Variable selection in high-dimensional linear model with possibly asymmetric

Gabriela Ciuperca

Version of Record: https://www.sciencedirect.com/science/article/pii/S0167947320302036
Manuscript_5378d156dd9a38c560b308878a481ab7

Variable selection in high-dimensional linear model with possibly asymmetric errors

Gabriela CIUPERCA[1]

Institut Camille Jordan, Université Claude Bernard Lyon 1, 69622 Villeurbanne, France

Abstract

In many application areas, the problem of the automatic variable selection in a linear model with asymmetric errors is encountered, when the number of explanatory variables diverges with the sample size. For this high-dimensional model, the penalized least squares method is not appropriate and the quantile framework makes the inference more difficult because of the non differentiability of the loss function. An estimation method by penalizing the expectile process with an adaptive LASSO penalty is proposed and studied. Two cases are considered: first with the number of model parameters is assumed to be much smaller than the sample size and afterwards it could be of the same order; the two cases being distinct by the adaptive penalties considered. For each case, the rate convergence is obtained and the oracle properties of the adaptive LASSO expectile estimator are established. The proposed estimators are evaluated through Monte Carlo simulations and compared with the adaptive LASSO quantile estimator. The proposed estimation method is also applied to real data in genetics.

Keywords: adaptive LASSO, expectile, high-dimension, oracle properties.

1. Introduction

The focus of the present paper is to better detect significant variables in a linear model, with the possibility that the number of explanatory variables varies with the sample size and when the error distribution is asymmetrical. For this type of law, the use of the least squares(LS) estimation method is not appropriate because of the estimator's accuracy (see Liao et al. (2019)). One possibility would be to use the quantile method, but it has the disadvantage that the loss function is not derivable, which complicates the theoretical study but also the computational methods. A very interesting possibility is to consider the expectile method, introduced by Newey and Powell (1987), under assumption that the first

[1]Correspondence to: Université Claude Bernard Lyon 1, UMR 5208, Institut Camille Jordan, Bat. Braconnier, 43, blvd du 11 novembre 1918, F - 69622 Villeurbanne Cedex, France
Email address: Gabriela.Ciuperca@univ-lyon1.fr

moments of ε exist. This method has the advantage over the quantile model that the loss function is differentiable, which makes the theoretical study more amenable and considerably facilitates the numerical computation (see also Gu and Zou (2016)). Readers can find in Newey and Powell (1987), Schnabel and Eilers (2009) the theoretical and computational advantages of expectile method over quantile regression. Two of these advantages are: the efficiency of the expectile estimator and the fact that its asymptotic variance can be calculated without going via the density of the model errors. The advantage of the quantile estimation method is that it is more robust to outliers than the expectile method, the last being also more sensitive to the extreme values in the response variable (see also Zhang and Li (2017)).

In application fields (genetics, chemistry, biology, industry, finance), with the development in recent years of storage and/or measurement tools, we are confronted to study the influence of a very large number of variables on a studied process. That is why, we consider in the present work the following linear model:

$$Y_i = \mathbf{X}_i^t \boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \cdots, n, \tag{1}$$

with the vector parameter $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\boldsymbol{\beta^0}$ its true value (unknown). The size p of $\boldsymbol{\beta^0}$ can depend on n but the components β_j^0 don't depend on n, for any $j = 1, \cdots, p$. The vector $\mathbf{X}_i^t = (X_{i1}, \cdots, X_{ip})^t$ contains the values of the p explanatory deterministic variables and Y_i the values of response variable for observation i. The values $(Y_i, \mathbf{X}_i)$ are known for any $i = 1, \cdots, n$. Throughout the paper, all vectors are column. If p is very large, in order to find the explanatory variables that significantly influence the response variable Y, an automatic selection should be made without performing hypothesis tests. Concerning the hypothesis testing of coefficients in high dimensional linear regression model, a lot of progress has been made in recent years. For Gaussian errors, with possibility that the number of variables exceeds the sample size, Zhang and Zhang (2014) propose a method for constructing confidence interval for individual coefficients, which can be used to select variables after proper thresholding. Dezeure et al. (2015) present a review of frequentist methods for constructing p-values and confidence intervals in a high dimensional linear model. Always for this model, with zero mean errors, Dezeure et al. (2017), Shah and Bühlmann (2018) propose bootstrap methodology.

For model (1), let the index set of the non-null true parameters,

$$\mathcal{A} \equiv \left\{ j \in \{1, \cdots, p\};\, \beta_j^0 \neq 0 \right\}.$$

Since $\boldsymbol{\beta^0}$ is unknown then, the set $\mathcal{A}$ is also unknown. We assume, without reducing generality, that $\mathcal{A} = \{1, \cdots, p_0\}$ and its complementary set is $\mathcal{A}^c = \{p_0 + 1, \cdots, p\}$, with $p_0 \leq p$. Hence the first p_0 explanatory variables have a significant influence on the response variable and the last $p - p_0$ variables are irrelevant. Thus, the true parameter vector can be written as $\boldsymbol{\beta^0} = \left(\boldsymbol{\beta}_{\mathcal{A}}^0, \boldsymbol{\beta}_{\mathcal{A}^c}^0\right) = \left(\boldsymbol{\beta}_{\mathcal{A}}^0, \mathbf{0}_{p-p_0}\right)$, with $\mathbf{0}_{p-p_0}$ a $(p - p_0)$-vector with all components zero. The number p_0 of the nonzero coefficients can depend on n.

For a vector $\boldsymbol{\beta}$ we use the notational convention $\boldsymbol{\beta}_{\mathcal{A}}$ for its subvector containing the corresponding components of $\mathcal{A}$. For $i = 1, \cdots, n$, we denote by $\mathbf{X}_{i,\mathcal{A}}$ the p_0-vector with the components X_{ij}, $j = 1, \cdots, p_0$. We also use the notation $|\mathcal{A}|$ or $Card(\mathcal{A})$ for the cardinality of $\mathcal{A}$.
In order to find the elements of $\mathcal{A}$, one of the most used techniques is the adaptive LASSO method, introduced for p fixed by Zou (2006) by penalizing the squares sum with an weighted L_1 penalty. This type of parameter estimator is interesting if it satisfies the oracle properties, i.e. the two following properties occur:

- *sparsity of estimation*: the non-zero parameters are estimated as non-zero and the null parameters are shrunk directly as 0, with a probability converging to 1 when $n \to \infty$;
- *asymptotic normality* of non-zero parameter estimators.

In order to distinguish between different types of adaptive LASSO estimators, we will use the term "adaptive LASSO LS-estimator" to refer to the minimizer of the LS sum penalized with adaptive LASSO.
Let us give some papers from very rich literature that consider the adaptive LASSO LS-estimator when p depends on n: Huang et al. (2008), Wang and Kulasekera (2012), Yang and Wu (2016). If the moments of the errors do not exist or the distribution of ε presents outliers, then the LS framework is not appropriate. One possibility is to consider the quantile model with the adaptive LASSO penalty. The recent literature is also very rich: Fan et al. (2014), Kaul and Koul (2015), Tang et al. (2013), Ciuperca (2019), Zheng et al. (2013), Zheng et al. (2015), to give just a few examples. As stated before, the non-differentiability of the loss function for quantile method complicates the theoretical study and its computational implementation, which is a very important aspect in high-dimensionality. Let us mention another work of Fan et al. (2017) which is also devoted to the high dimensional regression in absence of symmetry of the model errors and which proposes a penalized Huber loss but in which the sparsity of the robust approximate LASSO estimator is not studied. The robust approximate LASSO estimator is consistent with the same convergence rate as the optimal rate under the light tail situation.
In the present paper we consider the expectile loss function for a high-dimensional model. In order to introduce the expectile method, for a fixed $\tau \in (0,1)$, let us consider the function $\rho_\tau(.)$ of the form

$$\rho_\tau(x) = |\tau - \mathbb{1}_{x<0}|x^2, \qquad \text{with} \quad x \in \mathbb{R}.$$

For the error and the design of model (1) we make the following basic assumptions.
The errors ε_i satisfy the following assumption:

(A1) $(\varepsilon_i)_{1\leqslant i\leqslant n}$ are i.i.d. such that $\mathbb{E}[\varepsilon_i^4] < \infty$ and $\mathbb{E}[\varepsilon\big(\tau\mathbb{1}_{\varepsilon>0} + (1-\tau)\mathbb{1}_{\varepsilon<0}\big)] = 0$, that is its τ-th expectile is zero: $\mathbb{E}[\rho'_\tau(\varepsilon)] = 0$.

While, the design $(\mathbf{X}_i)_{1\leqslant i\leqslant n}$ satisfies the following assumption:

(A2) there exists two positive constants m_0, M_0 such that, $0 < m_0 \leq \mu_{min}\big(n^{-1}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^t\big)$ $\leq \mu_{max}\big(n^{-1}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^t\big) \leq M_0 < \infty$.

For a positive definite matrix, we denote by $\mu_{min}(.)$ and $\mu_{max}(.)$ its largest and smallest eigenvalues, respectively. Let us consider ε the generic variable for the sequence $(\varepsilon_i)_{1\leqslant i\leqslant n}$. Assumption (A1) is commonly required for the expectile models, see Zhao et al. (2018), Gu and Zou (2016), Liao et al. (2019), Newey and Powell (1987), while assumption (A2) is standard in linear model for the parameter identifiability (considered also by Gao and Huang (2010), Wang and Wang (2014), Fan et al. (2017), Zou and Zhang (2009)). Other assumptions will be stated about design in the following two sections, depending on the size p which varies in turn with n.

Quite in general, it is wise to use the expectile method when the moments of ε exist but its distribution is asymmetric. For $\tau = 0.5$, we get the classical method of least squares. For model (1), consider the expectile process

$$Q_n(\boldsymbol{\beta}) \equiv \sum_{i=1}^n \rho_\tau(Y_i - \mathbf{X}_i^t\boldsymbol{\beta}),$$

and the one with LASSO adaptive penalty:

$$R_n(\boldsymbol{\beta}) \equiv \sum_{i=1}^n \rho_\tau(Y_i - \mathbf{X}_i^t\boldsymbol{\beta}) + n\lambda_n \sum_{j=1}^p \widehat{\omega}_{n,j}|\beta_j|. \tag{2}$$

The adaptive weights $\widehat{\omega}_{n,j}$ will be defined later depending on the size of p in respect to n. The tuning parameter λ_n is a positive deterministic sequence which together with $\widehat{\omega}_{n,j}$ controls the overall model complexity. Hence, we should choose λ_n and $\widehat{\omega}_{n,j}$ such that $n\lambda_n\widehat{\omega}_{n,j} \xrightarrow[n\to\infty]{\mathbb{P}} 0$ for non-null parameters and $n\lambda_n\widehat{\omega}_{n,j} \xrightarrow[n\to\infty]{\mathbb{P}} \infty$ for null coefficients. In order to automatically detect the null and non-zero components of $\boldsymbol{\beta}$, we proceed in a similar way as for the adaptive LASSO LS-estimation introduced by Zou (2006), and we consider the adaptive LASSO expectile estimator of $\boldsymbol{\beta}$:

$$\widehat{\boldsymbol{\beta}}_n \equiv \underset{\boldsymbol{\beta}\in\mathbb{R}}{\arg\min}\, R_n(\boldsymbol{\beta}). \tag{3}$$

The components of $\widehat{\boldsymbol{\beta}}_n$ are $\widehat{\boldsymbol{\beta}}_n = \big(\widehat{\beta}_{n,1}, \cdots, \widehat{\beta}_{n,p}\big)$. Similarly to $\mathcal{A}$, let's define the index set:

$$\widehat{\mathcal{A}}_n \equiv \{j \in \{1, \cdots, p\};\ \widehat{\beta}_{n,j} \neq 0\},$$

with the non-zero components of the adaptive LASSO expectile estimator.
The estimator $\widehat{\boldsymbol{\beta}}_n$ will satisfy the *oracle properties* if:

- *sparsity*: $\lim_{n\to\infty} \mathbb{P}\big[\mathcal{A} = \widehat{\mathcal{A}}_n\big] = 1$.
- *asymptotic normality*: for any vector $\mathbf{u} \in \mathbb{R}^{p_0}$ with bounded norm, we have that: $\sqrt{n}(\mathbf{u}^t \boldsymbol{\Upsilon}_{n,\mathcal{A}}^{-1}\mathbf{u})^{-1/2}\mathbf{u}^t(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^0)_{\mathcal{A}}$ converges in distribution to a zero-mean Gaussian law, with the p_0-squared matrix: $\boldsymbol{\Upsilon}_{n,\mathcal{A}} \equiv n^{-1}\sum_{i=1}^n \mathbf{X}_{i,\mathcal{A}}\mathbf{X}_{i,\mathcal{A}}^t$.

For p fixed, the properties of the estimator $\widehat{\boldsymbol{\beta}}_n$ have been studied by Liao et al. (2019) where it is shown that the convergence rate of $\widehat{\boldsymbol{\beta}}_n$ towards $\boldsymbol{\beta^0}$ is of order $n^{-1/2}$ and that $\widehat{\boldsymbol{\beta}}_n$ satisfies the oracle properties. The case where p is fixed has also been studied by Zhao and Zhang (2018), where consider a penalized linear expectile regression with SCAD penalty function and obtain a $n^{-1/2}$-consistent estimator with oracle properties. In the present paper we assume that p depends on n, more precisely, $p = O(n^c)$, with the constant $c \in [0, 1]$. The size p_0 and the set $\mathcal{A}$ can also depend on n.
The case when p depends on n was also considered in Zhao et al. (2018) by considering the SCAD penalty for the expectile process. They propose an algorithm that converges, with probability converging to one as $n \to \infty$, to the oracle estimator after several iterations. Always for p depending on n, for (ε_i) sub-Gaussian errors, Gu and Zou (2016) penalize the expectile process with LASSO or nonconvex penalties. They find the convergence rate of the penalized estimator, propose an algorithm for finding this estimator and implement the algorithm in the R language in package *SALES*. The paper of Spiegel et al. (2017) introduces several approaches depending on selection criteria and shrinkage methods to perform model selection in semiparametric expectile regression.

Let us give some general notations. For a vector $\mathbf{v}$, we denote its transpose by $\mathbf{v}^t$, by $\|\mathbf{v}\|_1$, $\|\mathbf{v}\|_2$ and $\|\mathbf{v}\|_\infty$ the L_1, L_2, L_∞ norms, respectively. The number p of the explanatory variables and p_0 of the significant variables can depend on n, but for convenience, we do not write the subscript n. Throughout the paper, C denotes a positive generic constant that does not dependent on n, which value may differ from one formula to another.

In order to study the properties of the adapted LASSO expectile estimator $\widehat{\boldsymbol{\beta}}_n$, we introduce the following functions, using the same notations as in Liao et al. (2019):

$$g_\tau(x) \equiv \rho'_\tau(x-t)|_{t=0} = 2\tau x \mathbb{1}_{x\geq 0} + 2(1-\tau)x\mathbb{1}_{x<0},$$
$$h_\tau(x) \equiv \rho''_\tau(x-t)_{t=0} = 2\tau\mathbb{1}_{x\geq 0} + 2(1-\tau)\mathbb{1}_{x<0}$$

The paper is organized as follow. In Section 2 we study the asymptotic behaviour of the adaptive LASSO expectile estimator when $p = O(n^c)$, with $0 \leq c < 1/2$. We obtain the convergence rate of the $\widehat{\boldsymbol{\beta}}_n$ and the oracle properties. A similar study is realized in Section 3, when $c \in [1/2, 1]$. In Section 4, a simulation study and an application to real data are presented. All the proofs are relegated in Section 5.

2. Case $c < 1/2$, parameter number less than the sample size

In this section we study the asymptotic behavior of the adaptive LASSO expectile estimator when the number p of model parameter is $p = O(n^c)$, with $0 \leq c < 1/2$. If $c = 0$, that is p fixed, then we get the particular case studied by Liao et al. (2019).

An additional assumption to (A2) on the design is requested:

(A3) $p^{1/2} n^{-1/2} \max_{1 \leqslant i \leqslant n} \|\mathbf{X}_i\|_2 \underset{n \to \infty}{\longrightarrow} 0$.

The upper bound $1/2$ for c is deduced taking into account assumptions (A2) and (A3). Since $tr\big(\mathbf{X}_i^t \mathbf{X}_i\big) = tr\big(\mathbf{X}_i \mathbf{X}_i^t\big) = \big\|\mathbf{X}_i\big\|_2^2 \leq \max_{1 \leqslant i \leqslant n} \|\mathbf{X}_i\|_2^2$, taking into account assumption (A2) we have that $\max_{1 \leqslant i \leqslant n} \|\mathbf{X}_i\|_2^2 \geq Cp$. This last relation, together with assumption (A3) involve $pn^{-1/2} \to 0$, as $n \to \infty$, from where $c < 1/2$.
Because $p < n$, the regression parameters are identifiable and we can calculate the expectile estimator:

$$\widetilde{\boldsymbol{\beta}}_n \equiv \arg\min_{\boldsymbol{\beta} \in \mathbb{R}} Q_n(\boldsymbol{\beta}),$$

the components of $\widetilde{\boldsymbol{\beta}}_n$ being $\widetilde{\boldsymbol{\beta}}_n = \big(\widetilde{\beta}_{n,1}, \cdots, \widetilde{\beta}_{n,p}\big)$. This estimator will intervene in the adaptive weight of the penalty,

$$\widehat{\omega}_{n,j} = |\widetilde{\beta}_{n,j}|^{-\gamma}, \text{ for } j = 1, \cdots, p. \tag{4}$$

Conditions on the constant $\gamma > 0$ will be specified in Theorem 2.2. By the following theorem we obtain the convergence rate of the expectile and adaptive LASSO expectile estimators. We obtain that the convergence rate depends on the size p of the vector $\boldsymbol{\beta}$.

Theorem 2.1 *Under assumptions (A1), (A2) and (A3) we have:*
(i) $\|\widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^{\mathbf{0}}\|_2 = O_{\mathbb{P}}\left(\sqrt{\frac{p}{n}}\right)$.
(ii) If the tuning parameter sequence $(\lambda_n)_{n \in \mathbb{N}}$ *satisfies* $p_0^{1/2} n^{(1-c)/2} \lambda_n \to 0$, *as* $n \to \infty$, *then,* $\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^{\mathbf{0}}\|_2 = O_{\mathbb{P}}\left(\sqrt{\frac{p}{n}}\right)$.

Theorem 2.1 provides that the expectile and adaptive LASSO expectile estimators have the same convergence rate. Concerning the adaptive LASSO expectile estimator, the same convergence rate has been obtained for other adaptive LASSO estimators: by the likelihood method for a generalized linear model when $p < n$ in Wang and Wang (2014), by the least squares approximation method in Leng and Li (2010).
By the following theorem, considering a supplementary condition on λ_n, c and γ, in addition to that considered for the convergence rate in Theorem 2.1, we show that the adaptive LASSO expectile estimator $\widehat{\boldsymbol{\beta}}_n$ satisfies the oracle properties. If $\tau = 0.5$, that is, for the adaptive LASSO LS-estimator, the variance of the normal limit law is the variance of ε. In fact, we obtained for $\tau = 0.5$, the same asymptotic normality as in Zou and Zhang (2009).

Theorem 2.2 *Suppose that assumptions (A1), (A2) and (A3) hold and that the tuning parameter satisfies* $\lambda_n n^{(1-c)(1+\gamma)/2} \to \infty$, $p_0^{1/2} n^{(1-c)/2} \lambda_n \to 0$, *as* $n \to \infty$. *Then:*
(i) $\mathbb{P}\big[\widehat{\mathcal{A}}_n = \mathcal{A}\big] \to 1$, *for* $n \to \infty$.
(ii) For any vector $\mathbf{u}$ *of size* p_0 *such that* $\|\mathbf{u}\|_2 = 1$, *we have:* $n^{1/2}(\mathbf{u}^t \boldsymbol{\Upsilon}_{n,\mathcal{A}}^{-1} \mathbf{u})^{-1/2} \mathbf{u}^t (\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^0)_{\mathcal{A}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\big(0, \frac{\mathbb{V}ar[g_\tau(\varepsilon)]}{\mathbb{E}^2[h_\tau(\varepsilon)]}\big)$.

The convergence rate of $\widehat{\boldsymbol{\beta}}_n$ does not depend on the power γ, but otherwise, for holding the oracle properties, the choice of γ is very important. Concerning the suppositions and results stated in Theorem 2.2, let's make some remarks on the regularization parameter λ_n, the constants c, γ and the sizes p_0, p.

Remark 2.1 *1) If* $p_0 = O(p)$, *then if we want that* $\lambda_n n^{(1-c)(1+\gamma)/2} \to \infty$ *and* $\lambda_n p_0^{1/2} n^{(1-c)/2} \to 0$ *occur, we must choose the constant* γ *and sequence* (λ_n) *such that:* $\gamma > c/(1-c)$ *and* $n^{-1/2}\lambda_n \underset{n\to\infty}{\longrightarrow} 0$.
2) If $p_0 = O(1)$, *we must choose the constant* $\gamma > 0$ *and the tuning parameter such that* $n^{(1-c)/2}\lambda_n \underset{n\to\infty}{\longrightarrow} 0$.
3) If we want that $\lambda_n p_0^{1/2} n^{(1-c)/2} \to 0$ *holds, it is necessary that* $\lambda_n \to 0$, *as* $n \to \infty$.
4) If $c = 0$ *then the conditions on* λ_n *become:* $n^{1/2}\lambda_n \to 0$ *and* $n^{(1+\gamma)/2}\lambda_n \to \infty$, *conditions considered by Liao et al. (2019) for a linear model with* p *fixed. We also find the same variance of Gaussian distribution in Liao et al. (2019).*

Remark 2.2 *If* $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, *then* $\tau = 1/2$ *and the variance of the Normal limit law of the adaptive LASSO expectile estimators for the non-zero parameters, given by Theorem 2.2(ii), is equal to* σ^2. *The analogous result for the adaptive LASSO quantile estimator obtained by Ciuperca (2019) gives a variance of* $\pi\sigma^2/2$. *Thus, adaptive LASSO expectile estimator is more efficient than the adaptive LASSO quantile estimator.*

3. Case $c \in [1/2, 1]$

In this section, after we propose an adaptive weight, we study the asymptotic behavior of the estimator $\widehat{\boldsymbol{\beta}}_n$ when the number of regressors $p = O(n^c)$, with $1/2 \leq c \leq 1$, with the possibility that p is of the same order as the number of observations, with $p \leq n$.

Instead of assumption (A3), we consider:

(A4) There exists a constant $M > 0$ such that $\max_{1\leqslant i\leqslant n} \|\mathbf{X}_i\|_\infty < M$.

The same assumption (A4) was considered for a generalized linear model in Wang and Wang (2014) where the adaptive LASSO likelihood method is proposed.

In addition to the weight $\widehat{\omega}_{n,j}$ given by relation (4), in this section we also propose:

$$\widehat{\omega}_{n,j} = \min(|\overset{\vee}{\beta}_{n,j}|^{-\gamma}, n^{1/2}), \tag{5}$$

with $\overset{\vee}{\beta}_{n,j}$ an estimator of $\boldsymbol{\beta}_j^0$ consistent with $a_n \to 0$ the convergence rate: $\|\overset{\vee}{\boldsymbol{\beta}}_n - \boldsymbol{\beta^0}\|_2 = O_{\mathbb{P}}(a_n)$. These weights are proposed because when the estimator $\overset{\vee}{\beta}_{n,j}$ takes the value 0, then we consider $n^{1/2}$ as adaptive weight. An example of such estimator is the LASSO expectile estimator, proposed by Gu and Zou (2016), defined as:

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \left(n^{-1}\sum_{i=1}^{n} \rho_\tau(Y_i - \mathbf{X}_i^t\boldsymbol{\beta}) + \nu_n\|\boldsymbol{\beta}\|_1\right),$$

with the deterministic sequence $\nu_n \in (0,\infty)$, $\nu_n \to 0$ as $n \to \infty$. If ε_i is sub-Gaussian and $\mathbb{E}[g(\varepsilon)] = 0$, under our assumptions (A2), (A4), if $\kappa = \inf_{\mathbf{d}\in\mathcal{C}} \frac{\|\mathbb{X}\mathbf{d}\|_2^2}{\|n\mathbf{d}\|_2^2} \in (0,\infty)$, with $\mathbb{X}$ the matrix $n \times p$ of design and the set $\mathcal{C} \equiv \{\mathbf{d} \in \mathbb{R}^p; \|\mathbf{d}_{\mathcal{A}^c}\|_1 \leq 3\|\mathbf{d}_{\mathcal{A}}\|_1 \neq 0\}$, then $\|\overset{\vee}{\boldsymbol{\beta}}_n - \boldsymbol{\beta^0}\|_2 = O_{\mathbb{P}}\big(p_0^{1/2}\nu_n\big)$. Thus, the sequence (a_n) is in this case $a_n = p_0^{1/2}\nu_n$ (see Theorem 1 of Gu and Zou (2016)).

Another possibility of estimator $\overset{\vee}{\boldsymbol{\beta}}_n$ in (5) could be the expectile estimator calculated by (3). In this case, we study the convergence rate of $\widetilde{\boldsymbol{\beta}}_n$ when $p = O(n^c)$, $1/2 \leq c \leq 1$, $p \leq n$. Then, in a similar way to Theorem 2.1(i), we have the following Lemma.

Lemma 3.1 *Under assumptions (A1), (A2) and (A4), we have that* $\|\widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta^0}\|_1 = O_{\mathbb{P}}(a_n)$, *with the sequence* $(a_n)_{n\in\mathbb{N}}$ *such that* $a_n \to 0$ *and* $n^{1/2}a_n \to \infty$.

By this lemma we deduce that another estimator $\overset{\vee}{\beta}_{n,j}$ in (5) can be the expectile estimator $\widetilde{\beta}_{n,j}$, for $j = 1, \cdots, n$. We will compare in the next section, by simulations, these two possible choices of estimators in (5). Still in the next section, we will compare adaptive weights (5) with (4).
The form of the random process $R_n(\boldsymbol{\beta})$ defined by (2) and the adaptive LASSO expectile estimator $\widehat{\boldsymbol{\beta}}_n$ given by (3) remain the same, only the adaptive weight $\widehat{\omega}_{n,j}$ can change, using either (5) or (4). It would be desirable for $\widehat{\boldsymbol{\beta}}_n$ to satisfy the oracle proprties. For the sparsity property of $\widehat{\boldsymbol{\beta}}_n$ its convergence in L_1 norm is required. In the following theorem, $(b_n)_{n\in\mathbb{N}}$ is a deterministic sequence converging to 0 as $n \to \infty$. This theorem holds for the two possible choices of $\widehat{\omega}_{n,j}$.

Theorem 3.1 *Under assumptions (A1), (A2) and (A4), the tuning parameter* $(\lambda_n)_{n\in\mathbb{N}}$ *and sequence* (b_n) *satisfying* $\lambda_n p_0^{1/2} b_n^{-1} \to 0$, *as* $n \to \infty$, *we have,* $\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta^0}\|_1 = O_{\mathbb{P}}(b_n)$.

The result of Theorem 3.1 indicates that the convergence rate of the adaptive LASSO expectile estimator $\widehat{\boldsymbol{\beta}}_n$ depends on the chosen sequence $(\lambda_n)_{n\in\mathbb{N}}$. On the other hand, the convergence rate of $\widehat{\boldsymbol{\beta}}_n$ doesn't depend on the convergence rate (a_n) of the estimators $\overset{\vee}{\boldsymbol{\beta}}_n$ or $\widetilde{\boldsymbol{\beta}}_n$. The only thing that matters (see relation (23) of the proof in Section 5) is that $\overset{\vee}{\boldsymbol{\beta}}_n$ converges in probability to $\boldsymbol{\beta^0}$.
The result of Theorem 3.1 now allows us to state oracle properties. For a quantile model, Zheng et al. (2013) obtains that the convergence rate, in L_2 norm, of the adaptive LASSO quantile estimator is $(p_0/n)^{-1/2}$ and that it also satisfies the oracle properties. The sparsity of the adaptive LASSO quantile estimator will be shown in our Section 4 by simulations, where we obtain that compared to the adaptive LASSO expectile estimator, it would be necessary to have a larger number n of the observations when the model errors have an asymmetric distribution. From where, a supplementary interest in considering the expectile method instead of the quantile.

Theorem 3.2 *Suppose that assumptions (A1), (A2) and (A4) hold, the tuning parameter* (λ_n) *and sequence* (b_n) *satisfy* $\lambda_n p_0^{1/2} b_n^{-1} \to 0$, $\lambda_n b_n^{-1} \min\left(n^{1/2}, a_n^{-\gamma}\right) \to \infty$, *as* $n \to \infty$. *Then:*
(i) $\mathbb{P}\big[\widehat{\mathcal{A}}_n = \mathcal{A}\big] \to 1$, *for* $n \to \infty$.
(ii) For any vector $\mathbf{u}$ *of size* p_0 *such that* $\|\mathbf{u}\|_1 = 1$, *we have* $n^{1/2}(\mathbf{u}^t \boldsymbol{\Upsilon}_{n,\mathcal{A}}^{-1}\mathbf{u})^{-1/2}\mathbf{u}^t(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^0)_{\mathcal{A}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\big(0, \frac{\mathbb{V}ar[g_\tau(\varepsilon)]}{\mathbb{E}^2[h_\tau(\varepsilon)]}\big)$.

Hence, for $p = O(n^c)$, with $c \in [1/2, 1]$, the variance of the normal limit distribution is the same as that obtained when $p = o(n^{1/2})$. As for the case $c < 1/2$, studied in Section 2, the convergence rate of the adaptive LASSO espectile estimator doesn't depend on the power γ in the adaptive weight. However, γ intervenes in the imposed conditions so that the oracle properties are satisfied. If $\tau = 0.5$, that is for the adaptive LASSO LS-estimator, we obtained the same asymptotic normal distribution as that given by Huang et al. (2008)) for their adaptive LASSO LS-estimator.
Regarding the tuning parameter sequence we make the following remark, useful for simulations and applications on real data.

Remark 3.1 *The supposition* $\lambda_n p_0^{1/2} b_n^{-1} \to 0$ *made in the Theorems 3.1 and 3.2, implies that the tuning sequence* $\lambda_n \to 0$, *as* $n \to \infty$.

Remark 3.2 *As for the non-penalized expectile estimator, by Theorems 3.1 and 3.2, we obtained that the asymptotic variance of the adaptive LASSO expectile estimators can be calculated without going via the density of the model errors. Contrarily, the variance of the adaptive LASSO quantile estimator depends on the density function of the error (see Ciuperca (2016)). More precisely, the variance of normal distribution limit of the adaptive*

LASSO expectile estimators can be estimated without estimation of the density function of the error ε.

4. Numerical study

In this section we first perform a numerical simulation study to illustrate our theoretical results on the adaptive LASSO expectile estimation and to compare it with the estimations obtained by the adaptive LASSO quantile method. Afterwards, an application on real data is presented.
We use the following R language packages: package *SALES* with function *ernet* for the expectile regression and the package *quantreg* with function *rq* for quantile regression.
Given assumption (A1), the index τ is:

$$\tau = \frac{\mathbb{E}\left[\varepsilon \mathbb{1}_{\varepsilon<0}\right]}{\mathbb{E}\left[\varepsilon(\mathbb{1}_{\varepsilon<0} - \mathbb{1}_{\varepsilon>0})\right]}. \tag{6}$$

In the simulation study, the index τ is fixed, but it depends of the law of the model errors ε, such that assumption (A1) is satisfied.
Taking into account the suppositions imposed on the tuning parameter in Theorems 2.2 and 3.2, we consider $\lambda_n = n^{-2/5}$ for the expectile framework. For the quantile method, the tuning parameter is $n^{2/5}$ and the weight in the penalty have the power 1.225 (see Ciuperca (2016)).
In order to better compare the two methods, when the penalty weight for the adaptive LASSO expectile method is of type (4), then for adaptive LASSO quantile method is also of type (4). Similarly, if the adaptive weight is of type (5), it will be for both frameworks. In the weights (4) or (5) of the expectile penalty, the used estimations are obtained, respectively, by expectile or LASSO expectile method. For the adaptive LASSO quantile method, the estimation in the weights are obtained, respectively, by quantile or LASSO quantile method.

4.1. Simulation study: fixed p_0 case

In this subsection, we will study the numerical behavior of the adaptive LASSO expectile method and we will compare it with the simulation results obtained by the adaptive LASSO quantile method. For model (1), we consider $p_0 = 6$ and $\mathcal{A} = \{1, \cdots, 6\}$. In the all simulations of this subsection we take, $\beta_1^0 = 1$, $\beta_2^0 = 4$, $\beta_3^0 = -3$, $\beta_4^0 = 5$, $\beta_5^0 = 6$, $\beta_6^0 = -1$, while n and p are varied.
For the errors ε_i, three distributions are considered: $\mathcal{N}(0,1)$ which is symmetrical, $\mathcal{E}xp(-1)$ and $\mathcal{N}(-1.2, 0.4^2) + \chi^2(1)$, the last two being asymmetrical. In Figure 1 we give the histogram for 10^5 realizations of a $\mathcal{N}(-1.2, 0.4^2)+\chi^2(1)$ random variable, with empirical mean equal to -0.2 and median -0.628. The exponential law $\mathcal{E}xp(-1)$ has the density function $\exp(-(x+1))\mathbb{1}_{x>-1}$. For each value of n, p and distribution of ε, 1000 Monte Carlo replications are realized for two possible values for γ. In Tables 1 and 2 we give the average

Table 1: Sparsity study for expectile method with adaptive LASSO penalty, with weights of type (5), $\varepsilon \sim \mathcal{N}(0,1)$, two values for γ. Perfect method, if: $Card(\mathcal{A} \cap \widehat{\mathcal{A}}_n) = p_0 = 6$ and $Card(\widehat{\mathcal{A}}_n \setminus \mathcal{A}) = 0$.

n	p	$Card(\mathcal{A} \cap \widehat{\mathcal{A}}_n)$		$Card(\widehat{\mathcal{A}}_n \setminus \mathcal{A})$	
		$\gamma = 1/8$	$\gamma = 2$	$\gamma = 1/8$	$\gamma = 2$
50	10	5.99	5.74	0.22	0
	25	5.99	5.7	1	0.001
	50	5.99	5.64	2.13	0.002
100	10	6	5.99	0.13	0
	25	6	5.99	0.56	0
	100	6	5.98	2.5	0

of the 1000 Monte Carlo replications for the cardinalities (number of the true non-zeros estimated as non-zero) $Card(\mathcal{A} \cap \widehat{\mathcal{A}}_n)$ and $Card(\widehat{\mathcal{A}}_n \setminus \mathcal{A})$ (number of the false non-zero) by the expectile (ES) and quantile (Q) penalized methods, each with LASSO adaptive penalty. The weights of type (5) are based on the respective LASSO estimations. For a perfect method, we should have: $Card(\mathcal{A} \cap \widehat{\mathcal{A}}_n) = p_0 = 6$ and $Card(\widehat{\mathcal{A}}_n \setminus \mathcal{A}) = 0$. In Table 1, for standard Gaussian errors, the values considered for γ are $1/8$ and 2. Since for $\gamma = 1/8$ the number of false non-zeros, which in addition increases with n, is much larger than for $\gamma = 2$ and for $\gamma = 2$ the number of the true non-zeros decreases with n, these values will be dropped, other two values will be considered in Table 2. These results will be consolidated by the study on sparsity, for the three distributions of ε and different values for n and p. In Table 2, taking $\gamma \in \{5/8, 1\}$, for the adaptive LASSO expectile method with weights (5), all significant variables are detected when n is large enough, the number p of variables not coming into play. On the other hand, the smaller n, the worse the results obtained by the adaptive LASSO quantile method. In Table 3, in the adaptive weights of the expectile or quantile frameworks, we consider the expectile and quantile estimators, respectively. Comparing Tables 2 and 3 we deduce that by the quantile framework for the three error distributions and by the expectile framework for exponential and mixing error distributions, the best results are for weights of type (4) when $p \ll n$. Conversely, when $p \simeq n$, the performance is worse by using these weights for the two frameworks, the number of null parameters estimated as non null is large by the expectile method, specially for asymmetric errors. This conclusion is consolidated by Figures 2-8. When $p = n$, the R function *rq* cannot calculate the quantile estimations, since the matrix design is singular, from where the symbol "NA" (not available) in Table 3. Thus, the results obtained in Tables 2 and 3 show the advantage of using the adaptive LASSO expectile method compared to the adaptive LASSO quantile method.

We observe that the penalized expectile method better detects non-zero parameters compared to the penalized quantile method, especially for asymmetric errors. Concerning the two values considered for γ, when $\gamma = 5/8$, there are a little more true non-zeros detected, while, when $\gamma = 1$, there are fewer false non-zeros. This trend will be also confirmed by the following numerical studies.

Table 2: Sparsity study for expectile (ES) and quantile (Q) methods, with the adaptive LASSO weights of type (5). $|\mathcal{A}| = p_0 = 6$ and 1000 Monte Carlo replications. Perfect method, if: $Card(\mathcal{A} \cap \widehat{\mathcal{A}}_n) = p_0 = 6$ and $Card(\widehat{\mathcal{A}}_n \setminus \mathcal{A}) = 0$.

ε	n	p	$Card(\mathcal{A} \cap \widehat{\mathcal{A}}_n)$			$Card(\widehat{\mathcal{A}}_n \setminus \mathcal{A})$		
			ES		Q	ES		Q
			$\gamma = 5/8$	$\gamma = 1$		$\gamma = 5/8$	$\gamma = 1$	
$\mathcal{N}(-1.2, 0.4^2) + \chi^2(1)$	50	10	5.89	5.79	5.58	0.12	0.04	0.07
		25	5.84	5.75	5.58	0.60	0.25	0.30
		50	5.84	5.71	5.52	1.22	0.49	0.70
	100	10	5.99	5.99	5.97	0.05	0.004	0.03
		25	5.99	5.99	5.98	0.24	0.07	0.14
		100	5.99	5.98	5.96	1.01	0.22	0.45
	200	10	6	6	6	0.01	0.003	0.01
		100	6	6	6	0.38	0.03	0.13
		200	6	6	6	0.65	0.06	0.25
$\mathcal{E}xp(-1)$	50	10	5.98	5.95	5.95	0.01	0.005	0.03
		25	5.97	5.93	5.93	0.03	0.01	0.09
		50	5.97	5.93	5.90	0.11	0.04	0.22
	100	10	6	6	6	0	0	0.004
		25	6	6	6	0.003	0.001	0.02
		100	6	6	6	0.01	0.002	0.10
	200	10	6	6	6	0	0	0.001
		100	6	6	6	0	0	0.01
		200	6	6	6	0.003	0	0.03
$\mathcal{N}(0, 1)$	50	10	5.98	5.94	5.95	0	0.001	0.04
		25	5.97	5.95	5.94	0.048	0.01	0.16
		50	5.97	5.91	5.92	0.11	0.01	0.28
	100	10	6	6	6	0	0	0.005
		25	6	6	6	0	0	0.04
		100	6	6	6	0.01	0.01	0.16
	200	10	6	6	6	0	0	0
		100	6	6	6	0	0	0.03
		200	6	6	6	0.001	0	0.06

Table 3: Sparsity study for expectile (ES) and quantile (Q) methods, with adaptive LASSO weights of type (4). $|\mathcal{A}| = p_0 = 6$ and 1000 Monte Carlo replications. Perfect method, if: $Card(\mathcal{A} \cap \widehat{\mathcal{A}}_n) = p_0 = 6$ and $Card(\widehat{\mathcal{A}}_n \setminus \mathcal{A}) = 0$.

ε	n	p	$Card(\mathcal{A} \cap \widehat{\mathcal{A}}_n)$			$Card(\widehat{\mathcal{A}}_n \setminus \mathcal{A})$		
			ES		Q	ES		Q
			$\gamma = 5/8$	$\gamma = 1$		$\gamma = 5/8$	$\gamma = 1$	
$\mathcal{N}(-1.2, 0.4^2) + \chi^2(1)$	50	10	5.97	5.89	5.88	0.26	0.001	0.19
		25	5.94	5.84	5.80	1.56	0.02	1.35
		50	5.76	5.65	NA	9	9.9	NA
	100	10	5.99	6	5.99	0.12	0.05	0.08
		25	5.99	5.99	5.99	0.85	0.33	0.44
		100	5.92	5.8	NA	15	16	NA
	200	10	6	6	6	0.05	0.01	0.03
		100	6	6	6	1.73	0.44	1.33
		200	5.97	5.89	NA	20	18	NA
$\mathcal{E}xp(-1)$	50	10	5.99	5.99	5.99	0.03	0.01	0.06
		25	5.99	5.99	5.98	0.23	0.07	0.49
		50	5.88	5.76	NA	3.76	4.6	NA
	100	10	6	6	6	0.002	0	0.02
		25	6	6	6	0.03	.002	0.1
		100	5.95	5.86	NA	4.7	5.8	NA
	200	10	6	6	6	0	0	0.003
		100	6	6	6	0.02	0	0.21
		200	5.99	5.95	NA	3.2	2.9	NA
$\mathcal{N}(0, 1)$	50	10	6	5.99	5.99	0.02	0.007	0.06
		25	5.99	5.99	5.98	0.21	0.08	0.57
		50	5.86	5.76	NA	3.9	4.7	NA
	100	10	6	6	6	0.005	0	0.01
		25	6	6	6	0.03	0.001	0.14
		100	5.95	5.88	NA	4.5	5.4	NA
	200	10	6	6	6	0	0	0.005
		100	6	6	6	0.03	0	0.3
		200	5.98	5.96	NA	3	2.8	NA

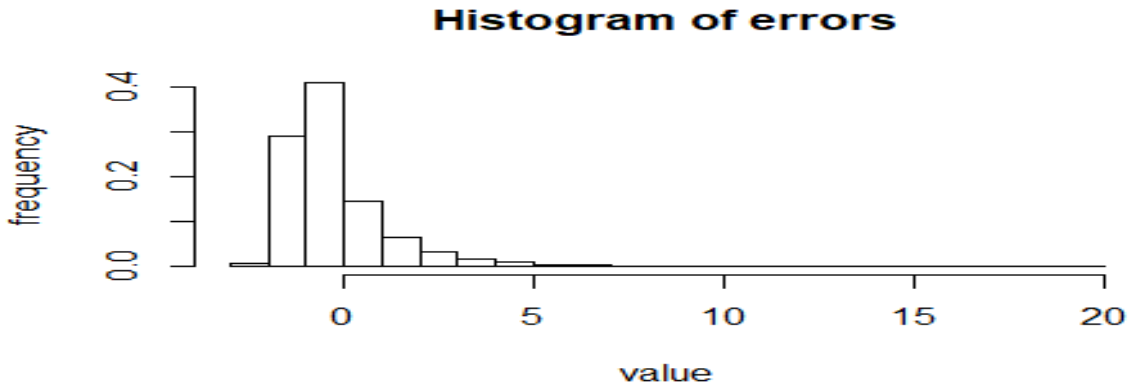

Figure 1: *Histogram for* 10^5 *realizations of a* $\mathcal{N}(-1.2, 0.4^2) + \chi^2(1)$ *random variable.*

Table 4: Study of the sparsity evolution and of the estimation accuracy for the expectile (ES) and quantile framework, with p and p_0 depending on n: $p = [n/2]$, $p_0 = 2[n^{1/2}]$. The adaptive weights are of type (5).

ε	n	$100p_0^{-1}Card(\mathcal{A} \cap \widehat{\mathcal{A}}_n)$			$100(n-p_0)^{-1}Card(\widehat{\mathcal{A}}_n \setminus \mathcal{A})$			$mean(\lvert\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta^0}\rvert)$			$mean(\lvert(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta^0})_{\mathcal{A}}\rvert)$		
		ES		Q	ES		Q	ES		Q	ES		Q
		$\gamma = 5/8$	$\gamma = 1$		$\gamma = 5/8$	$\gamma = 1$		$\gamma = 5/8$	$\gamma = 1$		$\gamma = 5/8$	$\gamma = 1$	
$\mathcal{N}(0,1)$	50	99.3	98.9	99.3	0.20	0.08	0.27	0.11	0.10	0.11	0.19	0.19	0.19
	100	99.9	99.9	99.9	0.02	0.002	0.09	0.04	0.04	0.04	0.11	0.10	0.11
	400	100	100	100	0	0	0.002	0.009	0.009	0.01	0.04	0.04	0.05
$\varepsilon \sim \mathcal{E}xp(-1)$	50	99.4	98.9	98.8	0.21	0.08	0.31	0.11	0.10	0.11	0.20	0.19	0.20
	100	99.9	99.9	99.9	0.02	0.008	0.08	0.04	0.04	0.04	0.11	0.11	0.11
	400	100	100	100	0.001	0	0.002	0.01	0.01	0.01	0.04	0.04	0.05
$\mathcal{N}(-1.2, 0.4^2) + \chi^2(1)$	50	99.8	98.3	95.6	1.24	0.14	1.03	0.15	0.04	0.23	0.26	0.126	0.40
	100	99.9	99.9	99.6	0.49	0.14	0.31	0.06	0.06	0.07	0.16	0.15	0.18
	400	100	100	100	0.04	0.002	0.02	0.01	0.01	0.01	0.07	0.07	0.07

4.2. Simulation study: case when p_0 *varies with* n

In this subsection, we always compare expectile and quantile penalized methods, but when the values considered for p vary with n. Moreover, the number of non-zero parameters can increase as n increases. In Table 4 we take $p = [n/2]$, $p_0 = 2[n^{1/2}]$, with $[x]$ the entire part of x, the power $\gamma \in \{5/8, 1\}$ and $\varepsilon \sim \mathcal{E}xp(-1)$. The true value of the non-null parameter vector is $\boldsymbol{\beta}^0_{\mathcal{A}} = (1, \cdots, p_0)$. We assess model selection by calculating the percentage $(100p_0^{-1}Card(\mathcal{A} \cap \widehat{\mathcal{A}}_n))$ of the non-zero parameters with a non-zero estimation and the percentage of false significant variables $(100(n-p_0)^{-1}Card(\widehat{\mathcal{A}}_n \setminus \mathcal{A}))$, by the two estimation methods. We also give the accuracy of the complete estimation vectors $(mean(|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta^0}|))$ and of the estimations of non-zero parameters $(mean(|(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta^0})_{\mathcal{A}}|))$ (average absolute estimation error) obtained on 1000 Monte Carlo replications. More precisely, if M is the Monte Carlo replication number and $\widehat{\boldsymbol{\beta}}^{(m)}_{n,j}$ is the estimation of β^0_j obtained for the Monte Carlo

Table 5: Study of the sparsity evolution and of the estimation accuracy for the expectile (ES) and quantile framework, with p and p_0 depending an n: $p = [n(\log n)^{-1}]$, $p_0 = 2[n^{1/4}]$, $\varepsilon \sim \mathcal{E}xp(-1)$. The adaptive weights are of type (5).

$\boldsymbol{\beta}^0_{\mathcal{A}}$	n	$100p_0^{-1}Card(\mathcal{A} \cap \widehat{\mathcal{A}}_n)$			$100(n-p_0)^{-1}Card(\widehat{\mathcal{A}}_n \setminus \mathcal{A})$			$mean(\lvert\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^{\mathbf{0}}\rvert)$			$mean(\lvert(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^{\mathbf{0}})_{\mathcal{A}}\rvert)$		
		ES		Q	ES		Q	ES		Q	ES		Q
		$\gamma = 5/8$	$\gamma = 1$		$\gamma = 5/8$	$\gamma = 1$		$\gamma = 5/8$	$\gamma = 1$		$\gamma = 5/8$	$\gamma = 1$	
$(1, 2, \cdots, p_0)$	50	99.8	99.5	99.7	0.04	0.004	0.10	0.07	0.06	0.06	0.21	0.20	0.19
	100	100	100	100	0.002	0.001	0.01	0.03	0.03	0.03	0.12	0.11	0.11
	400	100	100	100	0	0	0.0002	0.006	0.006	0.006	0.05	0.05	0.05
$(1, \cdots, 1)$	50	99.3	97.7	97.3	0.04	0.006	0.11	0.10	0.11	0.09	0.35	0.30	0.26
	100	99.98	99.97	99.98	0.005	0	0.02	0.05	0.06	0.04	0.20	0.22	0.14
	400	100	100	100	0	0	0	0.01	0.01	0.007	0.09	0.09	0.05

replication with the number m, then, $mean(|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^{\mathbf{0}}|) = (Mp)^{-1} \sum_{m=1}^{M} \sum_{j=1}^{p} |\widehat{\boldsymbol{\beta}}_{n,j}^{(m)} - \beta_j^0|$. Similarly we calculate $mean(|(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^{\mathbf{0}})_{\mathcal{A}}|) = (Mp)^{-1} \sum_{m=1}^{M} \sum_{j=1}^{p_0} |\widehat{\boldsymbol{\beta}}_{n,j}^{(m)} - \beta_j^0|$. The results are similar by the two estimation methods, while by the adaptive LASSO expectile method the results being more accurate for $\gamma = 1$ than for $\gamma = 5/8$. In Table 5, taking $p = [n(\log n)^{-1}]$, the value of p is increased compared to that considered in Table 4. Furthermore, the sparsity of the model is more accentuated by considering $p_0 = 2[n^{1/4}]$. Two values for $\boldsymbol{\beta}^0_{\mathcal{A}}$ are considered: $(1, 2, \cdots, p_0)$ and $\mathbf{1}_{p_0} = (1, \cdots, 1)$ while for the model errors, only the exponential distribution $\varepsilon \sim \mathcal{E}xp(-1)$ is made. For both values of $\boldsymbol{\beta}^{\mathbf{0}}$, the expectile and quantile methods with adaptive LASSO penalty give very good results for identifying of null and non-null parameters.
Comparing Tables 4 and 5, for n fixed and exponential law, we deduce that the penalized expectile estimation quality of the model does not vary for two different p. Furthermore, the quality is better if p_0 decreases and when $\gamma = 1$. We also deduce that the quality of automatic detection increases with n.

In Table 6 we give the median run time, for 1000 simulation replicates, of R functions: "*ernet*" for calculate the adaptive LASSO expectile estimation and "*rq*" for adaptive LASSO quantile estimation, for a linear model with $|\mathcal{A}| = p_0 = 6$, $\varepsilon \sim \mathcal{E}xp(-1)$, $\gamma = 1$ and adaptive LASSO weights of type (4). The algorithm used for "*ernet*" has been proposed by Gu and Zou (2016) and combines the cyclic coordinate descent and proximal gradient algorithms. The calculation time of the adaptive LASSO quantile estimate is much longer than that of the penalized expectile method, especially when the number p of parameters is close to the sample size n.

4.3. Sparsity study function of γ

In Figure 2 we present the results of the percentages of the true non-zero and false zero estimators by adaptive LASSO expectile method, for a model with $n = p = 100$ and Gaussian estimators. The true parameters are $\beta_1^0 = 1$, $\beta_2^0 = 4$, $\beta_3^0 = -3$, $\beta_4^0 = 5$, $\beta_5^0 = 6$, $\beta_6^0 = -1$ and $\beta_j^0 = 0$ for any $j > 6$. The results are better when in the adaptive weights

Table 6: Run time (in seconds) of R functions: "ernet" corresponding to the expectile method and "rq" for quantile method, both being LASSO penalized with LASSO penalty, for a model: $|\mathcal{A}| = p_0 = 6$, $\varepsilon \sim \mathcal{E}xp(-1)$, $\gamma = 1$, adaptive LASSO weights of type (4).

n	p	*time by adaptive LASSO expectile (by "ernet")*	*time by adaptive LASSO quantile (by "rq")*
50	10	$9.9 \cdot 10^{-4}$	$9.9 \cdot 10^{-4}$
	25	$9.9 \cdot 10^{-4}$	$9.9 \cdot 10^{-4}$
200	10	$9.9 \cdot 10^{-4}$	10^{-3}
	100	$2 \cdot 10^{-3}$	10^{-2}
500	10	$9.9 \cdot 10^{-4}$	$1.9 \cdot 10^{-3}$
	250	$2.9 \cdot 10^{-3}$	$1.1 \cdot 10^{-1}$
1000	10	10^{-3}	$2.99 \cdot 10^{-3}$
	500	10^{-2}	0.93
10^4	10	$4.9 \cdot 10^{-3}$	0.03
	10^3	0.29	38.9

we take the LASSO expectile estimator. Following this results, right now, for weight (5) we consider only LASSO expectile estimator.
In Figures 3-8, for $n = 100$, $p \in \{10, 100\}$, $\varepsilon \sim \mathcal{N}(0,1)$ or $\mathcal{E}xp(-1)$ or $\mathcal{N}(-1.2, 0.4^2) + \chi^2(1)$, we present the percentage of true and false zero, when for the adaptive weight we take (5) by dotted line, and (4) with solid line. We deduce that for $p \ll n$ we must take weights (4), for $p \simeq n$ weights (5), choosing in the two cases $\gamma \in [1/2, 1]$.
In Figures 9 and 10, for $\varepsilon \sim \mathcal{N}(-1.2, 0.4^2) + \chi^2(1)$, for the same n (either 50 or 100), we vary p. Then, when $p = n$ we consider adaptive weight (5) with dotted line and in the other case we take (4) with solid line. We deduce that we must take different weights, function that p is of the same order as n or much smaller.

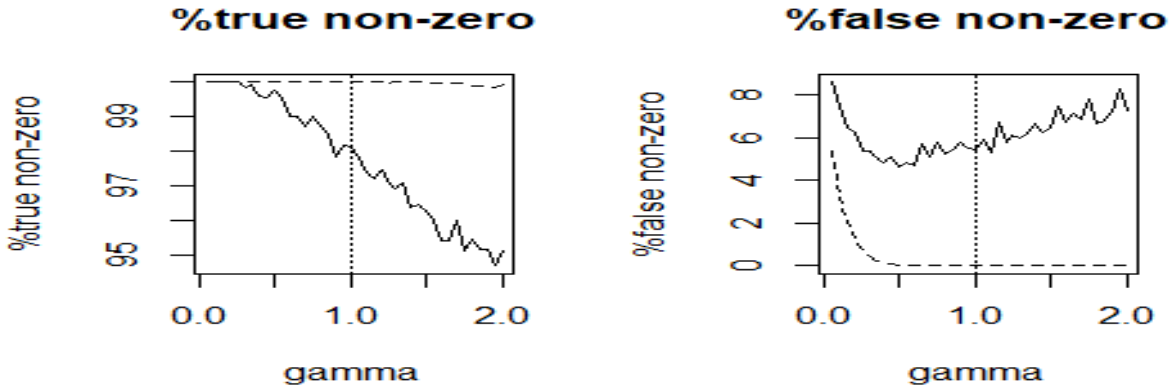

Figure 2: *Percentage of true and false non zero, for* $\varepsilon \sim \mathcal{N}(0,1)$, $n = 100$ *and* $p = 100$. *In adaptive weights (5), two estimators: LASSO expectile with dotted line and expectile with solid line.*

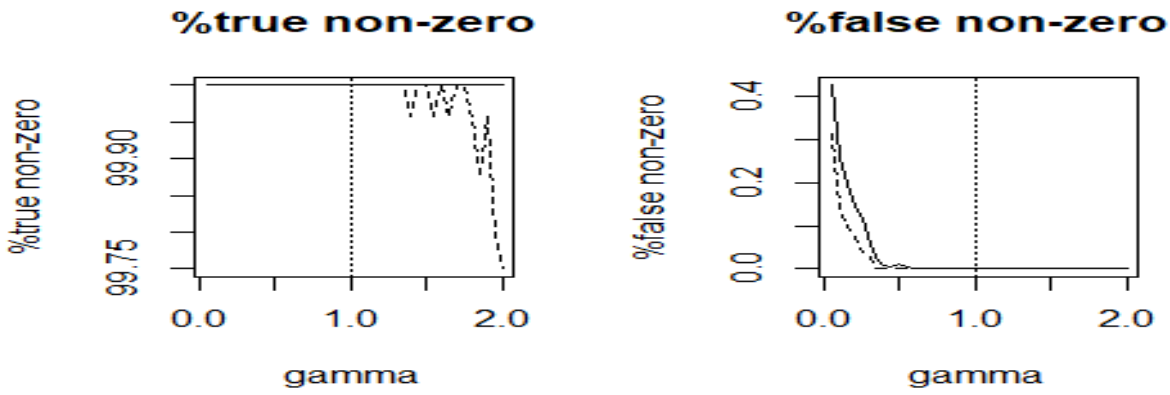

Figure 3: *Percentage of true and false non zero, for* $\varepsilon \sim \mathcal{N}(0,1)$, $n = 100$ *and* $p = 10$. *Two adaptive weights: (5) with dotted line and (4) with solid line.*

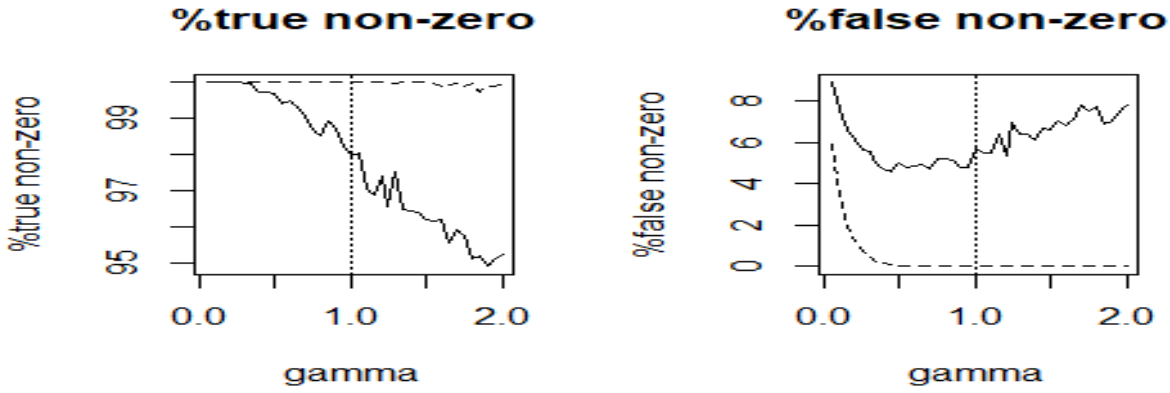

Figure 4: *Percentage of true and false non zero, for* $\varepsilon \sim \mathcal{N}(0,1)$, $n = 100$ *and* $p = 100$. *Two adaptive weights: (5) with dotted line and (4) with solid line.*

4.4. Conclusion of simulation study

For adaptive LASSO expectile estimation, when $p \ll n$, the best results are obtained for adapted weights (4) and when $p \simeq n$ for weights (5). The penalized expectile method is more accurate when n is small and the error distributions are asymmetric.
When p and p^0 increase with n, the adaptive LASSO expectile estimations identify very well the null and non-zero parameters, while the adaptive LASSO quantile estimations happen to have the same performances only for very large n. The calculation time of the adaptive LASSO quantile estimate is much longer than that of the adaptive LASSO

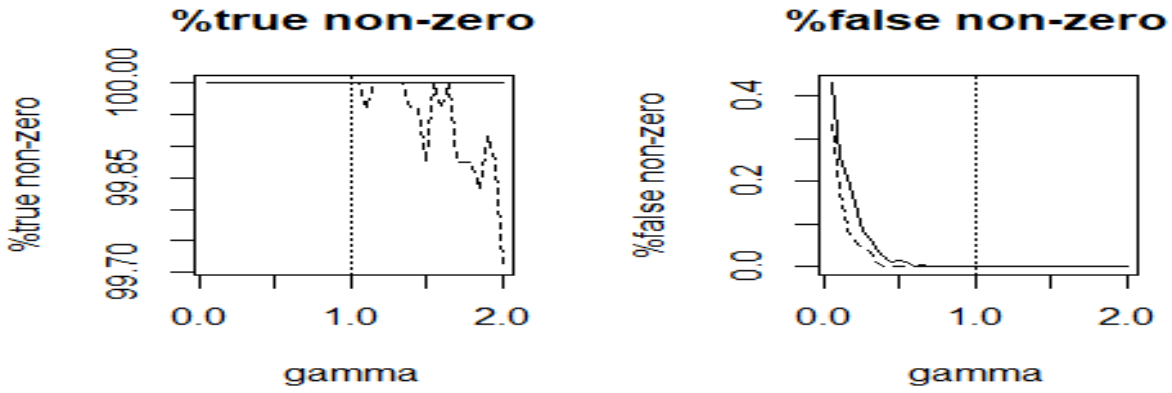

Figure 5: *Percentage of true and false non zero, for* $\varepsilon \sim \mathcal{E}xp(-1)$, $n = 100$ *and* $p = 10$. *Two adaptive weights: (5) with dotted line and (4) with solid line.*

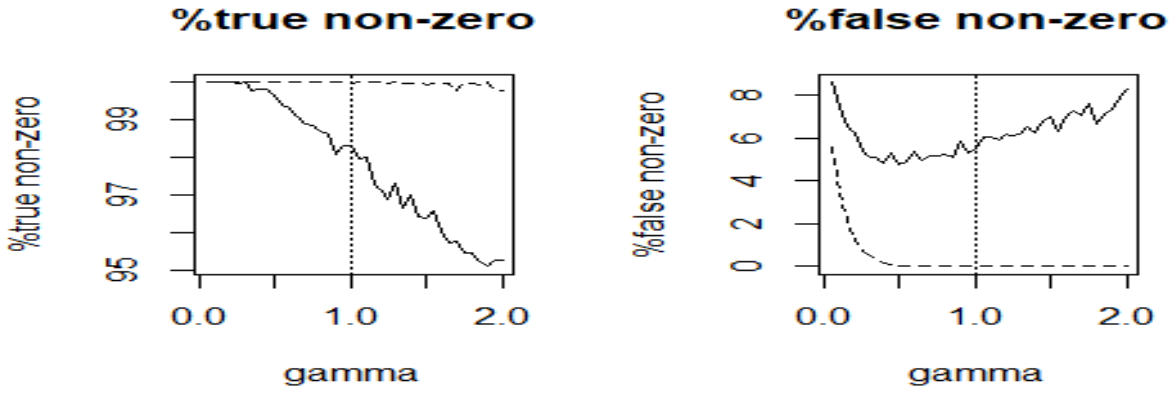

Figure 6: *Percentage of true and false non zero, for* $\varepsilon \sim \mathcal{E}xp(-1)$, $n = 100$ *and* $p = 100$. *Two adaptive weights: (5) with dotted line and (4) with solid line.*

expectile method, especially when the number p of parameters is close to n. On the other hand, the adaptive LASSO quantile method identifies true zeros less well when the number of observations is small. Moreover, another disadvantage of the penalized quantile method is that when the number of parameters is close to the number of observations, the R function cannot calculate numerically the quantile estimations.
In all cases, the most appropriate the power γ which intervenes in adaptive weights of the penalized expectile method is $\gamma \in [1/2, 1]$.

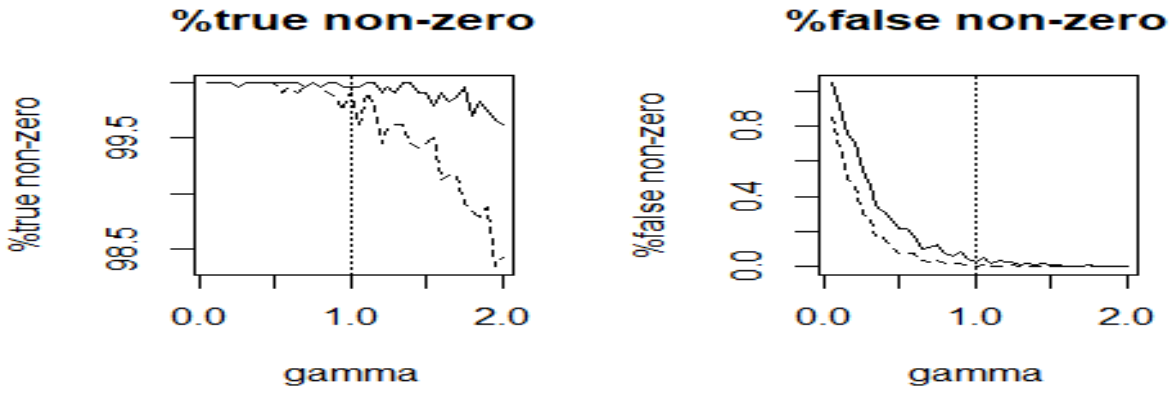

Figure 7: *Percentage of true and false non zero, for* $\varepsilon \sim \mathcal{N}(-1.2, 0.4^2) + \chi^2(1)$, $n = 100$ *and* $p = 10$. *Two adaptive weights: (5) with dotted line and (4) with solid line.*

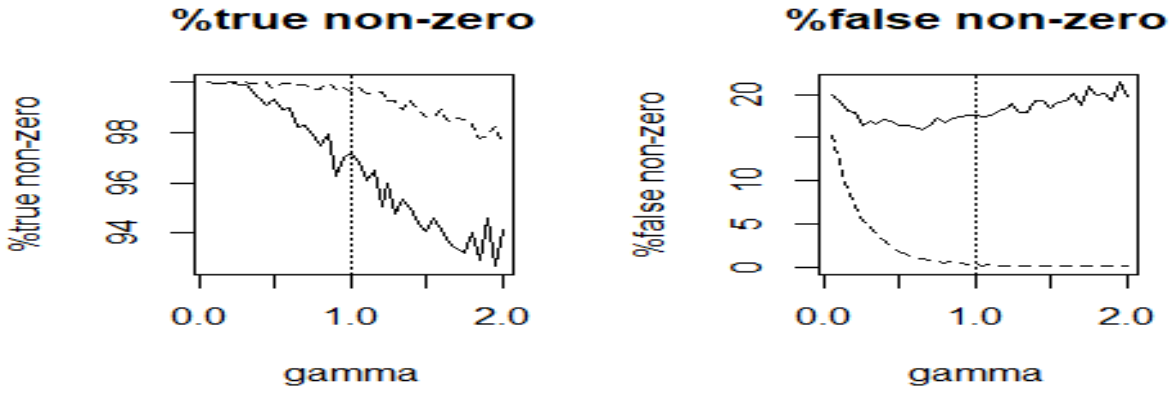

Figure 8: *Percentage of true and false non zero, for* $\varepsilon \sim \mathcal{N}(-1.2, 0.4^2) + \chi^2(1)$, $n = 100$ *and* $p = 100$. *Two adaptive weights: (5) with dotted line and (4) with solid line.*

4.5. Application to real data

We use the data *eyedata* of R package *flare* which contains $n = 120$ observations (rats) for the response variable of gene *TRIM32* and 200 explanatory variables, other genes probes, from the microarray experiments of mammalian-eye tissue samples in Scheetz et al. (2006). The objective is to find genes that are correlated with the TRIM32 gene, known to cause Bardet–Biedl syndrome, a genetically disease of multiple organ systems including the retina.

In this paper, the expectile index τ is supposed to be known, such that for model error ε we

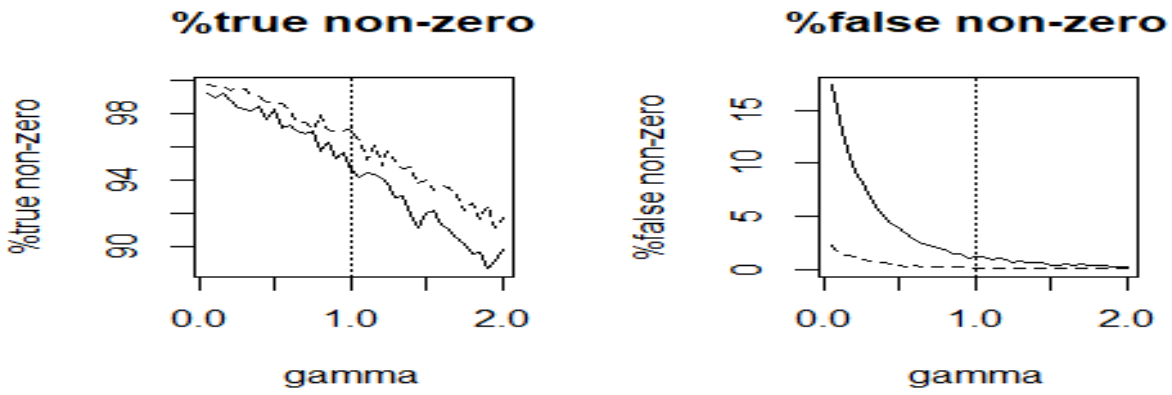

Figure 9: *Percentage of true and false non zero, for* $\varepsilon \sim \mathcal{N}(-1.2, 0.4^2) + \chi^2(1)$, $n = 50$. *Two value for the number of parameters:* $p = 10$ *together adaptive weight (5), with dotted line and* $p = 50$ *together adaptive weight (4), with solid line.*

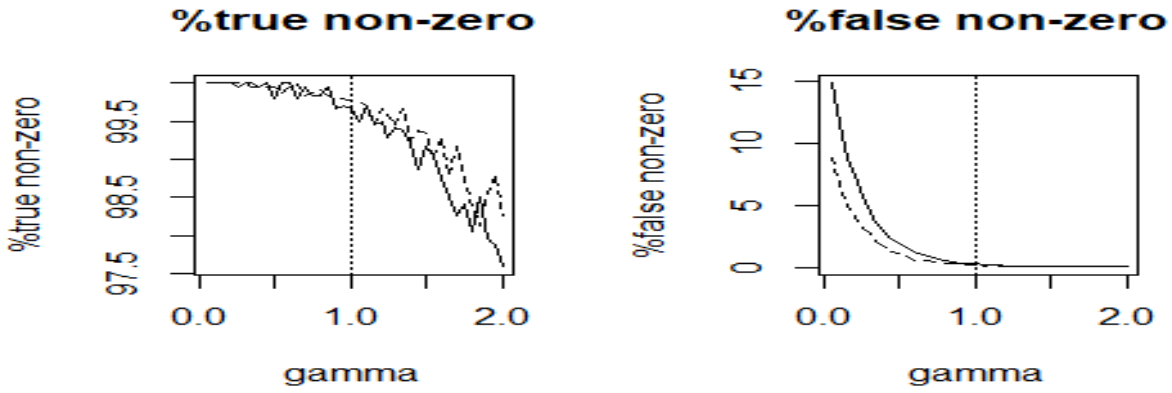

Figure 10: *Percentage of true and false non zero, for* $\varepsilon \sim \mathcal{N}(-1.2, 0.4^2) + \chi^2(1)$, $n = 100$. *Two value for the number of parameters:* $p = 50$ *together adaptive weight (5), represented with dotted line and* $p = 100$ *together adaptive weight (4), represented with solid line.*

have $\mathbb{E}[g_\tau(\varepsilon)] = 0$. On the other hand, in applications, an estimate should be given for τ, with the remark that same related question also exists in quantile regression for estimating the quantile level $\tilde{\tau}$ such that $I\!P[\varepsilon < 0] = \tilde{\tau}$. But to estimate τ we must first calculate the model errors and to calculate the errors we must know τ. Therefore, we will not give an exact estimator of τ but we will try to prefix the value of τ such that the dispersion around their median of the values of $(y_i)_{1 \leqslant i \leqslant n}$ is taken into account. We propose the

following approach in order to give a prefixed value for the expectile index τ. The values of the explained variable are first transformed, for $i = 1, \cdots, n$, $\widetilde{y}_i = (y_i - M_n)/D_n$, with $M_n = median(y_1, \cdots, y_n)$ and $D_n = n^{-1} \sum_{i=1}^{n} |y_i - M_n|$ the empirical mean of the absolute values of the distances to the median of the y_i. Afterwards, based on relation (6), we calculate the empirical estimation of τ for $\widetilde{y}_i$:

$$\widehat{\tau} = \frac{n^{-1} \sum_{i=1}^{n} \widetilde{y}_i \mathbb{1}_{\widetilde{y}_i<0}}{n^{-1}\big(\sum_{i=1}^{n} \widetilde{y}_i \mathbb{1}_{\widetilde{y}_i<0} - \sum_{i=1}^{n} \widetilde{y}_i \mathbb{1}_{\widetilde{y}_i>0}\big)}.$$

Then, in model (1), the response variable is $\widetilde{Y}_i = (Y_i - M_n)/D_n$. For this application, we get $\widehat{\tau} = 0.568$ and $\gamma = 5/8$.
We first consider two linear models (the first on the first hundred explanatory variables and the second on the other hundred explanatory variables), for which we consider adaptive weights of type (5), for $\gamma = 5/8$, $\lambda_n = n^{2/5}$. Afterwards we consider a linear model with explanatory variables the relevant regressors selected for the two preceding regressions. For this model, we select the relevant variables by the adaptive LASSO expectile method with the weights of type (4). Then, we obtain that the genes whose expressions influence gene TRIM32 are 87, 153, 180, 185, 200 with the labels: "21092", "25141", "28680", "28967" and "30141". The obtained estimations for the coefficients of these four explanatory variables are respectively: *-1.06, 2.97, 1.29, -1.727* and *-0.23*. In Figure 11 we illustrate the histogram and the box-plot for response variable $\widetilde{Y}_i$. We observe that there are outliers.
If a classical LS regression of the TRIM32 variable in respect to the five selected covariates is performed, we obtain a model with an adjusted $R^2 = 0.75$. The all five variables are significant and the residuals have Gaussian distribution (the p-value by Shapiro test equal to 0.69). In the sub-figure of the right-hand side of Figure 11, we also present, the forecasts beside of the true values of TRIM32. We observe that the scater graph is on the first bisectrix.
By the adaptive LASSO quantile method, no variable is selected among the 200 explanatory ones.
In literature works that model the same data, variable number 153, tagged "25141", has been selected as the sole regressor by the bayesian shrinkage in Song and Liang (2017) and by a globally adaptive quantile method in Zheng et al. (2015) for quantile index between 0.45 and 0.55. In this last paper, there are other covariates that appear to be significant for other quantile index values. These variables are: "11711", "24565", "25141", "25367", "21092", "29045", "25439", "22140", "15863" and "6222". If we make a classic regression for these ten regressors, we obtain, with a risk of 0.05 that only the variables "25141", "21092", "29045", "15863", "6222" are significant, in a model of lower quality (adjusted $R^2 = 0.72$, residual standard error=0.52) than the one with the five explanatory variables found in the present paper.
In order to complete the comparison of the three methods, we split the database of 120 observations into two: one for learning and the other for testing. The three methods

Table 7: Means of residual absolute values (RAV) by adaptive LASSO expectile estimation, LS on explanatory variables of Zheng et al. (2015), and by adaptive LASSO quantile estimation on all data, learning and test data.

estimation method	RAV on all data	RAV on learning data	RAV on test data
adaptive LASSO expectile	0.057	0.056	0.068
Zheng et al. (2015) method	0.059	0.059	0.064
adaptive LASSO quantile	0.75	0.787	0.771

are each calibrated on 105 observations and are tested on 15 observations. The empirical means of residual absolute values (RAV) are presented in Table 7. The forecast accuracy of the explained variable is very similar by the adaptive LASSO expectile and by the LS estimation on explanatory variables of Zheng et al. (2015), while by the adaptive LASSO quantile method it is less good, also because of the fact that no explanatory variable is significant for the latter. From this table, we also deduce that the method proposed in the present paper is also robust because we get the same accuracy on the test and learning set.

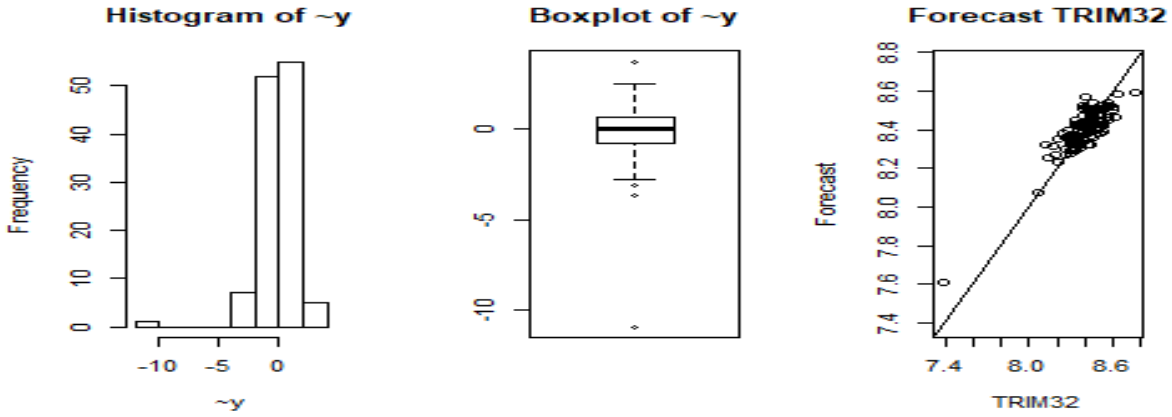

Figure 11: *Histogram and boxplot of* $(\widetilde{y}_i)_{1\leqslant i\leqslant 120}$. *Scater graph between forecast and the true value of TRIM32.*

5. Proofs

In this section we give the proofs of the results presented in Sections 2 and 3.

5.1. Result proofs in Section 2

Before presenting the proof of Theorem 2.1, let's recall a result given in Gu and Zou (2016).

Lemma 1 of Gu and Zou (2016). *For any $z, z_0 \in \mathbb{R}$ and $\tau \in (0,1)$ we have:*

$$\min(\tau, 1-\tau)(z-z_0)^2 \leq \rho_\tau(z) - \rho_\tau(z_0) - g_\tau(z_0)(z-z_0) \leq \max(\tau, 1-\tau)(z-z_0)^2.$$

Proof of Theorem 2.1.
(i) In order to show the convergence rate of the expectile estimator, we show that for all $\epsilon > 0$, there exists $B_\epsilon > 0$, large enough when n is large, such that:

$$\mathbb{P}\big[\inf_{\mathbf{u}\in\mathbb{R}^p,\ \|\mathbf{u}\|_2=1} Q_n\big(\boldsymbol{\beta^0} + B_\epsilon\sqrt{\frac{p}{n}}\mathbf{u}\big) > Q_n(\boldsymbol{\beta^0})\big] \geq 1-\epsilon. \tag{7}$$

This part of proof is similar to that of Lemma 1.2 in Zhao et al. (2018), for the convergence rate of the oracle estimator. Let $B > 0$ be a constant to be determined later and $\mathbf{u}$ a vector in $\mathbb{R}^p$ with the norm $\|\mathbf{u}\|_2 = 1$. Let's study the difference:

$$Q_n\big(\boldsymbol{\beta^0} + B\sqrt{\frac{p}{n}}\mathbf{u}\big) - Q_n(\boldsymbol{\beta^0}) = \sum_{i=1}^n \big[\rho_\tau\big(\varepsilon_i - B\sqrt{\frac{p}{n}}\mathbf{X}_i^t\mathbf{u}\big) - \rho_\tau(\varepsilon_i)\big]$$

$$= \sum_{i=1}^n \big[\rho_\tau\big(\varepsilon_i - B\sqrt{\frac{p}{n}}\mathbf{X}_i^t\mathbf{u}\big) - \rho_\tau(\varepsilon_i) - \mathbb{E}\big[\rho_\tau\big(\varepsilon_i - B\sqrt{\frac{p}{n}}\mathbf{X}_i^t\mathbf{u}\big) - \rho_\tau(\varepsilon_i)\big]\big]$$
$$+ \sum_{i=1}^n \mathbb{E}\big[\rho_\tau\big(\varepsilon_i - B\sqrt{\frac{p}{n}}\mathbf{X}_i^t\mathbf{u}\big) - \rho_\tau(\varepsilon_i)\big] \equiv \Delta_1 + \Delta_2.$$

We first study the term Δ_2. By Taylor expansion, we have: $\mathbb{E}[\rho_\tau(\varepsilon - t) - \rho_\tau(\varepsilon)] = \mathbb{E}\big[-g_\tau(\varepsilon)t + 2^{-1}h_\tau(\varepsilon)t^2\big] + o(t^2) = 2^{-1}\mathbb{E}\big[h_\tau(\varepsilon)\big]t^2 + o(t^2)$. By the Cauchy-Schwarz inequality, we have that $|\mathbf{X}_i^t\mathbf{u}|^2 \leq \|\mathbf{X}_i\|_2^2\|\mathbf{u}\|_2^2$ and then, using assumption (A3), we obtain that $(pn^{-1})^{1/2}\max_{1\leqslant i\leqslant n}|\mathbf{X}_i^t\mathbf{u}| \underset{n\to\infty}{\longrightarrow} 0$. Thus,

$$\Delta_2 = \frac{1}{2}\sum_{i=1}^n \bigg[B^2\frac{p}{n}\big(\mathbf{X}_i^t\mathbf{u}\big)^2\mathbb{E}\big[h_\tau(\varepsilon)\big] + o\bigg(B^2\frac{p}{n}\big(\mathbf{X}_i^t\mathbf{u}\big)^2\mathbb{E}\big[h_\tau(\varepsilon)\big]\bigg)\bigg].$$

On the other hand,

$$2\min(\tau, 1-\tau) \leq \mathbb{E}\big[h_\tau(\varepsilon)\big] = 2\tau\mathbb{E}[\mathbb{1}_{\varepsilon\geq 0}] + 2(1-\tau)\mathbb{E}[\mathbb{1}_{\varepsilon<0}] \leq 2\max(\tau, 1-\tau).$$

Then

$$0 < \Delta_2 = B^2\frac{p}{n}\mathbb{E}\big[h_\tau(\varepsilon)\big]\sum_{i=1}^n \big(\mathbf{X}_i^t\mathbf{u}\big)^2\big(1+o(1)\big) = O(B^2p). \tag{8}$$

We are now studying the term Δ_1. Let us consider the random variable:

$$D_i \equiv \rho_\tau\big(\varepsilon_i - B\sqrt{\frac{p}{n}}\mathbf{X}_i^t\mathbf{u}\big) - \rho_\tau(\varepsilon_i) + B\sqrt{\frac{p}{n}}g_\tau(\varepsilon_i)\mathbf{X}_i^t\mathbf{u}. \tag{9}$$

Then, we can write Δ_1 as:

$$\Delta_1 = \sum_{i=1}^{n} \big[- B\sqrt{\frac{p}{n}} g_\tau(\varepsilon_i)\mathbf{X}_i^t\mathbf{u} + D_i - \mathbb{E}[D_i]\big].$$

Using assumption (A1) we have, $\mathbb{E}\big[(pn^{-1})^{1/2} g_\tau(\varepsilon_i)\mathbf{X}_i^t\mathbf{u}\big] = 0$ and $\mathbb{V}\text{ar}\,\big[(pn^{-1})^{1/2}\sum_{i=1}^n g_\tau(\varepsilon_i)\mathbf{X}_i^t\mathbf{u}\big] = pn^{-1}\mathbf{u}^t\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^t\mathbf{u}\mathbb{V}\text{ar}\,[g_\tau(\varepsilon)] = O(p)$. Then, we have:

$$B\sqrt{\frac{p}{n}}\sum_{i=1}^{n} g_\tau(\varepsilon_i)\mathbf{X}_i^t\mathbf{u} = O_{\mathbb{P}}\big(p^{1/2}B\big). \tag{10}$$

By Lemma 1 of Gu and Zou (2016) for $z = \varepsilon_i - B(pn^{-1})^{1/2}\mathbf{X}_i^t\mathbf{u}$ and $z_0 = \varepsilon_i$, we can write D_i also in the form: $D_i = B^2pn^{-1}|\mathbf{X}_i^t\mathbf{u}|^2V_i$, with V_i a random variable between $\min(\tau, 1-\tau)$ and $\max(\tau, 1-\tau)$ with probability one:

$$\mathbb{P}\big[\min(\tau, 1-\tau) \le V_i \le \max(\tau, 1-\tau)\big] = 1. \tag{11}$$

Then, $\mathbb{V}\text{ar}\,[D_i] \le \mathbb{E}[D_i^2] = B^4(p/n)^{1/2}|\mathbf{X}_i^t\mathbf{u}|^4\mathbb{E}\big[V_i^2\big]$. But $\mathbb{E}\big[V_i^2\big] \le 1$ and thus $\mathbb{V}\text{ar}\,[D_i] \le B^4p^2n^{-2}|\mathbf{X}_i^t\mathbf{u}|^4$. On the other hand, the random variables D_i defined by (9), are independent. Then,

$$\sum_{i=1}^{n}\big[D_i - \mathbb{E}[D_i]\big] = O_{\mathbb{P}}\bigg(\sqrt{\sum_{i=1}^{n}\mathbb{V}\text{ar}\,[D_i]}\bigg) \le O_{\mathbb{P}}\bigg(\sqrt{\sum_{i=1}^{n}\mathbb{E}[D_i^2]}\bigg) = O_{\mathbb{P}}\bigg(B^2\frac{p}{\sqrt{n}}\bigg). \tag{12}$$

Relations (10) and (12), imply, since $pn^{-1} \to 0$, when $n \to \infty$, that:

$$\Delta_1 = O_{\mathbb{P}}(Bp^{1/2}) + O_{\mathbb{P}}(B^2pn^{-1/2}) = O_{\mathbb{P}}(Bp^{1/2}).$$

Then, this last relation together relation (8) imply $\Delta_2 > |\Delta_1|$, with probability converging to one, for B large enough. Relation (7) follows, which implies the convergence rate of the expectile estimator.

(ii) For p-vector $\mathbf{u} = (u_1, \cdots, u_p)$, with $\|\mathbf{u}\|_2 = 1$ and $B > 0$ a constant, let us consider the difference

$$R_n\big(\boldsymbol{\beta^0}+B\sqrt{\frac{p}{n}}\mathbf{u}\big)-R_n(\boldsymbol{\beta^0}) = Q_n\big(\boldsymbol{\beta^0}+B\sqrt{\frac{p}{n}}\mathbf{u}\big)-Q_n(\boldsymbol{\beta^0})+n\lambda_n\sum_{j=1}^{p}\widehat{\omega}_{n,j}\big[\big|\boldsymbol{\beta}_j^0+B\sqrt{\frac{p}{n}}u_j\big|-|\boldsymbol{\beta}_j^0|\big]. \tag{13}$$

The first term of the right-hand side of (13) becomes by the above proof for *(i)*,

$$Q_n\big(\boldsymbol{\beta^0} + B\sqrt{\frac{p}{n}}\mathbf{u}\big) - Q_n(\boldsymbol{\beta^0}) = B^2\frac{p}{n}\mathbb{E}\big[h_\tau(\varepsilon)\big]\sum_{i=1}^{n}(\mathbf{X}_i^t\mathbf{u})^2\big(1+o_{\mathbb{P}}(1)\big) = O_{\mathbb{P}}(B^2p). \tag{14}$$

Furthermore, for the penalty of (13) we have:

$$n\lambda_n \sum_{j=1}^{p} \widehat{\omega}_{n,j}\big[|\boldsymbol{\beta}_j^0+B\sqrt{\frac{p}{n}}u_j|-|\boldsymbol{\beta}_j^0|\big] \geq n\lambda_n \sum_{j=1}^{p_0} \widehat{\omega}_{n,j}\big[|\boldsymbol{\beta}_j^0+B\sqrt{\frac{p}{n}}u_j|-|\boldsymbol{\beta}_j^0|\big] \geq -n\lambda_n \sum_{j=1}^{p_0} \widehat{\omega}_{n,j} B\sqrt{\frac{p}{n}}|u_j|.$$

By the Cauchy-Schwarz inequality and afterwards by *(i)*, we have

$$\geq -n\lambda_n B\sqrt{\frac{p}{n}}\Big(\sum_{j=1}^{p_0}\widehat{\omega}_{n,j}^2\Big)^{1/2}\|\mathbf{u}\|_2 = -B\sqrt{\frac{p}{n}}np_0^{1/2}\lambda_n = -Bp_0^{1/2}n^{(1+c)/2}\lambda_n. \tag{15}$$

Since $p_0^{1/2}n^{(1-c)/2}\lambda_n \to 0$, as $n \to \infty$, we obtain that relation (14) dominates (15) and the assertion regarding the convergence rate of $\widehat{\boldsymbol{\beta}}_n$ results. ■

Proof of Theorem 2.2.
(i) Let us consider the parameter set: $\mathcal{V}_p(\boldsymbol{\beta^0}) \equiv \big\{\boldsymbol{\beta} \in \mathbb{R}^p; \|\boldsymbol{\beta}-\boldsymbol{\beta^0}\|_2 \leq B\sqrt{\frac{p}{n}}\big\}$, with $B>0$ large enough and $\mathcal{W}_n \equiv \big\{\boldsymbol{\beta} \in \mathcal{V}_p(\boldsymbol{\beta^0}); \|\boldsymbol{\beta}_{\mathcal{A}^c}\|_2 > 0\big\}$. According to Theorem 2.1, the estimator $\widehat{\boldsymbol{\beta}}_n$ belongs to the set $\mathcal{V}_p(\boldsymbol{\beta^0})$ with a probability converging to 1 as $n \to \infty$. In order to show the sparsity property of claim *(i)*, we will show that, $\lim_{n\to\infty}\mathbb{P}\big[\widehat{\boldsymbol{\beta}}_n \in \mathcal{W}_n\big] = 0$. Note that if $\boldsymbol{\beta} \in \mathcal{W}_n$, then $p > p_0$.
Let us consider two parameter vectors: $\boldsymbol{\beta} = (\boldsymbol{\beta}_{\mathcal{A}}, \boldsymbol{\beta}_{\mathcal{A}^c}) \in \mathcal{W}_n$ and $\boldsymbol{\beta}^{(1)} = (\boldsymbol{\beta}_{\mathcal{A}}^{(1)}, \boldsymbol{\beta}_{\mathcal{A}^c}^{(1)}) \in \mathcal{V}_p(\boldsymbol{\beta^0})$, such that $\boldsymbol{\beta}_{\mathcal{A}}^{(1)} = \boldsymbol{\beta}_{\mathcal{A}}$ and $\boldsymbol{\beta}_{\mathcal{A}^c}^{(1)} = \mathbf{0}_{p-p_0}$. For this parameters, we will study the following difference:

$$n^{-1}\big[R_n(\boldsymbol{\beta}) - R_n(\boldsymbol{\beta}^{(1)})\big] = n^{-1}\big[Q_n(\boldsymbol{\beta}) - Q_n(\boldsymbol{\beta}^{(1)})\big] + \lambda_n \sum_{j=p_0+1}^{p} \widehat{\omega}_{n,j}|\beta_j|. \tag{16}$$

First, note that by elementary calculations, we get: $\rho_\tau(\varepsilon - t) = \rho_\tau(\varepsilon) - g_\tau(\varepsilon)t + \frac{1}{2}h_\tau(\varepsilon)t^2 + o_{\mathbb{P}}(t^2)$ for $t \to 0$. Then, for the first term of the right-hand side of relation (16), we have

$$\begin{aligned}
&n^{-1}\sum_{i=1}^{n}\big[\rho_\tau(Y_i - \mathbf{X}_i^t\boldsymbol{\beta}^{(1)}) - \rho_\tau(Y_i - \mathbf{X}_i^t\boldsymbol{\beta})\big] \\
&= n^{-1}\sum_{i=1}^{n}\big[\rho_\tau\big(\varepsilon_i - \mathbf{X}_{i,\mathcal{A}}^t(\boldsymbol{\beta}_{\mathcal{A}} - \boldsymbol{\beta}_{\mathcal{A}}^0)\big) - \rho_\tau\big(\varepsilon_i - \mathbf{X}_{i,\mathcal{A}}^t(\boldsymbol{\beta}_{\mathcal{A}} - \boldsymbol{\beta}_{\mathcal{A}}^0) - \mathbf{X}_{i,\mathcal{A}^c}^t\boldsymbol{\beta}_{\mathcal{A}^c}\big)\big] \\
&= n^{-1}\sum_{i=1}^{n}\big[g(\varepsilon_i)\mathbf{X}_{i,\mathcal{A}}^t\big(\boldsymbol{\beta}_{\mathcal{A}} - \boldsymbol{\beta}_{\mathcal{A}}^0\big) + \frac{h(\varepsilon_i)}{2}\big(\mathbf{X}_{i,\mathcal{A}}^t(\boldsymbol{\beta}_{\mathcal{A}} - \boldsymbol{\beta}_{\mathcal{A}}^0)\big)^2 + o_{\mathbb{P}}\big(\mathbf{X}_{i,\mathcal{A}}^t(\boldsymbol{\beta}_{\mathcal{A}} - \boldsymbol{\beta}_{\mathcal{A}}^0)\big)^2\big] \\
&- n^{-1}\sum_{i=1}^{n}\big[g(\varepsilon_i)\mathbf{X}_i^t\big(\boldsymbol{\beta}_{\mathcal{A}} - \boldsymbol{\beta}_{\mathcal{A}}^0, \boldsymbol{\beta}_{\mathcal{A}^c}\big) + \frac{h(\varepsilon_i)}{2}\big(\mathbf{X}_i^t(\boldsymbol{\beta}_{\mathcal{A}} - \boldsymbol{\beta}_{\mathcal{A}}^0, \boldsymbol{\beta}_{\mathcal{A}^c})\big)^2 + o_{\mathbb{P}}\big(\mathbf{X}_i^t(\boldsymbol{\beta}_{\mathcal{A}} - \boldsymbol{\beta}_{\mathcal{A}}^0, \boldsymbol{\beta}_{\mathcal{A}^c})\big)^2\big].
\end{aligned}$$

By similar arguments used in the proof of Theorem 2.1*(i)* we have

$$n^{-1}\sum_{i=1}^{n} g(\varepsilon_i)\mathbf{X}_{i,\mathcal{A}}^t\big(\boldsymbol{\beta}_{\mathcal{A}}-\boldsymbol{\beta}_{\mathcal{A}}^0\big) = \mathbb{E}\big[n^{-1}\sum_{i=1}^{n} g(\varepsilon_i)\mathbf{X}_{i,\mathcal{A}}^t\big(\boldsymbol{\beta}_{\mathcal{A}}-\boldsymbol{\beta}_{\mathcal{A}}^0\big)\big] + O_{\mathbb{P}}\bigg(\mathbb{V}\mathrm{ar}\,\big[n^{-1}\sum_{i=1}^{n} g(\varepsilon_i)\mathbf{X}_{i,\mathcal{A}}^t\big(\boldsymbol{\beta}_{\mathcal{A}}-\boldsymbol{\beta}_{\mathcal{A}}^0\big)\big]\bigg)^{1/2}$$

$$= O_{\mathbb{P}}\bigg(\mathbb{V}\mathrm{ar}\,\big[n^{-1}\sum_{i=1}^{n} g(\varepsilon_i)\mathbf{X}_{i,\mathcal{A}}^t\big(\boldsymbol{\beta}_{\mathcal{A}}-\boldsymbol{\beta}_{\mathcal{A}}^0\big)\big]\bigg)^{1/2} \leq O_{\mathbb{P}}\bigg(\mathbb{E}[g^2(\varepsilon_i)]\frac{1}{n^2}\sum_{i=1}^{n}\|\mathbf{X}_{i,\mathcal{A}}\|_2^2\|\boldsymbol{\beta}_{\mathcal{A}}-\boldsymbol{\beta}_{\mathcal{A}}^0\|_2^2\bigg)^{1/2}$$

$$= O_{\mathbb{P}}\bigg(\frac{1}{n}\frac{p}{n}\bigg)^{1/2} = O_{\mathbb{P}}\bigg(\frac{p^{1/2}}{n}\bigg).$$

Proceeding similarly as above, we get:

$$n^{-1}\sum_{i=1}^{n} g(\varepsilon_i)\mathbf{X}_{i}^t\big(\boldsymbol{\beta}_{\mathcal{A}}-\boldsymbol{\beta}_{\mathcal{A}}^0,\boldsymbol{\beta}_{\mathcal{A}^c}\big) = O_{\mathbb{P}}\bigg(\frac{p^{1/2}}{n}\bigg).$$

Taking into account relation (11), we deduce that: $0 < n^{-1}\sum_{i=1}^{n} h(\varepsilon_i)\big(\mathbf{X}_i^t(\boldsymbol{\beta}_{\mathcal{A}}-\boldsymbol{\beta}_{\mathcal{A}}^0,\boldsymbol{\beta}_{\mathcal{A}^c})\big)^2 = O_{\mathbb{P}}(\|\boldsymbol{\beta}_{\mathcal{A}}-\boldsymbol{\beta}_{\mathcal{A}}^0\|_2^2) = O_{\mathbb{P}}(pn^{-1})$ and also that,

$$0 < n^{-1}\sum_{i=1}^{n} h(\varepsilon_i)\big(\mathbf{X}_{i,\mathcal{A}}^t(\boldsymbol{\beta}_{\mathcal{A}}-\boldsymbol{\beta}_{\mathcal{A}}^0)\big)^2 = O_{\mathbb{P}}\bigg(\frac{p}{n}\bigg).$$

By these relations, we obtain that the first term of the right-hand side of relation (16) is of order pn^{-1}. For the penalty of the right-hand side of relation (16), taking into account Theorem 2.1*(i)* and since $\boldsymbol{\beta}\in\mathcal{W}_n$ we obtain:

$$\lambda_n\sum_{j=p_0+1}^{p}\widehat{\omega}_{n,j}|\beta_j| \geq C\lambda_n\bigg(\frac{p}{n}\bigg)^{(1-\gamma)/2}.$$

Using the supposition $\lambda_n(pn^{-1})^{-(1+\gamma)/2}\underset{n\to\infty}{\longrightarrow}\infty$, that is $\lambda_n n^{(1-c)(1+\gamma)/2}\underset{n\to\infty}{\longrightarrow}\infty$, we have that in the right-hand side of relation (16), it's the penalty that dominates. Then, since $n^{-1}\big[Q_n(\boldsymbol{\beta})-Q_n(\boldsymbol{\beta}^{(1)})\big] = O_{\mathbb{P}}(pn^{-1})$, we have,

$$n^{-1}\big[R_n(\boldsymbol{\beta})-R_n(\boldsymbol{\beta}^{(1)})\big] \geq C\lambda_n\bigg(\frac{p}{n}\bigg)^{(1-\gamma)/2}. \tag{17}$$

But, on the other hand, since $\boldsymbol{\beta}_{\mathcal{A}^c}^{(1)} = \mathbf{0}_{p-p_0}$, by similar arguments as above, we have, $n^{-1}\big[R_n(\boldsymbol{\beta}^{\mathbf{0}})-R_n(\boldsymbol{\beta}^{(1)})\big] = O_{\mathbb{P}}(pn^{-1})$. From the last relation together relation (17), since $\lambda_n(pn^{-1})^{-(1+\gamma)/2}\underset{n\to\infty}{\longrightarrow}\infty$, we deduce, $\lim_{n\to\infty}\mathbb{P}[\widehat{\boldsymbol{\beta}}_n\in\mathcal{W}_n] = 0$.

(ii) Given the previous result we consider the parameter vector $\boldsymbol{\beta}$ of the form: $\boldsymbol{\beta} = \boldsymbol{\beta^0} + (pn^{-1})^{1/2}\boldsymbol{\delta}$, with $\boldsymbol{\delta} = (\boldsymbol{\delta}_{\mathcal{A}}, \boldsymbol{\delta}_{\mathcal{A}^c})$, $\boldsymbol{\delta}_{\mathcal{A}^c} = \mathbf{0}_{p-p_0}$, $\|\boldsymbol{\delta}_{\mathcal{A}}\|_2 \leq C$. We study then the following difference:

$$\frac{1}{n}R_n\big(\boldsymbol{\beta^0} + \sqrt{\frac{p}{n}}\boldsymbol{\delta}\big) - \frac{1}{n}R_n\big(\boldsymbol{\beta^0}\big) = \frac{1}{n}\sum_{i=1}^{n}\big[\rho_\tau\big(Y_i - \mathbf{X}_i^t(\boldsymbol{\beta^0} + \sqrt{\frac{p}{n}}\boldsymbol{\delta})\big) - \rho_\tau(\varepsilon_i)\big] + \mathcal{P}. \quad (18)$$

For the penalty $\mathcal{P} = \lambda_n \sum_{j=1}^{p_0} \widehat{\omega}_{n,j}\big(|\beta_j| - |\beta_j^0|\big)$ of the right-hand side of relation (18) we have, by 2.1*(i)*, $\widehat{\omega}_{n,j} = |\widetilde{\beta}_{n,j}|^{-\gamma} = O_{\mathbb{P}}(1)$ and by the triangular inequality $\big||\beta_j| - |\beta_j^0|\big| \leq |\beta_j - \beta_j^0|$. Then, as in the proof of Theorem 2.1, by relation (15), we obtain:

$$\mathcal{P} = O_{\mathbb{P}}\bigg(\lambda_n p_0^{1/2}\bigg(\frac{p}{n}\bigg)^{1/2}\bigg) = O_{\mathbb{P}}\bigg(\lambda_n p_0^{1/2} n^{(c-1)/2}\bigg). \quad (19)$$

For the first term of the right-hand side of relation (18) we have:

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\big[\rho_\tau\big(Y_i - \mathbf{X}_{i,\mathcal{A}}^t(\boldsymbol{\beta}_{\mathcal{A}}^0 + \sqrt{\frac{p}{n}}\boldsymbol{\delta}_{\mathcal{A}})\big) - \rho_\tau(\varepsilon_i)\big] \\
&\quad = -\frac{1}{n}\sum_{i=1}^{n} g(\varepsilon_i)\big(\mathbf{X}_{i,\mathcal{A}}^t\boldsymbol{\delta}_{\mathcal{A}}\big)\sqrt{\frac{p}{n}} + \frac{1}{2n}\sum_{i=1}^{n}\big[\frac{p}{n}\|\mathbf{X}_{i,\mathcal{A}}^t\boldsymbol{\delta}_{\mathcal{A}}\|_2^2 h(\varepsilon_i) + o_{\mathbb{P}}\big(\|\mathbf{X}_{i,\mathcal{A}}^t\boldsymbol{\delta}_{\mathcal{A}}\|^2\big)\big] \\
&\quad = \bigg(-\frac{1}{n}\sqrt{\frac{p}{n}}\sum_{i=1}^{n} g(\varepsilon_i)\big(\mathbf{X}_{i,\mathcal{A}}^t\boldsymbol{\delta}_{\mathcal{A}}\big) + \frac{1}{2n}\frac{p}{n}\sum_{i=1}^{n}\big(\boldsymbol{\delta}_{\mathcal{A}}^t\mathbf{X}_{i,\mathcal{A}}\mathbf{X}_{i,\mathcal{A}}^t\boldsymbol{\delta}_{\mathcal{A}}h(\varepsilon_i)\big)\bigg)\big(1 + o_{\mathbb{P}}(1)\big) \\
&\quad = \bigg(-\frac{1}{n}\sqrt{\frac{p}{n}}\sum_{i=1}^{n} g(\varepsilon_i)\big(\mathbf{X}_{i,\mathcal{A}}^t\boldsymbol{\delta}_{\mathcal{A}}\big) \\
&\qquad\qquad + \frac{1}{2n}\frac{p}{n}\sum_{i=1}^{n}\big(\boldsymbol{\delta}_{\mathcal{A}}^t\mathbf{X}_{i,\mathcal{A}}\mathbf{X}_{i,\mathcal{A}}^t\boldsymbol{\delta}_{\mathcal{A}}\big(\mathbb{E}[h(\varepsilon_i)] + h(\varepsilon_i) - \mathbb{E}[h(\varepsilon_i)]\big)\big)\bigg)\big(1 + o_{\mathbb{P}}(1)\big) \\
&\quad = \bigg(-\frac{1}{n}\sqrt{\frac{p}{n}}\sum_{i=1}^{n} g(\varepsilon_i)\big(\mathbf{X}_{i,\mathcal{A}}^t\boldsymbol{\delta}_{\mathcal{A}}\big) + \frac{1}{2n}\frac{p}{n}\sum_{i=1}^{n}\big(\boldsymbol{\delta}_{\mathcal{A}}^t\mathbf{X}_{i,\mathcal{A}}\mathbf{X}_{i,\mathcal{A}}^t\boldsymbol{\delta}_{\mathcal{A}}\mathbb{E}[h(\varepsilon_i)]\big)\bigg)\big(1 + o_{\mathbb{P}}(1)\big),
\end{aligned} \quad (20)$$

which has as minimizer the solution of

$$-\frac{1}{n}\sqrt{\frac{p}{n}}\sum_{i=1}^{n} g(\varepsilon_i)\mathbf{X}_{i,\mathcal{A}} + \boldsymbol{\Upsilon}_{n,\mathcal{A}}\sqrt{\frac{p}{n}}\boldsymbol{\delta}_{\mathcal{A}}\mathbb{E}[h(\varepsilon)] = \mathbf{0}_{p_0},$$

from where, we get,

$$\sqrt{\frac{p}{n}}\boldsymbol{\delta}_{\mathcal{A}} = \frac{\boldsymbol{\Upsilon}_{n,\mathcal{A}}^{-1}}{\mathbb{E}[h(\varepsilon)]}\frac{1}{n}\sum_{i=1}^{n} g(\varepsilon_i)\mathbf{X}_{i,\mathcal{A}}.$$

We deduce that, the minimum value of (20) is of order $O_{\mathbb{P}}\big(pn^{-1}\|\boldsymbol{\delta}_{\mathcal{A}}\|_2\big) = O_{\mathbb{P}}\big(pn^{-1}\big) = O_{\mathbb{P}}\big(n^{c-1}\big)$. Taking into account the supposition $\lambda_n p_0^{1/2} n^{(1-c)/2} \underset{n\to\infty}{\longrightarrow} 0$ and relation (19), we have that, $\mathcal{P} = o_{\mathbb{P}}(pn^{-1})$. Then, in the right-hand side of relation (18), the first term is the dominant one.
Let us now consider the following random variable sequence:

$$W_i \equiv g(\varepsilon_i)\mathbf{u}^t \frac{\boldsymbol{\Upsilon}_{n,\mathcal{A}}^{-1}}{\mathbb{E}[h(\varepsilon)]}\mathbf{X}_{i,\mathcal{A}},$$

with $\mathbf{u} \in \mathbb{R}^{p_0}$, $\|\mathbf{u}\|_2 = 1$. For the random variable W_i, we have that $\mathbb{E}[W_i] = 0$ and $\mathbb{V}\mathrm{ar}\,[W_i] = \mathbb{E}^{-2}[h(\varepsilon)]\boldsymbol{\Upsilon}_{n,\mathcal{A}}^{-1}\mathbf{u}^t\mathbf{X}_{i,\mathcal{A}}\mathbf{X}_{i,\mathcal{A}}^t\mathbf{u}\mathbb{V}\mathrm{ar}\,[g(\varepsilon_i)]$. Thus, taking into account assumption (A1), we get:

$$\sum_{i=1}^{n}\mathbb{V}\mathrm{ar}\,[W_i] = n\frac{\mathbf{u}^t\boldsymbol{\Upsilon}_{n,\mathcal{A}}^{-1}\mathbf{u}}{\mathbb{E}^2[h(\varepsilon)]}\mathbb{V}\mathrm{ar}\,[g(\varepsilon)],$$

which implies

$$\sqrt{n}\frac{\mathbb{E}[h(\varepsilon)]}{\sqrt{\mathbb{V}\mathrm{ar}\,[g(\varepsilon)]}}\frac{\mathbf{u}^t\big(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^{\mathbf{0}}\big)_{\mathcal{A}}}{\big(\mathbf{u}^t\boldsymbol{\Upsilon}_{n,\mathcal{A}}^{-1}\mathbf{u}\big)^{1/2}} \underset{n\to\infty}{\xrightarrow{\mathcal{L}}} \mathcal{N}(0,1).$$

The proof of claim *(ii)* is finished. ■

5.2. Result proofs in Section 3

Proof of Lemma 3.1.
The proof is similar to that of Theorem 2.1(i). Consequently, we give only the main results. Let us consider the p-vector $\mathbf{u} = (u_1, \cdots, u_p)$, with $\|\mathbf{u}\|_1 = 1$. By Holder's inequality: $|\mathbf{X}_i^t\mathbf{u}| \leq \|\mathbf{X}_i\|_\infty\|\mathbf{u}\|_1$, using also assumption (A4) we obtain that $\max_{1\leqslant i\leqslant n}|\mathbf{X}_i^t\mathbf{u}| < \infty$. For a constant $B > 0$, we will study the difference: $Q_n(\boldsymbol{\beta}^{\mathbf{0}} + Ba_n\mathbf{u}) - Q_n(\boldsymbol{\beta}^{\mathbf{0}})$. In this case, $\Delta_2 = O(a_n^2\sum_{i=1}^n(\mathbf{X}_i^t\mathbf{u})^2) = O(na_n^2)$. Since $\mathbb{V}\mathrm{ar}\left[a_n\sum_{i=1}^n g_\tau(\varepsilon_i)\mathbf{X}_i^t\mathbf{u}\right] = a_n^2\mathbf{u}^t\sum_{i=1}^n\mathbf{X}_i\mathbf{X}_i^t\mathbf{u}\mathbb{V}\mathrm{ar}\,[g_\tau(\varepsilon)] = O(na_n^2)$, then $Ba_n\sum_{i=1}^n g_\tau(\varepsilon_i)\mathbf{X}_i^t\mathbf{u} = O_{\mathbb{P}}\big(Bn^{1/2}a_n\big)$. Taking into account this last relation, we obtain for $D_i \equiv \rho_\tau\big(\varepsilon_i - Ba_n\mathbf{X}_i^t\mathbf{u}\big) - \rho_\tau(\varepsilon_i) + Ba_n g_\tau(\varepsilon_i)\mathbf{X}_i^t\mathbf{u}$, that: $\sum_{i=1}^n(D_i - \mathbb{E}[D_i]) = O_{\mathbb{P}}(B^2n^{1/2}a_n^2)$. Thus, $\Delta_1 = O_{\mathbb{P}}(Bn^{1/2}a_n) + O_{\mathbb{P}}(B^2n^{1/2}a_n^2) = O_{\mathbb{P}}(Bn^{1/2}a_n)$ and $\Delta_2 = O(B^2na_n^2)$, from where, since $a_n \to 0$, $n^{1/2}a_n \to \infty$, we get that $\Delta_2 > \Delta_1$ with probability converging to one, for B large enough.

■

Proof of Theorem 3.1.
The proof is similar to that of Theorem 2.1(ii). Consequently, we give only the main results. Otherwise, instead of the Cauchy-Schwarz inequality we use Holder's inequality:

$|\mathbf{X}_i^t\mathbf{u}| \leq \|\mathbf{X}_i\|_\infty \|\mathbf{u}\|_1$ and then we obtain: $0 < \Delta_2 = O(B^2 b_n^2 n \|\mathbf{u}\|_1^2)$.
For a p-vector $\mathbf{u} = (u_1, \cdots, u_p)$, with $\|\mathbf{u}\|_1 = 1$ and a constant $B > 0$, let the difference

$$R_n\big(\boldsymbol{\beta^0} + Bb_n\mathbf{u}\big) - R_n(\boldsymbol{\beta^0}) = Q_n\big(\boldsymbol{\beta^0} + Bb_n\mathbf{u}\big) - Q_n(\boldsymbol{\beta^0}) + n\lambda_n \sum_{j=1}^{p} \widehat{\omega}_{n,j}\big[\big|\boldsymbol{\beta}_j^0 + Bb_n u_j\big| - |\boldsymbol{\beta}_j^0|\big]. \tag{21}$$

By a similar approach made for the terms Δ_1 and Δ_2 of the proof of Theorem 2.1, we obtain:

$$Q_n\big(\boldsymbol{\beta^0} + Bb_n\mathbf{u}\big) - Q_n(\boldsymbol{\beta^0}) = O_{\mathbb{P}}(B^2 b_n^2 n \|\mathbf{u}\|_1^2). \tag{22}$$

For the penalty of the right-hand side of relation (21) we have:

$$n\lambda_n \sum_{j=1}^{p} \widehat{\omega}_{n,j}\big[\big|\boldsymbol{\beta}_j^0 + Bb_n u_j\big| - |\boldsymbol{\beta}_j^0|\big] \geq n\lambda_n \sum_{j=1}^{p_0} \widehat{\omega}_{n,j}\big[\big|\boldsymbol{\beta}_j^0 + Bb_n u_j\big| - |\boldsymbol{\beta}_j^0|\big] \geq -n\lambda_n \sum_{j=1}^{p_0} \widehat{\omega}_{n,j} Bb_n |u_j|,$$

by the Cauchy-Schwarz inequality and afterwards by the estimator consistency of $\overset{\vee}{\boldsymbol{\beta}}_n$, we have

$$\geq -n\lambda_n Bb_n \left(\sum_{j=1}^{p_0} \widehat{\omega}_{n,j}^2\right)^{1/2} \|\mathbf{u}\|_2 \geq -BCb_n n p_0^{1/2} \lambda_n \|\mathbf{u}\|_1^2 = -Bn\lambda_n p_0^{1/2} b_n. \tag{23}$$

Since $\lambda_n p_0^{1/2} b_n^{-1} \to 0$, as $n \to \infty$, then relation (22) dominates (23) and the theorem follows. ■

Proof of Theorem 3.2.
(i) Let $j \in \mathcal{A}^c$ be, then $j > p_0$. Thus, the derivative of the random process $R_n(\boldsymbol{\beta})$ in respect to β_j is:

$$\frac{\partial R_n(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} g_\tau\big(Y_i - \mathbf{X}_i^t\boldsymbol{\beta}\big) X_{ij} + n\lambda_n \widehat{\omega}_{n,j} \operatorname{sgn}(\beta_j). \tag{24}$$

For the first term of the right-hand side of relation (24), we have, $g_\tau\big(Y_i - \mathbf{X}_i^t\boldsymbol{\beta}\big) = g_\tau\big(\varepsilon_i - \mathbf{X}_i^t(\boldsymbol{\beta} - \boldsymbol{\beta^0})\big)$. We denote $\eta_i = \mathbf{X}_i^t(\boldsymbol{\beta} - \boldsymbol{\beta^0})$. By elementary calculations, we can show that, for $t \to 0$, we have: $g_\tau(\varepsilon - t) = g_\tau(\varepsilon) - h_\tau(\varepsilon)t + o_{\mathbb{P}}(t)$. On the other hand, by the Holder's inequality, we have: $|\eta_i| = \big|\mathbf{X}_i^t(\boldsymbol{\beta} - \boldsymbol{\beta^0})\big| \leq \|\mathbf{X}_i\|_\infty \|\boldsymbol{\beta} - \boldsymbol{\beta^0}\|_1$. Using assumption (A4) and Theorem 3.1, we obtain that $\eta_i \to 0$, when $n \to \infty$. Hence, $g_\tau(\varepsilon - \eta_i) = g_\tau(\varepsilon) - h_\tau(\varepsilon)\eta_i + o_{\mathbb{P}}(\eta_i)$, which implies:

$$\sum_{i=1}^{n} g_\tau\big(Y_i - \mathbf{X}_i^t\boldsymbol{\beta}\big) X_{ij} = \sum_{i=1}^{n} g_\tau(\varepsilon_i) X_{ij} - \sum_{i=1}^{n} \mathbf{X}_i^t(\boldsymbol{\beta} - \boldsymbol{\beta^0}) h_\tau(\varepsilon_i) X_{ij} + \sum_{i=1}^{n} o_{\mathbb{P}}\big(\mathbf{X}_i^t(\boldsymbol{\beta} - \boldsymbol{\beta^0})\big) X_{ij}.$$

By the Central Limit Theorem, tacking into account assumption (A4), we have that: $\sum_{i=1}^{n} g_\tau(\varepsilon_i)X_{ij} = O_{\mathbb{P}}(n^{1/2})$. On the other hand, $0 < h_\tau(\varepsilon_i) < 2$ with probability 1. Using the Holder's inequality, we have, with probability one,

$$\Big|\sum_{i=1}^{n} \mathbf{X}_i^t(\boldsymbol{\beta}-\boldsymbol{\beta^0})h_\tau(\varepsilon_i)X_{ij}\Big| \leq \sum_{i=1}^{n}\Big|\mathbf{X}_i^t(\boldsymbol{\beta}-\boldsymbol{\beta^0})h_\tau(\varepsilon_i)X_{ij}\Big| \leq \sum_{i=1}^{n}\|\mathbf{X}_i\|_\infty\big\|h_\tau(\varepsilon_i)X_{ij}(\boldsymbol{\beta}-\boldsymbol{\beta^0})\big\|_1,$$

from where, tacking into account assumption (A4), we have: $\sum_{i=1}^{n}\mathbf{X}_i^t(\boldsymbol{\beta}-\boldsymbol{\beta^0})h_\tau(\varepsilon_i)X_{ij} = O_{\mathbb{P}}(nb_n)$. Thus,

$$\sum_{i=1}^{n} g_\tau\big(Y_i - \mathbf{X}_i^t\boldsymbol{\beta}\big)X_{ij} = O_{\mathbb{P}}(nb_n). \tag{25}$$

For the penalty of relation (24) we have: $n\lambda_n\widehat{\omega}_{n,j} = O_{\mathbb{P}}\big(n\lambda_n \min\big(n^{1/2}, a_n^{-\gamma}\big)\big)$. Since, $\lambda_n b_n^{-1}\min\big(n^{1/2}, a_n^{-\gamma}\big) \to \infty$, as $n \to \infty$, also taking into account relation (25) we have that:

$$\frac{\partial R_n(\boldsymbol{\beta})}{\partial\beta_j}\begin{cases} > 0, & \text{if } \beta_j > 0,\\ \\ < 0, & \text{if } \beta_j < 0.\end{cases}$$

The function $R_n(\boldsymbol{\beta})$ is continuous in $\boldsymbol{\beta}$. Then, the solution of (24) must be equal to 0. From where $\widehat{\boldsymbol{\beta}}_{n,\mathcal{A}^c} = \mathbf{0}_{p-p_0}$, with probability converging to 1. This relation implies $\widehat{\mathcal{A}}_n \subseteq \mathcal{A}$ with probability converging to 1 when $n \to \infty$.
On the basis of this result, from now on we consider the parameters $\boldsymbol{\beta}$ of the form $\boldsymbol{\beta} = \big(\boldsymbol{\beta}_{\mathcal{A}}, \mathbf{0}_{p-p_0}\big)$. We must show now that $\mathcal{A} \subseteq \widehat{\mathcal{A}}_n$. By Theorem 3.1 we have $\|\widehat{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}^0_{\mathcal{A}}\|_1 = O_{\mathbb{P}}\big(b_n\big)$, from where for any $j = 1, \cdots, p_0$, we obtain, $\widehat{\beta}_{n,j} \xrightarrow[n\to\infty]{\mathbb{P}} \beta_j^0 \neq 0$. Thus, since $b_n \underset{n\to\infty}{\longrightarrow} 0$, we have that $\widehat{\beta}_{n,j} \neq 0$ with probability converging to 1, from where $\mathcal{A} \subseteq \widehat{\mathcal{A}}_n$.

(ii) Given the previous result *(i)* and Theorem 3.1, we consider the parameters $\boldsymbol{\beta}$ of the form: $\boldsymbol{\beta} = \boldsymbol{\beta^0} + b_n\boldsymbol{\delta}$, with $\boldsymbol{\delta} = (\boldsymbol{\delta}_{\mathcal{A}}, \boldsymbol{\delta}_{\mathcal{A}^c})$, $\boldsymbol{\delta}_{\mathcal{A}^c} = \mathbf{0}_{p-p_0}$, $\|\boldsymbol{\delta}_{\mathcal{A}}\|_1 \leq C$. For the penalty $\mathcal{P}$ of the right-hand side of relation (18) we have: $\big|\mathcal{P}\big| = \lambda_n\big|\sum_{j=1}^{p_0}\widehat{\omega}_{n,j}|\beta_j| - |\beta_j^0|]\big| \leq \lambda_n\sum_{j=1}^{p_0}\widehat{\omega}_{n,j}\big|\beta_j - \beta_j^0\big| \leq \lambda_n\big(\sum_{j=1}^{p_0}\widehat{\omega}_{n,j}^2\big)^{1/2}\|(\boldsymbol{\beta}-\boldsymbol{\beta^0})_{\mathcal{A}}\|_2 = O_{\mathbb{P}}\big(\lambda_n b_n p_0^{1/2}\big)$. For the main part, we have:

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\big[\rho_\tau\big(Y_i - \mathbf{X}_{i,\mathcal{A}}^t(\boldsymbol{\beta}^0_{\mathcal{A}} + b_n\boldsymbol{\delta}_{\mathcal{A}})\big) - \rho_\tau(\varepsilon_i)\big]\\ &= \bigg(-\frac{1}{n}b_n\sum_{i=1}^{n}g(\varepsilon_i)\big(\mathbf{X}_{i,\mathcal{A}}^t\boldsymbol{\delta}_{\mathcal{A}}\big) + \frac{1}{2n}b_n^2\sum_{i=1}^{n}\big(\boldsymbol{\delta}_{\mathcal{A}}^t\mathbf{X}_{i,\mathcal{A}}\mathbf{X}_{i,\mathcal{A}}^t\boldsymbol{\delta}_{\mathcal{A}}\mathbb{E}[h(\varepsilon_i)]\big)\bigg)\big(1 + o_{\mathbb{P}}(1)\big).\end{aligned}$$

The end of the proof is similar to that of Theorem 3.1*(ii)*. ■

Ciuperca, G., 2019. Adaptive group LASSO selection in quantile models. Statist. Papers 60 (1), 173–197.

Ciuperca, G., 2016. Adaptive LASSO model selection in a multiphase quantile regression. Statistics 50 (5),1100–1131.

Dezeure, R., Bühlmann, P., Zhang, C.H., 2017. High-dimensional simultaneous inference with the bootstrap. TEST 26 (4), 685–719.

Dezeure, R., Bühlmann, P., Meier, L., Meinshausen, N., 2015. High-Dimensional Inference: Confidence Intervals, p -Values and R -Software hdi. Statist. Sci. 30 (4), 533–558.

Fan, J., Fan, Y., Barut, E., 2014. Adaptive robust variable selection. Ann. Statist. 42 (1), 324–351.

Fan, J., Li, Q., Wang, Y., 2017. Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions, J. R. Statist. Soc. B 79, 247–265.

Gao, X., Huang, J., 2010. Asymptotic analysis of high-dimensional LAD regression with Lasso. Statist. Sinica 20 (4), 1485–1506.

Gu, Y., Zou, H., 2016. High-dimensional generalizations of asymmetric least squares regression and their applications. Ann. Statist. 44 (6), 2661–2694.

Huang, J., Ma, S., Zhang, C., 2008. Adaptive Lasso for sparse high-dimensional regression models. Statist. Sinica 18 (4), 1603–1618.

Kaul, A., Koul, H.L., 2015. Weighted l_1-penalized corrected quantile regression for high dimensional measurement error models. J. Multivariate Anal. 140, 72–91.

Leng, C., Li, B., 2010. Least squares approximation with a diverging number of parameters. Statist. Probab. Lett. 80, 254–261.

Liao, L., Park, C., Choi, H., 2019. Penalized expectile regression: an alternative to penalized quantile regression. Ann. Inst. Statist. Math. 71 (2), 409–438.

Newey, W.K., Powell, J.L., 1987. Asymmetric least squares estimation and testing. Econometrica 55 (4), 818–847.

Scheetz, T.E., Kim, K.A., Swiderski, R., Philp, A., Braun, T., Knudtson, K., Dorrance, A., DiBona, G., Huang, J., Casavant, T., Sheffield, V. , Stone E., 2006. Regulation of gene expression in the mammalian eye and its relevance to eye disease. Proceedings of the National Academy of Sciences of the United States of America 103, 14429–14434.

Schnabel, S.K., Eilers, P.H.C., 2009. Optimal expectile smoothing. Comput. Statist. Data Anal. 53 (12), 4168–4177.

Shah, R.D., Bühlmann, P., 2018. Goodness-of-fit tests for high dimensional linear models. J. R. Statist. Soc. B 80, 113–135.

Song, Q., Liang, F., 2017. Nearly optimal Bayesian Shrinkage for High Dimensional Regression. https://arxiv.org/abs/1712.08964.

Spiegel, E., Sobotka, F., Kneib, T., 2017. Model selection in semiparametric expectile regression. Electron. J. Stat. 11 (2), 3008–3038.

Tang, Y., Song, X., Wang, H.J., Zhu, Z., 2013. Variable selection in high-dimensional quantile varying coefficient models. J. Multivariate Anal. 122, 115–132.

Wang, D., Kulasekera, K.B., 2012. Parametric component detection and variable selection in varying-coefficient partially linear models. J. Multivariate Anal. 112, 117–129.

Wang, M., Wang, X., 2014. Adaptive Lasso estimators for ultrahigh dimensional generalized linear models. Statist. Probab. Lett. 89, 41–50.

Yang, Y., Wu, L., 2016. Nonnegative adaptive lasso for ultra-high dimensional regression models and a two-stage method applied in financial modeling. J. Statist. Plann. Inference 174, 52–67.

Zhang, F., Li, Q., 2017. A continuous threshold expectile model. Comput. Statist. Data Anal. 116, 49–66.

Zhang, C.H., Zhang, S.S., 2014. Confidence intervals for low dimensional parameters in high dimensional linear models. J. R. Statist. Soc. B 76, 217–242.

Zhao, J., Chen, Y., Zhang, Y., 2018. Expectile regression for analyzing heteroscedasticity in high dimension. Statist. Probab. Lett. 137, 304–311.

Zhao, J., Zhang, Y., 2018. Variable selection in expectile regression. Comm. Statist. Theory Methods 47 (7), 1731–1746.

Zheng, Q., Gallagher, C., Kulasekera, K.B., 2013. Adaptive penalized quantile regression for high dimensional data. J. Statist. Plann. Inference 143, 1029–1038.

Zheng, Q., Peng, L., He, X., 2015. Globally adaptive quantile regression with ultra-high dimensional data. Ann. Statist. 43 (5), 2225–2258.

Zou, H., 2006. The adaptive Lasso and its oracle properties. J. Amer. Statist. Assoc. 101 (476), 1418–1428.

Zou, H., Zhang, H.H., 2009. On the adaptive elastic-net with a diverging number of parameters. Ann. Statist. 37 (4), 1733–1751.