Choosing the order of a non parametric estimator by LASSO regression

Yves Ngounou, Denys Pommeret

Choosing the order of a non parametric estimator by LASSO regression

Yves Ngounou and Denys Pommeret

Aix-Marseille University, Ecole Centrale, CNRS, I2M, Campus de Luminy, 13288 Marseille cedex 9, France

November 27, 2018

Abstract

The use of LASSO technics permit to select automatically regressors in linear or generalized linear models. We combine here this method to get a selection operator of the number of components for non parametric estimators. The method is compared to different data driven technics. Elastic Net method is also studied. Simulation show the potential of this approach. A real data in actuarial science is studied.

1 General problem

Let μ be a reference measure with an associated dense orthogonal basis $\mathcal{Q} = \{Q_k; k \in \mathbb{N}^d\}$, satisfying $Q_0 = 1$ and such that

$$\int_{\mathbb{R}^d} Q_j(x)Q_k(x)\nu(dx) \quad = \quad \delta_{jk},$$

with $\delta_{jk} = 1$ if $j = k$ and 0 otherwise. We consider $g \in L^2(\mu)$ such that for all $x \in \mathbb{R}^d$

$$g(x) = \sum_{i\in\mathbb{N}^d} g_i Q_k(x) \quad \text{with} \quad g_i := \int_{\mathbb{R}^d} Q_k(x)g_i(x)\mu(dx).$$

For an integer $K > 0$ we define the Kth order approximation of g by

$$g^{(K)}(x) \quad = \quad \sum_{|i|\leq K} g_i Q_k(x), \tag{1}$$

where $|i| = i_1 + \cdots + i_d$. If we observe n random vectors $X_1, \cdots, X_n$ such that $\mathbb{E}(Q_k(X_i)) = g_i$, we can estimate this quantity by

$$\widehat{g}_i \quad = \quad \sum_{j=1}^{n} g_i(X_j)$$

and we deduce a Kth order non parametric estimator of g as

$$\widehat{g}^{(K)}(x) \quad = \quad \sum_{|i|\leq K} \widehat{g}_i Q_k(x).$$

Our problem is to choose the order K of this estimator. There exists various methods as cross validation (see), minimax approach (see), MISE (see). In this paper we consider the estimation problem as a regression model adn we apply a penalized technic to select automatically the components.

Remark 1 *We can also consider $K \in \mathbb{N}^d$. Then we have to use an order on $\mathbb{R}^d$. Here we consider a canonical basis, $e = (e_1, \cdots, e_d)$ and we assume that $x < y$ if $|x| < |y|$ or if $|x| = |y|$ and if j is the greater non null component index of y then $x_j < y_j$.*

2 A regression model

We describe here the bivariate case (but the multivariate is similar). We consider the selection of the non null coefficients in the sequence $g_1, \cdots, g_k$. If we use a product of univariate orthogonal basis, writing $X = (Y, Z)$ and $Q_{i,j}(x) = P_i(y)R_j(z)$ we get

$$g_{i,j} = \mathbb{E}(Q_{i,j}(X)) = \mathbb{E}(P_i(Y)R_j(Z)).$$

It is easily seen that for all integers u, s,

$$\begin{aligned} \mathbb{E}(P_s(Y)|Z) = 0 &\Rightarrow g_{u,s} = 0 \\ \mathbb{E}(R_u(Z)|Y) = 0 &\Rightarrow g_{u,s} = 0 \end{aligned}$$

and that

$$\begin{aligned} \mathbb{E}(P_s(Y)|Z) = 0 &\Leftrightarrow \mathbb{E}(Z|P_s(Y)) = 0 \\ \mathbb{E}(R_u(Z)|Y) = 0 &\Leftrightarrow \mathbb{E}(Y|R_u(Z)) = 0. \end{aligned}$$

Then (1) is related to the regression problem

$$\mathbb{E}(Z|P_1(Y), \cdots, P_K(Y)) = \sum_{i=1}^{K} \beta_i P_i(Y), \tag{2}$$

with $\beta_i = \mathbb{E}(Z|P_i(Y))$. The LASSO procedure can be used to detect which coefficients β_i are null. The model to be considered is the regression of Y on $R_t(Z)$, $t = 1, 2, \cdots, K$. Or by symmetry the regression of Z on $P_t(Y)$, $t = 1, 2, \cdots, K$. The Elastic Net approach could also be used to select more than n components.

3 Bootstrap approach

Consider the regression model (2) and write $g_K(Y) = \mathbb{E}(Z|P_1(Y), \cdots, P_K(Y))$. Assume that Y is a random variable with known distribution with respect to ν, say f_Y. Let P_n be a sequence of orthonormal functions with respect to ν. Then we have

$$K \leq K_0 \Leftrightarrow \mathbb{E}(\frac{g_K(Y)P_k(Y)}{f_Y(Y)}h(Y)) = 0 \ \forall k > K_0,$$

where h denotes the density of ν. Starting from this characterization of the order approximation, we now want to find the distribution of Y. One way is to consider an empirical density $\widehat{f_Y}$ and to write

$$K \leq K_0 \Leftrightarrow \mathbb{E}(\frac{g_K(Y)P_k(Y)}{\widehat{f_Y}(Y)}h(Y)) = 0 \ \forall k > K_0,$$

which is true asymptotically. Another way is to change the distribution of Y, keeping its relation with Z. We then propose a weighted bootstrap, such that the bootrstraped distribution of Y^* is close the ν. We then draw Y with a weighted equal to $h(Y)/\widehat{f}_Y(Y)$.

4 Numerical Analyses

4.1 Monte-Carlo Experiment

Empirical levels

Empirical powers

4.2 Real data sets

References

[1] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716–723.

[2] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B*, **57**, 289–300.

[3] Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *J. Amer. Statist. Assoc.*, **99**, 96–104.

[4] Guan, Z., Wu, B. and Zhao, H. (2008) Nonparametric estimator of false discovery rate based on Bernstein polynomials. *Statistica Sinica*, **18**, p. 905–923.

[5] Inglot, T. and Ledwina, T. (2006). Data-driven score tests for homoscedastic linear regression model: asymptotic results. *Probab. Math. Statist.*, **26**, 41–61.

[6] Janic-Wróblewska, A. and Ledwina, T. (2000). Data driven rank test for two-sample problem. *Scand. J. Statist.* **27**, 281–297.

[7] Kallenberg, W. C. and Ledwina, T. (1995). Consistency and Monte Carlo simulation of a data driven version of smooth goodness-of-fit tests. The Annals of Statistics, 23(5), 1594-1608.

[8] Ledwina, T. (1994). Data-driven version of Neyman's smooth test of Fit. *J. Amer. Statist. Assoc.* **89**, 1000–1005.

[9] Neyman, J. (1937). Smooth Test for Goodness of Fit, *Skandinavisk Aktuarietidskrift*, **20**, 149–199.

[10] Szegö, G. (1939) *Orthogonal Polynomials*. Amer. Math. Soc., Colloquium Publications Volume XXIII.

Appendix

The first orthonormal Hermite polynomials are

$$\begin{aligned} Q_1(x) &= x, \\ Q_2(x) &= \frac{1}{\sqrt{2}}(x^2 - 1) \\ Q_3(x) &= \frac{1}{\sqrt{6}}(x^3 - 3x) \\ Q_4(x) &= \frac{1}{\sqrt{24}}(x^4 - 6x^2 + 3) \\ Q_5(x) &= \frac{1}{\sqrt{120}}(x^5 - 10x^3 + 15x) \end{aligned}$$