Sparse estimation in high-dimensional models

Alexandre Tsybakov
CREST and University of Paris 6

The aim of this short course is to give an introduction to statistical estimation in high-dimensional models (where the dimension p of the vector of unknown parameters is larger than the sample size n) under sparsity scenario. The model is called sparse if the number of non-zero coordinates of the vector of unknown parameters is much smaller than p. The quality of sparse estimation is usually assessed in terms of *model selection consistency* (i.e., recovering of the set of non-zero coordinates) and *sparsity oracle inequalities* (SOI) for the prediction risk. One of the most important issues is to build methods that attain optimal performances with respect to these two criteria under minimal assumptions on the dictionary (for example, in linear regression, this requirement is translated as minimal assumptions on the design matrix X). Sparse statistical estimation is closely related to the problem of compressive sensing in approximation theory, but is more complex because the noise is added. It is also related to the problem of aggregation of estimators since, using sparse estimation methods obeying the SOI, we can construct aggregates that are simultaneously optimal for convex, linear and model selection type aggregation.

First, an overview of the most popular methods of sparse statistical estimation will be given. They are mainly of the two types. Some of them, like the BIC, enjoy nice theoretical properties without any assumption on the dictionary but are computationally infeasible starting from relatively modest dimensions p. Others, like the Lasso or the Dantzig selector, are easily realizable for very large p but their theoretical performance is conditioned by severe restrictions on the dictionary. We will discuss and compare various types of such restrictions emphasizing that the Lasso and the Dantzig selector can be studied by similar methods and exhibit similar behavior.

We will then focus on *Sparse Exponential Weighting*, a new method of sparse recovery in regression, density and classification models realizing a compromise between theoretical properties and computational efficiency. The theoretical performance of the method is comparable with that of the BIC in terms of SOI for the prediction risk. No assumption on the dictionary is required, except for the standard normalization. At the same time, the method is computationally feasible for relatively large dimensions p. It is constructed using the exponential weighting with suitably chosen priors, and its analysis is based on the PAC-Bayesian ideas in statistical learning. We will develop a general technique to derive sparsity oracle inequalities from the PAC-Bayesian bounds.

Bibiliography

BICKEL, P.J., RITOV, Y. and TSYBAKOV, A.B. (2008) Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, to appear:
`http://www.imstat.org/aos/future_papers.html`

BUNEA, F., TSYBAKOV, A.B. and WEGKAMP, M.H. (2007a) Aggregation for Gaussian regression. *Annals of Statistics*, v.35, 1674-1697.

BUNEA, F., TSYBAKOV, A.B. and WEGKAMP, M.H. (2007b) Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics*, v.1, 169-194.

BUNEA, F., TSYBAKOV, A.B. and WEGKAMP, M.H. (2007c) Sparse density estimation with ℓ_1 penalties. *COLT-2007*, 530-543.

DALALYAN, A. and TSYBAKOV, A.B. (2008) Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Machine Learning*, published on-line:
`http://dx.doi.org/10.1007/s10994-008-5051-0`

JUDITSKY, A., RIGOLLET, P. and TSYBAKOV, A.B. (2008) Learning by mirror averaging. *Annals of Statistics*, to appear: `http://www.imstat.org/aos/future_papers.html`.