Introduction to functional data analysis

Universitat Autònoma de Barcelona

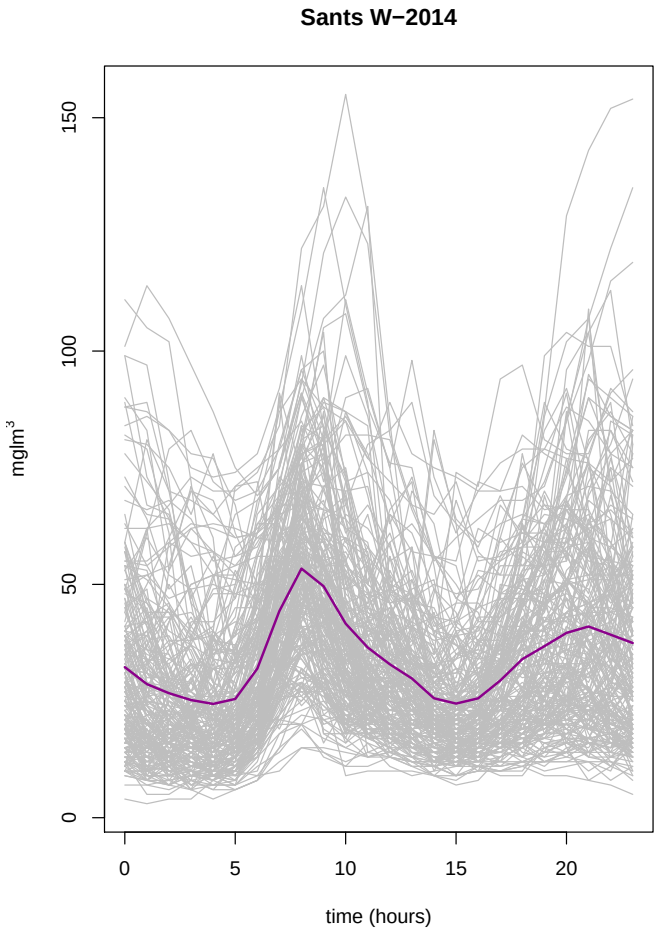
MBD 2018

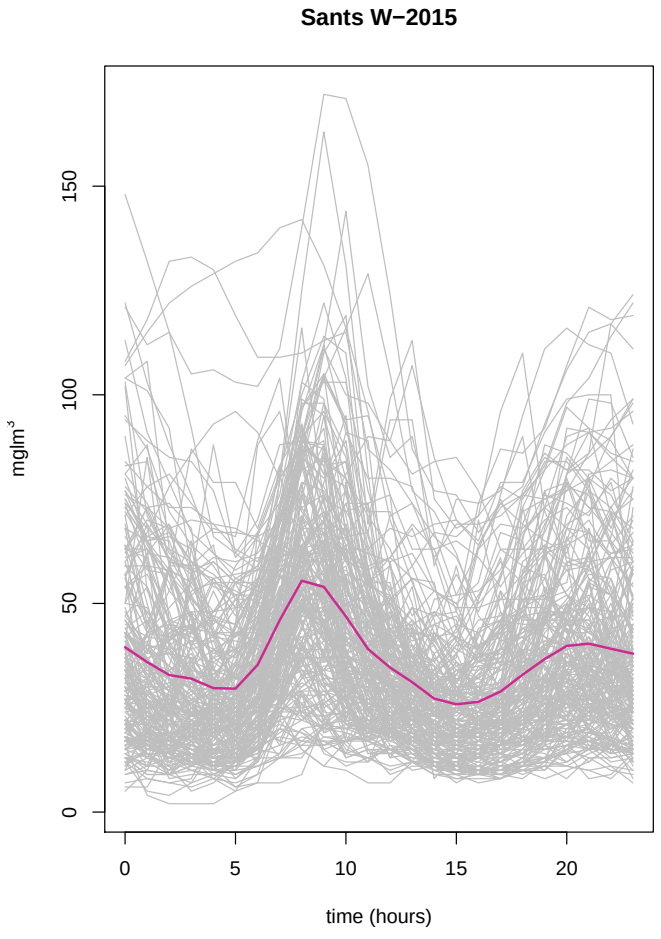
1. Examples of funcional data
2. What do we understand by Functional Data?
3. Descriptive Statistics
4. Depth Measures for Functional Data
5. A two-sample problem for Functional Data
   - On Permutation tests
   - Shilling's type test
   - Depths-based tests
   - Empirical comparison of powers
   - NO_2 Contamination in Barcelona
6. Functional Linear Models

Introduction to FDA

- Observing and saving complete functions as the result of random experiments is possible by the development of real-time measurement instruments and data storage resources.
- Example: For patients involved in a clinical trial, the blood pressure is monitored in continuous-time during 24 hours.
- Samples where a whole curve is observed at each sampling unit are referred to as Functional Data.
- Random functions are the statistical atoms in these cases.

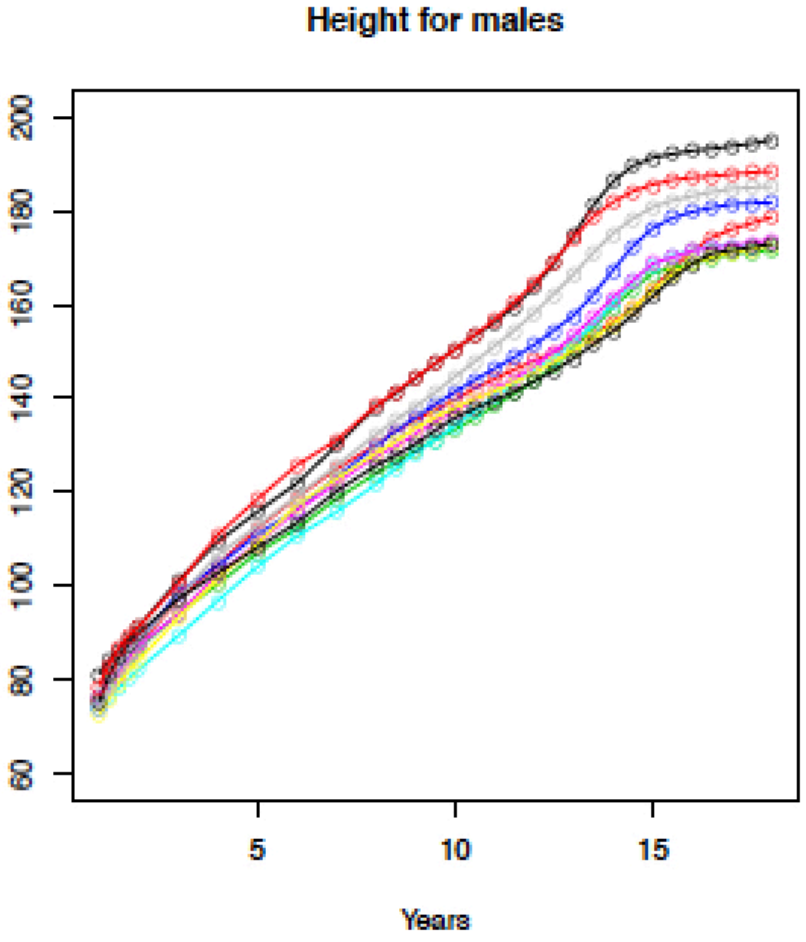
NO_2 levels on working days taken on the same station, in 2014 and 2015. Did (the law) of contamination- levels change?

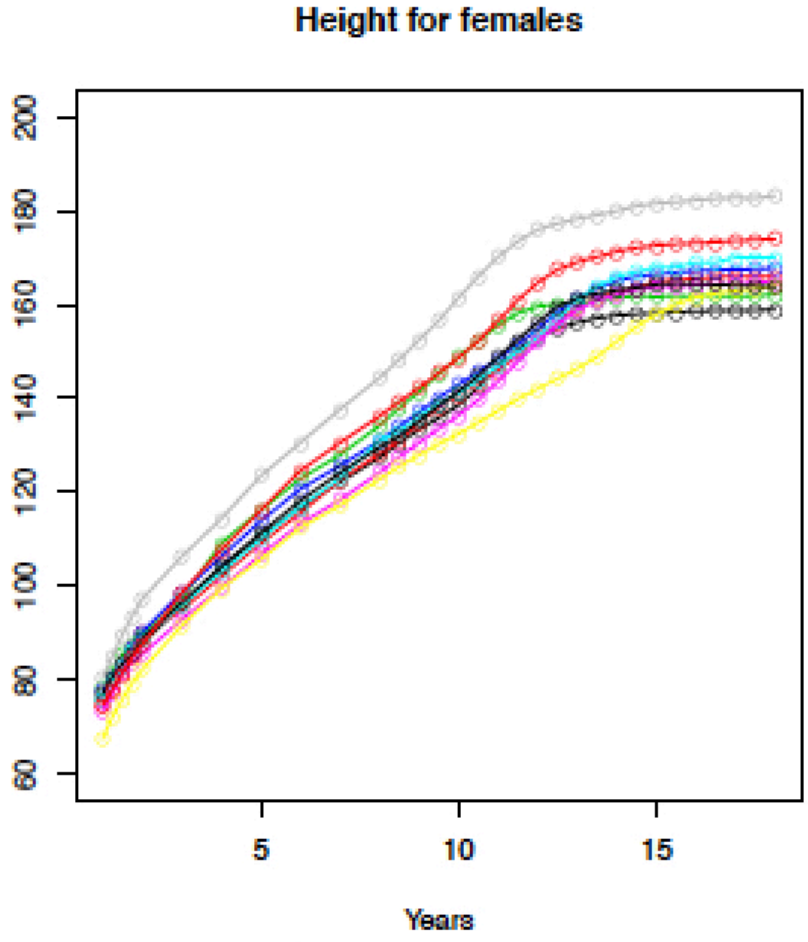




Growth Curves

Heights of 39 boys and 54 girls measured at a set of 31 ages (from 1 to 18 years) in the Berkeley Growth Study. Only 10 are shown.

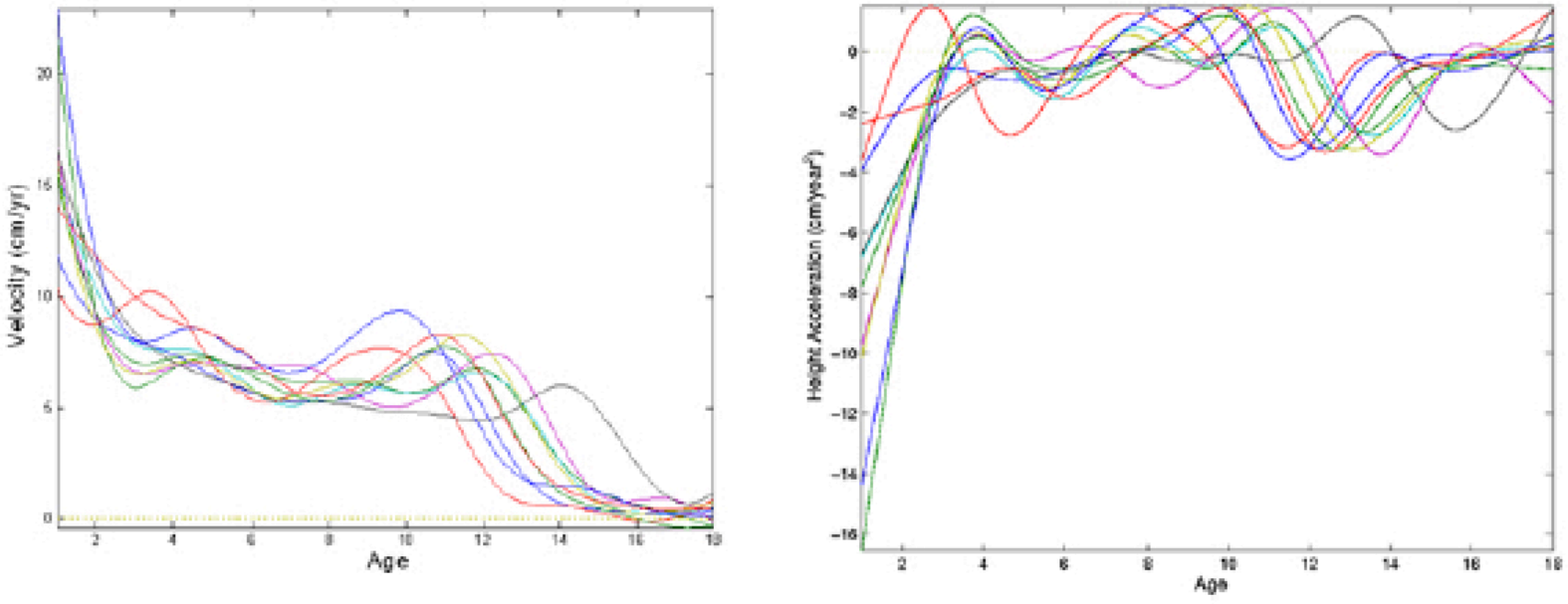





```
> library(fda); data(growth); demo(growth)
```

Growth Curves

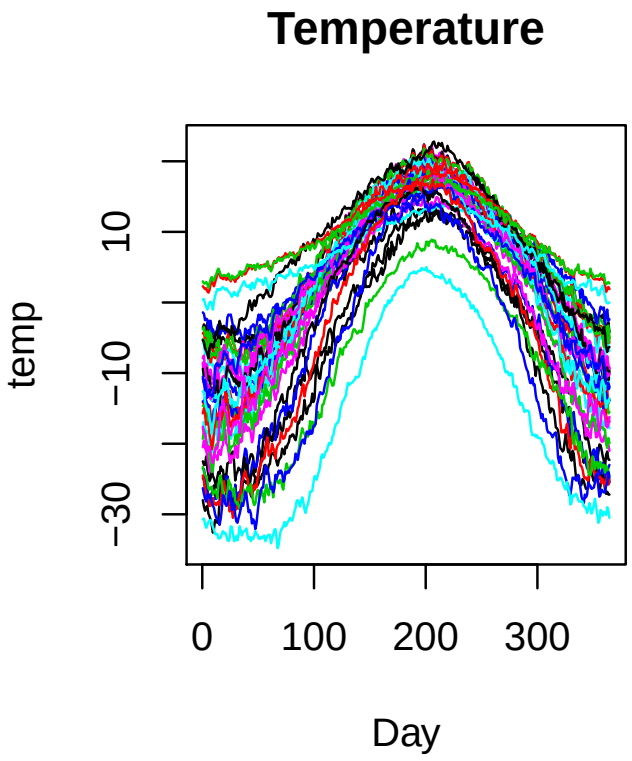
The derivatives of functional data are new functional data, that can be as informative as the original functions (or even more!).

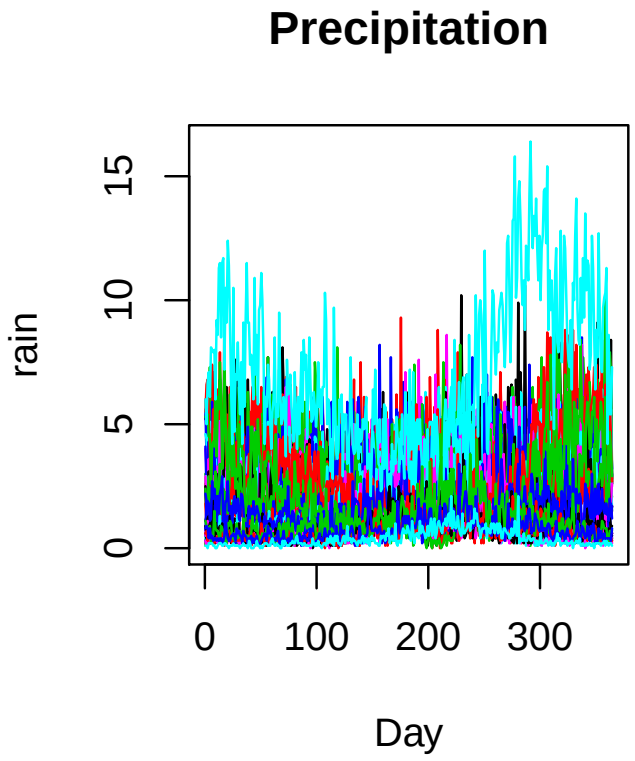


Height velocities and accelerations of 10 girls in the Berkeley Growth Study.

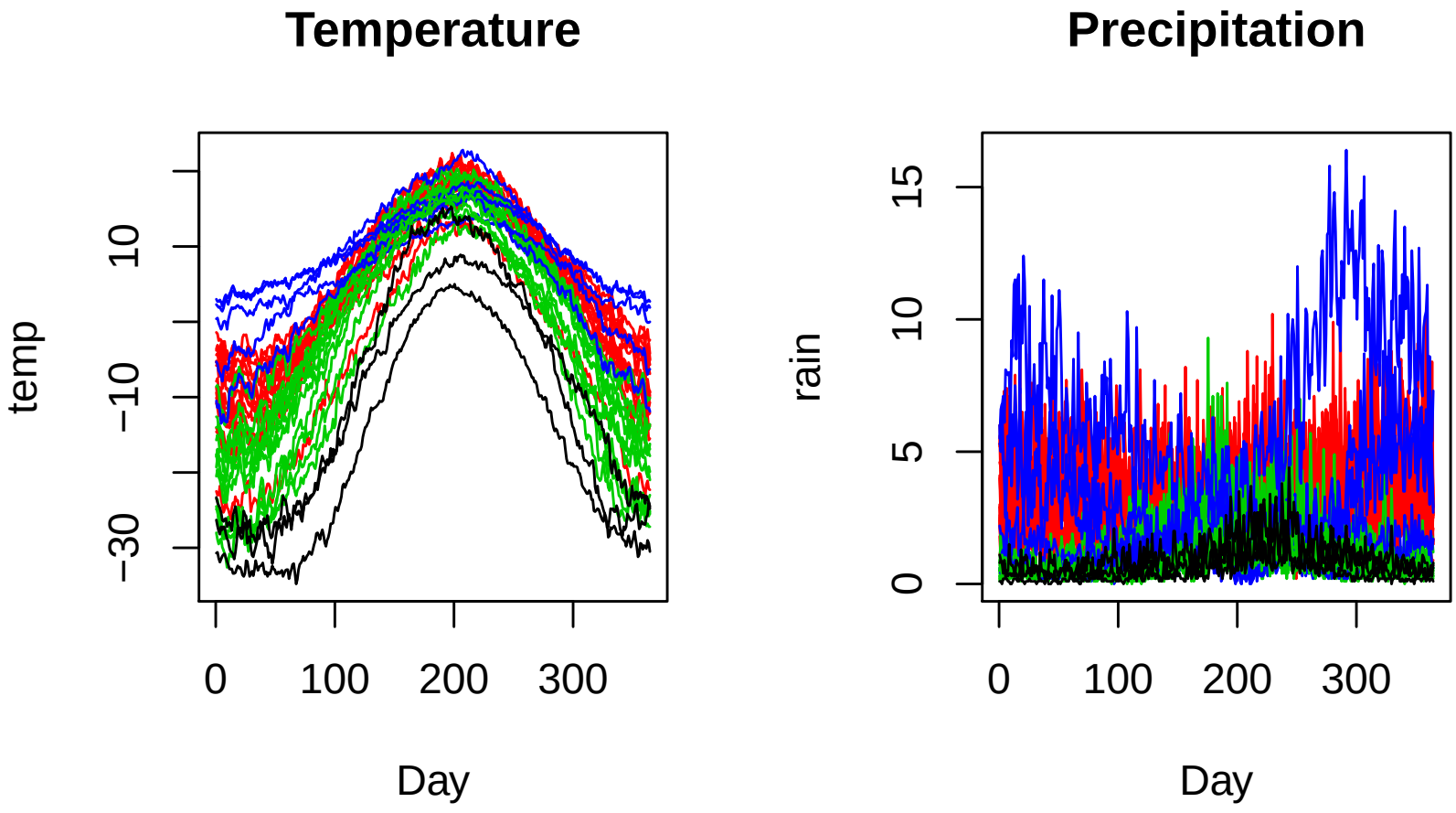
Canadian Weather Data

This is a classical dataset, that consists of average (from 1960 to 1994) of daily temperature and precipitation records in 35 weather stations across Canada.



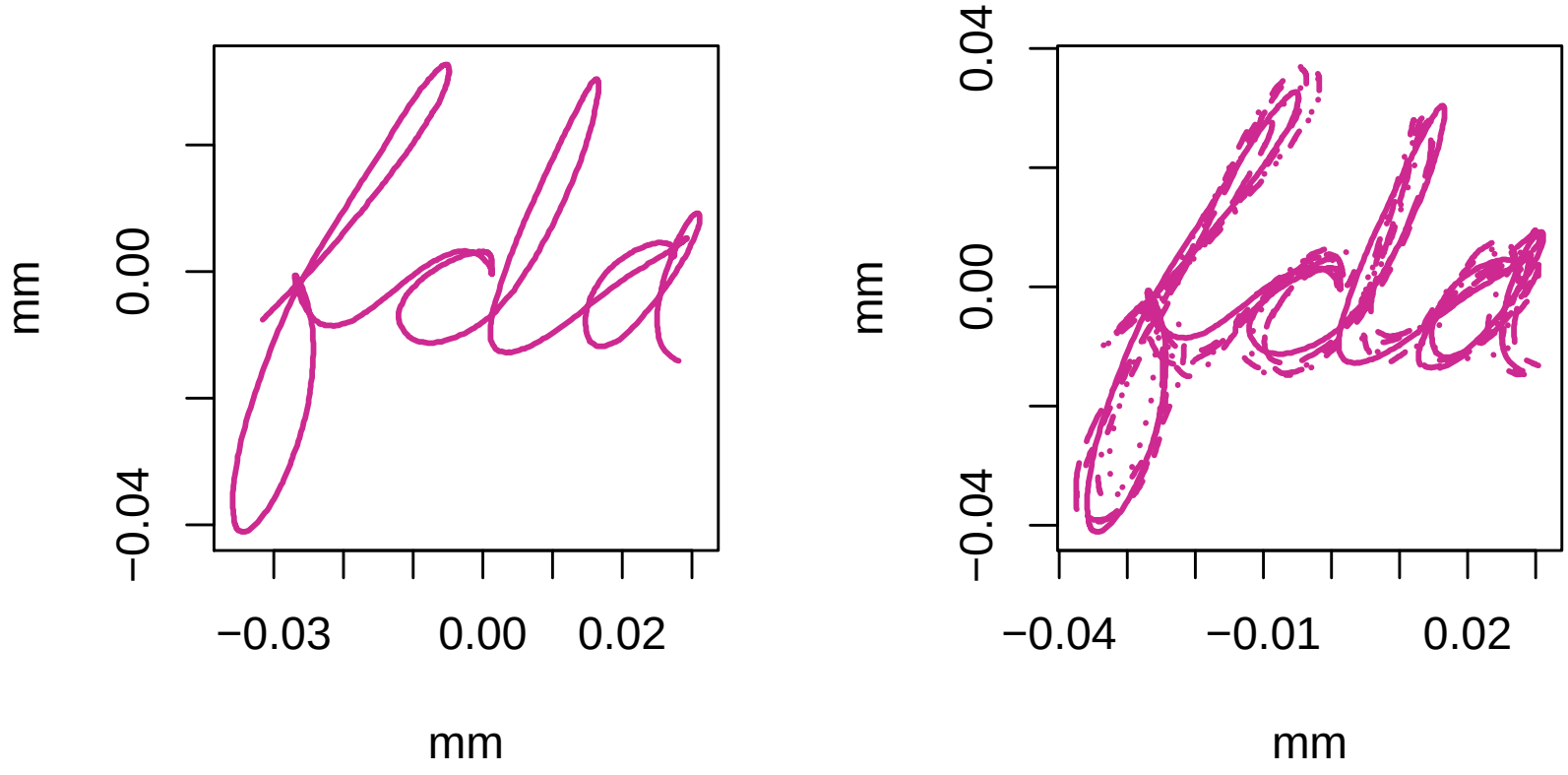


Different colors correspond to different geographical regions : Atlantic, Pacific, Central and North.



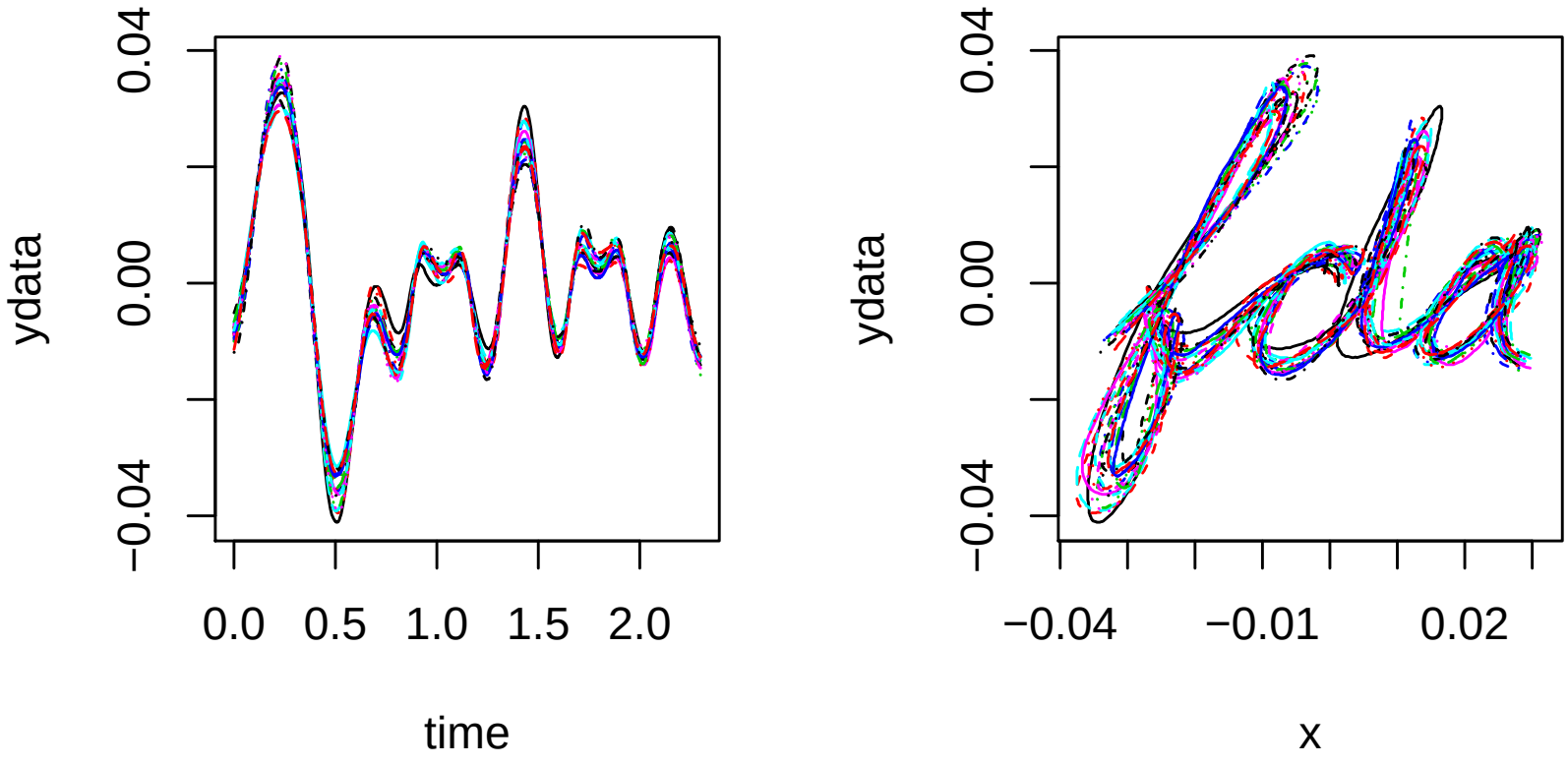
Interest is in variation and relationships between smooth underlying processes.

Multivariate functional data often arise from tracking movements of points through space.



This figure corresponds to the $X - Y$ coordinates of 1 and 5 samples superimposed. The role of time is lost in this plots.

We can include the information of the time taken by Jim Ramsay to actually do the writting.



Each replication is represented by 1401 coordinate values. The scripts have been extensively pre-processed. They have been adjusted to a common length that corresponds to 2.3 seconds or 2300 milliseconds, and they have already been registered so that important features in each script are aligned.

Some references

Functional data Analysis deals with the statistical description and modelling of random functions.

A broad outlook of FDA can be found in the books by

- Ramsay and Silverman (2002, 2005)
- Ramsay, Hooker and Graves (2009)
- Ferraty and Vieu (2006)
- Horvàth and Kokoszka (2012)
- Kokoszka and Reimherr (2017)
- `R` packages `fda` (Ramsay & Silverman) and `fda.usc` (Febrero Bande & de la Fuente)
- `R task view` `https://cran.r-project.org/web/views/FunctionalData.html`

Functional data

We call functional data to a collection of observations of a random process that is registered as a function $X_i : I \subset \mathbb{R} \to \mathbb{R}^n$.

In practice, the concept of a continuum measurement doesn't exist. The process has to be monitored in a discrete grid $t_1, t_2, ..., t_L$.

So at the end, one always has a vectorial observation $(x(t_1), ..., x(t_L))$. There are several reasons to consider this as functional data:

1. the continuous nature of the phenomenon at hand,
2. the theoretical possibility of observing the phenomenon in a much finer grid,
3. the choice of a functional model to approximately represent it.

The key assumption is smoothness:

$$y_{i,j} = x_i(t_{i,j}) + \epsilon_{i,j}$$

with t in a continuum (usually time) and $x_i(t)$ functions belonging to some "decent" functional space.

• In order to make inference on this type of data, we shall study later the space $L^2(T, \lambda)$ of square integrable functions defined on a set T with respect to some measure λ (usually Lebesgue measure).

• But before that, we shall discuss several aspects of descriptive statistics for functional data.

What are we interested in?

- Characteristics of data sets
  - mean
  - variation, covariation
  - a concept of "order" $\Rightarrow$ "depth"
- Relationships of functional data to
  - covariate
  - responses
- Relationships between derivatives of functions.
- Timing of events in functions.

What are the challenges?

- Estimation of functional data from noisy, discrete observations,
- Numerical representation of infinite-dimensional objects
- Representation of variation in infinite dimensions
- Description of statistical relationships between infinite dimensional objects.
- $n < p = \infty$, and use of smoothness.
- Measures of variation in estimates.

Statistical tools for FDA (so far...)

- Regression models: lm, glm, non-parametric regression, ANOVA...
- Multivariate Analysis: PCA, MDS, Clustering, Depth measures, ...
- Time Series, Spatial Statistics, ...
- Things that exploit the functional nature of the data
  - Principal differential analysis: a kind of PCA on the derivatives of the observed functions
  - Registration: a pre-process step that consists of a change of variables each observed function ir order to made them as similar as possible.

From Discrete to Functional Data

We need to represent data recorded at discrete times as a continous functions in order to

- Allow evaluation of records at any time point (specially if observation times are not the same across records).
- Evaluate rates of change (derivatives!)
- Reduce noise.
- Allow registration onto a common time-scale.

From Discrete to Functional Data: representing non-parametric continuous time functions

Basis-expansion methods

$$x(t) = \sum_{i=1}^{K} \phi_i(t) c_i$$

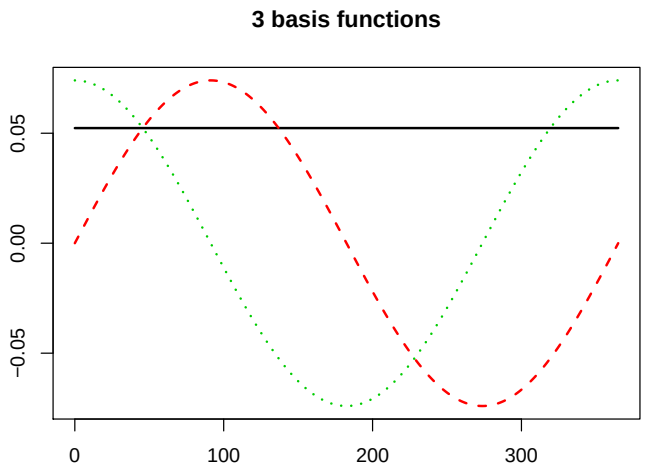
for pre-defined ϕ_i and coefficients c_i. Several basis systems available:

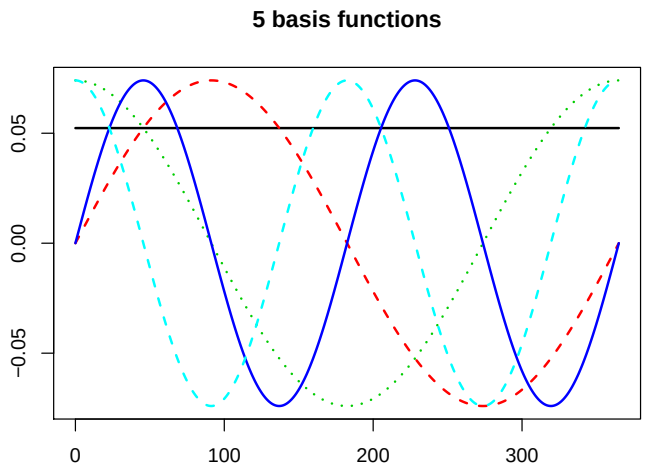
- Fourier basis: ideal for periodic data, such as weather cycles, signal processing...
- B-splines
- wavelets, Monomials $(1, x, x^2, \dots)$, Power (t^{λ_i}, λ_i not necessarily integer), Exponential ($\exp(\lambda_i t)$).

Fourier Basis

$$1, \sin(\omega t), \cos(\omega t), \sin(2\omega t), \cos(2\omega t), \ldots \sin(m\omega t), \cos(m\omega t), \ldots$$

$\omega = 2\pi/P$ defines the period of oscillation of the 1st sine-cosine pair.



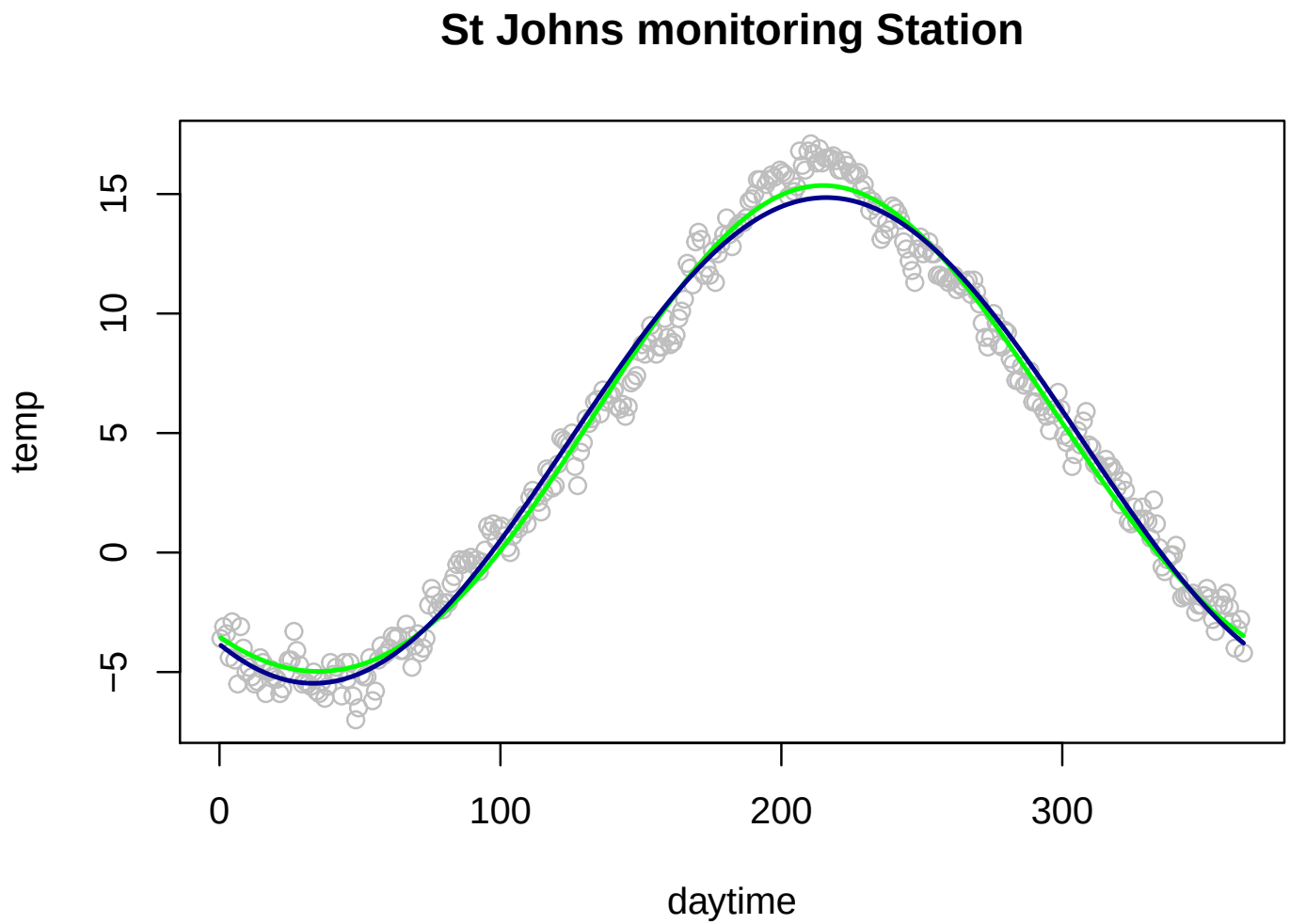


Natural for describing periodic data, such as annual wether cycles, but they always yield periodic functions, hence, it can be a problem in certain contexts.

Example: Canadian Weather Data, St John's Station

Linear regression on temperature with the first 3 (blue) and 5 (green) basis functions.

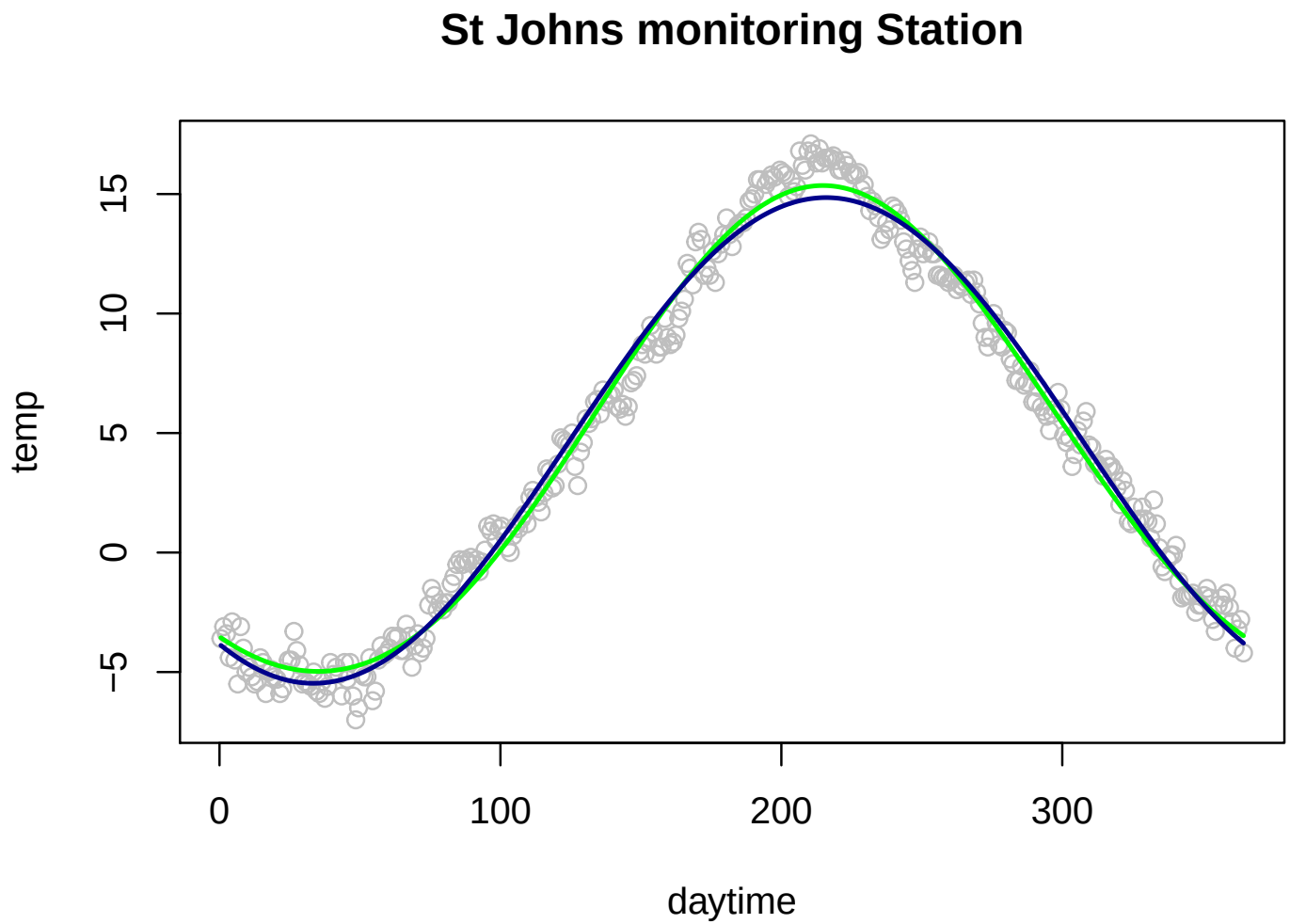
The basis functions are evaluated at the observation times



Example: Canadian Weather Data, St John's Station

Linear regression on temperature with the first 3 (blue) and 5 (green) basis functions.

The basis functions are evaluated at the observation times



Spline Basis (B-splines)

- Splines are polynomial segments joined end-to-end.
- Segments are constrained to be smooth at the joins.
- The points at which segments join are called *knots*.
- Systems are defined by
  - The order m (order=degree+1) of the polynomial.
  - The location of the knots.
- Number of basis functions:

$$\textit{order} + \textit{number of interior knots}$$

- Sum of all B-splines in a basis is always 1; can fit any polynomial of order m.

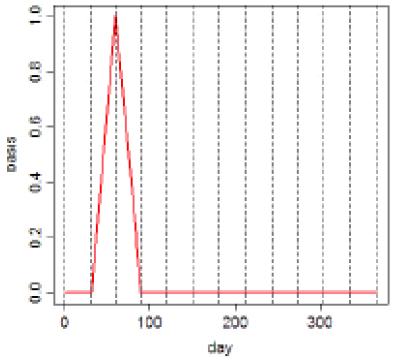
Properties of B-Splines

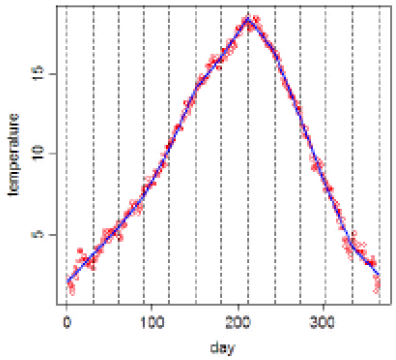
- Order m splines: derivatives up to $m-2$ are continuous.
- Most popular choice is order 4, implying continuous second derivatives. Second derivatives have straight-line segments.
- B-spline basis functions are positive over at most m adjacent intervals $\Rightarrow$ fast computation for even thousands of basis functions.
- Support on m adjacent intervals: highly sparce design matrix.
- Flexibility comes from knots & derivatives from order.
- Frequently, fewer knots do reasonably well (approximation properties can be formalized).

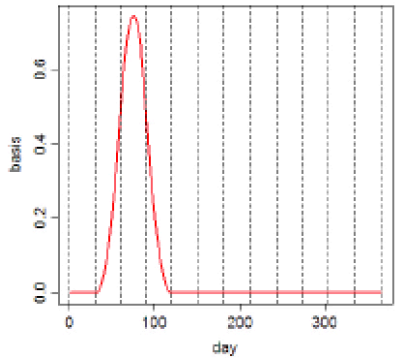
See de Boor, 2001, *A practical guide to splines*.

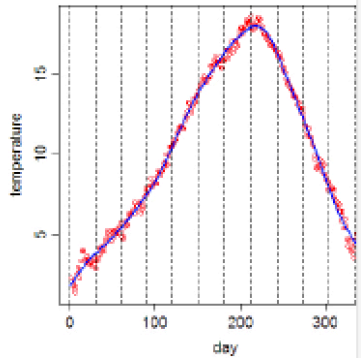
Example: Canadian Weather Data, Vancouver

Splines of order 1 and 3, knots at months.

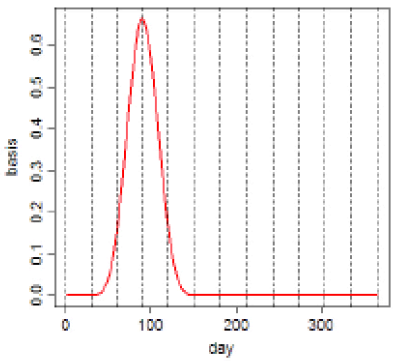


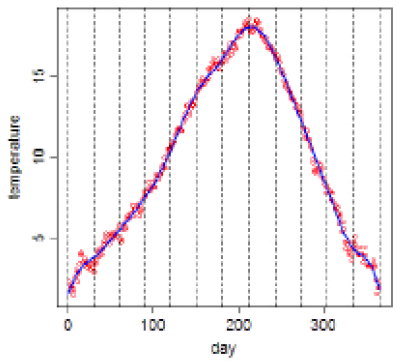


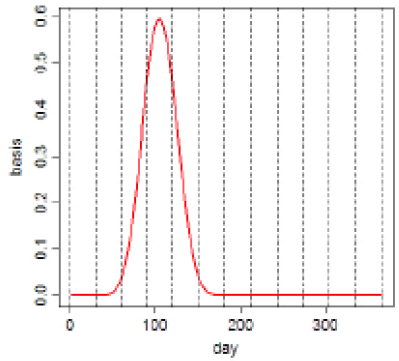


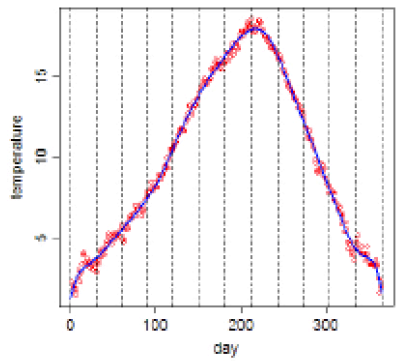


Splines of order 4 and 5, knots at months.









B-splines: Choosing Knots and Order

- The order of the spline should be at least $k + 2$ if you are interested in k derivatives..
- Knots are often equally spaced (a useful default).
- But there are two important rules:
  - Place more knots where you know there is strong curvature, and fewer where the function changes slowly.
  - Be sure there is at least one data point in every interval.
- Sum of all B-splines in a basis is always 1; can fit any polynomial of order m.

Smoothing Penalties

Assume we have observations for a single curve

$$y_i := x(t_i) + \varepsilon \quad i = 1, \ldots, N$$

and we want to estimate

$$x(t) \approx \sum_{j=1}^{K} c_j \phi_j(t).$$

Minimize the sum of squared errors:

$$SSE = \sum_{i=1}^{N} (y_i - x(t_i))^2 = \sum_{i=1}^{n} (y_i - c^T \Phi(t_i))^2$$

just as in linear regression.

If the $N \times K$ matrix $\mathbf{\Phi}$ contains the values $\phi_k(t_j)$ and $\boldsymbol{y} = (y_i, \ldots, y_n)$

$$SSE(c) = (\boldsymbol{y} - \mathbf{\Phi}c)^T((\boldsymbol{y} - \mathbf{\Phi}c)$$

The error sum of squares is minimized by the OLS estimate

$$\hat{c} = (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\boldsymbol{y}$$

and the estimate for $x(t)$ is

$$\hat{x}(t) = \mathbf{\Phi}(t)(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\boldsymbol{y}.$$

- Least squares is optimal if residuals are iid N(0,σ^2).
- If the variance is non-constant, weighted least squares can help.

$$WMSE(x) = \sum w_i (y_i - x(t_i))^2$$

$$\text{If } \boldsymbol{W} = \mathrm{diag}(w_i), \quad \hat{x}(t) = \boldsymbol{\Phi}(t)\hat{c} = \Phi(t)(\boldsymbol{\Phi}^T \boldsymbol{W} \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^T \boldsymbol{W} \boldsymbol{y}$$

and at the observation points

$$\hat{y} = \boldsymbol{\Phi}(t)(\boldsymbol{\Phi}^T \boldsymbol{W} \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^T \boldsymbol{W} \boldsymbol{y} = \boldsymbol{S} \boldsymbol{y}$$

$\boldsymbol{S}$ is called the smoothing matrix. So all we need to do is choosing the weights appropriately. This amounts to choosing the type of basis and the number of basis functions.
Trade off:

- Too many basis functions over-fit the data and reflect errors of measurement: small bias, large variance.
- Too few basis functions fail to capture interesting features of the curves: small variance, large bias.
- As often, cross-validation is a good idea.

An alternative: Kernel methods

One‘bump” function $K_i(t)$ at each data point calculate a weighted average

$$x(t) = \frac{\sum y_i K_i(t/h)}{\sum K_i(t/h)}$$

and use the band-with h to regulate the smoothness. Hence, we need a fair definition of “smooth”: what we really want is to eliminate “wiggles” in the data. Define the differential operator.

$$Dx(t) = \frac{d}{dt}x(t)$$

Higher order derivatives are defined in terms of powers:

$$D^2x(t) = \frac{d^2}{dt^2}x(t), \ldots, D^kx(t) = \frac{d^k}{dt^k}x(t)$$

$Dx(t)$ is the slope of $x(t)$, $D^2x(t)$ is its curvature. And we measure the size of the curvature by

$$J_2 = \int (D^2x(t))^2 dt$$

The smoothing Spline Theorem

Consider the penalized squared error[1] $PENSSE_\lambda x(t)$ defined by

$$\sum (y_i - x(t_i))^2 + \lambda \int (D^2 x(t))^2 d = (\boldsymbol{y} - x(\boldsymbol{t}))^T (\boldsymbol{y} - x(\boldsymbol{t})) + \lambda J_2$$

- λ is a smoothing parameter measuring compromise between fit and smoothness.
- As λ increases, roughness is increasingly penalized and $x(t)$ will become linear.
- As λ decreases, the penalty is reduced and allows $x(t)$ to fit the data better.
- A remarkable theorem tells us that the function $x(t)$ that minimizes $PENSSE_\lambda(x)$ is a spline function of order 4 (piecewise cubic) with a knot at each sample point t_j

[1] see Chapter 5 of Ramsay, Hooker, Graves(2009), Functional Data Analysis with R and Matlab, which is a sort of manual for package `fda`

Calculating the penalized fit

- The theorem tells us that $x(t)$ is of the form $x(t) = \mathbf{\Phi}(t)^T c$ where $\phi(t)$ is a vector of B-spline basis functions.
- The number of basis functions is $(n-2)+4 = n+2$ where n is the number of sampling points.

Using that $x(t) = \mathbf{\Phi}(t)c$, we have:

$$\int (D^2 x(t))^2 d = \int (c^T (D^2 \phi(t))(D^2 bm\phi(t))^T c dt = c^T R_2 C$$

where $[R_2]_{i,j} = \int D^2\phi_i(t) D^2\phi_j(t) dt$ is the penalty matrix (2 can be changed by m). The penalized least squares estimate of c is

$$\hat{c} = \left[\mathbf{\Phi}^T \mathbf{\Phi} + \lambda R_2\right] \mathbf{\Phi}^T \boldsymbol{y}$$

and the linear smoother is

$$\hat{y} = \phi \left[\mathbf{\Phi}^T \mathbf{\Phi} + \lambda R_2\right] \mathbf{\Phi}^T \boldsymbol{y} = \boldsymbol{S}(\lambda)$$

Again, (log)λ is chosen with cross-validation.

Descriptive Statistics

Estimation of mean and covariance functions

Consider square integrable random curves

$$X = \{X(t), t \in [0, T]\}$$

and let X_1,X_2,..., X_m be an i.i.d. sample drawn from X.

Mean function $\mu = E[X(t)]$ estimated by

$$\hat{\mu}(t) = \frac{1}{m} \sum_{i=1}^{m} X_i(t)$$

Covariance function $c(t, s) = E[(X(t) - \mu(t))(X(s) - \mu(s))]$ estimated by

$$\hat{c}(t, s) = \frac{1}{m} \sum_{i=1}^{m} (X_i(t) - \hat{\mu}(t)) (X_i(s) - \hat{\mu}(s))$$

The Covariance operator

Associated with $c(t,s)$ is the covariance operator $C : L^2(T) \to L^2(T)$

$$f \mapsto C(f) : T \to \mathbb{R}$$

$$t \mapsto C(f)(t) = \int_T C(t,s)f(s)ds$$

estimated by

$$\hat{C}(x) = \frac{1}{m}\sum_{i=1}^{m} \langle X_i - \hat{\mu}, x\rangle (X_i - \hat{\mu})$$

Let λ_j and v_j, $j \geq 1$ be the eigenvalues and eigenfunctions of the covariance operator C:

$$C(v_j) = \lambda_j v_j \quad j \geq 1.$$

Karhunen-Loève expansion

Assuming that $\mathbf{E}\int_T (X(t)-\mu(t))^2 dt = \mathbf{E}(||X-\mu^2||) < \infty$ the random element $X(t)$ can be expressed as

$$X(t) = \mu(t) + \sum_{j=1}^{\infty} Z_j v_j(t) \quad \text{where} \quad Z_j = \langle X-\mu, v_j \rangle \quad j \geq 1$$

This is known as the Karhunen-Loève expansion of X, and $\{Z_j, j \geq 1\}$ is a sequence of centred random variables with $\mathbf{E}(Z_j^2) = \lambda_j$, and are uncorrelated, that is $\mathbf{E}(Z_j Z_h) = 0$ for $h \neq j$.

Hence, when X is Gaussian, the Z_j are independent Gaussian random variables.

Moreover

$$\mathbf{E}\left(\int_T (X(t)-\mu(t))^2 dt\right) = \mathbf{E}||X-\mu||^2 = \sum_{j=1}^{\infty} \lambda_j$$

Functional principal component analysis (FPCA)

Let $\{X_n(t)\}$ be the sample of curves. We can understand FPC as the coordinates maximizing variability: p uncorrelated random variables $Y_i = \langle X(t), v_i(t) \rangle$ for $i = 1 \dots p$ such that

$$Var\left(\sum_{i=1}^{p} \langle X(t), v_i(t) \rangle\right)$$

is maximized subject to the constraint that $|v_i| = 1$.
It is possible to identify the principal components as the eigenfunctions with higher eigenvalues of the covariance operator C and build a data decomposition through the eigenvalues. We can also see that they are optimal orthonormal basis.

Parenthesis: Principal components in $\mathbb{R}^p$

A principal component analysis (or PCA) is a way of simplifying a complex multivariate dataset. It helps to expose the underlying sources of variation in the data.

- Uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables (PC)
- This transformation is defined in such a way that the first principal component has the largest possible variance and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components. The resulting vectors are an uncorrelated orthogonal basis set.
- PCA is sensitive to the relative scaling of the original variables.

Examples in $\mathbb{R}^p$

`http://setosa.io/ev/principal-component-analysis/`

PCA con be easily performed in `R` with the command `prcomp`.

See the accompanying file `pc.R`

Karhunen-Loève expansion and FPCA

• Assuming $\lambda_1 > \lambda_2 > \cdots > \lambda_j > \lambda_{j+1} > \ldots$ and fixing an integer p, the orthonormal basis $v_1, \ldots, v_p$ solving the minimization problem

$$\min_{v_1,\ldots,v_p} \mathbf{E}\left(||X - \mu - \sum_{j=1}^{p} \langle X - \mu, v_j \rangle ||^2 \right)$$

is formed by the first p eigenfunctions of the covariance operator, $v_1, \ldots, v_p$, and the minumum is $\sum_{j \geq p+1} \lambda_j$.

• That is, the Karhunen-Loève expansion of $X - \mu$ truncated at p gives the best p-dimensional approximation to $X - \mu$ in the sense of mean square error.

$$X(t) \approx \mu(t) + \sum_{=1}^{p} Z_j v_j(t) \qquad t \in T$$

• The eigenfunctions v_j are also known as principal functions

• The $Z_j = \langle X - \mu, v_j \rangle$ are known as Scores or Functional Principal Components.

Sampling version of FPCA

The empirical version of the previous result is obtained by replacing the covariance operator C by the estimated covariance function

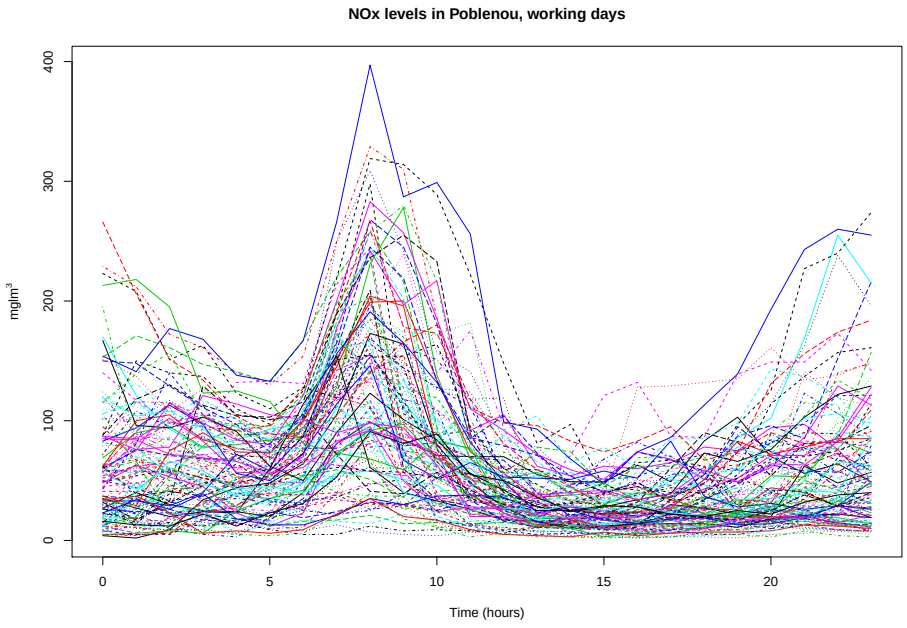
$$\hat{c}(t,s) = \frac{1}{n}\sum_{i=1}^{n}\left(X_i(t) - \bar{X}(t)\right)\left(X_i(s) - \bar{X}(s)\right)$$

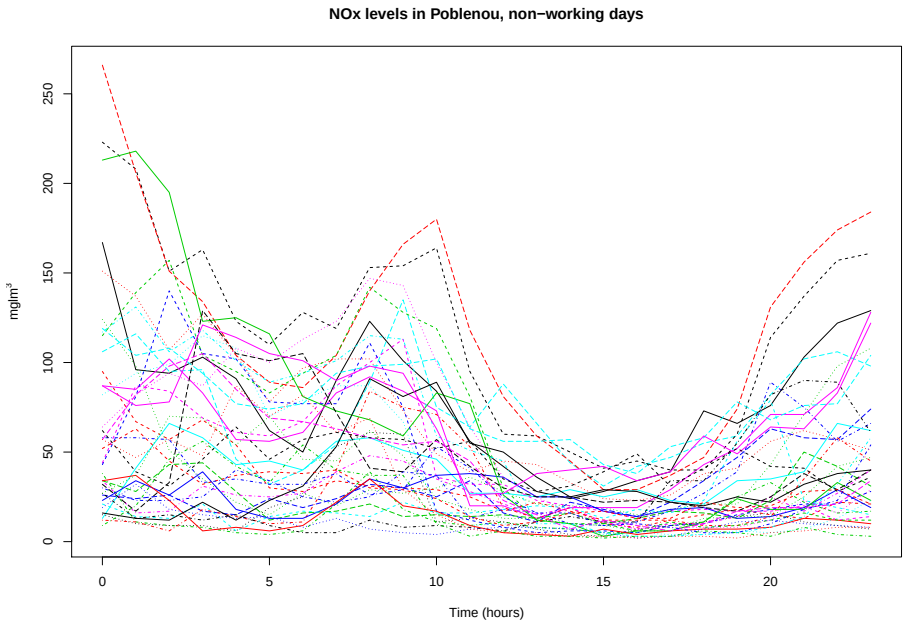
In practice, the eigensystem of the operator with kernel $\hat{C}(t,s)$ is obtained by matrix diagonalization.

$$X_i(t) = \bar{X}(t) + \sum_{j=1}^{p}\hat{Z}_{i,j}\hat{v}_j(t) \quad t \in T, i = 1,\ldots,n$$

FPCA con be easily performed in `R` within the libraries `fda` and `fda.usc`.

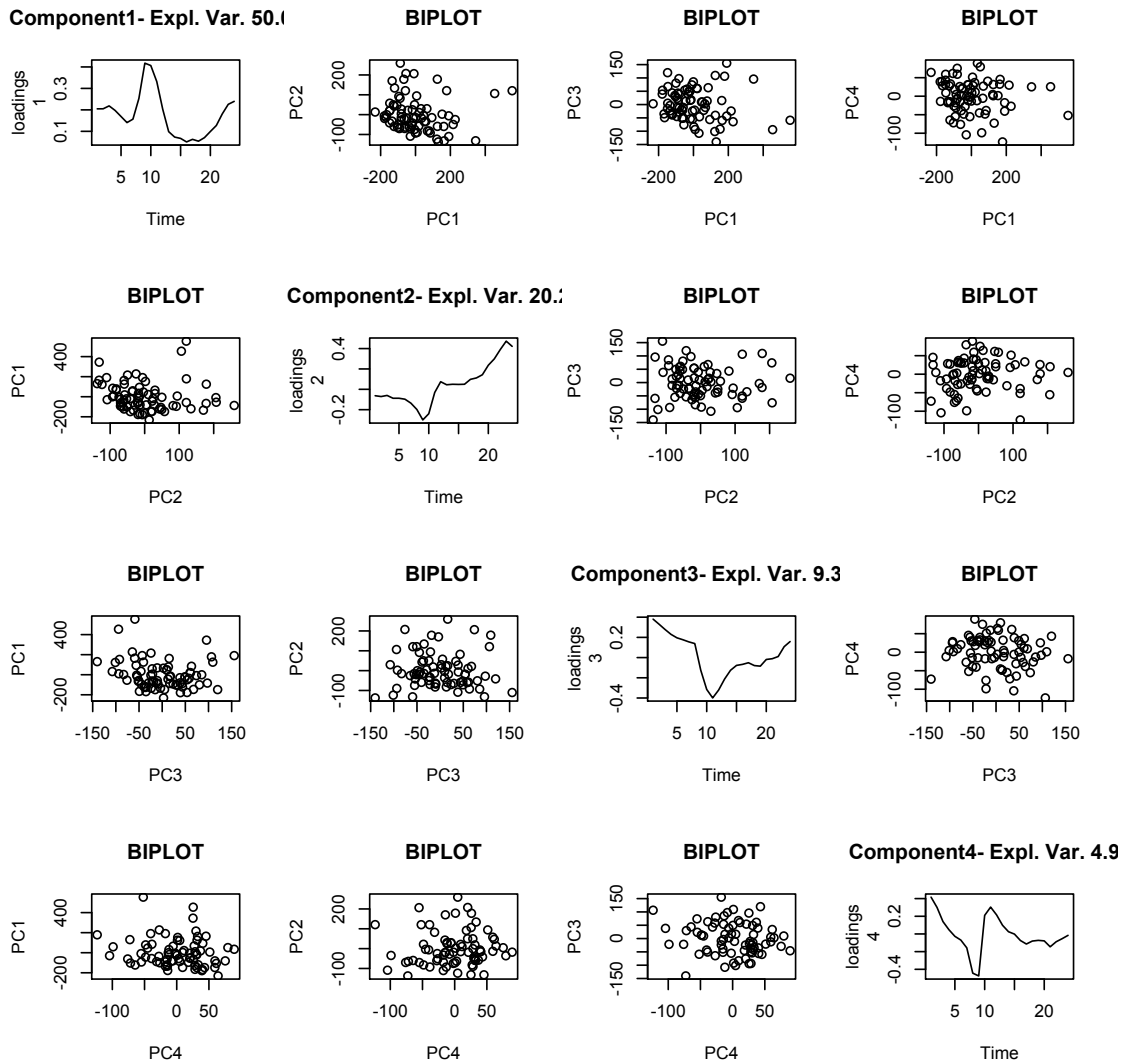
Example: NOx levels in Poblenou





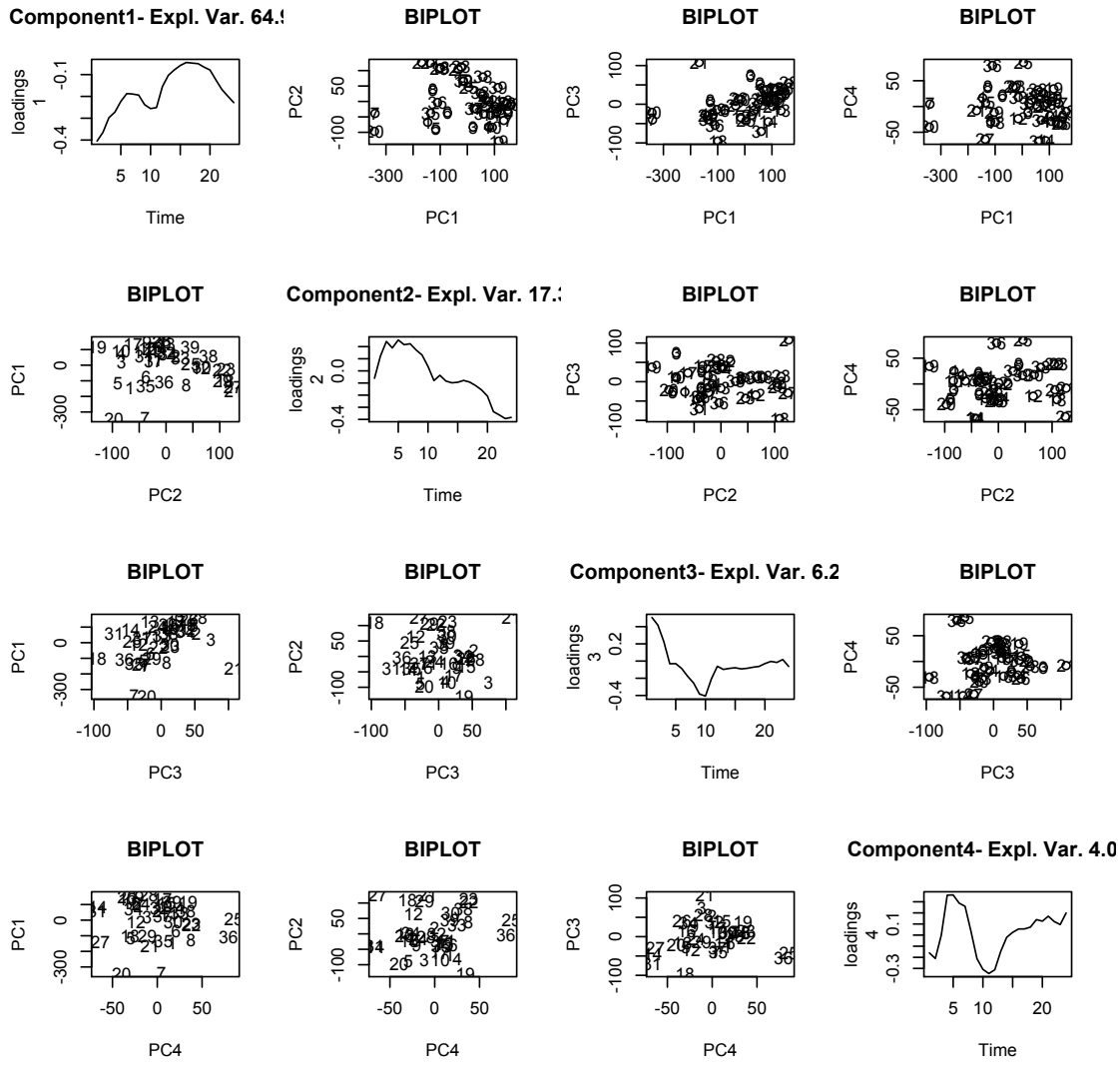
With 4 components 84.53 % of the variability in working-days is explained.

PC1	PC2	PC3	PC4
50.01	20.21	9.35	4.96



With 4 components 92.59 % of the variability of non-working days is explained.

PC1	PC2	PC3	PC4
64.92	17.32	6.28	4.08



Ordering:
a first step towards inference

Depth Measures for Functional Data

Another way of approaching the functional data analysis is the introduction of the notion of depth, that is related to a generalisation of the concept of ordering for functional data.

The idea is to assign an order to each element of the sample related to its centrality for the whole set.

There are many measures of depth, and there is no agreement about their advantages.[2]

We will concentrate on the Fraiman-Muñiz depth, but before that, let's review other notions of depth.

[2] a good review is done in Estrada's Thesis, ULA (2014)

The Fraiman-Muñiz Depth[3].

Consider first a univariate sample

$$U_1, ..., U_n \qquad \text{and} \qquad U_{(1)}, ..., U_{(n)}$$

be the corresponding order statistics. For no ties, we have that $U_i = U_{(j)}$ and then the natural depth of U_i is said to be

$$D_n(U_i) = \frac{1}{2} - \left| \frac{1}{2} - \left(\frac{j}{n} - \frac{1}{2n} \right) \right|$$

This notion of depth assigns minimal and equal depth to the two extreme values of the sample, maximum to the innermost point(s) and changes linearly with the position the datum occupies in the sample.
Observe that $2nD_n$ produces the ordering

$$1 \quad 3 \quad 5 \quad 7 \quad \ldots \quad 7 \quad 5 \quad 3 \quad 1$$

[3] Fraiman& Muñiz (2001) Trimmed means for functional data, *TEST* 10.

For the case of functional data, consider the sample $X=\{X_i(t)\}$ defined in a common interval $\mathcal{J}$.
For each t we can compute the natural depth, $D_n(X_i(t))$, and then the depth for each $X_i(t)$ is:

$$I(X_i, X) = \int_{\mathcal{J}} D_n(X_i(t))dt$$

Where, in practice, the integral is replaced by a sum over the t for the time grid.
The deeper curve will be the X_i that with the biggest $I(X_i, X)$.

Defining a Wilcoxon test based on this ordering is a natural option for the 2-sample problem[4]

[4] Lopez-Pintado & Romo (2007), Depth-based inference for functional data, *CSDA* 51 suggest the use of this procedure with their band-depth.

Some other depth Measures for Functional Data

- Band depth[5]

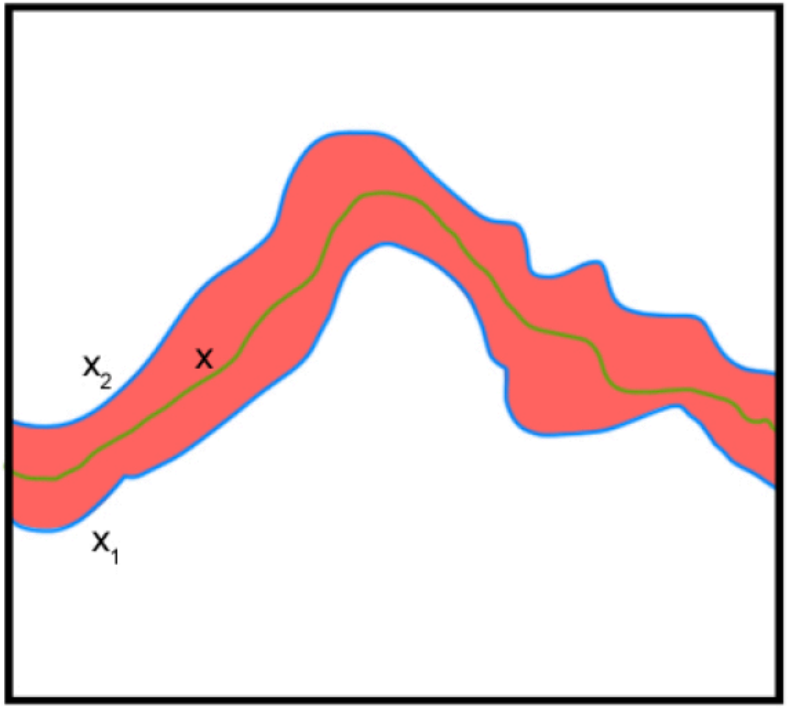
  1. The band determined by the curves $X_{i_1}, \ldots, X_{i_k}$ is

  $$B_{i_1,\ldots,i_k} = \{(t, y), \min_{1 \le r \le k} X_{i_r} \le y \le \max_{1 \le r \le k} X_{i_r}\}$$

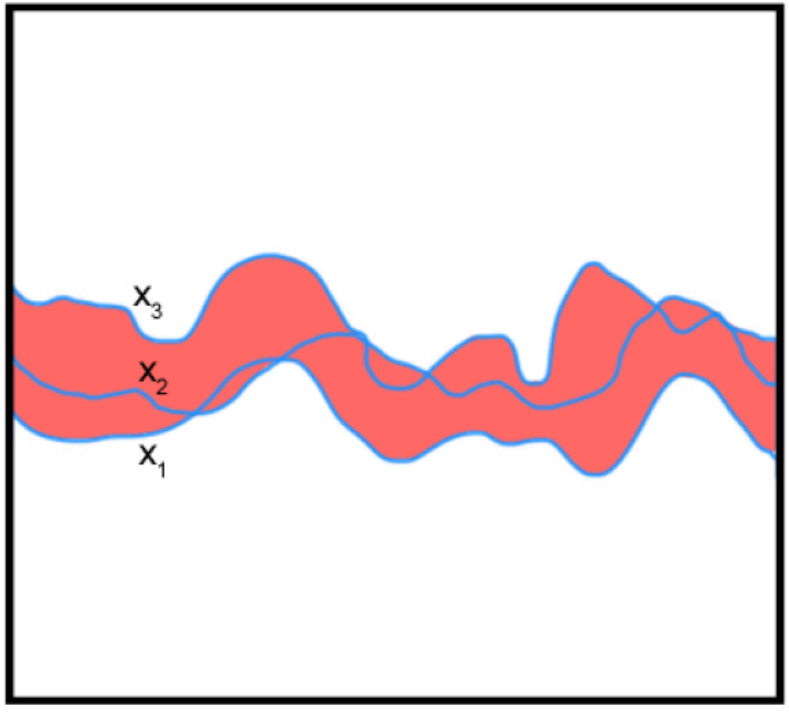
  Define S_i^j as the proportion of the bands formed by j curves that contain the graph of $X_i(t)$, and the depth of X_i as the sum $\sum_{j=1}^{J} S_i^j$ for a fixed J.

  2. Generalize GS_i^j considering the amount of time (Lebesgue measure of the set) that the curve X_i is inside the band, and again the sum of these will be the depth.

[5] Lopez-Pintado & Romo (2009), On the concept of depth for functional data *JASA* 104



$x \in$ band defined by x_1 and x_2



Band defined by x_1, x_2, x_3

Some other depth Measures for Functional Data

- h-modes[6]

For each curve X_i, its h-mode is defined as

$$\hat{f}_h(X_i) = \frac{1}{n}\sum_{j=1}^{n} K_h(||X_j - X_i||)$$

where $K_h(t) = 1/hK(t/h)$ and K is the Gaussian kernel.

The deepest curve will be that for which $\hat{f}_h$ is maximum. The choice of h is arbitrary...

[6] Cuevas, Febrero, Fraiman (2007), Robust estimation and classification for functional data via projection-based depth notions, *Computational Statistics* 22

Some other depth Measures for Functional Data

Projection over principal components[7]

Project X_i on the first h principal components, and take a weighted average of the univariate depth of the projection points.

Random Projections[8]

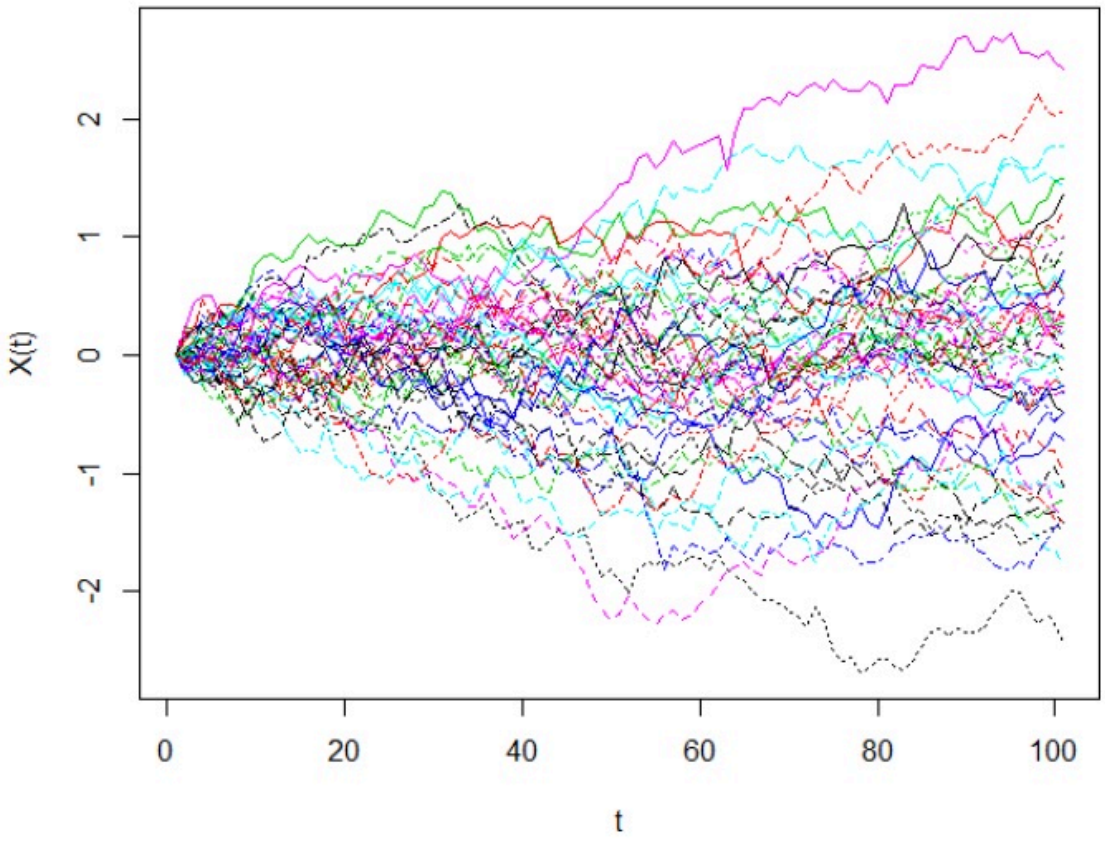
Given a random direction a, project the curve $_i$ on a (this requires some regularity condition to ensure that the inner product is well defined), and cosider the univariate depth of the projected points.
The depth of the curve is then defined as the average of the univariate depths of several random projections.

---

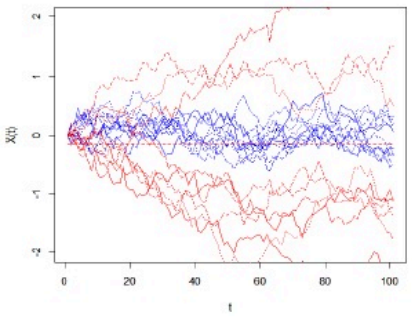
[7] Estrada, Quiroz (2014)

[8] Cuevas, Febrero, Fraiman (2007), *op.cit.*

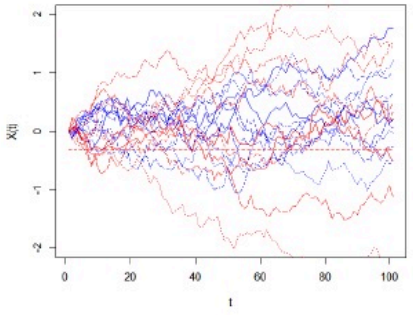
Example: 50 curves from Brownian Motion



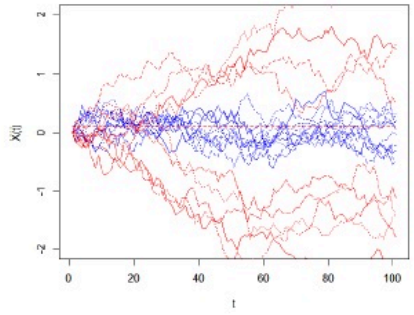
Deepest and shallowest curves



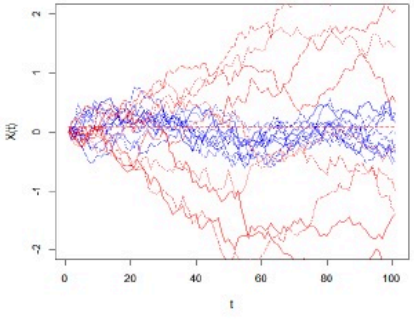
(a) Fraiman y Muñiz



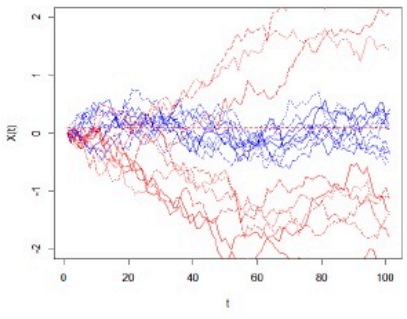
(b) Prof. por bandas (V1)



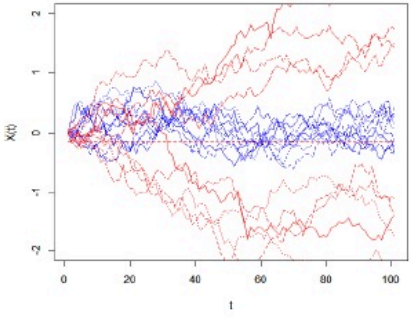
(c) Prof. por bandas (V2)



(d) h-moda



(e) Proyección aleatoria

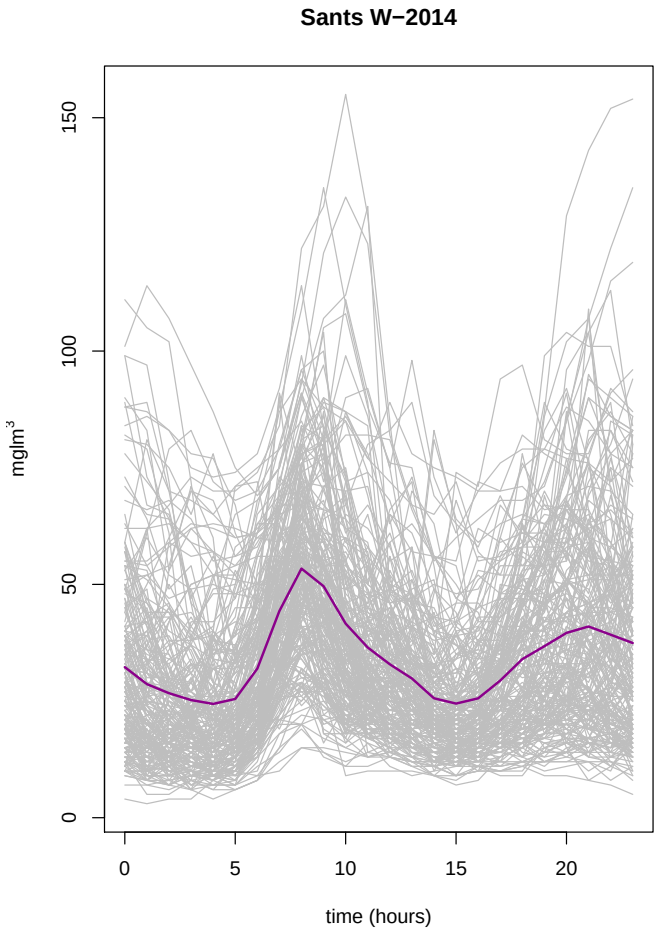


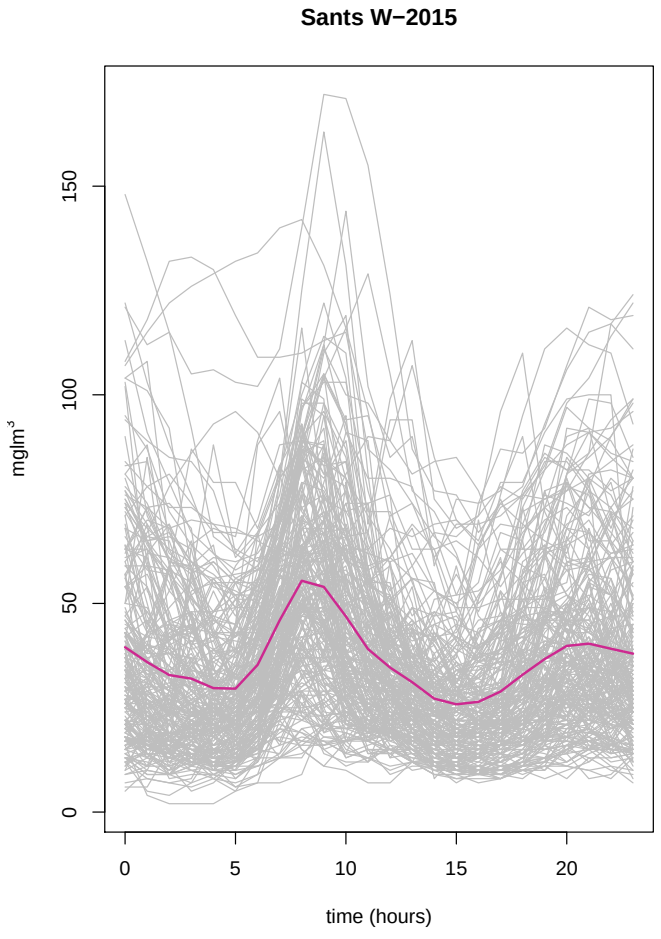
(f) Proyección PCA

Bands: $J = 2$, Modes: $h = 0.1$, 2 principal components (83.74% & 7.75 % resp.), and 50 random projections.

An example of Inference: The two-sample problem

NO_2 levels on working days taken on the same station, in 2014 and 2015. Did (the law) of contamination- levels change?





On Permutation tests

Permutation tests are one of the most important tools of non-parametric statistical methodology.

A permutation test gives a simple way to compute the sampling distribution for any test statistic, when the distribution is invariant under permutations for the null hypothesis.

To estimate the sampling distribution of the test statistic we randomly shuffle the exposures to make up a large enough sample of data sets.

If the null hypothesis is true, this shuffled data sets should have all the same distribution, and thus should look like the original data, otherwise they should look different.

Obtaining p-values

Let $N = m + n$ and $Z_1, \ldots, Z_N$ be the combined sample obtained by concatenating the X and Y samples. Relevant p-values for the statistics we shall consider here can be obtained as follows:

1. *From the joint sample, Z, select a random subset of size m. Declare that the chosen elements belong to the X sample and the remaining n to the Y sample.*
2. *Compute the statistic of interest for this artificial pair of samples.*
3. *Repeat steps (1) and (2) a large number B of times, and*
4. *From the B values computed above, extract an approximate p-value for the observed statistic (the one calculated with the original X and Y samples)*

Two sample problem for functional data

Let $X_1(t), \cdots, X_m(t)$ denote an i.i.d. sample of real valued curves defined on some interval J.

Let $\mathcal{L}(X)$ be the common probability law of these curves.

Likewise, let $Y_1(t), \cdots, Y_n(t)$, be another i.i.d. sample of curves, independent of the X sample and also defined on J, with probability law $\mathcal{L}(Y)$.

We want to test the null hypothesis,

$$H_0 : \mathcal{L}(X) = \mathcal{L}(Y) \text{ against the general alternative } \mathcal{L}(X) \neq \mathcal{L}(Y)$$

We shall discuss different permutation tests[9]:

- A functional Shilling test
- A Wilcoxon type test
- Another test based on depths, that uses meta analysis ideas to assess significance
- This last test can be re-defined in terms of FDRs (... I think, anyone wants to give it a thought?).

and compare their performance with the classical test by Kokoszka and Horváth, based on the principal components of the pooled covariance operator of the two samples.

[9] A. Cabaña, *et.al*. Permutation tests in the two-sample problem for functional data, in *Functional Statistics and related Fields*, G. Aneiros, E. Bongiorno,R. Cao & P.Vieu, Ed. Springer, (2017)

Schilling test

This is a generalisation to functional data of the k -nearest neighbours multivariate two-sample test of Schilling[10].

Let $N = m + n$ and $Z_1, \ldots, Z_N$ be the combined sample obtained by concatenating the X and Y samples.
Define the indicator function

$$I_i(r) = \begin{cases} 1 \text{ if } Z_i \text{ and its } r\text{-nn are from the same sample} \\ 0 \text{ else} \end{cases}$$

Nearest neighbours are based on L^2 distance, and with probability 1 the are uniquely defined. In case of ties, we would decide at random.

[10] Schilling (1986) Multivariate two-sample tests based on nearest neighbors, *JASA* 81, 799-806

In practice, if the functions have been registered in a common grid, say $0 = t_0 < t_1 < \ldots t_L = T$, a reasonable approximation to the distance between functions Z_i and Z_j would be

$$d_{i,j} = \sum_{l=1}^{L} \Delta_l (Z_i(t_l) - Z_j(t_l))^2, \text{ where } \Delta_l = t_l - t_{l-1}$$

If the grid used is equally spaced, Δ_l can be omitted and the curves can be treated as points in $\mathbf{R}^L$ in order to compute faster the k-nearest-neighbours.

When no common grid is available, represent the functions in the joint sample in terms of local polynomials, or some other basis functions, and the k-nearest-neighbours are identified by a quadratic algorithm (in the joint sample size N).

Define the statistic:

$$T_{N,k} = \frac{1}{Nk} \sum_{i=1}^{N} \sum_{r=1}^{k} I_i(r)$$

Under H_0 we expect $T_{N,k}$ to be small.

Observe that the expected value

$$\mathbf{E}\, T_{N,k} = \mathbf{E}\, I_i(r) = \frac{m(m-1) + n(n-1)}{N(N-1)}$$

while the variance depends on the amount of pairs of points that are mutual neighbours and the amount of pairs that share a common neighbour.
In this case, the belonging of one element to either $\{X_m(t)\}$ or $\{Y_n(t)\}$ is an arbitrary choice, for they represent just two subgroups of a larger sample generated from a single random experiment. Thus, we can estimate the distribution of the statistic through permutations on the labelling.

Algorithm

1. For the concatenation of the samples Z, keeping the natural ordering, we compute the $m \times n$ matrix of distances between its elements.
2. Consider we are running the test for k neighbours. We build a $N \times k$ matrix that in the i-th row contains the indices of the k nearest elements to the curve Z_i.
3. In order to compute the statistic it is enough to count how many of the indices in each of the first m rows is equal or less than m, (i.e. are originally X) and how many of the indices corresponding to $m+1, \ldots, N$ is greater than m.
4. Obtain the distribution of $T_{N,k}$ using the permutation procedure.

In order to compute the distribution using the permutation procedure, we do the following:

- To the elements of the original nearest neighbour table, apply a random permutation τ.
- In the r-th row, if $\tau(r) \leq m$, count the number of elements that are less than or equal to m, else, count the number of elements greater than m.
- with the numbers obtained in (ii) compute $T_{N,k}$.

All these operations can be performed in time $O(Nk)$, thus the computational cost of an iteration is basically linear in N. As for the initial cost of setting the neighbours indices table, several sub-quadratic algorithms have been developed for this problem, since the fundamental contribution of Friedman *et.al.*[11].

[11] Friedman, Baskett & Shustek (1975) An algorithm for finding nearest neighbours, *IEEE Trans. Com.* 24

Back to the 2- sample problem: Meta-analysis based tests

Let p_i be an empirical p-value related to the *centrality* of the variable Y_i on the sample X,

$$\hat{p}_* = \frac{\sum_{j=1}^{m} I\left(D\left(X_j|X\right) \leq D\left(Y_*|X\right)\right)}{m+1} \qquad \begin{cases} p_i = \hat{p}_i \text{ if } \hat{p}_i \neq 0 \\ p_i = \frac{1}{m+1} \text{ if } \hat{p}_i = 0 \end{cases}$$

Under H_0 (ignoring ties), each p_i is uniformly distributed in $\{1/m, 2/m, \ldots, 1\}$, but they are not independent.

For symmetry, we consider as well the q_i p-values related to the *centrality* of the variable X_i on the sample Y.

$$\hat{q}_* = \frac{\sum_{j=1}^{n} I\left(D\left(Y_j|Y\right) \leq D\left(X_*|Y\right)\right)}{n+1} \qquad \begin{cases} q_i = \hat{q}_i \text{ if } \hat{q}_i \neq 0 \\ q_i = \frac{1}{n+1} \text{ if } \hat{q}_i = 0 \end{cases}$$

Observe that when H_0 does not hold, the depth of a curve in a family where it does not belong will be very low, and so would be its associated p-value. In that case the corresponding statistic will be very big.

Consider[12]

$$S_Y = \sum_{i=1}^{m} -\ln(p_i) \qquad S_X = \sum_{i=1}^{n} -\ln(q_i)$$

We want to associate a p-value to the pair (S_X, S_Y).

[12] Hedges & Olkin (1985), *Statistical Methods for Meta-Analysis* Academic Press

When the two samples display a difference in "scale", most, of the curves of the (say) X sample, turn out to be central with respect to the Y sample and S_X will not show a significant value. In such a situation, typically, several curves of the Y sample will turn out to be clearly outlying respect to the X sample, and the maximum will reach a significant value.

Hence, one possibility is considering

$$S = \max\{S_X, S_Y\}$$

In order to assess significance to the observed value of S, apply the permutation procedure.

Better empirical performance is obtained using following result:

Lemma: Combining p-values

Let p_X (p_Y) denote the p-value of S_X (S_Y), under the null permutation distribution, as obtained from procedure p-value if all subsets of size m were used (instead of just a sample of size B) and assuming the null hypothesis. Then

1. $\Pr(p_X \leq t) \leq t$ for any $t \in (0, 1)$, and the same holds for p_Y.
2. $\Pr(2 \min(p_X, p_Y) \leq t) \leq t$ for any $t \in (0, 1)$.

The null permutation distribution of S_X is a discrete distribution that can not be assumed uniform on its range (some values of S_X can appear more frequently than others when subsets are chosen at random). This is why (1) is not obvious.

Part (2) of the Lemma tells us that an appropriate p-value for $2\min(p_X, p_Y)$ is the observed value of this statistic itself.

Thus, our second way of getting a p-value from S_X and S_Y is the following:

Compute, approximately, p_X and p_Y for S_X and S_Y, respectively, using the procedure p-value described and use $2\min(p_X, p_Y)$ as p-value.

In the experiments described next, p_X and p_Y was computed using independent "draws" of subsets of the joint sample, a procedure that yields good power results.

Empirical comparison of powers

In order to fix a standard, we have also computed the empirical power for Horvath and Kokoszka's test for equality of mean functions. The null hypothesis is rejected for large values of

$$U_{m,n} = \frac{mn}{m+n} \int \left(\bar{X}_m(t) - \bar{Y}_n(t)\right)^2 dt$$

Under some regularity conditions, the asymptotic distribution of $\bar{X}_m(t) - \bar{Y}_n(t)$ is a Gaussian process Γ whose covariance can be approximated by the pooled covariance operator of the two samples, hence, the distribution of $U_{m,n}$ can be approximated by the first d terms in the Karhunen-Loève expansion of $\int_0^1 \Gamma^2(s)ds \approx \sum_{i=1}^{d} \lambda_i N_i^2$, λ_i are the (ordered) eigenvalues of the pooled covariance estimator, and N_i are i.i.d. N(0,1).

A simulation experiment

We have simulated samples of functional data as realisations from a geometric Brownian motion process

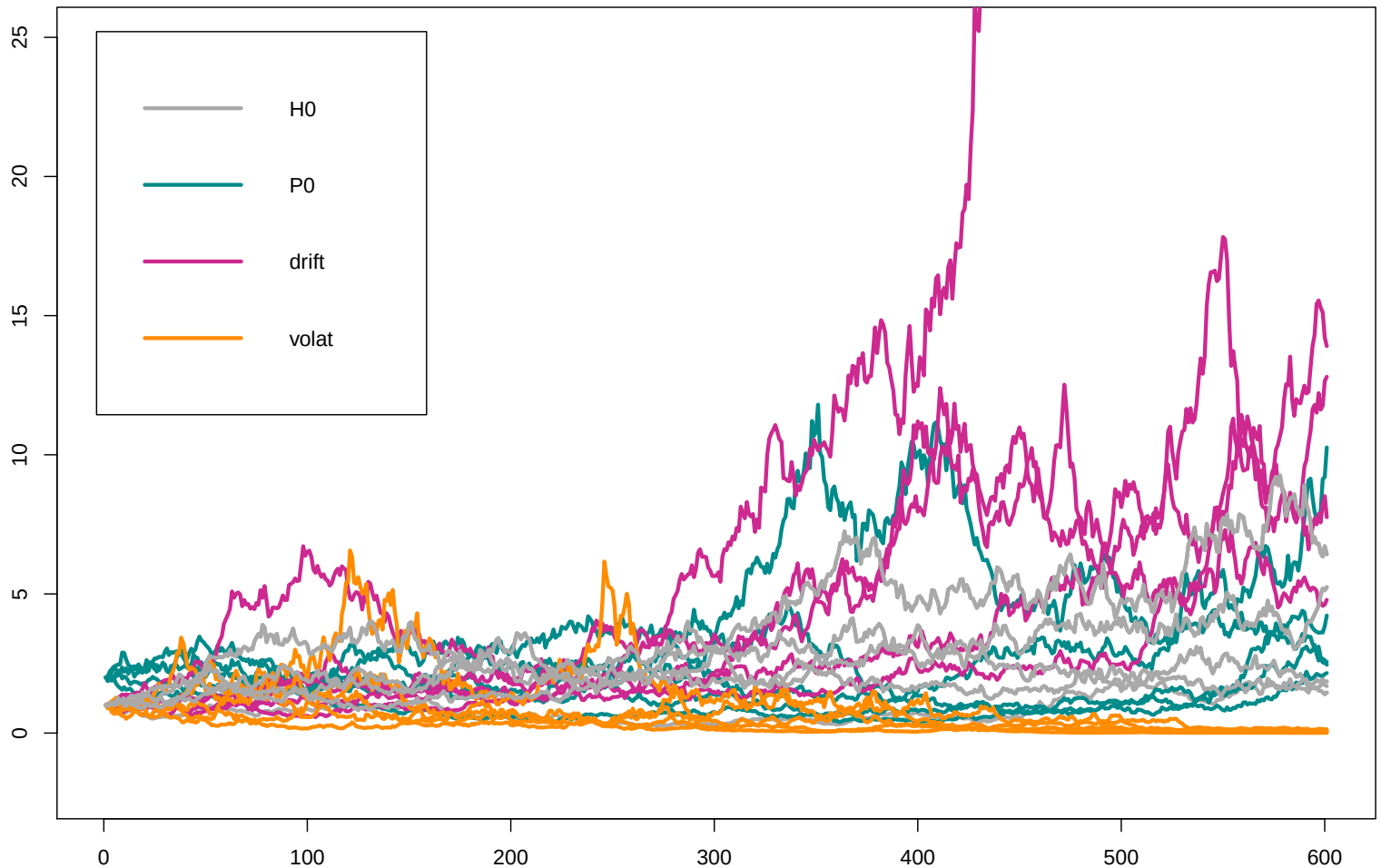
$$f(t) = X_0 \exp\left\{ rt - \frac{t\sigma^2}{2} + \sigma w_t \right\}$$

where r and σ are, respectively, the trend (drift) and volatility coefficients, w_t is a standard Wiener process and X_0 is the initial value.
We have tested

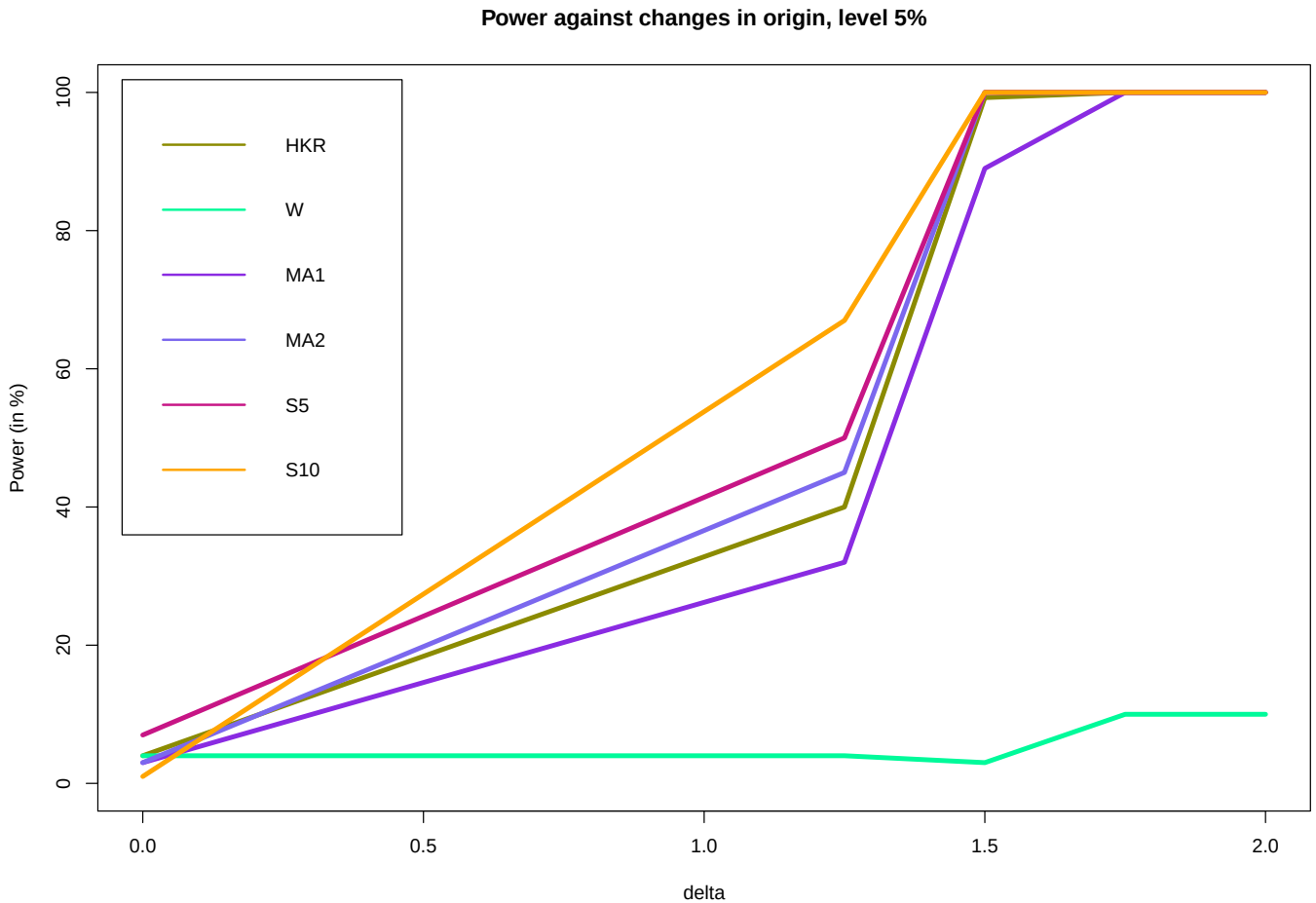
$$H_0 : \mathcal{L}(X) = \mathcal{L}(Y), \quad \text{against} \quad \mathcal{L}(X) \neq \mathcal{L}(Y)$$

where X was simulated with $r = 1, \sigma = 1, X_0 = 0$ and Y is any of the 'contaminated' samples with only one of the parameters variying at a time.

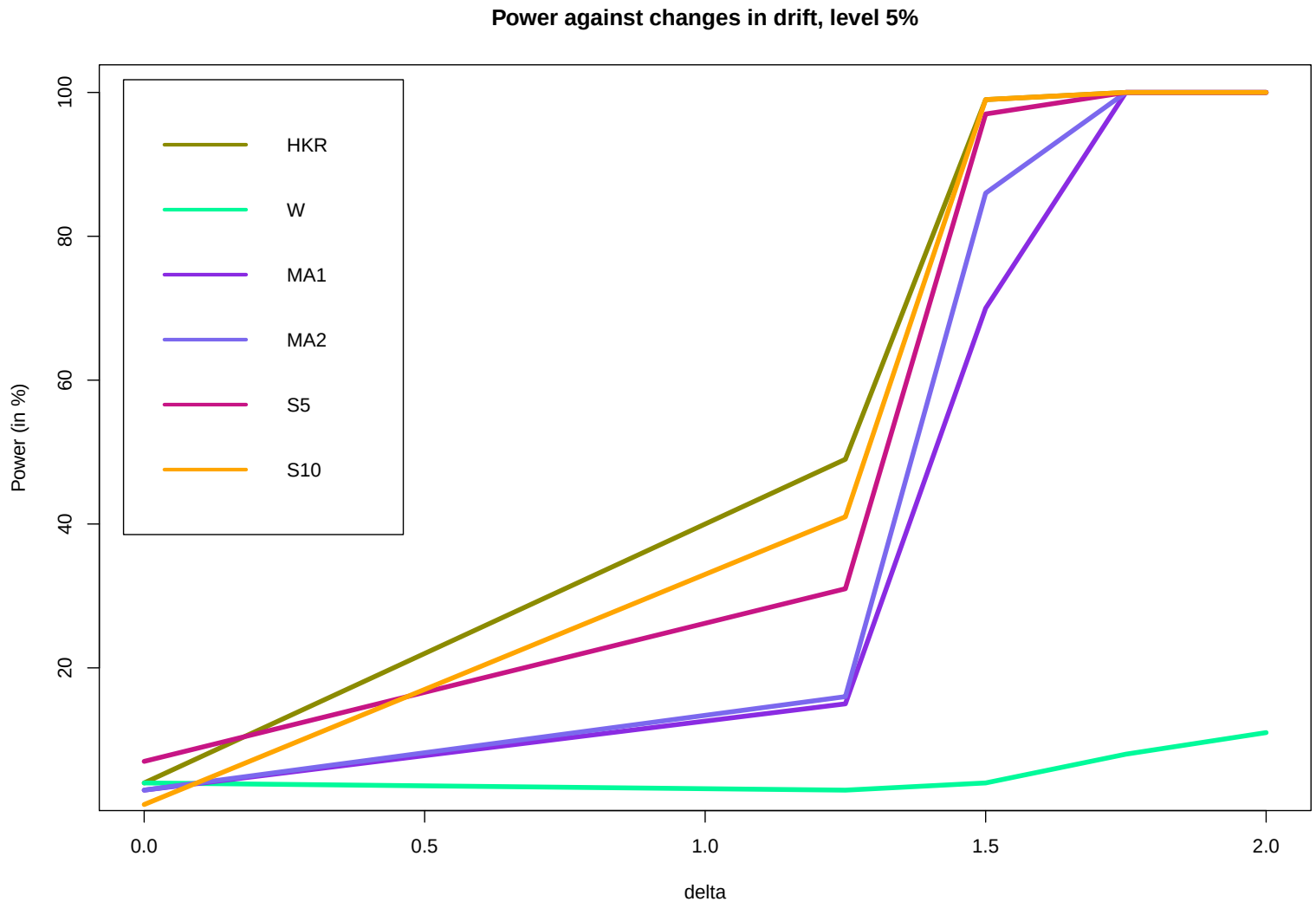
A few curves



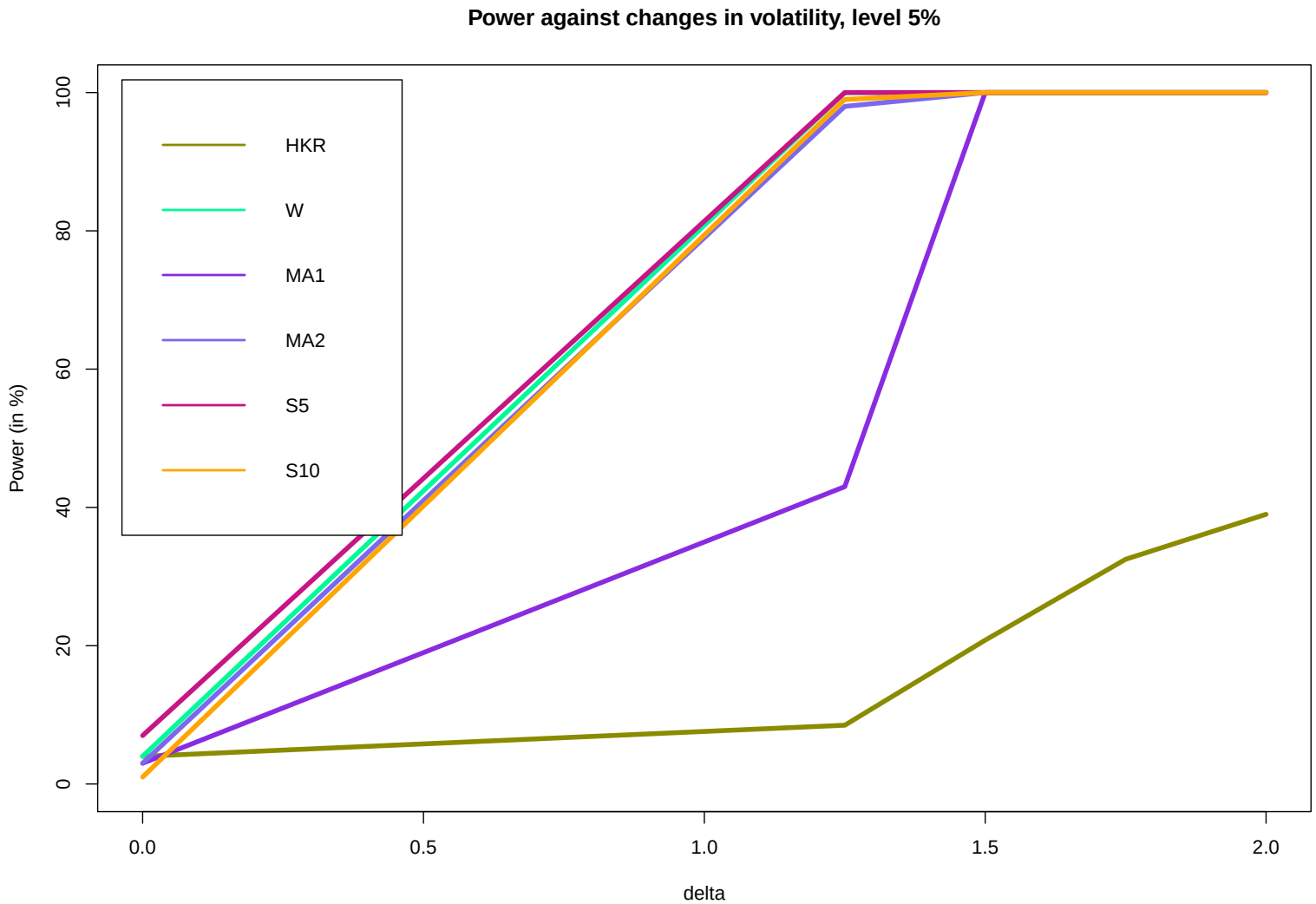
Empirical power of tests, Geometric Brownian motion



Empirical power of tests, Geometric Brownian motion



Empirical power of tests, Geometric Brownian motion



Comparing the NO_2 levels in different neighbourhoods

We have hourly measurements of nitrogen dioxide, a known pollutant, in four neighbourhoods in Barcelona.
The measurements were taken along the years 2014 and 2015 in automatic monitoring stations[13]

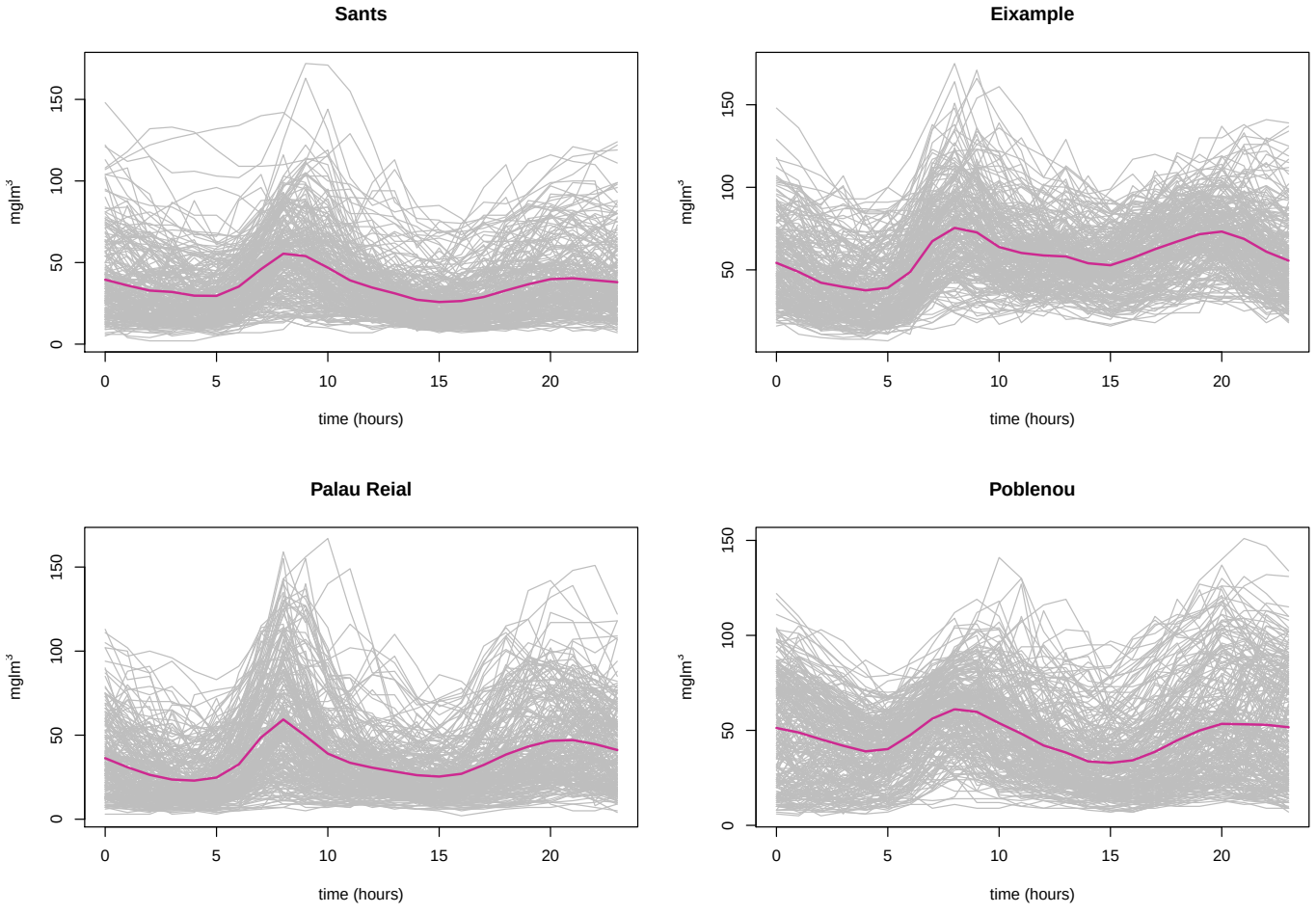
[13] available from `http://dtes.gencat.cat/icqa`.

We have split the data sets in workingdays ($\approx$220 curves) and non-working days ($\approx$120 curves), each year.

Questions of interest are

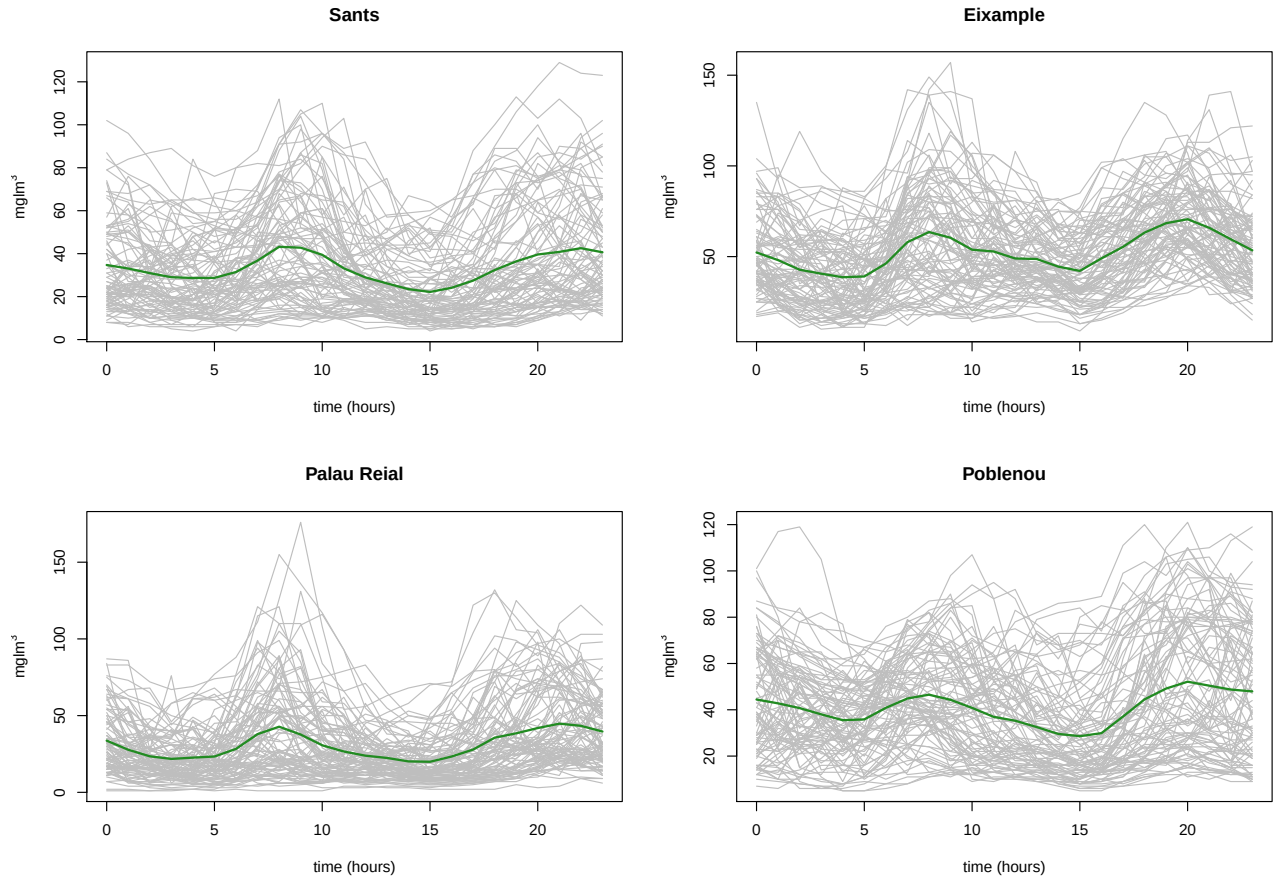
- Is the level of polutants signifficantly different during working and non-working days?
- Did the levels of NO_2 changed from one year to the next in each of the neighbourhoods?
- Comparing the pollution levels among the different neighbourhoods.

NO_2 levels on working days, 2015



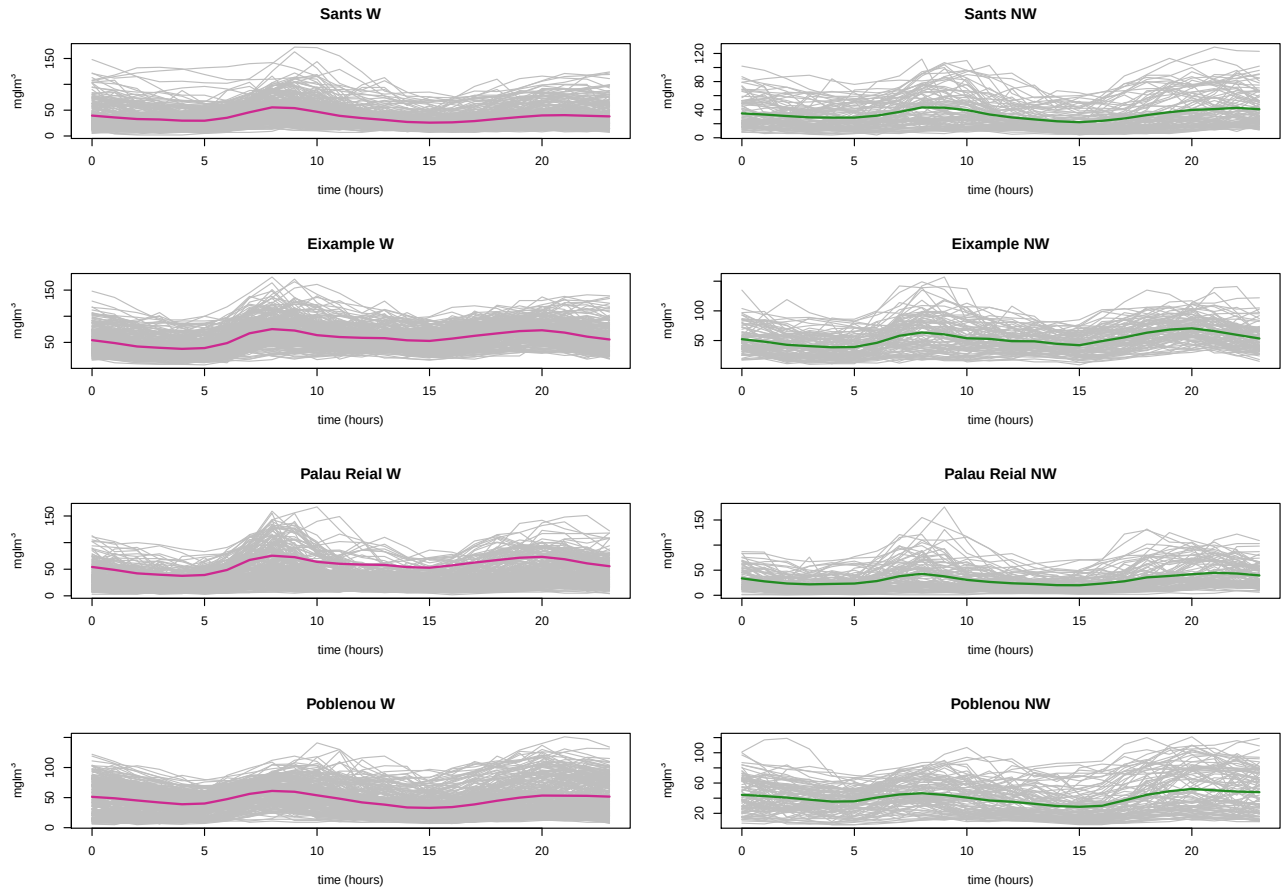
Schilling 10 and HK show marginal or no differences between pairs of neighbourhoods, the other tests show low p-values, indicating pairwise differences.

NO_2 levels on non- working days, 2015

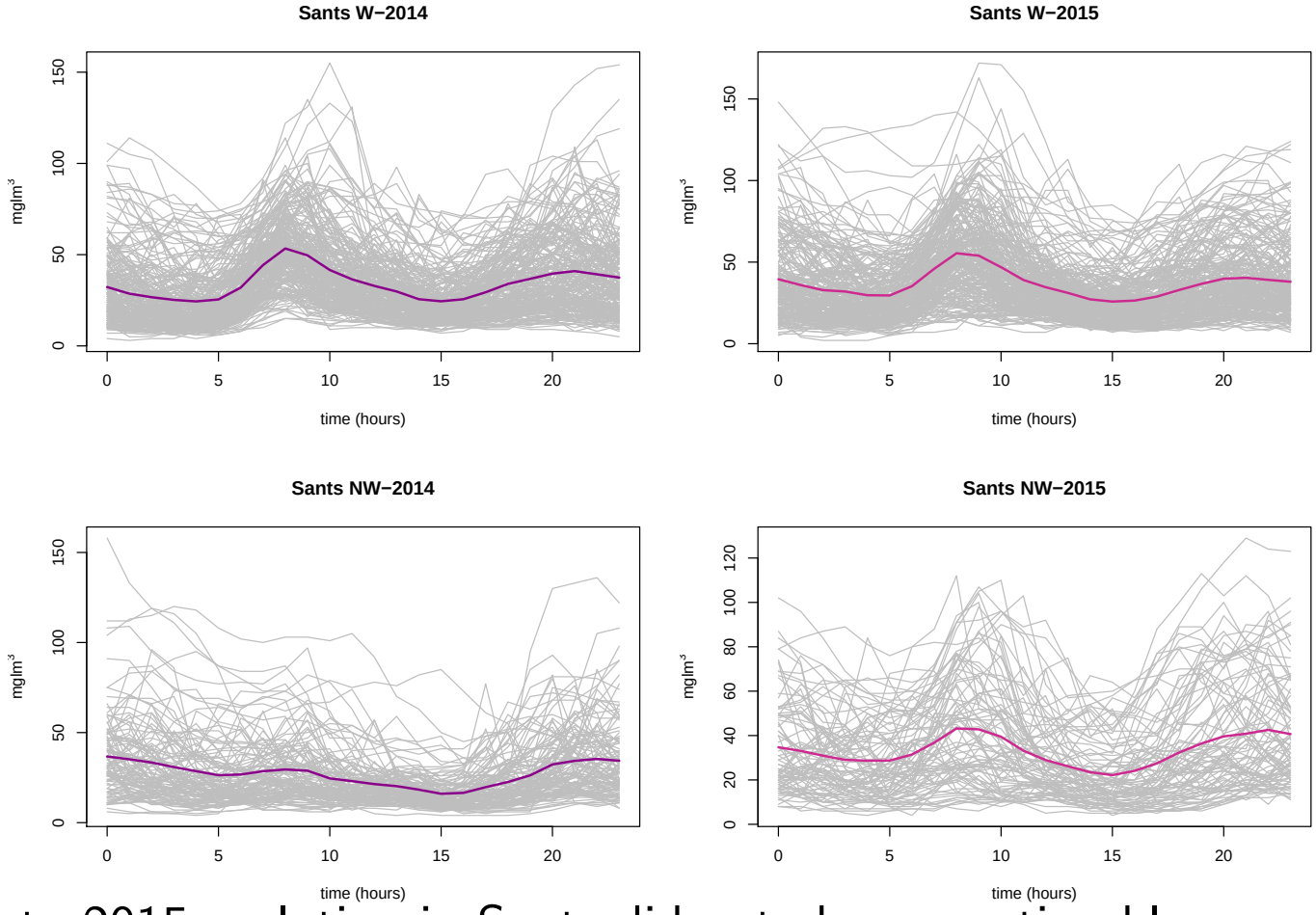


Moderate p-values (≈ 0.08) are found for Sants-Poblenou in 2015, all other pairs are significantly different.

NO_2 leves on working and non- working days, 2015



All tests show evidence of differences between working and non-working days, with Wilcoxon and Schilling-10 showing the strongest evidence of differences.



From 2014 to 2015, polution in Sants did not change noticeably on non-working days, but significant changes are found from one year to the next,on working days, with the Wilcoxon and Schilling procedures being the ones that find stronger evidence of change.

Functional Linear Models

Functional Linear Regression

We turn now to exploring predictive relationships, and begin by generalising linear models.

$$y_i = \beta_0 + \beta x_i + \epsilon_i$$

We have different scenarios for y_i and x_i:

- Functional covariate(s), scalar response
- Scalar covariate(s), functional response
- Functional covariate(s), functional response

Scalar response models

We observe y_i, $x_i(t)$ and want to model the dependence of y on x.

Example: predict (*log*) annual precipitation from the temperature profile.

An option[14] would be to choose $t_1, \ldots, t_k$ and set

$$y_i = \beta_0 + \sum \beta_j x_i(t_j) + \epsilon_i = \beta_0 + \boldsymbol{x}\beta + \epsilon_i$$

but, how many $t_1, \ldots, t_k$ and which ones? how do we avoid colinearity?
If we let $t_1 \ldots$ become increasingly dense, then the model in the limit becomes

$$y_i = \beta_0 + \boldsymbol{x}\beta + \epsilon_i = \beta_0 + \int \beta(t) x_i(t) dt + \epsilon_i$$

This is a sort of multivariate model with sums replaced by integrals.

The parameter is the regression function β, an infinite dimensional object which must be estimated from a finite sample.

[14] Mc Keague(2010)

Identification

A basic problem is that in linear regression we must have fewer covariates than observations.
And in this case, having $y_i, x_i(t)$ means there are infinitely many covariates.

The idea would be to estimate β by minimising the squared error

$$\beta(t) = \mathit{argmin} \sum \left(y_i - \beta_0 - \int \beta(t) x_i(t) dt \right)^2$$

but without any restrictions on the functional parameter, we can have a perfect fit (all residuals are 0) and the resulting estimates are erratic noisy functions which do not provide useful insight.

We need to ask for additional constraints: **we want β to be smooth!**

Adding a smoothing penalty

$$PENSSE_\lambda = \sum_{i=1}^{n} \left(y_i - \beta_0 - \int \beta(t) x_i(t) dt \right)^2 + \lambda \int [(L\beta)(t)]^2 dt$$

where L is a differential operator that acts on function β. A common choice of L is the second derivative $(L\beta)(t) = \beta''(t)$. In case of periodic data, usually harmonic acceleration$L = \omega^2 D + D^3$ that places no penalty on a sine wave and increases the penalty in higher order harmonics of Fourier approximation.

We have done similar things in the past: the choice of λ is critical: if it is too large, β will be too smooth, and crucial aspects of the shape will be lost; if it is too small, the estimator of β will reflect random errors.

As (almost) always, cross-validation will be the way of choosing λ.

But first, let us see how to estimate β with no penalty and for fixed λ.

Computing $\hat{\beta}$ and $\hat{y}$

Since β is a function, we need a basis expansion $\beta(t) = \sum c_j \phi_j(t) := \Phi^t \boldsymbol{c}$.

$$y_i = \beta_0 + \int \beta(t) x_i(t) dt + \epsilon_i = \beta_0 + \left[\int \Phi(t) x_i(t) dt \right] \boldsymbol{c} + \epsilon_i = \beta_0 + \boldsymbol{x}_i \boldsymbol{c} + \epsilon_i$$

where $\boldsymbol{x}_i = \int \Phi(t) x_i(t) dt$. Now consider the matrix $Z = [1 \quad x_i]$

$$\boldsymbol{y} = Z \begin{bmatrix} \beta_0 \\ \boldsymbol{c} \end{bmatrix} + \boldsymbol{\epsilon}$$

and with smoothing penalty matrix R

$$[\hat{\beta}_0 \quad \hat{\boldsymbol{c}}]^t = (Z^t Z + \lambda R)^{-1} Z^t \boldsymbol{y}$$

and then

$$\hat{\boldsymbol{y}} = \int \hat{\beta}(t) x_i(t) dt = \boldsymbol{Z} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\boldsymbol{c}} \end{bmatrix} := H_\lambda \boldsymbol{y}$$

Choosing the smoothing parameter

Ramsay, Hooker and Graves[15] recommend using leave-one-out Crossvalidation:

$$CV(\lambda) = \sum_{i=1}^{N} \left(y_i - \beta_0^{(-i)} - \int x_i(t)\beta_\lambda^{(-i)} dt \right)^2 = \sum_{i=1}^{N} \left(\frac{y_i - \hat{y}_i}{1 - H_{ii}} \right)^2$$

All these quantities can be computed with the `fRegress` command in `fda` library. See the file `FDA2(LM).R` for details within an example with Canadian Weather Data.

There's another library (`refund`) associated to the book by Kokotzka and Reimherr (2017).

[15] Functional Data Analysis with R and Matlab, Springer (2009)

Confidence Intervals

Assuming independent and normally distributed errors $\epsilon_i \sim N(0, \sigma^2)$, then, just as in standard Ridge Regression we have that

$$\mathbf{Var}\begin{bmatrix} \hat{\beta}_0 \\ \hat{\boldsymbol{c}} \end{bmatrix} = \left[(Z^t Z + \lambda R)^{-1} Z^t\right] \sigma^2 I \left[Z(Z^t Z + \lambda R)^{-1}\right]$$

Estimate

$$\hat{\sigma}^2 = \frac{SSE}{n - df} \qquad df = trace(H_\lambda)$$

and (pointwise) 95% confidence intervals for $\beta(t)$ are

$$\Phi(t)\hat{\boldsymbol{c}} \pm 1.96\sqrt{\Phi(t)^t \mathbf{Var}(\hat{\boldsymbol{c}})\Phi(t)}$$

Extensions to multiple functional covariates follows the same lines:

$$y_i = \beta_0 + \sum_{j=1}^{p} \int \beta_j x_{i,j}(t)dt + \epsilon_i$$

Functional Principal Component Regression

An alternative to the previous setting is writing $x_i(t)$ in terms of its principal components (or rather a few ones).

$$x_i(t) = \sum_j d_{i,j}\xi_j(t) \qquad d_{i,j} = \int x_i(t)\xi_j(t)dt$$

The model

$$y_i = \beta_0 + \sum_j \beta_j d_{i,j} + \epsilon_i$$

reduces to a standard linear regression problem.
Assuming the number of PCs is fixed, avoids cross-validation.
In fact, this was one of the first proposed methods.

fPCA and Regression Interpretation

$$y_i = \beta_0 + \sum_j \beta_j d_{i,j} + \epsilon_i$$

reduces to

$$y_i = \beta_0 + \sum_j \int \beta_j x_i(t) \xi_j(t) dt + \epsilon_i$$

and we can interpret

$$\beta(t) = \sum_j \beta_j \xi_j(t)$$

and write

$$y_i = \beta_0 + \sum_j \int \beta(t) x_i(t) dt + \epsilon_i$$

Confidence intervals derive from the variance of the $d_{i,j}$.

Summary

Essentially, the solution in the case of scalar response- functional covariate reduces to

1. Perform fPCA and use PC scores in a multivariate method.
2. Turn sums into integrals and add a smoothing penalty.

- Both methods can be adapted to functional response models.

$$y_i(t) = \beta_0(t) + \int \beta_1(s,t) x_i(s) ds + \epsilon(t)$$

Usually, penalise β_1 in each direction separately.

- Confidence intervals, cariance calculations follow the same principles.
- They can also be adapted in functional versions of GLMs, GAMs,...