Journal de la Société Française de Statistique
Vol. 160 No. 3 (2019)

Some thoughts about variable selection without information on the errors' structure

Titre: Quelques pensées sur la sélection de variables sans information sur la structure des erreurs

Lucien Birgé[1]

As mentioned in Sylvain's long and thorough study, I worked with Pascal Massart on the subject of calibration of the penalty terms for model selection but that was actually many years ago. Since then I worked on a somewhat different subject and forgot a large part of this old work. Reading Sylvain's paper reminded me of a few things and I also learned much from it in particular that a lot of progress has been made on the subject. Although I have been interested by another (but as we shall see not so different) type of problem, I occasionally thought about a particular case of this old stuff, namely complete variable selection. Let me first describe the mathematical framework that I was interested in.

One observes n real random variables $Y_1,\ldots,Y_n$ with the following structure:

$$Y_i = \theta_i + \sigma\varepsilon_i, \quad i = 1,\ldots,n \quad \text{or equivalently} \quad Y = \theta + \sigma\varepsilon \in \mathbb{R}^n,$$

where the ε_i are i.i.d., $\theta \in \mathbb{R}^n$ and σ being unknown parameters. An implicit assumption is that the number of θ_i that are *large* is relatively small compared to n. By "large" I mean distinguishable from the size σ of the noise (assuming that the distribution of the ε_i has been suitably normalized). The risk of an estimator $\widehat{\theta}$ of θ is then usually measured by $\mathbb{E}_\theta\left[\left\|\widehat{\theta}-\theta\right\|^2\right]$, which is quite reasonable when the $\varepsilon_i \sim \mathscr{N}(0,1)$ but not necessarily in a general case since, depending on the distribution of the ε_i, the Euclidean distance between $\widehat{\theta}$ and θ may not reflect the distance (total variation or Hellinger distance) between the associated distributions $\mathbf{P}_{\widehat{\theta}}$ and $\mathbf{P}_\theta$ of Y.

The case we studied in Birgé and Massart (2001, 2007) was that of $\varepsilon_i \sim \mathscr{N}(0,1)$. If θ belongs to a linear subspace of $\mathbb{R}^n$ the natural estimator of θ is the least squares estimator. Given a subset m of $\{1,\ldots,n\}$ with cardinality $|m|$, we may introduce as a model the set $\Theta_m = \{\theta \in \mathbb{R}^n \,|\, \theta_i = 0 \text{ for } i \notin m\}$. The least squares method provides an estimator for each of these models and a good way of selecting one is via penalization, penalizing the least squares on model m by a penalty of the form $C\sigma|m|$, proportional to the dimension of the model Θ_m. This estimator, with a good choice of C, actually results in a thresholding of the observations Y_i of the following form: let $T_n = K\sigma\left(1+\sqrt{2\log n}\right)$ with $K > 1$ and set $\widehat{\theta}_i = Y_i$ if $|Y_i| \geq T_n$, $\widehat{\theta}_i = 0$ otherwise. This is an example of the *hard-thresholding method* with threshold T_n depending on n and σ and popularized by

[1] Sorbonne Université, LPSM, France

Donoho and Johnstone for wavelet estimators in the early 90's. This estimator has been studied by Donoho and Johnstone (1994) and Birgé and Massart (2001) among many others and its optimality properties are well-known. When σ is unknown, it should be somehow estimated and the procedure modified in a suitable way. It is in this case that Pascal and I introduced in Birgé and Massart (2007) the concepts of minimum and optimal penalties which form the starting point for Sylvain's paper. The analysis of the performance of thresholding procedures and the construction of an optimal penalty rely heavily on large deviations arguments for the ε_i. What if the ε_i are not sub-gaussian, for instance Cauchy? And what if the true distribution of the ε_i is unknown?

If σ and the distribution of ε_i is known, but not Gaussian, say Cauchy, the least squares estimator does not coincide with the maximum likelihood estimator and an idea would be to use instead a penalized maximum likelihood. Again, one should find a suitable method to estimate σ which is likely to depend on the distribution of ε_i. Another fact which is worth noticing is that techniques based on the likelihood will generally lead to a control of the Hellinger distance between $\mathbf{P}_{\widehat{\theta}}$ and $\mathbf{P}_\theta$. A risk of the form $\mathbb{E}_\theta\left[\left\|\widehat{\theta}-\theta\right\|^2\right]$ is not appropriate anymore. It is also well-known that the maximum likelihood estimator is not robust. A poor modelling of the distribution of ε_i may therefore lead to bad results.

An alternative approach to regression with non-Gaussian errors (including the Cauchy case) has been introduced in Baraud (2011, Section 8) based on a variant of T-estimators and new robust tests between two distributions. It nevertheless still suffers from limitations inherent to T-estimation, in particular boundedness of the parameter space and does not apply to the case of $\Theta=\mathbb{R}^n$.

Following Baraud's paper, a novel approach to regression with unknown form of the errors has been developed in Baraud et al. (2017) and Baraud and Birgé (2018) using ρ-estimators. In these papers we explain how to design estimators of the joint distribution of the Y_i by model selection, even if σ is unknown and the density s of the ε_i is only approximatively known, the results being a consequence of the natural robustness of the ρ-estimator. The idea is rather simple: discretize the parameter σ and introduce a set $\mathscr{Q}$ of different guesses q for s. To each choice of σ and q, build a robust estimator $\widehat{\theta}_{\sigma,q}$ of θ and use a penalized method to choose one among them. In this framework, we measure the risk of a procedure by the Hellinger distance between $\mathbf{P}_{\widehat{\theta}}$ and $\mathbf{P}_\theta$.

The important point here is that the penalty function we use does not depend on an unknown parameter but only on some universal constant and known quantities. We only know an upper bound K for this constant, which results from our computations and even if we can show in examples that this form of penalty leads to optimal results up to constants it is quite likely that this bound K is much too large for practical purposes and that the constant driving the penalty should be empirically calibrated.

The use of ρ-estimators allows a considerable flexibility for estimating the distribution of Y in this regression framework. If the distribution of ε_i is known, up to a scaling factor σ, as is classically assumed, and the model is correct, it is maybe possible to define a concept of minimal penalty corresponding to the minimal value of the universal constant in the penalty that is required for the method to work. But this minimal penalty is likely to depend on the distribution of ε_i. If this distribution is only approximatively known, the situation seems even more complex.

Nevertheless, although the ρ-estimator is more a theoretical rather than a practical tool in many (not all) situations, it would be interesting to know whether some fine tuning, analogous to the one explained in Sylvain's paper, could be done for ρ-estimators in a regression framework.

References

Baraud, Y. (2011). Estimator selection with respect to Hellinger-type risks. *Probab. Theory Related Fields*, 151(1-2):353–401.

Baraud, Y. and Birgé, L. (2018). Rho-estimators revisited: General theory and applications. *Ann. Statist.*, 46(6B):3767–3804.

Baraud, Y., Birgé, L., and Sart, M. (2017). A new method for estimation and model selection: ρ-estimation. *Invent. Math.*, 207(2):425–517.

Birgé, L. and Massart, P. (2001). Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, 3(3):203–268.

Birgé, L. and Massart, P. (2007). Minimal penalties for Gaussian model selection. *Probability Theory and Related Fields*, 138(1-2):33–73.

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455.