

Sparse Principal Component Analysis

Promoteur : Rainer von Sachs (STAT), e-mail: rvs@stat.ucl.ac.be

Autre contact : Hilmar Böhm (STAT), e-mail: boehm@stat.ucl.ac.be

Localisation : Institut de statistique, Voie du Roman Pays 20

Principal component analysis (PCA) of possibly high-dimensional statistical signals is a tool to reduce the dimensionality of the data. Basically, PCA is a rotation of the original multivariate data vector into a basis which diagonalises the variance-covariance matrix of the data. The principal components arise as linear combinations of the original data with certain weights.

Though wildly used in many fields of applications (medical engineering, statistical signal processing, financial data modelling, psychology, ...) the choice of the (ideally "small") number of principal components to include into the description of the data without losing too much information is somewhat arbitrary and mostly only based on empirical criteria.

The same problem arises in a related question: how many and which of the original components of the multivariate data vector should be taken into account when trying to identify the most relevant contributions to the above-mentioned linear combination (making up the first PCA, the second, and so on)? This leads to the recent approach of "Sparse PCA" (see, e.g., Zou, Hastie and Tibshirani, 2004: Sparse Principal Component Analysis; Technical Report, Statistics Department, Stanford University).

The main task of this project is to investigate from an empirical point of view the behaviour of PCA and Sparse PCA for some selected models, related to our main application, multivariate analysis of EEG-data. This has to be done by systematic simulation studies (using the programming environment MATLAB), in collaboration with the more theoretical Ph.D. project of H. Böhm (STAT). Our goal is to find an ideally automatic (i.e. data-driven) approach to solve the above-mentioned problems of dimension-reduction.

Prérequis :

- Bonne compréhension en statistique multivariée
- Intérêt à la modélisation des signaux statistiques
- Maîtrise des langages de programmation, notamment MATLAB