Seminar: High-dimensional data and non-linear dimensionality reduction

Santiago Gallón*[a]

[a] *Departamento de Matemáticas y Estadística, Facultad de Ciencias Económicas, Universidad de Antioquia, Medellín, Colombia.*

2015

Description

Technological advances allow to scientists to collect high-dimensional data sets in which the number of variables p is considerably large with respect to number of samples n, doing it becomes the rule rather than exception in a plenty of areas like information technology, image processing, genomics, astronomy, finance, marketing among many others. This new development raises significant statistical challenges for data analysis due to the classical statistical tools cannot be used for high-dimensional problems (e.g. Bühlmann and van de Geer [1], and Cai and Shen [2]). In order to handle high-dimensional data, dimensionality reduction technique becomes crucial. Dimensionality reduction is a method to represent high-dimensional data by their low-dimensional embeddings. In other words, dimensionality reduction is the mapping of data from a high-dimensional space to a low-dimensional one such that uninformative variance in the data is discarded, or such that a subspace in which the data lives is detected. This technique has proved an important tool for instance in the fields of data analysis, data mining, data visualization, and machine learning (Lee and Verleysen [3], and Wang [4]).

Traditional methods like principal component analysis (PCA) and classical multidimensional scaling (cMDS) suffer from being based on linear models. However, they are not effective if data do not well reside on superplanes. Since the late 1990s, many nonlinear dimensionality reduction methods, also called manifold learning, have been developed assuming that each observed high-dimensional data resides on a low-dimensional manifold (Wang [4]).

In this seminar we give an overview of several methods for dimensionality reduction. The goal is to provide a self-contained overview of key concepts underlying many of these nonlinear dimensionality reduction algorithms. For each method, a short description of intuitive ideas, necessary mathematical details, and algorithmic implementation is provided. Knowledge of basic probability, analysis, and linear algebra is required.

* santiago.gallon@udea.edu.co

Contents

References

[1] P. Bühlmann and S. van de Geer. *Statistics for High Dimensional Data Methods: Methods, Theory and Applications.* Springer Series in Statistics. Springer, Berlin, 2011.

[2] T. Cai and X. Shen, editors. *High-Dimensional Data Analysis*, volume 2 of *Frontiers of Statistics.* Higher Education Press, Beijing, 2011.

[3] J. A. Lee and M. Verleysen. *Nonlinear Dimensionality Reduction.* Information Science and Statistics. Springer, New York, 2007.

[4] J. Wang. *Geometric structure of high-dimensional data and dimensionality reduction.* Higher Education Press, Beijing; Springer, Heidelberg, 2012.