STATISTICS PILLAR

MASTER IN BIG DATA IN BUSINESS

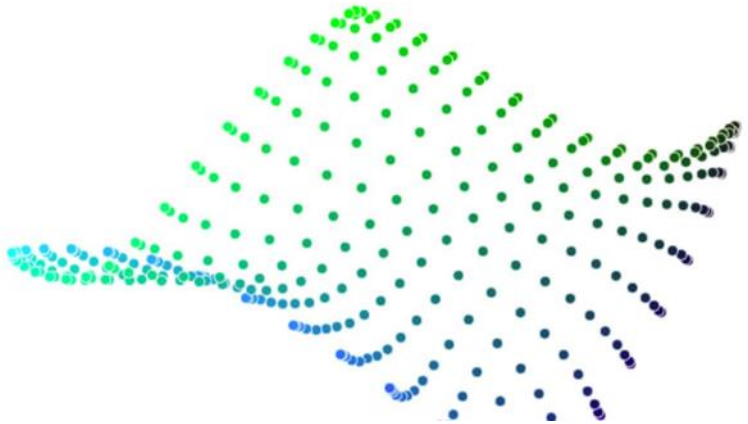
TOR VERGATA
UNIVERSITÀ DEGLI STUDI DI ROMA

CONTENTS

- This pillar focuses on the statistical analysis of High-Dimensional Data (HDD), a framework where the number of variables is larger than the number of observations.
- The goal of the pillar is to endow students with concepts and methods in both supervised and unsupervised statistical learning and in the analysis of large dimensional dynamic systems.

WHAT IS HDD ANALYSIS?

Nowadays HDD are pandemic in almost of any branch of knowledge, including basic sciences, biology medicine, business, engineering, economics, finance, medicine, etc. For instance, HDD are used to assess the effectiveness of new therapies, to monitor biological and climate phenomena, to optimise processes in industry and in administrations, to analyse consumer behaviour, to forecast macroeconomic and financial variables, etc.



COURSES

Course name	Professors' names	Scientific Disciplinary Sector (SDS)	Theoretical classes	Practicalclasses	ECTS credits	Teaching Term
Supervised learning	Prof. Cubadda, Prof. Parisi	Economic Statistics	36	18	6	I term
Unsupervised learning	Prof. A. Farcomeni, Prof.	Statistics	36	18	6	I term
High Dimensional Time Series	Prof. S. Grassi	Economic Statistics	18	9	3	II term

SUPERVISED LEARNING

The course provides an introduction to supervised learning, focusing on both regression and classification problems. Empirical applications will be illustrated using updated software tools.

LEARNING OUTCOMES:

- Understand regression models.
- Understand supervised classification methods.
- Gain practical experience in predicting a target variable using many predictors.

METHODOLOGY: Theoretical lessons and practice using R and Matlab.

SUPERVISED LEARNING - TEACHERS

- Gianluca Cubadda, Full Professor of Economic Statistics and Dean of the Faculty of Economics, University of Rome "Tor Vergata"
- Antonio Parisi, Researcher in Economic Statistics and teacher of the courses in *statistical computing, MATLAB, and statistics for economic applications*.
- Andrea De Mauro, Regional Leader of Analytics at Procter & Gamble (P&G).

STATISTICAL TOOLS

- The lessons will take place in room S1, at the Faculty of Economics of the University of Rome "Tor Vergata".
- The classroom is equipped with all the necessary requirements, on the PCs are installed specific software necessary for exercise hours of each course (lab).
- Software used during the modules of the master include: MatLab, R, Python, Hadoop, Java, Multichain, C, among others.

UNSUPERVISED LEARNING

The course covers the main statistical techniques used to find groups in the data (i.e., identify discrete structures not directly observed) and, even when there is only one group, outliers.

Furthermore, it discusses dimensionality reduction methods used to summarize data in few dimensions, create rankings/scores, compress information. Principles of robust estimation are also introduced.

As an example, consider a company and its customer database where for each customer (unit) a number of characteristics (variables) linked with customer behaviour are measured: number of visits, total amount spent, overall approval of services, etc. Unsupervised learning techniques can help us find answers to questions like: are there different types of customers? If yes, how many and what are their profiles? Are there few very unusual customers?

Dimensionality reduction techniques can help us find answers to questions like: how can we rank customers with respect to propensity of making business with us? How can we plot the data summarizing all of the information? What kind of information is available in the data available?

UNSUPERVISED LEARNING (2)

LEARNING OUTCOMES

- ability to use statistical learning techniques in the presence of unmeasured data labels
- ability to perform anomaly detection at basic level
- ability to evaluate and compare the resulting grouping structure
- ability to reduce data dimensionality for descriptive and scoring purposes

METHODOLOGY

Emphasis is on principles and specific models/techniques. Each method is introduced by examples and described in mathematical formulas. Some math is essential but very few derivations are made. Models and techniques are discussed from a theoretical and practical point of view, describing their definition/properties and their implementation by using statistical software R. Particular importance is given to the interpretation of results. Computer laboratory hours give to the students the possibility to practice what they learn.

UNSUPERVISED LEARNING (3)

PROJECT

Group assignments aim at assessing the capabilities of analyzing data, as well as the ability to communicate the relevant findings. The students are expected to produce a technical report no longer than 6 pages.

HIGH DIMENSIONAL TIME SERIES

Introduction to multivariate time series analysis.

The course covers the basic aspects of multivariate time series analysis, with the focus on modeling and forecasting a large set of variables. The topic in the course will be:

- Vector autoregressions;
- Estimation of Large VAR models using Bayesian VAR and Compressed VAR;
- Estimation of High-dimensional covariance matrices with applications in finance;
- Shrinkage estimators;
- Factor models and their applications.

Empirical applications will be illustrated using updated software tools.

HIGH DIMENSIONAL TIME SERIES (2)

LEARNING OUTCOMES

- Study the basic theory of multivariate processes.
- Learn Vector Auto-Regressive (VAR) models and their dimensionality reduction and data compression
- Learn large volatility models, Factor Models, and their applications.

METHODOLOGY: theoretical lessons and classes