Model-Consistent Sparse Estimation through the Bootstrap

Francis Bach

Model-Consistent Sparse Estimation through the Bootstrap

Francis Bach
Willow Project-team
Laboratoire d'Informatique de l'Ecole Normale Supérieure
(CNRS/ENS/INRIA UMR 8548)
45, rue d'Ulm, 75230 Paris, France
francis.bach@mines.org

January 20, 2009

Abstract

We consider the least-square linear regression problem with regularization by the ℓ^1-norm, a problem usually referred to as the Lasso. In this paper, we first present a detailed asymptotic analysis of model consistency of the Lasso in low-dimensional settings. For various decays of the regularization parameter, we compute asymptotic equivalents of the probability of correct model selection. For a specific rate decay, we show that the Lasso selects all the variables that should enter the model with probability tending to one exponentially fast, while it selects all other variables with strictly positive probability. We show that this property implies that if we run the Lasso for several bootstrapped replications of a given sample, then intersecting the supports of the Lasso bootstrap estimates leads to consistent model selection. This novel variable selection procedure, referred to as the Bolasso, is extended to high-dimensional settings by a provably consistent two-step procedure.

1 Introduction

Regularization by the ℓ^1-norm has attracted a lot of interest in recent years in statistics, machine learning and signal processing. In the context of least-square linear regression, the problem is usually referred to as the *Lasso* [36] or *basis pursuit* [14]. Much of the early effort has been dedicated to algorithms to solve the optimization problem efficiently, either through first-order methods [20, 19], or through homotopy methods that leads to the entire regularization path (i.e., the set of solutions for all values of the regularization parameters) at the cost of a single matrix inversion [29, 35, 16].

A well-known property of the regularization by the ℓ^1-norm is the *sparsity* of the solutions, i.e., it leads to loading vectors with many zeros, and thus performs model selection on top of

regularization. Recent works [44, 40, 45, 37] have looked precisely at the model consistency of the Lasso, i.e., if we know that the data were generated from a sparse loading vector, does the Lasso actually recover the sparsity pattern when the number of observations grows? In the case of a fixed number of covariates (i.e., low-dimensional settings), the Lasso does recover the sparsity pattern if and only if a certain simple condition on the generating covariance matrices is satisfied [40]. In particular, in low correlation settings, the Lasso is indeed consistent. However, in presence of strong correlations between relevant variables and irrelevant variables, the Lasso cannot be model-consistent, shedding light on potential problems of such procedures for variable selection. Various extensions of the Lasso have been designed to fix its inconsistency, based on thresholding [34], data-dependent weights [45, 40, 26] or two-step procedures [31]. The main contribution of this paper is to propose and analyze an alternative approach based on resampling. Note that recent work [33] has also looked at resampling methods for the Lasso, but focuses on resampling the weights of the ℓ^1-norm rather than resampling the observations (see Section 3 for more details).

In this paper, we first derive a detailed asymptotic analysis of sparsity pattern selection of the Lasso estimation procedure, that extends previous analysis [44, 40, 45] by focusing on a specific decay of the regularization parameter. Namely, in *low-dimensional* settings where the number of variables p is much smaller than the number of observations n, we show that when the decay of n is proportional to $n^{-1/2}$, then the Lasso will select all the variables that should enter the model (the *relevant* variables) with probability tending to one exponentially fast with n, while it selects all other variables (the *irrelevant* variables) with strictly positive probability. If several datasets generated from the same distribution were available, then the latter property would suggest to consider the intersection of the supports of the Lasso estimates for each dataset: all relevant variables would always be selected for all datasets, while irrelevant variables would enter the models randomly, and intersecting the supports from sufficiently many different datasets would simply eliminate them. However, in practice, only one dataset is given; but resampling methods such as the *bootstrap* are exactly dedicated to mimic the availability of several datasets by resampling from the same unique dataset [17]. In this paper, we show that when using the bootstrap and intersecting the supports, we actually get a consistent model estimate, *without* the consistency condition required by the regular Lasso. We refer to this new procedure as the *Bolasso* (**bo**otstrap-enhanced **l**east **a**b**s**olute **s**hrinkage **o**perator). Finally, our Bolasso framework could be seen as a voting scheme applied to the supports of the bootstrap Lasso estimates; however, our procedure may rather be considered as a consensus combination scheme, as we keep the (largest) subset of variables on which *all* regressors agree in terms of variable selection, which is in our case provably consistent and also allows to get rid of a potential additional hyperparameter.

We consider the two usual ways of using the bootstrap in regression settings, namely bootstrapping pairs and bootstrapping residuals [17, 18]. In Section 3, we show that the two types of bootstrap lead to consistent model selection in low-dimensional settings. Moreover, in Section 5, we provide empirical evidence that in high-dimensional settings, bootstrapping pairs does not lead to consistent estimation, while bootstrapping residuals still does. While we are currently unable to prove the consistency of bootstrapping residuals in high-dimensional settings, we prove in Section 4 the model consistency of a related two-step procedure: the Lasso is run once on the original data, with a larger regularization parameter, and then bootstrap

replications (pairs or residuals) are run within the support of the first Lasso estimation. We show in Section 4 that this procedure is consistent. In order to do so, we consider new sufficient conditions for the consistency of the Lasso, which do not rely on sparse eigenvalues [34, 41], low correlations [12, 27] or finer conditions [6, 15, 42]. In particular, our new assumptions allow to prove that the Lasso will select not only a few variables when the regularization parameter is properly chosen, but always the same variables with high probability.

In Section 5.1, we derive efficient algorithms for the bootstrapped versions of the Lasso. When bootstrapping pairs, we simply run an efficient homotopy algorithm, such as *Lars* [16], multiple times; however, when bootstrapping residuals, more efficient ways may be designed to obtain a running time complexity which is less than running Lars multiple times. Finally, in Section 5.2 and Section 5.3, we illustrate our results on synthetic examples, in low-dimensional and high-dimensional settings. This work is a follow-up to earlier work [1]: in particular, it refines and extends the analysis to high-dimensional settings and boostrapping of the residuals.

Notations For $x \in \mathbb{R}^p$ and $q > 0$, we denote by $\|x\|_q$ its ℓ^q-norm, defined as $\|x\|_q^q = \sum_{i=1}^p |x_i|^q$. We also denote by $\|x\|_\infty = \max_{i\in\{1,\dots,p\}} |x_i|$ its ℓ^∞-norm. For rectangular matrices A, we denote by $\|A\|_2$ its largest singular value, $\|A\|_\infty$ the largest magnitude of all its elements, and $\|A\|_F = (\mathrm{tr} A^\top A)^{1/2}$ its Frobenius norm. We let denote $\lambda_{\max}(Q)$ and $\lambda_{\min}(Q)$ the largest and smallest eigenvalue of a symmetric matrix Q.

For $a \in \mathbb{R}$, $\mathrm{sign}(a)$ denotes the sign of a, defined as $\mathrm{sign}(a) = 1$ if $a > 0$, -1 if $a < 0$, and 0 if $a = 0$. For a vector $v \in \mathbb{R}^p$, $\mathrm{sign}(v) \in \{-1, 0, 1\}^p$ denotes the vector of signs of elements of v. Given a set H, 1_H is the indicator function of the set H. We also denote, for $w \in \mathbb{R}^p$, by $\mathrm{m}(w) = \min_{j\in\{1,\dots,p\},\ w_j\neq 0} |w_j|$, the smallest (in magnitude) of non-zero elements of w.

Moreover, given a vector $v \in \mathbb{R}^p$ and a subset I of $\{1, \dots, p\}$, v_I denotes the vector in $\mathbb{R}^{\mathrm{Card}(I)}$ of elements of v indexed by I. Similarly, for a matrix $A \in \mathbb{R}^{p\times p}$, $A_{I,J}$ denotes the submatrix of A composed of elements of A whose rows are in I and columns are in J. Moreover, $|J|$ denotes the cardinal of the set J. For a positive definite matrix Q of size p, and two disjoint subsets of indices A and B included in $\{1, \dots, p\}$, we denote $Q_{A,A|B}$ the matrix $Q_{A,A} - Q_{A,B}Q_{B,B}^{-1}Q_{B,A}$, which is the conditional covariance of variables indexed by A given variables indexed by B, for a Gaussian vector with covariance matrix Q. Finally, we let denote $\mathbb{P}$ and $\mathbb{E}$ general probability measures and expectations.

Least-square regression with ℓ^1-norm penalization Throughout this paper, we consider n pairs of observations $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \dots, n$. The data are given in the form of a vector $y \in \mathbb{R}^n$ and a design matrix $X \in \mathbb{R}^{n\times p}$. We consider the normalized square loss function

$$\frac{1}{2n}\sum_{i=1}^n (y_i - w^\top x_i)^2 = \frac{1}{2n}\|y - Xw\|_2^2,$$

and the regularization by the ℓ^1-norm. That is, we look at the following convex optimization problem [36, 14]:

$$\min_{w\in\mathbb{R}^p} \frac{1}{2n}\|y - Xw\|_2^2 + \mu\|w\|_1, \tag{1.1}$$

where $\mu \geqslant 0$ is the regularization parameter. We denote by $\hat{w}$ any global minimum of Eq. (1.1), and $\hat{J} = \{j \in \{1, \dots, p\},\ \hat{w}_j \neq 0\}$ the support of $\hat{w}$.

In this paper, we consider two settings, depending on the value of the ratio of p/n. When this ratio is much smaller than one, as in Section 2, we refer to this setting as low-dimensional estimation, while in other cases, where this ratio is potentially much larger than one, we refer to this setting as a high-dimensional problem (see Section 4).

2 Low-Dimensional Asymptotic Analysis

We make the following "fixed-design" assumptions:

(**A1**) *Linear model with i.i.d. additive noise*: $y = X\mathbf{w} + \varepsilon$, where ε is a vector with independent components, identical distributions and zero mean; $\mathbf{w}$ is sparse, with $\mathbf{s} = \operatorname{sign}(\mathbf{w})$ and support $\mathbf{J} = \{j, \mathbf{w}_j \neq 0\}$.

(**A2**) *Subgaussian noise*: there exists $\tau > 0$ such that for all $j \in \{1, \dots, p\}$ and $s \in \mathbb{R}$, $\mathbb{E}e^{s\varepsilon_j} \leqslant e^{\frac{1}{2}\tau^2 s^2}$. Moreover, the variances of ε_j are greater than $\sigma^2 > 0$.

(**A3**) *Bounded design*: For all $i \in \{1, \dots, n\}$, $\|x_i\|_\infty \leqslant M$.

(**A4**) *Full rank design*: The matrix $Q = \frac{1}{n} X^\top X \in \mathbb{R}^{p \times p}$ is invertible.

Throughout this paper, we consider normalized constants $\tilde{\mathbf{w}} = \mathbf{w}M/\sigma$ (normalized population loading vector), $\tilde{\mu} = \mu/M\sigma$ (normalized regularization parameter), $\tilde{\lambda} = \lambda_{\min}(Q)/M^2$ (condition number of the matrix of second-order moments), and $\tilde{\tau} = \tau/\sigma$ (always larger than one, and equal to one if and only if the noise is Gaussian, see Appendix A.2).

With our assumptions, the problem in Eq. (1.1) is equivalent to

$$\min_{w \in \mathbb{R}^p} \frac{1}{2}(w - \mathbf{w})^\top Q (w - \mathbf{w}) - q^\top (w - \mathbf{w}) + \mu \|w\|_1, \tag{2.1}$$

where $Q = \frac{1}{n} X^\top X \in \mathbb{R}^{p \times p}$ and $q = \frac{1}{n} X^\top \varepsilon \in \mathbb{R}^p$. Note that under assumption (**A4**), there is a unique solution to Eq. (1.1) and Eq. (2.1), because the associated objective functions are then strongly convex. Moreover, assumption (**A4**) implies that $p \geqslant n$, that is, we consider in this section, only "low-dimensional" settings (see Section 4 for extensions to high-dimensional settings).

In this section, we detail the asymptotic behavior of the (unique) Lasso estimate $\hat{w}$, both in terms of the difference in norm with the population value $\mathbf{w}$ (i.e., regular consistency) and of the *sign pattern* $\operatorname{sign}(\hat{w})$, for all asymptotic behaviors of the regularization parameter μ. Note that information about the sign pattern includes information about the *support* $\hat{J}$, i.e., the indices $j \in \{1, \dots, p\}$ for which $\hat{w}_j$ is different from zero; moreover, when $\hat{w}$ is consistent, consistency of the sign pattern is in fact equivalent to the consistency of the support. We assume that p is fixed and n tends to infinity, the regularization parameter μ being considered as a function of n (though we still derive non-asymptotic bounds).

Note that for some of our results to be non trivial, we require that p is not only small compared to n, but that a power of p is small compared to n. Technically, this is due to the application of multivariate Berry-Esseen inequalities (reviewed in Appendix A.1), which could probably be improved to obtain smaller powers.

We consider five mutually exclusive possible situations which explain various portions of the regularization path; many of these results appear elsewhere [40, 44, 21, 45, 2, 27] but some of the finer results presented below are new (in particular most non-asymptotic results and the $n^{-1/2}$-decay of the regularization parameter in Section 2.4). These results are illustrated on synthetic examples in Section 5.2.

Note that all exponential convergences have a rate that depends on $\mathrm{m}(\mathbf{w})$, i.e., the smallest (in magnitude) non zero element of the generating sparse vector $\mathbf{w}$. Thus, we assume a sharp threshold in order to have a fast rate of convergence. Considering situations without such a threshold, which would notably require to estimate errors in model estimation (and not simply exactly correct or incorrect), is out of the scope of this paper (see, e.g., [41]).

2.1 Heavy regularization

If μ is large enough, then $\hat{w}$ is equal to zero with probability tending to one exponentially fast in n. Indeed, we have (see proof in Appendix D.1):

Proposition 2.1. *Assume (**A1-4**). If $\tilde{\mu} \geqslant 2\|\tilde{\mathbf{w}}\|_1$, then the probability that $\hat{w} = 0$ is greater than $1 - 2p\exp\left(-\frac{n\tilde{\mu}^2}{8\tilde{\tau}^2}\right)$.*

A well-known property of homotopy algorithms for the Lasso (see, e.g., [16]) is that if μ is large enough, then $\hat{w} = 0$. This proposition simply provides a uniform probabilistic bound.

2.2 Fixed regularization

If μ tends to a finite strictly positive constant μ_0, then $\hat{w}$ converges in probability to the unique global minimum of the noiseless objective function $\frac{1}{2}(w - \mathbf{w})^\top Q(w - \mathbf{w}) + \mu_0\|w\|_1$. Thus, the estimate $\hat{w}$ never converges in probability to $\mathbf{w}$, while the sign pattern tends to the one of the previous global minimum, which may or may not be the same as the one of the noiseless problem $\mathbf{w}$. It is thus possible, though not desirable, to have sign consistency without regular consistency. See [2] for examples and simulations of a similar behavior for the group Lasso.

All convergences are exponentially fast in n (proof in Appendix D.2). Note that here and in the next regime (Proposition 2.3), we do not take into account the pathological cases where the sign pattern of the limit in unstable, i.e., the limit is exactly at a hinge point of the regularization path. When this occurs, all associated sign patterns are attained with positive probability (see also Section 4).

Proposition 2.2. *Assume (**A1-4**). Let $\mu_0 > 0$ and $\tilde{\mu}_0 = \mu_0/M/sigma$. Let w_0 be the unique solution of $\min_{v\in\mathbb{R}^p} \frac{1}{2}(v - \mathbf{w})^\top Q(v - \mathbf{w}) + \mu_0\|v\|_1$. Then, if $|\tilde{\mu} - \tilde{\mu}_0| \leqslant \frac{\tilde{\lambda}}{4p^{1/2}}\beta$, we have:*

$$\mathbb{P}(\|\hat{w} - w_0\|_2 \geqslant \beta\sigma/M) \leqslant 2p\exp\left(-\frac{\tilde{\lambda}^2\beta^2}{32\tilde{\tau}^2}\frac{n}{p}\right) \leqslant 2p\exp\left(-\frac{(\tilde{\mu} - \tilde{\mu}_0)^2}{2\tilde{\tau}^2}n\right).$$

Moreover, assume the minimum v occurs away from a hinge point of the regularization path, i.e., there exists $\eta > 0$ such that for all $j \in \{1, \dots, p\}$, $v_j = 0$ implies $|(Q(w_0 - \mathbf{w}))_j| \leqslant \mu_0 - \eta M\sigma$. If $|\tilde{\mu} - \tilde{\mu}_0| \leqslant \tilde{\lambda} \min\{\eta/4, \mathrm{m}(w_0 M/\sigma)\}$, then

$$\mathbb{P}(\mathrm{sign}(\hat{w}) \neq \mathrm{sign}(w_0)) \leqslant 2p \exp\left(-\frac{\tilde{\lambda}^2}{\tilde{\tau}^2} \min\{\eta^2/4, \mathrm{m}(w_0 M/\sigma)^2\} \frac{n}{p}\right).$$

The proposition above makes no claim in the situation where μ tends to zero. As we now show, this depends on the rate of decay of μ, slower, faster, or exactly at the rate $n^{-1/2}$.

2.3 High regularization

If μ tends to zero slower than $n^{-1/2}$, then $\hat{w}$ converges in probability to $\mathbf{w}$ (regular consistency) and the sign pattern converges to the sign pattern of the global minimum of a local noiseless objective function $\frac{1}{2}\Delta^\top Q\Delta + \Delta_{\mathbf{J}}^\top \mathrm{sign}(\mathbf{w}_{\mathbf{J}}) + \|\Delta_{\mathbf{J}^c}\|_1$, the convergence being exponential in $\mu^2 n$ (see proof in Appendix D.3). The local noiseless problem in Eq. (2.2) is simply obtained by a first-order expansion of the Lasso objective function around $\mathbf{w}$ [21, 40].

Proposition 2.3. *Assume (**A1-4**). Let Δ be the unique solution of*

$$\min_{\Delta \in \mathbb{R}^p} \frac{1}{2}\Delta^\top Q\Delta + \Delta_{\mathbf{J}}^\top \mathrm{sign}(\mathbf{w}_{\mathbf{J}}) + \|\Delta_{\mathbf{J}^c}\|_1. \tag{2.2}$$

Assume that $\tilde{\mu} \leqslant \frac{\mathrm{m}(\tilde{\mathbf{w}})\tilde{\lambda}}{2p^{1/2}}$. We have:

$$\mathbb{P}(\|\hat{w} - \mathbf{w} - \mu\Delta\|_2 \geqslant \beta\sigma/M) \leqslant 2p \exp\left(-\frac{\tilde{\lambda}^2\beta^2}{8\tilde{\tau}^2}\frac{n}{p}\right).$$

Moreover, assume the minimum Δ of Eq. (2.2) occurs away from a hinge point of the regularization path, i.e., there exists $\eta > 0$ such that for all $j \in \mathbf{J}^c$, $\Delta_j = 0$ implies $|(Q\Delta)_j| \leqslant 1 - \eta$. Then,

$$\mathbb{P}(\mathrm{sign}(\hat{w}) \neq \mathrm{sign}(\mathbf{w} + \mu\Delta)) \leqslant 2p \exp\left(-\frac{\mathrm{m}(\tilde{\mathbf{w}})\tilde{\lambda}^2}{8\tilde{\tau}^2}\frac{n}{p}\right) + 2p \exp\left(-A\tilde{\mu}^2\frac{n}{p}\right),$$

where $A = \tilde{\tau}^{-2}\tilde{\lambda} \min\{\tilde{\lambda}\mathrm{m}(M^2\Delta)^2/2, \eta^2/8\}$.

Note that the sign pattern of $\mathbf{w} + \mu\Delta$ is equal to the population sign vector $\mathbf{s} = \mathrm{sign}(\mathbf{w})$ if and only if the problem in Eq. (2.2) has a solution where $\Delta_{\mathbf{J}^c}$ is equal to zero. A short calculation shows that this occurs if and only if the following consistency condition is satisfied [32, 44, 40, 45, 37]:

$$\|Q_{\mathbf{J}^c,\mathbf{J}} Q_{\mathbf{J},\mathbf{J}}^{-1} \mathrm{sign}(\mathbf{w}_{\mathbf{J}})\|_\infty \leqslant 1. \tag{2.3}$$

Thus, if Eq. (2.3) is satisfied strictly—which implies that we are not at a hinge point of Eq. (2.2)—the probability of correct sign estimation is tending to one, and to zero if Eq. (2.3) is not satisfied (see [40] for precise statements when there is equality). Moreover, when

Eq. (2.3) is satisfied strictly, Proposition 2.3 gives an upper bound on the probability of not selecting the correct pattern $\mathbf{J}$.

The first three regimes are not unique to low-dimensional settings; we show in Section 4 the corresponding proposition related to Proposition 2.3, for high-dimensional settings. However, the last two regimes (μ tending to zero at rate $n^{-1/2}$ or faster) are specific to low-dimensional settings.

2.4 Medium regularization

If $\mu n^{1/2}$ is bounded from above and from below, then we show that the sign pattern of $\hat{w}$ agrees on $\mathbf{J}$ with the one of $\mathbf{w}$ with probability tending to one exponentially fast in n (Proposition 2.4), while for all sign patterns consistent on $\mathbf{J}$ with the one of $\mathbf{w}$, the probability of obtaining this pattern is tending to a limit in $(0,1)$ (in particular strictly positive); that is, all sign patterns consistent with $\mathbf{w}$ on the relevant variables (i.e., the ones in $\mathbf{J}$) are possible with positive probability (Proposition 2.5). The convergence of this probability follows a rate of $n^{-1/2}$ (see proof in Appendix D.4 and D.5). Note the difference with earlier results [1] obtained for random designs.

Proposition 2.4. *Assume (**A1-4**) and $\tilde{\mu} \leqslant \frac{\mathrm{m}(\tilde{\mathbf{w}})\tilde{\lambda}}{2p^{1/2}}$. Then for any sign pattern $s \in \{-1,0,1\}^p$ such that $s_{\mathbf{J}} = \mathrm{sign}(\mathbf{w}_{\mathbf{J}})$, there exists $f(s, n^{1/2}\mu p^{1/2}) \in (0,1)$, such that:*

$$|\mathbb{P}(\mathrm{sign}(\hat{w}) = s) - f(s, n^{1/2}\mu p^{1/2})| \leqslant \frac{4C_1^{\mathrm{BE}}\tilde{\tau}^3}{\tilde{\lambda}^{1/2}}\frac{p^2}{n^{1/2}} + 2p\exp\left(-\frac{\mathrm{m}(\tilde{\mathbf{w}})\tilde{\lambda}^2}{8\tilde{\tau}^2}\frac{n}{p}\right).$$

Proposition 2.5. *Assume (**A1-4**) and $\tilde{\mu} \leqslant \frac{\mathrm{m}(\tilde{\mathbf{w}})\tilde{\lambda}}{2p^{1/2}}$. Then, for any pattern $s \in \{-1,0,1\}^p$ such that $s_{\mathbf{J}} \neq \mathrm{sign}(\mathbf{w}_{\mathbf{J}})$,*

$$\mathbb{P}(\mathrm{sign}(\hat{w}) = s) \leqslant 2p\exp\left(-\frac{\mathrm{m}(\tilde{\mathbf{w}})\tilde{\lambda}^2}{8\tilde{\tau}^2}\frac{n}{p}\right).$$

The positive real numbers C_1^{BE} and C_2^{BE} are universal constants related to multivariate Berry-Esseen inequalities (see Appendix A.1 for more details). From the proof in Appendix D.4, the constant $f(s,c)$ has specific behaviors when $c = \mu n^{1/2}p^{1/2}$ is small or large: on the one hand, if c tends to infinity, then we tend to the bevahior of the previous section, that is, $f(s,c)$ tends to one if s is the limiting pattern in Proposition 2.3 and zero otherwise. On the other hand, if c tends to 0, $f(s,c)$ tends to one if s has no zeros, and zero otherwise (see next section).

The last two propositions state that the relevant variables are *stable*, i.e., we get all relevant variables with probability tending to one *exponentially fast*, while we get exactly get all other patterns with probability tending to a limit *strictly* between zero and one. This stability of the relevant variables is the source of the intersection arguments outlined in Section 3.

Note that Proposition 2.4 makes non-trivial statements only for n larger than p^4; the fourth power is due to the application of Berry-Esseen inequalities, and could be improved.

2.5 Low regularization

If μ tends to zero faster than $n^{-1/2}$, then $\hat{w}$ is consistent (i.e., converges in probability to $\mathbf{w}$) but the support of $\hat{w}$ is equal to $\{1, \dots, p\}$ with probability tending to one (the signs of variables in $\mathbf{J}^c$ may then be arbitrarily negative or positive). That is, the ℓ^1-norm has no sparsifying effect. We obtain two different bounds, with different scalings in p and n (see proof in Appendix D.6):

Proposition 2.6. *Assume (**A1-4**) and* $\tilde{\mu} \leqslant \frac{\mathrm{m}(\tilde{\mathbf{w}})\tilde{\lambda}}{2p^{1/2}}$. *Then the probability of having at least one zero variable is smaller than* $3^p \left(C_1^{\mathrm{BE}} \frac{4\tilde{\tau}^3}{\tilde{\lambda}^{1/2}} \frac{p^2}{n^{1/2}} + \frac{\tilde{\mu} n^{1/2}}{\tilde{\lambda}^{1/2}} \right)$ *and* $\frac{\tilde{\mu} n^{1/2} p}{\tilde{\lambda}^{1/2}} + \frac{10 C_2^{\mathrm{BE}}}{\tilde{\tau}^3 \tilde{\lambda}} \frac{p^{7/2}}{\tilde{\mu} n} + C_2^{\mathrm{BE}} \frac{4\tilde{\tau}^3}{\tilde{\lambda}^{1/2}} \frac{p^3}{n^{1/2}} + 2|\mathbf{J}| \exp\left(-\frac{\mathrm{m}(\tilde{\mathbf{w}})\tilde{\lambda}^2}{8\tilde{\tau}^2} \frac{n}{p} \right)$.

The first bound simply requires that μ tends to zero faster than $n^{-1/2}$, but the constant is exponential in p, while the second bound required that μ does not tend to zero too fast, i.e., between $n^{-1/2}$ and n^{-1} (with constants polynomial in p). As shown in Appendix D.3, the two bounds correspond to two different applications of Berry-Esseen inequalities, one for all the possible 3^p sign patterns, one using a detailed analysis of the non-selection of a given variable (see Section 2.6). We are currently exploring the possibility of having a bound that shares the positive aspects of our two bounds—polynomial in p and without the term $(\tilde{\mu} n)^{-1}$.

Among the five previous regimes, the only ones with consistent estimates (in norm) and a sparsity-inducing effect are μ tending to zero and $\mu n^{1/2}$ tending to a finite or infinite limit. When $\mu n^{1/2}$ tends to infinity, we can only hope for model consistent estimates if the consistency condition in Eq. (2.3) is satisfied. This somewhat disappointing result for the Lasso has led to various improvements on the Lasso to ensure model consistency even when Eq. (2.3) is not satisfied [40, 45, 31]. Those are based on adaptive weights based on the non regularized least-square estimate or two-step procedures. We propose in Section 3 alternative ways which are based on resampling. Before doing so, we derive in the next section finer results that allows to consider the presence or absence in the support set $\hat{J}$ of a specific variable without considering all corresponding consistent sign patterns.

2.6 Probability of not selecting a given variable

We can lower and upper bound the probability of not selecting a certain irrelevant variable in $\mathbf{J}^c$ (see proof in Appendix D.7)—see Proposition 2.5 for a related proposition for relevant variables in $\mathbf{J}$:

Proposition 2.7. *Assume (**A1-4**) and* $\tilde{\mu} \leqslant \frac{\mathrm{m}(\tilde{\mathbf{w}})\tilde{\lambda}}{2p^{1/2}}$. *Let* $j \in \mathbf{J}^c$. *We have:*

$$\mathbb{P}(j \in \hat{J}) \geqslant \frac{\tilde{\mu} n^{1/2}/4}{1 + \tilde{\mu} n^{1/2}/2\tilde{\lambda}^{1/2}} \exp\left(-\frac{2\tilde{\mu}^2}{\tilde{\lambda}^2} np \right) - \frac{10 C_2^{\mathrm{BE}}}{\tilde{\tau}^3 \tilde{\lambda}^1} \frac{p^{5/2}}{\tilde{\mu} n} - C_2^{\mathrm{BE}} \frac{4\tilde{\tau}^3}{\tilde{\lambda}^{1/2}} \frac{p^2}{n^{1/2}},$$

$$\mathbb{P}(j \in \hat{J}) \leqslant \frac{\tilde{\mu} n^{1/2}}{\tilde{\lambda}^{1/2}} + \frac{8 C_2^{\mathrm{BE}}}{\tilde{\tau}^3 \tilde{\lambda}} \frac{p^{5/2}}{\tilde{\mu} n} + C_2^{\mathrm{BE}} \frac{4\tilde{\tau}^3}{\tilde{\lambda}^{1/2}} \frac{p^2}{n^{1/2}}.$$

This novel proposition allows to consider "marginal" probabilities of selecting (or not selecting) a given variable, without considering all consistent sign patterns associated with the selection (or non-selection) of that variable). Note that it makes interesting claims only when $\mu n^{1/2}$ is bounded from above and below (for the lower bound) and when $\mu n^{1/2}$ tends to zero, while μn tends to infinity (for the upper bound).

3 Support Estimation by Intersection

The results from Section 2.4 exactly show that under suitable choices of the regularization parameter μ, the relevant variables are stable while the irrelevant are unstable, leading to several intersecting arguments to keep only the relevant variables. We first consider the irrealistic situation where we have multiple independent copies, then we consider splitting a dataset in several pieces, and we finally present two usual types of bootstrap (pairs and residuals). Note that an alternative approach is to resample the columns of the design matrix instead of its rows, i.e., draw random weights for each variable from a well-chosen distribution [33].

The analysis of support estimation is essentially the same for all methods and is based on the following argument: we consider m "replications", and $\hat{J}^1, \dots, \hat{J}^m$ the associated active sets. The replications are assumed independent given the original data (i.e., the vector of noise ε). We let denote $\hat{J}^\cap = \bigcap_{i=1}^m \hat{J}^i$ the estimate of the active set (given the original data). Once the active set is found, the final estimate of w is obtained by the unregularized least-square estimate, restricted to the estimated active set.

We can upper bound the probability of incorrect pattern selection as follows:

$$\begin{aligned}
\mathbb{P}(\hat{J}^\cap \neq \mathbf{J}) &\leqslant \mathbb{P}(\mathbf{J}^c \cap \hat{J}^\cap \neq \varnothing) + \mathbb{P}(\mathbf{J} \cap (\hat{J}^\cap)^c \neq \varnothing), \\
&\leqslant \sum_{j \in \mathbf{J}^c} \mathbb{P}(\forall i \in \{1, \dots, m\}, j \in \hat{J}^i) + \mathbb{P}\left(\bigcup_{i=1}^m \left[(\hat{J}^i)^c \cup \mathbf{J} \right] \neq \varnothing \right), \\
&\leqslant \sum_{j \in \mathbf{J}^c} \mathbb{E}(\mathbb{P}(j \in \hat{J}^* | \varepsilon)^m) + m \mathbb{P}((\hat{J}^*)^c \cup \mathbf{J} \neq \varnothing),
\end{aligned}$$

where $\hat{J}^*$ denotes a generic support obtained from one replication. We now need to upper bound the probability $\mathbb{P}((\hat{J}^*)^c \cup \mathbf{J} \neq \varnothing)$ of forgetting at least a relevant variable $j \in \mathbf{J}$, and also the probability $\mathbb{P}(j \in \hat{J}^* | \varepsilon)$ that a replication does not include a given irrelevant variable $j \in \mathbf{J}^c$ (given the original data). The first term will always drop as the number of replications gets larger, while the second term increases, leading to a natural trade-off for the choice of the number m of replications. This is to be contrasted with usual applications of the boostrap where m is taken as large as computationally feasible.

3.1 Multiple independent copies

Let us assume for a moment that we have m independent copies of similar datasets, with potentially different fixed designs but same noise distribution. We then have m different active sets and we denote by $\hat{J}^\cap$ the intersection of the m active sets. We have the following upper bound on the probability of non selecting the correct pattern (see proof in Appendix D.8)

Proposition 3.1. *Assume (**A1-4**) for m independent datasets with same noise distribution, and $\tilde{\mu} \leqslant \frac{\mathrm{m}(\tilde{\mathbf{w}})\tilde{\lambda}}{2p^{1/2}}$. If $c = \tilde{\mu} n^{1/2} p^{1/2} > 0$, $\tilde{\mu} \leqslant \frac{\mathrm{m}(\tilde{\mathbf{w}})\tilde{\lambda}}{2p^{1/2}}$ and $n \geqslant p^6 g(c)$, then there exists $f(c) > 0$ such that*

$$\mathbb{P}(\hat{J}^{\cap} \neq \mathbf{J}) \leqslant p e^{-f(c) m p^{-1/2}} + 2pm \exp\left(-\frac{\mathrm{m}(\tilde{\mathbf{w}})\tilde{\lambda}^2}{8\tilde{\tau}^2} \frac{n}{p} \right).$$

From the proof of Proposition 3.1 in Appendix D.8, we can get the detailed behavior of $f(c)$ around $c = 0$ and $c = \infty$: it goes to zero in both cases, i.e., we actually need (in the bound) a regularization parameter that is proportional to $n^{-1/2}$.

Moreover, we get an exponential convergence rate in n and m, where we have two parts: one that states that the number of copies should be as large as possible to remove irrelevant variables (left part), and one that states that m should not be too large, otherwise, some relevant variables would start to disappear (right part). Note that best scaling (for the bound) is $m \approx n$, leading to a probability of incorrect selection that goes to zero exponentially fast in n.

Of course, in practice, one is not given multiple independent copies of the same datasets, but a single one. One strategy is to split it in different pieces, as described in Section 3.2; this however relies on having enough data to get a large number of pieces, which is unlikely to happen in practice. Our main goal is this paper is to show that by using the bootstrap, we can mimic the availability of having multiple copies. This will come at a price, namely an overall convergence rate of $n^{-1/2}$ instead of exponential in n

3.2 Splitting into pieces

We can cut the dataset into m pieces of the same size, a procedure reminiscent of cross-validation. However, it requires extra-assumption on the design, i.e., we need to assume that the smallest eigenvalues of the data matrices of length n/m are still strictly positive (see proof in Appendix D.9):

Proposition 3.2. *Assume (**A1-4**) for m disjoint subdatasets of the original dataset, and $\tilde{\mu} \leqslant \frac{\mathrm{m}(\tilde{\mathbf{w}})\tilde{\lambda}}{2p^{1/2}}$. If $c = \tilde{\mu} n^{1/2} m^{-1/2} p^{1/2} > 0$, there exists $f(c), a(c) > 0$ such that:*

$$\mathbb{P}(\hat{J}^{\cap} \neq \mathbf{J}) \leqslant p \left(1 - e^{-f(c)p^{-1/2}} + h(c) \frac{p^{5/2} m^{1/2}}{n^{1/2}} \right)^m + 2pm \exp\left(-\frac{\mathrm{m}(\tilde{\mathbf{w}})\tilde{\lambda}^2}{8\tilde{\tau}^2} \frac{n}{mp} \right).$$

The proposition above requires that m/n tends to zero, i.e., there should not be too many pieces (which is also required to allow invertibility of the sub-designs). Note that several independent partitions could be considered, and would lead to results similar to the ones for the bootstrap presented in the next two sections [33].

3.3 Random pair bootstrap

Given the n observations $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \dots, n$, put together into matrices $X \in \mathbb{R}^{n \times p}$ and $y \in \mathbb{R}^n$, we consider m *bootstrap* replications of the n data points [17]; that is, for

$k = 1, \dots, m$, we consider a *ghost sample* $(x_i^k, y_i^k) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \dots, n$, given by matrices $X^k \in \mathbb{R}^{n \times p}$ and $y^k \in \mathbb{R}^n$. For each $k \in \{1, \dots, m\}$, the n pairs (x_i^k, y_i^k), $i = 1, \dots, n$, are sampled uniformly and independently at random *with replacement* from the n original pairs in (X, y). Some pairs (x_i, y_i) are not selected, some selected once, some selected twice, and so on. Note that we could consider bootstrap replications with more or less points than n, but for simplicity, we keep it the same as the original number of data points.

The following proposition shows that we obtain a consistent model estimate by intersecting the active sets $\hat{J}^1, \dots, \hat{J}^m$ obtained from running the Lasso on each bootstrap sample $(X^1, y^1), \dots, (X^m, y^m)$, a procedure we refer to as the *Bolasso* (see proof in Appendix E):

Proposition 3.3. *Assume (**A1-4**). If $c = \tilde{\mu} n^{1/2} p^{1/2} > 0$, there exists strictly positive constants $A_0, \dots, A_7$ that may depend on c such that if $np^{-6} \geqslant A_6$ and $mp^{-1} \geqslant A_7$, we have, for boostrapping* pairs*:*

$$\mathbb{P}(\hat{J}^{\cap} \neq \mathbf{J}) \leqslant mp \exp\left(-A_0 \frac{n^{1/2}}{p^{1/2}}\right) + A_4 \left(A_3 \frac{p^3}{n^{1/2}} + \frac{\log m}{m}\right)^{1 + A_5 \left(2 \log\left(A_3 \frac{p^3}{n^{1/2}} + \frac{\log m}{m}\right)\right)^{-1/2}}.$$

Note that in Proposition 3.3, for any $\eta > 0$, if n and m are large enough, then we get an upper bound on the probability of incorrect model selection of the form $B_1 m e^{-B_2 n^{1/2}} + \left(\frac{B_3}{n^{1/2}} + B_4 \frac{\log m}{m}\right)^{1+\eta}$, where $B_1, \dots, B_4$ are positive constants. Note that in [1], we have derived a bound with better behavior in n, i.e., with $\eta = 0$. However, the bound in [1] holds for random designs and has constants which scale *exponentially* in p and not polynomially. We are currently trying to improve on the bound in Proposition 3.3 to remove the extra factor $\eta > 0$.

As before, the number of replications should be as large as possible to remove irrelevant variables, and m should not be too large, otherwise, some relevant variables would start to disappear from the intersection. Note that best scaling (for the bound) is $m \approx n^{1/2}$, leading to an overall probability of incorrect model selection that tends to zero at rate $n^{-1/2}$, instead of the exponential rate for the irrealistic situation of having multiple copies (Section 3.1).

We have not explored yet the optimality (in the minimax sense) of the bound given in Proposition 3.3. While we believe that a rate of $n^{-1/2}$ cannot be improved upon, the rate p^6 should be improved with further research.

Finally, we have explored in [1] the possibility of considering softer ways of performing the intersection, i.e., by keeping all variables that appear in a certain proportion of the active sets corresponding to the various replications. This is important in cases where the decay of the loading vectors does not have sharp threshold as assumed in most analyses (this paper included). However, it adds an extra hyper-parameter and the theoretical analysis of such schemes is out of the scope of this paper.

3.4 Boostrapping residuals

An alternative to resampling pairs (x_i, y_i) is to resample only the estimated centered residuals [17, 18]. This is well adapted to fixed-design assumptions, in particular because the design matrix X remains the same for all replications. Note however, that the consistency of this

resampling scheme usually relies more heavily on the homoscedasticity assumption (**A2**) that we make in this paper [18]. Moreover, since the Lasso estimate is biased, the behavior differs slightly from bootstrapping pairs, as shown empirically in Section 5.

Bootstrapping residuals works as follows; we let denote $\tilde{\varepsilon}_i = y_i - \hat{w}^\top x_i = \varepsilon_i - (\hat{w} - \mathbf{w})^\top x_i$ the vector of estimated residuals, and $\hat{\varepsilon}_i$ the centered residuals equal to $\hat{\varepsilon}_i = \tilde{\varepsilon}_i - \frac{1}{n}\sum_{k=1}^n \tilde{\varepsilon}_k$. When bootstrapping residuals, for each $i \in \{1, \dots, n\}$, we keep x_i unchanged and we use as data $y_i^* = \hat{w}^\top x_i + \hat{\varepsilon}_{i^*}$, where i^* is a random index in $\{1, \dots, n\}$—the sampling is uniform and the n indices are drawn independently.

We obtain a similar bound than when bootstrapping pairs (see proof in Appendix F.2):

Proposition 3.4. *Assume (**A1-4**). If $c = \tilde{\mu} n^{1/2} p^{1/2} > 0$, there exists strictly positive constants $A_0, \dots, A_7$ that may depend on c such that if $np^{-6} \geqslant A_6$ and $mp^{-1} \geqslant A_7$, we have, for boostrapping* residuals*:*

$$\mathbb{P}(\hat{J}^\cap \neq \mathbf{J}) \leqslant mp \exp\left(-A_0 \frac{n^{1/2}}{p^{1/2}}\right) + A_4 \left(A_3 \frac{p^3}{n^{1/2}} + \frac{\log m}{m}\right)^{1 + A_5 \left(2 \log\left(A_3 \frac{p^3}{n^{1/2}} + \frac{\log m}{m}\right)\right)^{-1/2}}.$$

The bound in Proposition 3.4 is the same as bootstrapping pairs, but as shown in Appendix F.2, the constants are slightly better). However, as shown in Section 5.3, the behaviors of the two methods differ notably: random-pair bootstrap does not lead to good selection performance in high-dimensional settings, while residual bootstrap does. While we are currently unable to proof the consistency of bootstrapping residuals in high-dimensional settings, we prove in Section 4 the model consistency of a related two-step procedure, where the bootstrap replications are performed within the support of the Lasso estimate on the full data.

4 High-Dimensional Analysis

In high-dimensional settings, i.e., when p may be larger than n, we need to change assumption (**A4**) regarding the invertibility of the empirical second order moment, which cannot hold. Various assumptions have been used for the Lasso, based on low correlations [27], sparse eigenvalues [34] or more general conditions [6, 15]. In this paper, we introduce a novel assumption, which not only allows us to consider that the support of the Lasso estimate has a bounded size, but also implies that we obtain the same sign pattern with high probability. The analysis carried out in low-dimensional settings in Section 2.3 is thus also valid in high-dimensional settings.

4.1 High-dimensional assumptions

Our analysis relies on the analysis carried out in Section 2.3 for "high" regularization, i.e., when μ tends to zero slower than $n^{-1/2}$. In this setting, we have shown that the Lasso estimate asymptotically behaves as $\mathbf{w} + \mu\Delta$, where Δ is the unique minimum of

$$\min_{\Delta \in \mathbb{R}^p} \frac{1}{2}\Delta^\top Q \Delta + \Delta_{\mathbf{J}}^\top \operatorname{sign}(\mathbf{w}_{\mathbf{J}}) + \|\Delta_{\mathbf{J}^c}\|_1. \tag{4.1}$$

We let denote $\mathbf{K} \subset \mathbf{J}^c$ the "extended" support of a solution $\Delta_{\mathbf{J}^c}$ of Eq. (4.3) and $\mathbf{L} = \mathbf{J} \cap \mathbf{K}$: that is, we not only keep all indices corresponding to non zero elements of $\Delta_{\mathbf{J}^c}$, but also the ones for which the optimality condition in Eq. (C.1) is an equality (i.e., if we are at a hinge point of the regularization path, we take all involved variables)

We consider the vector $\mathbf{t} \in \{-1, 0, +1\}^p$ defined by $t_{\mathbf{J}} = \operatorname{sign}(\mathbf{w}_{\mathbf{J}})$ and $\mathbf{t}_{\mathbf{J}^c} = \operatorname{sign}(\Delta_{\mathbf{J}^c})$. If we assume that $\lambda_{\min}(Q_{\mathbf{L},\mathbf{L}}) > 0$, then the solution to Eq. (4.1) is unique [22], and is such that $\Delta_{\mathbf{L}} = -Q_{\mathbf{L},\mathbf{L}}^{-1}\mathbf{t}_{\mathbf{L}}$ and the optimality conditions for Eq. (4.1) are simply

$$\operatorname{sign}(-[Q_{\mathbf{L},\mathbf{L}}^{-1}\mathbf{t}_{\mathbf{L}}]_{\mathbf{K}}) = \mathbf{t}_{\mathbf{K}} \text{ and } \|Q_{\mathbf{L}^c\mathbf{L}}Q_{\mathbf{L},\mathbf{L}}^{-1}\mathbf{t}_{\mathbf{L}}\|_\infty \leqslant 1.$$

We make the following assumptions (note that (**A6**) is essentially equivalent to the lack of hinge point which is also made in Proposition 2.3):

(**A5**) *Unicity of local noiseless problem*: the matrix $\mathbf{Q}_{\mathbf{L},\mathbf{L}}$ is invertible.

(**A6**) *Stability of local noiseless problem*: $\|Q_{\mathbf{L}^c\mathbf{L}}Q_{\mathbf{L},\mathbf{L}}^{-1}\mathbf{t}_{\mathbf{L}}\|_\infty < 1$.

We let denote

$$\boldsymbol{\theta} = \min\left\{1 - \|Q_{\mathbf{L}^c\mathbf{L}}Q_{\mathbf{L},\mathbf{L}}^{-1}\mathbf{t}_{\mathbf{L}}\|_\infty, \min_{k\in\mathbf{K}} |(Q_{\mathbf{L},\mathbf{L}}^{-1}\mathbf{t}_{\mathbf{L}})_k Q_{k,k}|\right\}, \tag{4.2}$$

the quantity that will characterize the *stability* of the local noiseless problem; if (**A5-6**) are satisfied, then $\boldsymbol{\theta} > 0$. As shown in Proposition 4.1, the quantity $\boldsymbol{\theta}$ dictates the speed of convergence of the probability of not getting $\mathbf{t}$ as a sign pattern for the Lasso problem in Eq. (1.1) or Eq. (2.1).

Comparison with consistency condition We now relate (**A6**) with the consistency condition for the Lasso in Eq. (2.3): if Eq. (2.3) satisfied, then $\mathbf{K} = \varnothing$ and the condition (**A6**) simply becomes:

$$\|Q_{\mathbf{J}^c,\mathbf{J}}Q_{\mathbf{J},\mathbf{J}}^{-1}\operatorname{sign}(\mathbf{w}_{\mathbf{J}})\|_\infty < 1,$$

which is exactly a strict version of Eq. (2.3)—an assumption commonly made for high-dimensional analysis of the Lasso [44, 37]. Note that we then have the simplified expression $\boldsymbol{\theta} = 1 - \|Q_{\mathbf{J}^c\mathbf{J}}Q_{\mathbf{J},\mathbf{J}}^{-1}\operatorname{sign}(\mathbf{w}_{\mathbf{J}})\|_\infty$.

The main goal of this paper is to design a consistent procedure even when Eq. (2.3) is not satisfied. As we have seen, (**A6**) is weaker than the usual assumptions made for the Lasso consistency; in Figure 1 (left and middle), we compare empirically the two conditions for random i.i.d. Gaussian designs, showing that our set of assumptions is weaker, but of course breaks down when n is too small (too few observations) or the cardinal of $\mathbf{J}$ is too large (too many relevant variables). We are currently exploring theoretical proofs of this behavior, extending the current analysis of [37] for Eq. (2.3); in particular, we aim at determining the various scalings between p, n and the number of relevant variables for which a Gaussian ensemble leads to consistent variable selection with high probability (according to our assumptions which are weaker than in [37]). Moreover, in the right plot of Figure 1, we show values of $\log \boldsymbol{\theta}$ for various n and $|\mathbf{J}|$, which characterize the convergence rate of our bound. Relying on $\boldsymbol{\theta}$ which is bounded from below is clearly a weakness of our approach to high-dimensional estimation; we are currently exploring refined conditions where we relax the stability, i.e., we allow several (but not too many) patterns to be selected with overwhelming probability.

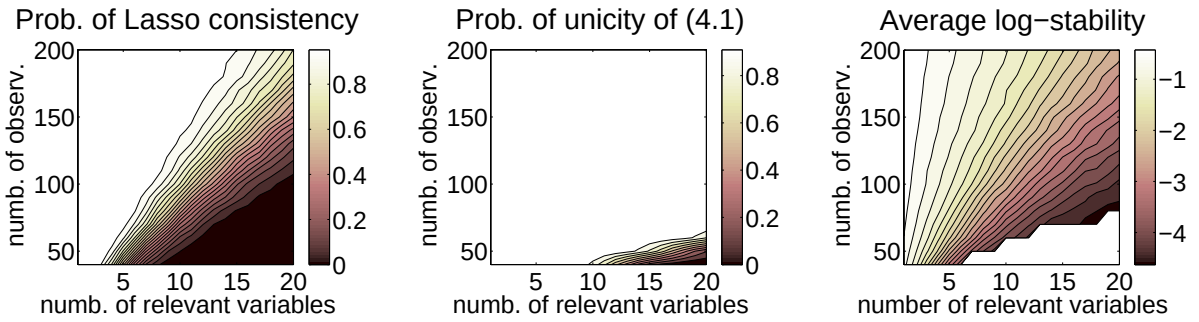

Figure 1: Consistency conditions for random Gaussian designs, $p = 128$, n from 40 to 200 and $|\mathbf{J}|$ from 1 to 20 (all probabilities and averages obtained from 1000 replications). Left: probability that Eq. (2.3) is satisfied. Middle: probability that (**A6**) is satisfied. Right: expectation of $\log \boldsymbol{\theta}$ (plotted only for the ones for which the local problem is unique with high probability).

Checking assumptions (A5-6) In Eq. (4.1), we can optimize in closed form with respect to $\Delta_{\mathbf{J}}$ as $\Delta_{\mathbf{J}} = Q_{\mathbf{J},\mathbf{J}}^{-1}(-\operatorname{sign}(\mathbf{w}_{\mathbf{J}}) - Q_{\mathbf{J},\mathbf{J}^c}\Delta_{\mathbf{J}^c})$, leading to an optimization problem for $\Delta_{\mathbf{J}^c}$:

$$\min_{\Delta \in \mathbb{R}^p} \frac{1}{2}\Delta_{\mathbf{J}^c}^\top Q_{\mathbf{J}^c,\mathbf{J}^c|\mathbf{J}}\Delta_{\mathbf{J}^c} - \Delta_{\mathbf{J}^c}^\top Q_{\mathbf{J}^c,\mathbf{J}}Q_{\mathbf{J},\mathbf{J}}^{-1}\operatorname{sign}(\mathbf{w}_{\mathbf{J}}) + \|\Delta_{\mathbf{J}^c}\|_1, \tag{4.3}$$

which can be solved using existing code for the Lasso. We are currently working on deriving sufficient conditions which do not depend on the sign pattern of the population loading $\mathbf{w}$ (but only on the sparsity pattern, or even its cardinality), as usually done for the consistency condition in Eq. (2.3) [44, 40].

4.2 Stability of sign selection

With assumptions (**A5**) and (**A6**), we can show that with high-probability, when the regularization parameter is asymptotically greater than $n^{-1/2}$, then the sign of the Lasso estimate is exactly $\mathbf{t}$ (see proof in Appendix G):

Proposition 4.1. *Assume (**A1-3**), (**A5-6**), and* $\tilde{\mu} \leqslant \frac{\tilde{\lambda}_{\mathbf{L}}\mathrm{m}(\tilde{\mathbf{w}})}{2|\mathbf{L}|^{1/2}}$. *Then:*

$$\mathbb{P}(\operatorname{sign}(\hat{w}) \neq \mathbf{t}) \leqslant 2p\exp\left(-\frac{n\tilde{\mu}^2\boldsymbol{\theta}^2\tilde{\lambda}_{\mathbf{L}}}{8\tilde{\tau}^2|\mathbf{L}|}\right) + 2|\mathbf{J}|\exp\left(-\frac{n\mathrm{m}(\tilde{\mathbf{w}})^2\tilde{\lambda}_{\mathbf{L}}^2}{4\tilde{\tau}^2|\mathbf{L}|}\right). \tag{4.4}$$

Note that if $\boldsymbol{\theta}$ is bounded away from zero, then we simply need that $\log p = o(n)$ for our result to hold. Moreover, in Eq. (4.4), we can see that $\boldsymbol{\theta}$ dictates the asymptotic behavior of our bound. If it is too small, then in order to have a meaningful bound for this design matrix, we would need to consider sign patterns which are close to $\mathbf{t}$ and show that the sign pattern of the Lasso estimate $\hat{w}$ is with high probability within these sign patterns.

4.3 High-dimensional Bolasso

Proposition 4.1 suggests to run the Lasso once with a larger regularization parameter (i.e., multiplied by $\log p$) and run the various resampling schemes within the active set of the original Lasso estimation (which is very likely to be the support associated with $\mathbf{t}$). More precisely, we have the proposition (see proof in Appendix G):

Proposition 4.2. *Assume (**A1-3**) and (**A5-6**). If $c = \tilde{\mu} n^{1/2} |\mathbf{L}|^{1/2} > 0$, there exists strictly positive constants $A_0, \dots, A_7$ that may depend on c such that if $n|\mathbf{L}|^{-6} \geqslant A_6$ and $m|\mathbf{L}|^{-1} \geqslant A_7$, we have, for boostrapping* residuals*:*

$$\mathbb{P}(\hat{J}^{\cap} \neq \mathbf{J}) \leqslant 2p \exp\left(-\frac{c^2 (\log p)^2 \boldsymbol{\theta}^2 \tilde{\lambda}_{\mathbf{L}}}{8\tilde{\tau}^2 |\mathbf{L}|^2} \right) + 2|\mathbf{J}| \exp\left(-\frac{n \mathrm{m}(\tilde{\mathbf{w}})^2 \tilde{\lambda}_{\mathbf{L}}^2}{4\tilde{\tau}^2 |\mathbf{L}|} \right) +$$
$$mp \exp\left(-A_0 \frac{n^{1/2}}{|\mathbf{L}|^{1/2}} \right) + A_4 \left(A_3 \frac{|\mathbf{L}|^3}{n^{1/2}} + \frac{\log m}{m} \right)^{1 + A_5 \left(2 \log\left(A_3 \frac{|\mathbf{L}|^3}{n^{1/2}} + \frac{\log m}{m} \right) \right)^{-1/2}}.$$

Note that the constants depend polynomially on $|\mathbf{L}|$ and $\lambda_{\min}(Q_{\mathbf{L},\mathbf{L}})$, and do not depend on p. This is thus a high-dimensional result where p may grow large compared to n. If we relax (**A6**), then the original Lasso estimate would have a small set of allowed patterns with high probability (instead of simply one), and a union bound considering all those would need be considered.

5 Algorithms and Simulations

In this section, we describe efficient algorithms for the boostrapped versions of the Lasso that we present in this paper and we illustrate the various consistency results obtained in previous sections, in low-dimensional and high-dimensional settings.

5.1 Efficient Path Algorithms

We first consider efficient algorithms for the boostrapping procedures, based on homotopy methods [35, 16, 23]. Similar developments could be made for first-order methods [20, 19]. For the regular Lasso, one can find the solutions of Eq. (1.1) for all values of the regularization parameter μ that correspond to less than k selected covariates in time which is empirically $O(pn + k^2 n)$: indeed, computing $\frac{1}{n} X^\top y$ once is $O(pn)$, while computing the relevant elements of $Q = \frac{1}{n} X^\top X$ and updating various quantities is $O(k^2 n)$. Note that our analysis suggests to stop the path when the solution of the problem is not unique anymore, i.e., when the design matrix of selected variables become rank-deficient.

Bootstrapping pairs When bootstrapping pairs, we require m applications of the regular Lasso procedure with different design matrices, so we get a complexity of $O(mpn + mk^2 n)$, and since the designs are different, there is no immediate possibility of sharing computations between different bootstrap replications

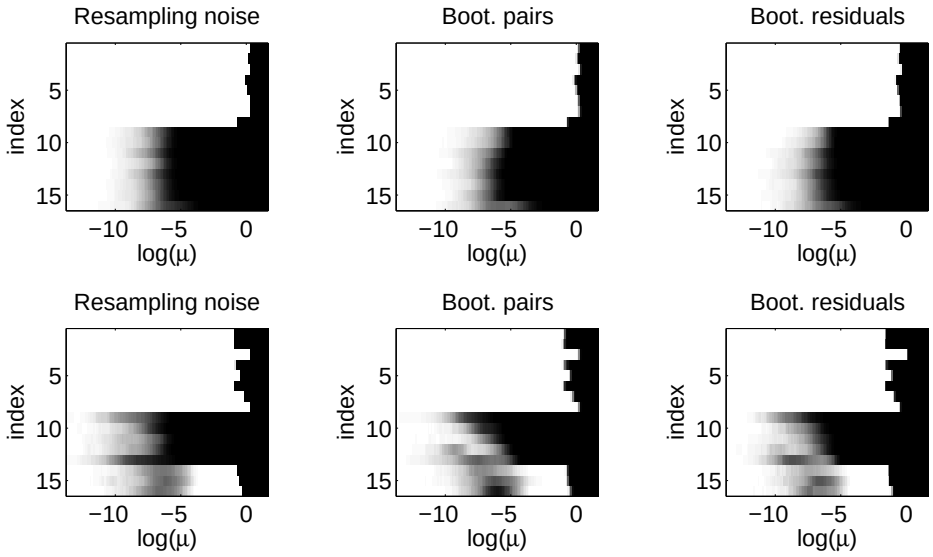

Figure 2: Probability of selecting each variable vs. regularization parameter μ (low-dimensional setting) for various resampling schemes, *before intersecting*. White values correspond to probability equal to one, and black values correspond to probability equal to zero (model consistency corresponds to while on the top 8 variables and black on the rest). Top: consistency condition of the Lasso is satisfied, Bottom: consistency condition not satistfied. Note the similar behavior of resampling noise (which requires knowing the generating distribution) and the two forms of bootstrapping (which do not). See text for details.

Bootstrapping residuals When bootstrapping residuals, we first run the Lasso once, with complexity $O(pn+k^2n)$. Then, for all values of the regularization parameter, naively, we would have to run the Lasso m times. In order to avoid running the Lasso as many times as m times the number of values of μ we want to consider, one can first notice that there are at most $O(k)$ break points in the original Lasso estimation, and that between break points, one has to minimize an objective function which is composed of a ℓ^1-penalty, a quadratic term and a linear term whose coefficients depend affinely in μ. This implies that the path is also piecewise linear within this segment and can be followed using an homotopy algorithm very similar to the one for the regular Lasso. Thus it makes $O(mpn + mk^2n)$ per segments when restarting an homotopy method for this segment, i.e., an overall complexity of $O(mkpn + mk^3n)$. This can be put down by computing a joint path that goes through all $O(k)$ segments sequentially instead of in parallel, in total time $O(mkpn + mk^2n)$. Moreover, since when bootstrapping residuals, the design matrix is the same for all replications and computations of submatrices of Q may be cached, to obtain a complexity of $O(mkpn + k^2n)$.

Similarly, when bootstrapping after projections onto the active set of a single global Lasso run, one can get even get a lower complexity of $O(pn + mk^2n)$, i.e., one Lasso followed by m Lasso on a reduced data set. This requires however updates (when the first Lasso estimation switches active sets) such as the ones proposed in [23].

5.2 Experiments - Low-Dimensional Settings

We first consider a low-dimensional design matrix, with $p = 16$, $n = 1024$ and 8 relevant variables (i.e., $\mathbf{J} = \{1, \dots, 8\}$). The design is sampled from a normal distribution with independent rows, sampled i.i.d. from a fixed covariance matrix. We consider two covariance matrices, one that leads to design matrices which do not satisfy the consistency condition of the Lasso in Eq. (2.3), and one that leads to Lasso-consistent design matrices.

In Figure 2, we plot the marginal probabilities (computed from 512 independent replications) of selecting any given of the $p = 16$ variables for all values of the regularization parameter μ and for the various resampling schemes (resampling noise, bootstrapping pairs or bootstrapping residuals), *without intersecting* (i.e., we are just reporting counts from 512 replications from a single dataset). Note that the left column (resampling noise) exactly corresponds to the various regimes of the Lasso presented in Section 2 (these require full knowledge of the generating distributions and are only displayed for illustration purposes): for large values of μ, no variable is selected (Proposition 2.1), then a fixed pattern is selected (μ tending to zero faster than $n^{-1/2}$, Proposition 2.2), then all patterns including the relevant variables (μ of order $n^{-1/2}$, Propositions 2.4 and 2.5), and finally, for small values of μ, all variables are selected (Proposition 2.6). Note that in the top plots, as expected (since Eq. (2.3) is not satisfied), some portions of the regularization paths lead to the correct pattern, while in the bottom plots, as expected (since Eq. (2.3) is satisfied), there is no consistent model selection. It is important to note that using the bootstrap leads to similar behavior than resampling the noise, but does not require extra knowledge (i.e., a single dataset is needed). Note finally, that bootstrapping residuals does alter slightly the regularization paths—because of the bias of the Lasso estimate—and the selected patterns (see other evidence of this behavior in Figure 3 and Figure 4).

In Figure 3, we compute the marginal probability of selecting the variables for the Lasso (left column) and the various ways of using the Bolasso (boostrapping pairs or residuals), i.e., *after intersecting*. Those are obtained by running the Bolasso with 512 replications, 128 times on the same design but with different noisy observations (thus, a total of 512×128 Lasso runs are used for each of the plots on the middle and right columns of Figure 3). On the top plots, the Lasso consistency condition in Eq. (2.3) is satisfied and the two versions of the Bolasso increase the width of the consistency region of the Lasso, while on the bottom plots, it is not, and the Bolasso creates a consistency region. Note that bootstrapping residuals modifies the early parts of the regularization path (i.e., large values of μ), illustrating the effect of the bias of the Lasso when bootstrapping residuals.

In Figure 4, we consider the effect of the number m of bootstrap replications, in the same two situations. Increasing m seems always beneficial. Note that (1) when $m = 1$ (essentially the Lasso), we get some strictly positive probabilities of good pattern selection even in the inconsistent case, illustrating Proposition 2.4, and (2) if m was too large, some of the relevant variables would start to leave the intersection of active sets (but this has not happened in our simulations with only 512 replications).

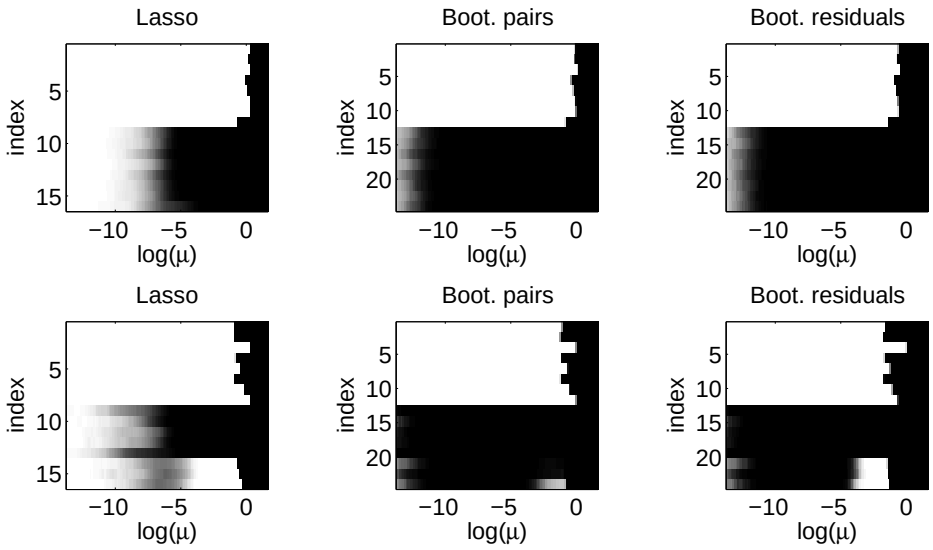

Figure 3: Probability of selecting each variable vs. regularization parameter μ (low-dimensional setting) for the Lasso (left column) and the Bolasso (middle and right columns). White values correspond to probability equal to one, and black values correspond to probability equal to zero (model consistency corresponds to while on the top 8 variables and black on the rest). Top: consistency condition of the Lasso is satisfied, Bottom: consistency condition not satistfied. See text for details.

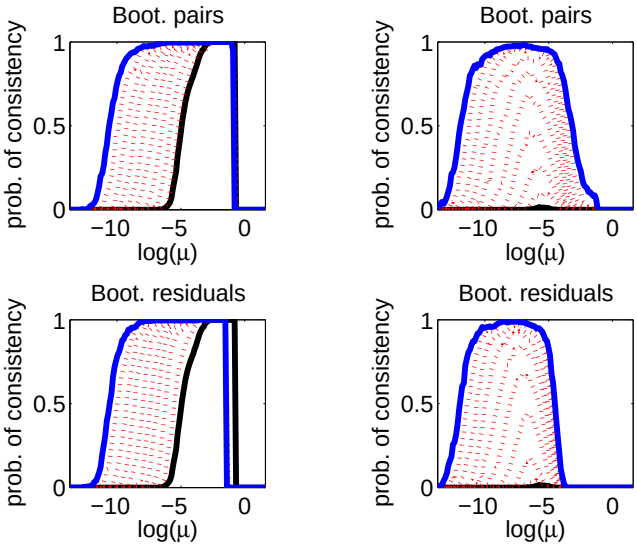

Figure 4: Probability of correct pattern selection with various numbers m of replications in $\{1$ in plain black$, 2, 4, 8, 16, 32, 64, 128, 256,$ all in dashed red$, 512$ in plain blue$\}$ (low-dimensional setting). Top: consistency condition of the Lasso is satisfied, Bottom: consistency condition not satistfied. Note that only one replication (plain black) is very similar to the regular Lasso.

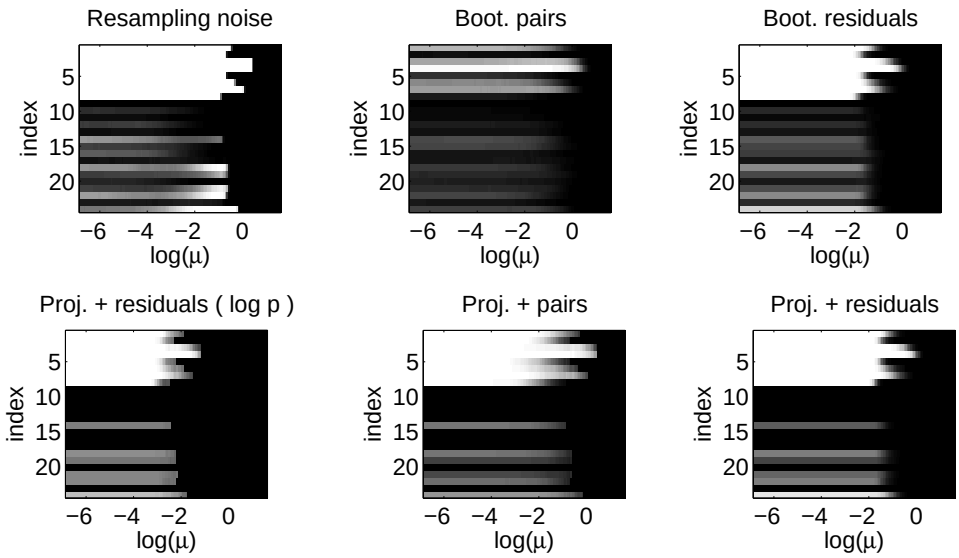

Figure 5: Probability of selecting each variable vs. regularization parameter μ (high-dimensional setting) for various resampling schemes, *before intersecting*. Only the first 8 variables and the 16 variables which violates condition in Eq. (2.3) the most are plotted. White values correspond to probability equal to one, and black values correspond to probability equal to zero (model consistency corresponds to while on the top 8 variables and black on the rest). Note the similar behavior of resampling noise (which requires knowing the generating distribution) and all forms of bootstrapping (except for bootsrapping pairs, in the top-middle plot).

5.3 Experiments - High-Dimensional Settings

We now consider a "high-dimensional" design matrix (i.e. such that $p > n$), with $p = 128$, $n = 64$ and 8 relevant variables (i.e., $\mathbf{J} = \{1, \dots, 8\}$). The design matrix is sampled from a normal distribution with i.i.d. elements. For the sampled design matrix, the condition in Eq. (2.3) is not satisfied, as for most designs with such p, n and $|\mathbf{J}|$, as shown in Figure 1 in Section 4, but assumptions (**A5-6**) are.

We performed the same simulations than in Section 5.2, with additional bootstrapping procedures, namely after projecting into the original Lasso estimate, with the same regularization parameter (no consistency result) or with a parameter multiplied by $\log p$ (consistency result in Proposition 4.2).

In Figure 5, we consider marginal probabilities *before intersection*, to study the general behavior of various resampling schemes. We see that bootstrapping procedures behave rather differently than resampling the noise (unlike in low-dimensional settings), and that boostrapping pairs does lose some of the relevant variables while boostrapping residuals does not. After projection, all resampling procedures behave correctly. In Figure 6, we compare the Lasso and the Bolasso (for several ways of performing the bootstrap): boostrapping residuals consistently leads to better performance. Note that while the top right plot behaves correctly, we currently have no proofs for it. In Figure 7, we consider the effect of various numbers of replications. Note that in the bottom-right plot, 512 replications are indeed too many (i.e., when too many replications are used, we start to lose some of the relevant variables).

6 Conclusion

We have presented a detailed analysis of the variable selection properties of a boostrapped version of the Lasso. The model estimation procedure, referred to as the Bolasso, is provably consistent under general assumptions, in low-dimensional and high-dimensional settings. We have considered the two types of bootstrap for linear regression, and have shown empirically and theoretically better properties for the bootstrap of residuals. This work brings to light that poor variable selection results of the Lasso may be easily enhanced thanks to a simple parameter-free resampling procedure. Our contribution also suggests that the use of bootstrap samples by L. Breiman in Bagging/Arcing/Random Forests [10] may have been so far slightly overlooked and considered a minor feature, while using boostrap samples may actually be a key computational feature in such algorithms for good model selection performances, and eventually good prediction performances on real datasets.

The current work could be extended in various ways: first, we have not proved yet that bootstrapping residuals, while giving nice empirical performance, is consistent in terms of model selection. Second, a similar analysis could be applied to other settings than least-square regression with the ℓ^1-norm, namely regularization by block ℓ^1-norms [39], multiple kernel learning [39], more general hierarchical norms [43, 3], and other losses such as general convex classification losses; in particular, an extension of our results to well-specified generalized linear models is straightforward, as they are locally equivalent to a problem like in Eq. (2.1), i.e., locally they are equivalent to minimizing $\frac{1}{2}(w - \mathbf{w})^\top Q(w - \mathbf{w}) - q^\top(w - \mathbf{w}) + \mu\|w\|_1$, with q being random and having as covariance matrix a multiple of Q.

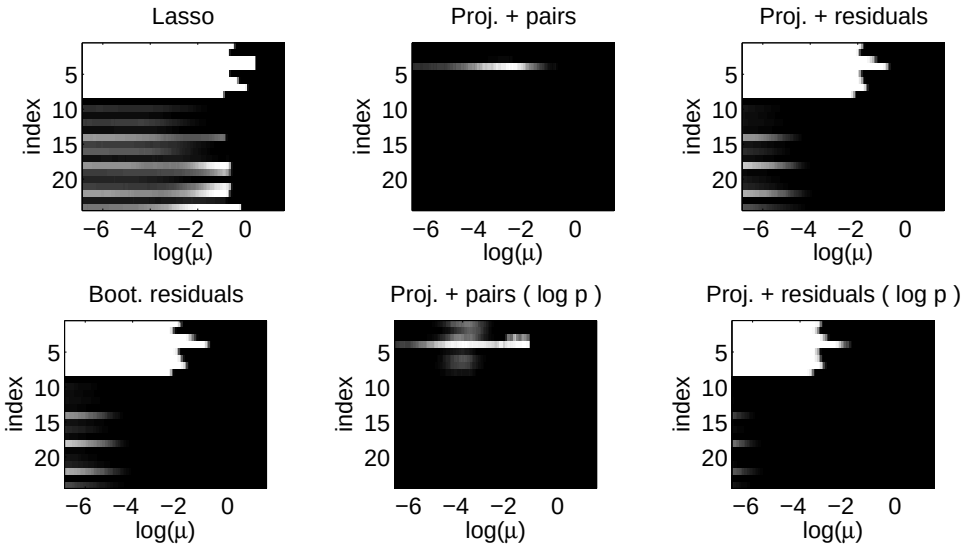

Figure 6: Probability of selecting each variable vs. regularization parameter μ (high-dimensional setting) for the Lasso (left column) and the Bolasso (middle and right columns). White values correspond to probability equal to one, and black values correspond to probability equal to zero (model consistency corresponds to while on the top 8 variables and black on the rest). See text for details.

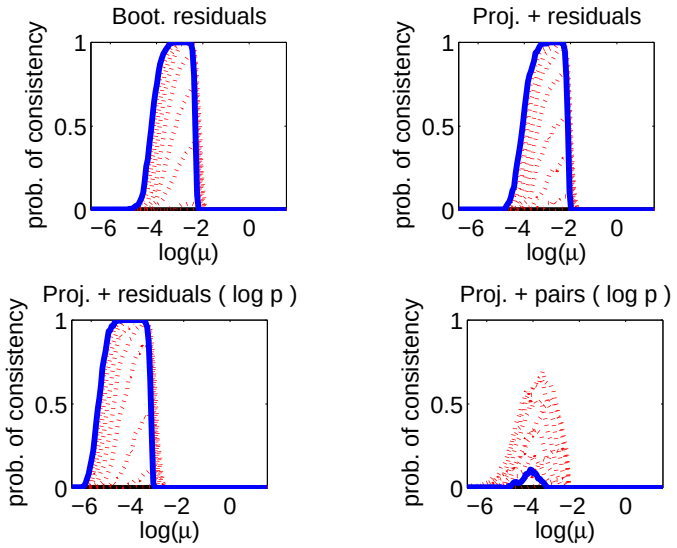

Figure 7: Probability of correct pattern selection with various numbers m of replications in $\{1$ in plain black$, 2, 4, 8, 16, 32, 64, 128, 256,$ all in dashed red$, 512$ in plain blue$\}$ (high-dimensional setting). Top: consistency condition of the Lasso is satisfied, Bottom: consistency condition not satistfied. Note that only one replication (plain black) is very similar to the regular Lasso.

Moreover, extensions to general misspecified models or models with heteroscedastic additive noise could be carried through. Also, theoretical and practical connections could be made with other work on resampling methods and boosting [11]. In particular, using the bootstrap to both select the model and estimate the regularization parameter is clearly of interest. Finally, applications of such resampling techniques for signal processing and compressed sensing [4, 13] remain to be explored, both in the context of basis pursuit (ℓ^1-norm regularization, [14]) and matching pursuit (greedy selection, [28]).

A Probability results

In this appendix, we review concentration inequalities that we will need throughout the proofs.

A.1 Multivariate Berry-Esseen Inequalities

If $X_1, \dots, X_n \in \mathbb{R}^p$ are n independent (but not indentically distributed) random vectors, with finite third-order moments, and normalized second-order moments, i.e., such that $\operatorname{var}(n^{-1/2}\sum_{i=1}^n X_i) = I$, then for all convex sets C, we have the multivariate Berry-Esseen inequality [5, 24]:

$$\left| \mathbb{P}\left(\frac{1}{n^{1/2}} \sum_{i=1}^n X_i \in C \right) - \mathbb{P}(u \in C) \right| \leqslant C_1^{\mathrm{BE}} \frac{p^{1/2}}{n^{1/2}} \left(\frac{1}{n} \sum_{i=1}^n \mathbb{E}\|X_i\|_2^3 \right), \tag{A.1}$$

where u is a standard normal random vector and C_1^{BE} is a universal constant.

We can also derive from [24] another version for expectation of bounded Lipschitz functions, i.e, if $f(x)$ is bounded by M_1 and Lipschitz, with Lipschitz constant M_2, then, we have:

$$\left| \mathbb{E} f\left(n^{-1/2} \sum_{i=1}^n X_i \right) - \mathbb{E} f(u) \right| \leqslant C_2^{\mathrm{BE}} (M_1 + M_2) \frac{p^{1/2}}{n^{1/2}} \left(\frac{1}{n} \sum_{i=1}^n \mathbb{E}\|X_i\|_2^3 \right), \tag{A.2}$$

where C_2^{BE} is a universal constant. Note that better bounds (with better scalings in p) exist in the i.i.d. case [5]. Any improvement on Berry-Esseen inequalities would lead to an improvement of our results.

In this paper, we will consider convex sets corresponding to selecting a given sign pattern (among the 3^p available ones), making use of Eq. (A.1). When considering leaving out a given variable (like in Appendix D.7), we will design a specific Lipschitz function and apply Eq. (A.2).

A.2 Concentration Inequalities for Subgaussian Variables

We consider n independent real random variables $Y_1, \dots, Y_n$, which are subgaussian with zero mean and uniform subgaussian constant, i.e., there exists $\tau > 0$ such that for all $i \in \{1, \dots, n\}$ and all $s \in \mathbb{R}$, $\mathbb{E}(e^{sY_i}) \leqslant e^{s^2\tau^2/2}$. Then, we have [30, 8]:

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n Y_i \geqslant t \right) \leqslant e^{-nt^2/2\tau^2}. \tag{A.3}$$

Note that the variance of Y is always then less than τ^2 (with equality if and only if Y is normally distributed). We will also use Hoeffding inequality for bounded variables, which amounts to use the fact that if $|Y| \leqslant M$, then Y is subgaussian with constant $\tau^2 = M^2/4$ [8, 30].

We will also use concentration inequalities for quadratic forms [38] in independent random subgaussian variables, with universal strictly positive constants C_1^{q}, C_2^{q}, C_3^{q}: for all symmetric matrices A, if $|A|$ denotes the matrix of absolute values of elements of A, then

$$\mathbb{P}\big(Y^\top AY - \mathbb{E}(Y^\top AY) \geqslant t\big) \leqslant C_1^{\mathrm{q}} \exp\left(-\min\left\{\frac{C_2^{\mathrm{q}} t\tau^{-2}}{\||A|\|_2}, \frac{C_3^{\mathrm{q}} t^2\tau^{-4}}{\|A\|_F^2}\right\}\right). \tag{A.4}$$

B Perturbation of positive matrices

In this appendix, we review known results of perturbation of positive matrices. Let Q and R be two positive matrices of size p, A and B two disjoint subsets of $\{1, \dots, p\}$ such that $A \cup B = \{1, \dots, p\}$. We have [25]:

$$\begin{aligned}
\|Q^{-1} - R^{-1}\|_2 &\leqslant \frac{1}{\lambda_{\min}(Q)\lambda_{\min}(R)} \|Q - R\|_2, \\
\|Q^{1/2} - R^{1/2}\|_2 &\leqslant \frac{\max\{\lambda_{\max}(Q), \lambda_{\max}(R)\}^{1/2}}{2\max\{\lambda_{\min}(Q), \lambda_{\min}(R)\}} \|Q - R\|_2, \\
\|Q^{-1/2} - R^{-1/2}\|_2 &\leqslant \frac{1}{2\max\{\lambda_{\min}(Q), \lambda_{\min}(R)\}^{3/2}} \|Q - R\|_2,
\end{aligned}$$

$$\|Q_{A,B}Q_{B,B}^{-1} - R_{A,B}R_{B,B}^{-1}\|_2 \leqslant \frac{\|Q_{A,B} - R_{A,B}\|_2}{\lambda_{\min}(Q_{B,B})} + \frac{\lambda_{\max}(R_{A,A})^{1/2}}{\lambda_{\min}(R_{B,B})^{3/2}} \|Q_{B,B} - R_{B,B}\|_2.$$

C Optimization lemmas

The following three lemmas give error bounds on the Lasso estimates and conditions for a sign pattern $s \in \{-1, 0, 1\}^p$ to be the one of the unique solution $\hat{w}$ to Eq. (1.1) or Eq. (2.1).

Lemma C.1. *Assume (**A1**) and (**A4**). Let $s \in \{0, -1, 1\}^p$ and $J = \{j, s_j \neq 0\}$. Then s is selected (i.e., $\operatorname{sign}(\hat{w}) = s$) if and only if:*

$$\|Q_{J^c,J}Q_{J,J}^{-1}q_J - q_{J^c} - Q_{J^c,J^c}\mathbf{w}_{J^c} - \mu Q_{J^c,J}Q_{J,J}^{-1}s_J\|_\infty \leqslant \mu, \tag{C.1}$$

$$\operatorname{sign}(\mathbf{w}_J + Q_{J,J}^{-1}q_J - \mu Q_{J,J}^{-1}s_J) = s_J. \tag{C.2}$$

The solution then satisfies $\hat{w}_J = \mathbf{w}_J + Q_{J,J}^{-1}(q_J - \mu s_J)$.

Proof. Following standard results in non-smooth convex optimization [9, 7], w is optimal for Eq. (1.1) or Eq. (2.1), if and only if, for all $j \in \{1, \dots, p\}$ such that $w_j \neq 0$, then, $[Q(w - \mathbf{w})]_j - q_j + \mu\operatorname{sign}(w_j) = 0$, and for all other j, $|[Q(w - \mathbf{w})]_j - q_j| \leqslant \mu$. We thus get $\hat{w}_J = \mathbf{w}_J + Q_{J,J}^{-1}(q_J - \mu s_J)$, and the result follows from expressing that $\hat{w}_J$ should have the right sign on J—Eq. (C.2)—and that the directional derivatives along other directions are positive—Eq. (C.1). □

When the sign pattern is consistent on $\mathbf{J}$ with $\mathbf{w}$, then we can further refine the conditions of Lemma C.1:

Lemma C.2. *Assume (**A1**) and (**A4**). Let $s \in \{0,-1,1\}^p$ such that $s_{\mathbf{J}} = \operatorname{sign}(\mathbf{w}_{\mathbf{J}})$ and let $J = \{j, s_j \neq 0\} \supset \mathbf{J}$. Then s is selected if and only if:*

$$\|Q_{J^c,J}Q_{J,J}^{-1}q_J - q_{J^c} - \mu Q_{J^c,J}Q_{J,J}^{-1}s_J\|_\infty \leqslant \mu, \tag{C.3}$$

$$\operatorname{sign}(\mathbf{w}_{\mathbf{J}} + (Q_{J,J}^{-1}q_J - \mu Q_{J,J}^{-1}s_J)_{\mathbf{J}}) = \operatorname{sign}(\mathbf{w}_{\mathbf{J}}), \tag{C.4}$$

$$\operatorname{sign}(Q_{J,J}^{-1}q_J - \mu Q_{J,J}^{-1}s_J)_{J\setminus\mathbf{J}} = s_{J\setminus\mathbf{J}}. \tag{C.5}$$

The solution then satisfies $w_J = \mathbf{w}_J + Q_{J,J}^{-1}(q_J - \mu s_J)$.

Lemma C.3. *Assume (**A1**) and (**A4**). We have $\|\hat{w}-\mathbf{w}\|_2 \leqslant \frac{p^{1/2}\mu+\|q\|_2}{\lambda_{\min}(Q)}$ and $\|Q^{1/2}(\hat{w}-\mathbf{w})\|_2 \leqslant \frac{p^{1/2}\mu+\|q\|_2}{\lambda_{\min}(Q)^{1/2}}$.*

Proof. From optimality conditions, we have $\|Q(\hat{w}-\mathbf{w})-q\|_\infty \leqslant \mu$, from which we get $\|Q^{1/2}(\hat{w}-\mathbf{w})\|_2 \leqslant \lambda_{\min}(Q)^{-1/2}\left(\|Q(\hat{w}-\mathbf{w})-q\|_2 + \|q\|_2\right)$. The results follow from the identity $\|a\|_2 \leqslant p^{1/2}\|a\|_\infty$ for any $a \in \mathbb{R}^p$. □

The following lemma relates the solutions of Eq. (2.1) for different values of Q and q. This will be used in Appendix D.7 to prove the Lipschitz continuity of the solution of Eq. (2.1) as a function of q.

Lemma C.4. *If $\hat{w}$ is solution of Eq. (2.1) for Q, q, and $\hat{w}'$ is solution for Q', q', then we have:*

$$\|Q^{1/2}(\hat{w}-\hat{w}')\|_2 \leqslant 2\|Q^{-1}(q-q')\|_2 + \frac{2\|(Q')^{-1/2}(Q-Q')Q^{-1/2}\|_2}{\lambda_{\min}(Q')^{1/2}}\left[p^{1/2}\mu + \|q'\|_2\right].$$

Let $\gamma = \frac{Q^{1/2}(\hat{w}-\mathbf{w}-Q^{-1}q)}{\mu}$, then if $Q = Q'$, $\|\gamma-\gamma'\|_2 \leqslant \frac{3\|Q^{-1}(q-q')\|_2}{\mu}$, and if $q = q'$, then $\|Q^{-1/2}\gamma - (Q')^{-1/2}\gamma'\|_2 \leqslant \frac{2\|(Q')^{-1/2}(Q-Q')Q^{-1/2}\|_2}{\mu\lambda_{\min}(Q)^{1/2}\lambda_{\min}(Q')^{1/2}}\left[p^{1/2}\mu + \|q'\|_2\right]$.

Proof. We let denote $J(w) = \frac{1}{2}(w-\mathbf{w})^\top Q(w-\mathbf{w}) - q^\top(w-\mathbf{w}) + \mu\|w\|_1$ the Lasso cost function. A short calculation shows that for all z such that $z^\top Qz = 1$, $J(\hat{w}'+\alpha z) - J(\hat{w}')$ is larger than

$$\frac{\alpha^2}{2} - \left(\|Q^{-1/2}(q-q')\|_2 + \|(Q')^{1/2}(\hat{w}'-\mathbf{w})\|_2\|(Q')^{-1/2}(Q-Q')Q^{-1/2}\|_2\right)\alpha.$$

If the last expression is nonnegative, since J is convex, the (unique, because Q is invertible) minimum $\hat{w}$ of J must occur within the convex set $\{w, \|Q^{1/2}(w-w')\|_2 \leqslant \alpha\}$. The first result follows, using Lemma C.3. Other results are direct consequences of using results from Appendix B. □

D Proofs for low-dimensional results

Note that assumption (**A3**) implies a bound on the largest eigenvalue of the matrix Q, i.e., $\lambda_{\max}(Q) \leqslant p\|Q\|_\infty \leqslant pM^2$. Moreover, we have for all $J \subset \{1,\dots,p\}$, $k \in \{1,\dots,n\}$ and $j \in J^c$:

$$|x_{kj} - Q_{j,J}Q_{J,J}^{-1}x_{kJ}| \leqslant M + \frac{M}{\lambda_{\min}(Q)^{1/2}} \times |J|^{1/2}M \leqslant \frac{2M|J|^{1/2}}{\tilde{\lambda}^{1/2}},$$

which leads to

$$\mathbb{P}(\|q_{J^c} - Q_{J^c,J}Q_{J,J}^{-1}q_J\|_\infty \geqslant tM\sigma) \leqslant 2p\exp\left(-\frac{t^2\tilde{\lambda}}{8\tilde{\tau}^2}\frac{n}{|J|}\right). \tag{D.1}$$

D.1 Proof of Proposition 2.1

The null vector $\hat{w} = 0$ is solution of Eq. (2.1), if and only if $\|Q\mathbf{w}+q\|_\infty \leqslant \mu$, which is the case, as soon as $\mu \geqslant M^2\|\mathbf{w}\|_1 + \|q\|_\infty$ (because $\|Q\mathbf{w}\|_\infty \leqslant M^2\|\mathbf{w}\|_1$), and, thus with the additional assumption $\mu \geqslant 2M^2\|\mathbf{w}\|_1$, as soon as $\|q\|_\infty \leqslant \mu/2$. We have, by the union bound:

$$\mathbb{P}(\|q\|_\infty \leqslant \mu/2) \geqslant 1 - \sum_{j=1}^p \mathbb{P}(|q_j| \geqslant \mu/2) \geqslant 1 - 2p\exp\left(-\frac{n\tilde{\mu}^2}{8\tilde{\tau}^2}\right),$$

because we have $\mathbb{E}(e^{sx_{ij}\varepsilon_i}) \leqslant e^{s^2\tau^2M^2/2}$ for all $s \in \mathbb{R}$ and $j \in \{1,\dots,p\}$, and by Eq. (A.3).

D.2 Proof of Proposition 2.2

If $J(w) = \frac{1}{2}(w-\mathbf{w})^\top Q(w-\mathbf{w}) - q^\top(w-\mathbf{w}) + \mu\|w\|_1$ is the Lasso cost function, we have, for all $z \in \mathbb{R}^p$ such that $\|z\|_2 = 1$ and $\alpha > 0$,

$$\begin{aligned} J(w_0+\alpha z) &\geqslant J(w_0) + \lambda_{\min}(Q)\alpha^2/2 - q^\top\alpha z + (\mu-\mu_0)(\|w_0+\alpha z\|_1 - \|w_0\|_1), \\ &\geqslant J(w_0) + \lambda_{\min}(Q)\alpha^2/2 - \alpha(\|q\|_2 + |\mu-\mu_0|p^{1/2}), \end{aligned}$$

which implies $\|\hat{w} - w_0\|_2 \leqslant 2\lambda_{\min}(Q)^{-1}\|q\|_2 + 2|\mu-\mu_0|\lambda_{\min}(Q)^{-1}p^{1/2}$. The first inequality follows from $\mathbb{P}(\|q\|_2 \geqslant t) \leqslant 2p\exp(-t^2n/2pM^2\tau^2)$, applied with $t = \lambda_{\min}(Q)\beta\sigma/4M$.

We let denote s and J the sign and support patterns of v. We have from optimality conditions of the noiseless problem, $(w_0)_J = \mathbf{w}_J - \mu_0 Q_{J,J}^{-1}s_J$ and $\|(Q(w_0-\mathbf{w}))_{J^c}\|_\infty \leqslant \mu_0 - \eta M\sigma$. We now need sufficient conditions for Eq. (C.1) and Eq. (C.2) in Lemma C.1. For Eq. (C.2), we need that $\operatorname{sign}((w_0)_J + Q_{J,J}^{-1}q_J + (\mu_0-\mu)Q_{J,J}^{-1}s_J) = s_J$. If $|\mu-\mu_0| \leqslant \frac{\lambda_{\min}(Q)\mathrm{m}(w_0)}{2p^{1/2}} = \frac{M\sigma\tilde{\lambda}\mathrm{m}(w_0M/\sigma)}{2p^{1/2}}$, then

$$\|(\mu_0-\mu)Q_{J,J}^{-1}s_J\|_\infty \leqslant |\mu-\mu_0|\lambda_{\min}(Q)^{-1}p^{1/2} \leqslant \mathrm{m}(w_0)/2,$$

and then Eq. (C.2) is satisfied as soon as $(Q_{J,J}^{-1}q_J)_j s_j \geqslant -\mathrm{m}(w_0)/2$, for all $j \in J$, which occurs with probability greater than $1 - p\exp\left(\frac{-n\mathrm{m}(w_0M/\sigma)^2\tilde{\lambda}^2}{8\tilde{\tau}^2p}\right)$.

For Eq. (C.1), we assume that $\|q_{J^c} - Q_{J^c,J}Q_{J,J}^{-1}q_J\|_\infty \leqslant \eta M\sigma/2$, which occurs with probability obtained from Eq. (D.1) (with $t = \eta/2$). Also, $|\mu - \mu_0| \leqslant \frac{\eta\lambda_{\min}(Q)^{1/2}M\sigma}{4p^{1/2}M} \leqslant \frac{\eta M\sigma/2}{2p^{1/2}\frac{M^{1/2}}{\lambda_{\min}(Q)^{1/2}}} \leqslant \frac{M\sigma\eta/2}{1+p^{1/2}\frac{M^{1/2}}{\lambda_{\min}(Q)^{1/2}}}$. This implies

$$\|q_{J^c} - Q_{J^c,J}Q_{J,J}^{-1}q_J\|_\infty \leqslant \eta M\sigma - |\mu - \mu_0|(1 + p^{1/2}M^{1/2}\lambda_{\min}(Q)^{-1/2}),$$

$$\|q_{J^c} - Q_{J^c,J}Q_{J,J}^{-1}q_J\|_\infty + \|Q(w_0 - \mathbf{w})\|_\infty + |\mu - \mu_0|p^{1/2}M^{1/2}\lambda_{\min}(Q)^{-1/2} \leqslant \mu,$$

because $\|Q_{J,J}^{-1}s_J\|_\infty \leqslant Mp^{1/2}\lambda_{\min}(Q)^{-1/2}$; hence the desired result.

D.3 Proof of Proposition 2.3

Note that $\mu \leqslant \frac{\mathrm{m}(\mathbf{w})\lambda_{\min}(Q)}{p^{1/2}}$, and $\|\Delta\|_2 \leqslant \lambda_{\min}(Q)^{-1}p^{1/2}$ implies that $\mathrm{sign}(\mathbf{w_J}+\mu\Delta_\mathbf{J}) = \mathrm{sign}(\mathbf{w_J})$. Thus, if $\|z\|_2 = 1$ and $\alpha > 0$, we have:

$$\begin{aligned} J(\mathbf{w} + \mu\Delta + \alpha z) &\geqslant J(\mathbf{w} + \mu\Delta) + \lambda_{\min}(Q)\alpha^2/2 - q^\top\alpha z + (\mu\Delta)^\top Q\alpha z + \\ &\quad \mu(\|\mathbf{w} + \mu\Delta + \alpha z\|_1 - \|\mathbf{w} + \mu\Delta\|_1), \\ &\geqslant J(\mathbf{w} + \mu\Delta) + \lambda_{\min}(Q)\alpha^2/2 - q^\top\alpha z, \end{aligned}$$

which implies $\|\hat{w} - \mathbf{w} - \mu\Delta\|_2 \leqslant 2\lambda_{\min}(Q)^{-1}\|q\|_2$, ans thus the first inequality.

We let denote s the sign pattern of Δ and J its support. Since, by assumption $\tilde{\mu} \leqslant \frac{\mathrm{m}(\tilde{\mathbf{w}})\tilde{\lambda}}{2p^{1/2}}$, we have $\mu\|Q_{JJ}^{-1}s_J)_\mathbf{J}\|_\infty \leqslant \mathrm{m}(\mathbf{w})/2$, if $\|(Q_{J,J}^{-1}q_J)_\mathbf{J}\|_2 \leqslant \frac{1}{2}\mathrm{m}(\mathbf{w})$, which occurs with probability greater than $1 - 2|\mathbf{J}|\exp(-n\frac{\mathrm{m}(\tilde{\mathbf{w}})\tilde{\lambda}^2 n}{8\tilde{\tau}^2 p})$, then Eq. (C.4) is satisfied.

If $\|q_{J^c} - Q_{J^c,J}Q_{J,J}^{-1}q_J\|_\infty \leqslant \mu\eta$, then Eq. (C.3) is satisfied, and this occurs with probability greater than $1 - 2p\exp\left(-\frac{\tilde{\lambda}n\eta^2\tilde{\mu}^2}{8|J|\tilde{\tau}^2}\right)$ (from Eq. (D.1)). Finally, if for all $j \in J\backslash\mathbf{J}$, $(Q_{J,J}^{-1}q_J)_j s_j \geqslant -\mu|\Delta_j|$, then Eq. (C.5) follows. This occurs with probability greater than $1 - p\exp(-\tilde{\lambda}^2\mathrm{m}(M^2\Delta)^2\tilde{\mu}^2 n/2\tilde{\tau}^2 p)$. The result follows by the union bound.

D.4 Proof of Proposition 2.4

The optimality condition in Eq. (C.4) from Lemma C.2 is satisfied as long as $\tilde{\mu} \leqslant \frac{\mathrm{m}(\tilde{\mathbf{w}})\tilde{\lambda}}{2p^{1/2}}$, and $\|(Q_{J,J}^{-1}q_J)_\mathbf{J}\|_2 \leqslant \frac{1}{2}\mathrm{m}(\mathbf{w})$, which occurs with probability greater than $1-2|\mathbf{J}|\exp(-n\mathrm{m}(\tilde{\mathbf{w}})\tilde{\lambda}^2 n/8\tilde{\tau}^2 p)$, while the intersection of events in Eq. (C.3) and Eq. (C.5), by the Berry-Esseen inequalities, converges to the probability that $\mathbb{P}(u \in C)$, where u is normal with zero mean and covariance matrix Q and C is the convex set defined as the intersection of

$$\left\{\left[(Q_{J,J}^{-1}u_J)_{J\backslash\mathbf{J}} - \mu n^{1/2}\sigma^{-1}(Q_{J,J}^{-1}s)_{J\backslash\mathbf{J}}\right] \circ s_{J\backslash\mathbf{J}} \geqslant 0\right\}$$

and

$$\left\{\|u_{J^c} - Q_{J^c,J}Q_{J,J}^{-1}u_J - \mu n^{1/2}\sigma^{-1}Q_{J^c,J}Q_{J,J}^{-1}s_J\|_\infty \leqslant \mu n^{1/2}\sigma^{-1}\right\}.$$

The set C and its complement have non-empty interior and since Q is full-rank, the probability is strictly inside the interval $(0,1)$. Moreover, by Eq. (A.1), the error bound is upperbounded by C_1^{BE} times

$$\frac{p^{1/2}}{n^{1/2}}\left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}\|(\sigma^2 Q)^{-1/2}\varepsilon_i x_i\|_2^3\right) \leqslant \frac{p^{1/2}}{n^{1/2}}\frac{4Mp^{1/2}\tau^3}{\sigma^3\lambda_{\min}(Q)^{1/2}}\left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}\|Q^{-1/2}x_i\|_2^2\right), \\ \leqslant \frac{p^{1/2}}{n^{1/2}}\frac{4Mp^{1/2}\tau^3}{\sigma^3\lambda_{\min}(Q)^{1/2}}p = \frac{p^2}{n^{1/2}}\frac{4\tilde{\tau}^3}{\tilde{\lambda}^{1/2}},$$

because $\mathbb{E}|\varepsilon_i|^3 \leqslant 4\tau^3$, which leads to the desired result.

D.5 Proof of Proposition 2.5

We simply use Lemma C.3: $\|\hat{w}-\mathbf{w}\|_2 \leqslant \frac{p^{1/2}\mu+\|q\|_2}{\lambda_{\min}(Q)}$. Thus if $\mathrm{m}(\mathbf{w}) > \frac{p^{1/2}\mu+\|q\|_2}{\lambda_{\min}(Q)}$, the result follows from concentration inequalities in Appendix A.

D.6 Proof of Proposition 2.6

We have, by considering all patterns consistent with the total absence of zeros:

$$\mathbb{P}(\exists j \in \{1,\dots,p\}, \hat{w}_j = 0) \leqslant \sum_{s, \exists j \in \{1,\dots,p\}, s_j = 0} \mathbb{P}(\mathrm{sign}(\hat{w}) = s).$$

We now consider such a pattern and its support (strictly included in $\{1,\dots,p\}$). From optimality conditions in Eq. (C.1), we get that $\mathrm{sign}(\hat{w}) = s$ implies that $\|q_j - Q_{j,J}Q_{J,J}^{-1}q_J - \mu Q_{j,J}Q_{J,J}^{-1}s_J\|_\infty \leqslant \mu$ for some $j \in J^c \neq \varnothing$. Note that the covariance matrix of $q_j - Q_{j,J}Q_{J,J}^{-1}q_J$ is equal to $\sigma^2 Q_{j,j|J}/n$ and has a lowest eigenvalue greater than $\sigma^2\lambda_{\min}(Q)/n$. Thus, by the Berry-Esseen inequality,

$$\mathbb{P}(\mathrm{sign}(\hat{w}) = s) \leqslant C_1^{\mathrm{BE}}\frac{4\tilde{\tau}^3}{\tilde{\lambda}^{1/2}}\frac{p^2}{n^{1/2}} + \frac{\tilde{\mu}n^{1/2}}{\tilde{\lambda}^{1/2}},$$

which implies the desired result, since there are at most 3^p allowed patterns.

We can get a better bound (with respect to p), but with a weaker dependence in μ, that is, we consider:

$$\mathbb{P}(\exists j \in \{1,\dots,p\}, \hat{w}_j = 0) \leqslant \mathbb{P}(\exists j \in \mathbf{J}, \hat{w}_j = 0) + \mathbb{P}(\exists j \in \mathbf{J}^c, \hat{w}_j = 0).$$

The first term is upper bounded by $2|\mathbf{J}|\exp(-n\mathrm{m}(\tilde{\mathbf{w}})\tilde{\lambda}^2 n/8\tilde{\tau}^2 p)$, while the second one is upper-bounded using Proposition 2.7. This leads to the global desired upper bound, which scales better in p but worse in n. In particular, it requires that μn tends to infinity, i.e., μ is not too small.

D.7 Proof of Proposition 2.7

We first start with a simple elementary lemma:

Lemma D.1. *If $u \in \mathbb{R}$ is a standard normal random variable, then*

$$\alpha \geqslant \mathbb{P}(|u-\beta| \leqslant \alpha) \geqslant \frac{\alpha}{1+\alpha} e^{-\beta^2/2}.$$

When minimizing Eq. (2.1), with the constraint that $w_j = 0$, we get the solution (with the notation $j^c = \{j\}^c$):

$$w_{j^c} = \mathbf{w}_{j^c} + Q^{-1}_{j^c,j^c} q_{j^c} + \mu Q^{-1/2}_{j^c,j^c} \gamma^j_{\mu,Q}(q_{j^c}),$$

for a certain $\gamma^j_{\mu,Q}(q_{j^c})$ such that $\|Q^{1/2}_{j^c,j^c}\gamma^j_{\mu,Q}(q_{j^c})\|_\infty \leqslant 1$. It is optimal for the full problem if and only if (because $\mathbf{w}_j = 0$), $|Q_{j,j^c}(w_{j^c} - \mathbf{w}_{j^c}) - q_j| \leqslant \mu$, i.e.,

$$|-q_j + Q_{j,j^c}Q^{-1}_{j^c,j^c}q_{j^c} + \mu Q_{j,j^c}Q^{-1/2}_{j^c,j^c}\gamma^j_{\mu,Q}(q_{j^c})| \leqslant \mu.$$

We consider the "soft indicator" function (triangle-shaped) $f^j_{\mu,Q}(q)$ of q defined as

$$f^j_{\mu,Q}(q) = \left(1 - \mu^{-1}\left|q_j - Q_{j,j^c}Q^{-1}_{j^c,j^c}q_{j^c} - Q_{j,j^c}\mu Q^{-1/2}_{j^c,j^c}\gamma^j_{\mu,Q}(q_{j^c})\right|\right)_+.$$

The function $f^j_{\mu,Q}$ is upper bounded by 1, moreover, from Lemma C.4, $\gamma^j_{\mu,Q}$ is Lipschitz with constant $L = \mu^{-1}\frac{3}{\lambda_{\min}(Q)^{1/2}}$. Thus, $f^j_{\mu,Q}$ is Lipshitz with constant $2\mu^{-1}\frac{M^{1/2}}{\lambda_{\min}(Q)^{1/2}} + LM \leqslant 5\mu^{-1}\frac{M}{\lambda_{\min}(Q)^{1/2}}$.

Moreover (by design) we have

$$f^j_{\mu,Q}(q) \leqslant 1_{|-q_j + Q^{-1/2}_{j,j^c}Q^{-1/2}_{j^c,j^c}q_{j^c} + Q_{j,j^c}\mu Q^{-1/2}_{j^c,j^c}\gamma^j_{\mu,Q}(q_{j^c})| \leqslant \mu},$$

thus $\mathbb{E}f^j_{\mu,Q}(q) \leqslant \mathbb{P}(j \notin \hat{J})$. This implies by the Berry-Esseen bound (see Appendix A.1), that, if q_{G} denotes the Gaussian approximation:

$$\mathbb{P}(j \notin \hat{J}) \geqslant \mathbb{E}f^j_{\mu,Q}(q_{\mathrm{G}}) - C_2^{\mathrm{BE}}\frac{p^{1/2}}{n^{1/2}}\left(\frac{5p^{1/2}n^{-1/2}}{\tilde{\mu}\tilde{\lambda}^{1/2}} + 1\right)\frac{4\tilde{\tau}^3 p^{3/2}}{\tilde{\lambda}^{1/2}},$$

because the average third order moment of the normalized variable is equal to $\frac{4\tilde{\tau}^3 p^{3/2}}{\tilde{\lambda}^{1/2}}$ and the Lipshitz constant of the function of the normalized variable is equal to $\frac{4\mu^{-1}M}{\lambda_{\min}(Q)^{1/2}} \times n^{-1/2}\sigma p^{1/2} M = \frac{5p^{1/2}n^{-1/2}}{\tilde{\mu}\tilde{\lambda}^{1/2}}$.

Moreover, we can lower bound, for any q,

$$\mathbb{E}f^j_{\mu,Q}(q) \geqslant \frac{1}{2}\mathbb{P}(|-q_j + Q_{j,j^c}Q^{-1}_{j^c,j^c}q_{j^c} + \mu Q_{j,j^c}Q^{-1/2}_{j^c,j^c}\gamma^j_{\mu,Q}(q_{j^c})| \leqslant \mu/2).$$

When applied to the Gaussian limiting distribution q_{G}, we know that the random variable $n^{1/2}\sigma^{-1}(-q_j + Q_{j,j^c}Q^{-1}_{j^c,j^c}q_{j^c})$ is asymptotically normal with mean zero and covariance $\kappa^2 =$

$Q_{j,j|j^c}$. We get by applying Lemma D.1 with $\beta = \frac{\mu n^{1/2}\sigma^{-1}}{\kappa} Q_{j,j^c} Q_{j^c,j^c}^{-1/2} \gamma^j_{\mu,Q}(q_{j^c})$ and $\alpha = \frac{\mu n^{1/2}\sigma^{-1}}{\kappa}$, which are such that $|\beta| \leqslant \frac{\mu n^{1/2}\sigma^{-1} M p^{1/2}}{\kappa \lambda_{\min}(Q)^{1/2}}$:

$$\mathbb{E}\left[f^j_{\mu,Q}(q_{\mathrm{G}})|(q_{\mathrm{G}})_{j^c}\right] \geqslant \frac{\frac{\mu n^{1/2}\sigma^{-1}}{2\kappa}}{1+\frac{\mu n^{1/2}\sigma^{-1}}{2\kappa}} \frac{1}{2} \exp\left[-\frac{\mu^2 n}{2\sigma^2\kappa^2} M \lambda_{\min}(Q)^{-1} p\right],$$

which leads to, with $\kappa \leqslant M$ and $\kappa \geqslant \lambda_{\min}(Q)^{1/2}$,

$$\mathbb{E} f^j_{\mu,Q}(q_{\mathrm{G}}) \geqslant \frac{\frac{\mu n^{1/2}\sigma^{-1}}{2M}}{1+\frac{\mu n^{1/2}\sigma^{-1}}{2\lambda_{\min}(Q)^{1/2}}} \frac{1}{2} \exp\left[-\frac{\mu^2 np}{2\sigma^2} \frac{M^2}{\lambda_{\min}(Q)^2}\right].$$

Similarly, we can get an upper bound on the probability of not selecting the variable j. We consider the same technique, but we now need to upperbound a probability of the type $\mathbb{E} f^j_{\mu,Q}(q_{\mathrm{G}})$, which leads to the desired result.

D.8 Proof of Proposition 3.1

Following the analysis in Section 3, we need to upper bound $\mathbb{P}(j \in \hat{J}^*)^m$ and $\mathbb{P}((\hat{J}^*)^c \cup \mathbf{J} \neq \varnothing)$. We obtain $\mathbb{P}((\hat{J}^*)^c \cup \mathbf{J} \neq \varnothing) \leqslant 2p \exp\left(-\frac{\mathrm{m}(\tilde{\mathbf{w}})\tilde{\lambda}^2}{8\tilde{\tau}^2 p} n\right)$ from Proposition 2.5. From Proposition 2.7, we get

$$\mathbb{P}(j \in \hat{J}^*) \geqslant \frac{\tilde{\mu} n^{1/2}/4}{1+\tilde{\mu} n^{1/2}/2\tilde{\lambda}^{1/2}} \exp\left(-\frac{\tilde{\mu}^2}{2\tilde{\lambda}^2} np\right) - \frac{10 C_2^{\mathrm{BE}}}{\tilde{\tau}^3 \tilde{\lambda}^1} \frac{p^3}{\tilde{\mu} n p^{1/2}} - \frac{4 C_2^{\mathrm{BE}} \tilde{\tau}^3 p^{5/2}}{\tilde{\lambda}^{1/2}}.$$

We let $h(c) = \frac{1}{2} \frac{c/4}{1+c/2\tilde{\lambda}^{1/2}} \exp\left(-\frac{2c^2}{\tilde{\lambda}^2}\right)$, and $g(c) = \left(\frac{8 C_2^{\mathrm{BE}}}{\tilde{\tau}^3\tilde{\lambda}^1}\frac{1}{c} + \frac{4 C_2^{\mathrm{BE}}}{\tilde{\tau}^3\tilde{\lambda}^{1/2}}\right)^2 h(c)^{-2}$, and $f(c) = -\log(1-h(c))$, to get the desired result.

D.9 Proof of Proposition 3.2

Using the same reasoning as in Appendix D.8, we get the same $f(c)$ and $a(c) = \frac{10 C_2^{\mathrm{BE}}}{\tilde{\tau}^3\tilde{\lambda}^1}\frac{1}{c} + \frac{4 C_2^{\mathrm{BE}}}{\tilde{\tau}^3\tilde{\lambda}^{1/2}}$.

E Proofs for boostrapping pairs

E.1 Concentration inequalities

We now assume that we have a bootstrap sample X^* and y^*, which leads to Q^* and q^*. We now derive concentration inequalities for q^* and Q^*, that we use in Appendix E.2.

For all $a, b \in \{1, \dots, p\}$, Q^*_{ab} is an average of variables bounded by M^2. Thus, by Hoeffding's inequality [8] and the union bound:

$$\mathbb{P}(\|Q^* - Q\|_\infty \geqslant t M^2) \leqslant 2p^2 \exp\left(-2nt^2\right). \tag{E.1}$$

Similarly, we bound the deviation between q and q^*:

$$\mathbb{P}(\|q - q^*\|_\infty \geqslant tM\sigma|\varepsilon) \leqslant 2p \exp\left(-2nt^2 \frac{\sigma^2}{\|\varepsilon\|_\infty^2}\right). \tag{E.2}$$

Also, by the central limit theorem, given ε, $n^{1/2}(q^* - q)$ converges in distribution to a normal variable with mean zero and covariance matrix

$$\sigma^2 \widetilde{Q} = \mathbb{E}\left[(\varepsilon_1^*)^2 x_1^* (x_1^*)^\top | \varepsilon\right] - \mathbb{E}[\varepsilon_1^* x_1^* | \varepsilon] \mathbb{E}[\varepsilon_1^* x_1^* | \varepsilon]^\top = \frac{1}{n} \sum_{i=1}^n \varepsilon_i^2 x_i x_i^\top - qq^\top.$$

We can derive concentration inequalities of $\widetilde{Q}$ around Q, by using Appendix A.2 and Eq. (A.4):

Lemma E.1. *Assume (**A**1-4). We have:*

$$\mathbb{P}(\|\widetilde{Q} - Q\|_\infty \geqslant tM^2) \leqslant 2p^2 C_1^{\mathrm{q}} \exp\left(-\min\left\{\frac{C_2^{\mathrm{q}} tn}{\tilde{\tau}^2}, \frac{C_3^{\mathrm{q}} nt^2}{\tilde{\tau}^4}\right\}\right) + 2p \exp\left(-\frac{nt}{2\tilde{\tau}^2}\right).$$

Proof. From Eq. (A.4) applied with a diagonal matrix for each pair of coordinates a, b (and using the union bound):

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n \sigma^{-2}\varepsilon_i^2 x_i x_i^\top - Q\right\|_\infty \geqslant tM^2\right) \leqslant 2p^2 C_1^{\mathrm{q}} \exp\left(-\min\left\{\frac{C_2^{\mathrm{q}} tn}{\tilde{\tau}^2}, \frac{C_3^{\mathrm{q}} nt^2}{\tilde{\tau}^4}\right\}\right).$$

If we use the inequality $\mathbb{P}(\|q\|_\infty \geqslant z) \leqslant 2p \exp(-nz^2/2M^2\tau^2)$, with $z = (t/2)^{1/2}\sigma M$, we get the desired result. □

E.2 Proof of Proposition 3.3

Following the analysis from Section 3, we need to upper bound $\mathbb{P}(j \in \hat{J}^*|\varepsilon)$ (probability of including a certain irrelevant variable into one of the replicated active sets), and $\mathbb{P}((\hat{J}^*)^c \cup \mathbf{J} \neq \varnothing)$ (probability of missing none of the relevant variables). We first prove two lemmas about each of them.

Lemma E.2. *Assume (**A**1-4), $\tilde{\mu} \leqslant \frac{\mathrm{m}(\tilde{\mathbf{w}})\tilde{\lambda}}{2p^{1/2}}$ and $\frac{n}{p} \geqslant \frac{256\tilde{\tau}^2}{\mathrm{m}(\mathbf{w})^2\tilde{\lambda}^2}$. We have:*

$$\mathbb{P}((\hat{J}^*)^c \cup \mathbf{J} \neq \varnothing) \leqslant 2p^2 \exp\left(-\frac{\tilde{\lambda}^2}{2}\frac{n}{p^2}\right) + 8pn^{1/2} \exp\left(-\frac{\mathrm{m}(\tilde{\mathbf{w}})\tilde{\lambda}}{8\tilde{\tau}}\frac{n^{1/2}}{p^{1/2}}\right).$$

Proof. This lemma shows that all relevant variables will be selected with overwhelming probability. From Lemma C.3, we have that $\mathbf{J} \subset \hat{J}^*$ as soon as $\|\hat{w} - \mathbf{w}\|_2 \leqslant \frac{p^{1/2}\mu + \|q^*\|_2}{\lambda_{\min}(Q^*)}$. Thus, if $\mathrm{m}(\mathbf{w}) > \frac{2\mu p^{1/2}}{\lambda_{\min}(Q)}$, $\lambda_{\min}(Q^*) \geqslant \lambda_{\min}(Q)/2$, $\|q - q^*\|_2 \leqslant \mathrm{m}(\mathbf{w})\lambda_{\min}(Q)/8$, and $\|q\|_2 \leqslant$

$\mathrm{m}(\mathbf{w})\lambda_{\min}(Q)/8$, then $\mathbf{J} \subset \hat{J}^*$. Thus, we have:

$$\begin{aligned}\mathbb{P}((\hat{J}^*)^c \cup \mathbf{J} \neq \varnothing) &\leqslant \mathbb{P}\left(\lambda_{\min}(Q^*) \leqslant \frac{\lambda_{\min}(Q)}{2}\right) + \mathbb{P}\left(\|q\|_2 \geqslant \frac{\mathrm{m}(\mathbf{w})\lambda_{\min}(Q)}{8}\right) \\ &\quad + \mathbb{P}\left(\|q - q^*\|_2 \geqslant \frac{\mathrm{m}(\mathbf{w})\lambda_{\min}(Q)}{8}\right), \\ &\leqslant 2p^2 \exp\left(-\frac{n\tilde{\lambda}^2}{2p^2}\right) + 2p \exp\left(-\frac{n\mathrm{m}(\tilde{\mathbf{w}})^2\tilde{\lambda}^2}{128p\tilde{\tau}^2}\right) \\ &\quad + 2p\mathbb{E}\exp\left(-\frac{n\mathrm{m}(\tilde{\mathbf{w}})^2\tilde{\lambda}^2\sigma^2}{32p\|\varepsilon\|_\infty^2}\right).\end{aligned}$$

We thus need to bound, for some $A > 0$,

$$\begin{aligned}\mathbb{E}\exp\left(-\tfrac{A}{\|\varepsilon\|_\infty^2}\right) &= \mathbb{E}\exp\left(-\tfrac{A}{\|\varepsilon\|_\infty^2}\right)1_{\|\varepsilon\|_\infty \leqslant M} + \mathbb{E}\exp\left(-\tfrac{A}{\|\varepsilon\|_\infty^2}\right)1_{\|\varepsilon\|_\infty > M}, \\ &\leqslant \exp\left(-\tfrac{A}{M^2}\right) + \mathbb{P}(\|\varepsilon\|_\infty > M), \\ &\leqslant \exp\left(-\tfrac{A}{M^2}\right) + 2n\exp(-\tfrac{M^2}{2\tau^2}) \leqslant 3n^{1/2}\exp(-\tfrac{A^{1/2}}{\tau\sqrt{2}}),\end{aligned}$$

for $A/M^2 = A^{1/2}/\tau 2^{1/2} + \log(n^{1/2})$, leading to

$$2p\mathbb{E}\exp\left(-\frac{n\mathrm{m}(\tilde{\mathbf{w}})^2\tilde{\lambda}^2\sigma^2}{32p\|\varepsilon\|_\infty^2}\right) \leqslant 6pn^{1/2}\exp\left(-\frac{n^{1/2}\mathrm{m}(\tilde{\mathbf{w}})\tilde{\lambda}}{8\tilde{\tau}p^{1/2}}\right).$$

The condition $n \geqslant \frac{256\tilde{\tau}^2 p}{\mathrm{m}(\mathbf{w})^2\tilde{\lambda}^2}$ allows to combine two terms into one, leading to the desired result. □

Lemma E.3. *Assume (A1-4) and $j \in \mathbf{J}^c$. Moreover, assume that $\|q\|_\infty \leqslant \beta_1 M\sigma/2$ and $\|Q - \widetilde{Q}\|_\infty \leqslant \frac{\beta_2 M^2}{2}$, with $\beta_1 \geqslant \tilde{\mu}$, $\beta_1\beta_2 \leqslant \frac{\tilde{\mu}\tilde{\lambda}^2}{40p^2}$, and $\beta_2 \leqslant \tilde{\lambda}$. We have:*

$$\begin{aligned}\mathbb{P}(j \notin \hat{J}^*|\varepsilon) \geqslant \frac{\frac{\tilde{\mu}n^{1/2}}{32}}{1 + \frac{\tilde{\mu}n^{1/2}}{4\tilde{\lambda}^{1/2}}} \exp\left[-\frac{1}{2}\left(\frac{8\tilde{\mu}n^{1/2}p^{1/2}}{\tilde{\lambda}} + \frac{|q_j - Q_{j,j^c}Q_{j^c,j^c}^{-1}q_{j^c}|}{\sigma n^{-1/2}Q_{jj|j^c}^{1/2}}\right)^2\right] \\ - \frac{16C_2^{\mathrm{BE}}}{\tilde{\tau}^3\tilde{\lambda}^1}\frac{p^{5/2}}{\tilde{\mu}n} - \frac{10C_2^{\mathrm{BE}}}{\tilde{\tau}^3\tilde{\lambda}^{1/2}}\frac{p^2}{n^{1/2}} - 2p\exp\left(-\frac{n\beta_1^2\sigma^2}{4\|\varepsilon\|_\infty^2}\right) - 2p^2\exp\left(-\frac{n\beta_2^2}{2}\right).\end{aligned}$$

Proof. We follow the same approach as in the proof of Proposition 2.7 in Appendix D.7. We first assume that $\|q - q^*\|_\infty \leqslant \beta_1 M\sigma/2$ and $\|Q^* - Q\|_\infty \leqslant \beta_2 M^2/2$ (on top of the assumptions made on $\widetilde{Q}$ and q). Following the same reasoning as in Appendix D.7, j is not included if

$$|-q_j^* + Q_{j,j^c}^*(Q_{j^c,j^c}^*)^{-1}q_{j^c}^* + \mu Q_{j,j^c}^*(Q_{j^c,j^c}^*)^{-1/2}\gamma_{\mu,Q^*}(q_{j^c}^*)| \leqslant \mu.$$

In order to apply Berry-Esseen inequality given ε, we first need to upper bound $|Q_{j,j^c}^*(Q_{j^c,j^c}^*)^{-1}q_{j^c}^* - \widetilde{Q}_{j,j^c}\widetilde{Q}_{j^c,j^c}^{-1}q_{j^c}^*|$, using Appendix B, by

$$\|Q^* - \widetilde{Q}\|_\infty\|q^*\|_2 \times \left(\frac{2p^{1/2}}{\lambda_{\min}(Q)} + \frac{4Mp}{\lambda_{\min}(Q)^{3/2}}\right) \leqslant \beta_1\beta_2\frac{6p^{3/2}M\sigma}{\tilde{\lambda}^{3/2}}.$$

Also, we need to bound, by Lemma C.4 and Appendix B:

$$\begin{aligned}
&|Q^*_{j,j^c}(Q^*_{j^c,j^c})^{-1/2}\gamma^j_{\mu,Q^*}(q^*_{j^c}) - Q_{j,j^c}(Q_{j^c,j^c})^{-1/2}\gamma^j_{\mu,Q}(q^*_{j^c})| \\
\leqslant\ & p^{1/2}\frac{\beta_2 M^2}{2}\lambda_{\min}(Q)^{-1}p^{1/2} + M^2 p^{1/2}\frac{4p\|Q^*-Q\|_\infty}{\mu\lambda_{\min}(Q)^2}(p^{1/2}\mu + \|q\|_2) \\
\leqslant\ & p\beta_2\tilde{\lambda}^{-1}\left(1+\frac{4p}{\tilde{\lambda}}+\frac{4}{\tilde{\lambda}}\frac{\beta_1 p}{\tilde{\mu}}\right) \leqslant p^2\beta_2\frac{5}{\tilde{\lambda}^2}\left(1+\frac{\beta_1}{\tilde{\mu}}\right) \leqslant \frac{10\beta_1\beta_2 p^2}{\tilde{\lambda}^2\tilde{\mu}}.
\end{aligned}$$

Since $\beta_1 \geqslant \tilde{\mu}$, $\beta_1\beta_2 \leqslant \frac{\tilde{\mu}\tilde{\lambda}^2}{40p^2}$, and $\beta_2 \leqslant \tilde{\lambda}$, we thus have:

$$|Q^*_{j,j^c}(Q^*_{j^c,j^c})^{-1}q^*_{j^c} - \widetilde{Q}_{j,j^c}\widetilde{Q}^{-1}_{j^c,j^c}q^*_{j^c}| \leqslant \mu/4,$$

$$|Q^*_{j,j^c}(Q^*_{j^c,j^c})^{-1/2}\gamma^j_{\mu,Q^*}(q^*_{j^c}) - Q_{j,j^c}Q^{-1/2}_{j^c,j^c}\gamma^j_{\mu,Q}(q^*_{j^c})| \leqslant \mu/4.$$

If we let denote A the event $\{|-q^*_j + \widetilde{Q}_{j,j^c}\widetilde{Q}^{-1}_{j^c,j^c}q^*_{j^c} + \mu Q_{j,j^c}Q^{-1/2}_{j^c,j^c}\gamma^j_{\mu,Q}(q^*_{j^c})| \leqslant \mu/2\}$ and B the event $\{\|q-q^*\|_\infty \leqslant \beta_1 M\sigma/2\} \cap \{\|Q^*-Q\|_\infty \leqslant \beta_2 M^2/2\}$, this implies that

$$\mathbb{P}(j \notin \hat{J}^*|\varepsilon) \geqslant \mathbb{P}(A|\varepsilon) - \mathbb{P}(B^c|\varepsilon). \tag{E.3}$$

We have, by concentration inequalities from Appendix E.1:

$$\mathbb{P}(B^c|\varepsilon) \leqslant 2p\exp\left(-\frac{n\beta_1^2\sigma^2}{4\|\varepsilon\|_\infty^2}\right) + 2p^2\exp\left(-\frac{n\beta_2^2}{2}\right). \tag{E.4}$$

Overall, if we assume the various bounds on q, q^*, Q^* and $\widetilde{Q}$, to have j excluded from the active set for the bootstrap sample, it is sufficient that A is satisfied. As in Appendix D.7, we consider a smooth version of the indicator function, and we get that the probability of A, given ε, is greater than

$$\frac{1}{2}\mathbb{P}\left(|u_j - \widetilde{Q}_{j,j^c}\widetilde{Q}^{-1}_{j^c,j^c}u_{j^c} - \mu Q_{j,j^c}Q^{-1/2}_{j^c,j^c}\gamma^j_{\mu,Q}(u_{j^c})| \leqslant \mu/4\right) - R, \tag{E.5}$$

where u is normal with mean q and covariance matrix $\sigma^2\widetilde{Q}/n$, and, from Proposition 2.7, $R \leqslant \frac{16C_2^{\mathrm{BE}}}{\tilde{\tau}^3\tilde{\lambda}^1}\frac{p^{5/2}}{\tilde{\mu}n} + \frac{10C_2^{\mathrm{BE}}}{\tilde{\tau}^3\tilde{\lambda}^{1/2}}\frac{p^2}{n^{1/2}}$ (note that we have used that $\widetilde{Q}$ is close to Q).

We have that given u_{j^c}, $-u_j + \widetilde{Q}_{j,j^c}\widetilde{Q}^{-1}_{j^c,j^c}u_{j^c}$ is normal with mean $-q_j + \widetilde{Q}_{j,j^c}\widetilde{Q}^{-1}_{j^c,j^c}q_{j^c}$ and covariance matrix $\sigma^2\widetilde{Q}_{j,j|j^c}/n$. Thus, we get, using Lemma D.1:

$$\begin{aligned}
&\frac{1}{2}\mathbb{P}\left(|-u_j + \widetilde{Q}_{j,j^c}\widetilde{Q}^{-1}_{j^c,j^c}u_{j^c} + \mu Q_{j,j^c}Q^{-1/2}_{j^c,j^c}\gamma^j_{\mu,Q}(u_{j^c})| \leqslant \mu/4\right) \qquad \text{(E.6)} \\
=\ & \frac{1}{2}\mathbb{E}\mathbb{P}\left(|-u_j + \widetilde{Q}_{j,j^c}\widetilde{Q}^{-1}_{j^c,j^c}u_{j^c} + \mu Q_{j,j^c}Q^{-1/2}_{j^c,j^c}\gamma^j_{\mu,Q}(u_{j^c})| \leqslant \mu/4 | u_{j^c}\right) \\
\geqslant\ & \frac{1}{2}\frac{\frac{\mu/4}{2\sigma n^{-1/2}\widetilde{Q}^{1/2}_{jj|j^c}}}{1+\frac{\mu/4}{2\sigma n^{-1/2}\widetilde{Q}^{1/2}_{jj|j^c}}} \times \\
& \exp\left[\frac{-1}{2}\left(\left|\frac{\mu Q_{j,j^c}(Q_{j^c,j^c})^{-1/2}\gamma^j_{\mu,Q}(u_{j^c})}{\sigma n^{-1/2}\widetilde{Q}^{1/2}_{j,j|j^c}}\right| + \left|\frac{q_j - \widetilde{Q}_{j,j^c}\widetilde{Q}^{-1}_{j^c,j^c}q_{j^c}}{\sigma n^{-1/2}\widetilde{Q}^{1/2}_{j,j|j^c}}\right|\right)^2\right].
\end{aligned}$$

We have using our assumptions regarding q and $\widetilde{Q}$: $\lambda_{\min}(Q)/2 \leqslant \widetilde{Q}_{jj|j^c} \leqslant 2M^2$ and $\left|Q_{j,j^c}(Q_{j^c,j^c})^{-1/2}\gamma^j_{\mu,Q}(u_j\right.$ $M\lambda_{\min}(Q)^{-1/2}p^{1/2}$. Moreover,

$$\begin{aligned}
|\widetilde{Q}_{j,j^c}\widetilde{Q}^{-1}_{j^c,j^c}q_{j^c} - Q_{j,j^c}Q^{-1}_{j^c,j^c}q_{j^c}| &\leqslant \frac{p^{1/2}\beta_1 M\sigma}{2}\left(\frac{p^{1/2}\beta_2 M^2}{\lambda_{\min}(Q)} + \frac{4M^3 p\beta_2}{\lambda_{\min}(Q)^{3/2}}\right) \\
&\leqslant \frac{3p^{3/2}M^3\beta_2\beta_1 M\sigma}{\lambda_{\min}(Q)^{3/2}} \leqslant \mu/8, \\
|Q^{-1/2}_{jj|j^c} - \widetilde{Q}^{-1/2}_{jj|j^c}| &\leqslant 4\lambda_{\min}(Q)^{-3/2}\|Q-\widetilde{Q}\|_2 \leqslant \frac{2\beta_2 M^2}{\lambda_{\min}(Q)^{3/2}}.
\end{aligned}$$

This leads to a lower bound of the form:

$$\frac{\frac{\tilde{\mu}n^{1/2}}{32}}{1+\frac{\tilde{\mu}n^{1/2}}{4\tilde{\lambda}^{1/2}}}\exp\left[-\frac{1}{2}\left(\frac{8\tilde{\mu}n^{1/2}p^{1/2}}{\tilde{\lambda}} + \frac{|q_j - Q_{j,j^c}Q^{-1}_{j^c,j^c}q_{j^c}|}{\sigma n^{-1/2}Q^{1/2}_{j,j|j^c}}\right)^2\right]. \tag{E.7}$$

By combining Eq. (E.3), Eq. (E.4), Eq. (E.5), Eq. (E.6) and Eq. (E.7), we get the desired result. □

We can now consider the full bound using the analysis outlined in Section 3, using Lemma E.1, E.2 and E.3:

$$\begin{aligned}
\mathbb{P}(\hat{J}^{\cap} \neq \mathbf{J}) &\leqslant m\mathbb{P}((\hat{J}^*)^c \cup \mathbf{J} \neq \varnothing) + \sum_{j\in\mathbf{J}^c}\mathbb{E}(\mathbb{P}(j\in\hat{J}^*|\varepsilon)^m), \\
&\leqslant 2p^2 m\exp\left(-\frac{n\tilde{\lambda}^2}{2p^2}\right) + 8pn^{1/2}m\exp\left(-\frac{n^{1/2}\mathrm{m}(\tilde{\mathbf{w}})\tilde{\lambda}}{8\tilde{\tau}p^{1/2}}\right) + 2p\exp\left(-\frac{n\beta_1^2}{2\tilde{\tau}^2}\right) \\
&\quad + 2p^2 C_1^{\mathrm{q}}\exp\left(-\min\left\{\frac{C_2^{\mathrm{q}}\beta_2 n}{2\tilde{\tau}^2}, \frac{C_3^{\mathrm{q}}n\beta_2^2}{2\tilde{\tau}^4}\right\}\right) + 2p\exp\left(-\frac{n\beta_2}{4\tilde{\tau}^2}\right) \\
&\quad + \sum_{j\in\mathbf{J}^c}\mathbb{E}\left[\mathbb{P}(j\in\hat{J}^*|\varepsilon)^m 1_{\|q\|_\infty\leqslant\beta_1 M\sigma/2}1_{\|\widetilde{Q}-Q\|_\infty\leqslant\beta_2 M^2/2}\right].
\end{aligned}$$

We consider $\beta_1 = \tilde{\mu}p^{-1}n^{3/10}$ and $\beta_2 = p^{-1}n^{-3/10}$. We truncate $\|\varepsilon\|_\infty$ at $n^{1/10}\sigma$ and $\frac{|q_{j|j^c}|}{\sigma n^{-1/2}Q^{1/2}_{jj|j^c}}$ at z and use Berry-Esseen inequality for $q_{j|j^c}$, to obtain:

$$\begin{aligned}
\mathbb{P}(\hat{J}^{\cap} \neq \mathbf{J}) \leqslant mp\exp\left(-A_0\frac{n^{1/2}}{p^{1/2}}\right) + A_1\frac{p^{3/2}}{n^{1/2}} + \exp(-z^2/2) + \\
+ p\left(1 - \frac{A_2(c)}{p^{1/2}}\exp\left[-\frac{1}{2}\left(\frac{8c}{\tilde{\lambda}} + z\right)^2\right] + A_3(c)\frac{p^3}{n^{1/2}}\right)^m,
\end{aligned}$$

with

$$2p\exp\left(-\frac{n\tilde{\lambda}^2}{2p^2}\right) + 8n^{1/2}\exp\left(-\frac{n^{1/2}\mathrm{m}(\tilde{\mathbf{w}})\tilde{\lambda}}{8\tilde{\tau}p^{1/2}}\right) \leqslant p\exp\left(-A_0\frac{n^{1/2}}{p^{1/2}}\right),$$

$$2p\exp\left(-\frac{n\tilde{\mu}^2 pn^{3/5}}{2p^2\tilde{\tau}^2}\right)+2p^2C_1^{\mathrm{q}}\exp\left(-\min\left\{\frac{C_2^{\mathrm{q}}p^{-1}n^{-3/10}n}{2\tilde{\tau}^2},\frac{C_3^{\mathrm{q}}np^{-2}n^{-3/5}}{2\tilde{\tau}^4}\right\}\right)$$
$$+2p\exp\left(-\frac{np^{-1}n^{-3/10}}{4\tilde{\tau}^2}\right)+\frac{\tilde{\tau}^3p^{3/2}}{\tilde{\lambda}^{3/2}}n^{-1/2}+2n\exp(-n^{1/5}/2\tilde{\tau}^2)\leqslant A_1\frac{p^{3/2}}{n^{1/2}},$$

$$A_3(c)\frac{p^3}{n^{1/2}}=\frac{16C_2^{\mathrm{BE}}}{\tilde{\tau}^3\tilde{\lambda}^1}\frac{p^3}{\tilde{\mu}np^{1/2}}+\frac{10C_2^{\mathrm{BE}}}{\tilde{\tau}^3\tilde{\lambda}^{1/2}}\frac{p^2}{n^{1/2}}+2p\exp\left(-\frac{n\tilde{\mu}^2n^{3/5}pp^{-3}}{4n^{1/5}}\right)+2p^2\exp\left(-\frac{np^{-2}n^{-3/5}}{2}\right).$$

$$A_2(c)p^{-1/2}\leqslant\frac{\frac{\tilde{\mu}n^{1/2}}{32}}{1+\frac{\tilde{\mu}n^{1/2}}{4\tilde{\lambda}^{1/2}}}.$$

All these constraints lead to the constraint tha np^{-6} should be larger than a function of c. We now need to optimize over z the following quantity:

$$p\left(1-\frac{A_2(c)}{p^{1/2}}\exp\left[-\frac{1}{2}\left(\frac{8c}{\tilde{\lambda}}+z\right)^2\right]+A_3(c)\frac{p^3}{n^{1/2}}\right)^m+e^{-z^2/2}.$$

If we select z such that $\frac{8c}{\tilde{\lambda}}+z=\left(2\log\frac{\frac{A_2(c)}{p^{1/2}}}{A_3(c)p^3n^{-1/2}+\frac{\log m}{m}}\right)^{1/2}$, which is possible if m and n large enough, i.e., if $m\geqslant e^{(\frac{8c}{\tilde{\lambda}})^2}(\frac{A_2(c)}{p^{1/2}})^{-2}$ and $n\geqslant e^{(\frac{8c}{\tilde{\lambda}})^2}(A_3(c)p^3)^2(\frac{A_2(c)}{p^{1/2}})^{-2}$, then we have the bound:

$$\left(1-\frac{A_2(c)}{p^{1/2}}\exp(-\frac{1}{2}(\frac{8c}{\tilde{\lambda}}+z)^2)+A_3(c)p^3n^{-1/2}\right)^m\leqslant\frac{1}{m},$$

and

$$e^{-z^2/2}\leqslant\frac{A_3(c)p^3n^{-1/2}+\frac{\log m}{m}}{\frac{A_2(c)}{p^{1/2}}}e^{\frac{(\frac{8c}{\tilde{\lambda}})^2}{2}}\exp\left(-\frac{\frac{8c}{\tilde{\lambda}}}{2}\left(2\log\frac{\frac{A_2(c)}{p^{1/2}}}{A_3(c)p^3n^{-1/2}+\frac{\log m}{m}}\right)^{1/2}\right).$$

This leads to the desired bound.

F Proofs for bootstrapping residuals

We use the following notation for the solution of the Lasso: $\hat{w}-\mathbf{w}=Q^{-1}q+\mu\hat{\alpha}$, where $\|Q\hat{\alpha}\|_\infty\leqslant 1$. We also denote $\Pi_X=X(X^\top X)^{-1}X^\top\in\mathbb{R}^{n\times n}$ the projection matrix on the data, which leads to the following expression for the non-centered estimated residuals:

$$\tilde{\varepsilon}=y-X\hat{w}=X(\mathbf{w}-\hat{w})+\varepsilon=(\mathrm{I}-\Pi_X)\varepsilon-\mu X\hat{\alpha}.$$

We let denote $\hat{\nu}=\frac{1}{n}\sum_{i=1}^n\tilde{\varepsilon}_i$. The boostrapped responses are thus $y_i^*=\tilde{\varepsilon}_{i^*}-\hat{\nu}+\hat{w}^\top x_i$. The bootstrapped residuals are thus $y_{i^*}+(\hat{w}-\mathbf{w})^\top x_i$, i.e.:

$$\varepsilon_i^*=[\Pi_X\varepsilon+\mu X\hat{\alpha}]_i+\tilde{\varepsilon}_{i^*}-\hat{\nu}.$$

We have the following expectations:

$$\begin{aligned}
\mathbb{E}(\tilde{\varepsilon}_{k^*}|\varepsilon) &= \frac{1}{n}1^\top\tilde{\varepsilon} = \frac{1}{n}1^\top(\mathrm{I}-\Pi_X)\varepsilon - \frac{\mu}{n}1^\top X\hat{\alpha} = \hat{\nu},\\
\mathbb{E}(\varepsilon^*|\varepsilon) &= \Pi_X\varepsilon + X\mu\hat{\alpha},\\
\mathrm{var}(\varepsilon_k^*|\varepsilon) &= \mathrm{var}(\tilde{\varepsilon}_{k^*}|\varepsilon) = \frac{1}{n}\tilde{\varepsilon}^\top\tilde{\varepsilon} - \hat{\nu}^2 = \frac{1}{n}\varepsilon^\top(\mathrm{I}-\Pi_X)\varepsilon + \mu^2\hat{\alpha}^\top Q\hat{\alpha} - \hat{\nu}^2,\\
\mathbb{E}(q^*|\varepsilon) &= \frac{1}{n}\sum_{k=1}^n \mathbb{E}(\tilde{\varepsilon}_{k^*}|\varepsilon)x_k = q + \mu Q\hat{\alpha},\\
\frac{\sigma^2}{n}\widetilde{Q} = \mathrm{var}(q^*|\varepsilon) &= \frac{1}{n^2}\sum_{k=1}^n \mathrm{var}(\hat{\varepsilon}_{k^*}|\varepsilon)x_kx_k^\top = \frac{Q}{n}\left[\frac{1}{n}\varepsilon^\top(\mathrm{I}-\Pi_X)\varepsilon + \mu^2\hat{\alpha}^\top Q\hat{\alpha} - \hat{\nu}^2\right].
\end{aligned}$$

We let denote $\gamma = \frac{\sigma^{-2}}{n}\varepsilon^\top(\mathrm{I}-\Pi_X)\varepsilon + \sigma^{-2}\mu^2\hat{\alpha}^\top Q\hat{\alpha} - \sigma^{-2}\hat{\nu}^2$ so that $\mathrm{var}(q^*|\varepsilon) = \sigma^2\gamma Q/n$.

F.1 Concentration inequalities

We need concentration inequalities for q^* around q (given ε) and of s around 1, as well as $\hat{\nu}$ around zero.

Lemma F.1. *Assume (A1-4) and $t \geqslant \frac{2\tilde{\mu}p}{\tilde{\lambda}}$. We have:*

$$\mathbb{P}\left(|\hat{\nu}| \geqslant t\sigma\right) \leqslant 2\exp\left(\frac{-nt^2\tilde{\lambda}}{32p^2\tilde{\tau}^2}\right).$$

Proof. We have: $\frac{1}{n}1^\top(\mathrm{I}-\Pi_X)\varepsilon = \frac{1}{n}\sum_{i=1}^n \varepsilon_i[(\mathrm{I}-\Pi_X)1]_i$ with $[(\mathrm{I}-\Pi_X)1]_i = 1 - x_i^\top Q^{-1}\left(\frac{1}{n}\sum_{k=1}^n x_k\right)$ is such that

$$|[(\mathrm{I}-\Pi_X)1]_i| \leqslant 1 + \lambda_{\min}(Q)^{-1}M^2p \leqslant 2\lambda_{\min}(Q)^{-1}M^2p = \frac{2p}{\tilde{\lambda}}.$$

Thus, we get:

$$\mathbb{P}\left(\left|\frac{1}{n}1^\top(\mathrm{I}-\Pi_X)\varepsilon\right| \geqslant t\sigma\right) \leqslant 2\exp\left(\frac{-nt^2\lambda_{\min}(Q)^2\sigma^2}{8\tau^2M^4p^2}\right) = 2\exp\left(\frac{-nt^2\tilde{\lambda}}{8p^2\tilde{\tau}^2}\right)$$

We also have $\left|\frac{\mu}{n}1^\top X\hat{\alpha}\right| = \left|\mu n^{-1}\sum_{k=1}^n x_i^\top\hat{\alpha}\right| \leqslant \mu pM\lambda_{\min}(Q)^{-1}$, hence the desired result with the extra condition on t. □

Lemma F.2. *Assume (A1-4). We have:*

$$\mathbb{P}(\|q^* - q - \mu Q\hat{\alpha}\|_\infty \geqslant tM\sigma|\varepsilon) \leqslant 2p\exp\left(\frac{-2nt^2}{\left(\frac{2p}{\tilde{\lambda}}\|\varepsilon\|_\infty/\sigma + \tilde{\mu}\frac{n^{1/2}p^{1/2}}{\tilde{\lambda}^{1/2}}\right)^2}\right).$$

Proof. We have $q^* = q + \mu Q\hat{\alpha} + \frac{1}{n}\sum_{i=1}^n x_i\hat{\varepsilon}_{i^*}$. Moreover, we have

$$\|\hat{\varepsilon}\|_\infty \leqslant \|\tilde{\varepsilon}\|_\infty \leqslant \left(1 + \frac{M^2 p}{\lambda_{\min}(Q)}\right)\|\varepsilon\|_\infty + \mu\|X\hat{\alpha}\|_2 \leqslant \frac{2p}{\tilde{\lambda}}\|\varepsilon\|_\infty + \frac{\tilde{\mu} n^{1/2} p^{1/2}\sigma}{\tilde{\lambda}^{1/2}}.$$

We get the result by Hoeffding's inequality. □

Lemma F.3. *Assume (A1-4), $\frac{p}{n} \leqslant \frac{t}{4}$ and $p\frac{\tilde{\mu}^2}{\tilde{\lambda}} \leqslant t/4$. We have:*

$$\mathbb{P}(|\gamma - 1| \geqslant t) \leqslant 2C_1^{\mathrm{q}} \exp\left(-\min\left\{\frac{C_2^{\mathrm{q}} t\tilde{\tau}^{-2}}{\frac{1}{n} + \frac{p^2}{n^2\tilde{\lambda}}}, \frac{C_3^{\mathrm{q}} n t^2 \tilde{\tau}^{-4}}{(1 - p/n)}\right\}\right) + 2\exp\left(\frac{-nt\tilde{\lambda}}{64p^2\tilde{\tau}^2}\right)$$

Proof. We first need to derive concentration for $\frac{1}{n}\varepsilon^\top(\mathrm{I} - \Pi_X)\varepsilon$ using Eq. (A.4). We have: $\lambda_{\max}(\frac{1}{n}|\mathrm{I} - \Pi_X|) \leqslant \frac{1}{n} + \frac{p}{n}\|\Pi_X\|_\infty \leqslant \frac{1}{n} + \frac{p^2}{n^2\tilde{\lambda}}$ and $\|\mathrm{I} - \Pi_X\|_F^2 = \frac{1}{n}$, because $|(\Pi_X)_{ij}| = \frac{1}{n}|x_i^\top Q^{-1} x_j| \leqslant p/n\tilde{\lambda}$. We thus obtain from Eq. (A.4):

$$\mathbb{P}\left(\left|\frac{\sigma^{-2}}{n}\varepsilon^\top(\mathrm{I} - \Pi_X)\varepsilon - (1 - p/n)\right| \geqslant t\right) \leqslant 2C_1^{\mathrm{q}} \exp\left(-\min\left\{\frac{C_2^{\mathrm{q}} t\tilde{\tau}^{-2}}{\frac{1}{n} + \frac{p^2}{n^2\tilde{\lambda}}}, \frac{C_3^{\mathrm{q}} n t^2 \tilde{\tau}^{-4}}{(1 - p/n)}\right\}\right).$$

Together with $\hat{\alpha}^\top Q\hat{\alpha} \leqslant p/\lambda_{\min}(Q)$, we get the desired result. □

F.2 Proof of Proposition 3.4

Following the analysis from Section 3, we need to upper bound $\mathbb{P}(j \in \hat{J}^*|\varepsilon)$ (probability of including a certain irrelevant variable into one of the replicated active sets), and $\mathbb{P}((\hat{J}^*)^c \cup \mathbf{J} \neq \varnothing)$ (probability of missing none of the relevant variables). We first prove two lemmas about each of them.

Lemma F.4. *Assume (A1-4) and $\tilde{\mu} \leqslant \frac{\mathrm{m}(\tilde{\mathbf{w}})\tilde{\lambda}}{p^{1/2}}$. We have:*

$$\mathbb{P}((\hat{J}^*)^c \cup \mathbf{J} \neq \varnothing) \leqslant 2p\exp\left(-\frac{n\mathrm{m}(\tilde{\mathbf{w}})^2\tilde{\lambda}^2}{32p\tilde{\tau}^2}\right) + 2n\exp\left(-\frac{n^{1/2}}{2\tilde{\tau}^2 p^{1/2}}\right) + 2p\exp\left(\frac{-n\tilde{\lambda}^2\mathrm{m}(\tilde{\mathbf{w}})^2}{8\left(\frac{2p^{3/4}}{\tilde{\lambda}}n^{1/4} + \tilde{\mu}\frac{n^{1/2}p^{1/2}}{\tilde{\lambda}^{1/2}}\right)^2}\right).$$

Proof. This lemma shows that all relevant variables will be selected with overwhelming probability. From Lemma C.3, we have that $\mathbf{J} \subset \hat{J}^*$ as soon as $\|\hat{w} - \mathbf{w}\|_2 \leqslant \frac{p^{1/2}\mu + \|q^*\|_2}{\lambda_{\min}(Q)}$. Thus, if $\mathrm{m}(\mathbf{w}) > \frac{\mu p^{1/2}}{\lambda_{\min}(Q)}$, $\|q - q^*\|_2 \leqslant \mathrm{m}(\mathbf{w})\lambda_{\min}(Q)/4$ and $\|q\|_2 \leqslant \mathrm{m}(\mathbf{w})\lambda_{\min}(Q)/4$, then $\mathbf{J} \subset \hat{J}^*$. Thus, we have (using results from Appendix F.1):

$$\begin{aligned}\mathbb{P}((\hat{J}^*)^c \cup \mathbf{J} \neq \varnothing) &\leqslant \mathbb{P}\left(\|q\|_2 \geqslant \frac{\mathrm{m}(\mathbf{w})\lambda_{\min}(Q)}{4}\right) + \mathbb{P}\left(\|q - q^*\|_2 \geqslant \frac{\mathrm{m}(\mathbf{w})\lambda_{\min}(Q)}{4}\right), \\ &\leqslant 2p\exp\left(-\frac{n\mathrm{m}(\tilde{\mathbf{w}})^2\tilde{\lambda}^2}{32p\tilde{\tau}^2}\right) + 2p\exp\left(\frac{-n\tilde{\lambda}^2\mathrm{m}(\tilde{\mathbf{w}})^2}{32\left(\frac{2p}{\tilde{\lambda}}\|\varepsilon\|_\infty/\sigma + \tilde{\mu}\frac{n^{1/2}p^{1/2}}{\tilde{\lambda}^{1/2}}\right)^2}\right).\end{aligned}$$

If we truncate $\|\varepsilon\|_\infty$ at $\sigma n^{1/4}p^{-1/4}$, then we have the bound

$$2p\exp\left(-\frac{n\mathrm{m}(\tilde{\mathbf{w}})^2\tilde{\lambda}^2}{32p\tilde{\tau}^2}\right)+\exp\left(-\frac{n^{1/2}}{2\tilde{\tau}^2p^{1/2}}\right)+2p\exp\left(\frac{-n\tilde{\lambda}^2\mathrm{m}(\tilde{\mathbf{w}})^2}{32\left(\frac{2p}{\tilde{\lambda}}n^{1/4}p^{-1/4}+\tilde{\mu}\frac{n^{1/2}p^{1/2}}{\tilde{\lambda}^{1/2}}\right)^2}\right),$$

hence the desired result. □

Lemma F.5. *Assume (**A1-4**) and $j\in\mathbf{J}^c$. We have:*

$$\mathbb{P}(j\notin\hat{J}^*|\varepsilon)\geqslant-\frac{16C_2^{\mathrm{BE}}}{\tilde{\tau}^3\tilde{\lambda}^1}\frac{p^{5/2}}{\tilde{\mu}n}-\frac{10C_2^{\mathrm{BE}}}{\tilde{\tau}^3\tilde{\lambda}^{1/2}}\frac{p^2}{n^{1/2}}+\frac{1}{2}\frac{\frac{\mu/4}{2\sigma n^{-1/2}\widetilde{Q}_{jj|j^c}^{1/2}}}{1+\frac{\mu/4}{2\sigma n^{-1/2}\widetilde{Q}_{jj|j^c}^{1/2}}}\times$$
$$\exp\left[\frac{-1}{2}\left(\left|\frac{\mu Q_{j,j^c}(Q_{j^c,j^c})^{-1/2}\gamma_{\mu,Q}^j(u_{j^c})}{\sigma n^{-1/2}\widetilde{Q}_{j,j|j^c}^{1/2}}\right|+\left|\frac{q_j-Q_{j,j^c}Q_{j^c,j^c}^{-1}q_{j^c}}{\sigma n^{-1/2}\widetilde{Q}_{j,j|j^c}^{1/2}}\right|\right)^2\right].$$

Proof. We follow the same approach as in the proof of Proposition 2.7 in Appendix D.7 and of Proposition 3.3 in Appendix E.2: j is not included if (note that $\widetilde{Q}=sQ$ and Q are proportional matrices)

$$|-q_j^*+\widetilde{Q}_{j,j^c}(\widetilde{Q}_{j^c,j^c})^{-1}q_{j^c}^*+\mu Q_{j,j^c}(Q_{j^c,j^c})^{-1/2}\gamma_{\mu,Q}(q_{j^c}^*)|\leqslant\mu.$$

As before, we consider a smooth version of the indicator function, and we get that the probability of not selecting j, given ε, is greater than

$$\frac{1}{2}\mathbb{P}\left(|u_j-\widetilde{Q}_{j,j^c}\widetilde{Q}_{j^c,j^c}^{-1}u_{j^c}-\mu Q_{j,j^c}Q_{j^c,j^c}^{-1/2}\gamma_{\mu,Q}^j(u_{j^c})|\leqslant\mu/4\right)-R,\tag{F.1}$$

where u is normal with mean q and covariance matrix $\sigma^2\widetilde{Q}/n$, and, from Proposition 2.7, $R\leqslant\frac{16C_2^{\mathrm{BE}}}{\tilde{\tau}^3\tilde{\lambda}^1}\frac{p^{5/2}}{\tilde{\mu}n}+\frac{10C_2^{\mathrm{BE}}}{\tilde{\tau}^3\tilde{\lambda}^{1/2}}\frac{p^2}{n^{1/2}}$.

We have that given u_{j^c}, $-u_j+\widetilde{Q}_{j,j^c}\widetilde{Q}_{j^c,j^c}^{-1}u_{j^c}$ is normal with mean $-q_j+\widetilde{Q}_{j,j^c}\widetilde{Q}_{j^c,j^c}^{-1}q_{j^c}$ and covariance matrix $\sigma^2\widetilde{Q}_{j,j|j^c}/n$. Thus, we get, using Lemma D.1:

$$\begin{aligned}
&\frac{1}{2}\mathbb{P}\left(|-u_j+\widetilde{Q}_{j,j^c}\widetilde{Q}_{j^c,j^c}^{-1}u_{j^c}+\mu Q_{j,j^c}Q_{j^c,j^c}^{-1/2}\gamma_{\mu,Q}^j(u_{j^c})|\leqslant\mu/4\right)\qquad(\text{F.2})\\
=\;&\frac{1}{2}\mathbb{E}\mathbb{P}\left(|-u_j+\widetilde{Q}_{j,j^c}\widetilde{Q}_{j^c,j^c}^{-1}u_{j^c}+\mu Q_{j,j^c}Q_{j^c,j^c}^{-1/2}\gamma_{\mu,Q}^j(u_{j^c})|\leqslant\mu/4|u_{j^c}\right)\\
\geqslant\;&\frac{1}{2}\frac{\frac{\mu/4}{2\sigma n^{-1/2}\widetilde{Q}_{jj|j^c}^{1/2}}}{1+\frac{\mu/4}{2\sigma n^{-1/2}\widetilde{Q}_{jj|j^c}^{1/2}}}\times\\
&\exp\left[\frac{-1}{2}\left(\left|\frac{\mu Q_{j,j^c}(Q_{j^c,j^c})^{-1/2}\gamma_{\mu,Q}^j(u_{j^c})}{\sigma n^{-1/2}\widetilde{Q}_{j,j|j^c}^{1/2}}\right|+\left|\frac{q_j-Q_{j,j^c}Q_{j^c,j^c}^{-1}q_{j^c}}{\sigma n^{-1/2}\widetilde{Q}_{j,j|j^c}^{1/2}}\right|\right)^2\right].
\end{aligned}$$

By combining Eq. (F.1) and Eq. (F.2), we get the desired result. Note that if $|s-1| \leqslant 1/2$, we have

$$\left| \frac{\mu Q_{j,j^c}(Q_{j^c,j^c})^{-1/2}\gamma^j_{\mu,Q}(u_{j^c})}{\sigma n^{-1/2}\widetilde{Q}^{1/2}_{j,j|j^c}} \right| \leqslant 2\mu M \lambda_{\min}(Q)^{-1} p^{1/2}\sigma^{-1}n^{1/2} = 2\tilde{\mu}n^{1/2}p^{1/2}.$$

and

$$\frac{\frac{\mu/4}{2\sigma n^{-1/2}\widetilde{Q}^{1/2}_{jj|j^c}}}{1+\frac{\mu/4}{2\sigma n^{-1/2}\widetilde{Q}^{1/2}_{jj|j^c}}} \geqslant \frac{\frac{\mu/4}{4\sigma n^{-1/2}M}}{1+\frac{\mu/4}{\sigma n^{-1/2}\tilde{\lambda}^{1/2}}} = \frac{\frac{\tilde{\mu}n^{1/2}}{16}}{1+\frac{\tilde{\mu}n^{12}}{4\tilde{\lambda}^{1/2}}}.$$

□

We can now consider the full bound using the analysis outlined in Section 3, using Lemma F.4, F.5, and Appendix F.1, together with truncating on the events $\|\varepsilon\|_\infty \leqslant \sigma n^{1/4}$ and $|\gamma - 1| \leqslant n^{-1/3}$ (note that we can apply from Lemma F.3 for n large enough). First, we need a bound on

$$\begin{aligned}\mathbb{P}\left(\left|\frac{q_j - Q_{j,j^c}Q^{-1}_{j^c,j^c}q_{j^c}}{\sigma n^{-1/2}\widetilde{Q}^{1/2}_{j,j|j^c}}\right| \geqslant z\right) &\leqslant \mathbb{P}\left(\left|\frac{q_j - Q_{j,j^c}Q^{-1}_{j^c,j^c}q_{j^c}}{\sigma n^{-1/2}Q^{1/2}_{j,j|j^c}}\right| \geqslant z(1-t)^{1/2}\right) + P(|\gamma-1|\leqslant t),\\ &\leqslant e^{-z^2(1-t)/2} + \frac{\tilde{\tau}^3 p^{3/2}}{\tilde{\lambda}^{3/2}}n^{-1/2} + P(|\gamma-1|\leqslant t).\end{aligned}$$

We have following the reasoning from Section 3:

$$\begin{aligned}\mathbb{P}(\hat{J}^\cap \neq \mathbf{J}) &\leqslant \sum_{j\in\mathbf{J}^c}\mathbb{E}\mathbb{P}(j\in\hat{J}^*|\varepsilon)^m) + m\mathbb{P}((\hat{J}^*)^c\cup\mathbf{J}\neq\varnothing),\\ &\leqslant p\left(1-\frac{A_2(c)}{p^{1/2}}\exp\left[-\frac{1}{2}(B_0+z)^2\right] + A_3(c)\frac{p^3}{n^{1/2}}\right)^m + A_1\frac{p^{3/2}}{n^{1/2}} + e^{-z^2(1-t)/2},\end{aligned}$$

with

$$\begin{aligned}&2p\exp\left(-\frac{n\mathrm{m}(\tilde{\mathbf{w}})^2\tilde{\lambda}^2}{32p\tilde{\tau}^2}\right) + 2n\exp\left(-\frac{n^{1/2}}{2\tilde{\tau}^2p^{1/2}}\right)\\ &\qquad + 2p\exp\left(\frac{-n\tilde{\lambda}^2\mathrm{m}(\tilde{\mathbf{w}})^2}{32\left(\frac{2p^{3/4}}{\tilde{\lambda}}n^{1/4} + \tilde{\mu}\frac{n^{1/2}p^{1/2}}{\tilde{\lambda}^{1/2}}\right)^2}\right) \leqslant p\exp\left(-A_0\frac{n^{1/2}}{p^{1/2}}\right)\end{aligned}$$

$$\begin{aligned}&2C_1^{\mathrm{q}}\exp\left(-\min\left\{\frac{C_2^{\mathrm{q}}t\tilde{\tau}^{-2}}{\frac{1}{n}+\frac{p^2}{n^2\tilde{\lambda}}}, \frac{C_3^{\mathrm{q}}nt^2\tilde{\tau}^{-4}}{(1-p/n)}\right\}\right) + 2\exp\left(\frac{-nt\tilde{\lambda}}{64p^2\tilde{\tau}^2}\right)\\ &\qquad + \frac{\tilde{\tau}^3p^{3/2}}{\tilde{\lambda}^{3/2}}n^{-1/2} + 2n\exp(-n^{1/2}/2\tilde{\tau}^2) \leqslant A_1\frac{p^{3/2}}{n^{1/2}},\end{aligned}$$

$$A_3(c)\frac{p^3}{n^{1/2}} = \frac{16C_2^{\mathrm{BE}}}{\tilde{\tau}^3\tilde{\lambda}^1}\frac{p^3}{\tilde{\mu}np^{1/2}} + \frac{10C_2^{\mathrm{BE}}}{\tilde{\tau}^3\tilde{\lambda}^{1/2}}\frac{p^2}{n^{1/2}}$$

$$A_2(c)p^{-1/2} \leqslant \frac{\frac{\tilde{\mu}n^{1/2}}{16}}{1+\frac{\tilde{\mu}n^{12}}{4\tilde{\lambda}^{1/2}}} \text{ and } B_0 \leqslant 2\tilde{\mu}n^{1/2}p^{1/2}.$$

All these constraints lead to the constraint tha np^{-6} should be larger than a function of c. The rest of proof follows along the lines of the proof of Proposition 3.3 (note that the term $e^{-z^2(1-t)/2}$ instead of $e^{-z^2/2}$ only affects the constant A_5).

G Proofs of high-dimensional results

G.1 Proof of Proposition 4.1

From Lemma C.2, we obtain optimality conditions for the solution of Eq. (2.1) to have the sign pattern $\mathbf{t}$:

$$\begin{aligned}
&\|Q_{\mathbf{L}^c,\mathbf{L}}Q_{\mathbf{L},\mathbf{L}}^{-1}q_{\mathbf{L}} - q_{\mathbf{L}^c} - \mu Q_{\mathbf{L}^c,\mathbf{L}}Q_{\mathbf{L},\mathbf{L}}^{-1}\mathbf{t}_{\mathbf{L}}\|_\infty \leqslant \mu,\\
&\mathrm{sign}(\mathbf{w}_{\mathbf{J}} + (Q_{\mathbf{L},\mathbf{L}}^{-1}q_{\mathbf{L}} - \mu Q_{\mathbf{L},\mathbf{L}}^{-1}\mathbf{t}_{\mathbf{L}})_{\mathbf{J}}) = \mathbf{t}_{\mathbf{J}},\\
&\mathrm{sign}[(Q_{\mathbf{L},\mathbf{L}}^{-1}q_{\mathbf{L}} - \mu Q_{\mathbf{L},\mathbf{L}}^{-1}\mathbf{t}_{\mathbf{L}})_{\mathbf{K}}] = \mathbf{t}_{\mathbf{K}}.
\end{aligned}$$

It is thus sufficient for $\mathbf{t}$ to be the sign pattern that

$$\forall k \in \mathbf{L}^c, \quad |Q_{k,\mathbf{L}}Q_{\mathbf{L},\mathbf{L}}^{-1}q_{\mathbf{L}} - q_k| \leqslant \mu\boldsymbol{\theta}, \tag{G.1}$$

$$\forall k \in \mathbf{J}, \quad |(Q_{\mathbf{LL}}^{-1}q_{\mathbf{L}})_k| \leqslant \frac{1}{2}\mu\mathrm{m}(\mathbf{w}), \tag{G.2}$$

$$\forall k \in \mathbf{K}, \quad |(Q_{\mathbf{LL}}^{-1}q_{\mathbf{L}})_k| \leqslant \mu\boldsymbol{\theta}Q_{kk}^{-1}. \tag{G.3}$$

Eq. (G.1) occurs with probability greater than $1 - 2|\mathbf{L}^c|\exp\left(-\frac{n\tilde{\mu}^2\boldsymbol{\theta}^2\tilde{\lambda}_{\mathbf{L}}}{8\tilde{\tau}^2|\mathbf{L}|}\right)$. Eq. (G.2) occurs with probability greater than $1-2|\mathbf{J}|\exp\left(-\frac{n\mathrm{m}(\tilde{\mathbf{w}})^2\tilde{\lambda}_{\mathbf{L}}^2}{4\tilde{\tau}^2|\mathbf{L}|}\right)$, and Eq. (G.3) occurs with probability greater than $1 - 2|\mathbf{K}|\exp\left(-\frac{n\tilde{\mu}^2\boldsymbol{\theta}^2\tilde{\lambda}_{\mathbf{L}}^2}{4\tilde{\tau}^2|\mathbf{L}|}\right)$. This leads to the desired result by the union bound.

G.2 Proof of Proposition 4.2

The bound is obtained simply from Proposition 4.1, Proposition Proposition 3.3 and Proposition Proposition 3.4, using the union bound.

Acknowledgements

I would like to thank Zaïd Harchaoui, Jean-Yves Audibert and Sylvain Arlot for fruitful discussions related to this work. This work was supported by a grant from the Agence Nationale de la Recherche, France (MGA project, BLAN07-3-198092).

References

[1] F. Bach. Bolasso: model consistent Lasso estimation through the bootstrap. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2008.

[2] F. Bach. Consistency of the group Lasso and multiple kernel learning. *Journal of Machine Learning Research*, 8:1179–1225, 2008.

[3] F. Bach. Exploring large feature spaces with hierarchical multiple kernel learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.

[4] R. Baraniuk. Compressive sensing. *IEEE Signal Processing Magazine*, 24(4):118–121, 2007.

[5] V. Bentkus. On the dependence of the Berry–Esseen bound on dimension. *Journal of Statistical Planning and Inference*, 113:385–402, 2003.

[6] P. J. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 2008. To appear.

[7] J. F. Bonnans, J. C. Gilbert, C. Lemaréchal, and C. A. Sagastizbal. *Numerical Optimization Theoretical and Practical Aspects*. Springer, 2003.

[8] S. Boucheron, G. Lugosi, and O. Bousquet. Concentration inequalities. In *Advanced Lectures on Machine Learning*, volume 3176 of *Lecture Notes in Artificial Intelligence*. Springer, 2004.

[9] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge Univ. Press, 2003.

[10] L. Breiman. Arcing classifier. *Annals of Statistics*, 26(3):801–849, 1998.

[11] P. Bühlmann. Boosting for high-dimensional linear models. *Annals of Statistics*, 34(2):559–583, 2006.

[12] F. Bunea, A. Tsybakov, and M. Wegkamp. Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics*, 1:169–194, 2007.

[13] E. Candès and M. Wakin. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 25(2):21–30, 2008.

[14] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43(1):129–159, 2001.

[15] A. Cohen, W. Dahmen, and R. DeVore. Compressed sensing and best k-term approximation. Technical report, IGPM Report, RWTH-Aachen, 2006.

[16] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32:407, 2004.

[17] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, 1998.

[18] D. Freedman. Bootstrapping regression models. *Annals of Statistics*, 9(6):1218–1228, 1981.

[19] J. Friedman, T. H. T, and R. Tibshirani. Pathwise coordinate optimization. *Annals of Applied Statistics*, 1(2):302–332, 2007.

[20] W. Fu. Penalized regressions: the bridge vs. the Lasso. *Journal of Computational and Graphical Statistics*, 7(3):397–416, 1998).

[21] W. Fu and K. Knight. Asymptotics for Lasso-type estimators. *Annals of Statistics*, 28(5):1356–1378, 2000.

[22] J.-J. Fuchs. On sparse representations in arbitrary redundant bases. *IEEE Transactions on Information Theory*, 50(6):1341–1344, 2004.

[23] P. Garrigues and L. E. Ghaoui. An homotopy algorithm for the Lasso with online observations. In *Advances in Neural Information Processing Systems (NIPS) 21*, 2009.

[24] F. Götze. On the rate of convergence in the multivariate central limit theorem. *Annals of Probability*, 19(2):724–739, 1991.

[25] R. Horn and C. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

[26] J. Huang, S. Ma, and C.-H. Zhang. Adaptive Lasso for sparse high-dimensional regression models. *Statistica Sinica*, 18:1603–1618, 2008.

[27] K. Lounici. Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electronic Journal of Statistics*, 2, 2008.

[28] S. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41:3397–3415, 1993.

[29] H. M. Markowitz. The optimization of a quadratic function subject to linear constraints. *Naval Research Logistics Quarterly*, 3:111–133, 1956.

[30] P. Massart. *Concentration Inequalities and Model Selection: Ecole d'été de Probabilités de Saint-Flour 23*. Springer, 2003.

[31] N. Meinshausen. Relaxed Lasso. *Computational Statistics and Data Analysis*, 52(1):374–393, September 2007.

[32] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34(3):1436–1462, 2006.

[33] N. Meinshausen and P. Bühlmann. Stability selection. Technical Report 0809.2932, ArXiv, 2008.

[34] N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*, 37(1):246–270, 2008.

[35] M. R. Osborne, B. Presnell, and B. A. Turlach. On the lasso and its dual. *Journal of Computational and Graphical Statistics*, 9(2):319–337, 2000.

[36] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of The Royal Statistical Society Series B*, 58(1):267–288, 1996.

[37] M. J. Wainwright. Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1-constrained quadratic programming. Technical Report 709, Department of Statistics, UC Berkeley, 2006.

[38] F. T. Wright. A bound on tail probabilities for quadratic forms in independent random variables whose distributions are not necessarily symmetric. *Annals of Probability*, 1(6):1068–1070, 1973.

[39] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of The Royal Statistical Society Series B*, 68(1):49–67, 2006.

[40] M. Yuan and Y. Lin. On the non-negative garrotte estimator. *Journal of The Royal Statistical Society Series B*, 69(2):143–161, 2007.

[41] C.-H. Zhang and J. Huang. The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Annals of Statistics*, 36(4):1567–1594, 2008.

[42] T. Zhang. Some sharp performance bounds for least squares regression with ℓ^1-regularization. *Annals of Statistics*, 2009. to appear.

[43] P. Zhao, G. Rocha, and B. Yu. Grouped and hierarchical model selection through composite absolute penalties. *Annals of Statistics*, To appear, 2008.

[44] P. Zhao and B. Yu. On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.

[45] H. Zou. The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, December 2006.