Robust Regression through the Huber's criterion and adaptive lasso penalty

Sophie Lambert-Lacroix, Laurent Zwald

Electronic Journal of Statistics
Vol. 5 (2011) 1015–1053
ISSN: 1935-7524
DOI: 10.1214/11-EJS635

Robust regression through the Huber's criterion and adaptive lasso penalty

Sophie Lambert-Lacroix

UJF-Grenoble 1 / CNRS / UPMF / TIMC-IMAG
UMR 5525, Grenoble, F-38041, France
e-mail: Sophie.Lambert@imag.fr

and

Laurent Zwald

LJK - Université Joseph Fourier BP 53,
Université Joseph Fourier
38041 Grenoble cedex 9, France
e-mail: Laurent.Zwald@imag.fr

Abstract: The Huber's Criterion is a useful method for robust regression. The adaptive least absolute shrinkage and selection operator (lasso) is a popular technique for simultaneous estimation and variable selection. The adaptive weights in the adaptive lasso allow to have the oracle properties. In this paper we propose to combine the Huber's criterion and adaptive penalty as lasso. This regression technique is resistant to heavy-tailed errors or outliers in the response. Furthermore, we show that the estimator associated with this procedure enjoys the oracle properties. This approach is compared with LAD-lasso based on least absolute deviation with adaptive lasso. Extensive simulation studies demonstrate satisfactory finite-sample performance of such procedure. A real example is analyzed for illustration purposes.

Keywords and phrases: Adaptive lasso, concomitant scale, Huber's criterion, oracle property, robust estimation.

Received June 2010.

Contents

1. Introduction

Data subject to heavy-tailed errors or outliers are commonly encountered in applications which may appear either in response variables or in the predictors. We consider here the regression problem with responses subject to heavy-tailed errors or outliers. In this case, the Ordinary Least Square (OLS) estimator is reputed to be not efficient. To overcome this problem, the least absolute deviation (LAD) or Huber type estimator for instance can be useful. On the other hand, an important topic in linear regression analysis is variable selection. Variable selection is particularly important when the true underlying model has sparse representation. To enhance the prediction performance of the fitted model and get an easy interpretation of the model, we need to identify significant predictors. Scientists prefer a simpler model because it puts more light on the relationship between the response and covariates. We consider the important problem of robust model selection.

The lasso penalty is a regularization technique for simultaneous estimation and variable selection ([32]). It consists to add a l_1 penalty to the least square criterion. This penalty forces to shrink some coefficients. In [4], the authors show that since lasso uses the same tuning parameters for all the regression coefficients, the resulting estimators may suffer an appreciable bias. Recently, [20, 18, 38] and [39] show that the underlying model must satisfy a nontrivial condition for the lasso estimator be consistent in variable selection. Consequently, in some cases, lasso estimator cannot be consistent in variable selection. In a first attempt to avoid this, [4] proposes the SCAD penalty and shows its consistency in variable selection. The main drawback of the SCAD penalty is due to its non-convexity: it typically leads to optimisation problems suffering from the local minima problem. In a second attempt, [39] provides a convex

optimisation problem leading to a consistent in variable selection estimator. He assigns adaptive weights for penalizing *differently* coefficients in the l_1 penalty and calls this new penalty the adaptive lasso. Owing to the convexity of this penalty, it typically leads to convex optimisation problems. As a consequence, they do not suffer from the local minima problem. These adaptive weights in the penalty allow to have the oracle properties. Moreover, the adaptive lasso can be solved by the same efficient algorithm (LARS) for solving lasso (see [39]).

In [36], the authors propose to treat the problem of robust model selection by combining LAD loss and adaptive lasso penalty. So they obtain an estimator which is robust against outliers and also enjoys a sparse representation. Unfortunately, the LAD loss (l_1 criterion) is not adapted for small errors: it penalizes strongly the small residuals. In particular when the error has no heavy tail and does not suffers from outliers, this estimator is expected to be less efficient than the OLS estimator with adaptive lasso. In practice, we do not know in which case we are. So it is important to consider some methods having good performances in both situations.

That is why we can prefer to consider Huber's criterion with concomitant scale (see [12]). The Huber's criterion is a hybrid of squared error for relatively small errors and absolute error for relative large ones. In this paper, we propose to combine Huber's criterion with concomitant scale and adaptive lasso. We show that the resulting estimators enjoy the oracle properties. This approach is compared with LAD-lasso based on least absolute deviation with adaptive lasso. Extensive simulation studies demonstrate satisfactory finite-sample performance of such procedure.

The rest of the article is organized as follows. In Section 2, we recall the lasso-type method and introduce the Huber's criterion with adaptive lasso penalty. In Section 3, we give its statistical properties. Section 4 is devoted to simulation. This study compares the Huber's criterion with adaptive lasso with two others methods: least square criterion with adaptive lasso and the LAD-lasso approach. In Section 5, we analyze Chinese stock market data for illustration purposes. We relegate technical proofs to the Appendix.

2. The lasso-type method

2.1. Lasso-type estimator

Let us consider the linear regression model

$$y_i = \alpha^* + \mathbf{x}_i^T \beta^* + \sigma\epsilon_i, \quad i = 1, \ldots, n, \tag{2.1}$$

where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$ is the p-dimensional centered covariable (that is $\sum_{i=1}^n \mathbf{x}_i = 0$), α^* is the constant parameter and $\beta^* = (\beta_1^*, \ldots, \beta_p^*)^T$ are the associated regression coefficients. We suppose that $\sigma > 0$ and ϵ_i are independent and identically-distributed random errors with mean 0 and variance 1, when it exists. Indeed in the sequel we do not need existence of variance.

Let $\mathcal{A} = \{1 \leq j \leq p,\ \beta_j^* \neq 0\}$ and $p_0 = |\mathcal{A}|$. In variables selection context, we usually assume that $\beta_j^* \neq 0$, for $j \leq p_0$ and $\beta_j^* = 0$, for $j > p_0$ for some $p_0 \geq 0$. In this case the correct model has p_0 significant regression variables. We denote by $\beta_{\mathcal{A}}$ the vector given by the coordinates of β the index of which are in $\mathcal{A}$.

When $p_0 = p$, the unknown parameters in the model (2.1) are usually estimated by minimizing the ordinary least square criterion. To shrink unnecessary coefficients to 0, [32] proposed to introduce a constraint on the l_1-norm of the coefficients. This leads to the primal formulation of the lasso. The link with the following dual criterion is studied in [21]:

$$\sum_{i=1}^{n}(y_i - \alpha - \mathbf{x}_i^T\beta)^2 + \lambda_n \sum_{j=1}^{p} |\beta_j| .$$

$\lambda_n > 0$ is the tuning parameter. Notice that the intercept α does not appear in the penalty term since it seems not reasonable to constrain it.

Fan and Li [4] studied a class of penalization methods including the lasso one. They showed that the lasso method leads to estimators that may suffer an appreciable bias. Furthermore they conjectured that the oracle properties do not hold for the lasso. Hence Zou [39] proposes to consider the following modified lasso criterion, called adaptive lasso,

$$Q^{adl}(\alpha, \beta) = \sum_{i=1}^{n}(y_i - \alpha - \mathbf{x}_i^T\beta)^2 + \lambda_n \sum_{j=1}^{p} \hat{w}_j^{adl}|\beta_j|,$$

where $\hat{\mathbf{w}}^{adl} = (\hat{w}_1^{adl}, \ldots, \hat{w}_p^{adl})$ is a known weights vector. We report to the Subsection 2.4 for the hyperparameter choice. This modification allows to produce sparse solutions more effectively than lasso. Precisely, Zou [39] shows that with a proper choice of λ_n and of $\hat{\mathbf{w}}^{adl}$ the adaptive lasso enjoys the oracle properties.

2.2. Robust lasso-type estimator: The LAD-lasso

When the regression error has very heavy tail or suffers from outliers, the finite sample performance of lasso can be poor. A first attempt to solve this problem has been done in [36]. This paper provides a procedure inspired by the convex function

$$Q^{ladl}(\alpha, \beta) = \sum_{i=1}^{n} |y_i - \alpha - \mathbf{x}_i^T\beta| + \lambda_n \sum_{j=1}^{p} \hat{w}_j^{ladl}|\beta_j| .$$

Note that the intercept α is not included in the study of [36] but to fairly compare the methods we consider this intercept term in this paper. As in [39], the authors show that with a proper choice of λ_n and of $\hat{\mathbf{w}}^{ladl} = (\hat{w}_1^{ladl}, \ldots, \hat{w}_p^{ladl})$, the adaptive LAD-lasso enjoys the oracle properties. We report again to the Subsection 2.4 for the hyperparameter choice. Moreover the obtained estimator is robust to heavy tailed errors since the squared loss has been replaced by the L_1 loss. Unfortunately, this loss is not adapted for small errors: it penalizes

strongly the small residuals. In particular when the error has no heavy tail and does not suffer from outliers, this estimator is expected to be less efficient than the adaptive lasso. That is why we can prefer to consider criterion like Huber's one.

2.3. The Huber's Criterion with adaptive lasso

To be robust to the heavy-tailed errors or outliers in the response, another possibility is to use the Huber's criterion as loss function as introduced in [12]. For any positive real M, let us introduce the following function

$$\mathcal{H}_M(z) = \begin{cases} z^2 & |z| \leq M, \\ 2M|z| - M^2 & |z| > M. \end{cases}$$

This function is quadratic in small values of z but grows linearly for large values of z. The parameter M describes where the transition from quadratic to linear takes place. The Huber's criterion can be written as

$$\sum_{i=1}^{n} \mathcal{H}_M \left(\frac{y_i - \alpha - \mathbf{x}_i^T \beta}{s} \right),$$

where $s > 0$ is a scale parameter for the distribution. That is if each y_i is replaced by cy_i for $c > 0$ then an estimate $\hat{s}$ should be replaced by $c\hat{s}$. Usually, the parameter s is denoted by σ. To avoid confusions, we adopt here another notation since one can choose σ as scale parameter but various choices are possible. For example any multiple of σ is a scale parameter and those are not only.

Let us remark that with this loss function errors smaller than sM get squared while larger errors increase this criterion only linearly. In [12], the parameter M is viewed as a shape parameter that one chooses to control the amount of robustness. The Huber's criterion becomes more similar to least square for larger values of M while it becomes more similar to LAD criterion for small values of M. In this case, the Huber's method is more robust against outliers (as LAD method) but less efficient for normally distributed data. In [12], Huber proposes to fix $M = 1.345$ to get as much robustness as possible while being efficient for normally distributed data. Even if we adopt this approach, there remains the scale parameter to estimate.

As far as we know, all the algorithms of the literature designed with a penalty first estimate the unknown parameter σ defined in (2.1) and plug it as a scale s in the Huber's criterion. Among all the possible estimations of the standard deviation σ of the data, there is no rule how to choose it. A popular choice (see e.g. [29, 25, 14, 28]) is the Median Absolute Deviation (MAD). It get a simple explicit formula, needs little computational time and is very robust as witnessed by its bounded influence function and it 50% breakdown point. Note that it has a low (37%) gaussian efficiency ([33]). However, this kind of approach is critizeable (see e.g. [10]). Indeed, the Huber's criterion is designed to work with a scale parameter s which, ideally, is *not* the standard deviation of the data σ (as

shown by lemma 1 below). Moreover, σ is only a nuisance parameter: the location one is generally more important. So, focusing attention on a good estimation of σ introduces difficulties in a superfluous step of the procedure: Huber's criterion only needs a well designed scale s. [12] proposed to *jointly* estimate s and the parameters of the model in several ways. The common property shared by these methods is that they do not need an estimation of the parameter σ. We retain the Huber's Criterion with concomitant scale defined by,

$$\mathcal{L}_{\mathcal{H}}(\alpha,\beta,s)=\begin{cases} ns+\sum_{i=1}^{n}\mathcal{H}_M\left(\frac{y_i-\alpha-\mathbf{x}_i^T\beta}{s}\right)s & \text{if} \quad s>0,\\ 2M\sum_{i=1}^{n}|y_i-\alpha-\mathbf{x}_i^T\beta| & \text{if} \quad s=0,\\ +\infty & \text{if} \quad s<0, \end{cases} \tag{2.2}$$

which are to minimize with respect to $s\geq 0$, α and β. To our knowledge, the statistical properties of this loss function have never been studied. Theorem 3.1 below shows that, as for the MAD, the provided scale estimation is a robust transformation of the residuals of a location estimate. However, in the case of MAD, the location esimate is the OLS estimator which can have very poor quality especially if collinearity is involved. In this case, the corresponding residuals are irrelevant. On the contrary, the scale used by the Huber's Criterion with concomitant scale is obtained from more relevant residuals.

Let us define, for $s>0$,

$$F(s)=\mathbb{E}\left[\frac{1}{n}\mathcal{L}_{\mathcal{H}}(\alpha^*,\beta^*,s)\right]=s+s\mathbb{E}\left[\mathcal{H}_M\left(\frac{\sigma\epsilon}{s}\right)\right]$$

and

$$s^*=\underset{s>0}{\operatorname{argmin}}\,F(s)\,. \tag{2.3}$$

We have the following lemma (its proof is given in Appendix 6.3).

Lemma 1. *If $M>1$ and* **(N2)** *(defined below page 10) holds, then there exists a unique $s^*>0$ satisfying (2.3). Moreover, it satisfies*

$$s^*=\mathbb{E}\left[\sigma\epsilon\mathcal{H}'_M\left(\frac{\sigma\epsilon}{s^*}\right)-s^*\mathcal{H}_M\left(\frac{\sigma\epsilon}{s^*}\right)\right]\,. \tag{2.4}$$

Consequently, the $\hat{s}$ obtained by the minimisation of the Huber loss function with concomitant scale is a scale estimation of the scale parameter s^*. Generally, it is a poor estimation of the standard deviation σ of the data. As explained previously, the algorithms of the literature use an estimation of σ as scale parameter which is not necessarily well suited for the Huber criterion. It is noticeable that the scale parameter s^* is the standard deviation σ of the noise only if the loss is quadratic (i.e. $M=+\infty$).

Let us now briefly comment the way we estimate the intercept α^* of the model. In practice, it is usual to center the y and the x and to remove it from the optimization problem. This procedure is equivalent to minimize the quadratic loss over α. However, since we use the Huber loss (and not the quadratic loss), this procedure is not any more equivalent to minimize the loss function. So, we

minimize the Huber's loss over α. Consequently, in our procedure, the intercept α^* is no more estimated by a mean.

In this paper, we want to combine the Huber's criterion and adaptive penalty as lasso. In particular, that allows to have the oracle properties (see Section 3). So, we consider the following criterion

$$Q^{\mathcal{H}adl}(\alpha, \beta, s) = \mathcal{L}_{\mathcal{H}}(\alpha, \beta, s) + \lambda_n \sum_{j=1}^{p} \hat{w}_j^{\mathcal{H}adl} |\beta_j|$$

where $\hat{\mathbf{w}}^{\mathcal{H}adl} = (\hat{w}_1^{\mathcal{H}adl}, \ldots, \hat{w}_p^{\mathcal{H}adl})$ is a known weights vector which will be defined at Section 2.4. As can be seen, the criterion $Q^{\mathcal{H}adl}$ combines the Huber's criterion and adaptive lasso penalty. Hence, the resulting estimator is expected to be robust against outlier and also to enjoy sparse representation. Let us remark that $Q^{\mathcal{H}adl}(\alpha, \beta, s)$ is a convex function and thus the optimization problem does not suffer from the multiple local minima issue. Its global minimizer can be efficiently solved. We give an algorithm in Appendix 6.4.

2.4. Tuning parameter estimation

We now consider the problem of tuning parameter estimation. For the adaptive lasso method, Zou [39] proposes to use the following estimation for weights vector. Let $\hat{\beta}$ be a root-n-consistent estimator to β^*; for instance, one can use the estimate obtained by OLS $\hat{\beta}^{ols}$. Let $\gamma > 0$ be a constant to be determined. He defines the weights vector estimation as $\hat{w}_j^{adl} = |\hat{\beta}_j^{ols}|^{-\gamma}$, $j = 1, \ldots, p$. Then he uses two dimensional cross-validation to find an optimal pair of (γ, λ_n).

For the LAD-lasso, Wang et al. [36] consider similar estimation for weights vector. Let $\hat{\beta}^{lad}$ the unpenalized LAD estimator of β^*. They propose to use weights vector estimation as $\hat{w}_j^{ladl} = |\hat{\beta}_j^{lad}|^{-1}$, $j = 1, \ldots, p$, and fix $\lambda_n = \log(n)$.

For Huber's Criterion with concomitant scale and adaptive penalty, we propose to use similar way. We denote by $\hat{\beta}^{\mathcal{H}}$ the unpenalized Huber's estimator with concomitant scale. So, the weights vector is estimated by $\hat{w}_j^{\mathcal{H}adl} = |\hat{\beta}_j|^{-\gamma}$, $j = 1, \ldots, p$. It remains to evaluate the constants λ_n and M. As in [12], we fix $M = 1.345$. Next, we can use cross-validation to find optimal values for λ_n. Let us note that the theoretical part is given for these forms of weights vector and for the numerical results we fix γ equal to 1.

Following the remarks of an anonymous referee, for these three methods a BIC-type model selection procedure has been designed to choose the regularization parameter λ_n. Let us now describe precisely the BIC criterions we used for each method. The collections of estimators obtained using adaptive-lasso or LAD-lasso methods are naturally log-likelihood estimators. The corresponding collections of models (containing probability density functions) are nested. We use the classical BIC criterions. In [30], relying on the Kullback-Leibler divergence, it is recommended to select λ_n^{adl} minimizing

$$\log\left(\sum_{i=1}^{n}\left(y_i - \widehat{\alpha}_{\lambda_n}^{adl} - \mathbf{x}_i^T \widehat{\beta}_{\lambda_n}^{adl}\right)^2\right) + k_{\lambda_n} \frac{\log(n)}{n},$$

over λ_n for adaptative-lasso and minimizing

$$\log\left(\sum_{i=1}^{n}|y_i-\widehat{\alpha}_{\lambda_n}^{ladl}-\mathbf{x}_i^T\widehat{\beta}_{\lambda_n}^{ladl}|\right)+k_{\lambda_n}\frac{\log(n)}{2n},$$

over λ_n for LAD-lasso. Let us note that k_{λ_n} denotes the model dimension. Following [35] and [37], we use for k_{λ_n} the number of non-zero coefficients of $\widehat{\beta}_{\lambda_n}^{adl}$ (resp. $\widehat{\beta}_{\lambda_n}^{ladl}$) for adaptive-lasso (resp. LAD-lasso). In the underlying models *all* the residuals are supposed to have the same distribution: gaussian or double exponential.

Let us now design a BIC-type procedure taking advantage of the two previous ones. The flexibility of BIC criterions would allow to gather the collection of estimators obtained by adaptive-lasso and LAD-lasso. The corresponding collection of models are no more nested but BIC criterion have been designed to work in this framework (see [30]). Thus one easily get a BIC criterion to select the final estimator in this augmented collection of estimators. However, datasets more likely contain *some* outliers. We thus propose a BIC type procedure relying on the Huber's loss. In this way, the weight associated to each residual is adapted: they are not treated *all* in the same way. By analogy with the two previous ones, we propose to select $\lambda_n^{\mathcal{H}adl}$ in the collection of estimators $(\widehat{\alpha}_{\lambda_n}^{\mathcal{H}adl},\widehat{\beta}_{\lambda_n}^{\mathcal{H}adl},\widehat{s}_{\lambda_n}^{\mathcal{H}adl})$ by minimizing

$$\log\left(\mathcal{L}_{\mathcal{H}}\left(\widehat{\alpha}_{\lambda_n}^{\mathcal{H}adl},\widehat{\beta}_{\lambda_n}^{\mathcal{H}adl},\widehat{s}_{\lambda_n}^{\mathcal{H}adl}\right)\right)+k_{\lambda_n}\frac{\log(n)}{2n},$$

over λ_n. As previously, k_{λ_n} denotes the number of non-zero coefficients of $\widehat{\beta}_{\lambda_n}^{\mathcal{H}adl}$. Since the scale s is not penalized in $Q^{\mathcal{H}adl}$, it remains to replace the quadratic loss of adaptative-lasso (or ℓ_1 loss of LAD-lasso) by the loss

$$\min_{s\geq 0}\mathcal{L}_{\mathcal{H}}(\widehat{\alpha}_{\lambda_n}^{\mathcal{H}adl},\widehat{\beta}_{\lambda_n}^{\mathcal{H}adl},s)$$

of Huber-lasso within the logarithm of the BIC criterion.

2.5. Some remarks on scale invariance

An estimator $\hat{e}(y_1,\ldots,y_n)$ calculated from the data $(\mathbf{x}_i,y_i)_{1\leq i\leq n}$ is said to be *scale invariant* if

$$\forall\, c>0,\ \hat{e}(cy_1,\ldots,cy_n)=c\hat{e}(y_1,\ldots,y_n)\,.$$

Note that, in this definition, the design is fixed. This means that if each y_i is replaced by cy_i for $c>0$ then the estimate $\hat{e}$ should be replaced by $c\hat{e}$. It is important to consider location estimators $(\hat{\alpha},\hat{\beta})$ satisfying this property since they provide a coherent interpretation of the results. Indeed, if we change the scale of our measurments y by an arbitrary change of units, the selected variables are the same and the prediction changes accordingly. Note that others equivariance notions have also been introduced in [17].

The following easy property can be used to assert that an estimator is scale invariant. Let A be a cone of $\mathbb{R}^q$ (this means that $cx \in A$ when $c > 0$ and $x \in A$). If

$$\hat{e}(y_1, \ldots, y_n) = \underset{\gamma \in A}{\operatorname{argmin}} Q(y_1, \ldots, y_n, \gamma)$$

with

$$Q(cy_1, \ldots, cy_n, c\gamma) = g(c)Q(y_1, \ldots, y_n, \gamma), \quad g(c) \geq 0,$$

then $\hat{e}(y_1, \ldots, y_n)$ is scale invariant.

Let us note that the lasso procedure (OLS with lasso penalty) is not scale invariant. On the other hand, the LAD criterion or the Huber's one with concomitant scale is always scale invariant when combining with lasso penalty. But, when an adaptive penalty is introduced in the previous criteria (even if $\gamma = 1$), the scale invariant property is lost. On the other hand, if we consider all the procedure (λ_n choice by BIC-type criterion or cross validation technique and estimation method), the adaptive methods are all scale invariant.

3. Theoretical properties

Let $\mathbf{X}$ denotes the design matrix i.e. the $n \times p$ matrix the i^{th} rows of which is $\mathbf{x}_i^T$. We will use some of the following assumptions on this design matrix.

(D1) $\max_{1 \leq i \leq n} \|\mathbf{x}_i\| / \sqrt{n} \to 0$ as $n \to \infty$.
(D2) $\mathbf{X}^T\mathbf{X}/n \to V$ as $n \to \infty$ with $V_{1,1} > 0$, where $V_{1,1}$ is the first $p_0 \times p_0$ bloc of V, corresponding to the covariables associated with non zero coefficients.

Assumption **(D1)** and **(D2)** are classical. For instance, **(D1)** is supposed in theorem 4.1 of [17] to ensure the asymptotic normality of the regression quantile estimation. It can also be seen as a "compacity assumption": it is satisfied if the variables are supposed to be bounded. In [39], the author needs only the assumption **(D2)** since he uses a least square criterion as loss function.

Let us denote by ϵ a variable with the same law as ϵ_i, $i = 1, \ldots, n$. The following assumptions on the errors are used in the following:

(N0) The distribution of the errors does not charge the points $\pm Ms^*$:

$$\mathbb{P}[\sigma\epsilon = \pm Ms^*] = 0.$$

(N1) The variable ϵ is symmetric (i.e. ϵ has the same distribution as $-\epsilon$).
(N2) For all $a > 0$, $\mathbb{P}[\epsilon \in [-a, a]] > 0$.

Note that **(N0)** holds if ϵ is absolutely continuous with respect to the Lebesgue's measure and **(N2)** is satisfied if, moreover, the density is continuous and strictly positive at the origin (which is assumption A of [36]). Condition **(N1)** is natural without prior knowledge on the distribution of the errors and **(N2)** ensures that the noise is not degenerated. It is noticeable that there is no integrability condition assumed on the errors ϵ. The theorems ensuring the convergence of the penalized least squared estimators (e.g. [16] and [39]) usually assume that ϵ has a finite variance.

Let $(\hat{\alpha}^{\mathcal{H}adl}, \hat{\beta}^{\mathcal{H}adl}, \hat{s}^{\mathcal{H}adl})$ be defined by the minimizer of $Q^{\mathcal{H}adl}(\cdot)$ where $\hat{w}_j^{\mathcal{H}adl} = 1/|\hat{\beta}_j|^\gamma$ with $\gamma > 0$ and $\hat{\beta}$ a root-n-consistent estimator to β^* (i.e. $\sqrt{n}(\hat{\beta} - \beta^*) = \mathcal{O}_P(1)$). We denote $\mathcal{A}_n = \{1 \leq j \leq p,\ \hat{\beta}_j^{\mathcal{H}adl} \neq 0\}$. Let us remark that if $\lambda_n > 0$, the argminimum $(\hat{\alpha}^{\mathcal{H}adl}, \hat{\beta}^{\mathcal{H}adl}, \hat{s}^{\mathcal{H}adl})$ exists since the criterion $Q^{\mathcal{H}adl}(\cdot)$ is a convex and coercive function.

Theorem 3.1 below states that the estimation of the scale proposed by the Huber criterion with concomitant is robust to large residuals. The robustness comes from the fact that only the smallest residuals are taking into account to obtain the scale estimation. For (α, β) fixed, let us sort the absolute values of the residuals, $r_i(\alpha, \beta) = y_i - \alpha - x_i^T\beta$, corresponding to (α, β):

$$|r_{(1)}(\alpha, \beta)| \leq |r_{(2)}(\alpha, \beta) \leq \cdots \leq |r_{(n)}(\alpha, \beta)|\,.$$

For a real number x, $\lceil x \rceil$ denotes the smallest integer larger than x. Then we have the following theorem (its proof is postponed in Appendix 6.1).

Theorem 3.1. *When $M \neq 1$, there exists a unique $\hat{s}^{\mathcal{H}adl}$. Moreover if $M \leq 1$ then $\hat{s}^{\mathcal{H}adl} = 0$ and $(\hat{\alpha}^{\mathcal{H}adl}, \hat{\beta}^{\mathcal{H}adl})$ is obtained by minimising the penalised ℓ_1 loss (as LAD). If $M > 1$,*

$$\left(\hat{s}^{\mathcal{H}adl}\right)^2 = \frac{1}{n - M^2(n-k)} \sum_{i=1}^{k} r_{(i)}^2(\hat{\alpha}^{\mathcal{H}adl}, \hat{\beta}^{\mathcal{H}adl})\,, \tag{3.1}$$

where $k \in [\lceil n\left(1 - \frac{1}{M^2}\right)\rceil, n]$ is such that

$$\frac{r_{(k)}^2(\hat{\alpha}^{\mathcal{H}adl}, \hat{\beta}^{\mathcal{H}adl})}{M^2} < \frac{\sum_{i=1}^{k} r_{(i)}^2(\hat{\alpha}^{\mathcal{H}adl}, \hat{\beta}^{\mathcal{H}adl})}{n - M^2(n-k)} \leq \frac{r_{(k+1)}^2(\hat{\alpha}^{\mathcal{H}adl}, \hat{\beta}^{\mathcal{H}adl})}{M^2}. \tag{3.2}$$

Note that the criterion (3.2) determines which residuals are small enough to be used in the estimation (3.1) of the scale. It relies on the energy of the smallest residuals. This way to be robust is different as the one used by a MAD type estimation of the standard deviation σ where the median (of the residuals) is used. Note that when $M = +\infty$, $k = n$ and $\hat{s}^{\mathcal{H}adl}$ is the classical arithmetical mean of the squared residuals. This is the maximum likelihood estimator of σ^2 in the gaussian case. Unfortunately, in this case, the likelihood is not concave. Huber loss criterion with $M = +\infty$ provides a convex objective function the minimum of which leads to the same estimation. Moreover, we have for any (α, β):

$$\min_{s>0} \mathcal{L}_{\mathcal{H}}(\alpha, \beta, s) = 2\sqrt{n - (n-k)M^2}\sqrt{\sum_{i=1}^{k} r_{(i)}^2(\alpha, \beta)} + 2M \sum_{i=k+1}^{n} |r_{(i)}(\alpha, \beta)|\,. \tag{3.3}$$

So the loss function linearly penalizes the largest residuals. We put attention to the reader on the fact that small residuals are put together through a L_2 norm and not the classical squared L_2 norm loss. This is natural since we consider

a scale invariant procedure. Consequently, the objective function $Q^{\mathcal{H}adl}$ with $M = +\infty$ is not equal to Q^{adl}.

In the following theorem we show that, with a proper choice of λ_n, the proposed estimator enjoys the oracle properties. Its proof is postponed in Appendix 6.2.

Theorem 3.2. *Suppose that* $\lambda_n/\sqrt{n} \to 0$ *and* $\lambda_n n^{(\gamma-1)/2} \to \infty$. *Let us also assume that conditions* $M > 1$, $p_0 > 0$, **(N0)**, **(N1)**, **(N2)**, **(D1)** *and* **(D2)** *hold. Moreover, for* $j = 1, \ldots, p$, *the weights in* $Q^{\mathcal{H}adl}$ *are* $\hat{w}_j^{\mathcal{H}adl} = 1/|\hat{\beta}_j|^\gamma$ *where* $\hat{\beta}$ *is a root-n-consistent estimator of* β^*. *Then, any minimizer* $(\hat{\alpha}^{\mathcal{H}adl}, \hat{\beta}^{\mathcal{H}adl}, \hat{s}^{\mathcal{H}adl})$ *of* $Q^{\mathcal{H}adl}$ *satisfies the following:*

- *Consistency in variable selection:* $\mathbb{P}[\mathcal{A}_n = \mathcal{A}] \to 1$ *as* $n \to +\infty$.
- *Asymptotic normality:*

$$\sqrt{n}\left(\hat{\alpha}^{\mathcal{H}adl} - \alpha^*, \hat{\beta}_{\mathcal{A}}^{\mathcal{H}adl} - \beta_{\mathcal{A}}^*, \hat{s}^{\mathcal{H}adl} - s^*\right) \longrightarrow_d \mathcal{N}_{p_0+2}\left(0, \Sigma^2\right),$$

where Σ^2 *is the squared block diagonal matrix*

$$\Sigma^2 = \mathtt{diag}\left(\frac{\mathbb{E}\left[\mathcal{H}'_M\left(\frac{\sigma\epsilon}{s^*}\right)^2\right]}{4A_{s^*}^2}, \frac{\mathbb{E}\left[\mathcal{H}'_M\left(\frac{\sigma\epsilon}{s^*}\right)^2\right]}{4A_{s^*}^2} V_{1,1}^{-1}, \frac{\mathbb{E}\left[Z^2\right]}{4D_{s^*}^2}\right)$$

and where

$$D_{s^*} = \frac{1}{s^{*3}}\mathbb{E}\left[\sigma^2\epsilon^2 \mathbb{1}_{|\sigma\epsilon| \le Ms^*}\right], \; A_{s^*} = \frac{1}{s^*}\mathbb{P}\left[|\sigma\epsilon| \le Ms^*\right],$$

$$Z = 1 + \mathcal{H}_M\left(\frac{\sigma\epsilon}{s^*}\right) - \frac{\sigma\epsilon}{s^*}\mathcal{H}'_M\left(\frac{\sigma\epsilon}{s^*}\right).$$

[3] already studied asymptotic normality of any minimizer of the Huber loss without concomitant scale nor penalty. It is noticeable that the previous result holds under the same assumptions: the introduction of the concomitant scale in the criterion does not lead to supplementary hypotheses.

Unlike the plug-in methods of the scale, this result provides a *simultaneous* convergence of location and scale estimations. Moreover, the asymptotic variance Σ^2 is block diagonal. This means that these estimations are asymptotically independent.

In [39], the author has already treated the case where the loss is quadratic. Theorem 3.2 generalizes theorem 2 of [39] to deal with a robust loss (i.e. the case $M < +\infty$). It is noticeable that, in the quadratic case where $M = +\infty$, the asymptotic variance matrix $\mathbb{E}[\mathcal{H}'_{Ms}(\sigma\epsilon)^2]V_{1,1}^{-1}/(4A_1(Ms)^2)$ obtained in theorem 3.2 is equal to $\sigma^2 V_{1,1}^{-1}$ and we recover the asymptotic variance of theorem 2 of [39].

Let us also remark that assumptions on λ_n are the same as in [39]. They are used in the penalty term control. The data-dependent $\hat{\mathbf{w}}^{\mathcal{H}adl}$ is the key in Theorem 3.2. As the sample size grows, the corresponding weights get inflated (to

infinity) for zero-coefficient whereas they converge to finite constant for nonzero-coefficient. Consequently, as explained in [39], we can simultaneously unbiasedly (asymptotically) estimate large coefficient and remove zero-coefficient.

For simplicity, Theorem 3.2 is stated assuming that the preliminary estimator $\hat{\beta}$ is a root-n-consistent estimator to β^*. Using results over loss function given in the proof of Theorem 3.2, we can prove that the unpenalized Huber's estimator $\hat{\beta}_{\mathcal{H}}$ satisfies this property and, consequently, can be used to determine the weights $\hat{\mathbf{w}}^{\mathcal{H}adl}$. As noticed in [39], examining carefully the provided proof, this assumption on $\hat{\beta}$ can be greatly weakened.

Same kind of results can be proved in the random design setting using similar techniques.

4. Simulation results

In this section, the algorithm minimising the objective function Q^{adl} (resp. Q^{ladl} and $Q^{\mathcal{H}adl}$) is called `ad-lasso` (resp. `LAD-ad-lasso` and `Huber-ad-lasso`). The adaptive weights are obtained from the corresponding unpenalized estimator and $\gamma = 1$. Our aim is to compare the finite sample performances of these procedures. The purpose of these simulations is to see how our estimator performs in the absence of outliers (where robustness is not essential) and in their presence (where robustness is needed) and in comparison with other robust (`LAD-ad-lasso`) or non-robust methods (`ad-lasso`). Paragraph 4.1 presents the studied models. The way simulations are conducted is described in 4.2 and an insight of conclusions is provided in paragraph 4.3.

4.1. Models used for simulations

The models used to compare the performances of the algorithms are inspired by those presented in [32], [39] and [36]. They all have the form $\underline{y} = \mathbb{1}_n + \mathbf{X}\beta^* + \sigma\underline{\epsilon}$, where $\mathbb{1}_n$ denotes the vector of $\mathbb{R}^n$ composed of ones and $\underline{y}$ (resp. $\underline{\epsilon}$) represents the response (resp. error) vector $(y_1, \ldots, y_n)^T$ (resp. $(\epsilon_1, \ldots, \epsilon_n)^T$). The vector of true coefficients is $\beta^* = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$. As compared with (2.1), this means that the intercept of the model is $\alpha^* = 1$ and the number of variables (without the intercept) is $p = 8$. The number of influencing variables is $p_0 = 3$. The design matrix $\mathbf{X}$ is constructed as follows. The rows of $\mathbf{X}$ are given by n independent gaussian vectors $N_8(0, \Sigma_r)$. They are normalized such that the corresponding p-dimensional covariables are centered (as assumed in (2.1)). For some $r \geq 0$, the variance matrix of the variables is defined by $\Sigma_{r,i,j} = r^{|i-j|}$ for $1 \leq i, j \leq p$.

- Model 1: *low correlation, gaussian noise.* $\underline{\epsilon} \sim N_n(0, I_n)$, $\sigma = 2$ and $r = 0.5$.
- Model 2: *high correlation, gaussian noise.* $\underline{\epsilon} \sim N_n(0, I_n)$, $\sigma = 2$ and $r = 0.95$.
- Model 3: *large outliers.* $\epsilon = V/\sqrt{\text{var}(V)}$, $\sigma = 9.67$ and $r = 0.5$. V is a mixture of gaussians: with probability 0.9, $V \sim N(0, 1)$ and, otherwise, $V \sim N(0, 225)$. Note that $\sqrt{\text{var}(V)} = 4.83$.

- Model 4: *sensible outliers*. $\epsilon = D/\sqrt{\text{var}(D)}$, $\sigma = 9.67$ and $r = 0.5$. The distribution of D is a standard double exponential i.e. its density is $x \in \mathbb{R} \to e^{-|x|}/2$ and $\text{var}(D) = 2$.

These four models can be divided into two types. The first type contains light tailed errors models (1 and 2) whereas the second type is composed of heavy tailed errors models (3 and 4). Models 1 and 2 allow to quantify the deterioration of the performances of the robust methods `LAD-ad-lasso` and `Huber-ad-lasso` in the absence of outliers. Thinking about the maximum likelihood approach, the loss used in `ad-lasso` (resp. `Huber-ad-lasso`, `LAD-ad-lasso`) is well designed for models 1 and 2 (resp. 3,4); see [12] for a justification of this claim for `Huber-ad-lasso`.

4.2. Assessing prediction methods

4.2.1. Prediction accuracy

The following procedure has been designed to compare the performances of the various algorithms in the fixed design setting. The performances are measured both by the prediction errors and the model selection ability. For any considered underlying models, the distribution of the design matrix is given. It is used to generate the covariates used in the training and test datasets. So we generate a first set of n training designs $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ and a second set of m =10 000 test designs $(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_{n+m})$.

These two sets are centered in mean to stick on the theoretical definition (2.1) of the model. For the `ad-lasso`, [39] recommends to standardize the design in mean (i.e. ensures that $\sum_{i=1}^{n} \mathbf{x}_i = 0$) which only leads to $\hat{\alpha} = \overline{y}$ and does not affect the $\hat{\beta}$ value (since the intercept is not penalized and the squared loss is used). Concerning the `LAD-ad-lasso`, the intercept is not included in the study of [36] and no recommendation are provided for the beforehand normalizations of the data. Moreover, their simulations are performed on stochastically *independent* covariables. If $(\hat{\alpha}^{ladl}, \hat{\beta}^{ladl})$ denotes the `LAD-ad-lasso` estimator, $\hat{\alpha}^{ladl}$ is the median of the n residuals $y_i - \mathbf{x}_i^T \hat{\beta}^{ladl}$ for $i = 1 \cdots n$. In a general way, for the `LAD-ad-lasso` or `Huber-ad-lasso`, such a normalization has some effect over $\hat{\beta}$ but it is not so clear how it works (since the loss function is no more the squared loss). In this paper, we follow the normalization of design in mean provided in [39] for the three procedures.

Since the theoretical results are established in fix design framework, the training and test design are fixed once and for all: they will be used for *all* the data generations. 100 training sets of size n are generated according to definition (2.1) of the model. All the algorithms have been runned on the 100 training sets of size n =50, 100, 200 and their prediction capacity have been evaluated on the test design set of size m =10 000.

In order to compare the prediction accuracy, the Relative Prediction Errors (RPEs) already considered in [39] are computed. Let us now precisely recall the definition of this index. In our model, it is assumed that the true regression

TABLE 1
Means of the RPE (standard deviation of the 100 RPE) with cross-validation. Best mean indicated in italic

	Model 1	Model 2	Model 3	Model 4
ad-lasso (n=50)	*0.140(0.101)*	*0.174(0.105)*	0.142(0.112)	0.196(0.115)
LAD-ad-lasso (n=50)	0.233(0.154)	0.231(0.135)	0.025(0.031)	*0.138(0.106)*
Huber-ad-lasso (n=50)	0.156(0.107)	0.179(0.109)	*0.024(0.037)*	0.144(0.094)
ad-lasso (n=100)	*0.075(0.053)*	*0.088(0.055)*	0.087(0.061)	0.086(0.051)
LAD-ad-lasso (n=100)	0.103(0.072)	0.122(0.074)	0.008(0.009)	0.065(0.043)
Huber-ad-lasso (n=100)	0.076(0.054)	0.093(0.056)	*0.008(0.011)*	*0.064(0.037)*
ad-lasso (n=200)	*0.027(0.019)*	*0.034(0.021)*	0.039(0.027)	0.036(0.022)
LAD-ad-lasso (n=200)	0.038(0.023)	0.050(0.029)	0.003(0.002)	*0.023(0.015)*
Huber-ad-lasso (n=200)	0.029(0.019)	0.039(0.024)	*0.002(0.002)*	0.024(0.017)

TABLE 2
Means of the RPE (standard deviation of the 100 RPE) with BIC. Best mean indicated in italic

	Model 1	Model 2	Model 3	Model 4
ad-lasso (n=50)	*0.125(0.094)*	*0.167(0.111)*	0.153(0.105)	0.174(0.086)
LAD-ad-lasso (n=50)	0.215(0.165)	0.254(0.152)	*0.016(0.016)*	*0.126(0.090)*
Huber-ad-lasso (n=50)	0.188(0.159)	0.214(0.139)	0.016(0.028)	0.143(0.087)
ad-lasso (n=100)	*0.060(0.045)*	*0.093(0.072)*	0.079(0.051)	0.091(0.054)
LAD-ad-lasso (n=100)	0.096(0.068)	0.122(0.065)	0.004(0.003)	*0.065(0.043)*
Huber-ad-lasso (n=100)	0.072(0.066)	0.098(0.070)	*0.004(0.003)*	0.066(0.042)
ad-lasso (n=200)	*0.020(0.014)*	0.045(0.034)	0.046(0.036)	0.041(0.025)
LAD-ad-lasso (n=200)	0.035(0.021)	0.055(0.035)	0.001(0.001)	*0.023(0.017)*
Huber-ad-lasso (n=200)	0.025(0.019)	*0.039(0.029)*	*0.001(9e-04)*	0.025(0.018)

function is $m(\mathbf{x}) = \alpha^* + \mathbf{x}^T\beta^*$ for any $\mathbf{x} \in \mathbb{R}^p$. Consequently, any estimator $(\hat{\alpha}, \hat{\beta})$ of (α^*, β^*) leads to the estimation $m_n(\mathbf{x}) = \hat{\alpha} + \mathbf{x}^T\hat{\beta}$ of the regression function. The following decomposition of the excess risk is classical when we are interested in regression functions (see e.g. [9]):

$$\mathbb{E}_{x,y}\left[(y - m_n(x))^2\right] - \mathbb{E}_{x,y}\left[(y - m(x))^2\right] = \mathbb{E}_x\left[(m_n(x) - m(x))^2\right].$$

Since in our model $\mathbb{E}_{x,y}\left[(y - m(x))^2\right] = \sigma^2$, the relative prediction error $\mathbb{E}_x\left[(m_n(x) - m(x))^2\right]/\sigma^2$ is reported. In our simulations, it is estimated by the corresponding mean $\sum_{j=1}^{m}(m_n(\mathbf{x}_{n+j}) - m(\mathbf{x}_{n+j}))^2/(m\sigma^2)$ over the test sample. Tables 1 and 2 provide the mean and the standard deviation of the 100 obtained RPE. Figures 1 and 2 provide the corresponding boxplots.

4.2.2. Selection ability

The model selection ability of the algorithms are reported in the same manner as done by [36], [32] and [4] in Tables 3, 4, 5 and 6 for cross-validation and in Tables 7, 8, 9 and 10 for BIC. In order to provide the indicators defined below, a coefficient is considered to be zero if it absolute value is strictly less than 10^{-5} (i.e. its five first decimals vanish). In all cases, amongst the 100 obtained estimators, the first column (C) counts the number of well chosen models i.e. the cases where the first, second and fifth coordinates of $\hat{\beta}$ are non-zeros *and*

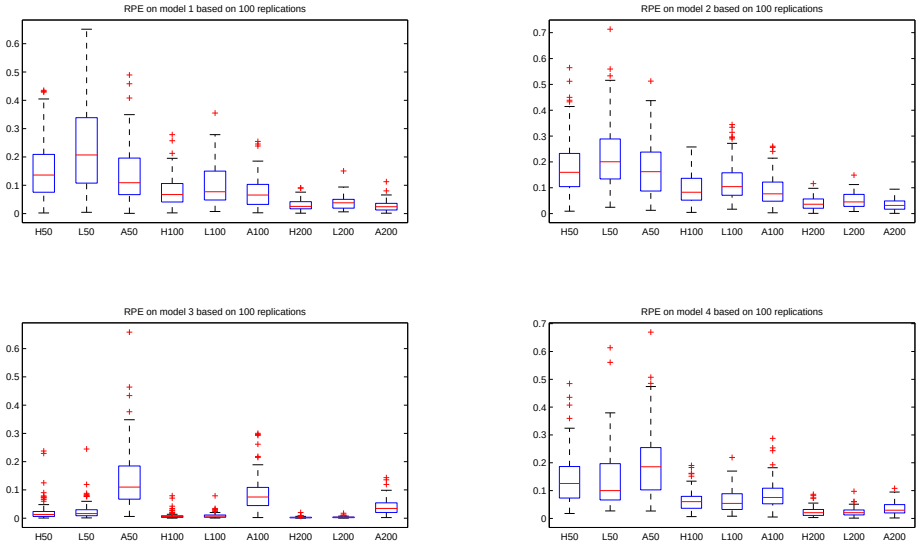

FIG 1. $n = 50, 100, 200$*: RPE of* `Huber-ad-lasso` *(H),* `LAD-ad-lasso` *(L) and* `ad-lasso` *(A) with cross-validation.*

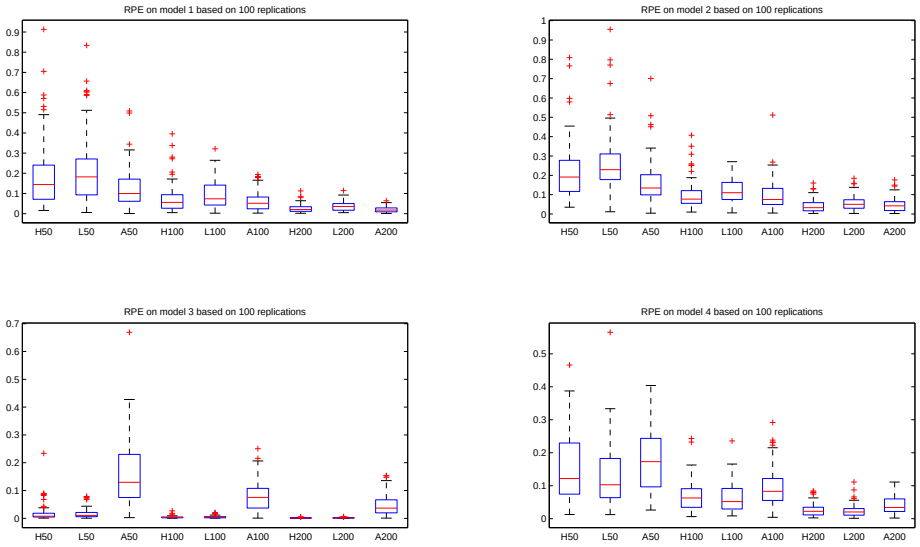

FIG 2. $n = 50, 100, 200$*: RPE of* `Huber-ad-lasso` *(H),* `LAD-ad-lasso` *(L) and* `ad-lasso` *(A) with BIC.*

TABLE 3
Selection model ability on Model 1 based on 100 replications and cross-validation

	C	O	U	SV	MM	IZ	CZ	Z
ad-lasso (n=50)	34	66	0	4	[-1.24 ; 1.01]	0	3.75	3.75
LAD-ad-lasso(n=50)	27	70	3	4	[-1.24 ; 1.33]	0.03	3.44	3.47
Huber-ad-lasso (n=50)	35	64	1	4	[-1.14 ; 1.08]	0.01	3.75	3.76
ad-lasso (n=100)	40	60	0	4	[-1.08 ; 0.78]	0	3.79	3.79
LAD-ad-lasso (n=100)	35	65	0	4	[-1.15 ; 0.83]	0	3.62	3.62
Huber-ad-lasso (n=100)	38	62	0	4	[-0.79 ; 0.73]	0	3.88	3.88
ad-lasso (n=200)	49	51	0	4	[-0.52 ; 0.53]	0	4.02	4.02
LAD-ad-lasso (n=200)	42	58	0	4	[-0.43 ; 0.56]	0	4.03	4.03
Huber-ad-lasso (n=200)	47	53	0	4	[-0.42 ; 0.50]	0	4.04	4.04

TABLE 4
Selection model ability on Model 2 based on 100 replications and cross-validation

	C	O	U	SV	MM	IZ	CZ	Z
ad-lasso (n=50)	14	18	68	3	[-4.12 ; 3.39]	0.84	3.58	4.42
LAD-ad-lasso (n=50)	8	25	67	3	[-4.28 ; 4.08]	0.87	3.41	4.28
Huber-ad-lasso (n=50)	12	16	72	3	[-3.53 ; 3.26]	0.91	3.63	4.54
ad-lasso (n=100)	19	42	39	4	[-3.58 ; 2.80]	0.43	3.85	4.28
LAD-ad-lasso (n=100)	12	45	43	4	[-3.68 ; 3.73]	0.48	3.51	3.99
Huber-ad-lasso (n=100)	14	39	47	4	[-2.71 ; 2.89]	0.49	3.73	4.22
ad-lasso (n=200)	33	51	16	4	[-1.66 ; 1.79]	0.16	3.85	4.01
LAD-ad-lasso (n=200)	25	51	24	4	[-1.51 ; 2.58]	0.25	3.66	3.91
Huber-ad-lasso (n=200)	26	55	19	4	[-1.45 ; 2.02]	0.19	3.76	3.95

TABLE 5
Selection model ability on Model 3 based on 100 replications and cross-validation

	C	O	U	SV	MM	IZ	CZ	Z
ad-lasso (n=50)	5	26	69	4	[-6.39 ; 6.52]	1.14	3.21	4.35
LAD-ad-lasso (n=50)	18	59	23	4	[-1.66 ; 2.08]	0.28	3.10	3.38
Huber-ad-lasso (n=50)	17	62	21	4	[-1.56 ; 1.75]	0.28	3.29	3.57
ad-lasso (n=100)	7	31	62	4	[-3.75 ; 5.16]	0.82	3.60	4.42
LAD-ad-lasso (n=100)	39	58	3	4	[-1.28 ; 0.97]	0.04	3.39	3.43
Huber-ad-lasso (n=100)	39	56	5	4	[-1.20 ; 0.97]	0.06	3.57	3.63
ad-lasso (n=200)	6	65	29	4	[-2.16 ; 3.13]	0.30	3.31	3.61
LAD-ad-lasso (n=200)	39	61	0	4	[-0.57 ; 0.68]	0	3.17	3.17
Huber-ad-lasso (n=200)	42	58	0	4	[-0.68 ; 0.61]	0	3.32	3.32

TABLE 6
Selection model ability on Model 4 based on 100 replications and cross-validation

	C	O	U	SV	MM	IZ	CZ	Z
ad-lasso (n=50)	4	8	88	3	[-7.32 ; 6.20]	1.54	3.62	5.16
LAD-ad-lasso (n=50)	1	32	67	3	[-7.66 ; 5.74]	1.04	3.39	4.43
Huber-ad-lasso (n=50)	3	20	77	3	[-7.05 ; 5.07]	1.20	3.60	4.80
ad-lasso (n=100)	8	32	60	4	[-3.93 ; 3.68]	0.80	3.44	4.24
LAD-ad-lasso (n=100)	7	51	42	4	[-3.52 ; 2.92]	0.51	3.18	3.69
Huber-ad-lasso (n=100)	12	37	51	4	[-3.47 ; 3.14]	0.60	3.37	3.97
ad-lasso (n=200)	18	47	35	4	[-2.50 ; 2.88]	0.39	3.66	4.05
LAD-ad-lasso (n=200)	16	66	18	4	[-2.12 ; 2.01]	0.19	3.24	3.43
Huber-ad-lasso (n=200)	26	60	14	4	[-1.94 ; 2.32]	0.15	3.62	3.77

TABLE 7
Selection model ability on Model 1 based on 100 replications and BIC's model selection

	C	O	U	SV	MM	IZ	CZ	Z
ad-lasso (n=50)	70	29	1	3	[-1.27 ; 0.93]	0.01	4.59	4.60
LAD-ad-lasso(n=50)	49	45	6	3	[-1.24 ; 1.15]	0.06	4.25	4.31
Huber-ad-lasso (n=50)	62	32	6	3	[-1.57 ; 1.33]	0.06	4.40	4.46
ad-lasso (n=100)	74	26	0	3	[-0.76 ; 0.64]	0	4.70	4.70
LAD-ad-lasso (n=100)	60	40	0	3	[-0.90 ; 0.76]	0	4.39	4.39
Huber-ad-lasso (n=100)	79	21	0	3	[-0.67 ; 1.40]	0	4.71	4.71
ad-lasso (n=200)	92	8	0	3	[-0.38 ; 0.34]	0	4.92	4.92
LAD-ad-lasso (n=200)	74	26	0	3	[-0.43 ; 0.47]	0	4.70	4.70
Huber-ad-lasso (n=200)	92	8	0	3	[-0.38 ; 0.41]	0	4.91	4.91

TABLE 8
Selection model ability on Model 2 based on 100 replications and BIC's model selection

	C	O	U	SV	MM	IZ	CZ	Z
ad-lasso (n=50)	13	11	76	3	[-4.49 ; 3.01]	0.89	4.27	5.16
LAD-ad-lasso (n=50)	10	15	75	3	[-4.28 ; 4.71]	1.02	3.86	4.88
Huber-ad-lasso (n=50)	19	6	75	3	[-4.06 ; 3.95]	0.97	4.10	5.07
ad-lasso (n=100)	30	9	61	3	[-1.88 ; 2.68]	0.70	4.55	5.25
LAD-ad-lasso (n=100)	22	12	66	3	[-2.61 ; 3.38]	0.78	4.23	5.01
Huber-ad-lasso (n=100)	26	10	64	3	[-2.43 ; 4.60]	0.71	4.40	5.11
ad-lasso(n=200)	45	21	34	3	[-1.56 ; 2.77]	0.38	4.47	4.85
LAD-ad-lasso (n=200)	42	23	35	3	[-2.28 ; 2.62]	0.38	4.41	4.79
Huber-ad-lasso (n=200)	57	15	28	3	[-1.23 ; 1.81]	0.29	4.70	4.99

TABLE 9
Selection model ability on Model 3 based on 100 replications and BIC's model selection

	C	O	U	SV	MM	IZ	CZ	Z
ad-lasso (n=50)	4	7	89	2	[-6.94 ; 6.12]	1.75	4.45	6.20
LAD-ad-lasso (n=50)	52	27	21	3	[-1.20 ; 2.00]	0.26	4.62	4.88
Huber-ad-lasso (n=50)	70	8	22	3	[-1.50 ; 1.50]	0.29	4.81	5.10
ad-lasso (n=100)	10	5	85	2	[-2.39 ; 3.20]	1.25	4.53	5.78
LAD-ad-lasso (n=100)	82	16	2	3	[-0.83 ; 1.07]	0.02	4.80	4.82
Huber-ad-lasso (n=100)	90	8	2	3	[-0.68 ; 1.00]	0.02	4.89	4.91
ad-lasso (n=200)	27	17	56	3	[-2.53 ; 2.73]	0.73	4.48	5.21
LAD-ad-lasso (n=200)	98	2	0	3	[-0.69 ; 0.52]	0	4.97	4.97
Huber-ad-lasso (n=200)	100	0	0	3	[0 ; 0]	0	5.00	5.00

TABLE 10
Selection model ability on Model 4 based on 100 replications and BIC's model selection

	C	O	U	SV	MM	IZ	CZ	Z
ad-lasso (n=50)	2	0	98	1	[-5.79 ; 4.54]	2	4.60	6.60
LAD-ad-lasso (n=50)	6	6	88	2	[-7.09 ; 4.21]	1.49	4.53	6.02
Huber-ad-lasso (n=50)	3	3	94	2	[-5.93 ; 4.20]	1.74	4.59	6.33
ad-lasso (n=100)	9	8	83	2	[-4.06 ; 3.55]	1.29	4.41	5.70
LAD-ad-lasso (n=100)	20	8	72	2	[-2.47 ; 2.86]	0.96	4.57	5.53
Huber-ad-lasso (n=100)	18	8	74	2	[-3.12 ; 2.65]	1.07	4.52	5.59
ad-lasso (n=200)	27	9	64	3	[-2.19 ; 2.85]	0.75	4.65	5.40
LAD-ad-lasso (n=200)	42	18	40	3	[-1.38 ; 2.24]	0.42	4.60	5.02
Huber-ad-lasso (n=200)	36	17	47	3	[-1.54 ; 1.63]	0.52	4.66	5.18

the third, fourth, sixth, seventh and eightth ones are zeros. To go further in the model selection ability analysis, we consider another measurements. To begin with, the second column (O) reports the number of overfitting models i.e. those selecting all the non-zeros coefficients *and* at least one zero coefficient. Next, the third column (U) reports the number of chosen underfitting models i.e. those not selecting at least one non-zero coefficient. In this way, all the 100 models are counted one time. Columns (0) and (U) aim to explain the results obtained in (C). The fourth column (SV) reports the median number of selected variables. Finally, the fifth column (MM) reports the minimal and maximal values of estimations of the coefficients of non-influencing variables. For each n and each column, the best performance is indicated in bold. The sixth (resp. seventh) column (IZ) (resp. (CZ)) provides the average number of correctly (resp. mistakenly) estimated 0's. The last column (Z) is the average number of estimated 0's.

Models selection abilities are closely related to the accuracy of estimations of the coefficients. This fact is illustated by columns (MM) and by boxplots of the coefficients estimations of Figures 3 and 4.

4.2.3. Hyperparameter choices

Concerning the hyperparameter choices, the regularization parameters are chosen by either BIC criterion or a 5−fold cross validation on *each* of the 100 training sets. The same grid has always been used. It is composed of 0 and 100 points log-linearly spaced between 0.01 and 1400. For `Huber-ad-lasso`, the simulation studies report the performances obtained with $M = 1.345$. This value has been recommended by Huber in [12]. Let us remark that it is possible to chose the M parameter from the data (for example by cross-validation simultaneous with the tuning parameter). But in practice we do not observe some improvement to make it data adaptive. It is also noticeable that, within each of the four considered models, the two model selection procedures have been tested on the same simulated datasets.

4.3. Comparison results

To begin with, let us point out some fact observed both with a cross-validation and BIC model selection. Gaussian tailed errors models (1 and 2) emphasize that the use of Huber loss with concomitant scale instead of the squared loss does not lead to a significant loss of performances in the absence of outlier. Indeed, the relative prediction errors and model selection ability of `Huber-ad-lasso` are closed to the ones of `ad-lasso` (see Tables 1, 2, 3 and 7). It is noticeable that, as attended, the results of `LAD-ad-lasso` are the worst from a prediction and model selection point of view. Let us also stress that in the case of large correlations between the covariables (Model 2), the use of Huber loss does not solve the poor behavior of the quadratic loss: in this case, in comparison with the low correlated case, we even observe a slightly more marked deterioration of the

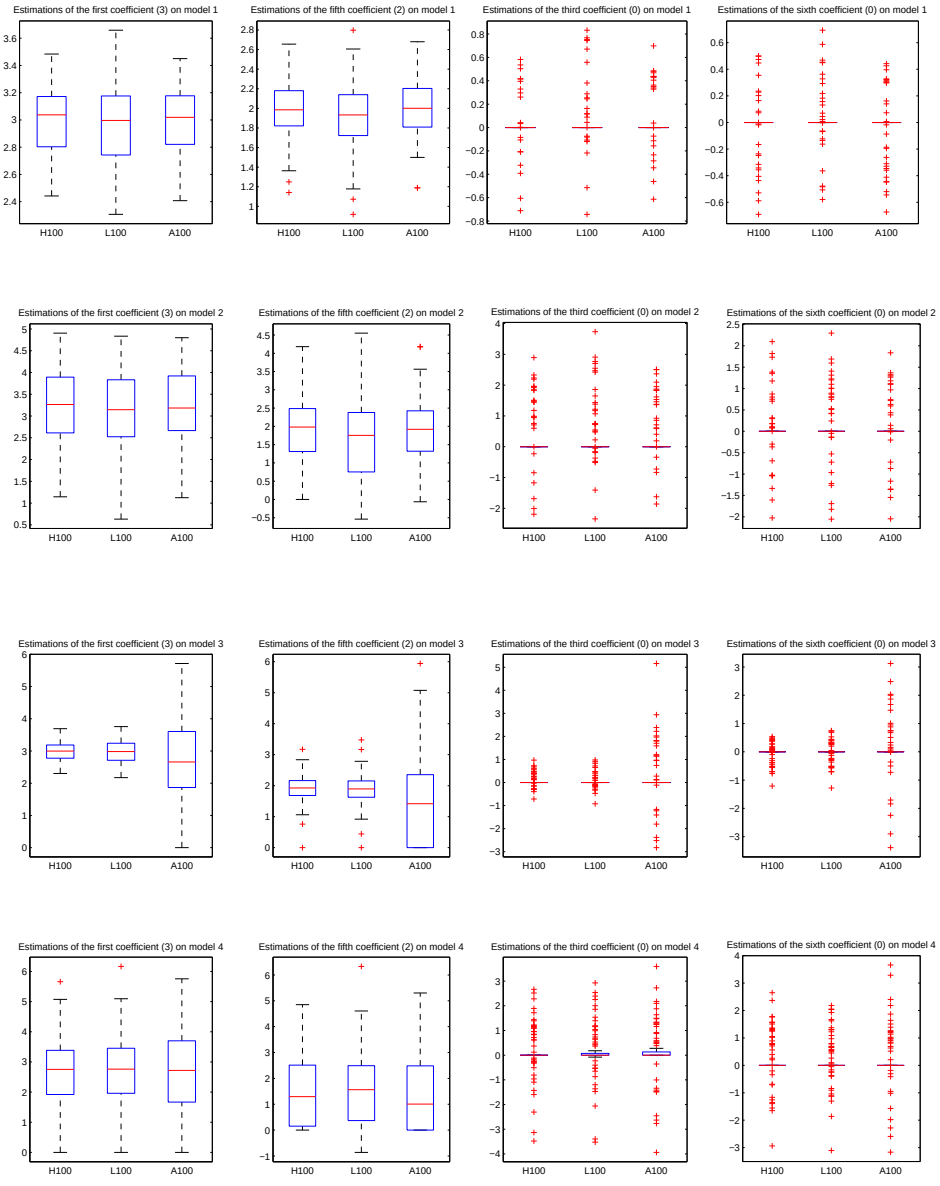

FIG 3. *For $n = 100$, estimations of influencing and non-influencing coefficients by* `Huber-ad-lasso` *(H),* `LAD-ad-lasso` *(L) and* `ad-lasso` *(A) with cross-validation.*

performances of `Huber-ad-lasso` with respect to `adaptative-lasso`. To settle this, one has to change the penalty: this is a point for a future work. Heavy tailed errors models (3 and 4) emphasize better performances of `Huber-ad-lasso` with respect to `ad-lasso`. More precisely, the quadratic loss is particularly unsuited

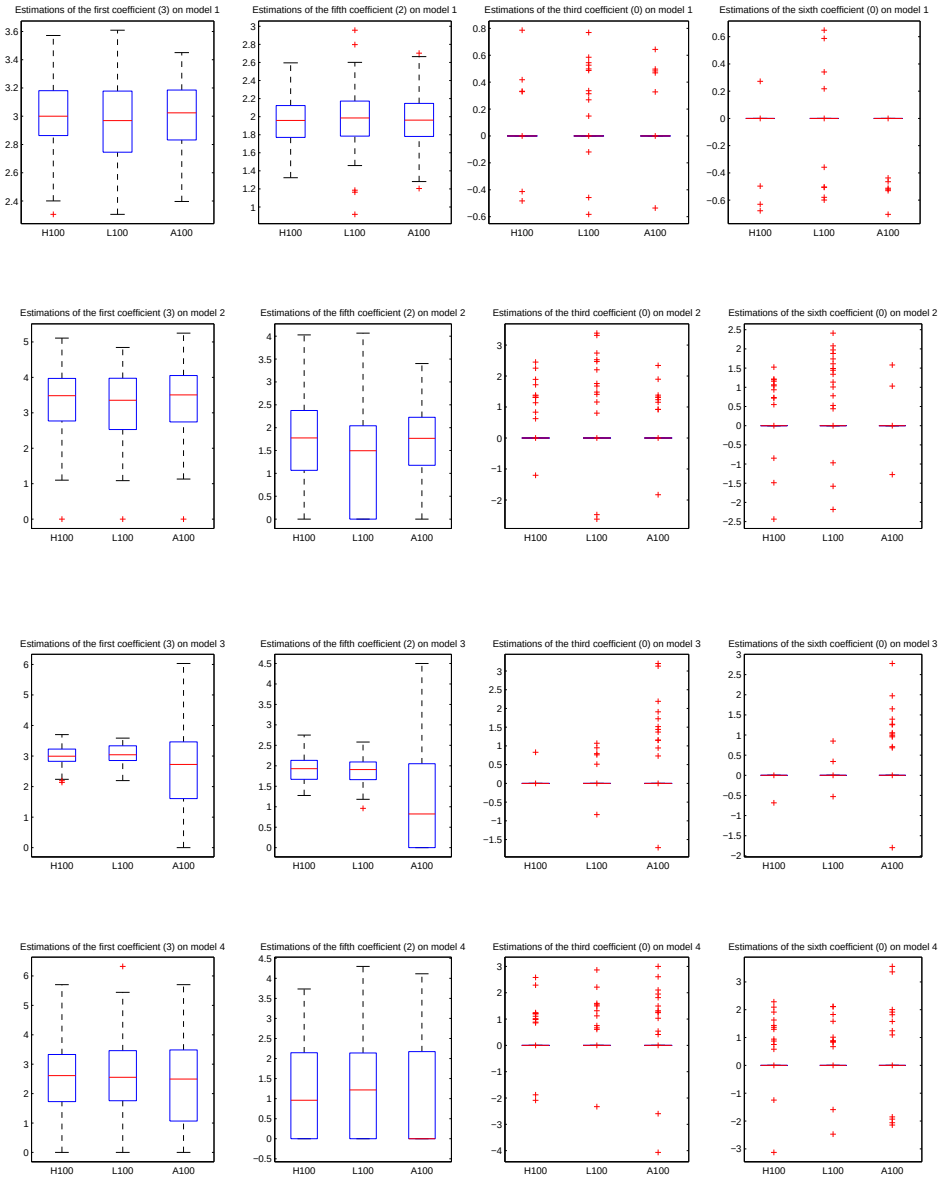

FIG 4. *For $n = 100$, estimations of influencing and non-influencing coefficients by* `Huber-ad-lasso` *(H),* `LAD-ad-lasso` *(L) and* `ad-lasso` *(A) with BIC.*

in presence of a small quantity of potentially large outliers (Model 3). Indeed, the relative prediction errors of `ad-lasso` are around ten times as big as the ones of `Huber-ad-lasso` (see Tables 1 and 2) in this Model 3. The standard deviation is also far bigger (see Figures 1 and 2). In this case, `ad-lasso` leads

to a poor estimation of the coefficients (see Figures 3 and 4). In presence of sensible outliers (Model 4), the quadratic loss get worst results than Huber loss but the gap is smaller. Let us now compare the `Huber-ad-lasso` with a robust-designed algorithm: the `LAD-ad-lasso`. In the gaussian tails errors models (1 and 2) `Huber-ad-lasso` always get better results than the `LAD-ad-lasso` from prediction error *and* model selection ability point of view (see Tables 1, 2, 3, 7, 4 and 8). Tables 4 and 8 emphasizes that the benefit decreases when the correlations between variables increase.

Next, let us point out some differences due to the model selection procedure. These differences occur for the comparison of `Huber-ad-lasso` and `LAD-ad-lasso` in the heavy tails models (model 3 and 4). When cross-validation is used, in model 3, `Huber-ad-lasso` and `LAD-ad-lasso` get similar performances in terms of prediction error and model selection ability (see Tables 1 and 5) while in model 4 `Huber-ad-lasso` has a better model selection ability than `LAD-ad-lasso`. As a contrary, when a BIC-type criterion is used, the results are less surprising. Indeed, the nature of the noise directly determines the more performant algorithm (from a prediction error and model selection ability points of view). Precisely, as attended, the mixture of gaussians (resp. double exponential) noise of model 3 (resp. 4) favours `Huber-ad-lasso` (resp. `LAD-ad-lasso`) over `LAD-ad-lasso` (resp. `Huber-ad-lasso`).

Finally, let us provide some general concluding remarks highlighted by these experiments. If the model selection is performed by cross-validation, it is preferable to use `Huber-ad-lasso` rather than `ad-lasso` or `LAD-ad-lasso`. However, it is noticeable that for each of the four models, each value of n and each method, the performances in terms of model selection ability are better with a BIC-type criterion rather than with cross-validation. Moreover, roughly speaking, BIC-type procedures are 5 times faster than the corresponding cross-validation procedure.

5. A real example: The Chinese stock market data

In this section we consider for illustrating purposes the Chinese stock market data analyzed in [36]. These data set is derived from CCER China Stock, which was partially developed by the China Center for Economic Research (CCER) at Peking University. It contains a total of 2 247 records; each record corresponds to one yearly observation of one company. In these data set, 1 092 records come from year 2002 and the rest comes from year 2003. As in [36] we use year 2002 records as the training data and year 2003 records as the testing data. The response variable is the return on equity (denoted by ROE) of the following year (denoted by ROE_{t+1}). The explanatory variables all measured at year t include ROE of the current year (denoted by ROE_t), asset turnover ratio (ATO), profit margin (PM), debt-to-asset ratio or leverage (LEV), sales growth rate (GROWTH), price-to-book ratio (PB), account receivables/revenues (ARR), inventory/asset (INV), and the logarithm of total assets (ASSET).

The reliability of the usual OLS-based estimation and model selection methods (e.g. lasso) is severely challenged for these data set (see [36]), whereas the

TABLE 11
Estimation results for Chinese stock market (`H-ad-l` *for* `Huber-ad-lasso`*,* `L-ad-l` `LAD-ad-lasso` *and* `ad-l` *for* `ad-lasso`*)*

	cross-validation			BIC criterion					
	`H-ad-l`	`L-ad-l`	`ad-l`	`H-ad-l`	`L-ad-l`	`ad-l`	`LAD`	`OLS`	`Huber`
Int	0.042	0.035	-0.028	0.034	0.034	-0.028	0.035	-0.028	0.033
ROEt	0.069	0.190	-0.114	0.135	0.191	-0.177	0.190	-0.181	0.145
ATO	0.054	0.058	0.128	0.064	0.060	0.159	0.058	0.165	0.066
PM	0.124	0.134	0.062	0.155	0.142	0.188	0.134	0.209	0.159
LEV	-0.014	-0.023	-0.240	-0.024	-0.020	-0.245	-0.023	-0.245	-0.029
GR	0.002	0.019	0.026	0.015	0.016	0.032	0.019	0.033	0.018
PB	0.000	0.001	0.017	0.000	0.000	0.018	0.001	0.018	0.001
ARR	-0.000	-0.002	-0.000	-0.000	-0.000	0.000	-0.002	0.009	-0.001
INV	0.000	0.012	0.209	0.000	0.000	0.344	0.012	0.354	0.034
ASSET	0.007	0.016	0.089	0.015	0.013	0.100	0.016	0.101	0.019
MAPE	0.116	0.120	0.229	0.119	0.120	0.233	0.120	0.234	0.120
STD	0.021	0.021	0.020	0.021	0.021	0.020	0.021	0.021	0.021

robust selection methods (such as LAD or Huber based methods) become more attractive. We run the three methods `ad-lasso`, `LAD-ad-lasso` and `Huber-ad-lasso` for the both hyperparameters choice methods (cross-validation and BIC criterion). Each method is used to select the best model based on the training dataset of 2002. The prediction accuracies of these methods are measured by the mean absolute prediction error (MAPE) and the corresponding standard deviation (STD), evaluated on the testing data for 2003. The STD denotes the standard deviation of the absolute values of the prediction errors divided by the square root of the number of testing data. Let us notice that contrary to [36], we do not penalize the intercept term. For comparison purposes, the results of the full model based on the LAD, OLS ad Huber estimators are also reported (see Table 11).

We find similar results as in [36]. The MAPE of the OLS (with or without variables selection) is as large and is substantially worse than all other robust methods (like LAD or Huber). That justifies the use of the robust methods for this data set. According to the reported standard error of the MAPE estimate, we can clearly see that a difference cannot be statistically significant between all the robust methods. Based on a substantially simplified model, the prediction accuracy of the robust lasso estimator remains very satisfactory. In particular, the `Huber-ad-lasso` leads to the same set of selected variables for the both hyperparameters choice methods whereas this is not the case for `LAD-ad-lasso`.

6. Appendix

6.1. *Proof of Theorem 3.1*

We have $\hat{s}^{\mathcal{H}adl} = \operatorname{argmin}_{s\geq 0} \mathcal{L}_{\mathcal{H}}(\hat{\alpha}^{\mathcal{H}adl}, \hat{\beta}^{\mathcal{H}adl}, s)$. The continuity and convexity of the objective function defining the estimators with respect to the involved variables implies that the new objective function $f(s) = \mathcal{L}_{\mathcal{H}}(\hat{\alpha}^{\mathcal{H}adl}, \hat{\beta}^{\mathcal{H}adl}, s)$ is convex continuous on $\mathbb{R}_+$. The function f is of class $\mathcal{C}^1$ on $\mathbb{R}_+^*$ by composition

of functions of class $\mathcal{C}^1$. Moreover, as s goes to 0, $f'(s) \to n - M^2\#\{1 \leq i \leq n, Y_i - \hat{\alpha}^{\mathcal{H}adl} - \mathbf{x}_i^T\hat{\beta}^{\mathcal{H}adl} \neq 0\}$. Thus, from the mean value theorem, f is right differentiable at 0 with $f'(0) = n - M^2\#\{1 \leq i \leq n, Y_i - \hat{\alpha}^{\mathcal{H}adl} - \mathbf{x}_i^T\hat{\beta}^{\mathcal{H}adl} \neq 0\}$. From now on, we denote $r_{(i)} = r^2_{(i)}(\hat{\alpha}^{\mathcal{H}adl}, \hat{\beta}^{\mathcal{H}adl})$ for $i = 1, \ldots, n$.

The set $\mathbb{R}_+$ can be cut up into intervals such that, on each interval, the function f get the form $\frac{a}{s} + bs + c$ for various constants a, b and c. Precisely, we have:

$$f(s) = \begin{cases} 2M\sum_{i=1}^n |r_{(i)}| & \text{if case 1,} \\ 2M\sum_{i=1}^n |r_{(i)}| + ns(1 - M^2) & \text{if case 2,} \\ 2M\sum_{i=k+1}^n |r_{(i)}| + \frac{1}{s}\sum_{i=1}^k r^2_{(i)} + s(n - M^2(n-k)) & \text{if case 3,} \\ ns + \frac{1}{s}\sum_{i=1}^n r^2_{(i)} & \text{if case 4,} \end{cases}$$

where

$$\begin{aligned} &\text{case 1:} \quad s = 0, \\ &\text{case 2:} \quad 0 < s \leq \tfrac{|r_{(1)}|}{M}, \\ &\text{case 3:} \quad \exists\, k \in [1, n-1]\,, \tfrac{|r_{(k)}|}{M} < s \leq \tfrac{|r_{(k+1)}|}{M}, \\ &\text{case 4:} \quad s > \tfrac{|r_{(n)}|}{M}\,. \end{aligned}$$

The function $s \to a/s + bs + c$ with $a \geq 0$ is decreasing on $\mathbb{R}_+$ if $b \leq 0$ whereas it is decreasing on $[0, \sqrt{a/b}]$ and increasing on $[\sqrt{a/b}, \infty[$ if $b > 0$ and so its minimum is reached at $\sqrt{a/b}$.

When $M < 1$, all the b coefficients are strictly positive but all the functions $s \to a/s + bs + c$ are considered on their increasing part. By continuity of f on $\mathbb{R}_+$, the function f is increasing in this case. Thus, $\hat{s}^{\mathcal{H}adl} = 0$.

When $M = 1$, the function f is constant on $[0, |r_{(1)}|]$ and increasing on $[|r_{(1)}|, +\infty[$. Thus, if $r_{(1)} \neq 0$, the argmin is not unique but we can choose $\hat{s}^{\mathcal{H}adl} = 0$. Thus, if $M \leq 1$, $\hat{s}^{\mathcal{H}adl} = 0$. The definition of $(\hat{\alpha}^{\mathcal{H}adl}, \hat{\beta}^{\mathcal{H}adl}, \hat{s}^{\mathcal{H}adl})$ and (2.2) implies that $(\hat{\alpha}^{\mathcal{H}adl}, \hat{\beta}^{\mathcal{H}adl})$ is an argmin of the penalized loss $2M\sum_{i=1}^n |y_i - \alpha - \mathbf{x}_i^T\beta|$. The first point of Theorem 3.1 is established.

Next, we suppose that $M > 1$. For the intervals having $k \leq n\left(1 - 1/(M^2)\right)$, the b coefficients are negative ($b \leq 0$). Thus, by continuity of f on $\mathbb{R}_+$, the function f is decreasing on these intervals. The interval $\left[|r_{(k)}|/M, |r_{(k+1)}|/M\right]$ with $k = \lceil n\left(1 - 1/M^2\right)\rceil$ is the first one where the function f can be minimal. By convexity of f, among all the intervals, there is only one where f is a valley. Expression and unicity of $\hat{s}^{\mathcal{H}adl}$ come from f expression over this interval. □

6.2. Proof of Theorem 3.2

In the Step 1, we prove the asymptotic normality of this estimator and, in the Step 2, its consistency in variable selection.

Step 1. Let us first prove the asymptotic normality. This proof in an adaptation to our case of the proof given by [39]. The only difference concerns the

treatment of the loss function. So in the following, we will use notations similar to the ones of [39]. We will point out the difference between the both proofs.

Let us define $U_n(u) = Q^{\mathcal{H}adl}\left((\alpha^*, \beta^*, s^*)^T + u/\sqrt{n}\right) - Q^{\mathcal{H}adl}(\alpha^*, \beta^*, s^*)$ with $u = (u_0, \ldots, u_{p+1})^T \in \mathbb{R}^{p+2}$. Obviously, $U_n(u)$ is minimized at

$$\hat{u}^{(n)} = \sqrt{n}\left(\hat{\alpha}^{\mathcal{H}adl} - \alpha^*, \hat{\beta}^{\mathcal{H}adl} - \beta^*, \hat{s}^{\mathcal{H}adl} - s^*\right)^T.$$

The principle of the proof of [39] is to study the epi-limit of U_n. The notion of epi-convergence has been introduced to ensure variational properties. It provides the natural conditions, minimal in some sense, under which, if U_n epi-converges to U, one can guarantee the convergence of the corresponding minimizations problems. See theorem 1.10, section 2.2 and page 39 of [2] for a precise definition in the case of deterministic functions. As noticed by [6], if we restrict ourselves to lower semicontinuous extended real-valued functions from $\mathbb{R}^q$ with non-empty epigraphs, epi-convergence is induced by an explicit distance. Consequently, [6] defines epi-convergence in distribution (denoted $\rightarrow_{e-d}$) for random lower semicontinuous extended real-valued variables from $\mathbb{R}^q$ as the convergence in distribution for this distance. We refer to [15] and [23] for other equivalent definitions. It has already been used in statistics to get expedient proofs for asymptotic normality theorems. For example, it has been used in [16]. Theorem 4.1 of [17] (asymptotic normality of the regression quantile estimates) is proved using epi-convergence. In these cases, it leads to a simpler argument than the one used previously.

The function $U_n(u)$ can be decomposed as the sum of two terms, one du to the loss function given by $J_n(u) = \mathcal{L}_{\mathcal{H}}\left((\alpha^*, \beta^*, s^*)^T + u/\sqrt{n}\right) - \mathcal{L}_{\mathcal{H}}(\alpha^*, \beta^*, s^*)$ and one du to the penalty term given by $P_n(u) = \lambda_n \sum_{j=1}^p \hat{w}_j^{\mathcal{H}adl}(|\beta_j^* + u_j/\sqrt{n}| - |\beta_j^*|)$. This penalty term is the same as in [39]. Let $P(u) = 0$ if u is such that $u_j = 0$ for all $j \notin \mathcal{A}$ and $P(u) = +\infty$ otherwise. [39] claims that, under assumptions $\lambda_n/\sqrt{n} \to 0$, $\lambda_n n^{(\gamma-1)/2} \to \infty$ and $\hat{\beta}$ is a root-n-consistent estimator of β^*, P_n epi-converges to P.

Concerning the term J_n, we need the following lemma. Let $u_{1:p}$ be the vector defined by $(u_1, \ldots, u_p)^T$. We denote by $\rightarrow_{f-d}$ the finite dimensional convergence in distribution.

Lemma 2. *Under conditions* $M > 1$, **(N0)**, **(N1)**, **(N2)**, **(D1)** *and* **(D2)**, *we have* $J_n \rightarrow_{f-d} J$ *and* $J_n \rightarrow_{e-d} J$, *where* J *is defined by*

$$J(u) = A_{s^*}\left(u_{1:p}^T V u_{1:p} + u_0^2\right) + D_{s^*} u_{p+1}^2 - W^T u\,.$$

In this expression, $W \sim \mathcal{N}_{p+2}\left(0, \Sigma_1^2\right)$ *with* Σ_1^2 *the squared block diagonal matrix of size* $p+2$ *defined by*

$$\Sigma_1^2 = \mathtt{diag}\left(\mathbb{E}\left[\mathcal{H}'_M\left(\sigma\epsilon/s^*\right)^2\right], \mathbb{E}\left[\mathcal{H}'_M\left(\sigma\epsilon/s^*\right)^2\right] V, \mathbb{E}\left[Z^2\right]\right)$$

and Z *is as in Theorem 3.2*.

Using the definiton of [6], the notion of epi-convergence in distribution of convex lower semicontinuous random variables is a particular case of weak convergence of a net as stated in definition 1.33 of [34]. Consequently, we can use Slutsky's theorem page 32 and example 1.4.7 of [34] to ensure that epi-convergence of J_n and P_n implies that $(J_n, P_n) \to_{e-d} (J, P)$ since P is deterministic. However, the metric space involved by epi-convergence is not a topological space i.e. the sum of epi-limits is not necessary the epi-limit of the sum (see [19] and [27] for counter-examples). Consequently, the epi-convergence of the couple (J_n, P_n) is not enough to ensure the epi-convergence of the sum. One has to consider a stronger convergence of a coordinate. Let $\to_{u-d}$ denotes the convergence in distribution with respect to the topology of uniform convergence on compacts. Since $J_n \to_{f-d} J$ and J_n, J are finite (for n sufficiently large) convex functions, [15], page 13 claims that $J_n \to_{u-d} J$. Gathering $(J_n, P_n) \to_{e-d} (J, P)$, $J_n \to_{u-d} J$ and continuity of J, theorem 4 of [15] ensures $U_n \to_{e-d} U$, where $U(u) = A_{s^*}\left(u_{1:p}^T V u_{1:p} + u_0^2\right) + D_{s^*} u_{p+1}^2 - W^T u$ if $u_j = 0 \, \forall j \notin \mathcal{A}$ and $j = 1, \ldots, p$. $U(u) = +\infty$ otherwise. Since U_n and U are convex lower semicontinuous functions defined *on the whole set* $\mathbb{R}^{p+2}$, gathering the previous *epi-convergence* with the definition of $\hat{u}^{(n)}$ as an *argmin*, theorem 5 of [15] ensures that

$$\hat{u}^{(n)} = \sqrt{n}\left(\hat{\alpha}^{\mathcal{H}adl} - \alpha^*, \hat{\beta}_{\mathcal{A}}^{\mathcal{H}adl} - \beta_{\mathcal{A}}^*, \hat{\beta}_{\mathcal{A}^c}^{\mathcal{H}adl}, \hat{s}^{\mathcal{H}adl} - s^*\right) \to_d \left(\frac{W_0}{2A_{s^*}}, \frac{1}{2A_{s^*}} V_{1,1}^{-1} W_{\mathcal{A}}, 0, \frac{W_{p+1}}{2D_{s^*}}\right). \quad (6.1)$$

Indeed, $V_{1,1}$ is supposed positive definite in assumption **(D2)** and Theorem 3.2 assumes that the noise satisfies **(N2)**. Consequently, U get a unique argminimum and the asymptotic normality part is proved.

Step 2. Let us now show the consistency in variable selection part. A finite intersection of random sets the probability of which tends to 1 as n tends to infinity also tends to 1 as n tends to infinity. Consequently, it suffices to show that $\mathbb{P}[\mathcal{A} \subset \mathcal{A}_n] \to 1$ and $\mathbb{P}[\mathcal{A}^c \subset {\mathcal{A}_n}^c] \to 1$ as n tends to infinity.

The first claim is an easy consequence of (6.1). Indeed, (6.1) implies that $\forall j \in \mathcal{A}$, $\hat{\beta}_j^{\mathcal{H}adl} \xrightarrow{\mathbb{P}} \beta_j^*$ with $\beta_j^* \neq 0$. Thus,

$$\forall \epsilon > 0, \exists N_\epsilon, \forall n \geq N_\epsilon, \mathbb{P}\left[|\hat{\beta}_j^{\mathcal{H}adl} - \beta_j^*| \leq |\beta_j^*|/2\right] \geq 1 - \epsilon.$$

Moreover, $|\hat{\beta}_j^{\mathcal{H}adl} - \beta_j^*| \leq |\beta_j^*|/2$ implies that $|\hat{\beta}_j^{\mathcal{H}adl}| \geq |\beta_j^*|/2$. Thus $\forall \epsilon > 0$, $\exists N_\epsilon$, $\forall n \geq N_\epsilon$, $\mathbb{P}[|\hat{\beta}_j^{\mathcal{H}adl}| \geq |\beta_j^*|/2] \geq 1 - \epsilon$ and the first claim is proved.

Let j such that $\beta_j^* = 0$. To prove the second claim, we have to show that $\mathbb{P}[\hat{\beta}_j^{\mathcal{H}adl} \neq 0] \to 0$ as n tends to infinity. If $\hat{\beta}_j^{\mathcal{H}adl} \neq 0$, $\hat{s}^{\mathcal{H}adl} > 0$, the objective function $Q^{\mathcal{H}adl}$ get a partial derivative with respect to β_j at $\hat{\beta}_j^{\mathcal{H}adl}$. Moreover, this function is minimal at $(\hat{\alpha}^{\mathcal{H}adl}, \hat{\beta}^{\mathcal{H}adl}, \hat{s}^{\mathcal{H}adl})$ thus the limit of the Newton's

difference quotient defining the partial derivative is null:

$$\{\hat{\beta}_j^{\mathcal{H}adl} \neq 0\} \subset \left\{ \sum_{i=1}^{n} \mathbf{x}_{i,j} \mathcal{H}'_M \left(\frac{y_i - \hat{\alpha}^{\mathcal{H}adl} - \mathbf{x}_i^T \hat{\beta}^{\mathcal{H}adl}}{\hat{s}^{\mathcal{H}adl}} \right) \mathbb{1}_{\hat{s}^{\mathcal{H}adl}>0} \right. \\ \left. = \lambda_n \hat{w}_j^{\mathcal{H}adl} \text{sign}(\hat{\beta}_j^{\mathcal{H}adl}) \mathbb{1}_{\hat{s}^{\mathcal{H}adl}>0} \right\}. \quad (6.2)$$

Now, the probability of this event is given by

$$\mathbb{P}\left[\hat{s}^{\mathcal{H}adl} = 0\right] + \mathbb{P}\left[\hat{s}^{\mathcal{H}adl} > 0 \text{ and } \right. \\ \left. \sum_{i=1}^{n} \mathbf{x}_{i,j} \mathcal{H}'_M \left(\frac{y_i - \hat{\alpha}^{\mathcal{H}adl} - \mathbf{x}_i^T \hat{\beta}^{\mathcal{H}adl}}{\hat{s}^{\mathcal{H}adl}} \right) = \lambda_n \hat{w}_j^{\mathcal{H}adl} \text{sign}(\hat{\beta}_j^{\mathcal{H}adl}) \right]. \quad (6.3)$$

Moreover, (6.1) ensures that $\sqrt{n}(\hat{s}^{\mathcal{H}adl} - s^*) = \mathcal{O}_P(1)$, thus $\hat{s}^{\mathcal{H}adl} \xrightarrow{\mathbb{P}} s^*$. The condition $M > 1$ ensures $s^* > 0$ through Lemma 1. Noting that $\mathbb{P}\left[\hat{s}^{\mathcal{H}adl} = 0\right] \leq \mathbb{P}\left[|\hat{s}^{\mathcal{H}adl} - s^*| \geq s^*/2\right]$, we have $\mathbb{P}\left[\hat{s}^{\mathcal{H}adl} = 0\right] \to 0$, as n tends to infinity. As in [39], we have for $j \in \mathcal{A}^c$, $\lambda_n |\hat{w}_j^{\mathcal{H}adl}|/\sqrt{n}$ goes to $+\infty$ as n goes to $+\infty$. Indeed, $\sqrt{n}/(\lambda_n |\hat{w}_j^{\mathcal{H}adl}|) = |\sqrt{n}\hat{\beta}_j|^\gamma/(\lambda_n n^{(\gamma-1)/2}) \xrightarrow{\mathbb{P}} 0$ since the numerator is uniformly tight ($j \in \mathcal{A}^c$ and $\hat{\beta}$ is root-n-consistent) and Theorem 2 supposes that the denominator tends to $+\infty$. So, if we show that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{x}_{i,j} \mathcal{H}'_M \left(\frac{y_i - \hat{\alpha}^{\mathcal{H}adl} - \mathbf{x}_i^T \hat{\beta}^{\mathcal{H}adl}}{\hat{s}^{\mathcal{H}adl}} \right) = \mathcal{O}_P(1), \quad (6.4)$$

we get the result since

$$\mathbb{P}\left[\hat{\beta}_j^{\mathcal{H}adl} \neq 0\right] \leq \mathbb{P}\left[\hat{s}^{\mathcal{H}adl} > 0 \text{ and } \mathcal{O}_P(1) = \lambda_n |\hat{w}_j^{\mathcal{H}adl}|/\sqrt{n}\right] + \mathbb{P}\left[\hat{s}^{\mathcal{H}adl} = 0\right] \to 0.$$

Let us now prove (6.4). Using the definition (2.1) of the model and $\hat{u}^{(n)}$, we have

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{x}_{i,j} \mathcal{H}'_M \left(\frac{y_i - \hat{\alpha}^{\mathcal{H}adl} - \mathbf{x}_i^T \hat{\beta}^{\mathcal{H}adl}}{\hat{s}^{\mathcal{H}adl}} \right) = \\ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{x}_{i,j} \mathcal{H}'_M \left(\frac{\sigma\epsilon_i - \frac{\hat{u}_0^{(n)} + \mathbf{x}_i^T \hat{u}_{1:p}^{(n)}}{\sqrt{n}}}{\frac{\hat{u}_{p+1}^{(n)}}{\sqrt{n}} + s^*} \right). \quad (6.5)$$

Equation (6.1) ensures that $\hat{u}^{(n)} = \mathcal{O}_P(1)$ since it converges in distribution to a finite everywhere random variable (theorem 2.4 of [33]). Combined with $\hat{s}^{\mathcal{H}adl} \xrightarrow{\mathbb{P}} s^* > 0$, this leads to $\forall\, \epsilon > 0$, $\exists M_\epsilon$, $\exists\, N_\epsilon$, $\forall\, n \geq N_\epsilon$,

$$\mathbb{P}\left[\left(-\frac{\sqrt{n} s^*}{2} \right) \vee (-M_\epsilon) < \hat{u}_{p+1}^{(n)} < M_\epsilon \text{ and } \|\hat{u}_{0:p}^{(n)}\| \leq M_\epsilon \right] \geq 1 - \frac{\epsilon}{2}. \quad (6.6)$$

Let us denote

$$F_n(u) = J_n(u) + \sum_{i=1}^{n} \mathcal{H}'_M\left(\frac{\sigma\epsilon_i}{s^*}\right)\left(\frac{u_0}{\sqrt{n}} + \frac{\mathbf{x}_i^T u_{1:p}}{\sqrt{n}}\right) - \frac{u_{p+1}}{\sqrt{n}}\sum_{i=1}^{n} Z_i,$$

where for $1 \leq i \leq n$, $Z_i = 1+\mathcal{H}_M(\sigma\epsilon_i/s^*)-\sigma\epsilon_i/s^*\mathcal{H}'_M(\sigma\epsilon_i/s^*)$. From Lemma 3, we have $F_n(u) \xrightarrow{\mathbb{P}} f(u)$, where

$$f(u) = A_{s^*}\left(\|V^{1/2}u_{1:p}\|_2^2 + u_0^2\right) + D_{s^*}u_{p+1}^2.$$

Consequently, equation (6.6) involves that $\forall\,\epsilon > 0$, $\exists M_\epsilon$, $\exists\, N_\epsilon$, $\forall\, n \geq N_\epsilon$,

$$\mathbb{P}\left[\left\|\frac{\partial F_n}{\partial u_{1:p}}(\hat{u}) - \frac{\partial f}{\partial u_{1:p}}(\hat{u})\right\| \leq \sup_{\mathcal{E}}\|\nabla F_n(u) - \nabla f(u)\|\right] \geq 1 - \frac{\epsilon}{2},$$

where $\mathcal{E} = \{\|u_{0:p}\| \leq M_\epsilon,\ (-\frac{\sqrt{n}s^*}{2})\vee(-M_\epsilon) \leq u_{p+1} \leq M_\epsilon\}$. Du to the convexity of F_n and f, the pointwise convergence in probability of Lemma 3 ensures convergence in probability uniformly over any compact set of F_n and its derivative. This result is available in a deterministic framework in theorems 10.8 and 25.7 of [26]. Its generalization for convergence in probability is done in theorem II.1 (appendix II) of [1] for F_n (see also [24], section 6 for a direct proof). In [1], the convergence in probability is shown by extracting from any subsequence of F_n a subsequence converging almost surely. The subsequence is extracted using a countable dense set of points in $\mathbb{R}^{p+2}$ and a diagonal argument. Theorem 10.8 of [26] ensures the almost sure convergence of the extracted subsequence. The same argument can be used to ensure the convergence in probability uniformly over any compact set of ∇F_n using Theorem 25.7 of [26] instead of Theorem 10.8. See [5] page 73 for a detailed treatment of the involved diagonal argument. Consequently,

$$\sup_{\mathcal{E}}\|\nabla F_n(u) - \nabla f(u)\| \xrightarrow{\mathbb{P}} 0, \text{ and so } \left\|\frac{\partial F_n}{\partial u_{1:p}}(\hat{u}) - \frac{\partial f}{\partial u_{1:p}}(\hat{u})\right\| \xrightarrow{\mathbb{P}} 0.$$

This means that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{x}_i\left[\mathcal{H}'_M\left(\frac{\sigma\epsilon_i - n^{-1/2}\left(\hat{u}_0^{(n)} + \mathbf{x}_i^T\hat{u}_{1:p}^{(n)}\right)}{s^* + \frac{\hat{u}_{p+1}^{(n)}}{\sqrt{n}}}\right) - \mathcal{H}'_M\left(\frac{\sigma\epsilon_i}{s^*}\right)\right] + 2A_{s^*}V\hat{u}_{1:p}^{(n)} \xrightarrow{\mathbb{P}} 0\,. \quad (6.7)$$

We have that

$$H_n \triangleq \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{x}_i\mathcal{H}'_M\left(\frac{\sigma\epsilon_i}{s^*}\right) = \mathcal{O}_P(1).$$

Indeed, $\mathrm{Var}(H_n) = \mathbb{E}[\mathcal{H}'_M(\sigma\epsilon/s^*)^2]\mathbf{X}^T\mathbf{X}/n$ since the random variables $\epsilon_1, \ldots, \epsilon_n$ are independent and $\mathbb{E}[\mathcal{H}'_M(\sigma\epsilon/s^*)] = 0$ under condition **(N1)**. Consequently, $\mathrm{tr}(\mathrm{Var}(H_n)) = \mathbb{E}[\mathcal{H}'_M(\sigma\epsilon/s^*)^2]\mathrm{tr}(\mathbf{X}^T\mathbf{X})/n$ and condition **(D2)** ensures that $(\mathrm{tr}(\mathrm{Var}(H_n)))_{n\geq 1}$ is a convergent and thus bounded sequence. Moreover, condition **(N1)** implies that $\mathbb{E}[H_n] = 0$. Since Markov's inequality entails

$\mathbb{P}[\|H_n\| \geq M] \leq \mathrm{tr}(\mathrm{Var}(H_n))/M^2$, we get that $\sup_{n\geq 1}\mathbb{P}[\|H_n\| \geq M] \to 0$ as M tends to infinity which is the definition of $H_n = \mathcal{O}_P(1)$. Since $\hat{u}_{1:p}^{(n)} = \mathcal{O}_P(1)$ (a consequence of (6.1)) and $H_n = \mathcal{O}_P(1)$, (6.7) leads to (6.4). □

6.3. Proof of technical Lemmas

6.3.1. Proof of Lemma 1

Let us firstly show that s^* is given by the equation (2.4). Convexity of $\mathcal{H}_M(.)$ and theorem 1.1.6 page 5 of [11] imply that for all $t \in \mathbb{R}$, the function $f_t : s \to s + s\mathcal{H}_M(\sigma t/s)$ is convex on $]0,+\infty[$. Since the law of ϵ defines a positive measure, this implies that the function $F(s) = \mathbb{E}[F_\epsilon(s)]$ is convex. Moreover, for all $t \in \mathbb{R}$ the function f_t is of class $\mathcal{C}^1$ on the open interval $]0,+\infty[$ and $f_t'(s) = 1 + \mathcal{H}_M(\sigma t/s) - (\sigma t/s)\mathcal{H}_M'(\sigma t/s)$. However, the function $g : x \in \mathbb{R} \to 1 + \mathcal{H}_M(x) - x\mathcal{H}_M'(x)$ satisfies $\sup_{x\in\mathbb{R}}|g(x)| \leq 1$. Consequently, the following uniform bound holds: $\forall s > 0,\ |f_t'(s)| \leq 1$. This allows to use the dominated convergence theorem impling that F is of class $\mathcal{C}^1$ on the open interval $]0,+\infty[$ with $F'(s) = \mathbb{E}[F_\epsilon'(s)]$. Consequently, s^* satisfies (2.3) if and only if it satisfies (2.4).

In order to finish the proof of lemma 1, it suffices to show that the function F' has a unique zero on $]0,+\infty[$. The function g is continuous and such that $g(x) \to 1 - M^2$ as $x \to +\infty$ and $g(0) = 1$. Thus, justifying the use of the dominated convergence theorem by the same uniform bound as previously, we get $F'(s) \to 1 - M^2$ as s tends to 0 and $F'(s) \to 1$ as s tends to infinity. The continuity of F' and the intermediate value theorem implies that F' has a zero on $]0,+\infty[$ if $M > 1$. To see that this zero is unique, we show that the function F' is strictly increasing. Let $s_1 > s_2 > 0$. The following decomposition holds since the function g is constant on $]-\infty,-M] \cup [M,+\infty[$:

$$\begin{aligned} F'(s_1) - F'(s_2) &= \mathbb{E}\left[\left(g\left(\frac{\sigma\epsilon}{s_1}\right) - g\left(\frac{\sigma\epsilon}{s_2}\right)\right)\right] \\ &= \sigma^2\left(\frac{1}{s_2^2} - \frac{1}{s_1^2}\right)\mathbb{E}\left[\epsilon^2 \mathbb{1}_{|\epsilon|\leq \frac{M}{\sigma}s_2}\right] + \mathbb{E}\left[\left(M^2 - \frac{\sigma^2\epsilon^2}{s_1^2}\right)\mathbb{1}_{\frac{M}{\sigma}s_2<|\epsilon|\leq\frac{M}{\sigma}s_1}\right]. \end{aligned}$$

Under condition **(N2)**, the last quantity is strictly positive. This equality also implies that condition **(N2)** is necessary to ensure that the function F' is strictly increasing. Actually, suppose that **(N2)** does not hold i.e. there exists $a > 0$ such that $\mathbb{P}[|\epsilon| \leq a] = 0$. If we consider $s_1 = a\sigma/M$ and $s_2 = a\sigma/(2M)$, we have $s_1 > s_2$. Moreover, the last equality implies that $F'(s_1) = F'(s_2)$ since the integral of a function over a null set is equal to 0. Consequently, if **(N2)** does not hold, the function F' is not strictly increasing. This concludes the proof. □

6.3.2. Proof of Lemma 2

First we need to the following lemma.

Lemma 3. *If* **(D1)**, **(D2)**, **(N0)**, **(N1)**, **(N2)** *hold and* $M > 1$, *for every* u *fixed in* $\mathbb{R}^{p+2}$,

$$J_n(u) + \sum_{i=1}^{n} \mathcal{H}'_M\left(\frac{\sigma\epsilon_i}{s^*}\right)\left(\frac{u_0}{\sqrt{n}} + \frac{\mathbf{x}_i^T u_{1:p}}{\sqrt{n}}\right) - \frac{u_{p+1}}{\sqrt{n}}\sum_{i=1}^{n} Z_i \xrightarrow{\mathbb{P}} A_{s^*}\left(\|V^{1/2}u_{1:p}\|_2^2 + u_0^2\right) + D_{s^*}^2 u_{p+1}, \quad (6.8)$$

where D_{s^*}, A_{s^*} *are defined as in Theorem 3.2.*

Gathering this lemma and Lemma 4 (with $\mathbf{w}_i^{(n)} = \binom{1}{\mathbf{x}_i}$, $\delta_i = \mathcal{H}'_M(\sigma\epsilon_i/s^*)$ and $W_i = -Z_i$) and the Cramer-Wold device, we get $J_n \to_{f-d} J$. Let us show that the conditions of Lemma 4 are satisfied. Condition **(T3)** holds with $\mathbf{w}_\infty = \binom{1}{0}$ since the covariables are supposed centered in the definition (2.1) of the model. Moreover, **(D2)** (resp. **(D1)**) leads to condition **(T1)** (resp. **(T2)**). Assumption **(R)** is satisfied. Indeed, the $(\epsilon_i)_{1\le i\le n}$ are supposed to be independent and identically-distributed random variables. $\mathbb{E}[\delta] = 0$ since $\mathcal{H}'_M(.)$ is odd and **(N1)** holds. $\mathbb{E}[W] = 0$ since the property (2.4) of s^* is available ($M > 1$ and **(N2)** are supposed). Finally, the functions $\mathcal{H}'_M(.)$ and $g : x \in \mathbb{R} \to 1 + \mathcal{H}_M(x) - x\mathcal{H}'_M(x)$ are bounded ensuring the finite variance hypothesis of **(R)**.

In order to show that the asymptotic variance Σ^2 involved in Lemma 4 is a diagonal matrix, one has to note that $\mathrm{Cov}(\delta, W) = \mathbb{E}[\delta W] = 0$. This holds since the function $x \to \mathcal{H}'_M(x)\, g(x)$ is odd ($\mathcal{H}'_M(.)$ is odd and $g(.)$ is even) and **(N1)** is supposed.

To conclude the proof of Lemma 2, since J is convex, continuous and finite everywhere random variable, part (a) of Theorem 5 of [15] implies that $J_n \to_{e-d} J$ since the finite-dimensional convergence holds. □

6.3.3. Proof of Lemma 3

The condition $M > 1$ and **(N2)** ensure $s^* > 0$ through Lemma 1. Thus, for n sufficiently large (with respect to a non-random bound depending on u_{p+1}), $u_{p+1}/\sqrt{n} + s^* > 0$. Consequently, the definitions (2.1) of the model and (2.2) of the function $\mathcal{L}_\mathcal{H}$ imply that, for n sufficiently large (with respect to a non-random bound depending on u_{p+1}),

$$\mathcal{L}_\mathcal{H}\left((\alpha^*, \beta^*, s^*)^T + \frac{u}{\sqrt{n}}\right) = \sum_{i=1}^{n}\left(s^* + \frac{u_{p+1}}{\sqrt{n}}\right)\left(1 + \mathcal{H}_M\left(\frac{\sigma\epsilon_i - \frac{u_0}{\sqrt{n}} - \frac{\mathbf{x}_i^T u_{1:p}}{\sqrt{n}}}{s^* + \frac{u_{p+1}}{\sqrt{n}}}\right)\right) \triangleq \sum_{i=1}^{n} d_{n,i}(u).$$

Simple calculations lead to the following explicit expression of the derivative of $d_{n,i}$ at 0. For $1 \le i \le n$,

$$\frac{\partial d_{n,i}}{\partial u}(0) = \frac{1}{\sqrt{n}}\left(-\mathcal{H}'_M\left(\frac{\sigma\epsilon_i}{s^*}\right), -\mathcal{H}'_M\left(\frac{\sigma\epsilon_i}{s^*}\right)\mathbf{x}_i^T, Z_i\right)^T.$$

We have to show the convergence in probability of

$$D_n \triangleq \sum_{i=1}^{n} \left(d_{n,i}(u) - d_{n,i}(0) - u^T \frac{\partial d_{n,i}}{\partial u}(0) \right)$$

since this quantity is equal (for n sufficiently large) to the left-hand side quantity in (6.8). The structure of the rest of the proof is inspired by [3].

Step 1. To begin with, we show that $\mathrm{Var}(D_n) \to 0$ as $n \to +\infty$. Using that the errors $(\epsilon_i)_{1 \leq i \leq n}$ are independent random variables, we get

$$\mathrm{Var}(D_n) \leq \sum_{i=1}^{n} \mathbb{E}\left[\left(d_{n,i}(u) - d_{n,i}(0) - u^T \frac{\partial d_{n,i}}{\partial u}(0) \right)^2 \right].$$

Moreover, convexity of $\mathcal{H}_M(.)$ combined with proposition 2.2.1 page 160 of [11] implies that for all $1 \leq i \leq n$, $d_{n,i}$ is a convex function. As such, it lies above all of its tangents:

$$0 \leq d_{n,i}(u) - d_{n,i}(0) - u^T \frac{\partial d_{n,i}}{\partial u}(0) \leq u^T \left(\frac{\partial d_{n,i}}{\partial u}(u) - \frac{\partial d_{n,i}}{\partial u}(0) \right).$$

Furthermore, easy calculations imply that the the previous upper bound is equal to

$$\frac{u_{p+1}}{\sqrt{n}} \left(g\left(\frac{\sigma\epsilon_i - \frac{u_0 + \mathbf{x}_i^T u_{1:p}}{\sqrt{n}}}{s^* + \frac{u_{p+1}}{\sqrt{n}}} \right) - g\left(\frac{\sigma\epsilon_i}{s^*} \right) \right) - \left(\mathcal{H}'_M\left(\frac{\sigma\epsilon_i - \frac{u_0 + \mathbf{x}_i^T u_{1:p}}{\sqrt{n}}}{s^* + \frac{u_{p+1}}{\sqrt{n}}} \right) - \mathcal{H}'_M\left(\frac{\sigma\epsilon_i}{s^*} \right) \right) \left(\frac{u_0 + \mathbf{x}_i^T u_{1:p}}{\sqrt{n}} \right),$$

where $g(x) = 1 + \mathcal{H}_M(x) - x\mathcal{H}'_M(x)$. Using that $(a+b+c)^2 \leq 4(a^2+b^2) + 2c^2$, $\mathrm{Var}(D_n)$ is bounded above by the sum of I_1, I_2 and I_3 where

$$I_1 \triangleq 4\frac{u_{p+1}^2}{n} \sum_{i=1}^{n} \mathbb{E}\left[\left(g\left(\frac{\sigma\epsilon_i - \frac{u_0 + \mathbf{x}_i^T u_{1:p}}{\sqrt{n}}}{s^* + \frac{u_{p+1}}{\sqrt{n}}} \right) - g\left(\frac{\sigma\epsilon_i}{s^*} \right) \right)^2 \right],$$

$$I_2 \triangleq 4\frac{u_0^2}{n} \sum_{i=1}^{n} \mathbb{E}\left[\left(\mathcal{H}'_M\left(\frac{\sigma\epsilon_i - \frac{u_0 + \mathbf{x}_i^T u_{1:p}}{\sqrt{n}}}{s^* + \frac{u_{p+1}}{\sqrt{n}}} \right) - \mathcal{H}'_M\left(\frac{\sigma\epsilon_i}{s^*} \right) \right)^2 \right]$$

and the term I_3 is defined by

$$I_3 \triangleq 2\sum_{i=1}^{n} \mathbb{E}\left[\left(\frac{\mathbf{x}_i^T u_{1:p}}{\sqrt{n}} \right)^2 \left(\mathcal{H}'_M\left(\frac{\sigma\epsilon_i - \frac{u_0 + \mathbf{x}_i^T u_{1:p}}{\sqrt{n}}}{s^* + \frac{u_{p+1}}{\sqrt{n}}} \right) - \mathcal{H}'_M\left(\frac{\sigma\epsilon_i}{s^*} \right) \right)^2 \right].$$

In order to show that each term I_i $(1 \leq i \leq 3)$ tends to 0 as n tends to infinity, we notice that they all have the same following general form. Let φ denote a lipschitz continuous function from $\mathbb{R}$ to $\mathbb{R}$ with lipschitz constant L such that, for some constants K_1 and K_2, $\varphi(x) = K_1$ if $x \geq M$ and $\varphi(x) = K_2$ if $x \leq -M$. Let

$$\Gamma(h, h') = \mathbb{E}\left[\left(\varphi\left(\frac{\sigma\epsilon + h}{s^* + h'}\right) - \varphi\left(\frac{\sigma\epsilon}{s^*}\right)\right)^2\right]$$

be a function of two variables associated to φ. We note that I_1 (resp. I_2) are proportionnal to $\Gamma\left(-(u_0 + \mathbf{x}_i^T u_{1:p})/\sqrt{n}, u_{p+1}/\sqrt{n}\right)$ with $\varphi = g, L = 2M$ and $K_1 = K_2 = 1 - M^2$ (resp. $\varphi = \mathcal{H}'_M(\cdot), L = 2, K_1 = 2M$ and $K_2 = -2M$).

Let us now show a general property satisfied by functions Γ. Spliting the expectation over the sets $\{|\sigma\epsilon| \leq Ms^*\}$ and $\{|\sigma\epsilon| > Ms^*\}$, we easily get

$$\begin{aligned}\Gamma(h, h') \leq \frac{2L^2}{s^{*2}(s^* + h')^2}\left(h'^2\sigma^2\mathbb{E}\left[\epsilon^2 \mathbb{1}_{|\sigma\epsilon| \leq Ms^*}\right] + h^2 s^*\right) \\ + 2\left(\|\varphi\|_\infty^2 + K_1^2\right)\mathbb{P}\left[\sigma\epsilon > Ms^*, \frac{\sigma\epsilon + h}{s^* + h'} \leq M\right] \\ + 2\left(\|\varphi\|_\infty^2 + K_2^2\right)\mathbb{P}\left[\sigma\epsilon < -Ms^*, \frac{\sigma\epsilon + h}{s^* + h'} > -M\right].\end{aligned}$$

Moreover,

$$\mathbb{P}\left[\sigma\epsilon > Ms^* \text{ and } \frac{\sigma\epsilon + h}{s^* + h'} \leq M\right] = \mathbb{P}\left[\sigma\epsilon \leq M(s^* + h') - h\right] - \mathbb{P}\left[\sigma\epsilon \leq Ms^*\right]$$

and

$$\begin{aligned}\mathbb{P}\left[\sigma\epsilon < -Ms^* \text{ and } \frac{\sigma\epsilon + h}{s^* + h'} > -M\right] = \\ \mathbb{P}\left[\sigma\epsilon < -Ms^*\right] - \mathbb{P}\left[\sigma\epsilon \leq -M(s^* + h') - h\right].\end{aligned}$$

Assumption **(N0)** ensures that the distribution function of $\sigma\epsilon$ is continuous at $\pm Ms^*$. Consequently, Γ is continuous at 0 with $\Gamma(0) = 0$: as (h, h') tends to 0, $\Gamma(h, h') \to 0$. Moreover, assumption **(D1)** ensures that

$$\max_{1 \leq i \leq n}\left\|\left(-\frac{u_0 + \mathbf{x}_i^T u_{1:p}}{\sqrt{n}}, \frac{u_{p+1}}{\sqrt{n}}\right)\right\|_2 \to 0$$

as n tends to infinity. Consequently, the following assertion holds:

$$\forall\, \epsilon > 0,\ \exists N_\epsilon,\ \forall n \geq N_\epsilon,\ \forall\, 1 \leq i \leq n,\ \Gamma\left((u_0 + \mathbf{x}_i^T u_{1:p})/\sqrt{n}, u_{p+1}/\sqrt{n}\right) \leq \epsilon\,.$$

It directly entails that I_1 and I_2 tend to 0 as n goes to infinity. Moreover, combined with assumption **(D2)**, it also implies $I_3 \to 0$. Consequently, $\mathrm{Var}(D_n) \to 0$.

Step 2. Now, we show that $\mathbb{E}[D_n] \to A_{s^*}(u_0^2 + u_{1:p}^T V u_{1:p}) + D_{s^*} u_{p+1}^2$ as $n \to +\infty$. Since the function $\mathcal{H}'_M(.)$ is odd, the explicit expression of the derivative of $d_{n,i}$ (given at the begining of the proof) combined with assumption **(N1)** and property (2.4) of s^* imply that $\mathbb{E}\left[\frac{\partial d_{n,i}}{\partial u}(0)\right] = 0$ (note that $Z_i = g(\sigma\epsilon_i/s^*)$). Consequently, the definition of D_n leads to $\mathbb{E}[D_n] = \sum_{i=1}^n \mathbb{E}[d_{n,i}(u) - d_{n,i}(0)]$. Moreover, the random variables $(\sigma\epsilon_i)_{1\le i\le n}$ are identically-distributed thus

$$\mathbb{E}[D_n] = \sum_{i=1}^n \mathbb{E}\left[\left(s^* + \frac{u_{p+1}}{\sqrt{n}}\right) + \left(s^* + \frac{u_{p+1}}{\sqrt{n}}\right)\mathcal{H}_M\left(\frac{\sigma\epsilon - \frac{u_0 + \mathbf{x}_i^T u_{1:p}}{\sqrt{n}}}{s^* + \frac{s'}{\sqrt{n}}}\right) - \left(s^* + s^*\mathcal{H}_M\left(\frac{\sigma\epsilon}{s^*}\right)\right)\right].$$

Let us introduce the following function:

$$\Psi(h,h') = (s^* + h') + (s^* + h')\mathbb{E}\left[\mathcal{H}_M\left((\sigma\epsilon + h)/(s^* + h')\right)\right].$$

It is defined on the open set $\mathbb{R}\times]-s^*;+\infty[$. In order to get the asymptotic limit of $\mathbb{E}[D_n]$, a Taylor expansion of degree 2 of Ψ around 0 is required. This means that we look for constants $A_{s^*}, B_{s^*}, C_{s^*}, D_{s^*}, E_{s^*}$ such that $\Psi(h,h') = \Psi(0) + B_{s^*}h + C_{s^*}h' + A_{s^*}h^2 + D_{s^*}h'^2 + E_{s^*}hh' + \xi(h,h')$ with $\xi(h,h')/\|(h,h')\|^2 \to 0$ as $(h,h') \to 0$. To get such a Taylor expansion, one has to be neat since $\mathcal{H}_M(\cdot)$ is not two times differentiable at $\pm M$. In the case of the Huber loss without concomitant scale, [3] already solves this technical difficulty by an argument of approximation of convex functions. In the case of Huber loss with concomitant scale, roughly speaking, one has to deal with a supplementary variable (s). However, the function $\mathcal{L}_{\mathcal{H}}(\alpha,\beta,s)$ is still convex w.r.t (α,β,s). Thus, following readily the proof of Lemma 1 of [3], we get the desired expansion of Ψ in return for a bit more work to deal with the supplementary terms. We notice that the obtained proofs rely on the convexity of the involved functions. Now, we claim that this expansion has nothing to do with convexity. Using successively the dominated convergence theorem, we show that the function Ψ is two times differentiable at 0. The rigorous reasoning relies on the fact that the derivatives with respect to h and h' of the integrated function are related to functions g and $\mathcal{H}'_M(\cdot)$. Precisely, using that these functions are lipschitz continuous and bounded, we can derive the uniform bounds justifying the uses of the dominated convergence theorem. Consequently, the classical Taylor-Young theorem is available to get the desired Taylor expansion. Moreover, the coefficients $A_{s^*}, B_{s^*}, C_{s^*}, D_{s^*}, E_{s^*}$ are now explicit. Note that the proof relying on dominated convergence theorem only supposes that **(N0)** holds to provide satisfying coefficients. Under the supplementary hypothesis **(N1)**, it is clear that $E_{s^*} = 0$ and $B_{s^*} = 0$ since $\mathcal{H}'_M(\cdot)$ is odd. Since we also suppose $M > 1$ and that **(N2)** holds, s^* satisfies (2.4) which is equivalent to $C_{s^*} = 0$. Finally, under assumptions **(N0)**, **(N1)** and **(N2)**, we get

$$\mathbb{E}\left[D_n\right] = A_{s^*} \sum_{i=1}^{n} \left(\frac{u_0 + \mathbf{x}_i^T u_{1:p}}{\sqrt{n}} \right)^2 + D_{s^*} u_{p+1}^2 + \sum_{i=1}^{n} \xi \left(-\frac{u_0 + \mathbf{x}_i^T u_{1:p}}{\sqrt{n}}, \frac{u_{p+1}}{\sqrt{n}} \right), \tag{6.9}$$

where $\xi(h, h')/\|(h, h')\|^2 \to 0$ as $(h, h') \to 0$. Gathering this property of ξ with assumption **(D1)**, we get $\forall\, \eta > 0, \exists N_\eta, \forall n \geq N_\eta$,

$$\sum_{i=1}^{n} \left| \xi \left(-\frac{u_0 + \mathbf{x}_i^T u_{1:p}}{\sqrt{n}}, \frac{u_{p+1}}{\sqrt{n}} \right) \right| \leq \eta \left(u_{p+1}^2 + \sum_{i=1}^{n} \left(\frac{u_0 + \mathbf{x}_i^T u_{1:p}}{\sqrt{n}} \right)^2 \right). \tag{6.10}$$

Assumption **(D2)** leads to $\sum_{i=1}^{n} \left((u_0 + \mathbf{x}_i^T u_{1:p})/\sqrt{n} \right)^2 \to u_0^2 + u_{1:p}^T V u_{1:p}$ since $\sum_{i=1}^{n} \mathbf{x}_i = 0$ in the definition (2.1) of the model. Noting that

$$\left(\sum_{i=1}^{n} \left((u_0 + \mathbf{x}_i^T u_{1:p})/\sqrt{n} \right)^2 \right)_{n \in \mathbb{N}}$$

is a bounded sequence (since it converges), (6.10) ensures that the rest

$$\sum_{i=1}^{n} \xi \left(-(u_0 + \mathbf{x}_i^T u_{1:p})/\sqrt{n}, u_{p+1}/\sqrt{n} \right)$$

tends to 0 as n tends to infinity. Consequently, (6.9) leads to the desired convergence of the expectation.

Step 3. Bienaymé-Chebyshev inequality and step 1 lead to $D_n - \mathbb{E}\left[D_n\right] \xrightarrow{\mathbb{P}} 0$. Gathering that with step 2 and using Slutsky's theorem, the proof of lemma 3 is done. □

6.3.4. Lemma 4 and its proof

The following lemma extends the classical Central Limit Theorem. It is an extension of corollary 1 page 637 of [13].

Lemma 4. *Let us consider a sequence of deterministic vectors of* $\mathbb{R}^p$ *with double subscript* $(\mathbf{w}_i^{(n)})_{(n \geq 1, 1 \leq i \leq n)}$ *such that, as* n *tends to infinity,*

(T1) $\frac{1}{n} \sum_{i=1}^{n} \mathbf{w}_i^{(n)} \mathbf{w}_i^{(n)^T} \to V$.
(T2) $\frac{1}{\sqrt{n}} \max_{1 \leq i \leq n} \|\mathbf{w}_i^{(n)}\| \to 0$.
(T3) $\frac{1}{n} \sum_{i=1}^{n} \mathbf{w}_i^{(n)} \to \mathbf{w}_\infty$.

Moreover,

(R) $(\delta_i)_{(i \geq 1)}$ *(resp.* $(W_i)_{(i \geq 1)}$*) is a real-valued (resp.* $\mathbb{R}^k$*-valued) random sequence such that* $(\delta_i, W_i)_{(i \geq 1)}$ *are independent and identically-distributed random variables of* $\mathbb{R}^{k+1}$ *with mean* 0 *(i.e.* $\mathbb{E}\left[(\delta, W)\right] = 0$*) and finite variance.*

Then

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\delta_i \mathbf{w}_i^{(n)}, W_i\right) \to_d \mathcal{N}_{p+k}(0,\Sigma^2)\,,$$

where Σ^2 *is the following* $(p+k)\times(p+k)$ *matrix:*

$$\Sigma^2 = \begin{bmatrix} \Sigma^2_{1,1} & \Sigma^2_{1,2} \\ \Sigma^{2T}_{1,2} & \Sigma^2_{2,2} \end{bmatrix},$$

with

$$\Sigma^2_{1,1} = \mathbb{E}\left[\delta^2\right] V, \quad \Sigma^2_{2,2} = Var(W), \quad \Sigma^2_{1,2} = \mathbf{w}_\infty Cov(\delta, W).$$

Corollary 1 page 637 of [13] studied the case where $W \equiv 0$. With our notations, it states that, if assumptions **(T1)** and **(R)** are true, then $1/\sqrt{n}\sum_{i=1}^{n}\delta_i \mathbf{w}_i^{(n)} \to_d \mathcal{N}_p(0,\Sigma^2_{1,1})$. Consequently, our assumption **(T2)** seems unnecessary. However, as the following example exhibits, this assumption **(T2)** is necessary. In the case $p=1$, let us define the following double subscript sequence: $\mathbf{w}_1^{(n)} = \sqrt{n}$ and, for $i \geq 2$, $\mathbf{w}_i^{(n)} = 0$. It satisfies **(T1)** but not **(T2)** and, for $n \geq 1$, $1/\sqrt{n}\sum_{i=1}^{n}\delta_i \mathbf{w}_i^{(n)} = \delta_1$ which is not gaussian. Consequently, the conclusion of corollary 1 page 637 of [13] is wrong in this case.

Proof. Following the proof of the classical Central Limit Theorem, we consider the characteristic function of the random variable

$$Z_n \triangleq 1/\sqrt{n}\sum_{i=1}^{n}\left(\delta_i \mathbf{w}_i^{(n)}, W_i\right) : \varphi_n(t) \triangleq \mathbb{E}\left[\exp i\left\langle t, Z_n\right\rangle\right]$$

with t an element of $\mathbb{R}^{p+k}$. We write $t = (\underline{t}_p, \underline{t}_k)^T$ where $\underline{t}_p \in \mathbb{R}^p$ and $\underline{t}_k \in \mathbb{R}^k$. Condition **(R)** implies that

$$\varphi_n(t) = \prod_{j=1}^{n}\phi\left(\left\langle \frac{\underline{t}_p}{\sqrt{n}}, \mathbf{w}_j^{(n)}\right\rangle, \frac{\underline{t}_k}{\sqrt{n}}\right),$$

where ϕ is the characteristic function of the random variable (δ, W). Condition **(R)** implies that $\mathbb{E}\left[(\delta, W)\right] = 0$ and $\mathbb{E}\|(\delta, W)\|^2 < +\infty$. Let $G = \mathrm{Var}\,(\delta, W)$ be the $(k+1)\times(k+1)$ variance matrix of the random variable (δ, W). The dominated convergence and Taylor-Young (at 0) theorems entail that $\phi(u, \underline{x}_k) = 1 - \frac{1}{2}(u, \underline{x}_k)^T G(u, \underline{x}_k)^T + \xi(u, \underline{x}_k)$ for all $u \in \mathbb{R}$ and $\underline{x}_k \in \mathbb{R}^k$, where ξ is a complex-valued function satisfying $\xi(u, \underline{x}_k)/\|(u, \underline{x}_k)\|^2 \to 0$ as the vector $(u, \underline{x}_k)$ tends to 0 in $\mathbb{R}^{k+1}$. The following equality holds:

$$\phi\left(\left\langle \frac{\underline{t}_p}{\sqrt{n}}, \mathbf{w}_j^{(n)}\right\rangle, \frac{\underline{t}_k}{\sqrt{n}}\right) = 1 + z_n^j$$

with

$$z_n^j = -\frac{1}{2}\mathbb{E}\left[\delta^2\right]\left\langle \frac{\underline{t}_p}{\sqrt{n}}, \mathbf{w}_j^{(n)}\right\rangle^2 - \frac{1}{2n}\underline{t}_k^T \mathrm{Var}(W)\underline{t}_k$$
$$- \left\langle \frac{\underline{t}_p}{\sqrt{n}}, \mathbf{w}_j^{(n)}\right\rangle \mathrm{Cov}(\delta, W)\frac{\underline{t}_k}{\sqrt{n}} + \xi\left(\left\langle \frac{\underline{t}_p}{\sqrt{n}}, \mathbf{w}_j^{(n)}\right\rangle, \frac{\underline{t}_k}{\sqrt{n}}\right).$$

Condition **(T2)** and the fact that the function ξ does not depend on j and satisfies $\xi(u, \underline{x}_k) \to 0$ as $(u, \underline{x}_k) \to 0$, ensure that $\max_{1\leq j\leq n} |z_n^j| \to 0$ as n tends to infinity. Consequently, for n sufficiently large, $\max_{1\leq j\leq n} |z_n^j| < 1$. However, the Newton-Mercator series allows to define a branch of $\log(1+z)$ defined on the open unit disk. Thus, for n sufficiently large, $\varphi_n(t) = \exp\left(\sum_{j=1}^n \log(1 + z_n^j)\right)$. Using again the Newton-Mercator series, we get $\log(1+z) = z-z^2/2+h(z)$, with $h(z)/|z|^2 \to 0$ as $|z| \to 0$. Consequently, we have $\log(1 + z_n^j) = z_n^j - (z_n^j)^2/2 + h(z_n^j)$. Moreover, Condition **(T2)** and the fact that $\xi(u, \underline{x}_k)/\|(u, \underline{x}_k)\|^2 \to 0$ as $(u, \underline{x}_k) \to 0$ imply

$$\max_{1\leq j\leq n} \frac{|\xi\left(\left\langle \frac{\underline{t}_p}{\sqrt{n}}, \mathbf{w}_j^{(n)} \right\rangle, \frac{\underline{t}_k}{\sqrt{n}}\right)|}{\left\langle \frac{\underline{t}_p}{\sqrt{n}}, \mathbf{w}_j^{(n)} \right\rangle^2 + \frac{\|\underline{t}_k\|^2}{n}} \to 0$$

and

$$\max_{1\leq j\leq n} \frac{|(z_n^j)^2|}{\left\langle \frac{\underline{t}_p}{\sqrt{n}}, \mathbf{w}_j^{(n)} \right\rangle^2 + \frac{\|\underline{t}_k\|^2}{n}} \to 0$$

as $n \to +\infty$. The last limit, combined with the fact that $\frac{h(z)}{|z|^2} \to 0$ as $|z| \to 0$, entails

$$\max_{1\leq j\leq n} \frac{|h(z_n^j)|}{\left\langle \frac{\underline{t}_p}{\sqrt{n}}, \mathbf{w}_j^{(n)} \right\rangle^2 + \frac{\|\underline{t}_k\|^2}{n}} \to 0$$

as $n \to +\infty$. Consequently, there exists a sequence $(u_n)_{n\geq 1}$ which is independent of j and converges to 0 as n tends to infinity such that

$$\log(1 + z_n^j) = -\frac{1}{2}\mathbb{E}\left[\delta^2\right]\left\langle \frac{\underline{t}_p}{\sqrt{n}}, \mathbf{w}_j^{(n)} \right\rangle^2 - \frac{1}{2n}\underline{t}_k^T \mathrm{Var}(W)\underline{t}_k - \left\langle \frac{\underline{t}_p}{\sqrt{n}}, \mathbf{w}_j^{(n)} \right\rangle \mathrm{Cov}(\delta, W)\frac{\underline{t}_k}{\sqrt{n}} + a_n^j.$$

with

$$|a_n^j| \leq u_n \left(\left\langle \frac{\underline{t}_p}{\sqrt{n}}, \mathbf{w}_j^{(n)} \right\rangle^2 + \frac{\|\underline{t}_k\|^2}{n}\right).$$

Noting $A_n = 1/n \sum_{j=1}^n \mathbf{w}_j^{(n)} {\mathbf{w}_j^{(n)}}^T$, this leads to

$$\varphi_n(t) = \exp\left(-\frac{1}{2}\mathbb{E}\left[\delta^2\right]\underline{t}_p^T A_n \underline{t}_p - \frac{1}{2}\underline{t}_k^T \mathrm{Var}(W)\underline{t}_k - \left\langle \underline{t}_p, \frac{1}{n}\sum_{j=1}^n \mathbf{w}_j^{(n)} \right\rangle \mathrm{Cov}(\delta, W)\underline{t}_k + \sum_{j=1}^n a_n^j\right). \quad (6.11)$$

Gathering the previous bound on $|a_n^j|$, the facts that u_n converges to 0 and that the sequence

$$\left(\sum_{j=1}^n \left(\left\langle \underline{t}_p/\sqrt{n}, \mathbf{w}_j^{(n)} \right\rangle^2 + \|\underline{t}_k\|^2/n\right)\right)_{n\geq 1}$$

is bounded (since **(T1)** ensures its convergence), we easily get that $\sum_{j=1}^{n} a_n^j \to 0$. Moreover, using conditions **(T1)** and **(T3)** in (6.11), we obtain the asymptotic behavior of the characteristic function: $\varphi_n(t) \to \exp\left(-\frac{1}{2}t^T\Sigma^2 t\right)$, which is the characteristic function of a gaussian vector $\mathcal{N}_{p+k}(0,\Sigma^2)$. Now, the Lévy continuity theorem leads to Lemma 4. □

6.4. Computations: Software used for numerical optimization

When the regularization parameter is fixed, to solve all the involved optimization problems we used `CVX`, a package for specifying and solving convex programs [7, 8]. `CVX` is a set of Matlab functions using the methodology of disciplined convex programming. Disciplined convex programming imposes a limited set of conventions or rules, which are called the DCP ruleset. Problems which adhere to the ruleset can be rapidly and automatically verified as convex and converted to solvable form. Problems that violate the ruleset are rejected, even when convexity of the problem is obvious to the user. The version of `CVX` we use, is a preprocessor for the convex optimization solver SeDuMi (Self-Dual-Minimization [31]).

Let us now recall a well-known fact of convex analysis: the Huber function is the Moreau-Yosida regularization of the absolute value function ([11, 26, 29]). Precisely, it can be easily shown that the Huber function satisfies

$$\mathcal{H}_M(z) = \min_{v\in\mathbb{R}} \left((z-v)^2 + 2M|v|\right) .$$

This allows to write our optimization problem in a conforming manner to use `CVX`.

Note that [22] uses an expression of $\mathcal{H}_M(z)$ as the solution of a quadratic optimization problem (borrowed from the user guide of `CVX`) to write his problem in a conforming manner to use `CVX`. However, the expression of [22] involves more constraints and more variables than the previous formulation. We give here the way to use `CVX` in order to compute the estimators `alpha`$=\hat{\alpha}^{\mathcal{H}adl}$, `beta`$=\hat{\beta}^{\mathcal{H}adl}$ and `s`$=\hat{s}^{\mathcal{H}adl}$. The variable `X` represents the design matrix $\mathbf{X}$. The unpenalized estimator `betaUNP`$=\hat{\beta}_{\mathcal{H}}$ is calculated beforehand (using also `CVX`) and the regularisation parameter λ_n is fixed and denoted by `lambda`.

```
cvx_begin
variables alpha beta(p) s v(n);
minimize (n*s+quad_over_lin(y-alpha-X*beta-v,s)+2*M*norm(v,1)
+ lambda*norm(beta./betaUNP,1))
subject to
s > 0;
cvx_end
```

Let us remark that `betaUNP` is computed in the same way but deleting the term multiplied by `lambda`.

Acknowledgements

Part of this work was supported by the Interuniversity Attraction Pole (IAP) research network in Statistics P5/24. We are grateful to Anestis Antoniadis for constructive and fruitful discussions. We also thank Peter Hall and Irene Gannaz for giving us fruitful references. We are grateful to Hansheng Wang for providing the Chinese stock market data. We thank referees and Associate Editor whose careful and thoughtful comments helped to improve the paper.

References

[1] P. Andersen and R. Gill. Cox's regression model for counting processes: A large sample study. *Ann. Stat.*, 10:1100–1120, 1982. MR0673646

[2] H. Attouch. *Variational Convergence of Functions and Operators.* Pitman, Boston, 1984. MR0773850

[3] Z. Bai, C. Rao, and Y. Wu. M-estimation of multivariate linear regression parameters under a convex discrepancy function. *Statistica Sinica*, 2(1):237–254, 1992. MR1152307

[4] J. Fan and R. Li. Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American Statistical Association*, 96:1438–1360, 2001. MR1946581

[5] I. Gannaz. Estimation par ondelettes dans les modeles partiellement linéaires. *Thesis of the University Joseph Fourier (Grenoble I)*, 2007.

[6] C. J. Geyer. On the asymptotics of constrained M-estimation. *Ann. Stat.*, 22(4):1993–2010, 1994. MR1329179

[7] M. Grant and S. Boyd. Cvx: Matlab software for disciplined convex programming (web page and software). `http://stanford.edu/~boyd/cvx`, june 2009.

[8] M. Grant and S. Boyd. Graph implementations for nonsmooth convex programs, recent advances in learning and control (a tribute to m. vidyasagar), v. blondel, s. boyd, and h. kimura, editors, pages 95-110, lecture notes in control and information sciences, springer, 2008. MR2409077

[9] L. Gyorfi, M. Kohler, A. Krzyżak, and H. Walk. *A distribution-free theory of nonparametric regression.* Springer Series in Statistics. New York, NY: Springer. xvi, 647 p. , 2002. MR1920390

[10] R. W. Hill and P. W. Holland. Two robust alternatives to least-squares regression. *J. Am. Stat. Assoc.*, 72:828–833, 1977.

[11] J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex analysis and minimization algorithms I*. Grundlehren der Mathematischen Wissenschaften. 306. Berlin: Springer- Verlag. , 1991. MR1261420

[12] P. Huber. *Robust Statistics.* Wiley, New York, 1981. MR0606374

[13] R. Jennrich. Asymptotic properties of non-linear least squares estimators. *Ann. Math. Stat.*, 40:633–643, 1969. MR0238419

[14] T. Kim and C. Muller. Two stage Huber estimation. *Journal of statistical planning and inference*, pages 405–418, 2007. MR2298946

[15] K. Knight. Epi-convergence in distribution and stochastic equi-semicontinuity. In *Corpus-based work*, pages 33–50, 1997.
[16] K. Knight and W. Fu. Asymptotics for Lasso-type estimators. *Ann. Stat.*, 28(5):1356–1378, 2000. MR1805787
[17] R. Koenker. *Quantile regression.* Econometric Society Monographs 38. Cambridge: Cambridge University Press. xv, 349 p., 2005. MR2268657
[18] C. Leng, Y. Lin, and G. Wahba. A note on the Lasso and related procedures in model selection. *Stat. Sin.*, 16(4):1273–1284, 2006. MR2327490
[19] L. McLinden and R. C. Bergstrom. Preservation of convergence of convex sets and functions in finite dimensions. *Trans. Am. Math. Soc.*, 268:127–142, 1981. MR0628449
[20] N. Meinshausen and P. Buhlmann. High-dimensional graphs and variable selection with the Lasso. *Ann. Stat.*, 34(3):1436–1462, 2006. MR2278363
[21] M. Osborne, B. Presnell, and B. Turlach. On the lasso and its dual. *journal of Computational and Graphical Statistics*, 9:319–337, 2000. MR1822089
[22] A. B. Owen. A robust hybrid of lasso and ridge regression. Technical report, 2006.
[23] G. C. Pflug. Asymptotic dominance and confidence for solutions of stochastic programs. *Czech. J. Oper. Res.*, 1(1):21–30, 1992.
[24] D. Pollard. Asymptotics for least absolute deviation regression estimators. *Econometric Theory*, 7:186–199, 1991. MR1128411
[25] W. J. Rey. *Introduction to robust and quasi-robust statistical methods.* Universitext. Berlin etc.: Springer-Verlag. IX, 236 p. DM 36.00; $ 14.00 , 1983.
[26] R. Rockafellar. *Convex analysis.* Princeton Landmarks in Mathematics. Princeton, NJ: Princeton University Press. , 1970. MR1451876
[27] R. Rockafellar and R. J.-B. Wets. *Variational analysis.* Grundlehren der Mathematischen Wissenschaften. , 1998. MR1491362
[28] P. J. Rousseeuw and C. Croux. Alternatives to the median absolute deviation. *J. Am. Stat. Assoc.*, 88(424):1273–1283, 1993. MR1245360
[29] S. Sardy, P. Tseng, and A. Bruce. Robust wavelet denoising. *Signal Processing, IEEE Transactions on [see also Acoustics, Speech, and Signal Processing, IEEE Transactions on]*, 49(6):1146–1152, 2001.
[30] G. Schwarz. Estimating the dimension of a model. *Ann. Stat.*, 6:461–464, 1978. MR0468014
[31] J. F. Sturm. Using SeDuMi 1. 02, a MATLAB toolbox for optimization over symmetric cones. 1999. MR1778433
[32] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1996. MR1379242
[33] A. Van der Vaart. *Asymptotic statistics.* Cambridge Series in Statistical and Probabilistic Mathematics, 3. Cambridge, 1998. MR1652247
[34] A. van der Vaart and J. A. Wellner. *Weak convergence and empirical processes. With applications to statistics.* Springer Series in Statistics. New York, NY: Springer. , 1996. MR1385671

[35] H. Wang, and C. Leng. Unified Lasso Estimation via Least Squares Approximation. *J. Am. Stat. Assoc.*, 102:1039–1048, 2007. MR2411663

[36] H. Wang, G. Li, and G. Jiang. Robust regression shrinkage and consistent variable selection through the LAD-Lasso. *Journal of Business & Economic Statistics*, 25(3):347–355, 2007. MR2380753

[37] H. Wang, R. Li, and C. Tsai. Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, 94(3):553–568, 2007. MR2410008

[38] P. Zhao and B. Yu. On Model Selection Consistency of Lasso. *Technical report, University of California, Berkeley. Dept. of Statistics*, 2006. MR2274449

[39] H. Zou. The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006. MR2279469