

Sélection de variables efficace en grande dimension

Projet

De nos jours, toujours plus de capteurs sont utilisés afin de mesurer l'influence d'un nombre colossal (big data) de variables sur le phénomène d'intérêt. C'est par exemple le cas sur les avions lors des vols d'essai d'Airbus, ou sur les voitures lors de tests, ou sur les malades lorsqu'on effectue des tests sur le génome visant à identifier la source d'une maladie. Une question essentielle et commune à ces différents contextes est d'identifier un petit nombre de variables qui suffiraient à "expliquer" le phénomène étudié.

À cette fin, des procédures algorithmiquement efficaces telles que le LASSO ont été introduites et étudiées. Toutefois si le LASSO fonctionne bien sous certaines conditions (petit nombre de variables), ses performances pour retrouver les variables influentes se dégradent fortement lorsque le nombre de variables incluses dans l'étude devient grand.

Récemment, la procédure SLOPE a été introduite pour pallier certains défauts du LASSO tout en restant algorithmiquement efficace. Par contre, ses performances sont encore mal connues.

Les objectifs du stage sont :

1. comprendre le LASSO pour identifier les variables influentes en grande dimension,
2. comprendre SLOPE dans le même contexte,
3. comparer ces deux procédures sur simulations et sur le plan théorique,
4. proposer une analyse de la performance de SLOPE dans un cas simple (design orthogonal).

Prérequis

Outre de bonnes compétences en probabilités et statistique, le candidat devra savoir programmer en R afin de pouvoir évaluer les performances des approches envisagées sur simulations ainsi que sur de vraies données.

Encadrant : Alain Celisse, maître de conférences.

Length: 4-6 mois.

Opportunity: Le stage pourrait donner lieu à une thèse.

Laboratory: MODAL équipe-projet Inria, Lille.

Contact: Alain Celisse (celisse@math.univ-lille1.fr).

Bibliography

1. Bogdan van den Berg, Sabatti, Su, Candès. SLOPE: Adaptive Variable Selection via Convex Optimization. arXiv, 2013.
2. Wainwright. Sharp Thresholds for High-Dimensional and Noisy Sparsity Recovery Using ℓ^1 constrained Quadratic Programming. IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 55, NO. 5, MAY 2009