An Introduction to Functional Data Analysis

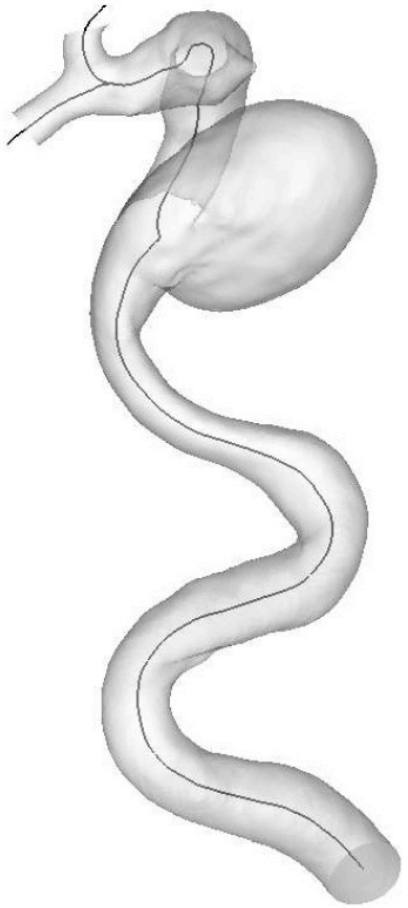
Piercesare Secchi

Applied Statistics – year 2018/2019

MOX, Department of Mathematics, Politecnico di Milano
piercesare.secchi@polimi.it

Introduction

Functional data: where do they come from?



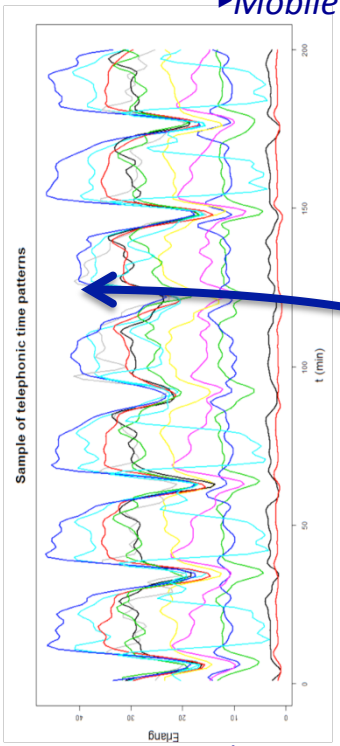
Explosive growth in recording complex and high-dimensional data having a functional nature (i.e., representable by curves, surfaces, dynamic curves and surfaces)

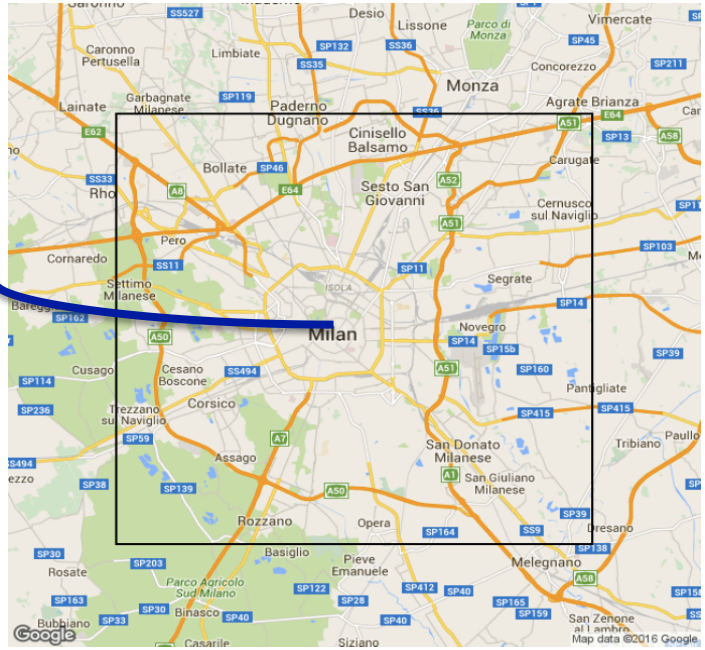
2D and 3D images and measures captured in time and space

▸*Reconstruction of an inner carotid artery with aneurysm, from angiographic images*

Sangalli, Secchi, Vantini, Veneziani (2009) J. R. Stat. Soc. Ser. C

▸*Mobile network data observed on a spatial lattice*

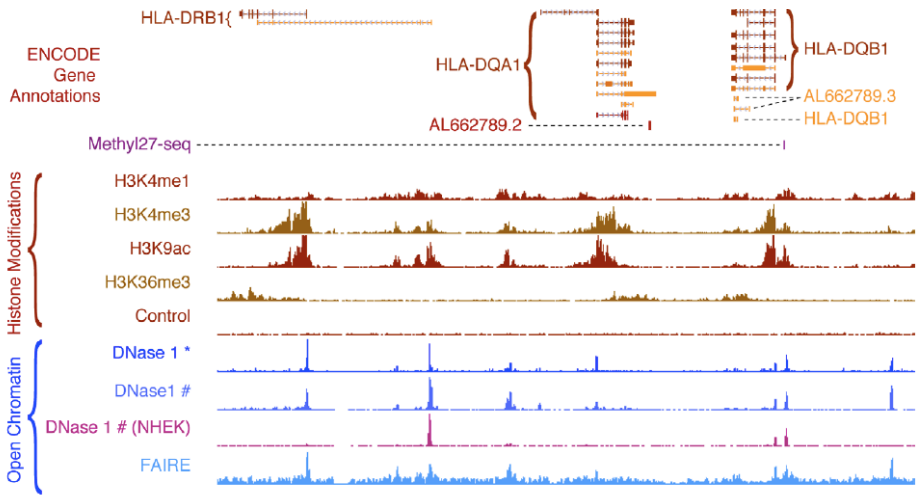




Secchi, Vantini, Vitelli (2015), Statistical Methods and Applications

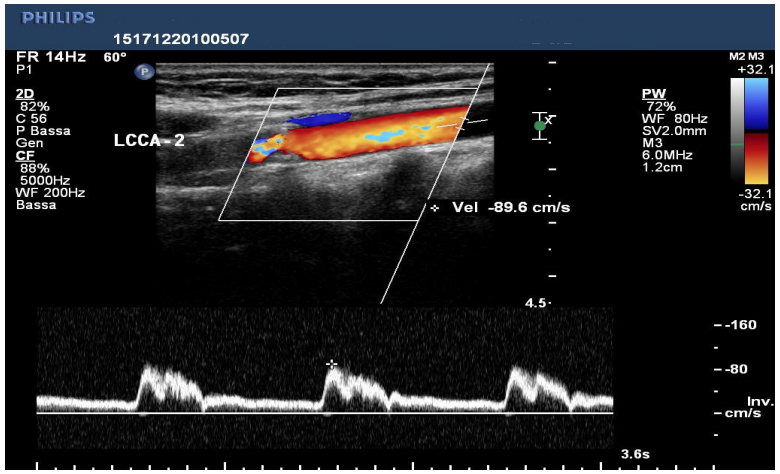
Functional data: where do they come from?

- Measurements of gene expression levels

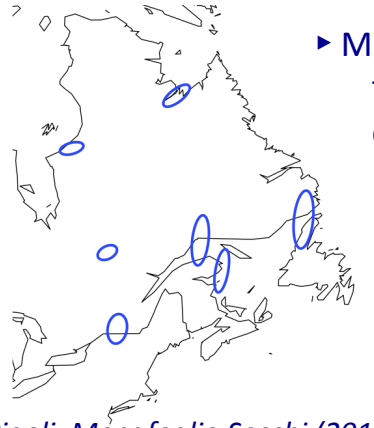


Cremona et al. (2015) BMC Bioinformatics

- ECD images for blood flow velocity field estimation



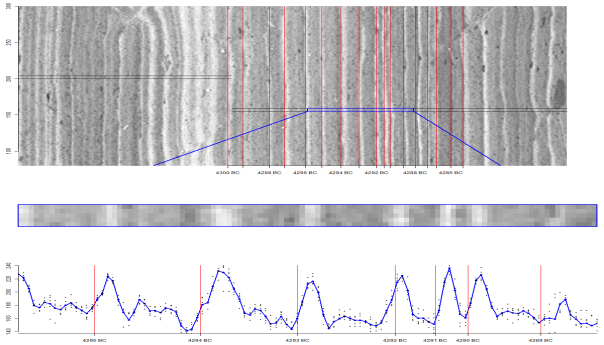
Azzimonti et al. (2014), JASA



- Manifold valued object data: temperature-precipitaion covariances in Quebec

Pigoli, Menafoglio Secchi (2016), JMVA

- Identification of past seasonal climates through the analysis of varves



Abramowicz et al. i (2016), SERRA

Functional data: where do they come from?

The analysis of complex and high dimensional data poses new and challenging problems in research

It is fueling one of the most fascinating and fast growing research fields of modern statistics

What are functional data?

- Informally, **functional data** are entities that can be described through a function, e.g., a curve, a surface, a image
- A **functional dataset** consists of a sample of functional observations
- Even though observations are actually discrete, the observed values reflect a **smooth variation of the phenomenon**. One might be interested not only in **point-wise** values, but also in **differential properties** of the data

Example: Berkeley Growth study
Observation of the height of 10 girls measured along 31 ages

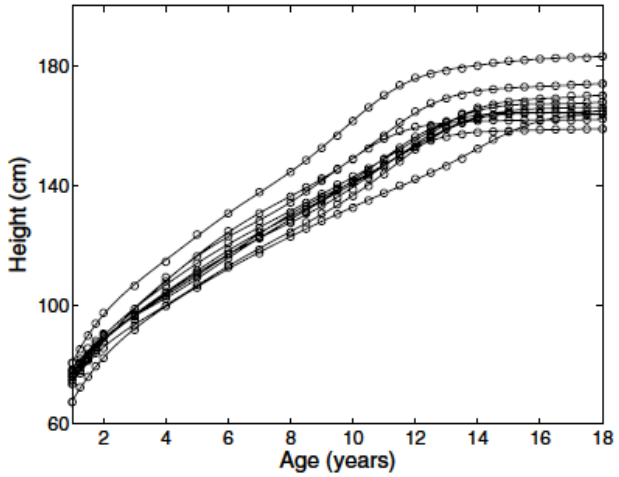


Figure 1.1. The heights of 10 girls measured at 31 ages. The circles indicate the unequally spaced ages of measurement.

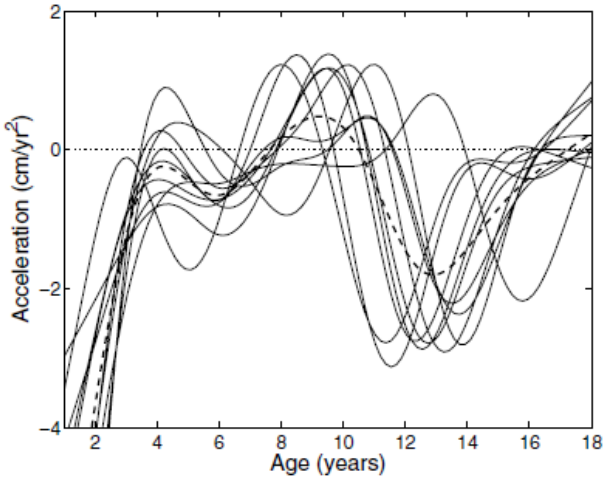


Figure 1.2. The estimated accelerations of height for 10 girls, measured in centimeters per year. The heavy dashed line is the cross-sectional mean, and is a rather poor summary of the curves.

Ramsay Silverman 2005 Springer

Berkeley Growth Curves as functional data

- Data reflect **smooth** variation of height over time: *h(t)*
- Some interesting features are only visible if **derivatives** are analyzed (e.g., mid-spurt and pubertal growth spurt)
- The grid spacing on the **time axis** is non-uniform. The underlying function might have been observed on different time points for different individuals
- **Large *p* small *n* problems**: classical multivariate methods fail when the number of variables is larger than the sample size (in this case, p=31, n=10)

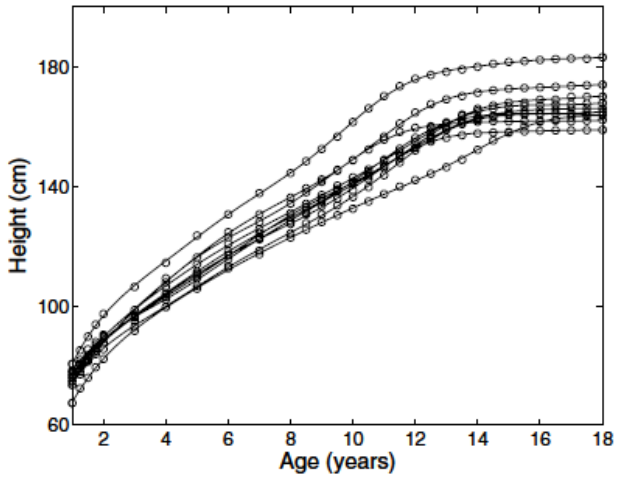


Figure 1.1. The heights of 10 girls measured at 31 ages. The circles indicate the unequally spaced ages of measurement.

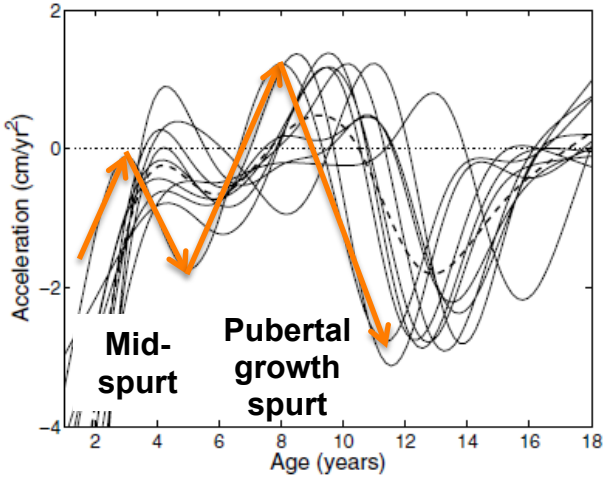
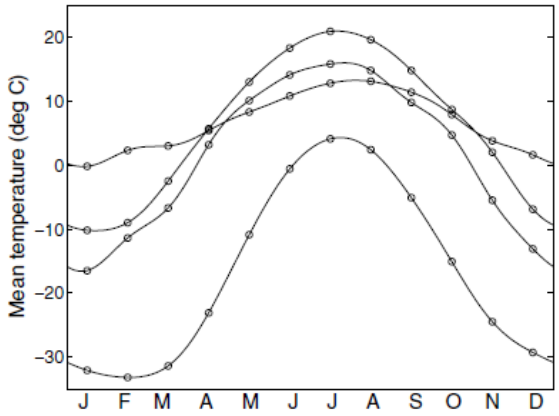


Figure 1.2. The estimated accelerations of height for 10 girls, measured in centimeters per year. The heavy dashed line is the cross-sectional mean, and is a rather poor summary of the curves.

Ramsay Silverman 2005 Springer

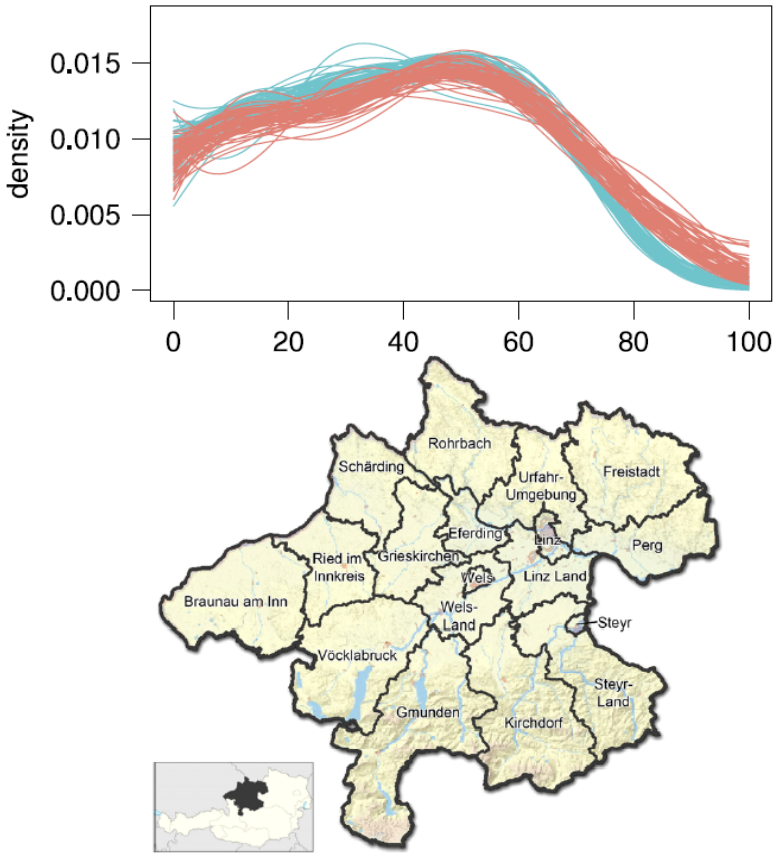
More simple examples of functional data

Example: Temperature curves in four locations in Canada (periodic data)



Ramsay Silverman 2005 Springer

Example: Density functions of Age Distribution in Austria (constrained data)



Hron et al. 2015, CSDA

Functional Data Analysis

- **Functional Data Analysis** is concerned with the statistical analysis of functional data

- **Typical goals of FDA**
  - Represent the data in ways that aid further analysis
  - Display the data to highlight their salient features
  - Study the main sources of pattern and variation among the data
  - Explain an outcome (response) using input/independent variable information. Here either the input or the output (or both) might be functional.
  - Classify the data or compare groups of data with respect to certain type of variations

- In this **short course**, we will be concerned with:
  - Representing data: given raw/discrete observations, represent the data through a functional form
  - Reducing the dimensionality of the representation space and highlights the main sources of variability (as in Principal Component Analysis)
  - Aligning (registration) and clustering (unsupervised classification) data

Course Agenda

1. **Hilbert space model for functional data**
    - 1.1. Basics notions on Hilbert spaces
    - 1.2. Hilbert space embedding for functional data
    - 1.3. Formal definition of functional data
2. **Smoothing and interpolation of functional data**
    - 2.1. Basis function
    - 2.2. Least square smoothing
    - 2.3. Smoothing with a differential penalization
3. **FDA & Dimensionality reduction in Hilbert spaces**
    - 3.1. Functional Principal Components in Hilbert spaces
    - 3.2. Examples in L2
4. **Data alignment and clustering**
    - 4.1 Phase and amplitude variability
    - 4.2 Landmark and continuous registration
    - 4.3 Decoupling phase and amplitude variability
    - 4.4 K-mean alignment

General references on FDA

Books:

- Ramsay, J.O. and Silverman, B.W. (2005). *Functional Data Analysis*, Springer, 2nd ed.

- Ramsay, J.O. and Silverman, B.W. (2002*). Applied Functional Data Analysis*, Springer.

- Ramsay, J.O., Hooker, G. and Graves, S. (2009). *Functional Data Analysis with R and Matlab*, Springer.

- Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice*, Springer.

- Horvath, L. and Kokoszka P. (2012). Inference for Functional Data with Applications, Springer.

Software: (available from CRAN)

- R package fda (corresponding Matlab code available from http://www.psych.mcgill.ca/misc/fda/)

- R package Refund

- Matlab code PACE

- R package mgcv

- R package fdakma (alignment and clustering)

- R package fdaPDE (surfaces)

1. Hilbert space model for functional data

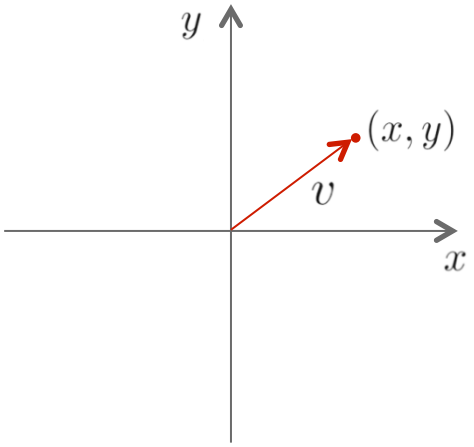
Course Agenda

1. **Hilbert space model for functional data**
    - 1.1. Basics notions on Hilbert spaces
    - 1.2. Hilbert space embedding for functional data
    - 1.3. Formal definition of functional data
2. **Smoothing and interpolation of functional data**
    - 2.1. Basis function
    - 2.2. Least square smoothing
    - 2.3. Smoothing with a differential penalization
3. **FDA & Dimensionality reduction in Hilbert spaces**
    - 3.1. Functional Principal Components in Hilbert spaces
    - 3.2. Examples in L2
4. **Data alignment and clustering**
    - 4.1 Phase and amplitude variability
    - 4.2 Landmark and continuous registration
    - 4.3 Decoupling phase and amplitude variability
    - 4.4 K-mean allignment
5. **Linear models**
    - 4.1. Functional Linear Models in Hilbert spaces
    - 4.2. Examples

The notion of **Hilbert space** generalizes the concept of Euclidean space to spaces of any (even infinite) dimension

- Vectorial structure (linear combinations)
- Distance, angles, projections (measure of dependence, best approximations)

Euclidean space $\mathbb{R}^2$



- Sum: $v_1 + v_2 = (x_1 + x_2, y_1 + y_2)$
- Product by a constant: $c \cdot v = (c \cdot x, c \cdot y)$

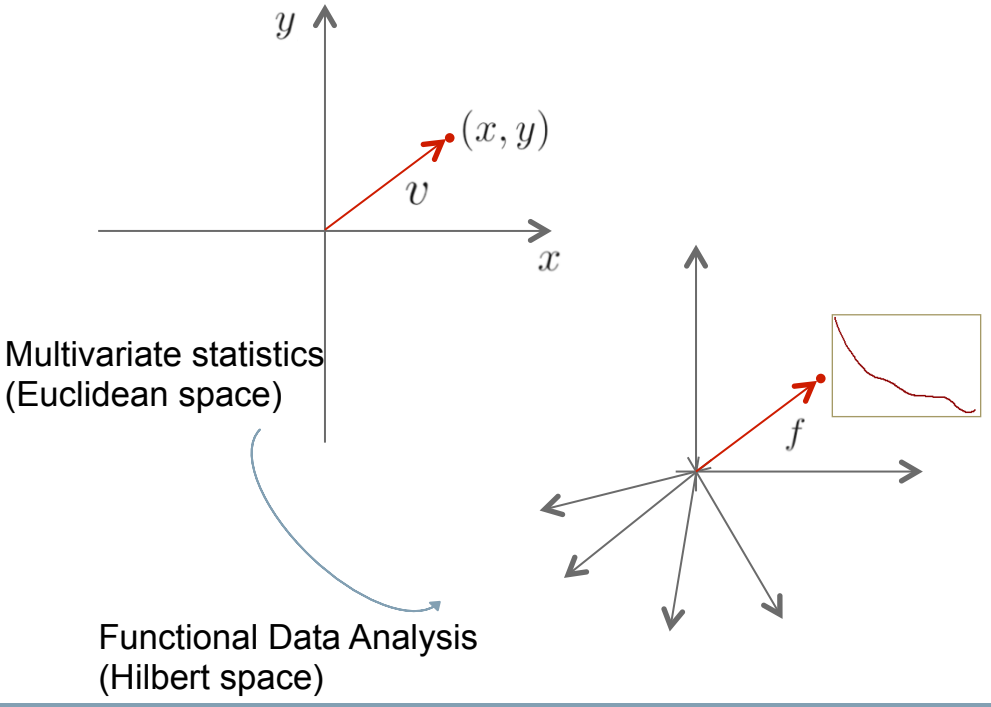
} Operations (+, ·)

- Norm (lenght of a vector): $\|v\| = (x^2 + y^2)^{1/2}$
- Distance: $\|v_1 - v_2\|^2 = (x_1 - x_2)^2 + (y_1 - y_2)^2$
- Angle: $\vartheta = \arccos \frac{\langle v_1, v_2 \rangle}{\|v_1\| \|v_2\|}$

} Inner product $\langle v_1, v_2 \rangle = (x_1 \cdot x_2) + (y_1 \cdot y_2)$

The notion of **Hilbert space** generalizes the concept of Euclidean space to spaces of any (even infinite) dimension

- Vectorial structure (linear combinations)
- Distance, angles, projections (measure of dependence, best approximations)



Why Hilbert spaces?

- We understand functional data as **points of a space of functions**
- Many methods of **multivariate statistics can be extended to data embedded in a Hilbert space**, through the notions of inner product and norm

Let *H* be a linear space. An inner product on *H* is a bilinear, symmetric, positive definite form

$$\langle \cdot, \cdot \rangle : H \times H \to \mathbb{R}$$

that satisfies

$$\text{(i)} \quad \langle \lambda x + y, z \rangle = \lambda \langle x, z \rangle + \langle y, z \rangle \qquad \forall \lambda \in \mathbb{R}, \ \ \forall x, y, z \in H$$

$$\text{(ii)} \quad \langle x, y \rangle = \langle y, x \rangle \qquad \forall x, y \in H$$

$$\text{(iii)} \quad \langle x, x \rangle \geq 0 \qquad \forall x \in H$$

$$\text{(iv)} \quad \langle x, x \rangle = 0 \iff x = 0$$

In particular:

- The inner product allows to measure lengths and angles
- It allows to define orthogonality: two vectors in *H* are orthogonal if $\langle x, y \rangle = 0$
- The inner product induces a norm and a metric
- The inner product allows generalizing the Pythagoras' Theorem:

$$\|x + y\|^2 = \|x\|^2 + \|y\|^2 \quad \text{if and only if} \quad \langle x, y \rangle = 0$$

A (real) Hilbert space *H* is an inner product space that is complete, in the norm induced by the inner product.

- A Hilbert space is complete in the sense that it contains all the limit points of its Cauchy sequences;
- A Hilbert space is separable if it contains a dense countable subset;

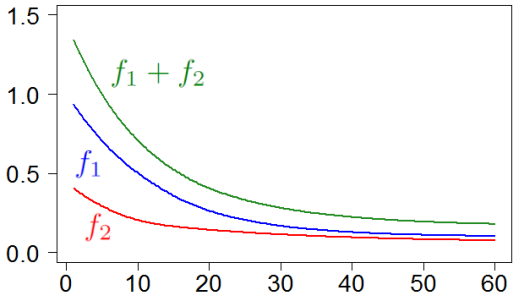
- Useful properties:
  - In a Hilbert space one has the notion of orthogonal projection and of best approximations
  - A Hilbert space *H* is separable iff it has an orthonormal basis $\{u_n\}_{n\in\mathbb{N}}$
  - If *H* is separable Hilbert space, $\{u_n\}_{n\in\mathbb{N}}$ is an orthonormal basis and $x \in H$. Then

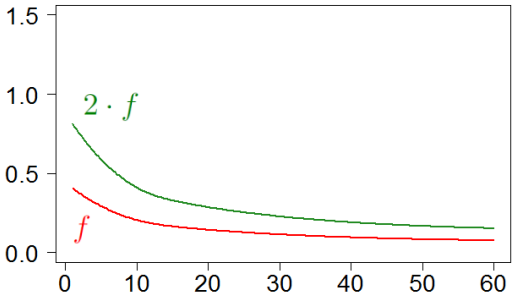
$$x = \sum_{n=1}^{\infty} \langle x, u_n \rangle u_n. \qquad \textbf{Basis expansion}$$

L^2: space of real-valued square-integrable functions

- Sum: $(f_1 + f_2)(t) = f_1(t) + f_2(t)$
- Product by a constant: $(c \cdot f)(t) = c \cdot f(t)$

} Operations (+, ·)





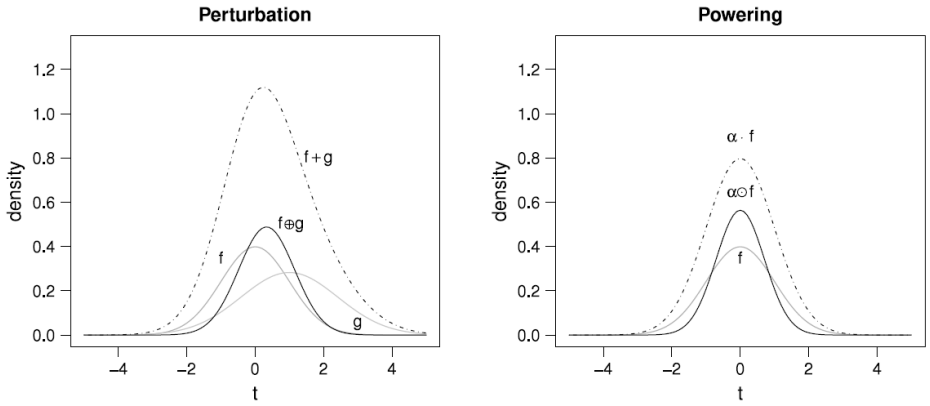
- Norm: $\|f\|^2 = \int (f(t))^2 dt$
- Distance: $\|f_1 - f_2\|^2 = \int (f_1(t) - f_2(t))^2 dt$
- Angle: $\vartheta = \arccos \frac{\langle f_1, f_2 \rangle}{\|f_1\| \|f_2\|}$

} Inner product $\langle f_1, f_2 \rangle = \int (f_1(t) \cdot f_2(t)) dt$

More precisely, L^2 is a quotient space with respect to the equivalence relation: $x = y$ if $\int [x(t) - y(t)]^2 dt = 0$

B^2: space of density functions on a close interval *I*, with log in L^2

- Equivalence relation: f, g are equivalent if they are proportional (*scale invariance*)
- Sum (perturbation): $(f \oplus g)(t) = \dfrac{f(t)g(t)}{\int_I f(s)g(s)\,ds}$
- Product by a constant (powering): $(\alpha \odot f)(t) = \dfrac{f(t)^\alpha}{\int_I f(s)^\alpha\,ds}, \quad t \in I.$
- Inner product: $\langle f, g \rangle_{\mathcal{B}} = \dfrac{1}{2\eta} \int_I \int_I \ln \dfrac{f(t)}{f(s)} \ln \dfrac{g(t)}{g(s)}\,dt\,ds$
- Norm: $\|f\|_{\mathcal{B}} = \left[\dfrac{1}{2\eta} \int_I \int_I \ln^2 \dfrac{f(t)}{f(s)}\,dt\,ds \right]^{1/2}$



Note: the geometry of L^2 wouldn't make sense for density functions

B^2: space of density functions on a close interval *I*, with log in L^2

- Equivalence relation: f, g are equivalent if they are proportional (*scale invariance*)
- Sum (perturbation): $(f \oplus g)(t) = \dfrac{f(t)g(t)}{\int_I f(s)g(s)\,ds}$
- Product by a constant (powering): $(\alpha \odot f)(t) = \dfrac{f(t)^\alpha}{\int_I f(s)^\alpha\,ds}, \quad t \in I.$
- Inner product: $\langle f, g\rangle_{\mathcal{B}} = \dfrac{1}{2\eta}\int_I\int_I \ln\dfrac{f(t)}{f(s)}\ln\dfrac{g(t)}{g(s)}\,dt\,ds$
- Norm: $\|f\|_{\mathcal{B}} = \left[\dfrac{1}{2\eta}\int_I\int_I \ln^2\dfrac{f(t)}{f(s)}\,dt\,ds\right]^{1/2}$

- B^2 is isomorphic to L^2 (in fact, all the Hilbert spaces are isomorphic). An isometric isomorphism is provided, e.g., by the **centred log-ratio transformation**

$$\mathrm{clr}(f)(t) = f_c(t) = \ln f(t) - \frac{1}{\eta}\int_I \ln f(s)\,ds.$$

Exercise: prove that

$$\mathrm{clr}(f \oplus g)(t) = f_c(t) + g_c(t), \qquad \mathrm{clr}(\alpha \odot f)(t) = \alpha \cdot f_c(t), \qquad \langle f, g\rangle_{\mathcal{B}} = \langle f_c, g_c\rangle_2 = \int_I f_c(t)g_c(t)\,dt.$$

B^2: space of density functions on a close interval *I*, with log in L^2

- Equivalence relation: f, g are equivalent if they are proportional (*scale invariance*)
- S
- P
- In
- N
- *B* is

- Hilbert space structure for functional compositional data (e.g., probability density functions)
- Account for the key properties of compositional data: scale invariance, relative scale, sub-compositional coherence
- Meaningful interpretations in mathematical statistics, e.g.,
  - Exponential families as affine finite-dimensional subspaces
  - Perturbation $\oplus$ as a Bayes update of information

c

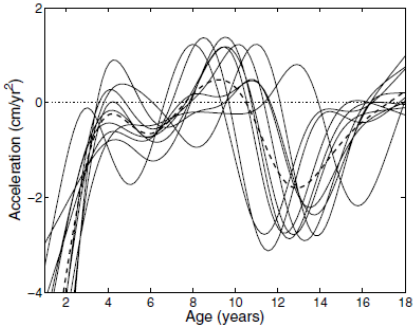
$$\operatorname{clr}(f)(t) = f_c(t) = \ln f(t) - \frac{1}{\eta}\int_I \ln f(s)\,ds.$$

Exercise: prove that

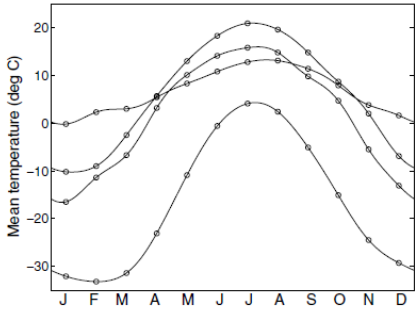
$$\operatorname{clr}(f \oplus g)(t) = f_c(t) + g_c(t), \qquad \operatorname{clr}(\alpha \odot f)(t) = \alpha \cdot f_c(t), \qquad \langle f, g\rangle_{\mathcal{B}} = \langle f_c, g_c\rangle_2 = \int_I f_c(t) g_c(t)\,dt.$$

1.2. Hilbert space embedding for functional data

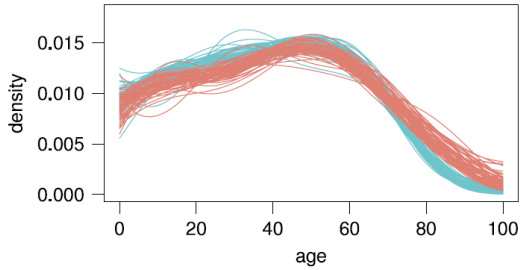
- As a first step of any functional data analysis, one need to **choose the embedding** for the data
- **Separable Hilbert spaces are a convenient choice** (projections, best approximations).
  Note: Not all the interesting spaces are Hilbert: e.g., the space of continuous functions is not a Hilbert space. Other interesting spaces: Riemannian manifolds (OODA)
- Examples of Hilbert spaces for FDA:
  - L^2, space of square integrable functions: OK for most data analyses (especially if data are unconstrained)
  - B^2, space of functional compositions: useful for density functions



Acceleration in Berkeley Growth data



Temperatures in Canada



Population Age densities

1.3. Formal definition of functional data

Functional random variables and functional data

- Let *H* be a Hilbert space, whose points are functions defined on a closed interval $T = [t_{min}, t_{max}]$ (e.g., range of time during which the data are collected)
- Hereafter, we will always consider functional data in Hilbert spaces

Definition 1:

A **functional random variable** is a random element on a probability space $(\Omega, \mathfrak{F}, \mathbb{P})$ in the space *H*: $X : \Omega \to H$

Definition 2:

A **functional datum** x is a realization of a functional random variable, i.e., for $\omega \in \Omega$

$$x = X(\omega) : T = [t_{min}, t_{max}] \to \mathbb{R}$$

Definition 3:

A **functional dataset** is a collection of functional data.

Let $X : \Omega \to H$ be a functional random variable in *H.* Hereafter, we always assume that $\mathbb{E}[\|X\|_H^4] < \infty$.

Definition 4:

We call Fréchet mean of X the (unique) element μ of *H* that solves

$$\underset{x \in H}{\operatorname{arginf}}\ \mathbb{E}[\|X - x\|_H^2].$$

- If $H=L^2$ (space of square-integrable functions), the Fréchet mean coincides a.e. with the point-wise mean

$$\mathbb{E}[X(t)] = \mu(t), \quad t \in T$$

- If $H=B^2$ (Bayes space of PDFs), the Fréchet mean can be computed as

$$\mu = clr^{-1}(\mathbb{E}[clr(X)])$$

(in particular, one can define the mean of the clr-transformed variable point-wise)

- In any *H*, one can **estimate the mean via the sample estimator**

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

In $H=L^2$, this is the point-wise sample mean

Let $X : \Omega \to H$ be a **zero-mean** functional random variable in *H*, such that $\mathbb{E}[\|X\|_H^4] < \infty$.

Definition 5:
We call covariance operator of X the operator from *H* to *H* defined as

$$Cx = \mathbb{E}[\langle X, x \rangle X], \quad x \in H$$

- If $H=L^2$ (space of square-integrable functions), the covariance operator can be equivalently defined through a kernel operator

$$[Cx](t) = \int_T c(s,t)x(s)d(s) \ , \quad x \in L^2$$

where the covariance kernel is precisely the point-wise covariance

$$c(s,t) = \mathbb{E}[X(s)X(t)]$$

- In $H=\mathbb{R}^p$, the covariance operator coincides with the linear operator defined by the covariance matrix

Let $X : \Omega \to H$ be a **zero-mean** functional random variable in *H*, such that $\mathbb{E}[\|X\|_H^4] < \infty$.

Definition 5:
We call covariance operator of X the operator from *H* to *H* defined as

$$Cx = \mathbb{E}[\langle X, x\rangle X], \quad x \in H$$

- In any *H*, the covariance operator can be estimated through the sample covariance operator

$$Sx = \frac{1}{N}\sum_{i=1}^{N}\langle X_i, x\rangle X_i, \quad x \in H$$

- If $H=L^2$, one can use the alternative definition

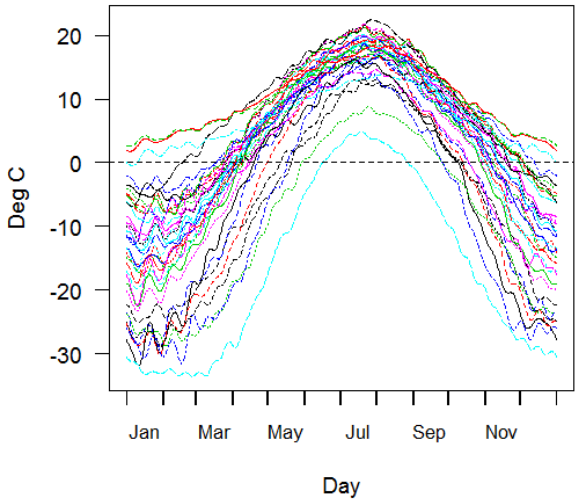
$$[Sx](t) = \int_T \widehat{c}(s,t)x(s)d(s)\,, \quad x \in L^2 \text{ with } \quad \widehat{c}(s,t) = \frac{1}{N}\sum_{i=1}^{N} X(s)X(t)$$

1.3. Formal definition of functional data

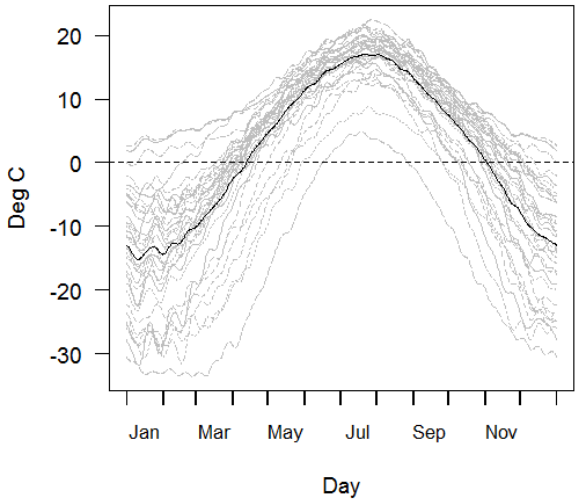
An example in L^2

Ramsay Silverman 2005 Springer

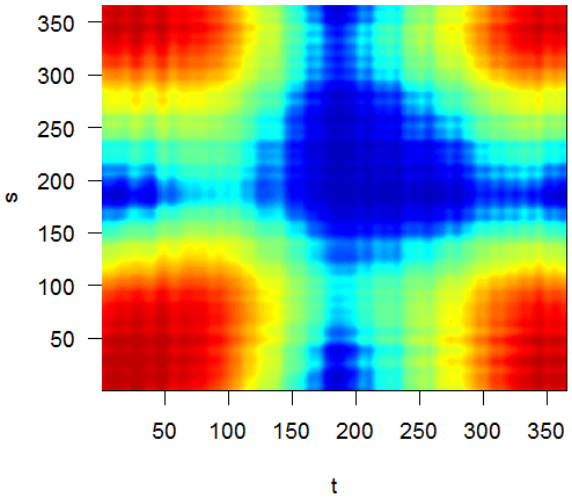
Dataset of Temperatures in Canada (35 observations)



Functional dataset



Sample mean



Sample covariance kernel

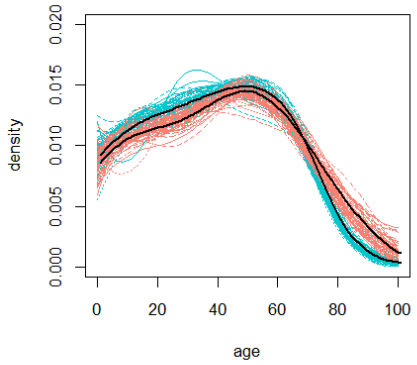
1.3. Formal definition of functional data

An example in B^2

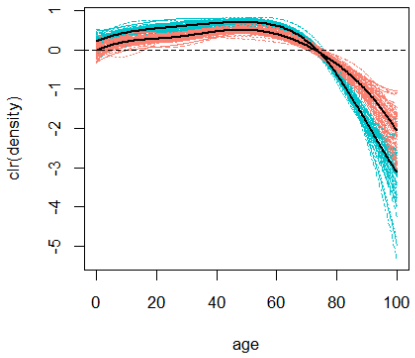
Hron et al. 2015, CSDA

Dataset of Age Densities (114 observations)

Functional dataset

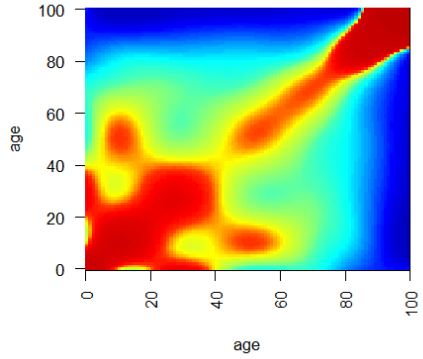


Functional dataset (after clr)

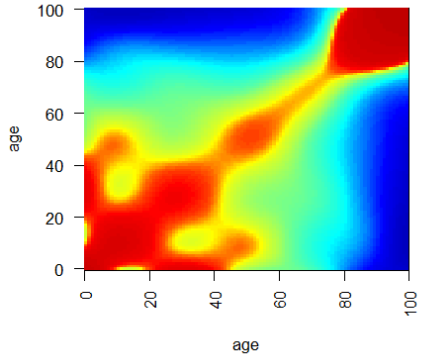


Sample covariance kernel (after clr)

males



females



2. Smoothing and interpolation of functional data

Course Agenda

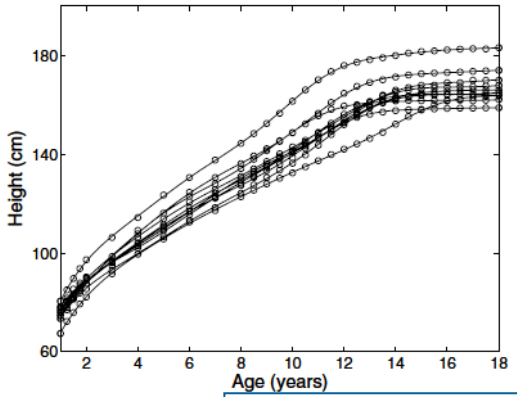
1. **Hilbert space model for functional data**
    - 1.1. Basics notions on Hilbert spaces
    - 1.2. Hilbert space embedding for functional data
    - 1.3. Formal definition of functional data
2. **Smoothing and interpolation of functional data**
    - 2.1. Basis function
    - 2.2. Least square smoothing
    - 2.3. Smoothing with a differential penalization
3. **FDA & Dimensionality reduction in Hilbert spaces**
    - 3.1. Functional Principal Components in Hilbert spaces
    - 3.2. Examples in L2
4. **Data alignment and clustering**
    - 4.1 Phase and amplitude variability
    - 4.2 Landmark and continuous registration
    - 4.3 Decoupling phase and amplitude variability
    - 4.4 K-mean alignment
5. **Linear models**
    - 4.1. Functional Linear Models in Hilbert spaces
    - 4.2. Examples

- Typical observations of functional data are **discrete** and **noisy**. Indeed, the record of each function x_i usually consists of n_i pairs (t_{ij}, y_{ij}), with $j=1,\ldots,n_i$.

- We model these pairs as $y_j = x(t_j) + \epsilon_j$.

  Note. The argument values t_{ij} may or may not be the same for each datum.

- For each *i,* we aim to reconstruct the underlying functional observation function x_i from the records (t_{ij}, y_{ij}), with $j=1,\ldots,n_i$
  Note: The assumptions on the properties of x_i (e.g., the smoothness) will reflect on the way we proceed to reconstruct the data

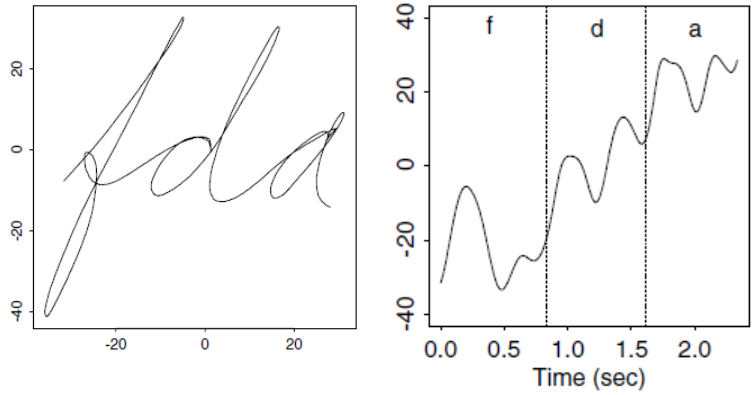


Example: Height of 10 girls in Berkeley Growth data
- Raw data are depicted as symbols
- Reconstructed functional data are depicted as lines

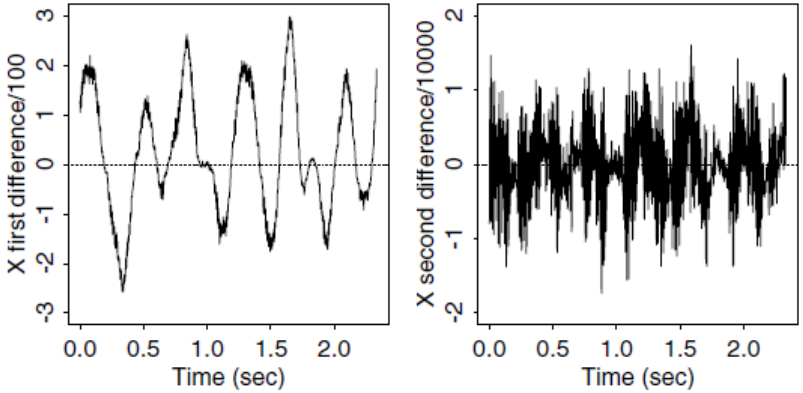
Ramsay Silverman 2005 Springer

- Depending on our prior knowledge on the measurement error (i.e., on the properties of the noise ϵ_j, we can decide to perform
  - Interpolation: the functional form reconstructed actually interpolates its discrete observations (noiseless measurements)
  - Smoothing: the functional form is smoother than the actual observations (noisy measurement)

- In most cases, smoothing is preferred to interpolation.
  Note. Differential operations (i.e., derivatives) amplify the effect of noise. Smoothing raw data actually enhances the estimation of derivatives



Functional datum: X-coordinate in hand-writing



First and second derivatives estimated via finite-differences

Ramsay Silverman 2005 Springer

If we aim to interpolate or smooth discrete data we typically perform the following steps:

- Choose a **target functional form** for x_i , that possibly depends on parameters
- **Estimate** the functional form, based on the pair (t_{ij}, y_{ij})

The **choice of the functional form** depends on various factors:

- **Features that we want to extract**: e.g., regularity of the functional form if the target is the differential information of the function (first, second derivatives)
- **Functional space embedding**: when we choose a Hilbert space embedding, we automatically identify possible orthonormal bases

In most cases:

- Hilbert space embedding is employed (especially, $H=L^2$)
- Functions are represented by basis functions

- Roughly speaking: a **system of basis functions** is a set of known functions that are linearly independent and allows us to approximate arbitrarily well any function as a linear combination of (a sufficiently large number of) *K* of these functions

- More precisely: given a system of basis functions ϕ_k , we will express a function x by the linear expansion

$$x(t) = \sum_{k=1}^{K} c_k \phi_k(t)$$

or in matrix notation

$$x = \mathbf{c}'\boldsymbol{\phi} = \boldsymbol{\phi}'\mathbf{c} \ .$$

Recall. In **Hilbert spaces**, we can always find an **orthonormal basis** that allows approximating, with any desired precision, any element of the space through the expansion

$$x = \sum_{n=1}^{K} \langle x, u_n \rangle u_n .$$

Note. In the following, we will mainly refer to L^2

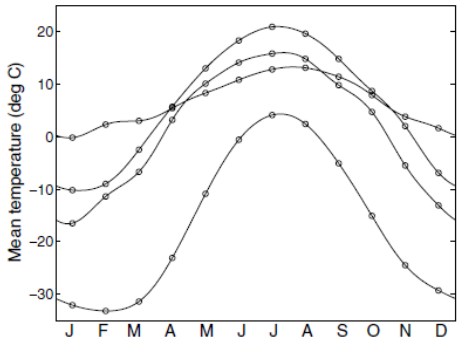
- One of the best known basis expansion in L^2 is provided by the **Fourier series**

$$\hat{x}(t) = c_0 + c_1 \sin \omega t + c_2 \cos \omega t + c_3 \sin 2\omega t + c_4 \cos 2\omega t + \ldots$$

i.e., with the previous notation

$$\phi_0(t) = 1, \phi_{2r-1}(t) = \sin r\omega t, \text{ and } \phi_{2r}(t) = \cos r\omega t.$$

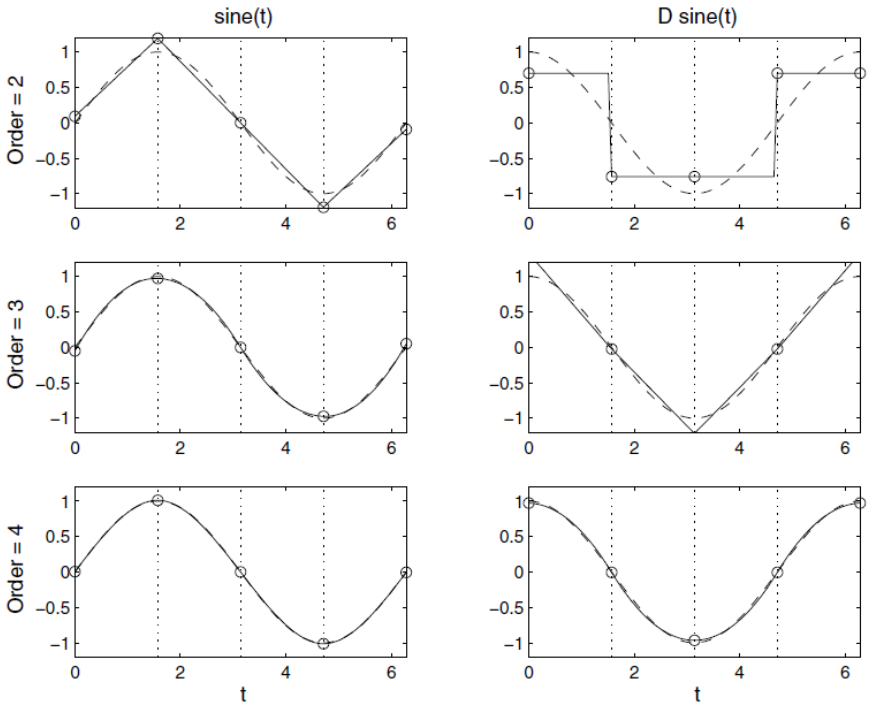
- **Properties**:
  - The basis is periodic, of period $2\pi/\omega$
  - If the values of t_j are equally spaced in T and the period is equal to the length of T than the basis is orthogonal (it can be made orthonormal via a proper rescaling)
- **Useful for:**
  - Extremely stable functions (i.e., no strong local features), for which uniformly smooth behavior is expected
  - Periodic data
- **Inappropriate for:**
  - Discontinuous functions (or with discontinuous derivatives)



Temperatures in Canada

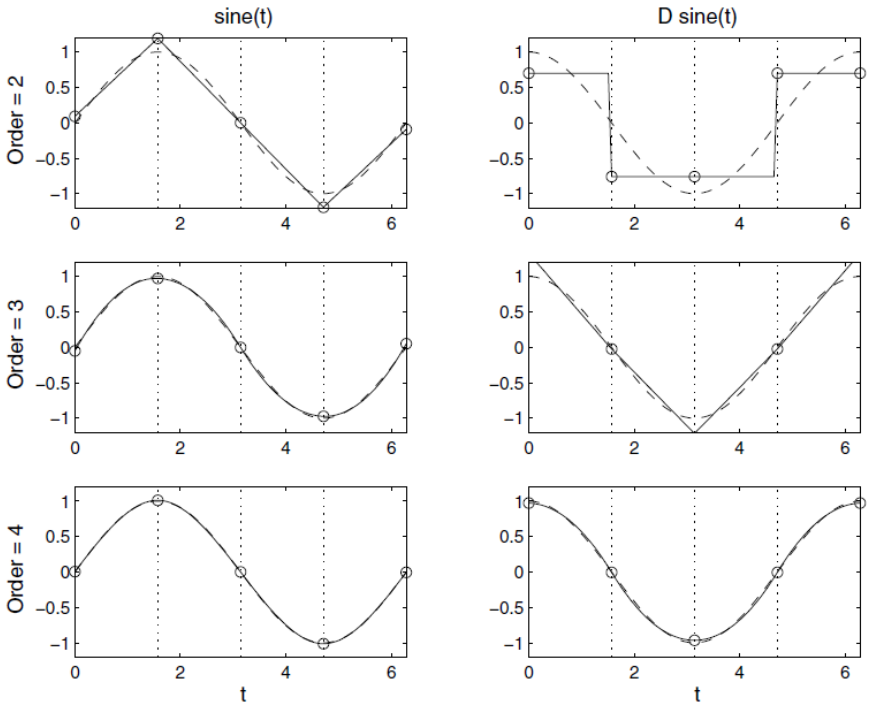
Ramsay Silverman 2005 Springer

- Spline functions are widely-used as approximation system for **non-periodic functional data**
- Construction of a ***m-order spline***:
  - Divide the interval of definition *T* into *L* subintervals, i.e. fix a set of **knots**
  - Over each interval, the spline is defined as a polynomial or order *m (# of constant to define the polynomial)*
  - The polynomials are constrained as to guarantee that **adjacent polynomials join with continuity** in their values and in those of the derivatives up to order *m-2*



Ramsay Silverman 2005 Springer

- To gain **flexibility** in a spline one can increase the number of its knots, e.g., by locating more knots where the function exhibits more variability
- The **number of parameters** required to define a spline function with non-overlapping knots is the order plus the number of interior knots *m+L-1*



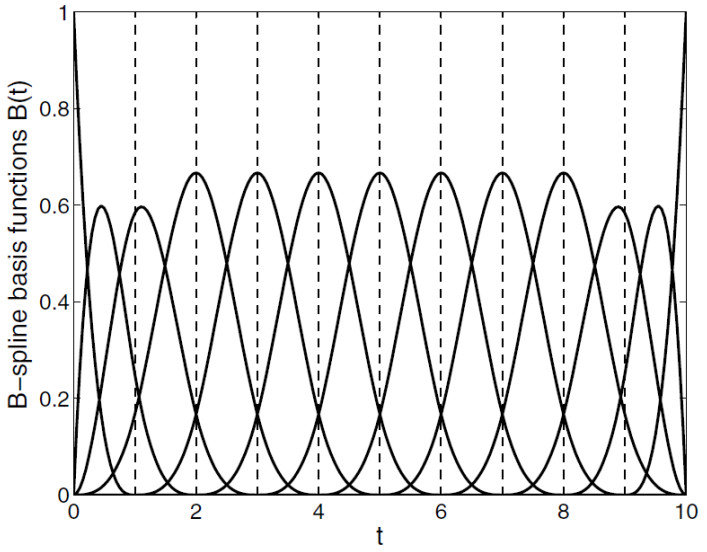
Ramsay Silverman 2005 Springer

- B-spline basis functions are systems of spline basis functions ϕ_k with the key properties:
  - Each basis function is a spline
  - A linear combination of the basis elements is a spline function
  - Any spline function can be expressed as a linear combination of these basis functions

- We call $B_k(t, \tau)$ a B-spline basis function in *t* with sequence of knots τ . A spline function is then defined as

$$S(t) = \sum_{k=1}^{m+L-1} c_k B_k(t, \tau)$$

- **Smoothing splines**: knots are placed at each argument value



Ramsay Silverman 2005 Springer

2.2. Least square smoothing

- We defined basis systems, that allows us to express a functional datum as a linear combinations of these basis elements

$$x(t) = \sum_{k=1}^{K} c_k \phi_k(t)$$

- Our next goal is to estimate the parameters c_k from the observed pairs (t_j, y_j) under the model

$$y_j = x(t_j) + \epsilon_j,$$

  or, in matrix form $x = \mathbf{c}'\boldsymbol{\phi} = \boldsymbol{\phi}'\mathbf{c}$.

Note 1. We can interpret this problem in the framework of classical linear models, and apply least square estimators.

Note 2. We smooth/interpolate one datum at a time, hence we here omit the index *i* of the statistical unit.

Goal: estimate the coefficient c_k of the linear model

$$x(t) = \sum_{k=1}^{K} c_k \phi_k(t) + \epsilon_j$$

from the pairs (t_j, y_j)

- **Solution 1**: We minimize the sum of squared errors between fitted values and observations:

$$\mathtt{SMSSE}(\mathbf{y}|\mathbf{c}) = \sum_{j=1}^{n} [y_j - \sum_{k}^{K} c_k \phi_k(t_j)]^2.$$

- From the theory of linear regression we know the solution of this problem

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

$$\hat{\mathbf{y}} = \mathbf{\Phi}\hat{\mathbf{c}} = \mathbf{\Phi}(\mathbf{\Phi}'\mathbf{\Phi})^{-1}\mathbf{\Phi}'\mathbf{y} \ .$$

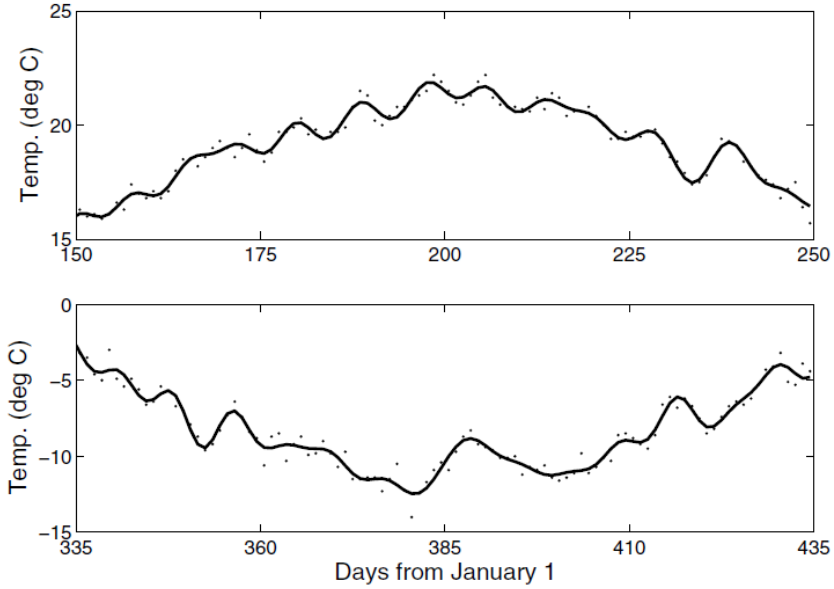
Note. $\hat{\mathbf{y}}$ is the projection of $\mathbf{y}$ over the space generated by the columns of $\mathbf{\Phi}$, that are the evaluations of the basis functions in the measurement points.

2.2. Least square smoothing

Ordinary least square fit

- Ordinary least squares are appropriate if the measurement error may be assumed to be iid
- The degree of smoothness of the estimated curve depends on the number of basis functions employed

Example. Smoothing of temperature data in Montreal using 109 Fourier basis functions



- We choose a Fourier basis because data are periodic
- We truncate the basis to 109 basis functions, that allows to catch (109-1)/2=54 different harmonic frequencies (about 1 per week)
- Performing OLS estimate means that the noise in the observations is iid across the days

Ramsay Silverman 2005 Springer

If data are not iid (e.g., there is autocorrelation in the measurement process), we can use a weighted least squares.

- **Solution 2**: We minimize the weighted sum of squared errors between fitted values and observations:

$$\mathrm{SMSSE}(\mathbf{y}|\mathbf{c}) = (\mathbf{y} - \mathbf{\Phi c})'\mathbf{W}(\mathbf{y} - \mathbf{\Phi c})$$

- Matrix $\mathbf{W}$ is assumed to be positive definite, and can be set e.g. to the covariance matrix of the errors $\mathbf{W} = \mathbf{\Sigma}_e^{-1}$.

- The solution of this minimization problem is found as

$$\hat{\mathbf{c}} = (\mathbf{\Phi}'\mathbf{W}\mathbf{\Phi})^{-1}\mathbf{\Phi}'\mathbf{W}\mathbf{y}$$

- Approximate point-wise confidence intervals can be built based upon the estimated model
- As in classical linear models, the variance of the estimator for the coefficients is

$$\mathrm{Var}[\mathbf{c}] = (\mathbf{\Phi}'\mathbf{W}\mathbf{\Phi})^{-1}\mathbf{\Phi}'\mathbf{W}\mathbf{\Sigma}_e\mathbf{W}\mathbf{\Phi}(\mathbf{\Phi}'\mathbf{W}\mathbf{\Phi})^{-1}$$

  which in case of unweighted least squares and iid errors reduces to

$$\mathrm{Var}[\mathbf{c}] = \sigma^2(\mathbf{\Phi}'\mathbf{\Phi})^{-1}$$

- The variance of the point-wise estimate of the curve is then obtained as the diagonal of the matrix

$$\mathrm{Var}[\hat{\mathbf{y}}] = \mathbf{\Phi}\mathrm{Var}[\mathbf{c}]\mathbf{\Phi}'$$

  which in case of unweighted least squares and iid errors reduces to

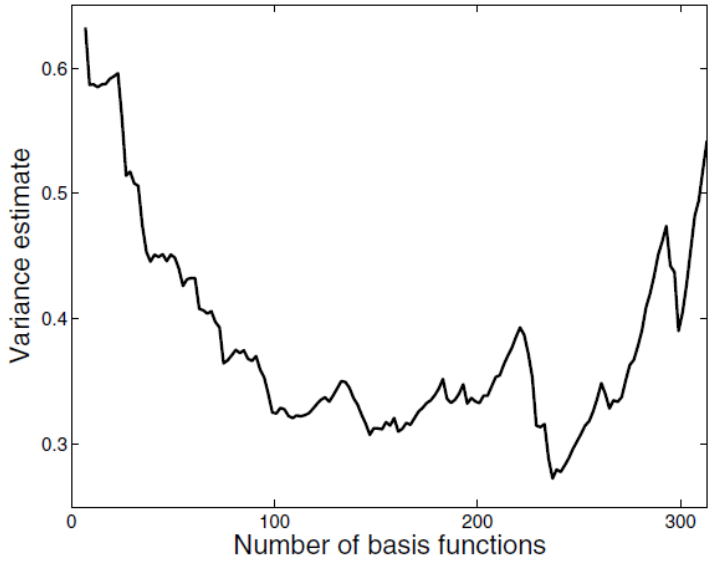
$$\mathrm{Var}[\hat{\mathbf{y}}] = \sigma^2\mathbf{\Phi}(\mathbf{\Phi}'\mathbf{\Phi})^{-1}\mathbf{\Phi}' = \sigma^2\mathbf{S}$$

- The variance of the errors can be estimated from the residual sum of squares

$$s^2 = \frac{1}{n-K}\sum_j^n (y_j - \hat{y}_j)^2$$

- To choose *K*, one may evaluate when a drop in sampling variance occurs for a range of candidate *K*

Example. Smoothing of temperature data in Montreal

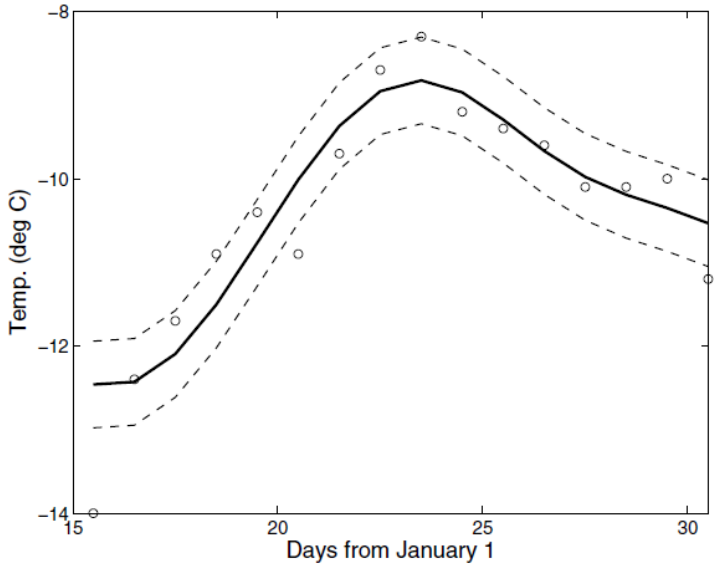


- A drop in variance is obtained around 100 basis functions
- We truncated the basis to 109 basis functions, that allowed to catch (109-1)/2=54 different harmonic frequencies (about 1 per week)
- Lower variances may be obtained but overfitting might occurs then

Ramsay Silverman 2005 Springer

- Based on previous expressions, approximate confidence limits may be built

Example. Smoothing of temperature data in Montreal



- Confidence limits are build summing/subtracting 2 standard deviations
- Quantiles of the normal can be used instead
- Confidence limits must be interpreted point-wise

Ramsay Silverman 2005 Springer

- A key point of least square smoothing is how to set the order of the basis expansion. Algorithms to set *K* can be borrowed from the context of linear regression (e.g., step-wise algorithms). Nevertheless, one should pay close attention to the fact that:
  - The larger *K*, the better the fit to the data, but higher risk to fit the noise (or non-interesting variations)
  - If *K* is too small we may miss important features of the underlying function that we wish to estimate

- In fact, as in linear regression, we have a bias/variance trade-off

$$\texttt{Bias}[\hat{x}(t)] = x(t) - \mathrm{E}[\hat{x}(t)],$$

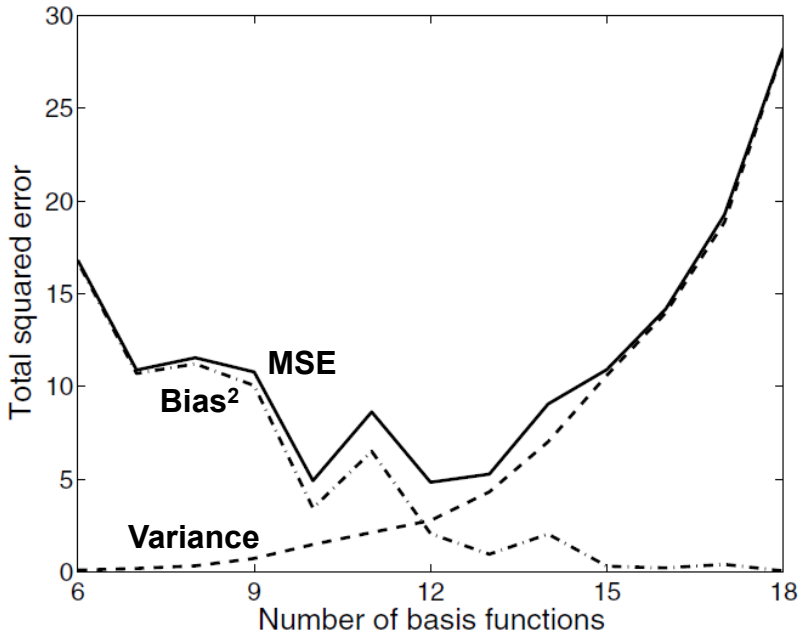
$$\texttt{Var}[\hat{x}(t)] = \mathrm{E}[\{\hat{x}(t) - \mathrm{E}[\hat{x}(t)]\}^2]$$

  - For large values of *K* the bias is small, the variance is high
  - For small values of *K*, the bias is high, the variance is low

- The mean-squared error summaries what we actually would like to minimize

$$\texttt{MSE}[\hat{x}(t)] = \mathrm{E}[\{\hat{x}(t) - x(t)\}^2] = \texttt{Bias}^2[\hat{x}(t)] + \texttt{Var}[\hat{x}(t)]$$

Example. Bias/variance trade of a simulated example inspired by the Berkeley Growth Study



Ramsay Silverman 2005 Springer

- We now focus on estimating a non-periodic function x on the basis of a vector $\mathbf{y}$ of discrete and noisy observations.
  Note: we are not (yet) assuming any functional form for x
- A way to approach the bias/variance trade-off is to impose a certain degree of smoothing on the curve (this reduces the variance at the expense of increasing bias)
- A popular way to do this is to quantify the notion of **roughness** through a **differential property of the curve**, and perform a **regression with** the corresponding **penalization**

Let us quantify *roughness* through the second derivative

$$\text{PEN}_2(x) = \int [D^2 x(s)]^2 \, ds$$

Measure of ***curvature*** of the function
(*$PEN_2(x)=0$* if *x* is a straight line)

Given λ, find x that minimizes

$$\text{PENSSE}_\lambda(x|\mathbf{y}) = [\mathbf{y} - x(\mathbf{t})]' \mathbf{W} [\mathbf{y} - x(\mathbf{t})]^2 + \lambda \times \text{PEN}_2(x)$$

Penalized SSE

- Let's give a closer look to the penalized SSE functional

$$\texttt{PENSSE}_\lambda(x|\mathbf{y}) = [\mathbf{y} - x(\mathbf{t})]'\mathbf{W}[\mathbf{y} - x(\mathbf{t})]^2 + \lambda \times \texttt{PEN}_2(x)$$

- Parameter λ is called *smoothing parameter* and controls the importance of the penalization with respect to the residual sum of squares:
  - If $\lambda \to \infty$ the functional gives emphasis to the penalization and the fitted curve will be a straight line ($\texttt{PEN}_2(x) = 0.$)
  - If $\lambda \to 0$ the curve approaches the smoothest twice-differentiable curve that interpolates the data

- **Key result** (de Boor, 2002): the curve x that minimizes $\texttt{PENSSE}_\lambda(x|\mathbf{y})$ is a cubic spline with knots at the data points t_j

  $\to$ *the functional form of* x *is a consequence of the objective function!*

- **Common computational technique:** use a four order B-spline basis (called *cubic spline*) and minimize the $\texttt{PENSSE}_\lambda(x|\mathbf{y})$ with respect to the coefficients of the expansion

$$x(t) = \sum_{k=1}^{K} c_k \phi_k(t)$$

λ set via GCV.

- We can re-express the penalization as

$$\begin{aligned} \texttt{PEN}_m(x) &= \int [D^m x(s)]^2 \, ds \\ &= \int [D^m \mathbf{c}'\phi(s)]^2 \, ds \\ &= \mathbf{c}'\mathbf{R}c \ , \end{aligned}$$

with $\mathbf{R} = \int D^m \phi(s) D^m \phi'(s) \, ds \ .$

- Plugging this expression in the objective function yields

$$\texttt{PENSSE}_m(\mathbf{y}|\mathbf{c}) = (\mathbf{y} - \mathbf{\Phi c})'\mathbf{W}(\mathbf{y} - \mathbf{\Phi c}) + \lambda \mathbf{c}'\mathbf{R}\mathbf{c} \ .$$

that is minimized for

$$\hat{\mathbf{c}} = (\mathbf{\Phi}'\mathbf{W}\mathbf{\Phi} + \lambda \mathbf{R})^{-1}\mathbf{\Phi}'\mathbf{W}\mathbf{y} \ .$$

Generalized cross-validation

$$\texttt{GCV}(\lambda) = \frac{n^{-1}\,\texttt{SSE}}{[n^{-1}\text{trace}\,(\mathbf{I} - \mathbf{S}_{\phi,\lambda})]^2}$$

with

$$\hat{\mathbf{y}} = \mathbf{\Phi}(\mathbf{\Phi}'\mathbf{W}\mathbf{\Phi} + \lambda\mathbf{R})^{-1}\mathbf{\Phi}'\mathbf{W}\mathbf{y} = \mathbf{S}_{\phi,\lambda}\mathbf{y}$$

2.3 Closing remarks

- We have seen basis expansions as a way to smooth raw functional observations
- Many other bases and smoothing techniques are available in the literature, e.g.,
  - Local polynomial smoothing: LS smoothing on neighborhoods of the point t_j through polynomial basis

$$\mathtt{SMSSE}_t(\mathbf{y}|\mathbf{c}) = \sum_{j=1}^{n} w_j(t)[y_j - \sum_{k=1}^{K} c_k \phi_k(t_j)]^2,$$

$$w_\ell(t) = \mathtt{Kern}\left(\frac{t_\ell - t_j}{h}\right)$$

Uniform:	$\mathtt{Kern}(u) = 0.5$ for $\lvert u\rvert \leq 1$,	0 otherwise
Quadratic:	$\mathtt{Kern}(u) = 0.75(1-u^2)$ for $\lvert u\rvert \leq 1$,	0 otherwise
Gaussian:	$\mathtt{Kern}(u) = (2\pi)^{-1/2}\exp(-u^2/2)$.	

2.3 Closing remarks

- We have seen basis expansions as a way to smooth raw functional observations
- Many other bases and smoothing techniques are available in the literature, e.g.,
  - Local polynomial smoothing: LS smoothing on neighborhoods of the point t_j through polynomial basis
  - Wavelet bases
- Ad hoc smoothing techniques need to be employed in case of constrained data, e.g.,
  - Monotonically increasing functions
  - Probability density functions
- Smoothing or interpolation is the very first step of a functional data analysis and all the subsequent results depend on this step.
  → One should pay close attention in applying the most appropriate technique for smoothing the data