Parameter tuning and cross-validation algorithms

Supervisor: Alain Celisse, Assistant Professor
Université Lille 1 – MODAL INRIA
celisse@math.univ-lille1.fr

Context and motivation

In real-life applications, estimators used by practitioners always depend on unkonwn parameters that have to be chosen, which is called "parameter tuning". For instance when estimating a density with a histogram (or a kernel estimator), the partition (resp. the bandwidth) has to be specified beforehand. In computer vision, other instances are LASSO and SVM algorithms: They both depend on a "regularization parameter" that has to be chosen by practitioners. Moreover the final performance of the estimator crucially depends on that unkonwn parameter.

This tuning step is often performed by use of Cross-Validation (CV) in practice, but without any real guaranty of performance. Indeed several CV algorithms exist and resulting performances are quite sensitive to the choice of one of them. However, only very few theoretical results have been derived that could provide guidelines to distinguish between CV algorithms.

From a practical point of view, a crucial question is to provide more insight in CV algorithms and then guidelines to practitioners for helping them in the tuning step.

Outline

Since CV is one of the most commonly used method for parameter tuning (and model selection) in practice, a crucial and challenging problem is to provide a unified theoretical understanding of CV algorithms.

The first step will consist in studying recent results about CV algorithms derived in the specific setting of projection estimators in density estimation. Since concentration inequalities are of great interest in this work, the student will spend some time on them.

The second step will be to extend existing results to kernel estimators in density estimation. A preliminary simulation study will be conducted to assess the potential gain in conveniently choosing the CV algorithm when estimating a density with kernels. Of course, other extensions are possible depending on the student skills.

In the last step, the goal will be to derive a unified strategy to assess the performance of CV algorithms applied to various estimators such as M-estimators.

Prerequisites

A good knowledge in Probability and basic concepts in Statistics are required. Skills on programming with for instance Matlab or R will be appreciated.