Functional Data Analysis in Statistical Processing of Cyclostationary Signals. Theory and Applications

Jacek Leśkow

Cracow Technical University
Poland

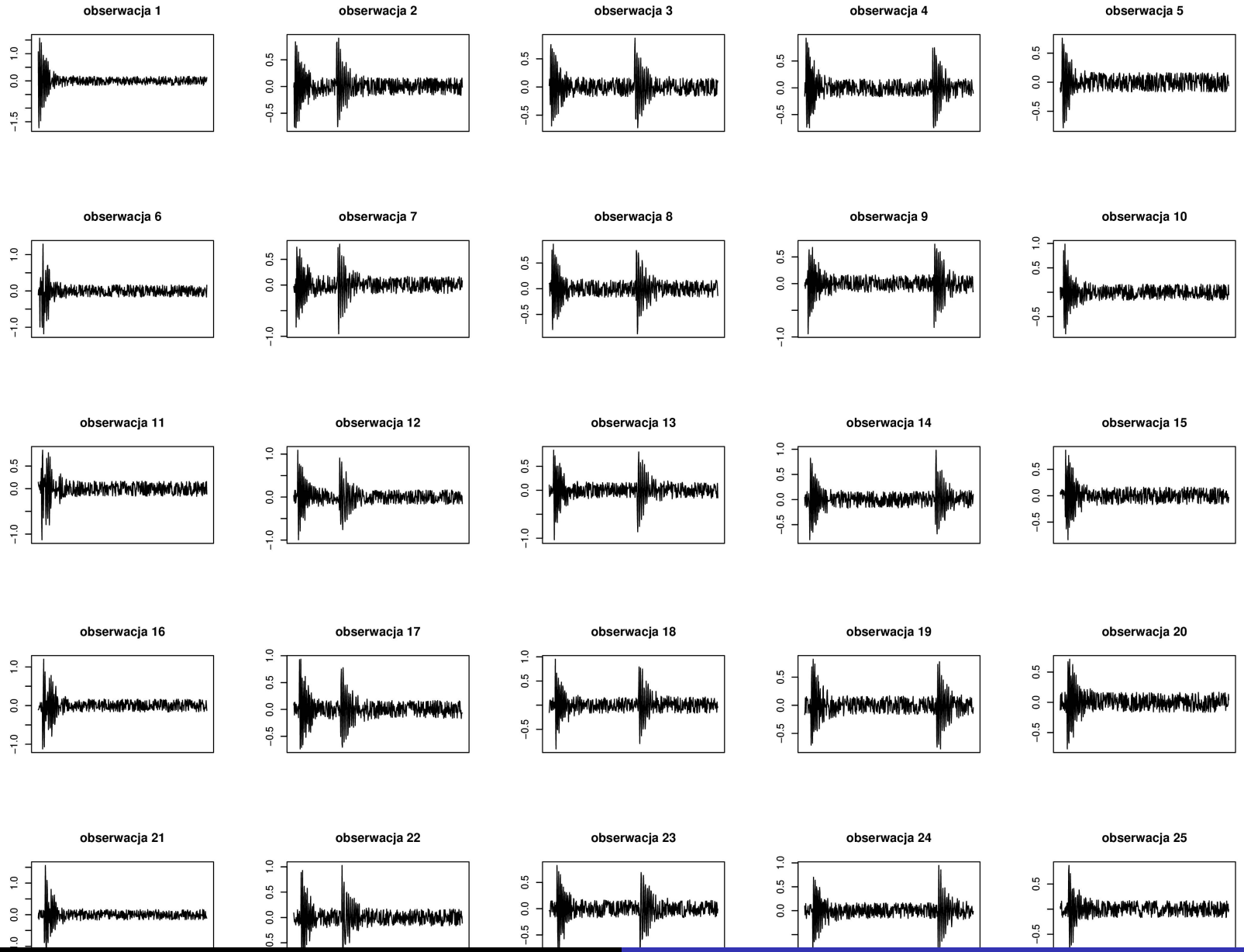
14th Grodek meeting, February 2021

Plan of the talk I

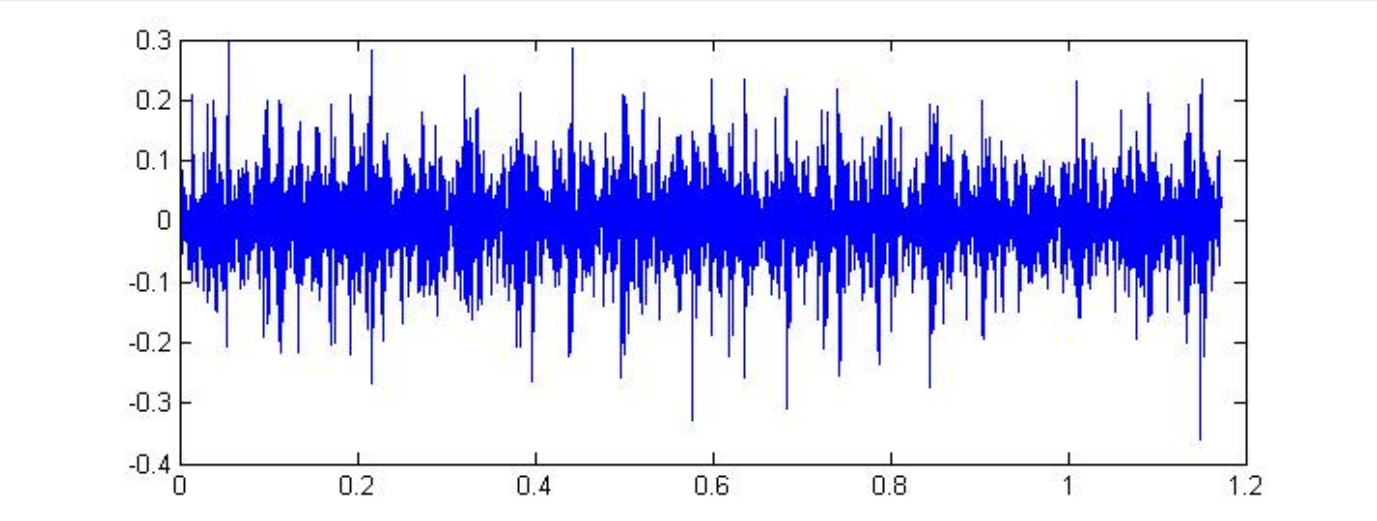
Abstract

Main goals of the talk:

- Review recent ideas on functional data analysis (FDA)
- Implementation of FDA in the context of PC/APC signals
- FDA language and estimation for PC/APC.


obserwacja 1
obserwacja 2
obserwacja 3
obserwacja 4
obserwacja 5
obserwacja 6
obserwacja 7
obserwacja 8
obserwacja 9
obserwacja 10
obserwacja 11
obserwacja 12
obserwacja 13
obserwacja 14
obserwacja 15
obserwacja 16
obserwacja 17
obserwacja 18
obserwacja 19
obserwacja 20
obserwacja 21
obserwacja 22
obserwacja 23
obserwacja 24
obserwacja 25

The segments usually come out of this.



Definition of APC

We say that $\{X(t);\ t \in \mathbb{Z}\}$ - APC, when $\mu_X(t) = E(X_t)$ and the autocovariance function

$$B_X(t,\tau) = \operatorname{cov}(X_t, X_{t+\tau})$$

are almost periodic function at t for every $\tau \in \mathbb{Z}$.
Function f is almost periodic in the norm $\|\cdot\|$ if
for each ϵ there exists an almost period P_ϵ such that

$$\|f(\cdot + P_\epsilon) - f(\cdot)\| < \epsilon$$

.

Introduction to FDA

To start, we will see any signal $\{X(t); t \in \mathbb{Z}\}$ as a collection of independent curves $\{y_i(u), i = 1, \ldots, N;\ u \in A\}$ belonging to a Hilbert space $\mathcal{H}$. For simplicity, assume that $\mathcal{H} = L^2[A]$ and $A = [0, 1]$.

Now, let us see the fundamental steps of the FDA approach to signal analysis.

Step 1 The stochastic model for the signal is the random element X from $(\Omega, \mathcal{F}, P)$ to $L^2[0, 1]$.

FDA - cont.

Step 2 Expectation of the random element

If X is integrable, there is a unique function $\mu \in L^2$ such that $\mathbb{E}\langle y, X\rangle = \langle y, \mu\rangle \ \forall\ y \in L^2$. It follows that $\mu(t) = \mathbb{E}[X(t)]$ for almost all $t \in [0,1]$.

Step 3 Covariance operator

For X intergrable and $\mathbb{E}X = 0$, the covariance operator of X is defined by

$$C(y) = \mathbb{E}[\langle X, y\rangle X], \quad y \in L^2.$$

Notice that

$$C(y)(t) = \mathbb{E}[\langle X, y\rangle X(t)] = \mathbb{E}\int X(s)y(s)ds X(t) =$$

$$= \int \underbrace{\mathbb{E}[X(s)X(t)]}_{=c(s,t)} y(s)ds = \int c(s,t)y(s)ds.$$

FDA - covariance

Step 4. Eigenvalues and eigenfunctions of the covariance operator

Let $v_j, \lambda_j, j \geq 1$ be the eigenfunctions and the eigenvalues of the covariance operator C. The relation $C(v_j) = \lambda_j v_j$ implies that

$$\lambda_j = \langle C(v_j), v_j \rangle = \langle \mathbb{E}[\langle X, v_j \rangle X], v_j \rangle = \mathbb{E}\langle X, v_j \rangle^2.$$

Having defined the mean and the covariance of the random element, we will proceed to the usual statistical questions, that is:

What is the approximate distribution of the linear statistics for samples generated by our random element ?
How to introduce the estimator of the covariance ?
Is there any chance for the dimensionality reduction ?

FDA - CLT

Suppose $\{X_n, n \geq 1\}$ is a sequence of iid mean zero random elements in a separable Hilbert space such that $\mathbb{E}\|X_i\|^2 < \infty$. Then

$$\frac{1}{\sqrt{N}} \sum_{n=1}^{N} X_n \xrightarrow{d} Z$$

where Z is a Gaussian random element with the covariance operator

$$C(x) = \mathbb{E}[\langle Z, x \rangle Z] = \mathbb{E}[\langle X_1, x \rangle X_1].$$

Notice that a normally distributed function Z with a covariance operator C admits the expansion (Karhunen-Lòeve representation)

$$Z \stackrel{d}{=} \sum_{j=1}^{\infty} \sqrt{\lambda_j} N_j v_j$$

where $N_j \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, λ_j, v_j - eigenvalues, eigenfunctions of the covariance operator C

FDA - estimation

$$\begin{aligned}
\mu(t) &= \mathbb{E}[X(t)] \qquad \text{(mean function)};\\
c(t,s) &= \mathbb{E}[(X(t)-\mu(t))(X(s)-\mu(s))] \qquad \text{(covariance function)};\\
C &= \mathbb{E}[\langle (X-\mu), \cdot \rangle (X-\mu)] \qquad \text{(covariance operator)}.
\end{aligned}$$

estimators:

$$\begin{aligned}
\hat{\mu}(t) &= \frac{1}{N}\sum_{i=1}^{N} X_i(t);\\
\hat{c}(t,s) &= \frac{1}{N}\sum_{i=1}^{N} (X_i(t)-\hat{\mu}(t))(X_i(s)-\hat{\mu}(s));\\
\hat{C}(x) &= \frac{1}{N}\sum_{i=1}^{N} \langle X_i-\hat{\mu}, x\rangle (X_i-\hat{\mu}), \quad x \in L^2.
\end{aligned}$$

FDA estimation - cont.

Assume that the observations have mean zero. We therefore have

$$\hat{c}(t,s) = \frac{1}{N}\sum_{i=1}^{N} X_i(t)X_i(s); \qquad \hat{C}(x) = \frac{1}{N}\sum_{i=1}^{N}\langle X_i, x\rangle X_i, \quad x \in L^2$$

therefore

$$\hat{C}(x)(t) = \int \hat{c}(t,s)x(s)ds, \quad x \in L^2.$$

Introduce the random functions

$$Z_N(t,s) = \sqrt{N}(\hat{c}(s,t) - c(s,t))$$

where $\hat{c}(s,t), c(s,t)$ are centered with the (sample) mean function.

FDA CLT for covariance

If the observations $X_1, X_2, \ldots, X_N$ are iid in L^2, and have the same distribution as X, which is assumed to be square integrable with $\mathbb{E}X(t) = 0$ and $\mathbb{E}\|X\|^4 < \infty$, then $Z_N(t,s)$ converges weakly in $L^2([0,1] \times [0,1])$ to a Gaussian process $\Gamma(t,s)$ with $\mathbb{E}\Gamma(t,s) = 0$ and

$$\mathbb{E}[\Gamma(t,s)\Gamma(t',s')] = \mathbb{E}[X(t)X(s)X(t')X(s')] - c(t,s)c(t',s').$$

For the lovers of spectrogram

If $X_1, X_2, \ldots, X_N$ represent functions of the frequency (vertical stripes), then FDA approach provides a simple description of the whole energy of the signal.

FDA and reduction of dimensionality

Let $\lambda_1 > \lambda_2 > \ldots$ be the eigenvalues of operator C. The eigenfunctions v_j are defined by $Cv_j = \lambda_j v_j$. The v_j are typically normalized, so that $\|v_j\| = 1$.

$$\hat{c}_j = \text{sign}(\langle \hat{v}_j, v_j \rangle)$$

$$\int \hat{c}(s,t)\hat{v}_j(s)ds = \hat{\lambda}_j \hat{v}_j(t), \quad j = 1, 2, \ldots, N.$$

Using the above ideas we will construct **optimal empirical orthonormal basis** for our signal $\{X(t);\ t \in \mathbb{Z}\}$ represented by random elements $X_1, \ldots, X_N$. In the context of the spectrogram $X_1, \ldots, X_N$ can be seen as the vertical stripes.

Suppose we observe functions $x_1, x_2, \ldots, x_N$. Fix an integer $\mathbb{Z} \ni p < N (p \ll N)$. We want to find an orthonormal basis $u_1, u_2, \ldots, u_p$ such that

$$\hat{S}^2 = \sum_{i=1}^{N} \left\| x_i - \sum_{k=1}^{p} \langle x_i, u_k \rangle u_k \right\|^2$$

is minimum.

Empirical basis

$$\boldsymbol{x_i} = [\langle x_i, u_1 \rangle, \langle x_i, u_2 \rangle, \ldots, \langle x_i, u_p \rangle]^T.$$

The functions u_j are called collectively the optimal empirical orthonormal basis or natural orthonormal components.

Empirical basis and covariance

The functions $u_1, u_2, \ldots, u_p$ minimizing $\hat{S}^2$ are equal (up to a sign) to the normalized eigenfunctions of the corresponding sample covariance operator.

We have

$$\hat{S}^2 = \sum_{i=1}^{N} \left(\|x_i\|^2 - \sum_{k=1}^{p} \langle x_i, u_k \rangle^2 \right)$$

$\hat{S}^2$ is minimum, when $\sum_{i=1}^{N} \sum_{k=1}^{p} \langle x_i, u_k \rangle^2$ is maximum.

$$\begin{aligned} \sum_{i=1}^{N} \sum_{k=1}^{p} \langle x_i, u_k \rangle^2 &= N \sum_{k=1}^{p} \langle \hat{C}(u_k), u_k \rangle \\ &= N \sum_{k=1}^{p} \sum_{j=1}^{\infty} \hat{\lambda}_j \langle u_k, \hat{v}_j \rangle^2 \leq N \sum_{k=1}^{p} \hat{\lambda}_k \end{aligned}$$

maximum is attained if $u_1 = \hat{v}_1, u_2 = \hat{v}_2, \ldots, u_p = \hat{v}_p$.

Dimensionality reduction can be achieved by

- Constructing the empirical basis
- Choosing the number of components p such that the model will exhaust the most important part of the energy (variance/covariance) of the signal
- Working with eigenvalues instead of many functions

Choosing p

To this end we can consider the function

$$CPV(p) = \frac{\sum_{i=1}^{p} \hat{\lambda}_i}{\sum_{i=1}^{N} \hat{\lambda}_i}$$

F-AR(1) model

Our starting point is again a sequence of Hilbert space valued random elements $X_1, \ldots, X_N$ that no longer are assumed independent. In the spectrogram representation, it is NOT realistic to assume that vertical stripes are independent.
Consider the model

F-AR(1)

$$X_n = \Psi(X_{n-1}) + \varepsilon_n$$

where $\Psi \in \mathcal{L}$ while $\mathcal{L}$ is the space of bounded continuous linear operators on L^2 equipped with the sup norm. Moreover, ε_n is a sequence of iid mean zero elements in L^2.

F-AR(1) model

It is known that under appropriate conditions (see Horvath, Kokoszka (2012)) we have that F-AR(1) is causal and strictly stationary.

Example of Ψ

Consider

$$\Psi(x)(t) \stackrel{def}{=} \int \psi(t,s)x(s)ds$$

where $x \in L^2$ and $\int\int \psi^2(t,s)dtds < 1$.

Estimation in F-AR(1)

Define the lag 1 autocovariance operator

$$C_1(x) = \mathbb{E}[\langle X_n, x\rangle\ X_{n+1}],\ x \in L^2$$

Like in the scalar case, we have the relationship

$$C_1 = \Psi C$$

where C is the covariance operator. Thus, to estimate Ψ we could define

$$\hat{\Psi} = \hat{C}_1 \hat{C}^{-1}.$$

Warning: getting $\hat{C}^{-1}$ may be difficult. However, we will use the empirical basis principle and take only p first components.

Estimation in F-AR(1)

Instead, we use

$$\widehat{IC}_p(x) = \sum_{j=1}^{p} \widehat{\lambda}_j^{-1} \langle x, \widehat{v}_j \rangle \widehat{v}_j.$$

We get: $\widehat{C}_1(x) = \frac{1}{N-1} \sum_{k=1}^{N-1} \langle X_k, x \rangle X_{k+1}$

For any $x \in L^2$ obtain

$$\widehat{C}_1 \widehat{IC}_p(x) = \frac{1}{N-1} \sum_{k=1}^{N-1} \sum_{j=1}^{p} \widehat{\lambda}_j^{-1} \langle x, \widehat{v}_j \rangle \langle X_k, \widehat{v}_j \rangle X_{k+1}.$$

The estimate

$$\widehat{\Psi}_p(x) = \frac{1}{N-1} \sum_{k=1}^{N-1} \sum_{j=1}^{p} \sum_{i=1}^{p} \widehat{\lambda}_j^{-1} \langle x, \widehat{v}_j \rangle \langle X_k, \widehat{v}_j \rangle \langle X_{k+1}, \widehat{v}_i \rangle \widehat{v}_i.$$

P-AR(1) and FP-AR model

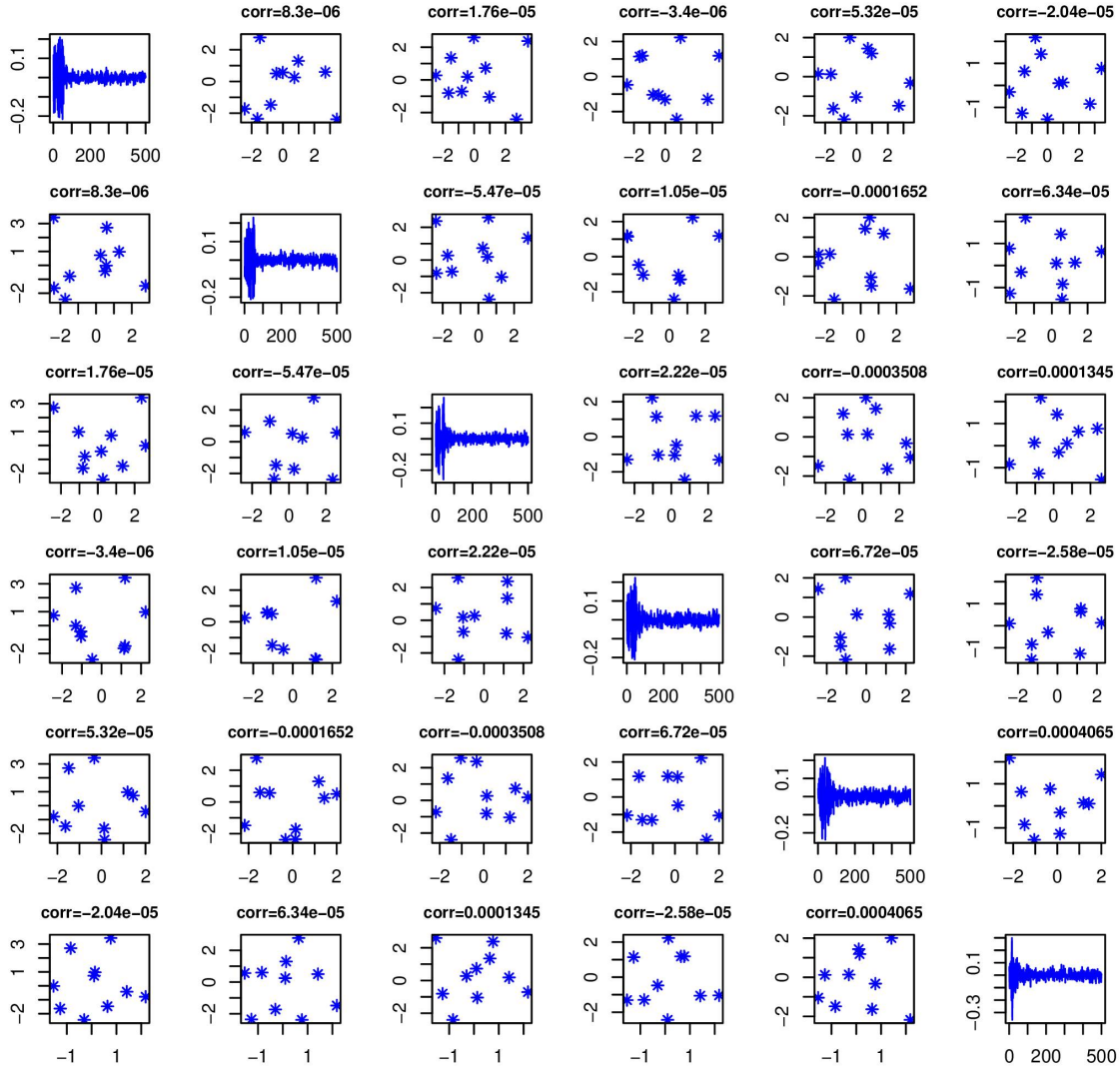
The cyclostationary generalization of the usual AR(1) model $y_t = \phi \cdot y_{t-1} + \epsilon_t$ is the P-AR(1) model $y_t = \phi(t) \cdot y_{t-1} + \epsilon_t$, where $\phi(t)$ is assumed to be periodic with the period P. The estimation P-AR(1) model can be solved by stacking up the original data into vectors of the length P and writing a vector AR(1) model for them. The same trick can be done in the F-AR(1) model to compensate for cyclostationarity of the functional data. Consider:

FP-AR(1)

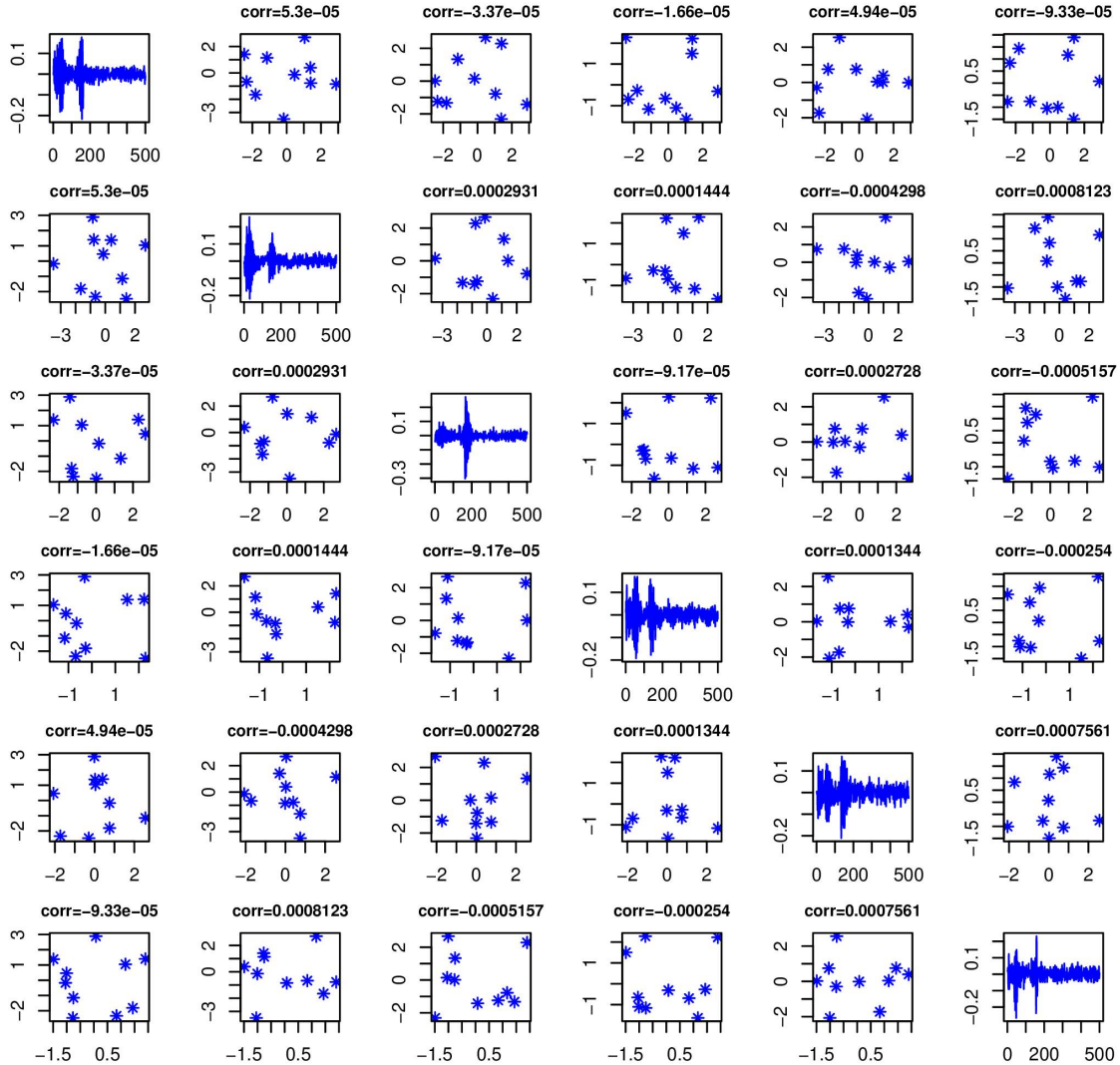
$$X_n = \Psi(n)(X_{n-1}) + \varepsilon_n$$

where $\Psi(n+P) = \Psi(n)$ and for each i $\Psi(i) \in \mathcal{L}$ while $\mathcal{L}$ is the space of bounded continuous linear operators on L^2 equipped with the sup norm. Moreover, ε_n is a sequence of iid mean zero elements in L^2.

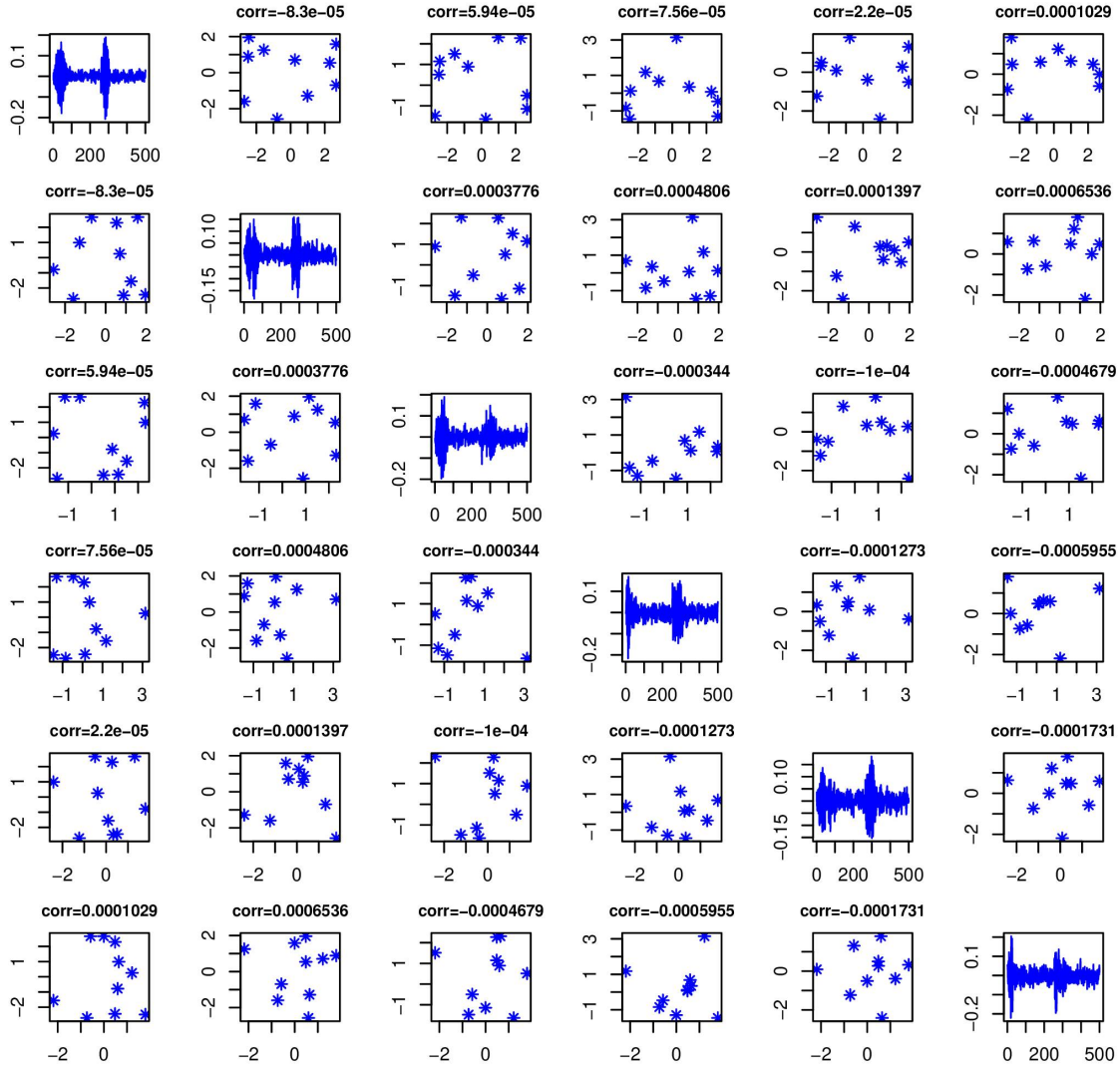
Eigenvalues and scores for the first group



Eigenvalues and scores for the second group



Eigenvalues and scores for the third group



Eigenvalues provide a signature for the PC functional signal

First group

eigenvalue	block bootstrap		CPV
0.006410	$8.330787 \cdot 10^{-3}$	$14.529901 \cdot 10^{-3}$	21%
0.005281	$5.299198 \cdot 10^{-3}$	$9.958018 \cdot 10^{-3}$	37%
0.004496	$3.760670 \cdot 10^{-3}$	$6.412833 \cdot 10^{-3}$	52%
0.004096	$2.061348 \cdot 10^{-17}$	$4.677168 \cdot 10^{-3}$	65%
0.003605	$1.263881 \cdot 10^{-17}$	$3.260474 \cdot 10^{-3}$	76%
0.002478	$9.668592 \cdot 10^{-18}$	$2.404282 \cdot 10^{-3}$	84%

Eigenvalues - cntd.

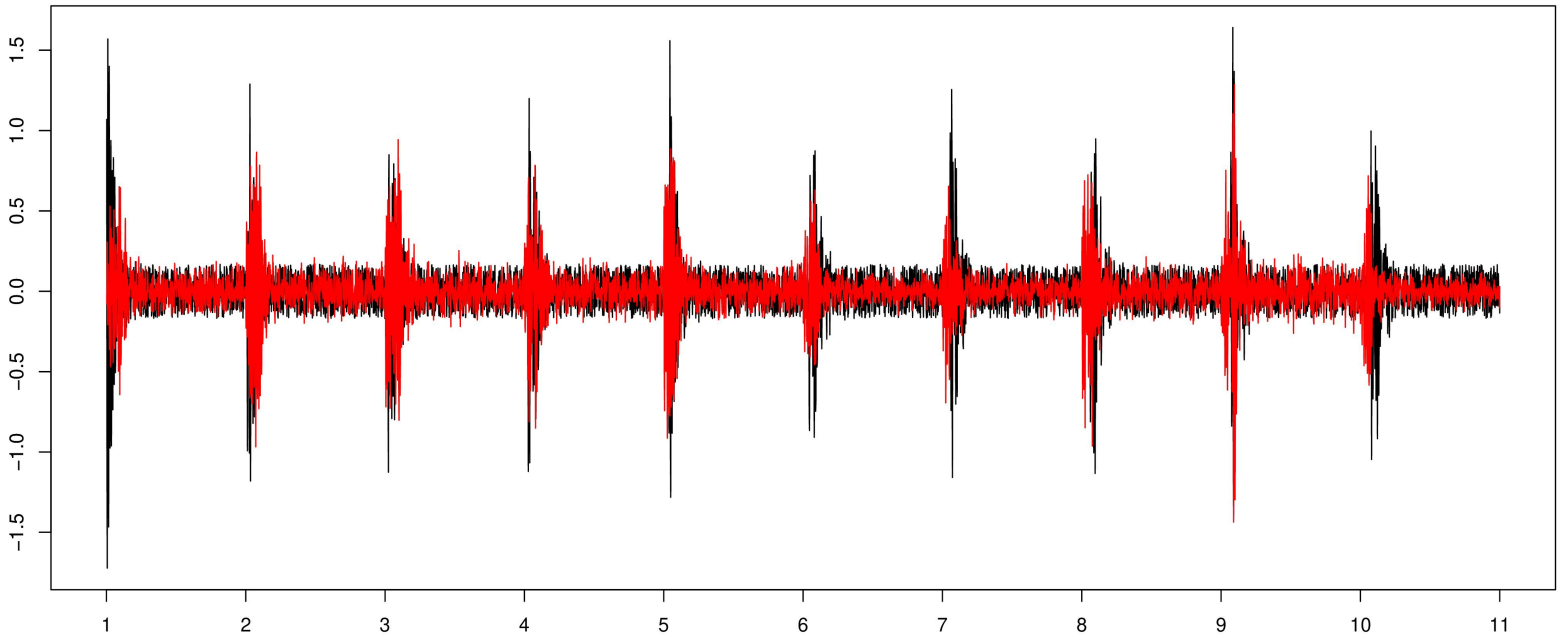
Second group

eigenvalue	block bootstrap		CPV
0.005929	$6.875680 \cdot 10^{-3}$	$12.830428 \cdot 10^{-3}$	19%
0.005410	$5.304248 \cdot 10^{-3}$	$9.620010 \cdot 10^{-3}$	35%
0.005081	$3.641856 \cdot 10^{-3}$	$6.456256 \cdot 10^{-3}$	51%
0.003749	$2.379073 \cdot 10^{-17}$	$4.252976 \cdot 10^{-3}$	63%
0.003033	$1.242109 \cdot 10^{-17}$	$3.477051 \cdot 10^{-3}$	73%
0.002640	$9.368904 \cdot 10^{-18}$	$2.669335 \cdot 10^{-3}$	81%

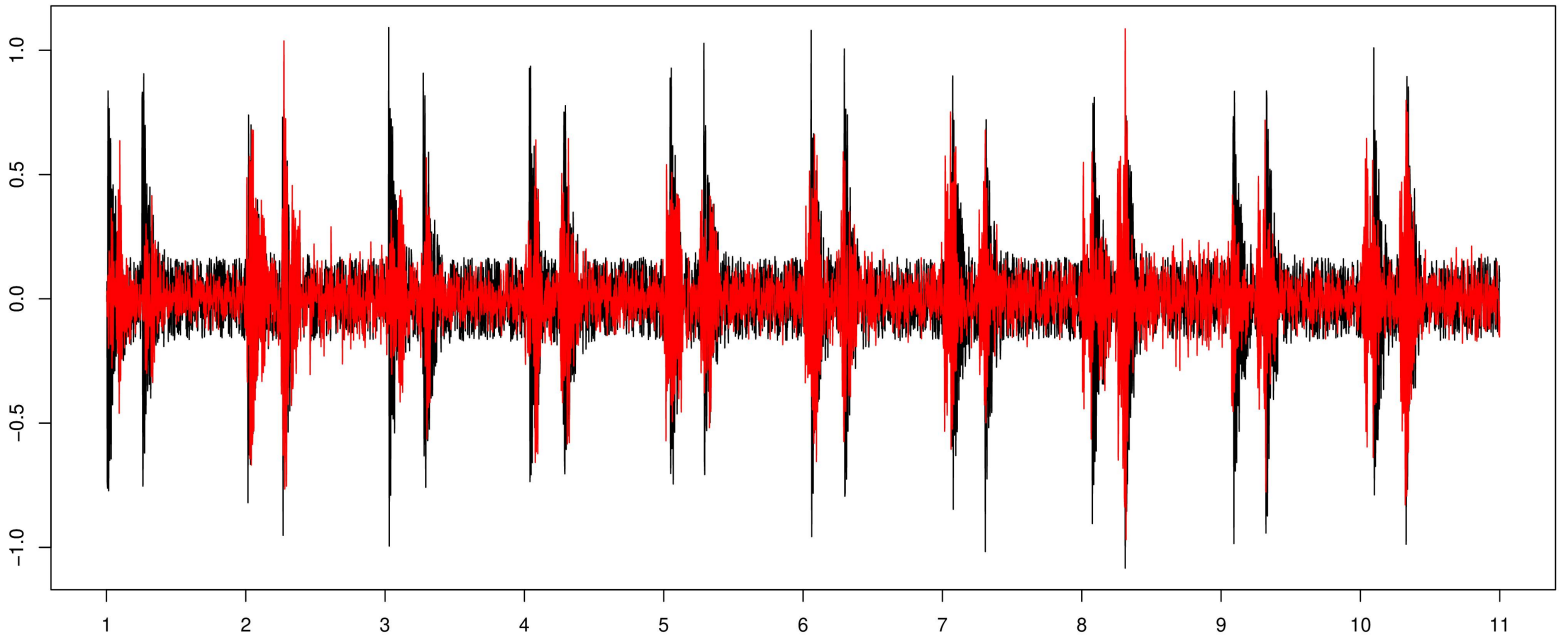
Third group

eigenvalue	block bootstrap		CPV
0.008586	$7.324938 \cdot 10^{-3}$	$14.848730 \cdot 10^{-3}$	29%
0.004144	$3.991475 \cdot 10^{-3}$	$8.019414 \cdot 10^{-3}$	43%
0.004021	$2.313940 \cdot 10^{-17}$	$5.214372 \cdot 10^{-3}$	56%
0.003285	$1.353050 \cdot 10^{-18}$	$3.755904 \cdot 10^{-3}$	67%
0.002625	$9.304354 \cdot 10^{-18}$	$2.966601 \cdot 10^{-3}$	76%
0.002261	$7.704712 \cdot 10^{-18}$	$2.324539 \cdot 10^{-3}$	84%

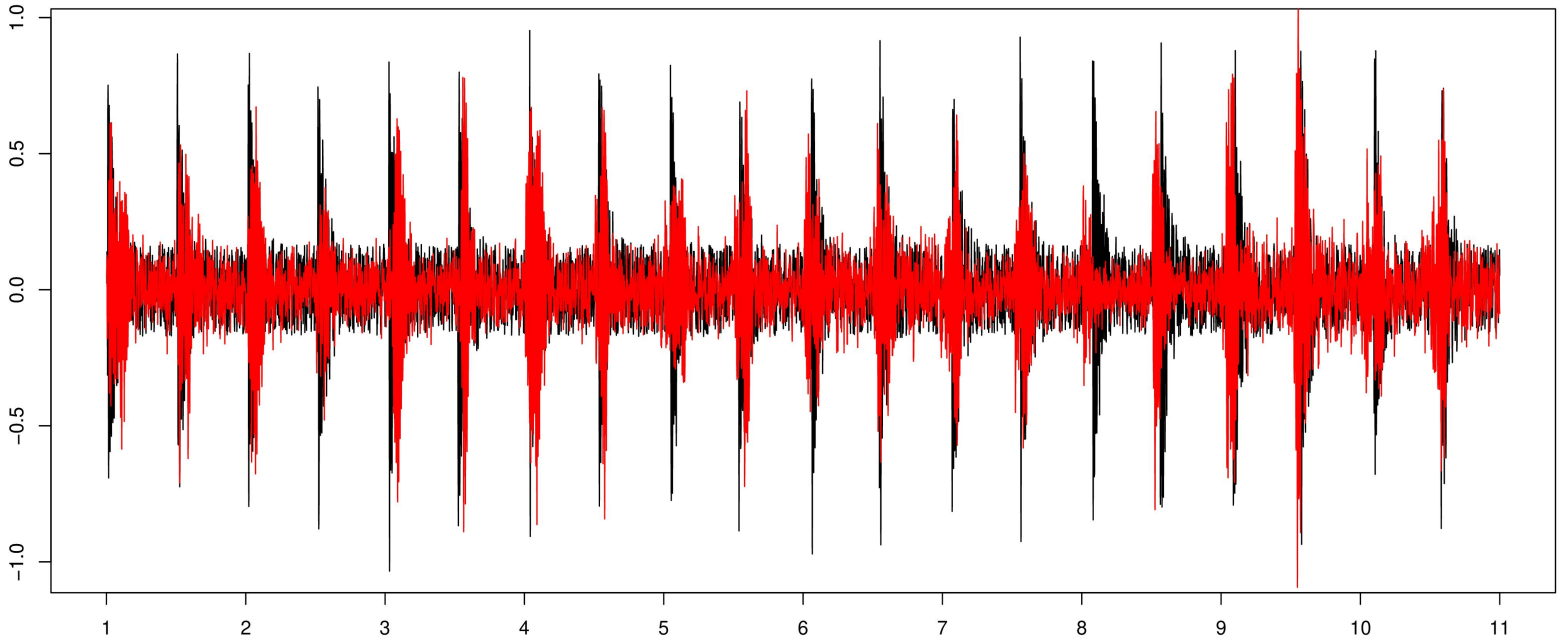
Signal reconstruction - first group



Signal reconstruction - second group



Signal reconstruction - third group



Some open questions:

- APC models from FDA perspective
- Solid limit theory approach for the estimators
- Validity of bootstrap

Thank you for your attention