DE-BIASING THE LASSO: OPTIMAL SAMPLE SIZE FOR GAUSSIAN DESIGNS

By

Adel Javanmard
Andrea Montanari

Technical Report No. 2015-18
September 2015

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305-4065

DE-BIASING THE LASSO: OPTIMAL SAMPLE SIZE FOR GAUSSIAN DESIGNS

By

Adel Javanmard
University of Southern California

Andrea Montanari
Stanford University

Technical Report No. 2015-18
September 2015

This research was supported in part by National Science Foundation grants CCF 1319979 and DMS 1106627, and Air Force Office of Scientific Research grant FA9550-13-1-0036.

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305-4065

http://statistics.stanford.edu

De-biasing the Lasso: Optimal Sample Size for Gaussian Designs

Adel Javanmard* and Andrea Montanari†

August 31, 2015

Abstract

Performing statistical inference in high-dimensional models is an outstanding challenge. A major source of difficulty is the absence of precise information on the distribution of high-dimensional regularized estimators.

Here, we consider linear regression in the high-dimensional regime $p \gg n$ and the Lasso estimator. In this context, we would like to perform inference on a high-dimensional parameters vector $\theta^* \in \mathbb{R}^p$. Important progress has been achieved in computing confidence intervals and p-values for single coordinates θ_i^*, $i \in \{1, \dots, p\}$. A key role in these new inferential methods is played by a certain de-biased (or de-sparsified) estimator $\widehat{\theta}^{\rm d}$ that is constructed from the Lasso estimator. Earlier work establishes that, under suitable assumptions on the design matrix, the coordinates of $\widehat{\theta}^{\rm d}$ are asymptotically Gaussian provided the true parameters vector θ^* is s_0-sparse with $s_0 = o(\sqrt{n}/\log p)$.

The condition $s_0 = o(\sqrt{n}/\log p)$ is considerably stronger than the one required for consistent estimation, namely $s_0 = o(n/\log p)$. Here we consider Gaussian designs with known or unknown population covariance. When the covariance is known, we prove that the de-biased estimator is asymptotically Gaussian under the nearly optimal condition $s_0 = o(n/(\log p)^2)$. Note that *earlier work was limited to* $s_0 = o(\sqrt{n}/\log p)$ *even for perfectly known covariance.*

The same conclusion holds if the population covariance is unknown but can be estimated sufficiently well, e.g. because its inverse is very sparse. For intermediate regimes, we describe the trade-off between sparsity in the coefficients θ^*, and sparsity in the inverse covariance of the design.

*Data Sciences and Operations Department, Marshall School of Business, University of Southern California, Email: ajavanma@marshall.usc.edu

†Department of Electrical Engineering and Department of Statistics, Stanford University. Email: montanar@stanford.edu

1 Introduction

1.1 Background

Consider random design model where we are given n i.i.d. pairs (y_1, x_1), (y_2, x_2), $\cdots$, (y_n, x_n) with $y_i \in \mathbb{R}$, and $x_i \in \mathbb{R}^p$. The response variable y_i is a linear function of x_i, contaminated by noise w_i independent of x_i

$$y_i = \langle \theta^*, x_i \rangle + w_i \,, \qquad w_i \sim \mathsf{N}(0, \sigma^2) \,. \tag{1}$$

Here $\theta^* \in \mathbb{R}^p$ is a vector of parameters to be estimated and $\langle \cdot \,, \cdot \rangle$ is the standard scalar product.

In matrix form, letting $y = (y_1, \ldots, y_n)^{\mathsf{T}}$ and denoting by X the matrix with rows $x_1^{\mathsf{T}}, \cdots, x_n^{\mathsf{T}}$ we have

$$y = X\,\theta^* + w \,, \qquad w \sim \mathsf{N}(0, \sigma^2 \mathrm{I}_{n\times n}) \,. \tag{2}$$

We are interested in the high-dimensional regime wherein the number of parameters p exceeds the sample size n. Over the last 20 years, impressive progress has been made in developing and understanding highly effective estimators in this regime [CT07, BRT09, BvdG11]. A prominent approach is the Lasso [Tib96, CD95] defined through the following convex optimization problem

$$\widehat{\theta}^{\text{Lasso}}(y, X; \lambda) \equiv \arg\max_{\theta \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|y - X\theta\|_2^2 + \lambda \|\theta\|_1 \right\} . \tag{3}$$

(We will omit the arguments of $\widehat{\theta}^{\text{Lasso}}(y, X; \lambda)$ whenever clear from the context.)

A far less understood question is how to perform statistical inference in the high-dimensional setting, for instance computing confidence intervals and p-values for quantities of interest. Progress in this direction was achieved only over the last couple of years. In particular, several papers [Büh13, ZZ14, JM14b, VdGBRD14, JM14a] develop methods to compute confidence intervals for single coordinates of the parameters vector θ^*. More precisely, these methods compute intervals $J_i(\alpha)$ depending on y, X, of nearly minimal size, with the coverage guarantee

$$\mathbb{P}\big(\theta_i^* \in J_i(\alpha)\big) \geq 1 - \alpha - o_n(1) \,. \tag{4}$$

The $o_n(1)$ term is explicitly characterized, and vanishes along sequence of instances of increasing dimensions under suitable condition on the design matrix X.

The fundamental idea developed in [ZZ14, JM14b, VdGBRD14, JM14a] is to construct a debiased (or de-sparsified) estimator that takes the form

$$\widehat{\theta}^{\text{d}} = \widehat{\theta}^{\text{Lasso}} + \frac{1}{n} M X^{\mathsf{T}} (y - X\widehat{\theta}^{\text{Lasso}}) \,, \tag{5}$$

where $M \in \mathbb{R}^{p\times p}$ is a matrix that is a function of X, but not of y. While the construction of M varies across different papers, the basic intuition is that M should be a good estimate of the precision matrix $\Omega = \Sigma^{-1}$, where $\Sigma = \mathbb{E}\{x_1 x_1^{\mathsf{T}}\}$ is the population covariance.

Assume θ^* is s_0-sparse, i.e. it has only s_0 non-zero entries. The key result that allows the construction of confidence intervals in [ZZ14, VdGBRD14, JM14a] is the following (holding under suitable conditions on the design matrix). If M is 'sufficiently close' to Ω, and the sparsity level is

$$s_0 \ll \frac{\sqrt{n}}{\log p} \,, \tag{6}$$

then $\widehat{\theta}^{\mathrm{d}}_i$ is approximately Gaussian with mean θ^*_i and variance of order σ^2/n.

The condition (6) comes as a surprise, and is somewhat disappointing. Indeed, consistent estimation using –for instance– the Lasso can be achieved under the much weaker condition $s_0 \ll n/\log p$. More specifically, in this regime, with high probability [CT07, BRT09, BvdG11]

$$\|\widehat{\theta}^{\mathrm{Lasso}} - \theta^*\|_2^2 \le \frac{Cs_0\sigma^2}{n}\,\log p\,. \tag{7}$$

This naturally leads to the following question:

> *Does the de-biased estimator have a Gaussian limit under the weaker condition $s_0 \ll n/\log p$?*

Let us emphasize that the key technical challenge here does not lie in the fact that M is not a good estimate of the precision matrix Ω. Of course, if M is not close to Ω, then $\widehat{\theta}^{\mathrm{d}}$ will not have a Gaussian limit. However *earlier proofs [ZZ14, VdGBRD14, JM14a] cannot establish the Gaussian limit for $s_0 \gtrsim \sqrt{n}/\log p$, even if Ω is known and we set $M = \Omega$.* Even the idealized case where the columns of X are known to be independent and identically distributed (i.e. $\Omega = \mathrm{I}$) is only understood in the asymptotic limit $s_0, n, p \to \infty$ with s_0/p, n/p having constant limits in $(0, 1)$ [JM14b].

In order to describe the challenge, let us set $M = \Omega$, and recall the common step of the proofs in [ZZ14, VdGBRD14, JM14a]. Using the definitions (2), (5), we get

$$\begin{aligned}\sqrt{n}(\widehat{\theta}^{\mathrm{d}} - \theta^*) &= \sqrt{n}(\widehat{\theta}^{\mathrm{Lasso}} - \theta^*) + \frac{1}{\sqrt{n}}\Omega X^{\mathsf{T}}(X\theta^* + w - X\widehat{\theta}^{\mathrm{Lasso}}) \\ &= \frac{1}{\sqrt{n}}\Omega X^{\mathsf{T}} w + \sqrt{n}(\Omega\widehat{\Sigma} - \mathrm{I})(\theta^* - \widehat{\theta}^{\mathrm{Lasso}})\,,\end{aligned} \tag{8}$$

where $\widehat{\Sigma} = X^{\mathsf{T}}X/n \in \mathbb{R}^{p\times p}$ is the empirical design covariance. Since $w \sim \mathsf{N}(0, \sigma^2\mathrm{I}_n)$, it is easy to see that vector $\Omega X^{\mathsf{T}} w/\sqrt{n}$ has Gaussian entries of variance of order one. In order for $\widehat{\theta}^{\mathrm{d}}$ to be approximately Gaussian, we need the second term (which can be interpreted as a bias) to vanish. Earlier papers [ZZ14, VdGBRD14, JM14a] address this by a simple ℓ_1-ℓ_∞ bound. Namely (denoting by $|Q|_\infty$ the maximum absolute value of any entry of matrix Q):

$$\begin{aligned}\left\|\sqrt{n}(\Omega\widehat{\Sigma} - \mathrm{I})(\theta^* - \widehat{\theta}^{\mathrm{Lasso}})\right\|_\infty &\le \sqrt{n}|\Omega\widehat{\Sigma} - \mathrm{I}|_\infty \|\theta^* - \widehat{\theta}^{\mathrm{Lasso}}\|_1 \\ &\le \sqrt{n} \times C\sqrt{\frac{\log p}{n}} \times Cs_0\sigma\sqrt{\frac{\log p}{n}} \\ &\le C^2\sigma\frac{s_0 \log p}{\sqrt{n}}\,,\end{aligned} \tag{9}$$

where the bound $|\Omega\widehat{\Sigma} - \mathrm{I}|_\infty \le C\sqrt{(\log p)/n}$ follows from standard concentration arguments, and the bound on $\|\theta^* - \widehat{\theta}^{\mathrm{Lasso}}\|_1$ is order-optimal and is proved, for instance, in [BRT09, BvdG11].

This simple argument implies that the de-biased estimator is approximately Gaussian if the upper bound in Eq. (9) is negligible, i.e. if $s_0 = o(\sqrt{n}/\log p)$. We see therefore that this requirement is not imposed as to control the error in estimating Ω. It instead follows from the simple ℓ_1-ℓ_∞ bound *even if Ω is known.*

1.2 Main results

The above exposition should clarify that the $\ell_1 - \ell_\infty$ bound is quite conservative. Considering the i-th entry in the bias vector $\mathsf{bias} = (\Omega\widehat{\Sigma} - \mathrm{I})(\theta^* - \widehat{\theta}^{\mathrm{Lasso}})$, the ℓ_1-ℓ_∞ bound controls it as $|\mathsf{bias}_i| \le \|(\Omega\widehat{\Sigma} - \mathrm{I})_{i,\cdot}\|_\infty \|\theta^* - \widehat{\theta}^{\mathrm{Lasso}}\|_1$. This bound would be accurate only if the signs of the entries $(\theta^*_j - \widehat{\theta}^{\mathrm{Lasso}}_j)$ were aligned to the signs $(\Omega\widehat{\Sigma} - \mathrm{I})_{i,j}$, $j \in \{1, \dots, p\}$. While intuitively this is quite unlikely, it is difficult to formalize this intuition; Note that in a random design setting, the terms $(\Omega\widehat{\Sigma} - \mathrm{I})_{i,\cdot}$ and $\theta^* - \widehat{\theta}^{\mathrm{Lasso}}$ are highly dependent: $\widehat{\theta}^{\mathrm{Lasso}}$ is a deterministic function of the random pair (X, w), while $(\Omega\widehat{\Sigma} - \mathrm{I}) = (\Omega X X^{\mathsf{T}}/n - \mathrm{I})$ is a function of X.

Our main result overcomes this technical hurdle via a careful analysis of such dependencies. We follow a leave-one-out proof technique. Roughly speaking, in order to understand the distribution of the i-th coordinate of the de-biased estimator $\widehat{\theta}^{\mathrm{d}}_i$, we consider a modified problem in which column i is removed from the design matrix X. We then study the consequences of adding back this column, and bound the effect of this perturbation. An outline of this proof strategy is provided in Section 4.1.

We state below a simplified version of our main result, referring to Theorem 3.6 below for a full statement, including technical conditions.

Theorem 1.1 (Known covariance)**.** *Consider the linear model (2) where X has independent Gaussian rows, with zero mean and covariance $\Sigma = \Omega^{-1}$. Assume that Σ satisfies the technical conditions stated in Theorem 3.6. Define the de-biased estimator $\widehat{\theta}^{\mathrm{d}}$ via Eq. (5) with $M = \Omega$ and $\widehat{\theta}^{\mathrm{Lasso}} = \widehat{\theta}^{\mathrm{Lasso}}(y, X; \lambda)$ with $\lambda = 8\sigma\sqrt{(\log p)/n}$.*

If $n, p \to \infty$ with $s_0 = o(n/(\log p)^2)$, then we have

$$\sqrt{n}(\widehat{\theta}^{\mathrm{d}} - \theta^*) = Z + o_P(1)\,, \qquad Z|X \sim \mathsf{N}(0, \sigma^2\Omega\widehat{\Sigma}\Omega)\,. \tag{10}$$

Here $o_P(1)$ is a (random) vector satisfying $\|o_P(1)\|_\infty \to 0$ in probability as $n, p \to \infty$, and $Z|X \sim \mathsf{N}(0, \sigma^2\Omega\widehat{\Sigma}\Omega)$ means that the conditional distribution of Z given X is centered Gaussian, with the stated covariance.

Remark 1.2. The more complete statement of this result, Theorem 3.6 provides explicit non-asymptotic bounds on the error term $o_P(1)$, In particular $\|o_P(1)\|_\infty$ turns out to be of order $\sqrt{s_0/n}\,(\log p)$ with probability converging to one as $n, p \to \infty$.

Remark 1.3. We believe that a generalization of this result should apply to a broad class of random designs with independent sub-Gaussian rows. The main technical challenge in extending the present techniques to the sub-Gaussian setting is in generalizing the leave-one-out construction. As discussed in Section 4.1, when studying the effect of modifying column i, we need to account for dependencies between columns. This is easier to do for Gaussian designs, where dependencies are fully described by the design covariance Σ.

Remark 1.4. Roughly speaking, Theorem 1.1 (and its complete version, Theorem 3.6) implies that –at least for Gaussian designs, and neglecting logarithmic factors– statistical inference can be performed from a *number of samples that scale as the number of non-zero parameters*. This should be contrasted with earlier results [ZZ14, VdGBRD14, JM14a], that require a number of samples scaling as the square of the number of parameters.

It is instructive to compare this with the past progress in sparse estimation and compressed sensing. In that context, earlier work based on incoherence conditions [DH01, DET06] implied

accurate reconstruction from a number of random samples scaling quadratically in the number of non-zero coefficients. Subsequent progress was based on the restricted isometry property [CRT06, CT07], and established accurate reconstruction from a linear number of measurements.

Remark 1.5. An alternative approach to avoid the ℓ_1-ℓ_∞ bund in Eq. (9) is to modify the definition of de-biased estimator in Eq. (5), using sample-splitting. Roughly speaking, we can split the same in two batches of size $n/2$. One batch is then used to estimate $\widehat{\theta}^{\text{Lasso}}$ and the other batch for y and X appearing in Eq. (5) (and possibly for computing M).

Appendix F discusses in greater detail this method. This approach is subject to variations due to the random splitting, and does not make use of part of half of the response variables. While it provides a viable alternative, it is not the focus of the present work.

1.3 Extensions and applications

Theorem 1.1 raises an important questions: *Can we compute confidence intervals under the weak sparsity condition $s_0 = o(n/(\log p)^2)$, even if the precision matrix Ω is unknown? Does the Gaussian limit hold even if M is an imperfect estimate of Ω?*

The general procedure we have in mind is the same as in [ZZ14, VdGBRD14, JM14a]. Namely, we construct a suitable de-biasing matrix M from the design matrix X, and an estimate $\widehat{\sigma}$ of the noise variance. Then, for a significance level $\alpha \in (0,1)$, we construct the following confidence interval for parameter θ_i:

$$J_i(\alpha) \equiv [\widehat{\theta}_i^{\text{d}} - \delta(\alpha,n), \widehat{\theta}_i^{\text{d}} + \delta(\alpha,n)] \tag{11}$$

$$\delta(\alpha,n) \equiv \Phi^{-1}(1-\alpha/2)\frac{\widehat{\sigma}}{\sqrt{n}}(M\widehat{\Sigma}M^{\mathsf{T}})_{i,i}^{1/2}\,, \tag{12}$$

where $\Phi(x) \equiv \int_{-\infty}^{x} e^{-t^2/}\mathrm{d}t/\sqrt{2\pi}$ is the Gaussian distribution. Section 3.3 presents a more technical discussion of this procedure. A straightforward generalization also allows to compute p-values, for the null hypothesis $H_{0,i} : \theta_i^* = 0$. Here we provide a brief examination of various types of issues arising in applications and corresponding solutions.

Assumptions on the design. The assumption of random design X with i.i.d. Gaussian rows is obviously highly idealized. However this naturally arises in the context of estimating Gaussian graphical models. This is itself a broad topic that attracted significant amount of work, since the seminal work of [MB06]. Remarkably recent contributions have shown the utility of de-biasing methods in this context [JvdG14, CRZZ15, JvdG15].

From an even broader point of view, let us emphasize that a substantial part of earlier results on de-biasing assumed random designs [ZZ14, VdGBRD14, JM14a], albeit with less restrictive assumptions. We believe that the proof technique developed in the present paper might be generalizable to such a broader setting.

Noise level and regularization. The construction of the confidence interval $J_i(\alpha)$ in Eqs. (11), (12) requires a suitable choice of the regularization parameter λ, and an estimate of the noise level $\widehat{\sigma}$. The same difficulty was present in [ZZ14, VdGBRD14, JM14a]. The approaches used there (for instance, using the scaled Lasso [SZ12]) can be followed in the present case as well. Under the

assumptions of Theorem 1.1, the same proofs of [JM14a] show that the additional error due to the choice of λ and $\widehat{\sigma}$ are negligible.

Estimation of Ω. Crucially, Theorem 1.1 (and its technical version, Theorem 3.6) assumes that de-biasing is performed using the precision matrix $M = \Omega$. While in general Ω is unknown, there are settings in which an estimate M that is accurate enough can be constructed. Let us briefly mention two such scenarios: the second one will be analyzed in greater technical detail in Section 3.3.

Semi-supervised learning. In this context, the statistician is given additional samples $\overline{x}_1, \overline{x}_2, \dots, \overline{x}_N \in \mathbb{R}^p$ with the same distribution as the $\{x_i\}_{1\le i\le n}$. For these 'unlabeled' samples, the response variable is unknown. There are indeed many applications in which acquiring the response variable is much more challenging than capturing the covariates [CSZ06], and therefore $N \gg n$. In this setting, we can estimate Ω more accurately from $\{\overline{x}_i\}_{1\le i\le N}$ (using –for instance– a high-dimensional covariance estimation method as in [MB06]), then use this estimate to construct M.

Very sparse precision matrix. If the precision matrix Ω is sufficiently structured, then it can be reliably estimated from the design matrix X. Both [ZZ14] and [VdGBRD14] assume that Ω is sparse, and use the node-wise Lasso to construct an estimate $\widehat{\Omega}$ [MB06]. They then set $M = \widehat{\Omega}$.

We followed the same procedure and hence generalized Theorem 1.1 to the setting of unknown, sparse precision matrix. We state here a simplified version of this result, deferring to Theorem 3.10 for a more technical statement including non-asymptotic probability bounds.

Theorem 1.6 (Unknown covariance)**.** *Consider the linear model (2) where X has independent Gaussian rows with precision matrix Ω, satisfying the technical conditions of Theorem 1.1 (stated in Theorem 3.6). Define the de-biased estimator $\widehat{\theta}^{\rm d}$ via Eq. (5) with $\widehat{\theta}^{\rm Lasso} = \widehat{\theta}^{\rm Lasso}(y, X; \lambda)$, $\lambda = 8\sigma\sqrt{(\log p)/n}$, and $M = \widehat{\Omega}$ computed through node-wise Lasso (see Section 3.3).*

Let s_Ω the maximum number of non-zero entries in any row of Ω. If $n, p \to \infty$ with $s_0 = o(n/(s_\Omega(\log p)^2))$, then we have

$$\sqrt{n}(\widehat{\theta}^{\rm d} - \theta^*) = Z + o_P(1)\,, \qquad Z|X \sim \mathsf{N}(0, \sigma^2\Omega\widehat{\Sigma}\Omega)\,, \tag{13}$$

where $o_P(1)$ is a (random) vector satisfying $\|o_P(1)\|_\infty \to 0$ in probability as $n, p \to \infty$.

Remark 1.7. As mentioned above, this version of the de-biased estimator can be constructed entirely from data. The only unspecified steps are the choice of the regularization parameter λ, and the estimation of the noise level σ. These can be addressed as in [ZZ14, VdGBRD14, JM14a] without changes in the sparsity condition.

Remark 1.8. The sparsity condition $s_0 = o(n/(s_\Omega(\log p)^2))$ nicely illustrates the practical improvement implied by our more refined analysis. If the sparsity of the precision matrix is of the same order as the sparsity of θ^*, i.e. $s_\Omega = \Theta(s_0)$ as in [ZZ14, VdGBRD14], we recover the condition $s_0 = o(\sqrt{n}/\log p)$ which is assumed in the results of [ZZ14, VdGBRD14]. (Note that [JM14a] obtain the same condition without sparsity assumption on Ω.) In this regime, our improved analysis does not bring any advantage, since the bottleneck is due to the inaccurate estimation of Ω.

On the other hand, if the precision matrix is much sparser, we obtain a much weaker condition on the coefficients θ^*. For instance if $s_\Omega = \Theta(s_0^{1-b})$, then we get the condition $s_0 = o(n^{1/(2-b)}/(\log p)^{2/(2-b)})$, and for $b = 1$ (i.e. when Ω has $O(1)$ non-zeros per row), this reduces to $s_0 = o(n/(\log p)^2)$.

1.4 Organization

The rest of the paper is organized as follows. Section 2 discusses relations with earlier work in this area. We state formally our results in Section 3. This section contains some preliminary material, a complete statements of the two theorems discussed above (known and unknown covariance), and a numerical illustration. Section 4 presents the proof of Theorem 1.1, covering the case of known covariance (whose technical version is stated as Theorem 3.6). Section 5 presents the proof of Theorem 1.6, for unknown covariance (whose technical version is stated as Theorem 3.10).

Proofs of several technical lemmas are deferred to appendices.

2 Related work

A parallel line of research develops methods for performing valid inference after a low-dimensional model is selected for fitting high-dimensional data [LTTT14, FST14, TLTT14, CHS15]. The resulting significance statements are typically conditional on the selected model. In contrast, here we are interested in classical (unconditional) significance statements: the two approaches are broadly complementary.

Our proof is based on a leave-one out technique, and is partially inspired from ideas in mathematical spin glass theory [Tal10]. Similar techniques recently proved useful in analyzing robust regression [EKBBL13, Kar13].

The focus of the present paper is assessing statistical significance, such as confidence intervals, for single coordinates in the parameters vector θ^* and more generally for small groups of coordinates. Other inference tasks are also interesting and challenging in high-dimension, and were the object of recent investigations [BEM13, BC14, JBC15, JS15].

Sample splitting provides a general methodology for inference in high dimension [WR09, MB10]. As mentioned above, sample splitting can also be used to define a modified de-biased estimator, see Appendix F. However sample splitting techniques typically use only part of the data for inference, and are therefore sub-optimal. Also, the result depend on the random split of the data.

A method for inference without assumptions on the design matrix was developed in [Mei14]. The resulting confidence intervals are typically quite conservative.

The de-biasing method was developed independently from several points of view [Büh13, ZZ14, JM14b, VdGBRD14, JM14a]. The present authors were motivated by the AMP analysis of the Lasso [DMM09, BM11, BM12, BLM15], and by the Gaussian limits that this analysis implies. In particular [JM14b] used those techniques to analyze standard Gaussian designs (i.e. the case $\Sigma = \mathrm{I}$) in the asymptotic limit $n, p, s_0 \to \infty$ with s_0/p, n/p constant. In this limit, the de-biased estimator was proven to be asymptotically Gaussian provided $s_0 \le C\, n/\log(p/s_0)$ (for a universal constant C). This sparsity condition is even weaker than the one of Theorem 1.1 (or Theorem 3.6), but the result of [JM14b] only holds asymptotically. Also [JM14b] proved Gaussian convergence in a weaker sense than the one established here, implying coverage of the constructed confidence intervals only 'on average' ofer the coordinates $i \in \{1, \dots, p\}$.

A non-asymptotic result under weaker sparsity conditions, and for designs with dependent columns, was proved in [JM13]. However, this only establishes gaussianity of $\widehat{\theta}^{\mathrm{d}}_i$ for most of the coordinates $i \in \{1, \dots, p\}$. Here we prove a significantly stronger result holding uniformly over $i \in \{1, \dots, p\}$. In a recent and independent contribution, Cai and Guo [CG15] also investigate the regime of moderate sparsity $\sqrt{n}/\log p \lesssim s_0 \lesssim n/\log p$, and construct valid confidence intervals, of size $(s_0 \log p)/n$ (much

larger than the parametric rate $1/\sqrt{n}$ that we consider here). While this length is proved to be optimal, this lower bound is related to the estimation of the design covariance. Our Theorem 3.10 clarifies the trade-off between knowledge of the design covariance, and sparsity of the coefficients' vector.

Most of the work on statistical inference in high-dimensional models has been focused so far on linear regression. The de-biasing method admits a natural extension to generalized linear models that was analyzed in [VdGBRD14]. Robustness to model misspecification was studied in [BvdG15]. An R-package for inference in high-dimension that uses the node-wise Lasso is available [DBMM14]. An R implementation of the method [JM14a] (which does not make sparsity assumptions on Ω) is also available[1].

3 Results

3.1 General notations

We use e_i to refer to the i-th standard basis element, e.g., $e_i = (1, 0, \dots, 0)$. For a vector v, $\mathrm{supp}(v)$ represents the positions of nonzero entries of v. Further, for a vector v, $\mathrm{sign}(v)$ is the vector with entries $\mathrm{sign}(v)_i = +1$ if $v_i > 0$, $\mathrm{sign}(v)_i = -1$ if $v_i < 0$, and $\mathrm{sign}(v)_i = 0$ otherwise. For a matrix $M \in \mathbb{R}^{n\times p}$ and a set of indices $J \subseteq [p]$ we use M_J to denote the submatrix formed by columns in J. Likewise, for a vector θ and a subset S, θ_S is the restriction of θ to indices in S. For an integer $p \geq 1$, we use the notation $[p] = \{1, \cdots, p\}$ and the shorthand $\sim i$ for the set $[p]\backslash i$. We write $\|v\|_p$ for the standard ℓ_p norm of a vector v, i.e., $\|v\|_p = (\sum_i |v_i|^p)^{1/p}$ and $\|v\|_0$ for the umber of nonzero entries of v. For a matrix $A \in \mathbb{R}^{m\times n}$, $\|A\|_p$ denotes it ℓ_p operator norm; in particular, $\|A\|_\infty = \max_{1\leq i\leq m} \sum_{j=1}^n |A_{ij}|$. This is to be contrasted with the maximum absolute value of any entry of A that, as mentioned above, we denote by $|A|_\infty \equiv \max_{i\leq m, j\leq n} |A_{ij}|$. Finally, for two functions $f(n)$ and $g(n)$, the notation $f(n) \gg g(n)$ means that f 'dominates' g asymptotically, namely, for every fixed positive C, there exists $n(C)$ such that $f(n) \geq Cg(n)$ for $n > n(C)$. We also use $f(n) \lesssim g(n)$ to indicate that f is 'bounded' above by g asymptotically, i.e., $f(n) \leq Cg(n)$ for some positive constant C. We use the notations $f(n) \ll g(n)$ and $f(n) = o(g(n))$ interchangeably and $o_P(\,\cdot\,)$ to indicate asymptotic behavior in probability as the sample size n tends to infinity.

We will use $c, C, \dots$ to denote generic constants that can vary from one position to the other of the paper.

3.2 Preliminaries

For the sake of simplicity, we will often use $\widehat{\theta} = \widehat{\theta}(y, X; \lambda)$ instead of $\widehat{\theta}^{\mathrm{Lasso}}$ to denote the Lasso estimator.

We denote the rows of the design matrix X by $x_1, \dots, x_n \in \mathbb{R}^p$ and its columns by $\tilde{x}_1, \dots, \tilde{x}_p \in \mathbb{R}^n$. The empirical covariance of the design X is defined as $\widehat{\Sigma} \equiv (X^{\mathsf{T}}X)/n$. The population covariance will be denoted by Σ, we let $\Omega \equiv \Sigma^{-1}$ be the precision matrix.

[1] See http://web.stanford.edu/ montanar/sslasso/.

Definition 3.1. *Given a symmetric matrix $\widehat{\Sigma} \in \mathbb{R}^{p\times p}$ and a set $S \subseteq [p]$, the corresponding* compatibility constant *is defined as*

$$\phi^2(\widehat{\Sigma}, S) \equiv \min_{\theta\in\mathbb{R}^p} \left\{ \frac{|S|\,\langle \theta, \widehat{\Sigma}\,\theta\rangle}{\|\theta_S\|_1^2} :\ \ \theta \in \mathbb{R}^p,\ \ \|\theta_{S^c}\|_1 \le 3\|\theta_S\|_1 \right\}. \tag{14}$$

We say that $\widehat{\Sigma} \in \mathbb{R}^{p\times p}$ satisfies the compatibility condition *for the set $S \subseteq [p]$, with constant ϕ if $\phi(\widehat{\Sigma}, S) \ge \phi$. We say that it holds for the design matrix X, if it holds for $\widehat{\Sigma} = X^{\mathsf{T}}X/n$.*

It is also useful to recall the notion of restricted eigenvalue, introduced by Bickel, Ritov and Tsybakov [BRT09]. For integer $0 < s_0 < p$ and a positive number L, define the set $\mathcal{C}(s_0, L)$ to be the set of vectors in $\mathbb{R}^p$ that satisfy the following cone constraints:

$$\mathcal{C}(s_0, L) \equiv \{\theta \in \mathbb{R}^p :\ \exists S \subseteq [p],\ |S| = s_0,\ \|\theta_{S^c}\|_1 \le L\|\theta_S\|_1\}\,.$$

In the high-dimensional regime the empirical covariance $\widehat{\Sigma}$ is singular, however, we can ask for non-singularity of $\widehat{\Sigma}$ on cone-restricted directions, namely for vectors in $\mathcal{C}(s_0, L)$. Rudelson and Zhou [RZ13] prove a reduction principle that bounds the restricted eigenvalues of the empirical covariance in terms of those of the population covariance. We will use their result specified to the case of Gaussian matrices.

Lemma 3.2. *[RZ13, Theorem 3.1] Suppose that $\sigma_{\min}(\Sigma) > C_{\min} > 0$ and $\sigma_{\max}(\Sigma) < C_{\max} < \infty$. Let $X \in \mathbb{R}^{n\times p}$ have independent rows drawn from $\mathsf{N}(0, \Sigma)$. Set $0 < \delta < 1$, $0 < s_0 < p$, and $L > 0$. Define the following event*

$$\mathcal{B}_\delta(n, s_0, L) \equiv \left\{ X \in \mathbb{R}^{n\times p} :\ (1-\delta)\sqrt{C_{\min}} \le \frac{\|Xv\|_2}{\sqrt{n}\|v\|_2} \le (1+\delta)\sqrt{C_{\max}}\,,\ \forall v \in \mathcal{C}(s_0, L)\ \textit{s.t.}\ v \neq 0 \right\}, \tag{15}$$

There exists a sufficiently large constant $c_1 = c_1(L)$, such that, for sample size $n \ge c_1 s_0 \log(p/s_0)$ we have

$$\mathbb{P}(\mathcal{B}_\delta(n, s_0, L)) \ge 1 - 2e^{-\delta^2 n}\,.$$

Remark 3.3. Fix $S \subseteq [p]$ with $|S| = s_0$. Under the event $\mathcal{B}_\delta(n, s_0, 3)$, we have

$$\begin{aligned}
\phi^2(\widehat{\Sigma}, S) &\ge \min_{\theta\in\mathcal{C}(s_0,3)} \frac{s_0\langle\theta, \widehat{\Sigma}\theta\rangle}{\|\theta_S\|_1^2} \\
&\ge \min_{\theta\in\mathcal{C}(s_0,3)} \frac{\langle\theta, \widehat{\Sigma}\theta\rangle}{\|\theta_S\|_2^2} \\
&\ge (1-\delta)^2 C_{\min}\,,
\end{aligned}$$

where the second inequality follows from Cauchy-Schwartz inequality.

We next introduce an event $\tilde{\mathcal{B}}(n, p)$ as

$$\tilde{\mathcal{B}}(n, p) \equiv \left\{ \frac{1}{n}\|X^{\mathsf{T}}w\|_\infty \le 2\sigma\sqrt{\frac{\log p}{n}} \right\}. \tag{16}$$

On $\tilde{\mathcal{B}}(n, p)$ we can control the randomness part of the problem occurring because of the measurement noise. A well-known union bound argument shows that $\tilde{\mathcal{B}}(n, p)$ has large probability (see, for instance, [BvdG11]).

Lemma 3.4. *[BvdG11, Lemma 6.2] Suppose that $\widehat{\Sigma}_{ii} \le 1$ for $i \in [p]$. Then we have*

$$\mathbb{P}(\tilde{\mathcal{B}}(n,p)) \ge 1 - 2p^{-1}\,.$$

Finally the lemma below states a property of Gaussian design matrices which will be used repeatedly in our analysis.

Lemma 3.5. *Let $v_i = X\Omega e_i$. Then v and $X_{\sim i}$ are independent.*

Proof. Define $u = \Omega e_i$ and fix $j \neq i$. Recall that $\tilde{x}_\ell$ denotes the ℓ-th column of X. We write $v = \sum_{\ell=1}^p \tilde{x}_\ell u_\ell$ and

$$\begin{aligned}\mathbb{E}(v\tilde{x}_j^{\mathsf{T}}) &= \sum_{\ell=1}^p u_\ell \mathbb{E}(\tilde{x}_\ell \tilde{x}_j^{\mathsf{T}}) \\ &= \sum_{\ell=1}^p u_\ell \Sigma_{\ell j} \mathrm{I}_{n\times n} = \sum_{\ell=1}^p \Omega_{\ell i}\Sigma_{\ell j}\mathrm{I}_{n\times n} \\ &= (\Omega\Sigma)_{ij}\mathrm{I}_{n\times n} = 0\,,\end{aligned}$$

where the last step holds since $i \neq j$. Since v and $\tilde{x}_j$ are jointly Gaussian, this implies that they are independent. □

3.3 Statement of main theorems

In our first theorem, we assume that the precision matrix $\Omega \equiv \Sigma^{-1}$ is available and we set $M = \Omega$. We prove the corresponding de-biased estimator is asymptotically unbiased provided that $n \gg s_0(\log p)^2$.

Theorem 3.6 (Known covariance)**.** *Consider the linear model (2) where X has independent Gaussian rows, with zero mean and covariance Σ. Suppose that Σ satisfies the following conditions:*

(i) For $i \in [p]$, we have $\Sigma_{ii} \le 1$.

(ii) We have $\sigma_{\min}(\Sigma) > C_{\min} > 0$ and $\sigma_{\max}(\Sigma) < C_{\max}$ for some constants $C_{\min}$ and $C_{\max}$.

(iii) We have $\|\Sigma^{-1}\|_\infty \le \rho$, for some constant $\rho > 0$.

Let $\widehat{\theta}$ be the Lasso estimator defined by (3) with $\lambda = 8\sigma\sqrt{(\log p)/n}$. Further, let $\widehat{\theta}^{\mathrm{d}}$ be defined as per equation (5), with $M = \Omega \equiv \Sigma^{-1}$. Then, there exist constants c, C depending solely on $C_{\min}, C_{\max}, \delta$ and ρ, such that, for $n \ge \max(25\log p, cs_0\log(p/s_0))$ the following holds true:

$$\sqrt{n}(\widehat{\theta}^{\mathrm{d}} - \theta^*) = Z + R\,, \qquad Z|X \sim \mathsf{N}(0, \sigma^2\Omega\widehat{\Sigma}\Omega)\,, \tag{17}$$

$$\mathbb{P}\Big(\|R\|_\infty \ge C\sqrt{\frac{s_0}{n}}\log p\Big) \le 2pe^{-c_* n/s_0} + pe^{-n/1000} + 8p^{-1} + 2e^{-\delta^2 n}\,, \tag{18}$$

with $c_* \equiv (1-\delta)^2 C_{\min}/8$.

Let us remind that the notation $\|\Sigma\|_\infty$ is the maximum ℓ_1 norms of the rows of Σ. The proof of this theorem is presented in Section 4.

This theorem states that if the sample size satisfies $n = \Omega(s_0 \log p)$, then the maximum size of the 'bias' R_i over $i \in [p]$ is bounded by

$$\|R\|_\infty = O_P\Big(\sqrt{\frac{s_0}{n}} \log p\Big)\,.$$

On the other hand, each entry of the 'noise term' Z_i has variance $\sigma^2(\Omega\widehat{\Sigma}\Omega)_{ii}$. Applying Lemma 7.2 in [JM13], we have $|\Omega\widehat{\Sigma}\Omega - \Omega|_\infty = o_P(1)$ and thus $\min_{i\in[p]}(\Omega\widehat{\Sigma}\Omega)_{ii} \geq \min_{ii}\Omega_{ii} - o_P(1)$ is of order one because $\Omega_{ii} \geq C_{\max}^{-1}$. Hence, $|R_i|$ is much smaller than Z_i for $n \gg s_0(\log p)^2$. We summarize this observation in the remark below.

Remark 3.7. Under the assumptions of Theorem 3.6, if the sample size satisfies $n \gg s_0(\log p)^2$, then we have $\|R\|_\infty = o_P(1)$ and $\min_{i\in[p]}(\Omega\widehat{\Sigma}\Omega)_{ii} = \Omega(1)$, with high probability. Hence, $\widehat{\theta}^{\mathrm{d}}$ is an asymptotically unbiased estimator for θ^*.

Corollary 3.8. *Under the assumptions of Theorem 3.6, if $s_0 \ll n/(\log p)^2$, then $\widehat{\theta}^{\mathrm{d}}$ is normal distributed. More precisely, for all $x \in \mathbb{R}$, we have*

$$\lim_{n\to\infty} \sup_{\theta_0\in\mathbb{R}^p, \|\theta_0\|_0\leq s_0} \left|\mathbb{P}\left\{\frac{\sqrt{n}(\widehat{\theta}^{\mathrm{d}}_i - \theta_{0,i})}{\sigma(M\widehat{\Sigma}M)^{1/2}_{i,i}} \leq x\right\} - \Phi(x)\right| = 0\,. \tag{19}$$

Armed with a precise distributional characterization of $\widehat{\theta}^{\mathrm{d}}$, we can construct asymptotically valid confidence intervals for each parameter $\theta_{0,i}$ as per Eqs. (11), (12).

Furthermore, in the context of hypothesis testing, we can test the null hypothesis $H_{0,i} : \theta_0 = 0$ versus the alternative $H_{A,i} : \theta_{0,i} \neq 0$. We construct the two sided p-values

$$P_i = 2\left(1 - \Phi\Big(\frac{\sqrt{n}|\widehat{\theta}^{\mathrm{d}}_i|}{\sigma(M\widehat{\Sigma}M^{\mathsf{T}})^{1/2}_{i,i}}\Big)\right). \tag{20}$$

The decision rule follows immediately: we reject $H_{0,i}$ if $P_i \leq \alpha$. As already mentioned, in practice σ has to be replaced by an estimator $\widehat{\sigma}$, an issue already discussed in [JM14a].

Remark 3.9. It is worth noting that the sample splitting approach, discussed in Appendix F, does not require Condition *(iii)* in Theorem 3.6. However as pointed in the introduction, this approach suffers from variability due to the random splitting and does not fully use half of the response variables.

We next generalize our result to the case of unknown covariance, where following [ZZ14, VdGBRD14] we construct the de-biasing matrix M using node-wise Lasso on matrix X. For reader's convenience, we first describe this construction.

For $i \in [p]$, we define the vector $\hat{\gamma}_i = (\hat{\gamma}_{i,j})_{j\in[p]\setminus i} \in \mathbb{R}^{p-1}$ by performing sparse regression of the i-th column of X against all the other columns. Formally

$$\hat{\gamma}_i(\widetilde{\lambda}) = \arg\min_{\gamma\in\mathbb{R}^p}\left\{\frac{1}{2n}\|\tilde{x}_i - X_{\sim i}\gamma\|_2^2 + \widetilde{\lambda}\|\gamma\|_1\right\}, \tag{21}$$

where $X_{\sim i}$ is the sub-matrix obtained by removing the i-th column (and columns indexed by $[p] \setminus i$). Also define

$$\widehat{C} = \begin{bmatrix} 1 & -\hat{\gamma}_{1,2} & \cdots & -\hat{\gamma}_{1,p} \\ -\hat{\gamma}_{2,1} & 1 & \cdots & -\hat{\gamma}_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ -\hat{\gamma}_{p,1} & -\hat{\gamma}_{p,2} & \cdots & 1 \end{bmatrix}, \tag{22}$$

and let

$$\widehat{T}^2 = \mathrm{diag}(\hat{\tau}_1^2, \dots, \hat{\tau}_p^2), \qquad \hat{\tau}_i^2 = \frac{1}{n}(\tilde{x}_i - X_{\sim i}\hat{\gamma}_j)^{\mathsf{T}}\tilde{x}_i\,. \tag{23}$$

Finally, define $M = M(\widetilde{\lambda})$ by

$$M = \widehat{T}^{-2}\widehat{C}\,. \tag{24}$$

Theorem 3.10 (Unknown covariance)**.** *Consider the linear model (2) where X has independent Gaussian rows, with zero mean and covariance Σ. Suppose that Conditions $(i), (ii), (iii)$ in Theorem 3.6 hold true for Σ. We further let s_Ω be the maximum sparsity of the rows of $\Omega \equiv \Sigma^{-1}$, i.e.*

$$s_\Omega \equiv \max_{i\in[p]} |\{j \neq i, \Omega_{i,j} \neq 0\}|\,. \tag{25}$$

Let $\widehat{\theta}$ be the Lasso estimator defined by (3) with $\lambda = 8\sigma\sqrt{(\log p)/n}$, and let $\widehat{\theta}^{\mathrm{d}}$ be de-biased estimator with M given by (24) with $\widetilde{\lambda} = K\sqrt{\log p/n}$ (with K a suitably large universal constant).

Then, there exist constants c, C depending solely on $C_{\min}, C_{\max}, \delta$ and ρ, such that, for $n \geq cs_0 \log p$, the following holds true:

$$\begin{aligned} &\sqrt{n}(\widehat{\theta}^{\mathrm{d}} - \theta^*) = Z + R\,, \qquad Z|X \sim \mathsf{N}(0, \sigma^2 M\widehat{\Sigma}M^{\mathsf{T}})\,, \\ &\mathbb{P}\Big(\|R\|_\infty \geq C\sqrt{\frac{s_0 s_\Omega}{n}}\log p\Big) \leq 2pe^{-c_* n/s_0} + pe^{-n/1000} + 8p^{-1} + 2e^{-\delta^2 n} + c'e^{-c''n}\,, \end{aligned} \tag{26}$$

for some constants $c_*, c', c'' > 0$.

The proof of Theorem 3.10 is deferred to Section 5.

A large family of precision matrices satisfy conditions of Theorem 3.6 with s_Ω bounded. Examples include block diagonal Ω where the size of blocks are bounded, and inverse of circulant matrices, e.g. $\Sigma_{i,j} = r^{|i-j|}$, for some $r \in (0, 1)$.

3.4 Numerical illustration

Our goal in this section is to numerically corroborate the result of Theorem 3.6. Mores specifically, we aim to show that the de-biased estimator exhibits an unbiased Gaussian distribution provided that the sample size scales linearly with the number of nonzero parameters.

We generate data from linear model (1) with the following configuration. We fix $p = 3000$ and consider regression parameter θ_0 with support S_0 chosen uniformly at random from the index set $[p]$ and $\theta_{0,i} = 0.15$ for $i \in S_0$ and zero otherwise. The design matrix X has i.i.d. rows drawn from

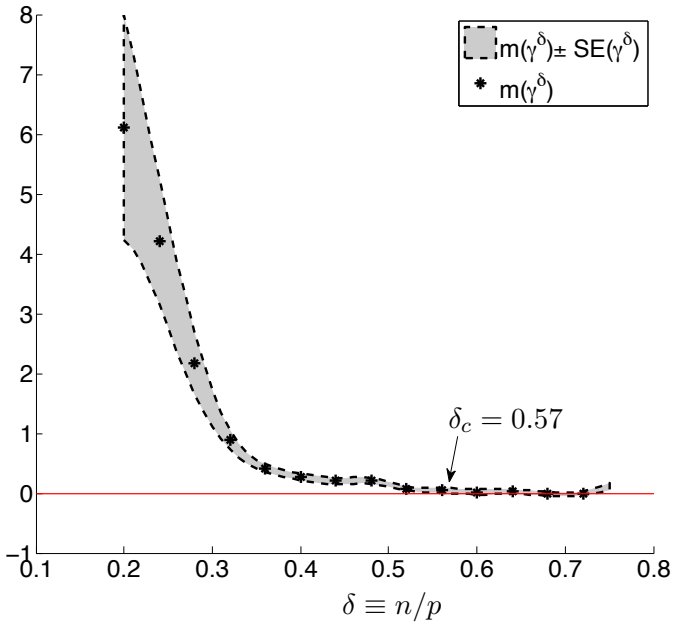

Figure 1: Empirical kurtosis of the (rescaled) de-biased Lasso estimator $T_i = \sqrt{n}(\widehat{\theta}_i^{\mathrm{d}} - \theta_i^*)/(\sigma[M\widehat{\Sigma}M]_{i,i}^{1/2})$. We plot the kurtosis $m(\gamma^\delta)$ (over coordinates and 100 independent realizations) versus δ along with the upper and lower one standard error curves, As a function of the number of samples per parameter δ.. Here, $\varepsilon = 0.2$ and $\delta_c = 0.57$ is our empirical estimate for the number of samples above which the de-biased estimator is approximately Gaussian.

$\mathsf{N}(0, \Sigma)$, where $\Sigma \in \mathbb{R}^{p\times p}$ is the circulant matrix with entries $\Sigma_{i,j} = 0.8^{|i-j|}$. The measurement noise w has i.i.d. standard normal entries.

Let $s_0 = |S_0|$ and $\varepsilon = s_0/p$ be the sparsity level and $\delta = n/p$ denote the under sampling rate. We vary ε in the set $\{0.1, 0.15, 0.2, 0.25, 0.3\}$ and for each value of ε we compute critical value of δ above which the unbiased estimator admits a Gaussian distribution. We will denote this critical value as δ_c and define it as follows. We vary δ and for each pair (ε, δ), compute the de-biased estimator (with $M = \Sigma^{-1}$) for 100 realizations of noise w. We then compute the empirical kurtosis of each coordinate $T_i = \sqrt{n}(\widehat{\theta}_i^{\mathrm{d}} - \theta_i^*)/(\sigma[M\widehat{\Sigma}M]_{i,i}^{1/2})$. For $i \in [p]$, let γ_i^δ denote the empirical kurtosis of T_i, where we make the dependence on δ explicit in the notation. Denote by $m(\gamma^\delta)$ and $\mathsf{SD}(\gamma^\delta)$ the mean and the standard deviation of $\gamma^\delta = (\gamma_1^\delta, \ldots, \gamma_p^\delta)$, respectively. We further define the standard error $\mathsf{SE}(\gamma^\delta) = \mathsf{SD}(\gamma^\delta)/\sqrt{p}$. We use one standard error rule to decide the value of δ_c. Namely,

$$\delta_c = \arg\min\{\delta \in (0,1),\ \text{s.t.},\ m(\gamma^\delta) \le \mathsf{SE}(\gamma^\delta) \le 0\}\,. \tag{27}$$

Figure 1 corresponds to $\varepsilon = 0.2$. The asterisks indicate $m(\gamma^\delta)$ and the dotted lines are $m(\gamma^\delta) \pm \mathsf{SE}(\gamma^\delta)$. By one standard error rule, the estimated value of δ_c works out at $\delta_c = 0.57$.

Figure 2 shows δ_c versus ε. The figure clearly verifies that δ_c scales at roughly linearly in ε (for small ε). In other words, in order for the de-biased estimator to have unbiased Gaussian distribution, the sample size n has only to scale roughly linearly in the support size s_0.

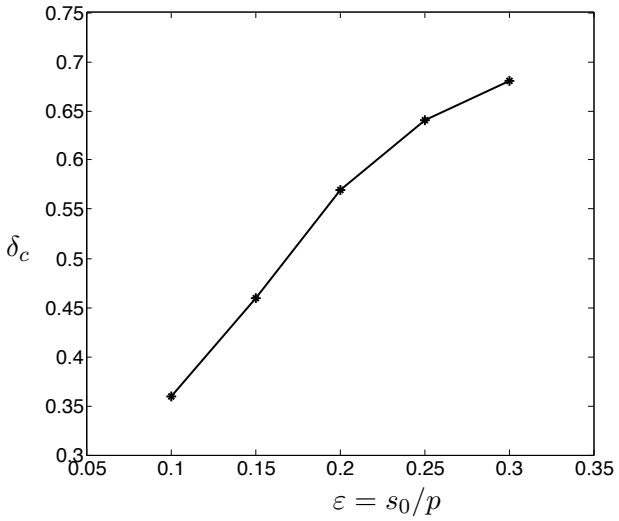

Figure 2: Critical number of samples per coordinate δ_c, versus fraction of non-zero coordinates ε. For $\delta > \delta_c(\varepsilon)$ the de-biased Lasso estimator is empirically Gaussian distributed in our experiment. The approximately linear relationship at small ε is in agreement with our theory.

4 Proof of Theorem 3.6 (known covariance)

4.1 Outline of the proof

Fix arbitrary integer $i \in [p]$. In our analysis, we focus on the i-th coordinate θ^*_i, and then discuss how the argument can be adjusted to apply to all the coordinates simultaneously. Our argument relies on a perturbation analysis. We let $\widehat{\theta}^{\rm p}$ be the Lasso estimator when one forces $\widehat{\theta}^{\rm p}_i = \theta^*_i$. With a slight abuse of notation, we use the representation $\theta = (\theta_i, \theta_{\sim i})$.[2] Adopting this convention, we have $\widehat{\theta}^{\rm p} = (\theta^*_i, \widehat{\theta}^{\rm p}_{\sim i})$ where

$$\widehat{\theta}^{\rm p}_{\sim i} = \arg\min_{\theta} \mathcal{L}_{y,X}(\theta^*_i, \theta)\,. \tag{28}$$

Throughout, we make the convention that $\mathcal{L}_{y,X}(\theta^*_i, \theta) \equiv \mathcal{L}_{y,X}((\theta^*_i, \theta))$.

We observe that $\widehat{\theta}^{\rm p}_{\sim i}$ can be written as a Lasso estimator. Specifically, by definition of Lasso cost function we have

$$\mathcal{L}_{y,X}(\theta^*_i, \theta) = \frac{1}{2n}\|y - \tilde{x}_i\theta^*_i - X_{\sim i}\theta\|_2^2 + \lambda|\theta^*_i| + \lambda\|\theta\|_1\,.$$

Letting $\tilde{y} = y - \tilde{x}_i\theta^*_i$, we obtain

$$\widehat{\theta}^{\rm p}_{\sim i} = \arg\min_{\theta} \mathcal{L}_{\tilde{y},X_{\sim i}}(\theta)\,. \tag{29}$$

[2] Or without loss of generality one can assume $i = 1$.

Let $v_i = X\Omega e_i$ and expand $\widehat{\theta}^{\mathrm{d}}_i - \theta^*_i$ as follows:

$$\begin{aligned}\sqrt{n}(\widehat{\theta}^{\mathrm{d}}_i - \theta^*_i) &\equiv \sqrt{n}\widehat{\theta}_i + \frac{1}{\sqrt{n}} e_i^{\mathsf{T}}\Omega X^{\mathsf{T}}(y - X\widehat{\theta}) - \sqrt{n}\theta^*_i \\ &= \sqrt{n}\widehat{\theta}_i + \frac{v^{\mathsf{T}}}{\sqrt{n}}\Big[w + \tilde{x}_i(\theta^*_i - \widehat{\theta}_i) + X_{\sim i}(\theta^*_{\sim i} - \widehat{\theta}_{\sim i})\Big] - \sqrt{n}\theta^*_i \\ &= \sqrt{n}\Big(1 - \frac{1}{n}\langle v_i, \tilde{x}_i\rangle\Big)(\widehat{\theta}_i - \theta^*_i) + \frac{v_i^{\mathsf{T}}}{\sqrt{n}}\Big[w + X_{\sim i}(\theta^*_{\sim i} - \widehat{\theta}_{\sim i})\Big]\,. \end{aligned} \tag{30}$$

We decompose the above expression into the following terms:

$$\begin{aligned} Z_i &\equiv \frac{v_i^{\mathsf{T}} w}{\sqrt{n}}\,, \\ R_i^{(1)} &\equiv \sqrt{n}\Big(1 - \frac{\langle v_i, \tilde{x}_i\rangle}{n}\Big)(\widehat{\theta}_i - \theta^*_i)\,, \\ R_i^{(2)} &\equiv \frac{v_i^{\mathsf{T}}}{\sqrt{n}} X_{\sim i}(\theta^*_{\sim i} - \widehat{\theta}^{\mathrm{p}}_{\sim i})\,, \\ R_i^{(3)} &\equiv \frac{v_i^{\mathsf{T}}}{\sqrt{n}} X_{\sim i}(\widehat{\theta}^{\mathrm{p}}_{\sim i} - \widehat{\theta}_{\sim i})\,. \end{aligned} \tag{31}$$

The bulk of the proof consists in treating each of the terms above separately. Term Z_i gives the Gaussian component Z in equation (17). In bounding $R_i^{(2)}$ we use the fact that v_i is independent of $X_{\sim i}$, as per Lemma 3.5. Moreover, since $\widehat{\theta}^{\mathrm{p}}_{\sim i}$ is a deterministic function of $(y, X_{\sim i})$, v_i is independent of $X_{\sim i}(\theta^*_{\sim i} - \widehat{\theta}^{\mathrm{p}}_{\sim i})$ as well. Bounding $R_i^{(3)}$ relies on a perturbation analysis showing that the solutions of Lasso $\widehat{\theta}$ and its perturbed form $\widehat{\theta}^{\mathrm{p}}$, are close to each other.

4.2 Technical steps

Let $Z = (Z_i)_{1\le i\le p}$. We rewrite Z as

$$Z = \frac{1}{\sqrt{n}}\Omega X^{\mathsf{T}} w\,.$$

Since $w \sim \mathsf{N}(0, \sigma^2\mathrm{I})$ is independent of X, we get

$$Z|X \sim \mathsf{N}(0, \sigma^2\Omega\widehat{\Sigma}\Omega)\,.$$

Let $R^{(1)} = (R_i^{(1)})_{i=1}^p, R^{(2)} = (R_i^{(2)})_{i=1}^p, R^{(3)} = (R_i^{(3)})_{i=1}^p \in \mathbb{R}^p$. In the following, we provide a detailed analysis to control the terms $R^{(1)}, R^{(2)}, R^{(3)}$.

• *Bounding term $R^{(1)}$:* Recalling the definition $v_i = X\Omega e_i$, we write

$$R_i^{(1)} = \sqrt{n}\Big(1 - \frac{1}{n} e_i^{\mathsf{T}}\Omega X^{\mathsf{T}} X e_i\Big)(\widehat{\theta}_i - \theta^*_i)\,.$$

Therefore,

$$\|R^{(1)}\|_\infty \le \sqrt{n}|\mathrm{I} - \Omega\widehat{\Sigma}|_\infty \|\widehat{\theta} - \theta^*\|_2\,.$$

For $A > 0$, let $\mathcal{G} = \mathcal{G}(A)$ be the event that

$$\mathcal{G}_n(A) \equiv \left\{ X \in \mathbb{R}^{n\times p} : \, |\Omega\widehat{\Sigma} - \mathrm{I}|_\infty \le A\sqrt{\frac{\log p}{n}} \right\}.$$

Using the result of [JM14a, Lemma 6.2] for $n \ge (A^2 C_{\min})/(4e^2 C_{\max}) \log p$ we have

$$\mathbb{P}(X \in \mathcal{G}_n(a)) \ge 1 - 2p^{-c}, \qquad c = \frac{A^2 C_{\min}}{24 e^2 C_{\max}} - 2.$$

By choosing $A \equiv 10e\sqrt{C_{\max}/C_{\min}}$ we get $c \ge 1$. Therefore, provided that $n \ge 25 \log p$,

$$\mathbb{P}(X \in \mathcal{G}_n(A)) \ge 1 - 2p^{-1}. \tag{32}$$

In addition, on the event $\mathcal{B} \equiv \mathcal{B}_\delta(n, s_0, 3) \cap \tilde{\mathcal{B}}(n,p)$ we have [BvdG11]

$$\|\widehat{\theta} - \theta^*\|_2 \le \frac{\sqrt{20}}{(1-\delta)^2 C_{\min}} \lambda \sqrt{s_0} = \frac{40\sigma}{(1-\delta)^2 C_{\min}} \sqrt{\frac{s_0 \log p}{n}}.$$

Combining the above bounds, we obtain that on event $\mathcal{G}_n(A) \cap \mathcal{B}$,

$$\|R^{(1)}\|_\infty \le \frac{40 A\sigma}{(1-\delta)^2 C_{\min}} \sqrt{\frac{s_0}{n}} \log p. \tag{33}$$

- *Bounding term $R^{(2)}$:* To lighten the notation, we define

$$\zeta_i \equiv \frac{1}{\sqrt{n}} X_{\sim i} (\theta^*_{\sim i} - \widehat{\theta}^{\mathrm{p}}_{\sim i}). \tag{34}$$

As discussed $\widehat{\theta}^{\mathrm{p}}_{\sim i}$ is a Lasso estimator with design matrix $X_{\sim i}$ and response vector $\tilde{y} = y - \tilde{x}_i \theta^*_i$, as per equation (29). We recall the following results on the prediction error of the Lasso estimator, which bounds $\|\zeta_i\|_2$.

Proposition 4.1. *[BvdG11, Theorem 6.1] Let $S \equiv \mathrm{supp}(\theta^*_{\sim i})$. Then on the event $\tilde{\mathcal{B}}(n,p)$, we have for $\lambda \ge 8\sigma\sqrt{(\log p)/n}$,*

$$\|\zeta_i\|_2^2 \le \frac{4\lambda^2 |S|}{\phi^2(S, \widehat{\Sigma}_{\sim i, \sim i})}.$$

From the definition of the compatibility constant (cf. Definition 3.1), it is clear that $\phi^2(S, \widehat{\Sigma}_{\sim i,\sim i}) \ge \phi^2(S, \widehat{\Sigma})$. Therefore, combining Proposition 4.1 and Remark 3.3, we arrive at the following corollary:

Corollary 4.2. *On the event $\mathcal{B} \equiv \mathcal{B}_\delta(n, s_0, 3) \cap \tilde{\mathcal{B}}(n,p)$, we have for $\lambda \ge 8\sigma\sqrt{(\log p)/n}$,*

$$\|\zeta_i\|_2^2 \le \frac{4\lambda^2 s_0}{(1-\delta)^2 C_{\min}}.$$

Employing Corollary 4.2, we derive a tail bound on $R_i^{(2)}$.
For $i \in [p]$ define the event

$$\mathcal{E}_i \equiv \left\{ \|\zeta_i\|_2^2 \leq \frac{4\lambda^2 s_0}{(1-\delta)^2 C_{\min}} \right\}. \tag{35}$$

By Corollary 4.2, we have $\mathcal{B} \subseteq \mathcal{E}_i$ for $i \in [p]$. Hence, for any value $t > 0$

$$\begin{aligned}
\mathbb{P}\Big(\|R^{(2)}\|_\infty \geq t; \mathcal{B} \Big) &\leq \mathbb{P}\Big(\max_{i \in [p]} |v_i^\mathsf{T} \zeta_i| \geq t; \mathcal{E}_i \Big) \\
&\leq p \max_{i \in [p]} \mathbb{E}\Big\{ \mathbb{I}(|v_i^\mathsf{T} \zeta_i| \geq t) \cdot \mathbb{I}(\mathcal{E}_i) \Big\} \\
&\leq 2p \max_{i \in [p]} \mathbb{E}\Big(\exp\Big[- \frac{t^2}{2\Omega_{ii} \|\zeta_i\|^2} \Big] \cdot \mathbb{I}(\mathcal{E}_i) \Big) \\
&\leq 2p \exp\Big(- \frac{c_* t^2}{s_0 \lambda^2 \Omega_{ii}} \Big),
\end{aligned}$$

with $c_* \equiv (1-\delta)^2 C_{\min}/8$. In the third inequality, we applied Fubini's theorem, and first integrate w.r.t v_i and then w.r.t ζ_i using the fact that v_i and ζ_i are independent. Note that $v_i \sim \mathsf{N}(0, \Omega_{ii} \mathrm{I}_{n \times n})$ and thus $v_i^\mathsf{T} \zeta_i | \zeta_i \sim \mathsf{N}(0, \Omega_{ii} \|\zeta_i\|^2)$. Further, on the event $\mathcal{E}_i$, $\|\zeta_i\|^2$ can be bounded as in equation 35.
Setting $t \equiv \sigma \sqrt{(128 s_0)/(c_* C_{\min} n)} \log p$, we get

$$\mathbb{P}\Big(\|R^{(2)}\|_\infty \geq \sigma \sqrt{\frac{128 s_0}{c_* C_{\min} n}} \log p; \mathcal{B} \Big) \leq 2p^{-1}. \tag{36}$$

• *Bounding term $R^{(3)}$:* In order to bound the last term, we first need to establish the following main lemma that bounds the distance between Lasso estimator and the solution of the perturbed problem. We refer to Section 4.3 for the proof of Lemma 4.3.

Lemma 4.3. (Perturbation bound) *Suppose that $\Sigma_{ii} \leq 1$, for $i \in [p]$. Set $\lambda = 8\sigma\sqrt{(\log p)/n}$ and let $\mathcal{B}(C_*) \equiv \tilde{\mathcal{B}}(n,p) \cap \mathcal{B}_\delta(n, (C_*+1)s_0, 3)$. The following holds true.*

$$\mathbb{P}\Big(\|\widehat{\theta}_{\sim i} - \widehat{\theta}^{\mathrm{p}}_{\sim i}\|_2 \geq C' \lambda; \mathcal{B}(C_*) \Big) \leq 2 \exp\Big(- \frac{c_* n}{s_0} \Big) + \exp\Big(- \frac{n}{1000} \Big), \tag{37}$$

where,

$$C' \equiv \frac{14(1+\rho)(1+\delta)\sqrt{C_{\max}}}{(1-\delta)^2 C_{\min}}, \qquad c_* \equiv \frac{1}{8}(1-\delta)^2 C_{\min},$$
$$C_* \equiv 16 \Big(\frac{1+\delta}{1-\delta} \Big)^2 \frac{C_{\max}}{C_{\min}}.$$

We are now ready to bound term $R^{(3)}$.

$$\begin{aligned}
|R_i^{(3)}| &\leq \frac{1}{\sqrt{n}} \|v^\mathsf{T} X_{\sim i}\|_\infty \|\widehat{\theta}^{\mathrm{p}}_{\sim i} - \widehat{\theta}_{\sim i}\|_1 \\
&\leq \frac{\sqrt{(C_*+1)s_0}}{\sqrt{n}} \|v^\mathsf{T} X_{\sim i}\|_\infty \|\widehat{\theta}^{\mathrm{p}}_{\sim i} - \widehat{\theta}_{\sim i}\|_2 \\
&\leq \sqrt{(C_*+1)s_0 n}\, |\Omega\widehat{\Sigma} - \mathrm{I}|_\infty \|\widehat{\theta}^{\mathrm{p}}_{\sim i} - \widehat{\theta}_{\sim i}\|_2,
\end{aligned}$$

where in the first inequality we used Proposition 4.7, which states that $\|\widehat{\theta}^{\mathrm{p}}_{\sim i}\|_0 \le C_* s_0$, under $\mathcal{B}$, with C_* given by equation (54). Therefore, by Lemma 4.3 and equation (32), we have

$$\mathbb{P}\Big(|R_i^{(3)}| \ge C''\sigma\sqrt{\frac{s_0}{n}}\log p; \mathcal{B}(C_*)\Big) \le 2\exp\Big(-\frac{c_* n}{s_0}\Big) + \exp\Big(-\frac{n}{1000}\Big) + 2p^{-2}\,,$$

with $C'' \equiv 8\sqrt{(C_*+1)}AC'$. Hence, by union bound over the p coordinates, we get

$$\mathbb{P}\Big(\|R^{(3)}\|_\infty \ge C''\sigma\sqrt{\frac{s_0}{n}}\log p; \mathcal{B}(C_*)\Big) \le 2p\exp\Big(-\frac{c_* n}{s_0}\Big) + p\exp\Big(-\frac{n}{1000}\Big) + 2p^{-1}\,. \tag{38}$$

We are now in position to prove the claim of Theorem 3.6.

Using equations (30) and (31), we have $\sqrt{n}(\widehat{\theta}^{\mathrm{d}} - \theta^*) = Z + R$, where $Z|X \sim \mathsf{N}(0, \sigma^2\Omega\widehat{\Sigma}\Omega)$ and $R = R^{(1)} + R^{(2)} + R^{(3)}$. Combining equations (33), (36) and (38), we get

$$\mathbb{P}\Big(\|R\|_\infty \ge C\sqrt{\frac{s_0}{n}}\log p; \mathcal{G}_n(A)\cap\mathcal{B}(C_*)\Big) \le 2p\exp\Big(-\frac{c_* n}{s_0}\Big) + p\exp\Big(-\frac{n}{1000}\Big) + 4p^{-1}\,, \tag{39}$$

where C is given by

$$C \equiv \sigma\Big(\frac{40A}{(1-\delta)^2 C_{\min}} + \sqrt{\frac{128}{c_* C_{\min}}} + 8\sqrt{(C_*+1)}AC'\Big)\,. \tag{40}$$

Further, for $n \ge \max(25\log p, c_1 C_* s_0 \log(p/s_0))$, we have

$$\begin{aligned}\mathbb{P}\Big((\mathcal{G}_n(A)\cap\mathcal{B}(C_*))^c\Big) &\le \mathbb{P}(\mathcal{G}_n(A)^c) + \mathbb{P}(\tilde{\mathcal{B}}(n,p)^c) + \mathbb{P}(\mathcal{B}_\delta(n,(C_*+1)s_0,3)^c)\\ &\le 2p^{-1} + 2p^{-1} + 2e^{-\delta^2 n} = 4p^{-1} + 2e^{-\delta^2 n}\,,\end{aligned} \tag{41}$$

where we used bound (32), Lemma 3.2 and Lemma 3.4.

The result follows from equations (39) and (41).

4.3 Proof of Lemma 4.3 (perturbation bound)

Lemma 4.4. *For all $\theta \in \mathbb{R}^{p-1}$ the following holds true.*

$$\frac{1}{2n}\|X_{\sim i}(\theta - \widehat{\theta}^{\mathrm{p}}_{\sim i})\|_2^2 \le \mathcal{L}_{y,X}(\theta_i^*, \theta) - \mathcal{L}_{y,X}(\theta_i^*, \widehat{\theta}^{\mathrm{p}}_{\sim i})\,. \tag{42}$$

Define

$$\begin{aligned}\mathcal{L}^+(\theta) &\equiv \min_{\theta_i\in\mathbb{R}} \mathcal{L}_{y,X}(\theta_i, \theta)\,,\\ \Delta(\theta) &\equiv \mathcal{L}_{y,X}(\theta_i^*, \theta) - \mathcal{L}^+(\theta)\,.\end{aligned}$$

Invoking Lemma 4.4, we obtain

$$\begin{aligned}\frac{1}{2n}\|X_{\sim i}(\widehat{\theta}_{\sim i} - \widehat{\theta}^{\mathrm{p}}_{\sim i})\|_2^2 &\le \mathcal{L}^+(\widehat{\theta}_{\sim i}) - \mathcal{L}^+(\widehat{\theta}^{\mathrm{p}}_{\sim i}) + \Delta(\widehat{\theta}_{\sim i}) - \Delta(\widehat{\theta}^{\mathrm{p}}_{\sim i})\\ &\le \Delta(\widehat{\theta}_{\sim i}) - \Delta(\widehat{\theta}^{\mathrm{p}}_{\sim i})\,,\end{aligned} \tag{43}$$

where in the last step we used the fact that $\widehat{\theta}_{\sim i}$ is the minimizer of $\mathcal{L}^+(\theta_{\sim i})$.[3]

Recall the definition of Huber function $\mathcal{H} : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}_+$.

$$\mathcal{H}(x;\alpha) = \begin{cases} \alpha|x| - \frac{\alpha^2}{2} & \text{if } |x| > \alpha\,, \\ \\ \frac{x^2}{2} & \text{if } |x| \le \alpha\,. \end{cases}$$

Lemma 4.5. *For $\theta \in \mathbb{R}^{p-1}$ define*

$$u(\theta) \equiv \frac{\tilde{x}_i^{\mathsf{T}}(w + X_{\sim i}(\theta^*_{\sim i} - \theta))}{\|\tilde{x}_i\|^2} \tag{44}$$

Also let $c_i \equiv \|\tilde{x}_i\|^2/n$. Then, the following holds true.

$$\Delta(\theta) = \mathsf{F}(u(\theta)) + \lambda|\theta^*_i|\,, \tag{45}$$

where

$$\mathsf{F}(u) = \frac{c_i}{2}u^2 - c_i\mathcal{H}(\theta^*_i + u; \lambda/c_i)\,. \tag{46}$$

Lemma 4.5 is proved in Appendix B.

Using equation(43) and by taylor expansion of $\Delta(\theta)$ around $\widehat{\theta}^{\mathrm{p}}_{\sim i}$ we obtain

$$\frac{1}{2n}\|X_{\sim i}(\widehat{\theta}_{\sim i} - \widehat{\theta}^{\mathrm{p}}_{\sim i})\|_2^2 \le \langle \nabla\Delta(\widehat{\theta}^{\mathrm{p}}_{\sim i}), \widehat{\theta}_{\sim i} - \widehat{\theta}^{\mathrm{p}}_{\sim i}\rangle + \frac{1}{2}\langle \widehat{\theta}_{\sim i} - \widehat{\theta}^{\mathrm{p}}_{\sim i}, \nabla^2\Delta(\bar{\theta})(\widehat{\theta}_{\sim i} - \widehat{\theta}^{\mathrm{p}}_{\sim i})\rangle\,, \tag{47}$$

for a vector $\bar{\theta}$ on the line segment between $\widehat{\theta}^{\mathrm{p}}_{\sim i}$ and $\widehat{\theta}_{\sim i}$. Using chain rule, gradient and Hessian of $\Delta(\theta)$ can be written as

$$\nabla\Delta(\theta) = -\mathsf{F}'(u(\theta))\frac{\tilde{x}_i^{\mathsf{T}}X_{\sim i}}{\|\tilde{x}_i\|^2}\,, \tag{48}$$

$$\nabla^2\Delta(\theta) = \mathsf{F}''(u(\theta))\frac{X_{\sim i}^{\mathsf{T}}\tilde{x}_i\tilde{x}_i^{\mathsf{T}}X_{\sim i}}{\|\tilde{x}_i\|^4}\,. \tag{49}$$

Since $\mathcal{H}(\cdot;\alpha)$ is convex in the first argument, we have $\mathsf{F}'' \le c_i$ uniformly. Denote by $\mathrm{P}_{\tilde{x}_i} \equiv \tilde{x}_i\tilde{x}_i^{\mathsf{T}}/\|\tilde{x}_i\|^2$, the projection on the direction of $\tilde{x}_i$. Then,

$$\nabla^2\Delta(\bar{\theta}) \preceq \frac{c_i}{\|\tilde{x}_i\|^2}X_{\sim i}^{\mathsf{T}}\mathrm{P}_{\tilde{x}_i}X_{\sim i} = \frac{1}{n}X_{\sim i}^{\mathsf{T}}\mathrm{P}_{\tilde{x}_i}X_{\sim i}\,. \tag{50}$$

Using equation (50) in bound (47), we get

$$\frac{1}{2n}\|\mathrm{P}^{\perp}_{\tilde{x}_i}X_{\sim i}(\widehat{\theta}_{\sim i} - \widehat{\theta}^{\mathrm{p}}_{\sim i})\|_2^2 \le \langle \nabla\Delta(\widehat{\theta}^{\mathrm{p}}_{\sim i}), \widehat{\theta}_{\sim i} - \widehat{\theta}^{\mathrm{p}}_{\sim i}\rangle\,. \tag{51}$$

Applying equation (48) and Cauchy-Schwartz inequality we get

$$\frac{1}{2n}\|\mathrm{P}^{\perp}_{\tilde{x}_i}X_{\sim i}(\widehat{\theta}_{\sim i} - \widehat{\theta}^{\mathrm{p}}_{\sim i})\|_2^2 \le |\mathsf{F}'(u(\widehat{\theta}^{\mathrm{p}}_{\sim i}))|\,\frac{\|X_{\sim i}(\widehat{\theta}_{\sim i} - \widehat{\theta}^{\mathrm{p}}_{\sim i})\|}{\|\tilde{x}_i\|}\,. \tag{52}$$

The following preposition upper bounds the gradient of F at the perturbed Lasso solution. We defer the proof of Proposition 4.6 to Appendix C.

[3]Note that $(\widehat{\theta}_i, \widehat{\theta}_{\sim i})$ is the minimizer of $\mathcal{L}_{y,X}(\theta_i, \theta_{\sim i})$.

Proposition 4.6. *For $i \in [p]$ define $\Sigma_{i|\sim i} \equiv \Sigma_{i,i} - \Sigma_{i,\sim i}(\Sigma_{\sim i,\sim i})^{-1}\Sigma_{\sim i,i}$. Let $\mathcal{B} \equiv \tilde{\mathcal{B}}(n,p) \cap \mathcal{B}_\delta(n, s_0, 3)$, where the events $\mathcal{B}_\delta(n, s_0, 3)$ and $\tilde{\mathcal{B}}(n,p)$ are given as per equations (15) and (16). The following holds true.*

$$\mathbb{P}\Big(|\mathsf{F}'(u(\widehat{\theta}^{\mathrm{p}}_{\sim i}))| \geq \frac{4+5\rho}{4}\lambda; \mathcal{B}\Big) \leq 2\exp\Big(-\frac{c_* n}{s_0 \Sigma_{i|\sim i}}\Big)\,.$$

where $c_* \equiv (1-\delta)^2 C_{\min}/8$.

We next upper bound the term $\|X_{\sim i}(\widehat{\theta}_{\sim i} - \widehat{\theta}^{\mathrm{p}}_{\sim i})\|$ that appears on the RHS of equation (52).

The following proposition states that the Lasso estimator is sparse. Its proof is given in Appendix D.

Proposition 4.7. *Consider the Lasso selector $\widehat{\theta}$ with $\lambda = 8\sigma\sqrt{\log p/n}$. On the event $\mathcal{B} \equiv \tilde{\mathcal{B}}(n,p) \cap \mathcal{B}_\delta(n, s_0, 3)$, the following holds:*

$$|\widehat{S}| \leq C_* s_0\,, \tag{53}$$

with

$$C_* \equiv \frac{16 C_{\max}}{(1-\delta)^2 C_{\min}}\,. \tag{54}$$

Corollary 4.8. *Set $\lambda = 8\sigma\sqrt{(\log p)/n}$. On the event $\mathcal{B}(C_*) \equiv \tilde{\mathcal{B}}(n,p) \cap \mathcal{B}_\delta(n, (C_*+1)s_0, 3)$, the following holds.*

$$\frac{1}{n}\|X_{\sim i}(\widehat{\theta}_{\sim i} - \widehat{\theta}^{\mathrm{p}}_{\sim i})\|^2 \leq (1+\delta)^2 C_{\max}\|\widehat{\theta}_{\sim i} - \widehat{\theta}^{\mathrm{p}}_{\sim i}\|^2\,. \tag{55}$$

Corollary 4.8 is proved in Appendix E.

We next lower bound $\|\tilde{x}_i\|_2$. Observe that the entries $\tilde{x}_{i\ell}^2 - 1$, $\ell \in [n]$, are zero-mean sub-exponential random variables. We obtain the following tail-bound inequality by applying Bernstein-type inequality for sub-exponential random variables. (See e.g. [JM14b, Equation (190)].)

$$\mathbb{P}\Big(\|\tilde{x}_i\| \leq \frac{\sqrt{n}}{5}\Big) \leq e^{-n/1000}\,. \tag{56}$$

Combining the results of Proposition (4.6) and equations (52), (55) and (56), we obtain that

$$\begin{aligned}
&\mathbb{P}\Big(\frac{1}{2n}\|\mathrm{P}^{\perp}_{\tilde{x}_i} X_{\sim i}(\widehat{\theta}_{\sim i} - \widehat{\theta}^{\mathrm{p}}_{\sim i})\|_2^2 \leq 7(1+\rho)(1+\delta)\sqrt{C_{\max}}\lambda\|\widehat{\theta}_{\sim i} - \widehat{\theta}^{\mathrm{p}}_{\sim i}\|; \mathcal{B}\Big) \\
&\leq 2\exp\Big(-\frac{c_* n}{s_0 \Sigma_{i|\sim i}}\Big) + \exp\Big(-\frac{n}{1000}\Big) \\
&\leq 2\exp\Big(-\frac{c_* n}{s_0}\Big) + \exp\Big(-\frac{n}{1000}\Big)\,,
\end{aligned} \tag{57}$$

where the last inequality follows from the assumption $\Sigma_{ii} \leq 1$ and noting that $\Sigma_{i|\sim i} \leq \Sigma_{ii}$, since $\Sigma_{\sim i,\sim i} \succeq 0$. The last step is to lower bound the LHS of equation (52). Write

$$\begin{aligned}
\mathrm{P}^{\perp}_{\tilde{x}_i} X_{\sim i}(\widehat{\theta}_{\sim i} - \widehat{\theta}^{\mathrm{p}}_{\sim i}) &= X_{\sim i}(\widehat{\theta}_{\sim i} - \widehat{\theta}^{\mathrm{p}}_{\sim i}) - \mathrm{P}_{\tilde{x}_i} X_{\sim i}(\widehat{\theta}_{\sim i} - \widehat{\theta}^{\mathrm{p}}_{\sim i}) \\
&= X_{\sim i}(\widehat{\theta}_{\sim i} - \widehat{\theta}^{\mathrm{p}}_{\sim i}) - \tilde{x}_i\Big\langle \frac{\tilde{x}_i}{\|\tilde{x}_i\|^2}, X_{\sim i}(\widehat{\theta}_{\sim i} - \widehat{\theta}^{\mathrm{p}}_{\sim i})\Big\rangle\,.
\end{aligned}$$

Define vector $\mu \in \mathbb{R}^p$ with

$$\mu_i \equiv -\Big\langle \frac{\tilde{x}_i}{\|\tilde{x}_i\|^2}, X_{\sim i}(\widehat{\theta}_{\sim i} - \widehat{\theta}^{\mathrm{p}}_{\sim i}) \Big\rangle\,, \qquad \mu_{\sim i} = \widehat{\theta}_{\sim i} - \widehat{\theta}^{\mathrm{p}}_{\sim i}\,.$$

Then $\mu \in \mathcal{C}((C_*+1)s_0, 3)$, by Proposition 4.7. Hence, on the event $\mathcal{B}(n, (C_*+1)s_0, 3)$, we have

$$\begin{aligned}\frac{1}{2n}\|\mathrm{P}^{\perp}_{\tilde{x}_i} X_{\sim i}(\widehat{\theta}_{\sim i} - \widehat{\theta}^{\mathrm{p}}_{\sim i})\|^2 &= \frac{1}{2n}\|X\mu\|^2 \\ &\geq \frac{1}{2}(1-\delta)^2 C_{\min}\|\mu\|^2 \\ &\geq \frac{1}{2}(1-\delta)^2 C_{\min}\|\widehat{\theta}_{\sim i} - \widehat{\theta}^{\mathrm{p}}_{\sim i}\|^2\,. \end{aligned} \tag{58}$$

Finally, note that $\mathcal{C}(s_0, 3) \subseteq \mathcal{C}((C_*+1)s_0, 3)$. Therefore, $\mathcal{B}_\delta(n, (C_*+1)s_0, 3) \subseteq \mathcal{B}_\delta(n, s_0, 3)$, by definition. Letting $\mathcal{B}(C_*) \equiv \tilde{\mathcal{B}}(n,p) \cap \mathcal{B}(n, (C_*+1)s_0, n)$, we have $\mathcal{B}(C_*) \subseteq \mathcal{B}$. Combining equations (57) and (58), we obtain

$$\mathbb{P}\Big(\|\widehat{\theta}_{\sim i} - \widehat{\theta}^{\mathrm{p}}_{\sim i}\| \geq \frac{14(1+\rho)(1+\delta)\sqrt{C_{\max}}}{(1-\delta)^2 C_{\min}}\lambda\,; \mathcal{B}(C_*)\Big) \leq 2\exp\Big(-\frac{c_* n}{s_0}\Big) + \exp\Big(-\frac{n}{1000}\Big)\,. \tag{59}$$

This completes the proof.

5 Proof of Theorem 3.10 (unknown covariance)

We decompose $\sqrt{n}(\widehat{\theta}^{\mathrm{d}} - \theta^*)$ into three terms:

$$\begin{aligned}\sqrt{n}(\widehat{\theta}^{\mathrm{d}} - \theta^*) &= \sqrt{n}(\widehat{\theta} - \theta^*) + \frac{1}{\sqrt{n}} M X^{\mathsf{T}}(y - X\widehat{\theta}) \\ &= \sqrt{n}(\mathrm{I} - M\widehat{\Sigma})(\widehat{\theta} - \theta^*) + \frac{1}{\sqrt{n}} M X^{\mathsf{T}} w \\ &= \underbrace{\sqrt{n}(\mathrm{I} - \Omega\widehat{\Sigma})(\widehat{\theta} - \theta^*)}_{I_1} + \underbrace{\sqrt{n}(\Omega - M)\widehat{\Sigma}(\widehat{\theta} - \theta^*)}_{I_2} + \underbrace{\frac{1}{\sqrt{n}} M X^{\mathsf{T}} w}_{I_3}\,. \end{aligned}$$

Note that the term I_1 is exactly the bias vector R of the de-biased estimator in case of known covariance (with $M = \Omega$). Therefore, by invoking the result of Theorem 3.6, we have

$$\mathbb{P}\Big(\|I_1\|_\infty \geq C\sqrt{\frac{s_0}{n}}\log p\Big) \leq 2pe^{-c_* n/s_0} + pe^{-n/1000} + 8p^{-1} + 2e^{-\delta^2 n}\,. \tag{60}$$

We next bound $\|I_2\|_\infty$.

$$\begin{aligned}\|I_2\|_\infty &= \sqrt{n}\Big\|(\Omega - M)\frac{1}{n}X^{\mathsf{T}}X(\widehat{\theta} - \theta^*)\Big\|_\infty \\ &\leq \sqrt{n}\max_{i\in[p]}\Big\|\frac{1}{\sqrt{n}}X(M-\Omega)e_i\Big\|_2\Big\|\frac{1}{\sqrt{n}}X(\widehat{\theta} - \theta^*)\Big\|_2 \end{aligned} \tag{61}$$

By applying [BvdG11, Theorem 6.1] (see also Corollary 4.2 in the present paper), on the event $\mathcal{B} \equiv \mathcal{B}_\delta(n, s_0, 3) \cap \tilde{\mathcal{B}}(n, p)$, we have for $\lambda \geq 8\sigma\sqrt{(\log p)}/n$,

$$\left\| \frac{1}{\sqrt{n}} X(\theta^* - \widehat{\theta}) \right\|_2 \leq \frac{2\lambda}{1-\delta} \sqrt{\frac{s_0}{C_{\min}}} \,. \tag{62}$$

To bound the other term, we recall definition of upper-RE condition.

Definition 5.1. *The matrix Γ satisfies an upper restricted eigenvalue condition with parameters $\alpha > 0$ and $\tau(n, p) > 0$ if*

$$v^{\mathsf{T}} \Gamma v \leq \alpha \|v\|_2^2 + \tau(n, p) \|v\|_1^2 \qquad \forall v \in \mathbb{R}^p \,.$$

Raskutti et al. [RWY10] showed that for Gaussian designs with population covariance Σ, the matrix $\widehat{\Sigma} \equiv (X^{\mathsf{T}} X)/n$ satisfies the upper-RE condition with $\alpha = 2\sigma_{\max}(\Sigma)$ and $\tau(n, p) = c(\log p)/n$, with probability at least $1 - c' \exp(-c'' n)$ for some constants $c, c', c'' > 0$.

Rewriting the above bound we obtain

$$\frac{1}{\sqrt{n}} \|X(M - \Omega) e_i\|_2 \leq \sqrt{2C_{\max}} \|(M - \Omega) e_i\|_2 + \sqrt{\frac{c \log p}{n}} \|(M - \Omega) e_i\|_1 \,. \tag{63}$$

As proved in [VdGBRD14, Theorem 2.4], we have the following bounds

$$\max_{i \in [p]} \|(M - \Omega) e_i\|_1 \lesssim s_\Omega \sqrt{\frac{\log p}{n}} \,, \tag{64}$$

$$\max_{i \in [p]} \|(M - \Omega) e_i\|_2 \lesssim \sqrt{\frac{s_\Omega \log p}{n}} \,, \tag{65}$$

Using equations (64), (65) in (63) and we get (recalling that $n \geq c s_0 \log p$)

$$\max_{i \in [p]} \left\| \frac{1}{\sqrt{n}} X(M - \Omega) e_i \right\|_2 \lesssim \sqrt{\frac{s_\Omega \log p}{n}} \,. \tag{66}$$

Combining equations (62) and (66) in (61), we get that on event $\mathcal{B}$,

$$\|I_2\|_\infty \lesssim \sigma\sigma \sqrt{\frac{s_\Omega s_0}{n}} \log p \,, \tag{67}$$

with probability at least $1 - c' \exp(-c'' n)$.

Finally, note that

$$I_3 | X \sim \mathsf{N}(0, \sigma^2 M \widehat{\Sigma} M^{\mathsf{T}}) \,.$$

The result follows by letting $Z \equiv I_3$ and $R \equiv I_1 + I_2$.

Acknowledgements

The authors would like to thank Jason D. Lee and Cun-Hui Zhang for stimulating discussions, and Zhao Ren for valuable comments to improve the presentation. A.M. was partially supported by NSF grants CCF-1319979 and DMS-1106627 and the AFOSR grant FA9550-13-1-0036.

A Proof of Lemma 4.4

For θ we have

$$\mathcal{L}_{y,X}(\theta_i^*, \theta) = \frac{1}{2n}\|y - \tilde{x}_i\theta_i^* - X_{\sim i}\theta\|^2 + \lambda\|\theta\|_1 + \lambda|\theta_i^*|$$

Let $\tilde{y} \equiv y - \tilde{x}_i\theta_i^*$. We then have

$$\begin{aligned}\mathcal{L}_{y,X}(\theta_i^*, \theta) &= \frac{1}{2n}\|\tilde{y} - X_{\sim i}\widehat{\theta}^{\mathrm{p}}_{\sim i} - X_{\sim i}(\theta - \widehat{\theta}^{\mathrm{p}}_{\sim i})\|^2 + \lambda\|\theta\|_1 + \lambda|\theta_i^*| \\ &= \mathcal{L}_{y,X}(\theta_i^*, \widehat{\theta}^{\mathrm{p}}_{\sim i}) + \frac{1}{2n}\|X_{\sim i}(\theta - \widehat{\theta}^{\mathrm{p}}_{\sim i})\|^2 - \frac{1}{n}\langle \tilde{y} - X_{\sim i}\widehat{\theta}^{\mathrm{p}}_{\sim i}, X_{\sim i}(\theta - \widehat{\theta}^{\mathrm{p}}_{\sim i})\rangle \\ &\quad + \lambda\|\theta\|_1 - \lambda\|\widehat{\theta}^{\mathrm{p}}_{\sim i}\|_1 \end{aligned} \tag{68}$$

Since $\widehat{\theta}^{\mathrm{p}}_{\sim i}$ is the minimizer of $\mathcal{L}_{y,X}(\theta_i^*, \theta)$ by KKT condition we have

$$\frac{1}{n}X_{\sim i}^{\mathsf{T}}(\tilde{y} - X_{\sim i}\widehat{\theta}^{\mathrm{p}}_{\sim i}) = \lambda\xi, \qquad \xi \in \partial\|\widehat{\theta}^{\mathrm{p}}_{\sim i}\|_1 \,. \tag{69}$$

Applying equation (69) in equation (68) we get

$$\begin{aligned}\mathcal{L}_{y,X}(\theta_i^*, \theta) - \mathcal{L}_{y,X}(\theta_i^*, \widehat{\theta}^{\mathrm{p}}_{\sim i}) &= \frac{1}{2n}\|X_{\sim i}(\theta - \widehat{\theta}^{\mathrm{p}}_{\sim i})\|^2 + \lambda\Big(\|\theta\|_1 - \|\widehat{\theta}^{\mathrm{p}}_{\sim i}\|_1 - \langle \xi, \theta - \widehat{\theta}^{\mathrm{p}}_{\sim i}\rangle\Big) \\ &\geq \frac{1}{2n}\|X_{\sim i}(\theta - \widehat{\theta}^{\mathrm{p}}_{\sim i})\|^2 \,,\end{aligned}$$

where the last step follows from the definition of a subgradient.

B Proof of Lemma 4.5

To lighten the notation, we drop the subscripts y, X in $\mathcal{L}_{y,X}(\cdot)$. Recall that $\Delta(\theta) \equiv \mathcal{L}_{y,X}(\theta_i^*, \theta) - \mathcal{L}^+(\theta)$. We start by expanding $\mathcal{L}(\theta_i, \theta)$.

$$\mathcal{L}(\theta_i, \theta) = \frac{1}{2n}\|y - \tilde{x}_i\theta_i - X_{\sim i}\theta\|_2^2 + \lambda|\theta_i^*| + \lambda\|\theta\|_1 \,.$$

Plugging in $y = \tilde{x}_i\theta_i^* + X_{\sim i}\theta_{\sim i}^* + w$ and rearranging the terms, we obtain

$$\begin{aligned}\mathcal{L}(\theta_i, \theta) =& \frac{1}{2n}\|w + X_{\sim i}(\theta_{\sim i}^* - \theta)\|_2^2 + \frac{1}{n}\langle \theta_i^* - \theta_i, \tilde{x}_i^{\mathsf{T}}(w + X_{\sim i}(\theta_{\sim i}^* - \theta))\rangle \\ &+ \frac{1}{2n}\|\tilde{x}_i\|^2(\theta_i^* - \theta_i)^2 + \lambda|\theta_i| + \lambda\|\theta\|_1 \,.\end{aligned} \tag{70}$$

Therefore,

$$\mathcal{L}(\theta_i^*, \theta) = \frac{1}{2n}\|w + X_{\sim i}(\theta_{\sim i}^* - \theta)\|_2^2 + +\lambda|\theta_i^*| + \lambda\|\theta\|_1 \,. \tag{71}$$

Combining equations (70) and (71), we rewrite $\mathcal{L}(\theta_i, \theta)$ as

$$\begin{aligned}\mathcal{L}(\theta_i, \theta) =& \mathcal{L}(\theta_i^*, \theta) + \frac{1}{n}\langle \theta_i^* - \theta_i, \tilde{x}_i^\mathsf{T}(w + X_{\sim i}(\theta_{\sim i}^* - \theta))\rangle \\ &+ \frac{1}{2n}\|\tilde{x}_i\|^2(\theta_i^* - \theta_i)^2 + \lambda|\theta_i| - \lambda|\theta_i^*| \\ =& \lambda|\theta_i| + \frac{1}{2n}\|\tilde{x}_i\|^2\Big(\theta_i - \theta_i^* - \frac{\tilde{x}_i^\mathsf{T}}{\|\tilde{x}_i\|^2}(w + X_{\sim i}(\theta_{\sim i}^* - \theta))\Big)^2 \\ &- \frac{1}{2n\|\tilde{x}_i\|^2}\Big(\tilde{x}_i^\mathsf{T}(w + X_{\sim i}(\theta_{\sim i}^* - \theta))\Big)^2 + \mathcal{L}(\theta_i^*, \theta) - \lambda|\theta_i^*|\,. \end{aligned} \tag{72}$$

Writing expression (72) in terms of $c_i \equiv \|\tilde{x}_i\|^2/n$ and $u(\theta)$, given by (44), we get

$$\mathcal{L}(\theta_i, \theta) = \lambda|\theta_i| + \frac{c_i}{2}(\theta_i - \theta_i^* - u(\theta))^2 - \frac{c_i}{2}u(\theta)^2 + \mathcal{L}(\theta_i^*, \theta) - \lambda|\theta_i^*|\,. \tag{73}$$

Let $\theta_i^{\mathrm{opt}} \equiv \arg\min_{\theta_i} \mathcal{L}(\theta_i, \theta)$. It is simple to see that $\theta_i^{\mathrm{opt}} = \eta(\theta_i^* + u(\theta); \lambda/c_i)$, where $\eta(x;\alpha)$ is the soft-thresholding function given by

$$\eta(x;\alpha) = \begin{cases} x - \alpha & x \geq \alpha\,, \\ 0 & |x| \leq \alpha\,, \\ x + \alpha & x \leq -\alpha\,. \end{cases}$$

By substituting for θ_i^{opt} in equation (73) and after some algebraic manipulations, we obtain

$$\mathcal{L}(\theta_i^{\mathrm{opt}}, \theta) = c_i\mathcal{H}(\theta_i^* + u(\theta); \lambda/c_i) - \frac{c_i}{2}u(\theta)^2 + \mathcal{L}(\theta_i^*, \theta) - \lambda|\theta_i^*|\,,$$

where $\mathcal{H}(x;\alpha)$ is the Huber function. By definition, $\mathcal{L}^+(\theta) = \mathcal{L}(\theta_i^{\mathrm{opt}}, \theta)$ and thus

$$\begin{aligned}\Delta(\theta) \equiv \mathcal{L}(\theta_i^*, \theta) - \mathcal{L}^+(\theta) &= \frac{c_i}{2}u(\theta)^2 - c_i\mathcal{H}(\theta_i^* + u(\theta); \lambda/c_i) + \lambda|\theta_i^*| \\ &= \mathsf{F}(u(\theta)) + \lambda|\theta_i^*|\,,\end{aligned}$$

where $\mathsf{F}(u)$ is given by equation (46).

C Proof of Preposition 4.6

Write

$$\mathsf{F}'(u) = c_i u - c_i\mathcal{H}'(\theta_i^* + u; \lambda/c_i)\,.$$

We note that $\mathcal{H}'(x;\alpha) = x - \eta(x;\alpha)$ and hence $|\mathcal{H}'(x;\alpha)| \leq \alpha$. Therefore,

$$|\mathsf{F}'(u)| \leq c_i|u| + c_i(\lambda/c_i) = \lambda + c_i|u|\,. \tag{74}$$

In the following we bound $c_i|u(\widehat{\theta^{\mathrm{p}}}_{\sim i})|$.

For $i \in [p]$ define

$$\Sigma_{i|\sim i} \equiv \Sigma_{i,i} - \Sigma_{i,\sim i}(\Sigma_{\sim i,\sim i})^{-1}\Sigma_{\sim i,i}\,.$$

Since $\tilde{x}_i$ and $X_{\sim i}$ are jointly Gaussian, we have

$$\tilde{x}_i = X_{\sim i}(\Sigma_{\sim i,\sim i})^{-1}\Sigma_{\sim i,i} + \Sigma_{i|\sim i}^{1/2} z\,, \tag{75}$$

where $z \in \mathbb{R}^n$ is independent of $X_{\sim i}$ with i.i.d standard normal coordinates. Further,

Recalling the definition of $c_i \equiv \|\tilde{x}_i\|/n$ and $u(\theta)$, given by equation (44), we write $c_i|u(\widehat{\theta}^{\mathrm{p}}_{\sim i})|$ as

$$\begin{aligned} c_i|u(\widehat{\theta}^{\mathrm{p}}_{\sim i})| &= \frac{1}{n}\Big|\tilde{x}_i^{\mathsf{T}}(w + X_{\sim i}(\theta^*_{\sim i} - \widehat{\theta}^{\mathrm{p}}_{\sim i}))\Big| \\ &\le \frac{1}{n}|\tilde{x}_i^{\mathsf{T}} w| + \frac{1}{n}\Sigma_{i|\sim i}^{1/2}\Big|z^{\mathsf{T}} X_{\sim i}(\theta^*_{\sim i} - \widehat{\theta}^{\mathrm{p}}_{\sim i})\Big| + \frac{1}{n}\Big|\Sigma_{i,\sim i}(\Sigma_{\sim i,\sim i})^{-1} X_{\sim i}^{\mathsf{T}} X_{\sim i}(\theta^*_{\sim i} - \widehat{\theta}^{\mathrm{p}}_{\sim i})\Big| \\ &\le \frac{1}{n}|\tilde{x}_i^{\mathsf{T}} w| + \frac{1}{n}\Sigma_{i|\sim i}^{1/2}\Big|z^{\mathsf{T}} X_{\sim i}(\theta^*_{\sim i} - \widehat{\theta}^{\mathrm{p}}_{\sim i})\Big| + \frac{1}{n}\|\Sigma_{i,\sim i}(\Sigma_{\sim i,\sim i})^{-1}\|_1 \|X_{\sim i}^{\mathsf{T}} X_{\sim i}(\theta^*_{\sim i} - \widehat{\theta}^{\mathrm{p}}_{\sim i})\|_\infty\,. \end{aligned} \tag{76}$$

The first inequality here follows from equation (75).

In the following we bound each term on the RHS of equation (76) individually.

On the event $\tilde{\mathcal{B}}(n,p)$, defined by equation (16), we have

$$\frac{1}{n}\|\tilde{x}_i^{\mathsf{T}} w\| \le \frac{1}{n}\|X^{\mathsf{T}} w\|_\infty \le 2\sigma\sqrt{\frac{\log p}{n}} = \frac{\lambda}{4}\,. \tag{77}$$

We use Corollary 4.2 to bound the second term of expression (76). We recall the event $\mathcal{B}_\delta(n, s_0, 3)$, given by equation (15) and let $\mathcal{B} \equiv \mathcal{B}_\delta(n, s_0, 3) \cap \tilde{\mathcal{B}}(n,p)$. Further, recall the notation $\zeta_i \equiv \|X_{\sim i}(\theta^*_{\sim i} - \widehat{\theta}^{\mathrm{p}}_{\sim i})\|/\sqrt{n}$ and the event $\mathcal{E}_i$ defined by equation (35). We write

$$\begin{aligned} \mathbb{P}\Big(\frac{1}{\sqrt{n}}\Sigma_{i|\sim i}^{1/2}\,|z^{\mathsf{T}}\zeta_i| \ge \lambda; \mathcal{B}\Big) &\le \mathbb{P}\Big(\frac{1}{\sqrt{n}}\Sigma_{i|\sim i}^{1/2}\,|z^{\mathsf{T}}\zeta_i| \ge \lambda; \mathcal{E}_i\Big) \\ &= \mathbb{E}\Big\{\mathbb{I}\Big(\frac{1}{\sqrt{n}}\Sigma_{i|\sim i}^{1/2}\,|z^{\mathsf{T}}\zeta_i| \ge \lambda\Big)\cdot \mathbb{I}(\mathcal{E}_i)\Big\} \\ &\le 2\mathbb{E}\Big(\exp\Big[-\frac{n\lambda^2}{2\Sigma_{i|\sim i}\|\zeta_i\|^2}\Big]\cdot \mathbb{I}(\mathcal{E}_i)\Big) \\ &\le 2\exp\Big(-\frac{c_* n}{s_0\Sigma_{i|\sim i}}\Big)\,, \end{aligned} \tag{78}$$

with $c_* \equiv (1-\delta)^2 C_{\min}/8$. Here, the penultimate inequality follows from Fubini's theorem where we first integrate w.r.t z and then w.r.t ζ_i. Note that z and ζ_i are independent. Therefore, $z^{\mathsf{T}}\zeta_i|\zeta_i \sim \mathsf{N}(0, \|\zeta_i\|^2)$. In the last step, we applied Corollary 4.2.

We next bound the third term on the RHS of equation (76). Note that the KKT conditions for optimization (29) reads

$$\frac{1}{n}X_{\sim i}^{\mathsf{T}}(w + X_{\sim i}(\theta^*_{\sim i} - \widehat{\theta}^{\mathrm{p}}_{\sim i})) = \lambda\xi\,, \tag{79}$$

for $\xi \in \partial\|\widehat{\theta}^{\mathrm{p}}_{\sim i}\|_1$. To lighten the notation, let

$$\nu_i \equiv \frac{1}{n}X_{\sim i}^{\mathsf{T}} X_{\sim i}(\theta^*_{\sim i} - \widehat{\theta}^{\mathrm{p}}_{\sim i})\,.$$

We know by equation (79),

$$\|\nu_i\|_\infty \le \frac{1}{n}\|X_{\sim i}^{\mathsf{T}} w\|_\infty + \lambda\|\xi\|_\infty\,.$$

On the event $\tilde{\mathcal{B}}(n,p)$ we have

$$\frac{1}{n}\|X_{\sim i}^{\mathsf{T}} w\|_\infty \le 2\sigma\sqrt{\frac{\log p}{n}} = \frac{\lambda}{4}\,.$$

Combining the above two inequalities we obtain

$$\|\nu\|_\infty \le 5\lambda/4\,. \tag{80}$$

We next employ Condition *(iii)* to bound $\|\Sigma_{i,\sim i}(\Sigma_{\sim i,\sim i})^{-1}\|_1$. By writing Σ^{-1} in terms of Schur complement, we have

$$\|\Sigma^{-1}e_i\|_1 = \Sigma^{-1}_{i|\sim i}\Big[1 + \|\Sigma_{i,\sim i}(\Sigma_{\sim i,\sim i})^{-1}\|_1\Big]\,.$$

By Condition *(iii)*, $\|\Sigma^{-1}e_i\|_1 \le \rho$. Further, by Condition *(i)*, $\Sigma_{i|\sim i} \le \Sigma_{i,i} \le 1$. Hence, we get

$$\|\Sigma_{i,\sim i}(\Sigma_{\sim i,\sim i})^{-1}\|_1 \le \rho - 1. \tag{81}$$

Using equations (77) to (81), we bound the RHS of equation (76) as follows. Under the event $\mathcal{B}$,

$$c_i|u(\widehat{\theta}^{\mathrm{p}}_{\sim i})| \le \frac{5\lambda}{4}\rho\,.$$

Applying the above bound to equation (74), we get the desired result.

D Proof of Proposition 4.7

This proposition is an improved version of Theorem 7.2 in [BRT09].

We first recall the definition of *restricted eigenvalues* as given by:

$$\phi_{\max}(k) \equiv \max_{1\le\|v\|_0\le k} \frac{\langle v, \widehat{\Sigma} v\rangle}{\|v\|_2^2}\,.$$

Clearly, $\phi_{\max}(k)$ is an increasing function of k.

Employing [Ver12, Remark 5.4], for any $1 \le k \le n$ and a fixed subset $J \subset [p]$ with $|J| = k$, we have

$$\mathbb{P}\Big(\sigma_{\max}(\widehat{\Sigma}_{J,J}) \ge C_{\max} + C\sqrt{\frac{k}{n}} + \frac{t}{\sqrt{n}}\Big) \le 2e^{-ct^2}\,,$$

for $t \ge 0$, where C and c depend only on $C_{\max}$. Therefore, by union bound over all possible subsets $J \subseteq [p]$ we obtain

$$\mathbb{P}\Big(\phi_{\max}(k) \ge C_{\max} + C\sqrt{\frac{k}{n}} + \frac{t}{\sqrt{n}}\Big) \le 2\binom{p}{k}e^{-ct^2} \le 2e^{-ct^2 + k\log p + k}\,, \tag{82}$$

for $t \ge 0$.

Let $\widehat{S} \equiv \mathrm{supp}(\widehat{\theta})$. Recall that the stationarity condition for the Lasso cost function reads $X^{\mathsf{T}}(y - X\widehat{\theta}) = n\lambda\, v(\widehat{\theta})$, where $v(\widehat{\theta}) \in \partial\|\widehat{\theta}\|_1$. Equivalently,

$$\frac{1}{n}X^{\mathsf{T}}X(\theta^* - \widehat{\theta}) = \lambda\, v(\widehat{\theta}) - \frac{1}{n}X^{\mathsf{T}}w\,.$$

On the event $\tilde{\mathcal{B}}(n,p)$, we have $\|X^{\mathsf{T}}w\|_\infty \le n\lambda/4$. Thus for all $i \in \widehat{S}$

$$\left|\frac{1}{n}[X^{\mathsf{T}}X(\theta^* - \widehat{\theta})]_i\right| \ge \frac{\lambda}{2}\,.$$

Squaring and summing the last identity over $i \in \widehat{S}$, we obtain that, for $h \equiv n^{-1/2}X(\theta^* - \widehat{\theta})$,

$$\frac{\lambda^2}{4}|\widehat{S}| \le \frac{1}{n}\sum_{i\in\widehat{S}}(e_i^{\mathsf{T}}X^{\mathsf{T}}h)^2 = \langle h, \frac{1}{n}X_{\widehat{S}}X_{\widehat{S}}^{\mathsf{T}}\,h\rangle \le \|\widehat{\Sigma}_{\widehat{S},\widehat{S}}\|_2^2\|h\|^2 \ \le \phi_{\max}(|\widehat{S}|)\|h\|_2^2\,. \tag{83}$$

By a similar argument as in Corollary 4.2, on the event $\mathcal{B} \equiv \tilde{\mathcal{B}}(n,p) \cap \mathcal{B}(n,s_0,3)$ we have

$$\|h\|_2^2 \le \frac{4\lambda^2 s_0}{(1-\delta)^2 C_{\min}}\,.$$

Thus,

$$|\widehat{S}| \le \frac{16\phi_{\max}(\widehat{S})}{(1-\delta)^2 C_{\min}}s_0\,. \tag{84}$$

Note that $|\widehat{S}| \le n$ by the fact that the columns of X are in generic positions. Using monotonicity property of $\phi_{\max}(\cdot)$, we have $\phi_{\max}(|\widehat{S}|) \le \phi_{\max}(n)$. Invoking equation (82) with $k = n$, we have $\phi_{\max}(n) \le c_1\sqrt{\log p}$ with high probability for some constant c_1.

Hence, by equation (84)

$$|\widehat{S}| \le \tilde{C}s_0\sqrt{\log p}\,, \qquad \tilde{C} \equiv \frac{16c_1}{(1-\delta)^2 C_{\min}}\,. \tag{85}$$

Now, we use this bound on $|\widehat{S}|$ along with equation (84) to get a better bound on $|\widehat{S}|$. Again by using the fact that $\phi_{\max}(k)$ is a non-decreasing function of k, we have

$$\phi_{\max}(|\widehat{S}|) \le \phi_{\max}(\tilde{C}s_0\sqrt{\log p}) \le C_{\max}\,, \tag{86}$$

with high probability where we used the assumption $n \gg s_0(\log p)^2$. Using this bound in equation (84), we get

$$|\widehat{S}| \le \frac{16C_{\max}}{(1-\delta)^2 C_{\min}}s_0\,.$$

The result follows.

E Proof of Corollary 4.8

Note that $\widehat{\theta}^{\mathrm{p}}_{\sim i}$ is the Lasso estimators corresponding to $(\tilde{y}, X_{\sim i})$, according to equation (29). As a corollary of Proposition 4.7, on event $\mathcal{B}$, $\|\widehat{\theta}^{\mathrm{p}}_{\sim i}\|_0 \le C_* s_0$, with $C_* \equiv (16C_{\max}/C_{\min})(1-\delta)^{-2}$. Also, $\|\widehat{\theta}_{\sim i}\|_0 \le s_0$. Therefore,$(0, \widehat{\theta}_{\sim i} - \widehat{\theta}^{\mathrm{p}}_{\sim i}) \in \mathcal{C}((C_*+1)s_0, 3)$ and, by definition, on event $\mathcal{B}_\delta(n, (C_*+1)s_0, 3)$, the claim holds true.

F Sample splitting techniques

In this appendix, we discuss how sample splitting can be used to modify the de-biased estimator as to go around the sparsity barrier at $s_0 = o(\sqrt{n}/\log p)$. This provides an alternative to the more careful analysis carried out in the main body of the paper, that we discuss for the sake of simplicity. As mentioned in the introduction, sample splitting has its own drawbacks, most notably the dependence of the results on the random data split, and the sub-optimal use of all the samples.

For the sake of notational simplicity we assume here that the number of samples is $2n$ and is randomly split in two batches of size n: (x_1, y_1), ..., (x_n, y_n), and $(\overline{x}_1, \overline{y}_1)$,..., $(\overline{x}_n, \overline{y}_n)$. Note that the change of notation only amounts to a constant multiplicative factor in the sample size, which is of no concern to us. In vector notation, these batches are denoted as (y, X) and $(\overline{y}, \overline{X})$. We then proceed as follows:

1. We use the second batch to compute the Lasso estimator, namely
$$\widehat{\theta}(\overline{y}, \overline{X}; \lambda) \equiv \arg\max_{\theta \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|\overline{y} - \overline{X}\theta\|_2^2 + \lambda \|\theta\|_1 \right\} . \tag{87}$$
2. We use the first batch to compute the debiasing matrix M, e.g. using the node-wise Lasso as in Section 3.3.
3. We use the first batch to implement the de-biasing, namely
$$\widehat{\theta}^{\text{split}} = \widehat{\theta}(\overline{y}, \overline{X}) + \frac{1}{n} M X^{\mathsf{T}} \big(y - X\widehat{\theta}(\overline{y}, \overline{X})\big) . \tag{88}$$

The main remark is that, thanks to the splitting, X is statistically independent from $\widehat{\theta}$, which greatly simplifies the analysis. Notice that we did not use the responses in y.

For the sake of simplicity, we shall analyze this procedure in the case in which the precision matrix Ω is known, and we hence set $M = \Omega$. The generalization to M constructed via the node-wise Lasso is straightforward as in the proof of Theorem 3.10.

The next statement implies that, for sparsity level $s_0 = o(n/(\log p)^2)$,the sample splitting de-biased estimator is asymptotically Gaussian.

Proposition F.1. *Consider the linear model (2) where X has independent Gaussian rows, with zero mean and covariance Σ. Suppose that Σ satisfies the technical conditions of Theorem 3.6*

Let $\widehat{\theta}$ be the Lasso estimator defined by (3) with $\lambda = 8\sigma\sqrt{(\log p)/n}$. Further, let $\widehat{\theta}^{\text{split}}$ be the modified (sample-splitting) de-biased estimator defined in Eq. (88) with $M = \Omega \equiv \Sigma^{-1}$. Then, there exist constants c, C depending solely on $C_{\min}, C_{\max}, \delta$ and ρ, such that, for $n \geq c \max(\log p, s_0 \log(p/s_0))$ the following holds true:

$$\sqrt{n}(\widehat{\theta}^{\text{d}} - \theta^*) = Z + R , \qquad Z|X \sim \mathsf{N}(0, \sigma^2 \Omega \widehat{\Sigma} \Omega) , \tag{89}$$

$$\lim_{n\to\infty} \mathbb{P}\Big(\|R\|_\infty \geq C \sqrt{\frac{s_0}{n}} \log p \Big) = 0 . \tag{90}$$

Proof. Proceeding as in the proof of Theorem 3.6, it is sufficient to bound the bias term of $\sqrt{n}(\widehat{\theta}^{\text{split}} - \theta^*)$, which is given by (cf. (8))

$$R \equiv \sqrt{n}(\Omega\widehat{\Sigma} - \mathrm{I})(\theta^* - \widehat{\theta}) . \tag{91}$$

To lighten the notation, let $u = \theta^* - \widehat{\theta}$. Expanding R we get

$$R = \sqrt{n}(\Omega\widehat{\Sigma} - \mathrm{I})u = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\Omega x_i x_i^{\mathsf{T}} - \mathrm{I})u\,. \tag{92}$$

To control $\|R\|_\infty$, we bound each component R_j individually. Let e_j be the j-th element of the standard basis with one at the j-th position and zero everywhere else. We write

$$R_j = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(e_j^{\mathsf{T}}\Omega x_i)(x_i^{\mathsf{T}}u) - u_j\,.$$

Let $Z_i \equiv (e_j^{\mathsf{T}}\Omega x_i)(x_i^{\mathsf{T}}u) - u_j$. The variables Z_i are independent since u is independent of X because of data splitting and the rows x_i are also independent. Furthermore, $\mathbb{E}(Z_i) = e_j^{\mathsf{T}}\Omega\Sigma u - u_j = 0$. We let $\|\cdot\|_{\psi_1}$ and $\|\cdot\|_{\psi_2}$ respectively denote the sub-exponential and sub-gaussian norms. As shown in [Ver12, Remark 5.18],

$$\|Z_i\|_{\psi_1} \le 2\|(e_j^{\mathsf{T}}\Omega x_i)(x_i^{\mathsf{T}}u)\|_{\psi_1}\,.$$

In addition, for any two random variables v and w, we have $\|vw\|_{\psi_1} \le 2\|v\|_{\psi_2}\|w\|_{\psi_2}$. Hence,

$$\begin{aligned}\|(e_j^{\mathsf{T}}\Omega x_i)(x_i^{\mathsf{T}}u)\|_{\psi_1} &\le 2\|e_j^{\mathsf{T}}\Omega x_i\|_{\psi_2}\|x_i^{\mathsf{T}}u\|_{\psi_2}\\ &= 2\|e_j^{\mathsf{T}}\Omega^{1/2}\|_2\|\Omega^{1/2}x_i\|_{\psi_2}^2\|\Omega^{-1/2}u\|_2\\ &\le 2\sqrt{C_{\max}/C_{\min}}\,\|\Omega^{1/2}x_i\|_{\psi_2}^2\|u\|_2\,.\end{aligned}$$

Given that $\Omega^{1/2}x_i \sim \mathsf{N}(0,\mathrm{I})$, we get $\|\Omega^{1/2}x_i\|_{\psi_2} = 1$. Hence, $\max_i\|Z_i\|_{\psi_1} \le C\|u\|_2$ with $C \equiv 4\sqrt{C_{\max}/C_{\min}}$. Applying Bernstein-type inequality [Ver12, Proposition 5.16], for every $t \ge 0$, we have

$$\mathbb{P}\Big\{\Big|\sum_{i=1}^{n}\frac{1}{\sqrt{n}}Z_i\Big| \ge t\Big\} \le 2\exp\Big[-c\min\Big(\frac{t^2}{C^2\|u\|_2^2}, \frac{t\sqrt{n}}{C\|u\|_2}\Big)\Big]\,, \tag{93}$$

where $c > 0$ is an absolute constant. Observe that on the event $\mathcal{B} \equiv \mathcal{B}_\delta(n, s_0, 3) \cap \tilde{\mathcal{B}}(n,p)$[4], we have

$$\|u\|_2^2 = \|\theta^* - \widehat{\theta}\|_2^2 \lesssim s_0\lambda^2\,.$$

Therefore, by using tail bound (93) and applying union bound over the p entries of R, we get (for $n \ge c\log p$ with c a suitable constant)

$$\|R\|_\infty \lesssim \sqrt{\frac{s_0}{n}}\log p\,,$$

with high probability. □

[4]See Section 3.2 for definition of $\mathcal{B}_\delta(n, s_0, 3)$ and $\tilde{\mathcal{B}}(n,p)$

References

[BC14] Rina Foygel Barber and Emmanuel Candes, *Controlling the false discovery rate via knockoffs*, arXiv:1404.5609 (2014). 7

[BEM13] Mohsen Bayati, Murat A Erdogdu, and Andrea Montanari, *Estimating lasso risk and noise level*, Advances in Neural Information Processing Systems, 2013, pp. 944–952. 7

[BLM15] Mohsen Bayati, Marc Lelarge, and Andrea Montanari, *Universality in polytope phase transitions and message passing algorithms*, The Annals of Applied Probability **25** (2015), no. 2, 753–822. 7

[BM11] M. Bayati and A. Montanari, *The dynamics of message passing on dense graphs, with applications to compressed sensing*, IEEE Trans. on Inform. Theory **57** (2011), 764–785. 7

[BM12] ———, *The LASSO risk for gaussian matrices*, IEEE Trans. on Inform. Theory **58** (2012), 1997–2017. 7

[BRT09] P. J. Bickel, Y. Ritov, and A. B. Tsybakov, *Simultaneous analysis of Lasso and Dantzig selector*, Amer. J. of Mathematics **37** (2009), 1705–1732. 2, 3, 9, 26

[Büh13] Peter Bühlmann, *Statistical significance in high-dimensional linear models*, Bernoulli **19** (2013), no. 4, 1212–1242. 2, 7

[BvdG11] Peter Bühlmann and Sara van de Geer, *Statistics for high-dimensional data*, Springer-Verlag, 2011. 2, 3, 9, 10, 16, 22

[BvdG15] Peter Bühlmann and Sara van de Geer, *High-dimensional inference in misspecified linear models*, arXiv:1503.06426 (2015). 8

[CD95] S.S. Chen and D.L. Donoho, *Examples of basis pursuit*, Proceedings of Wavelet Applications in Signal and Image Processing III (San Diego, CA), 1995. 2

[CG15] T Tony Cai and Zijian Guo, *Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity*, arXiv:1506.05539 (2015). 7

[CHS15] Victor Chernozhukov, Christian Hansen, and Martin Spindler, *Valid post-selection and post-regularization inference: An elementary, general approach*, arXiv:1501.03430 (2015). 7

[CRT06] E. Candes, J. K. Romberg, and T. Tao, *Stable signal recovery from incomplete and inaccurate measurements*, Communications on Pure and Applied Mathematics **59** (2006), 1207–1223. 5

[CRZZ15] Mengjie Chen, Zhao Ren, Hongyu Zhao, and Harrison Zhou, *Asymptotically normal and efficient estimation of covariate-adjusted gaussian graphical model*, Journal of the American Statistical Association (2015), no. just-accepted, 00–00. 5

[CSZ06] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien, *Semi-Supervised Learning*, MIT Press, Cambridge, 2006. 6

[CT07] E. Candés and T. Tao, *The Dantzig selector: statistical estimation when p is much larger than n*, Annals of Statistics **35** (2007), 2313–2351. 2, 3, 5

[DBMM14] Ruben Dezeure, Peter Bühlmann, Lukas Meier, and Nicolai Meinshausen, *High-dimensional inference: Confidence intervals, p-values and r-software hdi*, arXiv:1408.4026 (2014). 8

[DET06] David Donoho, Michael Elad, and Vladimir Temlyakov, *Stable recovery of sparse overcomplete representations in the presence of noise*, IEEE Trans. on Inform. Theory **52** (2006), no. 1, 6–18. 4

[DH01] David L Donoho and Xiaoming Huo, *Uncertainty principles and ideal atomic decomposition*, IEEE Trans. on Inform. Theory **47** (2001), no. 7, 2845–2862. 4

[DMM09] D. L. Donoho, A. Maleki, and A. Montanari, *Message Passing Algorithms for Compressed Sensing*, Proceedings of the National Academy of Sciences **106** (2009), 18914–18919. 7

[EKBBL13] Noureddine El Karoui, Derek Bean, Peter J Bickel, and Bin Lim, Chingwayand Yu, *On robust regression with high-dimensional predictors*, Proceedings of the National Academy of Sciences **110** (2013), no. 36, 14557–14562. 7

[FST14] William Fithian, Dennis Sun, and Jonathan Taylor, *Optimal inference after model selection*, arXiv:1410.2597 (2014). 7

[JBC15] Lucas Janson, Rina Foygel Barber, and Emmanuel Candès, *Eigenprism: Inference for high-dimensional signal-to-noise ratios*, arXiv:1505.02097 (2015). 7

[JM13] A. Javanmard and A. Montanari, *Nearly optimal sample size in hypothesis testing for high-dimensional regression*, 51st Annual Allerton Conference (Monticello, IL), June 2013, pp. 1427–1434. 7, 11

[JM14a] Adel Javanmard and Andrea Montanari, *Confidence intervals and hypothesis testing for high-dimensional regression*, The Journal of Machine Learning Research **15** (2014), no. 1, 2869–2909. 2, 3, 4, 5, 6, 7, 8, 11, 16

[JM14b] ———, *Hypothesis Testing in High-Dimensional Regression under the Gaussian Random Design Model: Asymptotic Theory*, IEEE Trans. on Inform. Theory **60** (2014), no. 10, 6522–6554. 2, 3, 7, 20

[JS15] Lucas Janson and Weijie Su, *Familywise error rate control via knockoffs*, arXiv:1505.06549 (2015). 7

[JvdG14] Jana Jankova and Sara van de Geer, *Confidence intervals for high-dimensional inverse covariance estimation*, arXiv:1403.6752 (2014). 5

[JvdG15] Jana Janková and Sara van de Geer, *Honest confidence regions and optimality in high-dimensional precision matrix estimation*, arXiv:1507.02061 (2015). 5

[Kar13] Noureddine El Karoui, *Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results*, arXiv:1311.2445 (2013). 7

[LTTT14] Richard Lockhart, Jonathan Taylor, Ryan J Tibshirani, and Robert Tibshirani, *A significance test for the lasso*, The Annals of Statistics **42** (2014), no. 2, 413. 7

[MB06] N. Meinshausen and P. Bühlmann, *High-dimensional graphs and variable selection with the lasso*, The Annals of Statistics **34** (2006), 1436–1462. 5, 6

[MB10] ———, *Stability selection*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **72** (2010), 417–473. 7

[Mei14] Nicolai Meinshausen, *Group bound: confidence intervals for groups of variables in sparse high dimensional regression without assumptions on the design*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) (2014). 7

[RWY10] G. Raskutti, M.J. Wainwright, and B. Yu, *Restricted eigenvalue properties for correlated gaussian designs*, Journal of Machine Learning Research **11** (2010), 2241–2259. 22

[RZ13] Mark Rudelson and Shuheng Zhou, *Reconstruction from anisotropic random measurements*, IEEE Trans. on Inform. Theory **59** (2013), no. 6, 3434–3447. 9

[SZ12] Tingni Sun and Cun-Hui Zhang, *Scaled sparse linear regression*, Biometrika **99** (2012), no. 4, 879–898. 5

[Tal10] M. Talagrand, *Mean field models for spin glasses: Volume i*, Springer-Verlag, Berlin, 2010. 7

[Tib96] R. Tibshirani, *Regression shrinkage and selection with the Lasso*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **58** (1996), 267–288. 2

[TLTT14] Jonathan Taylor, Richard Lockhart, Ryan J. Tibshirani, and Robert Tibshirani, *Exact post-selection inference for forward stepwise and least angle regression*, arXiv:1401.3889 (2014). 7

[VdGBRD14] Sara Van de Geer, Peter Bühlmann, Yaacov Ritov, and Ruben Dezeure, *On asymptotically optimal confidence regions and tests for high-dimensional models*, The Annals of Statistics **42** (2014), no. 3, 1166–1202. 2, 3, 4, 5, 6, 7, 8, 11, 22

[Ver12] R. Vershynin, *Introduction to the non-asymptotic analysis of random matrices*, Compressed Sensing: Theory and Applications (Y.C. Eldar and G. Kutyniok, eds.), Cambridge University Press, 2012, pp. 210–268. 26, 29

[WR09] Larry Wasserman and Kathryn Roeder, *High dimensional variable selection*, Annals of Statistics **37** (2009), no. 5A, 2178. 7

[ZZ14] Cun-Hui Zhang and Stephanie S Zhang, *Confidence intervals for low dimensional parameters in high dimensional linear models*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **76** (2014), no. 1, 217–242. 2, 3, 4, 5, 6, 7, 11