High-dimension Statistics

Alexandre Tsybakov - ENSAE-CREST

Cours : 14 heures - TP : 10 heures

Objectif

La science statistique s'est profondément transformée au cours de la dernière décennie grâce au développement des méthodes d'inférence statistique en grande dimension. Cette évolution récente découle de al nécessité de traiter les données nouvelles, telles que, pour chaque individu, on dispose d'un grand nombre de variables observées, qui est parfois plus grand que le nombre des individus dans l'échantillon. Bien évidemment, pas toutes les variables sont pertinentes et d'habitude il en existe très peu. La notion de parcimonie (sparsite) est donc fondamentale pour l'interprétation statistique de données en grande dimension. Le but de ce cours est de présenter quelques principes fondateurs qui émergent dans ce contexte. Ces principes sont communs à de nombreux problèmes apparus récemment, tels que la régression linéaire en grande dimension, l'estimation de grandes matrices de faible rang, ainsi que les modèles de réseaux, par exemple, les modèles stochastiques à blocs. L'accent sera mis sur la construction de méthodes optimales en vitesse de convergence et leurs propriétés d'oracle.

Plan

- Modèle de suite gaussienne. Sparsité et procédures de seuillage.
- Régression linéaire en grande dimension. Méthodes BIC, Lasso, Dantzig selector, square root Lasso.
- Propriétés d'oracle et sélection de variables.
- Estimation de grandes matrices de faible rang. Sparse PCA.
- Inférence sur les réseaux. Modèle stochastique à blocs (stochastic bloc model).

Références

C.Giraud. Introduction to high-dimensional statistics. Chapman and Hall, 2015.
A.B.Tsybakov. Apprentissage statistique et estimation non-paramétrique. Polycopié de l'Ecole Polytechnique, 2014.
S.van de Geer. Estimation and testing under sparsity. Lecture Notes in Mathematics 2159. Springer, 2016.