# V-fold cross-validation improved: V-fold penalization.

Sylvain Arlot

Laboratoire de Mathmatiques
Universit Paris-Sud - Btiment 425
F-91405 ORSAY, FRANCE
(e-mail: sylvain.arlot@math.u-psud.fr)

We investigate the efficiency of V-fold cross-validation (VFCV) for model selection from the non-asymptotic viewpoint, and suggest an improvement on it, which we call "V-fold penalization".

First, considering a particular (though simple) regression problem, we will show that VFCV with a bounded V is suboptimal for model selection. The main reason for this is that VFCV "overpenalizes" all the more that V is large. Hence, asymptotic optimality requires V to go to infinity. However, when the signal-to-noise ratio is low, it appears that overpenalizing is necessary, so that the optimal V is not always the larger one, despite of the variability issue. This is confirmed by some simulated data.

In order to improve on the prediction performance of VFCV, we propose a new model selection procedure, called "V-fold penalization" (penVF). It is a V-fold subsampling version of Efron's bootstrap penalties, so that it has the same computational cost as VFCV, while being more flexible. In a heteroscedastic regression framework, assuming the models to have a particular structure, penVF is proven to satisfy a non-asymptotic oracle inequality with a leading constant almost one. In particular, this implies adaptivity to the smoothness of the regression function, even with a highly heteroscedastic noise. Moreover, it is easy to overpenalize with penVF, independently from the V parameter. As shown by a simulation study, this results in a significant improvement on VFCV in several non-asymptotic situations.