Sign Consistency of the Generalized Elastic Net Estimator

Wencan Zhu, Eric Houngla Adjakossa, Céline Lévy-Leduc, Nils Ternès

Sign Consistency of the Generalized Elastic Net Estimator

Wencan Zhu, Eric Adjakossa, Céline Lévy-Leduc, Nils Ternès

Abstract

In this paper, we propose a novel variable selection approach in the framework of high-dimensional linear models where the columns of the design matrix are highly correlated. It consists in rewriting the initial high-dimensional linear model to remove the correlation between the columns of the design matrix and in applying a generalized Elastic Net criterion since it can be seen as an extension of the generalized Lasso. The properties of our approach called gEN (generalized Elastic Net) are investigated both from a theoretical and a numerical point of view. More precisely, we provide a new condition called GIC (Generalized Irrepresentable Condition) which generalizes the EIC (Elastic Net Irrepresentable Condition) of [1] under which we prove that our estimator can recover the positions of the null and non null entries of the coefficients when the sample size tends to infinity. We also assess the performance of our methodology using synthetic data and compare it with alternative approaches. Our numerical experiments show that our approach improves the variable selection performance in many cases.

Key words: Lasso; Model selection consistency; Irrepresentable Condition; Generalized Lasso; Elastic Net.

1 Introduction

Variable selection has become an important and actively used task for understanding or predicting an outcome of interest in many fields such as medicine [2, 3, 4, 5], social media [6, 7, 8], or finance [9, 10, 11]. Through decades, numerous variable selection methods have been developed such as subset selection [12] or regularization techniques [13].

Subset selection methods achieve sparsity by selecting the best subset of relevant variables using the Akaike information criterion [14] or the Bayesian information criterion [15] but are shown to be NP-hard and could be unstable in practice [16, 17].

The regularized variable selection techniques have become popular for their capability to overcome the above difficulties [18, 19, 20, 21]. Among them, the Lasso approach [18] is one of the most popular and can be defined as follows. Let $\mathbf{y}$ satisfy the following linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^{\star} + \boldsymbol{\epsilon}, \tag{1}$$

1

where $\mathbf{y} = (y_1, \ldots, y_n)' \in \mathbb{R}^n$ is the response variable, $'$ denoting the transposition, $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_p)$ is the design matrix with n rows of observations on p covariates, $\boldsymbol{\beta}^\star = (\beta_1^\star, \ldots, \beta_p^\star)' \in \mathbb{R}^p$ is a sparse vector, namely contains a lot of null components, and $\boldsymbol{\epsilon}$ is a Gaussian vector with zero-mean and a covariance matrix equal to $\sigma^2 \mathbb{I}_n$, $\mathbb{I}_n$ denoting the identity matrix in $\mathbb{R}^n$. The Lasso approach estimates $\boldsymbol{\beta}^\star$ with a sparsity enforcing constraint by minimizing the following penalized least-squares criterion:

$$L_\lambda^{Lasso}(\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1, \tag{2}$$

where $\|a\|_1 = \sum_{k=1}^{p} |a_k|$ denotes the ℓ_1 norm of the vector $(a_1, \ldots, a_p)'$, $\|b\|_2^2 = \sum_{k=1}^{n} b_k^2$ denotes the ℓ_2 norm of the vector $(b_1, \ldots, b_n)'$, and λ is a positive constant corresponding to the regularization parameter. The Lasso popularity largely comes from the fact that the resulting estimator

$$\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda) = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\text{Argmin}} \, L_\lambda^{Lasso}(\boldsymbol{\beta})$$

is sparse (has only a few nonzero entries), and sparse models are often preferred for their interpretability [22]. Moreover, $\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda)$ can be proved to be sign consistent under some assumptions, namely there exists λ such that

$$\lim_{n \to \infty} \mathbb{P}\left(sign\left(\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda)\right) = sign(\boldsymbol{\beta}^\star)\right) = 1,$$

where $sign(x) = 1$ if $x > 0$, -1 if $x < 0$ and 0 if $x = 0$. Before giving the conditions under which [22] prove the sign consistency of $\widehat{\boldsymbol{\beta}}^{Lasso}$, we first introduce some notations. Without loss of generality, we shall assume as in [22] that the first q components of $\boldsymbol{\beta}^\star$ are non null (*i.e.* the components that are associated to the active variables, and denoted as $\boldsymbol{\beta}_1^\star$) and the last $p - q$ components of $\boldsymbol{\beta}^\star$ are null (*i.e.* the components that are associated to the non active variables, and denoted as $\boldsymbol{\beta}_2^\star$). Moreover, we shall denote by $\mathbf{X}_1$ (resp. $\mathbf{X}_2$) the first q (resp. the last $p - q$) columns of $\mathbf{X}$. Hence, $C_n = n^{-1}\mathbf{X}'\mathbf{X}$, which is the empirical covariance matrix of the covariates, can be rewritten as follows:

$$C_n = \begin{bmatrix} C_{11}^n & C_{12}^n \\ C_{21}^n & C_{22}^n \end{bmatrix},$$

with $C_{11}^n = n^{-1}\mathbf{X}_1'\mathbf{X}_1$, $C_{12}^n = n^{-1}\mathbf{X}_1'\mathbf{X}_2$, $C_{21}^n = n^{-1}\mathbf{X}_2'\mathbf{X}_1$, $C_{22}^n = n^{-1}\mathbf{X}_2'\mathbf{X}_2$. It is proved by Zhao and Yu in [22] that $\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda)$ is sign consistent when the following Irrepresentable Condition (IC) is satisfied:

$$\left|\left(C_{21}^n (C_{11}^n)^{-1} \text{sign}(\boldsymbol{\beta}_1^\star)\right)_j\right| \leq 1 - \alpha, \text{ for all } j, \tag{3}$$

where α is a positive constant. In the case where $p \gg n$, Wainwright develops in [23] the necessary and sufficient conditions, for both deterministic and random designs, on p, q, and n for which it is possible to recover the positions of the null and non null components of $\boldsymbol{\beta}^\star$, namely its support, using the Lasso.

When there are high correlations between covariates, especially the active ones, the C_{11}^n matrix may not be invertible, and the Lasso estimator fails to be sign consistent. To

2

circumvent this issue, Zou and Hastie [20] introduced the Elastic Net estimator defined by:

$$\widehat{\boldsymbol{\beta}}^{EN}(\lambda, \eta) = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{Argmin}} \, L_{\lambda, \eta}^{EN}(\boldsymbol{\beta}), \tag{4}$$

where

$$L_{\lambda, \eta}^{EN}(\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 + \eta \|\boldsymbol{\beta}\|_2 \text{ with } \lambda, \eta > 0.$$

Yuan and Lin prove in [24] that when the following Elastic Net Condition (EIC) is satisfied the Elastic Net estimator defined by (4) is sign consistent when p and q are fixed: there exist positive λ and η such that

$$\left| \left(C_{21}^n \left(C_{11}^n + \frac{\eta}{n} \mathbb{I}_q \right)^{-1} \left(\operatorname{sign}(\boldsymbol{\beta}_1^\star) + \frac{2\eta}{\lambda} \boldsymbol{\beta}_1^\star \right) \right)_j \right| \le 1 - \alpha, \text{ for all } j. \tag{5}$$

Moreover, when p, q, and n go to infinity with $p \gg n$, Jia and Yu prove in [1] that the sign consistency of the Elastic Net estimator holds if additionally to Condition (5) n goes to infinity at a rate faster than $q \log(p - q)$.

In the case where the active and non active covariates are highly correlated, IC (3) and EIC (5) may be violated. To overcome this issue several approaches were proposed: the Standard PArtial Covariance (SPAC) method [25] and preconditioning approaches among others. Xue and Qu [25] developed the so-called SPAC-Lasso which enjoys strong sign consistency in both finite-dimensional ($p < n$) and high-dimensional ($p \gg n$) settings. However, the authors mentioned that the SPAC-Lasso method only selects the active variables that are not highly correlated to the non active ones, which may be a weakness of this approach. The preconditioning approaches consist in transforming the given data $\mathbf{X}$ and $\mathbf{y}$ before applying the Lasso criterion. For example, [26] and [27] proposed to left-multiply $\mathbf{X}$, $\mathbf{y}$ and thus $\boldsymbol{\epsilon}$ in Model (1) by specific matrices to remove the correlations between the columns of $\mathbf{X}$. A major drawback of the latter approach, called HOLP (High dimensional Ordinary Least squares Projection), is that the preconditioning step may increase the variance of the error term and thus may alter the variable selection performance.

Recently, [5] proposed another strategy under the following assumption:

(A1) $\mathbf{X}$ is assumed to be a random design matrix such that its rows $(\boldsymbol{x}_i)_{1 \le i \le n}$ are i.i.d. zero-mean Gaussian random vectors having a covariance matrix equal to Σ.

More precisely, they propose to rewrite Model (1) in order to remove the correlation existing between the columns of $\mathbf{X}$. Let $\Sigma^{-1/2} := U D^{-1/2} U^T$ where U and D are the matrices involved in the spectral decomposition of the symmetric matrix Σ given by: $\Sigma = U D U^T$, then, denoting $\widetilde{\mathbf{X}} = \mathbf{X}\Sigma^{-1/2}$, (1) can be rewritten as follows:

$$\mathbf{y} = \widetilde{\mathbf{X}}\widetilde{\boldsymbol{\beta}}^\star + \boldsymbol{\epsilon}, \tag{6}$$

where $\widetilde{\boldsymbol{\beta}}^\star = \Sigma^{1/2} \boldsymbol{\beta}^\star := U D^{1/2} U^T \boldsymbol{\beta}^\star$. With such a transformation, the covariance matrix of the n rows of $\widetilde{\mathbf{X}}$ is equal to identity and the columns of $\widetilde{\mathbf{X}}$ are thus uncorrelated. The advantage of such a transformation with respect to the preconditioning approach proposed by [27] is that the error term $\boldsymbol{\epsilon}$ is not modified thus avoiding an increase of the noise which can

overwhelm the benefits of a well conditioned design matrix. Their approach then consists in minimizing the following criterion with respect to $\widetilde{\boldsymbol{\beta}}$:

$$\left\| \mathbf{y} - \widetilde{\mathbf{X}}\widetilde{\boldsymbol{\beta}} \right\|_2^2 + \lambda \left\| \Sigma^{-1/2}\widetilde{\boldsymbol{\beta}} \right\|_1, \tag{7}$$

where $\widetilde{\mathbf{X}} = \mathbf{X}\Sigma^{-1/2}$ in order to ensure a sparse estimation of $\boldsymbol{\beta}^\star$ thanks to the penalization by the ℓ_1 norm. This criterion actually boils down to the Generalized Lasso proposed by [28]:

$$L_\lambda^{genlasso}(\widetilde{\boldsymbol{\beta}}) = \left\| \mathbf{y} - \widetilde{\mathbf{X}}\widetilde{\boldsymbol{\beta}} \right\|_2^2 + \lambda \left\| D\widetilde{\boldsymbol{\beta}} \right\|_1, \text{ with } \lambda > 0 \tag{8}$$

and $D = \Sigma^{-1/2}$.

Since, as explained in [28], some problems may occur when the rank of the design matrix is not full, we will consider in this paper the following criterion:

$$L_{\lambda,\eta}^{gEN}(\widetilde{\boldsymbol{\beta}}) = \left\| \mathbf{y} - \widetilde{\mathbf{X}}\widetilde{\boldsymbol{\beta}} \right\|_2^2 + \lambda \left\| \Sigma^{-1/2}\widetilde{\boldsymbol{\beta}} \right\|_1 + \eta \left\| \widetilde{\boldsymbol{\beta}} \right\|_2^2, \text{ with } \lambda, \eta > 0. \tag{9}$$

Since it consists in adding an L_2 penalty part to the Generalized Lasso as in the Elastic Net, we will call it generalized Elastic Net (gEN). We prove in Section 2 that under Assumption (A1) and the Generalized Irrepresentable Condition (GIC) (12) given below among others, $\widehat{\boldsymbol{\beta}}$ is a sign-consistent estimator of $\boldsymbol{\beta}^\star$ where $\widehat{\boldsymbol{\beta}}$ is defined by

$$\widehat{\boldsymbol{\beta}} = \Sigma^{-1/2}\widehat{\widetilde{\boldsymbol{\beta}}}, \tag{10}$$

with

$$\widehat{\widetilde{\boldsymbol{\beta}}} = \underset{\widetilde{\boldsymbol{\beta}}}{\text{Argmin}}\, L_{\lambda,\eta}^{gEN}\left(\widetilde{\boldsymbol{\beta}}\right), \tag{11}$$

$L_{\lambda,\eta}^{gEN}\left(\widetilde{\boldsymbol{\beta}}\right)$ being defined in Equation (9). The Generalized Irrepresentable Condition (GIC) can be stated as follows: There exist $\lambda, \eta, \alpha, \delta_4 > 0$ such that for all j,

$$\mathbb{P}\left(\left| \left((C_{21}^n + \frac{\eta}{n}\Sigma_{21})(C_{11}^n + \frac{\eta}{n}\Sigma_{11})^{-1}\left(\text{sign}(\boldsymbol{\beta}_1^\star) + \frac{2\eta}{\lambda}\boldsymbol{\beta}_1^\star\right) - \frac{2\eta}{\lambda}\Sigma_{21}\boldsymbol{\beta}_1^\star \right)_j \right| \le 1 - \alpha\right) = 1 - o\left(e^{-n^{\delta_4}}\right). \tag{12}$$

Note that GIC coincides with EIC when $\mathbf{X}$ is not random and $\Sigma = \mathbb{I}_p$. Moreover, GIC does not require C_{11}^n to be invertible. Since EIC and IC are both particular cases of GIC, if the IC or EIC holds, then there exist λ or η such that the GIC holds.

The rest of the paper is organized as follows. Section 2 is devoted to the theoretical results of the paper. More precisely, we prove that under some mild conditions $\widehat{\boldsymbol{\beta}}$ defined in (10) is a sign-consistent estimator of $\boldsymbol{\beta}^\star$. To support our theoretical results, some numerical experiments are presented in Section 3. The proofs of our theoretical results can be found in Section 5.

2 Theoretical results

The goal of this section is to establish the sign consistency of the Generalized Elastic Net estimator defined in (10). To prove this result, we shall use the following lemma.

Lemma 2.1. *Let* y *satisfying Model (1) under Assumption* (A1) *and* $\widehat{\boldsymbol{\beta}}$ *be defined in* (10). *Then,*

$$\mathbb{P}\left(sign\left(\widehat{\boldsymbol{\beta}}\right) = sign(\boldsymbol{\beta}^\star)\right) \geq \mathbb{P}\left(A_n \cap B_n\right), \tag{13}$$

where

$$A_n := \left\{ \left| \left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1} W_n(1) \right| < \sqrt{n} \left(|\boldsymbol{\beta}_1^\star| - \frac{\lambda}{2n} \left| \left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1} sign(\boldsymbol{\beta}_1^\star) \right| - \frac{\eta}{n} \left| \left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1} \Sigma_{11} \boldsymbol{\beta}_1^\star \right| \right) \right\},$$

$$B_n := \left\{ \left| \mathscr{C}_{21}^{n,\Sigma} \left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1} W_n(1) - W_n(2) \right| \leq \frac{\lambda}{2\sqrt{n}} \right.$$
$$\left. - \frac{\lambda}{2\sqrt{n}} \left| \mathscr{C}_{21}^{n,\Sigma} \left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1} \left(sign(\boldsymbol{\beta}_1^\star) + \frac{2\eta}{\lambda} \Sigma_{11} \boldsymbol{\beta}_1^\star \right) - \frac{2\eta}{\lambda} \Sigma_{21} \boldsymbol{\beta}_1^\star \right| \right\},$$

and

$$\mathscr{C}_{11}^{n,\Sigma} = C_{11}^n + \frac{\eta}{n} \Sigma_{11}, \ \mathscr{C}_{21}^{n,\Sigma} = C_{21}^n + \frac{\eta}{n} \Sigma_{21}, \ W_n = \frac{1}{\sqrt{n}} X' \epsilon = \begin{bmatrix} W_n(1) \\ W_n(2) \end{bmatrix}, \tag{14}$$

with

$$W_n(1) = \frac{1}{\sqrt{n}} X_1' \epsilon \ \text{and} \ W_n(2) = \frac{1}{\sqrt{n}} X_2' \epsilon.$$

The proof of Lemma 2.1 is given in Section 5.

The following theorem gives the conditions under which the sign consistency of the generalized Elastic Net estimator $\widehat{\boldsymbol{\beta}}$ defined in (10) holds.

Theorem 2.2. *Assume that* y *satisfies Model (1) under Assumption* (A1) *with* $p = p_n$ *is such that* $p_n \exp\left(n^{-\delta}\right)$ *tends to 0 as n tends to infinity for all positive δ. Assume also that there exist some positive constants* M_1, M_2, M_3 *and α satisfying*

$$M_1 < \frac{\boldsymbol{\beta}_{\min}^2}{9\sigma^2} \quad \text{and} \quad \frac{\sqrt{2 + \sqrt{2}}\sqrt{M_3}\sigma}{\alpha} < \frac{\boldsymbol{\beta}_{\min}}{3M_2\sqrt{q}}, \tag{15}$$

and that there exist $\lambda > 0$ *and* $\eta > 0$ *such that* (12) *and*

$$\frac{\lambda}{n} < \frac{2\boldsymbol{\beta}_{\min}}{3M_2\sqrt{q}}, \tag{16}$$

$$\frac{\lambda}{n} \geq \frac{2\sqrt{2 + \sqrt{2}}\sqrt{M_3}\sigma}{\alpha}, \tag{17}$$

$$\frac{\eta}{n} < \frac{1}{3M_2\lambda_{\max}(\Sigma_{11})} \times \frac{\boldsymbol{\beta}_{\min}}{\|\boldsymbol{\beta}_1^\star\|_2}, \tag{18}$$

hold as n tends to infinity, where $\boldsymbol{\beta}_{\min} = \min_{1 \leq j \leq q} \left| (\boldsymbol{\beta}_1^\star)_j \right|$. *Suppose also that there exist some positive constants* $\delta_1, \delta_2, \delta_3$ *such that, as* $n \to \infty$,

$$\mathbb{P}\left(\lambda_{\max}\left(H_A H_A'\right) \leq M_1\right) = 1 - o\left(e^{-n^{\delta_1}}\right), \tag{19}$$

5

$$\mathbb{P}\left(\lambda_{\max}\left(\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\right) \leq M_2\right) = 1 - o\left(e^{-n^{\delta_2}}\right), \tag{20}$$

$$\mathbb{P}\left(\lambda_{\max}\left(H_B H_B'\right) \leq M_3\right) = 1 - o\left(e^{-n^{\delta_3}}\right), \tag{21}$$

where $\lambda_{\max}(A)$ denotes the largest eigenvalue of A,

$$H_A = \frac{1}{\sqrt{n}}\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\mathbf{X}_1' \text{ and } H_B = \frac{1}{\sqrt{n}}\left(\mathscr{C}_{21}^{n,\Sigma}\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\mathbf{X}_1' - \mathbf{X}_2'\right),$$

$\mathscr{C}_{11}^{n,\Sigma}$ and $\mathscr{C}_{21}^{n,\Sigma}$ being defined in (14) and $\mathbf{X}_1$ (resp. $\mathbf{X}_2$) denoting the first q (resp. the last $p - q$) columns of $\mathbf{X}$. Then,

$$\mathbb{P}\left(sign\left(\widehat{\boldsymbol{\beta}}\right) = sign(\boldsymbol{\beta}^\star)\right) \to 1, \text{ as } n \to \infty,$$

where $\widehat{\boldsymbol{\beta}}$ is defined in (10).

Note that Conditions (16) and (17) are consistent thanks to (15).

The proof of Theorem 2.2 is given in Section 5 and a discussion on the assumptions of Theorem 2.2 is provided in Section 3.

3 Numerical experiments

The goal of this section is to discuss the assumptions and illustrate the results of Theorem 2.2. For this, we generated datasets from Model (1) where the matrix Σ appearing in (A1) is defined by

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}' & \Sigma_{22} \end{bmatrix}. \tag{22}$$

In (22), Σ_{11} is the correlation matrix of the active variables having its off-diagonal entries equal to α_1, Σ_{22} is the correlation matrix of the non active variables having its off-diagonal entries equal to α_3 and Σ_{12} is the correlation matrix between the active and the non active variables with entries equal to α_2. In the numerical experiments, $(\alpha_1, \alpha_2, \alpha_3) = (0.3, 0.5, 0.7)$. Moreover, $\boldsymbol{\beta}^\star$ appearing in Model (1) has q non zero components which are equal to b and $\sigma = 1$. The number of predictors p is equal to 200, 400, or 600 and the sample size n takes the same values for each value of p.

3.1 Discussion on the assumptions of Theorem 2.2

We first show that GIC defined in (12) can be satisfied even when EIC and IC, defined in (5) and (3) respectively, are not fulfilled. For this, we computed for different values of λ and η

the following values:

$$\text{IC} = \max_{j} \left(\left\| \left(C_{21}^n (C_{11}^n)^{-1} (\text{sign}(\boldsymbol{\beta}_1^\star)) \right)_j \right\| \right)$$

$$\text{EIC} = \min_{\lambda,\eta} \max_{j} \left(\left\| \left(C_{21}^n (C_{11}^n + \frac{\eta}{n} \mathbb{I}_q)^{-1} (\text{sign}(\boldsymbol{\beta}_1^\star) + \frac{2\eta}{\lambda} \boldsymbol{\beta}_1^\star) \right)_j \right\| \right)$$

$$\text{GIC} = \min_{\lambda,\eta} \max_{j} \left(\left\| \left((C_{21}^n + \frac{\eta}{n} \Sigma_{21})(C_{11}^n + \frac{\eta}{n} \Sigma_{11})^{-1} \left(\text{sign}(\boldsymbol{\beta}_1^\star) + \frac{2\eta}{\lambda} \boldsymbol{\beta}_1^\star \right) - \frac{2\eta}{\lambda} \Sigma_{21} \boldsymbol{\beta}_1^\star \right)_j \right\| \right) \quad (23)$$

and Figure 1 displays the boxplots of these criteria obtained from 100 replications. We can see from these figures that in all the considered cases GIC is satisfied (*i.e.* all values are smaller than 1) whereas EIC and IC are not. The values of p and n do not seem to have a big impact on EIC and IC. However, contrary to p, n seems to have an influence on GIC which increases with n when $b = 1$ and decreases when n increases when $b = 10$.

Figures 2 and 3 show the behavior of $\lambda_{\max} \left(H_A H_A' \right)$, $\lambda_{\max} \left(\left(\mathscr{C}_{11}^{n,\Sigma} \right)^{-1} \right)$ and $\lambda_{\max} \left(H_B H_B' \right)$ appearing in (19), (20) and (21) with respect to η for different values of n, p and for $q = 5$ or 10. These plots thus provide lower bounds for M_1, M_2 and M_3 appearing in the previous equations. Observe that (18) can be rewritten as:

$$\eta M_2 < \frac{n}{3\lambda_{\max}(\Sigma_{11})} \times \frac{\boldsymbol{\beta}_{\min}}{\|\boldsymbol{\beta}_1^\star\|_2}. \quad (24)$$

Based on the plots at the bottom right of Figures 2 and 3, we can see that there exist η's satisfying Condition 24 and thus (18) and that the interval in which the adapted η's lie is larger when $q = 5$ than when $q = 10$.

Based on the average of M_1 previously obtained, the left part of (15) is always satisfied as soon as $b > \sqrt{18}$. Based on the average of M_2 and M_3 previously obtained, the average of left-hand side and of the right-hand side of the right part of Equation (15) are displayed in Figures 4 and 5. We can see from these figures that it is only satisfied for large values of b. Moreover, it is more often satisfied when $q = 5$ than for $q = 10$.

We will show in the next section that even if the cases where all the conditions of the theorem are not fulfilled our method is robust enough to outperform the Elastic Net defined in (4) even in these cases.

3.2 Comparison with other methods

To assess the performance of our approach (gEN) in terms of sign-consistency with respect to other methods and to illustrate the results of Theorem 2.2, we computed the True Positive Rate (TPR), namely the proportion of active variables selected, and the False Positive Rate (FPR), namely the proportion of non active variables selected, of the Elastic Net and gEN estimators defined in (4) and (10), respectively.

Figures 6 and 8 display the empirical mean of the largest difference between the True Positive Rate and False Positive Rate over the replications. It is obtained by selecting for each replication the value of λ and η achieving the largest difference between the TPR and FPR and by averaging these differences. They also display the corresponding TPR and FPR

Figure 1: Boxplot of values defined in (23) and obtained from 100 replications.
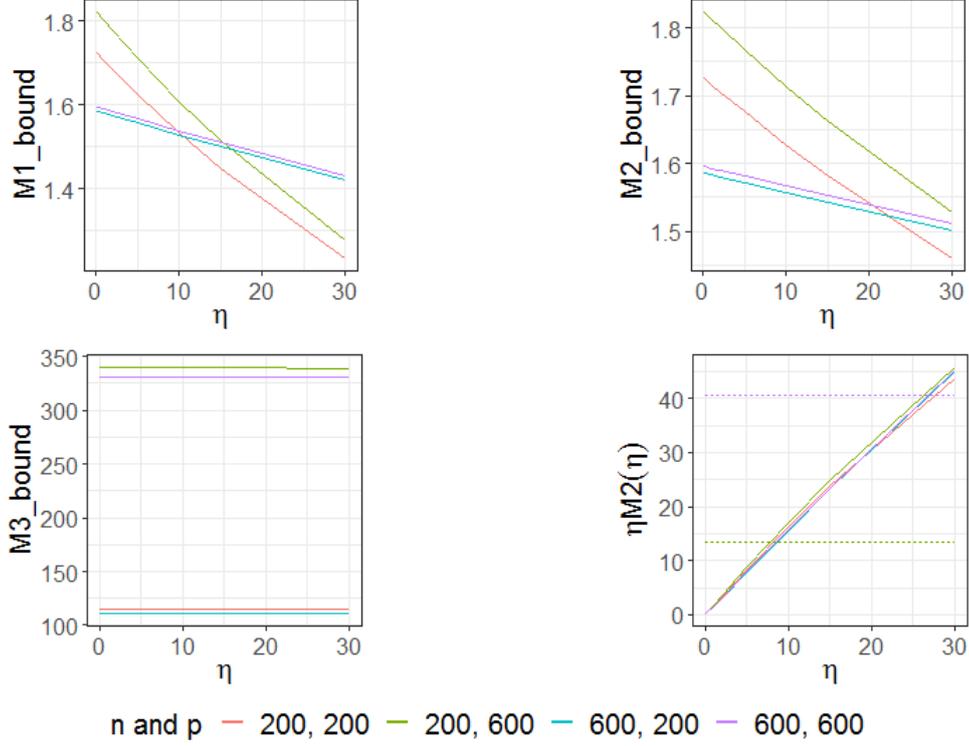
Figure 2: Top left: Average of $\lambda_{\max}\left(H_A H_A'\right)$ in (19) as a function of η. Top right: Average of $\lambda_{\max}\left(\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\right)$ in (20) as a function of η. Bottom left: Average of $\lambda_{\max}\left(H_B H_B'\right)$ in (21) as a function of η. Bottom right: Average of the left (resp. right) part of (24) in plain (resp. dashed) line. The averages are obtained from 10 replications. Here $q = 5$.

for gEN and Elastic Net for different values of n and p. We can see from these figures that the gEN and the Elastic Net estimators have a TPR equal to 1 but that the FPR of gEN is smaller than Elastic Net. We can see from these figures that the difference between the performance of gEN and Elastic Net is larger for high signal-to-noise ratios ($b = 10$). It has to be noticed that when TPR=1 for our approach it also means that the signs of the non null $\beta_i^\star$ are also properly retrieved.

4 Discussion

In this paper, we proposed a novel variable selection approach called gEN (generalized Elastic Net) in the framework of linear models where the columns of the design matrix are highly correlated and thus when the standard Lasso criterion usually fails. We proved that under mild conditions, among which the GIC, which is valid when other standard conditions like EIC or IC are not fulfilled, our method provides a sign-consistent estimator of $\boldsymbol{\beta}^\star$. For a more thorough discussion regarding the application of our approach in practical
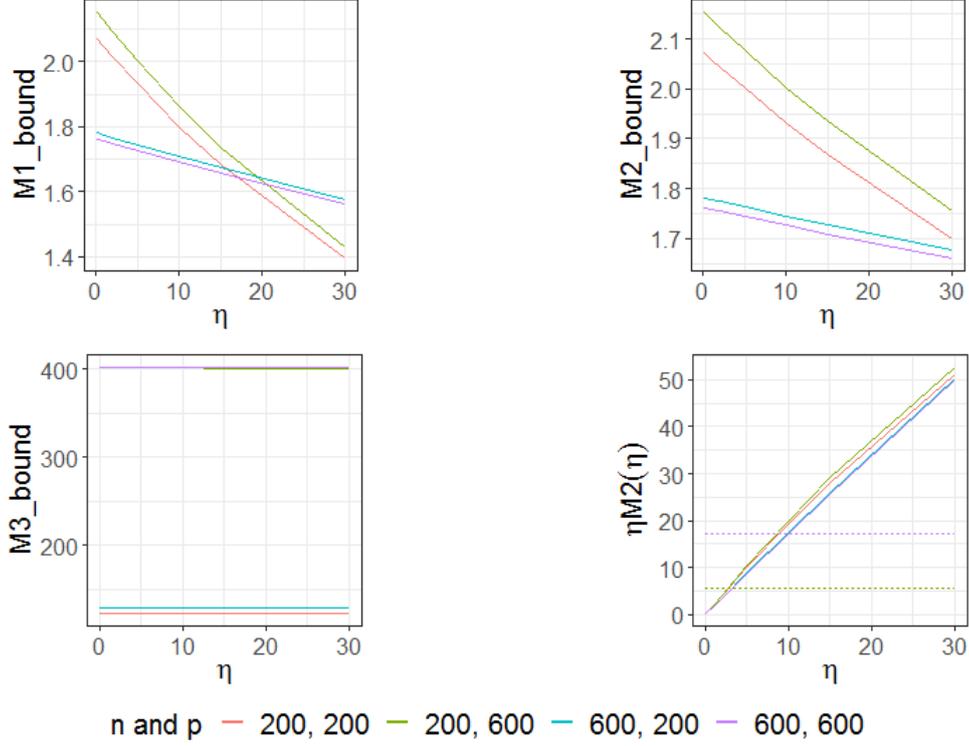
Figure 3: Top left: Average of $\lambda_{\max}\left(H_A H_A'\right)$ in (19) as a function of η. Top right: Average of $\lambda_{\max}\left(\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\right)$ in (20) as a function of η. Bottom left: Average of $\lambda_{\max}\left(H_B H_B'\right)$ in (21) as a function of η. Bottom right: Average of the left (resp. right) part of (24) in plain (resp. dashed) line. The averages are obtained from 10 replications. Here $q = 10$.

situations, we refer the reader to [5].

5 Proofs

5.1 Proof of Lemma 2.1

Note that (9) given by:

$$L^{gEN}(\widetilde{\boldsymbol{\beta}}) = \left\|\mathbf{y} - \widetilde{\mathbf{X}}\widetilde{\boldsymbol{\beta}}\right\|_2^2 + \lambda\left\|\Sigma^{-1/2}\widetilde{\boldsymbol{\beta}}\right\|_1 + \eta\left\|\widetilde{\boldsymbol{\beta}}\right\|_2^2$$

can be rewritten as

$$L^{gEN}(\widetilde{\boldsymbol{\beta}}) = \left\|\mathbf{y}^* - \widetilde{\mathbf{X}}^*\widetilde{\boldsymbol{\beta}}\right\|_2^2 + \lambda\left\|\Sigma^{-1/2}\widetilde{\boldsymbol{\beta}}\right\|_1,$$

where

$$\mathbf{y}^* = \begin{pmatrix} \mathbf{y} \\ 0 \end{pmatrix}, \quad \widetilde{\mathbf{X}}^* = \begin{pmatrix} \widetilde{\mathbf{X}} \\ \sqrt{\eta}\mathbb{I}_p \end{pmatrix}.$$
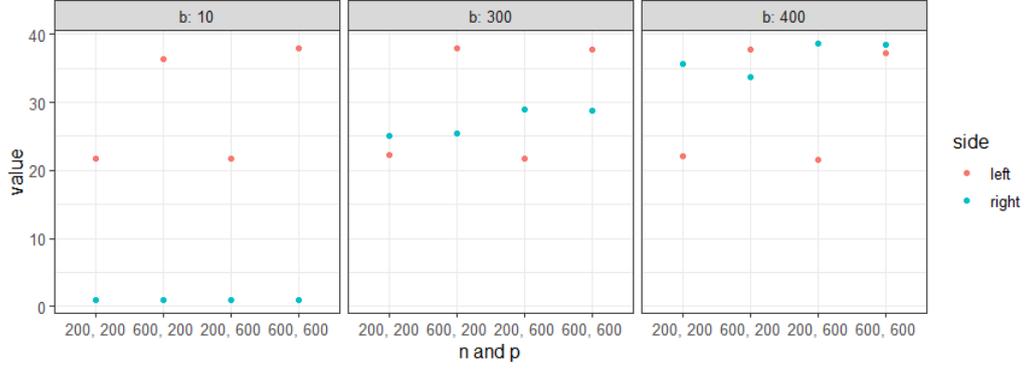
10

Figure 4: Average of the left-hand (resp. right-hand) side of the second part of (15) in red (resp. blue) for $q = 5$.
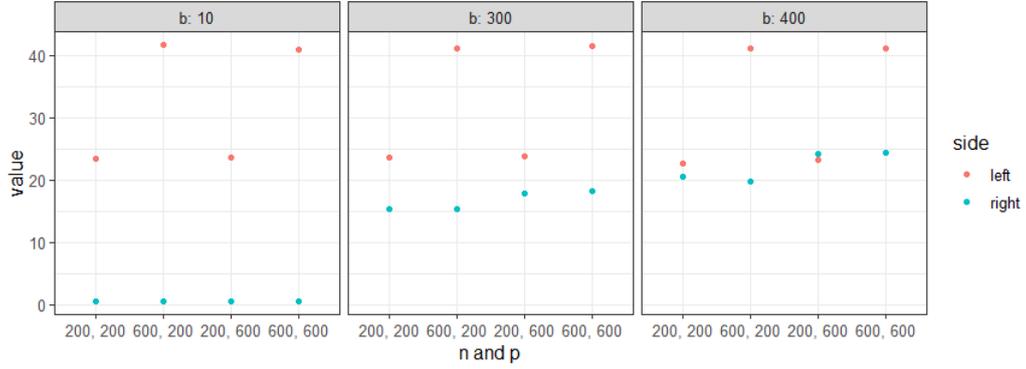

Figure 5: Average of the left-hand (resp. right-hand) side of the second part of (15) in red (resp. blue) for $q = 10$.

Then, $\widehat{\overline{\boldsymbol{\beta}}}$ satisfies

$$\widetilde{\mathbf{X}}^{*\prime}\left(\mathbf{y}^* - \widetilde{\mathbf{X}}^*\widehat{\overline{\boldsymbol{\beta}}}\right) = \frac{\lambda}{2}(\Sigma^{-1/2})^\prime z, \tag{25}$$

where $A^\prime$ denotes the transpose of the matrix A, and

$$\begin{cases} z_j = sign\left((\Sigma^{-1/2}\widehat{\overline{\boldsymbol{\beta}}})_j\right), & \text{if } (\Sigma^{-1/2}\widehat{\overline{\boldsymbol{\beta}}})_j \neq 0 \\ z_j \in [-1, 1], & \text{if } (\Sigma^{-1/2}\widehat{\overline{\boldsymbol{\beta}}})_j = 0 \end{cases}.$$

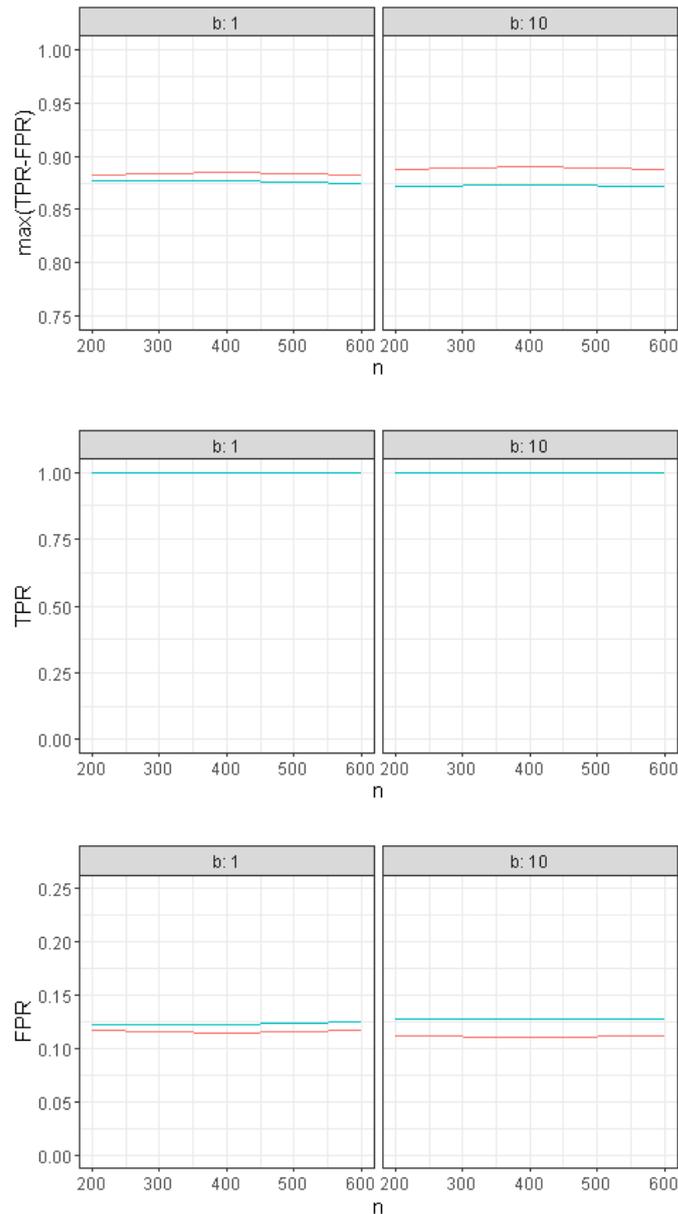Figure 6: Average of max(TPR-FPR) and the corresponding TPR and FPR for gEN (in red) and Elastic Net (in blue) with $p = 200$.

Figure 7: Average of max(TPR-FPR) and the corresponding TPR and FPR for gEN (in red) and Elastic Net (in blue) with $p = 400$.
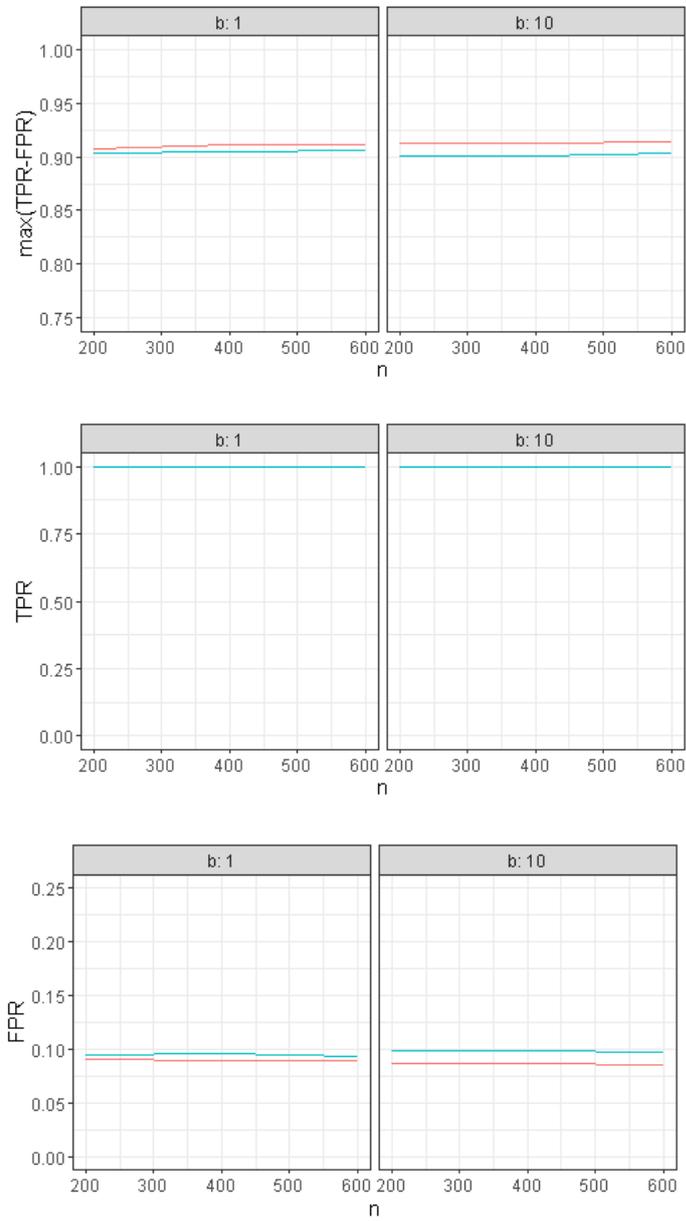
Figure 8: Average of max(TPR-FPR) and the corresponding TPR and FPR for gEN (in red) and Elastic Net (in blue) with $p = 600$.

13

Equation (25) can be rewritten as:

$$\mathbf{X}'\mathbf{y} - (\mathbf{X}'\mathbf{X} + \eta\Sigma)\,\widehat{\boldsymbol{\beta}} = \frac{\lambda}{2}z$$

which leads to

$$\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}^\star - \widehat{\boldsymbol{\beta}}) + \mathbf{X}'\boldsymbol{\epsilon} - \eta\Sigma\widehat{\boldsymbol{\beta}} = \frac{\lambda}{2}z,$$

by using that $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^\star + \boldsymbol{\epsilon}$. By using the following notations: $\widehat{\mathbf{u}} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star$,

$$C_n = \frac{1}{n}\mathbf{X}'\mathbf{X} \text{ and } W_n = \frac{1}{\sqrt{n}}\mathbf{X}'\boldsymbol{\epsilon},$$

Equation (25) becomes

$$\left(C_n + \frac{\eta}{n}\Sigma\right)\sqrt{n}\widehat{\mathbf{u}} + \frac{\eta}{\sqrt{n}}\Sigma\boldsymbol{\beta}^\star - W_n = -\frac{\lambda}{2\sqrt{n}}z. \tag{26}$$

With the following notations:

$$C_n = \begin{pmatrix} C_{11}^n & C_{12}^n \\ C_{21}^n & C_{22}^n \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \quad \widehat{\mathbf{u}} = \begin{pmatrix} \widehat{\mathbf{u}}_1 \\ \widehat{\mathbf{u}}_2 \end{pmatrix}, \quad W_n = \begin{pmatrix} W_n(1) \\ W_n(2) \end{pmatrix}, \quad \boldsymbol{\beta}^\star = \begin{pmatrix} \boldsymbol{\beta}_1^\star \\ \mathbf{0} \end{pmatrix},$$

the first components of Equation (26) are:

$$\left(C_{11}^n + \frac{\eta}{n}\Sigma_{11}\right)\sqrt{n}\widehat{\mathbf{u}}_1 + \left(C_{12}^n + \frac{\eta}{n}\Sigma_{12}\right)\sqrt{n}\widehat{\mathbf{u}}_2 + \frac{\eta}{\sqrt{n}}\Sigma_{11}\boldsymbol{\beta}_1^\star - W_n(1) = -\frac{\lambda}{2\sqrt{n}}sign(\boldsymbol{\beta}_1^\star). \tag{27}$$

If $\widehat{\mathbf{u}} = \begin{pmatrix} \widehat{\mathbf{u}}_1 \\ 0 \end{pmatrix}$, it can be seen as a solution of the generalized Elastic Net criterion where, by Equation (27), $\widehat{\mathbf{u}}_1$ is defined by:

$$\sqrt{n}\widehat{\mathbf{u}}_1 = \left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1} W_n(1) - \frac{\eta}{\sqrt{n}}\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\Sigma_{11}\boldsymbol{\beta}_1^\star - \frac{\lambda}{2\sqrt{n}}\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}sign(\boldsymbol{\beta}_1^\star), \tag{28}$$

where we used (14).

Note that the event A_n can be rewritten as follows:

$$\sqrt{n}\left(-|\boldsymbol{\beta}_1^\star| + \frac{\lambda}{2n}\left|\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}sign(\boldsymbol{\beta}_1^\star)\right| + \frac{\eta}{n}\left|\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\Sigma_{11}\boldsymbol{\beta}_1^\star\right|\right)$$

$$< \left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}W_n(1) <$$

$$\sqrt{n}\left(|\boldsymbol{\beta}_1^\star| - \frac{\lambda}{2n}\left|\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}sign(\boldsymbol{\beta}_1^\star)\right| - \frac{\eta}{n}\left|\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\Sigma_{11}\boldsymbol{\beta}_1^\star\right|\right)$$

which implies

$$\sqrt{n}\left(-|\boldsymbol{\beta}_1^\star| + \frac{\lambda}{2n}\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}sign(\boldsymbol{\beta}_1^\star) + \frac{\eta}{n}\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\Sigma_{11}\boldsymbol{\beta}_1^\star\right)$$

$$< \left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}W_n(1) <$$

$$\sqrt{n}\left(|\boldsymbol{\beta}_1^\star| + \frac{\lambda}{2n}\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}sign(\boldsymbol{\beta}_1^\star) + \frac{\eta}{n}\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\Sigma_{11}\boldsymbol{\beta}_1^\star\right),$$

using that $-|x| \leq x \leq |x|$, $\forall x \in \mathbb{R}$. Then, by using (28), we get that $\sqrt{n}|\widehat{\mathbf{u}}_1| < \sqrt{n}|\boldsymbol{\beta}_1^\star|$ and thus $|\widehat{\mathbf{u}}_1| < |\boldsymbol{\beta}_1^\star|$. Notice that $|\widehat{\mathbf{u}}_1| < |\boldsymbol{\beta}_1^\star|$ implies that $\widehat{\boldsymbol{\beta}}_1 \neq 0$ and that $sign(\widehat{\boldsymbol{\beta}}_1) = sign(\boldsymbol{\beta}_1^\star)$. Moreover, since $\widehat{\mathbf{u}}_2 = 0$, we get that $sign(\widehat{\boldsymbol{\beta}}) = sign(\boldsymbol{\beta}^\star)$.

The last components of (26) satisfy:

$$\left(C_{21}^n + \frac{\eta}{n}\Sigma_{21}\right)\sqrt{n}\widehat{\mathbf{u}}_1 + \left(C_{22}^n + \frac{\eta}{n}\Sigma_{22}\right)\sqrt{n}\widehat{\mathbf{u}}_2 + \frac{\eta}{\sqrt{n}}\Sigma_{21}\boldsymbol{\beta}_1^\star - W_n(2) = -\frac{\lambda}{2\sqrt{n}}z_2,$$

where by (25), $|z_2| \leq 1$. Hence,

$$\left|\left(C_{21}^n + \frac{\eta}{n}\Sigma_{21}\right)\sqrt{n}\widehat{\mathbf{u}}_1 + \frac{\eta}{\sqrt{n}}\Sigma_{21}\boldsymbol{\beta}_1^\star - W_n(2)\right| \leq \frac{\lambda}{2\sqrt{n}},$$

which can be rewritten as follows by using (28):

$$\left|\mathscr{C}_{21}\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\left(W_n(1) - \frac{\eta}{\sqrt{n}}\Sigma_{11}\boldsymbol{\beta}_1^\star - \frac{\lambda}{2\sqrt{n}}sign(\boldsymbol{\beta}_1^\star)\right) + \frac{\eta}{\sqrt{n}}\Sigma_{21}\boldsymbol{\beta}_1^\star - W_n(2)\right| \leq \frac{\lambda}{2\sqrt{n}}. \quad (29)$$

When the event B_n is satisfied:

$$-\frac{\lambda}{2\sqrt{n}} + \frac{\lambda}{2\sqrt{n}}\left|\mathscr{C}_{21}^{n,\Sigma}\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\left(sign(\boldsymbol{\beta}_1^\star) - \frac{2\eta}{\lambda}\Sigma_{11}\boldsymbol{\beta}_1^\star\right) - \frac{2\eta}{\lambda}\Sigma_{21}\boldsymbol{\beta}_1^\star\right|$$

$$\leq \mathscr{C}_{21}^{n,\Sigma}\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}W_n(1) - W_n(2)$$

$$\leq \frac{\lambda}{2\sqrt{n}} - \frac{\lambda}{2\sqrt{n}}\left|\mathscr{C}_{21}^{n,\Sigma}\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\left(sign(\boldsymbol{\beta}_1^\star) + \frac{2\eta}{\lambda}\Sigma_{11}\boldsymbol{\beta}_1^\star\right) - \frac{2\eta}{\lambda}\Sigma_{21}\boldsymbol{\beta}_1^\star\right|. \quad (30)$$

By using that $-|x| \leq x \leq |x|$ for all x in $\mathbb{R}$, we get that it implies that

$$\left|\mathscr{C}_{21}^{n,\Sigma}\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}W_n(1) - W_n(2) - \frac{\lambda}{2\sqrt{n}}\mathscr{C}_{21}^{n,\Sigma}\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\left(sign(\boldsymbol{\beta}_1^\star) + \frac{2\eta}{\lambda}\Sigma_{11}\boldsymbol{\beta}_1^\star\right) + \frac{\eta}{\sqrt{n}}\Sigma_{21}\boldsymbol{\beta}_1^\star\right| \leq \frac{\lambda}{2\sqrt{n}},$$

which corresponds to (29). Thus, if A_n and B_n are satisfied, we get that $sign\left(\widehat{\boldsymbol{\beta}}\right) = sign\left(\boldsymbol{\beta}\right)$, which concludes the proof.

5.2   Proof of Theorem 2.2

By Lemma 2.1,

$$\mathbb{P}\left(sign\left(\widehat{\boldsymbol{\beta}}\right) = sign(\boldsymbol{\beta}^\star)\right) \geq \mathbb{P}\left(A_n \cap B_n\right) \geq 1 - \mathbb{P}\left(A_n^c\right) - \mathbb{P}\left(B_n^c\right),$$

where A_n^c and B_n^c denote the complementary of A_n and B_n, respectively. Thus, to prove the theorem it is enough to prove that $\mathbb{P}\left(A_n^c\right) \to 0$ and $\mathbb{P}\left(B_n^c\right) \to 0$ as $n \to \infty$.

Recall that

$$A_n := \left\{\left|\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}W_n(1)\right| < \sqrt{n}\left(|\boldsymbol{\beta}_1^\star| - \frac{\lambda}{2n}\left|\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}sign(\boldsymbol{\beta}_1^\star)\right| - \frac{\eta}{n}\left|\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\Sigma_{11}\boldsymbol{\beta}_1^\star\right|\right)\right\}.$$

Let ζ and τ be defined by

$$\zeta = \left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1} W_n(1) \text{ and } \tau = \sqrt{n}\left(\left|\boldsymbol{\beta}_1^{\star}\right| - \frac{\lambda}{2n}\left|\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1} sign(\boldsymbol{\beta}_1^{\star})\right| - \frac{\eta}{n}\left|\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\Sigma_{11}\boldsymbol{\beta}_1^{\star}\right|\right).$$

Then,

$$\mathbb{P}(A_n) = \mathbb{P}\left(\forall j, |\zeta_j| < \tau_j\right).$$

Thus,

$$\mathbb{P}(A_n^c) = \mathbb{P}\left(\exists j, |\zeta_j| \geq \tau_j\right) \leq \sum_{j=1}^{q} \mathbb{P}\left(|\zeta_j| \geq \tau_j\right).$$

Note that

$$\mathbb{P}(|\zeta_j| \geq \tau_j) = \mathbb{P}\left(|\zeta_j| \geq \sqrt{n}\left(\left|(\boldsymbol{\beta}_1^{\star})_j\right| - \frac{\lambda}{2n}\left|\left(\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1} sign(\boldsymbol{\beta}_1^{\star})\right)_j\right| - \frac{\eta}{n}\left|\left(\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\Sigma_{11}\boldsymbol{\beta}_1^{\star}\right)_j\right|\right)\right)$$

$$= \mathbb{P}\left(|\zeta_j| + \frac{\lambda}{2\sqrt{n}}\left|\left(\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1} sign(\boldsymbol{\beta}_1^{\star})\right)_j\right| + \frac{\eta}{\sqrt{n}}\left|\left(\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\Sigma_{11}\boldsymbol{\beta}_1^{\star}\right)_j\right| \geq \sqrt{n}\left|(\boldsymbol{\beta}_1^{\star})_j\right|\right)$$

$$\leq \mathbb{P}\left(|\zeta_j| \geq \sqrt{n}\frac{\left|(\boldsymbol{\beta}_1^{\star})_j\right|}{3}\right) + \mathbb{P}\left(\frac{\lambda}{2\sqrt{n}}\left|\left(\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1} sign(\boldsymbol{\beta}_1^{\star})\right)_j\right| \geq \sqrt{n}\frac{\left|(\boldsymbol{\beta}_1^{\star})_j\right|}{3}\right)$$

$$+ \mathbb{P}\left(\frac{\eta}{\sqrt{n}}\left|\left(\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\Sigma_{11}\boldsymbol{\beta}_1^{\star}\right)_j\right| \geq \sqrt{n}\frac{\left|(\boldsymbol{\beta}_1^{\star})_j\right|}{3}\right). \tag{31}$$

Observe that

$$\zeta = \left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1} W_n(1) = \frac{1}{\sqrt{n}}\left(C_{11}^n + \frac{\eta}{n}\Sigma_{11}\right)^{-1} X_1'\epsilon = H_A\epsilon,$$

where

$$H_A = \frac{1}{\sqrt{n}}\left(C_{11}^n + \frac{\eta}{n}\Sigma_{11}\right)^{-1} X_1',$$

X_1 denoting the columns of the design matrix X associated to the q active covariates. Thus, for all j in $\{1, \ldots, q\}$,

$$\zeta_j = \sum_{k=1}^{n} (H_A)_{jk}\epsilon_k.$$

By using the Cauchy-Schwarz inequality,

$$|\zeta_j| = \left|\sum_{k=1}^{n} (H_A)_{jk}\epsilon_k\right| \leq \left(\sum_{k=1}^{n} (H_A)_{jk}^2\right)^{1/2}\left(\sum_{k=1}^{n} \epsilon_k^2\right)^{1/2}$$

$$= \sqrt{\left(H_A H_A'\right)_{jj}} \times \|\epsilon\|_2$$

$$\leq \sqrt{\lambda_{\max}\left(H_A H_A'\right)} \times \|\epsilon\|_2.$$

16

Hence, the first term in the r.h.s. of (31) satisfies the following inequalities:

$$\mathbb{P}\left(|\zeta_j| \geq \sqrt{n}\frac{|(\boldsymbol{\beta}_1^\star)_j|}{3}\right) \leq \mathbb{P}\left(\sqrt{\lambda_{\max}(H_A H_A')} \times \|\boldsymbol{\epsilon}\|_2 \geq \sqrt{n}\frac{|(\boldsymbol{\beta}_1^\star)_j|}{3}\right)$$

$$\leq \mathbb{P}\left(\lambda_{\max}(H_A H_A') \times \|\boldsymbol{\epsilon}\|_2^2 \geq n\frac{(\boldsymbol{\beta}_1^\star)_j^2}{9}\right). \tag{32}$$

Since by (19), there exist $M_1 > 0$ and $\delta_1 > 0$ such that

$$\mathbb{P}\left(\lambda_{\max}(H_A H_A') \leq M_1\right) = 1 - o\left(e^{-n^{\delta_1}}\right), \text{ as } n \to \infty,$$

we have:

$$\mathbb{P}\left(\lambda_{\max}(H_A H_A') \times \|\boldsymbol{\epsilon}\|_2^2 \geq n\frac{(\boldsymbol{\beta}_1^\star)_j^2}{9}\right) \leq \mathbb{P}\left(\|\boldsymbol{\epsilon}\|_2^2 \geq n\frac{(\boldsymbol{\beta}_1^\star)_j^2}{9M_1}\right) + \mathbb{P}\left(\lambda_{\max}(H_A H_A') > M_1\right)$$

$$\leq \mathbb{P}\left(\frac{\|\boldsymbol{\epsilon}\|_2^2}{\sigma^2} \geq \frac{n\boldsymbol{\beta}_{\min}^2}{9M_1\sigma^2}\right) + o\left(e^{-n^{\delta_1}}\right).$$

Using that $\dfrac{\|\boldsymbol{\epsilon}\|_2^2}{\sigma^2} \sim \chi^2(n)$, we get, by Lemma 1 of [29], that

$$\mathbb{P}\left(\lambda_{\max}(H_A H_A') \times \|\boldsymbol{\epsilon}\|_2^2 \geq n\frac{(\boldsymbol{\beta}_1^\star)_j^2}{9}\right) \leq \exp\left(-\frac{t}{2} + \frac{1}{2}\sqrt{n(2t-n)}\right) + o\left(e^{-n^{\delta_1}}\right), \tag{33}$$

since $t = \dfrac{n\boldsymbol{\beta}_{\min}^2}{9M_1\sigma^2} > \dfrac{n}{2}$ using that $\dfrac{2\boldsymbol{\beta}_{\min}^2}{9\sigma^2} > M_1$ by (15).

By putting together Equations (32) and (33) we get

$$\mathbb{P}\left(|\zeta_j| > \sqrt{n}\frac{|(\boldsymbol{\beta}_1^\star)_j|}{3}\right) \leq \exp\left(-\frac{t}{2} + \frac{1}{2}\sqrt{n(2t-n)}\right) + o\left(e^{-n^{\delta_1}}\right), \tag{34}$$

with $t = \dfrac{n\boldsymbol{\beta}_{\min}^2}{9M_1\sigma^2} > \dfrac{n}{2}$.

Let us now derive an upper bound for the second term in the r.h.s. of (31):

$$\mathbb{P}\left(\frac{\lambda}{2\sqrt{n}}\left|\left((\mathscr{C}_{11}^{n,\Sigma})^{-1} sign(\boldsymbol{\beta}_1^\star)\right)_j\right| \geq \sqrt{n}\frac{|(\boldsymbol{\beta}_1^\star)_j|}{3}\right).$$

By using the Cauchy-Schwarz inequality, we get that:

$$\left|\left((\mathscr{C}_{11}^{n,\Sigma})^{-1} sign(\boldsymbol{\beta}_1^\star)\right)_j\right| = \left|\sum_{k=1}^q \left((\mathscr{C}_{11}^{n,\Sigma})^{-1}\right)_{jk} (sign(\boldsymbol{\beta}_1^\star))_k\right|$$

$$\leq \sqrt{\sum_{k=1}^q \left((\mathscr{C}_{11}^{n,\Sigma})^{-1}\right)_{jk}^2} \times \|sign(\boldsymbol{\beta}_1^\star)\|_2 \leq \sqrt{\left(\mathscr{C}_{11}^{n,\Sigma}\right)_{jj}^{-2}} \times \sqrt{q}$$

$$\leq \lambda_{\max}\left((\mathscr{C}_{11}^{n,\Sigma})^{-1}\right) \times \sqrt{q}.$$

17

Then,

$$\mathbb{P}\left(\frac{\lambda}{2\sqrt{n}}\left|\left(\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}sign(\boldsymbol{\beta}_1^\star)\right)_j\right| \geq \sqrt{n}\frac{|(\boldsymbol{\beta}_1^\star)_j|}{3}\right)$$

$$\leq \mathbb{P}\left(\frac{\lambda}{2}\sqrt{\frac{q}{n}}\lambda_{\max}\left(\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\right) \geq \sqrt{n}\frac{|(\boldsymbol{\beta}_1^\star)_j|}{3}\right)$$

$$\leq \mathbb{P}\left(\lambda_{\max}\left(\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\right) \geq \frac{2n}{3\lambda\sqrt{q}}|(\boldsymbol{\beta}_1^\star)_j|\right)$$

$$\leq \mathbb{P}\left(\lambda_{\max}\left(\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\right) \geq \frac{2n}{3\lambda\sqrt{q}}\boldsymbol{\beta}_{\min}\right) = o\left(e^{-n^{\delta_2}}\right), \text{ as } n \to \infty, \tag{35}$$

since $\dfrac{2n}{3\lambda\sqrt{q}}\boldsymbol{\beta}_{\min} > M_2$ by (16). Let us now derive an upper bound for the third term in the r.h.s. of (31):

$$\mathbb{P}\left(\frac{\eta}{\sqrt{n}}\left|\left(\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\Sigma_{11}\boldsymbol{\beta}_1^\star\right)_j\right| > \sqrt{n}\frac{|(\boldsymbol{\beta}_1^\star)_j|}{3}\right).$$

We have

$$\left|\left(\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\Sigma_{11}\boldsymbol{\beta}_1^\star\right)_j\right| = \left|\sum_{k=1}^q\left(\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\Sigma_{11}\right)_{jk}(\boldsymbol{\beta}_1^\star)_k\right| \leq \sqrt{\sum_{k=1}^q\left(\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\Sigma_{11}\right)_{jk}^2} \times \|\boldsymbol{\beta}_1^\star\|_2$$

$$\leq \sqrt{\lambda_{\max}\left(\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\Sigma_{11}^2\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\right)} \times \|\boldsymbol{\beta}_1^\star\|_2 \leq \lambda_{\max}\left(\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\right)\lambda_{\max}(\Sigma_{11}) \times \|\boldsymbol{\beta}_1^\star\|_2.$$

Thus,

$$\mathbb{P}\left(\frac{\eta}{\sqrt{n}}\left|\left(\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\Sigma_{11}\boldsymbol{\beta}_1^\star\right)_j\right| \geq \sqrt{n}\frac{|(\boldsymbol{\beta}_1^\star)_j|}{3}\right)$$

$$\leq \mathbb{P}\left(\frac{\eta}{\sqrt{n}}\lambda_{\max}\left(\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\right)\lambda_{\max}(\Sigma_{11})\|\boldsymbol{\beta}_1^\star\|_2 \geq \sqrt{n}\frac{|(\boldsymbol{\beta}_1^\star)_j|}{3}\right)$$

$$\leq \mathbb{P}\left(\lambda_{\max}\left(\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\right) \geq \frac{n\boldsymbol{\beta}_{\min}}{3\eta\|\boldsymbol{\beta}_1^\star\|_2\lambda_{\max}(\Sigma_{11})}\right) = o\left(e^{-n^{\delta_2}}\right), \text{ as } n \to \infty, \tag{36}$$

since $\dfrac{n\boldsymbol{\beta}_{\min}}{3\eta\|\boldsymbol{\beta}_1^\star\|_2\lambda_{\max}(\Sigma_{11})} > M_2$ by (18).

By putting together Equations (34), (35) and (36), we get:

$$\mathbb{P}(A_n^c) \leq q\exp\left[-\frac{n}{2}\left(\kappa - \sqrt{2\kappa - 1}\right)\right] + q \times o\left(e^{-n^{\delta_1}}\right) + 2q \times o\left(e^{-n^{\delta_2}}\right), \tag{37}$$

with $\kappa = \dfrac{\boldsymbol{\beta}_{\min}^2}{9M_1\sigma^2}$. Note that $\kappa - \sqrt{2\kappa - 1} > 0$ since $\kappa = \dfrac{\boldsymbol{\beta}_{\min}^2}{9M_1\sigma^2} > 1$ by (15). Equation (37) then implies that

$$\mathbb{P}(A_n^c) \to 0 \text{ as } n \to \infty.$$

18

Let us now prove that $\mathbb{P}\left(B_n^c\right) \to 0$ as $n \to \infty$.
Recall that

$$B_n := \left\{ \left| \mathscr{C}_{21}^{n,\Sigma} \left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1} W_n(1) - W_n(2) \right| \leq \frac{\lambda}{2\sqrt{n}} \right.$$

$$\left. -\frac{\lambda}{2\sqrt{n}} \left| \mathscr{C}_{21}^{n,\Sigma} \left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1} \left(sign(\boldsymbol{\beta}_1^\star) + \frac{2\eta}{\lambda}\Sigma_{11}\boldsymbol{\beta}_1^\star \right) - \frac{2\eta}{\lambda}\Sigma_{21}\boldsymbol{\beta}_1^\star \right| \right\}.$$

Let

$$\boldsymbol{\psi} = \mathscr{C}_{21}^{n,\Sigma} \left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1} W_n(1) - W_n(2) = \frac{1}{\sqrt{n}} \left(\mathscr{C}_{21}^{n,\Sigma} \left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1} \mathbf{X}_1' - \mathbf{X}_2' \right) \boldsymbol{\epsilon} =: H_B \boldsymbol{\epsilon}$$

and

$$\boldsymbol{\mu} = \frac{\lambda}{2\sqrt{n}} - \frac{\lambda}{2\sqrt{n}} \left| \mathscr{C}_{21}^{n,\Sigma} \left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1} \left(sign(\boldsymbol{\beta}_1^\star) + \frac{2\eta}{\lambda}\Sigma_{11}\boldsymbol{\beta}_1^\star \right) - \frac{2\eta}{\lambda}\Sigma_{21}\boldsymbol{\beta}_1^\star \right|.$$

Then,

$$\mathbb{P}(B_n^c) = \mathbb{P}(\exists j, |\psi_j| > \mu_j) \leq \sum_{j=q+1}^{p} \mathbb{P}(|\psi_j| > \mu_j).$$

By using the Cauchy-Schwarz inequality, we get that:

$$|\psi_j| = \left| \sum_{k=1}^{n} (H_B)_{jk} \epsilon_k \right| \leq \left(\sum_{k=1}^{n} (H_B)_{jk}^2 \right)^{1/2} \times \|\boldsymbol{\epsilon}\|_2 = \sqrt{\left(H_B H_B'\right)_{jj}} \times \|\boldsymbol{\epsilon}\|_2 \leq \sqrt{\lambda_{\max}(H_B H_B')} \times \|\boldsymbol{\epsilon}\|_2,$$

$$(38)$$

where

$$H_B H_B' = \mathscr{C}_{21}^{n,\Sigma} \left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1} C_{11}^n \left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1} \mathscr{C}_{12}^{n,\Sigma} - \mathscr{C}_{21}^{n,\Sigma} \left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1} C_{12}^n - C_{21}^n \left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1} \mathscr{C}_{12}^{n,\Sigma} + C_{22}^n.$$

By (21), there exist $M_3 > 0$ and $\delta_3 > 0$ such that

$$\mathbb{P}\left(\lambda_{\max}\left(H_B H_B'\right) \leq M_3\right) = 1 - o\left(e^{-n^{\delta_3}}\right), \text{ as } n \to \infty.$$

By the GIC condition (12), there exist $\alpha > 0$ and $\delta_4 > 0$ such that for all j,

$$\mathbb{P}\left(\left| \mathscr{C}_{21}^{n,\Sigma} \left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1} \left(sign(\boldsymbol{\beta}_1^\star) + \frac{2\eta}{\lambda}\Sigma_{11}\boldsymbol{\beta}_1^\star \right) - \frac{2\eta}{\lambda}\Sigma_{21}\boldsymbol{\beta}_1^\star \right| \leq 1 - \alpha \right) = 1 - o\left(e^{-n^{\delta_4}}\right).$$

Thus, we get that:

$$
\begin{aligned}
\mathbb{P}(B_n^c) \;&\leq\; \sum_{j=q+1}^{p} \mathbb{P}\left(|\psi_j| > \mu_j\right) \\
&\leq\; \sum_{j=q+1}^{p} \mathbb{P}\left(|\psi_j| > \frac{\lambda\alpha}{2\sqrt{n}}\right) + (p-q)o\left(e^{-n^{\delta_4}}\right) \\
&\leq\; \sum_{j=q+1}^{p} \mathbb{P}\left(\sqrt{\lambda_{\max}(H_B H_B')} \times \|\boldsymbol{\epsilon}\|_2 > \frac{\lambda\alpha}{2\sqrt{n}}\right) + (p-q)o\left(e^{-n^{\delta_4}}\right), \text{ using Equation (38)} \\
&\leq\; \sum_{j=q+1}^{p} \mathbb{P}\left(\lambda_{\max}(H_B H_B') \times \|\boldsymbol{\epsilon}\|_2^2 > \frac{\lambda^2\alpha^2}{4n}\right) + (p-q)o\left(e^{-n^{\delta_4}}\right) \\
&\leq\; \sum_{j=q+1}^{p} \mathbb{P}\left(\frac{\|\boldsymbol{\epsilon}\|_2^2}{\sigma^2} > \frac{\lambda^2\alpha^2}{4nM_3\sigma^2}\right) + (p-q)o\left(e^{-n^{\delta_3}}\right) + (p-q)o\left(e^{-n^{\delta_4}}\right) \\
&\leq\; (p-q)\exp\left(-\frac{s}{2} + \frac{1}{2}\sqrt{n(2s-n)}\right) + (p-q)o\left(e^{-n^{\delta_3}}\right) + (p-q)o\left(e^{-n^{\delta_4}}\right) \\
&\leq\; (p-q)\exp\left(-\frac{n}{2}\left(\frac{s}{n} - \sqrt{2\frac{s}{n}-1}\right)\right) + (p-q)o\left(e^{-n^{\delta_3}}\right) + (p-q)o\left(e^{-n^{\delta_4}}\right) \\
&\leq\; (p-q)\exp\left(-\frac{n}{2}\right) + (p-q)o\left(e^{-n^{\delta_3}}\right) + (p-q)o\left(e^{-n^{\delta_4}}\right) \qquad (39)
\end{aligned}
$$

with $\dfrac{s}{n} = \dfrac{\lambda^2\alpha^2}{4n^2 M_3\sigma^2}$ since $\dfrac{\lambda^2\alpha^2}{4n^2 M_3\sigma^2} \geq 2 + \sqrt{2}$ by (17). Finally, Equation (39) implies that

$$
\mathbb{P}(B_n^c) \to 0, \text{ as } n \to \infty,
$$

which concludes the proof.

References

[1] Jinzhu Jia and B. Yu. On model selection consistency of the elastic net when $p \gg n$. Statistica Sinica, 20, 04 2010.

[2] Wenbin Lu, Hao Zhang, and Donglin Zeng. Variable selection for optimal treatment decision. Statistical methods in medical research, 22, 11 2011.

[3] Lacey Gunter, Ji Zhu, and Susan Murphy. Variable selection for qualitative interactions in personalized medicine while controlling the family-wise error rate. Journal of biopharmaceutical statistics, 21(6):1063–1078, 2011.

[4] Xuemin Gu, Guosheng Yin, and J Jack Lee. Bayesian two-step lasso strategy for biomarker selection in personalized medicine development for time-to-event endpoints. Contemporary clinical trials, 36(2):642–650, 2013.

[5] Wencan Zhu, Céline Lévy-Leduc, and Nils Ternès. A variable selection approach for highly correlated predictors in high-dimensional genomic data. Bioinformatics, 2021. doi: 10.1093/bioinformatics/btab114. Epub ahead of print. PMID: 33617644.

[6] Zeynep Tufekci. Big questions for social media big data: Representativeness, validity and other methodological pitfalls. Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014, 03 2014.

[7] Huijie Lin, Jia Jia, Liqiang Nie, Guangyao Shen, and Tat-Seng Chua. What does social media say about your stress?. In IJCAI, pages 3775–3781, 2016.

[8] Theodore S Tomeny, Christopher J Vargo, and Sherine El-Toukhy. Geographic and demographic correlates of autism-related anti-vaccine beliefs on twitter, 2009-15. Social science & medicine, 191:168–175, 2017.

[9] Georgios Sermpinis, Serafeim Tsoukas, and Ping Zhang. Modelling market implied ratings using lasso variable selection techniques. Journal of Empirical Finance, 48:19–35, 2018.

[10] Alessandra Amendola, Francesco Giordano, Maria Parrella, and Marialuisa Restaino. Variable selection in high-dimensional regression: A nonparametric procedure for business failure prediction. Applied Stochastic Models in Business and Industry, 33, 02 2017.

[11] Bartosz Uniejewski, Grzegorz Marcjasz, and Rafał Weron. Understanding intraday electricity markets: Variable selection and very short-term price forecasting using lasso. International Journal of Forecasting, 35(4):1533–1547, 2019.

[12] Norman R Draper and Harry Smith. Applied regression analysis, volume 326. John Wiley & Sons, 1998.

[13] Peter J Bickel, Bo Li, Alexandre B Tsybakov, Sara A van de Geer, Bin Yu, Teófilo Valdés, Carlos Rivero, Jianqing Fan, and Aad van der Vaart. Regularization in statistics. Test, 15(2):271–344, 2006.

[14] Hirotogu Akaike. Information theory and an extension of the maximum likelihood principle. In Selected papers of hirotugu akaike, pages 199–213. Springer, 1998.

[15] Gideon Schwarz et al. Estimating the dimension of a model. Annals of statistics, 6(2):461–464, 1978.

[16] William J Welch. Algorithmic complexity: three np-hard problems in computational statistics. Journal of Statistical Computation and Simulation, 15(1):17–25, 1982.

[17] Leo Breiman et al. Heuristics of instability and stabilization in model selection. Annals of Statistics, 24(6):2350–2383, 1996.

[18] Robert Tibshirani. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58:267–288, 01 1996.

[19] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. Statistical Learning with Sparsity: The Lasso and Generalizations. CRC Press, Boca Raton, 05 2015.

[20] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society Series B, 67:768–768, 02 2005.

[21] Tong Tong Wu, Yi Fang Chen, Trevor Hastie, Eric Sobel, and Kenneth Lange. Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics, 25(6):714–721, 01 2009.

[22] Peng Zhao and B. Yu. On model selection consistency of lasso. Journal of Machine Learning Research, 7:2541–2563, 12 2006.

[23] Martin J Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using l1-constrained quadratic programming (lasso). IEEE transactions on information theory, 55(5):2183–2202, 2009.

[24] Ming Yuan and Yi Lin. On the nonnegative garrote estimator. Journal of the Royal Statistical Society Series B, 69:143–161, 04 2007.

[25] Fei Xue and Annie Qu. Variable selection for highly correlated predictors. arXiv preprint arXiv:1709.04840, 2017.

[26] Jinzhu Jia and Karl Rohe. Preconditioning the lasso for sign consistency. Electron. J. Statist., 9(1):1150–1172, 2015.

[27] Xiangyu Wang and Chenlei Leng. High dimensional ordinary least squares projection for screening variables. J. R. Stat. Soc. Ser. B (Stat. Methodol.), 78(3):589–611, 2016.

[28] Ryan J. Tibshirani and Jonathan Taylor. The solution path of the generalized lasso. Ann. Statist., 39(3):1335–1371, 06 2011.

[29] Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. Annals of Statistics, pages 1302–1338, 2000.