LASSO, ITERATIVE FEATURE SELECTION AND THE CORRELATION SELECTOR: ORACLE INEQUALITIES AND NUMERICAL PERFORMANCES

PIERRE ALQUIER

Abstract. On this work we focus on the problem of regression estimation with quadratic loss. The setting used is the following: we observe pairs $(X_i, Y_i)_{1 \leq i \leq n}$ independent in $(\mathcal{X}, \mathbb{R})$, with no assumpation on $\mathcal{X}$. We assume that we have a (potentially very) large dictionary of functions $f_j : \mathcal{X} \to \mathbb{R}$ for $1 \leq j \leq m$ and want to predict Y, for a new pair (X, Y) with Y unknown, by

$$\sum_{j=1}^{m} \alpha_j f_j(X)$$

for a value $\alpha \in \mathbb{R}^m$. We propose a general family of algorithms for that task. This family is based on geometrical considerations: we build confidence regions for the optimal value (in a sense that we will define) $\overline{\alpha}$ and perform orthogonal projections on this region to build reliable estimators. We prove that a lot of estimators that have already been studied for this task (LASSO and Group LASSO [3, 4], Dantzig selector [2], Iterative Feature Selection [1], among others) belong to our general family of estimators. We also exhibit another particular member of this family that will be called Correlation Selector in this presentation. Using general properties of our family of algorithm we prove sparse oracle inequalities for these estimators, that means controls on the excess risk of prediction of the estimator $\hat{\alpha}$ in terms of the number of non-zero coefficients $\overline{\alpha}_j$. We also provide numerical simulations in order to compare the numerical performances of these estimators on a toy example.

References

[1] Alquier, P. Iterative feature selection in regression esimation. Annales de l'IHP, accepted; to be published in 2008.

[2] Candes, E., and Tao, T. The dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics 35* (2007).

[3] Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B 58*, 1 (1996), 267–288.

[4] Yuan, M., and Lin, Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society B 68*, 1 (2006), 49–67.

CREST, and, Laboratoire de Probabilites et Modeles Aleatoires (Universite Paris 7), 175, rue du Chevaleret, 75252 Paris Cedex 05, France.

URL: http://www.crest.fr/pageperso/alquier/alquier.htm

E-mail address: alquier@ensae.fr

Date: February 6, 2008.

2000 *Mathematics Subject Classification.* Primary 62G08; Secondary 62J07, 62G15, 68T05.

Key words and phrases. Regression estimation, statistical learning, confidence regions, shrinkage and thresholding methods, LASSO.