Detection of periodic signals in a sequence of functional data

Vaidotas Characiejus[a]

Joint work with Clément Cerovecki[b] and Siegfried Hörmann[c]

December 9, 2021

[a] Department of Mathematics and Computer Science, University of Southern Denmark, Denmark

[b] Département de mathématique, Université libre de Bruxelles, Belgium

[b] Department of Mathematics, Katholieke Universiteit Leuven, Belgium

[c] Institute of Statistics, Graz University of Technology, Austria

Outline

Motivation

Main results

Empirical study

Future work and summary

Motivation

Motivation

Problem

Periodic signals

- Periodicities are one of the most important characteristics of time series.
- The interest to detect, analyze and model periodicities goes back to the very origins of the field (Schuster [1898], Walker [1914], Yule [1927], Fisher [1929], etc.).

Our goal

- Major advances in data collection technology leads to new challenges and at the same time to new methodologies as well as a better understanding of the underlying periodic structure.
- The focus of the talk will be detection, analysis and estimation of periodic signals in a sequence of functional data.

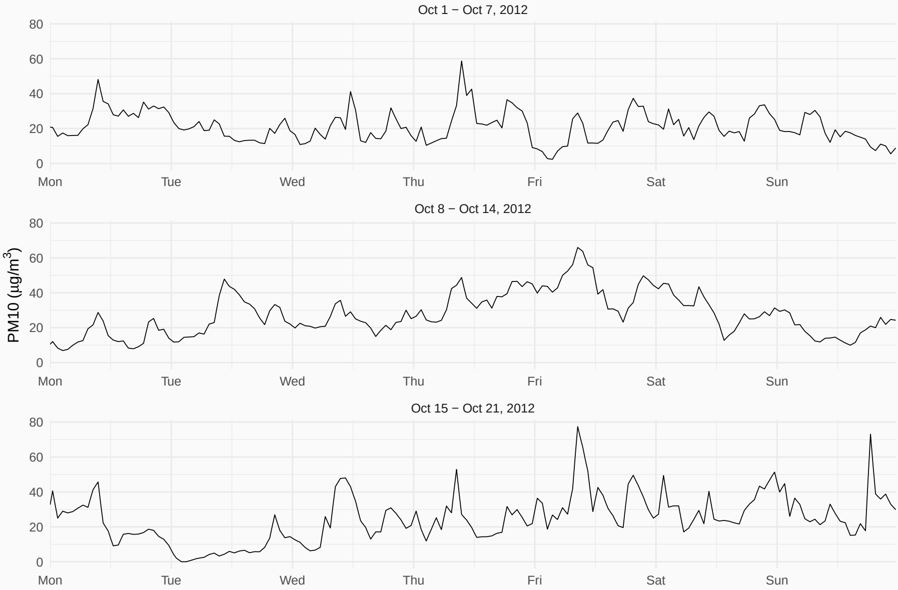
Motivation

Data example

PM10 data

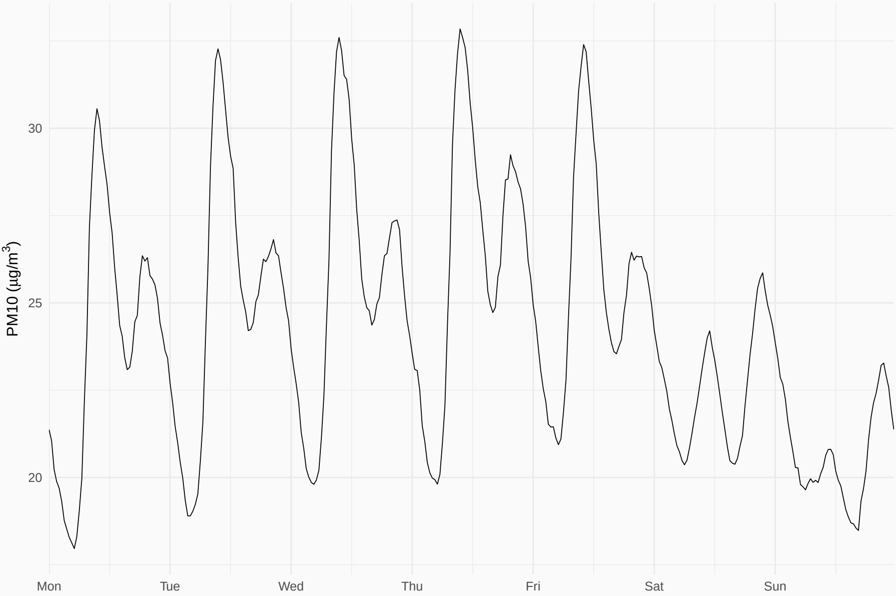
- Air quality data from Graz, Austria.
- The amount of particulate matter with a diameter of 10 µm or less (PM10) is measured.
- PM10 can settle in the bronchi and lungs and cause health problems.
- Starting on February 18, 2010, the amount of PM10 in $\mu g/m^3$ is recorded every 30 minutes resulting in 48 observations per day.
- Our data set contains observations from February 18, 2010 until January 27, 2021 (3997 observation days or almost 11 observation years in total).

Raw data

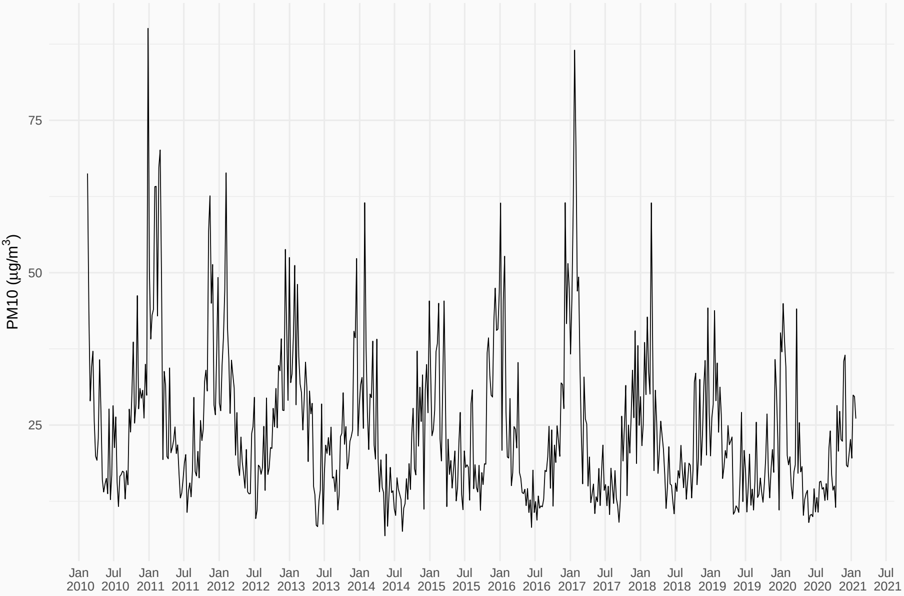


Oct 1 − Oct 7, 2012
Oct 8 − Oct 14, 2012
Oct 15 − Oct 21, 2012
PM10 (µg/m3)
0
20
40
60
80
Mon
Tue
Wed
Thu
Fri
Sat
Sun

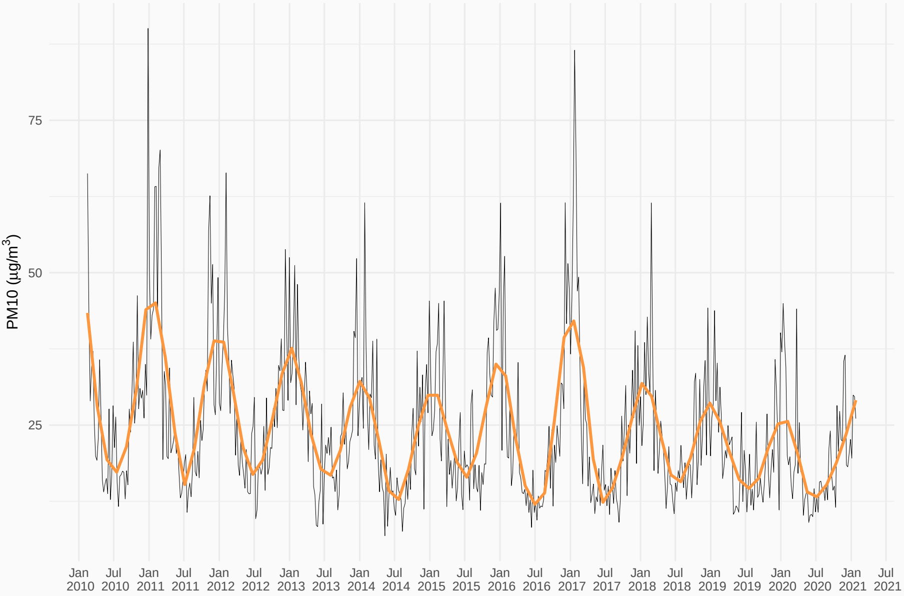
Weekly mean curve



Weekly averages



Weekly averages



Motivation

Our approach

The PM10 data as a sequence of curves

We investigate the PM10 data as a functional time series, i.e. as a sequence of daily curves.

Functional time series

- A functional time series is a sequence $\{X_t\}_{t\in\mathbb{Z}}$ such that each X_t is a curve $\{X_t(u)\}_{u\in[0,1]}$.
- We separate a continuous time process $\{\xi(u)\}_{u\in\mathbb{R}}$ using natural consecutive intervals, i.e.

$$X_t(u) = \xi(t+u)$$

  for $u \in [0,1]$ and $t \in \mathbb{Z}$.
- Such segmentation accounts for a periodic structure in the underlying continuous time process.
- There might still remain a periodic signal with respect to the discrete time parameter $t \in \mathbb{Z}$.

Model

Consider the time series $\{X_t\}_{t\in\mathbb{Z}}$ with values in a real separable Hilbert space $\mathbb{H}$ defined by

$$X_t = \mu + s_t + Y_t$$

for each $t \in \mathbb{Z}$, where

- $\mu \in \mathbb{H}$;
- $\{s_t\}_{t\in\mathbb{Z}} \subset \mathbb{H}$ is a deterministic sequence such that

  $$s_t = s_{t+T} \quad \text{and} \quad \sum_{t=1}^{T} s_t = 0$$

  for all $t \in \mathbb{Z}$ with some $T \geq 2$;
- $\{Y_t\}_{t\in\mathbb{Z}}$ is a stationary sequence of zero mean random elements with values in $\mathbb{H}$.

Hypothesis testing

- We develop a methodology to detect a periodic signals in Hilbert space value time series when $T \geq 2$ is not assumed to be known.
- Specifically, we want to test the following hypotheses:
  H_0 : observations are generated by a stationary sequence (no periodic component);
  H_1 : observations are generated by a stationary sequence with a superimposed deterministic periodic component with an unknown period $T \geq 2$.

Main results

Main results

Test statistic

Frequency domain approach

Our methodology is based on the frequency domain approach to the analysis of functional time series.

DFT and periodogram

Definition

The DFT of $X_1, \ldots, X_n$ is defined by

$$\mathcal{X}_n(\omega_j) = n^{-1/2} \sum_{t=1}^{n} X_t e^{-it\omega_j}$$

where $n \geq 1$, $i = \sqrt{-1}$, $j = -\lfloor (n-1)/2 \rfloor, \ldots, \lfloor n/2 \rfloor$ and $\omega_j = 2\pi j/n$.

Definition

The periodogram operator of $X_1, \ldots, X_n$ is defined by

$$I_n(\omega_j) = \mathcal{X}_n(\omega_j) \otimes \mathcal{X}_n(\omega_j) = \langle \cdot, \mathcal{X}_n(\omega_j) \rangle \mathcal{X}_n(\omega_j),$$

where $n \geq 1$, $j = -\lfloor (n-1)/2 \rfloor, \ldots, \lfloor n/2 \rfloor$ and $\omega_j = 2\pi j/n$.

Maximum of periodogram

The test statistic is given by

$$M_n = \max_{1 \leq j \leq q} \|I_n(\omega_j)\|_{op} = \max_{1 \leq j \leq q} \|\mathcal{X}_n(\omega_j)\|^2$$

for $n > 2$, where

i) $\|\cdot\|_{op}$ is the operator norm;

ii) $\omega_j = 2\pi j/n$ are the Fourier frequencies with $1 \leq j \leq q$;

iii) $q = \lfloor n/2 \rfloor$;

iv) $\|\cdot\|$ is the norm induced by the inner product of $\mathbb{H}$.

Maximum of periodogram

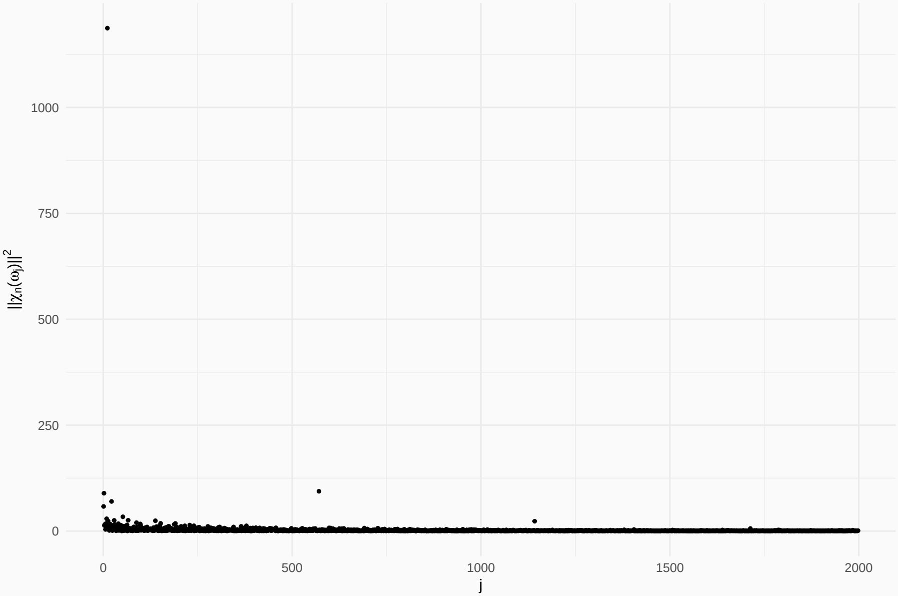
The test statistic is given by

$$M_n = \max_{1 \leq j \leq q} \|I_n(\omega_j)\|_{op} = \max_{1 \leq j \leq q} \|\mathcal{X}_n(\omega_j)\|^2$$

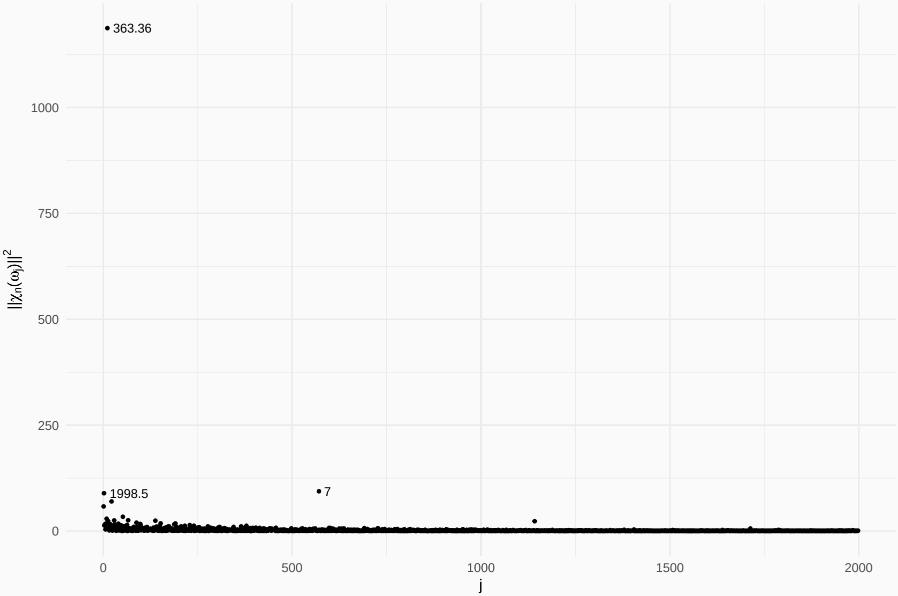
for $n > 2$.

- Small values of M_n indicate that there is no periodic component.
- Large values of M_n indicate that there is a periodic component.
- We need a criterion to decide when M_n is small and when M_n is large.

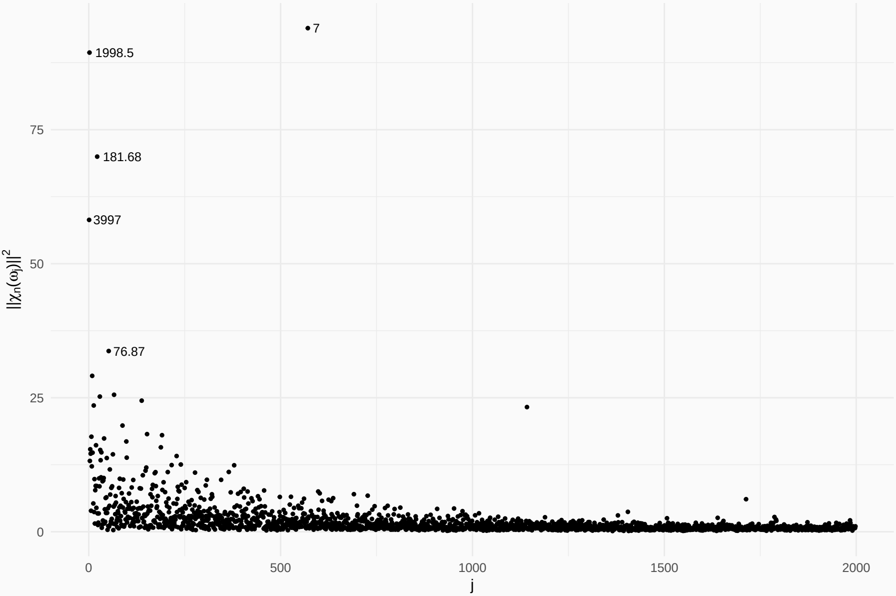
Periodogram of PM10



Periodogram of PM10



Periodogram of PM10 without the largest value



Results under Gaussianity in the univariate case

- The usefulness of the maximum of the periodogram for detecting periodicities is well known (Fisher [1929]).
- First results were established under the assumption of Gaussianity.
- An alternative approach is to establish the asymptotic distribution of the appropriately standardized M_n under some general conditions.

Results under Gaussianity in the univariate case (cont.)

If $X_1, \ldots, X_n$ are iid standard normal random variables,

$$M_n - \log q \xrightarrow{d} G \quad \text{as} \quad n \to \infty,$$

where $q = \lfloor n/2 \rfloor$ and G is the standard Gumbel distribution with the CDF given by

$$F(x) = \exp\{-\exp^{-x}\}$$

for $x \in \mathbb{R}$.

General results in the univariate case

- Walker [1965] conjectured that the same result holds provided that the moments up to some sufficiently high order exist.
- Walker [1965] also stated that no proof was known at the time and that the problem of constructing one is undoubtedly extremely difficult.
- Davis and Mikosch [1999] proved that the limit indeed remains the same provided that $E|X_1|^s < \infty$ with some $s > 2$ using a Gaussian approximation technique due to Einmahl [1989].

Main results

Asymptotic distribution of the test statistic

Our results

- Our main result is an extension of the result of Davis and Mikosch [1999] to real separable Hilbert spaces.
- The main ingredient of our proof is a powerful Gaussian approximation developed by Chernozhukov, Chetverikov, and Kato [2017].
- Our results allow us to propose several methodologies to detect periodic signals in Hilbert space valued time series when the length of the period is unknown.

Linear processes

Suppose that $\{Y_t\}_{t\in\mathbb{Z}}$ is a linear process with values in $\mathbb{H}$ given by

$$Y_t = \sum_{k=-\infty}^{\infty} a_k(\varepsilon_{t-k})$$

for each $t \in \mathbb{Z}$, where

- $\{a_k\}_{k\in\mathbb{Z}} \subset L(\mathbb{H})$ such that $\sum_{k=-\infty}^{\infty} \|a_k\|_{op} < \infty$;
- $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ are iid zero mean random elements with values in $\mathbb{H}$.

Notation for linear processes

- $A(\omega)$ denotes the impulse-response operator given by

$$A(\omega) = \sum_{k=-\infty}^{\infty} a_k e^{-it\omega}$$

  for $\omega \in [-\pi, \pi]$.
- $\{\lambda_k\}_{k \geq 1}$ are the eigenvalues of the autocovariance operator $E[\varepsilon_0 \otimes \varepsilon_0]$.

Assumptions

Assumption 1

i) $E\|\varepsilon_0\|^r < \infty$ where $r > 2$ if $\dim \mathbb{H} < \infty$ and $r \geq 4$ otherwise;

ii) the eigenvalues λ_k are distinct and the sequence $\{k\lambda_k\}_{k\geq 1}$ is ultimately non-increasing;

iii) some further conditions on the decay rate of $\{\lambda_k\}_{k\geq 1}$.

Assumption 2

i) $\sum_{k\neq 0} \log(|k|)\|a_k\|_{op} < \infty$;

ii) $A^{-1}(\omega)$ exists for each $\omega \in [-\pi, \pi]$;

iii) $\sup_{\omega\in[0,\pi]} \|A^{-1}(\omega)\|_{op} < \infty$.

Main result

Theorem

Under H_0 and Assumptions 1 and 2, we have that

$$\lambda_1^{-1}\left(\max_{1\le j\le q}\|A^{-1}(\omega_j)\mathcal{X}_n(\omega_j)\|^2 - b_n\right) \xrightarrow{d} G \quad \text{as} \quad n\to\infty,$$

where

- $q = \lfloor n/2 \rfloor$;
- $b_n = \lambda_1 \log q - \lambda_1 \sum_{j=2}^{\infty} \log(1-\lambda_j/\lambda_1)$;
- G is the standard Gumbel distribution with the CDF given by $F(x) = \exp\{-\exp\{-x\}\}$ for $x \in \mathbb{R}$.

Assumption for the FAR(1)

Suppose that $\{Y_t\}_{t\in\mathbb{Z}}$ is an FAR(1) model given by

$$Y_t = \rho(Y_{t-1}) + \varepsilon_t = \sum_{j=0}^{\infty} \rho^j(\varepsilon_{t-j})$$

for $t \in \mathbb{Z}$ with $\rho \in L(\mathbb{H})$ such that $\|\rho^{n_0}\|_{op} < 1$ with some $n_0 \geq 1$.

Assumption 3

$\hat{\rho}$ is an estimator of ρ such that $\|\hat{\rho} - \rho\|_{op} = o_p(\tau_n^{-1})$ as $n \to \infty$, where $\log n \leq a_n < \sqrt{n}$.

Test statistic

Theorem

Suppose that $\{\hat{\lambda}_j\}_{j\geq 1}$ *are the eigenvalues of* $(n-1)^{-1}\sum_{k=2}^{n}\hat{\varepsilon}_k\otimes\hat{\varepsilon}_k$, *where*

$$\hat{\varepsilon}_k = X_k - \hat{\rho}(X_{k-1}), \quad k = 2, \dots, n.$$

Under H_0 *and Assumptions 1, 2, and 3, we have that*

$$G_n = \hat{\lambda}_1^{-1}\max_{1\leq j\leq q}\|(I - e^{-i\omega_j}\hat{\rho})\mathcal{X}_n(\omega_j)\|^2 - \log q + \sum_{j=2}^{\tau_n}\log(1-\hat{\lambda}_j/\hat{\lambda}_1) \xrightarrow{d} G$$

as $n\to\infty$.

Representation of periodic signals

Lemma

Suppose that $\{s_t\}_{t\in\mathbb{Z}}$ is a deterministic sequence with values in $\mathbb{H}$ such that

$$s_t = s_{t+T} \quad \textit{and} \quad \sum_{t=1}^{T} s_t = 0$$

for all $t \in \mathbb{Z}$ with some $T \geq 2$. Then there exist $w_{11}, \ldots, w_{1\lfloor T/2\rfloor} \in \mathbb{H}$ and $w_{21}, \ldots, w_{2\lfloor T/2\rfloor} \in \mathbb{H}$ such that

$$s_t = \sum_{k=1}^{\lfloor T/2\rfloor} [\cos(2\pi kt/T) w_{1k} + \sin(2\pi kt/T) w_{2k}]$$

for all $t \in \mathbb{Z}$.

Consistency

Theorem

Suppose that

$$\|w_{11} + w_{21} - \cos(\omega_{\lfloor n/T \rfloor})\hat{\rho}(w_{11} + w_{21}) - \sin(\omega_{\lfloor n/T \rfloor})\hat{\rho}(w_{11} - w_{21})\|$$

is ultimately bounded away from zero in probability, where $w_{11}, w_{21} \in \mathbb{H}$ *come from the general expression of a periodic sequence* $\{s_t\}_{t \geq 1} \subset \mathbb{H}$*. Then under* H_1 *we have that*

$$G_n/\ell_n \xrightarrow{P} \infty \quad as \quad n \to \infty$$

for any positive sequence $\ell_n = o(n)$ *as* $n \to \infty$*.*

Empirical study

Empirical study

Simulation study

Simulation setting

- We simulate functional time series that are stationary and behaves similarly as the original PM10 data.
- The periodic component in the simulation study is given by

$$s_t(u) = a \cos(2\pi t/d),$$

  where $u \in [0, 1]$ and $d - 2$ is a Poisson distributed random variable P_λ with $\lambda = 5$ or $\lambda = 15$.
- a is equal to 0 (no periodic signal), 1 or 2.

Empirical rejection rates

		$a = 0$ ($\equiv H_0$)			$a = 1$			$a = 2$		
	α	0.1	0.05	0.01	0.1	0.05	0.01	0.1	0.05	0.01
$\lambda = 5$	$n = 100$	0.049	0.022	0.004	0.867	0.805	0.670	1.000	0.999	0.994
	$n = 200$	0.074	0.034	0.005	0.990	0.983	0.972	1.000	1.000	1.000
	$n = 500$	0.091	0.052	0.011	1.000	1.000	0.999	1.000	1.000	1.000
$\lambda = 15$	$n = 100$	0.067	0.030	0.004	0.260	0.172	0.072	0.837	0.773	0.629
	$n = 200$	0.069	0.030	0.006	0.585	0.488	0.312	0.987	0.975	0.926
	$n = 500$	0.093	0.044	0.007	0.990	0.979	0.946	1.000	1.000	1.000

Empirical study

The analysis of the PM10 data

Transforming data into curves

- The data is preprocessed in the following way:
  - the missing values are linearly interpolated;
  - the negative values are set to 0 so that the square root transformation can be performed;
  - the raw observations are transformed into curves using the `R` package `fda` and the function `Data2fd()` with 21 Fourier basis functions.
- We use the PCA based estimator of ρ (Bosq [2000]).
- The tuning parameter k_n which determines the number of principal components used in the estimation procedure is selected so that k_n principal components explain more than 99% of the variance in our dataset.

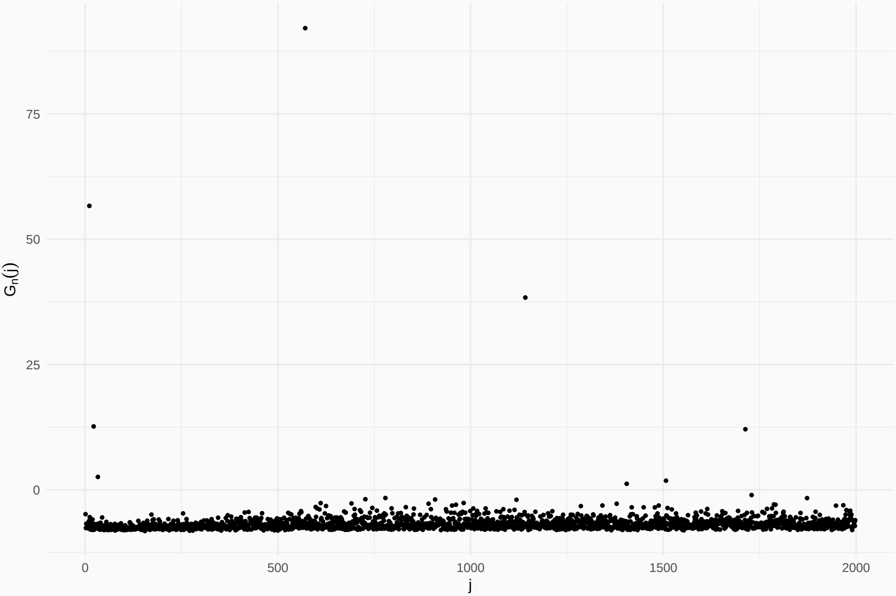
- Instead of just reporting the value of the test statistic or the p-value, we plot the points $(j, G_n(j))$ with $j = 1, \dots, q = 1998$ and

$$G_n(j) := \hat{\lambda}_1^{-1} \|(I - e^{-i\omega_j} \hat{\rho})(\mathcal{X}_n(\omega_j))\|^2 - \log q + \sum_{j=2}^{a_n} \log(1 - \hat{\lambda}_j / \hat{\lambda}_1),$$

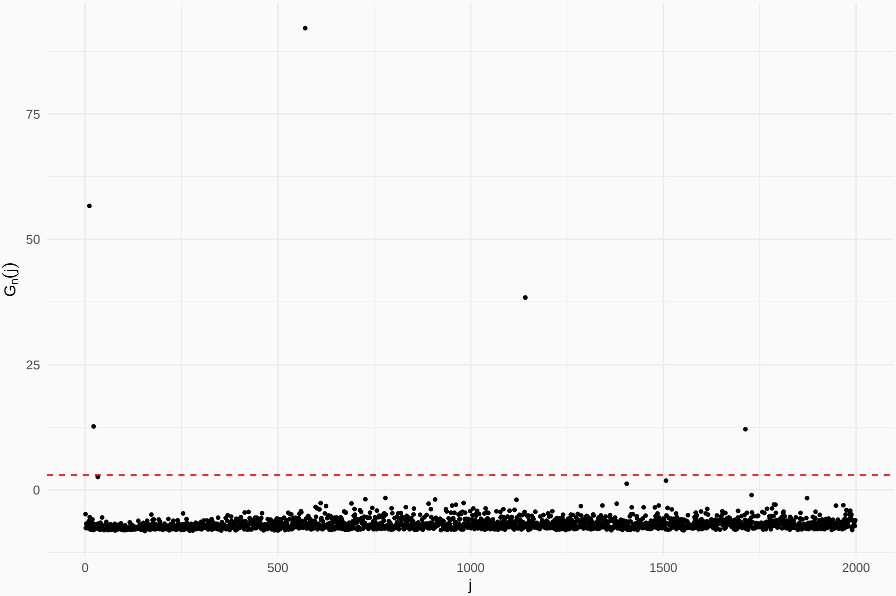
  where $n = 3997$.
- Observe that

$$G_n = \max_{1 \le j \le q} G_n(j).$$

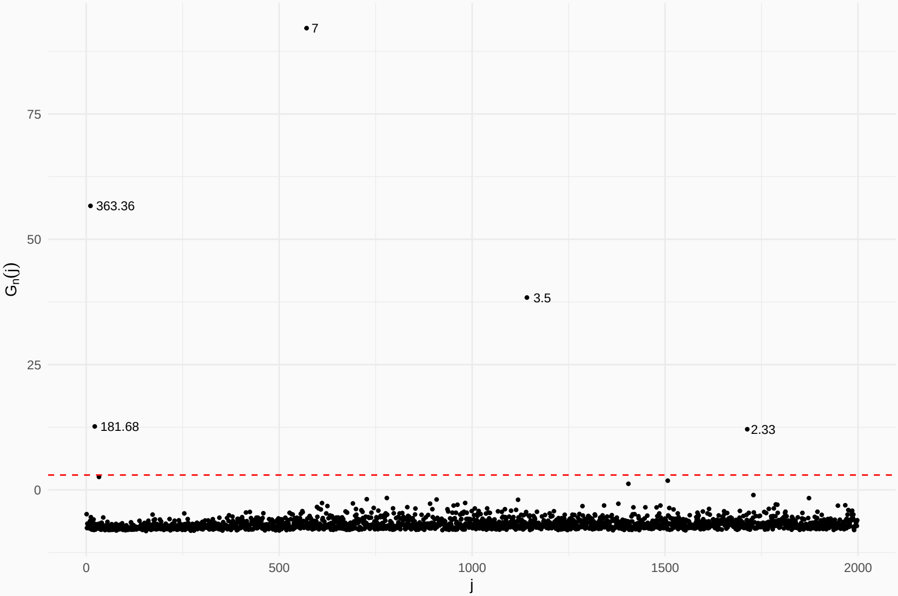
PM10 time series



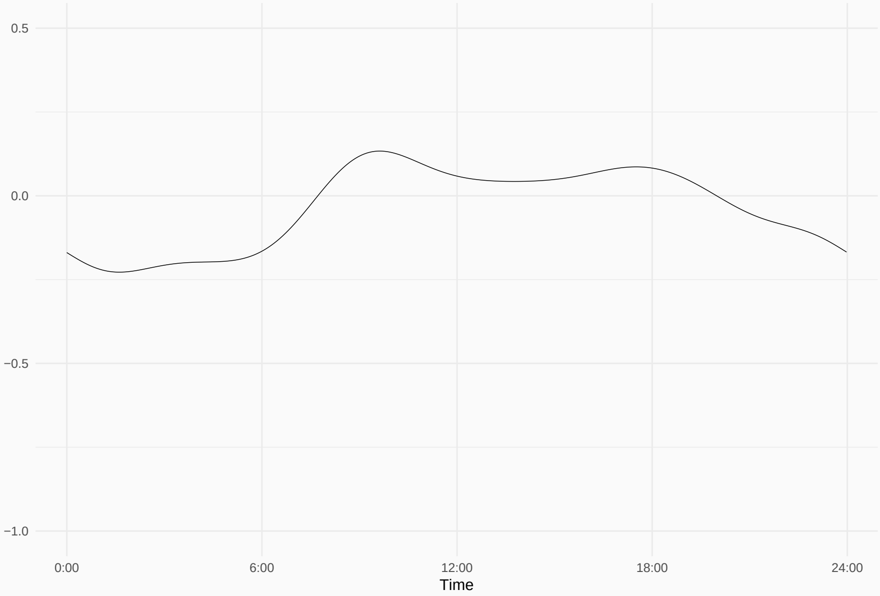
PM10 time series



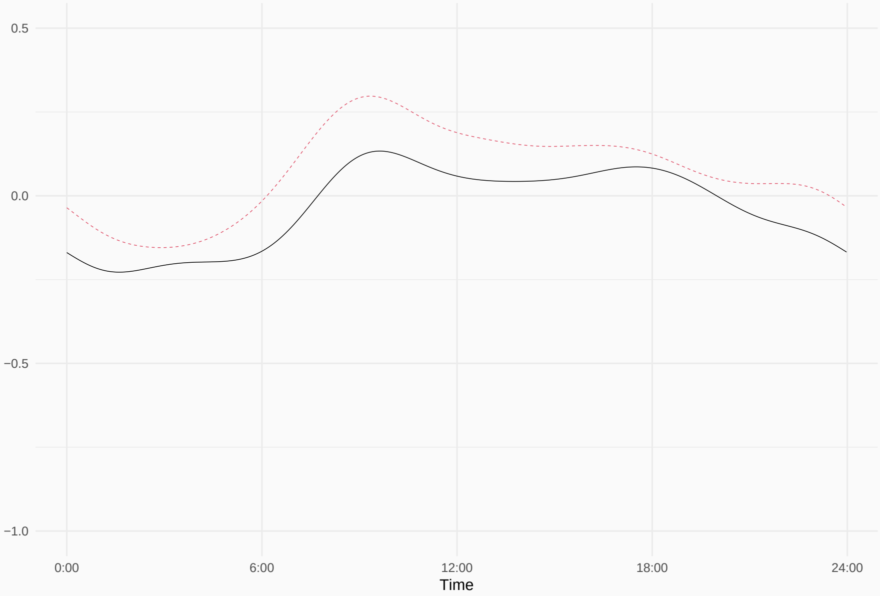
PM10 time series



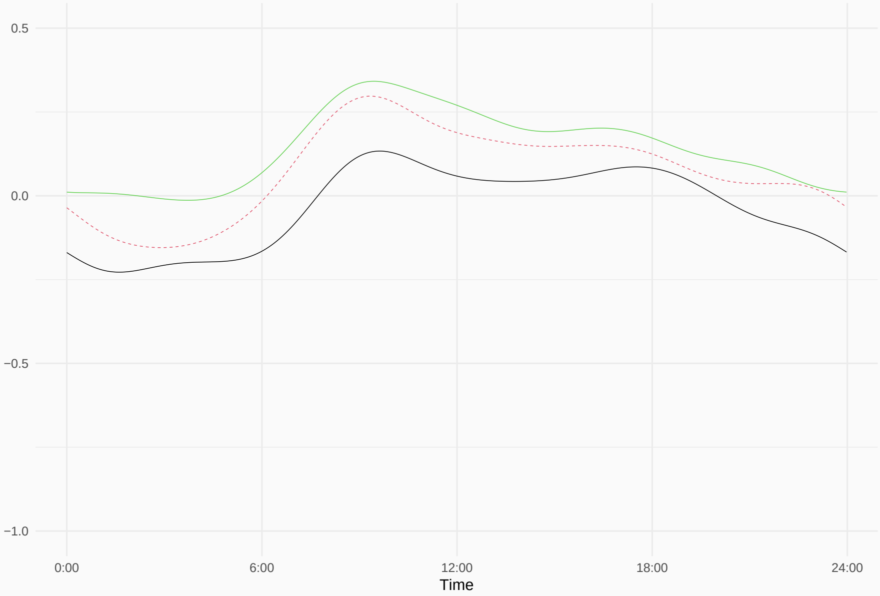
Weekly periodic component



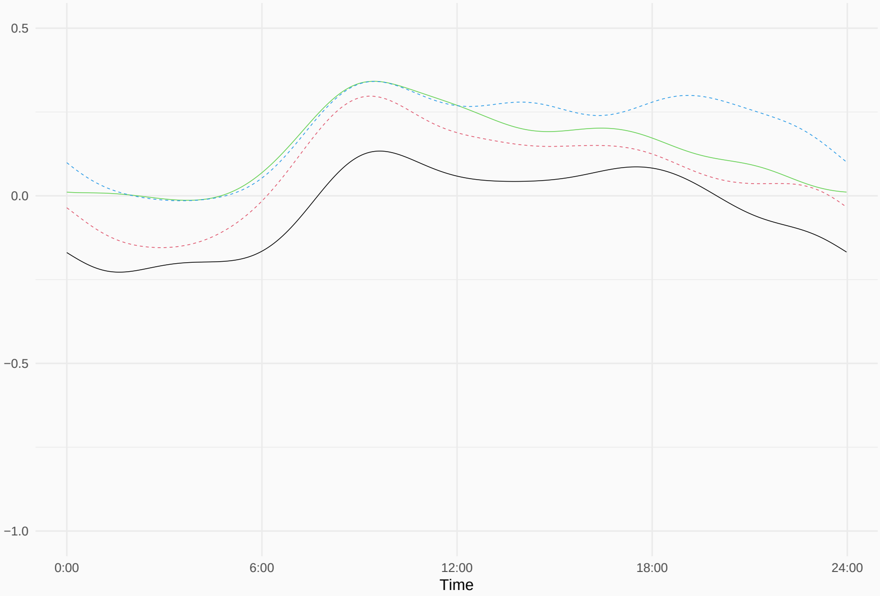
Weekly periodic component



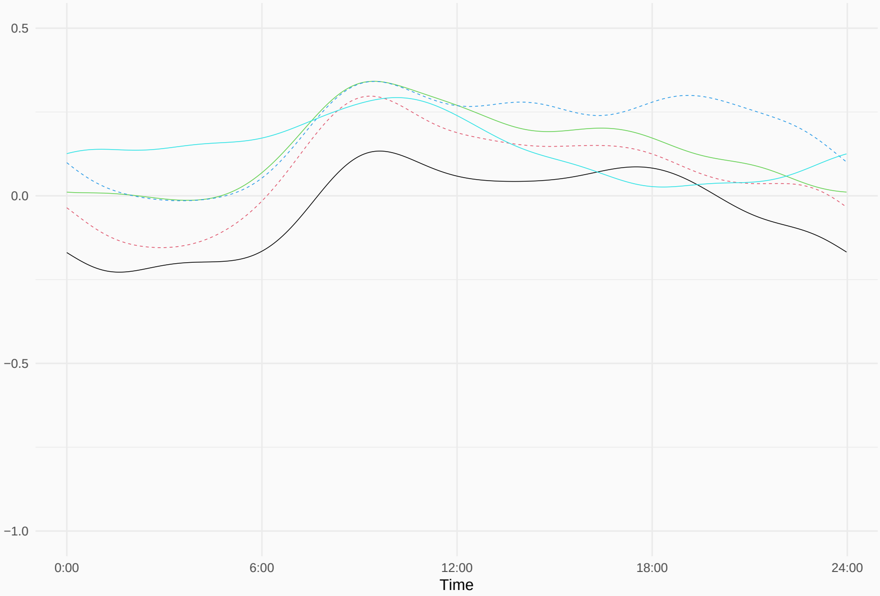
Weekly periodic component



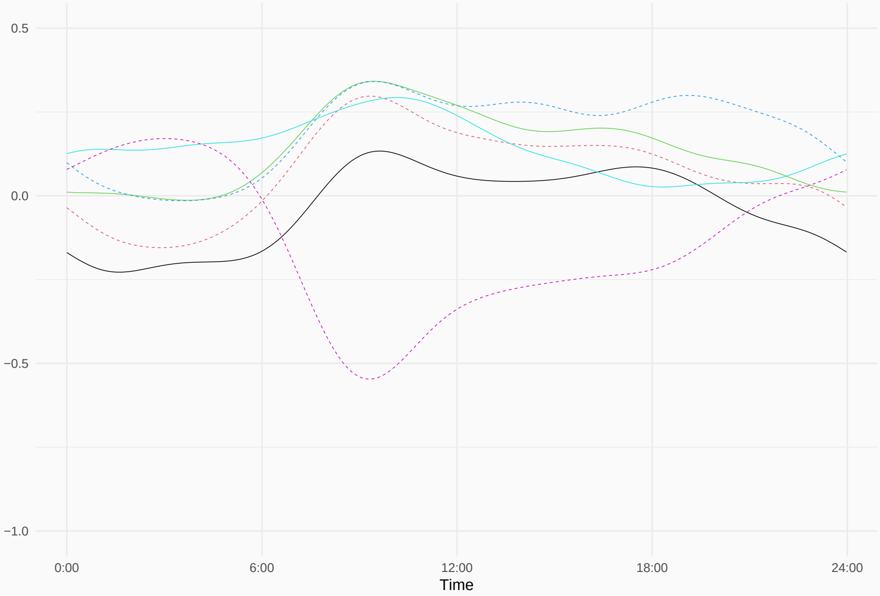
Weekly periodic component



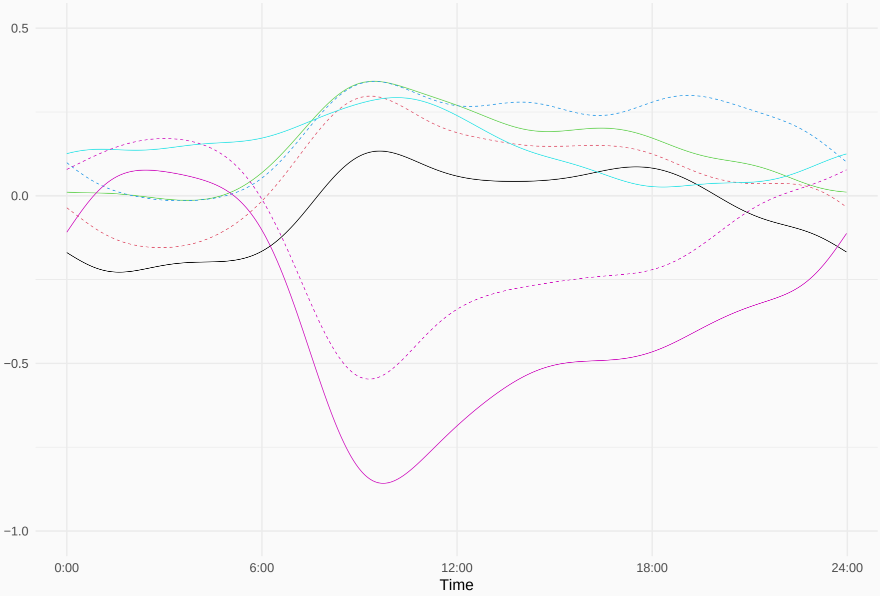
Weekly periodic component



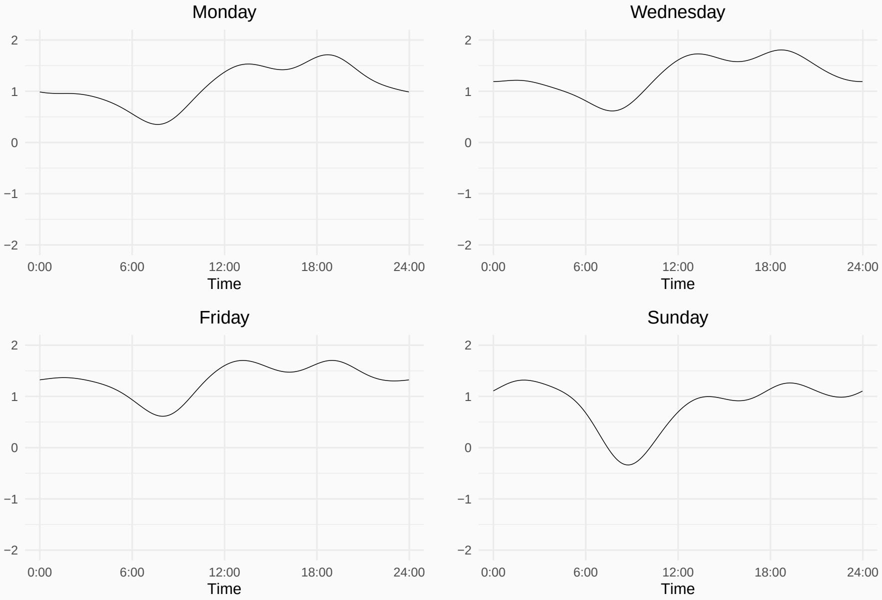
Weekly periodic component



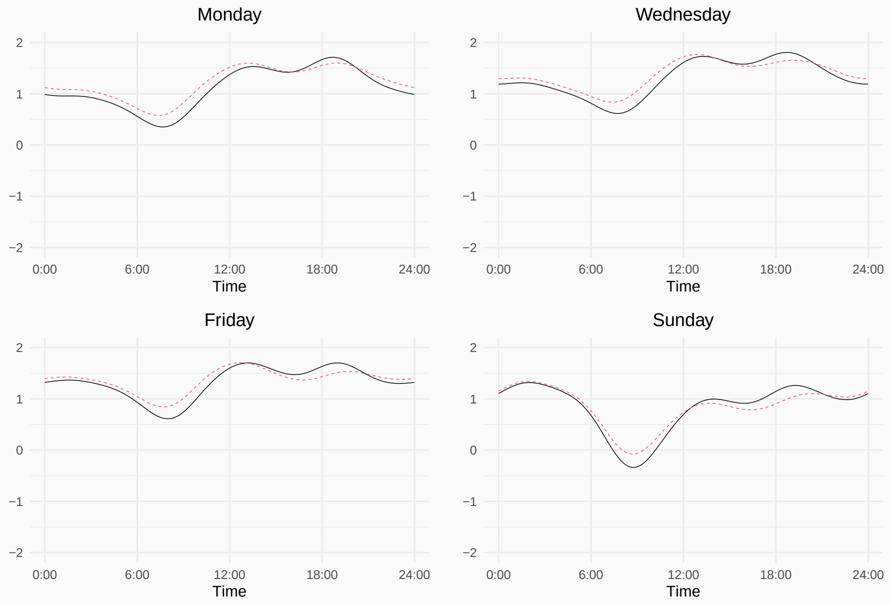
Weekly periodic component



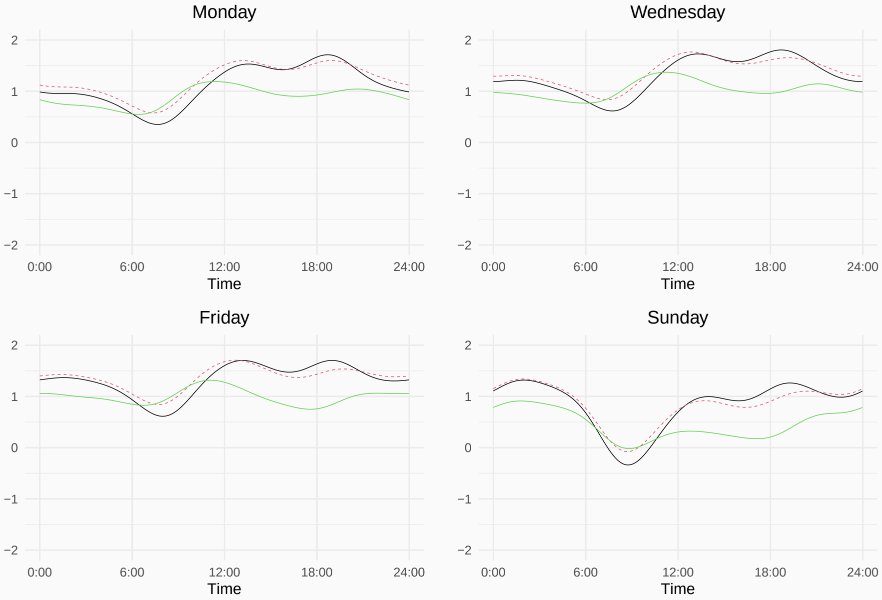
Periodic component



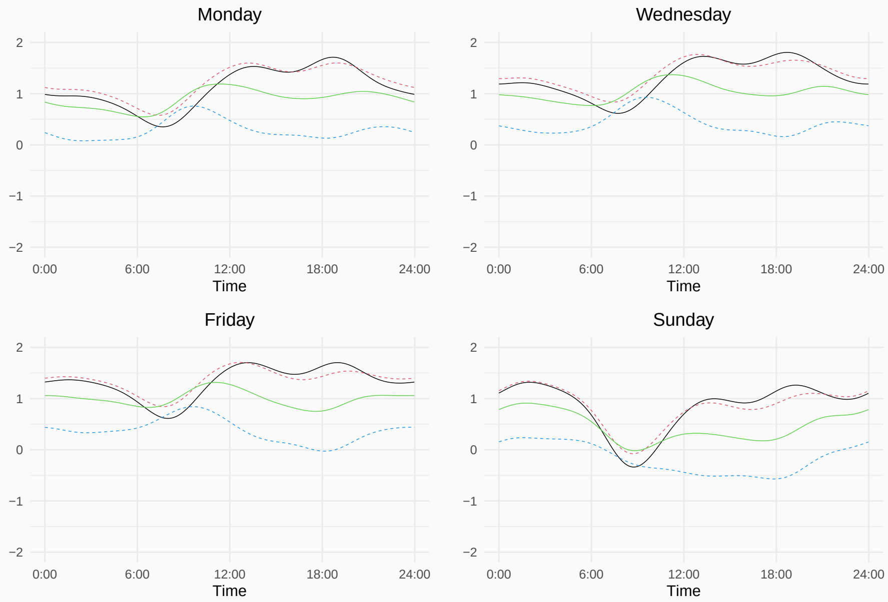
Periodic component



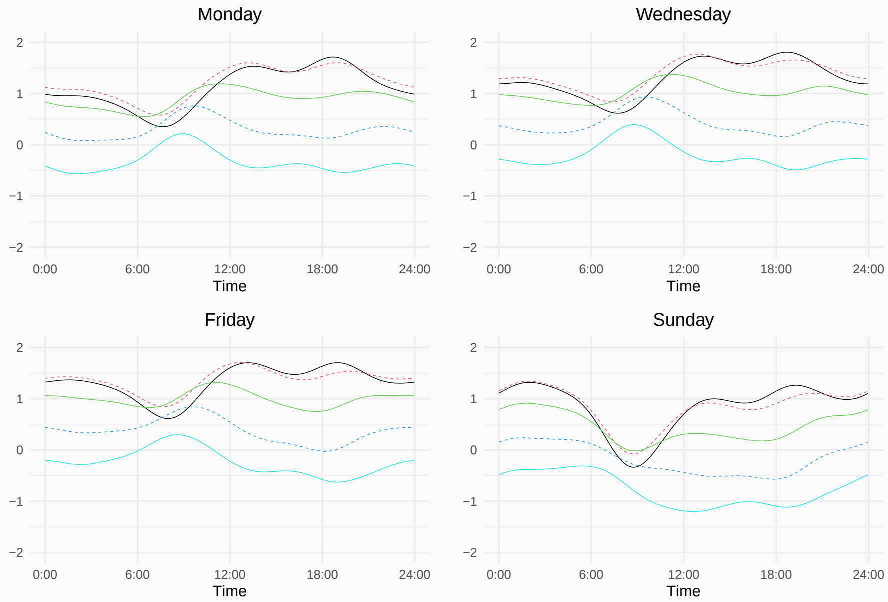
Periodic component



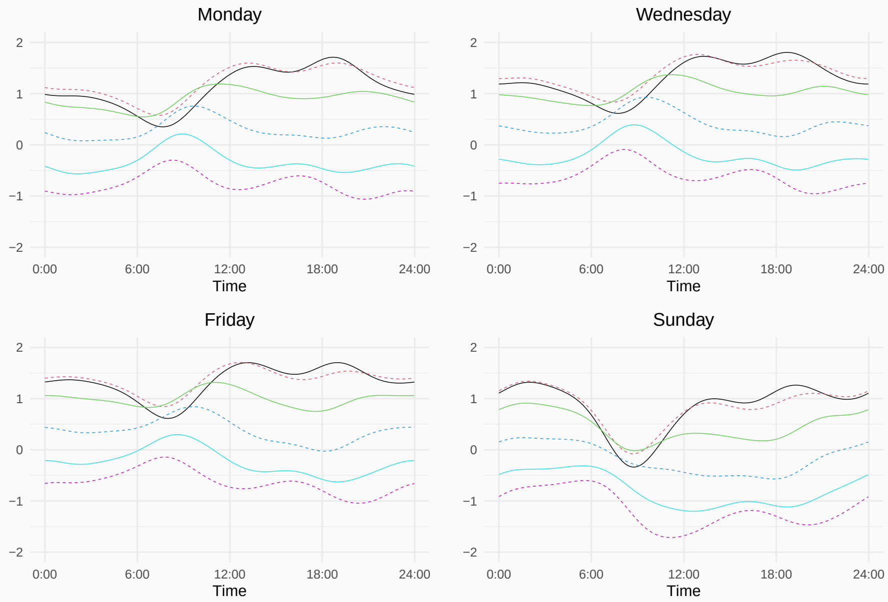
Periodic component



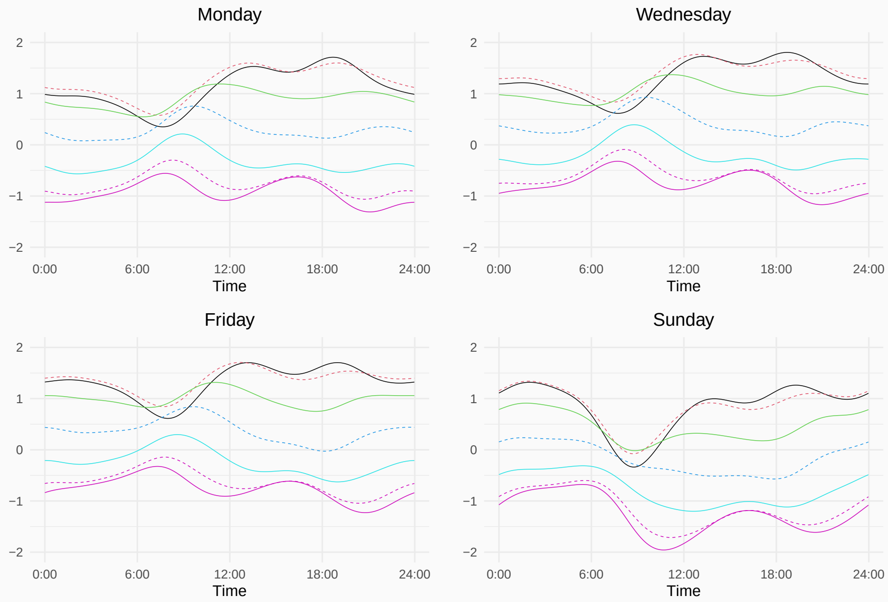
Periodic component



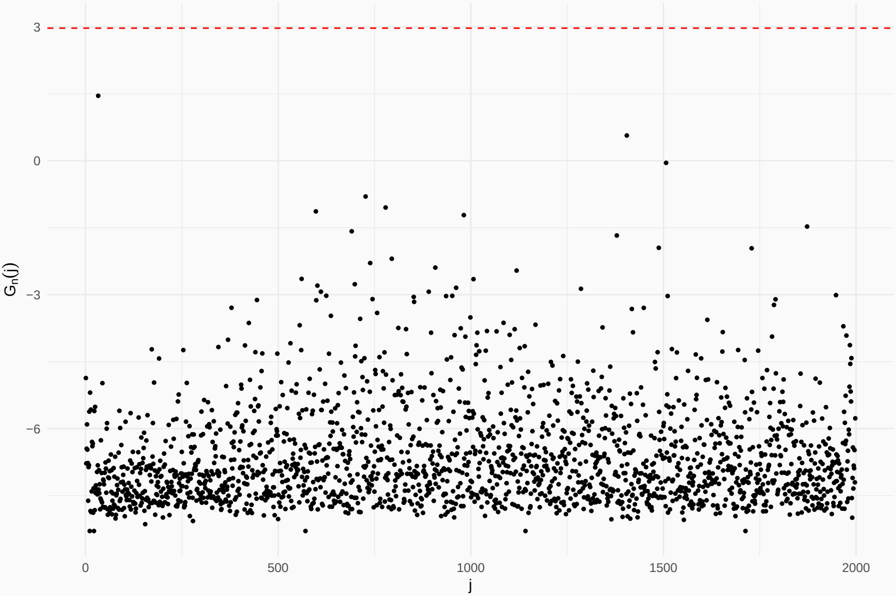
Periodic component



Periodic component



Deseasonalized data



Future work and summary

Concluding remarks

- A general test for periodic signals in Hilbert space valued time series when the length of the period is unknown.
- The appropriately standardized maximum of the periodogram converges in distribution to the standard Gumbel distribution.
- Very good finite sample performance.
- A weekly as well as a yearly periodic components are detected in the PM10 data.
- The periodic signals in the PM10 data are not pure sinusoids but actually superposition of several sinusoids.

`https://imada.sdu.dk/~characiejus/`

References

D. Bosq. *Linear Processes in Function Spaces*, volume 149 of *Lecture Notes in Statistics*. Springer-Verlag New York, 2000.

V. Chernozhukov, D. Chetverikov, and K. Kato. Central limit theorems and bootstrap in high dimensions. *The Annals of Probability*, 45:2309–2352, 2017.

R.A. Davis and T. Mikosch. The maximum of the periodogram of a non-Gaussian sequence. *The Annals of Probability*, 27: 522–536, 1999.

Uwe Einmahl. Extensions of results of Komlós, Major, and Tusnády to the multivariate case. *Journal of Multivariate Analysis*, 28(1):20 – 68, 1989.

Ronald A. Fisher. Tests of significance in harmonic analysis. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 125(796):54–59, 1929.

A. Schuster. On the investigation of hidden periodicities with application to a supposed 26 day period of meteorological phenomena. *Terrestrial Magnetism*, 3(1):13–41, 1898.

A.M. Walker. Some asymptotic results for the periodogram of a stationary time series. *Journal of the Australian Mathematical Society*, 5:107–128, 1965.

G. T. Walker. *Correlation in seasonal variations of weather, III : on the criterion for the reality of relationships or periodicities*, volume 21 of *Memoirs of the India Meteorological Department*. Meteorological Office, 1914.

G. Udny Yule. On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 226:267–298, 1927.