The BerHu penalty and the grouped effect

Sophie Lambert-Lacroix

UJF-Grenoble 1 / CNRS / UPMF / TIMC-IMAG

UMR 5525, Grenoble, F-38041, France

and

Laurent Zwald

LJK - Université de Grenoble

BP 53, 38041 Grenoble cedex 9, France

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. In the case of small sample size and large covariables numbers, this penalty is not very satisfactory variable selection method. In this paper, we introduce an adaptive reversed version of Huber's criterion as a penalty function. We call this penalty adaptive Berhu penalty. As for elastic net penalty, small coefficients contribute their ℓ_1 norm to this penalty while larger coefficients cause it to grow quadratically (as ridge regression). We show that the estimator associated with criterion such that ordinary least square or Huber's one combining with adaptive Berhu penalty enjoys the oracle properties. In addition, this procedure encourages a grouping effect. This approach is compared with adaptive elastic net regularization. Extensive simulation studies demonstrate satisfactory finite-sample performance of such procedure. A real example is analyzed for illustration purposes.

Keywords. Adaptive Berhu penalty; concomitant scale; elastic net penalty; Huber's criterion; oracle property; robust estimation.

Availability. The software that implements the procedures on which this paper focuses is developed in `Matlab`. It is available at `http://ljk.imag.fr/membres/Laurent.Zwald`.

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 eventually 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 ([25]). It consists to introduce ℓ_1 penalty. This penalty forces to shrink some coefficients. In [5], the authors show that since lasso uses the same tuning parameters for all the regression coefficients,

the resulting estimators may suffer an appreciable bias. Moreover in the case of the small sample n with larger number of covariables p, the lasso selects at most n variables. Recently, [18, 16, 33] and [34] 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. For instance, [34] assigns adaptive weights for penalizing differently coefficients in the ℓ_1 penalty and calls this new penalty the adaptive lasso. 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 [34]). Notice that recently (see [15]), this penalty has been combined with Huber's criterion. The estimator associated with this procedure enjoys oracle properties.

On the other hand, if there is a group of variables among which the pairwise correlations are very high, then the lasso penalty tends to select only any one variable from this group. Ridge regression (ℓ_2 penalty) does not make variables selection but tends instead to share the coefficients value among the group of correlated predictors. Moreover if there exist high correlations among predictors, the prediction performance of ridge regression dominated the lasso [25]. In order to overcome to this drawback of the lasso, [35] proposes a new regularization technique that combines the lasso and the ridge penalties. They call their method "elastic net" (en). The en penalty is the sum of the lasso and the ridge penalties. However even for usual case, it does not deemed to be an oracle procedure. In [6], the author proposes a new version of the elastic net called adaptive elastic net (adaptive en) which inherits some of the desirable properties of the adaptive lasso and elastic net. He proves its oracle properties. In [19], the author proposes to use a reversed version of Huber's criterion (called Berhu) as a penalty function. Let us recall that the Huber criterion (see [12]) is a hybrid of squared error for relatively small errors and absolute error for relative large ones. The Berhu penalty is such that relatively small coefficients contribute their ℓ_1 norm to this penalty whiles larger ones cause it to grow quadratically. This hybrid sets some coefficients to 0 as the lasso does while shrinking the larger coefficients in the same way as ridge regression. In [19], the author provides some way in order to optimize some objective function constituted of both the Huber criterion and the Berhu penalty in a no-adaptive form. Nevertheless nothing is shown about asymptotic feature.

In this paper we introduce an adaptive Berhu penalty with concomitant. We use it with the ordinary least square criterion or the Huber's one in order to take into account of data subject to heavy-tailed errors or outliers. We show that the estimator associated with such procedures enjoys the oracle properties (in the standard case of least square criterion and in the case of the Huber's one). In addition this procedure encourages a grouping effect in the following way. The spirit of the Berhu penalty with concomitant implicitly is to create one group with the largest coefficients. This group is penalized in a ℓ_2 way like the grouped lasso of [31] to avoid to remove anyone of these largest coefficients. The smallest coefficients are treated individually by an ℓ_1-penalty. The en procedure relies on the fact that, in order to have a grouped effect, we want to keep or delete together high correlated variables. We show that when combining with ordinary least squares criterion, the Berhu penalty leads to this "grouping effect property".

The rest of the article is organized as follows. In Section 2, we introduce the adaptive BerHu penalty and show that it induces a grouped effect. In Section 3, we give its statistical properties. Section 4 is devoted to simulation and illustration over real data. This study compares the least square criterion and the Huber's criterion with various penalties such as adaptive lasso, ridge, en

and adaptive Berhu. All technical proofs are relegated to the Appendix.

2 The Berhu penalty

2.1 The adaptive Berhu

Let us consider the linear regression model

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

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 (1) are usually estimated by minimizing the ordinary least squares criterion. To shrink unnecessary coefficients to 0, [25] proposed to introduce a constraint on the ℓ_1-norm of the coefficients:

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

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

Lots of reproaches have already been done to the Lasso (see e.g. [35]). In this paper, we focuse on the fact that when some variables are highly correlated, the ℓ_1 penalty tends to keep only one variable for each group. The literature already contains attempts to solve this problem. To begin with, grouped lasso procedures have been proposed first in [31] where the ℓ_1 penalty is imposed on predefined groups of coefficients. More precisely, the penalty is the l_1-norm of the vector composed of the ℓ_2-norm of each group of coefficients:

$$\sum_{i=1}^{n} (y_i - \alpha - \mathbf{x}_i^T \beta)^2 + \lambda_n \sum_{j=1}^{L} \sqrt{p_l} \|(\beta)_j\|_2,$$

where $(\beta)_j$ is the coordinates bloc corresponding to the j-th group. Consequently, the sparsity is encouraged at the group level (see also [32] and [9] page 91 for further references). In our framework it is diffcult to use the approach of group lasso since there is no obvious way for choosing the groups a priori. Next, [35] has proposed the Elastic Net. The naive Elastic Net is obtained by minimizing:

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

and the Elastic Net is a modification of this. In this procedure, the penalty imposed on the small coefficients is the sum of an ℓ_1-norm and a squared ℓ_2-norm. Moreover, ridge penalty reduces the variance of the estimates by imposing a small squared norm of all the coefficients. However, it suffices to constraint the largest coefficients to be small to get this reduction of variance: by definition, the smallest one do not need to be constrained to be small. Consequently, we consider a penalty which is quadratic only on the largest coefficient. Following [19], we focused on a penalty that acts separately on small and large coefficients. We consider the Berhu penalty defined by

$$\mathcal{B}_L(z) = \begin{cases} |z| & |z| \leq L, \\ \frac{z^2 + L^2}{2L} & |z| > L, \end{cases} \tag{3}$$

where L is any positive real. As Huber criterion, the Berhu function needs to be scaled. Precisely, the penalty can be defined by

$$\sum_{j=1}^{p} \mathcal{B}_L \left(\frac{\beta_j}{\tau} \right),$$

where τ is a scale parameter to be determined. To do that we can as in [19] replace the penalty term by

$$\text{pen}(\beta) = \min_{\tau > 0} \left(p\tau + \tau \sum_{j=1}^{p} \mathcal{B}_L \left(\frac{\beta_j}{\tau} \right) \right).$$

Fan and Li [5] 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 [34] proposes to consider the following modified lasso criterion, called adaptive lasso,

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

where $\hat{\mathbf{w}}^{adl} = (\hat{w}_1^{adl}, \ldots, \hat{w}_p^{adl})$ is a known weights vector. This modification allows to produce sparse solutions more effectively than lasso. Precisely, Zou [34] shows that with a proper choice of λ_n and of $\hat{\mathbf{w}}^{adl}$ the adaptive lasso enjoys the oracle properties. Such a penalty has been used in the en penalty (see [6]).

Here we propose to make the Berhu penalty adaptive. That is we consider the following penalty $\min_{\tau \in \mathbb{R}} P^{adb}(\beta, \tau)$ with

$$P^{adb}(\beta, \tau) = \begin{cases} \tau \left(\sum_{j=1}^{p} \frac{1}{\hat{w}_j^{adb}} + \sum_{j=1}^{p} \hat{w}_j^{adb} \mathcal{B}_L \left(\frac{\beta_j}{\tau} \right) \right) & if \quad \tau > 0, \\ 0 & if \quad \beta = 0, \ \tau = 0, \\ +\infty & if \quad \beta \neq 0, \ \tau = 0. \end{cases}$$

where $\hat{\mathbf{w}}^{adb} = (\hat{w}_1^{adb}, \ldots, \hat{w}_p^{adb})$ is a known weights vector. We will see at Section 3 that the resulting estimator enjoys the oracle properties. Let us notice that [19] introduced the Berhu penalty in his no-adaptive form and in the context of robust regression only. Moreover nothing is shown about asymptotic feature.

4

In the general case, the (adaptive) Behru penalty behaves like lasso on the smallest coefficients and does not delete the largest ones, whatever the correlation structure. That can be what we except to a right model selection procedure. This interpretation relies on the following calculation when β is fixed. For instance in the non adaptive case, let us sort the absolute values of the coordinates of β:

$$|\beta_{(p)}| \leq \cdots \leq |\beta_{(1)}|.$$

Let $k(\beta)$ denote the number of non-zeros coefficients of β. Then the minimum defined in $\text{pen}(\beta)$ is achieved at

$$\hat{\tau}(\beta) = \sqrt{\frac{1}{2Lp + L^2(q(\beta) - 1)} \sum_{j=1}^{q(\beta)-1} \beta_{(j)}^2},$$

if $\beta \neq 0$ and where $q(\beta)$ is the unique integer between 2 and $k(\beta) + 1$ such that $|\beta_{(q(\beta))}|/L \leq \hat{\tau}(\beta) \leq |\beta_{(q(\beta)-1)}|/L$. Consequently,

$$\text{pen}(\beta) = \sqrt{\frac{2p}{L} + q(\beta) - 1} \sqrt{\sum_{j=1}^{q(\beta)-1} \beta_{(j)}^2} + \sum_{j=q(\beta)}^{k(\beta)} |\beta_{(j)}| \,. \tag{4}$$

The en procedure (or its variant Elastic Corr-Net [4]) relies (explicitly for Elastic Corr-Net) on the fact that, in order to have a grouped effect, we want to keep or delete together high correlated variables. We will see that it is the case for Berhu procedure in Section 2.4. But we can note here different spirit of the Berhu penalty with concomitant: it implicitly creates one group with the largest coefficients (see (4)). This group is penalized in a ℓ_2 way like the grouped lasso of [31] to avoid to remove anyone of these largest coefficients. Let us note that as in the grouped lasso penalty, the ℓ_2-norm of the $q(\beta) - 1$ largest coefficients is scaled by the squared root of the number of such coefficients present in this group. The smallest coefficients are treated individually by an ℓ_1-penalty (see (4)). Consequently, whatever the structure of the correlation matrix, the Berhu penalty with concomitant tends to keep all the largest coefficients and to delete the smallest ones.

2.2   Robust estimation

To be robust to the heavy-tailed errors or outliers in the response, a 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 with concomitant scale defined by

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

which are to minimize with respect to $s \geq 0$, α and β. To get a robust scale invariant Lasso type procedure, [15] proposes to minimize simultaneously over s, α and β the function

$$\mathcal{L}_{\mathcal{H}}(\alpha, \beta, s) + \lambda_n \sum_{j=1}^{p} \hat{w}_j^{adh} |\beta_j|. \tag{5}$$

where $\hat{\mathbf{w}}^{adh} = (\hat{w}_1^{adh}, \ldots, \hat{w}_p^{adh})$ is a known weights vector. The loss function involving a concomitant estimation of the scale and location parameter was first proposed by Huber ([12]). We propose here to use the concomitant estimation of Huber with the Berhu penalty:

$$Q^{\mathcal{H}adb}(\alpha, \beta, s, \tau) = \mathcal{L}_{\mathcal{H}}(\alpha, \beta, s) + \lambda_n P^{adb}(\beta, \tau). \tag{6}$$

This criterion is minimized simultaneously over $\alpha \in \mathbb{R}, \beta \in \mathbb{R}^p, s \in \mathbb{R}_+$ and $\tau \in \mathbb{R}_+$. So we get another scale invariant robust location estimation. Contrary to the procedure proposed in [15], the largest coordinates of β are quadratically penalized.

2.3   Tuning parameter estimation

Let us now consider the problem of tuning parameter estimation. To run these procedures we have to determine the weights vector in the adaptive penalties, the regularization constant λ_n, the parameter M for Huber's criterion and L for Berhu's penalty. Usually the weights vector is given by (see [34, 15]) $\hat{w}_j^{adl} = |\hat{\beta}_j^{unpen}|^{-\gamma}$, $j = 1, \ldots, p$, where $\gamma > 0$ and $\hat{\beta}^{unpen}$ denotes the unpenalized estimator. For instance, in the least squares context $\hat{\beta}^{unpen}$ is the ordinary least squares estimator. In fact this estimator only must be root-n-consistent estimator of β^*. Let us note that the theoretical part is given for these forms of weights vector and γ is fixed to be equal to 1 for the numerical results. For Huber's Criterion with concomitant scale we need value for M. As in [12], we fix $M = 1.345$. For Berhu's penalty we fix as in [19], $L = M$. Let us note that we do not have any justification to do that. However in practice we have observed that these parameters have little impact on the results.

To find optimal values for λ_n, we use BIC-type criterions. When using least squares criterion we consider the classical BIC criterions ([22]), That is it is recommended to select λ_n minimizing

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

over λ_n, where k_{λ_n} denotes the model dimension. Following [28] and [30], we determine k_{λ_n} by the number of non-zero coefficients of the estimator. When using Huber's criterion, we consider the BIC-type procedure introduced in [15]: we select λ_n by minimizing

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

over λ_n. As previously, k_{λ_n} denotes the number of non-zero coefficients of $\hat{\beta}_{\lambda_n}$.

2.4 The Berhu penalty with concomitant induced a grouped effect

An algorithm is said to satisfy the grouping effect property if high correlated variables lead to similar estimations of the corresponding coefficients. Such a property was a motivation to introduce the Ridge Regression ([11]). Indeed, the normal equations associated to the Ordinary Least Square do not imply any stability of the coefficients associated to highly correlated variables. Now, adding a squared ℓ_2-norm penalty, the corresponding normal equations imply a stability of the coefficients associated to highly correlated variables. Such a reasoning leads to a bound quantifying the grouping effect of the Elastic Net ([35]). Such a property was generalized to the adaptive Elastic Net in [6] and also proved for the algorithm of [3].

The goal of the following theorem is to provide a quantitative description for the grouping effect of the Berhu penalty with concomitant.

Theorem 1. *Let $\gamma > 0$ and $(\hat{\alpha}^{adb}, \hat{\beta}^{adb}, \hat{\tau}^{adb})$ be a minimizer of*

$$\sum_{i=1}^{n} \left(y_i - \alpha - \mathbf{x}_i^T \beta\right)^2 + \lambda_n P^{adb}(\beta, \tau),$$

over $\alpha \in \mathbb{R}, \beta \in \mathbb{R}^p$ and $\tau \in \mathbb{R}_+$. We suppose that $\lambda_n > 0$, $\hat{\beta}_i^{adb} \neq 0$, $\hat{\beta}_j^{adb} \neq 0$. In this situation, the following bound holds:

$$|\hat{\beta}_i^{adb} \hat{w}_i^{adb} - \hat{\beta}_j^{adb} \hat{w}_j^{adb}| \leq \frac{2L\hat{\tau}}{\lambda_n} ||\underline{y}||_2 \sqrt{||x_i||_2^2 + ||x_j||_2^2 - 2C_{i,j} x_i^T x_j} \tag{7}$$

where $C_{i,j} = \min\left(1, \frac{|\hat{\beta}_j^{adb}|}{L\hat{\tau}^{adb}}, \frac{|\hat{\beta}_i^{adb}|}{L\hat{\tau}^{adb}}, \frac{|\hat{\beta}_i^{adb} \hat{\beta}_j^{adb}|}{(L\hat{\tau}^{adb})^2}\right)$.

To obtain this result for Huber's loss is a difficult task. That is an open question that is left for future work. Let us remark that when the variables are standardized in ℓ_2-norm, this leads to

$$|\hat{\beta}_i^{adb} \hat{w}_i^{adb} - \hat{\beta}_j^{adb} \hat{w}_j^{adb}| \leq \frac{2L\hat{\tau}}{\lambda} ||\underline{y}||_2 \sqrt{2\left(1 - C_{i,j} x_i^T x_j\right)}. \tag{8}$$

With $\gamma = 0$, we exactly get the grouping effect property in the non-adaptive case. Let now $\gamma \in \mathbb{R}_+$. The upper bound of equation (8) is a decreasing function of the correlation $x_i^T x_j$ between variables i and j (since $C_{i,j} > 0$). To ensure that the coefficients $\hat{\beta}_i^{adb}$ and $\hat{\beta}_j^{adb}$ become similar if the correlation increases, from (7), the initial estimator $\hat{\beta}^{unpen}$ used in the weights $\hat{w}_i^{adb}$ has to satisfy the grouping effect property. Consequently, this bound effectively provides a quantitative description for the grouping effect of the Berhu penalty with concomitant if, for example, the initial estimator is obtained with a ridge penalty.

As compared with the Elastic Net bounds provided by [35] and [6], we do not have to suppose that $\hat{\beta}_i^{adb}$ and $\hat{\beta}_j^{adb}$ have the same sign. Moreover, in our case, the grouping effect occurs more accurately for large coefficients (see Section 2.1) which is the natural situation where it has to happen. For the adaptive elastic net, [6] also have to suppose that the initial estimator satisfies the grouping effect property. Moreover, [36] recomands to choose a non-adaptive elastic net estimator as an initial estimator in the weights of the adaptive elastic net.

In the simulation study below, the initial estimator used for the weights of all adaptive methods is the corresponding unpenalized estimator. This choice avoids choosing a supplementary parameter (e.g. the regularization parameter of ridge regression) and also avoids numerical problems du to too small coefficients of the initial estimator. This unpenalized parameter does not satisfy the grouping effect property but comparisons between various methods remains fair. Moreover, in the simulation studies involving the Berhu penalty with concomitant, the variables were not normalized in ℓ_2-norm. Indeed, using the way we get the design matrix $\mathbf{X}$, explicit calculations when variables are normalized or not leads to the same order for the corresponding upper bounds.

3 Oracle Properties

In this section we give the asymptotic properties of the concomitant estimator of Huber with the Berhu penalty. We show that it enjoys the oracle properties. We have the same property by replacing Huber's loss by least squares one's. When necessary, we give the difference (for example for the assumptions) between the two loss functions.

Let $\mathbf{X}$ denotes the design matrix i.e. the $n \times p$ matrix the ith 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. It can be seen as a "compacity assumption": it is satisfied if the variables are supposed to be bounded. When considering least squares criterion as loss function, we need only the assumption **(D2)** (see for example [34]) while considering Huber's criterion we need the both **(D1)** and **(D2)** (see [15]).

Let us denote by ϵ a variable with the same law as ϵ_i, $i = 1, \ldots, n$. As in [15], we define

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

where for $s > 0$,

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

In addition, let us define $\tau^* > 0$ satisfying

$$\tau^* = \underset{\tau>0}{\operatorname{argmin}} \tau\left(\sum_{j=1}^{p}|\beta_j^*|^\gamma + \sum_{j=1}^{p_0}\frac{1}{|\beta_j^*|^\gamma}\mathcal{B}_L\left(\frac{\beta_j^*}{\tau}\right)\right). \tag{9}$$

The following assumptions on the errors are used in the following:

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

$$\mathbb{P}\left[\sigma\epsilon = \pm M s^*\right] = 0.$$

(N1) The variable ϵ is symmetric (i.e. ϵ has the same distribution as $-\epsilon$).

(N2) For all $a > 0$, $\mathbb{P}\left[\epsilon \in [-a, a]\right] > 0$.

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 [29]). 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. These three assumptions stand for the Huber's loss. For the penalized least squared estimators (e.g. [14] and [34]) we assume that ϵ_i are independent identically distributed random variables with mean 0 and has a finite variance.

Let $(\hat{\alpha}^{adb}, \hat{\beta}^{adb}, \hat{s}^{\mathcal{H}adb}, \hat{\tau}^{\mathcal{H}adb})$ be defined by the minimizer of $Q^{\mathcal{H}adb}(\cdot)$ where $\hat{w}_j^{adb} = 1/|\hat{\beta}_j^{unpen}|^\gamma$ with $\gamma > 1/3$ and $\hat{\beta}^{unpen}$ a root-n-consistent estimator of β^* (i.e. $\sqrt{n}(\hat{\beta} - \beta^*) = \mathcal{O}_P(1)$). We denote $\mathcal{A}_n = \{1 \leq j \leq p, \hat{\beta}_j^{\mathcal{H}adb} \neq 0\}$. Let us remark that if $\lambda_n > 0$, the argminimum $(\hat{\alpha}^{\mathcal{H}adb}, \hat{\beta}^{\mathcal{H}adb}, \hat{s}^{\mathcal{H}adb}, \hat{\tau}^{\mathcal{H}adb})$ exists since the criterion $Q^{\mathcal{H}adb}(\cdot)$ is a convex and coercive function.

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 5.3.

Theorem 2. *Suppose that $\lambda_n/n^{\gamma \wedge 1/2} \to 0$, $\lambda_n n^{(\gamma-1)/2} \to \infty$, $\lambda_n \to \infty$ and $\lambda_n > 1/3$. 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 $Q^{\mathcal{H}adb}$ are $\hat{w}_j^{adb} = 1/|\hat{\beta}_j^{unpen}|^\gamma$ where $\hat{\beta}^{unpen}$ is a root-n-consistent estimator of β^*. Then, any minimizer $(\hat{\alpha}^{\mathcal{H}adb}, \hat{\beta}^{\mathcal{H}adb}, \hat{s}^{\mathcal{H}adb}, \hat{\tau}^{\mathcal{H}adb})$ of $Q^{\mathcal{H}adb}$ satisfies the following:*

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

- *Asymptotic normality:*

$$\sqrt{n}\left(\hat{\alpha}^{\mathcal{H}adb} - \alpha^*, \hat{\beta}_{\mathcal{A}}^{\mathcal{H}adb} - \beta_{\mathcal{A}}^*, \hat{s}^{\mathcal{H}adb} - s^*, \frac{\sqrt{\lambda_n}}{\sqrt{n}}(\hat{\tau}^{\mathcal{H}adb} - \tau^*)\right) \longrightarrow_d \mathcal{N}_{p_0+3}\left(0, \Sigma^2\right),$$

where Σ^2 is the squared block diagonal matrix

$$\Sigma^2 = \mathtt{diag}\left(\frac{\mathbb{E}\left[\mathcal{H}_M'\left(\frac{\sigma\epsilon}{s^*}\right)^2\right]}{4A_{s^*}^2}, \frac{\mathbb{E}\left[\mathcal{H}_M'\left(\frac{\sigma\epsilon}{s^*}\right)^2\right]}{4A_{s^*}^2}V_{1,1}^{-1}, \frac{\mathbb{E}\left[Z^2\right]}{4D_{s^*}^2}, 0\right)$$

and where

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

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

Analogous results hold for the least squares loss function. In this case ($M = +\infty$), the asymptotic variance matrix $\mathbb{E}[\mathcal{H}_{Ms}'(\sigma\epsilon)^2]V_{1,1}^{-1}/(4A_{s^*}^2)$ obtained in Theorem 2 is equal to $\sigma^2 V_{1,1}^{-1}$ and we find the asymptotic variance of theorem 2 of [34].

4   Some numerical experiments

In this section, we consider the both criterions least squares and Huber's one combined with the following penalties: adaptive lasso, ridge, adaptive en and adaptive Berhu. We call these methods respectively `ad-lasso`, `ridge`, `ad-en`, `ad-Berhu` , `Huber-ad-lasso`, `Huber-ridge`, `Huber-ad-en` and `Huber-ad-Berhu`. The adaptive weights are obtained from the corresponding unpenalized estimator and $\gamma = 1$.

4.1   Simulation Results

Here our aim is to compare the finite sample performances of these procedures. Paragraph 4.1.1 presents the studied models. The way simulations are conducted is described in 4.1.2 and an insight of conclusions is provided in paragraph 4.1.3.

4.1.1   Models used for simulations

The models used to compare the performances of the algorithms are inspired by those presented in [35]. They involve groups of highly correlated variables: the block-variables model ([35], *example 4*). Let us remark that [35] considered a model without intercept. We now recall the definition of this model in a different way. Our formulation allows to clearly identify the groups of influencing correlated variables. 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 $(y_1, ..., y_n)^T$ (resp. $(\epsilon_1, ..., \epsilon_n)^T$). The design matrix $\mathbf{X}$ is constructed as follows. The rows of $\mathbf{X}$ are given by n independent gaussian vectors $\mathcal{N}_{40}(0, \Sigma)$. They are normalized such that the corresponding p-dimensional covariables are centered (as assumed in (1)). The variance matrix of the variables is a block diagonal matrix of size 40. The first block is the squared matrix of size 5 composed of 1 outside the diagonal and taking values 1.01 on the diagonal. The second and third blocks are the same as the first one. The last block is the identity matrix of size 25. The vector of true coefficients β^* is defined as follows: the 15 first coordinates are equal to 3 and the 25 last coefficients are 0. This means that, in this model, only the 15 first variables are influencing the response. The 25 others are pure noise. Amongst the 15 influencing variables, there is three groups of highly correlated variables: these groups are composed of the first five variables, the next five ones and the five last ones. The variables of different groups are independent. As compared with (1), this means that the intercept of the model is $\alpha^* = 1$ and the number of variables (without the intercept) is $p = 40$. Depending on the nature of the noise, various models are considered.

- Model 1: *block-variables model, gaussian noise.* In this case, the standard deviation of the noise is $\sigma = 15$ and the variables $\epsilon_1, \cdots, \epsilon_n$ are independent standard normal variables. Except for the part of the intercept parameter, this exactly example 4 of [35].

- Model 2: *block-variables model, mixture of gaussians.* In this case, the variables $\epsilon_1, \cdots, \epsilon_n$ are independent mixture of gaussians. Precisely, with probability 0.9, ϵ is a standard normal variable and with probability 0.1, ϵ is a centered normal with variance 225. The value $\sigma = 3.1009$ has been chosen such that the standard deviation of the noise is the same as in model 5. The common value is $\mathrm{std}(\sigma\epsilon) = 3.1009\sqrt{1 + 0.1(225 - 1)} = 15$.

- Model 3: *block-variables model, double-exponential noise.* $\epsilon = D/\sqrt{\mathrm{var}(D)}$ and $\sigma = 10.6$. The distribution of D is a standard double exponential i.e. its density is $x \in \mathbb{R} \to e^{-|x|}/2$ and $\mathrm{var}(D) = 2$.

These three models create a grouped variables situation. They allow us to illustrate the grouped selection ability of the penalties. They can be divided into two types. The first type contains light tailed errors models (1) whereas the second type is composed of heavy tailed errors models (2 and 3). Models 1 allows to quantify the deterioration of the performances of the robust algorithms in the absence of outliers. Thinking about the maximum likelihood approach, the least squares loss (resp. Huber's loss) is well designed for Models 1 (resp. 2,3).

4.1.2 Assessing prediction methods

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, we generate a first set of n training designs $(\mathbf{x}_1, \cdots, \mathbf{x}_n)$ and a second set of $m = 10\ 000$ test designs $(\mathbf{x}_{n+1}, \cdots, \mathbf{x}_{n+m})$. These two sets are centered in mean to stick on the theoretical definition (1) of the model (i.e. ensures that $\sum_{i=1}^n \mathbf{x}_i = 0$). 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 (1) of the model. All the algorithms have been run on the 100 training sets of size $n = 100, 200, 400$ and their prediction capacity have been evaluated on the test design set of size $m = 10\ 000$. To compare the prediction accuracy, the Relative Prediction Errors (RPEs) already considered in [34] are computed (see also [15] for explicit definition). Figures 1, 2 and 3, provide the boxplots associated with the 100 obtained RPE.

The model selection ability of the algorithms are reported in the same manner as done by [29], [25] and [5] in Tables 1, 2 and 3. Ridge penalty procedures are not reported since they do not constitute variables selection procedures. 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 15 first coordinates of $\hat{\beta}$ are non-zeros and the 25 last coefficients are zeros. To go further in the model selection ability analysis, we consider other measurements. The first (in the second column (O)) represents the number of overfitting models (i.e. those selecting all the non-zeros coefficients and at least one zero coefficient). The second (in 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 column (Z) is the average number of estimated zeros, the column (CZ) provides the average number of correctly estimated zeros and (TZ) recall the theoretical zeros number. The column (CNZ) is the average number of correctly estimated non zeros and (TNZ) recall the theoretical non zeros number. Models selection abilities are closely related to the accuracy of estimations of the coefficients. This fact is illustrated by boxplots of the coefficients estimations (see Figures 4, 5 and 6).

Concerning the hyperparameter choices, the regularization parameters associated with adaptive lasso or Berhu penalties are chosen by BIC criterion on each of the 100 training sets as described

at Section 2.3. The same grid has always been used for each method. It is composed of 100 points log-linearly spaced between 0 and 1400 for Berhu and 200 points log-linearly spaced between 0 and 10 000 for lasso. For Huber's loss, the simulation studies report the performances obtained with $M = 1.345$. This value has been recommended by Huber in [12]. For adaptive Berhu penalty, we report the performances obtained with $L = M = 1.345$ Let us remark that it is possible to chose the M and L parameters 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. For ridge-type procedures, the hyperparameter is chosen as usually by 5-fold cross-validation on each of the 100 training sets. The grid is composed of 100 points log-linearly spaced between 0 and 1400. For en-type procedure, we use the similar protocol as in [35]: we first pick a relatively small grid of values for $\lambda_{2,n}$ over $\{0, 0.01, 0.1, 1, 10, 100\}$ and 25 points log-linearly spaced between 0 and 5000 for $\lambda_{1,n}$. Then the both parameters are chosen simultaneously by 5-fold cross-validation.

4.1.3 Comparison results

Tables 1, 2 and 3 present the performances in terms of selection model ability. First we see that whatever the model the behavior of the methods are the same. The lasso and en penalties methods lead in general to underfitting models (columns U). It is surprising for the en penalty. Indeed the penalty imposed on the small coefficients is the sum of an ℓ_1-norm and a squared ℓ_2-norm. This implies that the obtained penalty is closer to differentiability than the ℓ_1-penalty. As shown in [1], if the penalty is far from differentiability, more small coefficients are deleted. For these examples, the en penalty as the same behavior as the lasso one. As a consequence, these penalties have a relatively high number of zeros with correct zeros number (columns Z) very close to the true one (columns TZ). But the correct non zeros number (columns CNZ) is very low in comparison with the true one (columns TNZ). The fact that en and lasso type methods underfit is reduced for Model 2 and for Huber loss. In all cases, these methods almost never identify the right model. The Behru penalty leads to some compromise between over and under fitting. We point out that contrary to en and lasso type methods, there is a case where Berhu type method identifies the right model a reasonable number of times: it is Model 2 with Huber loss. It is a little less good in terms of correct zeros but much better in terms of non zeros number.

This behavior occurs on the quality of estimation of the non zero coefficients (see Figures 4, 5 and 6)). Let us note that we have only considered the first coefficient β_1 and that the conclusions for the other non zero coefficient are the same. The ridge method is given here as a reference since it is known to lead good performances in presence of high correlation between the covariables. We observe that the Berhu penalty lead to good performance in terms of bias as ridge with higher variability than the ridge one. The bias and sometimes the variability are very high for the other methods du to their tendency to underfitting.

Figures 1, 2 and 3, provide the boxplots associated with RPE. As expected, the ordinary least squares loss leads to better performance for Model 1 (excepted for $n = 400$) and leads to less good performance for the Model 3 and especially for the Model 2. We observe that `ad-Berhu` and `Huber-ad-lasso` provide several extreme values du to numerical instabilities and are often more variable.

4.2 Prostate cancer data example

This data set comes from a prostate cancer study (see [24]) and analyzed earlier in the elastic net paper by [35, 6]. There are eight clinical covariates namely: logarithm of the cancer volume (lcavol), logarithm of the prostate weight (lweight), age, the logarithm of the amount of benign prostatic hyperplasia (lbph), seminal vesicle invasion (svi), logarithm of the capsular penetration (lcp), Gleason score (gleason) and percentage Gleason score 4 or 5 (pgg45). The response is the logarithm of prostate-specific antigen (lpsa). The predictors are are named as $1, \ldots, 8$ in results. OLS and the previous methods were applied to these data.

In [35], the data were divided into to parts: a training set with 67 observations and a test set with 30 observations while in [6], they have divided (randomly) the original data set in to training and testing set containing 60 and 37 observations respectively. To fairly compare the methods we propose to perform a resampling study: we have divided 100 times (randomly) the original data set into training and testing set containing 67 and 30 observations respectively. The hyperparameters are chosen as in the simulation study. We then compared the performances of the methods by computing their RPE on the 100 resampling testing sets (see Table 4). Contrary to what had been observed in [35, 6], our resampling study does not allow us to claim that one method emerges in terms of RPE: almost all these methods have similar RPE. We can only say perhaps `Huber-ad-lasso` is slightly less good. Let us notice that we observe a great variability in the choice of $\lambda_{2,n}$ for the adaptive en-type procedures (see first column of Table 4). This is also the case for `Huber-ad-lasso`. As a contrary, the choice of λ_n for Behru type procedures is more stable (it is comparable to the stability of ridge). Figure 7 show (except for OLS and ridge procedures) the histogram associated with the selected variables. We see that Berhu penalties leads to good models in terms of sparsity in comparison with en penalties. We observe that Behru type procedures are compromise between lasso type methods which select too few variables and en type methods which select too many variables.

5 Appendix

5.1 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 [23]).

Let us now recall a well-known fact of convex analysis: the Huber function is the Moreau-Yosida regularization of the absolute value function ([10, 20, 21]). Precisely, it can be easily shown that the Huber function satisfies

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

We can derive the same kind of formulation for the BerHu function leading to a characterization of the BerHu function as quadratic optimization problem. Indeed, the function (3) satisfies

$$\mathcal{B}_L(z) = \min_{w \geq L \vee |z|} \left(\frac{w^2}{2L} - w + |z| + \frac{L}{2} \right),$$

where $a \vee b$ denotes the maximum of the two real numbers a and b. The proof of this equality is trivial since it amounts to minimize a quadratic function on an interval.

This allows to write our optimization problem in a conforming manner to use CVX. Note that [19] 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 [19] 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) tau w(p);
minimize (n*s+quad_over_lin(y-alpha-X*beta-v,s)+2*M*norm(v,1)
+ mu*(tau*norm(betaUNP,1)+quad_over_lin(w./(sqrt(abs(betaUNP))),2*L*tau)
+norm(beta./betaUNP,1)-sum(w./abs(betaUNP))+0.5*L*tau*norm(1./betaUNP,1)))
subject to
s > 0;
tau > 0;
w >= L*tau;
w >= abs(beta);
cvx_end
```

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

5.2 Proof of Theorem 1

Since $\hat{\beta}_i^{adb} \neq 0$, we have $\hat{\beta}^{adb} \neq 0$ and $\hat{\tau}^{adb} > 0$. Consequently, the definition of partial derivatives involving Newton's quotient leads to the following KKT conditions by differentiating with respect to β_i, β_j and τ :

$$-2x_i^T \left(\underline{y} - \hat{\alpha}^{adb} \mathbb{1}_n - \mathbf{X}\hat{\beta}^{adb} \right) + \lambda_n \hat{w}_i^{adb} \mathcal{B}_L' \left(\frac{\hat{\beta}_i^{adb}}{\hat{\tau}^{adb}} \right) = 0, \tag{10}$$

$$-2x_j^T \left(\underline{y} - \hat{\alpha}^{adb} \mathbb{1}_n - \mathbf{X}\hat{\beta}^{adb} \right) + \lambda_n \hat{w}_j^{adb} \mathcal{B}_L' \left(\frac{\hat{\beta}_j^{adb}}{\hat{\tau}^{adb}} \right) = 0, \tag{11}$$

$$\sum_{j:|\hat{\beta}_j^{adb}|>L\hat{\tau}^{adb}} \hat{w}_j^{adb} \left(\frac{1}{2L} \left(\frac{\hat{\beta}_j^{adb}}{\hat{\tau}^{adb}} \right)^2 - \frac{L}{2} \right) = \sum_{j=1}^{p} \frac{1}{\hat{w}_j^{adb}}.$$

14

The last score equation implies that the set $G = \{j \in [1;p], |\hat{\beta}_j^{adb}| > L\hat{\tau}^{adb}\}$ is non-empty. Let us now distinguish some cases involving this set on indices. To begin with, if both the indexes i and j belong to G, equations (10) and (11) become

$$- 2x_i^T \left(\underline{y} - \hat{\alpha}^{adb}\mathbb{1}_n - \mathbf{X}\hat{\beta}^{adb} \right) + \lambda_n \hat{w}_j^{adb} \frac{\hat{\beta}_i^{adb}}{L\hat{\tau}^{adb}} = 0, \tag{12}$$

and

$$-2x_j^T \left(\underline{y} - \hat{\alpha}^{adb}\mathbb{1}_n - \mathbf{X}\hat{\beta}^{adb} \right) + \lambda_n \hat{w}_j^{adb} \frac{\hat{\beta}_i^{adb}}{L\hat{\tau}^{adb}} = 0.$$

Substracting the second one to the first one and using Cauchy-Schwarz inequality, we get :

$$|\hat{w}_i^{adb}\hat{\beta}_i^{adb} - \hat{w}_j^{adb}\hat{\beta}_j^{adb}| \leq \frac{2L\hat{\tau}^{adb}}{\lambda_n} \|x_i - x_j\|_2 \left\| \underline{y} - \hat{\alpha}^{adb}\mathbb{1}_n - \mathbf{X}\hat{\beta}^{adb} \right\|_2.$$

The definition of $(\hat{\alpha}^{adb}, \hat{\beta}^{adb}, \hat{\tau}^{adb})$ as a minimizer implies that, for all $\tau > 0$,

$$\left\| \underline{y} - \hat{\alpha}^{adb}\mathbb{1}_n - \mathbf{X}\hat{\beta}^{adb} \right\|_2 \leq \|\underline{y}\|_2 + \lambda_n \tau \sum_{j=1}^p \frac{1}{\hat{w}_j^{adb}}.$$

Now, letting τ tends to 0 in this inequality, we get:

$$\left\| \underline{y} - \hat{\alpha}^{adb}\mathbb{1}_n - \mathbf{X}\hat{\beta}^{adb} \right\|_2 \leq \|\underline{y}\|_2. \tag{13}$$

This leads to equation (7) of the Theorem 1 since $C_{i,j} = 1$ in this case.

Next, let us consider the case where only one index among $\{i, j\}$ belongs to G. If i and j are switched (if necessary), we can suppose that $i \in G$ and $j \notin G$. In this case, equations (10) and (11) become (12) and

$$-2x_j^T \left(\underline{y} - \hat{\alpha}^{adb}\mathbb{1}_n - \mathbf{X}\hat{\beta}^{adb} \right) + \lambda_n \hat{w}_j^{adb}\text{sign}\left(\hat{\beta}_i^{adb} \right) = 0.$$

These two equalities lead to

$$\hat{w}_i^{adb}\hat{\beta}_i^{adb} - \hat{w}_j^{adb}\hat{\beta}_j^{adb} = \frac{2L\hat{\tau}^{adb}}{\lambda - n} \left(x_i - \frac{|\hat{\beta}_j^{adb}|}{L\hat{\tau}^{adb}}x_j \right)^T \left(\underline{y} - \hat{\alpha}^{adb}\mathbb{1}_n - \mathbf{X}\hat{\beta}^{adb} \right).$$

Combining Cauchy-Schwarz inequality and inequality (13), this leads to

$$|\hat{w}_i^{adb}\hat{\beta}_i^{adb} - \hat{w}_j^{adb}\hat{\beta}_j^{adb}| \leq \frac{2L\hat{\tau}^{adb}}{\lambda_n} \|\underline{y}\|_2 \sqrt{\|x_i\|_2^2 + \|x_j\|_2^2 - 2\frac{|\hat{\beta}_j^{adb}|}{L\hat{\tau}^{adb}}x_i^T x_j},$$

where we have used $j \notin G$. This implies equation (7) of the Theorem 1 since $C_{i,j} = |\hat{\beta}_j^{adb}|/L\hat{\tau}^{adb}$ in this case.

Finally, when i and j do not belong to G, using similar arguments we obtain

$$|\hat{w}_i^{adb}\hat{\beta}_i^{adb} - \hat{w}_j^{adb}\hat{\beta}_j^{adb}| \leq \frac{2L\hat{\tau}^{adb}}{\lambda_n} \|\underline{y}\|_2 \sqrt{\|x_i\|_2^2 + \|x_j\|_2^2 - 2\frac{|\hat{\beta}_j^{adb}\hat{\beta}_i^{adb}|}{L^2\hat{\tau}^{adb2}}x_i^T x_j},$$

that implies equation (7) of the Theorem (7) since $C_{i,j} = |\hat{\beta}_j^{adb}\hat{\beta}_i^{adb}|/(L^2\hat{\tau}^{adb2})$ in this case. ∎

5.3 Proof of Theorem 2

The asymptotic normality of this estimator is proved in Step 1 and the consistency in variable selection in the Step 2. This proof in an adaptation to our case of the proof given by [34] or [15]. The difference with [15] concerns the treatment of the penalty term. So in the following, we will use notations similar to the ones of [15]. We will point out the difference between the both proofs.

Step 1. Let us first prove the asymptotic normality. Let us define $U_n(u) = Q^{\mathcal{H}adb}\left((\alpha^*, \beta^*, s^*, \tau^*)^T + u/\sqrt{n}\right) - Q^{\mathcal{H}adb}(\alpha^*, \beta^*, s^*, \tau^*)$ with $u = (u_0, \dots, u_{p+2})^T \in \mathbb{R}^{p+3}$. Obviously, $U_n(u)$ is minimized at

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

The principle of the proof of [34] or [15] is to study the epi-limit of U_n. Using the proof of theorem 3.2 in [15], we only need to study the epi-limit of the penalty term given by

$$P_n(u) = \lambda_n \left(\widetilde{P}^{adb}\left(\beta^* + \frac{u_{1:p}}{\sqrt{n}}, \tau^* + \frac{u_{p+2}}{\sqrt{\lambda_n}}\right) - \widetilde{P}^{adb}(\beta^*, \tau^*)\right),$$

where $\widetilde{P}^{adb}(\beta, \tau) = P^{adb}(\beta, \tau)$, if $\tau \geq 0$, ∞ if $\tau < 0$. The epi-limit of this term is given in the Lemma 1. This lemma together with lemma 2 of [15] indicates that $U_n \to_{e-d} U$, where $U(u) = A_{s^*}\left(u_{1:p}^T V u_{1:p} + u_0^2\right) + D_{s^*} u_{p+1}^2 - W^T u + u_{p+2}^2 C(u_{p+2})$, if $u_j = 0$, $\forall j \notin \mathcal{A}$, $+\infty$ otherwise. Under condition $\beta^* \neq 0$, equation (25) in Lemma 2 implies that $\sum_{j=1}^{p_0} |\beta_j^*|^{2-\gamma} \mathbb{1}_{|\beta_j^*| > L\tau^*} > 0$ thus the function $z \to z^2 C(z)$ is strictly convex. Moreover, $V_{1,1}$ is supposed positive definite in assumption **(D2)** and we assume that the noise satisfies **(N2)**. Consequently, U get a unique argmin and the asymptotic normality part is proved.

Step 2. Let us now show the consistency in variable selection part. It suffices to show that $\mathbb{P}\left[\mathcal{A} \subset \mathcal{A}_n\right] \to 1$ as n tends to infinity and $\mathbb{P}\left[\mathcal{A}^c \subset \mathcal{A}_n^c\right] \to 1$ as n tends to infinity. The first claim is an easy consequence of asymptotical normality obtained in Step 1.

Let us now show the second claim. Let j such that $\beta_j^* = 0$. We have to prove that $\mathbb{P}\left[\hat{\beta}_j^{\mathcal{H}adb} \neq 0\right] \to 0$ as n tends to infinity. As in [15], we have for a such j,

$$\mathbb{P}\left[\hat{\beta}_j^{\mathcal{H}adb} \neq 0\right] \leq \mathbb{P}\left[(\hat{s}^{\mathcal{H}adb}, \hat{\tau}^{\mathcal{H}adb}) = (0,0)\right] +$$

$$\mathbb{P}\left[\hat{\tau}^{\mathcal{H}adb} > 0 \text{ and } \hat{s}^{\mathcal{H}adb} > 0 \text{ and } \sum_{i=1}^n x_{i,j} \mathcal{H}_M'\left(\frac{y_i - \hat{\alpha}^{\mathcal{H}adb} - \mathbf{x}_i^T \hat{\beta}^{\mathcal{H}adb}}{\hat{s}^{\mathcal{H}adb}}\right) = -\lambda_n \hat{w}_j^{\mathcal{H}adb} \mathcal{B}_L'\left(\frac{\hat{\beta}_j^{\mathcal{H}adb}}{\hat{\tau}^{\mathcal{H}adb}}\right)\right].$$

Using similar arguments as in [15], we have, as n tends to infinity,

$$\mathbb{P}\left[(\hat{s}^{\mathcal{H}adb}, \hat{\tau}^{\mathcal{H}adb}) = (0,0)\right] \to 0.$$

Since $\forall x \in \mathbb{R}^*$, $|\mathcal{B}_L'(x)| \geq 1$, we have

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

$$\leq \mathbb{P}\left[\hat{s}^{\mathcal{H}adb} > 0 \text{ and } \frac{1}{\sqrt{n}}\left|\sum_{i=1}^n x_{i,j} \mathcal{H}_M'\left(\frac{y_i - \hat{\alpha}^{\mathcal{H}adb} - \mathbf{x}_i^T \hat{\beta}^{\mathcal{H}adb}}{\hat{s}^{\mathcal{H}adb}}\right)\right| \geq \frac{\lambda_n}{\sqrt{n}} \hat{w}_j^{\mathcal{H}adb}\right]$$

As in [15], we have

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_{i,j} \mathcal{H}'_M \left(\frac{y_i - \hat{\alpha}^{\mathcal{H}adb} - \mathbf{x}_i^T \hat{\beta}^{\mathcal{H}adb}}{\hat{s}^{\mathcal{H}adb}} \right) = O_P(1),$$

and $\sqrt{n}/(\lambda_n \hat{w}_j^{\mathcal{H}adb}) \xrightarrow{\mathbb{P}} 0$, that implies that $\mathbb{P}\left[\hat{\beta}_j^{\mathcal{H}adb} \neq 0 \right] \to 0$ as n tends to infinity. ∎

5.4 Technical lemma

5.4.1 Proof of lemma 1

Lemma 1. *Suppose that $\lambda_n/n^{\gamma \wedge 1/2} \to 0$, $\lambda_n n^{(\gamma-1)/2} \to \infty$, $\lambda_n \to \infty$, $\lambda_n > 1/3$ and $\beta^* \neq 0$. Then we have*

$$P_n(u) \to_{e-d} \begin{cases} u_{p+2}{}^2 C(u_{p+2}) & \text{if } u_j = 0, \ \forall j \notin \mathcal{A}, \\ +\infty & \text{otherwise}, \end{cases}$$

where

$$C(u_{p+2}) = \frac{1}{2L\tau^{*3}} \sum_{j=1}^{p_0} |\beta_j^*|^{2-\gamma} \mathbb{1}_{|\beta_j^*|>L\tau^*} + \frac{L^{1-\gamma}}{2\tau^{*(\gamma+1)}} \#\{1 \leq j \leq p, |\beta_j^*| = L\tau^*\} \mathbb{1}_{u_{p+2}<0}.$$

Since $\beta^* \neq 0$, Lemma 2 ensures that $\tau^* > 0$. Consequently, we have $P_n(u) = \sum_{j=1}^{p} P_{n,j}(u)$, where

$$P_{n,j}(u) = \begin{cases} \lambda_n \left(\frac{u_{p+2}}{\sqrt{\lambda_n} \hat{w}_j^{adb}} + \hat{w}_j^{adb} \left(\tau^* + \frac{u_{p+2}}{\sqrt{\lambda_n}} \right) \mathcal{B}_L \left(\frac{\beta_j^* + \frac{u_j}{\sqrt{n}}}{\tau^* + \frac{u_{p+2}}{\sqrt{\lambda_n}}} \right) - \tau^* \hat{w}_j^{adb} \mathcal{B}_L \left(\frac{\beta_j^*}{\tau^*} \right) \right) & \text{if } \quad u_{p+2} > -\sqrt{\lambda_n} \tau^*, \\ -\lambda_n \tau^* \left(\frac{1}{\hat{w}_j^{adb}} + \hat{w}_j^{adb} \mathcal{B}_L \left(\frac{\beta_j^*}{\tau^*} \right) \right) & \text{if } \quad u_{p+2} = -\sqrt{\lambda_n} \tau^*, \\ & \text{and } \quad u_j = -\sqrt{n} \beta_j^*, \\ +\infty & \text{otherwise.} \end{cases}$$

Step 1. First let us prove that

$$\sum_{j=1}^{p_0} P_{n,j}(u) \to_{e-d} u_{p+2}^2 C(u_{p+2}). \tag{14}$$

We show that, for every u fixed in $\mathbb{R}^{p+2}$, we have this convergence in probability. Since $\tau^* > 0$ and $\lambda_n \to +\infty$ as n tends to infinity, for n sufficiently large (with respect to a bound depending on u_{p+2}), $u_{p+2}/\sqrt{\lambda_n} + \tau^* > 0$ and

$$P_{n,j}(u) = \frac{u_{p+2}\sqrt{\lambda_n}}{\hat{w}_j^{adb}} + \lambda_n \hat{w}_j^{adb} \left(G_j \left(\frac{u_j}{\sqrt{n}}, \frac{u_{p+2}}{\sqrt{\lambda_n}} \right) - G_j(0) \right),$$

where

$$\forall j \in [1, p_0], \ G_j : (z_1, z_2) \to (z_2 + \tau^*) \mathcal{B}_L \left(\frac{z_1 + \beta_j^*}{z_2 + \tau^*} \right).$$

17

For $1 \le j \le p_0$ such that $|\beta_j^*| \neq L\tau^*$, G_j is two times differentiable at 0 and the Taylor-Young theorem entails that, $\forall (z_1, z_2) \in \mathbb{R}^2$,

$$
G_j(z_1, z_2) = G_j(0) + z_1 \mathcal{B}_L' \left(\frac{\beta_j^*}{\tau^*} \right) + z_2 B \left(\frac{\beta_j^*}{\tau^*} \right) + \frac{z_1^2}{2L\tau^*} \mathbb{1}_{|\beta_j^*| > L\tau^*}
$$
$$
+ \frac{z_2^2 \beta_j^{*2}}{2L\tau^{*3}} \mathbb{1}_{|\beta_j^*| > L\tau^*} - \frac{z_1 z_2 \beta_j^*}{L\tau^{*2}} \mathbb{1}_{|\beta_j^*| > L\tau^*} + \xi(z_1, z_2),
$$

where $\xi(z_1, z_2) / \|(z_1, z_2\|^2 \to 0$ as $(z_1, z_2) \to 0$, $B : z \in \mathbb{R} \to \mathcal{B}_L(z) - z \mathcal{B}_L'(z)$ and we have used that $\mathcal{B}_L'' \left(\beta_j^*/\tau^* \right) = \mathbb{1}_{|\beta_j^*| > L\tau*}/L$. Consequently, for $1 \le j \le p_0$ such that $|\beta_j^*| \neq L\tau^*$,

$$
P_{n,j}(u) = \frac{u_{p+2}\sqrt{\lambda_n}}{\hat{w}_j^{adb}} + \sqrt{\lambda_n} u_{p+2} \hat{w}_j^{adb} B \left(\frac{\beta_j^*}{\tau^*} \right) + \frac{u_{p+2}^2 |\beta_j^*|^{2-\gamma}}{2L\tau^{*3}} \mathbb{1}_{|\beta_j^*| > L\tau*} + a_{n,j}(u), \qquad (15)
$$

where

$$
a_{n,j}(u) = \frac{\lambda_n u_j \hat{w}_j^{adb}}{\sqrt{n}} \mathcal{B}_L' \left(\frac{\beta_j^*}{\tau^*} \right) + \frac{\lambda_n u_j^2 \hat{w}_j^{adb}}{2nL\tau^*} \mathbb{1}_{|\beta_j^*| > L\tau*} - \frac{\sqrt{\lambda_n} u_{p+2} \beta_j^* u_j \hat{w}_j^{adb}}{\sqrt{n} L\tau^{*2}} \mathbb{1}_{|\beta_j^*| > L\tau*} + \lambda_n \hat{w}_j^{adb} \xi \left(\frac{u_j}{\sqrt{n}}, \frac{u_{p+2}}{\sqrt{\lambda_n}} \right).
$$

Let us now consider $1 \le j \le p_0$ such that $|\beta_j^*| = L\tau^*$. When $\beta_j^* = L\tau^*$, for n sufficiently large (with respect to a bound depending on u),

$$
P_{n,j}(u) = \frac{\sqrt{\lambda_n} u_{p+2}}{\hat{w}_j^{adb}} + \lambda_n \hat{w}_j^{adb} \left(\left(\tau^* + \frac{u_{p+2}}{\sqrt{\lambda_n}} \right) \mathcal{B}_L \left(L\tau^* + \frac{\frac{u_j}{\sqrt{n}}}{\tau^* + \frac{u_{p+2}}{\sqrt{\lambda_n}}} \right) - L\tau^* \right)
$$

Let us consider n sufficiently large (with respect to a bound depending on u) such that $L\tau^* + u_j/\sqrt{n} > 0$ and $\tau^* + u_{p+2}/\sqrt{\lambda_n} > 0$. It is possible since $\tau^* > 0$. Thus, combined with the assumption $\lambda_n \to +\infty$ as n tends to ∞, the involved sequence tends to a strictly positive limit as n tends to ∞. Since $\lambda_n/n \to 0$ as n tends to ∞, two cases are possible. Either, $\sqrt{\lambda_n/n} u_j \le L u_{p+2}$ and

$$
b_{n,j}(u) = P_{n,j}(u) - \frac{\sqrt{\lambda_n} u_{p+2}}{\hat{w}_j^{adb}} = \frac{\lambda_n u_j \hat{w}_j^{adb}}{\sqrt{n}}, \qquad (16)
$$

or $\sqrt{\lambda_n/n} u_j > L u_{p+2}$ and

$$
b_{n,j}(u) = \frac{\lambda_n \hat{w}_j^{adb}}{\left(\tau^* + \frac{u_{p+2}}{\sqrt{\lambda_n}} \right)} \left(\frac{u_j^2}{2Ln} + \frac{\tau^* u_j}{\sqrt{n}} \right) + \frac{L \hat{w}_j^{adb}}{2 \left(\tau^* + \frac{u_{p+2}}{\sqrt{\lambda_n}} \right)} u_{p+2}^2. \qquad (17)
$$

Similarly, we get the same result if $\beta_j^* = -L\tau^*$. Gathering (15) and using $B(\pm L) = 0$, we have the following decomposition:

$$
\sum_{j=1}^{p_0} P_{n,j}(u) = \sum_{j=1}^{p_0} c_{n,j}(u) + \sum_{j=1}^{p_0} \left(a_{n,j}(u) \mathbb{1}_{|\beta_j^*| \neq L\tau^*} + b_{n,j}(u) \mathbb{1}_{|\beta_j^*| = L\tau^*} \right) + \frac{u_{p+2}^2}{2L\tau^{*3}} \sum_{j=1}^{p_0} |\beta_j^*|^{2-\gamma} \mathbb{1}_{|\beta_j^*| > L\tau*},
$$
$$(18)$$

where

$$c_{n,j}(u) = u_{p+2}\sqrt{\lambda_n}\sum_{j=1}^{p_0}\left(\frac{1}{\hat{w}_j^{adb}} + \hat{w}_j^{adb}B\left(\frac{\beta_j^*}{\tau^*}\right)\right).$$

We now study the convergence of each term. The $\sqrt{n}$-consistency of $\hat{\beta}^{unpen}$ implies that $\hat{w}_j^{adb} \xrightarrow{\mathbb{P}} 1/|\beta_j^*|^\gamma < +\infty$. Moreover, $\lambda_n/\sqrt{n} \to 0$ as n tends to infinity, thus, by Slutsky's theorem, the first three terms of $a_{n,j}(u)$ tends to 0 in probability for any $(u) \in \mathbb{R}^{p+2}$ fixed. Concerning the last term (the rest), we have that

$$\forall \epsilon > 0\,, \exists\, N_\epsilon(u)\,, \forall n \geq N_\epsilon(u),\ \lambda_n\xi\left(\frac{u_j}{\sqrt{n}}, \frac{u_{p+2}}{\sqrt{\lambda_n}}\right) \leq \epsilon\left(\frac{u_j^2\lambda_n}{n} + u_{p+2}^2\right).$$

Moreover, $(\lambda_n/n)_{n\geq 1}$ is a bounded sequence (since it converges to 0 as n tends to infinity). Thus, $\lambda_n\xi(u_j/\sqrt{n}, u_{p+2}/\sqrt{\lambda_n}) \to 0$ as n tends to ∞. Consequently, for any $u \in \mathbb{R}^{p+2}$ fixed, the forth term of $a_{n,j}$ tends to 0 in probability. Using Slutsky's lemma, this entails that, for any $u \in \mathbb{R}^{p+2}$ fixed, $a_{n,j}(u) \xrightarrow{\mathbb{P}} 0$. Concerning the term $b_{n,j}(u)$ As previously we have $\hat{w}_j^{adb} \xrightarrow{\mathbb{P}} 1/|\beta_j^*|^\gamma < +\infty$ and $\lambda_n/\sqrt{n} \to 0$ as n tends to infinity, so, if $\beta_j^* = L\tau^*$,

$$b_{n,j}(u) \xrightarrow{\mathbb{P}} \frac{L^{(1-\gamma)}}{2\tau^{*(\gamma+1)}}u_{p+2}^2\mathbb{1}_{u_{p+2}<0}\,.$$

Similarly, we get the same result if $\beta_j^* = -L\tau^*$. Concerning the term $c_{n,j}(u)$, Property (25) (see Lemma 2) is available since $\beta^* \neq 0$ and

$$c_{n,j}(u) = u_{p+2}\sqrt{\frac{\lambda_n}{n}}\sum_{j=1}^{p_0}\left(\sqrt{n}(|\hat{\beta}_j^{unpen}|^\gamma - |\beta_j^*|^\gamma) + B\left(\frac{\beta_j^*}{\tau^*}\right)\left(\sqrt{n}\frac{\left(|\hat{\beta}_j^{unpen}|^\gamma - |\beta_j^*|^\gamma\right)}{|\hat{\beta}_j^{unpen}|^\gamma|\beta_j^*|^\gamma}\right)\right). \quad (19)$$

Since $\beta_j^* \neq 0$, $x \to |x|^\gamma$ is differentiable at β_j^* and the Taylor-Young theorem entails that

$$\sqrt{n}\left(|\hat{\beta}_j^{unpen}|^\gamma - |\beta_j^*|^\gamma\right) = \gamma\mathrm{sign}(\beta_j^*)|\beta_j^*|^{\gamma-1}\sqrt{n}\left(\hat{\beta}_j^{unpen} - \beta_j^*\right) + \sqrt{n}\left(\hat{\beta}_j^{unpen} - \beta_j^*\right)\xi_j\left(\hat{\beta}_j^{unpen}\right)$$

with $\xi_j(x) \to 0$ as x tends to β_j^*. Now, the $\sqrt{n}$-consistency of $\hat{\beta}^{unpen}$ implies that the first term of this expansion is bounded in probability. It also entails that $\hat{\beta}_j^{unpen} \xrightarrow{\mathbb{P}} \beta_j^*$ which leads to $\xi_j\left(\hat{\beta}_j^{unpen}\right) \xrightarrow{\mathbb{P}} 0$ since $\xi_j(x) \to 0$ as x tends to β_j^*. Consequently, the second term of this expansion is also bounded in probability and, finally, $\sqrt{n}(|\hat{\beta}_j^{unpen}|^\gamma - |\beta_j^*|^\gamma) = \mathcal{O}_P(1)$. Since $\lambda_n/n \to 0$ as n tends to infinity, and $|\hat{\beta}_j^{unpen}|^\gamma \xrightarrow{\mathbb{P}} |\beta_j^*|^\gamma \neq 0$, so $c_{n,j}(u)$ converges in probability to 0. Combining (18) with all these convergences, the convergence in probability of (14) is proved. Using first theorem 2.7 (vi) of [26] and then that convergence in probability is stronger than convergence in distribution (theorem 2.7 (ii) of [26]), we get that convergence in probability implies finite-dimensional convergence in (14). Theorem 5 of [13] implies that (14) holds since the limit function $u \to u_{p+2}^2C(u_{p+2})$ is finite.

Step 2. Next, we treat the sum of terms $P_{n,j}$ for $j > p_0$, and first show that

$$(P_{n,p_0+1}, \cdots, P_{n,p}) \to_{e-d} \left(I_{B_{p_0+1}}, \cdots, I_{B_p}\right), \quad (20)$$

19

where $B_j = \{(u_{1:p}, u_{p+2}) \in \mathbb{R}^{p+1}, u_j = 0\}$ and for a set A, I_A denotes the indicator function of A (i.e. $I_A(x) = 0$ if $x \in A$ and $I_A(x) = +\infty$ otherwise). Let us put

$$q_{n,j}(u) = P_{n,j}(u) - \sqrt{\lambda_n} u_{p+2} |\hat{\beta}_j^{unpen}|^\gamma. \tag{21}$$

Since $\hat{\beta}^{unpen}$ is a $\sqrt{n}$-consistent estimator and $j \in [p_0 + 1, p]$, $n^{\gamma/2}|\hat{\beta}_j^{unpen}|^\gamma$ is a tight sequence. Moreover, we have $\lambda_n/n^\gamma \to 0$ as n tends to infinity, thus $\forall u_{p+2} \in \mathbb{R}$, $\sqrt{\lambda_n} u_{p+2} |\hat{\beta}_j^{unpen}|^\gamma = u_{p+2}\sqrt{\lambda_n n^{-\gamma}}(\sqrt{n}|\hat{\beta}_j^{unpen}|)^\gamma \xrightarrow{\mathbb{P}} 0$. Using first theorem 2.7 (vi) of [26], we get that convergence in probability implies finite-dimensional convergence: $\sqrt{\lambda_n} u_{p+2} |\hat{\beta}_j^{unpen}|^\gamma \to_{f-d} 0$. Since the involved limit function is finite and by convexity, theorem 5 of [13] ensures that we have the epiconvergence in distribution. Moreover, $\mathcal{B}_L(x) \geq |x|$ and $\mathcal{B}_L(0) = 0$, Lemma 3 with $q(x) = \mathcal{B}_L(x)$ leads to

$$d(q_{n,j}, I_{B_j}) \leq 2^{-\left[\tau^* \sqrt{\lambda_n}\right]+1} + \frac{2\sqrt{n}|\hat{\beta}_j^{unpen}|^\gamma}{\lambda_n},$$

where d is defined as in (26). We have $\lambda_n \to +\infty$ as n tends to infinity and $2^{-\left[\tau^* \sqrt{\lambda_n}\right]+1} \to 0$ as n tends to infinity since $\tau^* > 0$. Furthemore $2\sqrt{n}|\hat{\beta}_j^{unpen}|^\gamma/\lambda_n = 2(\sqrt{n}|\hat{\beta}_j^{unpen}|)^\gamma/\lambda_n/n^{(\gamma-1)/2}$ and since $\hat{\beta}^{unpen}$ is a $\sqrt{n}$-consistent estimator and $j \in [p_0 + 1, p]$, the numerator is a tight sequence and the denominator tends to $+\infty$ as n tends to infinity. Consequently, $2\sqrt{n}|\hat{\beta}_j|^\gamma/\lambda_n \xrightarrow{\mathbb{P}} 0$ and $d(q_{n,j}, I_{B_j}) \xrightarrow{\mathbb{P}} 0$. Finally, using part (ii) of lemma 1.10.2 page 57 of [27], we have $q_{n,j} \to_{e-d} I_{B_j}$. 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 [27]. Consequently, we can use Slutsky's theorem page 32, example 1.4.7 of [27] to ensure that

$$\left(\sqrt{\lambda_n} u_{p+2} |\hat{\beta}_j|^\gamma, q_{n,j}(u_{1:p}, u_{p+2})\right) \to_{e-d} (0, I_{B_j}) \tag{22}$$

since 0 is deterministic. Moreover, we have $\sqrt{\lambda_n} u_{p+2} |\hat{\beta}_j^{unpen}|^\gamma \to_{u-d} 0$ since we have shown the finite dimensional convergence in distribution and since $\sqrt{\lambda_n} u_{p+2} |\hat{\beta}_j|^\gamma$ and 0 are finite convex functions ([2] and [13]). We are now in position to use part (b) of theorem 4 of [13]: gathering (22), $\sqrt{\lambda_n} u_{p+2} |\hat{\beta}_j^{unpen}|^\gamma \to_{u-d} 0$, continuity of 0 and (21), it ensures that $P_{n,j} \to_{e-d} I_{B_j}$ holds. Since I_{B_j} is deterministic, theorem 18.10 (ii) of [26] ensures that the convergence in probability holds. Now, theorem 18.10 (vi) of [26] leads to the convergence in probability in (20). Moreover, convergence in probability is stronger than convergence in distribution thus (20) is proved.

For all $I \subset [p_0 + 1, p]$, $\text{dom}\left(\sum_{i \in I} I_{B_i}\right) = \{(u_{1:p}, u_{p+2}) \in \mathbb{R}^{p+1}, u_i = 0, \forall i \in I\}$. Thus, for all $I \subset [p_0 + 1, p]$ and $J \subset [p_0 + 1, p]$ satisfying $I \cap J = \emptyset$,

$$0 \in \text{int}\left(\text{dom}\left(\sum_{i \in I} I_{B_i}\right) - \text{dom}\left(\sum_{j \in J} I_{B_j}\right)\right),$$

where for f, a function defined on $\mathbb{R}^{p+1}$, $\text{dom}(f) = \{x \in \mathbb{R}^{p+1}/f(x) < +\infty\}$ and $A - B = \{a - b, a \in A, b \in B\}$. Using successively this fact, (20), Theorem 5 of [17] and theorem 18.10 (iii) (v) (vi) and 18.11 of [26], we get

$$\sum_{j=p_0+1}^p P_{n,j} \to_{e-d} \sum_{j=p_0+1}^p I_{B_j} \tag{23}$$

20

As previously, we can use Slutsky's theorem page 32, example 1.4.7 of [27] to ensure that (23) and (14) imply that

$$\left(\sum_{j=p_0+1}^{p} P_{n,j}(u), \sum_{j=1}^{p_0} P_{n,j}(u) \right) \to_{e-d} \left(\sum_{j=p_0+1}^{p} I_{B_j}, u_{p+2}^2 C(u_{p+2}) \right) \tag{24}$$

since $u_{p+2}^2 C(u_{p+2})$ is deterministic. Moreover, we have $\sum_{j=1}^{p_0} P_{n,j}(u) \to_{u-d} u_{p+2}^2 C(u_{p+2})$ since we have shown the finite dimensional convergence in distribution and $\sum_{j=1}^{p_0} P_{n,j}(u)$ and $u_{p+2}^2 C(u_{p+2})$ are finite (for n sufficiently large) convex functions ([2] and [13]). Using part (b) of theorem 4 of [13]: gathering (24), $\sum_{j=1}^{p_0} P_{n,j}(u_{1:p}, u_{p+2}) \to_{u-d} u_{p+2}^2 C(u_{p+2})$ and continuity of $u_{p+2}^2 C(u_{p+2})$, it ensures that Lemma 1 holds. ∎

5.4.2 Proof of lemma 2

Lemma 2. *If $\beta^* \neq 0$ then there exists a unique $\tau^* > 0$ satisfying equation (9) and*

$$\sum_{j=1}^{p} |\beta_j^*|^\gamma + \sum_{j=1}^{p_0} \frac{1}{|\beta_j^*|^\gamma} \left(\mathcal{B}_L \left(\frac{\beta_j^*}{\tau^*} \right) - \frac{\beta_j^*}{\tau^*} \mathcal{B}_L' \left(\frac{\beta_j^*}{\tau^*} \right) \right) = 0. \tag{25}$$

Proof. Let us denote by I the following function of τ

$$I(\tau) = \tau \left(\sum_{j=1}^{p} |\beta_j^*|^\gamma + \sum_{j=1}^{p_0} \frac{1}{|\beta_j^*|^\gamma} \mathcal{B}_L \left(\frac{\beta_j^*}{\tau} \right) \right).$$

This function is convex and $I'(\cdot)$ is continuous, increasing with $I'(\tau) \to \sum_{j=1}^{p} |\beta_j^*|^\gamma$ as $\tau \to +\infty$ and, if $\beta^* \neq 0$, $I'(\tau) \to -\infty$ as $\tau \to 0$. This leads to the existence of $\tau* > 0$ by the intermediate value theorem. The minimum of I is unique since I' is strictly increasing on each pieces $]0, |\beta_{(1)}^*|/L[$ and $[|\beta_{(k)}^*|/L, |\beta_{(k+1)}^*|/L[$ for $1 \leq k \leq p-1$, continuous and increasing on $\mathbb{R}_+^*$, strictly positive at $|\beta_{(p)}^*|/L$ since $I'(|\beta_{(p)}^*|/L) = \sum_{j=1}^{p} |\beta_j^*|^\gamma > 0$. Note that I' is constant on $[|\beta_{(p)}^*|/L, +\infty[$. This concludes the proof. ∎

5.4.3 Proof of lemma 3

For f, a function defined on S, we note $epi(f)$, its epigraph given by $epi(f) = \{(x,t) \in S \times \mathbb{R} / f(x) \leq t\}$.

Lemma 3. *Let q be a function such that $q(0) = 0$ and $\forall x \in \mathbb{R}$, $q(x) \geq |x|$. We use the notations of the proof of lemma 1. Let us recall that $q_{n,j}(u_{1:p}, u_{p+2}) = P_{n,j}(u_{1:p}, u_{p+2}) - \sqrt{\lambda_n} u_{p+2} |\hat{\beta}_j|^\gamma$ where*

$$P_{n,j}(u) = \begin{cases} \lambda_n \left(\frac{u_{p+2}}{\sqrt{\lambda_n}} |\hat{\beta}_j|^\gamma + \frac{1}{|\hat{\beta}_j|^\gamma} \left(\frac{u_{p+2}}{\sqrt{\lambda_n}} + \tau^* \right) q \left(\frac{\frac{\gamma_j}{\sqrt{n}} + \beta_j^*}{\frac{u_{p+2}}{\sqrt{\lambda_n}} + \tau^*} \right) - \frac{\tau^*}{|\hat{\beta}_j|^\gamma} q \left(\frac{\beta_j^*}{\tau^*} \right) \right) & \text{if } u_{p+2} > -\sqrt{\lambda_n} \tau^*, \\ -\lambda_n \tau^* \left(|\hat{\beta}_j|^\gamma + \frac{1}{|\hat{\beta}_j|^\gamma} q \left(\frac{\beta_j^*}{\tau^*} \right) \right) & \text{if } u_{p+2} = -\sqrt{\lambda_n} \tau^*, \\ & \quad and \gamma_j = -\sqrt{n} \beta_j^*, , \\ +\infty & \text{otherwise}. \end{cases}$$

21

Then, $\forall j \in [p_0 + 1, p]$,

$$d(q_{n,j}, I_{B_j}) \leq 2^{-\left[\tau^* \sqrt{\lambda_n}\right] + 1} + \frac{2\sqrt{n}|\hat{\beta}_j|^{\gamma}}{\lambda_n},$$

where

$$d(q_{n,j}, I_{B_j}) = \sum_{k=1}^{+\infty} \frac{1 \wedge d_k(epi(q_{n,j}), epi(I_{B_j}))}{2^k}, \tag{26}$$

d_k *is a semi-distance ("constrained Pompeiu-Haussdorf distance")*

$$d_k(epi(q_{n,j}), epi(I_{B_j})) = \max_{\|x\| \leq k} |d_{epi(q_{n,j})}(x) - d_{epi(I_{B_j})}(x)|, \tag{27}$$

and $d_S(x) = \min_{y \in S} \|x - y\|$ *for a subset S of $\mathbb{R}^{p+1}$.*

Proof. Let us note that distance d caracterises the epi-convergence of lower semi-continuous functions: a sequence $\{f_n\}$ of extended-real-valued lower semi-continuous functions from $\mathbb{R}^{p+1}$ *epi-converges* to a extended-real-valued lower semi-continuous function f if and only if $d(f_n, f) \to 0$ as n goes ton infinity. We recall that $B_j = \{(u_{1:p}, u_{p+2}) \in \mathbb{R}^{p+1}, u_j = 0\}$ and for a set A, I_A denotes the indicator function of A. Let us introduce the set $D_j = \{(u_{1:p}, u_{p+2}) \in \mathbb{R}^{p+1}, u_j = 0 \text{ and } u_{p+2} \geq -\sqrt{\lambda_n}\tau^*\}$. By using the triangular inequality,

$$d(q_{n,j}, I_{B_j}) \leq d(q_{n,j}, I_{D_j}) + d(I_{D_j}, I_{B_j}). \tag{28}$$

To begin with, let us show that

$$d(I_{D_j}, I_{B_j}) \leq 2^{-\left[\tau^* \sqrt{\lambda_n}\right]}. \tag{29}$$

Here we use a geometrical point of view. The epigraph of the indicator function I_A of a set A is the "half- cylinder with cross-section A" i.e. $A \times \mathbb{R}_+$. Consequently, the epigraph of I_{B_j} is an half-hyperplan supported by the u_j axis and the epigraph of I_{D_j} is the part of this half-hyperplan where, moreover, $u_{p+2} \geq -\sqrt{\lambda_n}\tau^*$. Note that this cut is perpendicular to the u_{p+2}-axis. So if we consider $x \in \mathbb{R}^{p+2}$ such that $x_{p+1} \geq -\sqrt{\lambda_n}\tau^*$, the distance between x and $\text{epi}(I_{D_j})$ is reached for a point in $\text{epi}(I_{B_j})$. Thus

$$\forall x, \|x\|_2 \leq k \text{ with } k \leq \sqrt{\lambda_n}\tau^*, \ d_{\text{epi}(I_{D_j})}(x) = d_{\text{epi}(I_{B_j})}(x), \tag{30}$$

and if $k \leq \sqrt{\lambda_n}\tau^*$ then $d_k(\text{epi}(I_{D_j}), \text{epi}(I_{B_j})) = 0$. Now the definition (26) of the distance d implies that

$$d(I_{D_j}, I_{B_j}) = \sum_{k \geq \left[\sqrt{\lambda_n}\tau^*\right]+1} \frac{1 \wedge d_k(\text{epi}(I_{D_j}), \text{epi}(I_{B_j}))}{2^k} \leq \sum_{k \geq \left[\sqrt{\lambda_n}\tau^*\right]+1} \frac{1}{2^k},$$

and (29) is proved.

Next, we show that

$$d(q_{n,j}, I_{D_j}) \leq \frac{2\sqrt{n}|\hat{\beta}_j^{unpen}|^{\gamma}}{\lambda_n} + 2^{-\left[\tau^* \sqrt{\lambda_n}\right]}. \tag{31}$$

22

For $j \in [p_0 + 1, p]$, $q(0) = 0$ implies that

$$q_{n,j}(u_{1:p}, u_{p+2}) = \frac{\lambda_n}{|\hat{\beta}_j^{unpen}|^\gamma} \left(\frac{u_{p+2}}{\sqrt{\lambda_n}} + \tau^* \right) q \left(\frac{u_j}{\sqrt{n} \left(\frac{u_{p+2}}{\sqrt{\lambda_n}} + \tau^* \right)} \right) + I_E \tag{32}$$

where we set $0/0 = 0$ and

$$E = \{(u_{1:p}, u_{p+2}), u_{p+2} > -\sqrt{\lambda_n}\tau^*\} \cup \{(u_{1:p}, u_{p+2}), u_{p+2} = -\sqrt{\lambda_n}\tau^* \text{ and } u_j = 0\}.$$

Consequently, $q_{n,j}(u_{1:p}, u_{p+2}) \leq I_{D_j}(u_{1:p}, u_{p+2})$. Indeed, it is clear if $(u_{1:p}, u_{p+2}) \notin D_j$. Moreover, if $(u_{1:p}, u_{p+2}) \in D_j$, $q_{n,j}(u_{1:p}, u_{p+2}) = 0$ since $q(0) = 0$. Consequently, $\text{epi}(I_{D_j}) \subset \text{epi}(q_{n,j})$, $d_{\text{epi}(I_{D_j})}(.) \geq d_{\text{epi}(q_{n,j})}(.)$ and

$$d_k(\text{epi}(q_{n,j}), \text{epi}(I_{D_j})) = \max_{\|x\| \leq k} \left(d_{\text{epi}(I_{D_j})}(x) - d_{\text{epi}(q_{n,j})}(x) \right).$$

Since $\forall t \in \mathbb{R}$, $q(t) \geq |t|$, it holds that, $\forall (t, \tau) \in \mathbb{R} \times \mathbb{R}_+^*$, $\tau q(t/\tau) \geq |t|$ and expression (32) entails

$$q_{n,j}(u_{1:p}, u_{p+2}) \geq F_{n,j}(u_{1:p}, u_{p+2}),$$

where $F_{n,j}(u_{1:p}, u_{p+2}) = \lambda_n |u_j| |\hat{\beta}_j^{unpen}|^{-\gamma} / \sqrt{n} + I_E$. Consequently, $\text{epi}(q_{n,j}) \subset \text{epi}(F_{n,j})$, $d_{\text{epi}(q_{n,j})}(.) \geq d_{\text{epi}(F_{n,j})}(.)$ and

$$d_k(\text{epi}(q_{n,j}), \text{epi}(I_{D_j})) \leq \max_{\|x\| \leq k} \left(d_{\text{epi}(I_{D_j})}(x) - d_{\text{epi}(F_{n,j})}(x) \right). \tag{33}$$

Now, $\text{epi}(F_{n,j}) = S_1 \cup S_2$ where

$$S_1 = \left\{ (u_{1:p}, u_{p+2}, t) \in \mathbb{R}^{p+2}, u_{p+2} > -\sqrt{\lambda_n}\tau^* \text{ and } \frac{\lambda_n |u_j|}{\sqrt{n} |\hat{\beta}_j^{unpen}|^\gamma} \leq t \right\},$$

and

$$S_2 = \{ (u_{1:p}, u_{p+2}, t) \in \mathbb{R}^{p+2}, u_{p+2} = -\sqrt{\lambda_n}\tau^*, u_j = 0, \text{ and } t \geq 0 \}.$$

Thus,

$$d_{\text{epi}(F_{n,j})}(x) = d_{S_1}(x) \wedge d_{S_2}(x). \tag{34}$$

Easy calculations lead to, $\forall x \in \mathbb{R}^{p+2}$,

$$d_{S_2}^2(x) = \inf_{z \in S_2} \sum_{i=1}^{p+2} (x_i - z_i)^2 = x_j^2 + (x_{p+1} + \sqrt{\lambda_n}\tau^*)^2 + x_{p+2}^2 \mathbb{1}_{x_{p+2} < 0}, \tag{35}$$

and

$$d_{S_1}^2(x) = \inf_{z \in S_2} \sum_{i=1}^{p+2} (x_i - z_i)^2 = d_{\text{epi}(f_{n,j})}^2(x_1, \cdots, x_p, x_{p+2}) + (x_{p+1} + \sqrt{\lambda_n}\tau^*)^2 \mathbb{1}_{x_{p+1} < -\sqrt{\lambda_n}\tau^*},$$

23

where $f_{n,j}(u_{1:p}) = \lambda_n |u_j| |\hat{\beta}_j^{unpen}|^{-\gamma}/\sqrt{n}$. If we consider $x \in \mathbb{R}^{p+2}$ such that $\|x\|_2 \leq k$ with $k \leq \sqrt{\lambda_n}\tau^*$, it satisfies that $x_{p+1} \geq -\sqrt{\lambda_n}\tau^*$ and thus $d_{S_1}^2(x) = d_{\mathrm{epi}(f_{n,j})}^2(x_1, \cdots, x_p, x_{p+2})$. Technical computations leads to

$$
d_{S_1}(x) = \begin{cases}
\sqrt{x_j^2 + x_{p+2}^2} & \text{if} \quad x_{p+2} \leq -\dfrac{\sqrt{n}|\hat{\beta}_j^{unpen}|^{\gamma}}{\lambda_n}|x_j|, \\[3mm]
\dfrac{|x_j| - x_{p+2}\frac{\sqrt{n}|\hat{\beta}_j^{unpen}|^{\gamma}}{\lambda_n}}{\sqrt{1 + \frac{n|\hat{\beta}_j^{unpen}|^{2\gamma}}{\lambda_n^2}}} & \text{if} \quad -\dfrac{\sqrt{n}|\hat{\beta}_j^{unpen}|^{\gamma}}{\lambda_n}|x_j| < x_{p+2} \leq \dfrac{\lambda_n}{\sqrt{n}|\hat{\beta}_j^{unpen}|^{\gamma}}|x_j|, \\[3mm]
0 & \text{if} \quad x_{p+2} > \dfrac{\lambda_n}{\sqrt{n}|\hat{\beta}_j^{unpen}|^{\gamma}}|x_j|.
\end{cases}
\tag{36}
$$

Using explicit expressions (36) and (35), we can show that for any $x \in \mathbb{R}^{p+2}$ such that $\|x\|_2 \leq k$ with $k \leq \sqrt{\lambda_n}\tau^*$,

$$
d_{S_1}(x) \leq d_{S_2}(x).
\tag{37}
$$

Gathering (37) with (34), for any $x \in \mathbb{R}^{p+2}$ such that $\|x\|_2 \leq k$ with $k \leq \sqrt{\lambda_n}\tau^*$,

$$
d_{\mathrm{epi}(F_{n,j})}(x) = d_{S_1}(x) = d_{\mathrm{epi}(f_{n,j})}(x_1, \cdots, x_p, x_{p+2}).
\tag{38}
$$

Combining (33), (38) and (30), if $k \leq \sqrt{\lambda_n}\tau^*$, we obtain

$$
d_k(\mathrm{epi}\,(q_{n,j})\,, \mathrm{epi}\,(I_{D_j})) \leq \max_{\|x\| \leq k} \left(d_{\mathrm{epi}(I_{B_j})}(x_1, \cdots, x_p, x_{p+2}) - d_{\mathrm{epi}(f_{n,j})}(x_1, \cdots, x_p, x_{p+2}) \right).
$$

The involved objective function does not depend on x_{p+1}. Moreover, using the form of the constraints, if $k \leq \sqrt{\lambda_n}\tau^*$, we get

$$
d_k(\mathrm{epi}\,(q_{n,j})\,, \mathrm{epi}\,(I_{D_j})) \leq \max_{x_1^2 + \cdots + x_p^2 + x_{p+2}^2 \leq k^2} \left(d_{\mathrm{epi}(I_{A_j})}(x_1, \cdots, x_p, x_{p+2}) - d_{\mathrm{epi}(f_{n,j})}(x_1, \cdots, x_p, x_{p+2}) \right).
$$

Moreover, since $\forall u_{1:p} \in \mathbb{R}^p$, $I_{A_j}(u_{1:p}) \geq f_{n,j}(u_{1:p})$, if $k \leq \sqrt{\lambda_n}\tau^*$,

$$
d_k(\mathrm{epi}\,(q_{n,j})\,, \mathrm{epi}\,(I_{D_j})) \leq d_k(\mathrm{epi}\,(f_{n,j})\,, \mathrm{epi}\,(I_{A_j})),
$$

and technical computations leads to

$$
d_k(\mathrm{epi}\,(f_{n,j})\,, \mathrm{epi}\,(I_{A_j})) = \frac{k\sqrt{n}|\hat{\beta}_j^{unpen}|^{\gamma}}{\lambda_n\sqrt{1 + \frac{n|\hat{\beta}_j^{unpen}|^{2\gamma}}{\lambda_n^2}}}.
$$

Finally, using the definition (26), we have

$$
d\left(q_{n,j}, I_{D_j}\right) \leq \sum_{k \leq \left[\sqrt{\lambda_n}\tau^*\right]} \frac{d_k(\mathrm{epi}\,(q_{n,j})\,, \mathrm{epi}\,(I_{D_j}))}{2^k} + \sum_{k \geq \left[\sqrt{\lambda_n}\tau^*\right] + 1} \frac{1}{2^k}.
$$

Gathering this inequality with the previous one and the fact that $\sum_{k \geq 1} \frac{k}{2^k} \leq 2$, (31) is proved. Using equation (28) with (29) and (31), the bound involved in Lemma 3 holds. ■

24

Acknowledgements

References

[1] A. Antoniadis and J. Fan. Regularization of Wavelet Approximations. *Journal of the American Statistical Association*, 96:939–967, 2001.

[2] M. A. Arcones. Weak convergence of convex stochastic processes. *Stat. Probab. Lett.*, 37(2):171–182, 1998.

[3] H. D. Bondell and B. J. Reich. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics*, 64(1):115–123, 2008.

[4] M. El Anbari and A. Mkhadri. Penalized regression combining the L1 norm and a correlation based penalty. Research Report RR-6746, INRIA, 2008.

[5] 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.

[6] S. Ghosh. Adaptive elastic net: An improvement of elastic net to achieve oracle properties. *Tech. rep., Department of Mathematical Sciences, Indiana University-Purdue University, Indianapolis.*, 2007.

[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.

[9] T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning*. Springer, corrected edition, July 2003.

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

[11] A. Hoerl and R. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67, 1970.

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

[13] K. Knight. Epi-convergence in distribution and stochastic equi-semicontinuity. In *Corpus-based work*, pages 33–50, 1997.

[14] K. Knight. and W. Fu. Asymptotics for Lasso-type estimators In *Ann. Stat.*, pages 1356–1378, 2000.

[15] S. Lambert-Lacroix and L. Zwald. Robust regression through the Huber's criterion and adaptive lasso penalty. *Electronic Journal of Statistics*, 5:1015–1053, 2011.

[16] 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.

[17] 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.

[18] N. Meinshausen and P. Buhlmann. High-dimensional graphs and variable selection with the Lasso. *Ann. Stat.*, 34(3):1436–1462, 2006.

[19] A. B. Owen. A robust hybrid of lasso and ridge regression. Technical report, 2006.

[20] R. Rockafellar. *Convex analysis.* Princeton Landmarks in Mathematics. Princeton, NJ: Princeton University Press. , 1970.

[21] 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.

[22] G. Schwarz. Estimating the dimension of a model. *Ann. Stat.*, 6:461–464, 1978.

[23] J. F. Sturm. Using SeDuMi 1. 02, a MATLAB toolbox for optimization over symmetric cones. 1999.

[24] T. A. Stamey, T. A., Kabalin, J. N., McNeal, J. E., Johnstone, I. M., Freiha, F., Redwine, E. A. and N. Yang. Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate: II. radical prostatectomy treated patients. *Journal of Urology.*, 141(5):1076–1083, 1989.

[25] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1996.

[26] A. Van der Vaart. *Asymptotic statistics.* Cambridge Series in Statistical and Probabilistic Mathematics, 3. Cambridge, 1998.

[27] 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.

[28] H. Wang and C. Leng. Unified Lasso Estimation via Least Squares Approximation. *JASA*, 102:1039–1048, 2007.

[29] 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.

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

[31] M. Yuan, M. Yuan, Y. Lin, and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68:49–67, 2006.

[32] P. Zhao, G. Rocha, and B. Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *Vol.*, (arXiv:0909.0411. IMS-AOS-AOS584), Sep 2009. Comments: Published in at http://dx.doi.org/10.1214/07-AOS584 the Annals of Statistics (http://www.imstat.org/aos/) by the Institute of Mathematical Statistics (http://www.imstat.org).

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

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

[35] H. Zou and T. Hastie. Regularization and variable selection via the Elastic Net. *Journal of the Royal Statistical Society B*, 67(2):301–320, 2005.

[36] H. Zou and H. H. Zhang. On the adaptive elastic net with a diverging number of parameters. *Ann. Stat.*, 37(4):1733–1751, 2009.

Tables and Figures

Table 1: Selection model ability on Model 1 based on 100 replications.

	C	O	U	Z	CZ	TZ	CNZ	TNZ
Least square criterion, $n = 100$								
ad-lasso	0	0	100	35.38	24.32	25	3.94	15
ad-en	0	0	100	27.72	19.69	25	6.97	15
ad-Berhu	0	51	49	6.50	5.79	25	14.27	15
Huber's criterion, $n = 100$								
ad-lasso	0	3	97	20.9	14.08	25	8.18	15
ad-en	0	3	97	26.99	18.81	25	6.82	15
ad-Berhu	0	33	67	9.89	8.57	25	13.68	15
Least square criterion, $n = 200$								
ad-lasso	0	0	100	35.15	24.59	25	4.44	15
ad-en	0	2	98	27.71	20.52	25	7.81	15
ad-Berhu	0	42	58	18.36	17.31	25	13.95	15
Huber's criterion, $n = 200$								
ad-lasso	0	7	93	29.52	19.85	25	5.33	15
ad-en	0	1	99	30.41	21.08	25	5.67	15
ad-Berhu	0	48	52	16.77	15.86	25	14.09	15
Least square criterion, $n = 400$								
ad-lasso	0	0	100	34.75	24.74	25	4.99	15
ad-en	0	0	100	27.06	21.17	25	9.11	15
ad-Berhu	1	32	67	21.46	19.29	25	12.83	15
Huber's criterion, $n = 400$								
ad-lasso	0	9	91	29.67	20.3	25	5.63	15
ad-en	0	0	100	30.01	21.59	25	6.58	15
ad-Berhu	0	35	65	19.47	18.42	25	13.95	15
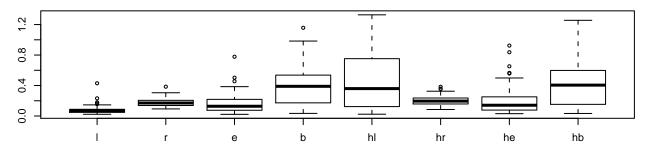
Table 2: Selection model ability on model 2 based on 100 replications.

	C	O	U	Z	CZ	TZ	CNZ	TNZ
Least square criterion, $n = 100$								
ad-lasso	0	0	100	35.03	24.18	25	4.15	15
ad-en	0	0	100	28.66	20.46	25	6.80	15
ad-Berhu	0	47	53	8.02	7.14	25	14.12	15
Huber's criterion, $n = 100$								
ad-lasso	0	22	78	10.79	7.38	25	11.59	15
ad-en	1	15	84	20.96	16.09	25	10.13	15
ad-Berhu	5	24	71	24.36	23.12	25	13.76	15
Least square criterion, $n = 200$								
ad-lasso	0	0	100	35.25	24.61	25	4.36	15
ad-en	0	0	100	28.21	20.68	25	7.47	15
ad-Berhu	0	49	51	17.17	16.59	25	14.11	15
Huber's criterion, $n = 200$								
ad-lasso	0	6	94	21.63	16.77	25	10.14	15
ad-en	1	15	84	25.38	20.31	25	9.93	15
ad-Berhu	4	11	85	25.36	23.78	25	13.42	15
Least square criterion, $n = 400$								
ad-lasso	0	0	100	35.05	24.69	25	4.64	15
ad-en	0	0	100	28.10	21.69	25	8.59	15
ad-Berhu	1	30	69	20.15	18.74	25	13.59	15
Huber's criterion, $n = 400$								
ad-lasso	0	14	86	21.30	17.76	25	11.46	15
ad-en	2	25	73	23.32	19.90	25	11.58	15
ad-Berhu	23	10	67	24.64	23.59	25	13.95	15
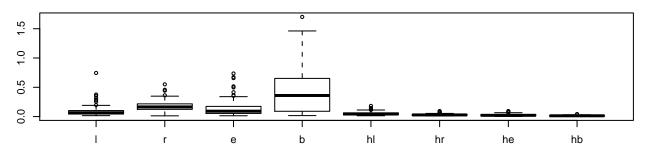
Table 3: Selection model ability on model 3 based on 100 replications.

|	C	O	U	Z	CZ	TZ	CNZ	TNZ
Least square criterion, $n = 100$								
ad-lasso	0	0	100	35.23	24.45	25	4.22	15
ad-en	0	0	100	29.11	20.97	25	6.86	15
ad-Berhu	0	53	47	6.68	6.02	25	14.34	15
Huber's criterion, $n = 100$								
ad-lasso	0	4	96	25.5	17.36	25	6.86	15
ad-en	0	0	100	29.40	20.60	25	6.20	15
ad-Berhu	0	34	66	14.34	13.24	25	13.90	15
Least square criterion, $n = 200$								
ad-lasso	0	0	100	35.32	24.74	25	4.42	15
ad-en	0	0	100	28.58	20.83	25	7.25	15
ad-Berhu	0	44	56	19.01	17.47	25	13.46	15
Huber's criterion, $n = 200$								
ad-lasso	0	3	97	31.95	21.85	25	4.9	15
ad-en	0	2	98	28.37	19.91	25	6.54	15
ad-Berhu	0	50	50	19.53	18.73	25	14.20	15
Least square criterion, $n = 400$								
ad-lasso	0	0	100	34.78	24.74	25	4.96	15
ad-en	0	0	100	26.54	20.61	25	9.07	15
ad-Berhu	0	22	78	20.88	19.29	25	13.41	15
Huber's criterion, $n = 400$								
ad-lasso	0	7	93	31.68	22.04	25	5.36	15
ad-en	0	3	97	26.25	19.65	25	8.4	15
ad-Berhu	2	23	75	22.14	21.02	25	13.88	15
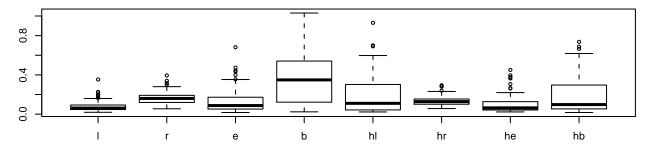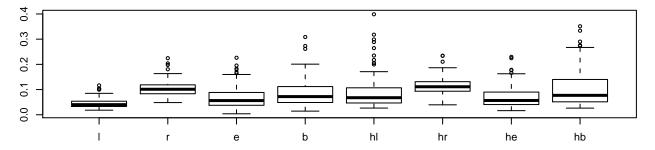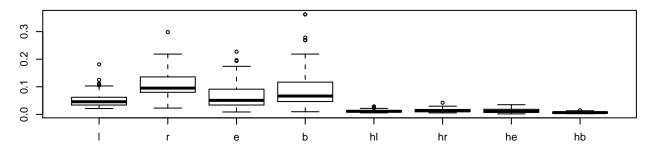
Figure 1: For $n = 100$, RPE for `ad-lasso` (l), `ridge` (r), `ad-en` (e), `ad-Berhu` (b), `Huber-ad-lasso` (hl), `Huber-ridge` (hr), `Huber-ad-en` (he), and `Huber-ad-Berhu` (hb). The boxplots are obtained without extreme values given by, for model 1 hl: 2.87; model 2 b: 2.95, hl: 2.94, he: 794.15; model 3 hl: 2.58.

31

RPE (model 1, n=200)

RPE (model 2, n=200)

RPE (model 3, n=200)
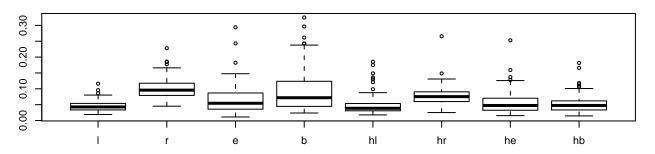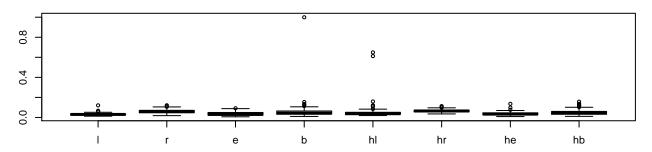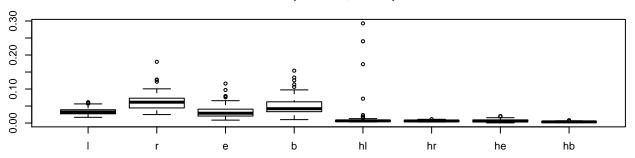
Figure 2: For $n = 200$, RPE for `ad-lasso` (l), `ridge` (r), `ad-en` (e), `ad-Berhu` (b), `Huber-ad-lasso` (hl), `Huber-ridge` (hr), `Huber-ad-en` (he), and `Huber-ad-Berhu` (hb). The boxplots are obtained without extreme values given by, for model 1 b: 2.95, hl: 2.48, 2.95,12.79, 2.86, 2.54, 2.96, 2.95; model 3 b: 2.95, 2.95, 2.95, hl: 2.95, 2.94, 2.95, 2.51, 49.03.
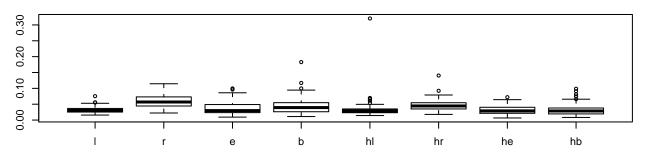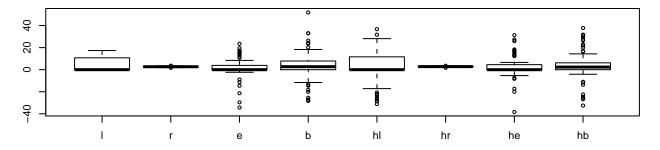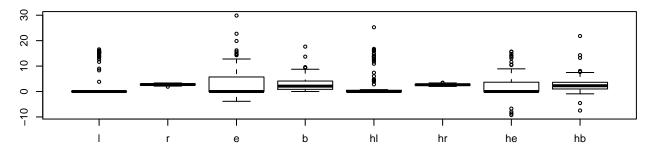
Figure 3: For $n = 400$, RPE for `ad-lasso` (l), `ridge` (r), `ad-en` (e), `ad-Berhu` (b), `Huber-ad-lasso` (hl), `Huber-ridge` (hr), `Huber-ad-en` (he), and `Huber-ad-Berhu` (hb). The boxplots are obtained without extreme values given by, for model 1 b: 2.95, 2.95, hl: 2.95, 2.49, 2.90, 2.95, 2.94, 2.95, 2.95, 2.93; model 2 b: 2.95, 2.95, 0.99; model 3 b: 2.95, 2.95, hl: 8.97, 2.95.
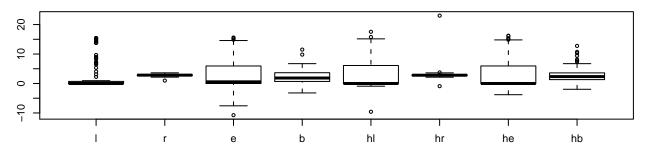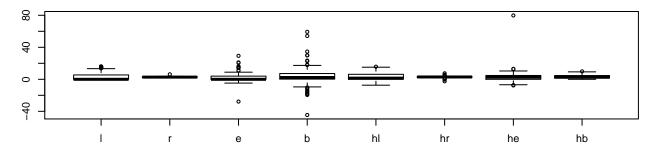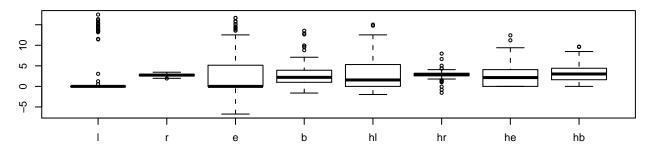
Figure 4: Model 1: estimations of first influencing coefficient (true value is equal to 3) by `ad-lasso` (l), `ridge` (r), `ad-en` (e), `ad-Berhu` (b), `Huber-ad-lasso` (hl), `Huber-ridge` (hr), `Huber-ad-en` (he), and `Huber-ad-Berhu` (hb).

34

Estimation of the first coefficient model 2, n=100

Estimation of the first coefficient model 2, n=200

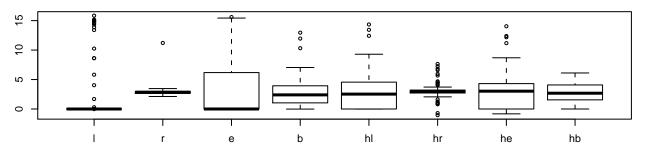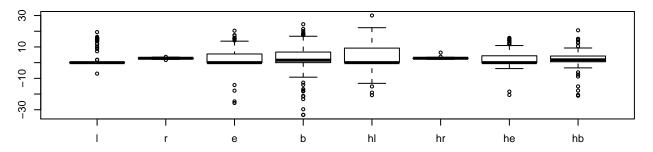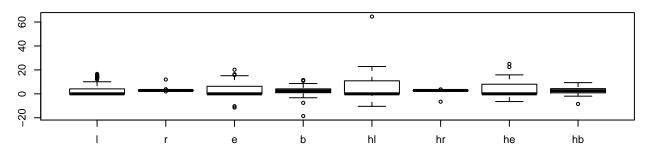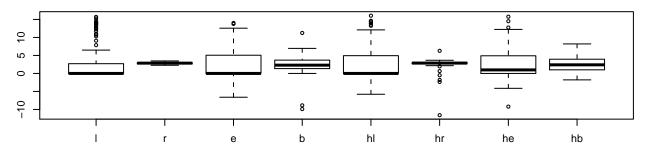**Estimation of the first coefficient model 2, n=400**

Figure 5: Model 2: estimations of first influencing coefficient (true value is equal to 3) by `ad-lasso` (l), `ridge` (r), `ad-en` (e), `ad-Berhu` (b), `Huber-ad-lasso` (hl), `Huber-ridge` (hr), `Huber-ad-en` (he), and `Huber-ad-Berhu` (hb).

Figure 6: Model 3: estimations of first influencing coefficient (true value is equal to 3) by `ad-lasso` (l), `ridge` (r), `ad-en` (e), `ad-Berhu` (b), `Huber-ad-lasso` (hl), `Huber-ridge` (hr), `Huber-ad-en` (he), and `Huber-ad-Berhu` (hb).

Table 4: Prostate cancer data: comparing methods

| Methods	mean of 100 parameters (std of the 100)	mean of 100 RPE (std of the 100)
OLS	none	0.6054(0.1397)
Least square criterion		
`ad-lasso`	$\lambda_n : 2.4177(1.7368)$	0.6357(0.1410)
`ridge`	$\lambda_n : 2.6104(2.3111)$	0.6145(0.1406)
`ad-en`	$\lambda_{1,n} : 1.1361(1.0048), \lambda_{2,n} : 2.5032(10.2605)$	0.6231(0.1351)
`ad-Berhu`	$\lambda_n : 1.9850(1.2782)$	0.6237(0.1423)
Huber's criterion		
`ad-lasso`	$\lambda_n : 26.2749(7.4369)$	0.7765(0.1879)
`ridge`	$\lambda_n : 3.7437(3.5792)$	0.6020(0.1327)
`ad-en`	$\lambda_{1,n} : 1.3885(1.5778), \lambda_{2,n} : 4.3222(14.2073)$	0.6185(0.1295)
`ad-Berhu`	$\lambda_n : 2.7456(1.9015)$	0.6322(0.1391)
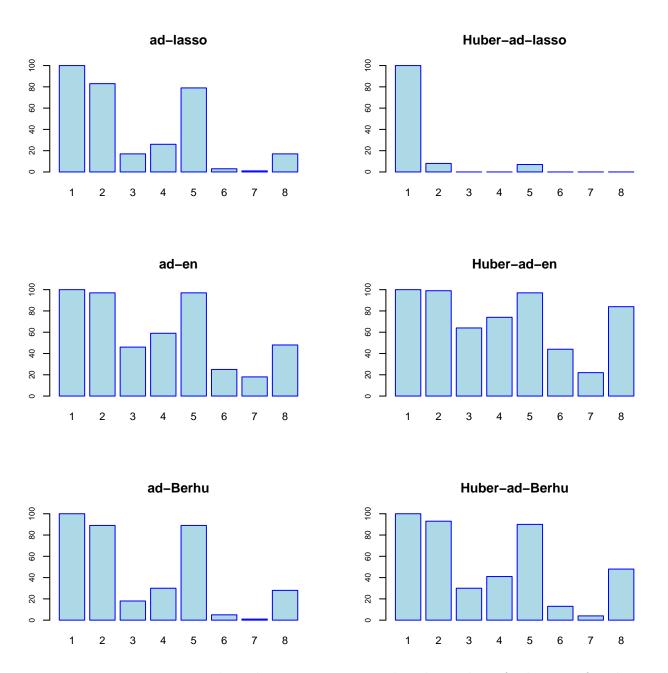
Figure 7: Prostate cancer data: histogram associated with number of selection of each variables in the re-sampling study.