POLITECNICO DI MILANO

Politecnico di Milano

Applied Statistics

May 2023

An introduction to functional data analysis

Laura M. SANGALLI

MOX - Dipartimento di Matematica, Politecnico di Milano

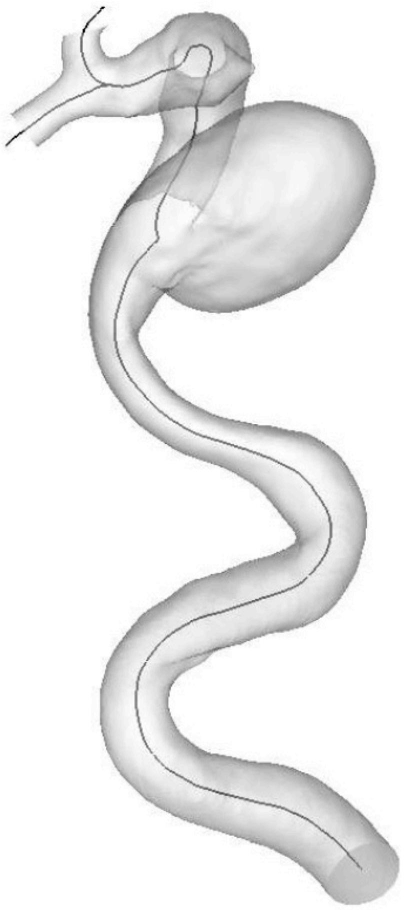
Part 1 - Introduction

Functional data: where they come from

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

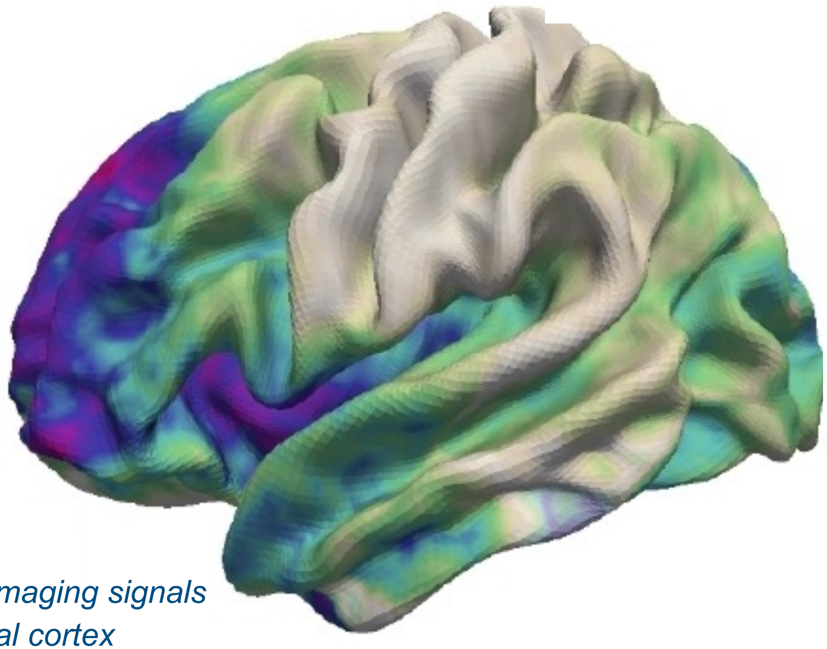
2D and 3D images and measures captured in time and space

- images of the internal structures of a body provided by diagnostic medical scanners



Study of the morphology of inner carotid arteries with aneurysms

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

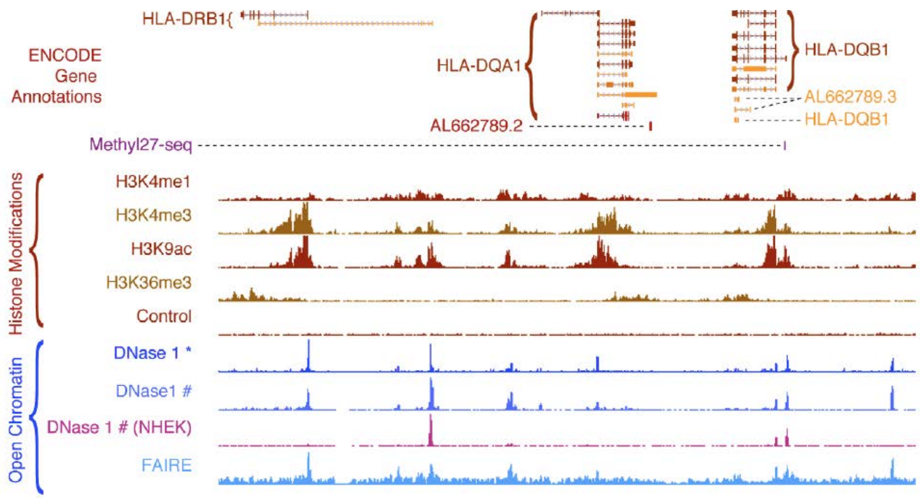


Study of neuroimaging signals over the cerebral cortex

Lila, Aston, Sangalli (2016) AoAS

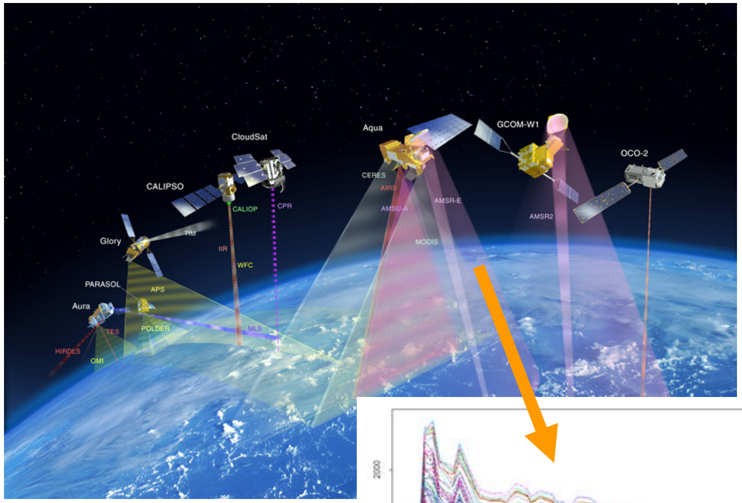
Functional data: where they come from

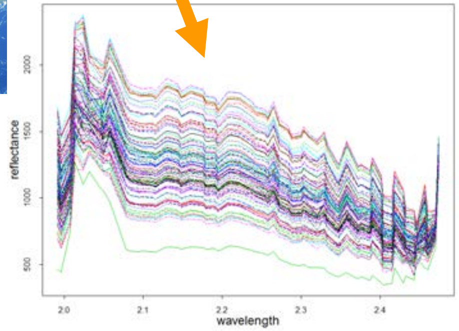
- measurements of gene expression levels via next generation sequencing data



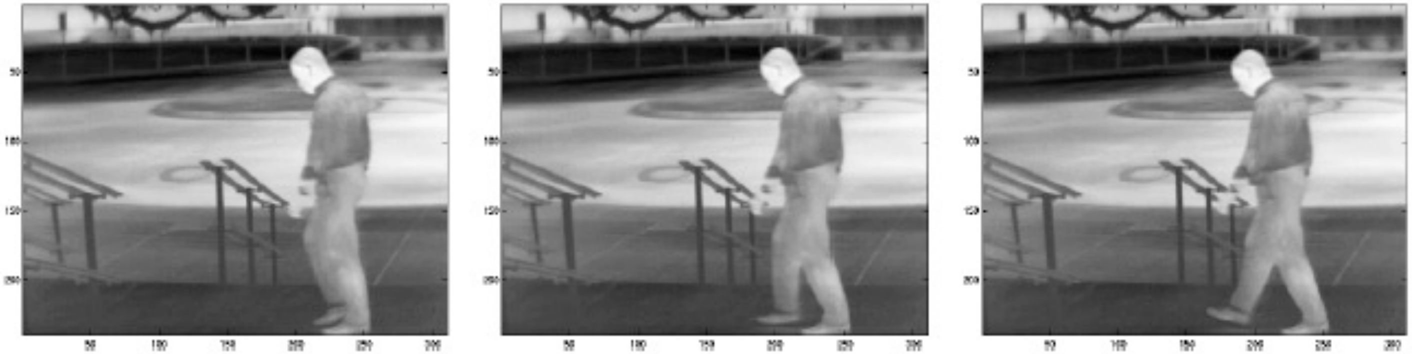
Cremona et al. (2015) BMC Bioinformatics

- multi-spectral data from satellite remote sensing

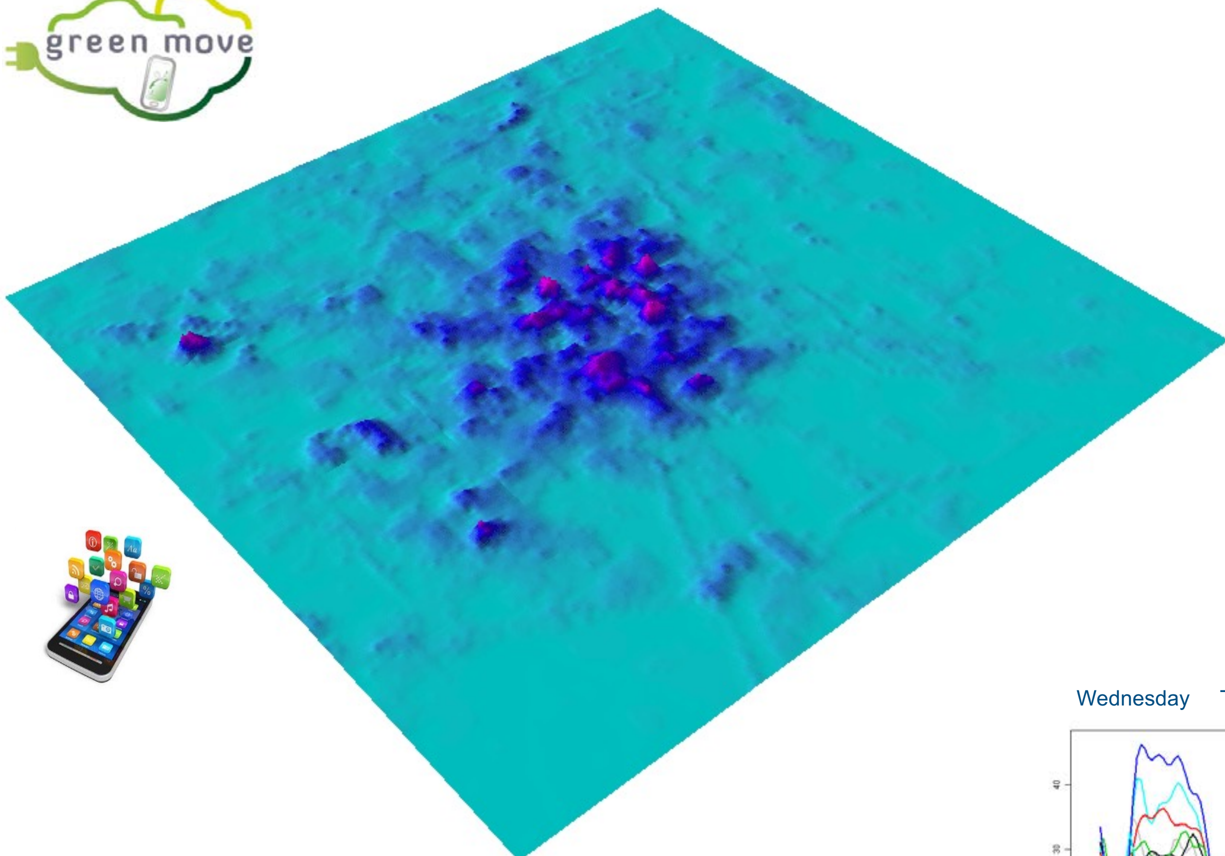


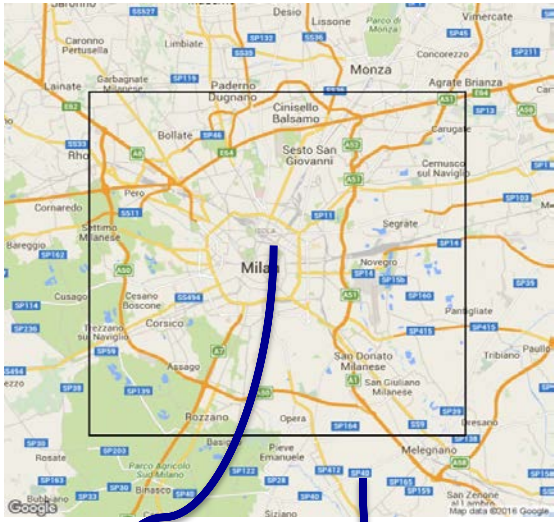


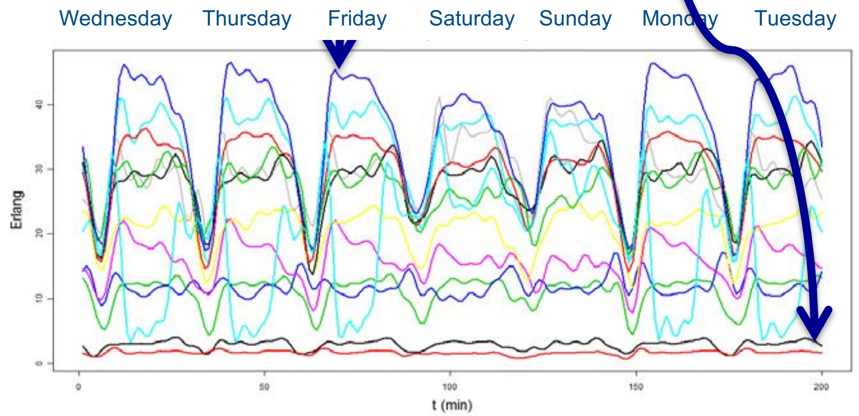
- images of steady or moving objects/individuals recorded by computer vision devices



Functional data: where they come from

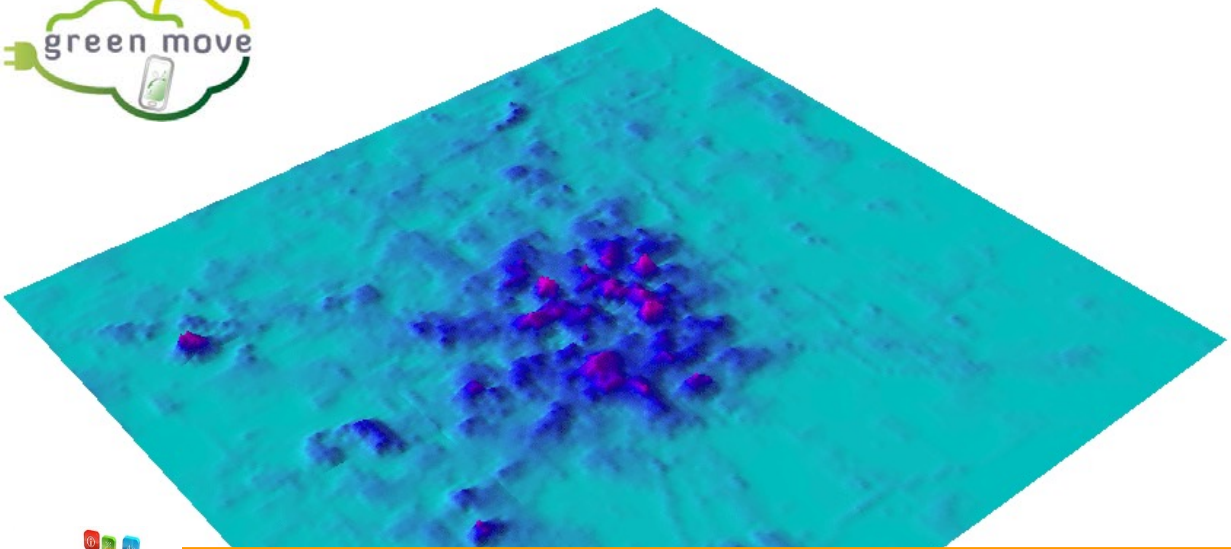


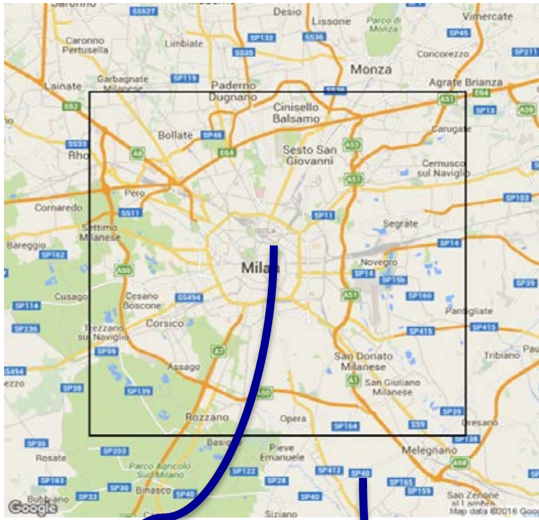




Secchi, Vantini, Vitelli (2015), SMA

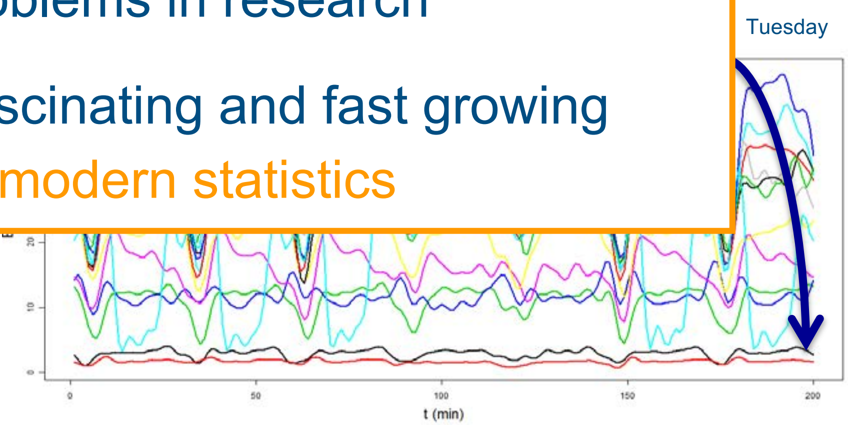
Functional data: where they come from





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

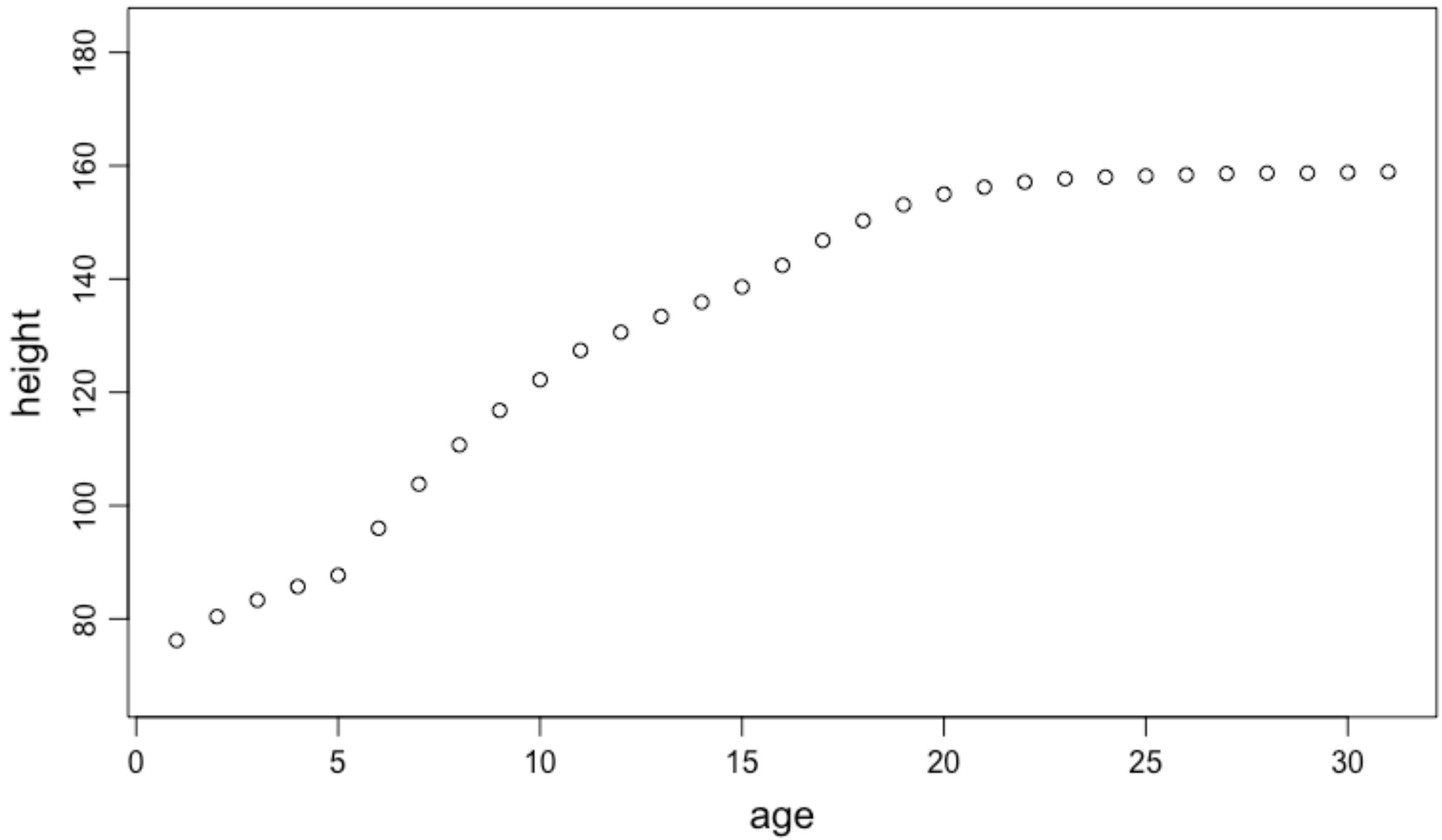
It is fueling one of the most fascinating and fast growing research fields of modern statistics



Secchi, Vantini, Vitelli (2015), SMA

What characterizes functional data?

Smoothness

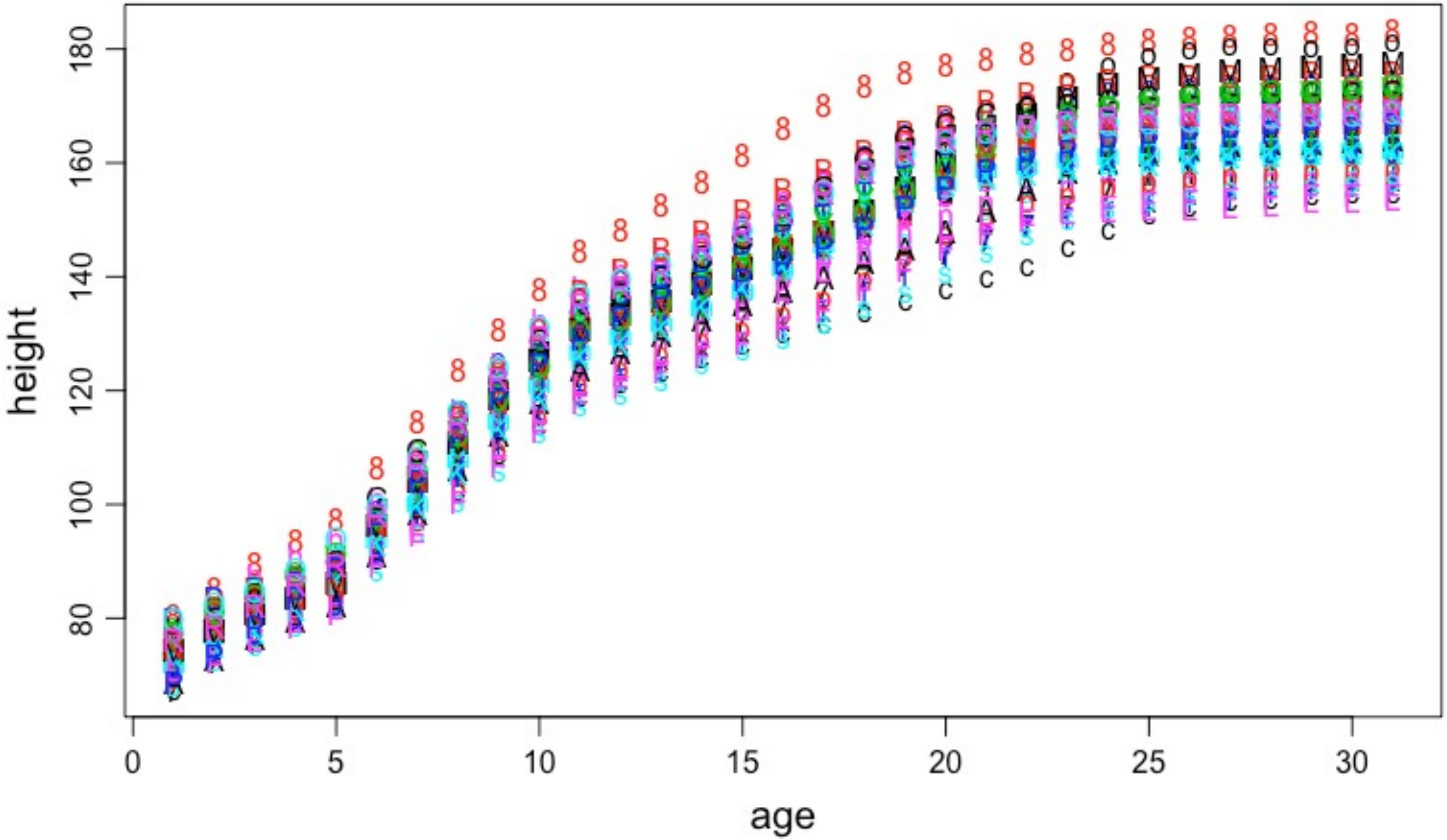


Ramsay Silverman 2005 Springer

Berkley Growth Study, Girl 1

What characterizes functional data?

Smoothness



Ramsay Silverman 2005 Springer

Berkley Growth Study, all girls

What characterizes functional data?

- Informally, **functional data** are entities that can be described through a function, e.g., a curve, a surface, an image
- A **functional dataset** consists of a sample of functional observations
- Even though observations are actually discrete and affected by noise, 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

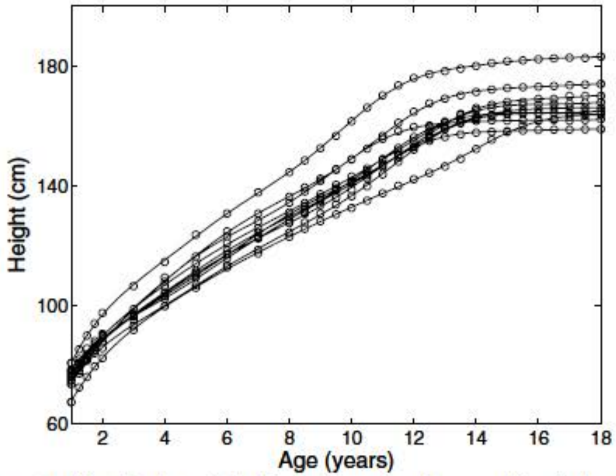


Figure 1.1. The heights of 10 girls measured at 31 ages. The circles indicate the unequally spaced ages of measurement.

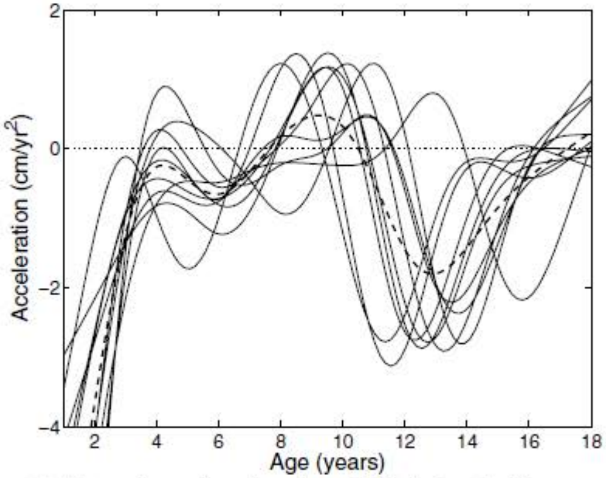


Figure 1.2. The estimated accelerations of height for 10 girls, measured in centimeters per year. The heavy dashed line is the cross-sectional mean, and is a rather poor summary of the curves.

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)

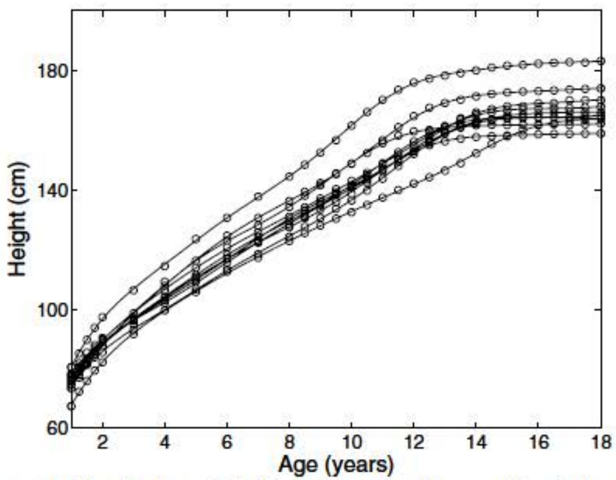


Figure 1.1. The heights of 10 girls measured at 31 ages. The circles indicate the unequally spaced ages of measurement.

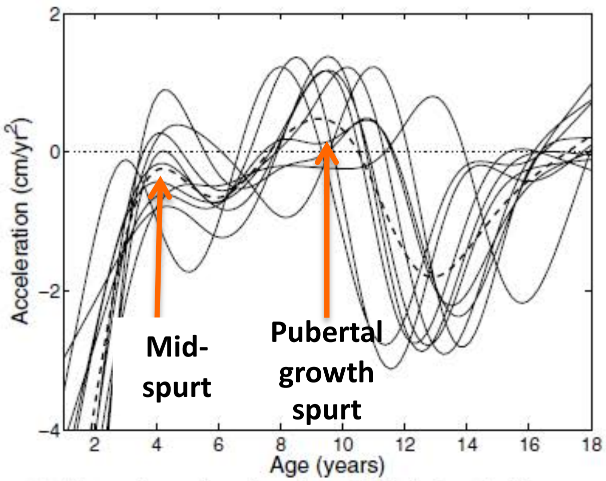


Figure 1.2. The estimated accelerations of height for 10 girls, measured in centimeters per year. The heavy dashed line is the cross-sectional mean, and is a rather poor summary of the curves.

Ramsay Silverman 2005 Springer

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.
- Kokoszka P. and Reimherr, M. (2017). Introduction to Functional Data Analysis. Chapman & Hall

Introductory paper:

- Sørensen, H., Goldsmith, J., Sangalli, L.M. (2013), "An introduction with medical applications to functional data analysis". Statistics in Medicine, 32, pp. 5222–5240.

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 fdaCluster (alignment and clustering)
- R package fdaPDE (functional data over complex multidimensional domains)

Basics notions on Hilbert spaces: a reminder

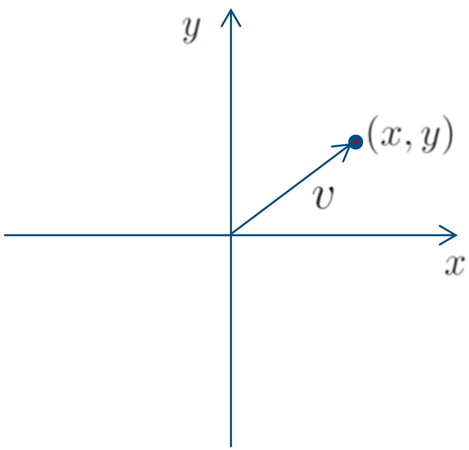
A Hilbert Space approach to the analysis of Functional Data

Courtesy of P. Secchi

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: $\vartheta = \arccos \dfrac{\langle v_1, v_2 \rangle}{\|v_1\| \|v_2\|}$

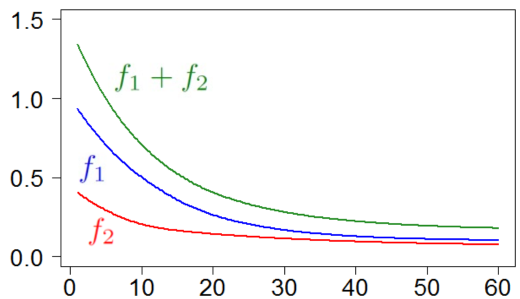
Inner product

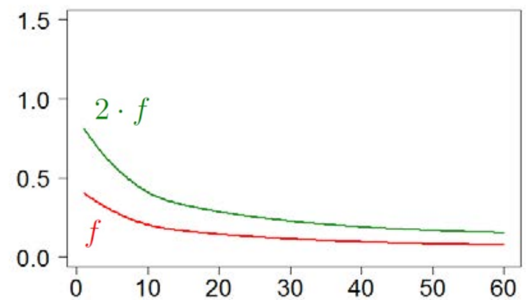
$\langle v_1, v_2 \rangle = (x_1 \cdot x_2) + (y_1 \cdot y_2)$

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: $\vartheta = \arccos \dfrac{\langle f_1, f_2 \rangle}{\|f_1\| \|f_2\|}$

Inner product

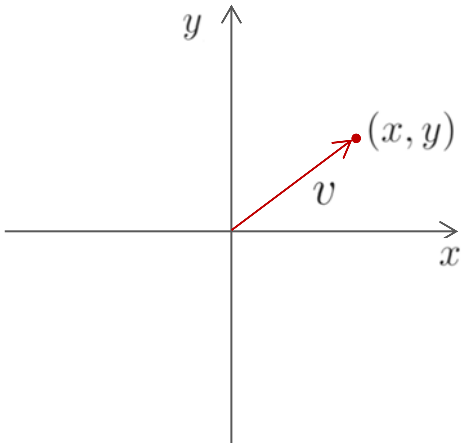
$$\langle f_1, f_2 \rangle = \int (f_1(t) \cdot f_2(t)) dt$$

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

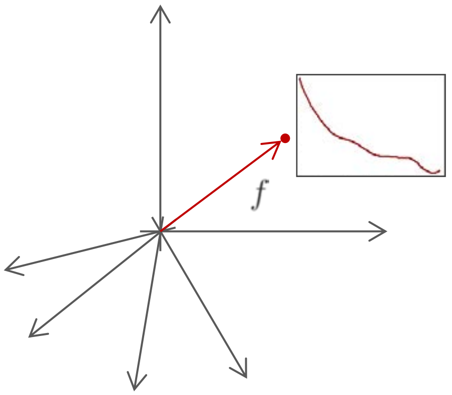
Embedding functional data in un appropriate Hilbert space enable us

- to understand functional data as **points of a space of functions**
- to uplift many methods of **multivariate statistics to functional data**, through the notions of inner product and norm

Multivariate statistics
(Euclidean space)



Functional Data Analysis
(Hilbert space)



Basics notions on Hilbert spaces: a reminder

Courtesy of P. Secchi

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

$$\begin{aligned}
&\text{(i)} \ \langle \lambda x + y, z \rangle = \lambda \langle x, z \rangle + \langle y, z \rangle \qquad \forall \lambda \in \mathbb{R}, \ \ \forall x, y, z \in H \\
&\text{(ii)} \ \langle x, y \rangle = \langle y, x \rangle \qquad \forall x, y \in H \\
&\text{(iii)} \ \langle x, x \rangle \geq 0 \qquad \forall x \in H \\
&\text{(iv)} \ \langle x, x \rangle = 0 \quad \iff \quad x = 0
\end{aligned}$$

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:

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

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

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

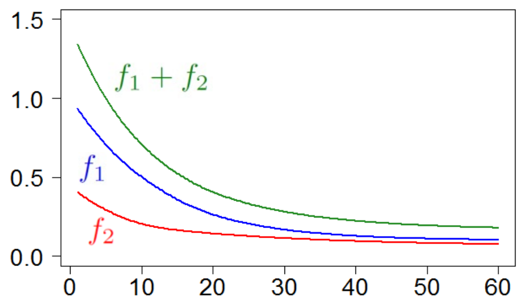
Basis expansion

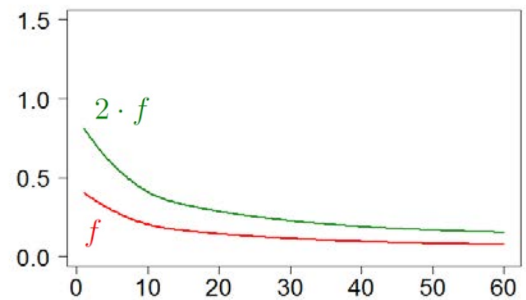
Courtesy of P. Secchi

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 $(+, \cdot)$





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

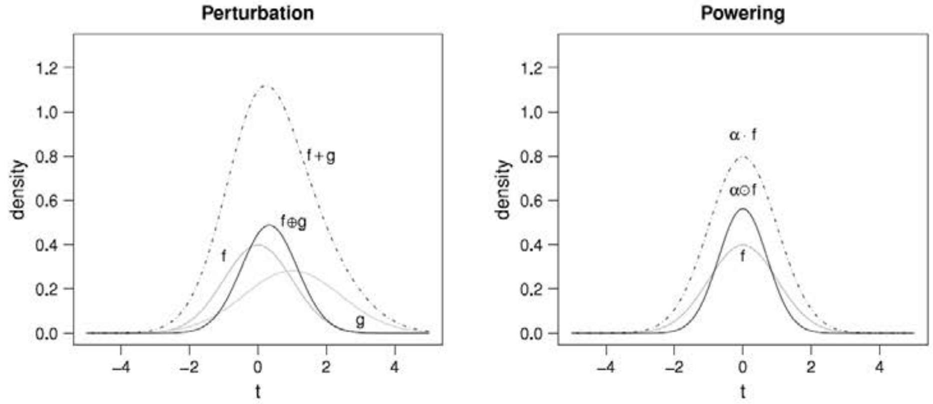
Inner product

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

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

B²: space of density functions on a closed 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 doesn't make sense for density functions

B^2: space of density functions on a closed 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**

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

Exercise: prove that

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

Hilbert space embedding for functional data

Courtesy of P. Secchi

- First step in fda:

  choose appropriate 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)
  - H^2, Sobolev space of functions that L^2 and whose derivative up to the second order are also in L^2
  - B^2, space of functional compositions: useful for density functions

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 defined on a probability space $(\Omega, \mathfrak{F}, \mathbb{P})$ with values in H $X : \Omega \rightarrow H$

Definition 2:
A **functional datum** x is a realization of a functional random variable, i.e., for $\omega \in \Omega$,

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

Definition 3:
A **functional dataset** is a collection of functional data.

Let $X : \Omega \to H$ be a functional random variable in *H.*

We assume $\mathbb{E}[\|X\|_H^4] < \infty$

Definition 4:
We call Fréchet mean of X the (unique) element μ of *H* that solves

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

- If *H=L*^{2} the Fréchet mean coincides a.e. with the point-wise mean

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

- In any *H*, we 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 $H=L^2$ 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

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

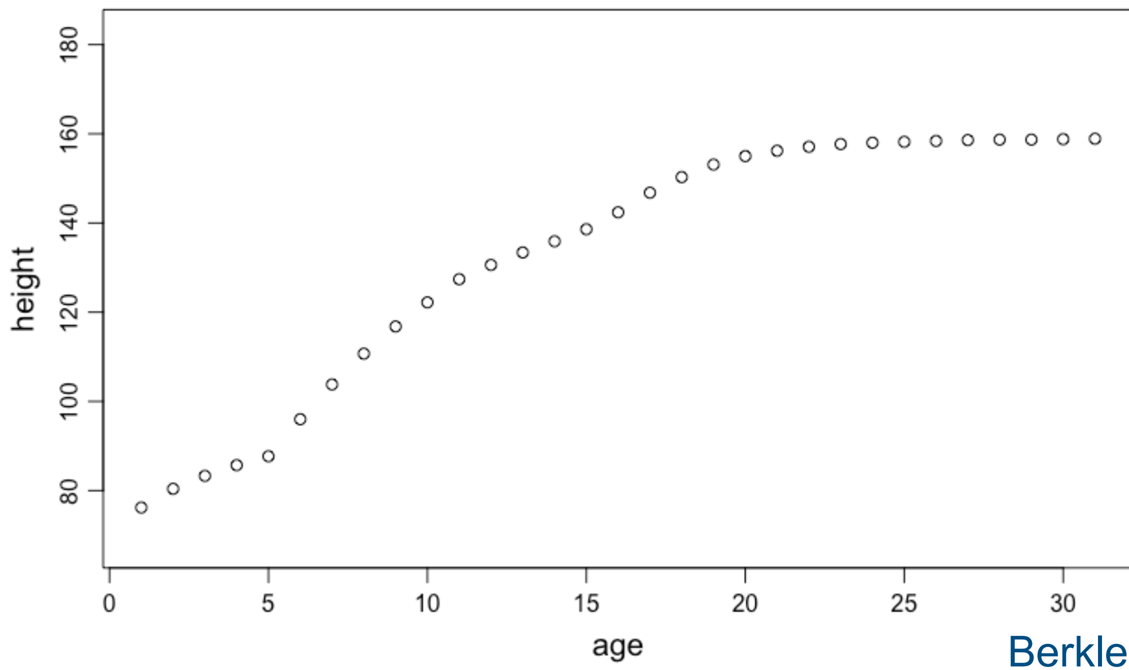
- In any *H*, the covariance operator can be estimated through the sample covariance operator

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

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

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

Smoothing



Berkley Growth Study, Girl 1

Noisy and discrete data $\rightarrow$ functional representations

Smoothing - curve fitting

Chapters 3, 4, 5, 6 of Ramsay and Silverman (2005), *Functional Data Analysis*, Springer