Journal of Machine Learning Research 15 (2014) 2869-2909 Submitted 3/14; Revised 4/14; Published 10/14

Confidence Intervals and Hypothesis Testing for High-Dimensional Regression

Adel Javanmard ADELJ@STANFORD.EDU
Department of Electrical Engineering
Stanford University
Stanford, CA 94305, USA

Andrea Montanari MONTANAR@STANFORD.EDU
Department of Electrical Engineering and Department of Statistics
Stanford University
Stanford, CA 94305, USA

Editor: Nicolai Meinshausen

Abstract

Fitting high-dimensional statistical models often requires the use of non-linear parameter estimation procedures. As a consequence, it is generally impossible to obtain an exact characterization of the probability distribution of the parameter estimates. This in turn implies that it is extremely challenging to quantify the *uncertainty* associated with a certain parameter estimate. Concretely, no commonly accepted procedure exists for computing classical measures of uncertainty and statistical significance as confidence intervals or p-values for these models.

We consider here high-dimensional linear regression problem, and propose an efficient algorithm for constructing confidence intervals and p-values. The resulting confidence intervals have nearly optimal size. When testing for the null hypothesis that a certain parameter is vanishing, our method has nearly optimal power.

Our approach is based on constructing a 'de-biased' version of regularized M-estimators. The new construction improves over recent work in the field in that it does not assume a special structure on the design matrix. We test our method on synthetic data and a high-throughput genomic data set about riboflavin production rate, made publicly available by Bühlmann et al. (2014).

Keywords: hypothesis testing, confidence intervals, LASSO, high-dimensional models, bias of an estimator

1. Introduction

It is widely recognized that modern statistical problems are increasingly high-dimensional, i.e., require estimation of more parameters than the number of observations/samples. Examples abound from signal processing (Lustig et al., 2008), to genomics (Peng et al., 2010), collaborative filtering (Koren et al., 2009) and so on. A number of successful estimation techniques have been developed over the last ten years to tackle these problems. A widely applicable approach consists in optimizing a suitably regularized likelihood function. Such estimators are, by necessity, non-linear and non-explicit (they are solution of certain optimization problems).

The use of non-linear parameter estimators comes at a price. In general, it is impossible to characterize the distribution of the estimator. This situation is very different from the one of classical statistics in which either exact characterizations are available, or asymptotically exact ones can be derived from large sample theory (Van der Vaart, 2000). This has an important and very concrete consequence. In classical statistics, generic and well accepted procedures are available for characterizing the uncertainty associated to a certain parameter estimate in terms of confidence intervals or p-values (Wasserman, 2004; Lehmann and Romano, 2005). However, no analogous procedures exist in high-dimensional statistics.

In this paper we develop a computationally efficient procedure for constructing confidence intervals and p-values for a broad class of high-dimensional regression problems. The salient features of our procedure are:

(*i*) Our approach guarantees nearly optimal confidence interval sizes and testing power.

(*ii*) It is the first one to achieve this goal under essentially no assumptions beyond the standard conditions for high-dimensional consistency.

(*iii*) It allows for a streamlined analysis with respect to earlier work in the same area.

For the sake of clarity, we will focus our presentation on the case of linear regression, under Gaussian noise. Section 4 provides a detailed study of the case of non-Gaussian noise. A preliminary report on our results was presented in NIPS 2013 (Javanmard and Montanari, 2013a), which also discusses generalizations of the same approach to generalized linear models, and regularized maximum likelihood estimation.

In a linear regression model, we are given n i.i.d. pairs $(Y_1, X_1), (Y_2, X_2), \dots, (Y_n, X_n)$, with vectors $X_i \in \mathbb{R}^p$ and response variables Y_i given by

$$Y_i \,=\, \langle \theta_0, X_i \rangle + W_i\,, \qquad W_i \sim \mathsf{N}(0, \sigma^2)\,. \tag{1}$$

Here $\theta_0 \in \mathbb{R}^p$ and $\langle \,\cdot\,,\,\cdot\, \rangle$ is the standard scalar product in $\mathbb{R}^p$. In matrix form, letting $Y = (Y_1, \dots, Y_n)^{\mathsf{T}}$ and denoting by $\mathbf{X}$ the design matrix with rows $X_1^{\mathsf{T}}, \dots, X_n^{\mathsf{T}}$, we have

$$Y \,=\, \mathbf{X}\,\theta_0 + W\,, \qquad W \sim \mathsf{N}(0, \sigma^2 \mathrm{I}_{n\times n})\,. \tag{2}$$

The goal is to estimate the unknown (but fixed) vector of parameters $\theta_0 \in \mathbb{R}^p$.

In the classic setting, $n \gg p$ and the estimation method of choice is ordinary least squares yielding $\widehat{\theta}^{\mathrm{OLS}} = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}Y$. In particular $\widehat{\theta}^{\mathrm{OLS}}$ is Gaussian with mean θ_0 and covariance $\sigma^2(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}$. This directly allows to construct confidence intervals.[1]

In the high-dimensional setting where $p > n$, the matrix $(\mathbf{X}^{\mathsf{T}}\mathbf{X})$ is rank deficient and one has to resort to biased estimators. A particularly successful approach is the LASSO (Tibshirani, 1996; Chen and Donoho, 1995) which promotes sparse reconstructions through an ℓ_1 penalty:

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

[1] For instance, letting $Q \equiv (\mathbf{X}^{\mathsf{T}}\mathbf{X}/n)^{-1}$, $\widehat{\theta}_i^{\mathrm{OLS}} - 1.96\sigma\sqrt{Q_{ii}/n}, \widehat{\theta}_i^{\mathrm{OLS}} + 1.96\sigma\sqrt{Q_{ii}/n}]$ is a 95% confidence interval (Wasserman, 2004).

Algorithm 1 Unbiased estimator for θ_0 in high-dimensional linear regression models

```
Input: Measurement vector $y$, design matrix $\mathbf{X}$, parameters $\lambda$, $\mu$.
Output: Unbiased estimator $\widehat{\theta}^u$.
Let $\widehat{\theta}^n = \widehat{\theta}^n(Y, \mathbf{X}; \lambda)$ be the LASSO estimator as per Equation (3).
Set $\widehat{\Sigma} \equiv (\mathbf{X}^{\mathsf{T}}\mathbf{X})/n$.
for $i = 1, 2, \dots, p$ do
    Let $m_i$ be a solution of the convex program:
        $\begin{aligned} &\text{minimize} && m^{\mathsf{T}}\widehat{\Sigma}m \\ &\text{subject to} && \|\widehat{\Sigma}m - e_i\|_\infty \le \mu\,, \end{aligned} \qquad (4)$
    where $e_i \in \mathbb{R}^p$ is the vector with one at the $i$-th position and zero everywhere else.
Set $M = (m_1, \dots, m_p)^{\mathsf{T}}$. If any of the above problems is not feasible, then set $M = \mathrm{I}_{p\times p}$.
Define the estimator $\widehat{\theta}^u$ as follows:
        $\widehat{\theta}^u = \widehat{\theta}^n(\lambda) + \frac{1}{n} M\mathbf{X}^{\mathsf{T}}(Y - \mathbf{X}\widehat{\theta}^n(\lambda)) \qquad (5)$
```

In case the right hand side has more than one minimizer, one of them can be selected arbitrarily for our purposes. We will often omit the arguments Y, $\mathbf{X}$, as they are clear from the context.

We denote by $S \equiv \mathrm{supp}(\theta_0)$ the support of $\theta_0 \in \mathbb{R}^p$, defined as

$$\mathrm{supp}(\theta_0) \equiv \{i \in [p] : \theta_{0,i} \neq 0\}\,,$$

where we use the notation $[p] = \{1, \dots, p\}$. We further let $s_0 \equiv |S|$. A copious theoretical literature (Candès and Tao, 2005; Bickel et al., 2009; Bühlmann and van de Geer, 2011) shows that, under suitable assumptions on $\mathbf{X}$, the LASSO is nearly as accurate as if the support S was known *a priori*. Namely, for $n = \Omega(s_0 \log p)$, we have $\|\widehat{\theta}^n - \theta_0\|_2^2 = O(s_0\sigma^2(\log p)/n)$.

As mentioned above, these remarkable properties come at a price. Deriving an exact characterization for the distribution of $\widehat{\theta}^n$ is not tractable in general, and hence there is no simple procedure to construct confidence intervals and p-values. A closely related property is that $\widehat{\theta}^n$ is biased, an unavoidable property in high dimension, since a point estimate $\widehat{\theta}^n \in \mathbb{R}^p$ must be produced from data in lower dimension $Y \in \mathbb{R}^n$, $n < p$. We refer to Section 2.2 for further discussion of this point.

In order to overcome this challenge, we construct a de-biased estimator from the LASSO solution. The de-biased estimator is given by the simple formula $\widehat{\theta}^u = \widehat{\theta}^n + (1/n)\,M\mathbf{X}^{\mathsf{T}}(Y - \mathbf{X}\widehat{\theta}^n)$, as in Equation (5). The basic intuition is that $\mathbf{X}^{\mathsf{T}}(Y - \mathbf{X}\widehat{\theta}^n)/(n\lambda)$ is a subgradient of the ℓ_1 norm at the LASSO solution $\widehat{\theta}^n$. By adding a term proportional to this subgradient, our procedure compensates the bias introduced by the ℓ_1 penalty in the LASSO.

We will prove in Section 2.1 that $\widehat{\theta}^u$ is approximately Gaussian, with mean θ_0 and covariance $\sigma^2(M\widehat{\Sigma}M)/n$, where $\widehat{\Sigma} = (\mathbf{X}^{\mathsf{T}}\mathbf{X}/n)$ is the empirical covariance of the feature vectors. This result allows to construct confidence intervals and p-values in complete analogy with classical statistics procedures. For instance, letting $Q \equiv M\widehat{\Sigma}M$, $[\widehat{\theta}^u_i - 1.96\sigma\sqrt{Q_{ii}/n}, \widehat{\theta}^u_i + 1.96\sigma\sqrt{Q_{ii}/n}]$ is a 95% confidence interval. The size of this interval is of order $\sigma/\sqrt{n}$, which is the optimal (minimum) one, i.e., the same that would have been obtained by knowing *a priori* the support of θ_0. In practice the noise standard deviation is not known, but σ can be replaced by any consistent estimator $\widehat{\sigma}$ (see Section 3 for more details on this).

A key role is played by the matrix $M \in \mathbb{R}^{p\times p}$ whose function is to 'decorrelate' the columns of $\mathbf{X}$. We propose here to construct M by solving a convex program that aims at optimizing two objectives. One one hand, we try to control $|M\widehat{\Sigma} - \mathrm{I}|_\infty$ (here and below $|\cdot|_\infty$ denotes the entrywise ℓ_∞ norm) which, as shown in Theorem 8, controls the non-Gaussianity and bias of $\widehat{\theta}^u$. On the other, we minimize $[M\widehat{\Sigma}M]_{i,i}$, for each $i \in [p]$, which controls the variance of $\widehat{\theta}^u_i$.

The idea of constructing a de-biased estimator of the form $\widehat{\theta}^u = \widehat{\theta}^n + (1/n)\, M\mathbf{X}^{\mathsf{T}}(Y - \mathbf{X}\widehat{\theta}^n)$ was used by the present authors in Javanmard and Montanari (2013b), that suggested the choice $M = c\Sigma^{-1}$, with $\Sigma = \mathbb{E}\{X_1X_1^{\mathsf{T}}\}$ the population covariance matrix and c a positive constant. A simple estimator for Σ was proposed for sparse covariances, but asymptotic validity and optimality were proven only for uncorrelated Gaussian designs (i.e., Gaussian $\mathbf{X}$ with $\Sigma = \mathrm{I}$). Van de Geer, Bülhmann, Ritov and Dezeure (van de Geer et al., 2014) used the same construction with M an estimate of Σ^{-1} which is appropriate for sparse inverse covariances. These authors prove semi-parametric optimality in a non-asymptotic setting, provided the sample size is at least $n = \Omega((s_0 \log p)^2)$.

From a technical point of view, our proof starts from a simple decomposition of the de-biased estimator $\widehat{\theta}^u$ into a Gaussian part and an error term, already used in van de Geer et al. (2014). However, departing radically from earlier work, we realize that M need not be a good estimator of Σ^{-1} in order for the de-biasing procedure to work. We instead set M as to minimize the error term and the variance of the Gaussian term. As a consequence of this choice, our approach applies to general covariance structures Σ. By contrast, earlier approaches applied only to sparse Σ, as in Javanmard and Montanari (2013b), or sparse Σ^{-1} as in van de Geer et al. (2014). The only assumptions we make on Σ are the standard compatibility conditions required for high-dimensional consistency (Bühlmann and van de Geer, 2011). A detailed comparison of our results with the ones of van de Geer et al. (2014) can be found in Section 2.3.

1.1 Organization of the Paper

Our presentation is organized as follows.

Section 2 considers a general debiased estimator of the form $\widehat{\theta}^u = \widehat{\theta}^n + (1/n)\, M\mathbf{X}^{\mathsf{T}}(Y - \mathbf{X}\widehat{\theta}^n)$. We introduce a figure of merit of the pair $M, \mathbf{X}$, termed the generalized coherence parameter $\mu_*(\mathbf{X}; M)$. We show that, if the generalized coherence is small, then the debiasing procedure is effective (for a given deterministic design), see Theorem 6. We then turn to random designs, and show that the generalized coherence parameter can be made as small as $\sqrt{(\log p)/n}$, though a convex optimization procedure for computing M. This results in a bound on the bias of $\widehat{\theta}^u$, cf. Theorem 8: the largest entry of the bias is of order $(s_0 \log p)/n$. This must be compared with the standard deviation of $\widehat{\theta}^u_i$, which is of order $\sigma/\sqrt{n}$. The conclusion is that, for $s_0 = o(\sqrt{n}/\log p)$, the bias of $\widehat{\theta}^u$ is negligible.

Section 3 applies these distributional results to deriving confidence intervals and hypothesis testing procedures for low-dimensional marginals of θ_0. The basic intuition is that $\widehat{\theta}^u$ is approximately Gaussian with mean θ_0, and known covariance structure. Hence standard optimal tests can be applied. We prove a general lower bound on the power of our testing procedure, in Theorem 16. In the special case of Gaussian random designs with i.i.d. rows, we can compare this with the upper bound proved in Javanmard and Montanari (2013b), cf. Theorem 17. As a consequence, the asymptotic efficiency of our approach is constant-optimal. Namely, it is lower bounded by a constant $1/\eta_{\Sigma,s_0}$ which is bounded away from 0, cf. Theorem 18. (For instance $\eta_{\mathrm{I},s_0} = 1$, and η_{Σ,s_0} is always upper bounded by the condition number of Σ.)

Section 4 uses the central limit theorem for triangular arrays to generalize the above results to non-Gaussian noise.

Section 5 illustrates the above results through numerical simulations both on synthetic and on real data. In the interest of reproducibility, an R implementation of our algorithm is available at `http://www.stanford.edu/~montanar/sslasso/`.

Note that our proofs require stricter sparsity s_0 (or larger sample size n) than required for consistent estimation. We assume $s_0 = o(\sqrt{n}/\log p)$ instead of $s_0 = o(n/\log p)$ (Candès and Tao, 2007; Bickel et al., 2009; Bühlmann and van de Geer, 2011). The same assumption is made in van de Geer et al. (2014), on top of additional assumptions on the sparsity of Σ^{-1}.

It is currently an open question whether successful hypothesis testing can be performed under the weaker assumption $s_0 = o(n/\log p)$. We refer to Javanmard and Montanari (2013c) for preliminary work in that direction. The barrier at $s_0 = o(\sqrt{n}/\log p)$ is possibly related to an analogous assumption that arises in Gaussian graphical models selection (Ren et al., 2013).

1.2 Further Related Work

The theoretical literature on high-dimensional statistical models is vast and rapidly growing. Estimating sparse linear regression models is the most studied problem in this area, and a source of many fruitful ideas. Limiting ourselves to linear regression, earlier work investigated prediction error (Greenshtein and Ritov, 2004), model selection properties (Meinshausen and Bühlmann, 2006; Zhao and Yu, 2006; Wainwright, 2009; Candès and Plan, 2009), ℓ_2 consistency (Candès and Tao, 2005; Bickel et al., 2009). . Of necessity, we do not provide a complete set of references, and instead refer the reader to Bühlmann and van de Geer (2011) for an in-depth introduction to this area.

The problem of quantifying statistical significance in high-dimensional parameter estimation is, by comparison, far less understood. Zhang and Zhang (2014) and Bühlmann (2013) proposed hypothesis testing procedures under restricted eigenvalue or compatibility conditions (Bühlmann and van de Geer, 2011). These papers provide deterministic guarantees but, in order to achieve a certain target significance level α and power $1-\beta$, they require $|\theta_{0,i}| \ge c \max\{\sigma s_0 \log p/\, n, \sigma/\sqrt{n}\}$. The best lower bound (Javanmard and Montanari, 2013b) shows that any such test requires instead $|\theta_{0,i}| \ge c(\alpha,\beta)\sigma/\sqrt{n}$. (The lower bound of Javanmard and Montanari 2013b is reproduced as Theorem 17 here, for the reader's convenience.)

In other words, the guarantees of Zhang and Zhang (2014); Bühlmann (2013) can be suboptimal by a factor as large as $\sqrt{s_0}$. Equivalently, in order for the coefficient $\theta_{0,i}$ to be detectable with appreciable probability, it needs to be larger than the overall ℓ_2 error. Here we will propose a test that, for random designs, achieves significance level α and power $1-\beta$ for $|\theta_{0,i}| \geq c'(\alpha,\beta)\sigma/\sqrt{n}$.

Lockhart et al. (2014) develop a test for the hypothesis that a newly added coefficient along the LASSO regularization path is irrelevant. This however does not allow to test arbitrary coefficients at a given value of λ, which is instead the problem addressed in this paper. These authors further assume that the current LASSO support contains the actual support $\text{supp}(\theta_0)$ and that the latter has bounded size.

Belloni et al. (2014, 2013) consider inference in a regression model with high-dimensional data. In this model the response variable relates to a scalar main regressor and a p-dimensional control vector. The main regressor is of primary interest and the control vector is treated as nuisance component. Assuming that the control vector is s_0-sparse, the authors propose a method to construct confidence regions for the parameter of interest under the sample size requirement $(s_0^2 \log p)/n \to 0$. The proposed method is shown to attain the semi-parametric efficiency bounds for this class of models. The key modeling assumption in this paper is that the scalar regressor of interest is random, and depends linearly on the p-dimensional control vector, with a sparse coefficient vector (with sparsity again of order $o(\sqrt{n/\log p})$. This assumption is closely related to the sparse inverse covariance assumption of van de Geer et al. (2014) (with the difference that only one regressor is tested).

Finally, resampling methods for hypothesis testing were studied in Meinshausen and Bühlmann (2010); Minnier et al. (2011). These methods are perturbation-based procedures to approximate the distribution of a general class of penalized parameter estimates for the case $n > p$. The idea is to consider the minimizer of a stochastically perturbed version of the regularized objective function, call it $\tilde{\theta}$, and characterize the limiting distribution of the regularized estimator $\widehat{\theta}$ in terms of the distribution of $\tilde{\theta}$. In order to estimate the latter, a large number of random samples of the perturbed objective function are generated, and for each sample the minimizer is computed. Finally the theoretical distribution of $\tilde{\theta}$ is approximated by the empirical distribution of these minimizers.

After the present paper was submitted for publication, we became aware that Dezeure and Bühlmann (2013) had independently worked on similar ideas.

1.3 Preliminaries and Notations

In this section we introduce some basic definitions used throughout the paper, starting with simple notations.

For a matrix A and set of indices I, J, we let $A_{I,J}$ denote the submatrix formed by the rows in I and columns in J. Also, $A_{I,\cdot}$ (resp. $A_{\cdot,I}$) denotes the submatrix containing just the rows (reps. columns) in I. Likewise, for a vector v, v_I is the restriction of v to indices in I. We use the shorthand $A^{-1}_{I,J} = (A^{-1})_{I,J}$. In particular, $A^{-1}_{i,i} = (A^{-1})_{i,i}$. The maximum and the minimum singular values of A are respectively denoted by $\sigma_{\max}(A)$ and $\sigma_{\min}(A)$. 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 number of nonzero entries of v. For a matrix A, $\|A\|_p$ is the ℓ_p operator norm, and $|A|_p$ is the elementwise ℓ_p norm. For a vector v, $\text{supp}(v)$ represents the positions of nonzero entries of v. Throughout, $\Phi(x) \equiv \int_{-\infty}^{x} e^{-t^2/2}\text{d}t/\sqrt{2\pi}$ denotes the CDF of the standard normal distribution. Finally, *with high probability* (w.h.p) means with probability converging to one as $n \to \infty$.

We let $\widehat{\Sigma} \equiv \mathbf{X}^{\mathsf{T}}\mathbf{X}/n$ be the sample covariance matrix. For $p > n$, $\widehat{\Sigma}$ is always singular. However, we may require $\widehat{\Sigma}$ to be nonsingular for a restricted set of directions.

Definition 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*

$$\phi^2(\widehat{\Sigma}, S) \equiv \min_{\theta\in\mathbb{R}^p} \Big\{ \frac{|S|\,\langle\theta, \widehat{\Sigma}\,\theta\rangle}{\|\theta_S\|_1^2} : \ \ \theta \in \mathbb{R}^p, \ \ \|\theta_{S^c}\|_1 \le 3\|\theta_S\|_1 \Big\} . \tag{6}$$

We say that $\widehat{\Sigma} \in \mathbb{R}^{p\times p}$ satisfies the compatibility condition *for the set $S \subseteq [p]$, with constant ϕ_0 if $\phi(\widehat{\Sigma}, S) \ge \phi_0$. We say that it holds for the design matrix $\mathbf{X}$, if it holds for $\widehat{\Sigma} = \mathbf{X}^{\mathsf{T}}\mathbf{X}/n$.*

In the following, we shall drop the argument $\widehat{\Sigma}$ if clear from the context. Note that a slightly more general definition is used normally (Bühlmann and van de Geer, 2011, Section 6.13), whereby the condition $\|\theta_{S^c}\|_1 \le 3\|\theta_S\|_1$, is replaced by $\|\theta_{S^c}\|_1 \le L\|\theta_S\|_1$. The resulting constant $\phi(\widehat{\Sigma}, S, L)$ depends on L. For the sake of simplicity, we restrict ourselves to the case $L = 3$.

Definition 2 *The* sub-Gaussian norm *of a random variable X, denoted by $\|X\|_{\psi_2}$, is defined as*

$$\|X\|_{\psi_2} = \sup_{q\ge 1} q^{-1/2}(\mathbb{E}|X|^q)^{1/q} .$$

For a random vector $X \in \mathbb{R}^n$, its sub-Gaussian norm is defined as

$$\|X\|_{\psi_2} = \sup_{x\in S^{n-1}} \|\langle X, x\rangle\|_{\psi_2} ,$$

where S^{n-1} denotes the unit sphere in $\mathbb{R}^n$.

Definition 3 *The* sub-exponential norm *of a random variable X, denoted by $\|X\|_{\psi_1}$, is defined as*

$$\|X\|_{\psi_1} = \sup_{q\ge 1} q^{-1}(\mathbb{E}|X|^q)^{1/q} .$$

For a random vector $X \in \mathbb{R}^n$, its sub-exponential norm is defined as

$$\|X\|_{\psi_1} = \sup_{x\in S^{n-1}} \|\langle X, x\rangle\|_{\psi_1} ,$$

where S^{n-1} denotes the unit sphere in $\mathbb{R}^n$.

2. Compensating the Bias of the LASSO

In this section we present our characterization of the de-biased estimator $\widehat{\theta}^u$ (Subsection 2.1). This characterization also clarifies in what sense the LASSO estimator is biased. We discuss this point in Subsection 2.2.

2.1 A De-biased Estimator for θ_0

As emphasized above, our approach is based on a de-biased estimator defined in Equation (5), and on its distributional properties. In order to clarify the latter, it is convenient to begin with a slightly broader setting and consider a general debiasing procedure that makes use of a an arbitrary $M \in \mathbb{R}^{p\times p}$. Namely, we define

$$\widehat{\theta}^*(Y, \mathbf{X}; M, \lambda) = \widehat{\theta}^n(\lambda) + \frac{1}{n} M\mathbf{X}^{\mathsf{T}}(Y - \mathbf{X}\widehat{\theta}^n(\lambda))\,. \tag{7}$$

For notational simplicity, we shall omit the arguments $Y, \mathbf{X}, M, \lambda$ unless they are required for clarity. The quality of this debiasing procedure depends of course on the choice of M, as well as on the design $\mathbf{X}$. We characterize the pair $(\mathbf{X}, M)$ by the following figure of merit.

Definition 4 *Given the pair $\mathbf{X} \in \mathbb{R}^{n\times p}$ and $M \in \mathbb{R}^{p\times p}$, let $\widehat{\Sigma} = \mathbf{X}^{\mathsf{T}}\mathbf{X}/n$ denote the associated sample covariance. Then, the* generalized coherence parameter of $\mathbf{X}, M$, *denoted by $\mu_*(\mathbf{X}; M)$, is*

$$\mu_*(\mathbf{X}; M) \equiv \big|M\widehat{\Sigma} - \mathrm{I}\big|_\infty\,. \tag{8}$$

The minimum (generalized) coherence *of* $\mathbf{X}$ *is* $\mu_{\min}(\mathbf{X}) = \min_{M\in\mathbb{R}^{p\times p}} \mu_*(\mathbf{X}; M)$. *We denote by $M_{\min}(\mathbf{X})$ any minimizer of $\mu_*(\mathbf{X}; M)$.*

Note that the minimum coherence can be computed efficiently since $M \mapsto \mu_*(\mathbf{X}; M)$ is a convex function (even more, the optimization problem is a linear program).

The motivation for our terminology can be grasped by considering the following special case.

Remark 5 *Assume that the columns of $\mathbf{X}$ are normalized to have ℓ_2 norm equal to $\sqrt{n}$ (i.e., $\|\mathbf{X}e_i\|_2 = \sqrt{n}$ for all $i \in [p]$), and $M = \mathrm{I}$. Then $(M\widehat{\Sigma} - \mathrm{I})_{i,i} = 0$, and the maximum $|M\widehat{\Sigma} - \mathrm{I}|_\infty = \max_{i\neq j} |(\widehat{\Sigma})_{ij}|$. In other words $\mu(\mathbf{X}; \mathrm{I})$ is the maximum normalized scalar product between distinct columns of $\mathbf{X}$:*

$$\mu_*(\mathbf{X}; \mathrm{I}) = \frac{1}{n} \max_{i\neq j} \big|\langle \mathbf{X}e_i, \mathbf{X}e_j\rangle\big|\,. \tag{9}$$

The quantity (9) is known as the *coherence parameter* of the matrix $\mathbf{X}/\sqrt{n}$ and was first defined in the context of approximation theory by Mallat and Zhang (1993), and by Donoho and Huo (2001).

Assuming, for the sake of simplicity, that the columns of $\mathbf{X}$ are normalized so that $\|\mathbf{X}e_i\|_2 = \sqrt{n}$, a small value of the coherence parameter $\mu_*(\mathbf{X}; \mathrm{I})$ means that the columns of $\mathbf{X}$ are roughly orthogonal. We emphasize however that $\mu_*(\mathbf{X}; M)$ can be much smaller than its classical coherence parameter $\mu_*(\mathbf{X}; \mathrm{I})$. For instance, $\mu_*(\mathbf{X}; \mathrm{I}) = 0$ if and only if $\mathbf{X}/\sqrt{n}$ is an orthogonal matrix. On the other hand, $\mu_{\min}(\mathbf{X}) = 0$ if and only if $\mathbf{X}$ has rank p.[2]

The following theorem is a slight generalization of a result of van de Geer et al. (2014). Let us emphasize that it applies to deterministic design matrices $\mathbf{X}$.

[2] Of course this example requires $n \geq p$. It is the simplest example that illustrates the difference between coherence and generalized coherence, and it is not hard to find related examples with $n < p$.

Theorem 6 *Let $\mathbf{X} \in \mathbb{R}^{n\times p}$ be any (deterministic) design matrix, and $\widehat{\theta}^* = \widehat{\theta}^*(Y,\mathbf{X};M,\lambda)$ be a general debiased estimator as per Equation (7). Then, setting $Z = M\mathbf{X}^{\mathsf{T}}W/\sqrt{n}$, we have*

$$\sqrt{n}(\widehat{\theta}^* - \theta_0) = Z + \Delta\,, \quad Z \sim \mathsf{N}(0,\sigma^2 M\widehat{\Sigma}M^{\mathsf{T}})\,, \quad \Delta = \sqrt{n}(M\widehat{\Sigma} - \mathrm{I})(\theta_0 - \widehat{\theta}^n)\,. \tag{10}$$

Further, assume that $\mathbf{X}$ satisfies the compatibility condition for the set $S = \mathrm{supp}(\theta_0)$, $|S| \le s_0$, with constant ϕ_0, and has generalized coherence parameter $\mu_* = \mu_*(\mathbf{X};M)$, and let $K \equiv \max_{i\in[p]} \widehat{\Sigma}_{i,i}$. Then, letting $\lambda = \sigma\sqrt{(c^2\log p)/n}$, we have

$$\mathbb{P}\Big(\|\Delta\|_\infty \ge \frac{4c\mu_*\sigma s_0}{\phi_0^2}\sqrt{\log p}\Big) \le 2p^{-c_0}\,, \quad c_0 = \frac{c^2}{32K} - 1\,. \tag{11}$$

Further, if $M = M_{\min}(\mathbf{X})$ minimizes the convex cost function $|M\widehat{\Sigma} - \mathrm{I}|_\infty$, then μ_* can be replaced by $\mu_{\min}(\mathbf{X})$ in Equation (11).

The above theorem decomposes the estimation error $(\widehat{\theta}^* - \theta_0)$ into a zero mean Gaussian term $Z/\sqrt{n}$ and a bias term $\Delta/\sqrt{n}$ whose maximum entry is bounded as per Equation (11). This estimate on $\|\Delta\|_\infty$ depends on the design matrix through two constants: the compatibility constant ϕ_0 and the generalized coherence parameter $\mu_*(\mathbf{X};M)$. The former is a well studied property of the design matrix (Bühlmann and van de Geer, 2011; van de Geer and Bühlmann, 2009), and assuming ϕ_0 of order one is nearly necessary for the LASSO to achieve optimal estimation rate in high dimension. On the contrary, the definition of $\mu_*(\mathbf{X};M)$ is a new contribution of the present paper.

The next theorem establishes that, for a natural probabilistic model of the design matrix $\mathbf{X}$, both ϕ_0 and $\mu_*(\mathbf{X};M)$ can be bounded with probability converging rapidly to one as $n,p \to \infty$. Further, the bound on $\mu_*(\mathbf{X},M)$ hold for the special choice of M that is constructed by Algorithm 1.

Theorem 7 *Let $\Sigma \in \mathbb{R}^{p\times p}$ be such that $\sigma_{\min}(\Sigma) \ge C_{\min} > 0$, and $\sigma_{\max}(\Sigma) \le C_{\max} < \infty$, and $\max_{i\in[p]} \Sigma_{ii} \le 1$. Assume $\mathbf{X}\Sigma^{-1/2}$ to have independent sub-Gaussian rows, with zero mean and sub-Gaussian norm $\|\Sigma^{-1/2}X_1\|_{\psi_2} = \kappa$, for some constant $\kappa \in (0,\infty)$.*

(a) For $\phi_0, s_0, K \in \mathbb{R}_{>0}$, let $\mathcal{E}_n = \mathcal{E}_n(\phi_0,s_0,K)$ be the event that the compatibility condition holds for $\widehat{\Sigma} = (\mathbf{X}^{\mathsf{T}}\mathbf{X}/n)$, for all sets $S \subseteq [p]$, $|S| \le s_0$ with constant $\phi_0 > 0$, and that $\max_{i\in[p]} \widehat{\Sigma}_{i,i} \le K$. Explicitly

$$\mathcal{E}_n(\phi_0,s_0,K) \equiv \Big\{\mathbf{X} \in \mathbb{R}^{n\times p}:\ \min_{S:\,|S|\le s_0} \phi(\widehat{\Sigma},S) \ge \phi_0,\ \max_{i\in[p]} \widehat{\Sigma}_{i,i} \le K,\ \widehat{\Sigma} = (\mathbf{X}^{\mathsf{T}}\mathbf{X}/n)\Big\}\,. \tag{12}$$

Then there exists $c_* \le 2000$ such that the following happens. If $n \ge \nu_0\, s_0\log(p/s_0)$, $\nu_0 \equiv 5\times 10^4 c_*(C_{\max}/C_{\min})^2\kappa^4$, $\phi_0 = \sqrt{C_{\min}}/2$, and $K \ge 1 + 20\kappa^2\sqrt{(\log p)/n}$, then

$$\mathbb{P}\big(\mathbf{X} \in \mathcal{E}_n(\phi_0,s_0,K)\big) \ge 1 - 4\,e^{-c_1 n}\,, \qquad c_1 \equiv \frac{1}{4c_*\kappa^4}\,. \tag{13}$$

(*b*) *For $a > 0$, let $\mathcal{G}_n = \mathcal{G}_n(a)$ be the event that the problem (4) is feasible for $\mu = a\sqrt{(\log p)/n}$, or equivalently*

$$\mathcal{G}_n(a) \equiv \Big\{\mathbf{X} \in \mathbb{R}^{n\times p} : \; \mu_{\min}(\mathbf{X}) < a\sqrt{\frac{\log p}{n}}\Big\}\,. \tag{14}$$

Then, for $n \ge a^2 C_{\min} \log p/(4e^2 C_{\max}\kappa^4)$

$$\mathbb{P}\big(\mathbf{X} \in \mathcal{G}_n(a)\big) \ge 1 - 2\,p^{-c_2}\,, \qquad c_2 \equiv \frac{a^2 C_{\min}}{24e^2\kappa^4 C_{\max}} - 2\,. \tag{15}$$

The proof of this theorem is given in Section 6.2 (for part (*a*)) and Section 6.3 (part (*b*)).

The proof that event $\mathcal{E}_n$ holds with high probability relies crucially on a theorem by Rudelson and Shuheng (2013, Theorem 6). Simplifying somewhat, the latter states that, if the restricted eigenvalue condition of Bickel et al. (2009) holds for the population covariance Σ, then it holds with high probability for the sample covariance $\widehat{\Sigma}$. (Recall that the restricted eigenvalue condition is implied by a lower bound on the minimum singular value,[3] and that it implies the compatibility condition van de Geer and Bühlmann, 2009.)

Finally, by putting together Theorem 6 and Theorem 7, we obtain the following conclusion. We refer to Section 6.4 for the proof of Theorem 8.

Theorem 8 *Consider the linear model* (1) *and let $\widehat{\theta}^u$ be defined as per Equation* (5) *in Algorithm 1, with $\mu = a\sqrt{(\log p)/n}$. Then, setting $Z = M\mathbf{X}^{\mathsf{T}}W/\sqrt{n}$, we have*

$$\sqrt{n}(\widehat{\theta}^u - \theta_0) = Z + \Delta\,, \quad Z|\mathbf{X} \sim \mathsf{N}(0, \sigma^2 M\widehat{\Sigma}M^{\mathsf{T}})\,, \quad \Delta = \sqrt{n}(M\widehat{\Sigma} - \mathrm{I})(\theta_0 - \widehat{\theta}^n)\,. \tag{16}$$

Further, under the assumptions of Theorem 7, and for $n \ge \max(\nu_0 s_0 \log(p/s_0), \nu_1 \log p)$, $\nu_1 = \max(1600\kappa^4, a^2/(4e^2\kappa^4))$, and $\lambda = \sigma\sqrt{(c^2 \log p)/n}$, we have

$$\mathbb{P}\bigg\{\|\Delta\|_\infty \ge \Big(\frac{16ac\,\sigma}{C_{\min}}\Big)\frac{s_0 \log p}{\sqrt{n}}\bigg\} \le 4\,e^{-c_1 n} + 4\,p^{-\tilde{c_0}\wedge c_2}\,. \tag{17}$$

where $\tilde{c_0} = (c^2/48) - 1$ and c_1, c_2 are given by Equations (13) and (15).

Finally, the tail bound (17) holds for any choice of M that is only function of the design matrix $\mathbf{X}$, and satisfies the feasibility condition in Equation (4), i.e., $|M\widehat{\Sigma} - \mathrm{I}|_\infty \le \mu$.

Assuming $\sigma, C_{\min}$ of order one, the last theorem establishes that, for random designs, the maximum size of the 'bias term' Δ_i over $i \in [p]$ is:

$$\|\Delta\|_\infty = O\Big(\frac{s_0 \log p}{\sqrt{n}}\Big) \tag{18}$$

On the other hand, the 'noise term' Z_i is roughly of order $\sqrt{[M\widehat{\Sigma}M^{\mathsf{T}}]_{ii}}$. Bounds on the variances $[M\widehat{\Sigma}M^{\mathsf{T}}]_{ii}$ will be given in Section 3.3 (cf. Equation 82 in the proof of Theorem 16) showing that, if M is computed through Algorithm 1, $[M\widehat{\Sigma}M^{\mathsf{T}}]_{ii}$ is of order one for a broad family of random designs. As a consequence $|\Delta_i|$ is much smaller than $|Z_i|$ whenever $s_0 = o(\sqrt{n}/\log p)$. We summarize these remarks below.

[3] Note, in particular, at the cost of further complicating the last statement, the condition $\sigma_{\min}(\Sigma) = \Omega(1)$ can be further weakened.

Remark 9 *Theorem 8 only requires that the support size satisfies* $s_0 = O(n/\log p)$*. If we further assume* $s_0 = o(\sqrt{n}/\log p)$*, then we have* $\|\Delta\|_\infty = o(1)$ *with high probability. Hence,* $\widehat{\theta}^u$ *is an asymptotically unbiased estimator for* θ_0*.*

A more formal comparison of the bias of $\widehat{\theta}^u$, and of the one of the LASSO estimator $\widehat{\theta}^n$ can be found in Section 2.2 below. Section 2.3 compares our approach with the related one in van de Geer et al. (2014).

As it can be seen from the statement of Theorem 6 and Theorem 7, the claim of Theorem 8 does not rely on the specific choice of the objective function in optimization problem (4) and only uses the constraint on $\|\widehat{\Sigma}m - e_i\|_\infty$. In particular it holds for any matrix M that is feasible. On the other hand, the specific objective function problem (4) minimizes the variance of the noise term $\mathrm{Var}(Z_i)$.

2.2 Discussion: Bias of the LASSO

Theorems 6 and 7 provide a quantitative framework to discuss in what sense the LASSO estimator $\widehat{\theta}^n$ is asymptotically biased, while the de-biased estimator $\widehat{\theta}^u$ is asymptotically unbiased.

Given an estimator $\widehat{\theta}^n$ of the parameter vector θ_0, we define its *bias* to be the vector

$$\mathsf{Bias}(\widehat{\theta}^n) \equiv \mathbb{E}\{\widehat{\theta}^n - \theta_0|\mathbf{X}\}\,. \tag{19}$$

Note that, if the design is random, $\mathsf{Bias}(\widehat{\theta}^n)$ is a measurable function of $\mathbf{X}$. If the design is deterministic, $\mathsf{Bias}(\widehat{\theta}^n)$ is a deterministic quantity as well, and the conditioning is redundant.

It follows from Equation (10) that

$$\mathsf{Bias}(\widehat{\theta}^u) = \frac{1}{\sqrt{n}}\mathbb{E}\{\Delta|\mathbf{X}\}\,. \tag{20}$$

Applying Theorem 8 with high probability, $\|\Delta\|_\infty = O(s_0 \log p/\sqrt{n})$. The next corollary establishes that this translates into a bound on $\mathsf{Bias}(\widehat{\theta}^u)$ for all $\mathbf{X}$ in a set that has probability rapidly converging to one as n, p get large.

Corollary 10 *Under the assumptions of Theorem 8, let* c_1, c_2 *be defined as per Equations (13), (15). Then we have*

$$\mathbf{X} \in \mathcal{E}_n(\sqrt{C_{\min}}/2, s_0, 3/2) \cap \mathcal{G}_n(a) \;\Rightarrow\; \|\mathsf{Bias}(\widehat{\theta}^u)\|_\infty \le \frac{160a}{C_{\min}} \cdot \frac{\sigma s_0 \log p}{n}\,, \tag{21}$$

$$\mathbb{P}\Big(\mathbf{X} \in \mathcal{E}_n(\sqrt{C_{\min}}/2, s_0, 3/2) \cap \mathcal{G}_n(a)\Big) \ge 1 - 4e^{-c_1 n} - 2\,p^{-c_2}\,. \tag{22}$$

The proof of this corollary can be found in Appendix B.1.

This result can be contrasted with a converse result for the LASSO estimator. Namely, as stated below, there are choices of the vector θ_0, and of the design covariance Σ, such that $\mathsf{Bias}(\widehat{\theta}^n)$ is the sum of two terms. One is of order order $\lambda = c\sigma\sqrt{(\log p)/n}$ and the second is of order $\|\mathsf{Bias}(\widehat{\theta}^u)\|_\infty$. If s_0 is significantly smaller than $\sqrt{n/\log p}$ (which is the main regime studied in the rest of the paper), the first term dominates and $\|\mathsf{Bias}(\widehat{\theta}^n)\|_\infty$ is much larger than $\|\mathsf{Bias}(\widehat{\theta}^u)\|_\infty$. On the other hand, if s_0 is significantly larger than $\sqrt{n/\log p}$

then $\|\mathsf{Bias}(\widehat{\theta}^n)\|_\infty$ is of the same order as $\|\mathsf{Bias}(\widehat{\theta}^u)\|_\infty$. This justifies referring to $\widehat{\theta}^u$ as an *unbiased estimator*.

Notice that, since we want to establish a negative result about the LASSO, it is sufficient to exhibit a specific covariance structure Σ satisfying the assumptions of the previous corollary. Remarkably it is sufficient to consider standard designs, i.e., $\Sigma = \mathrm{I}_{p\times p}$.

Corollary 11 *Under the assumptions of Theorem 8, further consider the case $\Sigma = \mathrm{I}$. Then, there exist a set of design matrices $\mathcal{B}_n \subseteq \mathbb{R}^{n\times p}$, and coefficient vectors $\theta_0 \in \mathbb{R}^p$, $\|\theta_0\|_0 \le s_0$, such that*

$$\mathbf{X}\in\mathcal{B}_n \;\Rightarrow\; \|\mathsf{Bias}(\widehat{\theta}^n)\|_\infty \ge \left|\frac{2}{3}\lambda - \|\mathsf{Bias}(\widehat{\theta}^*)\|_\infty\right|\,, \tag{23}$$

$$\mathbb{P}(\mathcal{B}_n) \ge 1 - 4\,e^{-c_1 n} - 2\,p^{-3}\,, \tag{24}$$

where $\widehat{\theta}^* = \widehat{\theta}^*(Y,\mathbf{X};\mathrm{I},\lambda)$, with $\lambda = c\sigma\sqrt{(\log p)/n}$. In particular, there exists $c_{**} \le 4800$ such that if $n \ge (3c_{**}s_0/c)^2\log p$ and $p \ge 13^{48/(c^2-48)}$, then the following hold true:

$$\|\mathsf{Bias}(\widehat{\theta}^*)\|_\infty \le \lambda/3\,, \tag{25}$$

$$\|\mathsf{Bias}(\widehat{\theta}^n)\|_\infty \ge \frac{c\sigma}{3}\sqrt{\frac{\log p}{n}} \gg \|\mathsf{Bias}(\widehat{\theta}^u)\|_\infty\,, \tag{26}$$

where $\widehat{\theta}^u$ is given by Equation (5) in Algorithm 1, with $\mu = 30\sqrt{(\log p)/n}$.

A formal proof of this statement is deferred to Appendix B.2, but the underlying mathematical mechanism is quite simple. Recall that the KKT condition for the LASSO estimator (3) reads

$$\frac{1}{n}\mathbf{X}^{\mathsf{T}}(Y - \mathbf{X}\widehat{\theta}^n) = \lambda\, v(\widehat{\theta}^n)\,, \tag{27}$$

with $v(\widehat{\theta}^n) \in \mathbb{R}^p$ being a vector in the subgradient of the ℓ_1 norm at $\widehat{\theta}^n$. Adding $\widehat{\theta}^n - \theta_0$ to both sides, and taking expectation over the noise, we get

$$\mathsf{Bias}(\widehat{\theta}^*) = \mathsf{Bias}(\widehat{\theta}^n) + \lambda\mathbb{E}\{v(\widehat{\theta}^n)|\mathbf{X}\}\,, \tag{28}$$

where $\widehat{\theta}^*$ is a debiased estimator of the general form (7), for $M = \mathrm{I}$. As shown formally in Appendix B.2, $\|\mathbb{E}\{v(\widehat{\theta}^n)|\mathbf{X}\}\|_\infty \ge 2/3$, which directly implies Equation (23) using triangle inequality.

2.3 Comparison with Earlier Results

In this Section we briefly compare the above debiasing procedure and in particular Theorems 6, 7 and 8 to the results of van de Geer et al. (2014). In the case of linear statistical models considered here, the authors of van de Geer et al. (2014) construct a debiased estimator of the form (7). However, instead of solving the optimization problem (4), they follow Zhang and Zhang (2014) and use the regression coefficients of the i-th column of $\mathbf{X}$ on the other columns to construct the i-th row of M. These regression coefficients are computed, once again, using the LASSO (node-wise LASSO).

It useful to spell out the most important differences between our contribution and the ones of van de Geer et al. (2014):

1. The case of fixed non-random designs is covered by van de Geer et al. (2014, Theorem 2.1), which should be compared to our Theorem 6. While in our case the bias is controlled by the generalized coherence parameter, a similar role is played in van de Geer et al. (2014) by the regularization parameters of the nodewise LASSO.
2. The case of random designs is covered by van de Geer et al. (2014, Theorem 2.2, Theorem 2.4), which should be compared with our Theorem 8. In this case, the assumptions underlying our result are less restrictive. More precisely:
   (a) van de Geer et al. (2014, Theorem 2.2, Theorem 2.4) assume $\mathbf{X}$ has i.i.d. rows, while we only assume the rows are independent.
   (b) van de Geer et al. (2014, Theorem 2.2, Theorem 2.4) assumes the rows of the inverse covariance matrix Σ^{-1} are sparse. More precisely, letting s_j be the number of non-zero entries of the j-th row of Σ^{-1}, van de Geer et al. (2014) assumes $\max_{j\in[p]} s_j = o(n/\log p)$, that is much smaller than p. We do not make any sparsity assumption on Σ^{-1}, and s_j can be as large as p.
   van de Geer et al. (2014, Theorem 2.4) also considers a slightly different setting, where $\mathbf{X}$ has bounded entries, under analogous sparsity assumptions.

   It is currently unknown whether the sparsity assumption in van de Geer et al. (2014) is required for that approach to work, or it is rather an artifact of the specific analysis. Indeed van de Geer et al. (2014, Theorem 2.1) can in principle be used to weaken this condition.

In addition our Theorem 8 provides the specific dependence on the maximum and minimum singular value of $\widehat{\Sigma}$.

Note that solving the convex problem (4) is not more burdensome than solving the nodewise LASSO as in Zhang and Zhang (2014); van de Geer et al. (2014), This can be confirmed by checking that the dual of problem (4) is an ℓ_1-regularized quadratic optimization problem. It has therefore the same complexity as the nodewise LASSO (but it is different from the nodewise LASSO).

3. Statistical Inference

A direct application of Theorem 8 is to derive confidence intervals and statistical hypothesis tests for high-dimensional models. Throughout, we make the sparsity assumption $s_0 = o(\sqrt{n}/\log p)$ and omit explicit constants that can be readily derived from Theorem 8.

3.1 Preliminary Lemmas

As discussed above, the bias term Δ is negligible with respect to the random term Z in the decomposition (16), provided the latter has variance of order one. Our first lemma establishes that this is indeed the case.

Lemma 12 *Let $M = (m_1, \dots, m_p)^{\mathsf{T}}$ be the matrix with rows m_i^{T} obtained by solving convex program (4) in Algorithm 1. Then for all $i \in [p]$,*

$$[M\widehat{\Sigma}M^{\mathsf{T}}]_{i,i} \ge \frac{(1-\mu)^2}{\widehat{\Sigma}_{i,i}}\,.$$

Lemma 12 is proved in Appendix A.1.

Using this fact, we can then characterize the asymptotic distribution of the residuals $(\widehat{\theta}^u - \theta_{0,i})$. Theorem 8 naturally suggests to consider the scaled residual $\sqrt{n}(\widehat{\theta}^u_i - \theta_{0,i})/(\sigma[M\widehat{\Sigma}M^{\mathsf{T}}]^{1/2}_{i,i})$. In the next lemma we consider a slightly more general scaling, replacing σ by a consistent estimator $\widehat{\sigma}$.

Lemma 13 *Consider a sequence of design matrices $\mathbf{X} \in \mathbb{R}^{n\times p}$, with dimensions $n \to \infty$, $p = p(n) \to \infty$ satisfying the following assumptions, for constants $C_{\min}, C_{\max}, \kappa \in (0,\infty)$ independent of n. For each n, $\Sigma \in \mathbb{R}^{p\times p}$ is such that $\sigma_{\min}(\Sigma) \ge C_{\min} > 0$, and $\sigma_{\max}(\Sigma) \le C_{\max} < \infty$, and $\max_{i\in[p]} \Sigma_{ii} \le 1$. Assume $\mathbf{X}\Sigma^{-1/2}$ to have independent sub-Gaussian rows, with zero mean and sub-Gaussian norm $\|\Sigma^{-1/2}X_1\|_{\psi_2} \le \kappa$,*

Consider the linear model (1) *and let $\widehat{\theta}^u$ be defined as per Equation* (5) *in Algorithm 1, with $\mu = a\sqrt{(\log p)/n}$ and $\lambda = \sigma\sqrt{(c^2 \log p)/n}$, with a, c large enough constants. Finally, let $\widehat{\sigma} = \widehat{\sigma}(y, \mathbf{X})$ be an estimator of the noise level satisfying, for any $\varepsilon > 0$,*

$$\lim_{n\to\infty} \sup_{\theta_0 \in \mathbb{R}^p;\, \|\theta_0\|_0 \le s_0} \mathbb{P}\Big(\Big|\frac{\widehat{\sigma}}{\sigma} - 1\Big| \ge \varepsilon\Big) = 0\,. \tag{29}$$

If $s_0 = o(\sqrt{n}/\log p)$ ($s_0 \ge 1$), then, for all $x \in \mathbb{R}$, we have

$$\lim_{n\to\infty} \sup_{\theta_0 \in \mathbb{R}^p;\, \|\theta_0\|_0 \le s_0} \left| \mathbb{P}\left\{ \frac{\sqrt{n}(\widehat{\theta}^u_i - \theta_{0,i})}{\widehat{\sigma}[M\widehat{\Sigma}M^{\mathsf{T}}]^{1/2}_{i,i}} \le x \right\} - \Phi(x) \right| = 0\,. \tag{30}$$

The proof of this lemma can be found in Section 6.5. We also note that the dependence of a, c on $C_{\min}, C_{\max}, \kappa$ can be easily reconstructed from Theorem 7.

The last lemma requires a consistent estimator of σ, in the sense of Equation (29). Several proposals have been made to estimate the noise level in high-dimensional linear regression. A short list of references includes Fan and Li (2001); Fan and Lv (2008); Städler et al. (2010); Zhang (2010); Sun and Zhang (2012); Belloni and Chernozhukov (2013); Fan et al. (2012); Reid et al. (2013); Dicker (2012); Fan et al. (2009); Bayati et al. (2013). Consistency results have been proved or can be proved for several of these estimators.

In order to demonstrate that the consistency criterion (29) can be achieved, we use the scaled LASSO (Sun and Zhang, 2012) given by

$$\{\widehat{\theta}^n(\widetilde{\lambda}), \widehat{\sigma}(\widetilde{\lambda})\} \equiv \arg\min_{\theta\in\mathbb{R}^p, \sigma>0} \Big\{ \frac{1}{2\sigma n}\|Y - \mathbf{X}\theta\|_2^2 + \frac{\sigma}{2} + \widetilde{\lambda}\|\theta\|_1 \Big\}\,. \tag{31}$$

This is a joint convex optimization problem which provides an estimate of the noise level in addition to an estimate of θ_0.

The following lemma uses the analysis of Sun and Zhang (2012) to show that $\widehat{\sigma}$ satisfies the consistency criterion (29).

Lemma 14 *Under the assumptions of Lemma 13, let $\widehat{\sigma} = \widehat{\sigma}(\widetilde{\lambda})$ be the scaled LASSO estimator of the noise level, see Equation (31), with $\widetilde{\lambda} = 10\sqrt{(2\log p)/n}$. Then $\widehat{\sigma}$ satisfies Equation (29).*

The proof of this lemma is fairly straightforward and can be found in Appendix C.

3.2 Confidence Intervals

In view of Lemma 13, it is quite straightforward to construct asymptotically valid confidence intervals. Namely, for $i \in [p]$ and significance level $\alpha \in (0,1)$, we let

$$\begin{aligned} J_i(\alpha) &\equiv [\widehat{\theta}^u_i - \delta(\alpha,n), \widehat{\theta}^u_i + \delta(\alpha,n)]\,, \\ \delta(\alpha,n) &\equiv \Phi^{-1}(1-\alpha/2)\,\frac{\widehat{\sigma}}{\sqrt{n}}[M\widehat{\Sigma}M^{\mathsf{T}}]^{1/2}_{i,i}\,. \end{aligned} \tag{32}$$

Theorem 15 *Consider a sequence of design matrices $\mathbf{X} \in \mathbb{R}^{n\times p}$, with dimensions $n \to \infty$, $p = p(n) \to \infty$ satisfying the assumptions of Lemma 13.*

Consider the linear model (1) *and let $\widehat{\theta}^u$ be defined as per Equation* (5) *in Algorithm 1, with $\mu = a\sqrt{(\log p)/n}$ and $\lambda = \sigma\sqrt{(c^2 \log p)/n}$, with a, c large enough constants. Finally, let $\widehat{\sigma} = \widehat{\sigma}(y,\mathbf{X})$ a consistent estimator of the noise level in the sense of Equation (29). Then the confidence interval $J_i(\alpha)$ is asymptotically valid, namely*

$$\lim_{n\to\infty} \mathbb{P}\Big(\theta_{0,i} \in J_i(\alpha)\Big) = 1-\alpha\,. \tag{33}$$

Proof The proof is an immediate consequence of Lemma 13 since

$$\begin{aligned} \lim_{n\to\infty} \mathbb{P}\Big(\theta_{0,i} \in J_i(\alpha)\Big) &= \lim_{n\to\infty} \mathbb{P}\left\{\frac{\sqrt{n}(\widehat{\theta}^u_i - \theta_{0,i})}{\widehat{\sigma}[M\widehat{\Sigma}M^{\mathsf{T}}]^{1/2}_{i,i}} \le \Phi^{-1}(1-\alpha/2)\right\} \\ &\quad - \lim_{n\to\infty} \mathbb{P}\left\{\frac{\sqrt{n}(\widehat{\theta}^u_i - \theta_{0,i})}{\widehat{\sigma}[M\widehat{\Sigma}M^{\mathsf{T}}]^{1/2}_{i,i}} \le -\Phi^{-1}(1-\alpha/2)\right\} \\ &= 1-\alpha\,. \end{aligned} \tag{34}$$

■

3.3 Hypothesis Testing

An important advantage of sparse linear regression models is that they provide parsimonious explanations of the data in terms of a small number of covariates. The easiest way to select the 'active' covariates is to choose the indexes i for which $\widehat{\theta}^n_i \neq 0$. This approach however does not provide a measure of statistical significance for the finding that the coefficient is non-zero.

More precisely, we are interested in testing an individual null hypothesis $H_{0,i} : \theta_{0,i} = 0$ versus the alternative $H_{A,i} : \theta_{0,i} \neq 0$, and assigning p-values for these tests. We construct a p-value P_i for the test $H_{0,i}$ as follows:

$$P_i = 2\bigg(1 - \Phi\bigg(\frac{\sqrt{n}\,|\widehat{\theta}^u_i|}{\widehat{\sigma}[M\widehat{\Sigma}M^{\mathsf{T}}]^{1/2}_{i,i}}\bigg)\bigg)\,. \tag{35}$$

The decision rule is then based on the p-value P_i:

$$\widehat{T}_{i,\mathbf{X}}(y) = \begin{cases} 1 & \text{if } P_i \le \alpha \qquad \text{(reject } H_{0,i})\,, \\ 0 & \text{otherwise} \qquad \text{(accept } H_{0,i})\,, \end{cases} \tag{36}$$

where α is the fixed target Type I error probability. We measure the quality of the test $\widehat{T}_{i,\mathbf{X}}(y)$ in terms of its significance level α_i and statistical power $1-\beta_i$. Here α_i is the probability of type I error (i.e., of a false positive at i) and β_i is the probability of type II error (i.e., of a false negative at i).

Note that it is important to consider the tradeoff between statistical significance and power. Indeed any significance level α can be achieved by randomly rejecting $H_{0,i}$ with probability α. This test achieves power $1-\beta=\alpha$. Further note that, without further assumption, no nontrivial power can be achieved. In fact, choosing $\theta_{0,i}\neq 0$ arbitrarily close to zero, $H_{0,i}$ becomes indistinguishable from its alternative. We will therefore assume that, whenever $\theta_{0,i}\neq 0$, we have $|\theta_{0,i}|>\gamma$ as well. We take a minimax perspective and require the test to behave uniformly well over s_0-sparse vectors. Formally, given a family of tests $T_{i,\mathbf{X}}:\mathbb{R}^n\to\{0,1\}$, indexed by $i\in[p]$, $\mathbf{X}\in\mathbb{R}^{n\times p}$, we define, for $\gamma>0$ a lower bound on the non-zero entries:

$$\alpha_{i,n}(T)\equiv\sup\Big\{\mathbb{P}_{\theta_0}(T_{i,\mathbf{X}}(y)=1):\ \theta_0\in\mathbb{R}^p,\ \|\theta_0\|_0\le s_0(n),\ \theta_{0,i}=0\Big\}\,. \tag{37}$$

$$\beta_{i,n}(T;\gamma)\equiv\sup\Big\{\mathbb{P}_{\theta_0}(T_{i,\mathbf{X}}(y)=0):\ \theta_0\in\mathbb{R}^p,\ \|\theta_0\|_0\le s_0(n),\ |\theta_{0,i}|\ge\gamma\Big\}\,. \tag{38}$$

Here, we made dependence on n explicit. Also, $\mathbb{P}_\theta(\,\cdot\,)$ denotes the induced probability for random design $\mathbf{X}$ and noise realization w, given the fixed parameter vector θ. Our next theorem establishes bounds on $\alpha_{i,n}(\widehat{T})$ and $\beta_{i,n}(\widehat{T};\gamma)$ for our decision rule (36).

Theorem 16 *Consider a sequence of design matrices $\mathbf{X}\in\mathbb{R}^{n\times p}$, with dimensions $n\to\infty$, $p=p(n)\to\infty$ satisfying the assumptions of Lemma 13.*

Consider the linear model (1) and let $\widehat{\theta}^u$ be defined as per Equation (5) in Algorithm 1, with $\mu=a\sqrt{(\log p)/n}$ and $\lambda=\sigma\sqrt{(c^2\log p)/n}$, with a,c large enough constants. Finally, let $\widehat{\sigma}=\widehat{\sigma}(y,\mathbf{X})$ a consistent estimator of the noise level in the sense of Equation (29), and $\widehat{T}$ be the test defined in Equation (36).

Then the following holds true for any fixed sequence of integers $i=i(n)$:

$$\lim_{n\to\infty}\alpha_{i,n}(\widehat{T})\le\alpha\,. \tag{39}$$

$$\liminf_{n\to\infty}\frac{1-\beta_{i,n}(\widehat{T};\gamma)}{1-\beta^*_{i,n}(\gamma)}\ge 1\,,\qquad 1-\beta^*_{i,n}(\gamma)\equiv G\bigg(\alpha,\frac{\sqrt{n}\,\gamma}{\sigma[\Sigma^{-1}_{i,i}]^{1/2}}\bigg)\,, \tag{40}$$

where, for $\alpha\in[0,1]$ and $u\in\mathbb{R}_+$, the function $G(\alpha,u)$ is defined as follows:

$$G(\alpha,u)=2-\Phi(\Phi^{-1}(1-\frac{\alpha}{2})+u)-\Phi(\Phi^{-1}(1-\frac{\alpha}{2})-u)\,.$$

Theorem 16 is proved in Section 6.6. It is easy to see that, for any $\alpha>0$, $u\mapsto G(\alpha,u)$ is continuous and monotone increasing. Moreover, $G(\alpha,0)=\alpha$ which is the trivial power obtained by randomly rejecting $H_{0,i}$ with probability α. As γ deviates from zero, we obtain nontrivial power. Notice that in order to achieve a specific power $\beta>\alpha$, our scheme requires $\gamma\ge c_\beta(\sigma/\sqrt{n})$, for some constant c_β that depends on β. This is because $\Sigma^{-1}_{i,i}\le\sigma_{\max}(\Sigma^{-1})\le(\sigma_{\min}(\Sigma))^{-1}=O(1)$.

3.3.1 Near optimality of the hypothesis testing procedure

The authors of Javanmard and Montanari (2013b) prove an upper bound for the minimax power of tests with a given significance level α, under random designs. For the reader's convenience, we recall this result here. (The following is a restatement of Javanmard and Montanari (2013b, Theorem 2.3), together with a standard estimate on the tail of chi-squared random variables.)

Theorem 17 ((Javanmard and Montanari, 2013b)) *Assume $\mathbf{X} \in \mathbb{R}^{n\times p}$ to be a random design matrix with i.i.d. Gaussian rows with zero mean and covariance Σ. For $i \in [p]$, let $T_{i,\mathbf{X}} : \mathbb{R}^n \to \mathbb{R}^n$ be a hypothesis testing procedure for testing $H_{0,i} : \theta_{0,i} = 0$, and denote by $\alpha_i(T)$ and $\beta_{i,n}(T;\gamma)$ its fraction of type I and type II errors, cf. Equations (37) and (38). Finally, for $S \subseteq [p] \setminus \{i\}$, define $\Sigma_{i|S} \equiv \Sigma_{ii} - \Sigma_{i,S}\Sigma^{-1}_{S,S}\Sigma_{S,i} \in \mathbb{R}$.*

For any $\ell \in \mathbb{R}$ and $|S| < s_0 < n$, if $\alpha_{i,n}(T) \le \alpha$, then

$$1 - \beta_{i,n}(T;\gamma) \le G\Big(\alpha, \frac{\gamma}{\sigma_{\rm eff}(\xi)}\Big) + e^{-\xi^2/8}\,, \tag{41}$$

$$\sigma_{\rm eff}(\xi) \equiv \frac{\sigma}{\Sigma^{1/2}_{i|S}(\sqrt{n-s_0+1}+\xi)}\,, \tag{42}$$

for any $\xi \in [0,(3/2)\sqrt{n-s_0+1}]$.

The intuition behind this bound is straightforward: the power of any test for $H_{0,i} : \theta_{0,i} = 0$ is upper bounded by the power of an oracle test that is given access to $\text{supp}(\theta_0) \setminus \{i\}$ and outputs a test for $H_{0,i}$. Computing the minimax power of such oracle reduces to a classical hypothesis testing problem.

Let us emphasize that the last theorem applies to *Gaussian* random designs. Since this theorem establishes a negative result (an upper bound on power), it makes sense to consider this somewhat more specialized setting.

Using this upper bound, we can restate Theorem 16 as follows.

Corollary 18 *Consider a Gaussian random design model that satisfies the conditions of Theorem 16, and let $\widehat{T}$ be the testing procedure defined in Equation (36), with $\widehat{\theta}^u$ as in Algorithm 1. Further, let*

$$\eta_{\Sigma,s_0} \equiv \min_{i\in[p];S}\Big\{\Sigma_{i|S}\,\Sigma^{-1}_{ii} :\; S \subseteq [p]\backslash\{i\},\, |S| < s_0\Big\}\,. \tag{43}$$

Under the sparsity assumption $s_0 = o(\sqrt{n}/\log p)$, the following holds true. If $\{T_{i,\mathbf{X}}\}$ is any sequence of tests with $\limsup_{n\to\infty}\alpha_{i,n}(T) \le \alpha$, then

$$\liminf_{n\to\infty} \frac{1-\beta_{i,n}(\widehat{T};\gamma)}{1-\beta_{i,n/\eta_{\Sigma,s_0}}(T;\gamma)} \ge 1\,. \tag{44}$$

In other words, the asymptotic efficiency of the test $\widehat{T}$ is at least $1/\eta_{\Sigma,s_0}$.

Hence, our test $\widehat{T}$ has nearly optimal power in the following sense. It has power at least as large as the power of any other test T, provided the latter is applied to a sample size decreased by a factor η_{Σ,s_0}.

Further, under the assumptions of Theorem 8, the factor η_{Σ,s_0} is a bounded constant. Indeed

$$\eta_{\Sigma,s_0} \le \Sigma^{-1}_{i,i}\Sigma_{i,i} \le \frac{\sigma_{\max}(\Sigma)}{\sigma_{\min}(\Sigma)} \le \frac{C_{\max}}{C_{\min}}\,, \tag{45}$$

since $\Sigma^{-1}_{ii} \le (\sigma_{\min}(\Sigma))^{-1}$, and $\Sigma_{i|S} \le \Sigma_{i,i} \le \sigma_{\max}(\Sigma)$ due to $\Sigma_{S,S} \succ 0$.

Note that n, γ and σ appears in our upper bound (41) in the combination $\gamma\sqrt{n}/\sigma$, which is the natural measure of the signal-to-noise ratio (where, for simplicity, we neglected $s_0 = o(\sqrt{n}/\log p)$ with respect to n). Hence, the above result can be restated as follows. The test $\widehat{T}$ has power at least as large as the power of any other test T, provided the latter is applied at a noise level augmented by a factor $\sqrt{\eta_{\Sigma,s_0}}$.

3.4 Generalization to Simultaneous Confidence Intervals

In many situations, it is necessary to perform statistical inference on more than one of the parameters simultaneously. For instance, we might be interested in performing inference about $\theta_{0,R} \equiv (\theta_{0,i})_{i\in R}$ for some set $R \subseteq [p]$.

The simplest generalization of our method is to the case in which $|R|$ stays finite as $n, p \to \infty$. In this case we have the following generalization of Lemma 13. (The proof is the same as for Lemma 13, and hence we omit it.)

Lemma 19 *Under the assumptions of Lemma 13, define*

$$Q^{(n)} \equiv \frac{\widehat{\sigma}^2}{n}\,[M\widehat{\Sigma}M^{\mathsf{T}}]\,. \tag{46}$$

Let $R = R(n)$ be a sequence of sets $R(n) \subseteq [p]$, with $|R(n)| = k$ fixed as $n, p \to \infty$, and further assume $s_0 = o(\sqrt{n}/\log p)$, with $s_0 \ge 1$. Then, for all $x = (x_1, \dots, x_k) \in \mathbb{R}^k$, we have

$$\lim_{n\to\infty} \sup_{\theta_0\in\mathbb{R}^p;\, \|\theta_0\|_0 \le s_0} \left|\mathbb{P}\left\{(Q^{(n)}_{R,R})^{-1/2}(\widehat{\theta}^u_R - \theta_{0,R}) \le x\right\} - \Phi_k(x)\right| = 0\,, \tag{47}$$

where $(a_1, \dots, a_k) \le (b_1, \dots, b_k)$ indicates that $a_1 \le b_1, \dots a_k \le b_k$, and $\Phi_k(x) = \Phi(x_1) \times \cdots \times \Phi(x_k)$.

This lemma allows to construct confidence regions for low-dimensional projections of θ_0, much in the same way as we used Lemma 13 to compute confidence intervals for one-dimensional projections in Section 3.2.

Explicitly, let $\mathcal{C}_{k,\alpha} \subseteq \mathbb{R}^k$ be any Borel set such that $\int_{\mathcal{C}_{k,\alpha}} \phi_k(x)\,\mathrm{d}x \ge 1 - \alpha$, where

$$\phi_k(x) = \frac{1}{(2\pi)^{k/2}}\,\exp\Big(-\frac{\|x\|^2}{2}\Big)\,,$$

is the k-dimensional Gaussian density. Then, for $R \subseteq [p]$, we define $J_R(\alpha) \subseteq \mathbb{R}^k$ as follows

$$J_R(\alpha) \equiv \widehat{\theta}^u_R + (Q^{(n)}_{R,R})^{1/2}\mathcal{C}_{k,\alpha}\,. \tag{48}$$

Then Lemma 19 implies (under the assumptions stated there) that $J_R(\alpha)$ is a valid confidence region

$$\lim_{n\to\infty} \mathbb{P}\big(\theta_{0,R} \in J_R(\alpha)\big) = 1 - \alpha\,. \tag{49}$$

A more challenging regime is the one of large-scale inference, that corresponds to $|R(n)| \to \infty$ with n. Even in the seemingly simple case in which a correct p-value is given for each individual coordinate, the problem of aggregating them has attracted considerable amount of work, see e.g., Efron (2010) for an overview.

Here we limit ourselves to designing a testing procedure for the family of hypotheses $\{H_{0,i} : \theta_{0,i} = 0\}_{i\in[p]}$ that controls the familywise error rate (FWER). Namely we want to define $T_{i,\mathbf{X}} : \mathbb{R}^n \to \{0,1\}$, for each $i \in [p]$, $\mathbf{X} \in \mathbb{R}^{n\times p}$ such that

$$\mathrm{FWER}(T,n) \equiv \sup_{\theta_0\in\mathbb{R}^p, \|\theta_0\|_0\le s_0} \mathbb{P}\Big\{\exists i \in [p] :\ \theta_{0,i} = 0,\, T_{i,\mathbf{X}}(y) = 1\Big\}\,, \tag{50}$$

Here $T = \{T_{i,\mathbf{X}}\}_{i\in[p]}$ represents the family of tests.

In order to achieve familywise error control, we adopt a standard trick based on Bonferroni inequality. Given p-values defined as per Equation (35), we let

$$\widehat{T}^{\mathrm{F}}_{i,\mathbf{X}}(y) = \begin{cases} 1 & \text{if } P_i \le \alpha/p \qquad (\text{reject } H_{0,i})\,,\\ 0 & \text{otherwise} \qquad\;\; (\text{accept } H_{0,i})\,. \end{cases} \tag{51}$$

Then we have the following error control guarantee.

Theorem 20 *Consider a sequence of design matrices $\mathbf{X} \in \mathbb{R}^{n\times p}$, with dimensions $n \to \infty$, $p = p(n) \to \infty$ satisfying the assumptions of Lemma 13.*

Consider the linear model (1) *and let $\widehat{\theta}^u$ be defined as per Equation* (5) *in Algorithm 1, with $\mu = a\sqrt{(\log p)/n}$ and $\lambda = \sigma\sqrt{(c^2\log p)/n}$, with a, c large enough constants. Finally, let $\widehat{\sigma} = \widehat{\sigma}(y,\mathbf{X})$ be a consistent estimator of the noise level in the sense of Equation (29), and $\widehat{T}$ be the test defined in Equation (51). Then:*

$$\limsup_{n\to\infty} \mathrm{FWER}(\widehat{T}^{\mathrm{F}}, n) \le \alpha\,. \tag{52}$$

The proof of this theorem is similar to the one of Lemma 13 and Theorem 16, and is deferred to Appendix D.

4. Non-Gaussian Noise

As can be seen from the proof of Theorem 8, $Z = M\mathbf{X}^{\mathsf{T}}W/\sqrt{n}$, and since the noise is Gaussian, i.e., $W \sim \mathsf{N}(0, \sigma^2\mathrm{I})$, we have $Z|\mathbf{X} \sim \mathsf{N}(0, \sigma^2 M\widehat{\Sigma}M^{\mathsf{T}})$. We claim that the distribution of the coordinates of Z is asymptotically Gaussian, even if W is non-Gaussian, provided the definition of M is modified slightly. As a consequence, the definition of confidence intervals and p-values in Corollary 15 and (35) remain valid in this broader setting.

In case of non-Gaussian noise, we write

$$\frac{\sqrt{n}(\widehat{\theta}^u_i - \theta_{0,i})}{\sigma[M\widehat{\Sigma}M^{\mathsf{T}}]^{1/2}_{i,i}} = \frac{1}{\sqrt{n}} \frac{m_i^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}W}{\sigma[m_i^{\mathsf{T}}\widehat{\Sigma}m_i]^{1/2}} + o(1)$$

$$= \frac{1}{\sqrt{n}} \sum_{j=1}^n \frac{m_i^{\mathsf{T}}X_jW_j}{\sigma[m_i^{\mathsf{T}}\widehat{\Sigma}m_i]^{1/2}} + o(1)\,.$$

Conditional on $\mathbf{X}$, the summands $\xi_j = m_i^{\mathsf{T}}X_jW_j/(\sigma[m_i^{\mathsf{T}}\widehat{\Sigma}m_i]^{1/2})$ are independent and zero mean. Further, $\sum_{j=1}^n \mathbb{E}(\xi_j^2|\mathbf{X}) = 1$. Therefore, if Lindeberg condition holds, namely for every $\varepsilon > 0$, almost surely

$$\lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^n \mathbb{E}(\xi_j^2 \mathbb{I}_{\{|\xi_j|>\varepsilon\sqrt{n}\}}|\mathbf{X}) = 0\,, \tag{53}$$

then $\sum_{j=1}^n \xi_j/\sqrt{n}|\mathbf{X} \xrightarrow{\mathrm{d}} \mathsf{N}(0,1)$, from which we can build the valid p-values as in (35).

In order to ensure that the Lindeberg condition holds, we modify the optimization problem (54) as follows:

$$\begin{aligned} &\text{minimize} && m^{\mathsf{T}}\widehat{\Sigma}m \\ &\text{subject to} && \|\widehat{\Sigma}m - e_i\|_\infty \le \mu \\ &&& \|\mathbf{X}m\|_\infty \le n^\beta \quad \text{for arbitrary fixed } 1/4 < \beta < 1/2 \end{aligned} \tag{54}$$

Next theorem shows the validity of the proposed p-values in the non-Gaussian noise setting.

Theorem 21 *Suppose that the noise variables W_i are independent with $\mathbb{E}(W_i) = 0$, $\mathbb{E}(W_i^2) = \sigma^2$, and $\mathbb{E}(|W_i|^{2+a}) \le C\,\sigma^{2+a}$ for some $a > 0$.*

Let $M = (m_1, \dots, m_p)^{\mathsf{T}}$ be the matrix with rows m_i^{T} obtained by solving optimization problem (54). Then under the assumptions of Theorem 8, and for sparsity level $s_0 = o(\sqrt{n}/\log p)$, an asymptotic two-sided confidence interval for $\theta_{0,i}$ with significance α is given by $I_i = [\widehat{\theta}^u_i - \delta(\alpha,n), \widehat{\theta}^u_i + \delta(\alpha,n)]$ where

$$\delta(\alpha,n) = \Phi^{-1}(1-\alpha/2)\widehat{\sigma}\,n^{-1/2}\sqrt{[M\widehat{\Sigma}M^{\mathsf{T}}]_{i,i}}\,. \tag{55}$$

Further, an asymptotically valid p-value P_i for testing null hypothesis $H_{0,i}$ is constructed as:

$$P_i = 2\bigg(1 - \Phi\bigg(\frac{\sqrt{n}|\widehat{\theta}^u_i|}{[M\widehat{\Sigma}M^{\mathsf{T}}]^{1/2}_{i,i}}\bigg)\bigg)\,.$$

Theorem 21 is proved in Section 6.7.

5. Numerical Experiments

We corroborate our theoretical results with numerical experiments on both synthetic and real data examples. We further compare performance of our approach with the previous proposals.

5.1 Synthetic Data

We consider linear model (2), where the rows of design matrix $\mathbf{X}$ are fixed i.i.d. realizations from $\mathsf{N}(0,\Sigma)$, where $\Sigma \in \mathbb{R}^{p\times p}$ is a circulant symmetric matrix with entries Σ_{jk} given as follows for $j \le k$:

$$\Sigma_{jk} = \begin{cases} 1 & \text{if } k = j\,, \\ 0.1 & \text{if } k \in \{j+1,\dots,j+5\} \\ & \text{or } k \in \{j+p-5,\dots,j+p-1\}\,, \\ 0 & \text{for all other } j \le k\,. \end{cases} \tag{56}$$

Regarding the regression coefficient, we consider a uniformly random support $S \subseteq [p]$, with $|S| = s_0$ and let $\theta_{0,i} = b$ for $i \in S$ and $\theta_{0,i} = 0$ otherwise. The measurement errors are $W_i \sim \mathsf{N}(0,1)$, for $i \in [n]$. We consider several configurations of (n,p,s_0,b) and for each configuration report our results based on 20 independent realizations of the model with fixed design and fixed regression coefficients. In other words, we repeat experiments over 20 independent realization of the measurement errors.

We use the regularization parameter $\lambda = 4\widehat{\sigma}\sqrt{(2\log p)/n}$, where $\widehat{\sigma}$ is given by the scaled LASSO as per equation (31) with $\widetilde{\lambda} = 10\sqrt{(2\log p)/n}$. Furthermore, parameter μ (cf. Equation 4) is set to

$$\mu = 2\sqrt{\frac{\log p}{n}}\,.$$

This choice of μ is guided by Theorem 7 (b).

Throughout, we set the significance level $\alpha = 0.05$.

Confidence intervals. For each configuration, we consider 20 independent realizations of measurement noise and for each parameter $\theta_{0,i}$, we compute the average length of the corresponding confidence interval, denoted by $\text{Avglength}(J_i(\alpha))$ where $J_i(\alpha)$ is given by equation (32) and the average is taken over the realizations. We then define

$$\ell \equiv p^{-1}\sum_{i\in[p]} \text{Avglength}(J_i(\alpha))\,. \tag{57}$$

We also consider the average length of intervals for the active and inactive parameters, as follows:

$$\ell_S \equiv s_0^{-1}\sum_{i\in S} \text{Avglength}(J_i(\alpha))\,, \quad \ell_{S^c} \equiv (p-s_0)^{-1}\sum_{i\in S^c} \text{Avglength}(J_i(\alpha))\,. \tag{58}$$

Similarly, we consider average coverage for individual parameters. We define the following three metrics:

$$\widehat{\mathsf{Cov}} \equiv p^{-1}\sum_{i\in[p]} \widehat{\mathbb{P}}[\theta_{0,i} \in J_i(\alpha)]\,, \tag{59}$$

$$\widehat{\mathsf{Cov}}_S \equiv s_0^{-1}\sum_{i\in S} \widehat{\mathbb{P}}[\theta_{0,i} \in J_i(\alpha)]\,, \tag{60}$$

$$\widehat{\mathsf{Cov}}_{S^c} \equiv (p-s_0)^{-1}\sum_{i\in S^c} \widehat{\mathbb{P}}[0 \in J_i(\alpha)]\,, \tag{61}$$

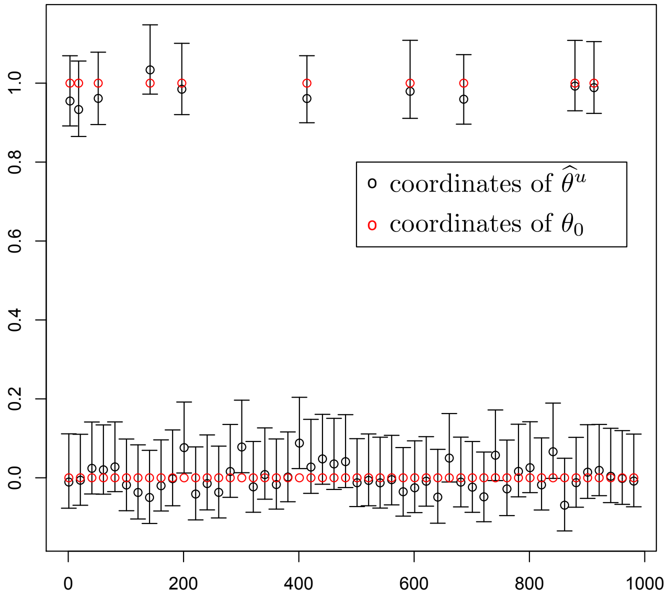

Figure 1: 95% confidence intervals for one realization of configuration $(n, p, s_0, b) = (1000, 600, 10, 1)$. For clarity, we plot the confidence intervals for only 100 of the 1000 parameters. The true parameters $\theta_{0,i}$ are in red and the coordinates of the debiased estimator $\widehat{\theta}^u$ are in black.

where $\widehat{\mathbb{P}}$ denotes the empirical probability computed based on the 20 realizations for each configuration. The results are reported in Table 1. In Figure 1, we plot the constructed 95%-confidence intervals for one realization of configuration $(n, p, s_0, b) = (1000, 600, 10, 1)$. For sake of clarity, we plot the confidence intervals for only 100 of the 1000 parameters.

False positive rates and statistical powers. Table 2 summarizes the false positive rates and the statistical powers achieved by our proposed method, the multisample-splitting method (Meinshausen et al., 2009), and the ridge-type projection estimator (Bühlmann, 2013) for several configurations. The results are obtained by taking average over 20 independent realizations of measurement errors for each configuration. As we see the multisample-splitting achieves false positive rate 0 on all of the configurations considered here, making no type I error. However, the true positive rate is always smaller than that of our proposed method. By contrast, our method achieves false positive rate close to the pre-assigned significance level $\alpha = 0.05$ and obtains much higher true positive rate. Similar to the multisample-splitting, the ridge-type projection estimator is conservative and achieves false positive rate smaller than α. This, however, comes at the cost of a smaller true positive rate than our method. It is worth noting that an ideal testing procedure should allow to control the level of statistical significance α, and obtain the maximum true positive rate at that level.

Here, we used the R-package hdi to test multisample-splitting and the ridge-type projection estimator.

Measure Configuration	ℓ	ℓ_S	ℓ_{S^c}	$\widehat{\mathsf{Cov}}$	$\widehat{\mathsf{Cov}}_S$	$\widehat{\mathsf{Cov}}_{S^c}$
$(1000, 600, 10, 0.5)$	0.1870	0.1834	0.1870	0.9766	0.9600	0.9767
$(1000, 600, 10, 0.25)$	0.1757	0.1780	0.1757	0.9810	0.9000	0.9818
$(1000, 600, 10, 0.1)$	0.1809	0.1823	0.1809	0.9760	1	0.9757
$(1000, 600, 30, 0.5)$	0.2107	0.2108	0.2107	0.9780	0.9866	0.9777
$(1000, 600, 30, 0.25)$	0.1956	0.1961	0.1956	0.9660	0.9660	0.9659
$(1000, 600, 30, 0.1)$	0.2023	0.2043	0.2023	0.9720	0.9333	0.9732
$(2000, 1500, 50, 0.5)$	0.1383	0.1391	0.1383	0.9754	0.9800	0.9752
$(2000, 1500, 50, 0.25)$	0.1356	0.1363	0.1355	0.9720	0.9600	0.9723
$(2000, 1500, 50, 0.1)$	0.1361	0.1361	0.1361	0.9805	1	0.9800
$(2000, 1500, 25, 0.5)$	0.1233	0.1233	0.1233	0.9731	0.9680	0.9731
$(2000, 1500, 25, 0.25)$	0.1208	0.1208	0.1208	0.9735	1	0.9731
$(2000, 1500, 25, 0.1)$	0.1242	0.1237	0.1242	0.9670	0.9200	0.9676

Table 1: Simulation results for the synthetic data described in Section 5.1. The results corresponds to 95% confidence intervals.

Let $Z = (z_i)_{i=1}^p$ denote the vector with $z_i \equiv \sqrt{n}(\widehat{\theta}^u_i - \theta_{0,i})/\widehat{\sigma}\sqrt{[M\widehat{\Sigma}M^{\mathsf{T}}]_{i,i}}$. Figure 2 shows the sample quantiles of Z versus the quantiles of the standard normal distribution for one realization of the configuration $(n, p, s_0, b) = (1000, 600, 10, 1)$. The scattered points are close to the line with unit slope and zero intercept. This confirms the result of Theorem 13 regarding the Gaussianity of the entries z_i.

For the same problem, in Figure 3 we plot the empirical CDF of the computed p-values restricted to the variables outside the support. Clearly, the p-values for these entries are uniformly distributed as expected.

5.2 Real Data

As a real data example, we consider a high-throughput genomic data set concerning riboflavin (vitamin B_2) production rate. This data set is made publicly available by Bühlmann et al. (2014) and contains $n = 71$ samples and $p = 4,088$ covariates corresponding to $p = 4,088$ genes. For each sample, there is a real-valued response variable indicating the logarithm of the riboflavin production rate along with the logarithm of the expression level of the $p = 4,088$ genes as the covariates.

Following Bühlmann et al. (2014), we model the riboflavin production rate as a linear model with $p = 4,088$ covariates and $n = 71$ samples, as in Equation (1). We use the R package glmnet (Friedman et al., 2010) to fit the LASSO estimator. Similar to the previous section, we use the regularization parameter $\lambda = 4\widehat{\sigma}\sqrt{(2\log p)/n}$, where $\widehat{\sigma}$ is given by the scaled LASSO as per equation (31) with $\widetilde{\lambda} = 10\sqrt{(2\log p)/n}$. This leads to the choice $\lambda = 0.036$. The resulting model contains 30 genes (plus an intercept term) corresponding to the nonzero parameters of the lasso estimator.

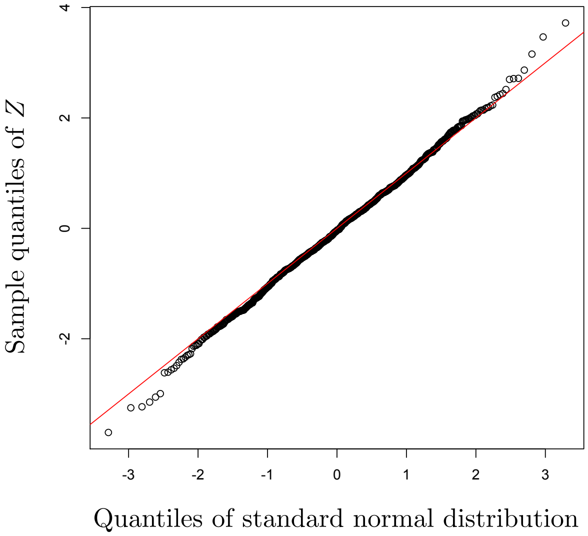

Figure 2: Q-Q plot of Z for one realization of configuration $(n, p, s_0, b) = (1000, 600, 10, 1)$.

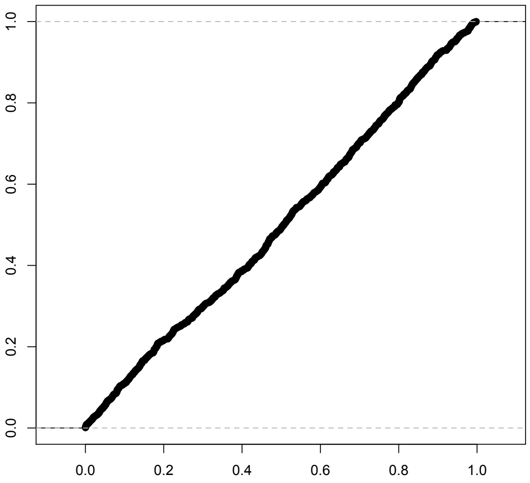

Figure 3: Empirical CDF of the computed p-values (restricted to entries outside the support) for one realization of configuration $(n, p, s_0, b) = (1000, 600, 10, 1)$. Clearly the plot confirms that the p-values are distributed according to uniform distribution.

	Our method		Multisample-splitting		Ridge projection estimator	
Configuration	FP	TP	FP	TP	FP	TP
$(1000, 600, 10, 0.5)$	0.0452	1	0	1	0.0284	0.8531
$(1000, 600, 10, 0.25)$	0.0393	1	0	0.4	0.02691	0.7506
$(1000, 600, 10, 0.1)$	0.0383	0.8	0	0	0.2638	0.6523
$(1000, 600, 30, 0.5)$	0.0433	1	0	1	0.0263	0.8700
$(1000, 600, 30, 0.25)$	0.0525	1	0	0.4	0.2844	0.8403
$(1000, 600, 30, 0.1)$	0.0402	0.7330	0	0	0.2238	0.6180
$(2000, 1500, 50, 0.5)$	0.0421	1	0	1	0.0301	0.9013
$(2000, 1500, 50, 0.25)$	0.0415	1	0	1	0.0292	0.8835
$(2000, 1500, 50, 0.1)$	0.0384	0.9400	0	0	0.02655	0.7603
$(2000, 1500, 25, 0.5)$	0.0509	1	0	1	0.0361	0.9101
$(2000, 1500, 25, 0.25)$	0.0481	1	0	1	0.3470	0.8904
$(2000, 1500, 25, 0.1)$	0.0551	1	0	0.16	0.0401	0.8203

Table 2: Simulation results for the synthetic data described in Section 5.1. The false positive rates (FP) and the true positive rates (TP) are computed at significance level $\alpha = 0.05$.

We use Equation (35) to construct p-values for different genes. Adjusting FWER to 5% significance level, we find two significant genes, namely genes YXLD-at and YXLE-at. By contrast, the multisample-splitting method proposed in Meinshausen et al. (2009) finds only the gene YXLD-at at the FWER-adjusted 5% significance level. Also the Ridge-type projection estimator, proposed in Bühlmann (2013), returns no significance gene. (See Bühlmann et al. 2014 for further discussion on these methods.) This indicates that these methods are more conservative and produce typically larger p-values.

In Figure 4 we plot the empirical CDF of the computed p-values for riboflavin example. Clearly the plot confirms that the p-values are distributed according to uniform distribution.

6. Proofs

This section is devoted to the proofs of theorems and main lemmas.

6.1 Proof of Theorem 6

Substituting $Y = \mathbf{X}\theta_0 + W$ in the definition (7), we get

$$\begin{aligned} \widehat{\theta}^* &= \widehat{\theta}^n + \frac{1}{n} M \mathbf{X}^{\mathsf{T}} \mathbf{X}(\theta_0 - \widehat{\theta}^n) + \frac{1}{n} M \mathbf{X}^{\mathsf{T}} W \\ &= \theta_0 + \frac{1}{\sqrt{n}}\, Z + \frac{1}{\sqrt{n}}\, \Delta\,, \end{aligned} \tag{62}$$

with Z, Δ defined as per the theorem statement. Further Z is Gaussian with the stated covariance because it is a linear function of the Gaussian vector $W \sim \mathsf{N}(0, \sigma^2\, \mathrm{I}_{p\times p})$.

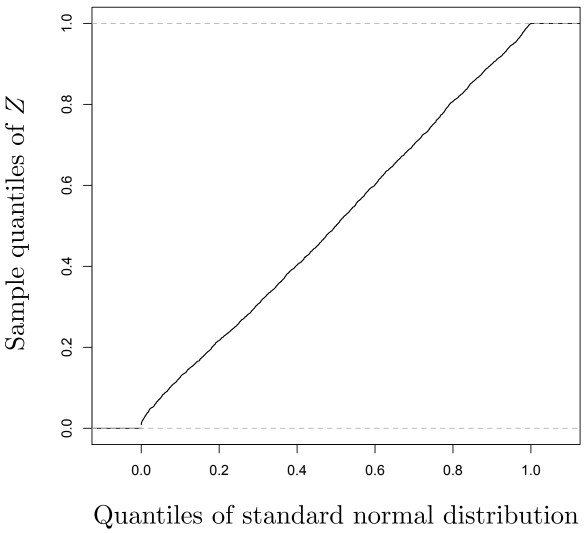

Figure 4: Empirical CDF of the computed p-values for riboflavin example. Clearly the plot confirms that the p-values are distributed according to uniform distribution.

We are left with the task of proving the bound (11) on Δ. Note that by definition (4), we have

$$\|\Delta\|_\infty \le \sqrt{n}\,|M\widehat{\Sigma} - \mathrm{I}|_\infty\,\|\widehat{\theta}^n - \theta_0\|_1 = \sqrt{n}\,\mu_*\|\widehat{\theta}^n - \theta_0\|_1\,. \tag{63}$$

By Bühlmann and van de Geer (2011, Theorem 6.1, Lemma 6.2), we have, for any $\lambda \ge 4\sigma\sqrt{2K\log(pe^{t^2/2})/n}$

$$\mathbb{P}\Big(\|\widehat{\theta}^n - \theta_0\|_1 \ge \frac{4\lambda s_0}{\phi_0^2}\Big) \le 2\,e^{-t^2/2}\,. \tag{64}$$

(More precisely, we consider the trivial generalization of Bühlmann and van de Geer 2011, Lemma 6.2 to the case $(\mathbf{X}^T\mathbf{X}/n)_{ii} \le K$, instead of $(\mathbf{X}^T\mathbf{X}/n)_{ii} = 1$ for all $i \in [p]$.)

Substituting Equation (63) in the last bound, we get

$$\mathbb{P}\Big(\|\Delta\|_\infty \ge \frac{4\lambda\mu_* s_0\sqrt{n}}{\phi_0^2}\Big) \le 2\,e^{-t^2/2}\,. \tag{65}$$

Finally, the claim follows by selecting t so that $e^{t^2/2} = p^{c_0}$.

6.2 Proof of Theorem 7.(a)

Note that the event $\mathcal{E}_n$ requires two conditions. Hence, its complement is given by

$$\mathcal{E}_n(\phi_0, s_0, K)^c = \mathcal{B}_{1,n}(\phi_0, s_0) \cup \mathcal{B}_{2,n}(K)\,, \tag{66}$$

$$\mathcal{B}_{1,n}(\phi_0, s_0) \equiv \Big\{\mathbf{X} \in \mathbb{R}^{n\times p} : \min_{S:\,|S|\le s_0} \phi(\widehat{\Sigma}, S) < \phi_0,\ \ \widehat{\Sigma} = (\mathbf{X}^{\mathsf{T}}\mathbf{X}/n)\Big\}\,, \tag{67}$$

$$\mathcal{B}_{2,n}(K) \equiv \Big\{\mathbf{X} \in \mathbb{R}^{n\times p} : \max_{i\in[p]} \widehat{\Sigma}_{i,i} > K,\ \ \widehat{\Sigma} = (\mathbf{X}^{\mathsf{T}}\mathbf{X}/n)\Big\}\,. \tag{68}$$

We will bound separately the probability of $\mathcal{B}_{1,n}$ and the probability of $\mathcal{B}_{2,n}$. The claim of Theorem 7.(a) follows by union bound.

6.2.1 Controlling $\mathcal{B}_{1,n}(\phi_0, s_0)$

It is also useful to recall the notion of restricted eigenvalue, introduced by Bickel, Ritov and Tsybakov (Bickel et al., 2009).

Definition 22 *Given a symmetric matrix $Q \in \mathbb{R}^{p\times p}$ an integer $s_0 \geq 1$, and $L > 0$, the restricted eigenvalue of Q is defined as*

$$\phi^2_{\rm RE}(Q, s_0, L) \equiv \min_{S\subseteq[p],|S|\le s_0} \min_{\theta\in\mathbb{R}^p} \Big\{ \frac{\langle\theta, Q\,\theta\rangle}{\|\theta_S\|_2^2} : \;\; \theta \in \mathbb{R}^p, \;\; \|\theta_{S^c}\|_1 \le L\|\theta_S\|_1 \Big\} . \tag{69}$$

Rudelson and Shuheng (2013) prove that, if the population covariance satisfies the restricted eigenvalue condition, then the sample covariance satisfies it as well, with high probability. More precisely, by Rudelson and Shuheng (2013, Theorem 6) we have

$$\mathbb{P}\Big(\phi_{\rm RE}(\widehat{\Sigma}, s_0, 3) \ge \frac{1}{2}\phi_{\rm RE}(\Sigma, s_0, 9)\Big) \ge 1 - 2e^{-n/(4c_*\kappa^4)} , \tag{70}$$

for some $c_* \le 2000$, $m \equiv 6\times 10^4 s_0 C^2_{\rm max}/\phi^2_{\rm RE}(\Sigma, s_0, 9)$, and every $n \ge 4c_* m\kappa^4 \log(120ep/m)$.

Note that $\phi_{\rm RE}(\Sigma, s_0, 9) \ge \sigma_{\rm min}(\Sigma)^{1/2} \ge \sqrt{C_{\rm min}}$, and by Cauchy-Schwartz,

$$\min_{S:|S|\le s_0} \phi(\widehat{\Sigma}, S) \ge \phi_{\rm RE}(\widehat{\Sigma}, s_0, 3) .$$

With the definitions in the statement (cf. Equation 13), we therefore have

$$\mathbb{P}\Big(\min_{S:|S|\le s_0} \phi(\widehat{\Sigma}, S) \ge \frac{1}{2}\sqrt{C_{\rm min}}\Big) \ge 1 - 2e^{-c_1 n} . \tag{71}$$

Equivalently, $\mathbb{P}(\mathcal{B}_{1,n}(\phi_0, s_0)) \le 2\,e^{-c_1 n}$.

6.2.2 Controlling $\mathcal{B}_{2,n}(K)$

By definition

$$\widehat{\Sigma}_{ii} - 1 = \frac{1}{n}\sum_{\ell=1}^n (\langle X_\ell, e_i\rangle^2 - 1) = \frac{1}{n}\sum_{\ell=1}^n u_\ell, \, . \tag{72}$$

Note that u_ℓ are independent centered random variables. Further, (recalling that, for any random variables U, V, $\|U+V\|_{\psi_1} \le \|U\|_{\psi_1} + \|V\|_{\psi_1}$, and $\|U^2\|_{\psi_1} \le 2\|U\|^2_{\psi_2}$) they are subexponential with subexponential norm

$$\begin{aligned} \|u_\ell\|_{\psi_1} &\le 2\|\langle X_\ell, e_i\rangle^2\|_{\psi_1} \le 4\|\langle X_\ell, e_i\rangle\|^2_{\psi_1} \\ &\le 4\|\langle \Sigma^{-1/2}X_\ell, \Sigma^{1/2}e_i\rangle\|^2_{\psi_1} \\ &\le 4\kappa^2\|\Sigma^{1/2}e_i\|_2^2 = 4\kappa^2\Sigma_{ii} = 4\kappa^2 . \end{aligned}$$

By Bernstein-type inequality for centered subexponential random variables, cf. Vershynin (2012), we get

$$\mathbb{P}\Big\{\frac{1}{n}\Big|\sum_{\ell=1}^{n}u_\ell\Big|\geq\varepsilon\Big\}\leq 2\exp\Big[-\frac{n}{6}\min\Big((\frac{\varepsilon}{4e\kappa^2})^2,\frac{\varepsilon}{4e\kappa^2}\Big)\Big]\,. \tag{73}$$

Hence, for all ε such that $\varepsilon/(e\kappa^2)\in[\sqrt{(48\log p)/n},4]$,

$$\mathbb{P}\Big(\max_{i\in[p]}\widehat{\Sigma}_{ii}\geq 1+\varepsilon\Big)\leq 2p\,\exp\Big(-\frac{n\varepsilon^2}{24e^2\kappa^4}\Big)\leq 2e^{-c_1n}\,, \tag{74}$$

which implies $\mathbb{P}(\mathbf{X}\in\mathcal{B}_{2,n}(K))\leq 2\,e^{-c_1n}$ for all $K-1\geq 20\kappa^2\sqrt{(\log p)/n}\geq\sqrt{(48e^2\kappa^4\log p)/n}$.

6.3 Proof of Theorem 7.(b)

Obviously, we have

$$\mu_{\min}(\mathbf{X})\leq\big|\Sigma^{-1}\widehat{\Sigma}-\mathrm{I}\big|\,, \tag{75}$$

and hence the statement follows immediately from the following estimate.

Lemma 23 *Consider a random design matrix $\mathbf{X}\in\mathbb{R}^{p\times p}$, with i.i.d. rows having mean zero and population covariance Σ. Assume that*

- (i) *We have $\sigma_{\min}(\Sigma)\geq C_{\min}>0$, and $\sigma_{\max}(\Sigma)\leq C_{\max}<\infty$.*
- (ii) *The rows of $X\Sigma^{-1/2}$ are sub-Gaussian with $\kappa=\|\Sigma^{-1/2}X_1\|_{\psi_2}$.*

Let $\widehat{\Sigma}=(\mathbf{X}^{\mathsf{T}}\mathbf{X})/n$ be the empirical covariance. Then, for any constant $C>0$, the following holds true.

$$\mathbb{P}\bigg\{\Big|\Sigma^{-1}\widehat{\Sigma}-\mathrm{I}\Big|_\infty\geq a\sqrt{\frac{\log p}{n}}\bigg\}\leq 2p^{-c_2}\,, \tag{76}$$

with $c_2=(a^2C_{\min})/(24e^2\kappa^4C_{\max})-2$.

Proof [Proof of Lemma 23] The proof is based on Bernstein-type inequality for sub-exponential random variables (Vershynin, 2012). Let $\tilde{X}_\ell=\Sigma^{-1/2}X_\ell$, for $\ell\in[n]$, and write

$$Z\equiv\Sigma^{-1}\widehat{\Sigma}-\mathrm{I}=\frac{1}{n}\sum_{\ell=1}^{n}\Big\{\Sigma^{-1}X_\ell X_\ell^{\mathsf{T}}-\mathrm{I}\Big\}=\frac{1}{n}\sum_{\ell=1}^{n}\Big\{\Sigma^{-1/2}\tilde{X}_\ell\tilde{X}_\ell^{\mathsf{T}}\Sigma^{1/2}-\mathrm{I}\Big\}\,.$$

Fix $i,j\in[p]$, and for $\ell\in[n]$, let $v_\ell^{(ij)}=\langle\Sigma_{i,\cdot}^{-1/2},\tilde{X}_\ell\rangle\langle\Sigma_{j,\cdot}^{1/2},\tilde{X}_\ell\rangle-\delta_{i,j}$, where $\delta_{i,j}=\mathbf{1}_{\{i=j\}}$. Notice that $\mathbb{E}(v_\ell^{(ij)})=0$, and the $v_\ell^{(ij)}$ are independent for $\ell\in[n]$. Also, $Z_{i,j}=(1/n)\sum_{\ell=1}^{n}v_\ell^{(ij)}$. By Vershynin (2012, Remark 5.18), we have

$$\|v_\ell^{(ij)}\|_{\psi_1}\leq 2\|\langle\Sigma_{i,\cdot}^{-1/2},\tilde{X}_\ell\rangle\langle\Sigma_{j,\cdot}^{1/2},\tilde{X}_\ell\rangle\|_{\psi_1}.$$

Moreover, for any two random variables X and Y, we have

$$\begin{aligned}\|XY\|_{\psi_1} &= \sup_{p\ge 1} p^{-1}\mathbb{E}(|XY|^p)^{1/p}\\ &\le \sup_{p\ge 1} p^{-1}\mathbb{E}(|X|^{2p})^{1/2p}\,\mathbb{E}(|Y|^{2p})^{1/2p}\\ &\le 2\Big(\sup_{q\ge 2} q^{-1/2}\mathbb{E}(|X|^q)^{1/q}\Big)\Big(\sup_{q\ge 2} q^{-1/2}\mathbb{E}(|Y|^q)^{1/q}\Big)\\ &\le 2\|X\|_{\psi_2}\,\|Y\|_{\psi_2}\,.\end{aligned}$$

Hence, by assumption (ii), we obtain

$$\begin{aligned}\|v_\ell^{(ij)}\|_{\psi_1} &\le 2\|\langle\Sigma_{i,\cdot}^{-1/2},\tilde{X}_\ell\rangle\|_{\psi_2}\|\langle\Sigma_{j,\cdot}^{1/2},\tilde{X}_\ell\rangle\|_{\psi_2}\\ &\le 2\|\Sigma_{i,\cdot}^{-1/2}\|_2\|\Sigma_{j,\cdot}^{1/2}\|_2\kappa^2 \le 2\sqrt{C_{\max}/C_{\min}}\,\kappa^2\,.\end{aligned}$$

Let $\kappa' = 2\sqrt{C_{\max}/C_{\min}}\kappa^2$. Applying Bernstein-type inequality for centered sub-exponential random variables (Vershynin, 2012), we get

$$\mathbb{P}\Big\{\frac{1}{n}\Big|\sum_{\ell=1}^n v_\ell^{(ij)}\Big| \ge \varepsilon\Big\} \le 2\exp\Big[-\frac{n}{6}\min\Big((\frac{\varepsilon}{e\kappa'})^2, \frac{\varepsilon}{e\kappa'}\Big)\Big]\,.$$

Choosing $\varepsilon = a\sqrt{(\log p)/n}$, and assuming $n \ge [a/(e\kappa')]^2\log p$, we arrive at

$$\mathbb{P}\Big\{\frac{1}{n}\Big|\sum_{\ell=1}^n v_\ell^{(ij)}\Big| \ge a\sqrt{\frac{\log p}{n}}\Big\} \le 2p^{-a^2/(6e^2\kappa'^2)}\,.$$

The result follows by union bounding over all possible pairs $i,j\in[p]$. ■

6.4 Proof of Theorem 8

Let

$$\Delta_0 \equiv \Big(\frac{16ac\,\sigma}{C_{\min}}\Big)\frac{s_0\log p}{\sqrt{n}} \tag{77}$$

be a shorthand for the bound on $\|\Delta\|_\infty$ appearing in Equation (17). Then we have

$$\begin{aligned}\mathbb{P}\Big(\|\Delta\|_\infty \ge \Delta_0\Big) \le&\mathbb{P}\Big(\big\{\|\Delta\|_\infty \ge \Delta_0\big\}\cap\mathcal{E}_n(\sqrt{C_{\min}}/2, s_0, 3/2)\cap\mathcal{G}_n(a)\Big)\\ &+\mathbb{P}\big(\mathcal{E}_n(\sqrt{C_{\min}}/2, s_0, 3/2)\big)+\mathbb{P}\big(\mathcal{G}_n^c(a)\big)\\ \le&\mathbb{P}\Big(\big\{\|\Delta\|_\infty \ge \Delta_0\big\}\cap\mathcal{E}_n(\sqrt{C_{\min}}/2, s_0, 3/2)\cap\mathcal{G}_n(a)\Big)+4\,e^{-c_1 n}+2\,p^{-c_2}\,,\end{aligned}$$

where, in the first equation $\mathcal{A}^c$ denotes the complement of event $\mathcal{A}$ and the second inequality follows from Theorem 7. Notice, in particular, that the bound (13) can be applied for

$K = 3/2$ since, under the present assumptions $20\kappa^2\sqrt{(\log p)/n} \le 1/2$. Finally

$$\mathbb{P}\Big(\big\{\|\Delta\|_\infty \ge \Delta_0\big\} \cap \mathcal{E}_n(\sqrt{C_{\min}}/2, s_0, 3/2) \cap \mathcal{G}_n(a)\Big) \\ \le \sup_{\mathbf{X}\in\mathcal{E}_n(\sqrt{C_{\min}}/2,s_0,3/2)\cap\mathcal{G}_n(a)} \mathbb{P}\Big(\|\Delta\|_\infty \ge \Delta_0 \Big| \mathbf{X}\Big) \le 2\,p^{-\tilde{c_0}}\,. \tag{78}$$

Here the last inequality follows from Theorem 6 applied per given $\mathbf{X} \in \mathcal{E}_n(\sqrt{C_{\min}}/2, s_0, 3/2) \cap \mathcal{G}_n(a)$ and hence using the bound (11) with $\phi_0 = \sqrt{C_{\min}}/2$, $K = 3/2$, $\mu_* = a\sqrt{(\log p)/n}$.

6.5 Proof of Lemma 13

We will prove that, under the stated assumptions

$$\limsup_{n\to\infty} \sup_{\|\theta_0\|_0 \le s_0} \mathbb{P}\left\{\frac{\sqrt{n}(\widehat{\theta}^u_i - \theta_{0,i})}{\widehat{\sigma}[M\widehat{\Sigma}M^{\mathsf{T}}]^{1/2}_{i,i}} \le x\right\} \le \Phi(x)\,. \tag{79}$$

A matching lower bound follows by a completely analogous argument.

Notice that by Equation (16), we have

$$\frac{\sqrt{n}(\widehat{\theta}^u_i - \theta_{0,i})}{\sigma[M\widehat{\Sigma}M^{\mathsf{T}}]^{1/2}_{ii}} = \frac{e_i^{\mathsf{T}} M \mathbf{X}^{\mathsf{T}} W}{\sigma[M\widehat{\Sigma}M^{\mathsf{T}}]^{1/2}_{ii}} + \frac{\Delta_i}{\sigma[M\widehat{\Sigma}M^{\mathsf{T}}]^{1/2}_{ii}}\,. \tag{80}$$

Let $V = \mathbf{X}M^{\mathsf{T}}e_i/(\sigma[M\widehat{\Sigma}M^{\mathsf{T}}]^{1/2}_{ii})$ and $\widetilde{Z} \equiv V^{\mathsf{T}}W$. We claim that $\tilde{Z} \sim \mathsf{N}(0,1)$. To see this, note that $\|V\|_2 = 1$, and V and W are independent. Hence,

$$\mathbb{P}(\widetilde{Z} \le x) = \mathbb{E}\{\mathbb{P}(V^{\mathsf{T}}W \le x|V)\} = \mathbb{E}\{\Phi(x)|V\} = \Phi(x)\,, \tag{81}$$

which proves our claim. In order to prove Equation (79), fix $\varepsilon > 0$ and write

$$\begin{aligned}\mathbb{P}\left(\frac{\sqrt{n}(\widehat{\theta}^u_i - \theta_{0,i})}{\widehat{\sigma}[M\widehat{\Sigma}M^{\mathsf{T}}]^{1/2}_{i,i}} \le x\right) &= \mathbb{P}\left(\frac{\sigma}{\widehat{\sigma}}\widetilde{Z} + \frac{\Delta_i}{\widehat{\sigma}[M\widehat{\Sigma}M^{\mathsf{T}}]^{1/2}_{i,i}} \le x\right) \\ &\le \mathbb{P}\Big(\frac{\sigma}{\widehat{\sigma}}\widetilde{Z} \le x + \varepsilon\Big) + \mathbb{P}\left(\frac{|\Delta_i|}{\widehat{\sigma}[M\widehat{\Sigma}M^{\mathsf{T}}]^{1/2}_{i,i}} \ge \varepsilon\right) \\ &\le \mathbb{P}\Big(\widetilde{Z} \le x + 2\varepsilon + \varepsilon|x|\Big) + \mathbb{P}\left(\frac{|\Delta_i|}{\widehat{\sigma}[M\widehat{\Sigma}M^{\mathsf{T}}]^{1/2}_{i,i}} \ge \varepsilon\right) \\ &\quad + \mathbb{P}\Big(\Big|\frac{\widehat{\sigma}}{\sigma} - 1\Big| \ge \varepsilon\Big)\,.\end{aligned}$$

By taking the limit and using assumption (29), we obtain

$$\limsup_{n\to\infty} \sup_{\|\theta_0\|_0 \le s_0} \mathbb{P}\left(\frac{\sqrt{n}(\widehat{\theta}^u_i - \theta_{0,i})}{\widehat{\sigma}[M\widehat{\Sigma}M^{\mathsf{T}}]^{1/2}_{i,i}} \le x\right) \le \\ \Phi(x + 2\varepsilon + \varepsilon|x|) + \limsup_{n\to\infty} \sup_{\|\theta_0\|_0 \le s_0} \mathbb{P}\left(\frac{|\Delta_i|}{\widehat{\sigma}[M\widehat{\Sigma}M^{\mathsf{T}}]^{1/2}_{i,i}} \ge \varepsilon\right)\,.$$

Since $\varepsilon > 0$ is arbitrary, it is therefore sufficient to show that the limit on the right hand side vanishes for any $\varepsilon > 0$.

Note that $[M\widehat{\Sigma}M^{\mathsf{T}}]_{i,i} \ge 1/(4\widehat{\Sigma}_{ii})$ for all n large enough, by Lemma 12, and since $\mu = a\sqrt{(\log p)/n} \to 0$ as $n, p \to \infty$. We have therefore

$$\begin{aligned}\mathbb{P}\left(\frac{|\Delta_i|}{\widehat{\sigma}[M\widehat{\Sigma}M^{\mathsf{T}}]_{i,i}^{1/2}} \ge \varepsilon\right) &\le \mathbb{P}\Big(\frac{2}{\widehat{\sigma}}\widehat{\Sigma}_{ii}^{1/2}\,|\Delta_i| \ge \varepsilon\Big)\\ &\le \mathbb{P}\Big(\frac{5}{\sigma}\,|\Delta_i| \ge \varepsilon\Big) + \mathbb{P}\Big(\frac{\widehat{\sigma}}{\sigma} \le \frac{1}{2}\Big) + \mathbb{P}(\widehat{\Sigma}_{ii} \ge \sqrt{2})\,.\end{aligned}$$

Note that $\mathbb{P}\big((\widehat{\sigma}/\sigma) \le 1/2\big) \to 0$ by assumption (29), and $\mathbb{P}(\widehat{\Sigma}_{ii} \ge \sqrt{2}) \to 0$ by Theorem 7.(b). Hence

$$\begin{aligned}\limsup_{n\to\infty} \sup_{\|\theta_0\|_0 \le s_0} \mathbb{P}\left(\frac{|\Delta_i|}{\widehat{\sigma}[M\widehat{\Sigma}M^{\mathsf{T}}]_{i,i}^{1/2}} \ge \varepsilon\right) &\le \limsup_{n\to\infty} \sup_{\|\theta_0\|_0 \le s_0} \mathbb{P}\Big(\|\Delta\|_\infty \ge \frac{\varepsilon\sigma}{5}\Big)\\ &\le \limsup_{n\to\infty} \big(4\,e^{-c_1 n} + 4\,p^{-(\tilde{c}_0 \wedge c_2)}\big) = 0\,,\end{aligned}$$

where the last inequality follows from Equation (17), recalling that $s_0 = o(\sqrt{n}/\log p)$ and hence $(16acs_0\log p)/(C_{\min}\sqrt{n}) \le \varepsilon/5$ for all n large enough.

This completes the proof of Equation (79). The matching lower bound follows by the same argument.

6.6 Proof of Theorem 16

We begin with proving Equation (39). Defining $Z_i \equiv \sqrt{n}(\widehat{\theta}_i^u - \theta_{0,i})/(\widehat{\sigma}[M\widehat{\Sigma}M^{\mathsf{T}}]_{i,i}^{1/2})$, we have

$$\begin{aligned}\lim_{n\to\infty} \alpha_{i,n}(\widehat{T}) &= \lim_{n\to\infty} \sup_{\theta_0} \Big\{\mathbb{P}(P_i \le \alpha) : i \in [p],\, \|\theta_0\|_0 \le s_0,\, \theta_{0,i} = 0\Big\}\\ &= \lim_{n\to\infty} \sup_{\theta_0} \Big\{\mathbb{P}\Big(\Phi^{-1}(1 - \frac{\alpha}{2}) \le \frac{\sqrt{n}|\widehat{\theta}_i^u|}{\widehat{\sigma}[M\widehat{\Sigma}M^{\mathsf{T}}]_{i,i}^{1/2}}\Big) : i \in [p],\, \|\theta_0\|_0 \le s_0,\, \theta_{0,i} = 0\Big\}\\ &= \lim_{n\to\infty} \sup_{\theta_0} \Big\{\mathbb{P}\Big(\Phi^{-1}(1 - \frac{\alpha}{2}) \le |Z_i|\Big) : i \in [p],\, \|\theta_0\|_0 \le s_0\Big\} \le \alpha\,,\end{aligned}$$

where the last inequality follows from Lemma 13.

We next prove Equation (40). Recall that $\Sigma^{-1}_{\cdot,i}$ is a feasible solution of (4), for $1 \le i \le p$ with probability at least $1 - 2p^{-c_2}$, as per Lemma 23). On this event, letting m_i be the solution of the optimization problem (4), we have

$$\begin{aligned}m_i^{\mathsf{T}}\widehat{\Sigma}m_i &\le \Sigma^{-1}_{i,\cdot}\widehat{\Sigma}\Sigma^{-1}_{\cdot,i}\\ &= (\Sigma^{-1}_{i,\cdot}\widehat{\Sigma}\Sigma^{-1}_{\cdot,i} - \Sigma^{-1}_{ii}) + \Sigma^{-1}_{i,i}\\ &= \frac{1}{n}\sum_{j=1}^{N}(V_j^2 - \Sigma^{-1}_{ii}) + \Sigma^{-1}_{i,i}\,,\end{aligned}$$

where $V_j = \Sigma^{-1}_{i,\cdot} X_j$ are i.i.d. random variables with $\mathbb{E}(V_j^2) = \Sigma^{-1}_{ii}$ and sub-Gaussian norm

$$\|V_j\|_{\psi_2} \le \|\Sigma^{-1/2}_{i,\cdot}\|_2 \|\Sigma^{-1/2} X_j\|_{\psi_2} \le \kappa \sqrt{\Sigma^{-1}_{i,i}}\,.$$

Letting $U_j = V_j^2 - \Sigma^{-1}_{ii}$, we have that U_j is zero mean and sub-exponential with $\|U_j\|_{\psi_1} \le 2\|V_j^2\|_{\psi_1} \le 4\|V_j\|^2_{\psi_2} \le 4\kappa^2 \Sigma^{-1}_{ii} \le 4\kappa^2 \sigma_{\min}(\Sigma)^{-1} \le 4\kappa^2 C^{-1}_{\min} \equiv \kappa'$. Hence, by applying Bernstein inequality (as, for instance, in the proof of Lemma 23), we have, for $\varepsilon \le e\kappa'$,

$$\mathbb{P}\Big(m_i^{\mathsf{T}} \widehat{\Sigma} m_i \ge \Sigma^{-1}_{i,i} + \varepsilon\Big) \le 2\, e^{-(n/6)(\varepsilon/e\kappa')^2} + 2\, p^{-c_2}\,.$$

We can make $c_2 \ge 2$ by a suitable choice of a and therefore, by Borel-Cantelli we have the following almost surely

$$\limsup_{n\to\infty}\, [m_i^{\mathsf{T}} \widehat{\Sigma} m_i - \Sigma^{-1}_{i,i}] \le 0\,. \tag{82}$$

Now we are ready to prove the lower bound for the power. Let $z_* \equiv \Phi^{-1}(1-\alpha/2)$. Then,

$$\begin{aligned}
&\liminf_{n\to\infty} \frac{1-\beta_{i,n}(\widehat{T};\gamma)}{1-\beta^*_{i,n}(\gamma)}\\
&= \liminf_{n\to\infty} \frac{1}{1-\beta^*_i(\gamma;n)} \inf_{\theta_0}\Big\{\mathbb{P}(P_i \le \alpha):\ \|\theta_0\|_0 \le s_0,\ |\theta_{0,i}| \ge \gamma\Big\}\\
&= \liminf_{n\to\infty} \frac{1}{1-\beta^*_{i,n}(\gamma)} \inf_{\theta_0}\Big\{\mathbb{P}\Big(z_* \le \frac{\sqrt{n}|\widehat{\theta}^u_i|}{\widehat{\sigma}[M\widehat{\Sigma}M^{\mathsf{T}}]^{1/2}_{i,i}}\Big):\ \|\theta_0\|_0 \le s_0,\ |\theta_{0,i}| \ge \gamma\Big\}\\
&= \liminf_{n\to\infty} \frac{1}{1-\beta^*_{i,n}(\gamma)} \inf_{\theta_0}\Big\{\mathbb{P}\Big(z_* \le \Big|Z_i + \frac{\sqrt{n}\theta_{0,i}}{\widehat{\sigma}[M\widehat{\Sigma}M^{\mathsf{T}}]^{1/2}_{i,i}}\Big|\Big):\ \|\theta_0\|_0 \le s_0,\ |\theta_{0,i}| \ge \gamma\Big\}\\
&\overset{(a)}{\ge} \liminf_{n\to\infty} \frac{1}{1-\beta^*_{i,n}(\gamma)} \inf_{\theta_0}\Big\{\mathbb{P}\Big(z_* \le \Big|Z_i + \frac{\sqrt{n}\gamma}{\sigma[\Sigma^{-1}_{i,i}]^{1/2}}\Big|\Big):\ \|\theta_0\|_0 \le s_0\Big\}\\
&= \liminf_{n\to\infty} \frac{1}{1-\beta^*_{i,n}(\gamma)} \Big\{1 - \Phi\Big(z_* - \frac{\sqrt{n}\gamma}{\sigma[\Sigma^{-1}_{i,i}]^{1/2}}\Big) + \Phi\Big(-z_* - \frac{\sqrt{n}\gamma}{\sigma[\Sigma^{-1}_{i,i}]^{1/2}}\Big)\Big\}\\
&= \liminf_{n\to\infty} \frac{1}{1-\beta^*_{i,n}(\gamma)} G\Big(\alpha, \frac{\sqrt{n}\gamma}{\sigma[\Sigma^{-1}_{i,i}]^{1/2}}\Big) = 1\,.
\end{aligned}$$

Here (a) follows from Equation (82) and the fact $|\theta_{0,i}| \ge \gamma$.

6.7 Proof of Theorem 21

Under the assumptions of Theorem 8 and assuming $s_0 = o(\sqrt{n}/\log p)$, we have

$$\sqrt{n}(\widehat{\theta}^u - \theta_0) = \frac{1}{\sqrt{n}} M \mathbf{X}^{\mathsf{T}} W + \Delta\,,$$

with $\|\Delta\|_\infty = o(1)$. Using Lemma 12, we have

$$\frac{\sqrt{n}(\widehat{\theta}^u_i - \theta_{0,i})}{\sigma[M\widehat{\Sigma}M^{\mathsf{T}}]^{1/2}_{i,i}} = Z_i + o(1)\,, \quad \text{with } Z_i \equiv \frac{1}{\sqrt{n}}\frac{m_i^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}W}{\sigma[m_i^{\mathsf{T}}\widehat{\Sigma}m_i]^{1/2}}\,.$$

The following lemma characterizes the limiting distribution of $Z_i|\mathbf{X}$ which implies the validity of the proposed p-value P_i and confidence intervals.

Lemma 24 *Suppose that the noise variables W_i are independent with $\mathbb{E}(W_i) = 0$, and $\mathbb{E}(W_i^2) = \sigma^2$, and $E(|W_i|^{2+a}) \le C\,\sigma^{2+a}$ for some $a > 0$. Let $M = (m_1, \dots, m_p)^{\mathsf{T}}$ be the matrix with rows m_i^{T} obtained by solving optimization problem (54). For $i \in [p]$, define*

$$Z_i = \frac{1}{\sqrt{n}}\frac{m_i^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}W}{\sigma[m_i^{\mathsf{T}}\widehat{\Sigma}m_i]^{1/2}}\,.$$

Under the assumptions of Theorem 8, for any sequence $i = i(n) \in [p]$, and any $x \in \mathbb{R}$, we have

$$\lim_{n\to\infty} \mathbb{P}(Z_i \le x|\mathbf{X}) = \Phi(x)\,.$$

Lemma 24 is proved in Appendix A.2.

Acknowledgments

A.J. is supported by a Caroline and Fabian Pease Stanford Graduate Fellowship. This work was partially supported by the NSF CAREER award CCF-0743978, the NSF grant DMS-0806211, and the grants AFOSR/DARPA FA9550-12-1-0411 and FA9550-13-1-0036.

Appendix A. Proof of Technical Lemmas

This appendix contains the proofs of several technical steps needed in establishing our theoretical results.

A.1 Proof of Lemma 12

Let $C_i(\mu)$ be the optimal value of the optimization problem (4). We claim that

$$C_i(\mu) \ge \frac{(1-\mu)^2}{\widehat{\Sigma}_{ii}}\,. \tag{83}$$

To prove this claim notice that the constraint implies (by considering its i-th component):

$$1 - \langle e_i, \widehat{\Sigma}m\rangle \le \mu\,.$$

Therefore if $\tilde{m}$ is feasible and $c \ge 0$, then

$$\langle \tilde{m}, \widehat{\Sigma}\tilde{m}\rangle \ge \langle \tilde{m}, \widehat{\Sigma}\tilde{m}\rangle + c(1-\mu) - c\langle e_i, \widehat{\Sigma}\tilde{m}\rangle \ge \min_m \Big\{\langle m, \widehat{\Sigma}m\rangle + c(1-\mu) - c\langle e_i, \widehat{\Sigma}m\rangle\Big\}\,.$$

Minimizing over all feasible $\tilde{m}$ gives

$$C_i(\mu) \geq \min_{m} \Big\{ \langle m, \widehat{\Sigma} m \rangle + c(1-\mu) - c\langle e_i, \widehat{\Sigma} m \rangle \Big\} \,. \tag{84}$$

The minimum over m is achieved at $m = ce_i/2$. Plugging in for m, we get

$$C_i(\mu) \geq c(1-\mu) - \frac{c^2}{4}\widehat{\Sigma}_{ii} \tag{85}$$

Optimizing this bound over c, we obtain the claim (83), with the optimal choice being $c = 2(1-\mu)/\widehat{\Sigma}_{ii}$.

A.2 Proof of Lemma 24

Write

$$Z_i = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \xi_j \qquad \text{with} \quad \xi_j \equiv \frac{m_i^{\mathsf{T}} X_j W_j}{\sigma [m_i^{\mathsf{T}} \widehat{\Sigma} m_i]^{1/2}} \,.$$

Conditional on $\mathbf{X}$, the summands ξ_j are zero mean and independent. Further, $\sum_{j=1}^{n} \mathbb{E}(\xi_j^2|\mathbf{X}) = n$. We next prove the Lindeberg condition as per Equation (53). Let $c_n \equiv (m_i^{\mathsf{T}} \widehat{\Sigma} m_i)^{1/2}$. By Lemma 12, we have $\liminf_{n\to\infty} c_n \geq c_\infty > 0$, almost surely. If all the optimization problems in (54) are feasible, then $|\xi_j| \leq c_n^{-1} \|\mathbf{X} m_i\|_\infty \|W\|_\infty/\sigma \leq c_n^{-1} n^{\beta} (\|W\|_\infty/\sigma)$. Hence,

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} \mathbb{E}\Big(\xi_j^2 \mathbb{I}_{\{|\xi_j| > \varepsilon\sqrt{n}\}} |\mathbf{X}\Big) &\leq \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} \mathbb{E}\Big(\xi_j^2 \mathbb{I}_{\{\|W\|_\infty/\sigma > \varepsilon c_n n^{1/2-\beta}\}} |\mathbf{X}\Big) \\
&= \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} \frac{m_i^{\mathsf{T}} X_j X_j^{\mathsf{T}} m_i}{m_i^{\mathsf{T}} \widehat{\Sigma} m_i} \mathbb{E}(\widetilde{W}_j^2 \, \mathbb{I}_{\{\|\widetilde{W}\|_\infty > \varepsilon c_\infty n^{1/2-\beta}\}}\Big) \\
&\leq \lim_{n\to\infty} \mathbb{E}(\widetilde{W}_1^2 \, \mathbb{I}_{\{|\widetilde{W}_1| > \varepsilon c_\infty n^{1/2-\beta}\}}\Big) \\
&\leq c'(\varepsilon) \lim_{n\to\infty} n^{-a(1/2-\beta)} \mathbb{E}\{|\widetilde{W}_1|^{2+a}\} = 0 \,.
\end{aligned}$$

where $\widetilde{W}_j = W_j/\sigma$ and the last limit follows since $\beta < 1/2$ and $a > 0$.

Using Lindeberg central limit theorem, we obtain $Z_i|\mathbf{X}$ converges weakly to standard normal distribution, and hence, $\mathbf{X}$-almost surely

$$\lim_{n\to\infty} \mathbb{P}(Z_i \leq x|\mathbf{X}) = \Phi(x) \,.$$

What remains is to show that with high probability all the p optimization problems in (54) are feasible. In particular, we show that $\Sigma_{i,\cdot}^{-1}$ is a feasible solution to the i-th optimization problem, for $i \in [p]$. By Lemma 23, $|\Sigma^{-1}\widehat{\Sigma} - \mathrm{I}|_\infty \leq \mu$, with high probability. Moreover,

$$\begin{aligned}
\sup_{j\in[p]} \|\Sigma_{i,\cdot}^{-1} X_j\|_{\psi_2} &= \sup_{j\in[p]} \|\Sigma_{i,\cdot}^{-1/2} \Sigma^{-1/2} X_j\|_{\psi_2} \\
&= \|\Sigma_{i,\cdot}^{-1/2}\|_2 \sup_{j\in[p]} \|\Sigma^{-1/2} X_j\|_{\psi_2} \\
&= [\Sigma_{i,i}^{-1}]^{1/2} \sup_{j\in[p]} \|\Sigma^{-1/2} X_j\|_{\psi_2} = O(1) \,.
\end{aligned}$$

Using tail bound for sub-Gaussian variables $\Sigma^{-1}_{i,\cdot}X_j$ and union bounding over $j \in [n]$, we get

$$\mathbb{P}(\|\mathbf{X}\Sigma^{-1}_{\cdot,i}\|_\infty > n^\beta) \le ne^{-cn^{2\beta}}\,,$$

for some constant $c > 0$. Note that $s_0 = o(\sqrt{n}/\log p)$ and $\beta > 1/4$ imply $p = e^{o(n^{2\beta})}$. Hence, almost surely, $\Sigma^{-1}_{i,\cdot}$ is a feasible solution to optimization problem (54), for all $i \in [p]$.

Appendix B. Corollaries of Theorem 8

In this appendix, we prove Corollary 10 and Corollary 11.

B.1 Proof of Corollary 10

By Theorem 6, for any $\mathbf{X} \in \mathcal{E}_n(\sqrt{C_{\min}}/2, s_0, 3/2) \cap \mathcal{G}_n(a)$, we have

$$\mathbb{P}\Big\{\|\Delta\|_\infty \ge L\,c\Big|\mathbf{X}\Big\} \le 2\,p^{1-(c^2/48)}\,, \quad L \equiv \frac{16a\sigma}{C_{\min}}\,\frac{s_0\log p}{\sqrt{n}}\,. \tag{86}$$

This is obtained by setting $\phi_0 = \sqrt{C_{\min}}/2$, $K = 3/2$, $\mu_* = a\sqrt{(\log p)/n}$ in Equation (11). Hence

$$\begin{aligned}\|\mathsf{Bias}(\widehat{\theta}^u)\|_\infty &\le \frac{1}{\sqrt{n}}\mathbb{E}\big\{\|\Delta\|_\infty\big|\mathbf{X}\big\}\\ &= \frac{L}{\sqrt{n}}\int_0^\infty \mathbb{P}\Big\{\|\Delta\|_\infty \ge L\,c\Big|\mathbf{X}\Big\}\,\mathrm{d}c\\ &\le \frac{2L}{\sqrt{n}}\int_0^\infty \min(1, p^{1-(c^2/48)})\,\mathrm{d}c \le \frac{10L}{\sqrt{n}}\,, \end{aligned}\tag{87}$$

which coincides with Equation (21). The probability estimate (22) simply follows from Theorem 7 using union bound.

B.2 Proof of Corollary 11

By Theorem 7.(a), we have

$$\mathbb{P}\Big(\mathbf{X} \in \mathcal{E}_n(1/2, s_0, 3/2)\Big) \ge 1 - 4\,e^{-c_1 n}\,. \tag{88}$$

Further, by Lemma 23, with $\widehat{\Sigma} \equiv \mathbf{X}^{\mathsf{T}}\mathbf{X}/n$, we have

$$\mathbb{P}\Big(\mu_*(\mathbf{X};\mathrm{I}) \le 30\sqrt{\frac{\log p}{n}}\Big) \ge 1 - 2\,p^{-3}\,. \tag{89}$$

Hence, defining

$$\mathcal{B}_n \equiv \mathcal{E}_n(1/2, s_0, 3/2) \cap \Big\{\mathbf{X} \in \mathbb{R}^{n\times p}:\ \mu_*(\mathbf{X};\mathrm{I}) \le 30\sqrt{\frac{\log p}{n}}\Big\} \tag{90}$$

we have the desired probability bound (24). Let $\widehat{\theta}^n = \widehat{\theta}^n(Y, \mathbf{X}; \mathrm{I}, \lambda)$. By Theorem 6, we have, for any $\mathbf{X} \in \mathcal{B}_n$

$$\widehat{\theta}^* = \theta_0 + \frac{1}{\sqrt{n}} Z + \frac{1}{\sqrt{n}} \Delta \,, \quad Z|\mathbf{X} \sim \mathsf{N}(0, \sigma^2 \widehat{\Sigma}) \,, \tag{91}$$

and further

$$\mathbb{P}\Big\{ \|\Delta\|_\infty \ge \frac{480 c \sigma s_0 \log p}{\sqrt{n}} \Big| \mathbf{X} \Big\} \le 2 p^{1-(c^2/48)} \,, \tag{92}$$

whence, proceeding as in the proof in the last section, we get, for some universal numerical constant $c_{**} \le 4800$,

$$\|\mathsf{Bias}(\widehat{\theta}^*)\|_\infty \le \frac{1}{\sqrt{n}} \mathbb{E}\Big\{ \|\Delta\|_\infty \Big| \mathbf{X} \Big\} \le c_{**} \sigma \frac{s_0 \log p}{n} \,. \tag{93}$$

Next by Equation (28) we have

$$\big\|\mathsf{Bias}(\widehat{\theta}^n)\big\|_\infty \ge \Big| \lambda \big\| \mathbb{E}\{ v(\widehat{\theta}^n) | \mathbf{X} \} \big\|_\infty - \big\| \mathsf{Bias}(\widehat{\theta}^*) \big\|_\infty \Big| \,. \tag{94}$$

Hence, in order to prove Equation (23), it is sufficient to prove that $\|\mathbb{E}\{v(\widehat{\theta}^n)|\mathbf{X}\}\|_\infty \ge 2/3$.

Note that $v(\widehat{\theta}^n)_i = 1$ whenever $\widehat{\theta}^n_i > 0$, and $|v(\widehat{\theta}^n)_i| \le 1$ for all coordinates i. Therefore, letting $b_0 \equiv 480 c \sigma (s_0 \log p)/n$ we have

$$\begin{aligned} 1 - \mathbb{E}\{ v(\widehat{\theta}^n)_i | \mathbf{X} \} &\le 2 \mathbb{P}\Big(\widehat{\theta}^n_i \le 0 \Big| \mathbf{X} \Big) \le 2 \mathbb{P}\Big(\widehat{\theta}^u_i \le \lambda \Big| \mathbf{X} \Big) \qquad (95) \\ &\le 2 \mathbb{P}\Big(\theta_{0,i} + \frac{1}{\sqrt{n}} Z_i + \frac{1}{\sqrt{n}} \Delta_i \le \lambda \Big| \mathbf{X} \Big) \\ &\le 2 \mathbb{P}\Big(\frac{1}{\sqrt{n}} Z_i \le \lambda + b_0 - \theta_{0,i} \Big| \mathbf{X} \Big) + 2 \mathbb{P}\Big(\|\Delta\|_\infty > \sqrt{n} b_0 \Big) \\ &= 2 \Phi\Big((\lambda + b_0 - \theta_{0,i}) \sqrt{n/(\sigma^2 \widehat{\Sigma}_{ii})} \Big) + 4 p^{1-(c^2/48)} \\ &\le 2 \Phi\Big((\lambda + b_0 - \theta_{0,i}) \sqrt{2n/(3\sigma^2)} \Big) + 4 p^{1-(c^2/48)} \qquad (96) \end{aligned}$$

with $\Phi(x)$ the standard normal distribution function. Here, we used the relation $\widehat{\theta}^u = \widehat{\theta} + \lambda v(\widehat{\theta})$ in Equation (95) and Equation (96) holds because $\max_{i \in [p]} \widehat{\Sigma}_{ii} \le 3/2$ on $\mathcal{B}_n$. We then choose θ_0 so that $\theta_{0,i} \ge b_0 + \lambda + \sqrt{30\sigma^2/n}$, for $i \in [p]$ in the support of θ_0. We therefore obtain

$$\mathbb{E}\{ v(\widehat{\theta}^n)_i | \mathbf{X} \} \ge 1 - 2\Phi(-\sqrt{20}) - 4 p^{1-(c^2/48)} \ge \frac{2}{3} \,, \tag{97}$$

where in the last step we used the assumption $p \ge 13^{48/(c^2-48)}$. This finishes the proof of Equation (23).

Equation (25) follows readily from Equation (93), substituting $\lambda = c\sigma\sqrt{(\log p)/n}$ and recalling the assumption $n \ge (3 c_{**} s_0 / c)^2 \log p$.

Finally, combining Equation (25) and Equation (23), we get

$$\|\mathsf{Bias}(\widehat{\theta}^n)\|_\infty \ge \frac{\lambda}{3} \,. \tag{98}$$

Therefore, Equation (26) is derived by substituting $\lambda = c\sigma\sqrt{(\log p)/n}$ and using Corollary 10, Equation (21) with $a = 30$.

Appendix C. Proof of Lemma 14

Let $\mathcal{E}_n = \mathcal{E}_n(\phi_0, s_0, K)$ be the event defined as per Theorem 7.(a). In particular, we take $\phi_0 = \sqrt{C_{\min}}/2$, and $K \ge 1 + 20\kappa^2\sqrt{(\log p)/n}$.[4] Further note that we can assume without loss of generality $n \ge \nu_0\, s_0 \log(p/s_0)$, since $s_0 = o(\sqrt{n}/\log p)$. Fixing $\varepsilon > 0$, we have

$$\begin{aligned}\mathbb{P}\Big(\Big|\frac{\widehat{\sigma}}{\sigma} - 1\Big| \ge \varepsilon\Big) &\le \sup_{\mathbf{X}\in\mathcal{E}_n} \mathbb{P}\Big(\Big|\frac{\widehat{\sigma}}{\sigma} - 1\Big| \ge \varepsilon \,\Big|\, \mathbf{X}\Big) + \mathbb{P}\big(\mathbf{X} \notin \mathcal{E}_n\big)\\ &\le \sup_{\mathbf{X}\in\mathcal{E}_n} \mathbb{P}\Big(\Big|\frac{\widehat{\sigma}}{\sigma} - 1\Big| \ge \varepsilon \,\Big|\, \mathbf{X}\Big) + 4\, e^{-c_1 n}\,,\end{aligned}$$

where $c_1 > 0$ is a constant defined as per Theorem 7.(a).

We are therefore left with the task of bounding the first term in the last expression above, uniformly over $\theta_0 \in \mathbb{R}^p$, $\|\theta_0\|_0 \le s_0$. For $\mathbf{X} \in \mathcal{E}_n$, we apply the result of Sun and Zhang (2012, Theorem 1). More precisely, using the notations of Sun and Zhang (2012), with $\lambda_0 = \widetilde{\lambda}$, $\xi = 3$, $T = \mathrm{supp}(\theta_0)$, $\kappa(\xi, T) \ge \phi_0$, we have $\eta_*(\widetilde{\lambda}, \xi) \le 4s_0\widetilde{\lambda}^2/\phi_0^2$. Further, let σ^* be the oracle estimator of σ introduced there. If $\|\mathbf{X}^{\mathsf{T}}W/(n\sigma^*)\|_\infty \le \widetilde{\lambda}/4$, using Equation (13) in Sun and Zhang (2012), we obtain

$$\Big|\frac{\widehat{\sigma}}{\sigma^*} - 1\Big| \le \frac{2\sqrt{s_0}\widetilde{\lambda}}{\sigma^*\phi_0} \le \frac{\varepsilon}{2}\,, \tag{99}$$

where the last inequality follows for all n large enough since $s_0 = o(\sqrt{n}/\log p)$.

Hence

$$\sup_{\mathbf{X}\in\mathcal{E}_n} \mathbb{P}\Big(\Big|\frac{\widehat{\sigma}}{\sigma} - 1\Big| \ge \varepsilon \,\Big|\, \mathbf{X}\Big) \le \sup_{\mathbf{X}\in\mathcal{E}_n} \mathbb{P}\Big(\, \|\mathbf{X}^{\mathsf{T}}W/n\|_\infty > \widetilde{\lambda}/4 \Big|\, \mathbf{X}\Big) + \sup_{\mathbf{X}\in\mathcal{E}_n} \mathbb{P}\Big(\Big|\frac{\sigma^*}{\sigma} - 1\Big| \ge \frac{\varepsilon}{10} \,\Big|\, \mathbf{X}\Big)\,, \tag{100}$$

where we note that the right hand side is independent of θ_0. The first term vanishes as $n \to \infty$ by a standard tail bound on the supremum of p Gaussian random variables. The second term also vanishes because it is controlled by the tail of a chi-squared random variable (see Sun and Zhang, 2012).

Appendix D. Proof of Theorem 20

Let $\mathcal{F}_{p,s_0} \equiv \{x \in \mathbb{R}^p :\, \|x\|_0 \le s_0\}$, and fix $\varepsilon \in (0, 1/10)$. By definition,

$$\begin{aligned}\mathrm{FWER}(\widehat{T}^{\mathrm{F}}, n) &= \sup_{\theta_0\in\mathcal{F}_{p,s_0}} \mathbb{P}\left\{\exists i \in [p]\setminus \mathrm{supp}(\theta_0),\ \text{ s.t. } \frac{\sqrt{n}\,|\widehat{\theta}^u_i - \theta_{0,i}|}{\widehat{\sigma}[M\widehat{\Sigma}M^{\mathsf{T}}]^{1/2}_{i,i}} \ge \Phi^{-1}\Big(1 - \frac{\alpha}{2p}\Big)\right\}\\ &\le \sup_{\theta_0\in\mathcal{F}_{p,s_0}} \mathbb{P}\left\{\exists i \in [p]\setminus \mathrm{supp}(\theta_0),\ \text{ s.t. } \frac{\sqrt{n}\,|\widehat{\theta}^u_i - \theta_{0,i}|}{\sigma[M\widehat{\Sigma}M^{\mathsf{T}}]^{1/2}_{i,i}} \ge (1-\varepsilon)\Phi^{-1}\Big(1 - \frac{\alpha}{2p}\Big)\right\}\\ &\qquad + \sup_{\theta_0\in\mathcal{F}_{p,s_0}} \mathbb{P}\Big(\Big|\frac{\widehat{\sigma}}{\sigma} - 1\Big| \ge \varepsilon\Big)\,.\end{aligned}$$

[4] For instance $K = 1.1$ will work for all n large enough since $(s_0 \log p)^2/n \to 0$, with $s_0 \ge 1$, by assumption.

Since the second term vanishes as $n \to \infty$ by Equation (29). Using Bonferroni inequality, letting $z_\alpha(\varepsilon) \equiv (1-\varepsilon)\Phi^{-1}\big(1-\frac{\alpha}{2p}\big)$, we have

$$\begin{aligned}
\limsup_{n\to\infty} \mathrm{FWER}(\widehat{T}^{\mathrm{F}}, n) &\le \limsup_{n\to\infty} \sum_{i=1}^{p} \sup_{\theta_0\in\mathcal{F}_{p,s_0}, \theta_{0,i}=0} \mathbb{P}\left\{\frac{\sqrt{n}\,|\widehat{\theta}^u_i - \theta_{0,i}|}{\sigma[M\widehat{\Sigma}M^{\mathsf{T}}]^{1/2}_{i,i}} \ge z_\alpha(\varepsilon)\right\}\\
&= \limsup_{n\to\infty} \sum_{i=1}^{p} \sup_{\theta_0\in\mathcal{F}_{p,s_0}, \theta_{0,i}=0} \mathbb{P}\left\{\left|\widetilde{Z}_i + \frac{\Delta_i}{\sigma[M\widehat{\Sigma}M^{\mathsf{T}}]^{1/2}_{ii}}\right| \ge z_\alpha(\varepsilon)\right\}\,,
\end{aligned}$$

where, by Theorem 8, $\widetilde{Z}_i \sim \mathsf{N}(0,1)$ and Δ_i is given by Equation (16). We then have

$$\begin{aligned}
\limsup_{n\to\infty} \mathrm{FWER}(\widehat{T}^{\mathrm{F}}, n) \le\; &\limsup_{n\to\infty} \sum_{i=1}^{p} \mathbb{P}\big\{|\widetilde{Z}_i| \ge z_\alpha(\varepsilon) - \varepsilon\big\}\\
&+ \limsup_{n\to\infty} \sum_{i=1}^{p} \sup_{\theta_0\in\mathcal{F}_{p,s_0}, \theta_{0,i}=0} \mathbb{P}\left\{\|\Delta\|_\infty \ge \frac{\varepsilon\sigma}{2\widehat{\Sigma}^{1/2}_{ii}}\right\}\\
\le\; & 2p\big(1-\Phi(z_\alpha(\varepsilon)-\varepsilon)\big) + \limsup_{n\to\infty}\, p \max_{i\in[p]} \mathbb{P}(\widehat{\Sigma}_{ii} \ge 2)\\
&+ \limsup_{n\to\infty} \sup_{\theta_0\in\mathcal{F}_{p,s_0}, \theta_{0,i}=0} p\,\mathbb{P}\left\{\|\Delta\|_\infty \ge \frac{\varepsilon\sigma}{4}\right\}\,, \qquad (101)
\end{aligned}$$

where in the first inequality, we used $[M\widehat{\Sigma}M^{\mathsf{T}}]_{i,i} \ge 1/(4\widehat{\Sigma}_{ii})$ for all n large enough, by Lemma 12, and since $\mu = a\sqrt{(\log p)/n} \to 0$ as $n, p \to \infty$. Now, the second term in the right hand side of Equation (101) vanishes by Theorem 7.(a), and the last term is zero by Theorem 8, since $s_0 = o(\sqrt{n}/\log p)$. Therefore

$$\limsup_{n\to\infty} \mathrm{FWER}(\widehat{T}^{\mathrm{F}}, n) \le 2p\big(1-\Phi(z_\alpha(\varepsilon)-\varepsilon)\big)\,. \qquad (102)$$

The claim follows by letting $\varepsilon \to 0$.

References

M. Bayati, M. A. Erdogdu, and A. Montanari. Estimating Lasso risk and noise level. In *Advances in Neural Information Processing Systems*, pages 944–952, 2013.

A. Belloni and V. Chernozhukov. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547, 2013.

A. Belloni, V. Chernozhukov, and Y. Wei. Honest confidence regions for a regression parameter in logistic regression with a large number of controls. *arXiv Preprint* arXiv:1304.3969, 2013.

A. Belloni, V. Chernozhukov, and C. Hansen. Inference on treatment effects after selection amongst high-dimensional controls. *arXiv Preprint* arXiv:1201.0224, The Review of Economic Studies, 2014.

P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37:1705–1732, 2009.

P. Bühlmann. Statistical significance in high-dimensional linear models. *Bernoulli*, 19(4): 1212–1242, 2013.

P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data.* Springer-Verlag, 2011.

P. Bühlmann, M. Kalisch, and L. Meier. High-dimensional statistics with a view toward applications in biology. *Annual Review of Statistics and Its Application*, 1(1):255–278, 2014.

E. Candès and T. Tao. The Dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, 35:2313–2351, 2007.

E. J. Candès and Y. Plan. Near-ideal model selection by ℓ_1 minimization. *The Annals of Statistics*, 37(5A):2145–2177, 2009.

E. J. Candès and T. Tao. Decoding by linear programming. *IEEE Trans. on Inform. Theory*, 51:4203–4215, 2005.

S. S. Chen and D. L. Donoho. Examples of basis pursuit. In *Proceedings of Wavelet Applications in Signal and Image Processing III*, San Diego, CA, 1995.

R. Dezeure and P. Bühlmann. *Private Communication*, 2013.

L. H. Dicker. Residual variance and the signal-to-noise ratio in high-dimensional linear models. *arXiv Preprint* arXiv:1209.0012, 2012.

D. L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory*, 47(7):2845–2862, 2001.

B. Efron. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*, volume 1. Cambridge University Press, 2010.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.

J. Fan, R. Samworth, and Y. Wu. Ultrahigh dimensional feature selection: beyond the linear model. *The Journal of Machine Learning Research*, 10:2013–2038, 2009.

J. Fan, S. Guo, and N. Hao. Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(1):37–65, 2012.

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. URL `http://www.jstatsoft.org/v33/i01/`.

E. Greenshtein and Y. Ritov. Persistence in high-dimensional predictor selection and the virtue of over-parametrization. *Bernoulli*, 10:971–988, 2004.

A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional statistical models. In *Advances in Neural Information Processing Systems*, pages 1187–1195, 2013a.

A. Javanmard and A. Montanari. Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory. *arXiv Preprint* arXiv:1301.4240, (To appear in *IEEE Transactions on Information Theory*), 2013b.

A. Javanmard and A. Montanari. Nearly optimal sample size in hypothesis testing for high-dimensional regression. In *51st Annual Allerton Conference*, pages 1427–1434, Monticello, IL, June 2013c.

Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, August 2009.

E. L. Lehmann and J. P. Romano. *Testing Statistical Hypotheses.* Springer, 2005.

R. Lockhart, J. Taylor, R. J. Tibshirani, and R. Tibshirani. A significance test for the Lasso. *Annals of Statistics*, 42(2):413–468, 2014.

M. Lustig, D. L. Donoho, J. M. Santos, and J. M. Pauly. Compressed sensing MRI. *IEEE Signal Processing Magazine*, 25:72–82, 2008.

S. G. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.

N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, 34:1436–1462, 2006.

N. Meinshausen and P. Bühlmann. Stability selection. *J. R. Statist. Soc. B*, 72:417–473, 2010.

N. Meinshausen, L. Meier, and P. Bühlmann. p-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488):1671–1681, 2009.

J. Minnier, L. Tian, and T. Cai. A perturbation method for inference on regularized regression estimates. *Journal of the American Statistical Association*, 106(496), 2011.

J. Peng, J. Zhu, A. Bergamaschi, W. Han, D.-Y. Noh, J. R. Pollack, and P. Wang. Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *The Annals of Applied Statistics*, 4(1):53–77, 2010.

S. Reid, R. Tibshirani, and J. Friedman. A study of error variance estimation in Lasso regression. *arXiv Preprint* arXiv:1311.5274, 2013.

Z. Ren, T. Sun, C.-H. Zhang, and H. H. Zhou. Asymptotic normality and optimalities in estimation of large gaussian graphical model. *arXiv Preprint* arXiv:1309.6024, 2013.

M. Rudelson and Z. Shuheng. Reconstruction from anisotropic random measurements. *IEEE Transactions on Information Theory*, 59(6):3434–3447, 2013.

N. Städler, P. Bühlmann, and S. van de Geer. ℓ_1-penalization for mixture regression models (with discussion). *Test*, 19(2):209–256, 2010.

T. Sun and C.-H. Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012.

R. Tibshirani. Regression shrinkage and selection with the Lasso. *J. Royal. Statist. Soc B*, 58:267–288, 1996.

S. van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electron. J. Statist.*, 3:1360–1392, 2009.

S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *arXiv Preprint* arXiv:1303.0518, (To appear in *Annals of Statistics*), 2014.

A. W. Van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge university press, 2000.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Y.C. Eldar and G. Kutyniok, editors, *Compressed Sensing: Theory and Applications*, pages 210–268. Cambridge University Press, 2012.

M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming. *IEEE Trans. on Inform. Theory*, 55:2183–2202, 2009.

L. Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer Verlag, 2004.

C-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010.

C.-H. Zhang and S. S. Zhang. Confidence intervals for low-dimensional parameters in high-dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76:217–242, 2014.

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