Computationally Tractable Methods for High-Dimensional Data

Peter Bühlmann

Many applications nowadays involve high-dimensional data with p variables (or covariates), sample size n and the relation that $p \gg n$. We focus on penalty-based estimation methods which are computationally feasible and have provable statistical and numerical properties. The Lasso (Tibshirani, 1996), an ℓ_1-penalty method, became very popular in recent years for estimation in high-dimensional generalized linear models. Extensions to other models or data-types call for more flexible convex penalty functions, for example to handle categorical data or for improved control of smoothness in additive models. The Group-Lasso (Yuan and Lin, 2006) and a new sparsity-smoothness penalty are general and useful penalty functions for many high-dimensional models beyond GLM's. Fast coordinatewise descent algorithms can be used for solving the corresponding convex optimization problems which allow to easily deal with large dimensionality p (e.g. $p \approx 10^6, n \approx 10^3$).

The talk includes: (i) a review of Lasso-type methods; (ii) new flexible penalty functions, fast algorithms (`R` package `grplasso`) and some comparisons with boosting; and (iii) some illustrations for bio-molecular data.