Functional Data Analysis

Łukasz Kidziński

Département de Mathématique

Université Libre de Bruxelles

Outline

- Principal component analysis
- Introduction to Functional Data Analysis
- (Static) Functional PCA
- Dynamic FPCA
- Examples

Statistics

"I can prove anything by statistics except the truth." - George Canning

Statistics

$$F = ma$$

Statistics

$$F = ma + \varepsilon$$

Statistics

Explaining the variance is one of the main concerns.

Multivariate statistics

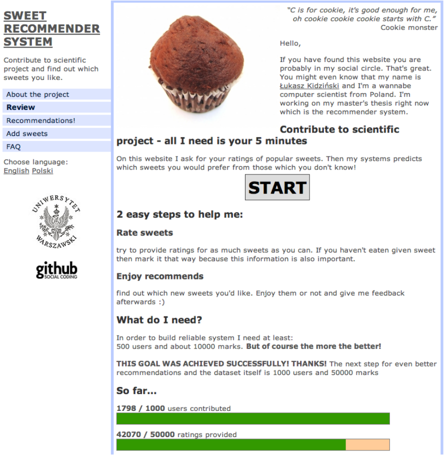


Figure 1 : Sweets Recommender System

	1	2	3	4	5	8	9	10	11	12	13	14	15	16	17	18
10	1	1	2	2	5	4	5	1	3	1	2	5	2	3	3	3
11	5	4	4	1	5	5	2		4	2	3	1	3	3	1	2
12	2	3	1	2	4	5	3	4	1	3	4	1	2	3	3	1
13	3	3		1	4	4	2		5	5	5	3	4	2	3	4
14	4	4	5	1	2	5	4	5	5	2	3	3	2	2	2	4
15	4	4		5	4	4	3			3	4		4	4		1
16	5	2	4	1	5	1	3		3	3	4		4	4	2	5
17	4	4		4	5	5	2		3	3	4		5	3	1	3
18	4	4	3	5	5	5	1		3	5	5	4	4	4	4	3
19	5	4	1	5	3	5	2	1	1	5	5	5	3	4	5	1
20	4	3		4	5	5	1		3	4	3	2	2	2	4	4
21	3	3	1	4	3	5	3	2	3	3	4	5	2	3	3	2

Table 1 : Ratings of some sweets. Users in rows and sweets in columns.

Principal Component Analysis

Let $X \in R^d$ zero-mean random vector

Let (λ_i, e_i) eigenvalues and eigenvectors of $C = \mathrm{Var}(X)$, in non-increasing order of λ_i.

We define *m*-th principal component

$$Y_m = X' e_m.$$

We have

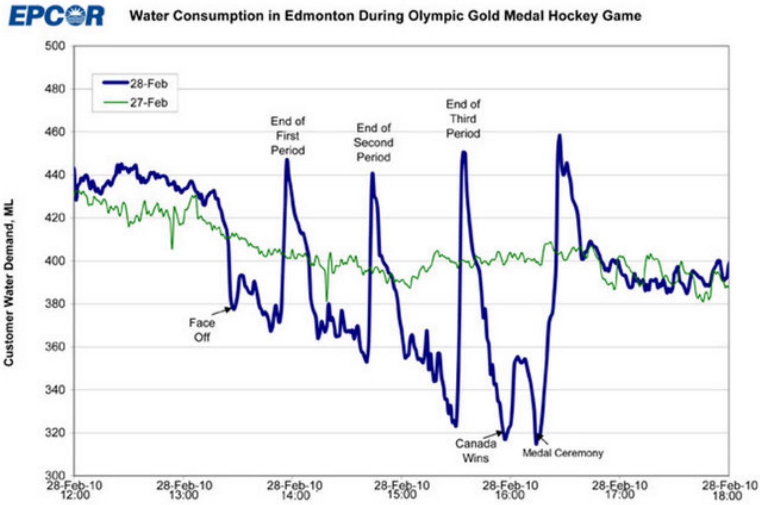
$$\mathrm{Var}(Y) = \mathrm{diag}(\lambda_1, \lambda_2, \ldots).$$

Moreover, (e_i) form an orthonormal basis and thus

$$X = \sum_{m=1}^{d} Y_m e_m.$$

Classical solution found by Pearson (1901) and Hotelling (1933).

Time series



EPCOR
Water Consumption in Edmonton During Olympic Gold Medal Hockey Game
Customer Water Demand, ML
500
480
460
440
420
400
380
360
340
320
300
28-Feb
27-Feb
End of First Period
End of Second Period
End of Third Period
Face Off
Canada Wins
Medal Ceremony
28-Feb-10 12:00
28-Feb-10 13:00
28-Feb-10 14:00
28-Feb-10 15:00
28-Feb-10 16:00
28-Feb-10 17:00
28-Feb-10 18:00

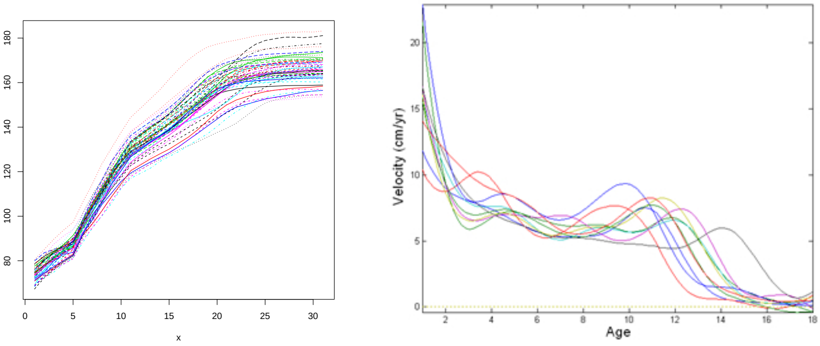


Figure 2 : Berkeley Growth Data: Heights of 20 girls taken from ages 0 through 18 (left). Growth process easier to visualize in terms of speed of growth (right). Tuddenham and Snyder (1954) and Ramsey and Silverman (1997)

EMG

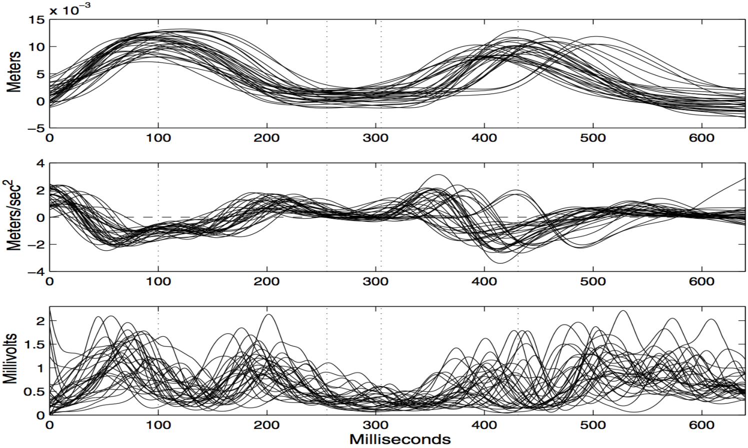


Figure 3 : Lower lip movement (top), acceleration (middle) and EMG of a facial muscle (bottom) of a speaker pronouncing the syllable “bob” for 32 replications. Malfait, Ramsay, and Froda (2001)

Minicircles

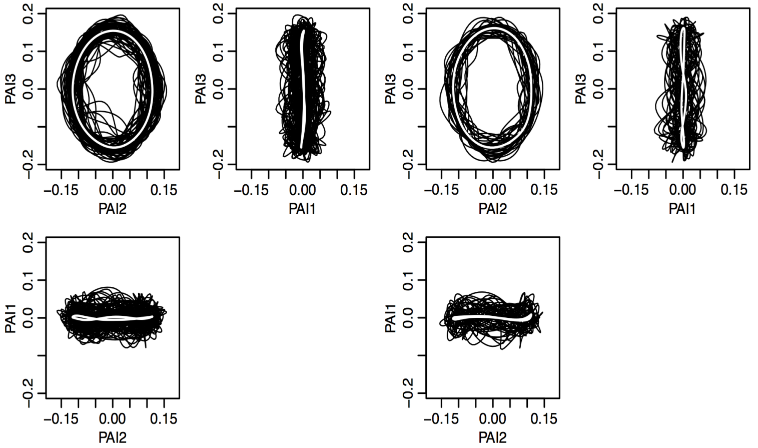


Figure 4 : Projections of DNA minicircles on the planes given by the principal axes of inertia (three panels on the left side: TATA curves, right: CAP curves). Mean curves are plotted in white. Panaretos, Kraus and Maddocks (2011)

Motivation

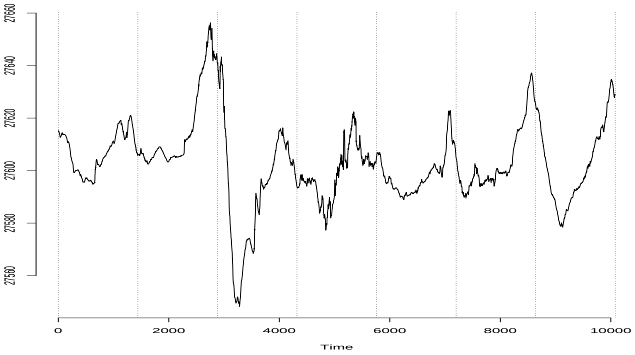


Figure 5 : Horizontal component of the magnetic field measured in one minute resolution at Honolulu magnetic observatory from 1/1/2001 00:00 UT to 1/7/2001 24:00 UT. 1440 measurements per day. Gabrys and Kokoszka (2007)

Functional data representation

A (theoretical) problem in FDA is the dimensionality of the model.

Functions are intrinsically infinite dimensional.

How can we store functional data?

We can consider a function as an infinite vector of coefficient in an orthonormal basis.

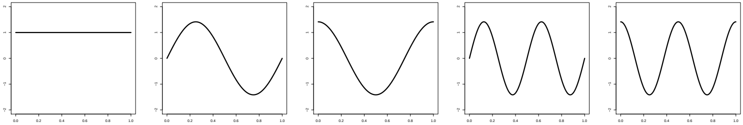


Figure 6 : First 5 Fourier basis functions

Example

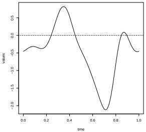


Figure 7 : A function with following Fourier coefficients:
-0.45, 0.55, 0.051, -0.45, 0.12, -0.07, -0.06, **0.13, -0.14**, 0.04, 0.00, ...

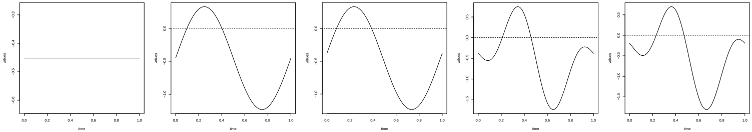


Figure 8 : Projections on the space spanned by 1,2,3,4,5 basis functions

Dimension reduction

Techniques for dimension reduction are even more important than in multivariate statistics.

Functional principal component analysis (FPCA) is considered as a key techniques in FDA.

The idea is to find the "optimal" basis. As a measure we take the L_2 distance from projection:

$$E\left\|X - \sum_{i=1}^{p}\langle X, e_i\rangle e_i\right\|^2 \leq E\left\|X - \sum_{i=1}^{p}\langle X, b_i\rangle b_i\right\|^2,$$

for any ONB (b_i).

Example $H = L^2([0,1])$

We need a notion of mean and variance. This can be easily establish for random variables in Hilbert space.

The mean is given by $(EX)(t) = E(X(t))$ for $t \in [0,1]$

The covariance operator is given by

$$C(y)(t) = \int_0^1 \mathrm{Cov}(X(t), X(s))y(s)ds.$$

(just a continous extension of $(Cv)_k = \sum_{l=1}^d Cov(X_k, X_l)v_l$.)

C is a symmetric, non-negative definite, Hilbert-Schmidt operator $\Longrightarrow$ we can find it's eigenfunctions.

FPCs

Suppose for the rest of the presentation that $\{X_t\}$ is zero mean and weakly stationary taking values in $H = L^2$.

Let $\lambda_1 \geq \lambda_2 \geq \dots$ (assumption) be the eigenvalues of C and let $e_1, e_2, \dots$ be the eigenfunctions (standardized to unit length).

The functional principal components are defined as

$$Y_m = \langle X, e_m \rangle.$$

We have

$$\mathrm{Var}(Y) = \mathrm{diag}(\lambda_1, \lambda_2, \dots).$$

Karhunen-Loève expansion: In a separable Hilbert space (e_i) form an ONB and thus

$$X = \sum_{i=1}^{\infty} Y_m e_i.$$

Example

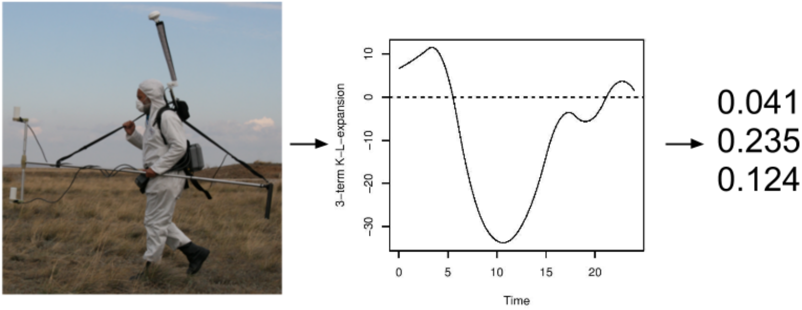


Figure 9 : Imagine you want to measure the intraday variation of the magnetic field in your garden but you can send only 3 real numbers per day.

Example

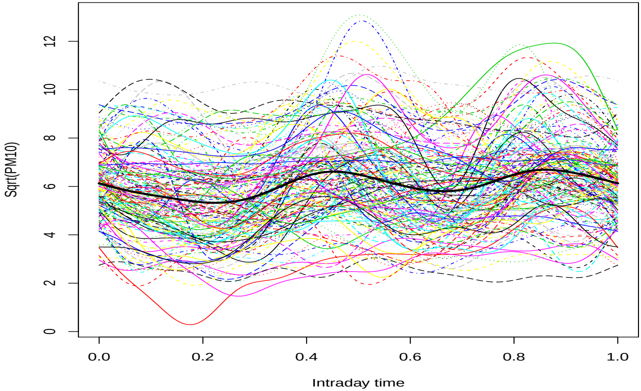


Figure 10 : We display $\sqrt{x_t(u)}$, $1 \leq t \leq 175$, where $x_t(u)$ are daily functional observations of PM10 represented with 15 Fourier basis functions. The solid black line represents the sample mean curve.

Example

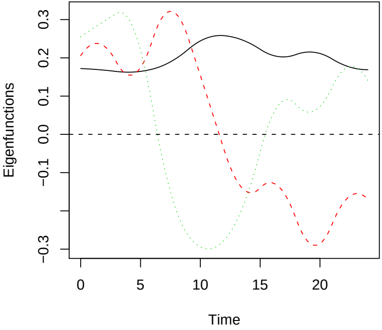


Figure 11 : The eigenfunctions e_1 (black), e_2 (red) and e_3 (green) belonging to the three largest eigenvalues of the corresponding covariance operator. First 3 principal components explain 84% of variability.

Example

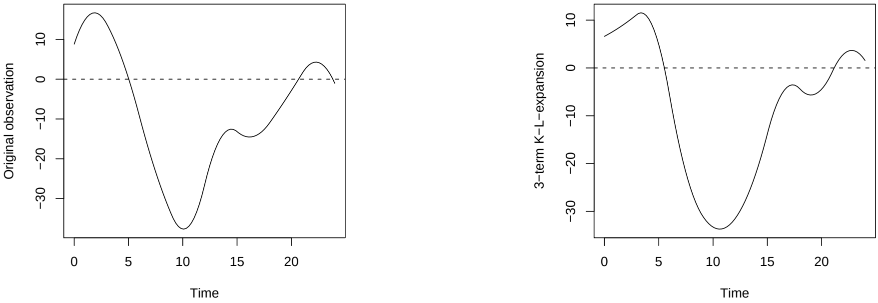


Figure 12 : A comparison of $x_1 - \bar{x}$ and $\sum_{i=1}^{3} \langle x_1 - \bar{x}, \hat{e}_i \rangle \hat{e}_i$.

Motivation

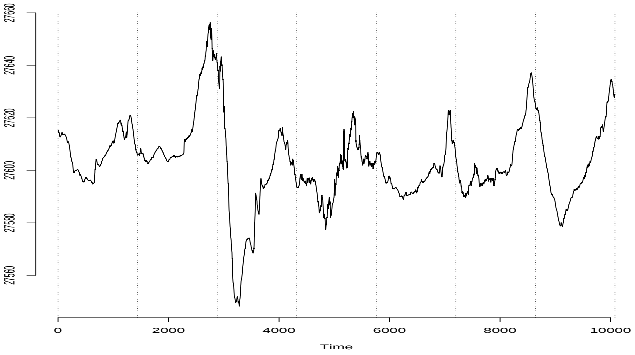


Figure 13 : Horizontal component of the magnetic field measured in one minute resolution at Honolulu magnetic observatory from 1/1/2001 00:00 UT to 1/7/2001 24:00 UT. 1440 measurements per day. Gabrys and Kokoszka (2007)

Dynamic FPCA: Motivation

Let $Y_1 = (Y_{11}, \ldots, Y_{p1})'$.

1. We ignore valueable information hidden in the serial dependence of the data.

   $\Longrightarrow$ more efficient compression may be possible.

2. Though

   $$\mathrm{Var}(Y_1) = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p),$$

   we have (unlike in the iid case)

   $$\mathrm{Cov}(Y_1, Y_h) \quad \text{non-diagonal (in general).}$$

   $\Longrightarrow$ needs to be studied as a multivariate series.

Dynamic FPCA: Motivation

Instead of

$$Y_{mt} = \langle X_t, e_m \rangle$$

we consider filter $(\phi_{mk})_{m \in N, k \in Z}$ and the convolution

$$Y_{mt} = \sum_{l \in Z} \langle X_{t-k}, \phi_{mk} \rangle.$$

We claim that there exists $(\phi_{mk})_{m \in N, k \in Z}$ s.t.

- For $k \neq m$, and any $s, t \in Z$

$$Cov(Y_{ks}, Y_{mt}) = 0,$$

- Reconstructed

$$\tilde{X}_t = \sum_{m=1}^{n} \sum_{t \in Z} Y_{mt} \phi_{mt}$$

  has the L_2 smaller than FPCA.

We propose a construction of such a filter $(\phi_{mk})_{m \in N, k \in Z}$.

Dynamic FPCA

Rather than only using

$$C(y)(t) = \int_0^1 \mathrm{Cov}(X_0(t), X_0(s))y(s)ds.$$

in the derivation functional PCs we want to use $(C_h \colon h \in \mathbb{Z})$ where

$$C_h(y)(t) = \int_0^1 \mathrm{Cov}(X_0(t), X_h(s))y(s)ds.$$

This leads to the concept of

SPECTRAL DENSITY OPERATORS

$$\Gamma_\theta^X = \sum_h C_h e^{-ih\theta}.$$

Converges if $\sum_h \|C_h\|_{\mathcal{S}} < \infty$. (Key assumption.)

Brillinger (1975) for multivariate data. Now, in the functional context, Panaretos and Tavakoli (2013), Hörmann, **Kidziński** and Hallin (2013).

Dynamic FPCA

Definition (Dynamic functional principal components)

Assume that $(X_t \colon t \in \mathbb{Z})$ is a mean zero, stationary process with values in $L^2([0,1])$ and summable cross-covariances. Then the *m-th dynamic functional principal component (DFPC) score of* X_t is defined by

$$Y_{mt} := \sum_{\ell \in \mathbb{Z}} \langle X_{t-\ell}, \phi_{m\ell} \rangle, \quad t \in \mathbb{Z},\ m \geq 1. \tag{1}$$

We call $\Phi_m := (\phi_{m\ell} \colon \ell \in \mathbb{Z})$ the *m-th DFPC filter.*

Properties

We have proven that

- Y_{mt} is real-valued
- If there is no temporal dependence (C_h) our method coincide with PCA
- For $m \neq m'$, the principal components Y_{mt} and $Y_{m's}$ are uncorrelated for all s, t.
- Y_{mt} is optimal in the sense that

$$E\|X_t - \sum_{m=1}^{p} X_{mt}\|^2 \leq E\|X_t - \sum_{m=1}^{p} \tilde{X}_{mt}\|^2.$$

for any different p-dimensional filter $(\phi'_{m\ell})_{1\leq m\leq p}$.

Implementation (in a nutshell)

(I) Transform discrete data into functions by basis functions approach:

$$\mathbf{X} \mapsto X(t) = Z_1 v_1(t) + \ldots + Z_d v_d(t) = \mathbf{v}'(t)\mathbf{Z}.$$

(II) We are working on finite dimensional space now. Operators have a corresponding matrix. This matrix acts on the coefficients of the curves.

For example

$$\mathfrak{G}_\theta^X = \left(\sum_{h \in \mathbb{Z}} C_h^{\mathbf{Z}} e^{-\mathbb{i} h\theta} \right) V',$$

when $V := (\langle v_i, v_j \rangle \colon 1 \leq i, j \leq d)$.

Implementation (in a nutshell)

(III) Eigenfunctions/values of Γ_θ^X correspond to eigenvectors/values of $\mathfrak{G}_\theta^X$: $\varphi_m(\theta) = \mathbf{v}'\boldsymbol{\varphi}_m(\theta)$. Then

$$\phi_{mk} = \frac{\mathbf{v}'}{2\pi}\int_{-\pi}^{\pi} \boldsymbol{\varphi}_m(s)e^{-\mathbb{i}ks}ds =: \mathbf{v}'\boldsymbol{\phi}_{mk},$$

and

$$Y_{mt} = \sum_{k\in\mathbb{Z}}\int_0^1 \mathbf{Z}'_{t-k}\mathbf{v}(u)\mathbf{v}'(u)\boldsymbol{\phi}_{mk}du = \sum_{k\in\mathbb{Z}}\mathbf{Z}'_{t-k}V\boldsymbol{\phi}_{mk}.$$

(IV) All involved quantities need to be estimated (consistency results have been established).

(V) Numerical integration is used.

Real data example

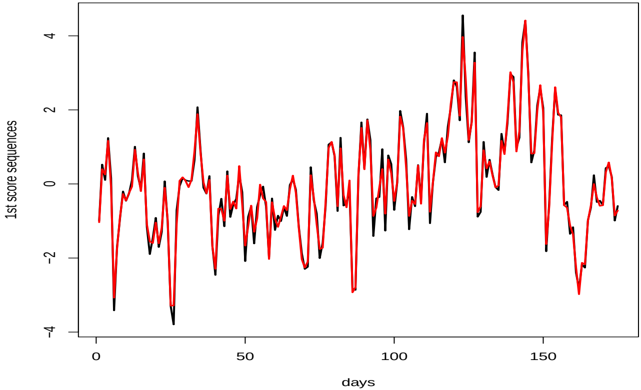


Figure 14 : The sequence of the first static FPC scores (red) [73%] and the dynamic ones (black) [80%].

Real data example

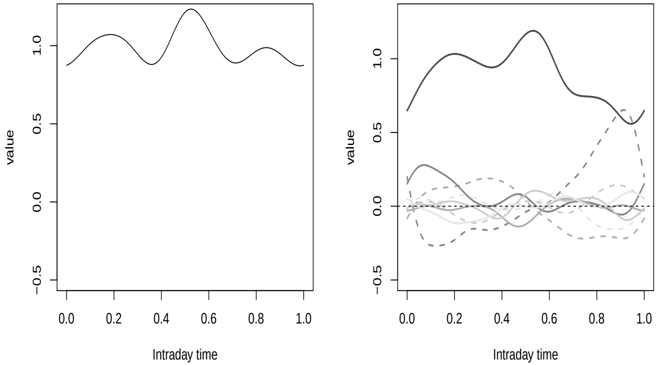


Figure 15 : First static FPC (left) and filters corresponding to first dynamic FPC. The filters $\hat{\phi}_{1t}(u)$, $t \geq 1$, are dashed and for $i \leq 0$ solid. The larger $|t|$, the lighter the curve.

Real data example

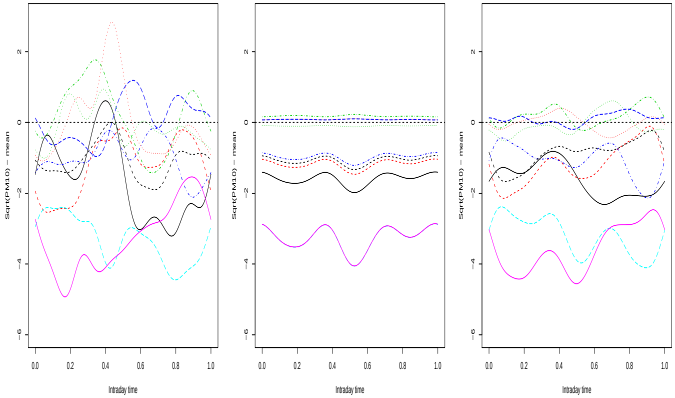


Figure 16 : We show 10 subsequent observations (left panel), the corresponding static KL expansion with one component (middle panel) and the dynamic KL expansion with one component (right panel).

Conclusion

- Nature is a Hilbert space!
- One can analyse a time series of anything
- Dynamic PCA in the time series context:
  - explains more variance,
  - gives uncorrelated processes.

Thank you for your attention

Dynamic FPCA

One can show that applying the functional filter $\Phi = (\phi_k)$ to (X_t),

$$X_t \mapsto Y_t = \sum_{k \in \mathbb{Z}} \phi_k(X_{t-k}),$$

with $\phi_k(x) = (\langle x, \phi_{1k} \rangle, \ldots, \langle x, \phi_{pk} \rangle)'$, results in spectral density matrix

$$\Gamma_\theta^Y = \begin{pmatrix} \langle \Gamma_\theta^X(\phi_1^\star(\theta)), \phi_1^\star(\theta) \rangle & \cdots & \langle \Gamma_\theta^X(\phi_p^\star(\theta)), \phi_1^\star(\theta) \rangle \\ \vdots & \ddots & \vdots \\ \langle \Gamma_\theta^X(\phi_1^\star(\theta)), \phi_p^\star(\theta) \rangle & \cdots & \langle \Gamma_\theta^X(\phi_p^\star(\theta)), \phi_p^\star(\theta) \rangle \end{pmatrix},$$

where $\phi_m^\star(\theta) := \sum_{k \in \mathbb{Z}} \phi_{mk} e^{\mathrm{i} k\theta}$.

If $\phi_m^\star(\theta)$ are eigenfunctions of Γ_θ^X, then cross spectra are zero.

Let $(\lambda_m(\theta), \varphi_m(\theta))$ be m-th eigenvalue and eigenvector of spectral density Γ_θ^X. We choose $\phi_{mk} \in H$ such that

$$\phi_{mk} = \frac{1}{2\pi} \int_{-\pi}^{\pi} \varphi_m(\theta) e^{-k\mathbb{i}\theta} d\theta.$$

Remarks

1. We can show that Γ_θ^X is
   - symmetric
   - non-negative
   - Hilbert Schmidt

   $\Longrightarrow$ Spectral decomposition of Γ_θ^X is possible.

2. We see that ϕ_{mk} are Fourier-coefficients $\varphi_m(\theta)$.

Cookies

Figure 17 : http://sweetrs.org/