate since similarity searching is governed only by the behavior of each series but doesn't take into account the cross-dependency among the series.

A dissimilarity measure based on mutual dependency

Suppose that we have stationary (standardized) time series.

Define $r_{xx}(h) = E(x_{it}x_{i,t-h})$ and $r_{xy}(h) = E(x_{it}y_{j,t-h})$.

We can build a measure of the dependency as follows:

- Let $\mathbf{B}(h) = \begin{bmatrix} r_{xx}(h) & r_{xy}(h) \\ r_{yx}(h) & r_{yy}(h) \end{bmatrix}$.
- Then the matrix

$$\mathbf{B}_k = \begin{bmatrix} \mathbf{B}(0) & \mathbf{B}(1) & \cdots & \mathbf{B}(k) \\ \mathbf{B}(-1) & \mathbf{B}(0) & \cdots & \mathbf{B}(k-1) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{B}(-k) & \mathbf{B}(-k+1) & \cdots & \mathbf{B}(0) \end{bmatrix}$$

is the covariance matrix of the vector stationary process $\mathbf{Z}_t = (x_t, y_t, x_{t-1}, y_{t-1}, ..., x_{t-k}, y_{t-k})^T$.

A dissimilarity measure based on mutual dependency

A convenient measure of dissimilarity based on their joint dependency is

$$S_{xy}(k) = |\mathbf{B}_k|^{1/2(k+1)}$$

- Notice that $0 \leq |\mathbf{B}_k| \leq 1$ with equality to one when $\mathbf{B}_k$ is diagonal.
- This measure will be non-negative, symmetric and will be zero if $x = y$.
- The dissimilarity will reach the largest value, one, when the two series are independent, and will be zero if they are identical.

Our approach

Objective

To present a new way to find clusters in large vectors of time series.

Procedure

The procedure has two steps:

- In the first, series are split by their dependency, which is measured by the determinant of their correlation matrix including lags.
- Second, inside each groups the series are split by putting together series with a similar autocorrelacion structure.

The clustering procedure: General description

First step

We use the dissimilarity defined by

$$PRM(X, Y) = |\mathbf{B}_k|$$

as input of an agglomerative hierarchical clustering.

The clustering procedure: General description

A problem with $|\mathbf{B}_k|$ for integrated series is that it will be always close to zero, even if the series are independent.

Note that

$$|\mathbf{B}_k| = |\mathbf{B}_{k-1}| \left(1 - R^2_{k,k+1}(x_t/y_t, x_{t-1}, ..., y_{t-k})\right),$$

where $R^2_{k,k+1}$ is the square of the multiple correlation coefficient in the linear fit $\widehat{x}_t = \sum_{j=1}^{k} b_j x_{t-j} + \sum_{j=0}^{k} c_j y_{t-j}$, and this correlation coefficient will be close to one for integrated series.

The clustering procedure: General description

Considering the vector of variables $(x_t, x_{t-1}, ..., x_{t-k}, y_t, y_{t-1}, ..., y_{t-k})$, we can separate the univariate results from x to the dependence between x and y.

Thus, we can write

$$\mathbf{B}_k = \begin{bmatrix} \mathbf{R}(x)_k & \mathbf{R}(x,y)_k \\ \mathbf{R}(y,x)_k & \mathbf{R}(y)_k \end{bmatrix}$$

where $\mathbf{R}(x)_k$ and $\mathbf{R}(y)_k$ are the $(k+1) \times (k+1)$ correlation matrices of series x and y, respectively, and $\mathbf{R}(x,y)_k$ includes the dependence between x and y.

The clustering procedure: General description

Note that

$$|\mathbf{B}_k| = \left|\mathbf{R}(x)_k\right| \left|\mathbf{R}(y)_k - \mathbf{R}(y,x)_k \mathbf{R}^{-1}(x)_k \mathbf{R}(x,y)_k\right|$$

It should be noticed that if x is integrated then $|\mathbf{R}(x)_k|$ will be close to zero and the product will be small whatever the second term is.

This suggest the *alternative measure*

$$PRR(X,Y) = |\mathbf{B}_k| \,/\, (|\mathbf{R}(x)_k| \cdot |\mathbf{R}(y)_k|),$$

which has not this limitation.

The clustering procedure: General description

First step - bis

We use the dissimilarity defined by

$$PRR(X, Y) = |\mathbf{B}_k| / (|\mathbf{R}(x)_k| \cdot |\mathbf{R}(y)_k|)$$

as input of an agglomerative hierarchical clustering.

The clustering procedure: General description

Second step

- To obtain vectors of autocorrelation measures for each time series.
- To calculate an appropriate distance among these vectors.
- To use those distances as input of an agglomerative hierarchical clustering.

The clustering procedure: General description

Second step

What is the meaning of autocorrelation measures?

- SAC (Simple AutoCorrelations).
- PAC (Partial AutoCorrelations).
- PRG (PR Global), $PR_k(x) = -\frac{T}{k+1} \log |\mathbf{R}(x)_K|$ proposed in Peña and Rodríguez (2006) as a goodness-of-fit test.
- PRV (PR Vector) that calculates the PR statistics for $k = 1, 2, \ldots, K$.
- PRD (PR Differences) which is similar to PRV but it calculates the differences of the consecutive measures.
- QAC (Quantile AutoCovariances) as in Lafuente-Rego and Vilar (2015).

The clustering procedure: General description

The nonlinear features of some time series, as for instance, volatility and nonlinear behavior are not indicated by the measures SAC, PAC or PR types statistics.

We know that these nonlinear features can be shown by the autocorrelation of the absolute values or the squared residuals of a linear fit.

Suppose that we fit an AR(p) model to the series where p is chosen by the AIC or BIC criterion and we obtain:

$$e_t = y_t - \widehat{\pi}_1 y_{t-1} - \ldots - \widehat{\pi}_p y_{t-p}.$$

The clustering procedure: General description

Then the vector of autocorrelation $\mathbf{r}'_S = (r_1(e_t^2), \ldots, r_k(e_t^2))$ or $\mathbf{r}'_A = (r_1(|e_t|), \ldots, r_k(|e_t|))$ may be useful to detect nonlinear effects.

Thus, we propose to have a vector of features $(\mathbf{r}', \mathbf{r}'_N)$ where $\mathbf{r}'_N$=($\mathbf{r}'_S$ or $\mathbf{r}'_A$) of dimension $2k$ and use these vectors to find clusters.

Similar ideas can be used with PAC, PRG, PRD and PRD.

Simulation study: Motivation

Since our procedure has two steps, then the simulation study will have two parts:

1. We will evaluate the performance of measures SAC, PAC, PRG, PRD and PRD in clustering independent series.
2. We will evaluate the performance of the dissimilarity measures based on mutual dependency, PRM and PRR, in clustering dependent series.

Simulation study: Independent scenarios

Set A - Independent series

- Scenario A.1: Gaussian white noise vs AR(1) model with $\phi = 0.5$.
- Scenario A.2: Gaussian white noise vs GARCH(1,1) model with $\omega = 0.1$, $\alpha = 0.7$ and $\beta = 0.2$.
- Scenario A.3: AR(1) model with $\phi = 0.5$ vs AR(1)-GARCH(1,1) model with $\phi = 0.5$, $\omega = 0.1$, $\alpha = 0.7$ and $\beta = 0.2$.

For the three scenarios, we generate independent fifteen series of length $T = 100$ and 200 for each group.

Scenarios, A and B, was used by Lafuente-Rego and Vilar (2015).

Simulation study: Independent scenarios

Set B - Independent series

- Scenario B.1: ARMA processes classification.
  1. AR(1) $X_t = 0.9X_{t-1} + \epsilon_t$
  2. MA(1) $X_t = -0.7\epsilon_{t-1} + \epsilon_t$
  3. AR(2) $X_t = 0.3X_{t-1} - 0.1X_{t-2} + \epsilon_t$
  4. MA(2) $X_t = 0.8\epsilon_{t-1} - 0.6\epsilon_{t-1} + \epsilon_t$
  5. ARMA(1,1) $X_t = 0.8X_{t-1} + 0.2\epsilon_{t-1} + \epsilon_t$
  6. I(1) $X_t = X_{t-1} + \epsilon_t$
- Scenario B.2: Non-linear processes classification.
  1. TAR $X_t = 0.5X_{t-1}I(X_{t-1} \leq 0) - 2X_{t-1}I(X_{t-1} > 0) + \epsilon_t$
  2. EXPAR $X_t = [0.3 - 10\exp(-X_{t-1}^2)]X_{t-1} + \epsilon_t$
  3. MA $X_t = -0.7\epsilon_{t-1} + \epsilon_t$
  4. NLMA $X_t = -0.5\epsilon_{t-1} + 0.8\epsilon_{t-1}^2 + \epsilon_t$

Simulation study: Independent scenarios

Set B - Independent series

- Scenario B.3: Conditional heteroscedastic processes classification.
  1. ARCH $\sigma_t^2 = 0.1 + 0.8a_{t-1}^2$
  2. GARCH $\sigma_t^2 = 0.1 + 0.1a_{t-1}^2 + 0.8\sigma_{t-1}^2$
  3. GJR-GARCH
     $\sigma_t^2 = 0.1 + [0.25 + 0.3I(a_{t-1} < 0)]a_{t-1}^2 + 0.5\sigma_{t-1}^2$
  4. EGARCH
     $\ln(\sigma_t^2) = 0.1 + 0.3\epsilon_{t-1} + 0.3\left[|\epsilon_{t-1}| - E(|\epsilon_{t-1}|)\right] + 0.4\ln(\sigma_{t-1}^2)$

  where $a_t = \sigma_t\epsilon_t$ and $X_t = 0.2a_{t-1} + a_t$.

For the three scenarios, we generate five independent series of each type having length $T = 100$ and 200.

Simulation study: Results

The following results are the means of the Gravilov index from 10000 replicates for the sets A and B with $T = 100$.

The similarity index used in Gavrilov et al. (2000) compares two different cluster partitions, $C = (C_1, \ldots, C_k)$ and $C' = (C'_1, \ldots, C'_{k'})$ using the following formulas:

$$Sim(C_i, C'_j) = 2\frac{\#(C_i \bigcap C'_j)}{\#(C_i) + \#(C'_j)},$$

and

$$Sim(C, C') = k^{-1} \sum_{i=1}^{k} \max_{1 \le j \le k'} Sim(C_i, C'_j).$$

The closer to one the index, the higher is the agreement between the two partitions.

Simulation study: Results

Scenarios A

Method	A.1	A.2	A.3
QAC	**0.950**	0.560	0.556
SAC	*0.846*	**0.687**	**0.635**
PAC	**0.936**	0.582	0.578
PRG	0.785	0.528	0.505
PRV	0.774	0.529	0.505
PRD	0.763	0.510	0.494

Main conclusions

- The results for the first scenario, A.1, are similar for the first three approaches.
- The PR based methods are slightly worse.
- The results point out that SAC have the better results for the second and third scenario, A.2 and A.3.
- It should be notice that three PR approaches have similar results and improved by PAC.

Simulation study: Results

Scenarios B

Method	B.1	B.2	B.3
QAC	**0.657**	**0.884**	0.472
SAC	**0.679**	0.681	0.475
PAC	**0.670**	*0.701*	0.459
PRG	0.509	0.555	0.447
PRV	0.505	0.562	0.447
PRD	0.497	0.576	0.436

Main conclusions

- The results for the first scenario, B.1, points out that PAC and SAC have the better performances.
- The results for the second scenario, B.2, show that QAC improves the other methods.
- The results for the third scenario, B.3, are similar for all approaches.
- As in set A, the PR approaches have similar results and are improved by PAC.

Simulation study: Dependent scenarios

Set D - Dependent series

- The models for the three populations are:
  1. AR(1) $X_t^{(1,i)} = 0.9X_{t-1}^{(1,i)} + \epsilon_t^{(1,i)}$ with $i = 1, 2, ..., 5$.
  2. AR(1) $X_t^{(2,i)} = 0.2X_{t-1}^{(2,i)} + \epsilon_t^{(2,i)}$ with $i = 1, 2, ..., 5$.
  3. AR(1) $X_t^{(3,i)} = 0.2X_{t-1}^{(3,i)} + \epsilon_t^{(3,i)}$ with $i = 1, 2, ..., 5$.

  That is, the second and the third models have the same autocorrelation structure.
- The five scenarios differs in the dependence structure of the innovations. In the following, we present the autocorrelation matrices of $(\epsilon_t^{(1,1)}, \epsilon_t^{(1,2)}, ..., \epsilon_t^{(3,5)})$.

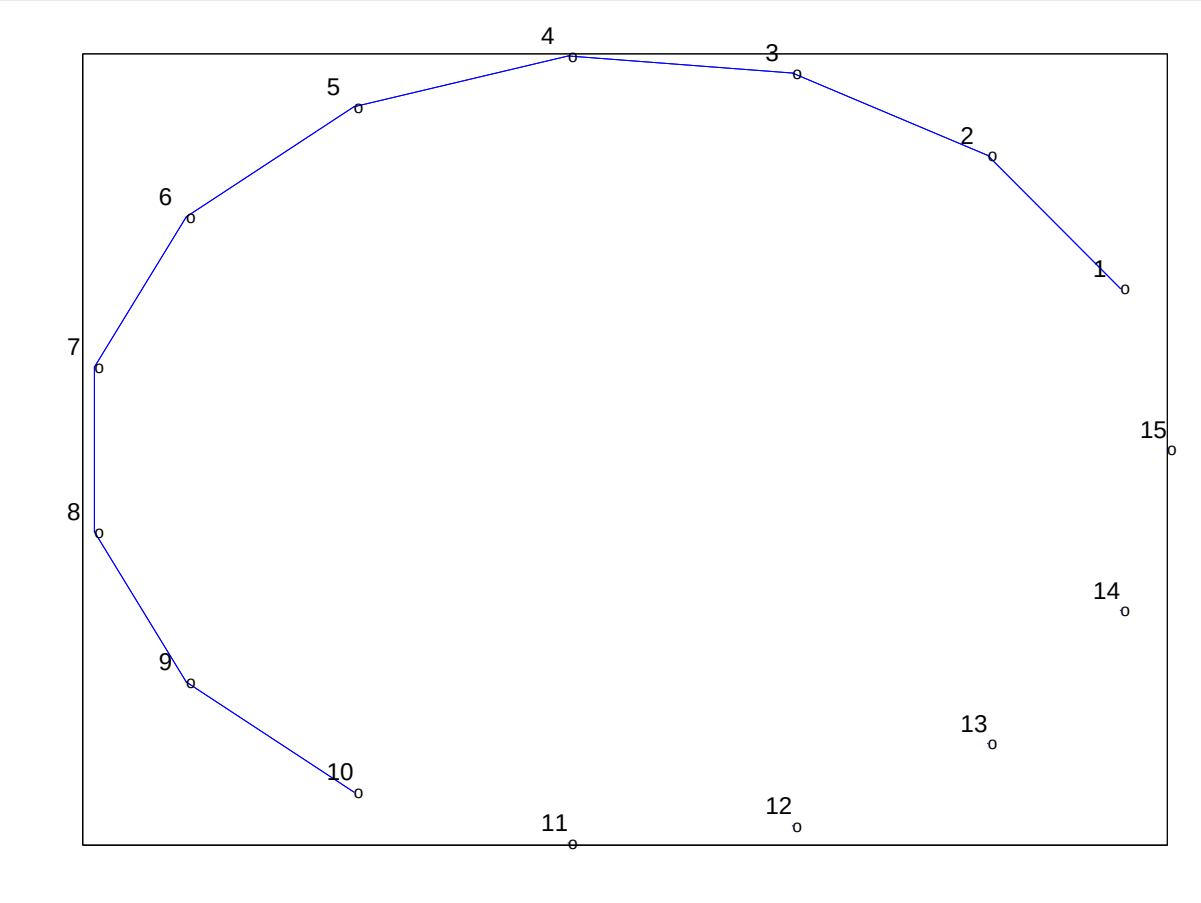
Simulation study: Dependent scenarios

Scenario D.1

$$R_{D.1} = \begin{pmatrix} 1 & .5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ & 1 & .5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & 1 & .5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & 1 & .5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & & 1 & .5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & & & 1 & .5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & & & & 1 & .5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & & & & & 1 & .5 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & & & & & & 1 & .5 & 0 & 0 & 0 & 0 & 0 \\ & & & & & & & & & 1 & 0 & 0 & 0 & 0 & 0 \\ & & & & & & & & & & 1 & 0 & 0 & 0 & 0 \\ & & & & & & & & & & & 1 & 0 & 0 & 0 \\ & & & & & & & & & & & & 1 & 0 & 0 \\ & & & & & & & & & & & & & 1 & 0 \\ & & & & & & & & & & & & & & 1 \end{pmatrix}$$

Simulation study: Dependent scenarios

Scenario D.1



Simulation study: Dependent scenarios

Scenario D.2

$$R_{D.2} = \begin{pmatrix}
1 & .5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 & 1 & .5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 & & 1 & .5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 & & & 1 & .5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 & & & & 1 & .5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & 1 & .5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & & 1 & .5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & & & 1 & .5 & 0 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & & & & 1 & .5 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & & & & & 1 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & & & & & & 1 & .5 & 0 & 0 & 0 \\
 & & & & & & & & & & & 1 & .5 & 0 & 0 \\
 & & & & & & & & & & & & 1 & .5 & 0 \\
 & & & & & & & & & & & & & 1 & .5 \\
 & & & & & & & & & & & & & & 1
\end{pmatrix}$$

Simulation study: Dependent scenarios

Scenario D.3

$$R_{D.3} = \begin{pmatrix} 1 & .9 & .9 & .9 & .9 & .9 & .9 & .9 & .9 & .9 & 0 & 0 & 0 & 0 & 0 \\ & 1 & .9 & .9 & .9 & .9 & .9 & .9 & .9 & .9 & 0 & 0 & 0 & 0 & 0 \\ & & 1 & .9 & .9 & .9 & .9 & .9 & .9 & .9 & 0 & 0 & 0 & 0 & 0 \\ & & & 1 & .9 & .9 & .9 & .9 & .9 & .9 & 0 & 0 & 0 & 0 & 0 \\ & & & & 1 & .9 & .9 & .9 & .9 & .9 & 0 & 0 & 0 & 0 & 0 \\ & & & & & 1 & .9 & .9 & .9 & .9 & 0 & 0 & 0 & 0 & 0 \\ & & & & & & 1 & .9 & .9 & .9 & 0 & 0 & 0 & 0 & 0 \\ & & & & & & & 1 & .9 & .9 & 0 & 0 & 0 & 0 & 0 \\ & & & & & & & & 1 & .9 & 0 & 0 & 0 & 0 & 0 \\ & & & & & & & & & 1 & 0 & 0 & 0 & 0 & 0 \\ & & & & & & & & & & 1 & 0 & 0 & 0 & 0 \\ & & & & & & & & & & & 1 & 0 & 0 & 0 \\ & & & & & & & & & & & & 1 & 0 & 0 \\ & & & & & & & & & & & & & 1 & 0 \\ & & & & & & & & & & & & & & 1 \end{pmatrix}$$

Simulation study: Dependent scenarios

Scenario D.4

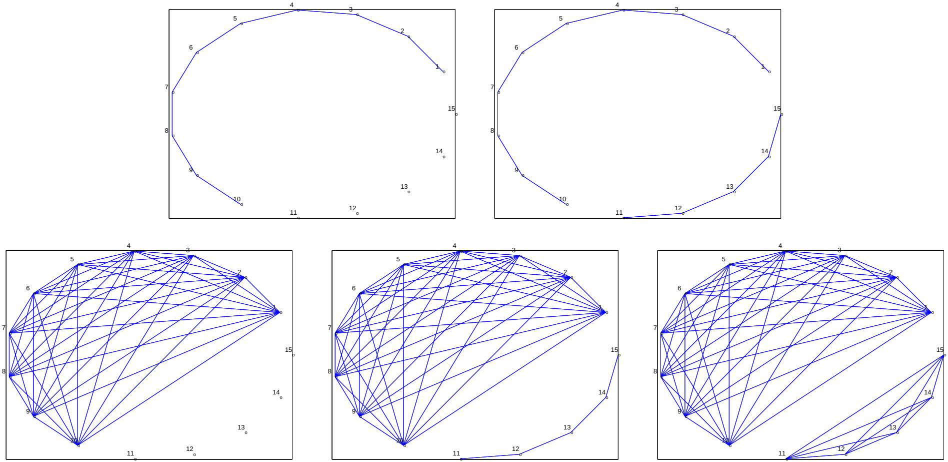
$$R_{D.4} = \begin{pmatrix}
1 & .9 & .9 & .9 & .9 & .9 & .9 & .9 & .9 & .9 & 0 & 0 & 0 & 0 & 0 \\
 & 1 & .9 & .9 & .9 & .9 & .9 & .9 & .9 & .9 & 0 & 0 & 0 & 0 & 0 \\
 & & 1 & .9 & .9 & .9 & .9 & .9 & .9 & .9 & 0 & 0 & 0 & 0 & 0 \\
 & & & 1 & .9 & .9 & .9 & .9 & .9 & .9 & 0 & 0 & 0 & 0 & 0 \\
 & & & & 1 & .9 & .9 & .9 & .9 & .9 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & 1 & .9 & .9 & .9 & .9 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & & 1 & .9 & .9 & .9 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & & & 1 & .9 & .9 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & & & & 1 & .9 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & & & & & 1 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & & & & & & 1 & .5 & 0 & 0 & 0 \\
 & & & & & & & & & & & 1 & .5 & 0 & 0 \\
 & & & & & & & & & & & & 1 & .5 & 0 \\
 & & & & & & & & & & & & & 1 & .5 \\
 & & & & & & & & & & & & & & 1
\end{pmatrix}$$

Simulation study: Dependent scenarios

Scenario D.5

$$R_{D.5} = \begin{pmatrix}
1 & .9 & .9 & .9 & .9 & .9 & .9 & .9 & .9 & .9 & 0 & 0 & 0 & 0 & 0 \\
 & 1 & .9 & .9 & .9 & .9 & .9 & .9 & .9 & .9 & 0 & 0 & 0 & 0 & 0 \\
 & & 1 & .9 & .9 & .9 & .9 & .9 & .9 & .9 & 0 & 0 & 0 & 0 & 0 \\
 & & & 1 & .9 & .9 & .9 & .9 & .9 & .9 & 0 & 0 & 0 & 0 & 0 \\
 & & & & 1 & .9 & .9 & .9 & .9 & .9 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & 1 & .9 & .9 & .9 & .9 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & & 1 & .9 & .9 & .9 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & & & 1 & .9 & .9 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & & & & 1 & .9 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & & & & & 1 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & & & & & & 1 & .9 & .9 & .9 & .9 \\
 & & & & & & & & & & & 1 & .9 & .9 & .9 \\
 & & & & & & & & & & & & 1 & .9 & .9 \\
 & & & & & & & & & & & & & 1 & .9 \\
 & & & & & & & & & & & & & & 1
\end{pmatrix}$$

Simulation study: Scenarios D.1 - D.5



Simulation study: Results

The following results are the means of the Gravilov index from 10000 replicates for the set D using the complete and single linkage

Method	D.1	D.2	D.3	D.4	D.5
QAC	0.4911	0.6657	0.5954	0.6654	0.6666
SAC	0.4425	0.6433	0.7167	0.6651	0.6648
PAC	0.4914	0.6657	0.8140	0.6784	0.6885
PRM	0.6984	0.6639	**1.0000**	**0.8420**	**1.0000**
PRR	0.5269	0.6540	**1.0000**	**0.8654**	**1.0000**

Method	D.1	D.2	D.3	D.4	D.5
QAC	0.5055	0.6642	0.6663	0.6657	0.6663
SAC	0.4447	0.6465	0.8178	0.6655	0.6661
PAC	0.4845	0.6430	0.8580	0.6611	0.6751
PRM	0.6340	0.5261	**1.0000**	0.5833	0.9022
PRR	**0.9414**	**0.8315**	**1.0000**	**0.9602**	**1.0000**

Simulation study: Results

Main conclusions

- The results of the three univariate methods are similar and they don't change much across linkage methods.
  - Notice that here a Gravilov index around 0.667 corresponds to approximately separate the first population from the third one in scenarios D.2, D.4 and D.5
- For scenarios D.3, D.4 and D-5 where there are some "strong" clusters, the complete linkage for both multivariate measures improve the univariate measures.
- For all scenarios, the single linkage and PRR is preferable to other considered alternatives.

A case-study with real data: Data description

Spanish mortality rates

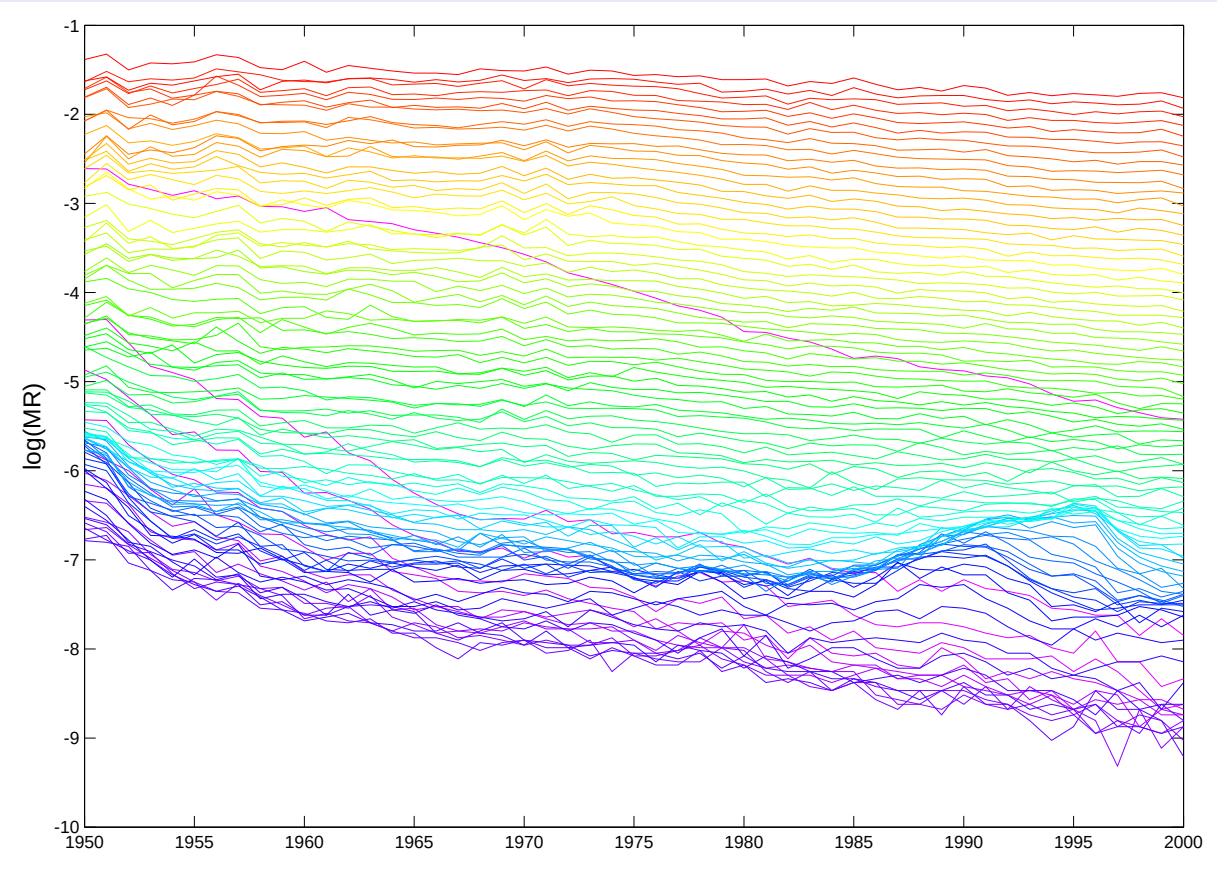
We consider the Spanish mortality rates by age (0 – 90 years) for both genders taken from the Human Mortality Database (http://www.mortality.org).

The data is available from 1908 to 2010. We skip the period 1908 – 1949.

This allows us to use the period 1950 – 2000 as a model adjustment period and 2001 – 2010 as a test period in the forecasting exercise.

A case-study with real data: Data description

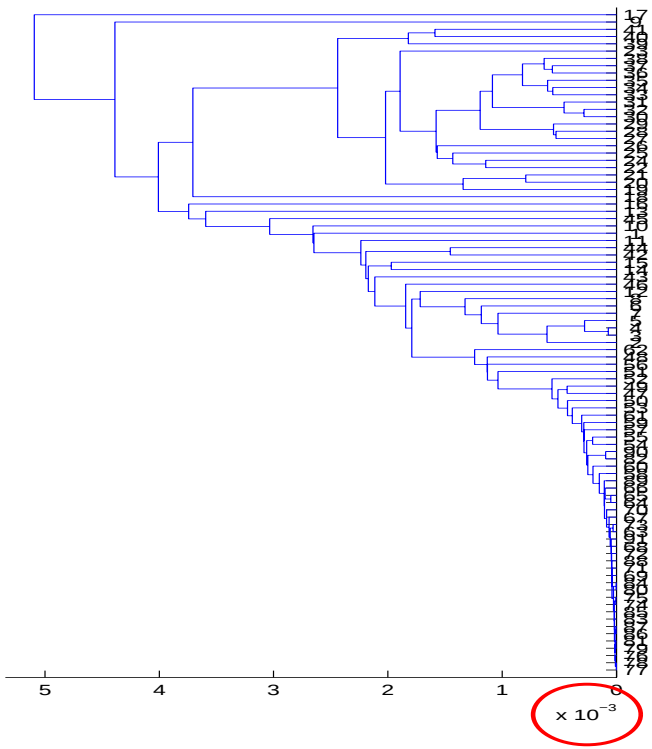
Spanish mortality rates

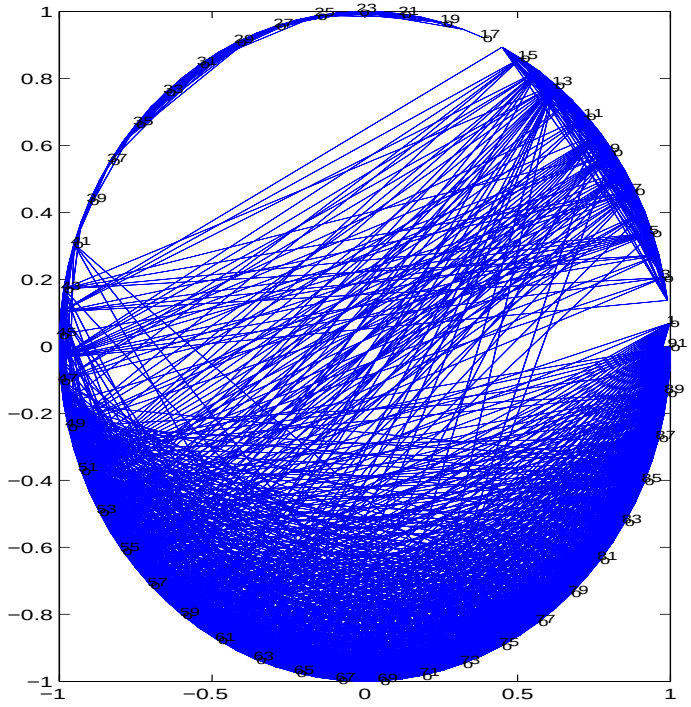


A case-study with real data: Data description

Spanish mortality rates

It is clear that these series has an strong negative trend. In fact they share a common trend.





A case-study with real data: Data description

Lee-Carter model

It is a well-known model which looks at the dependence between mortality time series. It relates the mortality rates by age with a single unobservable factor:

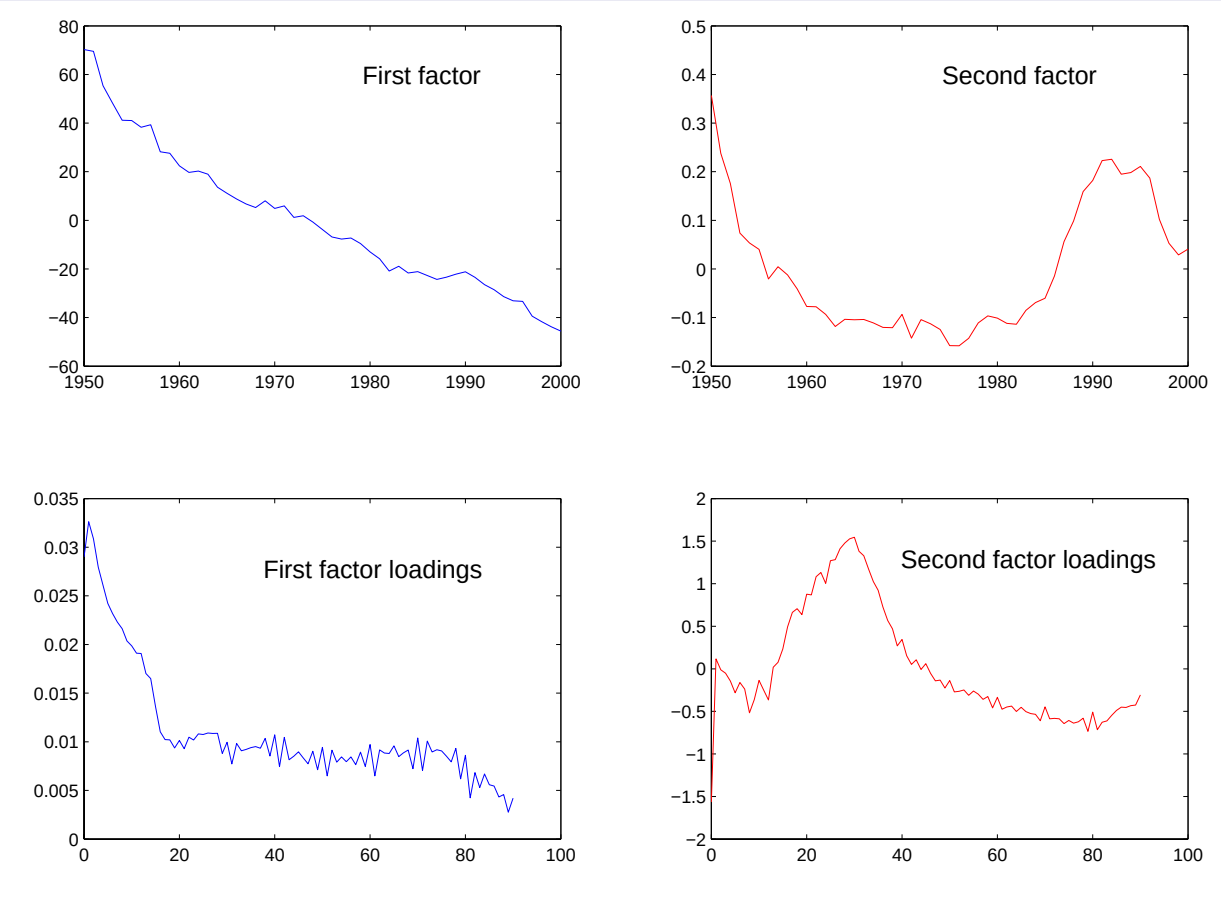
$$\begin{aligned} \ln(MR_{x,t}) &= a_x + b_x k_t + \varepsilon_{x,t} \\ k_t &= c + k_{t-1} + \eta_t \end{aligned} \quad ,$$

where a_x and b_x are parameters which depend on age, x; k_t is the unobservable factor which picks up the general characteristics of mortality in the year t, and $\varepsilon_{x,t}$ are the age-specific factors.

We will cluster the series of age-specific factors, $\varepsilon_{x,t}$.

A case-study with real data: Factors & Loadings

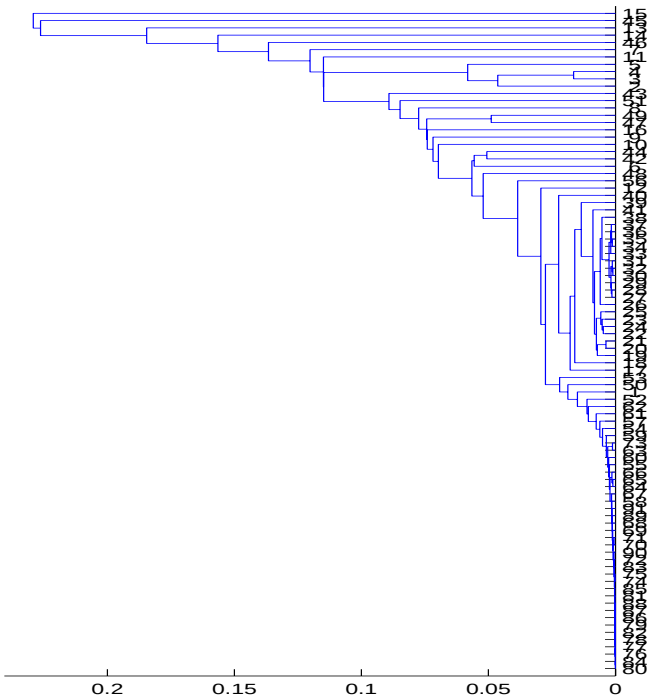
Spanish mortality rates

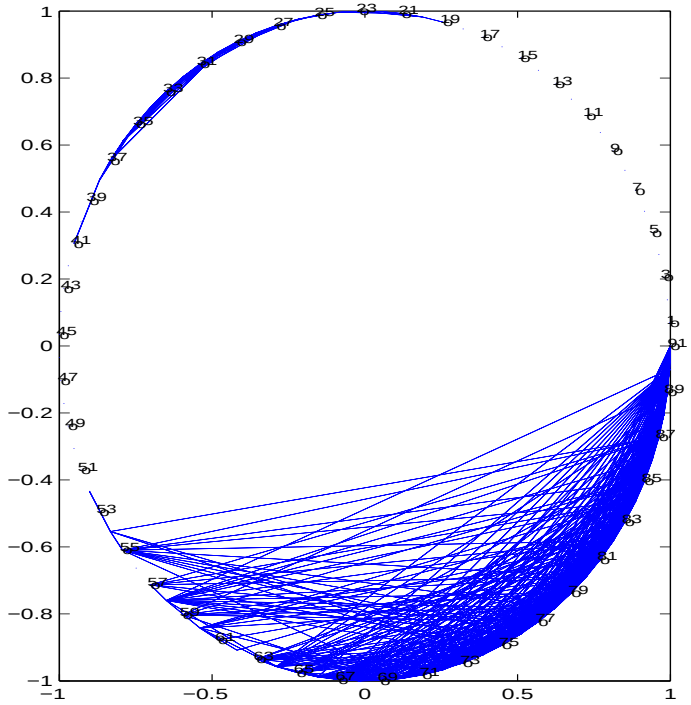


Spanish mortality rates: Clustering results

Spanish mortality rates

At the age-specific factors, we find two clusters and some “independent” series.





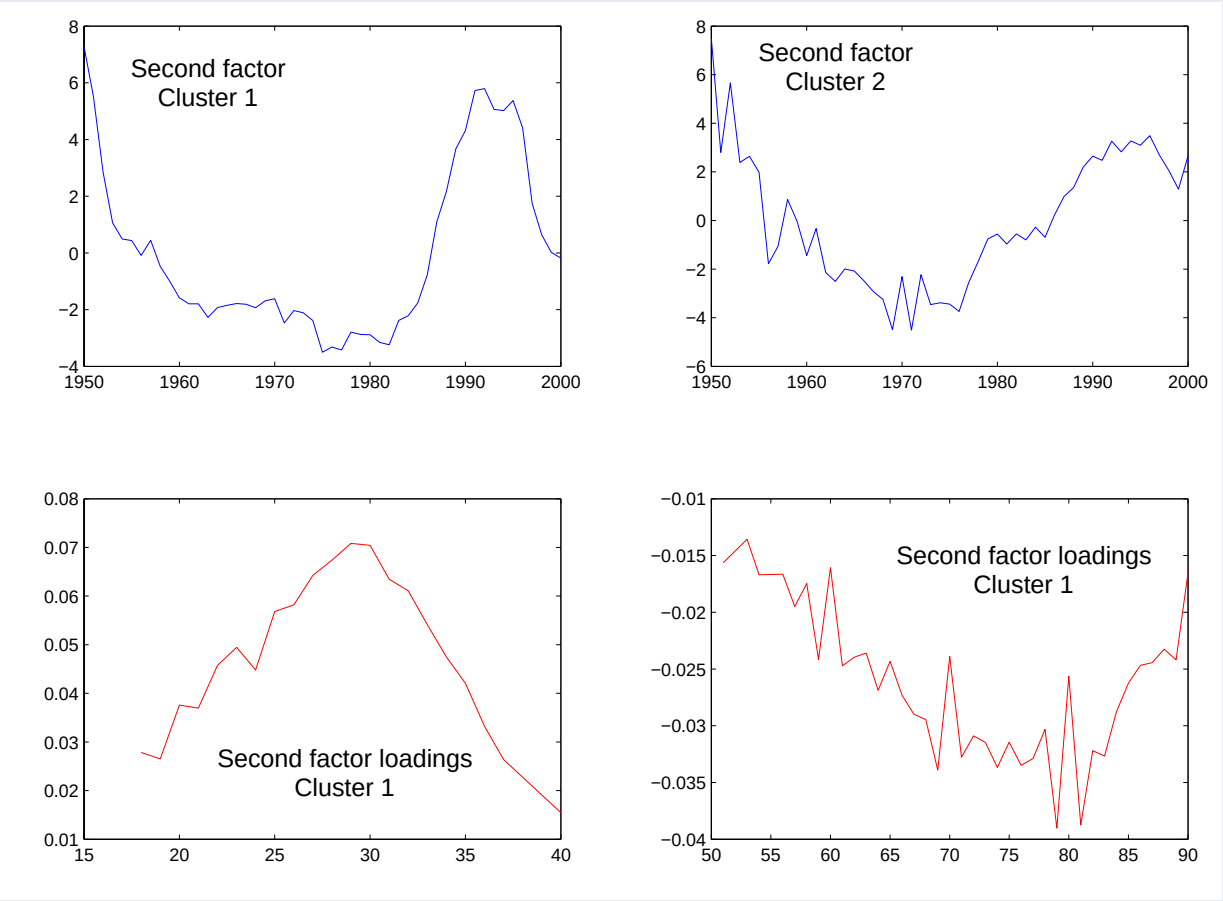
Spanish mortality rates: Clustering results

Here, we will compare the forecasting performance of three models:

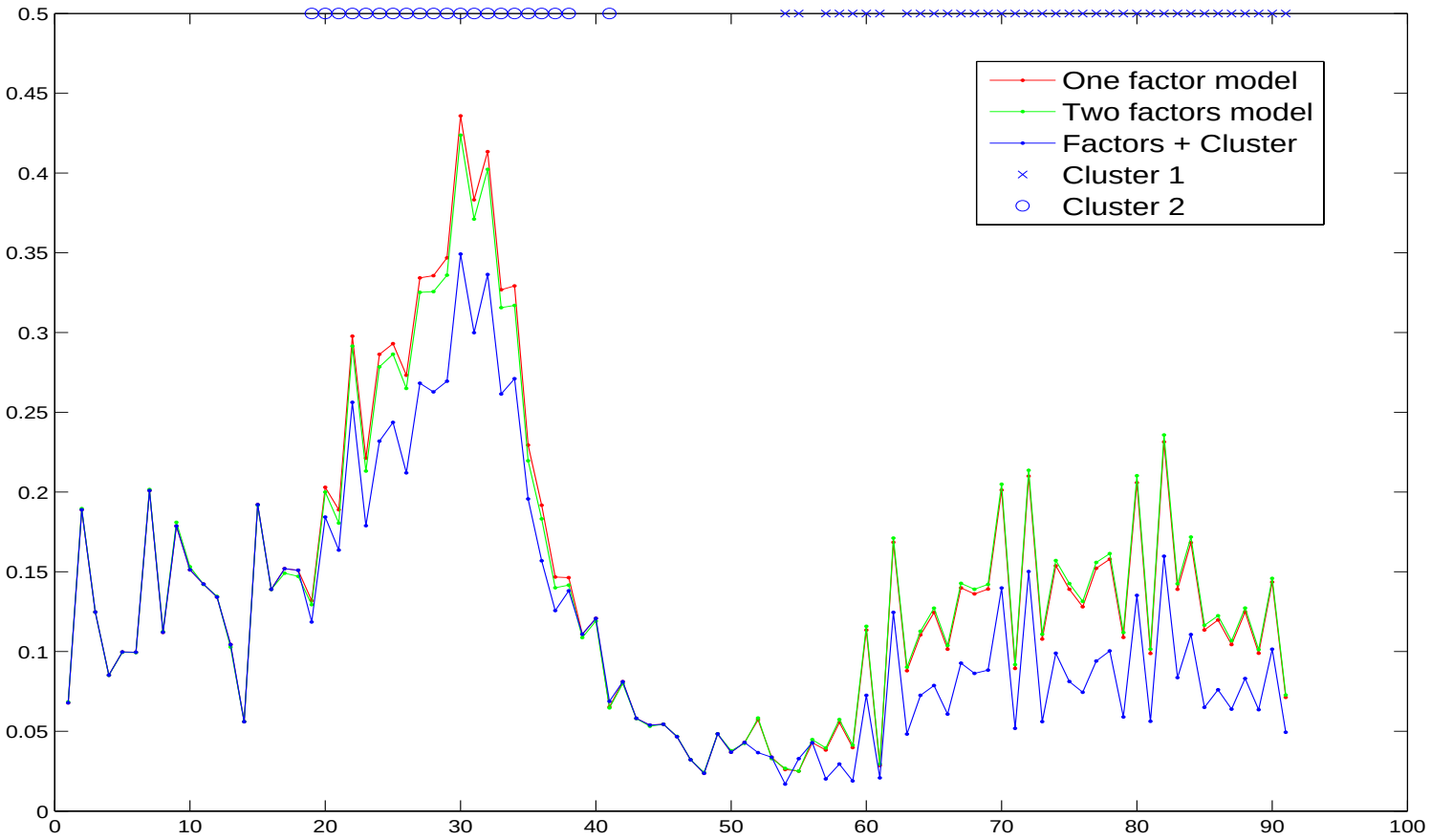
- A factorial model with a single unobservable factor, as in Lee-Carter (1992).
- A factorial model with two unobservable factors, as in Alonso, Peña and Rodríguez (2005).
- A factorial model with two unobservable factors where:
  - the first factor is estimated using all series.
  - the second factor is estimated using the two obtained clusters.

A case-study with real data: Factors & Loadings

Spanish mortality rates

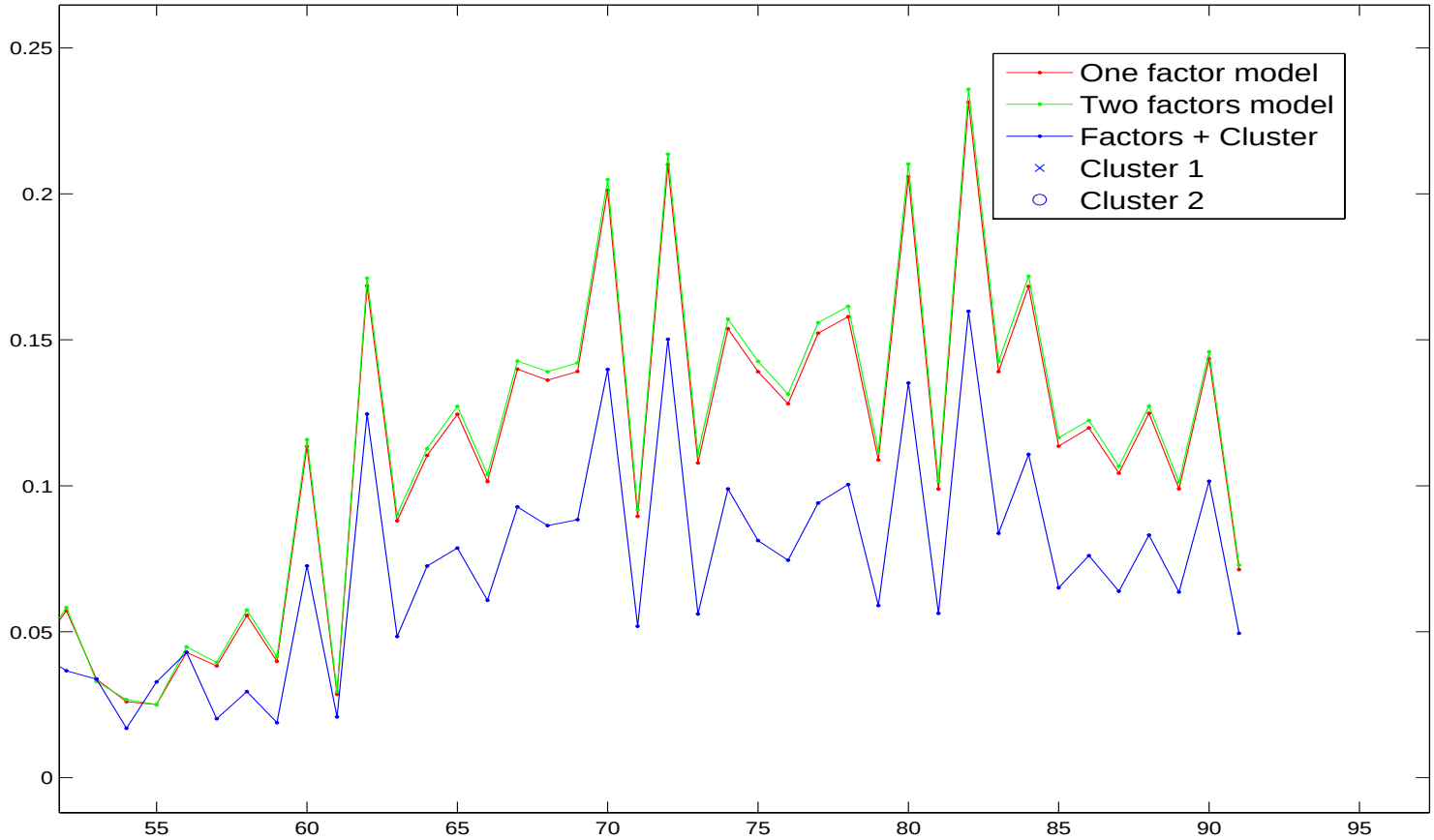


Mean absolute prediction errors



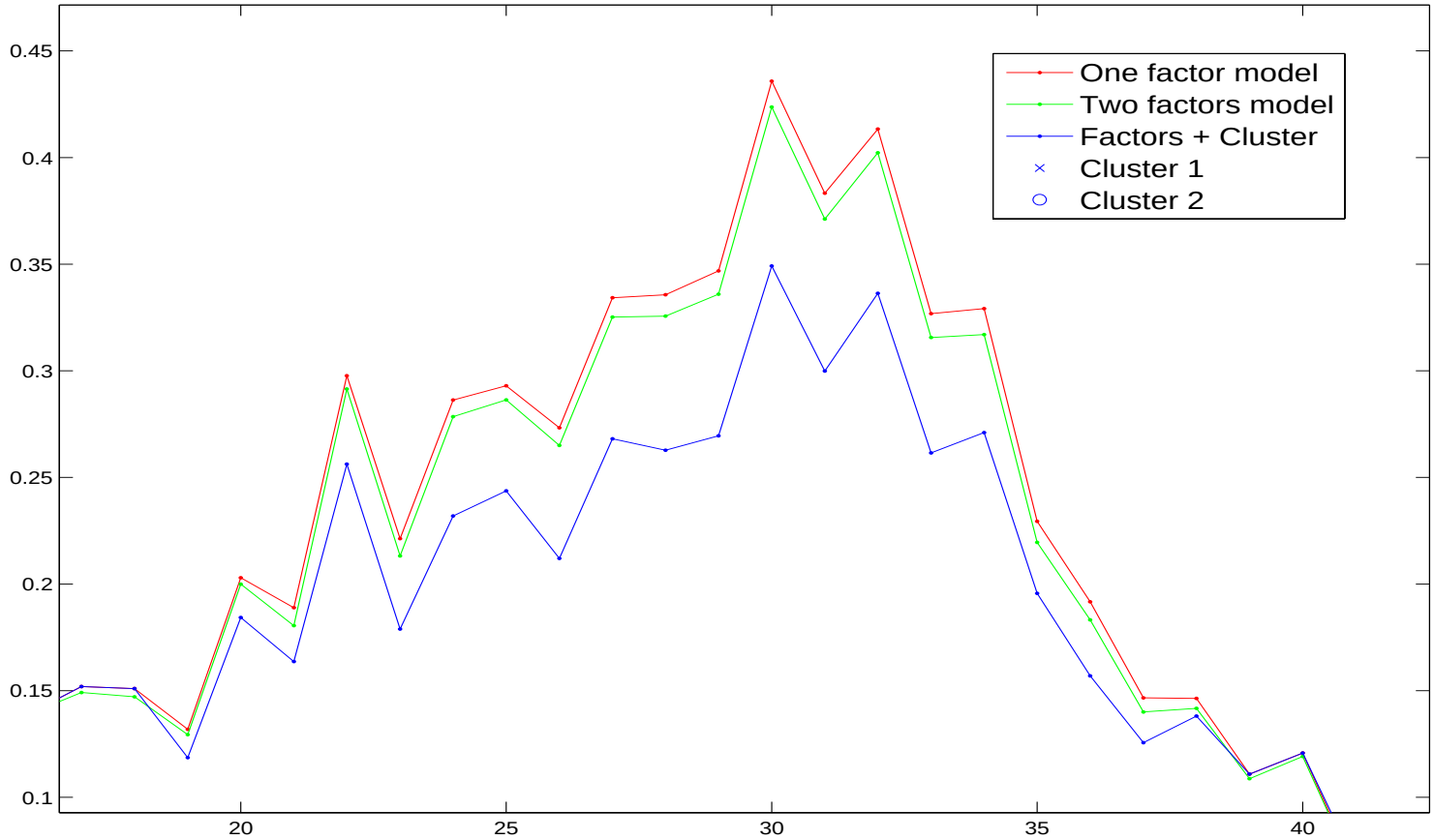
We observe improvements in almost all ages

Mean absolute prediction errors



We observe improvements in ages where two factors is worse than one factor

Mean absolute prediction errors

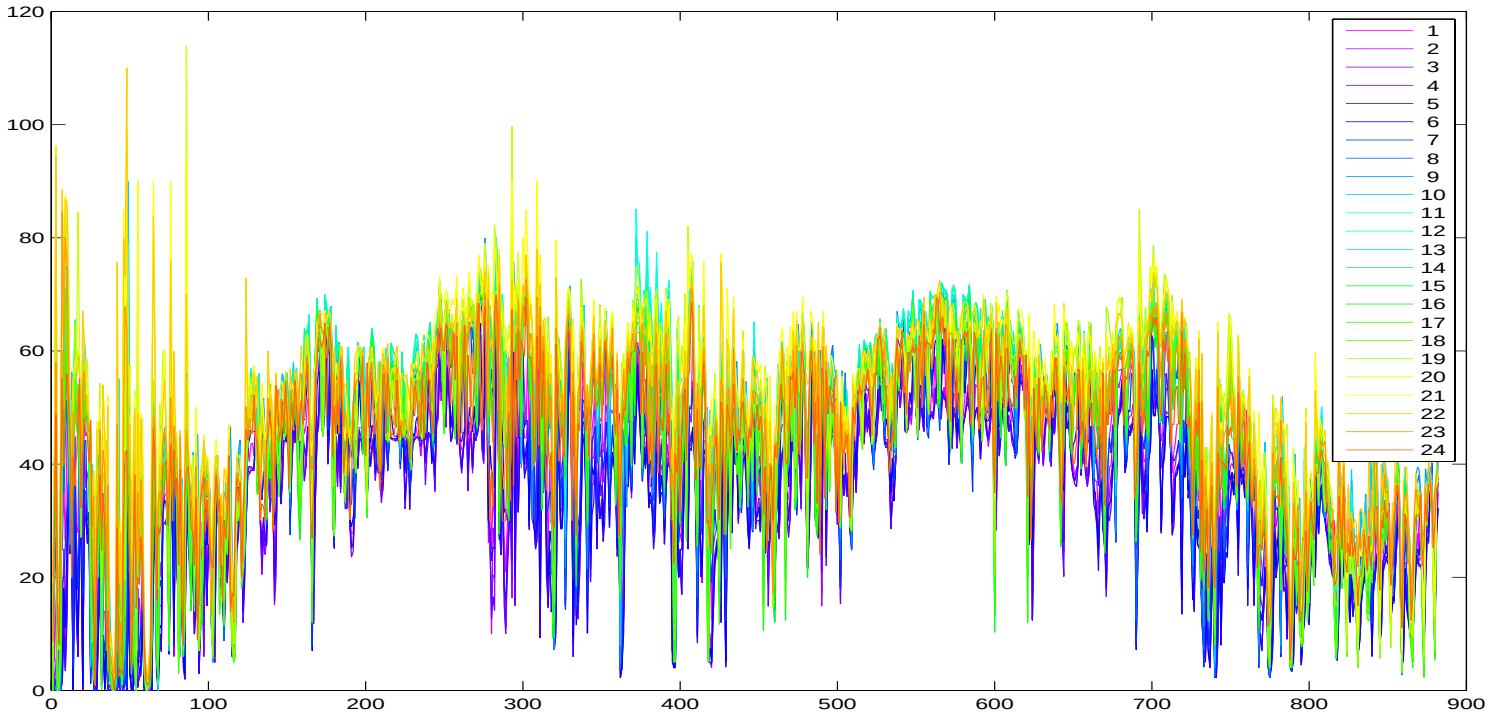


But also in ages where two factors is better than one factor

A case-study with real data: Data description

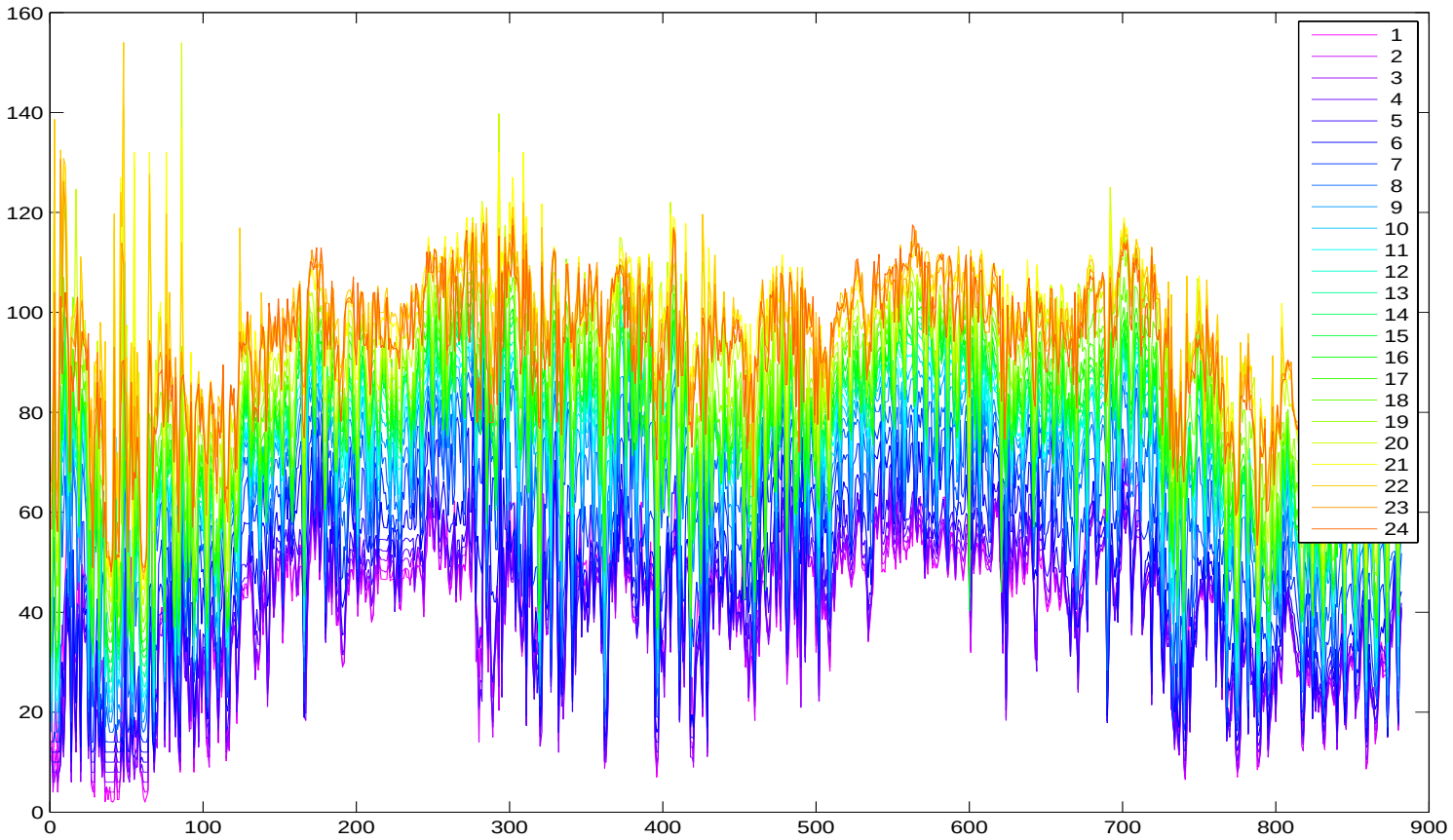
Spanish electricity prices

We study the 24 series of hourly prices for the Iberian electricity market from January 2014 to May 2016.



A case-study with real data: Data description

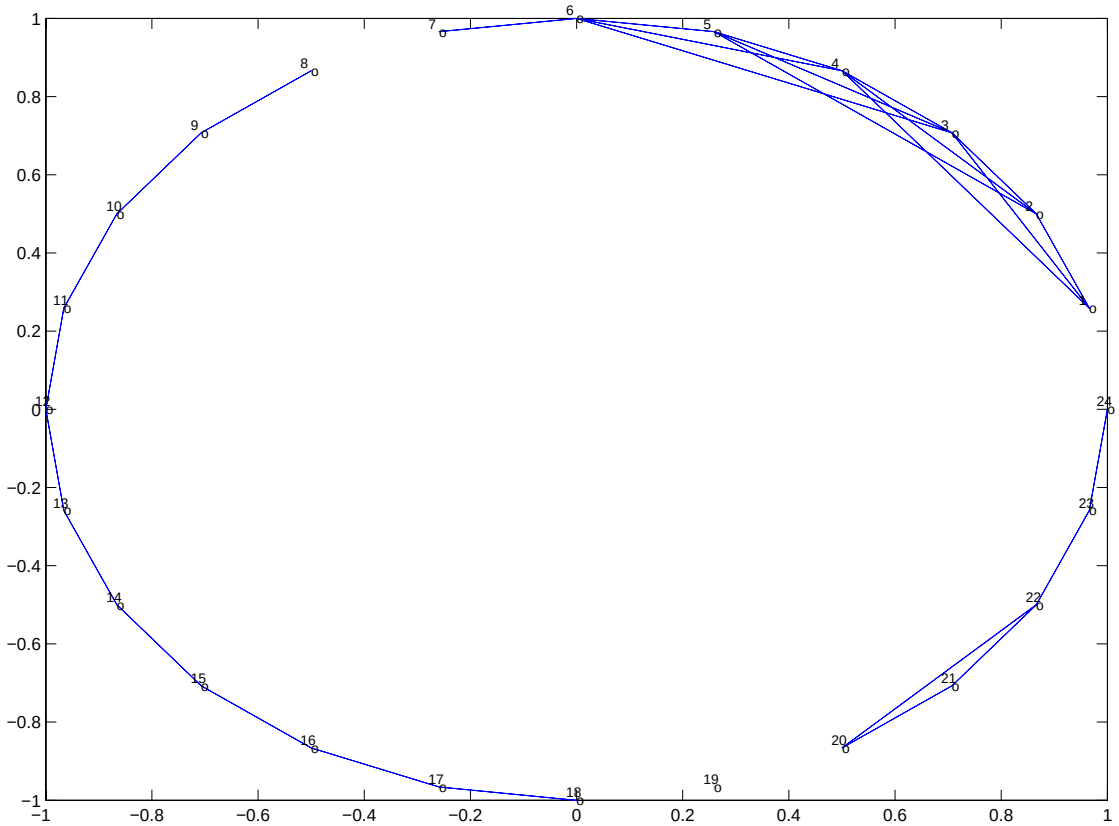
Spanish electricity prices - Translated for better visualization.



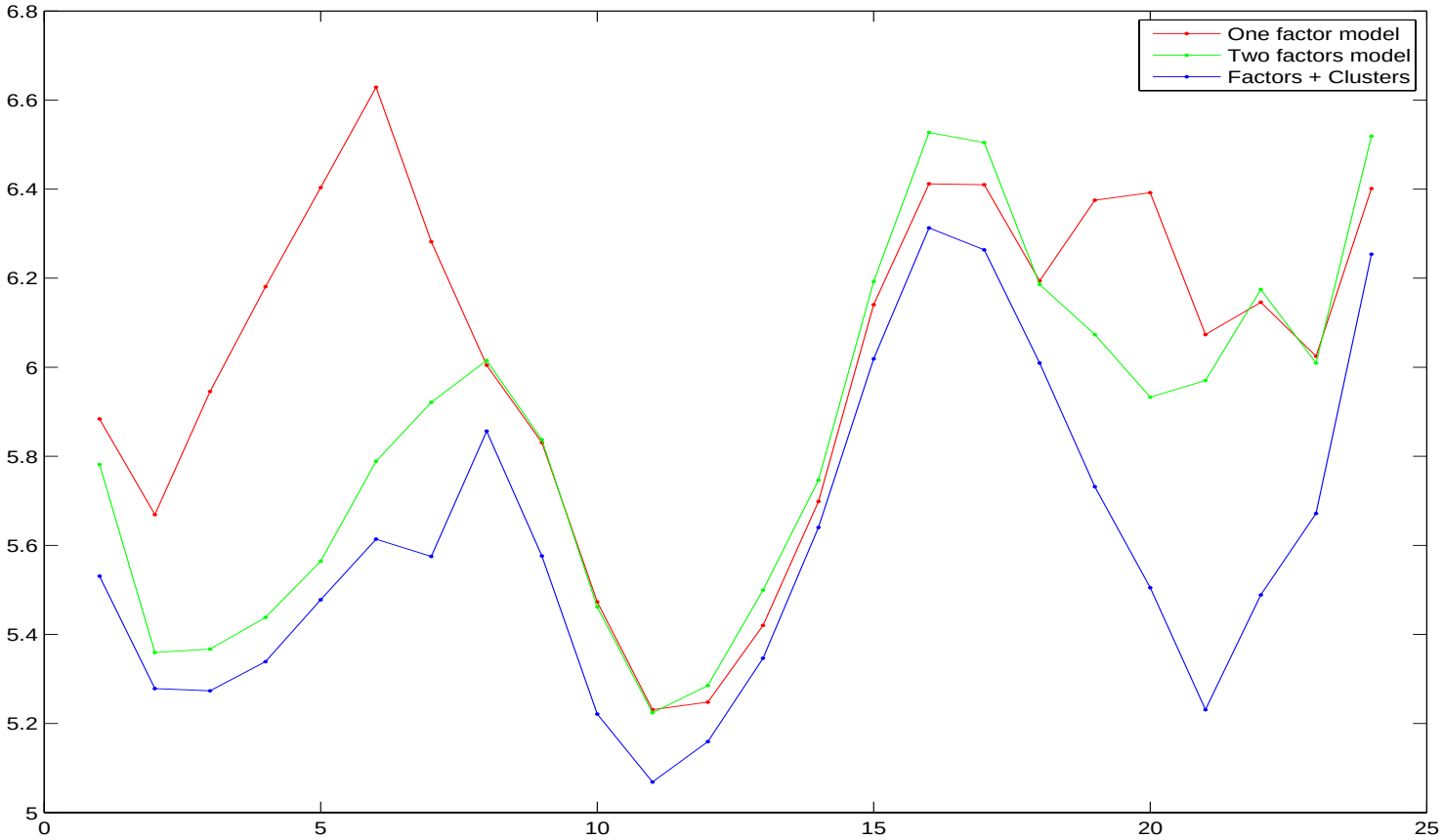
Spanish electricity prices: Clustering results

There are three clusters:

- Sleeping hours
- Working hours
- Arriving & staying at home.



Mean absolute prediction errors



We observe improvements in all hours for one-day-ahead forecast

Concluding remarks

- We propose a two-step clustering procedure for time series.
  - Based on (cross–)dependence measures.
  - Based on (auto–)dependence measures.
- The simulation studies point out that PRM and PRR have reasonable performance in clustering dependent time series and SAC, PAC and QAC in clustering independent time series.
- We find that the clustering procedure produces interpretable clusters and it helps us to improve forecasting in two real data examples.

Extensions

Clustering by predictability

"Weather prediction is ... the process of determining how the weather will change as time advances, and the problem of weather predictability becomes that of ascertaining whether such predictions are possible" Lorenz (1975)

- For AR(1) processes, DelSole (2004) derive the following measure of predictability:

$$M_k = -\frac{1}{2}\log(1-\phi^{2k})$$

for horizon k.

- $PR_k(x) = -\frac{T}{k+1}\log|\mathbf{R}(x)_K|$ can be interpreted as a weighted measure of predictability up to horizon k.

Extensions

Preliminary results

Method	Scenario S1	Scenario S2	Scenario S3
QAC	0.6372	0.6666	0.6674
SAC	0.6559	0.6667	0.6711
PAC	0.6561	0.6667	0.6667
PRG	0.7700	0.9996	0.9853
PRV	0.7634	0.9990	0.9937
PRD	0.7662	0.9994	0.9947


DelSole, T. Predictability and information theory. Part I: Measure of predictability, J. Atmos. Sci. **61**, 2425–2440 (2004a)


DelSole, T. Predictability and information theory. Part II: Part II: Imperfect Forecasts, J. Atmos. Sci. **62**, 3368–3381 (2004b)