Generalized Dynamic Principal Components

Daniel Peña and Victor Yohai

Universidad Carlos III de Madrid

and

Universidad de Buenos Aires

NBER/NSF Time Series Conference, Vienna
September 2015

Outline

1. Introduction
2. Generalyzed Dynamic Principal Components
3. Case k=1
4. Monte Carlo and Examples
5. Robust GDPC
6. Conclusions

Paper forthcoming in the *Journal of American Statistical Association*, 2016

1. Introduction

Dimension reduction is very important in vector time series because the number of parameters in a model grows with the square of the dimension m of the vector of time series.

Simplifying structures or *factors* reduce the number of parameters to model the series. Two related approaches:

1. Linear combinations with interesting properties:

Brillinger (1981), DPC; maximun/minimun predictabiliy-canonical analysis (Box and Tiao, 1977); Stationary combinations in nonstationary time series (Granger and Engle, 1987), the scalar component models, SCM, Tiao and Tsay (1989); white noise with stationary data (Ash and Reinsel, 1990), among others.

2. Dynamic Factor Models : Geweke, 1977, Engle and Watson, 1981, Peña and Box, 1987, Stock and Watson, 1988, West et al, 1999, Forni et al, 2000,2005, Bai and Ng,2002, 2006, Lam and Yao,2012, Tsai and Tsay, 2015, and many others.

Brillinger Dynamic Principal Components

Brillinger (1981) addressed the reconstruction problem as follows. Suppose zero mean m dimensional stationary process $\{\mathbf{z}_t\}$, $-\infty < t < \infty$. The dynamic principal components are defined by searching for $m \times 1$ vectors $\mathbf{c}_h, -\infty < h < \infty$ and $\beta_j, -\infty < j < \infty$, so that if we consider as first principal component the linear combination

$$f_t = \sum_{h=-\infty}^{\infty} \mathbf{c}_h' \mathbf{z}_{t-h}, \tag{1}$$

then

$$E\left[(\mathbf{z}_t - \sum_{j=-\infty}^{\infty} \beta_j f_{t+j})'(\mathbf{z}_t - \sum_{j=-\infty}^{\infty} \beta_j f_{t+j})\right]. \tag{2}$$

is minimum. Brillinger elegantly solved this problem by showing that $\mathbf{c}_k$ is the inverse Fourier transform of the principal components of the cross spectral matrices for each frequency, and β_j is the inverse Fourier transform of the conjugates of the same principal components. See Brillinger (1981) and Shumway and Stoffer (2000)

Limitation of Brillinger DPC:

1. They are noly justifed under stationarity.
2. They are not easy to robustify.
3. They are not reliable when m, the number of time series is large

and the ratio T/m is small because the covariance matrices and the spectral matrices are estimated with low precision and so are there eigenvalues.

Our procedure gives an optimal reconstruction of the vector of time series from a finite number of lags:

(1) the solution can be easily computed even if m is large.

(2) it does not require stationarity.

(3) It does not assume that the DPC is a linear combination of the series.

(4) it can be easily made robust by changing in the minimization criterion of the squared function by a bounded function.

2. Generalized Dynamic Principal Components (GDPC)

Suppose that we observe $z_{j,t}, 1 \leq j \leq m, 1 \leq t \leq T$, and consider two integer numbers $k_1 \geq 0$ and $k_2 \geq 0$. We can define the first dynamic principal component with $k = k_1 + k_2$ lags (first DPC_k) as a vector

$$\mathbf{f} = (f_{1-k_1}, f_{-k_1}, \ldots, f_0, f_1, \ldots, f_t, f_{t+1}, \ldots, f_T, f_{T+1}, \ldots, f_{T+k_2-1}, f_{T+k_2})$$

so that the reconstruction of the series from $\mathbf{f}$ is optimal

Given $\mathbf{f}$, the $m \times (k_1 + k_2)$ matrix of coefficients.

$$\beta = (\beta_{j,i})_{1 \leq j \leq m, -k_1+1 \leq i \leq k_2},$$

and

$$\alpha = (\alpha_1, \ldots, \alpha_m)$$

are used to reconstruct the values $z_{j,t}$ as

$$\widehat{z}_{j,t}(\mathbf{f}, \beta_j, \alpha_j) = \sum_{i=-k_1}^{k_2} \beta_{j,i} f_{t+i} + \alpha_j,$$

where β_j is the j-th row of β.

We can always assume one side lags

$$\widehat{z}_{j,t} = \sum_{i=-k_1}^{k_2} \beta_{j,i} f_{t+i} + \alpha_j.$$

Let $k = k_1 + k_2$ and put

$$f_t^* = f_{t-k_1}, 1 \leq t \leq T + k, \ \beta_{j.,h}^* = \beta_{j,h-k_1}, 0 \leq h \leq k$$

Also define

$$f_t^{**} = f_{t+k}^*, 1 - k \leq t \leq T, \beta_{j.,h}^{**} = \beta_{j,k-h}, 0 \leq h \leq k \tag{3}$$

then, the reconstructed series can also be obtained as

$$\widehat{z}_{j,t} = \sum_{i=-k_1}^{k} \beta_{j,i} f_{t+i+k_1} + \alpha_j = \sum_{h=0}^{k} \beta_{j,h,}^* f_{t+h}^* + \alpha_j = \sum_{h=0}^{k} \beta_{j,h,}^{**} f_{t-h}^{**} + \alpha_j$$

For this reason in the remaining of the paper we will assume without loss of generality $k_1 = 0$ and we will use k to denote the number of forward leads. Once obtained

Consider the loss function

$$MSE(\mathbf{f},\beta,\alpha) = \frac{1}{T}\sum_{j=1}^{m}\sum_{t=1}^{T}(z_{j,t} - \widehat{z}_{j,t}(\mathbf{f},\beta_j,\alpha_j))^2$$

$$= \frac{1}{T}\sum_{j=1}^{m}\sum_{t=1}^{T}(z_{j,t} - \sum_{i=0}^{k}\beta_{j,i+1}f_{t+i} - \alpha_j)^2. \qquad (1)$$

The values of $\mathbf{f} = (f_1, ..., f_{T+k})'$, $\beta = (\beta_{j,i})$ and $\alpha = (\alpha_1, ...\alpha_m)$ which minimize the mean square error, are

$$(\widehat{\mathbf{f}},\widehat{\beta},\widehat{\alpha}) = \arg\min_{\mathbf{f},\beta,\alpha} MSE(\mathbf{f},\beta,\alpha).$$

Clearly if $\mathbf{f}$ is optimal, $\gamma\mathbf{f}+\delta$ is optimal too. Then we can choose $\mathbf{f}$ so that

$$\frac{1}{T+k}\sum_{t=1}^{T+k} f_t^2 = 1.$$

and

$$\frac{1}{T+k}\sum_{t=1}^{T+k} f_t = 0.$$

Then, we call $\widehat{\mathbf{f}}$ the first DPC of order k of the observed series $\mathbf{z}_1, ..., \mathbf{z}_t$.

Note that the first DPC of order 0 corresponds to the first regular principal component of the data.

Moreover the matrix $\widehat{\beta}$ contains the coefficients to be used to reconstruct the m series from $\widehat{\mathbf{f}}$ in an optimal way.

Given $\widehat{\mathbf{f}}$

$$\begin{pmatrix} \widehat{\beta}_j \\ \widehat{\alpha}_j \end{pmatrix} = \left(\mathbf{F}(\widehat{\mathbf{f}})'\mathbf{F}(\widehat{\mathbf{f}})\right)^{-1} \mathbf{F}(\widehat{\mathbf{f}})' \, \mathbf{z}^{(j)},$$

and given $\widehat{\beta}_j$ and $\widehat{\alpha}_j$ we have

$$\mathbf{f} = \mathbf{D}(\mathbf{f},\widehat{\beta})^{-1} \sum_{j=1}^{m} \mathbf{C}_j(\mathbf{f},\widehat{\alpha})\widehat{\beta}_j.$$

The coefficients β_j and α_j, $1 \leq j \leq m$ can be obtained using the least squares estimator, where $\mathbf{z}^{(j)} = (z_{j,1}, ..., z_{j,T})'$ and $\mathbf{F}(\mathbf{f})$ is the $T \times (k+2)$ matrix with t-th row $(f_t, f_{t+1}, ..., f_{t+k}, 1)$.

Iterative algorithm:

step 1 Define $\beta_j^{(h)}$ and $\alpha_j^{(h)}$ by

$$\begin{pmatrix} \beta_j^{(h)} \\ \alpha_j^{(h)} \end{pmatrix} = \left(\mathbf{F}(\mathbf{f}^{(h)})' \, \mathbf{F}(\mathbf{f}^{(h)})\right)^{-1} \mathbf{F}(\mathbf{f}^{(h)})' \mathbf{z}^{(j)}$$

step 2 Then $\mathbf{f}^{(h+1)}$ can be defined by

$$\mathbf{f}^* = \mathbf{D}(\mathbf{f}^{(h)}, \beta^{(h)}, \alpha^{(h)})^{-1} C(\mathbf{f}^{(h)}, \beta^{(h)}, \alpha)\beta^{(h)}$$

and

$$\mathbf{f}^{(h+1)} = (\mathbf{T}+k)^{1/2}(\mathbf{f}^* - \overline{\mathbf{f}}^*)/||\mathbf{f}^* - \overline{\mathbf{f}}^*||.$$

The initial value $\mathbf{f}^{(0)}$ can be chosen equal to the standard (non dynamic) first principal component, completed with k zeros.

The second S -DPC is defined as the first S-DPC of the residuals $r_{j,t}(\mathbf{f}, \beta)$. Higher order S-DPC are defined in a similar manner.

Note:

$\mathbf{D}(\beta)$ is $(T + k)\text{x}(T + k)$, and $\mathbf{C}(\alpha)$ is $(T + k)\text{x}(k + 1)$

$\mathbf{F}(\mathbf{f})$ is the $T\text{x}(k + 2)$

Remark 1. Note that the dimension of the matrices to be inverted to compute $\mathbf{f}^{(h)}, \beta^{(h)}, \alpha^{(h)}$ are independent of the number of time series and therefore we can deal with large number of variables.

Remark 2. Note also that there are no restrictions on the values $\mathbf{f}$ and in particular we do not assume, as in Brillinger, that they must be linear combinations of the series. In this way the values of $\mathbf{f}$ can be adapted to the nonstationarity character of the time series.

3 Dynamic Principal Components when $k = 1$

To illustrate the computation of the first DPC, let us consider the simplest case of $k = 1$. Then, we search for $\widehat{\beta}$ and $\widehat{\mathbf{f}} = (\widehat{f}_1, ..., \widehat{f}_{T+1})'$ such that

$$(\widehat{\mathbf{f}}, \widehat{\beta}) = \arg\min_{1} \sum_{t=1}^{T} \sum_{j=1}^{m} (z_{j,t} - \beta_{j,1} f_t - \beta_{j,2} f_{t+1})^2.$$

It can be shown that with $0 \leq c < 1$

$$\widehat{f}_t = \frac{1}{\delta} \left[\sum_{j=1}^{m} \widehat{\beta}_{j,1} \sum_{q=1}^{T} c^{|t-q|} z_{j,q} + \sum_{j=1}^{m} \widehat{\beta}_{j,2} \sum_{q=2}^{T+1} c^{|t-q|} z_{j,q-1} \right] + R_t,$$

where $R_t \to 0$ except for t close to 1 or close to T.

The DPC is computed as a weighted average of the observations by a two side moving average. For $\widehat{f}_t$ the maximum weight is given to $z_{j,t}$ and $z_{j,t-1}$.

Suppose now that $\mathbf{z}_t$ is stationary, then except in both ends $\widehat{f}_t$ can be approximated by the stationary process

$$\widehat{f}_t^* = \frac{1}{\alpha}\left[\sum_{j=1}^{m}\widehat{\beta}_{j,1}\sum_{q=-\infty}^{\infty}c^{|t-q|}z_{j,q} + \sum_{j=1}^{m}\widehat{\beta}_{j,2}\sum_{q=-\infty}^{\infty}c^{|t-q|}z_{j,q-1}\right].$$

The DPC is approximated as linear combinations of the geometrically and symmetrically filtered series

$$z_{j,t} + \sum_{i=1}^{\infty}c^i(z_{j,t+i} + z_{j,t-i}), 1 \leq j \leq m$$

and

$$z_{j,t-1} + \sum_{i=1}^{\infty}c^i(z_{j,t-1+i} + z_{j,t-1-i}), 1 \leq j \leq m$$

This series give the largest weight to the periods t and $t-1$ respectively and the weights decrease geometrically when we move away of these values.

We conjecture that in the case of the first DPC of order k, a similar approximation outside both ends of $\widehat{f}_t$ by an stationary process can be obtained.

Monte Carlo results

We perform a Monte Carlo study using as vector series $\mathbf{z}_t = (z_{1,t}, z_{2,t}, ..., z_{m,t})'$, $1 \leq t \leq T$ generated as follows:

$$z_{i,t} = 10\sin(2pi(i/m))f_t + 10\cos(2pi(i/m))f_{t-1} + 10(i/m)f_{t-2} + u_{i,t}, 1 \leq i \leq m, 1 \leq t \leq T, \tag{12}$$

where $f_t, -2 \leq t \leq T$ and $u_{i,t}, \leq t \leq T$, $1 \leq i \leq m$ are i.i.d. random variables with distribution $\mathrm{N}(0,1)$. We compute three different principal components: (i) The ordinary principal component used in a dynamic way with k lags to reconstruct the original series (OPC_k) (ii) the dynamic principal component (DPC_k) proposed here, (iii) Brillinger dynamic principal components (BDPC_k) adapted for finite samples as follows:

m	T	$\mathbf{OPC}_2$	$\mathbf{DPC}_2$	BDPC_{10}
20	100	52.53	0.91	0.94
	200	55.86	0.92	0.95
100	100	54.89	0.95	0.99
	200	57.65	0.97	0.99
500	100	53.56	0.96	1.00
	200	57.14	0.98	1.00
1000	100	54.88	0.96	-
	200	59.09	1.00	-

Table 1: MSE of the Reconstructed Series for the Stationary Model with one Factor

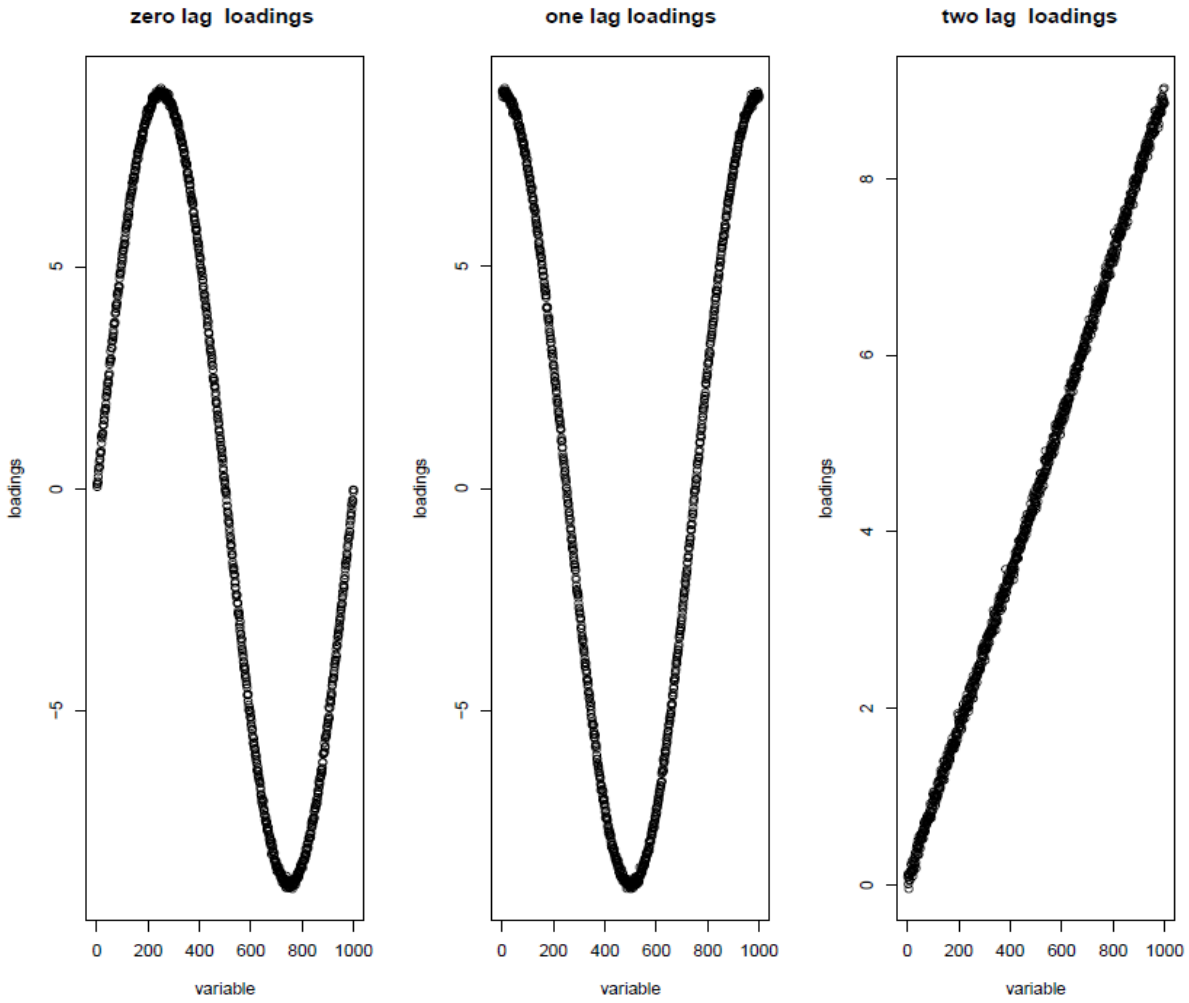


Figure 1: Loadings for one Replication of the Stationary Model with T=200 and m=1000

In this case we consider a VARI$(1,1)$ $m-$dimensional vector series $\mathbf{z}_t$ generated as follows. Consider an stationary VAR(1) model $\mathbf{x}_t = A\mathbf{x}_{t-1} + \mathbf{u}_t, 1 \leq t \leq T$, where the $\mathbf{u}_t$s are i.i.d. m-dimensional vectors with distribution $\mathrm{N}_m(\mathbf{0}, \mathbf{I})$ and let $\mathbf{z}_t = \mathbf{z}_{t-1} + \mathbf{x}_t$. We consider 1000 replications and in each replication we generate a new matrix A of the form $A = V\Lambda V'$, where V is an orthogonal matrix generated at random with uniform distribution and Λ is a diagonal matrix, where the diagonal elements are independent with uniform distribution in the interval $[0, 0.9]$.

m	$\mathbf{OPC}_{10}$	$\mathbf{DPC}_{10}$	BDPC_{10}
20	67	83	55
100	67	86	62
200	69	86	62

Table 2: Percentage of Explained Variance in the VARI(1,1) Model

Example 1(I)

4.1 Example 1

Six series of Industrial Production Index of France, Germany, Italy, United Kingdom, USA and Japan

Monthly data from January 1991 to December 2012

The series shows a break in 2008

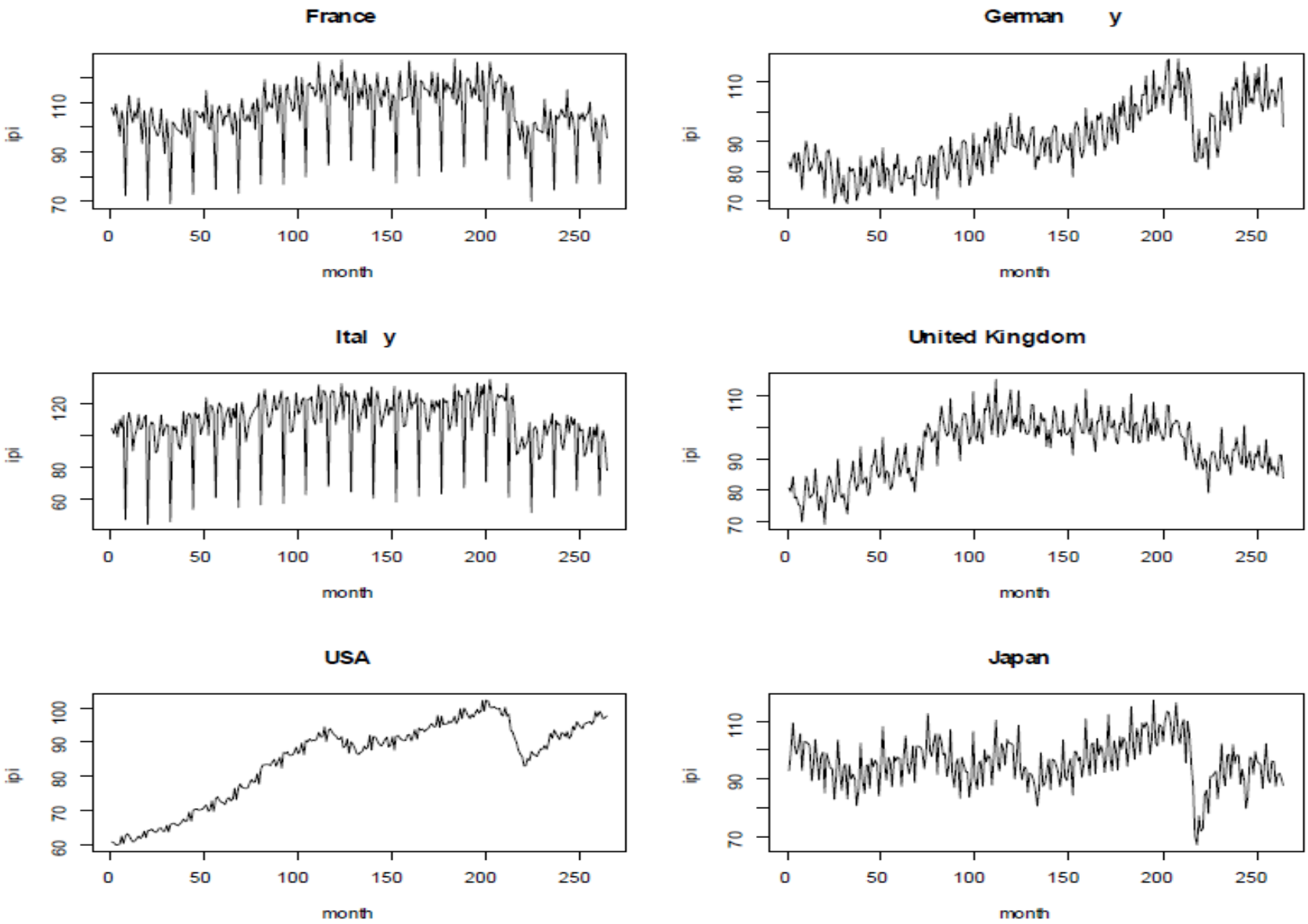


Figure 1: Industrial production Index of six countries 1991-2012

Example 1 (II)

Table 1. Explained variability of the OPC $(EV_{0,k})$ and DPC $(EV_{k,k})$ for the IPI series with different number of lags for the IPI7 data using DPC and OPC

k	$EV_{0,k}$	$EV_{k,k}$
0	63.07	63.07
1	66.19	82.47
5	76.66	90.05
10	77.98	94.81
12	80.00	96.67

EV_0= Ordinary (static) principal components
EV_k=Dynamic principal components

Adding lags: Small improvements in OPC but large in DPC

PC	PC(0)	PC(1)	DPC(0)	DPC(1)	
-0.456	-0.456	-0.001	-3.951	3.965	FR
-0.285	-0.275	-0.034	-1.509	1.492	GE
-0.719	-0.750	0.099	-6.548	6.577	IT
-0.298	-0.269	-0.092	-2.114	2.111	UK
-0.241	-0.198	-0.138	-0.787	0.760	US
-0.212	-0.212	-0.001	-1.885	1.894	JP

Table 1: Coefficients of the six countries in the Ordinary (OPC) and Dynamic Principal Components

PC	PC(0)	DPC(0)	
63	61	60	FR
40	37	23	GE
100	100	100	IT
41	36	32	UK
33	26	12	US
30	28	29	JP

Table 2: Relative coefficients of the six countries in the Ordinary (OPC) and Dynamic Principal Components

Example 1 (III)

For the ordinary PC the coefficients in the first column in Table 1 coincide with the weights given to each country for the definition of the OPC.

The first OPC gives the largest weight to Italy and then France, because of the strong seasonality of these series which have the largest variability.

The second and third columns show that for reconstructing the original variables including the lag of the OPC is practically irrelevant.

The fourth and fifth columns show that the DPC with one lag is almost equivalent to using the first difference of the DPC in the reconstruction of the series.

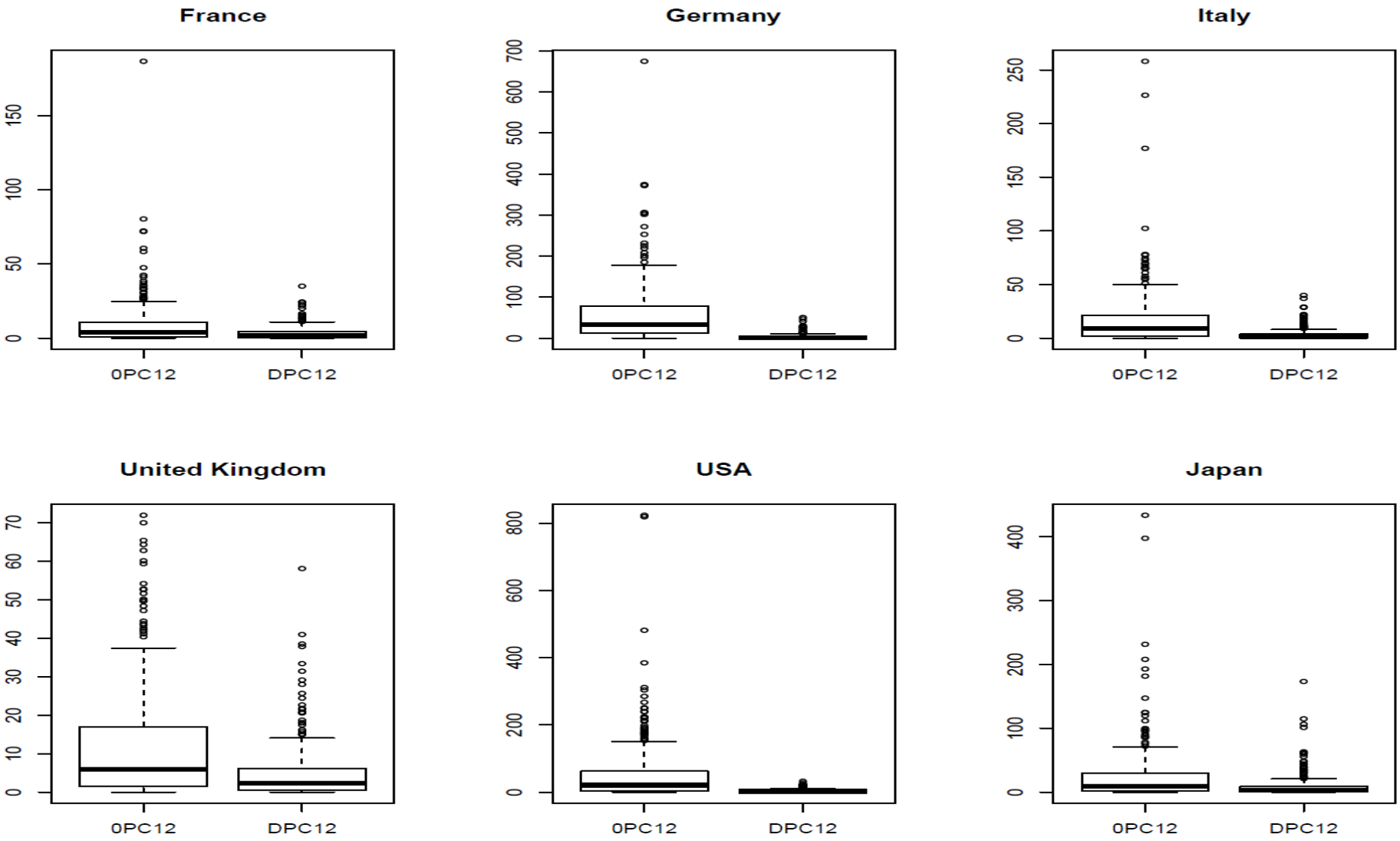


Figure 3: Boxplots of the Absolute Values of the Errors of the Reconstructed IPI Series

Summary of the example:

k	$EV_{0,k}$	$EV_{k,k}$
0	63.07	63.07
1	66.19	82.47
5	76.66	90.05
10	77.98	94.81
12	80.00	96.67

- Original data 264 x 6=1584 values
- DPC with 12 lags = 264+12+12 x 6=348 values
- Represent 348/1584=0,22 of the data
- With 22% of the data, 96,7% of the information

4.2 Example 2.

Data: The main components of the IBEX (general index of the Madrid stock market) 31 daily stock prices in the stock market in Madrid corresponding to the 251 trading days of the year 2004.

In this example the data set is composed of 30 daily stock prices in the stock market in Madrid corresponding to the 251 trading days of the year 2004. The source of the data is the Ministry of Economy, Spain. In Table 5 we show the percentage of the variance explained by the different procedures using the OPC_k,DPC_k BDPC_k procedures.

k	OPC_k	DPC_k	BDPC_k
0	60	60	-
1	60	82	-
5	61	87	-
10	62	88	60

Table 5: Explained variability of the OPC and DPC for the stock prices series with different number of lags

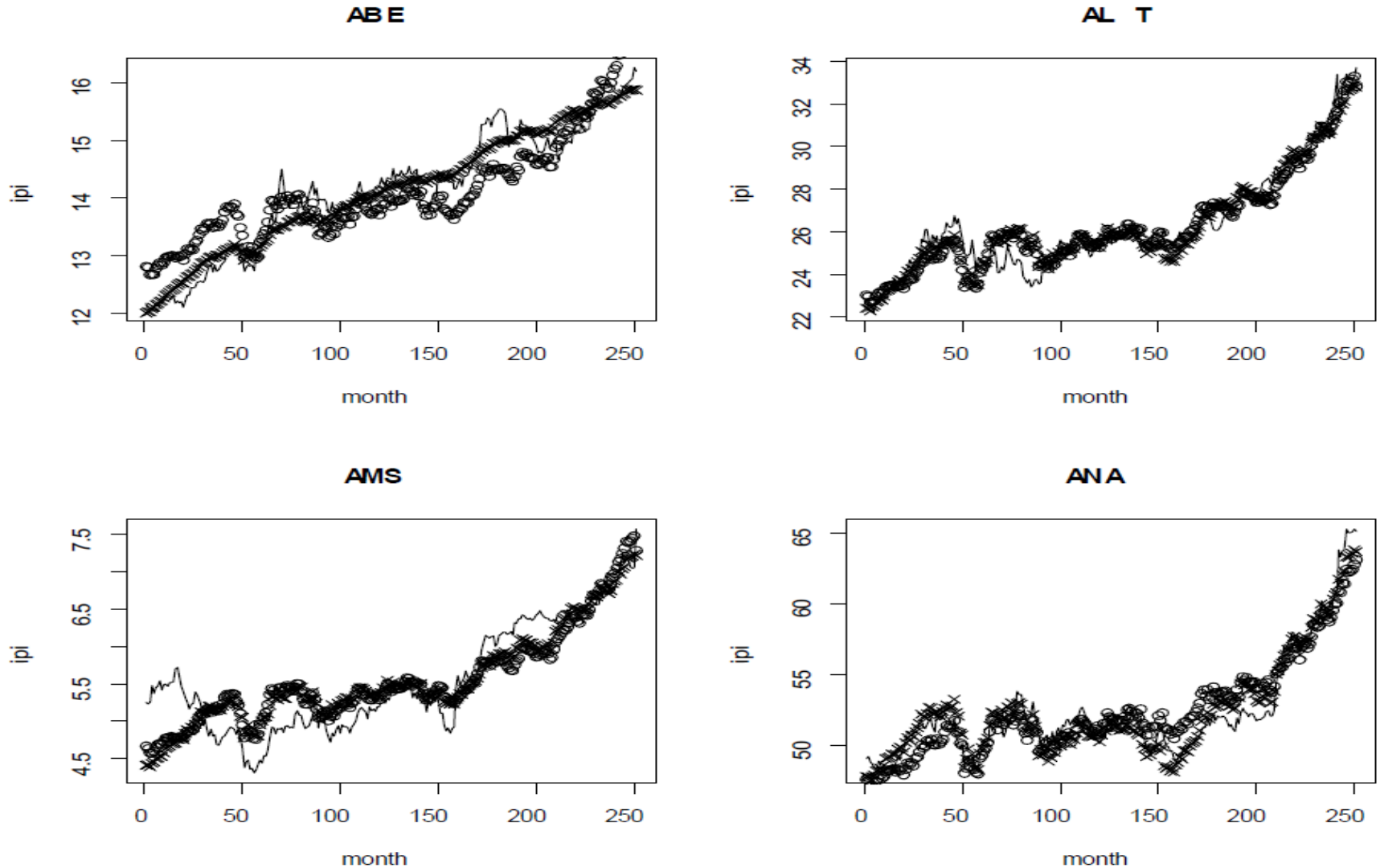


Figure 6: Values of the original and reconstructed of the first four chosen in alphabetic orders. The reconstruction was made with the OPC (o) and DPC (*) using one lag

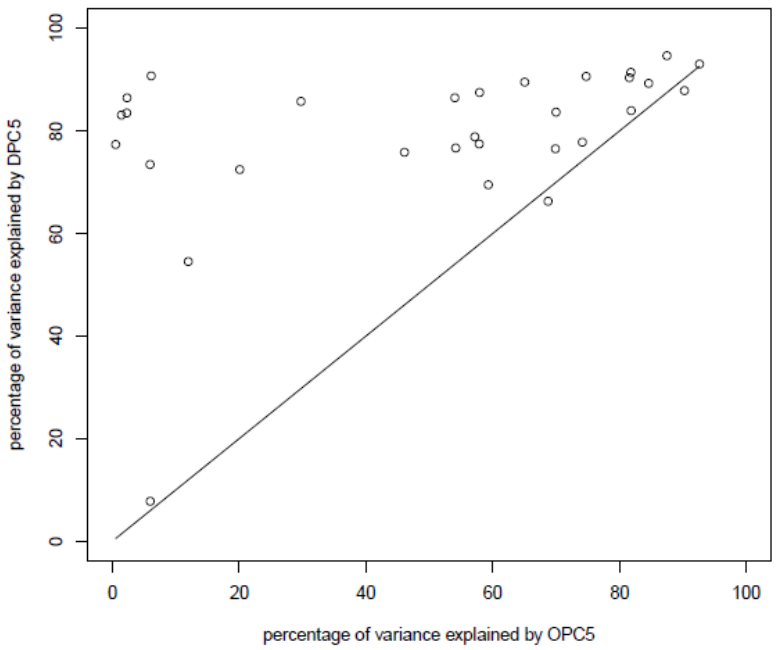
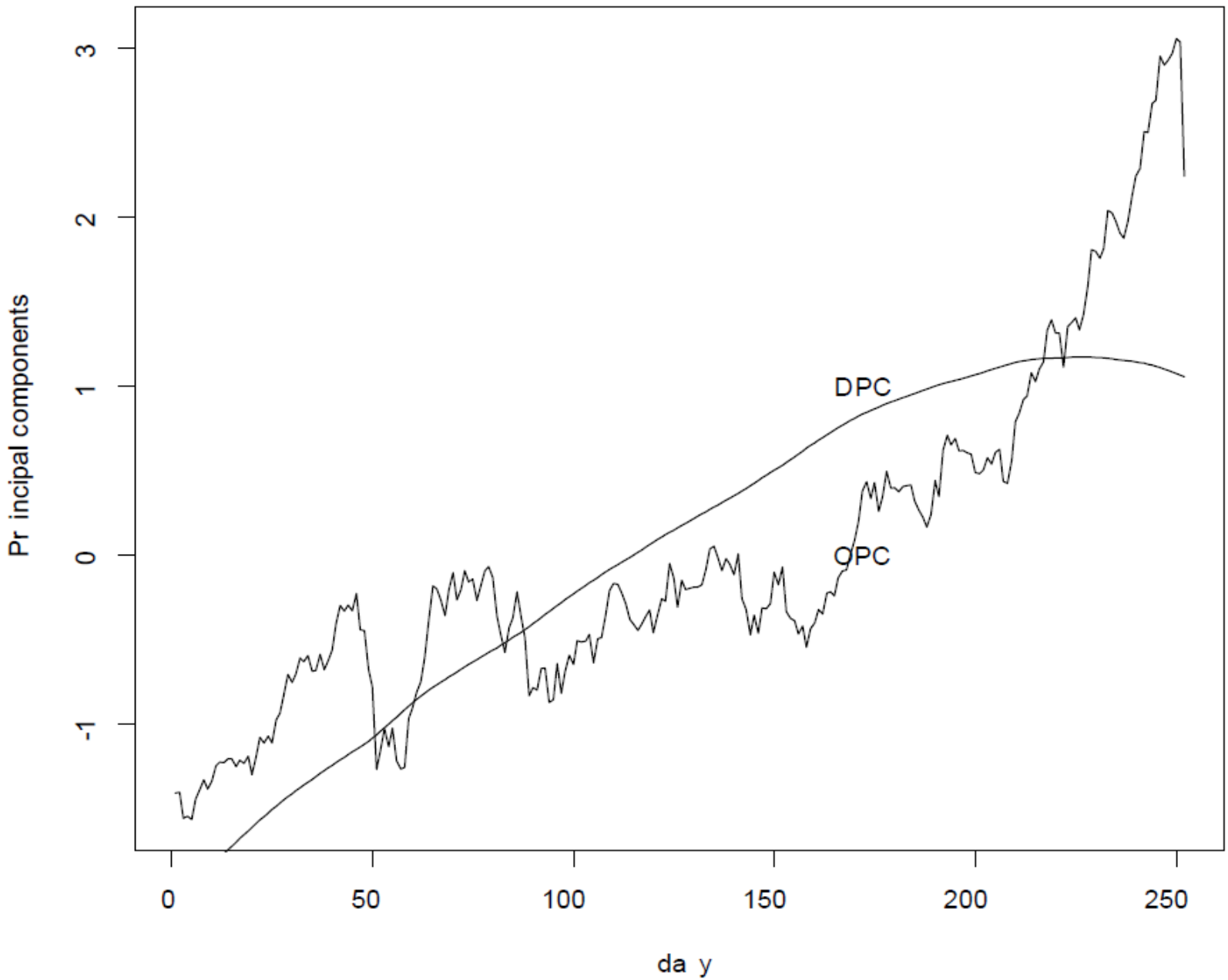


Figure 4: Percentage of the Variance Explained by the DPC_5 Against the Percentage Explained by the OPC_5 Procedure.


Pr incipal components
3
2
1
0
-1
DPC
CPC
0
50
100
150
200
250
da y

Summary of the example:

k	$EV_{0,k}$	$EV_{k,k}$
0	0.598	0.598
1	0.602	0.822
5	0.610	0.873
10	0.620	0.881

Table 2: Explained variability of the OPC and DPC for the istock prices series with different number of lags

- Original data 251 x 31 = 7781
- DPC with 5 lags = 251+5+5 x 31= 411
- Represent 411/7781=0,05 of the data
- With 5% of the data, 87,3% of the information

4.2.1 Robust Dynamic Principal Components

The DPC are not robust. In fact a very small fraction of outliers may have an unbounded influence on $(\mathbf{f}, \alpha, \beta)$.

For this reason we are going to study a robust alternative. One of the standard procedures to obtain robust estimates for many statistical models is to replace the minimization of the mean square scale for a the minimization of a robust M-scale.

The estimators defined by means of a robust M-scale are called S-estimators. In this section we extend the S-estimators for the case of the DPC.

In the case of time series with strong seasonality, a special care is required. The reason is that the values corresponding to a season very different to the other may be taken by the robust procedures as outliers. and therefore down-weighted. As a consequence, the reconstruction of these observations may be affected by large errors. Thus the procedure we present here assumes that the series have been adjusted by seasonality and therefore this problem is not present.

We can define the first S-DPC as follows Let $\mathbf{f}, \beta$ and α as before. For $1 \leq j \leq m$, let $\mathbf{r}_j(\mathbf{f}, \beta_j, \alpha_j) = (r_{j,t}(\mathbf{f}, \beta_j, \alpha_j))_{1 \leq t \leq T}$, where

$$r_{j,t}(\mathbf{f}, \beta_j, \alpha_j) = z_{j,t} - \sum_{i=0}^{k} \beta_{j,i} f_{t+i} - \alpha_j.$$

Define

$$SRS(f, \beta, \alpha) = \sum_{j=1}^{m} S^2(\mathbf{r}_j(\mathbf{f}, \beta_j, \alpha_j)) \tag{4}$$

$$(\widehat{\mathbf{f}}, \widehat{\beta}, \widehat{\alpha}) = \arg\min_{\mathbf{f}, \beta} SRS(f, \beta, \alpha)$$

Algorithm for Robust DPC:

Then to define the algorithm is enough to describe how to compute $(\mathbf{f}^{(h+1)}, \beta^{(h+1)}, s^{(h+1)})$ once $(\mathbf{f}^{(h)}, \beta^{(h)}, s^{(h)})$ is known. This is done in the following four steps:

step1 Compute

$$\mathbf{f}^* = \mathbf{D}(\mathbf{f}^{(h)}, \beta^{(h)}, \alpha^{(\mathbf{h})}, \mathbf{s}^{(h)})^{-1} C(\mathbf{f}^{(h)}, \beta^{(h)}, \alpha^{(h)} \mathbf{s}^{(h)}) \beta^{(h)}$$

step 2 Put $\mathbf{f}^{(h+1)} = (\mathbf{T}+k)^{1/2}(\mathbf{f}^* - \overline{\mathbf{f}}^*)/|||\mathbf{f}^* - \overline{\mathbf{f}}^*||.$

step3 Compute the j-th row

$$\begin{pmatrix} \beta_j^{(h+1)} \\ \alpha_j^{(h+1)} \end{pmatrix} = \left(\mathbf{F}(\mathbf{f}^{(h+1)})' W_j(\mathbf{f}^{(h)}, \beta^{(h)}, \alpha^{(h)}, \mathbf{s}^{(h)}) \mathbf{F}(\mathbf{f}^{(h+1)})\right)^{-1} \mathbf{F}(\mathbf{f}^{(h+1)}) \mathbf{W}_j(\mathbf{f}^{(h)}, \beta^{(h)}, \alpha^{(h)}, s^{(h)})' \mathbf{z}^{(j)}$$

for $1 \leq j \leq m$.

step4 Compute $s_j^{(h+1)} = S(\mathbf{r}_j(\mathbf{f}^{(h+1)}, \beta, \alpha_{h+1}))$.

6 Example

We will use the data of example 2 to illustrate the performance of the robust DPC. This dataset was modified as follows .

Each of the 7781 values composing the dataset was modified with 5% probability adding 20 to the true value.

The Table include MSE in the reconstruction of the series with the DPC.

However since the MSE is very sensitive to the presence of outliers, we also include de criterion SRS

We also compute the robust S DPC and the corresponding SRS values are included in the fourth column of Table.

k	MSE of the DPC$_k$	SRS of the DPC$_k$	SRS of the S-DPC$_k$
1	18.81	6.05	0.84
5	17.75	6.64	0.50
10	16.90	7.63	0.48

Table 6: MSE and SRS of the DPC$_k$ and S DPC$_k$ for the contaminated stock prices series

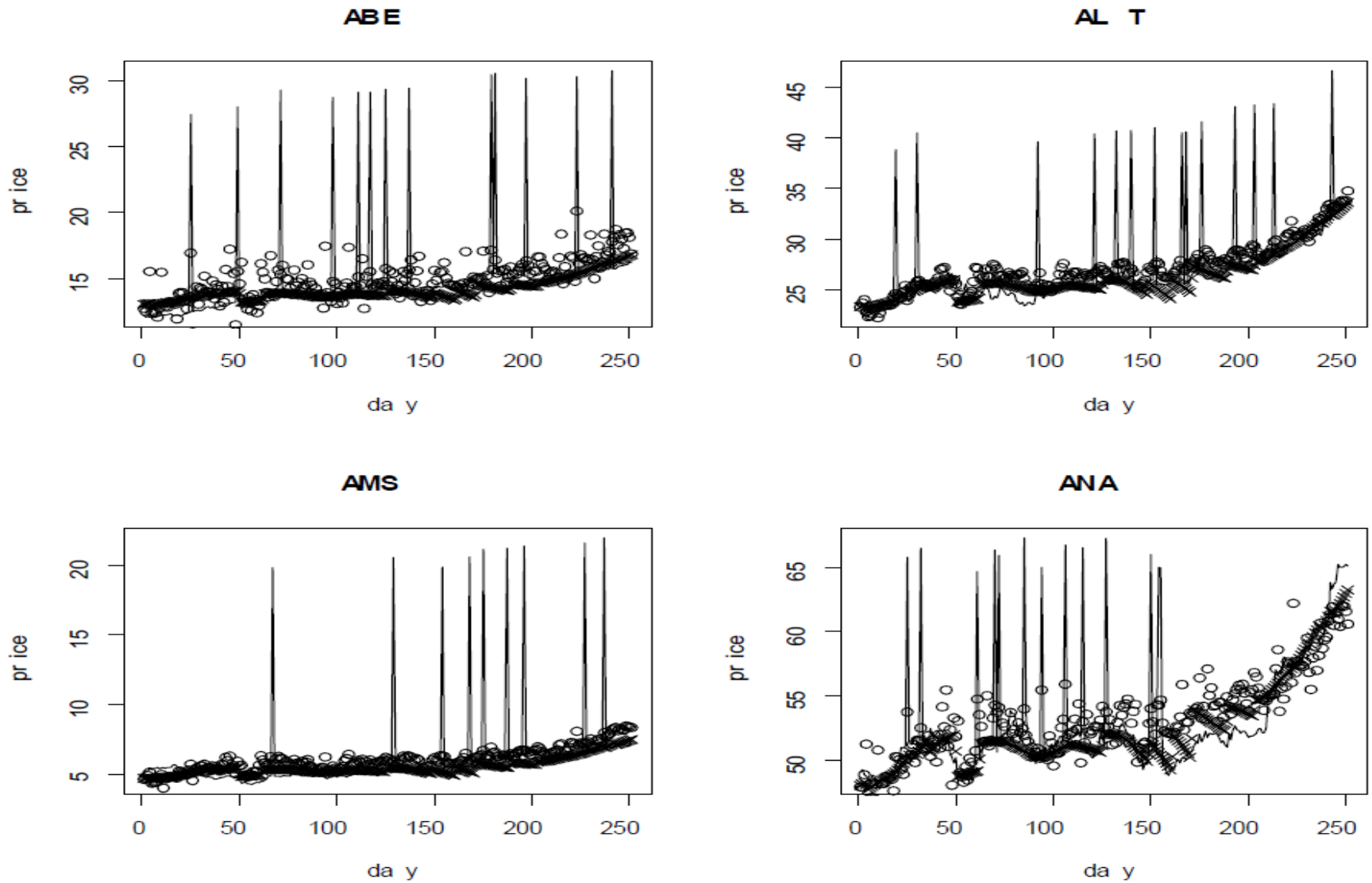


Figure 9: Contaminated Stock prices series and their reconstruction by DPC (o) and by robust DPC (x)

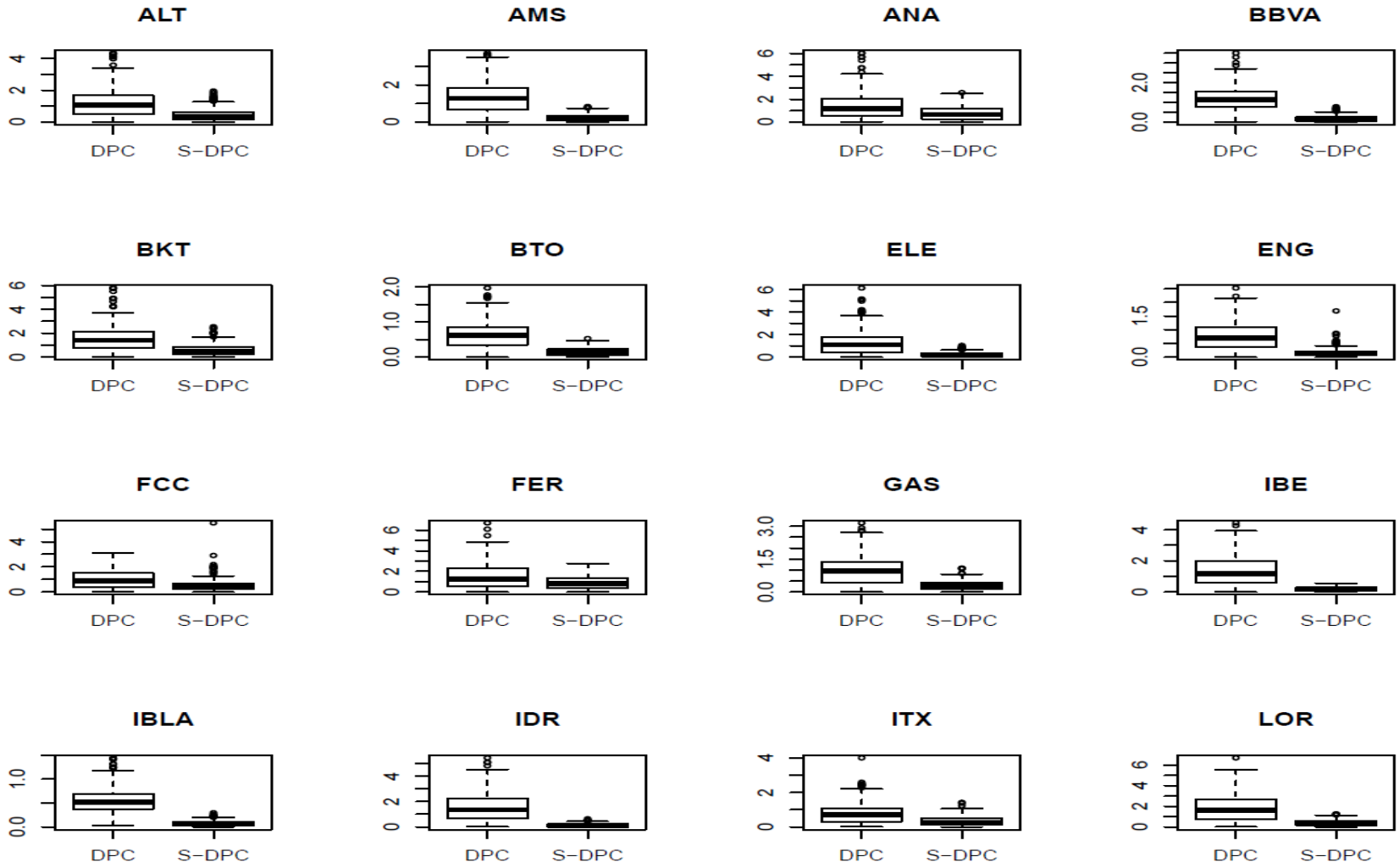


Figure 6: Boxplots of the Residual Absolute Values of the Stock Prices Obtained with the DPC_5 and S-$DPCC_5$ Procedures

Conclusions

- The GDPC we present are much more powerful for reconstruction than the OPC and can be applied for non stationary time series and very large data sets when Brillinger DPC fail.
- They can provide insights in the structure of the series
- The robust GDPC are able to filter contaminated data
- They can be applied in a broad range of problems.

Thank you for your attention!