Hypothesis Testing in High-Dimensional Regression under the Gaussian Random Design Model: Asymptotic Theory

Adel Javanmard* and Andrea Montanari *†

January 17, 2013

Abstract

We consider linear regression in the high-dimensional regime in which the number of observations n is smaller than the number of parameters p. A very successful approach in this setting uses ℓ_1-penalized least squares (a.k.a. the Lasso) to search for a subset of $s_0 < n$ parameters that best explain the data, while setting the other parameters to zero. A considerable amount of work has been devoted to characterizing the estimation and model selection problems within this approach.

In this paper we consider instead the fundamental, but far less understood, question of statistical significance.

We study this problem under the random design model in which the rows of the design matrix are i.i.d. and drawn from a high-dimensional Gaussian distribution. This situation arises, for instance, in learning high-dimensional Gaussian graphical models. Leveraging on an asymptotic distributional characterization of regularized least squares estimators, we develop a procedure for computing p-values and hence assessing statistical significance for hypothesis testing. We characterize the statistical power of this procedure, and evaluate it on synthetic and real data, comparing it with earlier proposals. Finally, we provide an upper bound on the minimax power of tests with a given significance level and show that our proposed procedure achieves this bound in case of design matrices with i.i.d. Gaussian entries.

1 Introduction

The Gaussian random design model for linear regression is defined as follows. We are given n i.i.d. pairs (y_1, x_1), (y_2, x_2), $\dots (y_n, x_n)$ with $y_i \in \mathbb{R}$ and $x_i \in \mathbb{R}^p$, $x_i \sim \mathsf{N}(0, \Sigma)$ for some covariance matrix $\Sigma \succ 0$. Further, y_i is a linear function of x_i, plus noise

$$y_i \,=\, \langle \theta_0, x_i \rangle + w_i \,, \qquad w_i \sim \mathsf{N}(0, \sigma^2) \,. \tag{1}$$

Here $\theta_0 \in \mathbb{R}^p$ is a vector of parameters to be learned and $\langle \, \cdot \, , \, \cdot \, \rangle$ is the standard scalar product. The special case $\Sigma = \mathrm{I}_{p\times p}$ is usually referred to as 'standard' Gaussian design model.

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

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

*Department of Electrical Engineering, Stanford University

†Department of Statistics, Stanford University

We are interested in high dimensional settings where the number of parameters exceeds the sample size, i.e., $p > n$, but the number of non-zero entries of θ_0 is smaller than p. In this situation, a recurring problem is to select the non-zero entries of θ_0 that hence can provide a succinct explanation of the data. The vast literature on this topic is briefly overviewed in Section 1.1.

In statistical applications, it is unrealistic to assume that the set of nonzero entries ofθ_0 can be determined with absolute certainty. The present paper focuses on the problem of quantifying the *uncertainty* associated to the entries of θ_0. More specifically, we are interested in testing null-hypotheses of the form:

$$H_{0,i}: \quad \theta_{0,i} = 0, \tag{3}$$

for $i \in [p] \equiv \{1, 2, \dots, p\}$ and assigning p-values for these tests. Rejecting $H_{0,i}$ is equivalent to stating that $\theta_{0,i} \neq 0$.

Any hypothesis testing procedure faces two types of errors: false positives or type I errors (incorrectly rejecting $H_{0,i}$, while $\theta_{0,i} = 0$), and false negatives or type II errors (failing to reject $H_{0,i}$, while $\theta_{0,i} \neq 0$). The probabilities of these two types of errors will be denoted, respectively, by α and β (see Section 2.1 for a more precise definition). The quantity $1 - \beta$ is also referred as the power of the test, and α its significance level. It is trivial to achieve α arbitrarily small if we allow for $\beta = 1$ (never reject $H_{0,i}$) or β arbitrarily small if we allow for $\alpha = 1$ (always reject $H_{0,i}$). This paper aims at optimizing the trade-off between power $1 - \beta$ and significance α.

Without further assumptions on the problem structure, the trade-off is trivial and no non-trivial lower bound on $1 - \beta$ can be established. Indeed we can take $\theta_{0,i} \neq 0$ arbitrarily close to 0, thus making $H_{0,i}$ in practice indistinguishable from its complement. We will therefore assume that, whenever $\theta_{0,i} \neq 0$, we have $|\theta_{0,i}| > \mu$ as well. The smallest value of μ such that the power and significance reach some fixed non-trivial value (e.g., $\alpha = 0.05$ and $1 - \beta \geq 0.9$) has a particularly compelling interpretation, and provides an answer to the following question:

> What is the minimum magnitude of $\theta_{0,i}$ to be able to distinguish it from the noise level, with a given degree of confidence?

Recently Zhang and Zhang [ZZ11] and Bühlmann [Büh12] proposed hypothesis testing procedures for design matrices $\mathbf{X}$ satisfying the restricted eigenvalue property [BRT09]. When specialized to the case of standard Gaussian designs $x_i \sim \mathsf{N}(0, \mathrm{I}_{p\times p})$, these methods require $|\theta_{0,i}| \geq \mu = c \max\{\sigma s_0 \log p/\, n, \sigma/\sqrt{n}\}$ to reject hypothesis $H_{0,i}$ with a given degree of confidence, where c is a constant independent of the problem dimensions (see Appendix A).

In this paper we prove that a significantly stronger test can be constructed, at least in an asymptotic sense, for some Gaussian designs. Indeed we show that $|\theta_{0,i}| \geq c\,\sigma/\sqrt{n}$ is sufficient for $H_{0,i}$ to be rejected. This is somewhat surprising. Even if $n = p$ and the measurement directions x_i are orthogonal, e.g.,[1] $\mathbf{X} = \sqrt{n}\mathrm{I}_{n\times n}$, we would need $|\theta_{0,i}| \geq c\,\sigma/\sqrt{n}$ to distinguish the i-th entry from noise.

As in [ZZ11, Büh12], our approach is based on the Lasso estimator [Tib96]

$$\widehat{\theta}(y, \mathbf{X}) = \arg\min_{\theta\in\mathbb{R}^p} \left\{ \frac{1}{2n}\|y - \mathbf{X}\theta\|^2 + \lambda\,\|\theta\|_1 \right\}. \tag{4}$$

[1]The factor $\sqrt{n}$ in $\mathbf{X} = \sqrt{n}\,\mathrm{I}_{n\times n}$ is required for a a fair comparison with the standard Gaussian design, where $\|x_i\|_2 \approx \sqrt{p}$ with high probability and hence the signal-to-noise ratio is p/σ^2.

Unlike [ZZ11, Büh12], the precise test, and its analysis are based on an exact asymptotic distributional characterization of high-dimensional estimators of the type (4). This characterization was proved in [BM11, BM12] for standard Gaussian designs, $\Sigma = \mathrm{I}_{p\times p}$. We further generalize these results to a broad class of covariance Σ, and general regularized least squares estimators, by using the non-rigorous replica method from statistical mechanics [MM09].

The contributions of this paper are organized as follows:

Upper bound on the minimax power. In Section 2 we introduce the problem formally, by taking a minimax point of view. We prove a general upper bound on the minimax power of tests with a given significance level α. We then specialize this bound to the case of standard Gaussian design matrices, showing formally that no test can detect $\theta_{0,i} \neq 0$ unless $|\theta_{0,i}| \geq \mu_{\mathrm{UB}} = c\sigma/\sqrt{n}$.

Hypothesis testing procedure for standard Gaussian designs. Building on the results of [BM12], we describe in Section 3.1 a test that is well suited for the case of standard Gaussian designs, $\Sigma = \mathrm{I}_{p\times p}$. We prove that this test achieves a 'nearly-optimal' power-significance trade-off in a properly defined asymptotic sense. Here 'nearly optimal' means that the trade-off has the same form, except that μ_{UB} is replaced by $\mu = C\mu_{\mathrm{UB}}$ with C a universal constant.

Generalization to nonstandard Gaussian designs. For $\Sigma \neq \mathrm{I}_{p\times p}$, no rigorous characterization analogous to the one of [BM12] is available. Using the non-rigorous replica method, we derive a conjecture for a broad class of covariance Σ and general regularized least squares estimators, that we will call the *standard distributional limit* (see Sections 3.2 and 4). Assuming that the standard distributional limit holds, we develop in Section 3.2 a hypothesis testing procedure for this more general case, that we refer to as SDL-TEST.

Validation. We validate our approach on both synthetic and real data in Sections 3.1.1, 3.2.1 and Section 5, comparing it with the method of [Büh12]. Simulations suggest that the latter is indeed overly conservative, resulting in suboptimal statistical power.

Proofs are deferred to Section 6.

This paper focuses on the asymptotic regime introduced in [Don06, DT05, DT09, DT10] and studied in [DMM09, DMM11, BM12]. The advantage of this approach is that the asymptotic characterization of [BM12] is sharp and appears to be accurate already at moderate sizes.

A forthcoming paper [JM13] will address the same questions in a non-asymptotic setting.

1.1 Further related work

As mentioned above, regularized least squares estimators were the object of intense theoretical investigation over the last few years. The focus of this work has been so far on establishing order optimal guarantees on: (1) The prediction error $\|\mathbf{X}(\widehat{\theta} - \theta_0)\|_2$ [GR04]; (2) The estimation error, typically quantified through $\|\widehat{\theta} - \theta_0\|_q$, with $q \in [1, 2]$ [CT07, BRT09, RWY09]; (3) The model selection (or support recovery) properties, e.g., by bounding $\mathbb{P}\{\mathrm{supp}(\widehat{\theta}) \neq \mathrm{supp}(\theta_0)\}$ [MB06, ZY06, Wai09]. For establishing estimation and support recovery guarantees, it is necessary to make specific assumptions on the design matrix $\mathbf{X}$, such as the restricted eigenvalue property of [BRT09] or the compatibility condition of [vdGB09]. Both [ZZ11] and [Büh12] assume conditions of this type for developing hypothesis testing procedures.

As mentioned above, our guarantees assume the Gaussian random design model. This was fruitfully studied in the context of standard linear regression [HKZ11], as well as sparse recovery. Donoho and Tanner [Don06, DT05, DT09, DT10] studied the noiseless case $\sigma = 0$, for standard Gaussian designs $\Sigma = \mathrm{I}_{p\times p}$, and reconstruction using basis pursuit, i.e., the $\lambda \to 0$ limit of the Lasso estimator (4). They considered the asymptotic behavior as $s_0, p, n \to \infty$ with $s_0/p \to \varepsilon \in (0,1)$ and $n/p \to \delta \in (0,1)$. They proved that, depending on the values of (ε, δ), the unknown vector θ_0 is either recovered exactly with probability converging to one or not recovered with probability converging to one, and characterized the boundary between these regimes.

Wainwright [Wai09] considered the Gaussian design model and established upper and lower thresholds $n_{\mathrm{UB}}(p, s_0; \Sigma)$, $n_{\mathrm{LB}}(p, s_0; \Sigma)$ for correct recovery of $\mathrm{supp}(\theta_0)$ in noise $\sigma > 0$, under an additional condition on $\mu \equiv \min_{i\in\mathrm{supp}(\theta_0)} |\theta_{0,i}|$. Namely, for $n, p, s_0 \to \infty$ with $n \geq n_{\mathrm{UB}}(p, s_0; \Sigma)$, $\mathbb{P}\{\mathrm{supp}(\widehat{\theta}) = \mathrm{supp}(\theta_0)\} \to 1$, while for $n, p, s_0 \to \infty$ with $n \leq n_{\mathrm{LB}}(p, s_0; \Sigma)$, $\mathbb{P}\{\mathrm{supp}(\widehat{\theta}) = \mathrm{supp}(\theta_0)\} \to 0$. For the special case of standard Gaussian designs, both $n_{\mathrm{LB}}(p, s_0; \Sigma = \mathrm{I})$ and $n_{\mathrm{UB}}(p, s_0; \Sigma = \mathrm{I})$ are asymptotically equivalent to $2s_0 \log(p)$, hence determining the threshold location. More generally $n_{\mathrm{UB}}(p, s_0; \Sigma) = O(s_0 \log p)$ for many covariance structures Σ, provided $\mu = \Omega(\sqrt{\log p/n})$. Correct support recovery depends, in a crucial way, on the irrepresentability condition of [ZY06].

In the regime $n = \Theta(s_0 \log p)$ that is relevant for exact support recovery, both type I and type II error rates tend to 0 rapidly as $n, p, s_0 \to \infty$. This makes it difficult to study the trade-off between statistical significance and power, and the optimality of testing procedures. Further, the techniques of [Wai09] (which are built on the results of [ZY06]) do not allow to estimate type I and type II error rates α and β but only their sum as $\alpha + \beta \leq p^{-c}$ for $c > 0$ depending on the level of regularization and the scaling of various dimensions.

Here we are interested in triples n, p, s_0 for which α and β stay bounded away from 0 and from the trivial baseline $\alpha + \beta = 1$. As shown in Section 2, any hypothesis testing method that achieves this requires $\mu = \Omega(n^{-1/2})$ and $n > s_0$. Since further we build on the Lasso estimator (4), we need $n = \Omega(s_0 \log(p/s_0))$ by the results of [DMM11, BM12]. In other words, the regime of interest for standard Gaussian designs is $c_1 s_0 \log(p/s_0) \leq n \leq c_2 s_0 \log(p)$. At the lower end the number of observations n is so small that essentially nothing can be inferred about $\mathrm{supp}(\theta_0)$ using optimally tuned Lasso estimator, and therefore a nontrivial power $1 - \beta > \alpha$ cannot be achieved. At the upper end, the number of samples is sufficient enough to recover $\mathrm{supp}(\theta_0)$ with high probability, leading to arbitrary small errors α, β. We consider the asymptotic scaling $s_0/p \to \varepsilon \in (0,1)$ which is indeed the asymptotic scaling covered by [BM12], and previously studied in [Don06, DT05, DT09, DT10] and [DMM09, DMM11]. In Section 3, we apply the results of [BM12] to this asymptotic regime, and develop a test that achieves non-trivial power, $\beta < 1 - \alpha$, provided $n \geq 2\,(1 + O(s_0/p))s_0 \log(p/s_0)$. We indeed show that our proposed testing procedure achieves nearly optimal power-significance trade-off.

Several papers study the estimation error under Gaussian design models, including [RWY10, CR11, CP10]. The last one, in particular, considers a much more general setting than the Gaussian one, but assumes again $n = \Omega(s_0 \log p)$.

As mentioned above, the asymptotic characterization of [BM12] only applies to standard Gaussian designs. A similar distributional limit is not available for general covariance matrices $\Sigma \neq \mathrm{I}_{p\times p}$. In Section 3.2 we conjecture such a limit on the basis of non-rigorous statistical physics calculations presented in Appendix B. Assuming that this conjecture holds, we derive a corresponding hypothesis testing procedure. It is worth mentioning that the replica method from statistical physics is already used by several groups for analyzing sparse regression problems, in particular by Ran-

gan, Fletcher, Goyal [RFG09], Kabashima, Tanaka and Takeda [KWT09, TK10], Guo, Baron and Shamai [GBS09], Wu and Verdú [WV11]. Earlier work applying the same method to the analysis of large CDMA systems includes [Tan02, GV05] (whose results were –in part– rigorously confirmed in [MT06, GW08]). This line of work is largely aimed at deriving asymptotically exact expressions for the risk of specific estimators, e.g., the Lasso or the Bayes optimal (minimum MSE) estimator. Most of the previous work in this line is limited to standard Gaussian setting. Exceptions include [TK10, TCSV11, KMV12] but they are limited either to the noiseless setting [TK10, KMV12] or to other matrix ensembles [TCSV11, KMV12]. To the best of our knowledge, the present paper is the first that applies the same techniques to high-dimensional hypothesis testing. Further, we consider a broader setting than the standard Gaussian one.

Let us finally mention that resampling methods provide an alternative path to assess statistical significance. A general framework to implement this idea is provided by the stability selection method of [MB10]. However, specializing the approach and analysis of [MB10] to the present context does not provide guarantees superior to [ZZ11, Büh12], that are more directly comparable to the present work.

1.2 Notations

We provide a brief summary of the notations used throughout the paper. For an $n \times p$ matrix $\mathbf{X}$, $\mathbf{X}_S$ denotes the $n \times |S|$ matrix with columns indices in S. Likewise, for a vector $\theta \in \mathbb{R}^p$, θ_S is the restriction of θ to indices in S. We denote the rows of the design matrix $\mathbf{X}$ by $x_1, \dots, x_n \in \mathbb{R}^p$. We also denote its columns by $\widetilde{x}_1, \dots, \widetilde{x}_p \in \mathbb{R}^n$. The support of a vector $\theta \in \mathbb{R}^p$ is denoted by $\mathrm{supp}(\theta)$, i.e., $\mathrm{supp}(\theta) = \{i \in [p], \theta_i \neq 0\}$. We use I to denote the identity matrix in any dimension, and $\mathrm{I}_{d\times d}$ whenever is useful to specify the dimension d. Throughout, $\Phi(x) \equiv \int_{-\infty}^{x} e^{-t^2/2}\mathrm{d}t/\sqrt{2\pi}$ is the CDF of the standard normal distribution.

2 Minimax formulation

2.1 Tests with guaranteed power

We consider the minimax criterion to measure the quality of a testing procedure. In order to define it formally, we first need to establish some notations.

A testing procedure for the family of hypotheses $H_{0,i}$, cf. Eq. (3), is given by a family of measurable functions

$$T_i : \quad \mathbb{R}^n \times \mathbb{R}^{n\times p} \to \{0,1\}\,. \tag{5}$$

$$(y, \mathbf{X}) \mapsto T_{i,\mathbf{X}}(y)\,. \tag{6}$$

Here $T_{i,\mathbf{X}}(y) = 1$ has the interpretation that hypothesis $H_{0,i}$ is rejected when the observation is $y \in \mathbb{R}^n$ and the design matrix is $\mathbf{X}$. We will hereafter drop the subscript $\mathbf{X}$ whenever clear from the context.

As mentioned above, we will measure the quality of a test T in terms of its significance level α (probability of type I errors) and power $1-\beta$ (β is the probability of type II errors). A type I error (false rejection of the null) leads one to conclude that a relationship between the response vector y and a column of the design matrix $\mathbf{X}$ exists when in reality it does not. On the other hand, a type II error (the failure to reject a false null hypothesis) leads one to miss an existing relationship.

Adopting a minimax point of view, we require that these metrics are achieved uniformly over s_0-sparse vectors. Formally, for $\mu > 0$, we let

$$\alpha_i(T) \equiv \sup\Big\{\mathbb{P}_\theta\big(T_{i,\mathbf{X}}(y) = 1\big) : \ \theta \in \mathbb{R}^p,\ \|\theta\|_0 \le s_0,\ \theta_i = 0\Big\}, \tag{7}$$

$$\beta_i(T;\mu) \equiv \sup\Big\{\mathbb{P}_\theta\big(T_{i,\mathbf{X}}(y) = 0\big) : \ \theta \in \mathbb{R}^p,\ \|\theta\|_0 \le s_0,\ |\theta_i| \ge \mu\Big\}. \tag{8}$$

In words, for any s_0-sparse vector with $\theta_i = 0$, the probability of false alarm is upper bounded by α. On the other hand, if θ is s_0-sparse with $|\theta_i| \ge \mu$, the probability of misdetection is upper bounded by β. Note that $\mathbb{P}_\theta(\cdot)$ is the induced probability distribution on $(y, \mathbf{X})$ for random design $\mathbf{X}$ and noise realization w, given the fixed parameter vector θ. Throughout we will accept randomized testing procedures as well[2].

Definition 2.1. *The* minimax power *for testing hypothesis $H_{0,i}$ against the alternative $|\theta_i| \ge \mu$ is given by the function $1 - \beta_i^{\rm opt}(\,\cdot\,;\mu) : [0,1] \to [0,1]$ where, for $\alpha \in [0,1]$*

$$1 - \beta_i^{\rm opt}(\alpha;\mu) \equiv \sup_T \Big\{1 - \beta_i(T;\mu) : \ \ \alpha_i(T) \le \alpha\Big\}. \tag{9}$$

Note that for standard Gaussian designs (and more generally for designs with exchangeable columns), $\alpha_i(T)$, $\beta_i(T;\mu)$ do not depend on the index $i \in [p]$. We shall therefore omit the subscript i in this case.

Remark 2.2. *The optimal power $\alpha \mapsto 1 - \beta_i^{\rm opt}(\alpha;\mu)$ is non-decreasing. Further, by using a test such that $T_{i,\mathbf{X}}(y) = 1$ with probability α independently of y, $\mathbf{X}$, we conclude that $1 - \beta_i(\alpha;\mu) \ge \alpha$.*

2.2 Upper bound on the minimax power

In this section we develop an upper bound for the minimax power $1 - \beta_i^{\rm opt}(\alpha;\mu)$ under the Gaussian random design model. Our basic tool is a simple reduction to the binary hypothesis testing problem.

Definition 2.3. *Let Q_0 be a probability distribution on $\mathbb{R}^p$ supported on $\Omega_0 \equiv \{\theta \in \mathbb{R}^p : \ \|\theta\|_0 \le s_0,\ \theta_i = 0\}$, and Q_1 a probability distribution supported on $\Omega_1 \equiv \{\theta \in \mathbb{R}^p : \ \|\theta\|_0 \le s_0,\ |\theta_i| \ge \mu\}$. For fixed design matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$, and $z \in \{0,1\}$, let $\mathbb{P}_{Q,z,\mathbf{X}}$ denote the law of y as per model (2) when θ_0 is chosen randomly with $\theta_0 \sim Q_z$.*

We denote by $1 - \beta_{i,\mathbf{X}}^{\rm bin}(\alpha;Q)$ the optimal power for the binary hypothesis testing problem $\theta_0 \sim Q_0$ versus $\theta_0 \sim Q_1$, namely:

$$\beta_{i,\mathbf{X}}^{\rm bin}(\alpha;Q) \equiv \inf_T \Big\{\ \mathbb{P}_{Q,1,\mathbf{X}}(T_{i,\mathbf{X}}(y) = 0) : \quad \mathbb{P}_{Q,0,\mathbf{X}}(T_{i,\mathbf{X}}(y) = 1) \le \alpha\ \Big\}. \tag{10}$$

The reduction is stated in the next lemma.

Lemma 2.4. *Let Q_0, Q_1 be any two probability measures supported, respectively, on Ω_0 and Ω_1 as per Definition 2.3. Then, the minimax power for testing hypothesis $H_{0,i}$ under the random design model, cf. Definition 2.1 is bounded as*

$$\beta_i^{\rm opt}(\alpha;\mu) \ge \inf\Big\{\mathbb{E}\beta_{i,\mathbf{X}}^{\rm bin}(\alpha_{\mathbf{X}};Q) : \quad \mathbb{E}(\alpha_{\mathbf{X}}) \le \alpha\ \Big\}. \tag{11}$$

[2]Formally, this corresponds to assuming $T_i(y) = T_i(y;U)$ with U uniform in $[0,1]$ and independent of the other random variables.

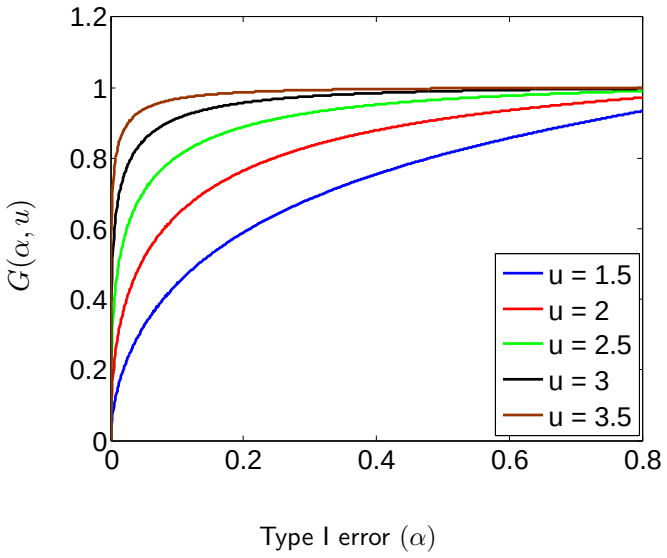


Figure 1: Function $G(\alpha, u)$ versus α for several values of u.

Here expectation is taken with respect to the law of $\mathbf{X}$ *and the* inf *is over all measurable functions* $\mathbf{X} \mapsto \alpha_{\mathbf{X}}$.

For the proof we refer to Section 6.1.

The binary hypothesis testing problem is addressed in the next lemma by reducing it to a simple regression problem. For $S \subseteq [p]$, we denote by P_S the orthogonal projector on the linear space spanned by the columns $\{\widetilde{x}_i\}_{i\in S}$. We also let $\mathrm{P}_S^{\perp} = \mathrm{I}_{n\times n} - \mathrm{P}_S$ be the projector on the orthogonal subspace. Further, for $\alpha \in [0,1]$ and $u \in \mathbb{R}_+$, define the function $G(\alpha, u)$ as follows.

$$G(\alpha, u) \equiv 2 - \Phi\Big(\Phi^{-1}(1-\frac{\alpha}{2}) + u\Big) - \Phi\Big(\Phi^{-1}(1-\frac{\alpha}{2}) - u\Big)\,. \tag{12}$$

In Fig. 1, the values of $G(\alpha, u)$ are plotted versus α for several values of u.

Lemma 2.5. *Let* $\mathbf{X} \in \mathbb{R}^{n\times p}$ *and* $i \in [p]$. *For* $S \subset [p] \setminus \{i\}$, $\alpha \in [0,1]$, *define*

$$1 - \beta^{\mathrm{oracle}}_{i,\mathbf{X}}(\alpha; S, \mu) \;\; = \;\; G\Big(\alpha, \frac{\mu \|\mathrm{P}_S^{\perp}\widetilde{x}_i\|_2}{\sigma}\Big)\,.$$

If $|S| < s_0$ *then for any* $\xi > 0$ *there exists distributions* Q_0, Q_1 *as per Definition 2.3, depending on* i, S, μ *but not on* $\mathbf{X}$, *such that* $\beta^{\mathrm{bin}}_{i,\mathbf{X}}(\alpha; Q) \geq \beta^{\mathrm{oracle}}_{i,\mathbf{X}}(\alpha; S, \mu) - \xi$.

The proof of this Lemma is presented in Section 6.2.

Using Lemma 2.4 and 2.5, we obtain the following upper bound on the optimal power of random Gaussian designs.

Theorem 2.6. *For $i \in [p]$, let $1-\beta_i^{\rm opt}(\alpha;\mu)$ be the minimax power of a Gaussian random design $\mathbf{X}$ with covariance matrix $\Sigma \in \mathbb{R}^{p\times p}$, as per Definition 2.1. For $S \subseteq [p]\setminus\{i\}$, define $\Sigma_{i|S} \equiv \Sigma_{ii} - \Sigma_{i,S}\Sigma_{S,S}^{-1}\Sigma_{S,i} \in \mathbb{R}$. Then, for any $\ell \in \mathbb{R}$ and $|S| < s_0$,*

$$1-\beta_i^{\rm opt}(\alpha;\mu) \le G\Big(\alpha, \frac{\mu}{\sigma_{\rm eff}(\ell)}\Big) + F_{n-s_0+1}(n-s_0+\ell)\,,$$
$$\sigma_{\rm eff}(\ell) \equiv \frac{\sigma}{\sqrt{\Sigma_{i|S}(n-s_0+\ell)}}\,,$$

where $F_k(x) = \mathbb{P}(Z_k \ge x)$, and Z_k is a chi-squared random variable with k degrees of freedom.

In other words, the statistical power is upper bounded by the one of testing the mean of a scalar Gaussian random variable, with effective noise variance $\sigma_{\rm eff}^2 \approx \sigma^2/[\Sigma_{i|S}(n-s_0)]$. (Note indeed that by concentration of a chi-squared random variable around their mean, ℓ can be taken small as compared to $n-s_0$.) The proof of this statement is to be found in Section 6.3.

The next corollary specializes the above result to the case of standard Gaussian designs. (The proof is immediate and hence we omit it.)

Corollary 2.7. *For $i \in [p]$, let $1-\beta_i^{\rm opt}(\alpha;\mu)$ be the minimax power of a standard Gaussian design $\mathbf{X}$ with covariance matrix $\Sigma = {\rm I}_{p\times p}$, cf. Definition 2.1. Then, for any $\xi \in [0, (3/2)\sqrt{n-s_0+1}]$ we have*

$$1-\beta_i^{\rm opt}(\alpha;\mu) \le G\Big(\alpha, \frac{\mu(\sqrt{n-s_0+1}+\xi)}{\sigma}\Big) + e^{-\xi^2/8}\,.$$

It is instructive to look at the last result from a slightly different point of view. Given $\alpha \in (0,1)$ and $1-\beta \in (\alpha,1)$, how big the entry μ needs to be so that $1-\beta_i^{\rm opt}(\alpha;\mu) \ge 1-\beta$? It is easy to check that, for any $\alpha > 0$, $u \mapsto G(\alpha,u)$ is continuous and monotone increasing with $G(\alpha,0) = \alpha$ and $\lim_{u\to\infty} G(\alpha,u) = 1$. It follows therefore from Corollary 2.7 that any pair (α,β) as above can be achieved if $\mu \ge \mu_{\rm UB} = c\sigma/\sqrt{n}$ for some $c = c(\alpha,\beta)$. Previous work [ZZ11, Büh12] is tailored to deterministic designs $\mathbf{X}$ and requires $\mu \ge c\max\{\sigma s_0 \log p/\, n, \sigma/\sqrt{n}\}$ to achieve the same goal (see Appendix A).

On the other hand, the upper bounds in Lemma 2.5 and Theorem 2.6 are obtained by assuming that the testing procedure knows ${\rm supp}(\theta)/\backslash\{i\}$, and this might be very optimistic. Surprisingly, these bounds turn to be tight, at least in an asymptotic sense, as demonstrated in the next Section.

3 The hypothesis testing procedure

3.1 Standard Gaussian designs

The authors of [BM12] characterize the high-dimensional behavior of the Lasso estimator for sequences of design matrices of increasing dimensions, with independent standard Gaussian entries. We build upon this result and propose a hypothesis testing procedure for testing null hypothesis $H_{0,i}$. We analyze it in the asymptotic setting and show that it achieves nearly optimal power-significance trade-off.

For given dimension p, an *instance* of the standard Gaussian design model is defined by the tuple (θ_0, n, σ), where $\theta_0 \in \mathbb{R}^p$, $n \in \mathbb{N}$, $\sigma \in \mathbb{R}_+$. We consider sequences of instances indexed by the problem dimension $\{(\theta_0(p), n(p), \sigma(p))\}_{p\in\mathbb{N}}$.

Definition 3.1. *The sequence of instances* $\{(\theta_0(p), n(p), \sigma(p))\}_{p\in\mathbb{N}}$ *indexed by* p *is said to be a* converging sequence *if* $n(p)/p \to \delta \in (0,\infty)$, $\sigma(p)^2/n \to \sigma_0^2$, *and the empirical distribution of the entries* $\theta_0(p)$ *converges weakly to a probability measure* p_{Θ_0} *on* $\mathbb{R}$ *with bounded second moment. Further* $p^{-1}\sum_{i\in[p]}\theta_{0,i}(p)^2 \to \mathbb{E}_{p_{\Theta_0}}\{\Theta_0^2\}$.

Note that this definition assumes that the coefficients $\theta_{0,i}$ are of order one, while the noise is scaled as $\sigma^2(p) = \Theta(n)$. Equivalently, we could have assumed $\theta_{0,i} = \Theta(1/\sqrt{n})$ and $\sigma^2(p) = \Theta(1)$, since the two settings only differ by a scaling of y. We favor the first scaling as it simplifies somewhat the notation in the following.

It is useful to recall the following result established in [BM12].

Proposition 3.2. *([BM12]) Let* $\{(\theta_0(p), n(p), \sigma(p))\}_{p\in\mathbb{N}}$ *be a converging sequence of instances of the standard Gaussian design model. Denote by* $\widehat{\theta} = \widehat{\theta}(y, \mathbf{X}, \lambda)$ *the Lasso estimator given as per Eq. (4) and define* $\widehat{\theta}^u \in \mathbb{R}^p$, $r \in \mathbb{R}^n$ *by letting*

$$\widehat{\theta}^u \equiv \widehat{\theta} + \frac{\mathsf{d}}{n}\,\mathbf{X}^{\mathsf{T}}(y - \mathbf{X}\widehat{\theta})\,, \qquad r \equiv \frac{\mathsf{d}}{\sqrt{n}}(y - \mathbf{X}\widehat{\theta})\,, \tag{13}$$

with $\mathsf{d} = (1 - \|\widehat{\theta}\|_0/n)^{-1}$. *Then, with probability one, the empirical distribution of* $\{(\theta_{0,i}, \widehat{\theta}_i^u)\}_{i=1}^p$ *converges weakly to the probability distribution of* $(\Theta_0, \Theta_0 + \tau Z)$, *for some* $\tau \in \mathbb{R}$, *where* $Z \sim \mathsf{N}(0,1)$, *and* $\Theta_0 \sim p_{\Theta_0}$ *is independent of* Z. *Furthermore, with probability one, the empirical distribution of* $\{r_i\}_{i=1}^p$ *converges weakly to* $\mathsf{N}(0,\tau^2)$.

In other words, $\widehat{\theta}^u$ *is an unbiased estimator of* θ_0, *and that its distribution is asymptotically normal*. Roughly speaking, the regression model (2) is asymptotically equivalent to a simpler sequence model

$$\widehat{\theta}^u = \theta_0 + \mathsf{noise} \tag{14}$$

with noise having zero mean. Further, the construction of $\widehat{\theta}^u$ has an appealing geometric interpretation. Notice that $\widehat{\theta}$ is necessarily biased towards small ℓ_1 norm. The minimizer in Eq. (4) must satisfy $(1/n)\,\mathbf{X}^{\mathsf{T}}(y - \mathbf{X}\widehat{\theta}) = \lambda g$, with g a subgradient of ℓ_1 norm at $\widehat{\theta}$. Hence, we can rewrite $\widehat{\theta}^u = \widehat{\theta} + \mathsf{d}\lambda g$. The bias is eliminated by modifying the estimator in the direction of increasing ℓ_1 norm. See Fig. 2 for an illustration.

Based on Proposition 3.2, we develop a hypothesis testing procedure as described in Table 1. The definitions of d and τ in step 2 are motivated by Proposition 3.2. Recall that $\mathsf{d}(y - \mathbf{X}\widehat{\theta})/\sqrt{n}$ is asymptotically normal with variance τ^2. In step 2, τ is estimated from data using median absolute deviation (MAD) estimator. This is a well-known estimator in robust statistic and is more resilient to outliers than the sample variance [HR09].

Finally, under the null hypothesis $H_{0,i}$, the quantity $\widehat{\theta}_i^u/(\tau)$ is asymptotically $\mathsf{N}(0,1)$. The definition of (two-sided) p-values P_i in step 4 follows. In the final step, the assigned p-value P_i is used to test the hypothesis $H_{0,i}$.

As before, we will measure the quality of the proposed test in terms of its significance level (size) α and power $1-\beta$. Recall that α and β respectively indicate the type I error (false positive) and type II error (false negative) rates. The following theorem establishes that the P_i's are indeed valid p-values, i.e., allow to control type I errors. Throughout $S_0(p) = \{i \in [p] : \theta_{0,i}(p) \neq 0\}$ is the support of $\theta_0(p)$.

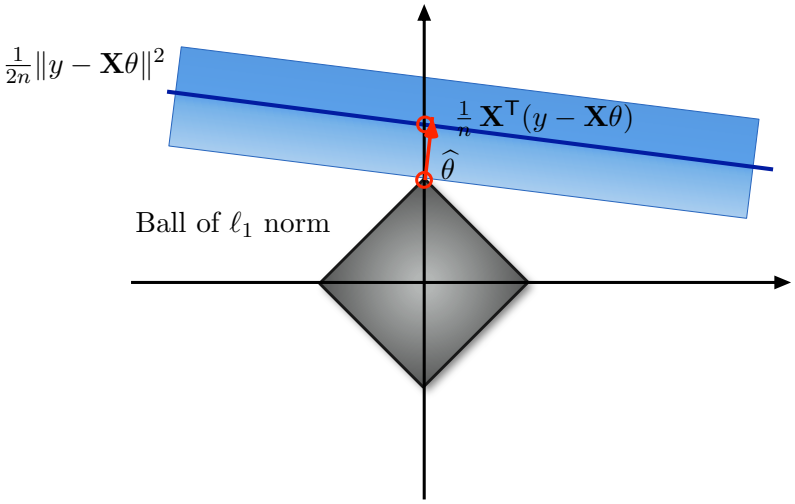


Figure 2: Geometric interpretation for construction of $\widehat{\theta}^u$. The bias in $\widehat{\theta}$ is eliminated by modifying the estimator in the direction of increasing its ℓ_1 norm

Theorem 3.3. *Let $\{(\theta_0(p), n(p), \sigma(p))\}_{p\in\mathbb{N}}$ be a converging sequence of instances of the standard Gaussian design model. Assume $\lim_{p\to\infty} |S_0(p)|/p = \mathbb{P}(\Theta_0 \neq 0)$. Then, for $i \in S_0^c(p)$, we have*

$$\lim_{p\to\infty} \mathbb{P}_{\theta_0(p)}(T_{i,\mathbf{X}}(y) = 1) = \alpha\,.$$

A more general form of Theorem 3.3 (cf. Theorem 3.6) is proved in Section 6. We indeed prove the stronger claim that the following holds true almost surely

$$\lim_{p\to\infty} \frac{1}{|S_0^c(p)|} \sum_{i\in S_0^c(p)} T_{i,\mathbf{X}}(y) = \alpha\,. \tag{15}$$

The result of Theorem 3.3 follows then by taking the expectation of both sides of Eq. (15) and using bounded convergence theorem and exchangeability of the columns of $\mathbf{X}$.

Our next theorem provides lower bound for the power of the proposed test. In order to obtain a non-trivial result, we need to make suitable sparsity assumptions on the parameter vectors $\theta_0 = \theta_0(p)$. In particular, we need to assume that the non-zero entries of θ_0 are lower bounded in magnitude. If this were not the case, it would be impossible to distinguish arbitrarily small parameters $\theta_{0,i}$ from $\theta_{0,i} = 0$. Similar assumptions are made in [MB06, ZY06, Wai09]. (The value of λ can also be predicted, but we omit it for brevity.)

Theorem 3.4. *Let $\{(\theta_0(p), n(p), \sigma(p))\}_{p\in\mathbb{N}}$ be a converging sequence of instances under the standard Gaussian design model. Assume that $|S_0(p)| \le \varepsilon p$, and for all $i \in S_0(p)$, $|\theta_{0,i}(p)| \ge \mu$ with $\mu = \mu_0\sigma(p)/\sqrt{n(p)}$. Then, for $i \in S_0(p)$, we have*

$$\lim_{p\to\infty} \mathbb{P}_{\theta_0(p)}(T_{i,\mathbf{X}}(y) = 1) \ge G\Big(\alpha, \frac{\mu_0}{\tau_*}\Big)\,,$$

Table 1: A hypothesis testing procedure for testing $H_{0,i}$ under standard Gaussian design model.

Testing hypothesis $H_{0,i}$ under standard Gaussian design model.

Input: regularization parameter λ, significance level α
Output: p-values P_i, test statistics $T_{i,\mathbf{X}}(y)$

1: Let
$$\widehat{\theta}(\lambda) = \mathrm{argmin}_{\theta\in\mathbb{R}^p}\ \Big\{\frac{1}{2n}\|y-\mathbf{X}\theta\|^2 + \lambda\|\theta\|_1\Big\}.$$

2: Let
$$\mathsf{d} = \left(1-\frac{1}{n}\|\widehat{\theta}(\lambda)\|_0\right)^{-1}, \qquad \tau = \frac{1}{\Phi^{-1}(0.75)}\frac{\mathsf{d}}{\sqrt{n}}|(y-\mathbf{X}\widehat{\theta}(\lambda))|_{(n/2)},$$

where for $v\in\mathbb{R}^K$, $|v|_\ell$ is the ℓ-th largest entry in the vector $(|v_1|,\cdots,|v_n|)$.

3: Let
$$\widehat{\theta}^u = \widehat{\theta}(\lambda) + \frac{\mathsf{d}}{n}\mathbf{X}^\mathsf{T}(y-\mathbf{X}\widehat{\theta}(\lambda)).$$

4: Assign the p-values P_i for the test $H_{0,i}$ as follows.
$$P_i = 2\Big(1-\Phi^{-1}\big(|\frac{\widehat{\theta}_i^u}{\tau}|\big)\Big).$$

5: The decision rule is then based on the p-values:
$$T_{i,\mathbf{X}}(y) = \begin{cases} 1 & \text{if } P_i\le\alpha \qquad \text{(reject the null hypothesis } H_{0,i}\text{)},\\ 0 & \text{otherwise} \qquad \text{(accept the null hypothesis)}.\end{cases}$$

where $\tau_* = \tau_*(\sigma_0,\varepsilon,\delta)$ is defined as follows
$$\tau_*^2 = \begin{cases}\dfrac{1}{1-M(\varepsilon)/\delta}, & \text{if } \delta > M(\varepsilon),\\ \infty, & \text{if } \delta\le M(\varepsilon).\end{cases} \tag{16}$$

Finally, $M(\varepsilon)$ is the minimax risk of the soft thresholding denoiser, with following parametric expression in terms of the parameter $\xi\in(0,\infty)$:
$$\varepsilon = \frac{2(\phi(\xi)-\xi\Phi(-\xi))}{\xi+2(\phi(\xi)-\xi\Phi(-\xi))}, \qquad M(\varepsilon) = \frac{2\phi(\xi)}{\xi+2(\phi(\xi)-\xi\Phi(-\xi))}. \tag{17}$$

Theorem 3.4 is proved in Section 6. We indeed prove the stronger claim that the following holds true almost surely:
$$\lim_{p\to\infty}\frac{1}{|S_0(p)|}\sum_{i\in S_0(p)} T_{i,\mathbf{X}}(y) \ge G\Big(\alpha,\frac{\mu_0}{\tau_*}\Big). \tag{18}$$

The result of Theorem 3.4 follows then by taking the expectation of both sides of Eq. (18) and using exchangeability of the columns of $\mathbf{X}$.

Again, it is convenient to rephrase Theorem 3.4 in terms of the minimum value of μ for which we can achieve statistical power $1-\beta \in (\alpha, 1)$ at significance level α. It is known that $M(\varepsilon) = 2\varepsilon \log(1/\varepsilon)\,(1+O(\varepsilon))$ [DMM11]. Hence, for $n \geq 2\, s_0 \log(p/s_0)\,(1+O(s_0/p))$, we have $\tau_*^2 = O(1)$. Since $\lim_{u\to\infty} G(\alpha, u) = 1$, any pre-assigned statistical power can be achieved by taking $\mu \geq C(\varepsilon, \delta)\sigma/\sqrt{n}$ which matches the fundamental limit established in the previous section.

3.1.1 Numerical experiments

As an illustration, we simulate from the linear model (1) with $w \sim \mathsf{N}(0, \mathrm{I}_{p\times p})$ and the following configurations.

Design matrix: For pairs of values $(n, p) = \{(300, 1000), (600, 1000), (600, 2000)\}$, the design matrix is generated from a realization of n i.i.d. rows $x_i \sim \mathsf{N}(0, \mathrm{I}_{p\times p})$.

Regression parameters: We consider active sets S_0 with $|S_0| = s_0 \in \{10, 20, 25, 50, 100\}$, chosen uniformly at random from the index set $\{1, \cdots, p\}$. We also consider two different strengths of active parameters $\theta_{0,i} = \mu$, for $i \in S_0$, with $\mu \in \{0.1, 0.15\}$.

We examine the performance of the proposed testing procedure (cf. Table 1) at significance levels $\alpha = 0.025, 0.05$. The experiments are done using glmnet-package in R that fits the entire Lasso path for linear regression models. Let $\varepsilon = s_0/p$ and $\delta = n/p$. We do not assume ε is known, but rather estimate it as $\bar{\varepsilon} = 0.25\,\delta/\log(2/\delta)$. The value of $\bar{\varepsilon}$ is half the maximum sparsity level ε for the given δ such that the Lasso estimator can correctly recover the parameter vector if the measurements were noiseless [DMM09, BM12]. Provided it makes sense to use Lasso at all, $\bar{\varepsilon}$ is thus a reasonable ballpark estimate.

Method	Type I err (mean)	Type I err (std.)	Avg. power (mean)	Avg. power (std)
Our testing Procedure $(1000, 600, 100, 0.1)$	0.05422	0.01069	0.44900	0.06951
Bühlmann's method $(1000, 600, 100, 0.1)$	0.01089	0.00358	0.13600	0.02951
Asymptotic Bound $(1000, 600, 100, 0.1)$	0.05	NA	0.37692	NA
Our testing Procedure $(1000, 600, 50, 0.1)$	0.04832	0.00681	0.52000	0.06928
Bühlmann's method $(1000, 600, 50, 0.1)$	0.01989	0.00533	0.17400	0.06670
Asymptotic Bound $(1000, 600, 50, 0.1)$	0.05	NA	0.51177	NA
Our testing Procedure $(1000, 600, 25, 0.1)$	0.06862	0.01502	0.56400	0.11384
Bühlmann's method $(1000, 600, 25, 0.1)$	0.02431	0.00536	0.25600	0.06586
Asymptotic Bound $(1000, 600, 25, 0.1)$	0.05	NA	0.58822	NA

Table 2: Comparison between our procedure (Table 1), Bühlmann's method [Büh12] and the asymptotic bound for our procedure (established in Theorem 3.4) on the setup described in Section 3.1.1. The significance level is $\alpha = 0.05$. The means and the standard deviations are obtained by testing over 10 realizations of the corresponding configuration. Here a quadruple such as $(1000, 600, 50, 0.1)$ denotes the values of $p = 1000$, $n = 600$, $s_0 = 50$, $\mu = 0.1$.

The regularization parameter λ is chosen to satisfy $\lambda \mathsf{d} = \kappa\tau$, where τ and d are determined in step 2 of the procedure. Here $\kappa = \kappa(\bar{\varepsilon})$ is the tuned parameter for the worst distribution p_{Θ_0} in

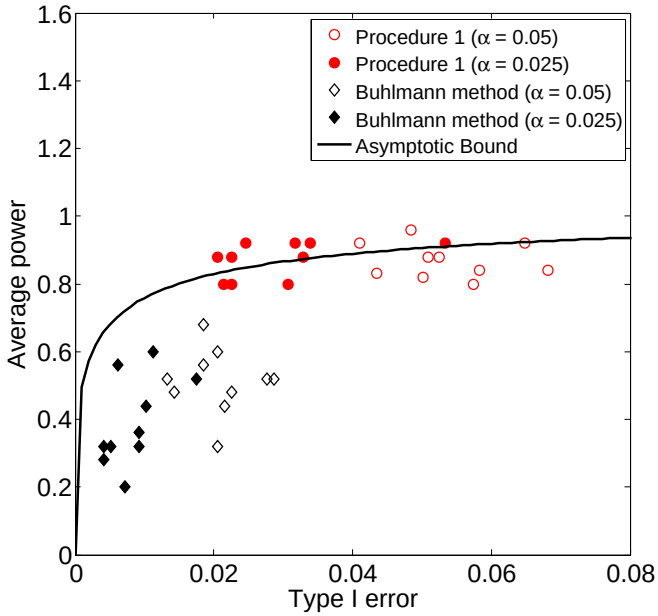


Figure 3: Comparison between our testing procedure (Table 1), Bühlmann's method [Büh12] and the asymptotic bound for our procedure (established in Theorem 3.4). Here, $p = 1000, n = 600, s_0 = 25, \mu = 0.15$.

the sense of minimax estimation error among the $\bar{\varepsilon}$-sparse distributions[3]. In [DMM09], its value is characterized for standard Gaussian design matrices.

It is worth noting that the proposed test can be used for any value of λ and its performance is robust for a wide range of values of λ. However, the above is an educated guess based on the analysis of [DMM09, BM12]. We also tried the values of λ proposed for instance in [vdGB09, Büh12] on the basis of oracle inequalities. Finally, note that τ and d implicitly depend upon λ. Since `glmnet` returns the entire Lasso path, the prescribed λ in above can simply be computed by applying the bisection method to equation $\lambda \mathsf{d} = \kappa\tau$.

Fig. 3 shows the results of our testing procedure and the method of [Büh12] for parameter values $p = 1000, n = 600, s_0 = 25, \mu = 0.15$, and significance levels $\alpha \in \{0.025, 0.05\}$. Each point in the plot corresponds to one realization of this configuration (there are a total of 10 realizations). We also depict the theoretical curve $(\alpha, G(\alpha, \mu_0/\tau_*))$, predicted by Theorem 3.4. As it can be seen the experiment results are in a good agreement with the theoretical curve.

We compare our procedure with the procedure proposed in [Büh12]. Table 2 summarizes the performances of the two methods for a few configurations (p, n, s_0, μ), and $\alpha = 0.05$. Simulation results for a larger number of configurations and $\alpha = 0.05, 0.025$ are reported in Tables 8 and 9 in Appendix C. As demonstrated by these results, the method of [Büh12] is very conservative. Namely, it achieves smaller type I error than the prescribed level α and this comes at the cost of a smaller statistical power than our testing procedure. This is to be expected since the approach of [Büh12] is

[3]A distribution p is ε-sparse if $p(\{0\}) \geq 1 - \varepsilon$.

tailored to adversarial design matrices $\mathbf{X}$.

3.2 Nonstandard Gaussian designs

In this section, we generalize our testing procedure to nonstandard Gaussian design models where the rows of the design matrix $\mathbf{X}$ are drawn independently form $\mathsf{N}(0,\Sigma)$. We will first consider the ideal case in which Σ is known. Later on, we will discuss the estimation of the covariance Σ (cf. SUBROUTINE in Table 4). Appendix D discusses an alternative implementation that does not estimate Σ but instead bounds the effect of unknown Σ.

For given dimension p, an *instance* of the nonstandard Gaussian design model is defined by the tuple $(\Sigma,\theta_0,n,\sigma)$, where $\Sigma\in\mathbb{R}^{p\times p}$, $\Sigma\succ 0$, $\theta_0\in\mathbb{R}^p$, $n\in\mathbb{N}$, $\sigma\in\mathbb{R}_+$. We will be interested in the asymptotic properties of sequences of instances indexed by the problem dimension $\{(\Sigma(p),\theta_0(p),n(p),\sigma(p))\}_{p\in\mathbb{N}}$. Motivated by Proposition 3.2, we define a property of a sequence of instances that we refer to as *standard distributional limit*.

Definition 3.5. *A sequence of instances $\{(\Sigma(p),\theta_0(p),n(p),\sigma(p))\}_{p\in\mathbb{N}}$ indexed by p is said to have an (almost sure)* standard distributional limit *if there exist $\tau,\mathsf{d}\in\mathbb{R}$, such that the following holds. Denote by $\widehat{\theta}=\widehat{\theta}(y,\mathbf{X},\lambda)$ the Lasso estimator given as per Eq. (4) and define $\widehat{\theta}^u\in\mathbb{R}^p$, $r\in\mathbb{R}^n$ by letting*

$$\widehat{\theta}^u\equiv\widehat{\theta}+\frac{\mathsf{d}}{n}\,\Sigma^{-1}\mathbf{X}^{\mathsf{T}}(y-\mathbf{X}\widehat{\theta})\,,\qquad r\equiv\frac{\mathsf{d}}{\sqrt{n}}(y-\mathbf{X}\widehat{\theta}). \tag{19}$$

Let $v_i=(\theta_{0,i},\widehat{\theta}^u_i,(\Sigma^{-1})_{i,i})$, for $1\le i\le p$, and $\nu^{(p)}$ be the empirical distribution of $\{v_i\}_{i=1}^p$ defined as

$$\nu^{(p)}=\frac{1}{p}\sum_{i=1}^{p}\delta_{v_i}\,, \tag{20}$$

where δ_{v_i} denotes the Dirac delta function centered at v_i. Then, with probability one, the empirical distribution $\nu^{(p)}$ converges weakly to a probability measure ν on $\mathbb{R}^3$ as $p\to\infty$. Here, ν is the probability distribution of $(\Theta_0,\Theta_0+\tau\Upsilon^{1/2}Z,\Upsilon)$, where $Z\sim\mathsf{N}(0,1)$, and Θ_0 and Υ are random variables independent of Z. Furthermore, with probability one, the empirical distribution of $\{r_i\}_{i=1}^p$ converges weakly to $\mathsf{N}(0,\tau^2)$.

Proving the standard distributional limit for specific families of instance sequences $\{(\Sigma(p),\theta_0(p),n(p),\sigma(p))\}_{p\in\mathbb{N}}$ is an outstanding mathematical challenge. In Section 4 we discuss a number of cases in which this can be established rigorously. We also outline a non-rigorous calculation using statistical physics methods that suggests significantly broader validity. Assuming validity of standard distributional limit, we generalize our hypothesis testing procedure to nonstandard Gaussian design models. In order to stress the use of the standard distributional limit, we refer to our test as SDL-TEST .

The hypothesis testing procedure (SDL-TEST) is described in Table 3. Our presentation of the SDL-TEST focuses on using exact covariance Σ to emphasize the validity of the proposed p-values.

Parameters d and τ in step 2 are defined in the same manner to the standard Gaussian designs. Notice that Definition 3.5 does not provide any explicit prescription for the value of d. Its definition in step 2 is indeed motivated by the general theory discussed in Section 4.

Table 3: SDL-TEST for testing hypothesis $H_{0,i}$ under nonstandard Gaussian design model

SDL-TEST: Testing hypothesis $H_{0,i}$ under nonstandard Gaussian design model.

Input: regularization parameter λ, significance level α, covariance matrix Σ
Output: p-values P_i, test statistics $T_{i,\mathbf{X}}(y)$

1: Let

$$\widehat{\theta}(\lambda) = \mathrm{argmin}_{\theta\in\mathbb{R}^p} \Big\{ \frac{1}{2n}\|y - \mathbf{X}\theta\|^2 + \lambda\|\theta\|_1 \Big\}.$$

2: Let

$$\mathsf{d} = \left(1 - \frac{1}{n}\|\widehat{\theta}(\lambda)\|_0\right)^{-1}, \qquad \tau = \frac{1}{\Phi^{-1}(0.75)}\frac{\mathsf{d}}{\sqrt{n}}|(y - \mathbf{X}\widehat{\theta}(\lambda))|_{(n/2)},$$

where for $v \in \mathbb{R}^K$, $|v|_\ell$ is the ℓ-th largest entry in the vector $(|v_1|, \cdots, |v_n|)$.

3: Let

$$\widehat{\theta}^u = \widehat{\theta}(\lambda) + \frac{\mathsf{d}}{n}\Sigma^{-1}\mathbf{X}^\mathsf{T}(y - \mathbf{X}\widehat{\theta}(\lambda)).$$

4: Assign the p-values P_i for the test $H_{0,i}$ as follows.

$$P_i = 2\Big(1 - \Phi^{-1}\big(|\frac{\widehat{\theta}_i^u}{\tau[(\Sigma^{-1})_{i,i}]^{1/2}}|\big)\Big).$$

5: The decision rule is then based on the p-values:

$$T_{i,\mathbf{X}}(y) = \begin{cases} 1 & \text{if } P_i \le \alpha \qquad \text{(reject the null hypothesis } H_{0,i}\text{)}, \\ 0 & \text{otherwise} \qquad \text{(accept the null hypothesis)}. \end{cases}$$

Under the assumption of a standard distributional limit and assuming null hypothesis $H_{0,i}$, the quantity $\widehat{\theta}_i^u/(\tau[(\Sigma^{-1})_{i,i}]^{1/2})$ is asymptotically $\mathsf{N}(0,1)$, whence the definition of (two-sided) p-values P_i follows as in step 4. In the final step, the assigned p-value P_i is used to test the hypothesis $H_{0,i}$.

The following theorem is a generalization of Theorem 3.3 to nonstandard Gaussian designs under the standard distributional limit.

Theorem 3.6. *Let $\{(\Sigma(p), \theta_0(p), n(p), \sigma(p))\}_{p\in\mathbb{N}}$ be a sequence of instances for which a standard distributional limit holds. Further assume $\lim_{p\to\infty}|S_0(p)|/p = \mathbb{P}(\Theta_0 \neq 0)$. Then,*

$$\lim_{p\to\infty}\frac{1}{|S_0^c(p)|}\sum_{i\in S_0^c(p)}\mathbb{P}_{\theta_0(p)}(T_{i,\mathbf{X}}(y) = 1) = \alpha\,.$$

The proof of Theorem 3.6 is deferred to Section 6. In the proof, we show the stronger result that

the following holds true almost surely

$$\lim_{p\to\infty} \frac{1}{|S_0^c(p)|} \sum_{i\in S_0^c(p)} T_{i,\mathbf{X}}(y) = \alpha\,. \tag{21}$$

The result of Theorem 3.6 follows then by taking the expectation of both sides of Eq. (21) and using bounded convergence theorem.

The following theorem characterizes the power of SDL-TEST for general Σ, and under the assumption that a standard distributional limit holds .

Theorem 3.7. *Let $\{(\Sigma(p), \theta_0(p), n(p), \sigma(p))\}_{p\in\mathbb{N}}$ be a sequence of instances with standard distributional limit. Assume (without loss of generality) $\sigma(p) = \sqrt{n(p)}$, and further $|\theta_{0,i}(p)|/[(\Sigma^{-1})_{i,i}]^{1/2} \geq \mu_0$ for all $i \in S_0(p)$. Then,*

$$\lim_{p\to\infty} \frac{1}{|S_0(p)|} \sum_{i\in S_0(p)} \mathbb{P}_{\theta_0(p)}(T_{i,\mathbf{X}}(y) = 1) \geq G\Big(\alpha, \frac{\mu_0}{\tau}\Big)\,.$$

Theorem 3.7 is proved in Section 6. We indeed prove the stronger result that the following holds true almost surely

$$\lim_{p\to\infty} \frac{1}{|S_0(p)|} \sum_{i\in S_0(p)} T_{i,\mathbf{X}}(y) \geq G\Big(\alpha, \frac{\mu_0}{\tau}\Big)\,.$$

We also notice that in contrast to Theorem 3.4, where τ_* has an explicit formula that leads to an analytical lower bound for the power (for a suitable choice of λ), in Theorem 3.7, τ depends upon λ implicitly and can be estimated from the data as in step 3 of SDL-TEST procedure. The result of Theorem 3.7 holds for any value of λ.

Notice that in general the exact covariance Σ is not available and we need to use an estimate of that in SDL-TEST. There are several high-dimensional covariance estimation methods that provide a consistent estimate $\widehat{\Sigma}$, under suitable structural assumptions on Σ. For instance, if Σ is sparse, $\widehat{\Sigma}$ can be constructed by thresholding the empirical covariance, cf. Table 4. Note that the Lasso is unlikely to perform well if the columns of $\mathbf{X}$ are highly correlated and hence the assumption of sparse Σ is very natural. If the inverse covariance Σ^{-1} is sparse, the graphical model method of [MB06] can be used instead.

Appendix D describes an alternative covariance-free procedure that only uses bounds on Σ where the bounds are estimated from the data. In our numerical experiments and comparisons with other methods, we use the estimated covariance returned by SUBROUTINE. The p-values computation appears to be fairly robust with respect to errors in the estimation of Σ.

3.2.1 Numerical experiments

We consider the same setup as the one in Section 3.1.1 except that the rows of the design matrix are independently $x_i \sim \mathsf{N}(0, \Sigma)$. Here $\Sigma \in \mathbb{R}^{p\times p}$ is a circulant matrix with $\Sigma_{ii} = 1$, $\Sigma_{jk} = 0.1$ for $j \neq k$, $|j-k| \leq 5$ and zero everywhere else. (The difference between indices is understood modulo p.)

Table 4: SUBROUTINE for estimating covariance Σ

SUBROUTINE: Estimating covariance matrix Σ

Input: Design matrix $\mathbf{X}$
Output: Estimate $\widehat{\Sigma}$
1: Let $S = (1/n)\mathbf{X}^{\mathsf{T}}\mathbf{X} \in \mathbb{R}^{p\times p}$.
2: Let σ_1 be the empirical variance of the entries in S, and let $\mathcal{A} = \{S_{ij} : |S_{ij}| \le 3\sigma_1\}$;
3: Fit a normal distribution to the entries in $\mathcal{A}$; let σ_2 be the variance of this distribution;
4: Construct the estimation $\widehat{\Sigma}$ as follows:

$$\widehat{\Sigma}_{ij} = S_{ij}\,\mathbb{I}(|S_{ij}| \ge 3\sigma_2). \tag{22}$$

In Fig. 4(a), we compare SDL-TEST with the procedure proposed in [Büh12]. While the type I errors of SDL-TEST are in good match with the chosen significance level α, Bühlmann's method is overly conservative. It results in significantly smaller type I errors than α and smaller average power in return. Table 5 summarizes the performances of the two methods for a few configurations (p, n, s_0, μ), and $\alpha = 0.05$. Simulation results for a larger number of configurations and $\alpha = 0.05, 0.025$ are reported in Tables 10 and 11 in Appendix C.

Let $Z = (z_i)_{i=1}^p$ denote the vector with $z_i \equiv \widehat{\theta}_i^u/(\tau[(\Sigma^{-1})_{ii}]^{1/2})$. In Fig. 4(b) we plot the normalized histograms of Z_{S_0} (in red) and $Z_{S_0^c}$ (in white), where Z_{S_0} and $Z_{S_0^c}$ respectively denote the restrictions of Z to the active set S_0 and the inactive set S_0^c. The plot clearly exhibits the fact that $Z_{S_0^c}$ has (asymptotically) standard normal distribution, and the histogram of Z_{S_0} appears as a distinguishable bump. This is the core intuition in defining SDL-TEST.

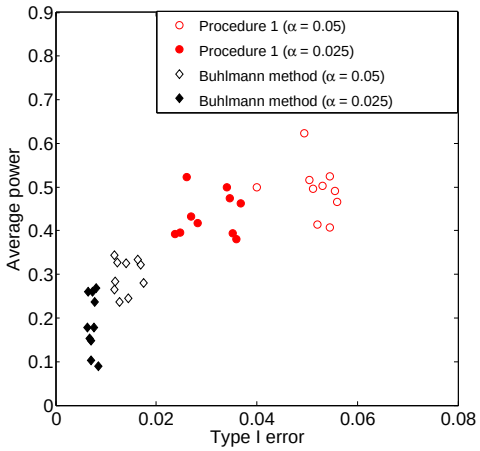
4 Generalization and discussion

In previous sections we described our hypothesis testing procedure (SDL-TEST) using the Lasso estimator. Lasso estimator is particularly useful when one seeks sparse parameter vectors θ satisfying Eq. (2). In general, other penalty functions than ℓ_1 norm might be used based on the prior knowledge about θ_0. Here, we consider regularized least squares estimators of the form

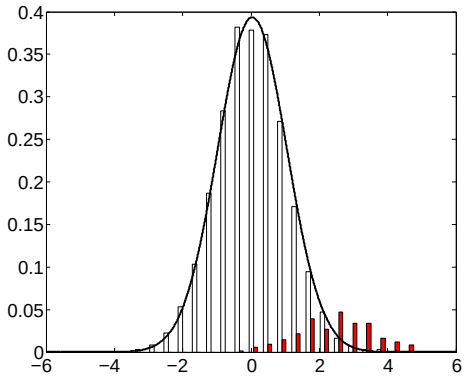
$$\widehat{\theta}(y, \mathbf{X}) = \arg\min_{\theta\in\mathbb{R}^p} \left\{ \frac{1}{2n}\|y - \mathbf{X}\theta\|^2 + J(\theta) \right\}, \tag{23}$$

with $J(\theta)$ being a convex separable penalty function; namely for a vector $\theta \in \mathbb{R}^p$, we have $J(\theta) = J_1(\theta_1) + \cdots + J_p(\theta_p)$, where $J_\ell : \mathbb{R} \to \mathbb{R}$ is a convex function. Important instances from this ensemble of estimators are Ridge-regression ($J(\theta) = \lambda\|\theta\|^2/2$), and the Lasso ($J(\theta) = \lambda\|\theta\|_1$).

Assuming the general penalty function $J(\theta)$, the standard distributional limit is defined similar to Definition 3.5 except that the Lasso estimator $\widehat{\theta} = \widehat{\theta}(y, \mathbf{X}, \lambda)$ is replaced by the estimator $\widehat{\theta} = \widehat{\theta}(y, \mathbf{X})$ given by Eq. (23).



(a) Comparison between SDL-TEST and Bühlmann's method [Büh12].



(b) Normalized histograms of Z_{S_0} (in red) and $Z_{S_0^c}$ (in white) for one realization.

Figure 4: Simulation results for the setting in Section 3.2.1 and the configuration $p = 2000$, $n = 600$, $s_0 = 50$, $\mu = 0.1$.

Method	Type I err (mean)	Type I err (std.)	Avg. power (mean)	Avg. power (std)
SDL-test $(1000, 600, 100, 0.1)$	0.06733	0.01720	0.48300	0.03433
Bühlmann's method $(1000, 600, 100, 0.1)$	0.00856	0.00416	0.11000	0.02828
Lower bound $(1000, 600, 100, 0.1)$	0.05	NA	0.45685	0.04540
SDL-test $(1000, 600, 50, 0.1)$	0.04968	0.00997	0.50800	0.05827
Bühlmann's method $(1000, 600, 50, 0.1)$	0.01642	0.00439	0.21000	0.04738
Lower bound $(1000, 600, 50, 0.1)$	0.05	NA	0.50793	0.03545
SDL-test $(1000, 600, 25, 0.1)$	0.05979	0.01435	0.55200	0.08390
Bühlmann's method $(1000, 600, 25, 0.1)$	0.02421	0.00804	0.22400	0.10013
Lower bound $(1000, 600, 25, 0.1)$	0.05	NA	0.54936	0.06176

Table 5: Comparison between SDL-TEST, Bühlmann's method [Büh12] and the lower bound for SDL-TEST power (cf. Theorem 3.7) on the setup described in Section 3.2.1. The significance level is $\alpha = 0.05$. The means and the standard deviations are obtained by testing over 10 realizations of the corresponding configuration. Here a quadruple such as $(1000, 600, 50, 0.1)$ denotes the values of $p = 1000$, $n = 600$, $s_0 = 50$, $\mu = 0.1$.

Generalizing SDL-TEST for convex separable penalty functions $J(\theta)$ is immediate. The only required modification is about the definition of d in step 2. We let d be the unique positive solution of the following equation

$$1 = \frac{1}{\mathsf{d}} + \frac{1}{n} \mathrm{Trace}\Big\{ (1 + \mathsf{d}\Sigma^{-1/2} \nabla^2 J(\widehat{\theta}) \Sigma^{-1/2})^{-1} \Big\}, \tag{24}$$

where $\nabla^2 J(\widehat{\theta})$ denotes the Hessian, which is diagonal since J is separable. If J is nondifferentiable,

then we formally set $[\nabla^2 J(\widehat{\theta})]_{ii} = \infty$ for all the coordinates i such that J is non-differentiable at $\widehat{\theta}_i$. It can be checked that this definition is well posed and that yields the previous choice for $J(\theta) = \lambda\|\theta\|_1$. We next discuss the validity of standard distributional limit and the rationale for the above choice of d.

1. For $\Sigma = \mathrm{I}_{p\times p}$ and $J(\theta) = \lambda\|\theta\|_1$ [BM11, BM12] proves that indeed the standard distributional limit holds.

2. For $\Sigma = \mathrm{I}_{p\times p}$ and general separable $J(\theta)$, a formal proof of the same statement does not exist. However, the results of Talagrand on the Shcherbina-Tirozzi model (cf. [Tal10, Theorem 3.2.14] and comment below) imply the claim for strictly convex $J(\theta)$ under the assumption that a set of three non-linear equations admit unique solution. This can be checked on a case-by-case basis. Additional evidence is provided by the remark that the AMP algorithm to compute $\widehat{\theta}$ satisfies the standard distributional limit at each iteration.

3. The work in [Ran11, JM12] extends the analysis of AMP to a larger class of algorithms called G-AMP. This suggests that the standard distributional limit can be shown to hold for special block-diagonal matrices Σ as well.

4. Finally, the broadest domain of validity of the standard distributional limit can be established using the replica method from statistical physics. This is a non-rigorous but mathematically sophisticated technique, originally devised to treat mean-field models of spin glasses [MPV87]. Over the last 20 years, its domain was considerably extended [MM09], and in fact it was already successfully applied to estimation problems under the noisy linear model (1) in the case $\Sigma = \mathrm{I}$ [Tan02, GV05, KWT09, RFG09]. While its validity was confirmed by rigorous analysis in a number of examples, developing an equally powerful probabilistic method remains an outstanding challenge.

Replica Method Claim 4.1. *Assume the sequence of instances $\{(\Sigma(p), \theta(p), n(p), \sigma(p))\}_{p\in\mathbb{N}}$ to be such that, as $p \to \infty$: (i) $n(p)/p \to \delta > 0$; (ii) $\sigma(p)^2/n(p) \to \sigma_0^2 > 0$; (iii) The sequence of functions*

$$\mathfrak{E}^{(p)}(a,b) \equiv \frac{1}{p}\,\mathbb{E}\min_{\theta\in\mathbb{R}^p}\left\{\frac{b}{2}\|\theta-\theta_0-\sqrt{a}\Sigma^{-1/2}Z\|_\Sigma^2 + J(\theta)\right\}, \tag{25}$$

with $\|v\|_\Sigma^2 \equiv \langle v, \Sigma v\rangle$ and $Z \sim \mathsf{N}(0, \mathrm{I}_{p\times p})$ admits a differentiable limit $\mathfrak{E}(a,b)$ on $\mathbb{R}_+ \times \mathbb{R}_+$, with $\nabla\mathfrak{E}^{(p)}(a,b) \to \nabla\mathfrak{E}(a,b)$. Then $\mathcal{S}$ has a standard distributional limit. Further let

$$\eta_b(y) \equiv \arg\min_{\theta\in\mathbb{R}^p}\left\{\frac{b}{2}\|\theta-y\|_\Sigma^2 + J(\theta)\right\}, \tag{26}$$

$$\mathsf{E}_1(a,b) \equiv \lim_{p\to\infty}\frac{1}{p}\mathbb{E}\left\{\left\|\eta_b(\theta_0+\sqrt{a}\Sigma^{-1/2}Z)-\theta_0\right\|_\Sigma^2\right\}, \tag{27}$$

where the the limit exists by the above assumptions on the convergence of $\mathfrak{E}^{(p)}(a,b)$. Then, the parameters τ and d of the standard distributional limit are obtained by solving Eq. (24) for d and

$$\tau^2 = \sigma_0^2 + \frac{1}{\delta}\,\mathsf{E}_1(\tau^2, 1/\mathsf{d})\,. \tag{28}$$

We refer to Appendix B for the related statistical physics calculations.

It is worth stressing that convergence assumption for the sequence $\mathfrak{E}^{(p)}(a,b)$ is quite mild, and is satisfied by a large family of covariance matrices. For instance, it can be proved that it holds for block-diagonal matrices Σ as long as the blocks empirical distribution converges.

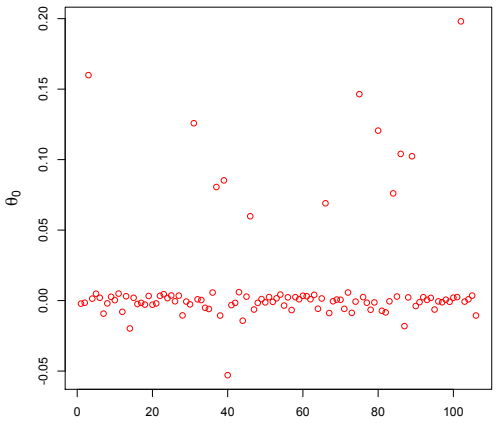


Figure 5: Parameter vector θ_0 for the communities data set.

5 Real data application

We tested our method on the UCI communities and crimes dataset [FA10]. This concerns the prediction of the rate of violent crime in different communities within US, based on other demographic attributes of the communities. The dataset consists of a response variable along with 122 predictive attributes for 1994 communities. Covariates are quantitative, including e.g., the fraction of urban population or the median family income. We consider a linear model as in (2) and hypotheses $H_{0,i}$. Rejection of $H_{0,i}$ indicates that the i-th attribute is significant in predicting the response variable.

We perform the following preprocessing steps: (i) Each missing value is replaced by the mean of the non missing values of that attribute for other communities. (ii) We eliminate 16 attributes to make the ensemble of the attribute vectors linearly independent. Thus we obtain a design matrix $\mathbf{X}_{\text{tot}} \in \mathbb{R}^{n_{\text{tot}} \times p}$ with $n_{\text{tot}} = 1994$ and $p = 106$; (iii) We normalize each column of the resulting design matrix to have mean zero and ℓ_2 norm equal to $\sqrt{n_{\text{tot}}}$.

In order to evaluate various hypothesis testing procedures, we need to know the true significant variables. To this end, we let $\theta_0 = (\mathbf{X}_{\text{tot}}^\mathsf{T}\mathbf{X}_{\text{tot}})^{-1}\mathbf{X}_{\text{tot}}^\mathsf{T}y$ be the least-square solution. We take θ_0 as the true parameter vector obtained from the whole data set. Fig. 5 shows the the entries of θ_0. Clearly, only a few entries have non negligible values which correspond to the significant attributes. In computing type I errors and powers, we take the elements in θ_0 with magnitude larger than 0.04 as active and the others as inactive.

We take random subsamples of size $n = 84$ from the communities. We compare SDL-TEST with Buhlmann's method over 20 realizations and significance levels $\alpha = 0.01, 0.025, 0.05$. Type I errors and powers are computed by comparing to θ_0. Table 6 summarizes the results. As it can be seen, Buhlmann's method is conservative yielding to zero type I error and smaller power than SDL-TEST in return.

In table 7, we report the relevant features obtained from the whole dataset as described above i.e., θ_0, and the relevant features predicted by SDL-TEST and the Bühlmann's method from one random subsample of communities of size $n = 84$. Features description is available at [FA10].

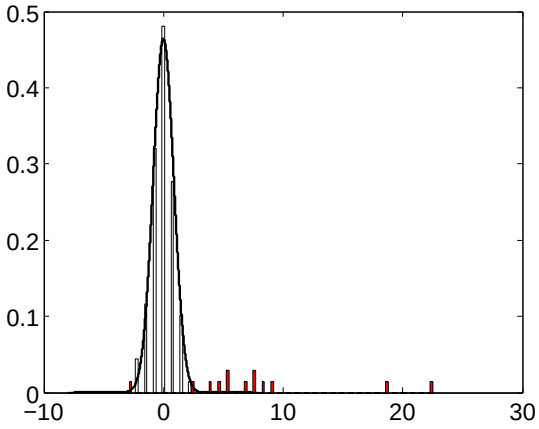


Figure 6: Normalized histogram of Z_{S_0} (in red) and $Z_{S_0^c}$ (in white) for the communities data set.

Method	Type I err (mean)	Avg. power (mean)
SDL-test ($\alpha = 0.05$)	0.0172043	0.4807692
Bühlmann's method	0	0.1423077
SDL-test ($\alpha = 0.025$)	0.01129032	0.4230769
Bühlmann's method	0	0.1269231
SDL-test ($\alpha = 0.01$)	0.008602151	0.3576923
Bühlmann's method	0	0.1076923

Table 6: Simulation results for the communities data set.

Finally, in Fig. 6 we plot the normalized histograms of Z_{S_0} (in red) and $Z_{S_0^c}$ (in white). Recall that $Z = (z_i)_{i=1}^p$ denotes the vector with $z_i \equiv \widehat{\theta}_i^u/(\tau[(\Sigma^{-1})_{ii}]^{1/2})$. Further, Z_{S_0} and $Z_{S_0^c}$ respectively denote the restrictions of Z to the active set S_0 and the inactive set S_0^c. This plot demonstrates that $Z_{S_0^c}$ has roughly standard normal distribution as predicted by the theory.

Relevant features		racePctHisp, PctTeen2Par, PctImmigRecent, PctImmigRec8, PctImmigRec10, PctNotSpeakEnglWell, OwnOccHiQuart, NumStreet, PctSameState85, LemasSwFTFieldPerPop, LemasTotReqPerPop, RacialMatchCommPol, PolicOperBudg
$\alpha = 0.01$	Relevant features (SDL-TEST)	racePctHisp, PctTeen2Par, PctImmigRecent, PctImmigRec8, PctImmigRec10, PctNotSpeakEnglWell, OwnOccHiQuart, NumStreet, PctSameState85, LemasSwFTFieldPerPop, LemasTotReqPerPop, RacialMatchCommPol, PolicOperBudg
	Relevant features (Bühlmann's method)	racePctHisp, PctSameState85
$\alpha = 0.025$	Relevant features (SDL-TEST)	racePctHisp, PctTeen2Par, PctImmigRecent, PctImmigRec8, PctImmigRec10, PctNotSpeakEnglWell, PctHousOccup, OwnOccHiQuart, NumStreet, PctSameState85, LemasSwFTFieldPerPop, LemasTotReqPerPop, RacialMatchCommPol, PolicOperBudg
	Relevant features (Bühlmann's method)	racePctHisp, PctSameState85
$\alpha = 0.05$	Relevant features (SDL-TEST)	racePctHisp, PctUnemployed, PctTeen2Par, PctImmigRecent, PctImmigRec8, PctImmigRec10, PctNotSpeakEnglWell, PctHousOccup, OwnOccHiQuart, NumStreet, PctSameState85, LemasSwornFT, LemasSwFTFieldPerPop, LemasTotReqPerPop, RacialMatchCommPol, PctPolicWhite
	Relevant features (Bühlmann's method)	racePctHisp, PctSameState85

Table 7: The relevant features (using the whole dataset) and the relevant features predicted by SDL-TEST and Bühlmann's method [Büh12] for a random subsample of size $n = 84$ from the communities. The false positive predictions are in red.

6 Proofs

6.1 Proof of Lemma 2.4

Fix $\alpha \in [0,1]$, $\mu > 0$, and assume that the minimum error rate for type II errors in testing hypothesis $H_{0,i}$ at significance level α is $\beta = \beta_i^{\text{opt}}(\alpha;\mu)$. Further fix $\xi > 0$ arbitrarily small. By definition there exists a statistical test $T_{i,\mathbf{X}} : \mathbb{R}^m \to \{0,1\}$ such that $\mathbb{P}_\theta(T_{i,\mathbf{X}}(y) = 1) \le \alpha$ for any $\theta \in \Omega_0$ and $\mathbb{P}_\theta(T_{i,\mathbf{X}}(y) = 0) \le \beta + \xi$ for any $\theta \in \Omega_1$ (with $\Omega_0, \Omega_1 \in \mathbb{R}^p$ defined as in Definition 2.3). Equivalently:

$$\begin{aligned} &\mathbb{E}\big\{\mathbb{P}_\theta(T_{i,\mathbf{X}}(y) = 1|\mathbf{X})\big\} \le \alpha, \qquad \text{for any } \theta \in \Omega_0, \\ &\mathbb{E}\big\{\mathbb{P}_\theta(T_{i,\mathbf{X}}(y) = 0|\mathbf{X})\big\} \le \beta + \xi, \quad \text{for any } \theta \in \Omega_1. \end{aligned}$$

We now take expectation of these inequalities with respect to $\theta \sim Q_0$ (in the first case) and $\theta \sim Q_1$ (in the second case) and we get, with the notation introduced in the Definition 2.3,

$$\begin{aligned} &\mathbb{E}\big\{\mathbb{P}_{Q,0,\mathbf{X}}(T_{i,\mathbf{X}}(y) = 1)\big\} \le \alpha\,, \\ &\mathbb{E}\big\{\mathbb{P}_{Q,1,\mathbf{X}}(T_{i,\mathbf{X}}(y) = 0)\big\} \le \beta + \xi\,. \end{aligned}$$

Call $\alpha_{\mathbf{X}} \equiv \mathbb{P}_{Q,0,\mathbf{X}}(T_{i,\mathbf{X}}(y) = 1)$. By assumption, for any test T, we have $\mathbb{P}_{Q,1,\mathbf{X}}(T_{i,\mathbf{X}}(y) = 0) \geq \beta^{\mathrm{bin}}_{i,\mathbf{X}}(\alpha_{\mathbf{X}}; Q)$ and therefore the last inequalities imply

$$\mathbb{E}\{\alpha_{\mathbf{X}}\} \leq \alpha\,,$$
$$\mathbb{E}\{\beta^{\mathrm{bin}}_{i,\mathbf{X}}(\alpha_{\mathbf{X}}; Q)\} \leq \beta + \xi\,.$$

The thesis follows since $\xi > 0$ is arbitrary.

6.2 Proof of Lemma 2.5

Fix $\mathbf{X}$, α, i, S as in the statement and assume, without loss of generality, $\mathrm{P}_S^{\perp}\widetilde{x}_i \neq 0$, and $\mathrm{rank}(\mathbf{X}_S) = |S| < n$. We take $Q_0 = \mathsf{N}(0, M\,\mathrm{I}_S)$ where $M \in \mathbb{R}_+$ and $\mathrm{I}_S \in \mathbb{R}^{p\times p}$ is the diagonal matrix with $(\mathrm{I}_S)_{ii} = 1$ if $i \in S$ and $(\mathrm{I}_S)_{ii} = 0$ otherwise. For the same covariance matrix I_S, we let $Q_1 = \mathsf{N}(\mu\, e_i, M\,\mathrm{I}_S)$ where e_i is the i-th element of the standard basis. Recalling that $i \notin S$, and $|S| < s_0$, the support of Q_0 is in Ω_0 and the support of Q_1 is in Ω_1.

Under $\mathbb{P}_{Q,0,\mathbf{X}}$ we have $y \sim \mathsf{N}(0, M\,\mathbf{X}_S\mathbf{X}_S^{\mathsf{T}} + \sigma^2\mathrm{I})$, and under $\mathbb{P}_{Q,1,\mathbf{X}}$ we have $y \sim \mathsf{N}(\mu\widetilde{x}_i, M\,\mathbf{X}_S\mathbf{X}_S^{\mathsf{T}} + \sigma^2\mathrm{I})$. Hence the binary hypothesis testing problem under study reduces to the problem of testing a null hypothesis on the mean of a Gaussian random vector with known covariance against a simple alternative. It is well known that the most powerful test [LR05, Chapter 8] is obtained by comparing the ratio $\mathbb{P}_{Q,0,\mathbf{X}}(y)/\mathbb{P}_{Q,1,\mathbf{X}}(y)$ with a threshold. Equivalently, the most powerful test is of the form

$$T_{i,\mathbf{X}}(y) = \mathbb{I}\Big\{\langle \mu\widetilde{x}_i, (M\mathbf{X}_S\mathbf{X}_S^{\mathsf{T}} + \sigma^2\mathrm{I})^{-1}y\rangle \geq c\Big\}\,,$$

for some $c \in \mathbb{R}$ that is to be chosen to achieve the desired significance level α. Letting

$$\alpha \equiv 2\Phi\Big(-\frac{c}{\mu\|(M\mathbf{X}_S\mathbf{X}_S^{\mathsf{T}} + \sigma^2\mathrm{I})^{-1/2}\widetilde{x}_i\|}\Big)\,,$$

it is a straightforward calculation to drive the power of this test as

$$G\Big(\alpha, \frac{\mu\widetilde{x}_i^{\mathsf{T}}(M\mathbf{X}_S\mathbf{X}_S^{\mathsf{T}} + \sigma^2\mathrm{I})^{-1}\widetilde{x}_i}{\|(M\mathbf{X}_S\mathbf{X}_S^{\mathsf{T}} + \sigma^2\mathrm{I})^{-1/2}\widetilde{x}_i\|}\Big)\,.$$

where the function $G(\alpha, u)$ is defined as per Eq. (12). Next we show that the power of this test converges to $1 - \beta^{\mathrm{oracle}}_{i,\mathbf{X}}(\alpha; S, \mu)$ as $M \to \infty$. Hence the claim is proved by taking $M \geq M(\xi)$ for some $M(\xi)$ large enough.

Let $X_S = U\Delta V^{\mathsf{T}}$ be a singular value decomposition of X_S. Therefore, columns of U form a basis for the linear subspace spanned by $\{\widetilde{x}_i\}_{i\in S}$. Let $\tilde{U}$ be such that its columns form a basis for the orthogonal subspace $\{\widetilde{x}_i\}_{i\notin S}$. Then,

$$\frac{\mu\widetilde{x}_i^{\mathsf{T}}(M\mathbf{X}_S\mathbf{X}_S^{\mathsf{T}} + \sigma^2\mathrm{I})^{-1}\widetilde{x}_i}{\|(M\mathbf{X}_S\mathbf{X}_S^{\mathsf{T}} + \sigma^2\mathrm{I})^{-1/2}\widetilde{x}_i\|} = \frac{\mu\widetilde{x}_i^{\mathsf{T}}\{U(M\Delta^2 + \sigma^2\mathrm{I})^{-1}U^{\mathsf{T}} + \sigma^{-2}\tilde{U}\tilde{U}^{\mathsf{T}}\}\widetilde{x}_i}{\|\{U(M\Delta^2 + \sigma^2\mathrm{I})^{-1/2}U^{\mathsf{T}} + \sigma^{-1}\tilde{U}\tilde{U}^{\mathsf{T}}\}\widetilde{x}_i\|}\,.$$

Clearly, as $M \to \infty$, the right hand side of the above equation converges to $(\mu/\sigma)\|\tilde{U}\widetilde{x}_i\| = (\mu/\sigma)\|\mathrm{P}_S^{\perp}\widetilde{x}_i\|$, and thus the power converges to $1 - \beta^{\mathrm{oracle}}_{i,\mathbf{X}}(\alpha; S, \mu) = G(\alpha, \mu\sigma^{-1}\|\mathrm{P}_S^{\perp}\widetilde{x}_i\|)$.

6.3 Proof of Theorem 2.6

Let $u_{\mathbf{X}} \equiv \mu \|\mathrm{P}_S^{\perp} \widetilde{x}_i\|_2/\sigma$. By Lemma 2.4 and 2.5, we have,

$$1 - \beta_i^{\mathrm{opt}}(\alpha;\mu) \le \sup\Big\{\mathbb{E}G(\alpha_{\mathbf{X}}, u_{\mathbf{X}}) \,:\, \mathbb{E}(\alpha_{\mathbf{X}}) \le \alpha\Big\}\,,$$

with the sup taken over measurable functions $\mathbf{X} \mapsto \alpha_{\mathbf{X}}$, and $G(\alpha, u)$ defined as per Eq. (12).

It is easy to check that $\alpha \mapsto G(\alpha, u)$ is concave for any $u \in \mathbb{R}_+$ and $u \mapsto G(\alpha, u)$ is non-decreasing for any $\alpha \in [0,1]$ (see Fig. 1). Further G takes values in $[0,1]$. Hence

$$\begin{aligned}
\mathbb{E}G(\alpha_{\mathbf{X}}, u_{\mathbf{X}}) &\le \mathbb{E}\big\{G(\alpha_{\mathbf{X}}, u_{\mathbf{X}})\mathbb{I}(u \le u_0)\big\} + \mathbb{P}(u_{\mathbf{X}} > u_0) \\
&\le \mathbb{E}\big\{G(\alpha_{\mathbf{X}}, u_0)\big\} + \mathbb{P}(u_{\mathbf{X}} > u_0) \\
&\le G(\mathbb{E}(\alpha_{\mathbf{X}}), u_0) + \mathbb{P}(u_{\mathbf{X}} > u_0) \\
&\le G(\alpha, u_0) + \mathbb{P}(u_{\mathbf{X}} > u_0)
\end{aligned}$$

Since $\widetilde{x}_i$ and $\mathbf{X}_S$ are jointly Gaussian, we have

$$\widetilde{x}_i = \Sigma_{i,S}\Sigma_{S,S}^{-1}\mathbf{X}_S + \Sigma_{i|S}^{1/2} z_i\,,$$

with $z_i \sim \mathsf{N}(0, \mathrm{I}_{n\times n})$ independent of $\mathbf{X}_S$. It follows that

$$u_{\mathbf{X}} = (\mu\Sigma_{i|S}^{1/2}/\sigma)\,\|\mathrm{P}_S^{\perp} z_i\|_2 \stackrel{\mathrm{d}}{=} (\mu\Sigma_{i|S}^{1/2}/\sigma)\, Z_{n-s_0+1}^{1/2}\,,$$

with Z_{n-s_0+1} a chi-squared random variable with $n - s_0 + 1$ degrees of freedom. The desired claim follows by taking $u_0 = (\mu/\sigma)\sqrt{\Sigma_{i|S}(n - s_0 + \ell)}$.

6.4 Proof of Theorem 3.4

Since $\{(\Sigma(p) = \mathrm{I}_{p\times p}, \theta_0(p), n(p), \sigma(p))\}_{p\in\mathbb{N}}$ has a standard distributional limit, the empirical distribution of $\{(\theta_{0,i}, \widehat{\theta}_i^u)\}_{i=1}^p$ converges weakly to $(\Theta_0, \Theta_0 + \tau Z)$ (with probability one). By the portmanteau theorem, and the fact that $\liminf_{p\to\infty} \sigma(p)/\sqrt{n(p)} = \sigma_0$, we have

$$\mathbb{P}(0 < |\Theta_0| < \mu_0\sigma_0) \le \lim_{p\to\infty} \frac{1}{p}\sum_{i=1}^{p} \mathbb{I}\left(0 < \theta_{0,i} < \mu_0 \frac{\sigma(p)}{\sqrt{n(p)}}\right) = 0. \tag{29}$$

In addition, since $\mu_0\sigma_0/2$ is a continuity point of the distribution of Θ_0, we have

$$\lim_{p\to\infty} \frac{1}{p}\sum_{i=1}^{p} \mathbb{I}(|\theta_{0,i}| \ge \frac{\mu_0\sigma_0}{2}) = \mathbb{P}(|\Theta_0| \ge \frac{\mu_0\sigma_0}{2}). \tag{30}$$

Now, by Eq. (29), $\mathbb{P}(|\Theta_0| \ge \mu_0\sigma_0/2) = \mathbb{P}(\Theta_0 \neq 0)$. Further, $\mathbb{I}(|\theta_{0,i}| \ge \mu_0\sigma_0/2) = \mathbb{I}(\theta_{0,i} \neq 0)$ for $1 \le i \le p$, as $p \to \infty$. Therefore, Eq. (30) yields

$$\lim_{p\to\infty} \frac{1}{p}|S_0(p)| = \lim_{p\to\infty} \frac{1}{p}\sum_{i=1}^{p} \mathbb{I}(\theta_{0,i} \neq 0) = \mathbb{P}(\Theta_0 \neq 0). \tag{31}$$

Hence,

$$\begin{aligned}
\lim_{p\to\infty}\frac{1}{|S_0(p)|}\sum_{i\in S_0(p)}T_{i,\mathbf{X}}(y) &= \lim_{p\to\infty}\frac{1}{|S_0(p)|}\sum_{i\in S_0(p)}\mathbb{I}(P_i\le\alpha)\\
&= \frac{1}{\mathbb{P}(\Theta_0\neq 0)}\lim_{p\to\infty}\frac{1}{p}\sum_{i=1}^{p}\mathbb{I}(P_i\le\alpha, i\in S_0(p))\\
&= \frac{1}{\mathbb{P}(\Theta_0\neq 0)}\lim_{p\to\infty}\frac{1}{p}\sum_{i=1}^{p}\mathbb{I}\Big(\Phi(1-\alpha/2)\le\frac{|\widehat{\theta}_i^u|}{\tau},\,|\theta_{0,i}|\ge\mu_0\frac{\sigma(p)}{\sqrt{n(p)}}\Big)\\
&\ge \frac{1}{\mathbb{P}(\Theta_0\neq 0)}\mathbb{P}\Big(\Phi(1-\alpha/2)\le\big|\frac{\Theta_0}{\tau}+Z\big|,\,|\Theta_0|\ge\mu_0\sigma_0\Big).
\end{aligned}\tag{32}$$

Note that τ depends on the distribution p_{Θ_0}. Since $|S_0(p)|\le\varepsilon p$, using Eq. (31), we have $\mathbb{P}(\Theta_0\neq 0)\le\varepsilon$, i.e, p_{Θ_0} is ε-sparse. Let $\tilde{\tau}$ denote the maximum τ corresponding to densities in the family of ε-sparse densities. As shown in [DMM09], $\tilde{\tau}=\tau_*\sigma_0$, where τ_* is defined by Eqs. (16) and (17). Consequently,

$$\begin{aligned}
\lim_{p\to\infty}\frac{1}{|S_0(p)|}\sum_{i\in S_0(p)}T_{i,\mathbf{X}}(y) &\ge \mathbb{P}\Big(\Phi(1-\alpha/2)\le\big|\frac{\mu_0}{\tau_*}+Z\big|\Big)\\
&= 1-\mathbb{P}\Big(-\Phi(1-\alpha/2)-\frac{\mu_0}{\tau_*}\le Z\le\Phi(1-\alpha/2)-\frac{\mu_0}{\tau_*}\Big)\\
&= 1-\{\Phi(\Phi(1-\alpha/2)-\mu_0/\tau_*)-\Phi(-\Phi(1-\alpha/2)-\mu_0/\tau_*)\}\\
&= G(\alpha,\mu_0/\tau_*)\,.
\end{aligned}\tag{33}$$

Now, we take the expectation of both sides of Eq. (33) with respect to the law of random design $\mathbf{X}$ and random noise w. Changing the order of limit and expectation by applying dominated convergence theorem and using linearity of expectation, we obtain

$$\lim_{p\to\infty}\frac{1}{|S_0(p)|}\sum_{i\in S_0(p)}\mathbb{E}_{\mathbf{X},w}\{T_{i,\mathbf{X}}(y)\}\ge G\Big(\alpha,\frac{\mu_0}{\tau_*}\Big)\,.$$

Since $T_{i,\mathbf{X}}(y)$ takes values in $\{0,1\}$, we have $\mathbb{E}_{\mathbf{X},w}\{T_{i,\mathbf{X}}(y)\}=\mathbb{P}_{\theta_0(p)}(T_{i,\mathbf{X}}(y)=1)$. The result follows by noting that the columns of $\mathbf{X}$ are exchangeable and therefore $\mathbb{P}_{\theta_0(p)}(T_{i,\mathbf{X}}(y)=1)$ does not depend on i.

6.5 Proof of Theorem 3.6

Since the sequence $\{\Sigma(p),\theta_0(p),n(p),\sigma(p)\}_{p\in\mathbb{N}}$ has a standard distributional limit, with probability one the empirical distribution of $\{(\theta_{0,i},\widehat{\theta}_i^u,(\Sigma^{-1})_{i,i})\}_{i=1}^p$ converges weakly to the distribution of $(\Theta_0,\Theta_0+\tau\Upsilon^{1/2}Z,\Upsilon)$. Therefore, with probability one, the empirical distribution of

$$\Big\{\frac{\widehat{\theta}_i^u-\theta_{0,i}}{\tau[(\Sigma^{-1})_{i,i}]^{1/2}}\Big\}_{i=1}^{p}$$

converges weakly to $\mathsf{N}(0,1)$. Hence,

$$\begin{aligned}
\lim_{p\to\infty}\frac{1}{|S_0^c(p)|}\sum_{i\in S_0^c(p)}T_{i,\mathbf{X}}(y) &= \lim_{p\to\infty}\frac{1}{|S_0^c(p)|}\sum_{i\in S_0^c(p)}\mathbb{I}(P_i\le\alpha)\\
&= \frac{1}{\mathbb{P}(\Theta_0=0)}\lim_{p\to\infty}\frac{1}{p}\sum_{i=1}^{p}\mathbb{I}(P_i\le\alpha,\, i\in S_0^c(p))\\
&= \frac{1}{\mathbb{P}(\Theta_0=0)}\lim_{p\to\infty}\frac{1}{p}\sum_{i=1}^{p}\mathbb{I}\Big(\Phi(1-\alpha/2)\le\frac{|\widehat{\theta}_i^u|}{\tau[(\Sigma^{-1})_{i,i}]^{1/2}},\,\theta_{0,i}=0\Big)\\
&= \frac{1}{\mathbb{P}(\Theta_0=0)}\,\mathbb{P}(\Phi(1-\alpha/2)\le|Z|,\Theta_0=0)\\
&= \mathbb{P}(\Phi(1-\alpha/2)\le|Z|)=\alpha.
\end{aligned}\tag{34}$$

Applying the same argument as in the proof of Theorem 3.4, we obtain the following by taking the expectation of both sides of the above equation

$$\lim_{p\to\infty}\frac{1}{|S_0(p)|}\sum_{i\in S_0(p)}\mathbb{P}_{\theta_0(p)}(T_{i,\mathbf{X}}(y)=1)=\alpha\,.$$

In particular, for the standard Gaussian design (cf. Theorem 3.3), since the columns of $\mathbf{X}$ are exchangeable we get $\lim_{p\to\infty}\mathbb{P}_{\theta_0(p)}(T_{i,\mathbf{X}}(y)=1)=\alpha$ for all $i\in S_0(p)$.

6.6 Proof of Theorem 3.7

Proof of Theorem 3.7 proceeds along the same lines as the proof of Theorem 3.4. Since $\{(\Sigma(p),\theta_0(p),n(p),\sigma(p))\}_{p\in\mathbb{N}}$ has a standard distributional limit, with probability one the empirical distribution of $\{(\theta_{0,i},\widehat{\theta}_i^u,(\Sigma^{-1})_{i,i})\}_{i=1}^p$ converges weakly to the distribution of $(\Theta_0,\Theta_0+\tau\Upsilon^{1/2}Z,\Upsilon)$. Similar to Eq. (31), we have

$$\lim_{p\to\infty}\frac{1}{p}|S_0(p)|=\mathbb{P}(\Theta_0\neq0).\tag{35}$$

Also

$$\begin{aligned}
\lim_{p\to\infty}\frac{1}{|S_0(p)|}\sum_{i\in S_0(p)}T_{i,\mathbf{X}}(y) &= \lim_{p\to\infty}\frac{1}{|S_0(p)|}\sum_{i\in S_0(p)}\mathbb{I}(P_i\le\alpha)\\
&= \frac{1}{\mathbb{P}(\Theta_0\neq0)}\lim_{p\to\infty}\frac{1}{p}\sum_{i=1}^{p}\mathbb{I}(P_i\le\alpha,i\in S_0(p))\\
&= \frac{1}{\mathbb{P}(\Theta_0\neq0)}\lim_{p\to\infty}\frac{1}{p}\sum_{i=1}^{p}\mathbb{I}\Big(\Phi(1-\alpha/2)\le\frac{|\widehat{\theta}_i^u|}{\tau[(\Sigma^{-1})_{i,i}]^{1/2}},\,\frac{|\theta_{0,i}|}{[(\Sigma^{-1})_{i,i}]^{1/2}}\ge\mu_0\Big)\\
&= \frac{1}{\mathbb{P}(\Theta_0\neq0)}\mathbb{P}\Big(\Phi(1-\alpha/2)\le\big|\frac{\Theta_0}{\tau\Upsilon^{1/2}}+Z\big|,\,\frac{|\Theta_0|}{\Upsilon^{1/2}}\ge\mu_0\Big)\\
&\ge \frac{1}{\mathbb{P}(\Theta_0\neq0)}\mathbb{P}\Big(\Phi(1-\alpha/2)\le\big|\frac{\mu_0}{\tau}+Z\big|\Big)\\
&= 1-\{\Phi(\Phi(1-\alpha/2)-\mu_0/\tau)-\Phi(-\Phi(1-\alpha/2)-\mu_0/\tau)\}\\
&= G(\alpha,\mu_0/\tau)\,.
\end{aligned}$$

Similar to the proof of Theorem 3.4, by taking the expectation of both sides of the above inequality we get

$$\lim_{p\to\infty} \frac{1}{|S_0(p)|} \sum_{i\in S_0(p)} \mathbb{P}_{\theta_0}(T_{i,\mathbf{X}}(y) = 1) \geq G\Big(\alpha, \frac{\mu_0}{\tau}\Big)\,.$$

Acknowledgements

This work was partially supported by the NSF CAREER award CCF-0743978, and the grants AFOSR FA9550-10-1-0360 and AFOSR/DARPA FA9550-12-1-0411.

A Statistical power of earlier approaches

In this appendix, we briefly compare our results with those of Zhang and Zhang [ZZ11], and Bühlmann [Büh12]. Both of these papers consider deterministic designs under restricted eigenvalue conditions. As a consequence, controlling both type I and type II errors requires a significantly larger value of μ/σ.

Following the treatment of [ZZ11], a necessary condition for rejecting $H_{0,j}$ with non-negligible probability is

$$|\theta_{0,j}| \geq c\tau_j\sigma(1+\epsilon'_n),$$

which follows immediately from [ZZ11, Eq. (23)]. Further τ_j and ε'_n are lower bounded in [ZZ11] as follows

$$\tau_j \geq \frac{1}{\|\widetilde{x}_j\|_2}\,,$$
$$\varepsilon'_n \geq C\eta^* s_0\sqrt{\frac{\log p}{n}}\,,$$

where for a standard Gaussian design $\eta^* \geq \sqrt{\log p}$. Using further $\|\widetilde{x}_j\|_2 \leq 2\sqrt{n}$ which again holds with high probability for standard Gaussian designs, we get the necessary condition

$$|\theta_{0,j}| \geq c' \max\Big\{\frac{\sigma s_0 \log p}{n}, \frac{\sigma}{\sqrt{n}}\Big\}\,,$$

for some constant c'.

In [Büh12], p-values are defined, in the notation of the present paper, as

$$P_j \equiv 2\Big\{1 - \Phi\big((a_{n,p;j}(\sigma)|\widehat{\theta}_{j,\mathrm{corr}}| - \Delta_j)_+\big)\Big\}\,,$$

with $\widehat{\theta}_{j,\mathrm{corr}}$ a 'corrected' estimate of $\theta_{0,j}$, cf. [Büh12, Eq. (2.14)]. The corrected estimate $\widehat{\theta}_{j,\mathrm{corr}}$ is defined by the following motivation. The ridge estimator bias, in general, can be decomposed into two terms. The first term is the estimation bias governed by the regularization, and the second term is the additional projection bias $\mathrm{P}_{\mathbf{X}}\theta_0 - \theta_0$, where $\mathrm{P}_{\mathbf{X}}$ denotes the orthogonal projector on the row space of $\mathbf{X}$. The corrected estimate $\widehat{\theta}_{j,\mathrm{corr}}$ is defined in such a way to remove the second bias term under the null hypothesis $H_{0,j}$. Therefore, neglecting the first bias term, we have $\widehat{\theta}_{j,\mathrm{corr}} = (\mathrm{P}_{\mathbf{X}})_{jj}\theta_{0,j}$.

Assuming the corrected estimate to be consistent (which it is in ℓ_1 sense under the assumption of the paper), rejecting $H_{0,j}$ with non-negligible probability requires

$$|\theta_{0,j}| \geq \frac{c}{a_{n,p;j}(\sigma)|(\mathrm{P}_{\mathbf{X}})_{jj}|} \max\{\Delta_j, 1\}\,,$$

Following [Büh12, Eq. (2.13)], and keeping the dependence on s_0 instead of assuming $s_0 = o((n/\log p)^{\xi})$, we have

$$\frac{\Delta_j}{a_{n,p;j}(\sigma)|(\mathrm{P}_{\mathbf{X}})_{jj}|} = C \max_{k\in[p]\setminus j} \frac{|(\mathrm{P}_{\mathbf{X}})_{jk}|}{|(\mathrm{P}_{\mathbf{X}})_{jj}|}\, \sigma s_0 \sqrt{\frac{\log p}{n}}\,.$$

Further, plugging for $a_{n,p;j}$ we have

$$\frac{1}{a_{n,p;j}(\sigma)|(\mathrm{P}_{\mathbf{X}})_{jj}|} = \frac{\sigma\sqrt{\Omega_{jj}}}{\sqrt{n}\,|(\mathrm{P}_{\mathbf{X}})_{jj}|}\,.$$

For a standard Gaussian design $(p/n)(\mathrm{P}_{\mathbf{X}})_{jk}$ is approximately distributed as u_1, where $u = (u_1, u_2, \dots, u_n) \in \mathbb{R}^n$ is a uniformly random vector with $\|u\| = 1$. In particular u_1 is approximately $\mathsf{N}(0, 1/n)$. A standard calculation yields $\max_{k\in[p]\setminus j} |(\mathrm{P}_{\mathbf{X}})_{jk}| \geq \sqrt{n\log p}/p$ with high probability. Furthermore, $|(\mathrm{P}_{\mathbf{X}})_{jj}|$ concentrates around n/p. Finally, by definition of Ω_{jj} (cf. [Büh12, Eq. (2.3)]) and using classical large deviation results about the singular values of a Gaussian matrix, we have $\Omega_{jj} \geq (n/p)^2$ with high probability. Hence, a necessary condition for rejecting $H_{0,j}$ with non-negligible probability is

$$|\theta_{0,j}| \geq C \max\Big\{\frac{\sigma s_0 \log p}{n}, \frac{\sigma}{\sqrt{n}}\Big\},$$

as stated in Section 1.

B Statistical physics calculation

In this section we outline the replica calculation leading to the Claim 4.1. We limit ourselves to the main steps, since analogous calculations can be found in several earlier works [Tan02, GV05, TK10]. For a general introduction to the method and its motivation we refer to [MPV87, MM09]. Also, for the sake of simplicity, we shall focus on characterizing the asymptotic distribution of $\widehat{\theta}^u$, cf. Eq. (19). The distribution of r is derived by the same approach.

Fix a sequence of instances $\{(\Sigma(p), \theta_0(p), n(p), \sigma(p))\}_{p\in\mathbb{N}}$. For the sake of simplicity, we assume $\sigma(p)^2 = n(p)\sigma_0^2$ and $n(p) = p\delta$ (the slightly more general case $\sigma(p)^2 = n(p)[\sigma_0^2 + o(1)]$ and $n(p) = p[\delta + o(1)]$ does not require any change to the derivation given here, but is more cumbersome notationally). Fix $\widetilde{g} : \mathbb{R}\times\mathbb{R}\times\mathbb{R} \to \mathbb{R}$ a continuous function convex in its first argument, and let $g(u, y, z) \equiv \max_{x\in\mathbb{R}}[ux - \widetilde{g}(x, y, z)]$ be its Lagrange dual. The replica calculation aims at estimating the following moment generating function (*partition function*)

$$\mathcal{Z}_p(\beta, s) \equiv \int \exp\Big\{ -\frac{\beta}{2n}\|y - \mathbf{X}\theta\|_2^2 - \beta J(\theta) - \beta s \sum_{i=1}^{p} [g(u_i, \theta_{0,i}, (\Sigma^{-1})_{ii}) - u_i\theta_i^u] \\ - \frac{\beta}{2n}(s\tilde{\mathsf{d}})^2 \|\mathbf{X}\Sigma^{-1}u\|_2^2\Big\} \mathrm{d}\theta\, \mathrm{d}u\,. \tag{36}$$

Here (y_i, x_i) are i.i.d. pairs distributed as per model (1) and $\theta^u = \theta + (\tilde{\mathsf{d}}/n)\,\Sigma^{-1}\mathbf{X}^{\mathsf{T}}(y - \mathbf{X}\theta)$ with $\tilde{\mathsf{d}} \in \mathbb{R}$ to be defined below. Further, $g : \mathbb{R} \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is a continuous function strictly convex in its first argument. Finally, $s \in \mathbb{R}_+$ and $\beta > 0$ is a 'temperature' parameter not to be confused with the type II error rate as used in the main text. We will eventually show that the appropriate choice of $\tilde{\mathsf{d}}$ is given by Eq. (24).

Within the replica method, it is assumed that the limits $p \to \infty$, $\beta \to \infty$ exist almost surely for the quantity $(p\beta)^{-1} \log \mathcal{Z}_p(\beta, s)$, and that the order of the limits can be exchanged. We therefore define

$$\mathfrak{F}(s) \equiv -\lim_{\beta\to\infty}\lim_{p\to\infty} \frac{1}{p\beta}\,\log \mathcal{Z}_p(\beta, s) \tag{37}$$

$$\equiv -\lim_{p\to\infty}\lim_{\beta\to\infty} \frac{1}{p\beta}\,\log \mathcal{Z}_p(\beta, s)\,. \tag{38}$$

In other words $\mathfrak{F}(s)$ is the exponential growth rate of $\mathcal{Z}_p(\beta, s)$. It is also assumed that $p^{-1}\log \mathcal{Z}_p(\beta, s)$ concentrates tightly around its expectation so that $\mathfrak{F}(s)$ can in fact be evaluated by computing

$$\mathfrak{F}(s) = -\lim_{\beta\to\infty}\lim_{p\to\infty} \frac{1}{p\beta}\,\mathbb{E}\log \mathcal{Z}_p(\beta, s)\,, \tag{39}$$

where expectation is being taken with respect to the distribution of (y_1, x_1), ..., (y_n, x_n). Notice that, by Eq. (38) and using Laplace method in the integral (36), we have

$$\mathfrak{F}(s) = \lim_{p\to\infty} \frac{1}{p} \min_{\theta, u\in\mathbb{R}^p} \Big\{ \frac{1}{2n}\|y - \mathbf{X}\theta\|_2^2 + J(\theta) + s\sum_{i=1}^{p}[g(u_i, \theta_{0,i}, (\Sigma^{-1})_{ii}) - u_i\theta_i^u] + \frac{1}{2n}(s\tilde{\mathsf{d}})^2\|\mathbf{X}\Sigma^{-1}u\|_2^2 \Big\}$$

Finally we assume that the derivative of $\mathfrak{F}(s)$ as $s \to 0$ can be obtained by differentiating inside the limit. This condition holds, for instance, if the cost function is strongly convex at $s = 0$. We get

$$\frac{\mathrm{d}\mathfrak{F}}{\mathrm{d}s}(s=0) \;= \lim_{p\to\infty} \tfrac{1}{p}\textstyle\sum_{i=1}^{p} \min_{u_i\in\mathbb{R}}[g(u_i, \theta_{0,i}, (\Sigma^{-1})_{ii}) - u_i\widehat{\theta}_i^u] \tag{40}$$

were $\widehat{\theta}^u = \widehat{\theta} + (\tilde{\mathsf{d}}/n)\,\Sigma^{-1}\mathbf{X}^{\mathsf{T}}(y - \mathbf{X}\widehat{\theta})$ and $\widehat{\theta}$ is the minimizer of the regularized least squares as in Section 3. Since, by duality $\widetilde{g}(x, y, z) \equiv \max_{u\in\mathbb{R}}[ux - g(u, y, z)]$, we get

$$\frac{\mathrm{d}\mathfrak{F}}{\mathrm{d}s}(s=0) \;= -\lim_{p\to\infty} \tfrac{1}{p}\textstyle\sum_{i=1}^{p} \widetilde{g}(\widehat{\theta}_i^u, \theta_{0,i}, (\Sigma^{-1})_{ii})\,. \tag{41}$$

Hence, by computing $\mathfrak{F}(s)$ using Eq. (39) for a complete set of functions $\widetilde{g}$, we get access to the corresponding limit quantities (41) and hence, via standard weak convergence arguments, to the joint empirical distribution of the triple $(\widehat{\theta}_i^u, \theta_{0,i}, (\Sigma^{-1})_{ii})$, cf. Eq. (20).

In order to carry out the calculation of $\mathfrak{F}(s)$, we begin by rewriting the partition function (36) in a more convenient form. Using the definition of θ^u and after a simple manipulation

$$\mathcal{Z}_p(\beta, s) = \int \exp\Big\{ -\frac{\beta}{2n}\|y - \mathbf{X}(\theta + s\tilde{\mathsf{d}}\Sigma^{-1}u)\|_2^2 - \beta J(\theta) + \beta s\langle u, \theta\rangle - \beta\, s\sum_{i=1}^{p} g(u_i, \theta_{0,i}, (\Sigma^{-1})_{ii})\Big\}\mathrm{d}\theta\,\mathrm{d}u\,.$$

Define the measure $\nu(\mathrm{d}\theta)$ over $\theta \in \mathbb{R}^p$ as follows

$$\nu(\mathrm{d}\theta) = \int \exp\Big\{ -\beta J(\theta - s\tilde{\mathsf{d}}\Sigma^{-1}u) + \beta s\langle \theta - s\tilde{\mathsf{d}}\Sigma^{-1}u, u\rangle - \beta s \sum_{i=1}^{p} g(u_i, \theta_{0,i}, (\Sigma^{-1})_{ii})\Big\}\, \mathrm{d}u\,. \tag{42}$$

Using this definition and with the change of variable $\theta' = \theta + s\tilde{\mathsf{d}}\Sigma^{-1}u$, we can rewrite Eq. (42) as

$$\begin{aligned}
\mathcal{Z}_p(\beta, s) &\equiv \int \exp\Big\{ -\frac{\beta}{2n}\|y - \mathbf{X}\theta\|_2^2\Big\}\, \nu(\mathrm{d}\theta) \\
&= \int \exp\Big\{ i\sqrt{\frac{\beta}{n}}\langle z, y - \mathbf{X}\theta\rangle\Big\}\, \nu(\mathrm{d}\theta)\, \gamma_n(\mathrm{d}z) \\
&= \int \exp\Big\{ i\sqrt{\frac{\beta}{n}}\langle w, z\rangle + i\sqrt{\frac{\beta}{n}}\langle z, \mathbf{X}(\theta_0 - \theta)\rangle\Big\}\, \nu(\mathrm{d}\theta)\, \gamma(\mathrm{d}z)\,,
\end{aligned} \tag{43}$$

where $\gamma_n(\mathrm{d}z)$ denotes the standard Gaussian measure on $\mathbb{R}^n$: $\gamma_n(\mathrm{d}z) \equiv (2\pi)^{-n/2}\exp(-\|z\|_2^2/2)\,\mathrm{d}z$.

The replica method aims at computing the expected log-partition function, cf. Eq. (39) using the identity

$$\mathbb{E}\log \mathcal{Z}_p(\beta, s) = \frac{\mathrm{d}}{\mathrm{d}k}\Big|_{k=0} \log \mathbb{E}\big\{\mathcal{Z}_p(\beta, s)^k\big\}\,. \tag{44}$$

This formula would require computing fractional moments of $\mathcal{Z}_p$ as $k \to 0$. The replica method consists in a prescription that allows to compute a formal expression for the k integer, and then extrapolate it as $k \to 0$. Crucially, the limit $k \to 0$ is inverted with the one $p \to \infty$:

$$\lim_{p\to\infty} \frac{1}{p}\mathbb{E}\log \mathcal{Z}_p(\beta, s) = \frac{\mathrm{d}}{\mathrm{d}k}\Big|_{k=0} \lim_{p\to\infty} \frac{1}{p}\log \mathbb{E}\big\{\mathcal{Z}_p(\beta, s)^k\big\}\,. \tag{45}$$

In order to represent $\mathcal{Z}_p(\beta, s)^k$, we use the identity

$$\Big(\int f(x), \rho(\mathrm{d}x)\Big)^k = \int f(x^1)f(x^2)\cdots f(x^k)\, \rho(\mathrm{d}x^1)\cdots\rho(\mathrm{d}x^k)\,. \tag{46}$$

In order to apply this formula to Eq. (43), we let, with a slight abuse of notation, $\nu^k(\mathrm{d}\theta) \equiv \nu(\mathrm{d}\theta^1)\times \nu(\mathrm{d}\theta^2)\times\cdots\times\nu(\mathrm{d}\theta^k)$ be a measure over $(\mathbb{R}^p)^k$, with $\theta^1, \dots, \theta^k \in \mathbb{R}^p$. Analogously $\gamma_n^k(\mathrm{d}z) \equiv \gamma_n(\mathrm{d}z^1)\times \gamma_n(\mathrm{d}z^2)\times\cdots\times\gamma_n(\mathrm{d}z^k)$, with $z^1, \dots, z^k \in \mathbb{R}^n$. With these notations, we have

$$\mathbb{E}\{\mathcal{Z}_p(\beta, s)^k\} = \int \mathbb{E}\exp\Big\{ i\sqrt{\frac{\beta}{n}}\langle w, \sum_{a=1}^{k} z^a\rangle + i\sqrt{\frac{\beta}{n}}\langle \mathbf{X}, \sum_{a=1}^{k} z^a(\theta_0 - \theta^a)^{\mathsf{T}}\rangle\Big\}\, \nu^k(\mathrm{d}\theta)\, \gamma_n^k(\mathrm{d}z)\,. \tag{47}$$

In the above expression $\mathbb{E}$ denotes expectation with respect to the noise vector w, and the design matrix $\mathbf{X}$. Further, we used $\langle\,\cdot\,,\,\cdot\,\rangle$ to denote matrix scalar product as well: $\langle A, B\rangle \equiv \mathrm{Trace}(A^{\mathsf{T}}B)$.

At this point we can take the expectation with respect to w, $\mathbf{X}$. We use the fact that, for any $M \in \mathbb{R}^{n\times p}$, $u \in \mathbb{R}^n$

$$\begin{aligned}
\mathbb{E}\big\{\exp\big(i\langle w, u\rangle\big)\big\} &= \exp\Big\{ -\frac{1}{2}n\sigma_0^2\,\|u\|_2^2\Big\}\,, \\
\mathbb{E}\big\{\exp\big(i\langle M, \mathbf{X}\rangle\big)\big\} &= \exp\Big\{ -\frac{1}{2}\,\langle M, M\Sigma\rangle\Big\}\,,
\end{aligned}$$

Using these identities in Eq. (47), we obtain

$$\mathbb{E}\{\mathcal{Z}_p^k\} = \int \exp\Big\{ -\frac{1}{2}\,\beta\sigma_0^2 \sum_{a=1}^k \|z^a\|_2^2 - \frac{\beta}{2n} \sum_{a,b=1}^k \langle z^a, z^b\rangle\, \langle(\theta^a-\theta_0), \Sigma(\theta^b-\theta_0)\rangle \Big\}\, \nu^k(\mathrm{d}\theta)\, \gamma_n^k(\mathrm{d}z)\,. \tag{48}$$

We next use the identity

$$e^{-xy} = \frac{1}{2\pi i} \int_{(-i\infty, i\infty)} \int_{(-\infty,\infty)} e^{-\zeta q + \zeta x - q y}\, \mathrm{d}\zeta\, \mathrm{d}q\,,$$

where the integral is over $\zeta \in (-i\infty, i\infty)$ (imaginary axis) and $q \in (-\infty, \infty)$. We apply this identity to Eq. (48), and introduce integration variables $Q \equiv (Q_{ab})_{1\le a,b\le k}$ and $\Lambda \equiv (\Lambda_{ab})_{1\le a,b\le k}$. Letting $\mathrm{d}Q \equiv \prod_{a,b} \mathrm{d}Q_{ab}$ and $\mathrm{d}\Lambda \equiv \prod_{a,b} \mathrm{d}\Lambda_{ab}$

$$\mathbb{E}\{\mathcal{Z}_p^k\} = \Big(\frac{\beta n}{4\pi i}\Big)^{k^2} \int \exp\Big\{ -p\mathcal{S}_k(Q,\Lambda)\Big\}\, \mathrm{d}Q\, \mathrm{d}\Lambda\,, \tag{49}$$

$$\mathcal{S}_k(Q,\Lambda) = \frac{\beta\delta}{2} \sum_{a,b=1}^k \Lambda_{ab} Q_{ab} - \frac{1}{p}\, \log \xi(\Lambda) - \delta\, \log \widehat{\xi}(Q)\,, \tag{50}$$

$$\xi(\Lambda) \equiv \int \exp\Big\{ \frac{\beta}{2} \sum_{a,b=1}^k \Lambda_{ab} \langle(\theta^a-\theta_0), \Sigma(\theta^b-\theta_0)\rangle \Big\}\, \nu^k(\mathrm{d}\theta)\,, \tag{51}$$

$$\widehat{\xi}(Q) \equiv \int \exp\Big\{ -\frac{\beta}{2} \sum_{a,b=1}^k (\sigma_0^2 \mathrm{I} + Q)_{a,b}\, z_1^a\, z_1^b \Big\}\, \gamma_1^k(\mathrm{d}z_1)\,. \tag{52}$$

Notice that above we used the fact that, after introducing Q, Λ, the integral over $(z^1, \dots, z^k) \in (\mathbb{R}^n)^k$ factors into n integrals over $(\mathbb{R})^k$ with measure $\gamma_1^k(\mathrm{d}z_1)$.

We next use the saddle point method in Eq. (49) to obtain

$$-\lim_{p\to\infty} \frac{1}{p} \log \mathbb{E}\{\mathcal{Z}_p^k\} \quad = \mathcal{S}_k(Q^*, \Lambda^*)\,, \tag{53}$$

where Q^*, Λ^* is the saddle-point location. The replica method provides a hierarchy of ansatz for this saddle-point. The first level of this hierarchy is the so-called *replica symmetric* ansatz postulating that Q^*, Λ^* ought to be invariant under permutations of the row/column indices. This is motivated by the fact that $\mathcal{S}_k(Q,\Lambda)$ is indeed left unchanged by such change of variables. This is equivalent to postulating that

$$Q_{ab}^* = \begin{cases} q_1 & \text{if } a = b, \\ q_0 & \text{otherwise,} \end{cases}\,, \qquad \Lambda_{ab}^* = \begin{cases} \beta\zeta_1 & \text{if } a = b, \\ \beta\zeta_0 & \text{otherwise,} \end{cases} \tag{54}$$

where the factor β is for future convenience. Given that the partition function, cf. Eq. (36) is the integral of a log-concave function, it is expected that the replica-symmetric ansatz yields in fact the correct result [MPV87, MM09].

The next step consists in substituting the above expressions for Q^*, Λ^* in $\mathcal{S}_k(\,\cdot\,,\,\cdot\,)$ and then taking the limit $k \to 0$. We will consider separately each term of $\mathcal{S}_k(Q,\Lambda)$, cf. Eq. (50).

Let us begin with the first term

$$\sum_{a,b=1}^{k} \Lambda^*_{ab} Q^*_{ab} = k\,\beta\zeta_1 q_1 + k(k-1)\beta\zeta_0 q_0\,. \tag{55}$$

Hence

$$\lim_{k\to\infty} \frac{\beta\delta}{2k} \sum_{a,b=1}^{k} \Lambda^*_{ab} Q^*_{ab} = \frac{\beta^2\delta}{2}\,(\zeta_1 q_1 - \zeta_0 q_0)\,. \tag{56}$$

Let us consider $\widehat{\xi}(Q^*)$. We have

$$\log\widehat{\xi}(Q^*) = -\frac{1}{2}\log\mathrm{Det}(\mathrm{I} + \beta\sigma^2\mathrm{I} + \beta Q^*) \tag{57}$$

$$= -\frac{k-1}{2}\log\big(1+\beta(q_1-q_0)\big) - \frac{1}{2}\log\big(1+\beta(q_1-q_0)+\beta k(\sigma^2+q_0)\big)\,. \tag{58}$$

In the limit $k \to 0$ we thus obtain

$$\lim_{k\to 0}\frac{1}{k}(-\delta)\log\widehat{\xi}(Q^*) = \frac{\delta}{2}\,\log\big(1+\beta(q_1-q_0)\big) + \frac{\delta}{2}\,\frac{\beta(\sigma^2+q_0)}{1+\beta(q_1-q_0)}\,. \tag{59}$$

Finally, introducing the notation $\|v\|^2_\Sigma \equiv \langle v, \Sigma v\rangle$, we have

$$\xi(\Lambda^*) \equiv \int \exp\Big\{\frac{\beta^2}{2}(\zeta_1-\zeta_0)\sum_{a=1}^{k}\|\theta^a-\theta_0\|^2_\Sigma + \frac{\beta^2\zeta_0}{2}\sum_{a,b=1}^{k}\langle(\theta^a-\theta_0),\Sigma(\theta^b-\theta_0)\rangle\Big\}\,\nu^k(\mathrm{d}\theta)\,,$$

$$= \mathbb{E}\int \exp\Big\{\frac{\beta^2}{2}(\zeta_1-\zeta_0)\sum_{a=1}^{k}\|\theta^a-\theta_0\|^2_\Sigma + \beta\sqrt{\zeta_0}\sum_{a=1}^{k}\langle Z,\Sigma^{1/2}(\theta^a-\theta_0)\rangle\Big\}\,\nu^k(\mathrm{d}\theta)\,,$$

where expectation is with respect to $Z \sim \mathsf{N}(0,\mathrm{I}_{p\times p})$. Notice that, given $Z \in \mathbb{R}^p$, the integrals over $\theta^1,\theta^2,\dots,\theta^k$ factorize, whence

$$\xi(\Lambda^*) = \mathbb{E}\left\{\left[\int \exp\Big\{\frac{\beta^2}{2}(\zeta_1-\zeta_0)\|\theta-\theta_0\|^2_\Sigma + \beta\sqrt{\zeta_0}\langle Z,\Sigma^{1/2}(\theta-\theta_0)\rangle\Big\}\,\nu(\mathrm{d}\theta)\right]^k\right\}. \tag{60}$$

Therefore

$$\lim_{k\to 0}\frac{(-1)}{pk}\log\xi(\Lambda^*) =$$

$$-\frac{1}{p}\mathbb{E}\left\{\log\left[\int \exp\Big\{\frac{\beta^2}{2}(\zeta_1-\zeta_0)\|\theta-\theta_0\|^2_\Sigma + \beta\sqrt{\zeta_0}\langle Z,\Sigma^{1/2}(\theta-\theta_0)\rangle\Big\}\,\nu(\mathrm{d}\theta)\right]\right\}. \tag{61}$$

Putting Eqs. (56), (59), and (61) together we obtain

$$
\begin{aligned}
-\lim_{p\to\infty}\frac{1}{p\beta}\mathbb{E}\log\mathcal{Z}_p &= \lim_{k\to 0}\frac{1}{k\beta}\mathcal{S}_k(Q^*,\Lambda^*) \qquad (62)\\
&= \frac{\beta\delta}{2}\left(\zeta_1 q_1-\zeta_0 q_0\right)+\frac{\delta}{2\beta}\,\log\big(1+\beta(q_1-q_0)\big)+\frac{\delta}{2}\,\frac{\sigma^2+q_0}{1+\beta(q_1-q_0)}\\
&\quad -\lim_{p\to\infty}\frac{1}{p\beta}\mathbb{E}\left\{\log\left[\int\exp\left\{\frac{\beta^2}{2}(\zeta_1-\zeta_0)\|\theta-\theta_0\|_\Sigma^2\right.\right.\right.\\
&\qquad\qquad \left.\left.\left.+\beta\sqrt{\zeta_0}\langle Z,\Sigma^{1/2}(\theta-\theta_0)\rangle\right\}\nu(\mathrm{d}\theta)\right]\right\}.
\end{aligned}
$$

We can next take the limit $\beta\to\infty$. In doing this, one has to be careful with respect to the behavior of the saddle point parameters q_0, q_1, ζ_0, ζ_1. A careful analysis (omitted here) shows that q_0, q_1 have the same limit, denoted here by q_0, and ζ_0, ζ_1 have the same limit, denoted by ζ_0. Moreover $q_1-q_0=(q/\beta)+o(\beta^{-1})$ and $\zeta_1-\zeta_0=(-\zeta/\beta)+o(\beta^{-1})$. Substituting in the above expression, and using Eq. (39), we get

$$
\begin{aligned}
\mathfrak{F}(s) =& \frac{\delta}{2}(\zeta_0 q-\zeta q_0)+\frac{\delta}{2}\,\frac{q_0+\sigma^2}{1+q} \qquad (63)\\
&+\lim_{p\to\infty}\frac{1}{p}\mathbb{E}\min_{\theta\in\mathbb{R}^p}\left\{\frac{\zeta}{2}\|\theta-\theta_0\|_\Sigma^2-\sqrt{\zeta_0}\langle Z,\Sigma^{1/2}(\theta-\theta_0)\rangle+\widetilde{J}(\theta;s)\right\},\\
\widetilde{J}(\theta;s) =& \min_{u\in\mathbb{R}^p}\left\{J(\theta-s\tilde{\mathsf{d}}\Sigma^{-1}u)-s\langle\theta-s\tilde{\mathsf{d}}\Sigma^{-1}u,u\rangle+s\sum_{i=1}^p g(u_i,\theta_{0,i},(\Sigma^{-1})_{ii})\right\}. \qquad (64)
\end{aligned}
$$

After the change of variable $\theta-s\tilde{\mathsf{d}}\Sigma^{-1}u\to\theta$, this reads

$$
\begin{aligned}
\mathfrak{F}(s) &= \frac{\delta}{2}(\zeta_0 q-\zeta q_0)+\frac{\delta}{2}\,\frac{q_0+\sigma_0^2}{1+q}-\frac{\zeta_0}{2\zeta} \qquad (65)\\
&\quad+\lim_{p\to\infty}\frac{1}{p}\mathbb{E}\min_{\theta,u\in\mathbb{R}^p}\left\{\frac{\zeta}{2}\left\|\theta-\theta_0-\frac{\sqrt{\zeta_0}}{\zeta}\Sigma^{-1/2}Z+s\tilde{\mathsf{d}}\Sigma^{-1}u\right\|_\Sigma^2+\widetilde{J}(\theta,u;s)\right\},\\
\widetilde{J}(\theta,u;s) &= J(\theta)-s\langle\theta,u\rangle+s\sum_{i=1}^p g(u_i,\theta_{0,i},(\Sigma^{-1})_{ii}). \qquad (66)
\end{aligned}
$$

Finally, we must set ζ,ζ_0 and q,q_0 to their saddle point values. We start by using the stationarity conditions with respect to q, q_0:

$$
\begin{aligned}
\frac{\partial\mathfrak{F}}{\partial q}(s) &= \frac{\delta}{2}\zeta_0-\frac{\delta}{2}\,\frac{q_0+\sigma_0^2}{(1+q)^2}, \qquad (67)\\
\frac{\partial\mathfrak{F}}{\partial q_0}(s) &= -\frac{\delta}{2}\,\zeta+\frac{\delta}{2}\,\frac{1}{1+q}. \qquad (68)
\end{aligned}
$$

We use these to eliminate q and q_0. Renaming $\zeta_0 = \zeta^2\tau^2$, we get our final expression for $\mathfrak{F}(s)$:

$$\begin{aligned}\mathfrak{F}(s) &= -\frac{1}{2}(1-\delta)\zeta\tau^2 - \frac{\delta}{2}\zeta^2\tau^2 + \frac{\delta}{2}\sigma_0^2\zeta \\ &\quad + \lim_{p\to\infty}\frac{1}{p}\mathbb{E}\min_{\theta,u\in\mathbb{R}^p}\left\{\frac{\zeta}{2}\left\|\theta-\theta_0-\tau\Sigma^{-1/2}Z+s\tilde{\mathsf{d}}\Sigma^{-1}u\right\|_\Sigma^2 + \widetilde{J}(\theta,u;s)\right\}, \qquad (69)\\ \widetilde{J}(\theta,u;s) &= J(\theta) - s\langle\theta,u\rangle + s\sum_{i=1}^p g(u_i,\theta_{0,i},(\Sigma^{-1})_{ii})\,. \qquad (70)\end{aligned}$$

Here it is understood that ζ and τ^2 are to be set to their saddle point values.

We are interested in the derivative of $\mathfrak{F}(s)$ with respect to s, cf. Eq. (41). Consider first the case $s=0$. Using the assumption $\mathfrak{E}^{(p)}(a,b)\to\mathfrak{E}(a,b)$, cf. Eq. (25), we get

$$\mathfrak{F}(s=0) = -\frac{1}{2}(1-\delta)\zeta\tau^2 - \frac{\delta}{2}\zeta^2\tau^2 + \frac{\delta}{2}\sigma_0^2\zeta + \mathfrak{E}(\tau^2,\zeta)\,. \qquad (71)$$

The values of ζ, τ^2 are obtained by setting to zero the partial derivatives

$$\begin{aligned}\frac{\partial\mathfrak{F}}{\partial\zeta}(s=0) &= -\frac{1}{2}(1-\delta)\tau^2 - \delta\zeta\tau^2 + \frac{\delta}{2}\sigma_0^2 + \frac{\partial\mathfrak{E}}{\partial\zeta}(\tau^2,\zeta)\,, \qquad (72)\\ \frac{\partial\mathfrak{F}}{\partial\tau^2}(s=0) &= -\frac{1}{2}(1-\delta)\zeta - \frac{\delta}{2}\,\zeta^2 + \frac{\partial\mathfrak{E}}{\partial\tau^2}(\tau^2,\zeta)\,, \qquad (73)\end{aligned}$$

Define, as in the statement of the Replica Claim

$$\begin{aligned}\mathsf{E}_1(a,b) &\equiv \lim_{p\to\infty}\frac{1}{p}\mathbb{E}\big\{\big\|\eta_b(\theta_0+\sqrt{a}\Sigma^{-1/2}Z)-\theta_0\big\|_\Sigma^2\big\}\,, \qquad (74)\\ \mathsf{E}_2(a,b) &\equiv \lim_{p\to\infty}\frac{1}{p}\mathbb{E}\big\{\operatorname{div}\eta_b(\theta_0+\sqrt{a}\Sigma^{-1/2}Z)\big\} \qquad (75)\\ &= \lim_{p\to\infty}\frac{1}{p\tau}\mathbb{E}\big\{\langle\eta_b(\theta_0+\sqrt{a}\Sigma^{-1/2}Z),\Sigma^{1/2}Z\rangle\big\}\,,\end{aligned}$$

where the last identity follows by integration by parts. These limits exist by the assumption that $\nabla\mathfrak{E}^{(p)}(a,b)\to\nabla\mathfrak{E}(a,b)$. In particular

$$\begin{aligned}\frac{\partial\mathfrak{E}}{\partial\zeta}(\tau^2,\zeta) &= \frac{1}{2}\mathsf{E}_1(\tau^2,\zeta) - \tau^2\,\mathsf{E}_2(\tau^2,\zeta) + \frac{1}{2}\,\tau^2\,, \qquad (76)\\ \frac{\partial\mathfrak{E}}{\partial\tau^2}(\tau^2,\zeta) &= -\frac{\zeta}{2}\mathsf{E}_2(\tau^2,\zeta) + \frac{1}{2}\,\zeta\,. \qquad (77)\end{aligned}$$

Substituting these expressions in Eqs. (72), (73), and simplifying, we conclude that the derivatives vanish if and only if ζ,τ^2 satisfy the following equations

$$\begin{aligned}\tau^2 &= \sigma_0^2 + \frac{1}{\delta}\,\mathsf{E}_1(\tau^2,\zeta)\,, \qquad (78)\\ \zeta &= 1 - \frac{1}{\delta}\,\mathsf{E}_2(\tau^2,\zeta)\,. \qquad (79)\end{aligned}$$

The solution of these equations is expected to be unique for J convex and $\sigma_0^2>0$.

Next consider the derivative of $\mathfrak{F}(s)$ with respect to s, which is our main object of interest, cf. Eq. (41). By differentiating Eq. (69) and inverting the order of derivative and limit, we get

$$\frac{\mathrm{d}\mathfrak{F}}{\mathrm{d}s}(s=0) = \lim_{p\to\infty}\frac{1}{p}\mathbb{E}\min_{u\in\mathbb{R}^p}\Big\{\zeta\tilde{\mathsf{d}}\langle u, \widehat{\theta}-\theta_0-\tau\Sigma^{-1/2}Z\rangle - \langle\widehat{\theta},u\rangle + \sum_{i=1}^{p} g(u_i,\theta_{0,i},(\Sigma^{-1})_{ii})\Big\}\,, \tag{80}$$

where $\widehat{\theta}$ is the minimizer at $s=0$, i.e., $\widehat{\theta}=\eta_\zeta(\theta_0+\tau\Sigma^{-1/2}Z)$, and ζ,τ^2 solve Eqs. (78), (79). At this point we choose $\tilde{\mathsf{d}}=1/\zeta$. Minimizing over u (recall that $\widetilde{g}(x,y,z)=\max_{u\in\mathbb{R}}[ux-g(u,y,z)]$), we get

$$\frac{\mathrm{d}\mathfrak{F}}{\mathrm{d}s}(s=0) = -\lim_{p\to\infty}\frac{1}{p}\mathbb{E}\,\widetilde{g}(\theta_{0,i}+\tau(\Sigma^{-1/2}Z)_i,\theta_{0,i},(\Sigma^{-1})_{ii})\,. \tag{81}$$

Comparing with Eq. (41), this proves the claim that the standard distributional limit does indeed hold.

Notice that τ^2 is given by Eq. (78) that, for $\mathsf{d}=1/\zeta$ does indeed coincide with the claimed Eq. (28). Finally consider the scale parameter $\mathsf{d}=\mathsf{d}(p)$ defined by Eq. (24). We claim that

$$\lim_{p\to\infty}\mathsf{d}(p)=\tilde{\mathsf{d}}=\frac{1}{\zeta}\,. \tag{82}$$

Consider, for the sake of simplicity, the case that J is differentiable and strictly convex (the general case can be obtained as a limit). Then the minimum condition of the proximal operator (26) reads

$$\theta=\eta_b(y) \quad\Leftrightarrow\quad b\Sigma(y-\theta)=\nabla J(\theta)\,. \tag{83}$$

Differentiating with respect to θ, and denoting by $\mathrm{D}\eta_b$ the Jacobian of η_b, we get $\mathrm{D}\eta_b(y)=(\mathrm{I}+b^{-1}\Sigma^{-1}\nabla^2 J(\theta))^{-1}$ and hence

$$\mathsf{E}_2(a,b) = \lim_{p\to\infty}\frac{1}{p}\mathbb{E}\,\mathrm{Trace}\Big\{(1+b^{-1}\Sigma^{-1/2}\nabla^2 J(\widehat{\theta})\Sigma^{-1/2})^{-1}\Big\}\,, \tag{84}$$

$$\widehat{\theta} \equiv \eta_b(\theta_0+\sqrt{a}\,\Sigma^{-1/2}\,Z)\,. \tag{85}$$

Hence, combining Eqs. (79) and (84) implies that $\tilde{\mathsf{d}}=\zeta^{-1}$ satisfies

$$1 = \frac{1}{\tilde{\mathsf{d}}}+\lim_{p\to\infty}\frac{1}{n}\mathbb{E}\,\mathrm{Trace}\Big\{(1+\tilde{\mathsf{d}}\Sigma^{-1/2}\nabla^2 J(\widehat{\theta})\Sigma^{-1/2})^{-1}\Big\}\,, \tag{86}$$

$$\widehat{\theta} \equiv \eta_{1/\tilde{\mathsf{d}}}(\theta_0+\tau\,\Sigma^{-1/2}\,Z)\,. \tag{87}$$

The claim (82) follows by comparing this with Eq. (24), and noting that, by the above $\widehat{\theta}$ is indeed asymptotically distributed as the estimator (23).

C Simulation results

Consider the setup discussed in Section 3.1.1. We compute type I error and the average power for SDL-TEST and Bühlmann's method [Büh12] for 10 realizations of each configuration. The experiment results for the case of identity covariance ($\Sigma = \mathrm{I}_{p\times p}$) are summarized in Tables 8 and 9. Table 8 and Table 9 respectively correspond to significance levels $\alpha = 0.05$, and $\alpha = 0.025$. The results are also compared with the asymptotic results given in Theorem 3.4.

The results for the case of circulant covariance matrix are summarized in Tables 10 and 11. Table 10 and Table 11 respectively correspond to significance levels $\alpha = 0.05$, and $\alpha = 0.025$. The results are also compared with the lower bound given in Theorem 3.7.

For each configuration, the tables contain the means and the standard deviations of type I errors and the powers across 10 realizations. A quadruple such as $(1000, 600, 50, 0.1)$ denotes the values of $p = 1000$, $n = 600$, $s_0 = 50$, $\mu = 0.1$.

Method	Type I err (mean)	Type I err (std.)	Avg. power (mean)	Avg. power (std)
Our testing Procedure $(1000, 600, 100, 0.1)$	0.05422	0.01069	0.44900	0.06951
Bühlmann's method $(1000, 600, 100, 0.1)$	0.01089	0.00358	0.13600	0.02951
Asymptotic Bound $(1000, 600, 100, 0.1)$	0.05	NA	0.37692	NA
Our testing Procedure $(1000, 600, 50, 0.1)$	0.04832	0.00681	0.52000	0.06928
Bühlmann's method $(1000, 600, 50, 0.1)$	0.01989	0.00533	0.17400	0.06670
Asymptotic Bound $(1000, 600, 50, 0.1)$	0.05	NA	0.51177	NA
Our testing Procedure $(1000, 600, 25, 0.1)$	0.06862	0.01502	0.56400	0.11384
Bühlmann's method $(1000, 600, 25, 0.1)$	0.02431	0.00536	0.25600	0.06586
Asymptotic Bound $(1000, 600, 25, 0.1)$	0.05	NA	0.58822	NA
Our testing Procedure $(1000, 300, 25, 0.1)$	0.05303	0.01527	0.36400	0.07648
Bühlmann's method $(1000, 300, 25, 0.1)$	0.01374	0.00279	0.13200	0.04638
Asymptotic Bound $(1000, 300, 25, 0.1)$	0.05	NA	0.26456	NA
Our testing Procedure $(1000, 300, 10, 0.1)$	0.05364	0.01430	0.34000	0.11738
Bühlmann's method $(1000, 300, 10, 0.1)$	0.01465	0.00424	0.16000	0.08433
Asymptotic Bound $(1000, 300, 10, 0.1)$	0.05	NA	0.33906	NA
Our testing Procedure $(1000, 600, 100, 0.15)$	0.04944	0.00791	0.70700	0.07409
Bühlmann's method $(1000, 600, 100, 0.15)$	0.00400	0.00247	0.18600	0.04742
Asymptotic Bound $(1000, 600, 100, 0.15)$	0.05	NA	0.69478	NA
Our testing Procedure $(1000, 600, 50, 0.15)$	0.06189	0.01663	0.83600	0.04300
Bühlmann's method $(1000, 600, 50, 0.15)$	0.00989	0.00239	0.35000	0.07071
Asymptotic Bound $(1000, 600, 50, 0.15)$	0.05	NA	0.84721	NA
Our testing Procedure $(1000, 600, 25, 0.15)$	0.0572	0.0190	0.8840	0.0638
Bühlmann's method $(1000, 600, 25, 0.15)$	0.0203	0.0052	0.3680	0.1144
Asymptotic Bound $(1000, 600, 25, 0.15)$	0.05	NA	0.9057	NA
Our testing Procedure $(1000, 300, 50, 0.15)$	0.05547	0.01554	0.45800	0.06957
Bühlmann's method $(1000, 300, 50, 0.15)$	0.01084	0.00306	0.19200	0.04541
Asymptotic Bound $(1000, 300, 50, 0.15)$	0.05	NA	0.31224	NA
Our testing Procedure $(1000, 300, 25, 0.15)$	0.05149	0.01948	0.55600	0.11384
Bühlmann's method $(1000, 300, 25, 0.15)$	0.00964	0.00436	0.32400	0.09324
Asymptotic Bound $(1000, 300, 25, 0.15)$	0.05	NA	0.51364	NA
Our testing Procedure $(2000, 600, 100, 0.1)$	0.05037	0.00874	0.44800	0.04940
Bühlmann's method $(2000, 600, 100, 0.1)$	0.01232	0.00265	0.21900	0.03143
Asymptotic Bound $(2000, 600, 100, 0.1)$	0.05	NA	0.28324	NA
Our testing Procedure $(2000, 600, 50, 0.1)$	0.05769	0.00725	0.52800	0.08548
Bühlmann's method $(2000, 600, 50, 0.1)$	0.01451	0.00303	0.27000	0.04137
Asymptotic Bound $(2000, 600, 50, 0.1)$	0.05	NA	0.46818	NA
Our testing Procedure $(2000, 600, 20, 0.1)$	0.05167	0.00814	0.58000	0.11595
Bühlmann's method $(2000, 600, 20, 0.1)$	0.01879	0.00402	0.34500	0.09846
Asymptotic Bound $(2000, 600, 20, 0.1)$	0.05	NA	0.58879	NA
Our testing Procedure $(2000, 600, 100, 0.15)$	0.05368	0.01004	0.64500	0.05104
Bühlmann's method $(2000, 600, 100, 0.15)$	0.00921	0.00197	0.30700	0.04877
Asymptotic Bound $(2000, 600, 100, 0.15)$	0.05	NA	0.54728	NA
Our testing Procedure $(2000, 600, 20, 0.15)$	0.04944	0.01142	0.89500	0.07619
Bühlmann's method $(2000, 600, 20, 0.15)$	0.01763	0.00329	0.64000	0.08756
Asymptotic Bound $(2000, 600, 20, 0.15)$	0.05	NA	0.90608	NA

Table 8: Comparison between our procedure (cf. Table 1), Bühlmann's method [Büh12] and the asymptotic bound for our procedure (cf. Theorem 3.4) on the setup described in Section 3.1.1. The significance level is $\alpha = 0.05$ and $\Sigma = \mathrm{I}_{p\times p}$.

Method	Type I err (mean)	Type I err (std.)	Avg. power (mean)	Avg. power (std)
Our testing Procedure $(1000, 600, 100, 0.1)$	0.02478	0.00954	0.35300	0.06550
Bühlmann's method $(1000, 600, 100, 0.1)$	0.00300	0.00196	0.06100	0.01792
Asymptotic Bound $(1000, 600, 100, 0.1)$	0.025	NA	0.27560	NA
Our testing Procedure $(1000, 600, 50, 0.1)$	0.02611	0.00784	0.39600	0.06168
Bühlmann's method $(1000, 600, 50, 0.1)$	0.00674	0.00282	0.11200	0.03425
Asymptotic Bound $(1000, 600, 50, 0.1)$	0.025	NA	0.40025	NA
Our testing Procedure $(1000, 600, 25, 0.1)$	0.02923	0.01189	0.46000	0.08692
Bühlmann's method $(1000, 600, 25, 0.1)$	0.01344	0.00436	0.13200	0.06546
Asymptotic Bound $(1000, 600, 25, 0.1)$	0.025	NA	0.47640	NA
Our testing Procedure $(1000, 300, 25, 0.1)$	0.02974	0.00949	0.22800	0.08011
Bühlmann's method $(1000, 300, 25, 0.1)$	0.00421	0.00229	0.08000	0.06532
Asymptotic Bound $(1000, 300, 25, 0.1)$	0.025	NA	0.18080	NA
Our testing Procedure $(1000, 300, 10, 0.1)$	0.02586	0.00794	0.22000	0.16865
Bühlmann's method $(1000, 300, 10, 0.1)$	0.00727	0.00177	0.08000	0.07888
Asymptotic Bound $(1000, 300, 10, 0.1)$	0.025	NA	0.24273	NA
Our testing Procedure $(1000, 600, 100, 0.15)$	0.025222	0.006064	0.619000	0.061725
Bühlmann's method $(1000, 600, 100, 0.15)$	0.000778	0.000750	0.105000	0.026352
Asymptotic Bound $(1000, 600, 100, 0.15)$	0.025	NA	0.5899	NA
Our testing Procedure $(1000, 600, 50, 0.15)$	0.02874	0.00546	0.75600	0.07706
Bühlmann's method $(1000, 600, 50, 0.15)$	0.00379	0.00282	0.22800	0.06052
Asymptotic Bound $(1000, 600, 50, 0.15)$	0.025	NA	0.77107	NA
Our testing Procedure $(1000, 600, 25, 0.15)$	0.03262	0.00925	0.79200	0.04131
Bühlmann's method $(1000, 600, 25, 0.15)$	0.00759	0.00223	0.28800	0.07729
Asymptotic Bound $(1000, 600, 25, 0.15)$	0.025	NA	0.84912	NA
Our testing Procedure $(1000, 300, 50, 0.15)$	0.02916	0.00924	0.36000	0.08380
Bühlmann's method $(1000, 300, 50, 0.15)$	0.00400	0.00257	0.10800	0.05432
Asymptotic Bound $(1000, 300, 50, 0.15)$	0.025	NA	0.22001	NA
Our testing Procedure $(1000, 300, 25, 0.15)$	0.03005	0.00894	0.42400	0.08884
Bühlmann's method $(1000, 300, 25, 0.15)$	0.00492	0.00226	0.21600	0.06310
Asymptotic Bound $(1000, 300, 25, 0.15)$	0.025	NA	0.40207	NA
Our testing Procedure $(2000, 600, 100, 0.1)$	0.03079	0.00663	0.33000	0.05033
Bühlmann's method $(2000, 600, 100, 0.1)$	0.00484	0.00179	0.11200	0.03615
Asymptotic Bound $(2000, 600, 100, 0.1)$	0.025	NA	0.19598	NA
Our testing Procedure $(2000, 600, 50, 0.1)$	0.02585	0.00481	0.41200	0.06197
Bühlmann's method $(2000, 600, 50, 0.1)$	0.00662	0.00098	0.20600	0.03406
Asymptotic Bound $(2000, 600, 50, 0.1)$	0.025	NA	0.35865	NA
Our testing Procedure $(2000, 600, 20, 0.1)$	0.02626	0.00510	0.47500	0.10607
Bühlmann's method $(2000, 600, 20, 0.1)$	0.00838	0.00232	0.23500	0.08182
Asymptotic Bound $(2000, 600, 20, 0.1)$	0.025	NA	0.47698	NA
Our testing Procedure $(2000, 600, 100, 0.15)$	0.02484	0.00691	0.52700	0.09522
Bühlmann's method $(2000, 600, 100, 0.15)$	0.00311	0.00154	0.22500	0.04007
Asymptotic Bound $(2000, 600, 100, 0.15)$	0.025	NA	0.43511	NA
Our testing Procedure $(2000, 600, 20, 0.15)$	0.03116	0.01304	0.81500	0.09443
Bühlmann's method $(2000, 600, 20, 0.15)$	0.00727	0.00131	0.54500	0.09560
Asymptotic Bound $(2000, 600, 20, 0.15)$	0.025	NA	0.84963	NA

Table 9: Comparison between our procedure (cf. Table 1), Bühlmann's method [Büh12] and the asymptotic bound for our procedure (cf. Theorem 3.4) on the setup described in Section 3.1.1. The significance level is $\alpha = 0.025$ and $\Sigma = \mathrm{I}_{p\times p}$.

Method	Type I err (mean)	Type I err (std.)	Avg. power (mean)	Avg. power (std)
SDL-test $(1000, 600, 100, 0.1)$	0.06733	0.01720	0.48300	0.03433
Bühlmann's method $(1000, 600, 100, 0.1)$	0.00856	0.00416	0.11000	0.02828
Lower bound $(1000, 600, 100, 0.1)$	0.05	NA	0.45685	0.04540
SDL-test $(1000, 600, 50, 0.1)$	0.04968	0.00997	0.50800	0.05827
Bühlmann's method $(1000, 600, 50, 0.1)$	0.01642	0.00439	0.21000	0.04738
Lower bound $(1000, 600, 50, 0.1)$	0.05	NA	0.50793	0.03545
SDL-test $(1000, 600, 25, 0.1)$	0.05979	0.01435	0.55200	0.08390
Bühlmann's method $(1000, 600, 25, 0.1)$	0.02421	0.00804	0.22400	0.10013
Lower bound $(1000, 600, 25, 0.1)$	0.05	NA	0.54936	0.06176
SDL-test $(1000, 300, 25, 0.1)$	0.05887	0.01065	0.28000	0.08433
Bühlmann's method $(1000, 300, 25, 0.1)$	0.01354	0.00413	0.12400	0.04402
Lower bound $(1000, 300, 25, 0.1)$	0.05	NA	0.31723	0.02572
SDL-test $(1000, 300, 10, 0.1)$	0.0595	0.0144	0.3400	0.1350
Bühlmann's method $(1000, 300, 10, 0.1)$	0.0163	0.0040	0.1200	0.1549
Lower bound $(1000, 300, 10, 0.1)$	0.05	NA	0.32567	0.04661
SDL-test $(1000, 600, 100, 0.15)$	0.04722	0.01164	0.72200	0.04638
Bühlmann's method $(1000, 600, 100, 0.15)$	0.00356	0.00195	0.19100	0.03213
Lower bound $(1000, 600, 100, 0.15)$	0.05	NA	0.69437	0.05352
SDL-test $(1000, 600, 50, 0.15)$	0.05579	0.01262	0.81400	0.07604
Bühlmann's method $(1000, 600, 50, 0.15)$	0.01095	0.00352	0.34000	0.05735
Lower bound $(1000, 600, 50, 0.15)$	0.05	NA	0.84013	0.03810
SDL-test $(1000, 600, 25, 0.15)$	0.05374	0.01840	0.85600	0.06310
Bühlmann's method $(1000, 600, 25, 0.15)$	0.01969	0.00358	0.46800	0.08011
Lower bound $(1000, 600, 25, 0.15)$	0.05	NA	0.86362	0.02227
SDL-test $(1000, 300, 50, 0.15)$	0.05411	0.01947	0.43800	0.09402
Bühlmann's method $(1000, 300, 50, 0.15)$	0.01011	0.00362	0.20200	0.05029
Lower bound $(1000, 300, 50, 0.15)$	0.05	NA	0.43435	0.03983
SDL-test $(1000, 300, 25, 0.15)$	0.05262	0.01854	0.53600	0.08044
Bühlmann's method $(1000, 300, 25, 0.15)$	0.01344	0.00258	0.33200	0.08230
Lower bound $(1000, 300, 25, 0.15)$	0.05	NA	0.50198	0.05738
SDL-test $(2000, 600, 100, 0.1)$	0.05268	0.01105	0.43900	0.04383
Bühlmann's method $(2000, 600, 100, 0.1)$	0.01205	0.00284	0.21200	0.04392
Lower bound $(2000, 600, 100, 0.1)$	0.05	NA	0.41398	0.03424
SDL-test $(2000, 600, 50, 0.1)$	0.05856	0.00531	0.50800	0.05350
Bühlmann's method $(2000, 600, 50, 0.1)$	0.01344	0.00225	0.26000	0.03771
Lower bound $(2000, 600, 50, 0.1)$	0.05	NA	0.49026	0.02625
SDL-test $(2000, 600, 20, 0.1)$	0.04955	0.00824	0.57500	0.13385
Bühlmann's method $(2000, 600, 20, 0.1)$	0.01672	0.00282	0.35500	0.08960
Lower bound $(2000, 600, 20, 0.1)$	0.05	NA	0.58947	0.04472
SDL-test $(2000, 600, 100, 0.15)$	0.05284	0.00949	0.61600	0.06802
Bühlmann's method $(2000, 600, 100, 0.15)$	0.00895	0.00272	0.31800	0.04131
Lower bound $(2000, 600, 100, 0.15)$	0.05	NA	0.64924	0.05312
SDL-test $(2000, 600, 20, 0.15)$	0.05318	0.00871	0.85500	0.11891
Bühlmann's method $(2000, 600, 20, 0.15)$	0.01838	0.00305	0.68000	0.12517
Lower bound $(2000, 600, 20, 0.15)$	0.05	NA	0.87988	0.03708

Table 10: Comparison between SDL-TEST, Bühlmann's method [Büh12] and the lower bound for the statistical power of SDL-TEST (cf. Theorem 3.7) on the setup described in Section 3.2.1. The significance level is $\alpha = 0.05$ and Σ is the described circulant matrix.

Method	Type I err (mean)	Type I err (std.)	Avg. power (mean)	Avg. power (std)
SDL-test $(1000, 600, 100, 0.1)$	0.03089	0.00982	0.36300	0.03802
Bühlmann's method $(1000, 600, 100, 0.1)$	0.00311	0.00115	0.07400	0.01430
Lower bound $(1000, 600, 100, 0.1)$	0.025	NA	0.33679	0.03593
SDL-test $(1000, 600, 50, 0.1)$	0.03411	0.01382	0.44800	0.07729
Bühlmann's method $(1000, 600, 50, 0.1)$	0.00747	0.00295	0.11400	0.03273
Lower bound $(1000, 600, 50, 0.1)$	0.025	NA	0.40814	0.03437
SDL-test $(1000, 600, 25, 0.1)$	0.03015	0.01062	0.44400	0.10741
Bühlmann's method $(1000, 600, 25, 0.1)$	0.00964	0.00279	0.14000	0.04320
Lower bound $(1000, 600, 25, 0.1)$	0.025	NA	0.43102	0.04210
SDL-test $(1000, 300, 25, 0.1)$	0.03723	0.01390	0.28000	0.08000
Bühlmann's method $(1000, 300, 25, 0.1)$	0.00605	0.00271	0.08400	0.06096
Lower bound $(1000, 300, 25, 0.1)$	0.025	NA	0.23702	0.05392
SDL-test $(1000, 300, 10, 0.1)$	0.02909	0.01029	0.19000	0.11972
Bühlmann's method $(1000, 300, 10, 0.1)$	0.00606	0.00172	0.11000	0.08756
Lower bound $(1000, 300, 10, 0.1)$	0.025	NA	0.24348	0.07433
SDL-test $(1000, 600, 100, 0.15)$	0.027000	0.006850	0.626000	0.065354
Bühlmann's method $(1000, 600, 100, 0.15)$	0.000667	0.000777	0.112000	0.08724
Lower bound $(1000, 600, 100, 0.15)$	0.025	NA	0.63952	0.06072
SDL-test $(1000, 600, 50, 0.15)$	0.02979	0.00967	0.71800	0.03824
Bühlmann's method $(1000, 600, 50, 0.15)$	0.00326	0.00274	0.21000	0.05437
Lower bound $(1000, 600, 50, 0.15)$	0.025	NA	0.75676	0.05937
SDL-test $(1000, 600, 25, 0.15)$	0.03262	0.00866	0.75600	0.12429
Bühlmann's method $(1000, 600, 25, 0.15)$	0.01077	0.00346	0.30400	0.08262
Lower bound $(1000, 600, 25, 0.15)$	0.025	NA	0.80044	0.05435
SDL-test $(1000, 300, 50, 0.15)$	0.03463	0.01473	0.39200	0.11478
Bühlmann's method $(1000, 300, 50, 0.15)$	0.00368	0.00239	0.15000	0.04137
Lower bound $(1000, 300, 50, 0.15)$	0.025	NA	0.36084	0.04315
SDL-test $(1000, 300, 25, 0.15)$	0.02800	0.00892	0.42400	0.09834
Bühlmann's method $(1000, 300, 25, 0.15)$	0.00513	0.00118	0.18800	0.07786
Lower bound $(1000, 300, 25, 0.15)$	0.025	NA	0.42709	0.03217
SDL-test $(2000, 600, 100, 0.1)$	0.03268	0.00607	0.32600	0.07412
Bühlmann's method $(2000, 600, 100, 0.1)$	0.00432	0.00179	0.14100	0.05065
Lower bound $(2000, 600, 100, 0.1)$	0.025	NA	0.32958	0.03179
SDL-test $(2000, 600, 50, 0.1)$	0.03108	0.00745	0.41800	0.04662
Bühlmann's method $(2000, 600, 50, 0.1)$	0.00687	0.00170	0.18800	0.06680
Lower bound $(2000, 600, 50, 0.1)$	0.025	NA	0.40404	0.06553
SDL-test $(2000, 600, 20, 0.1)$	0.02965	0.00844	0.38500	0.07091
Bühlmann's method $(2000, 600, 20, 0.1)$	0.00864	0.00219	0.22500	0.07906
Lower bound $(2000, 600, 20, 0.1)$	0.025	NA	0.47549	0.06233
SDL-test $(2000, 600, 100, 0.15)$	0.026737	0.009541	0.528000	0.062681
Bühlmann's method $(2000, 600, 100, 0.15)$	0.002947	0.000867	0.236000	0.035653
Lower bound $(2000, 600, 100, 0.15)$	0.025	NA	0.54512	0.05511
SDL-test $(2000, 600, 20, 0.15)$	0.03298	0.00771	0.79000	0.12202
Bühlmann's method $(2000, 600, 20, 0.15)$	0.00732	0.00195	0.53500	0.07091
Lower bound $(2000, 600, 20, 0.15)$	0.025	NA	0.81899	0.03012

Table 11: Comparison between SDL-test, Bühlmann's method [Büh12] and the lower bound for the statistical power of SDL-test (cf. Theorem 3.7) on the setup described in Section 3.2.1. The significance level is $\alpha = 0.025$ and Σ is the described circulant matrix.

D Alternative hypothesis testing procedure

SDL-TEST, described in Table 3, needs to compute an estimate of the covariance matrix Σ. Here, we discuss another hypothesis testing procedure which leverages on a slightly different form of the standard distributional limit, cf. Definition 3.5. This procedure only requires bounds on Σ that can be estimated from the data. Furthermore, we establish a connection with the hypothesis testing procedure of [Büh12]. We will describe this alternative procedure synthetically since it is not the main focus of the paper.

By Definition 3.5, if a sequence of instances $\mathcal{S} = \{(\Sigma(p), \theta(p), n(p), \sigma(p))\}_{p\in\mathbb{N}}$ has standard distributional limit, then with probability one the empirical distribution of $\{(\widehat{\theta}^u_i - \theta_i)/[(\Sigma^{-1})_{i,i}]^{1/2}\}_{i=1}^p$ converges weakly to $\mathsf{N}(0,\tau^2)$. We make a somewhat different assumption that is also supported by the statistical physics arguments of Appendix B. The two assumptions coincide in the case of standard Gaussian designs.

In order to motivate the new assumption, notice that the standard distributional limit is consistent with $\widehat{\theta}^u - \theta_0$ being approximately $\mathsf{N}(0, \tau^2\Sigma^{-1})$. If this holds, then

$$\Sigma(\widehat{\theta}^u - \theta_0) = \Sigma(\widehat{\theta} - \theta_0) + \frac{\mathsf{d}}{n}\mathbf{X}^{\mathsf{T}}(y - \mathbf{X}\widehat{\theta}) \approx \mathsf{N}(0, \tau^2\Sigma). \tag{88}$$

This motivates the definition of $\tilde{\theta}_i = \tau^{-1}(\Sigma_{i,i})^{-1/2}[\Sigma(\widehat{\theta}^u - \theta_0)]_i$. We then assume that the empirical distribution of $\{(\theta_{0,i}, \tilde{\theta}_i, D)\}_{i\in[p]}$ converges weakly to (Θ_0, Z, D), with $Z \sim \mathsf{N}(0,1)$ independent of Θ_0, D.

Under the null-hypothesis $H_{0,i}$, we get

$$\begin{aligned}
\tilde{\theta}_i &= \tau^{-1}(\Sigma_{i,i})^{-1/2}[\Sigma(\widehat{\theta}^u - \theta_0)]_i \\
&= \tau^{-1}(\Sigma_{i,i})^{-1/2}[\Sigma(\widehat{\theta} - \theta_0) + \frac{\mathsf{d}}{n}\mathbf{X}^{\mathsf{T}}(y - \mathbf{X}\widehat{\theta})]_i \\
&= \tau^{-1}(\Sigma_{i,i})^{1/2}\widehat{\theta}_i + \tau^{-1}(\Sigma_{i,i})^{-1/2}[\frac{\mathsf{d}}{n}\mathbf{X}^{\mathsf{T}}(y - \mathbf{X}\widehat{\theta})]_i + \tau^{-1}(\Sigma_{i,i})^{-1/2}\Sigma_{i,\sim i}(\widehat{\theta}_{\sim i} - \theta_{0,\sim i}),
\end{aligned}$$

where $\Sigma_{i,\sim i}$ denotes the vector $(\Sigma_{i,j})_{j\neq i}$. Similarly $\widehat{\theta}_{\sim i}$ and $\theta_{0,\sim i}$ respectively denote the vectors $(\widehat{\theta}_j)_{j\neq i}$ and $(\theta_{0,j})_{j\neq i}$. Therefore,

$$\tau^{-1}(\Sigma_{i,i})^{1/2}\widehat{\theta}_i + \tau^{-1}(\Sigma_{i,i})^{-1/2}[\frac{\mathsf{d}}{n}\mathbf{X}^{\mathsf{T}}(y - \mathbf{X}\widehat{\theta})]_i = \tilde{\theta}_i - \tau^{-1}(\Sigma_{i,i})^{-1/2}\Sigma_{i,\sim i}(\widehat{\theta}_{\sim i} - \theta_{0,\sim i}).$$

Following the philosophy of [Büh12], the key step in obtaining a p-value for testing $H_{0,i}$ is to find constants Δ_i, such that asymptotically

$$\xi_i \equiv \tau^{-1}(\Sigma_{i,i})^{1/2}\widehat{\theta}_i + \tau^{-1}(\Sigma_{i,i})^{-1/2}[\frac{\mathsf{d}}{n}\mathbf{X}^{\mathsf{T}}(y - \mathbf{X}\widehat{\theta})]_i \preceq |Z| + \Delta_i, \tag{89}$$

where $Z \sim \mathsf{N}(0,1)$, and $\preceq$ denotes "stochastically smaller than or equal to". Then, we can define the p-value for the two-sided alternative as

$$P_i = 2(1 - \Phi((|\xi_i| - \Delta_i)+)).$$

Control of type I errors then follows immediately from the construction of p-values:

$$\limsup_{p\to\infty} \mathbb{P}(P_i \le \alpha) \le \alpha, \qquad \text{if } H_{0,i} \text{ holds.}$$

In order to define the constant Δ_i, we use analogous argument to the one in [Büh12]:

$$|\tau^{-1}(\Sigma_{i,i})^{-1/2}\Sigma_{i,\sim i}(\widehat{\theta}_{\sim i}-\theta_{0,\sim i})| \le \max_{j\neq i}|\Sigma_{i,j}|\,(\tau^{-1}\Sigma_{i,i}^{-1/2})\,\|\widehat{\theta}-\theta_0\|_1.$$

Recall that $\widehat{\theta}=\widehat{\theta}(\lambda)$ is the solution of the Lasso with regularization parameter λ. Due to the result of [BRT09, vdGB09], using $\lambda = 4\sigma\sqrt{(t^2+2\log(p))/n}$, the following holds with probability at least $1-2e^{-t^2/2}$:

$$\|\widehat{\theta}-\theta_0\|_1 \le 4\lambda s_0/\phi_0^2, \tag{90}$$

where s_0 is the sparsity (number of active parameters) and ϕ_0 is the compatibility constant. Assuming for simplicity $\Sigma_{i,i}=1$ (which can be ensured by normalizing the columns of $\mathbf{X}$), we can define

$$\Delta_i \equiv \frac{4\lambda s_0}{\tau\phi_0^2}\max_{j\neq i}|\Sigma_{i,j}|\,.$$

Therefore, this procedure only requires to bound the off-diagonal entries of Σ, i.e., $\max_{j\neq i}|\Sigma_{i,j}|$. It is straightforward to bound this quantity using the empirical covariance, $\widehat{\Sigma}=(1/n)\mathbf{X}^{\mathsf{T}}\mathbf{X}$.

Claim D.1. *Consider Gaussian design matrix $\mathbf{X}\in\mathbb{R}^{n\times p}$, whose rows are drawn independently from $\mathsf{N}(0,\Sigma)$. For any fixed $i\in[p]$, the following holds true with probability at least $1-6p^{-1/3}$*

$$\max_{j\neq i}|\Sigma_{i,j}| \le \max_{j\neq i}|\widehat{\Sigma}_{i,j}| + 20\sqrt{\frac{\log p}{n}}\,. \tag{91}$$

Proof. Write

$$2\widehat{\Sigma}_{i,j} = \frac{2}{n}\widetilde{x}_j^{\mathsf{T}}\widetilde{x}_j = \frac{1}{n}\|\widetilde{x}_i+\widetilde{x}_j\|^2 - \frac{1}{n}\|\widetilde{x}_i\|^2 - \frac{1}{n}\|\widetilde{x}_j\|^2\,.$$

Note that $\|\widetilde{x}_i+\widetilde{x}_j\|^2 \sim 2(1+\Sigma_{i,j})Z_n$, and $\|\widetilde{x}_i\|^2, \|\widetilde{x}_j\|^2 \sim Z_n$, where Z_n is a chi-squared random variable with n degrees of freedom.

Hence, for any c we have

$$\begin{aligned}
\mathbb{P}(\widehat{\Sigma}_{i,j}\le\Sigma_{i,j}-c) &= \mathbb{P}\Big(\frac{2}{n}\widetilde{x}_i^{\mathsf{T}}\widetilde{x}_j \le 2\Sigma_{i,j}-2c\Big)\\
&\le \mathbb{P}\Big(\frac{1}{n}\|\widetilde{x}_i+\widetilde{x}_j\|^2 \le 2(\Sigma_{i,j}+1)-\frac{2c}{3}\Big)\\
&\quad+\mathbb{P}\Big(\frac{1}{n}\|\widetilde{x}_i\|^2 \ge 1+\frac{2c}{3}\Big)+\mathbb{P}\Big(\frac{1}{n}\|\widetilde{x}_j\|^2 \ge 1+\frac{2c}{3}\Big)\\
&\le \mathbb{P}\Big(Z_n \le n(1-\frac{c}{6})\Big)+2\,\mathbb{P}\Big(Z_n \ge n(1+\frac{2c}{3})\Big)\,,
\end{aligned} \tag{92}$$

where in the last step, we used $\Sigma_{i,j}\le\Sigma_{i,i}=1$.

Similarly,

$$\begin{aligned}
\mathbb{P}(\widehat{\Sigma}_{i,j}\ge\Sigma_{i,j}+c) &= \mathbb{P}\Big(\frac{2}{n}\widetilde{x}_i^{\mathsf{T}}\widetilde{x}_j \ge 2\Sigma_{i,j}+2c\Big)\\
&\le \mathbb{P}\Big(\frac{1}{n}\|\widetilde{x}_i+\widetilde{x}_j\|^2 \ge 2(\Sigma_{i,j}+1)+\frac{2c}{3}\Big)\\
&\quad+\mathbb{P}\Big(\frac{1}{n}\|\widetilde{x}_i\|^2 \le 1-\frac{2c}{3}\Big)+\mathbb{P}\Big(\frac{1}{n}\|\widetilde{x}_j\|^2 \le 1-\frac{2c}{3}\Big)\\
&\le \mathbb{P}\Big(Z_n \ge n(1+\frac{c}{6})\Big)+2\mathbb{P}\Big(Z_n \le n(1-\frac{2c}{3})\Big)\,.
\end{aligned} \tag{93}$$

Let $F_n(x) = \mathbb{P}(Z_n \geq x)$. Then, combining Eqs. (92), (93), we obtain

$$\mathbb{P}(|\widehat{\Sigma}_{i,j} - \Sigma_{i,j}| \geq c) \leq \{1 - F_n(n(1 - \frac{c}{6})) + F_n(n(1 + \frac{c}{6}))\} + 2\{1 - F_n(n(1 - \frac{2c}{3})) + F_n(n(1 + \frac{2c}{3}))\}$$

We upper bound the above probability using the concentration of a chi-squared random variable around its mean. Indeed, applying Chernoff tail bound for Z_n (similar to the one in Corollary 2.7) and taking $c = 20\sqrt{\log p/n}$, we have that for $\log p/n < 0.01$,

$$\mathbb{P}\Big(|\widehat{\Sigma}_{i,j} - \Sigma_{i,j}| \geq 20\sqrt{\log p/n}\Big) \leq 6p^{-4/3}\,.$$

Using union bound for $j \in [p]$, $j \neq i$, we get

$$\mathbb{P}\Big(\max_{j \neq i} |\widehat{\Sigma}_{i,j} - \Sigma_{i,j}| \leq 20\sqrt{\log p/n}\Big) \geq 1 - 6p^{-1/3}\,.$$

The result follows from the inequality $\max_{j\neq i} |\Sigma_{i,j}| - \max_{j\neq i} |\widehat{\Sigma}_{i,j}| \leq \max_{j\neq i} |\widehat{\Sigma}_{i,j} - \Sigma_{i,j}|$. □

References

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

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

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

[Büh12] P. Bühlmann, *Statistical significance in high-dimensional linear models*, arXiv:1202.1377, 2012.

[CP10] E. Candés and Y. Plan, *A probabilistic and RIPless theory of compressed sensing*, arXiv:1011.3854, 2010.

[CR11] E. Candés and B. Recht, *Simple bounds for low-complexity model reconstruction*, arXiv:1106.1474, 2011.

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

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

[DMM11] D.L. Donoho, A. Maleki, and A. Montanari, *The Noise Sensitivity Phase Transition in Compressed Sensing*, IEEE Trans. on Inform. Theory **57** (2011), 6920–6941.

[Don06] D. L. Donoho, *High-dimensional centrally symmetric polytopes with neighborliness proportional to dimension*, Discrete Comput. Geometry **35** (2006), 617–652.

[DT05] D. L. Donoho and J. Tanner, *Neighborliness of randomly-projected simplices in high dimensions*, Proceedings of the National Academy of Sciences **102** (2005), no. 27, 9452–9457.

[DT09] ———, *Counting faces of randomly projected polytopes when the projection radically lowers dimension*, Journal of American Mathematical Society **22** (2009), 1–53.

[DT10] ———, *Precise undersampling theorems*, Proceedings of the IEEE **98** (2010), 913–924.

[FA10] A. Frank and A. Asuncion, *UCI machine learning repository (communities and crime data set)*, http://archive.ics.uci.edu/ml, 2010, University of California, Irvine, School of Information and Computer Sciences.

[GBS09] D. Guo, D. Baron, and S. Shamai, *A single-letter characterization of optimal noisy compressed sensing*, Communication, Control, and Computing, 2009. Allerton 2009. 47th Annual Allerton Conference on, IEEE, 2009, pp. 52–59.

[GR04] E. Greenshtein and Y. Ritov, *Persistence in high-dimensional predictor selection and the virtue of over-parametrization*, Bernoulli **10** (2004), 971–988.

[GV05] D. Guo and S. Verdu, *Randomly Spread CDMA: Asymptotics via Statistical Physics*, IEEE Trans. on Inform. Theory **51** (2005), 1982–2010.

[GW08] D. Guo and C.C. Wang, *Multiuser detection of sparsely spread cdma*, Selected Areas in Communications, IEEE Journal on **26** (2008), no. 3, 421–431.

[HKZ11] D. Hsu, S. M. Kakade, and T. Zhang, *An Analysis of Random Design Regression*, arXiv:1106.2363, 2011.

[HR09] P.J. Huber and E. Ronchetti, *Robust statistics (second edition)*, J. Wiley and Sons, 2009.

[JM12] A. Javanmard and A. Montanari, *State Evolution for General Approximate Message Passing Algorithms, with Applications to Spatial Coupling*, arXiv:1211.5164, 2012.

[JM13] ———, *Hyothesis Testing and Sparse Support Recovery in the Gaussian Random Design Model*, In preparation, 2013.

[KMV12] Y. Kabashima and S. Chatterjee M. Vehkapera, *Typical l_1-recovery limit of sparse vectors represented by concatenations of random orthogonal matrices*, J. Stat. Mech. (2012), P12003.

[KWT09] Y. Kabashima, T. Wadayama, and T. Tanaka, *A typical reconstruction limit for compressed sensing based on L_p-norm minimization*, J.Stat. Mech. (2009), L09003.

[LR05] E.L. Lehmann and J.P. Romano, *Testing statistical hypotheses*, Springer, 2005.

[MB06] N. Meinshausen and P. Bühlmann, *High-dimensional graphs and variable selection with the lasso*, Ann. Statist. **34** (2006), 1436–1462.

[MB10] ———, *Stability selection*, J. R. Statist. Soc. B **72** (2010), 417–473.

[MM09] M. Mézard and A. Montanari, *Information, Physics and Computation*, Oxford, 2009.

[MPV87] M. Mézard, G. Parisi, and M. A. Virasoro, *Spin glass theory and beyond*, World Scientific, 1987.

[MT06] A. Montanari and D. Tse, *Analysis of belief propagation for non-linear problems: the example of CDMA (or: how to prove Tanaka's formula)*, Proceedings of IEEE Inform. Theory Workshop (Punta de l'Este, Uruguay), 2006.

[Ran11] S. Rangan, *Generalized Approximate Message Passing for Estimation with Random Linear Mixing*, IEEE Intl. Symp. on Inform. Theory (St. Perersbourg), August 2011.

[RFG09] S. Rangan, A. K. Fletcher, and V. K. Goyal, *Asymptotic Analysis of MAP Estimation via the Replica Method and Applications to Compressed Sensing*, NIPS (Vancouver), 2009.

[RWY09] G. Raskutti, M. J. Wainwright, and B. Yu, *Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls*, 47th Annual Allerton Conf. (Monticello, IL), September 2009.

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

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

[Tan02] T. Tanaka, *A Statistical-Mechanics Approach to Large-System Analysis of CDMA Multiuser Detectors*, IEEE Trans. on Inform. Theory **48** (2002), 2888–2910.

[TCSV11] A. Tulino, G. Caire, S. Shamai, and S. Verdú, *Support recovery with sparsely sampled free random matrices*, Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on, IEEE, 2011, pp. 2328–2332.

[Tib96] R. Tibshirani, *Regression shrinkage and selection with the Lasso*, J. Royal. Statist. Soc B **58** (1996), 267–288.

[TK10] K. Takeda and Y. Kabashima, *Statistical mechanical analysis of compressed sensing utilizing correlated compression matrix*, IEEE Intl. Symp. on Inform. Theory, june 2010.

[vdGB09] S.A. van de Geer and P. Bühlmann, *On the conditions used to prove oracle results for the lasso*, Electron. J. Statist. **3** (2009), 1360–1392.

[Wai09] M.J. Wainwright, *Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming*, IEEE Trans. on Inform. Theory **55** (2009), 2183–2202.

[WV11] Y. Wu and S. Verdú, *Optimal Phase Transitions in Compressed Sensing*, arXiv:1111.6822, 2011.

[ZY06] P. Zhao and B. Yu, *On model selection consistency of Lasso*, The Journal of Machine Learning Research **7** (2006), 2541–2563.

[ZZ11] C.-H. Zhang and S.S. Zhang, *Confidence Intervals for Low-Dimensional Parameters in High-Dimensional Linear Models*, arXiv:1110.2563, 2011.