Detection of periodic signals in functional time series

Vaidotas Characiejus[a]

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

CMStatistics 2020 / December 19, 2020

[a]Department of Statistics, University of California, Davis, USA

[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

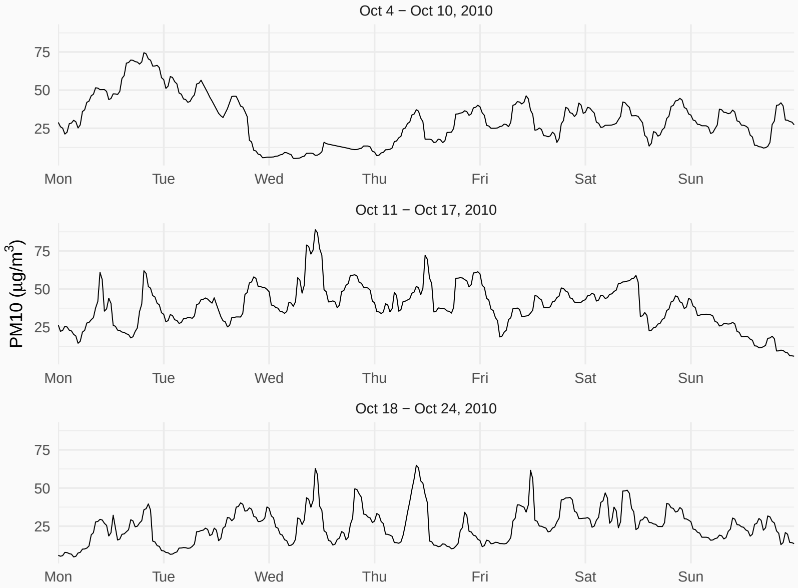
Outline

Motivation

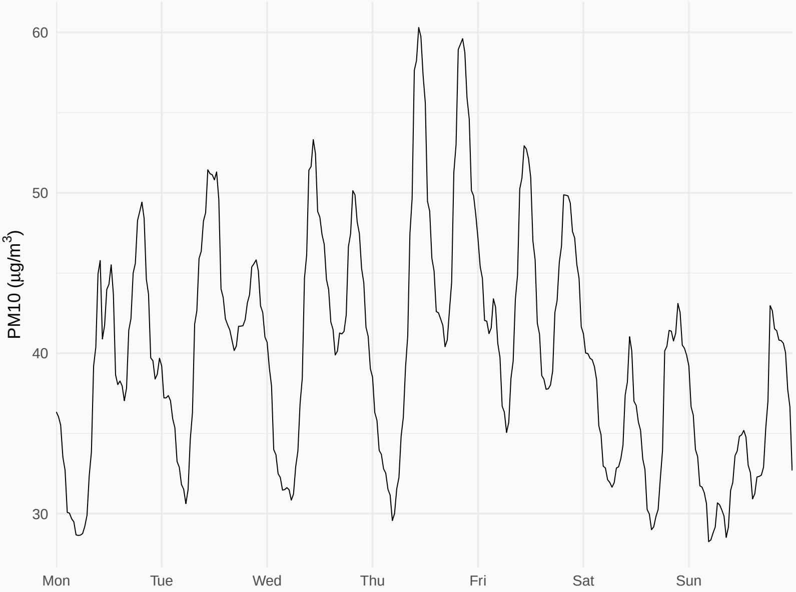
PM10 data

- 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.
- Data set consists of 182 observation days in the winter season of 2010–2011 (October – March) and the amount of PM10 in μg/m^3 is recorded every 30 minutes.

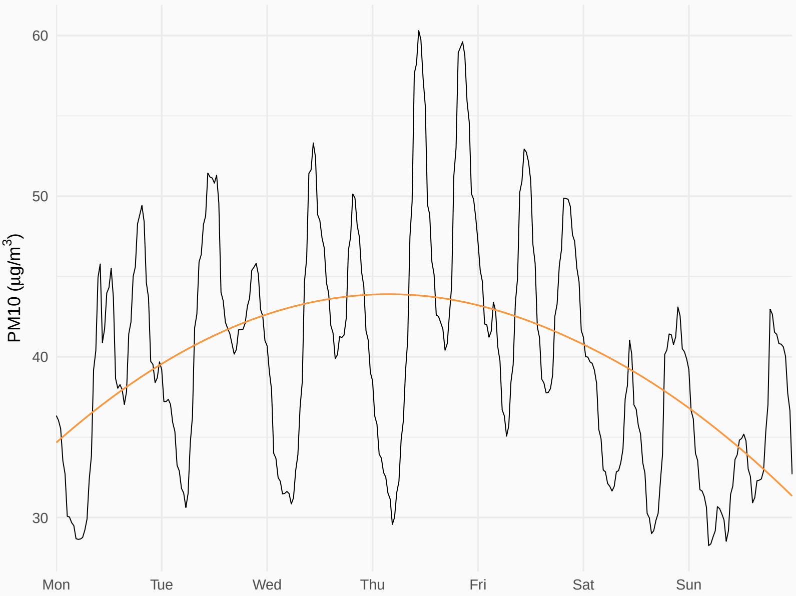
Raw PM10 data



Mean PM10 curves



Mean PM10 curves



Functional time series

- We model the PM10 data as a functional time series $X_1, \dots, X_{182}$ where each curve X_t with $t = 1, \dots, 182$ represents a single day.
- Such segmentation accounts for a daily 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 a time series with values in a separable Hilbert space $\mathbb{H}$ given by

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

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

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

for all $t \in \mathbb{Z}$ with some $d > 1$ and $\{X_t\}_{t\in\mathbb{Z}}$ is a stationary sequence of zero mean random elements with values in $\mathbb{H}$.

Problem

- Our goal is to develop a methodology to detect a periodic component when $d > 1$ is not assumed to be known.
- Specifically, we want to test the following hypotheses:
  $\mathcal{H}_0$: observations are generated by a stationary sequence (no periodic component);
  $\mathcal{H}_1$: observations are generated by a stationary sequence with a superimposed deterministic periodic component with an unknown period $d > 1$.

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

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

with $i = \sqrt{-1}$ for $\omega \in [-\pi, \pi]$ and $n \geq 1$.

Definition

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

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

for $\omega \in [-\pi, \pi]$ and $n \geq 1$.

Maximum of periodogram

The test statistic is given by

$$M_n = \max_{1 \le j \le q} |||I_n(\omega_j)|||_2 = \max_{1 \le j \le q} \|\mathcal{X}_n(\omega_j)\|^2$$

for $n > 2$, where

i) $|||\cdot|||_2$ is the Hilbert-Schmidt norm and $\|\cdot\|$ is the norm induced by the inner product of $\mathbb{H}$;

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

iii) $q = \lfloor (n-1)/2 \rfloor \sim n/2$ as $n \to \infty$.

Asymptotic results

Results in the univariate case

- The usefulness of the maximum of the periodogram for detecting periodicities is well known [Fisher (1929)].
- First results in the univariate case were established under the assumption of Gaussianity.
- Davis and Mikosch (1999) established the asymptotic distribution of the appropriately standardized M_n provided that $\mathrm{E}\,|X_1|^s < \infty$ with $s > 2$ using a Gaussian approximation technique due to Einmahl (1989).

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).

Linear processes

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

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

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

- $\{a_k\}_{k\in\mathbb{Z}} \subset \mathcal{L}(\mathbb{H})$ 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

- The impulse-response operator $A(\omega)$ defined 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 covariance operator $\mathrm{E}[\varepsilon_0 \otimes \varepsilon_0]$.

Assumptions

Assumption 1

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

ii) $\lambda_k > \lambda_{k+1}$ for $k \geq 1$;

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

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

Theorem

Suppose that Assumption 1 and Assumption 2 hold. Then

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

where

- $b_n = \lambda_1 \log q - \lambda_1 \sum_{j=2}^{\infty} \log(1-\lambda_j/\lambda_1)$;
- $q = \lfloor (n-1)/2 \rfloor$;
- $\mathcal{G}$ 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 $\{X_t\}_{t\in\mathbb{Z}}$ is an FAR(1) model given by

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

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

Assumption 3

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

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*

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

Under $\mathcal{H}_0$ *and Assumptions 1, 2, and 3, we have that*

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

as $n \to \infty$.

Empirical study

Simulation setting

- We simulate functional time series that are stationary and behave 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 $\lambda = 15$.
- We consider the situation when $a = 0, 1, 2$, where $a = 0$ corresponds to $\mathcal{H}_0$.

Empirical rejection rates

		$a=0$			$a=1$			$a=2$		
	α	0.1	0.05	0.01	0.1	0.05	0.01	0.1	0.05	0.01
$\lambda=5$	$n=100$	0.066	0.029	0.004	0.861	0.799	0.670	1.000	0.999	0.993
	$n=200$	0.082	0.038	0.006	0.989	0.983	0.970	1.000	1.000	1.000
	$n=500$	0.093	0.054	0.011	1.000	1.000	0.999	1.000	1.000	1.000
$\lambda=15$	$n=100$	0.082	0.041	0.005	0.249	0.165	0.071	0.818	0.758	0.606
	$n=200$	0.071	0.035	0.006	0.569	0.471	0.293	0.985	0.973	0.922
	$n=500$	0.096	0.045	0.007	0.990	0.978	0.942	1.000	1.000	1.000

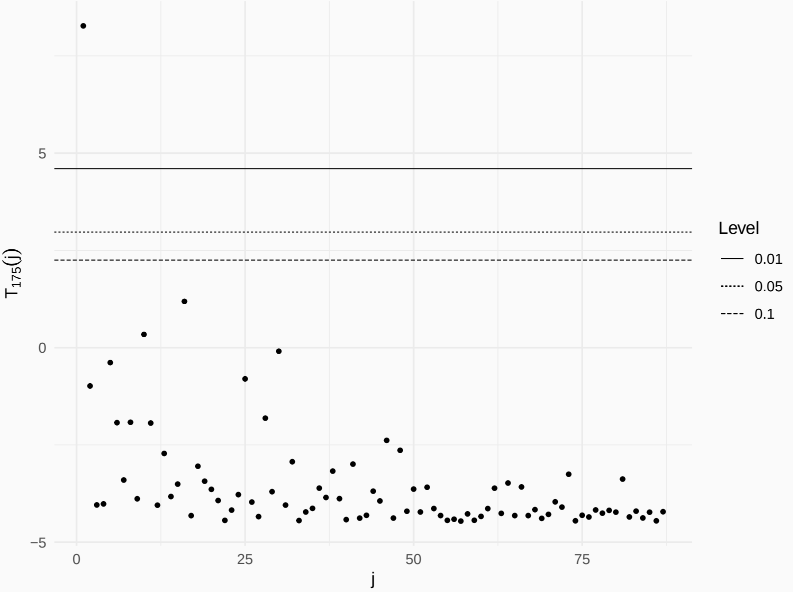
PM10 time series

- We consider the square-root transformation of the PM10 time series.
- We plot the values of the test statistic

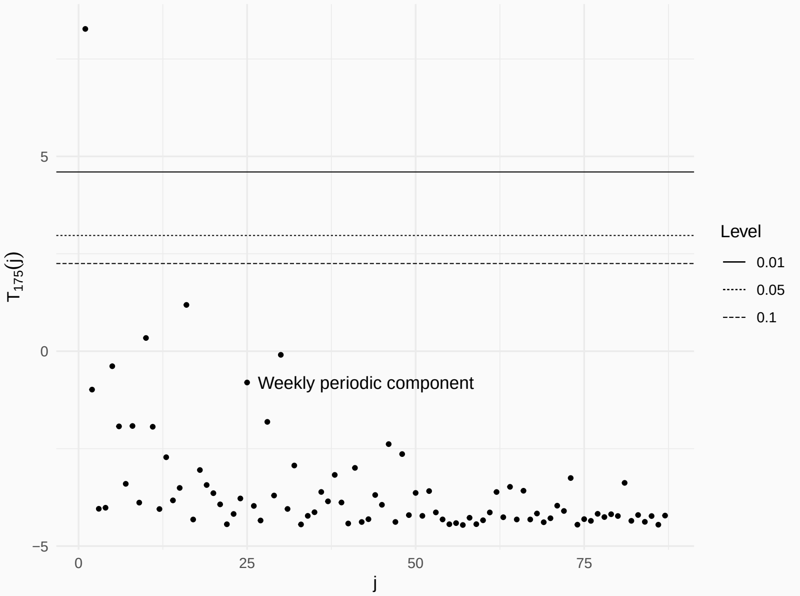
$$T_n(j) := \hat{\lambda}^{-1} \|(I - e^{-i\omega_j} \hat{\rho})\mathcal{Y}_n(\omega_j)\|^2 - \log q + \sum_{j=2}^{a_n} \log(1 - \hat{\lambda}_j / \hat{\lambda}_1)$$

for $j = 1, \ldots, q = 87$.

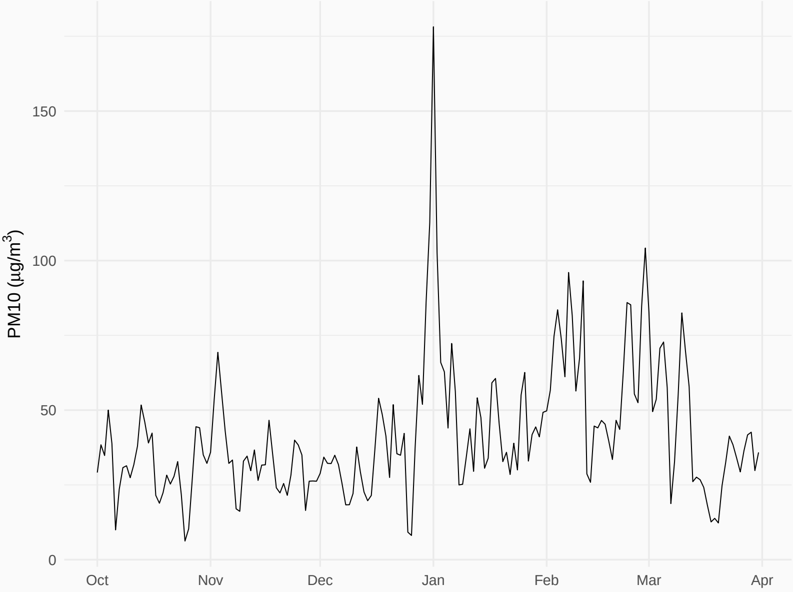
PM10 time series



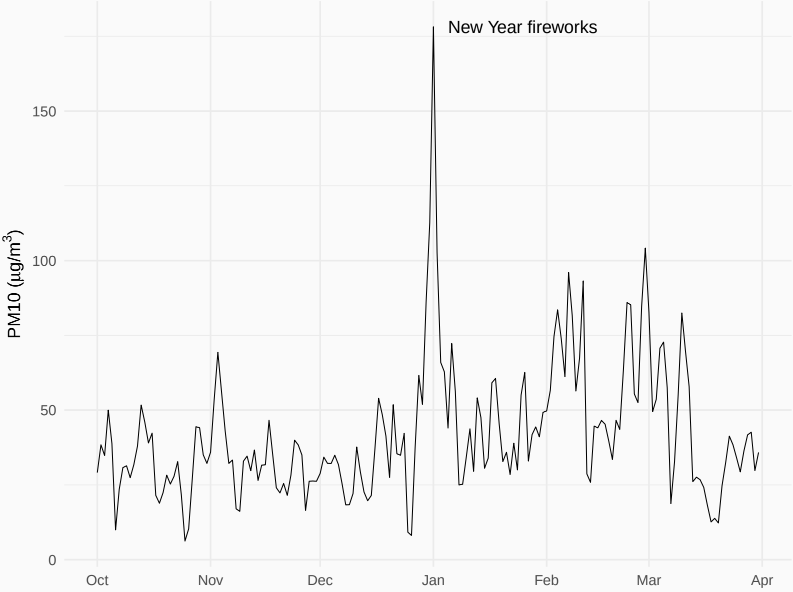
PM10 time series



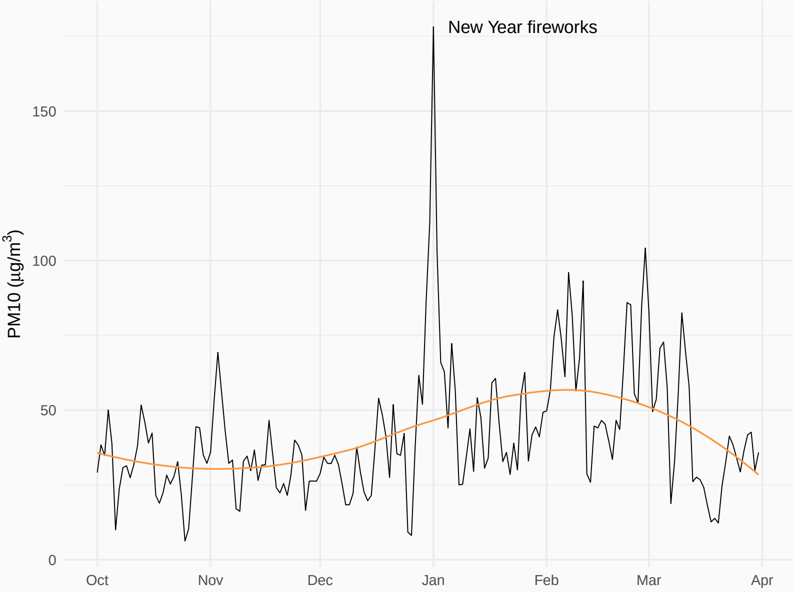
Daily means of PM10



Daily means of PM10



Daily means of PM10



Summary

Concluding remarks

- We propose a general test for periodicities in Hilbert space valued time series when the length of the period is unknown.
- The test is based on the maximum of the periodogram.
- We establish that the asymptotic distribution of the appropriately standardized test statistic is the Gumbel distribution.
- Empirical study shows that the test performs reasonably well and the analysis of the PM10 time series illustrates the usefulness of our approach as it is capable of detecting periodic signals which are not a priori expected.

`https://www.stat.ucdavis.edu/~vaidas/`