

The maximum of the periodogram of a sequence of functional data

Vaidotas Characiejus^a

Joint work with Clément Cerovecki^b and Siegfried Hörmann^c

CMStatistics 2022 / London, December 18, 2022

^aDepartment of Mathematics and Computer Science, University of Southern Denmark, Denmark

^bDépartement de mathématique, Université libre de Bruxelles, Belgium

^bDepartment of Mathematics, Katholieke Universiteit Leuven, Belgium

^cInstitute of Statistics, Graz University of Technology, Austria

Motivation and problem

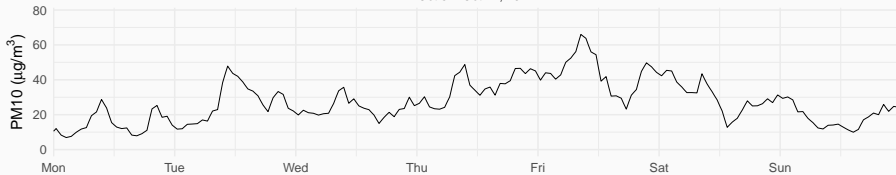
- Air quality data from Graz, Austria.
- The amount of particulate matter with a diameter of 10 μm or less (PM10) is measured.
- PM10 can settle in the bronchi and lungs and cause health problems.
- Starting on February 18, 2010, the amount of PM10 in $\mu\text{g}/\text{m}^3$ is recorded every 30 minutes resulting in 48 observations per day.

Raw data

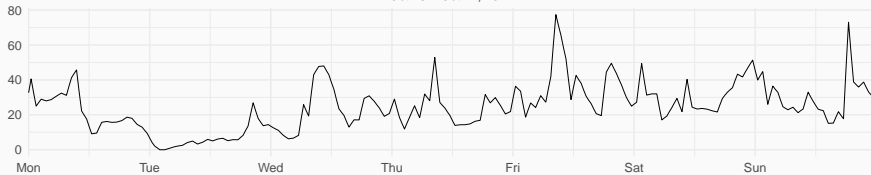
Oct 1 – Oct 7, 2012



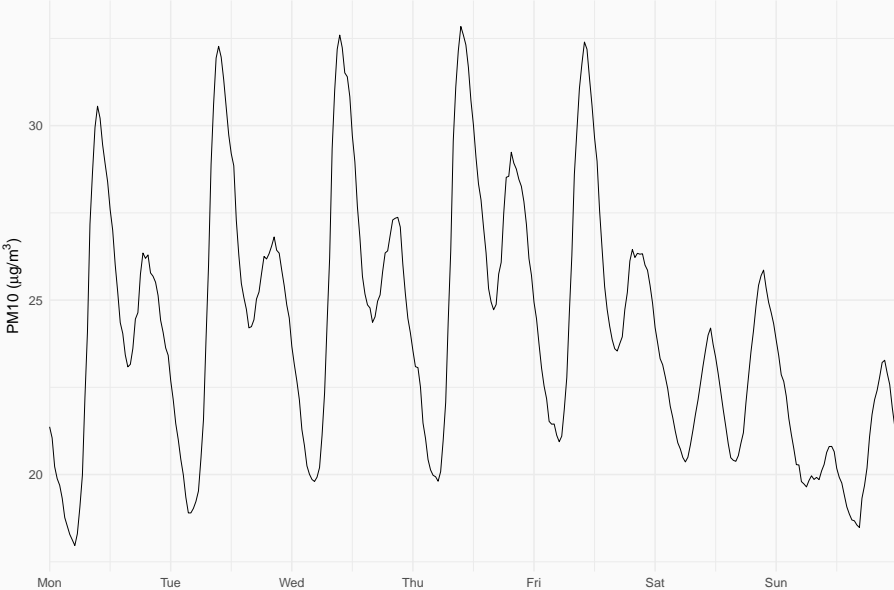
Oct 8 – Oct 14, 2012



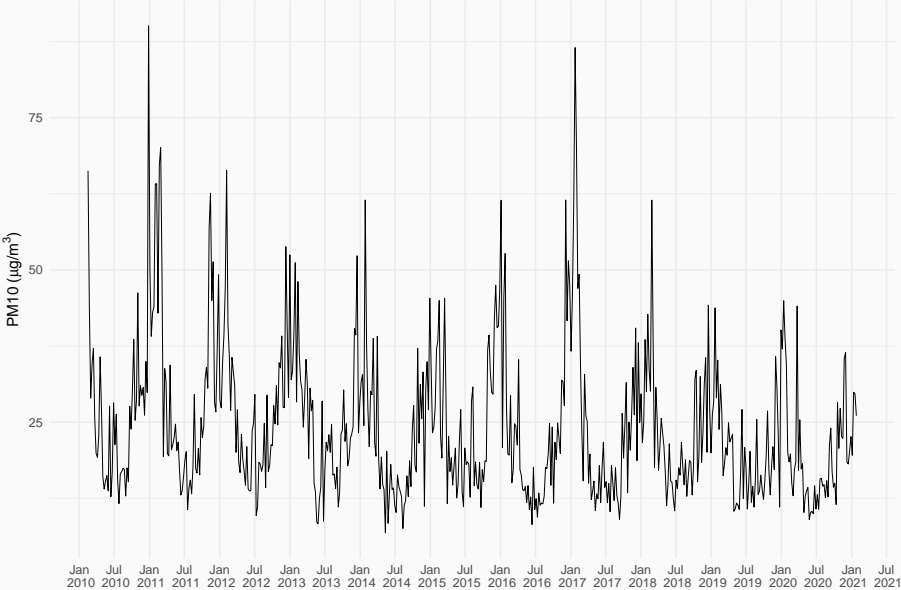
Oct 15 – Oct 21, 2012



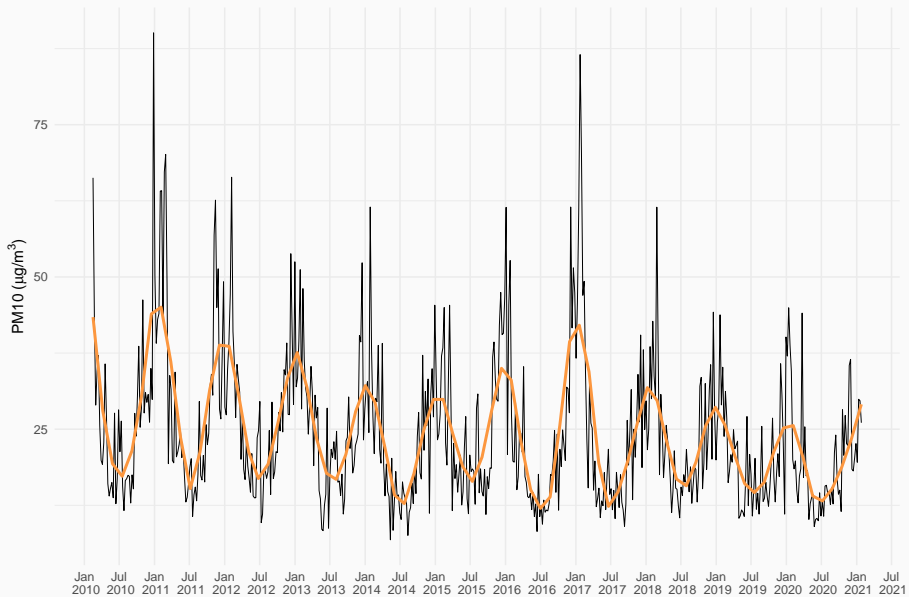
Weekly mean curve



Weekly averages



Weekly averages



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.

Model

$\{X_t\}_{t \in \mathbb{Z}}$ is a time series 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 test

$$H_0 : X_t = \mu + Y_t \quad \text{versus} \quad H_1 : X_t = \mu + S_t + Y_t$$

with an unknown $T \geq 2$.

Main results

Frequency domain approach

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

Definition

The discrete Fourier transform (DFT) of X_1, \dots, X_n is defined by

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

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

Maximum of periodogram

The test statistic is given by

$$\max_{1 \leq j \leq q} \|\mathcal{X}_n(\omega_j)\|^2$$

for $n > 1$, where

- i) $\omega_j = 2\pi j/n$ with $1 \leq j \leq q = \lfloor n/2 \rfloor$;
- ii) $\|\cdot\|$ is the norm of the complexification of \mathbb{H} .

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

- $\{\varepsilon_t\}_{t \in \mathbb{Z}}$ are iid zero mean random elements with values in \mathbb{H} ;
- $\{a_k\}_{k \in \mathbb{Z}} \subset L(\mathbb{H})$.

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 of $E[\varepsilon_0 \otimes \varepsilon_0]$ are distinct and the sequence $\{k\lambda_k\}_{k \geq 1}$ is ultimately non-increasing;
- iii) some technical 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 $\omega \in [-\pi, \pi]$, where $A(\omega) = \sum_{k=-\infty}^{\infty} a_k e^{-ik\omega}$ with $\omega \in [-\pi, \pi]$ is the transfer function;
- iii) $\sup_{\omega \in [0, \pi]} \|A^{-1}(\omega)\|_{op} < \infty$.

Theorem

Under H_0 and Assumptions 1 and 2, we have that

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

where

- $A(\omega_j) = \sum_{k=-\infty}^{\infty} a_k e^{-ik\omega_j}$ with $j = 1, \dots, q$;
- $b_n = \lambda_1 \log q - \lambda_1 \sum_{j=2}^{\infty} \log(1 - \lambda_j/\lambda_1)$;
- G is the standard Gumbel distribution with the CDF given by $F(x) = \exp\{-\exp\{-x\}\}$ for $x \in \mathbb{R}$.

$\{Y_t\}_{t \in \mathbb{Z}}$ 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 $\rho \in L(\mathbb{H})$.

Assumption 3

- i) There is an $n_0 \geq 1$ such that $\|\rho^{n_0}\| < 1$;
- ii) $\hat{\rho}$ is an estimator of ρ such that

$$\|\hat{\rho} - \rho\|_{op} = o_p(1/\tau'_n)$$

as $n \rightarrow \infty$ with $\tau'_n \geq \log n$.

Residuals and their eigenvalues

- $\{\hat{\epsilon}_k\}_{2 \leq k \leq n}$ are the residuals given by

$$\hat{\epsilon}_k = X_k - \hat{\rho}(X_{k-1})$$

for $k = 2, \dots, n$.

- $\{\hat{\lambda}_j\}_{j \geq 1}$ are the eigenvalues of

$$\frac{1}{n-1} \sum_{k=2}^n \hat{\epsilon}_k \otimes \hat{\epsilon}_k.$$

Theorem

Under H_0 and Assumptions 1 and 3,

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

as $n \rightarrow \infty$, where $\{\tau_n\}_{n \geq 1} \subset \mathbb{N}$ and $\{c_n\}_{n \geq 1} \subset \mathbb{R}$ are sequences that satisfy certain technical conditions.

Theorem

Under H_1 ,

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

for any positive sequence $\ell_n = o(n)$ as $n \rightarrow \infty$ provided certain technical conditions are satisfied.

Empirical study

- We plot the points $(j, G_n(j))$ with $j = 1, \dots, q = 1998$ and

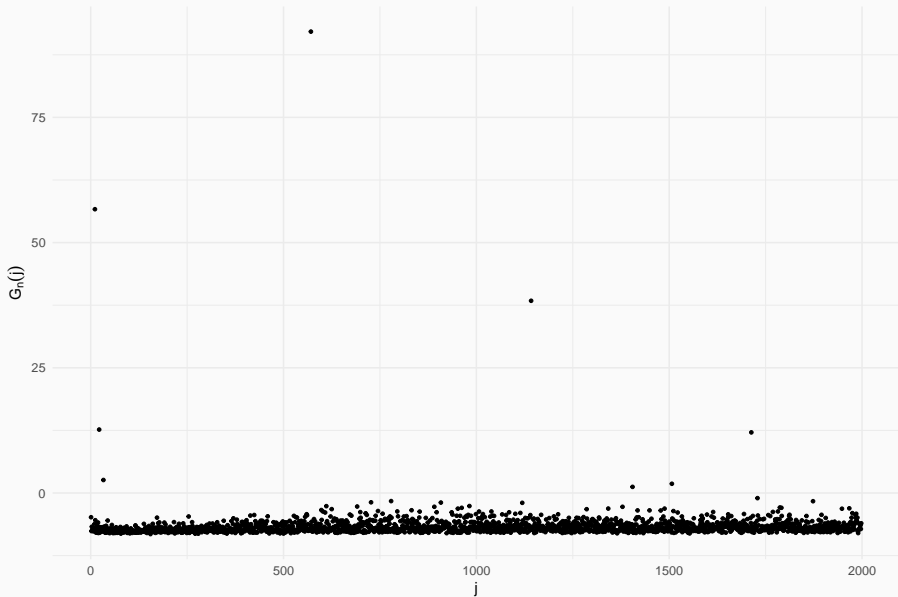
$$G_n(j) := \lambda_1^{-1} \|(I - e^{-i\omega_j} \hat{\rho})(\mathcal{X}_n(\omega_j))\|^2 - \log q + \max \left\{ \sum_{j=2}^{\tau_n} \log(1 - \hat{\lambda}_j / \hat{\lambda}_1), c_n \right\},$$

where $n = 3997$.

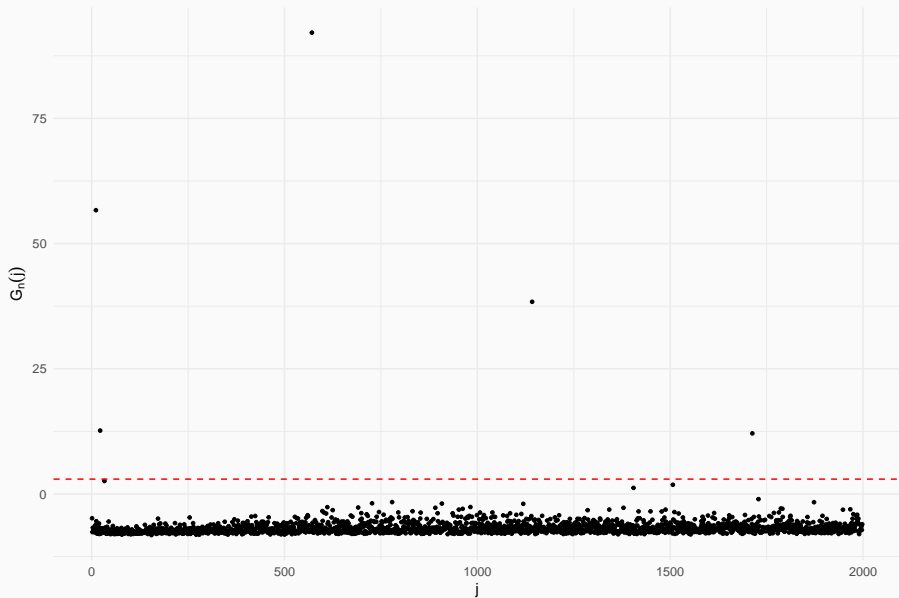
- Observe that

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

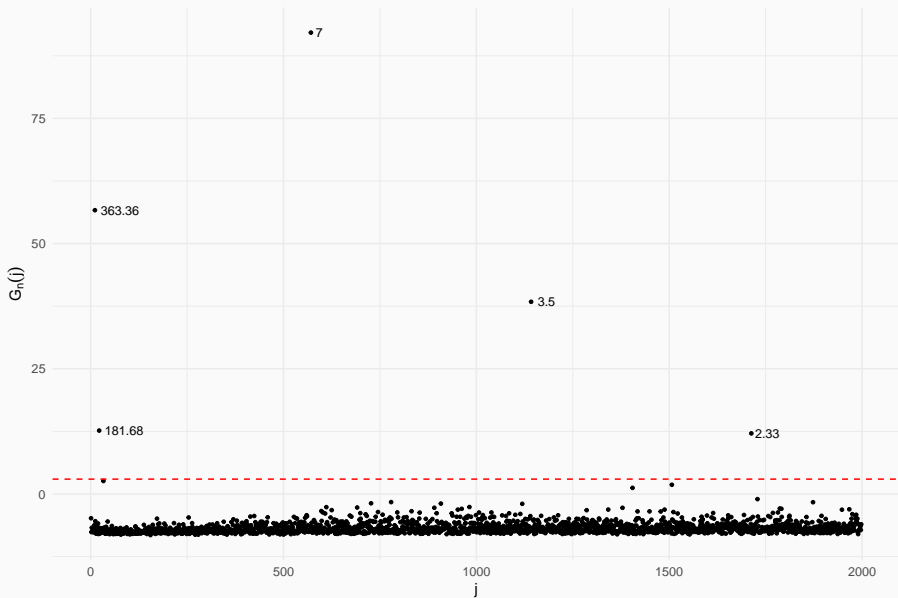
PM10 time series



PM10 time series

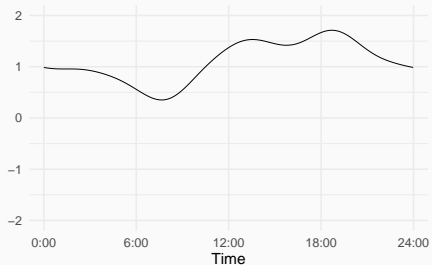


PM10 time series

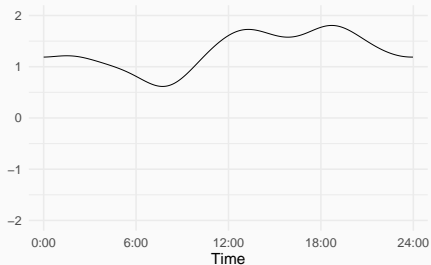


Yearly periodic component

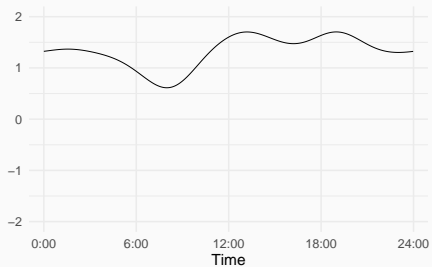
Monday



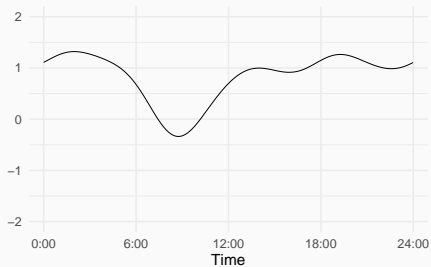
Wednesday



Friday

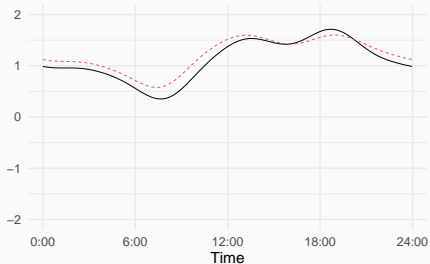


Sunday

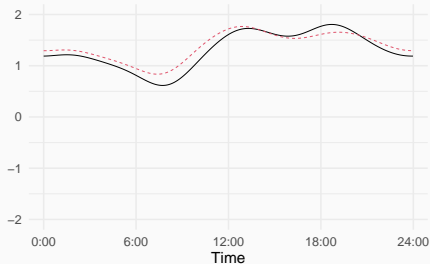


Periodic component

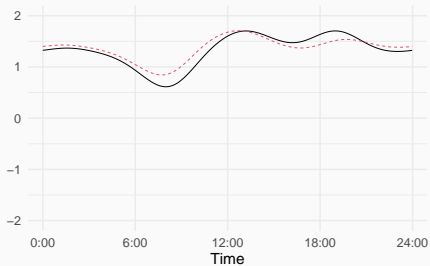
Monday



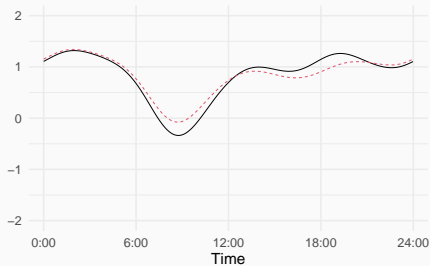
Wednesday



Friday

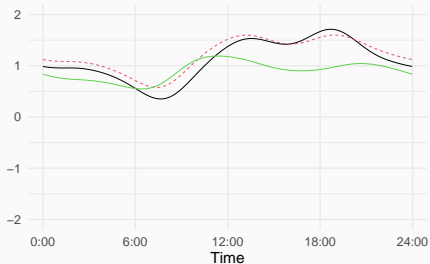


Sunday

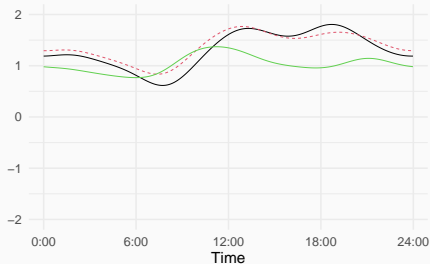


Periodic component

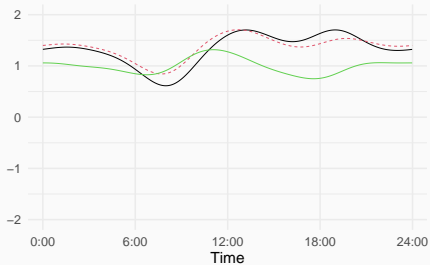
Monday



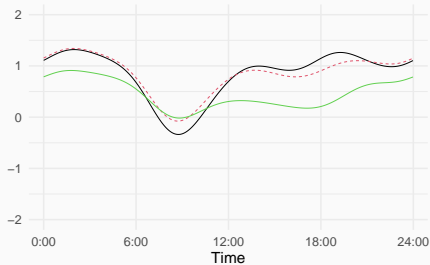
Wednesday



Friday

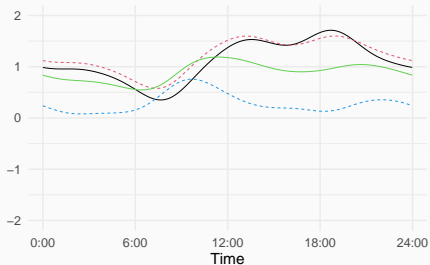


Sunday

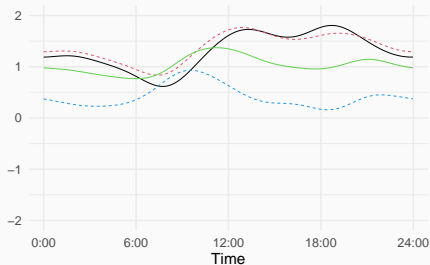


Periodic component

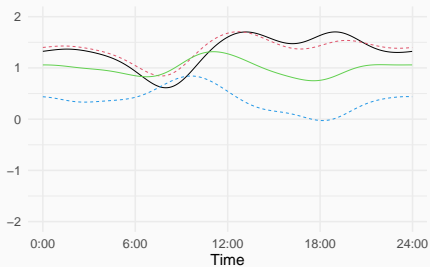
Monday



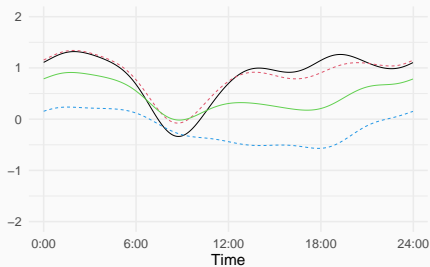
Wednesday



Friday

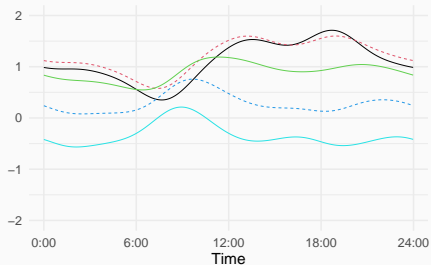


Sunday

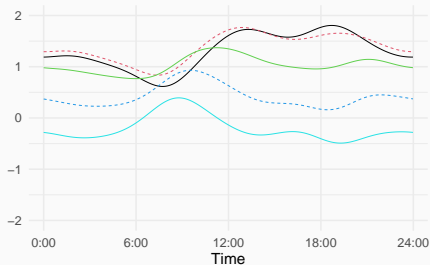


Periodic component

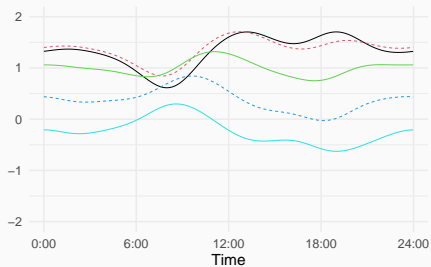
Monday



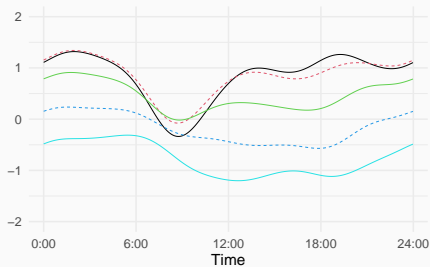
Wednesday



Friday

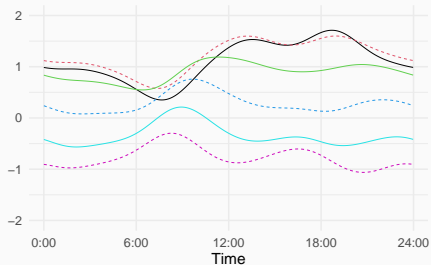


Sunday

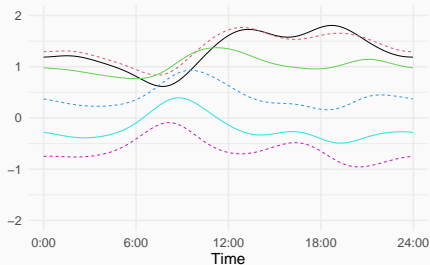


Periodic component

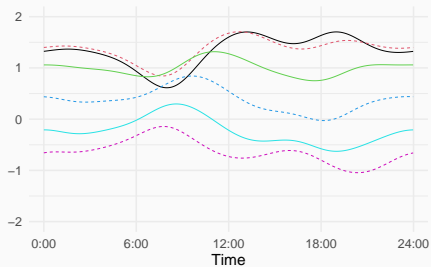
Monday



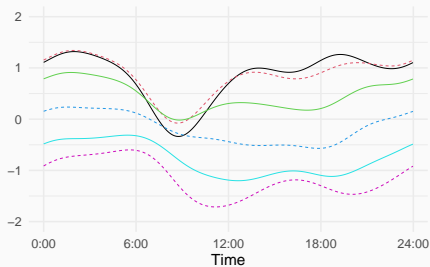
Wednesday



Friday

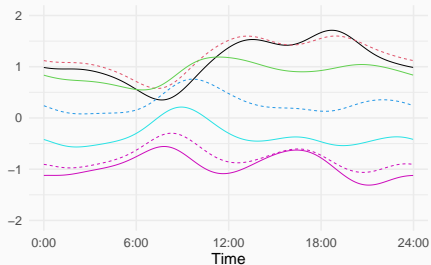


Sunday

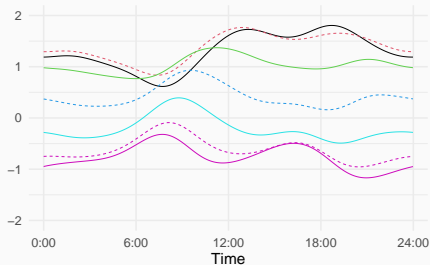


Periodic component

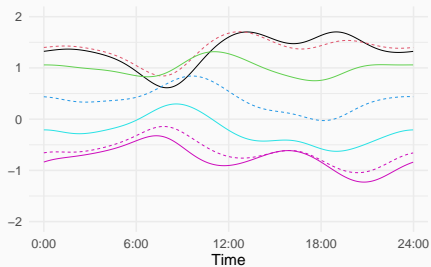
Monday



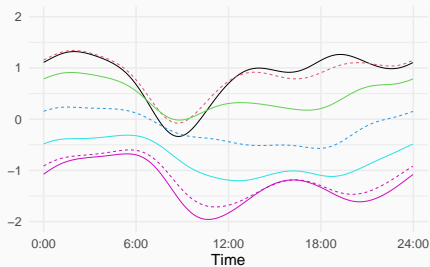
Wednesday



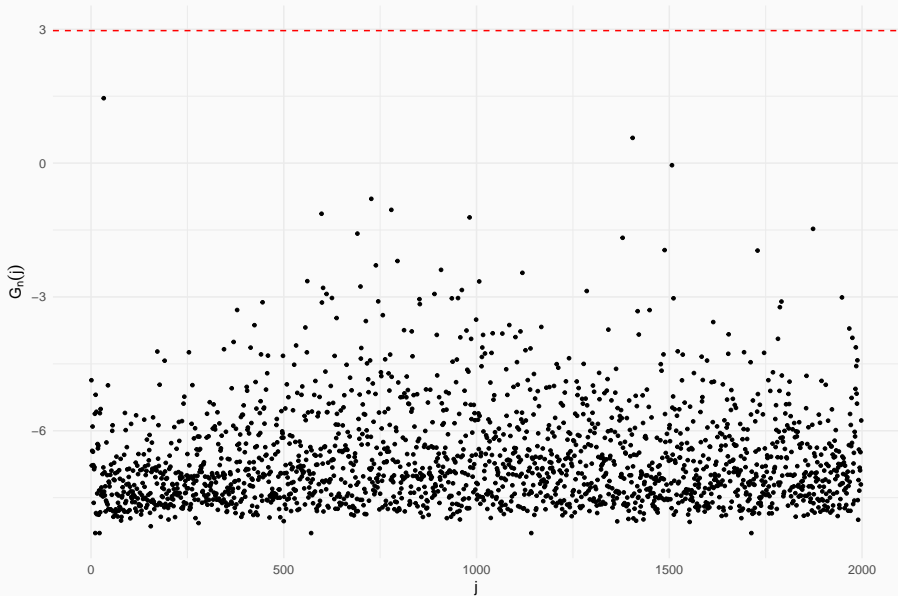
Friday



Sunday



Deseasonalized data



Summary

Summary

- 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.
- 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 are actually driven by several sinusoids.

<https://imada.sdu.dk/~characiejus/>