A NON ASYMPTOTIC THEORY FOR MODEL SELECTION

PASCAL MASSART

Model selection is a classical topic in statistics. The idea of selecting a model via penalizing a log-likelihood type criterion goes back to the early seventies with the pioneering works of Mallows and Akaike. One can find many consistency results in the literature for such criteria. These results are asymptotic in the sense that one deals with a given number of models and the number of observations tends to infinity. We shall give an overview of a non asymtotic theory for model selection which has emerged during these last ten years. In various contexts of function estimation it is possible to design penalized log-likelihood type criteria with penalty terms depending not only on the number of parameters defining each model (as for the classical criteria) but also on the "complexity" of the whole collection of models to be considered.
The performance of such a criterion is analyzed via non asymptotic risk bounds for the corresponding penalized estimator which express that it performs almost as well as if the "best model" (i.e. with minimal risk) were known. For practical relevance of these methods, it is desirable to get a precise expression of the penalty terms involved in the penalized criteria on which they are based. This is why this approach heavily relies on concentration inequalities, the prototype being Talagrand's inequality for empirical processes. Our purpose will be to give an account of the theory and discuss some selected applications.