Functional data analysis in signal processing - theory and applications

Jacek Leśkow

American University Kyiv
and
Cracow University of Technology

Chapel Hill, March 13, 2025

Plan of the talk I

Abstract

Main goals of the talk:

- Presenting cyclostationary signals
- Models for cyclostationary signals, resampling, mechanical signals
- Functional models

Motivating example

Engine signal (Lafon, Antoni, Sidahmed, Polac, Journal of Sound and Vibration (2011))

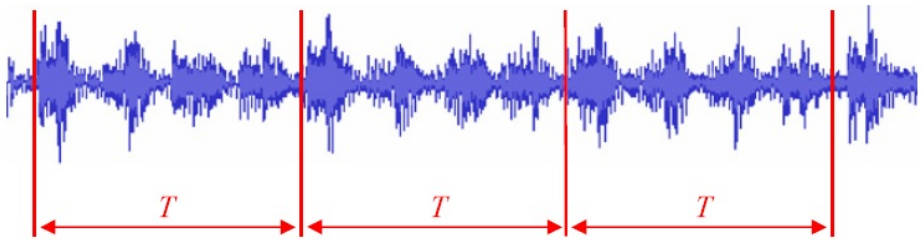
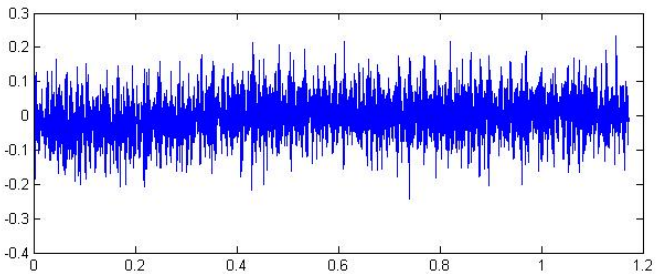
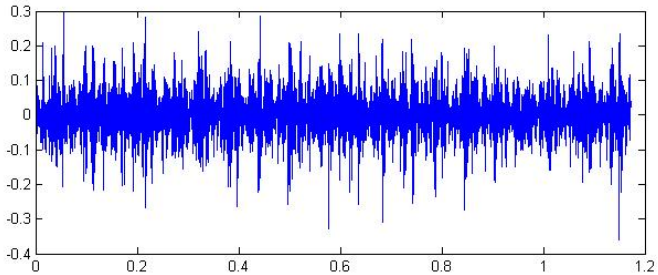


Fig. 1. Example of three cycles of a cyclostationary acoustical signal recorded in front of a 4-cylinder engine under steady operating conditions

Motivating example

Wheel bearing signal - normally operating and inner race default





Motivating example

Brown coal (lignite) excavating machine

Definition of nonstationary APC time series

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

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

.

Inference for APC

If $\{X(t);\ t \in \mathbb{Z}\}$ - APC then

$$\mu_X(t) = E(X_t)$$

and the autocovariance function

$$B_X(t,\tau) = \sum_{\lambda \in \Lambda} a(\lambda,\tau)\, e^{i\lambda t}.$$

Assume for simplicity that $\mu_X(t) \equiv 0$). Then

$$\hat{a}_n(\lambda,\tau) = \frac{1}{n-\tau} \sum_{t=1}^{n-\tau} X(t+\tau)\, X(t)\, e^{-i\lambda t}$$

Given some mixing assumptions $\hat{a}_n(\lambda,\tau)$ is asymptotically normal. Variance-covariance matrix is very complicated.

Theorem - Dehay, Dudek, Leśkow (2014))

Consider the following conditions:

(A1) $\sup_t E\left\{|X(t)|^{4+\delta}\right\} < \infty$ for some $\delta > 0$, the fourth moment is almost periodic in the following sense : the function $v \mapsto \mathrm{cov}\{X(u+v+\tau)X(u+v), X(v+\tau)X(v)\}$ is almost periodic for each u. Moreover the process X is α-mixing and the mixing coefficient satisfies $\int_0^\infty \alpha_X(t)^{\delta/(4+\delta)}\, dt < \infty$,

(A2) For each $\lambda \in \Lambda$ the following separability property is fulfilled

$$\sum_{\lambda' \in \Lambda \setminus \{\lambda\}} \left| \frac{a(\lambda', \tau)}{\lambda' - \lambda} \right| < \infty.$$

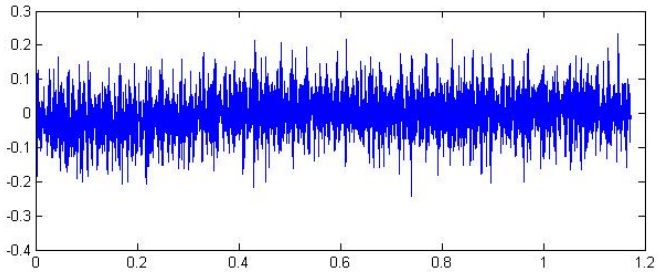
Under the following conditions we have

$$\sqrt{T}\{\widehat{a}_T(\lambda, \tau) - a(\lambda, \tau)\} \xrightarrow{\mathcal{L}} \mathcal{N}_2(0, V(\lambda, \tau)).$$

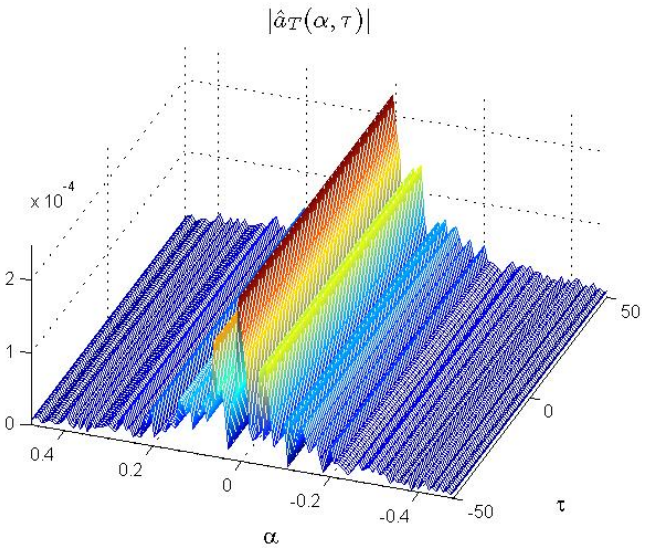
where the limiting V is very complicated!

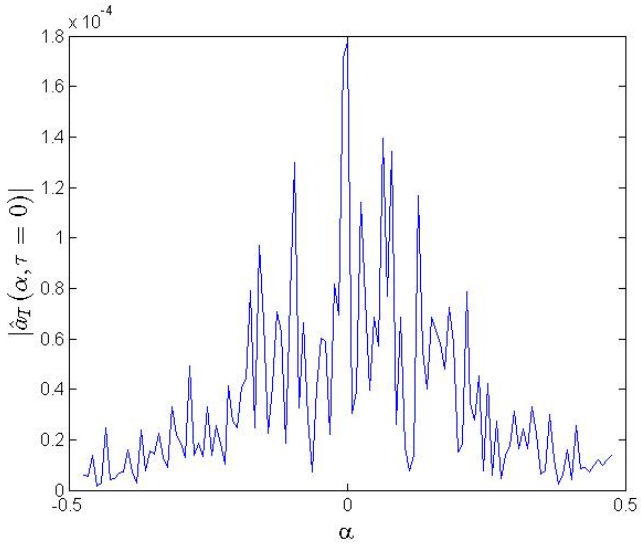
Ball bearing signals

Properly working ball bearing



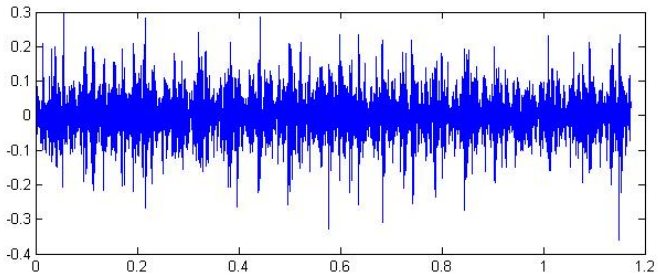
autocovariance structure



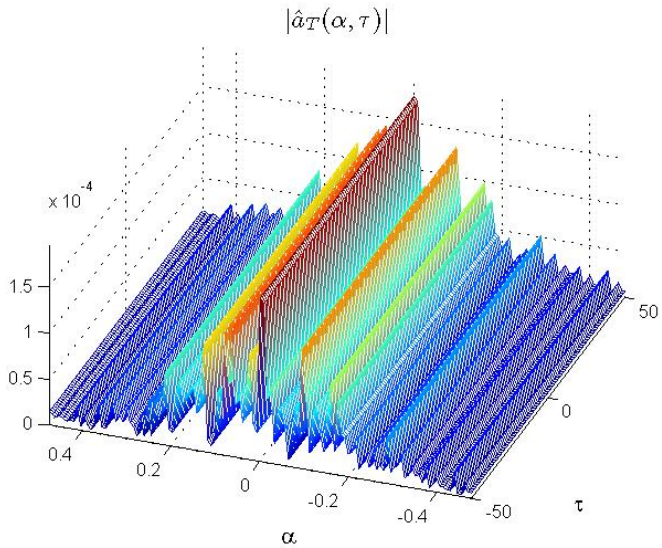


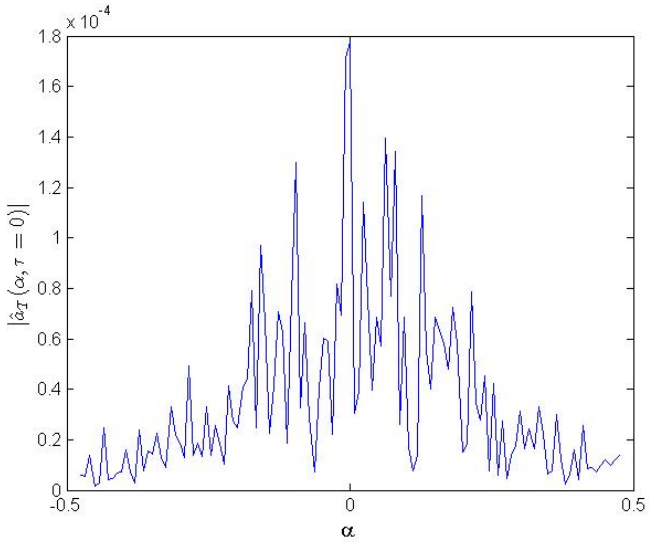
Ball bearing signals

Ball bearing with rolling element damaged



autocovariance structure





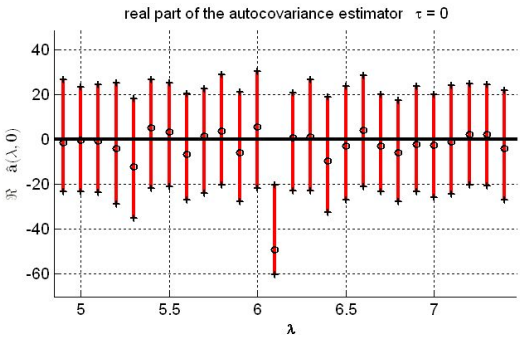
Resampling for time series

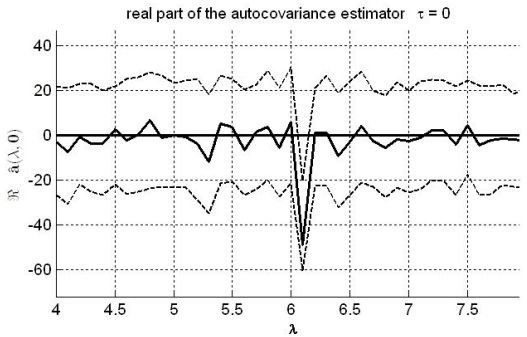
Quick facts:

- Nonparametric bootstrap - useless, as it destroys the longitudinal structure
- Stationary time series: moving block bootstrap (MBB)
- Cyclostationary time series: periodic block bootstrap (PBB), seasonal block bootstrap (GSBB), circular versions of MBB and GSBB, subsampling
- asymptotic independence: mixing or weak dependence

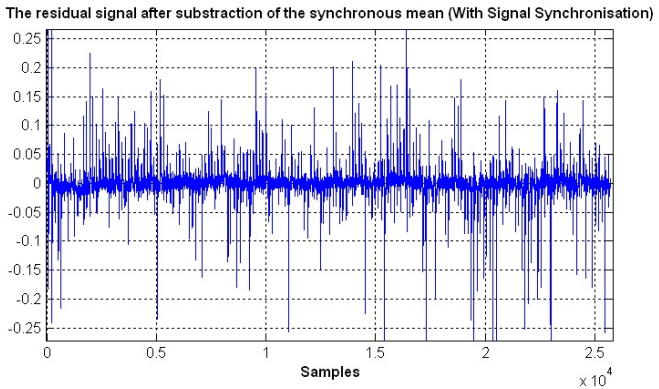
Resampling in practice

Typical cyclostationary signal analysis.





Underlying residual signal



PFAR(1) model

In our considerations we will assume that the observed signal

$$\{X(s) : s \in [0, S]\}$$

can be represented by a sequence of random functions (curves):

$$\{Z_1(\cdot), \ldots, Z_N(\cdot)\}.$$

Since we are going to work with cyclostationary signals, we will assume that these curves are correlated and are repeatable that is there is a period T such that $Z_{nT+\nu}$ will be similar to $Z_{(n-1)T+\nu}$ for all $n \geq 1$ and for each ν, where $\nu = 1, \ldots, T$. The sense of this similarity will be specified later.

Model space

Without a loss of generality, we will assume that the random functions $Z_i(s)$ are defined on a common interval $[0,1]$ so $s \in [0,1]$ throughout our presentation.
Without a loss of generality, we will identify $\mathcal{H}$ with $L^2[0,1]$. We will also use the symbol Φ to denote the linear operator on $\mathcal{H}$ with values in $\mathcal{H}$, that is $\Phi \in \mathcal{L}(\mathcal{H},\mathcal{H})$. On the Hilbert space $\mathcal{L}(\mathcal{H},\mathcal{H})$

Model

We will say that the functional time series $\{Z_i, i \geq 1\}$ fulfills the PFAR(1) model if for each $i = nT + \nu$ with $\nu = 1, \ldots T$ we have

$$Z_{nT+\nu} = \Phi_\nu(Z_{nT+(\nu-1)}) + \varepsilon_{nT+\nu}. \tag{1}$$

In the model (1) the operators $\Phi_\nu, \nu = 1, \ldots, T$ are Hilbert-Schmidt integral operators in L^2 with corresponding kernels ϕ_ν fulfilling the assumption

$$\Phi_\nu(x)(t) = \int \phi_\nu(t,s)x(s)ds.$$

Moreover, the sequence ε_i is a sequence of i.i.d. mean zero elements in $\mathcal{H}$.

Estimator

The estimator is based on the first p functional principal components and has the form:

$$\hat{\Phi}_{\nu,n} = \hat{C}_{1,\nu,n}(\hat{\Gamma}^{p}_{\nu,n})^{-1} \tag{2}$$

The above formula can be presented in a more technical form:

$$\widehat{\Phi}_{\nu,n}(x) = \frac{1}{n-1}\sum_{k=0}^{n-1}\sum_{j=1}^{p}\widehat{\lambda}^{-1}_{\nu,j}\langle x, \widehat{\eta}_{\nu,j}\rangle\langle Z_{kT+\nu}, \widehat{\eta}_{\nu,j}\rangle Z_{kT+\nu+1} \tag{3}$$

Assumptions

The statistical inference in our PFAR model will be based on the following assumptions .

Assumption A1

There exists an integer j_0 such that for each $\nu, \nu = 1, \ldots, T$ we have $||\Phi_\nu^{j_0}||_{\mathcal{L}} < 1$.

The assumption provides the causal representation for our periodic time series $\{Z_{nT+\nu}\}$ in the following form

$$Z_{nT+\nu} = \sum_{j=0}^{\infty} \Phi_\nu^j(\epsilon_{nT+\nu-j}), \quad \nu = 1, \ldots, T. \tag{4}$$

Assumptions

Assumption A2

Assume that for each n and ν we have $E\|Z_{nT+\nu}\|^4 \leq \infty$.

Assumptions

Assumption A3

Assume that for $n \to \infty$ we have that $p \to \infty$ but $\lim_{n\to\infty} \frac{p}{n} = 0$.

The condition means that the number of principal components grows with the growth of the sample size but the rate of growth is slower than the growth of the sample.

Mean square constistency

Theorem

Let the sequence $Z_{nT+\nu}$ follow the model (1) and let assumptions **A1**, **A2** and **A3** be fulfilled. Let the sequence of random operators $\hat{\Phi}_{\nu,n}$ be defined in (3). Then, the following holds true:

(i) $E\|\hat{\Phi}_{\nu,n} - \Phi_\nu\|^2 = O(n^{-1})$ for each $\nu = 1, \ldots, T$,

(ii) $E\|\hat{\eta}_{\nu,n} - \eta_\nu\|^2 = O(n^{-1})$ for each $\nu = 1, \ldots, T$,

(iii) $E|\hat{\lambda}_{\nu,n} - \lambda_\nu|^2 = O(n^{-1})$ for each $\nu = 1, \ldots, T$.

Asymptotic normality

Theorem

Let the sequence $Z_{nT+\nu}$ follow the model (1) and let assumptions **A1**, **A2** and **A3** be fulfilled. Then

$$\sqrt{n}(\hat{\lambda}_{\nu,n} - \lambda_\nu) \xrightarrow{d} \mathcal{N}(0, \sigma_\nu),$$

where $\mathcal{N}(0, \sigma_\nu)$ is a univariate normal distribution with the mean zero and the standard deviation σ_ν. The symbol $\xrightarrow{d}$ denotes the convergence in distribution and λ_ν is the eigenvalue of the operator Γ_ν.

Bootstrap algorithm

Step 1. Vector stationarization of Z_i

Let $\mathsf{Z_n} = (Z_{n\mathsf{T}+1}, \ldots, Z_{n\mathsf{T}+\mathsf{T}})'$. Since Z_i follows the PFAR(1) model thus $\mathsf{Z_n}$ is T-variate stationary. Let $\hat{\lambda}_{\nu,n}$ be the estimates of the theoretical eigenvalues λ_ν based on a sample $\mathsf{Z}_1, \ldots, \mathsf{Z_n}$.

Bootstrap algorithm

Step 2. Creating the first MBB resample

Assume that $n = kb$ and let $\{B_1, \ldots, B_k\}$ be the nonoverlapping blocks of the length b covering $Z_1, \ldots, Z_n$. We sample with replacement k blocks $\{B_1^{*1}, \ldots, B_k^{*1}\}$ to get the first MBB resample $Z_1^{*1}, \ldots, Z_n^{*1}$ and the first resample $\hat{\lambda}_{\nu,n}^{*1}$.

Step 3. Getting L replicates

Repeat Step 2 L times to get L replicates $\hat{\lambda}_{\nu,n}^{*1}, \ldots, \hat{\lambda}_{\nu,n}^{*L}$.

Consistency

Using the procedure described above we arrive at the following consistency theorem.

Theorem 4

The confidence intervals derived from the MBB replications $\hat{\lambda}^{*1}_{\nu,n}, \ldots, \hat{\lambda}^{*L}_{\nu,n}$ are asymptotically consistent that is for each $\nu = 1, \ldots, T$

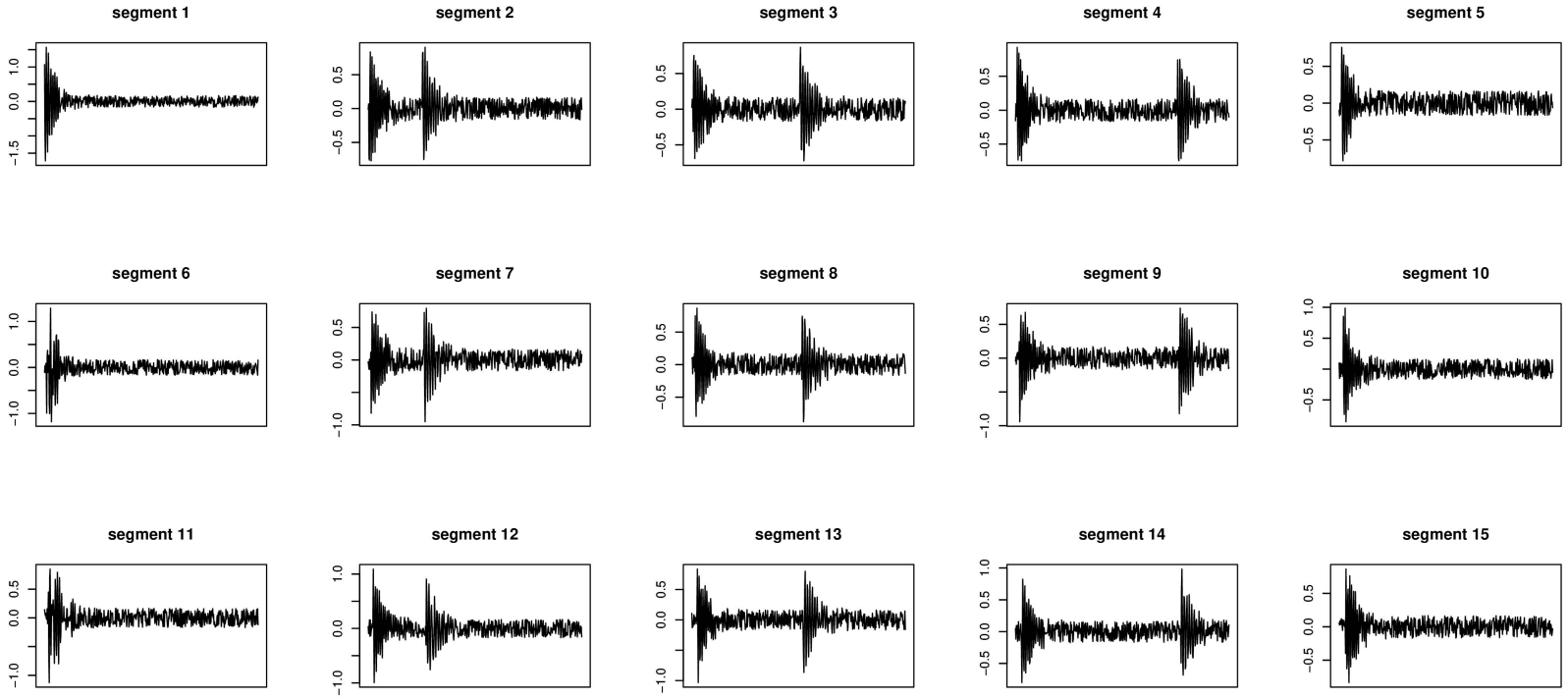
$$P\{\lambda_\nu \in (q_{\alpha/2}(L,\nu), q_{1-\alpha/2}(L,\nu))\} \longrightarrow 1 - \alpha$$

where $q_{\alpha/2}(L,\nu)$ is the $\alpha/2$-quantile generated by the bootstrap sample $\hat{\lambda}^{*1}_{\nu,n}, \ldots, \hat{\lambda}^{*L}_{\nu,n}$. The convergence is for $n \longrightarrow \infty$ under the assumption of stationarity and α-mixing.

Description of the data

We will analyze now another cyclostationary signal that is generated by the wheel bearing. We have a signal corresponding to 10 seconds of the wheel bearing operation, the frequency being 2,5 kHz. Having, therefore, 25000 time data points we have divided them into 50 equal segments with 500 data points in each segment. Therefore, each second contained 5 segments with periodicity observed between first segments of each second, second segment of each second and so on. Our choice is motivated by observed periodicity among data grouped into 50 segments.

Data organized into segments

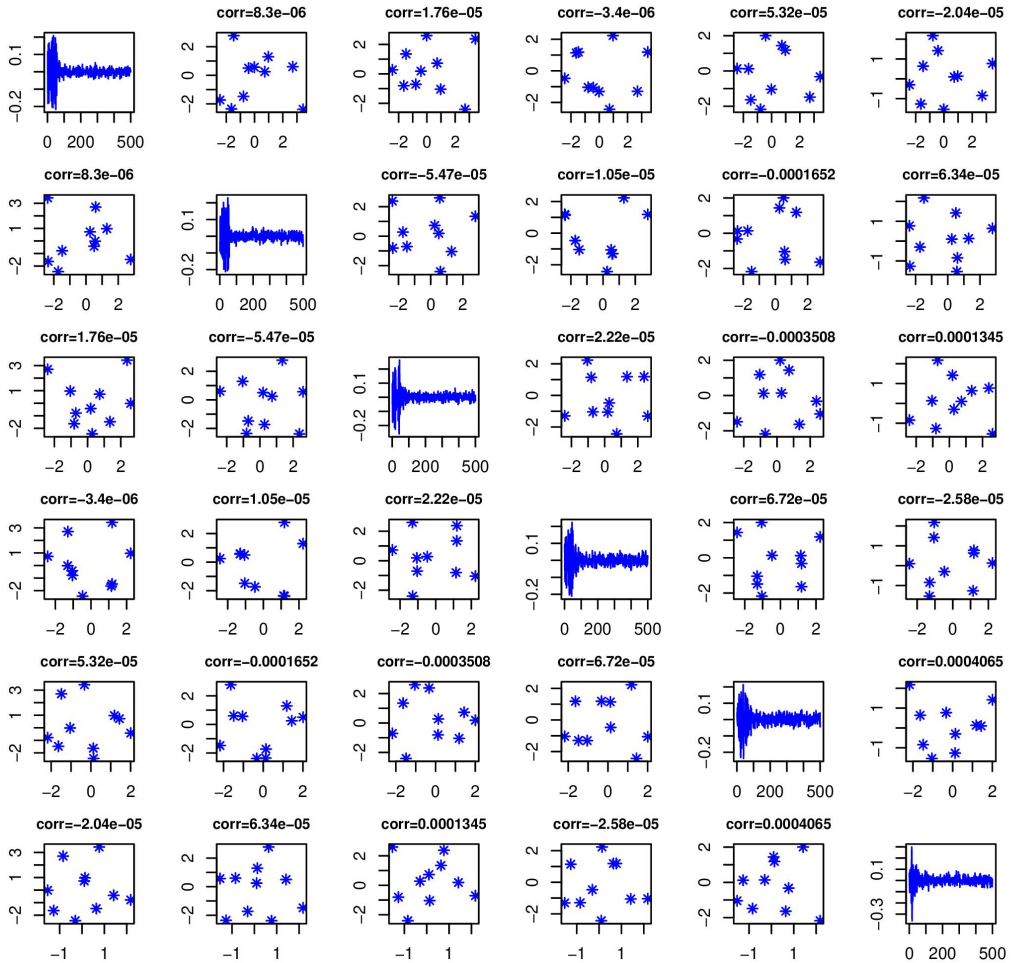


The segments with numbers 1,6,11 and so on look pretty much alike. The same can be said about segments with numbers 2,7,12 and so on. Clearly, we have periodicity equal to 5 and we can treat the data as the periodic functional time series with the period length 5. We will therefore group data according to columns and perform the estimation procedure using PFAR(1) model. We will have therefore five groups of curves. The first group contains 10 first segments of each of second, the second group contains the 10 second segments and so on.

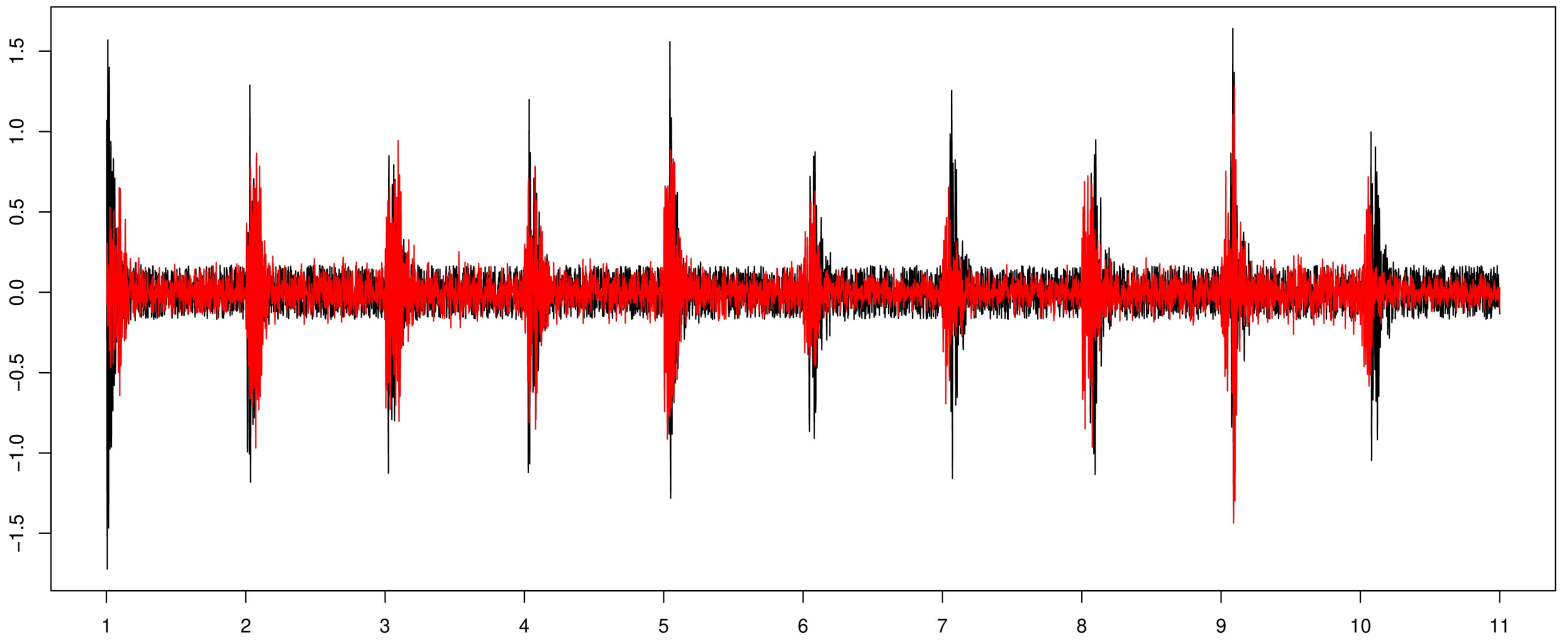
First group eigenvalues

eigenvalue	block bootstrap ($k = 5, n = 2, L = 1000$)		CPV
0.006410	$6.410022 \cdot 10^{-3}$	$15.042201 \cdot 10^{-3}$	21%
0.005281	$5.037865 \cdot 10^{-3}$	$8.210638 \cdot 10^{-3}$	37%
0.004496	$3.334865 \cdot 10^{-3}$	$6.538219 \cdot 10^{-3}$	52%
0.004096	$1.476030 \cdot 10^{-17}$	$4.322919 \cdot 10^{-3}$	65%
0.003605	$1.234785 \cdot 10^{-17}$	$3.604941 \cdot 10^{-3}$	76%
0.002478	$9.018584 \cdot 10^{-18}$	$2.555049 \cdot 10^{-3}$	84%

First group diagnostics



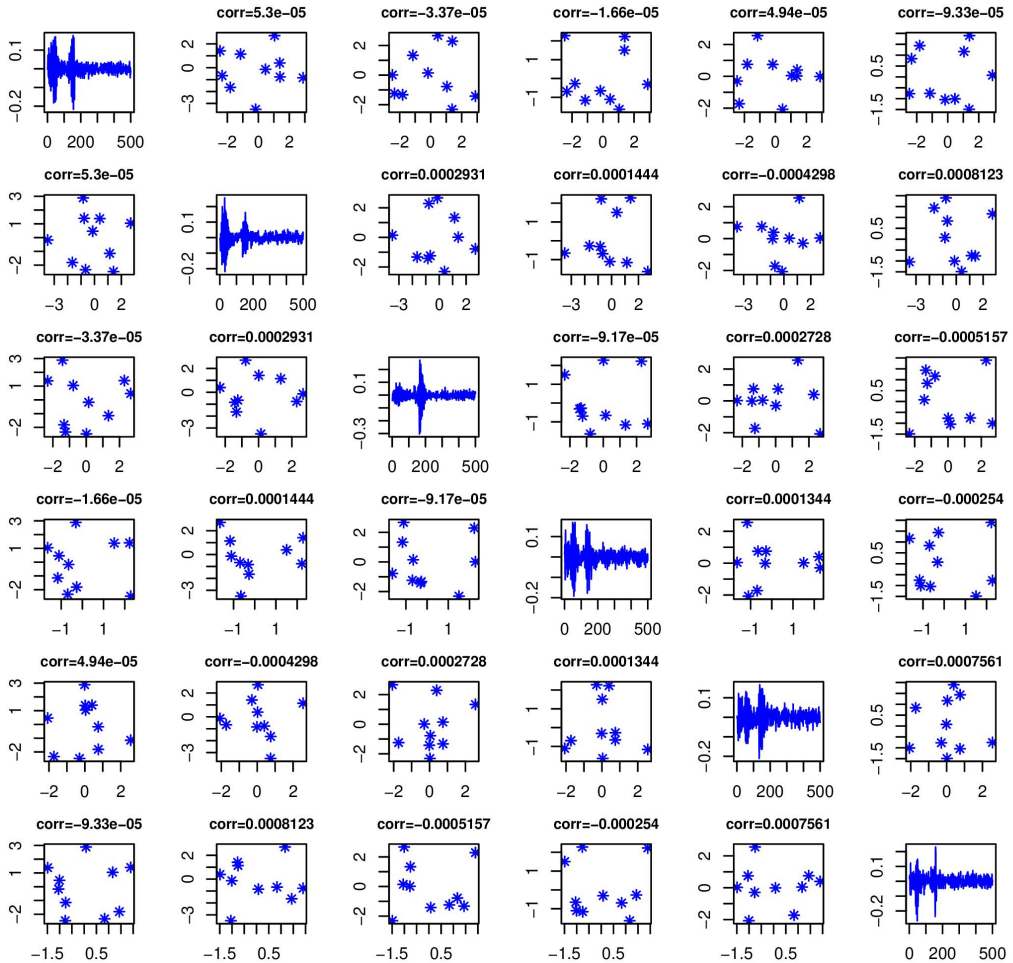
First group model fit



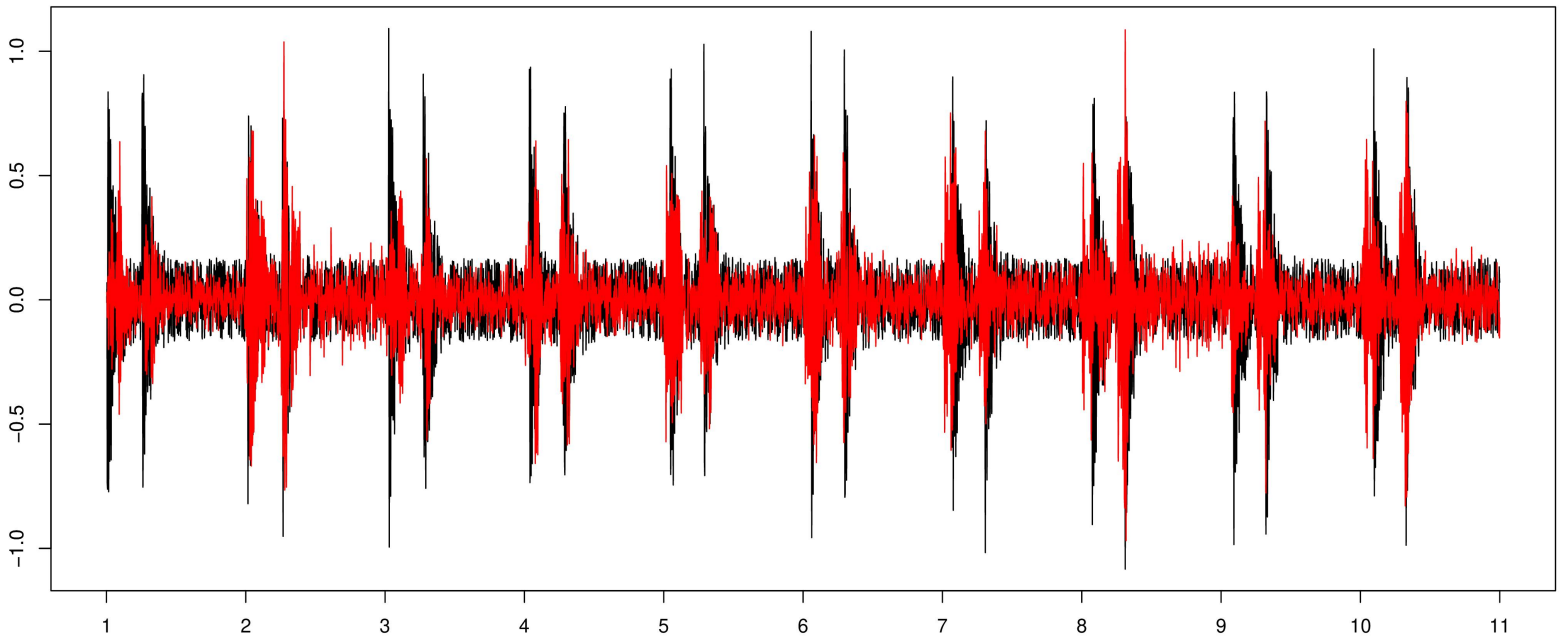
Second group eignevalues

eigenvalue	block bootstrap($k = 5, n = 2, L = 1000$)		CPV
0.005929	$5.929743 \cdot 10^{-3}$	$13.553896 \cdot 10^{-3}$	19%
0.005410	$5.082175 \cdot 10^{-3}$	$8.584699 \cdot 10^{-3}$	35%
0.005081	$3.424612 \cdot 10^{-3}$	$6.624032 \cdot 10^{-3}$	51%
0.003749	$1.548226 \cdot 10^{-17}$	$4.429500 \cdot 10^{-3}$	63%
0.003033	$1.262034 \cdot 10^{-17}$	$3.452723 \cdot 10^{-3}$	73%
0.002640	$9.301247 \cdot 10^{-18}$	$2.730186 \cdot 10^{-3}$	81%

Second group diagnostics



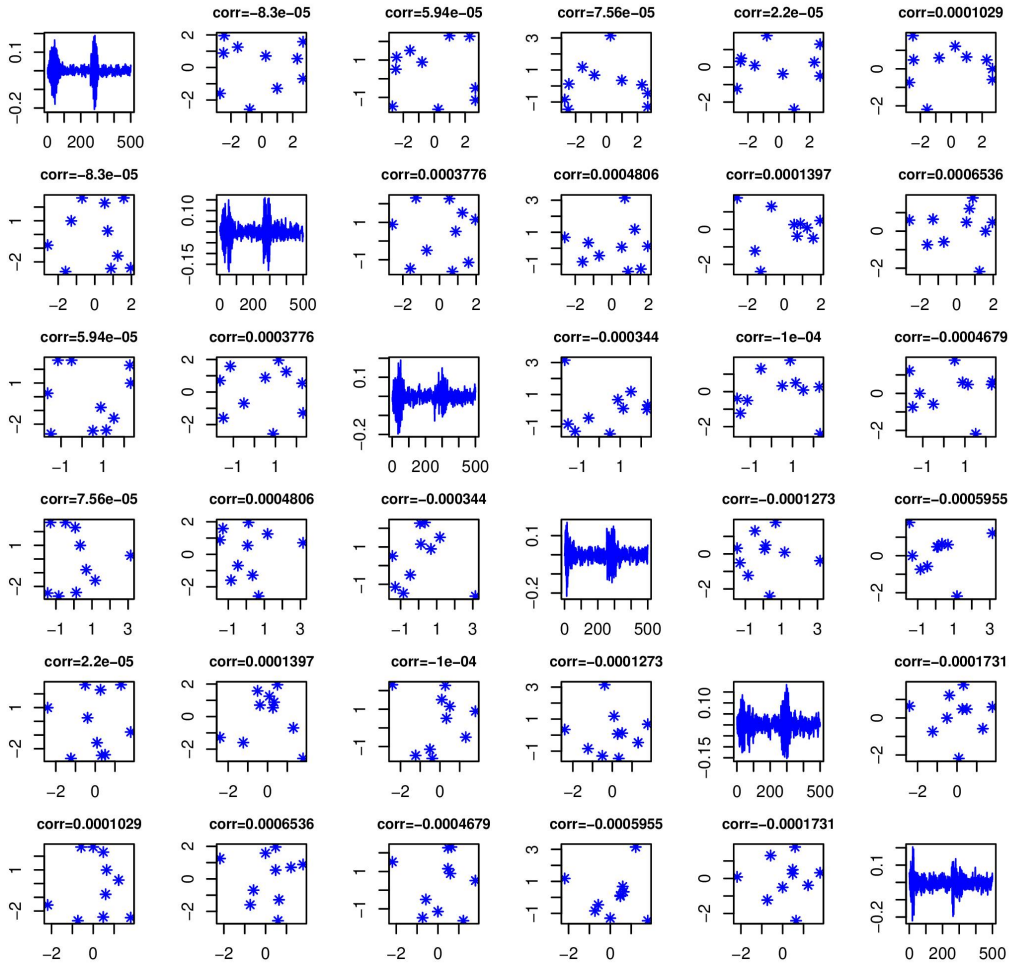
Second group model fit



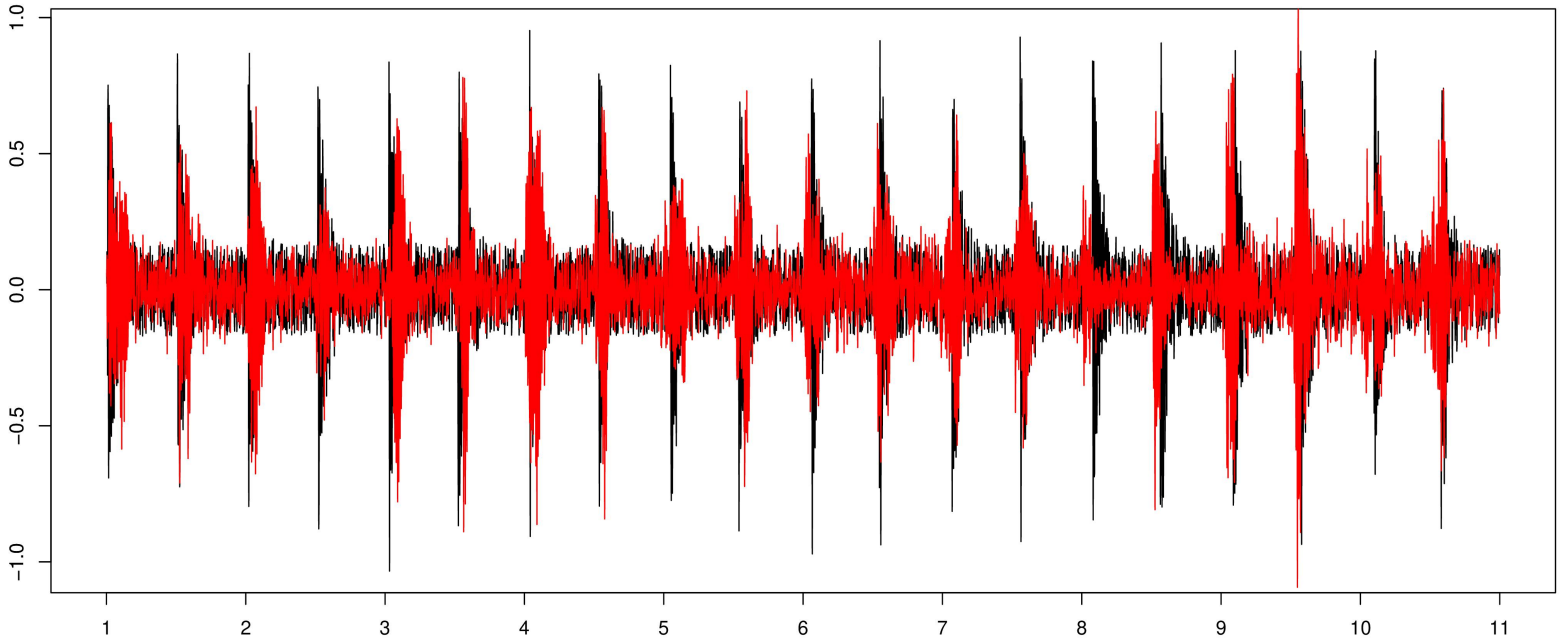
Third group eignevalues

eigenvalue	block bootstrap ($k = 5, n = 2, L = 1000$)		CPV
0.008586	$7.377949 \cdot 10^{-3}$	$15.318479 \cdot 10^{-3}$	29%
0.004144	$4.141210 \cdot 10^{-3}$	$7.915194 \cdot 10^{-3}$	43%
0.004021	$2.876588 \cdot 10^{-3}$	$5.730514 \cdot 10^{-3}$	56%
0.003285	$1.136990 \cdot 10^{-17}$	$4.039814 \cdot 10^{-3}$	67%
0.002625	$9.868712 \cdot 10^{-18}$	$2.934986 \cdot 10^{-3}$	76%
0.002261	$7.633720 \cdot 10^{-18}$	$2.431313 \cdot 10^{-3}$	84%

Third group diagnostics



Third group model fit



References- selected

- Cioch, W., Knapik, O. and Leśkow, J. (2013), *Finding a frequency signature for a cyclostationary signal with applications to wheel bearing diagnostics*, Mechanical Systems and Signal Processing, vol **38**, pp. 55 - 64.
- W. Cioch, J. Duda and P. Pawlik (2024) , *'CMAFI - Copula-based Multifeature Autocorrelation Fault Identification of rolling bearing*, *Mechanical Systems and Signal Processing*, 211
- Horman, S. and Kokoszka, P. (2010), *Weakly dependent functional data*, The Annals of Statistics, vol 38(3), pp. 1845 - 1884.
- Horvath, L. and Kokoszka, P. (2012) *Inference for Functional Data with Applications*, Springer Series in Statistics.
- Kokoszka,P and Reimherr. M. (2013), *Asymptotic normality of the principal components of functional time series*, Stoch.

Thanks for your attention!