DISCUSSION ON THE PAPER "LEAST ANGLE REGRESSION"

JEAN-MICHEL LOUBES AND PASCAL MASSART

The issue of model selection has drawn the attention of both applied and theoretical statisticians for a long time. Indeed, there has been an enormous range of contribution in model selection proposal, which includes the work by Akaike [1], Mallows [7], Foster and George [5], Birgé and Massart [3] or Abramovich, Benjamini, Donoho and Johnstone [2]. Over the last decade, modern computer-driven methods have been developed such as All Subsets, Forward Selection, Forward Stagewise or LASSO. Such methods are useful in the setting of standard linear model, where we observe noisy data, and wish to predict the response variable using only a few number of covariables, since they provide automatically linear models that fit the data. The procedure described in this paper is, on the one hand, numerically very efficient and on the other hand, very general, since, with slight modifications, it enables to recover the estimates given by the LASSO and Stagewise.

I. Estimation procedure

The "LARS" method is based on a recursive procedure selecting, at each step, the covariables having largest absolute correlation with the response y. In the case of an orthogonal design, the estimates can then be viewed as an l^1-penalized estimator. Consider the linear regression model where we observe y with some random noise ε, with orthogonal design assumptions:

$$y = X\beta + \varepsilon.$$

Using the soft-thresholding form of the estimator, we can write it, in an equivalent way, as the minimum of an ordinary least-squares and a l^1 penalty over the coefficients of the regression. As a matter of fact, at step $k = 1, \dots, m$, the estimators $\hat{\beta}^k = X^{-1}\hat{\mu}^k$ are given by

$$\hat{\mu}^k = \arg\min_{\mu \in \mathbb{R}^n} \left(\|Y - \mu\|_n^2 + 2\lambda_n^2(k)\|\mu\|_1 \right).$$

There is a trade-off between the two terms, balanced by the smoothing decreasing sequence $\lambda_n^2(k)$. The more stress is laid on the penalty, the more parsimonious the representation will be. The choice of the l^1 penalty enables to keep the largest coefficients, while the smallest one are shrinked towards zero in a soft-thresholding scheme. This point of view is investigated in the LASSO algorithm as well as in studying the False Discovery Rate (FDR).

So, choosing these weights in an optimal way determine the form of the estimator as well as its asymptotic behaviour. In the case of the algorithmic procedure, the suggested level is the $k+1$ order statistic:

$$\lambda_n^2(k) = |y|_{(k+1)}$$

As a result, it seems possible to study the asymptotic behaviour of the LARS estimates under some conditions on the coefficients of β. For instance, if there exists a roughness parameter $\rho \in [0, 2]$, such that $\sum_{j=1}^m |\beta_j|^\rho \leq M$, metric entropy theory results lead to upper bound for the mean squares error $\|\hat{\beta} - \beta\|^2$. Here we refer to the results obtained in [8]. Consistency should be followed by the asymptotic distribution, as it is done for the LASSO in [6].

The interest for such an investigation is double: first, it gives some insight to the properties of such estimators. Secondly it suggests an approach for choosing the threshold λ_n^2 which can justify the empirical cross-validation method, developed later in the paper. Moreover, the asymptotic distributions of the estimators are needed for inference.

Other choices of penalty and loss functions can also be considered. First of all for $\gamma \in (0,1]$, consider

$$J_\gamma(\beta) = \sum_{j=1}^{m} (X\beta)_j^\gamma.$$

If $\gamma < 1$, the penalty is not convex anymore, but there exist algorithms to solve the minimization problem. Constraints on the l^γ norm of the coefficients are equivalent to lacunarity assumptions and may make easier estimation of sparse signals, which it is often the case for high dimensional data for instance.

Moreover, replacing the quadratic loss function with a l^1 loss gives rise to a robust estimator, the penalized absolute deviation of the form

$$\tilde{\mu}^k = \arg\min_{\mu \in \mathbb{R}^n} \left(\|Y - \mu\|_{n,1} + 2\lambda_n^2(k)\|\mu\|_1 \right).$$

Hence, it is possible to get rid of the problem of variance estimation for the model with this estimates whose asymptotic behaviour can be derived from [8], in the regression framework.

Finally, a penalty over both the number of coefficients and the smoothness of the coefficients can be used to study, from a theoretical point of view, the asymptotics of the estimate. Such a penalty is analogous to complexity penalties studied in [9]:

$$\mu^\star = \arg\min_{\mu \in \mathbb{R}^n,\, k \in [1,m]} \left(\|Y - \mu\|_n^2 + 2\lambda_n^2(k)\|\mu\|_1 + \mathrm{pen}(k) \right).$$

II. Mallows' C_p

We would like now to discuss the crucial issue of selecting the number k of influential variables. To make clear this discussion, let us first assume the variance σ^2 of the regression errors to be known. Interestingly the penalized criterion which is proposed by the authors is exactly equivalent to Mallows' C_p when the design is orthogonal (this is indeed the meaning of their Theorem 3). More precisely, using the same notations as in the paper, let us focus on the following situation which illustrates what happens in the orthogonal case where LARS is equivalent to the Lasso. One observes some random vector y in $\mathbb{R}^n$, with expectation μ and covariance matrix $\sigma^2 I_n$. The variable selection problem that we want to solve here is to determine which components of y are influential. According to Lemma 1, given k, the kth LARS estimate $\widehat{\mu}_k$ of μ can be explicitly computed as a soft thresholding estimator. Indeed, considering the order statistics of the absolute values of the data denoted by

$$|y|_{(1)} \geq |y|_{(2)} \geq \dots \geq |y|_{(n)}$$

and defining the soft threshold function $\eta\,(.,t)$ with level $t \geq 0$ as

$$\eta\,(x,t) = x\mathbb{1}_{|x|>t}\left(1 - \frac{t}{|x|}\right),$$

one has

$$\widehat{\mu}_{k,i} = \eta\left(y_{i,}\, |y|_{(k+1)}\right).$$

In order to select k, the authors propose to minimize the C_p criterion

$$C_p\left(\widehat{\mu}_k\right) = \|y - \widehat{\mu}_k\|^2 - n\sigma^2 + 2k\sigma^2. \tag{1}$$

Our purpose is to analyze this proposal with the help of the results on penalized model selection criteria proved in [3] and [4]. In those papers some oracle type inequalities are proved for selection procedures among some arbitrary collection of projection estimators on linear models when the regression errors are Gaussian. In particular one can apply them to the variable subset selection problem above, assuming the random vector y to be Gaussian. If one decides to penalize in the same way the subsets of variables with the same cardinality, then the penalized criteria studied in [3] and [4] take the form

$$C_p'\left(\widetilde{\mu}_k\right) = \left\|y - \widetilde{\mu}_k\right\|^2 - n\sigma^2 + pen\left(k\right) \tag{2}$$

where, $pen\left(k\right)$ is some penalty to be chosen and $\widetilde{\mu}_k$ denotes the hard threshold estimator with components

$$\widetilde{\mu}_{k,i} = \eta'\left(y_{i,}\left|y\right|_{(k+1)}\right)$$

where

$$\eta'\left(x,t\right) = x\mathbb{1}_{|x|>t}.$$

The essence of the results proved by Birgé and Massart in this case is the following. Their analysis covers penalties of the form

$$pen\left(k\right) = 2k\sigma^2 C\left(\log\left(\frac{n}{k}\right) + C'\right)$$

(note that the FDR penalty proposed in [2] corresponds to the case $C = 1$). It is proved in [3] that if the penalty $pen\left(k\right)$ is heavy enough, i.e. $C > 1$ and C' is an adequate absolute constant, then the model selection procedure works in the sense that, up to constant, the selected estimator $\widetilde{\mu}_{\widetilde{k}}$ performs as well as the best estimator among the collection $\left\{\widetilde{\mu}_k, 1 \leq k \leq n\right\}$ in terms of quadratic risk. On the contrary, it is proved in [4] that if $C < 1$, then at least asymptotically, whatever C', the model selection does not work, in the sense that even if $\mu = 0$, the procedure will systematically choose large values of k, leading to a suboptimal order for the quadratic risk of the selected estimator $\widetilde{\mu}_{\widetilde{k}}$. So, to summarize, some $2k\sigma^2\log\left(n/k\right)$ term should be present in the penalty, in order to make the model selection criterion (2) work. In particular, the choice $pen\left(k\right) = 2k\sigma^2$ is not appropriate, which means that Mallows' C$_p$ does not work in this context. At a first glance, those results seem to indicate that some problems could occur with the use of the Mallows' C$_p$ criterion (1). But fortunately, this is not at all the case because some very interesting phenomenon occurs, due to the soft thresholding effect. As a matter of fact, if we compare the residual sums of squares of the soft threshold estimator $\widehat{\mu}_k$ and the hard threshold estimator $\widetilde{\mu}_k$, we easily get

$$\left\|y - \widehat{\mu}_k\right\|^2 - \left\|y - \widetilde{\mu}_k\right\|^2 = \sum_{i=1}^{n}\left|y\right|_{(k+1)}^2\mathbb{1}_{|y_i|>|y|_{(k+1)}} = k\left|y\right|_{(k+1)}^2$$

so that the "soft" C$_p$ criterion (1) can be interpreted as a "hard" criterion (2) with random penalty

$$pen\left(k\right) = k\left|y\right|_{(k+1)}^2 + 2k\sigma^2. \tag{3}$$

Of course this kind of penalty escapes *stricto sensu* to the analysis of [3] and [4] as described above since the penalty is not deterministic. But it is quite easy to realize that in this penalty, $\left|y\right|_{(k+1)}^2$ plays the role of the apparently "missing" logarithmic factor $2\sigma^2\log\left(n/k\right)$. Indeed, let us consider the pure noise situation where $\mu = 0$ in order to keep the calculations as simple as possible. Then, if we consider the order statistics of a sample $U_1, ..., U_n$ of the uniform distribution on $[0,1]$

$$U_{(1)} \leq U_{(2)} \leq ... \leq U_{(n)},$$

taking care of the fact that these statistics are taken according to the usual increasing order while the order statistics on the data are taken according to the reverse order, $\sigma^{-2}\left|y\right|_{(k+1)}^2$ has the same distribution as

$$Q^{-1}\left(U_{(k+1)}\right),$$

where Q denotes the tail function of the chi-square distribution with 1 degree of freedom. Now using the double approximations $Q^{-1}(u) \sim 2\log(|u|)$ as u goes to 0 and $U_{(k+1)} \approx (k+1)/n$ (which at least means that given k, $nU_{(k+1)}$ tends to $k+1$ almost surely as n goes to infinity but can also be expressed with much more precise probability bounds) we derive that $|y|^2_{(k+1)} \approx 2\sigma^2 \log(n/(k+1))$. The conclusion is that it is possible to interpret the "soft" C_p criterion (1) as a randomly penalized "hard" criterion (2). The random part of the penalty $k|y|^2_{(k+1)}$ cleverly plays the role of the unavoidable logarithmic term $2\sigma^2 k \log(n/k)$, allowing the hope that the usual $2k\sigma^2$ term will be heavy enough to make the selection procedure works as we believe it does. A very interesting feature of the penalty (3) is that its random part does not depend neither on the scale parameter σ^2, nor on the tail of the errors. This means that one could think to adapt the data driven strategy proposed in [4] to choose the penalty without knowing the scale parameter to this context, even if the errors are not Gaussian. This would lead to the following heuristics. For large values of k, one expects the quantity $-\|y-\widehat{\mu}_k\|^2$ to behave as an affine function of k with unknown slope. Just estimate this slope (for instance by making a regression of $-\|y-\widehat{\mu}_k\|^2$ with respect to k for large enough values of k) and plug the resulting estimate mulitplied by 2 in (1) to replace the $2k\sigma^2$ term. Of course, some more efforts would be requested to complete this analysis and provide rigorous oracle inequalities in the spirit of those given in [3] and [4] or [2] and also some simulations to check whether our proposal to estimate σ^2 is valid or not.

Our purpose was just here to mention some possible explorations starting from the present paper that we have found very stimulating. It seems to us that it solves practical questions of crucial interest and raises very interesting theoretical questions: consistency of LARS estimator, efficiency of Mallows' C_p in this context, use of random penalties in model selection for more general frameworks.

References

[1] AKAIKE, H. Maximum Likelihood identification of Gaussian autoregressive moving average models. *Biometrika*, 253-265 (1973).

[2] ABRAMOVICH, F., BENJAMINI, Y., DONOHO D. and JOHNSTONE, I. Adapting to unkown sparsity by controlling the false discovery rate. Technical Report n2000 − 19, Stanford University (2000).

[3] BIRGÉ, L. and MASSART, P. Gaussian model selection. *J. Eur. Math. Soc.* **3**, 203-268 (2001)

[4] BIRGÉ, L. and MASSART, P. A generalized C_p criterion for Gaussian model selection. Technical Report n°647 , Universités de Paris 6 & Paris 7 (2001).

[5] FOSTER, D. and GEORGE, E. The risk inflation criterion for multiple regression. *The Annals of Statistics*, 1947-1975, (1994).

[6] KNIGHT, K. and FU, B. Asymptotics for Lasso type estimators. *The Annals of Statistics*, 1356-1378 (2000).

[7] MALLOWS. Some comments on C_p. *Technometrics* , **15** 661-675 (1973).

[8] LOUBES, J-M. and VAN DE GEER, S. Adaptive Estimation using Thresholding type Penalties. *Statistica Neerlandica* , 1-26 (2002).

[9] VAN DE GEER, S. Least squares with complexity penalties. *Math Methods of Statistics*, 355-374 (2001).

CNRS and Laboratoire de Mathématiques, UMR 8628, Equipe de Probabilités, Statistique et Modélisation - Université Paris-Sud, Bât 425, 91405 Orday cedex

Laboratoire de Mathématiques, UMR 8628, Equipe de Probabilités, Statistique et Modélisation - Université Paris-Sud, Bât 425, 91405 Orday cedex

E-mail address: Jean-Michel.Loubes@math.u-psud.fr

E-mail address: Pascal.Massart@math.u-psud.fr