The adaptive BerHu penalty in robust regression

Sophie Lambert-Lacroix, Laurent Zwald

The adaptive BerHu penalty in robust regression

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. We intend to combine Huber's loss with an adaptive reversed version as a penalty function. The purpose is twofold: first we would like to propose an estimator that is robust to data subject to heavy-tailed errors or outliers. Second we hope to overcome the variable selection problem in presence of highly correlated predictors. For instance, in this framework, the adaptive least absolute shrinkage and selection operator (lasso) is not a very satisfactory variable selection method, although it is a popular technique for simultaneous estimation and variable selection. We call this new penalty adaptive BerHu penalty. As for elastic net penalty, small coefficients contribute through their ℓ_1 norm to this penalty while larger coefficients cause it to grow quadratically (as ridge regression). We will show that the estimator associated with Huber's loss combined with adaptive BerHu penalty enjoys theoretical properties in the fixed design context. This approach is compared to existing regularization methods such as adaptive elastic net and is illustrated via simulation studies and real data.

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

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

1 Introduction

Variable selection is a significant problem in linear regression analysis. Its goal is to identify relevant predictors to improve prediction performances and to achieve an easy interpretation of the model. Here we will focus on the variable selection problem in the robust context first, where data subject to heavy-tailed errors or outliers are encountered, and second in presence of highly correlated predictors. Occurences of these two contexts are very common in practice.

When there is a group of variables among which the pairwise correlations are very high, the adaptive lasso penalty, introduced in [39], tends to select a single variable from this group. This penalty was recently combined with Huber's criterion (see [18]). The estimator associated with this procedure enjoys oracle properties (as in [39]) but behaves poorly when there are highly correlated variables. Ridge regression (ℓ_2 penalty) does not lead to variable selection but tends to share the coefficients' value among the group of correlated predictors. Moreover if there are high correlations among predictors, the prediction performance of ridge regression dominates the lasso ([29]). In order to overcome this drawback, [40] proposes a new regularization technique that combines lasso and ridge penalties. This method is called the "elastic net" (Enet). This penalty is the sum of lasso

and ridge penalties. In [9], a new version of the elastic net called adaptive elastic net (ad-Enet) which inherits some of the desirable properties of the adaptive lasso and elastic net is proposed. Its oracle properties are proven. In [23], the author proposes to use a reversed version of Huber's loss (called BerHu) as a penalty function. Let us recall that the Huber loss (see [15]) 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 through their ℓ_1 norm to this penalty while 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 [23], a way is provided to optimize objective function constituted of both the Huber loss and the BerHu penalty. Nevertheless, nothing is shown about statistical properties. Since [23] presents the BerHu penalty in its non-adaptive form, the corresponding estimator does not generally satisfy the oracle property.

In this paper we introduce an adaptive concomitant BerHu penalty. We use it with Huber's loss in order to take into account data subject to heavy-tailed errors or outliers. Indeed, in this case, the Ordinary Least Square (OLS) estimator is considered to be not efficient. Nevertheless, our method is designed to maintain good properties in the Gaussian noise case. Moreover the adaptive concomitant BerHu penalty encourages a grouping effect in the following sense: the concomitant BerHu penalty implicitly creates one group with the largest coefficients. This group is penalized in a ℓ_2 way like the grouped lasso of [35] to avoid removing any of the largest coefficients. The smallest coefficients are treated individually by an ℓ_1-penalty.

In addition, contrary to [23], we show that the estimator associated with Huber's loss combined with adaptive concomitant BerHu penalty enjoys some "oracle" properties in the fixed design context as already satisfied by the adaptive lasso penalty (see [39, 18]). In addition to [39, 18], the adaptive concomitant BerHu penalty is designed to keep all the largest coefficients whatever the correlation structure. It should be noted that the needed assumptions to get oracle property for BerHu penalty are the same as in the lasso case. Contrary to elastic net (see [9]) no supplementary assumptions are needed. These asymptotic results are valid in the case where p is fixed and n goes to infinity. Recently [41] obtained oracle properties for ad-Enet when p diverges for the least squares loss. These proof techniques are reused by [6] for a more general penalty. These results are based on a sparsity inequality (see for instance theorem 3.1 in [41]). Unfortunately we can not use similar proof techniques for adaptive BerHu penalty. Extending this result to the case of p diverging is not so straightforward and is left for future work. Extensive simulation studies show that in presence of highly correlated variables, the lasso and the elastic net procedures tend to underfit, i.e. suppress some influencing variables whereas adaptive concomitant BerHu penalty tends to identify all influencing variables but keeps some non-influencing variables.

The article is subsequently organized as follows: in Section 2, we introduce the adaptive concomitant BerHu penalty, in Section 3, we give its statistical properties. Section 4 is devoted to simulations and illustration over real data. All technical proofs are relegated to the Appendix.

2 The BerHu penalty

2.1 The adaptive BerHu

We focus on the following model

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

where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$ denote centered covariables, α^* and $\beta^* = (\beta_1^*, \ldots, \beta_p^*)^T$ the regression coefficients. The constant σ is positive and ϵ_i are independent and identically-distributed random errors with mean 0 and variance 1, when it exists. Without loss of generality, we put $\beta_j^* \neq 0$, for $j \leq p_0$ and $\beta_j^* = 0$, for $j > p_0$ for some $p_0 \geq 0$. Let $\beta_{\mathcal{A}}$ be the vector given by the nonzero coordinates of β ($\mathcal{A} = [1, \ldots, p_0]$.)

The lasso has already been extensively critized (see e.g. [40]). In this paper, we focused 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 lists attempts to solve this problem. To begin with, grouped lasso procedures were proposed first in [35] 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 $\lambda_n > 0$ is the tuning parameter and $(\beta)_j$ is the coordinate block corresponding to the j-th group. Consequently, sparsity is encouraged at group level (see also [36] and [12] page 91 for further references). In our framework it is difficult to use the group lasso approach since there is no obvious way of choosing the groups a priori. Next, [40] proposed the Elastic Net. The naive Enet is obtained by minimizing in α and β:

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

and Enet 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, a ridge penalty reduces the variance of the estimates by imposing a small squared norm of all the coefficients. Following [23], we focused on a penalty that acts separately on small and large coefficients: only the largest coefficients have quadratic penalty. 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 a positive real. As Huber criterion, the BerHu function needs to be scaled. The penalty can be precisely defined by

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

where τ is a scale parameter to be determined. This scale parameter was introduced by [23] and plays a similar role to s in concomitant scale Huber's Criterion defined in (5) that was introduced by [15]. When this parameter is set to a value (for instance $\tau = 1$), the procedure is not scale invariant. So [23] proposes to replace the penalty term by

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

This makes it possible to avoid algorithms that alternate between estimating α and β for fixed τ and reciprocally. The reference [23] is the only existing literature on this kind of penalty as far as we know. However, [23] does not provide any statistical analysis of the resulting estimator.

Fan and Li [7] showed that the lasso method leads to estimators that may suffer an appreciable bias. Furthermore they conjectured that the oracle properties do not hold for the lasso. Hence Zou [39] proposes to consider the following modified lasso criterion, called adaptive lasso,

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

where $\hat{\mathbf{w}}^{adl} = (\hat{w}_1^{adl}, \ldots, \hat{w}_p^{adl})$ is a known weight vector. This modification makes it possible to produce sparse solutions more effectively than lasso. Indeed, Zou [39] 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 Enet penalty (see [9]). In the sequel, we will call it ad-Enet.

Here we propose to make the BerHu penalty adaptive ([23] only proposes the penalty $\text{pen}(\cdot)$ in a non-adaptive form). We thus consider the following penalty: $\text{pen}_{adb}(\beta) = \min_{\tau \geq 0} P^{adb}(\beta, \tau)$ 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 weight vector. We will see in Section 3 that the resulting estimator enjoys some "oracle " properties. Let us note that [23] introduced the BerHu penalty in its non-adaptive form and in the context of robust regression only. Moreover nothing is mentioned about asymptotic feature.

In general, the (adaptive) Berhu penalty behaves like the lasso on the smallest coefficients and does not delete the largest ones, whatever the correlation structure. This is expected of a correct model selection procedure. This interpretation relies on the following calculation when β is fixed (its proof is postponed in Section 5.2). Let us sort the absolute values of the coordinates of β:

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

Theorem 1. *Let $k(\beta)$ denote the number of non-zeros coefficients of β. Then*

$$pen_{adb}(\beta) = 2 \sqrt{\sum_{j=1}^{p} \frac{1}{\hat{w}_j} + \frac{L}{2} \sum_{j=1}^{q(\beta)-1} \hat{w}_{(j)}} \sqrt{\frac{1}{2L} \sum_{j=1}^{q(\beta)-1} \beta_{(j)}^2 \hat{w}_{(j)}} + \sum_{j=q(\beta)}^{k(\beta)} |\beta_{(j)}| \hat{w}_{(j)}, \tag{4}$$

where $q(\beta)$ is the unique integer between 2 and $k(\beta)+1$ such that the following inequalities hold

$$\frac{|\beta_{(q(\beta))}|}{L} \leq \sqrt{\frac{\sum_{j=1}^{q(\beta)-1} \beta_{(j)}^2 \hat{w}_{(j)}}{2L\sum_{j=1}^{p}\frac{1}{\hat{w}_j} + L^2\sum_{j=1}^{q(\beta)-1}\hat{w}_{(j)}}} \leq \frac{|\beta_{(q(\beta)-1)}|}{L}.$$

We can note here the spirit of the concomitant BerHu penalty: it implicitly creates one group with the largest coefficients (see first term (4)). This group is penalized in a ℓ_2 way like the grouped lasso of [35] to avoid removing any 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 second term (4)). Consequently, whatever the structure of the correlation matrix, the concomitant BerHu penalty tends to keep all the largest coefficients and to delete the smallest ones. [3] also groups together the largest coefficients but with the squared ℓ_2-norm instead of a ℓ_2-norm as in our case. It should be noted that in [3], nothing is mentioned about its asymptotic features.

Let us note that elastic net (adaptive version or not) is well adapted to the presence of highly-correlated predictors but not the lasso (adaptive version or not).

2.2 Robust estimation

Let us recall the definition of the loss function based on Huber's criterion introduced by [15]: for any positive real M,

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

This function is quadratic in small values of z (of absolute value less than M) and grows linearly for large values of z. The concomitant scale Huber loss is defined by

$$\mathcal{L}_{\mathcal{H}}(\alpha,\beta,s) = \begin{cases} ns + \sum_{i=1}^{n} \mathcal{H}_M\left(\frac{y_i-\alpha-\mathbf{x}_i^T\beta}{s}\right) s & \text{if} \quad s>0, \\ 2M\sum_{i=1}^{n}|y_i-\alpha-\mathbf{x}_i^T\beta| & \text{if} \quad s=0, \\ +\infty & \text{if} \quad s<0, \end{cases} \tag{5}$$

which is to minimize with respect to $s \geq 0$, α and β. The loss function involving a concomitant estimation of the scale and location parameter was first proposed by Huber ([15]). The parameter s is a scale parameter for the distribution. That is if each y_i is replaced by cy_i for $c>0$ then an estimate $\hat{s}$ should be replaced by $c\hat{s}$. When setting this parameter, the procedure is not scale invariant. The concomitant scale Huber criterion is jointly convex as a function of β and s. This removes the need for backfitting algorithms to estimate α and β; the parameter s is only a nuisance parameter. The introduction of these parameters is also crucial to have the asymptotic performance

To achieve a robust scale invariant Lasso type procedure, [18] proposes to minimize simultaneously over s, α and β the function

$$\mathcal{L}_{\mathcal{H}}(\alpha,\beta,s) + \lambda_n \sum_{j=1}^{p} \hat{w}_j^{adh}|\beta_j|. \tag{6}$$

where $\hat{\mathbf{w}}^{adh} = (\hat{w}_1^{adh}, \ldots, \hat{w}_p^{adh})$ is a known weight vector. We propose here to use the concomitant Huber estimation with the adaptive BerHu penalty:

$$Q^{\mathcal{H}adb}(\alpha, \beta, s, \tau) = \mathcal{L}_{\mathcal{H}}(\alpha, \beta, s) + \lambda_n P^{adb}(\beta, \tau). \tag{7}$$

This convex 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 [18], the largest coordinates of β are quadratically penalized. We consider two types of procedures: one with $s = 0$ in the equation (7) that leads to ℓ_1-loss and one with $s > 0$ that leads to Huber's loss. Let us observe that there is no closer form of the minimizers of (7). For numerical optimization, we use `CVX`, a set of matlab functions (see Appendix 5.1).

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 [39, 18]) $\hat{w}_j^{adl} = |\hat{\beta}_j^{unpen}|^{-\gamma}$, $j = 1, \ldots, p$, 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 need only 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 a value for M. As in [15], we fix $M = 1.345$. For BerHu's penalty we fix as in [23], $L = M$. Let us note that there is no justification for this. 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 criteria. When using least squares criterion we consider the classical BIC criterions ([26]), That is recommended to select λ_n minimizing

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

over λ_n, where k_{λ_n} is the number of non-zero coefficients of the estimator (see [32] and [34]) and corresponds to the model dimension. When using Huber's criterion, we consider the BIC-type procedure introduced in [18]: we select λ_n by minimizing

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

over λ_n. As previously, k_{λ_n} denotes the number of non-zero coefficients of $\widehat{\beta}_{\lambda_n}$. For ℓ_1-loss, it is recommended to select λ_n minimizing

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

3 Theoretical 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 in the fixed design context with p is fixed and n goes to infinity. We have the same property by replacing Huber's loss by least squares loss. When necessary, we give the difference (for example for the assumptions) between the two loss functions.

Let $\mathbf{X}$ denote the design matrix, i.e. the $n \times p$ matrix the i^{th} rows of which is $\mathbf{x}_i^T$. We will use 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 $p_0 \times p_0$ block of V, corresponding to the covariables associated with non zero coefficients.

Assumption **(D1)** and **(D2)** are classical. It can be seen as a "compactness assumption": for instance, 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 [39]) while considering Huber's criterion we need both **(D1)** and **(D2)** (see [18]).

Let us denote by ϵ a variable with the same law as ϵ_i, $i = 1, \ldots, n$. As in [18], 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

$$\tau^* = \underset{\tau>0}{\operatorname{argmin}}\, \tau\left(\sum_{j=1}^{p}|\beta_j^*|^\gamma + \sum_{j=1}^{p_0}\frac{1}{|\beta_j^*|^\gamma}\mathcal{B}_L\left(\frac{\beta_j^*}{\tau}\right)\right). \tag{8}$$

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

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

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

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

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

The assumptions **(N1)** or **(N2)** are usual in a robust context (see [15, 33]). The classical laws used to model data with outliers such as double exponential and Cauchy distributions satisfy these constraints. It should be noted that there is no integrability condition assumed on the errors ϵ. These three hypotheses are assumed for the Huber's loss. For the penalized least squared estimators (e.g.

[17] and [39]) we assume that ϵ_i are independent identically distributed random variables with mean 0 and have a finite variance.

Let $(\hat{\alpha}^{\mathcal{H}adb}, \hat{\beta}^{\mathcal{H}adb}, \hat{s}^{\mathcal{H}adb}, \hat{\tau}^{\mathcal{H}adb})$ be defined as the minimizer of $Q^{\mathcal{H}adb}(\cdot)$ where $\hat{w}_j^{adb} = 1/|\hat{\beta}_j^{unpen}|^\gamma$ with $\hat{\beta}^{unpen}$ 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 minimizer $(\hat{\alpha}^{\mathcal{H}adb}, \hat{\beta}^{\mathcal{H}adb}, \hat{s}^{\mathcal{H}adb}, \hat{\tau}^{\mathcal{H}adb})$ exists since the criterion $Q^{\mathcal{H}adb}(\cdot)$ is a convex and coercive function.

The following theorem shows that, with a proper choice of λ_n, the proposed estimator keeps the asymptotic properties already satisfied by adaptive lasso penalty as proved in [18]. It should be noted that the needed assumptions to get this property are the same as in [18]. No supplementary hypotheses are required to quadratically penalize large coefficients. It is not the case for Elastic Net as shown by [9] for which some assumptions must be added to move from the lasso penalty to the Elastic Net penalty. Moreover, contrary to [18], the proof requires more technicalities due to the use of the BerHu concomitant penalty (instead of the lasso penalty). More precisely, as in [39], the proof relies on the epi-convergence of the objective function. However, in [39] and [18], the proof of the epi-convergence is not provided. The proposed proof (postponed in Appendix 5.3) provides the necessary theoretical computations to obtain epi-convergence. This leads to a self-contained proof including both the adaptive BerHu concomitant penalty of the present work and the adaptive lasso penalty of [39].

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

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

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

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

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

and where

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

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

and where $\mathcal{H}'_M(\cdot)$ *is the derivative function of* $\mathcal{H}_M(\cdot)$.

In order to have this result, we have to add the constraint λ_n/n^γ tends to 0, as $n \to \infty$, compared to the work of [39]. This assumption implies, with $\lambda_n n^{(\gamma-1)/2} \to \infty$, that $\gamma > 1/3$. This was not the

case for [39]; currently the case $0 < \gamma \leq 1/3$ remains an open problem. To determine the weights $\hat{w}_j^{adb}$, we can use the unpenalized Huber estimator $\hat{\beta}^{unpen}$ that is a root-n-consistent estimator of β^*. When $M = +\infty$ (least squares loss), 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 $\sigma^2 V_{1,1}^{-1}$ as in theorem 2 of [39]. Otherwise we find the asymptotic variance of the unpenalized Huber estimator associated with the sub-model induced by $\beta^*_{\mathcal{A}}$. In that sense the procedure satisfies some "oracle" property. Let us remark that [20] proposed a generalized Huber criterion with bridge penalty for variable selection in linear regression when they are heavy-tailed errors or outliers in the response. This generalized Huber criterion is no longer convex. Our theoretical results do not remain valid for this loss function since they are based on convex minimization properties. Note that the theorem is not valid in the case $s = 0$. In this case the loss ℓ_1 is not differentiable and we can not use our proof technique based on the use of Taylor expansions.

Let us remark that our proof can be generalized to the random design setting using other technical arguments. One way to proceed would be to use the results of [33] that give asymptotic properties in random design. Another way would be to use the results of [38] that consider asymptotic properties of adaptive penalized quantile regression in random design. They underline the use of Bernstein's inequality.

4 Some numerical experiments

In this section, the estimators $\hat{\beta}^{\mathcal{H}adb}$ (with $s = 0$ and $s > 0$) are compared with four estimators of the literature: the adaptive lasso, ridge, adaptive elastic net and the adaptive lasso with Huber loss. We call these methods respectively `ad-lasso`, `ridge`, `ad-Enet`, `Huber-ad-lasso`, whereas our methods ($\hat{\beta}^{\mathcal{H}adb}$) are called `l1-loss-ad-BerHu` for $s = 0$ and `Huber-ad-BerHu` for $s > 0$. The adaptive weights are obtained from the corresponding unpenalized estimator and $\gamma = 1$.

4.1 Simulation Results

Our aim here 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

As in [40] (*example 4*), we use models that involve groups of highly correlated variables by block to compare the performances of the algorithms. Let us remark that [40] considered a model without intercept. Here we add an intercept. We now recall the definition of this model in a different way. Our formulation allows us to clearly identify the groups of influencing correlated variables. The models 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. The last block is the identity matrix of size 25. The vector of true coefficients β^* is defined as follows: the first 15 coordinates are equal to 3 and the last 25 coefficients are 0. This means that, in this model, only the first 15 variables influence the response. The 25 following are pure noise. Amongst the 15 influencing variables, there are three groups of highly correlated variables: these groups are composed of the first five variables, the next five and the last five. 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-variable model, gaussian noise.* Here 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 is exactly example 4 of [40].

- Model 2: *block-variable model, mixture of gaussians.* Here the variables $\epsilon_1, \cdots, \epsilon_n$ are an independent mixture of gaussians. Indeed, 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 1. The common value is $\text{std}(\sigma\epsilon) = 3.1009\sqrt{1 + 0.1(225 - 1)} = 15$.

- Model 3: *block-variable model, double-exponential noise.* $\epsilon = D/\sqrt{\text{var}(D)}$ and $\sigma = 212$. The distribution of D is a standard double exponential i.e. its density is $x \in \mathbb{R} \to e^{-|x|}/2$ and $\text{var}(D) = 2$.

These three models create a grouped variable situation. They can be divided into two types. The first type contains light tailed error models (Model 1) whereas the second type is composed of heavy tailed error models (Model 2 and Model 3). Model 1 makes it possible to quantify the deterioration of the performances of the robust algorithms in the absence of outliers. With regard to 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 (as in [18]). 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 so as to fit 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 a fixed design framework, the training and test designs 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 [39] are computed (see also [18] 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 in [33], [29] and [7] in Tables 1, 2 and 3. Ridge penalty procedures are not reported since they do not constitute variable selection procedures. To provide the indicators defined below, a coefficient is considered to be zero if its absolute value is strictly less than 10^{-5} (i.e. its first five decimals vanish). In all cases, amongst the 100 obtained estimators, the first column (C) has the number of well chosen models, i.e. the cases where first 15 coordinates of $\hat{\beta}$ are non-zeros and the last 25 coefficients are zeros. We consider other measurements to go further in our model selection ability analysis. The first (in the second column (O)) represents the number of overfitting models (i.e. those selecting all the non-zero 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 once. Columns (O) 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) recalls the number of theoretical zeros. Model selection abilities are closely related to the accuracy of estimations of the coefficients. This fact is illustrated by boxplots of the coefficient 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 in each of the 100 training sets as described in Section 2.3. The same grid has always been used. It is composed of 100 log-linearly spaced points between 0 and 1400 for BerHu and 200 log-linearly spaced points between 0 and 10 000 for the lasso. For Huber's loss, the simulation studies report the performances obtained with $M = 1.345$. This value was recommended by Huber in [15]. For adaptive BerHu penalty, we report the performances obtained with $L = M = 1.345$ Let us remark that it is possible to choose 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 any improvement to make it data adaptive. For ridge-type procedures the hyperparameter is chosen as usual by 5-fold cross-validation for each of the 100 training sets. The grid is composed of 100 points, log-linearly spaced between 0 and 1400. For adaptive Enet procedure, we use the similar protocol as in [40]: we first pick a relatively small grid of values for $\lambda_{2,n}$ over $\{0, 0.01, 0.1, 1, 10, 100\}$ and 25 log-linearly spaced points between 0 and 5000 for $\lambda_{1,n}$. Then 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. For Model 1 and 2,the lasso and Enet penalties methods lead in most cases to underfitting models (columns U). This is quite a surprising outcome for the Enet 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 penalty thus obtained is closer to differentiability than the ℓ_1-penalty. As shown in [1], if the penalty is far from differentiability, smaller coefficients are deleted. These penalties have a relatively high number of zeros (columns Z) with the number of correct zeros (columns CZ) very close to the true value (columns TZ). However, the number of incorrect zeros (columns Z$-$CZ) is almost the same whatever n and whatever the model: the lasso (resp. Enet) does not identify approximately 10 (resp. 8) influencing variables. For Model 3 the trend reverses. In particular these two methods tend to underestimate the number

of zeros and are very bad at recognizing correct zeros. In all cases these methods almost never identify the right model. The Behru penalty leads to some compromise between over and under fitting. Its number of incorrect zeros is approximatively 1 even in Model 3 : Berhu penalty deletes fewer non influencing variables than the lasso and Enet but keeps more non influencing variables. We point out that contrary to Enet and lasso type methods, in Model 2 the BerHu penalty allows us to identify the right model a reasonable number of times.

This behavior occurs on the quality of estimation of the non zero coefficients (see Figures 4, 5 and 6)). In these figures, we report, for $n = 200$, boxplots associated to the differences between $\hat{\beta}_1$ and β_1. Let us note that the conclusions for the other influencing coefficients remain the same. The ridge method is given here as a reference since it is known to lead to good performances in presence of high correlation between the covariables. We observe that the BerHu penalty leads to good performance in terms of bias, as the ridge method, with higher variability for Model 1 and 2. The bias and sometimes the variability are very high for the other methods due to their tendency to underfit. Most of the time (for Model 1 and 2), the median of the differences between $\hat{\beta}_1$ and β_1 is equal to $-3 = -\beta_1$ (that is $\hat{\beta}_1 = 0$). For model 3, `ad-lasso` has a very high variability in comparison with other methods.

Figures 1, 2 and 3 provide the boxplots associated with RPE. Let us recall that the Model 1 (without outliers) was introduced to quantify the potential fall of performances of Huber type methods. For n sufficiently large ($n \geq 200$), all the methods behave similarly. It should be noted that this model includes 40 variables, consequently $n = 100$ is very small. One can note that `Huber-ad-Berhu` has better performances than `l1-loss-ad-Berhu` in the Gaussian case (Model 1). Concerning the Model 2 (gaussian mixtures) and Model 3 (double-exponential noise), the RPE of Huber type methods are significantly better than those of the other methods as well as in term of bias and variance. We observe that `Huber-ad-lasso` provides several extreme values du to numerical instabilities and is often more variable.

4.1.4 Numerical illustration of Theorem 2 (consistency of variable selection)

Previous simulations illustrated the behavior of the methods in realistic cases that is to say for n not very large. In this subsection we aim at illustrating the consistency of the variable selection part of Theorem 2 when n becomes large. We introduce a new model, denoted by Model 4, which is a variant of Model 2 with fewer covariates: the number of covariates (without the intercept) is equal to 8, 2 blocks of 2 active variables (instead of 3 blocks of 5 active variables) and one block of 4 inactive variables (instead of 15 inactive variables). The simulations are conducted as previously, using $n = 100$, 200, 400, 1000 and 2000. Table 4 gives, for each value of n, C, the number of well chosen models from the 100 samples and the mean and standard deviation of the RPE. We observe, as stated in Theorem 2, that in the simulations C tends to 100 when n becomes large.

4.2 Prostate cancer data example

This data set comes from a prostate cancer study (see [28]) and was previoulsly analyzed in the elastic net paper by [40, 9]. There are eight clinical covariates, namely: logarithm of the cancer volume (lcavol), logarithm of the prostate weight (lweight), age, 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 variable is the logarithm of prostate-specific antigen (lpsa). The predictors are named $1, \ldots, 8$ in results. OLS and the previous methods were applied to these data.

In [40], the data were divided into two parts: a training set with 67 observations and a test set with 30 observations while in [9], they randomly divided the original data set into training and testing sets containing 60 and 37 observations respectively. To fairly compare the methods we propose to perform a resampling study: we randomly divided 100 times the original data set into training and testing sets 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 RPEs on the 100 resampling testing sets (see Table 5). Contrary to what had been observed in [40, 9], our resampling study does not allow us to claim that one method emerges in terms of RPE: almost all these methods have similar RPEs. We can only say that perhaps `Huber-ad-lasso` is slightly less good. It should be noted that we observe a great variability in the choice of $\lambda_{2,n}$ for the ad-Enet (see first column of Table 5). This is also the case for `Huber-ad-lasso`. On the contrary, the choice of λ_n for Behru type procedures is more stable (it is comparable to the stability of ridge). Figure 7 shows (except for OLS and ridge procedures) the histogram associated with the selected variables. We see that `Huber-ad-Berhu` leads to good models in terms of sparsity in comparison with `Enet`. We observe that `Huber-ad-Berhu` procedure is a compromise between lasso type methods which select few variables and ad-Enet type methods which select 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 [10, 11]. `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 [27]).

Let us now recall a well-known fact of convex analysis: the Huber function is the Moreau-Yosida regularization of the absolute value function ([13, 24, 25]). 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 [23] 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 [23] 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

Let us fix some $\beta \neq 0$. The following holds $\mathrm{pen}_{adb}(\beta) = \min_{\tau \geq 0} f(\tau)$ where $f(\tau) = P^{adb}(\beta, \tau)$. The function f is clearly convex continuous on $\mathbb{R}_+^*$. The set $\mathbb{R}_+$ can be cut up into intervals such that, on each interval the function f get the form $a/\tau + b\tau + c$ for various constants a, b and c. Precisely, we have:

$$f(\tau) = \begin{cases} +\infty & \text{if case 1,} \\ \frac{\sum_{j=1}^{m-1} \beta_{(j)}^2 \hat{w}_{(j)}}{2L\tau} + \left(\sum_{j=1}^{p} \frac{1}{\hat{w}_j} + \frac{L}{2} \sum_{j=1}^{k(\beta)-1} \hat{w}_{(j)} \right) \tau + \sum_{j=m}^{k(\beta)} |\beta_{(j)}| \hat{w}_{(j)} & \text{if case 2,} \\ \tau \sum_{j=1}^{p} \frac{1}{\hat{w}_j} + \sum_{j=1}^{p} |\beta_j| \hat{w}_j & \text{if case 3,} \end{cases}$$

where

$$\begin{aligned} &\text{case 1:} \quad \tau = 0, \\ &\text{case 2:} \quad \exists\, m \in [2, k(\beta)+1]\,, \frac{|\beta_{(m)}|}{L} < \tau \leq \frac{|\beta_{(m-1)}|}{L}, \\ &\text{case 3:} \quad \tau > \frac{|\beta_{(1)}|}{L}. \end{aligned}$$

The function $\tau \longrightarrow a/\tau + b\tau + c$ with $a \geq 0$ and $b > 0$ is decreasing on $]0, \sqrt{a/b}[$ and increasing on $[\sqrt{a/b}, +\infty[$. So, its minimum on $\mathbb{R}_+^*$ is reached at $\sqrt{a/b}$ and is equal to $2\sqrt{ab} + c$. The function f

is a valley on $\mathbb{R}_+$ since f is convex, $f(\tau) \longrightarrow +\infty$ as $\tau \longrightarrow 0^+$ and $f(\tau) \longrightarrow +\infty$ as $\tau \longrightarrow +\infty$. By convexity of f, among the intervals $]|\beta_{(m)}|/L, |\beta_{(m-1)}|/L]$ for $m = 2, \ldots, k(\beta)+1$, there is only one where f is a valley.

Consequently, the minimum of f on $\mathbb{R}_+$ is reached at

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

and is equal to the expression given in Theorem 1.

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 [39] or [18]. The difference with [18] concerns the treatment of the penalty term. So in the following, we will use notations similar to the ones of [18]. 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 $u = (u_0, \ldots, u_{p+2})^T \in \mathbb{R}^{p+3}$. 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 [39] or [18] is to study the epi-limit of U_n. Epi-convergence provides natural conditions under which, if U_n epi-converges to U, one can guarantee the convergence of the minimizer of U_n to the minimizer of U. [8] defines epi-convergence in distribution (denoted by $\rightarrow_{e-d}$) for random lower semicontinuous extended real-valued variables from $\mathbb{R}^q$ as the convergence in distribution for the distance d defined in Lemma 3.

Using the proof of theorem 3.2 in [18], 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 [18] indicates that $U_n \rightarrow_{e-d} U$, 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 (20) in Lemma 2 implies that $\sum_{j=1}^{p_0} |\beta_j^*|^{2-\gamma} \mathbb{1}_{|\beta_j^*| > L\tau^*} > 0$ thus the function $z \rightarrow z^2 C(z)$ 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 minimizer. Theorem 5 of [16] ensures that $\hat{u}^{(n)}$ converges in distribution to the minimizer of U. The asymptotic normality part is proved. It should be noted that [39] does not provide proof for the epi-convergence for the lasso penalty: our proof is self-contained.

Step 2. Let us now show the consistency in variable selection part. It suffices to show that $\mathbb{P}[\mathcal{A} \subset \mathcal{A}_n] \rightarrow 1$ as n tends to infinity and $\mathbb{P}[\mathcal{A}^c \subset {\mathcal{A}_n}^c] \rightarrow 1$ as n tends to infinity. The first claim is an easy consequence of asymptotical normality obatined 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 [18], we have for a such j,

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

Using similar arguments as in [18], we have, as n tends to infinity,

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

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 [18], 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 $\lambda_n / n^{\gamma \wedge 1/2} \to 0$, $\lambda_n n^{(\gamma-1)/2} \to \infty$, $\lambda_n \to \infty$ 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

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

Proof. The proof is divided into two steps.

Step 1. First let us prove that

$$\sum_{j=1}^{p_0} P_{n,j}(u) \to_{e-d} u_{p+2}^2 C(u_{p+2})\,. \tag{9}$$

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

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

For $1 \leq j \leq p_0$ 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 $1 \leq j \leq p_0$ such that $|\beta_j^*| \neq L\tau^*$,

$$P_{n,j}(u) = \frac{u_{p+2}\sqrt{\lambda_n}}{\hat{w}_j^{adb}} + \sqrt{\lambda_n} u_{p+2} \hat{w}_j^{adb} B\left(\frac{\beta_j^*}{\tau^*}\right) + \frac{u_{p+2}^2 |\beta_j^*|^{2-\gamma}}{2L\tau^{*3}}\mathbb{1}_{|\beta_j^*|>L\tau*} + a_{n,j}(u)\,, \tag{10}$$

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 $1 \leq j \leq p_0$ 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, $\sqrt{\lambda_n/n}u_j \leq Lu_{p+2}$ and

$$b_{n,j}(u) = P_{n,j}(u) - \frac{\sqrt{\lambda_n}u_{p+2}}{\hat{w}_j^{adb}} = \frac{\lambda_n u_j \hat{w}_j^{adb}}{\sqrt{n}}, \tag{11}$$

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

$$b_{n,j}(u) = \frac{\lambda_n \hat{w}_j^{adb}}{\left(\tau^* + \frac{u_{p+2}}{\sqrt{\lambda_n}}\right)}\left(\frac{u_j^2}{2Ln} + \frac{\tau^* u_j}{\sqrt{n}}\right) + \frac{L\hat{w}_j^{adb}}{2\left(\tau^* + \frac{u_{p+2}}{\sqrt{\lambda_n}}\right)}u_{p+2}^2. \tag{12}$$

Similiarly, we get the same result if $\beta_j^* = -L\tau^*$. Gathering (10) and using $B(\pm L) = 0$, we have the following decomposition:

$$\sum_{j=1}^{p_0} P_{n,j}(u) = \sum_{j=1}^{p_0} c_{n,j}(u) + \sum_{j=1}^{p_0}\left(a_{n,j}(u)\mathbb{1}_{|\beta_j^*|\neq L\tau^*} + b_{n,j}(u)\mathbb{1}_{|\beta_j^*|=L\tau^*}\right) + \frac{{u_{p+2}}^2}{2L{\tau^*}^3}\sum_{j=1}^{p_0}|\beta_j^*|^{2-\gamma}\mathbb{1}_{|\beta_j^*|>L\tau*}, \tag{13}$$

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 (20) (see Lemma 2) is available since $\beta^* \neq 0$ and

$$c_{n,j}(u) = u_{p+2}\sqrt{\frac{\lambda_n}{n}}\sum_{j=1}^{p_0}\left(\sqrt{n}(|\hat{\beta}_j^{unpen}|^\gamma - |\beta_j^*|^\gamma) + B\left(\frac{\beta_j^*}{\tau^*}\right)\left(\sqrt{n}\frac{\left(|\hat{\beta}_j^{unpen}|^\gamma - |\beta_j^*|^\gamma\right)}{|\hat{\beta}_j^{unpen}|^\gamma|\beta_j^*|^\gamma}\right)\right). \tag{14}$$

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

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

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 (13) with all these convergences, the convergence in probability of (9) is proved. Using first theorem 2.7 (vi) of [30] and then that convergence in probability is stronger than convergence in distribution (theorem 2.7 (ii) of [30]), we get that convergence in probability implies finite-dimensional convergence in (9). Theorem 5 of [16] implies that (9) holds since the limit function $u \to u_{p+2}^2 C(u_{p+2})$ is finite.

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

$$(P_{n,p_0+1}, \cdots, P_{n,p}) \to_{e-d} \left(I_{B_{p_0+1}}, \cdots, I_{B_p}\right), \tag{15}$$

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

$$q_{n,j}(u) = P_{n,j}(u) - \sqrt{\lambda_n} u_{p+2}|\hat{\beta}_j^{unpen}|^\gamma. \tag{16}$$

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 [30], 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 [16] 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 (21). 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 [31], 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 [31]. Consequently, we can use Slutsky's theorem page 32, example 1.4.7 of [31] to ensure that

$$\left(\sqrt{\lambda_n} u_{p+2}|\hat{\beta}_j|^\gamma, q_{n,j}(u_{1:p}, u_{p+2})\right) \to_{e-d} (0, I_{B_j}) \tag{17}$$

since 0 is deterministic. Moreover, we have $\sqrt{\lambda_n}u_{p+2}|\hat{\beta}_j^{unpen}|^\gamma \rightarrow_{u-d} 0$ 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 [16]). We are now in position to use part (b) of theorem 4 of [16]: gathering (17), $\sqrt{\lambda_n}u_{p+2}|\hat{\beta}_j^{unpen}|^\gamma \rightarrow_{u-d} 0$, continuity of 0 and (16), it ensures that $P_{n,j} \rightarrow_{e-d} I_{B_j}$ holds. Since I_{B_j} is deterministic, theorem 18.10 (ii) of [30] ensures that the convergence in probability holds. Now, theorem 18.10 (vi) of [30] leads to the convergence in probability in (15). Moreover, convergence in probability is stronger than convergence in distribution thus (15) 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, (15), Theorem 5 of [21] and theorem 18.10 (iii) (v) (vi) and 18.11 of [30], we get

$$\sum_{j=p_0+1}^{p} P_{n,j} \rightarrow_{e-d} \sum_{j=p_0+1}^{p} I_{B_j} \tag{18}$$

As previously, we can use Slutsky's theorem page 32, example 1.4.7 of [31] to ensure that (18) and (9) imply that

$$\left(\sum_{j=p_0+1}^{p} P_{n,j}(u), \sum_{j=1}^{p_0} P_{n,j}(u)\right) \rightarrow_{e-d} \left(\sum_{j=p_0+1}^{p} I_{B_j}, u_{p+2}^2 C(u_{p+2})\right) \tag{19}$$

since $u_{p+2}^2 C(u_{p+2})$ is deterministic. Moreover, we have $\sum_{j=1}^{p_0} P_{n,j}(u) \rightarrow_{u-d} u_{p+2}^2 C(u_{p+2})$ 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 [16]). Using part (b) of theorem 4 of [16]: gathering (19), $\sum_{j=1}^{p_0} P_{n,j}(u_{1:p}, u_{p+2}) \rightarrow_{u-d} {u_{p+2}}^2 C(u_{p+2})$ 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 (8) and*

$$\sum_{j=1}^{p} |\beta_j^*|^\gamma + \sum_{j=1}^{p_0} \frac{1}{|\beta_j^*|^\gamma}\left(\mathcal{B}_L\left(\frac{\beta_j^*}{\tau^*}\right) - \frac{\beta_j^*}{\tau^*}\mathcal{B}_L'\left(\frac{\beta_j^*}{\tau^*}\right)\right) = 0. \tag{20}$$

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*

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

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

$$d(q_{n,j}, I_{B_j}) = \sum_{k=1}^{+\infty} \frac{1 \wedge d_k(epi(q_{n,j}), epi(I_{B_j}))}{2^k}, \tag{21}$$

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

$$d_k(epi(q_{n,j}), epi(I_{B_j})) = \max_{\|x\| \leq k} |d_{epi(q_{n,j})}(x) - d_{epi(I_{B_j})}(x)|, \tag{22}$$

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,

$$d(q_{n,j}, I_{B_j}) \leq d(q_{n,j}, I_{D_j}) + d(I_{D_j}, I_{B_j}). \tag{23}$$

To begin with, let us show that

$$d(I_{D_j}, I_{B_j}) \leq 2^{-\left[\tau^* \sqrt{\lambda_n}\right]}. \tag{24}$$

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 $\mathrm{epi}(I_{D_j})$ is reached for a point in $\mathrm{epi}(I_{B_j})$. Thus

$$\forall\, x,\ \|x\|_2 \leq k \,\text{with}\, k \leq \sqrt{\lambda_n}\tau^*,\ d_{\mathrm{epi}(I_{D_j})}(x) = d_{\mathrm{epi}(I_{B_j})}(x), \tag{25}$$

and if $k \leq \sqrt{\lambda_n}\tau^*$ then $d_k(\mathrm{epi}(I_{D_j}), \mathrm{epi}(I_{B_j})) = 0$. Now the definition (21) of the distance d implies that

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

and (24) is proved.

Next, we show that

$$d(q_{n,j}, I_{D_j}) \leq \frac{2\sqrt{n}|\hat{\beta}_j^{unpen}|^\gamma}{\lambda_n} + 2^{-\left[\tau^*\sqrt{\lambda_n}\right]}\,. \tag{26}$$

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

$$q_{n,j}(u_{1:p}, u_{p+2}) = \frac{\lambda_n}{|\hat{\beta}_j^{unpen}|^\gamma}\left(\frac{u_{p+2}}{\sqrt{\lambda_n}} + \tau^*\right) q\left(\frac{u_j}{\sqrt{n}\left(\frac{u_{p+2}}{\sqrt{\lambda_n}} + \tau^*\right)}\right) + I_E \tag{27}$$

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, $\mathrm{epi}\left(I_{D_j}\right) \subset \mathrm{epi}\left(q_{n,j}\right)$, $d_{\mathrm{epi}\left(I_{D_j}\right)}(.) \geq d_{\mathrm{epi}(q_{n,j})}(.)$ and

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

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 (27) 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, $\mathrm{epi}\left(q_{n,j}\right) \subset \mathrm{epi}\left(F_{n,j}\right)$, $d_{\mathrm{epi}(q_{n,j})}(.) \geq d_{\mathrm{epi}(F_{n,j})}(.)$ and

$$d_k(\mathrm{epi}\left(q_{n,j}\right), \mathrm{epi}\left(I_{D_j}\right)) \leq \max_{\|x\| \leq k}\left(d_{\mathrm{epi}(I_{D_j})}(x) - d_{\mathrm{epi}(F_{n,j})}(x)\right)\,. \tag{28}$$

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

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

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,

$$d_{\text{epi}(F_{n,j})}(x) = d_{S_1}(x) \wedge d_{S_2}(x)\,. \tag{29}$$

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

$$d^2_{S_2}(x) = \inf_{z \in S_2} \sum_{i=1}^{p+2} (x_i - z_i)^2 = x_j^2 + (x_{p+1} + \sqrt{\lambda_n}\tau^*)^2 + x_{p+2}^2 \mathbb{1}_{x_{p+2}<0}, \tag{30}$$

and

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

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 $d^2_{S_1}(x) = d^2_{\text{epi}(f_{n,j})}(x_1, \cdots, x_p, x_{p+2})$. Technical computations leads to

$$d_{S_1}(x) = \begin{cases} \sqrt{x_j^2 + x_{p+2}^2} & \text{if} \quad x_{p+2} \leq -\frac{\sqrt{n}|\hat{\beta}_j^{unpen}|^\gamma}{\lambda_n}|x_j|, \\ \frac{|x_j| - x_{p+2}\frac{\sqrt{n}|\hat{\beta}_j^{unpen}|^\gamma}{\lambda_n}}{\sqrt{1+\frac{n|\hat{\beta}_j^{unpen}|^{2\gamma}}{\lambda_n^2}}} & \text{if} \quad -\frac{\sqrt{n}|\hat{\beta}_j^{unpen}|^\gamma}{\lambda_n}|x_j| < x_{p+2} \leq \frac{\lambda_n}{\sqrt{n}|\hat{\beta}_j^{unpen}|^\gamma}|x_j|, \\ 0 & \text{if} \quad x_{p+2} > \frac{\lambda_n}{\sqrt{n}|\hat{\beta}_j^{unpen}|^\gamma}|x_j|. \end{cases} \tag{31}$$

Using explicit expressions (31) and (30), we can show that for any $x \in \mathbb{R}^{p+2}$ such that $\|x\|_2 \leq k$ with $k \leq \sqrt{\lambda_n}\tau^*$,

$$d_{S_1}(x) \leq d_{S_2}(x). \tag{32}$$

Gathering (32) with (29), for any $x \in \mathbb{R}^{p+2}$ such that $\|x\|_2 \leq k$ with $k \leq \sqrt{\lambda_n}\tau^*$,

$$d_{\text{epi}(F_{n,j})}(x) = d_{S_1}(x) = d_{\text{epi}(f_{n,j})}(x_1, \cdots, x_p, x_{p+2}). \tag{33}$$

Combining (28), (33) and (25), if $k \leq \sqrt{\lambda_n}\tau^*$, we obtain

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

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

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

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

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

and technical computations leads to

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

Finally, using the definition (21), we have

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

Gathering this inequality with the previous one and the fact that $\sum_{k\geq 1}\frac{k}{2^k} \leq 2$, (26) is proved. Using equation (23) with (24) and (26), the bound involved in Lemma 3 holds. ■

Acknowledgements

Part of this work was supported by the Interuniversity Attraction Pole (IAP) research network in Statistics P5/24 and by MSTIC project of the Joseph-Fourier University. We are grateful to Anestis Antoniadis for constructive and fruitful discussions.

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] A. Argyriou and R. Foygel and N. Srebro, Sparse Prediction with the k-Support Norm. *NIPS*, 1466–1474, 2012.

[4] 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.

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

[6] M. El Anbari and A. Mkhadri. On the adaptive Grill estimator with diverging number of parameters. *Communications in Statistics Theory & Methods* , to appear, 2013.

[7] 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.

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

[9] 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.

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

[11] 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.

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

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

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

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

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

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

[18] 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.

[19] 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.

[20] B. Li and Q. Yu. Robust and sparse bridge regression. *Statistics and Its Interface.*, 2:481–491, 2009.

[21] 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.

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

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

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

[25] 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.

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

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

[28] 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.

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

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

[31] 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.

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

[33] 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.

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

[35] 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.

[36] 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).

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

[38] Q. Zheng, KB. Kulasekera, and C. Gallagher Adaptive penalized quantile regression for high dimensional data. *Journal of Statistical Planning and Inference*, 142 (6):1029–1038, 2013.

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

[40] 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.

[41] 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 (resp. O and U) is the number of well-chosen models (resp. number of overfitting models and underfitting models). Z is the average number of estimated zeros, CZ provides the average number of correctly estimated zeros, (TZ) recall the theoretical zeros number.

	C	O	U	Z(std)	CZ(std)	TZ	λ_n (std)
$n = 100$							
`ad-lasso`	0	0	100	35.38(1.87)	24.32(1.33)	25	λ_n : 2683(1576)
`ad-Enet`	0	0	100	27.72(5.65)	19.69(4.60)	25	$\lambda_{1,n}$: 1252(1197)
							$\lambda_{2,n}$: 1.4(3.1)
`l1-loss-ad-Berhu`	0	35	65	5.16(4.06)	4.07(4.19)	25	λ_n : 1.33(3.39)
`Huber-ad-lasso`	0	3	97	20.9(12.10)	14.08(8.35)	25	λ_n : 59.68(103.74)
`Huber-ad-BerHu`	0	33	67	9.89(6.20)	8.57(6.91)	25	λ_n : 9.57(12.44)
$n = 200$							
`ad-lasso`	0	0	100	35.15(1.45)	24.59(0.73)	25	λ_n : 2718(1562)
`ad-Enet`	0	2	98	27.71(5.92)	20.52(4.71)	25	$\lambda_{1,n}$: 1321(965)
							$\lambda_{2,n}$: 2.5(3.9)
`l1-loss-ad-Berhu`	0	28	72	12.81(5.50)	11.54(6.01)	25	λ_n : 7.59(6.96)
`Huber-ad-lasso`	0	7	93	29.52(9.75)	19.85(6.76)	25	λ_n : 135.26(169.02)
`Huber-ad-BerHu`	0	48	52	16.77(4.65)	15.86(5.19)	25	λ_n : 27.24(16.67)
$n = 400$							
`ad-lasso`	0	0	100	34.75(1.61)	24.74(0.70)	25	λ_n : 3289(2136)
`ad-Enet`	0	0	100	27.06(4.81)	21.17(3.52)	25	$\lambda_{1,n}$: 1260(777)
							$\lambda_{2,n}$: 6(4.74)
`l1-loss-ad-Berhu`	0	38	62	16.27(4.69)	15.10(5.14)	25	λ_n : 14.10(10.05)
`Huber-ad-lasso`	0	9	91	29.67(11.90)	20.3(8.07)	25	λ_n : 277.06(545.91)
`Huber-ad-BerHu`	0	35	65	19.47(4.06)	18.42(4.19)	25	λ_n : 41.78(18.75)

Table 2: Selection model ability on Model 2 based on 100 replications. C (resp. O and U) is the number of well-chosen models (resp. number of overfitting models and underfitting models). Z is the average number of estimated zeros, CZ provides the average number of correctly estimated zeros, (TZ) recall the theoretical zeros number.

	C	O	U	Z(std)	CZ(std)	TZ	λ_n (std)
$n = 100$							
`ad-lasso`	0	0	100	35.03(2.15)	24.18(1.67)	25	λ_n : 2727(2115)
`ad-Enet`	0	0	100	28.66(5.85)	20.46(4.63)	25	$\lambda_{1,n}$: 1432(1374)
							$\lambda_{2,n}$: 1.6(3.2)
`l1-loss-ad-Berhu`	3	23	74	23.30(3.50)	22(2.88)	25	λ_n : 11.65(3.98)
`Huber-ad-lasso`	0	22	78	10.79(8.18)	7.38(5.86)	25	λ_n : 4.19(16.56)
`Huber-ad-BerHu`	5	24	71	24.36(2.64)	23.12(2.30)	25	λ_n : 28.90(8.52)
$n = 200$							
`ad-lasso`	0	0	100	35.25(1.53)	24.61(0.80)	25	λ_n : 3151(1919)
`ad-Enet`	0	0	100	28.21(5.16)	20.68(4.08)	25	$\lambda_{1,n}$: 1362(975)
							$\lambda_{2,n}$: 2.3(3.66)
`l1-loss-ad-Berhu`	8	19	73	24.47(3.82)	23.14(3.54)	25	λ_n : 14.45(5.16)
`Huber-ad-lasso`	0	6	94	21.63(9.03)	16.77(6.94)	25	λ_n : 14.09(16.25)
`Huber-ad-BerHu`	4	11	85	25.36(2.91)	23.78(2.59)	25	λ_n : 34.43(12.01)
$n = 400$							
`ad-lasso`	0	0	100	35.05(1.45)	24.69(0.66)	25	λ_n : 3362(2142)
`ad-Enet`	0	0	100	28.10(4.03)	21.69(3.30)	25	$\lambda_{1,n}$: 1471(898)
							$\lambda_{2,n}$: 5.8(4.69)
`l1-loss-ad-Berhu`	16	7	77	25.65(1.67)	24.31(1.30)	25	λ_n : 19.30(6.40)
`Huber-ad-lasso`	0	14	86	21.30(9.61)	17.76(7.68)	25	λ_n : 104.02(333.07)
`Huber-ad-BerHu`	23	10	67	24.64(4.30)	23.59(4.08)	25	λ_n : 37.91(16.07)

Table 3: Selection model ability on Model 3 based on 100 replications. C (resp. O and U) is the number of well-chosen models (resp. number of overfitting models and underfitting models). Z is the average number of estimated zeros, CZ provides the average number of correctly estimated zeros, (TZ) recall the theoretical zeros number.

	C	O	U	Z(std)	CZ(std)	TZ	λ_n (std)
$n = 100$							
`ad-lasso`	0	66	34	1.97(1.48)	1.51(1.32)	25	λ_n : 3239(2999)
`ad-Enet`	0	83	17	4.46(11.93)	2.87(7.44)	25	$\lambda_{1,n}$: 3648(2052)
							$\lambda_{2,n}$: 65.9(45.8)
`l1-loss-ad-Berhu`	0	0	100	39.46(1.21)	24.99(0.10)	25	λ_n : 284(355)
`Huber-ad-lasso`	0	7	93	24.93(11.20)	18.34(8.25)	25	λ_n : 3357(3500)
`Huber-ad-BerHu`	0	1	99	33.45(4.01)	24.62(0.90)	25	λ_n : 969(333)
$n = 200$							
`ad-lasso`	0	64	36	2.19(5.55)	1.49(3.52)	25	λ_n : 2516(3064)
`ad-Enet`	0	84	16	4.64(12.5)	2.94(7.8)	25	$\lambda_{1,n}$: 2147(2266)
							$\lambda_{2,n}$: 47.9(46.5)
`l1-loss-ad-Berhu`	0	0	100	38.54(2.72)	24.87(0.66)	25	λ_n : 299(357)
`Huber-ad-lasso`	0	14	86	27.97(12.14)	19.92(8.55)	25	λ_n : 3886(3415)
`Huber-ad-BerHu`	0	2	98	30.51(4.34)	23.70(1.32)	25	λ_n : 1091(274)
$n = 400$							
`ad-lasso`	0	74	26	2.23(6.77)	1.50(4.26)	25	λ_n : 2041(2869)
`ad-Enet`	0	91	9	2.64(9.50)	1.71(5.94)	25	$\lambda_{1,n}$: 2451(2186)
							$\lambda_{2,n}$: 50.6(46.9)
`l1-loss-ad-Berhu`	0	0	100	37.69(4.49)	24.77(1.35)	25	λ_n : 311(370)
`Huber-ad-lasso`	0	23	77	25.83(14.31)	18.34(10.13)	25	λ_n : 3373(3153)
`Huber-ad-BerHu`	0	3	97	28.45(4.48)	22.70(2.28)	25	λ_n : 1085(316)

Table 4: Selection model ability and RPE on Model 4. C is the number of well-chosen models

n	100	200	400	1000	2000
C	42	46	69	93	95
RPE mean (std)	0.004(0.002)	0.002(0.001)	0.001(7.7e-04)	4.1e-04(2.9e-04)	1.7e-04(1.1e-04)

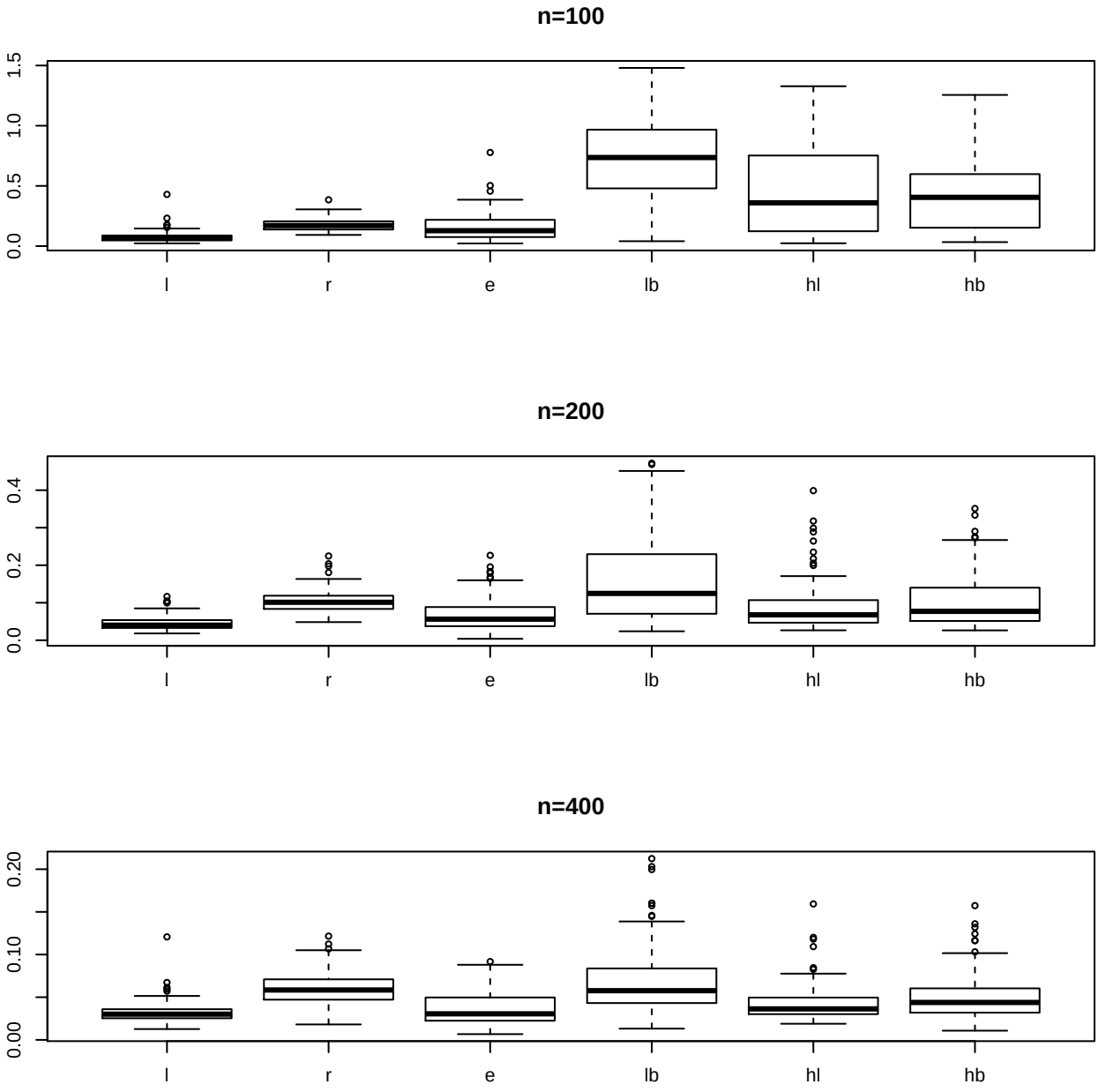

Figure 1: RPE for Model 1 and for `ad-lasso` (l), `ridge` (r), `ad-Enet` (e), `l1-loss-ad-Berhu` (lb), `Huber-ad-lasso` (hl), and `Huber-ad-BerHu` (hb). The boxplots are obtained without extreme values given by, for $n = 100$, hl: 2.87; $n = 200$, hl: 2.48, 2.95, 12.79, 2.86, 2.54, 2.96, 2.95; $n = 400$, hl: 2.95, 2.49, 2.90, 2.95, 2.94, 2.95, 2.95, 0.61, 0.64, 2.93.

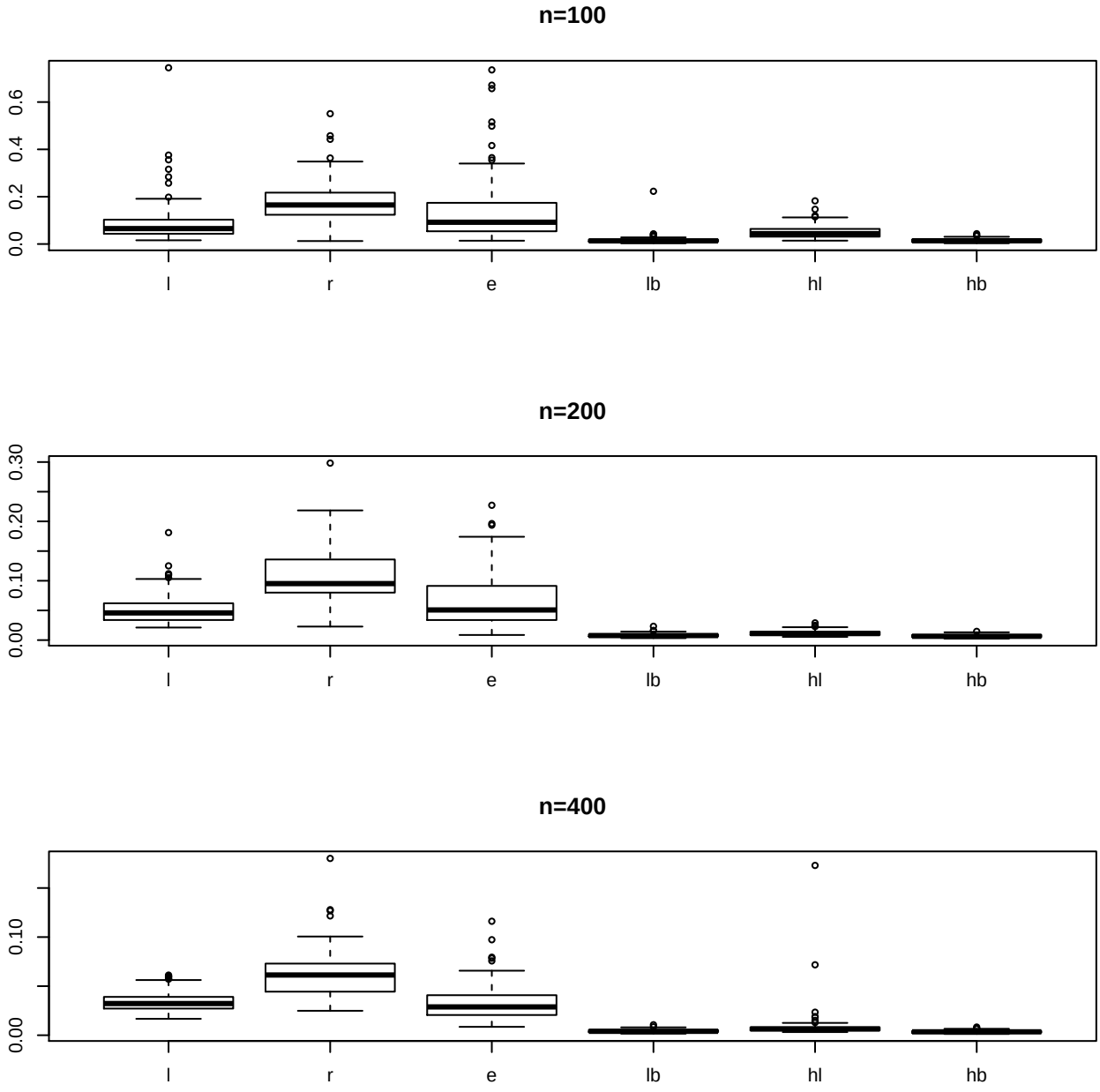

Figure 2: RPE for Model 2 and for `ad-lasso` (l), `ridge` (r), `ad-Enet` (e), `l1-loss-ad-Berhu` (lb), `Huber-ad-lasso` (hl), and `Huber-ad-BerHu` (hb). The boxplots are obtained without extreme values given by, for $n = 100$, hl: 2.94; $n = 200$, hb: 2.24, lb: 0.48; $n = 400$, hl: 0.29, 0.24.

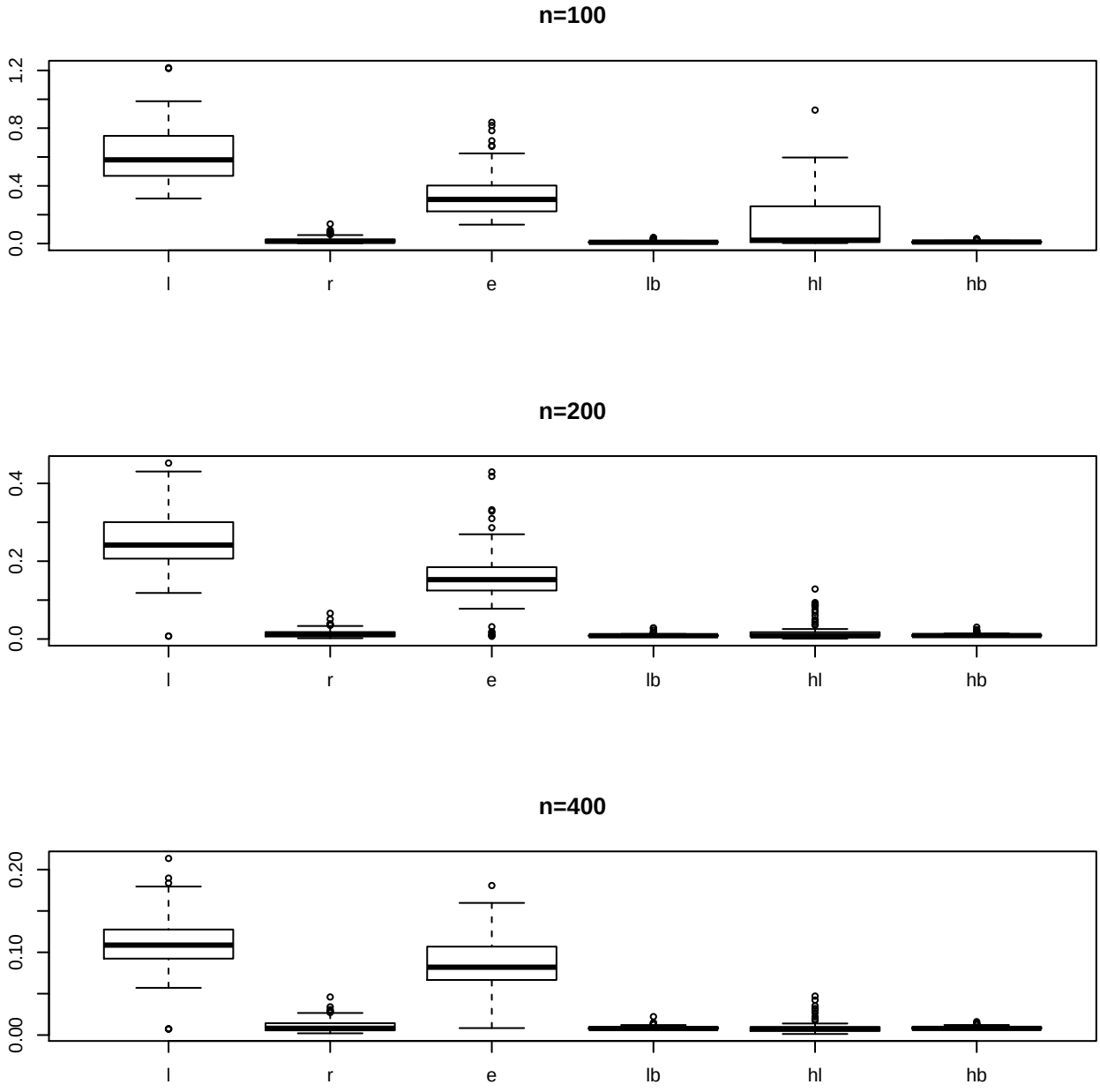

Figure 3: RPE for Model 3 and for `ad-lasso` (l), `ridge` (r), `ad-Enet` (e), `l1-loss-ad-Berhu` (lb), `Huber-ad-lasso` (hl), and `Huber-ad-BerHu` (hb). The boxplots are obtained without extreme values given by, for $n = 200$, hl: 1.52.

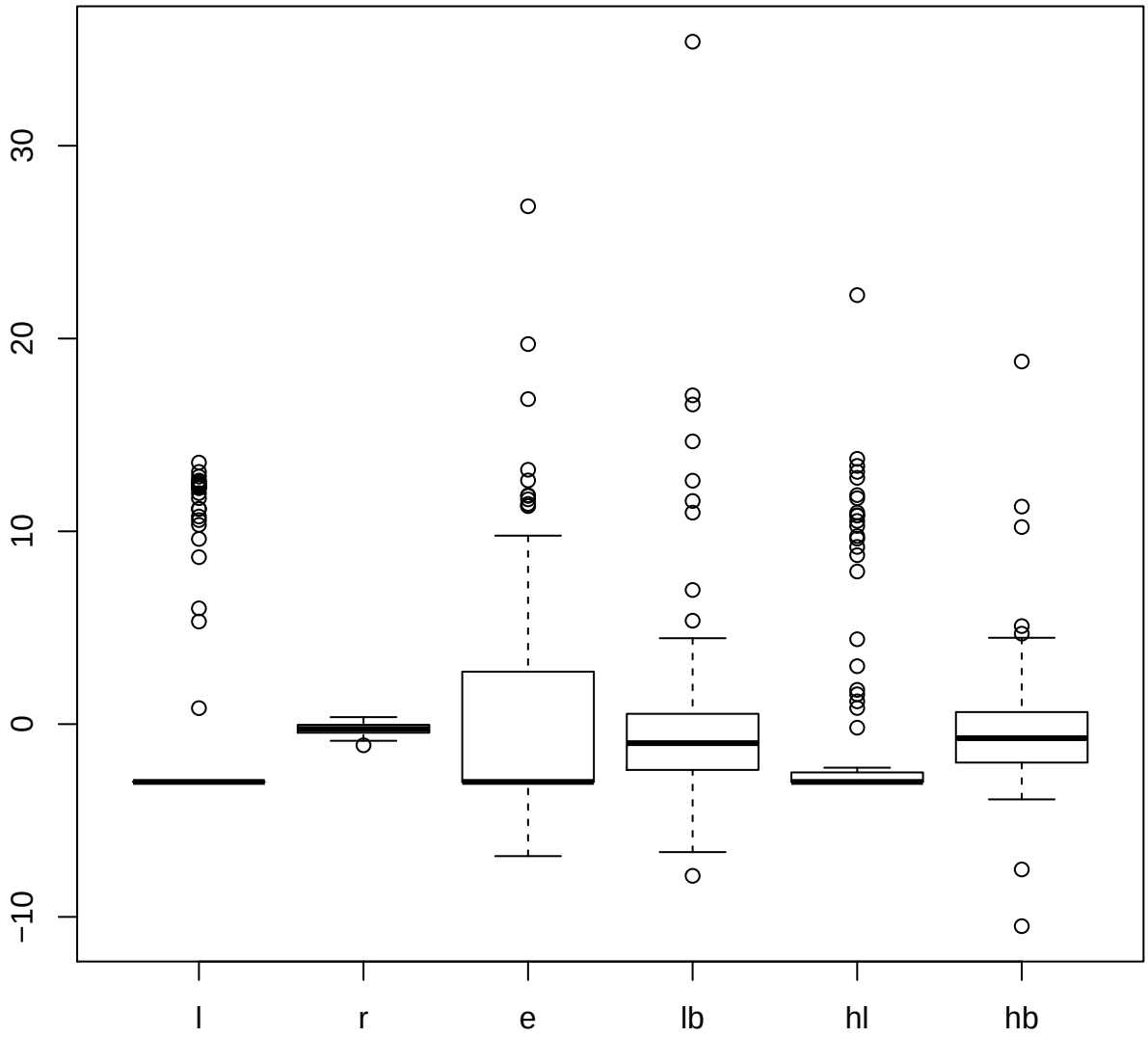

Figure 4: Model 1: Estimation of the difference between first estimated and true coefficients, $n = 200$ for `ad-lasso` (l), `ridge` (r), `ad-Enet` (e), `l1-loss-ad-Berhu` (lb), `Huber-ad-lasso` (hl), and `Huber-ad-BerHu` (hb).

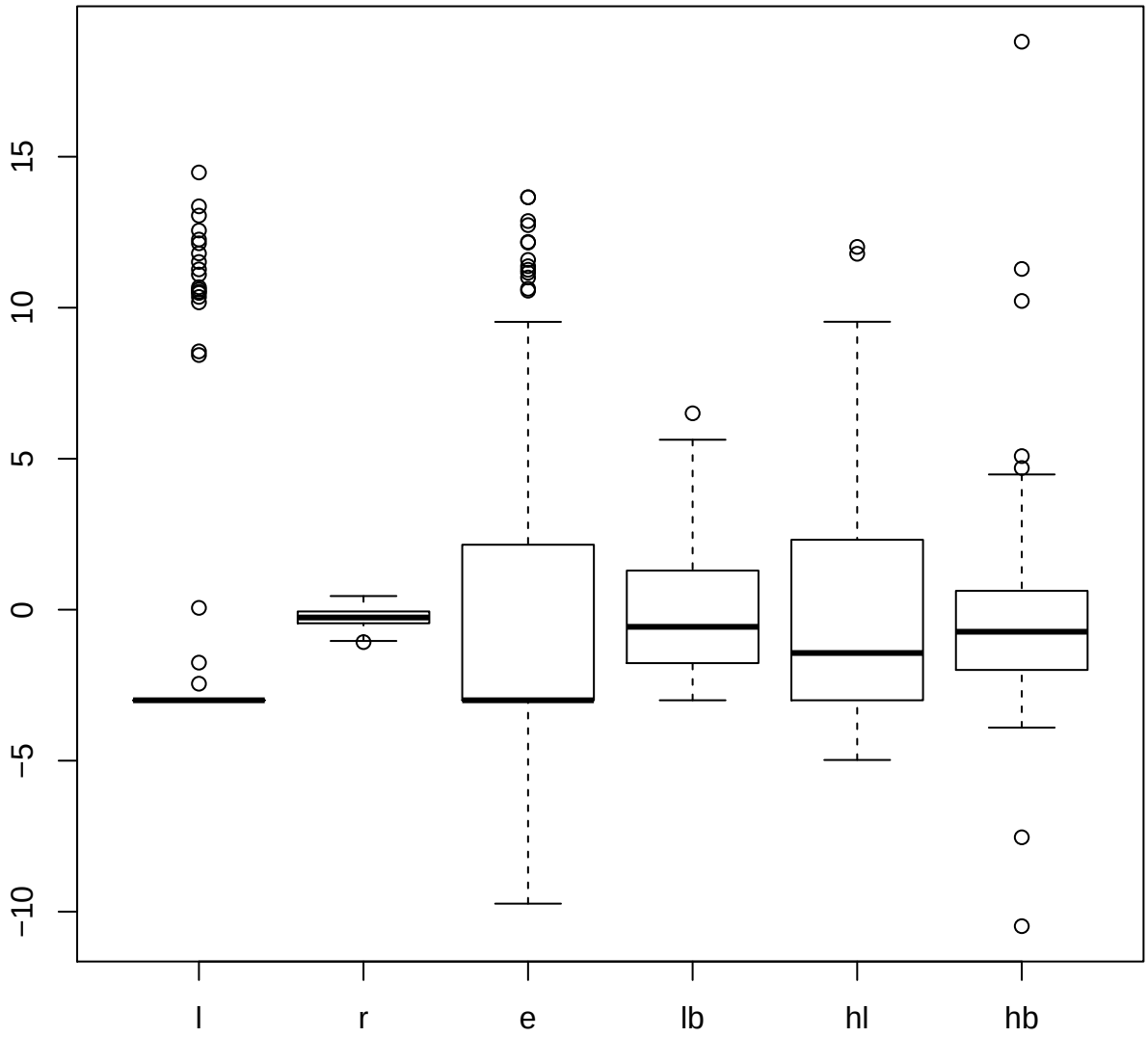

Figure 5: Model 2: Estimation of the difference between first estimated and true coefficients, $n = 200$ for `ad-lasso` (l), `ridge` (r), `ad-Enet` (e), `l1-loss-ad-Berhu` (lb), `Huber-ad-lasso` (hl), and `Huber-ad-BerHu` (hb).

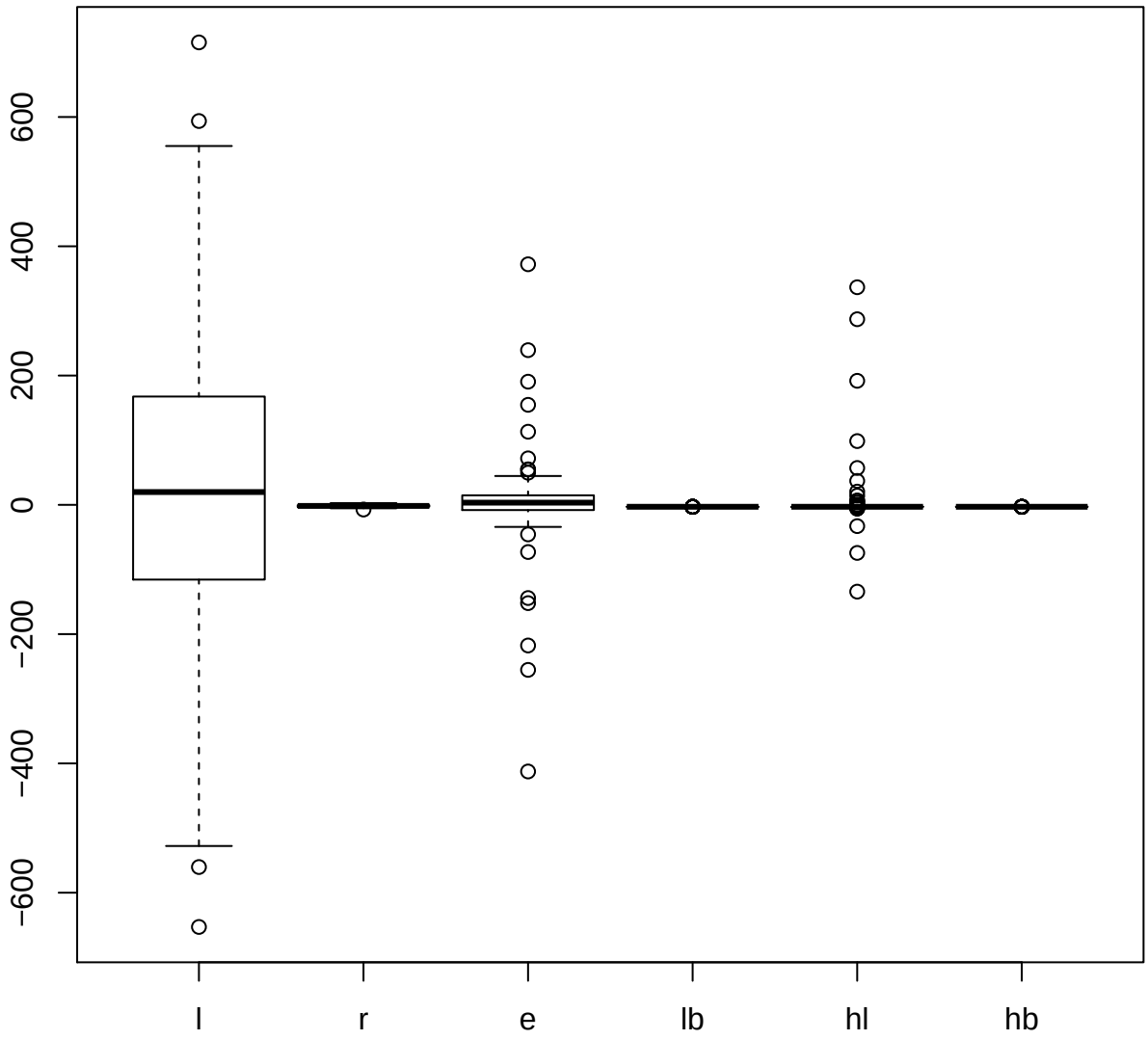

Figure 6: Model 3: Estimation of the difference between first estimated and true coefficients, $n = 200$ for `ad-lasso` (l), `ridge` (r), `ad-Enet` (e), `l1-loss-ad-Berhu` (lb), `Huber-ad-lasso` (hl), and `Huber-ad-BerHu` (hb). The boxplots are obtained without extreme values given by 61.55 for hl.

Table 5: Prostate cancer data: comparing methods

Methods	mean of 100 parameters (std)	mean of 100 RPE (std)
OLS	none	0.6054(0.1397)
`ad-lasso`	λ_n : 2.4177(1.7368)	0.6357(0.1410)
`ridge`	λ_n : 2.6104(2.3111)	0.6145(0.1406)
`ad-Enet`	$\lambda_{1,n}$: 1.1361(1.0048), $\lambda_{2,n}$: 2.5032(10.2605)	0.6231(0.1351)
`l1-loss-ad-Berhu`	λ_n : 1.0173(0.8794)	0.6410(0.1440)
`Huber-ad-lasso`	λ_n : 26.2749(7.4369)	0.7765(0.1879)
`Huber-ad-BerHu`	λ_n : 2.7456(1.9015)	0.6322(0.1391)

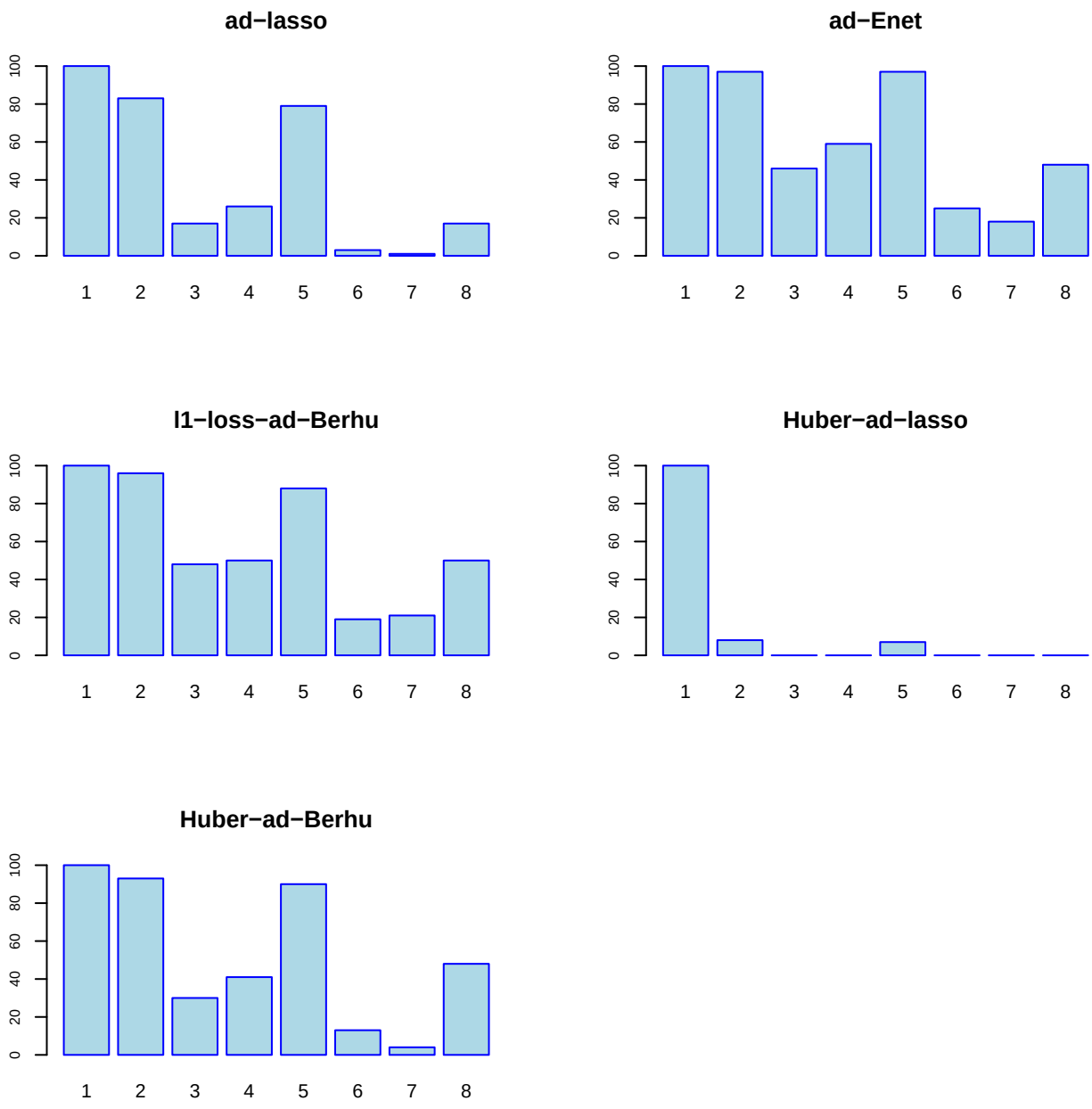

Figure 7: Prostate cancer data: histogram associated with number of selection of each variables in the re-sampling study.