High Dimensional Statistics and Big Data (61227)

Instructors: P.PAPASTAMOULIS

Elective Course, 4th semester, 5 ECTS units
Course level: Graduate (MSc)
Language: Greek

Course Description

Modern statistical applications deal with the analysis of high-dimensional datasets. However, inference is extremelly challenging due to the curse of dimensionality: as the number of variables increases, the data may become quite noisy and it is difficult to separate the actual underlying information. The course will present modern statistical techniques particularly suited to problems with high dimensionality, such as:

- Large scale hypothesis testing
- Statistical applications in bioinformatics
- Applications of (generalized) linear models in big data problems
- Regularization techniques (LASSO, Ridge regression)
- Classification and model-based clustering for multivariate data with dimensionality reduction techniques
- Statistical network analysis

Prerequisites

- Essential
  – Hypothesis Testing - Linear Models
  – Generalized Linear Models
- Some basic knowledge on
  – Computational Statistics
  – R programming
- Students will also benefit from the following courses (not required)
  – Applied Bayesian Statistics
  – Statistical Learning

Target Learning Outcomes

After completing the course, the students will:

- know the challenges that big data impose to the statistician
- properly deal with large scale hypothesis testing
- put their hands on many different types of data from cutting edge research areas such as bioinformatics and network analysis

- learn many novel statistical ideas and methods developed in the last 20 years
- know how to apply the relevant methods using R and Bioconductor.

Recommended Bibliography

[1] Christophe Giraud (2015). Introduction to High-Dimensional Statistics. Philadelphia: Chapman and Hall/CRC

[2] Tony Cai, Xiaotong Shen, eds. 2011. High-dimensional data analysis. Frontiers of Statistics. Singapore: World Scientific

[3] Hastie, R. Tibshirani and R. Friedman (2009). Elements of Statistical Learning. Springer

[4] Efron, Bradley. Large scale inference: Empirical Bayes Methods for Estimation, Testing and Prediction. Cambridge University Press, 2010

[5] McLachlan, G. and Peel, D (2000). Finite Mixture Models. Cambridge University Press, 2010

[6] Wasserman, S. and G. Robins (2005). An introduction to random graphs, dependence graphs, and p⋆ . Models and methods in social network analysis.

[7] Hoff, P. D., A. E. Raftery, and M. S. Handcock (2002). Latent space approaches to social network analysis. Journal of the american Statistical association.

Teaching and Learning Activities

One three-hour lecture per week, study exercises and programming exercises as homework (some to be submitted).

Assessment and Grading Methods

The final grade will be a weighted average of 2 or 3 assignments combined with a presentation/oral examination.