Statistical methods of data science

An introduction to Functional Data Analysis

Alessandra Menafoglio[1*]

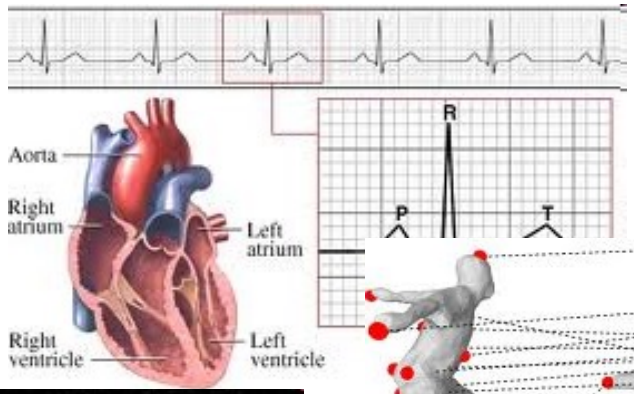
[1]MOX, Department of Mathematics, Politecnico di Milano

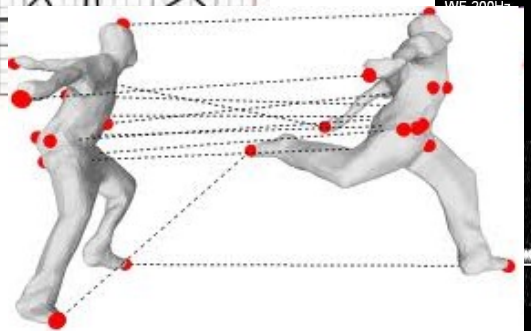
*alessandra.menafoglio@polimi.it

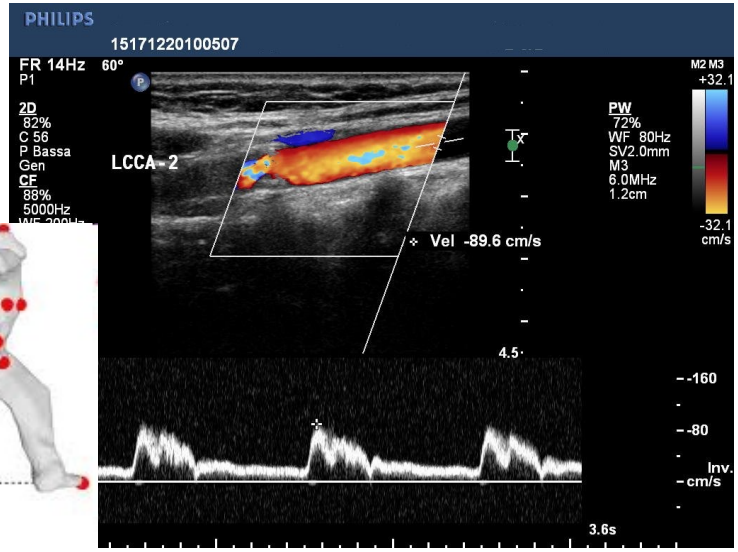
Introduction

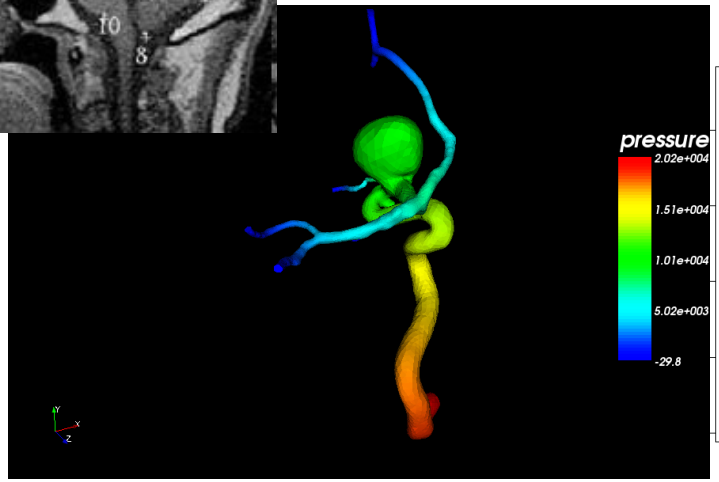
The Data Deluge Era

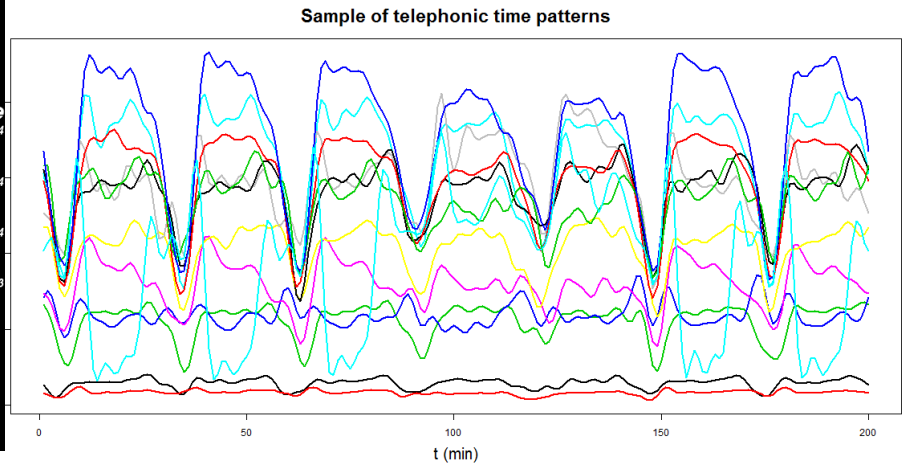
Big Data as complex observations











Big Data

Big Data ≠ Big Information

About us

Politecnico di Milano

Department of Mathematics «F. Brioschi»

MOX – Laboratory for Modeling and Scientific Computing

Statistics Group

mox.polimi.it/research-areas/statistics/

The research activity of the Statistics group at MOX is focused on the development of statistical models and methods for high dimensional and complex data, driven by industrial problems and problems arising in the life, environmental and social sciences.

Keywords:
Statistical Modelling of Complex and High Dimensional Data, ***Functional Data Analysis****, Geostatistics, Nonparametrics, Penalized Regression, Compositional Data, Big data, Data Mining, Health Analytics.*

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

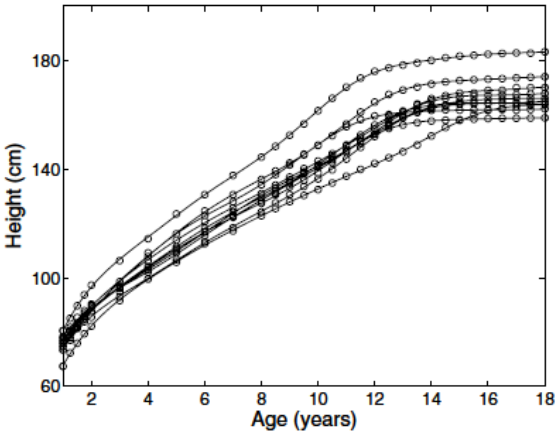


Figure 1.1. The heights of 10 girls measured at 31 ages. The circles indicate the unequally spaced ages of measurement.

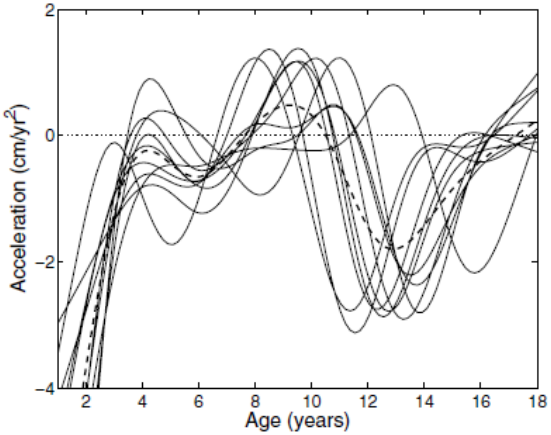
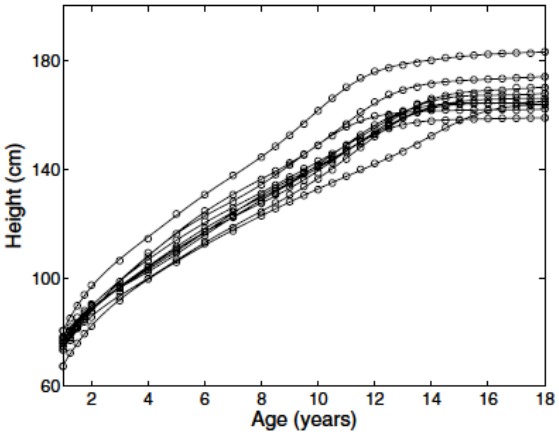


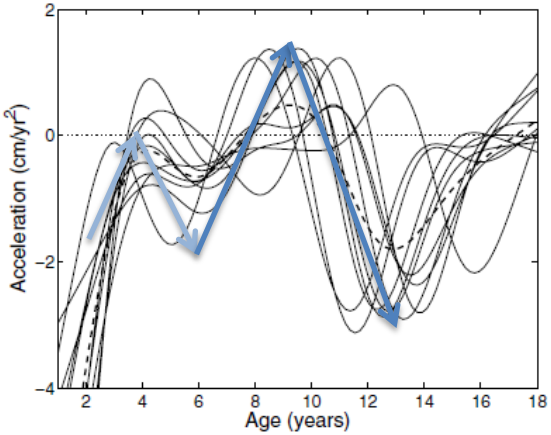
Figure 1.2. The estimated accelerations of height for 10 girls, measured in centimeters per year. The heavy dashed line is the cross-sectional mean, and is a rather poor summary of the curves.

Taken from Ramsay & Silverman (2002)

Berkeley Growth Curves as functional data

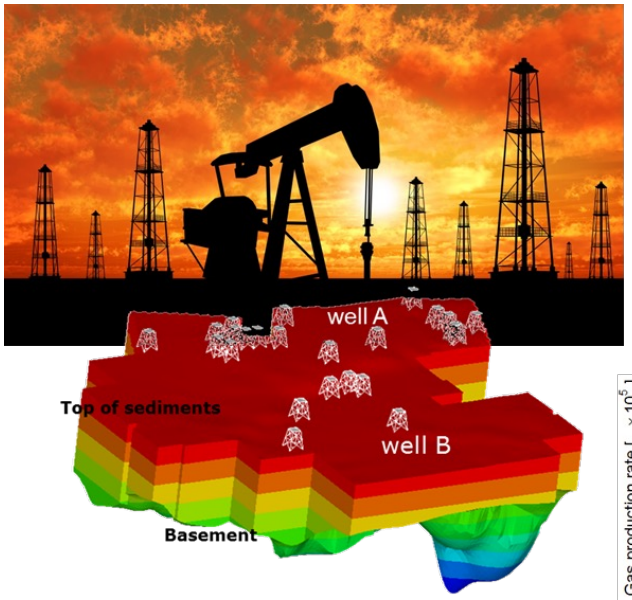
- Data reflect **smooth** variation of height over time: *h(t)*
- Some interesting features are only visible if **derivatives** are analyzed (e.g., mid-spurt and pubertal growth spurt)
- The grid spacing on the **time axis** is non-uniform. The underlying function might have been observed on different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

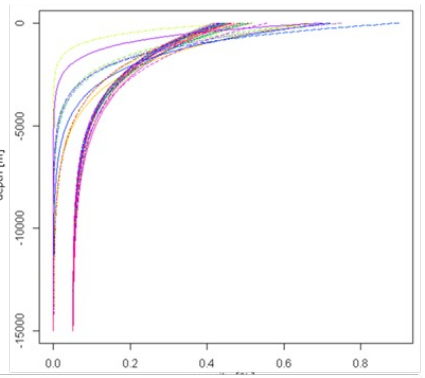


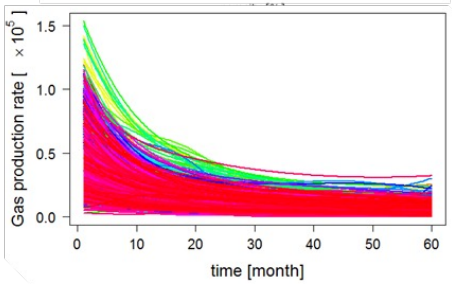


More examples of functional data

Example: Oil & gas industry
Data on sediment compaction curves and gas rate production curves





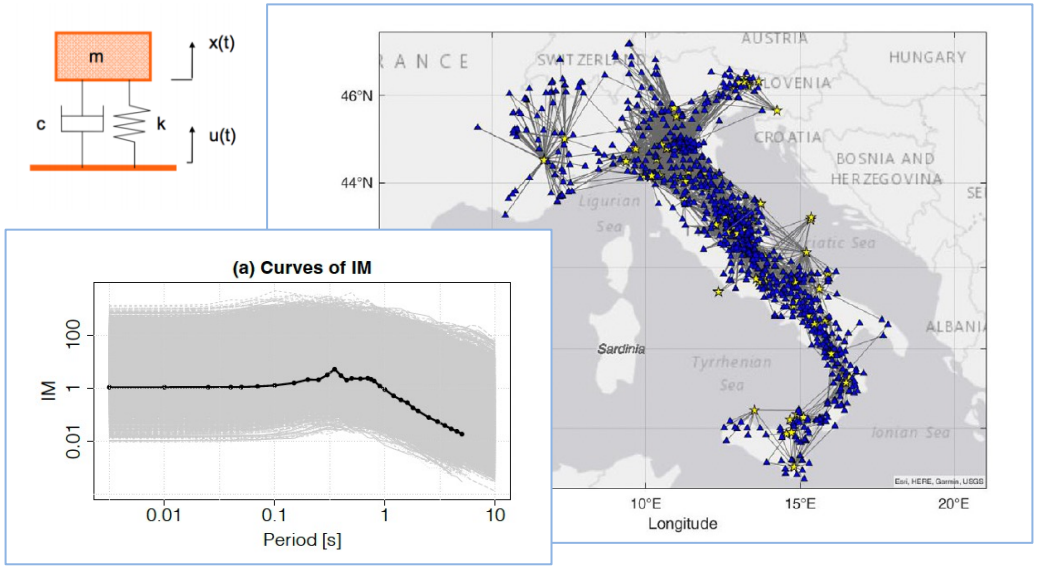


More examples of functional data

Example: Ground Motion Models on functional intensity measures in earthquakes events

Intensity Measure:

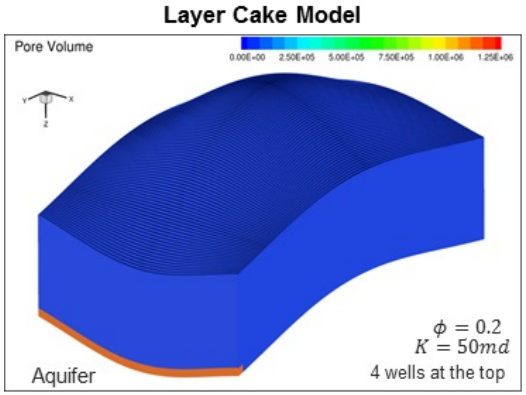
Acceleration response spectrum, i.e., Maximum of the acceleration of the oscillator in a single degree-of-freedom system with natural period T_n



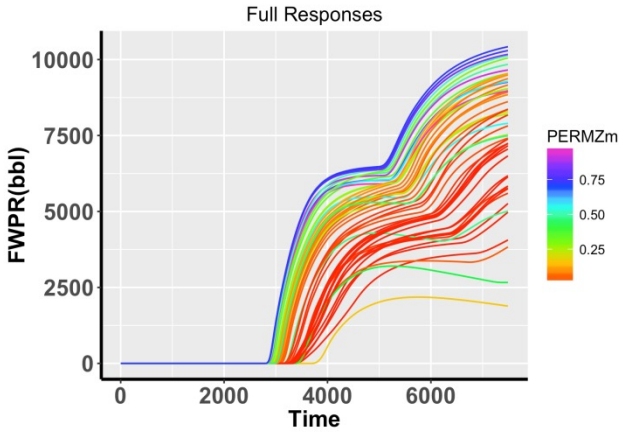
Taken from Bortoltti et al. (JASA, 2024)

More examples of functional data

Example: Response of a numerical model of fluid flow in a reservoir



Simulation of fluid flow in subsurface

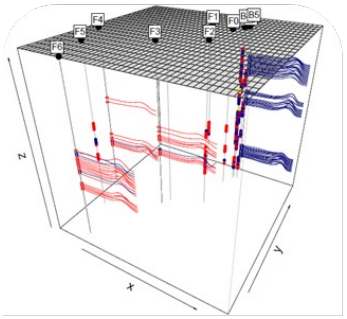


Field Water Production Rate

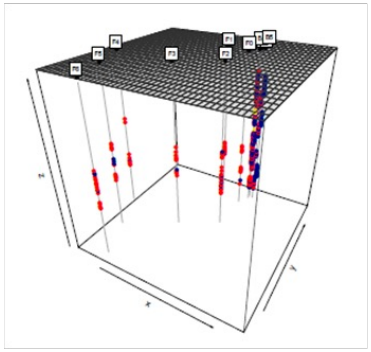
The limit of classical approaches

- Classical approaches would advocate the reduction of the data to simple indicators, which can be analyzed with classical multivariate methods.
- This inevitably yields an information loss.

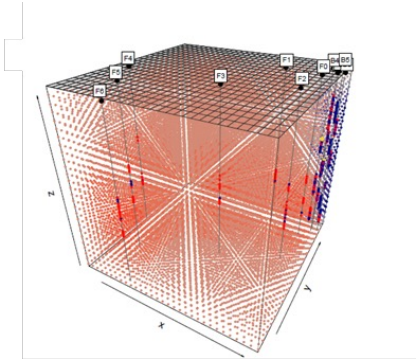
Conceptual example: spatial prediction of functional data through data reduction



Dataset of complex objects



Reduced dataset

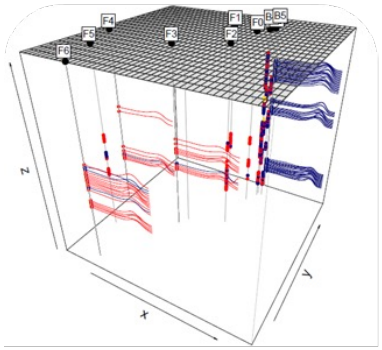


Prediction

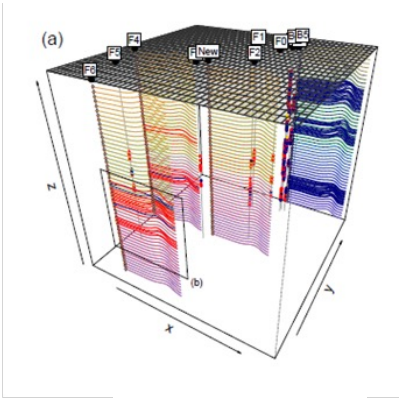
The limit of classical approaches

- Classical approaches would advocate the reduction of the data to simple indicators, which can be analyzed with classical multivariate methods.
- This inevitably yields an information loss.
- In Functional Data Analysis, the **"atom"** of the statistical analysis is the entire function, rather than a limited number of selected features of the data. This potentially allows to **exploit the entire information content embedded within the data**

Conceptual example: spatial prediction of functional data through FDA



Dataset of complex objects



FDA Prediction

The limit of classical approaches

- Classical approaches would advocate the reduction of the data to simple indicators, which can be analyzed with classical multivariate methods.
- This inevitably yields an information loss.
- In Functional Data Analysis, the **"atom"** of the statistical analysis is the entire function, rather than a limited number of selected features of the data. This potentially allows to **exploit the entire information content embedded within the data**

Functional Data Analysis is concerned with the statistical analysis of functional data. FDA offers a toolkit enabling for a varied range of analyses, including

- Exploratory Data Analysis: Data representation, visualization, outlier detection
- Dimensionality reduction
- Statistical inference (especially testing)
- Supervised & unsupervised classification
- Regression
- Spatial statistics

About the course

Statistical methods of data science

The course offers an introduction to Functional Data Analysis (FDA), the area of statistics focusing on modeling and analyzing complex observations like curves, surfaces, or images. It covers classical and advanced FDA methodologies, with emphasis on the key topics of smoothing, dimensionality reduction, anomaly detection, and spatial analysis. The course integrates theoretical sessions and the analysis of case studies, demonstrating concepts through real data examples.

Calendar

- Thu 11/04, h. 10:00-12:00, 13:00-15:00 (4 hrs), Room B-101
- Tue 23/04, h. 10:00-12:00, 13:00-15:00 (4 hrs), Room C3.02
- Tue 30/04, h. 10:00-12:00, 13:00-15:00 (4 hrs), Room C3.02
- Tue 7/5, h. 10:00-13:00 (3 hrs), Room C3.02?

Material available at: https://github.com/AMenafoglio/PhD-FDA_2024

About the course

Course Agenda

0. Introduction
1. Hilbert space model for functional data
2. Smoothing and interpolation of functional data
3. FDA & Dimensionality reduction in Hilbert spaces
4. Linear models
5. Anomaly detection through control chart schemes
6. Spatial statistics for functional data

Main References*

- Horvath, L. and Kokoszka, P. (2012). Inference for Functional Data with Applications. Springer Series in Statistics. Springer.
- Hron K, Menafoglio A, Templ M, Hruzova K, Filzmoser P (2015) Simplicial principal component analysis for density functions in Bayes spaces. Comput Stat Data Anal. doi:10.1016/j.csda.2015.07.007
- Ramsay, J. and Silverman, B. (2005). Functional data analysis (Second ed.). Springer, New York
- Ramsay J. O.; Wickham H.; Graves S.; Hooker G. (2010). fda: Functional Data Analysis. R package version 2.2.5.

**Additional references on specific topics will be given within the course material*