Oberwolfach Seminar: Statistics for High-Dimensional Data

Organizers: Peter Bühlmann and Sara van de Geer
May 27th - June 2nd, 2012

High-dimensional statistics deals with data where the number of variables p is much larger than the number of observations n. This means that classical statistical methods cannot be applied directly, and one needs a certain amount of complexity regularization to avoid overfitting. In the last decade, it has been shown that by using ℓ_1-type regularization methods, one can obtain computationally feasible algorithms, good and optimal theoretical properties, and practically meaningful results in the analysis of complex high-dimensional data. In this seminar, we discuss this popular and widely used ℓ_1-approach, and some of its nephews such as boosting and thresholding. Our goal is to provide an overview of the recent methodology, theory and computational aspects, with an emphasis on the Lasso and its extensions.

The Lasso estimator $\hat{\beta}$ is a least squares estimator with a penalty proportional to the sum of the absolute values of the coefficients:

$$\hat{\beta} := \arg\min_{\beta \in \mathbb{R}^p} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1.$$

Here, Y is an n-vector of responses and X is an $n \times p$ design matrix, and $\|\beta\|_1 := \sum_{j=1}^p |\beta_j|$ is the ℓ_1-norm of the regression coefficients. The parameter $\lambda > 0$ is a tuning parameter. Large values of λ will put many of the estimated coefficients to zero.

The Lasso has been extensively studied, including the following aspects:

- computational issues, algorithms, stability properties;
- theoretical conditions for obtaining a small prediction error;
- theoretical conditions for performing variable selection.

The Lasso is designed for (approximately) sparse situations, that is, the case where the underlying true regression has most of its coefficients (approximately) equal to zero. Extensions of the Lasso include the group Lasso, and versions for multiple regression, for panel data, and for high-dimensional additive models. The ℓ_1-penalty has been modified to handle diverse forms of structural sparsity, where zero coefficients occur in groups or concentrate in high levels of resolution, etc. We will present an overview of these extensions and address the above mentioned aspects.

A different direction concerns the extension to high-dimensional generalized linear models. We will treat for example, logistic regression, quantile regression, and regression with hinge-loss. Also non-convex problems will be studied, such as mixture models and mixed effects models.

We will moreover consider graphical models. The GLasso is introduced, as well as the many regressions approach. We include some recent work on directed acyclic graphs.

Key concepts in the mathematical theory include restricted eigenvalues, compatibility, restricted isometry and irrepresentability. We will discuss these issues in detail. Furthermore, the mathematical treatment is to some extent based on results from empirical process theory. We will provide the necessary background, which includes exponential probability inequalities, concentration and contraction inequalities.

We will also discuss how computational algorithms can be constructed exploiting the Karush-Kuhn-Tucker conditions, and we will illustrate the methodology and above mentioned models with real data applications.

The seminar is based on the book:

Statistics for High-Dimensional Data: Methods, Theory and Applications,
by Peter Bühlmann and Sara van de Geer (2011), Springer Series in Statistics
(http://www.springer.com/statistics/statistical+theory+and+methods/book/978-3-642-20191-2)

Prerequisites: The participants should have basic knowledge in mathematical statistics, on the level of e.g. the book *Mathematical Statistics: Basic and Selected Topics. 1 (Second ed.)*, by Peter J. Bickel and Kjell A. Doksum (2007), Pearson Prentice-Hall.