ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2020)
82, *Part* 3, *pp.* 685–718

A flexible framework for hypothesis testing in high dimensions

Adel Javanmard

University of Southern California, Los Angeles, USA

and Jason D. Lee

Princeton University, USA

[Received September 2018. Final revision March 2020]

Summary. Hypothesis testing in the linear regression model is a fundamental statistical problem. We consider linear regression in the high dimensional regime where the number of parameters exceeds the number of samples ($p > n$). To make informative inference, we assume that the model is approximately sparse, i.e. the effect of covariates on the response can be well approximated by conditioning on a relatively small number of covariates whose identities are unknown. We develop a framework for testing very general hypotheses regarding the model parameters. Our framework encompasses testing whether the parameter lies in a convex cone, testing the signal strength, and testing arbitrary functionals of the parameter. We show that the procedure proposed controls the type I error, and we also analyse the power of the procedure. Our numerical experiments confirm our theoretical findings and demonstrate that we control the false positive rate (type I error) near the nominal level and have high power. By duality between hypotheses testing and confidence intervals, the framework proposed can be used to obtain valid confidence intervals for various functionals of the model parameters. For linear functionals, the length of confidence intervals is shown to be minimax rate optimal.

Keywords: Bias; Confidence intervals; False positive rate; High dimensional inference; Hypothesis testing; Statistical power

1. Introduction

Consider the high dimensional regression model where we are given n independent and identically distributed pairs $(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)$ with $y_i \in \mathbb{R}$ and $x_i \in \mathbb{R}^p$ denoting the response values and the feature vectors respectively. The linear regression model posits that response values are generated as

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

Here $\theta_0 \in \mathbb{R}^p$ is a vector of parameters to be estimated. In matrix form, letting $y = (y_1, \ldots, y_n)^{\mathrm{T}}$ and denoting by X the matrix with rows $x_1^{\mathrm{T}}, \ldots, x_n^{\mathrm{T}}$ we have

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

We are interested in high dimensional models where the number of parameters p may far exceed the sample size n. To make informative inference feasible in this setting, we assume

Address for correspondence: Adel Javanmard, Data Sciences and Operations Department, Marshall School of Business, University of Southern California, 3670 Trousdale Parkway, Los Angeles, CA 90089, USA.
E-mail: ajavanma@usc.edu

 1369–7412/20/82685

sparsity structure for the model, i.e. θ_0 has only a few ($s_0 < n$) non-zero entries, whose identities are unknown.

Our goal in this paper is to understand various parameter structures of the high dimensional model. Specifically, we develop a flexible framework for testing null hypotheses of the form

$$H_0 : \theta_0 \in \Omega_0 \quad \textit{versus} \quad H_A : \theta_0 \notin \Omega_0, \tag{3}$$

for a general set $\Omega_0 \subset \mathbb{R}^p$. Remarkably, we make no additional assumptions (such as convexity or connectedness) on Ω_0.

In Section 5, we shall relax the sparsity assumption on the model parameters to *approximate sparsity*. Consider the linear model $y = X\theta_* + w$, where $\theta_* \in \mathbb{R}^p$ is not necessarily sparse. Approximate sparsity posits that, even if the true signal $X\theta_*$ cannot be written as a sparse linear combination of the covariates, there is at least one sparse linear combination of the covariates that grows close to the true signal. Formally, we assume that there is a vector $\theta_0 \in \mathbb{R}^p$ such that $\|\theta_0\|_0 = s_0$, and $\|X\theta_* - X\theta_0\| = o_P(1)$. Note that this notion of approximate sparsity is similar to but stronger than that introduced in Bunea *et al.* (2007) and Belloni *et al.* (2012). (In Belloni *et al.* (2012) the approximate sparsity assumption allows $\|X\theta_* - X\theta_0\| = O_P(\sqrt{s_0})$, whereas here we are imposing the stronger requirement $\|X\theta_* - X\theta_0\| = o_P(1)$.)

In addition, in Section 6 we extend our analysis to non-Gaussian heteroscedastic noise.

1.1. Motivation

High dimensional models are ubiquitous in many areas of applications. Examples range from signal processing (e.g. compressed sensing), to recommender systems (collaborative filtering), to statistical network analysis, to predictive analytics, etc. The widespread interest in these applications has spurred remarkable progress in the area of high dimensional data analysis (Candès and Tao, 2007; Bickel *et al.*, 2009; Bühlmann and van de Geer, 2011). Given that the number of parameters goes beyond the sample size, there is no hope of designing reasonable estimators without making further assumptions on the structure of model parameters. A natural such assumption is sparsity, which posits that only s_0 of the parameters $\theta_{0,i}$ are non-zero, and $s_0 \leqslant n$. A prominent approach in this setting for estimating the model parameters is via the lasso estimator (Tibshirani, 1996; Chen and Donoho, 1995) defined by

$$\hat{\theta}^n(y, X; \lambda) \equiv \arg\max_{\theta \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|y - X\theta\|_2^2 + \lambda \|\theta\|_1 \right\}. \tag{4}$$

(We shall omit the arguments of $\hat{\theta}^n(y, X; \lambda)$ whenever clear from the context.)

To date, the majority of work on high dimensional parametric models has focused on point estimation such as consistency for prediction (Greenshtein and Ritov, 2004), oracle inequalities and estimation of parameter vectors (Candès and Tao, 2007; Bickel *et al.*, 2009; Raskutti *et al.*, 2009), model selection (Meinshausen and Bühlmann, 2006; Zhao and Yu, 2006; Wainwright, 2009) and variable screening (Fan and Lv, 2008). Bunea *et al.* (2007) extended the oracle inequalities for the lasso to the setting of weak sparsity and weak approximation, where the effect of covariates on the response can be controlled up to a small approximation error by conditioning on a relatively small number of covariates, whose identities are unknown. The minimax rate for estimating the parameters in the high dimensional linear model was studied in Ye and Zhang (2010) and Raskutti *et al.* (2011), assuming that the true model parameters belong to some l_q-ball.

Despite this remarkable progress, the fundamental problem of statistical significance is far less

understood in the high dimensional setting. Uncertainty assessment is particularly important when one seeks subtle statistical patterns about the model parameters θ_0.

Below, we discuss some important examples of high dimensional inference that can be performed when provided a methodology for testing hypotheses of form (3).

1.1.1. Example 1 (testing $\theta_{\min}$ condition)

Support that recovery in high dimension concerns the problem of finding a set $\hat{S} \subseteq \{1, 2, \ldots, p\}$, such that $\mathbb{P}(\hat{S} = S)$ is large, where $S \equiv \{i : \theta_{0,i} \neq 0, 1 \leqslant i \leqslant p\}$. Work on support recovery requires that the non-zero parameters are sufficiently large to be detected. Specifically, for exact support recovery meaning that $\mathbb{P}(\hat{S} \neq S) \to 1$, it is assumed that $\min_{i \in S} |\theta_{0,i}| = \Omega[\sqrt{\{\log(p)/n\}}]$. This assumption is often referred to as the $\theta_{\min}$-condition and has been shown to be necessary for exact support recovery (Zhao and Yu, 2006; Fan and Li, 2001; Wainwright, 2009; Meinshausen and Bühlmann, 2006).

Relaxing the task of exact support recovery, let α and β be the type I and type II error rates in detecting non-zero (active) parameters of the model. In Javanmard and Montanari (2014a), it is shown that, even for Gaussian design matrices, any hypothesis testing rule with non-trivial power $1 - \beta > \alpha$ requires $\min_{i \in S} |\theta_{0,i}| = \Omega(1/\sqrt{n})$. Although the $\theta_{\min}$-assumption is commonplace, it is not verifiable in practice and hence it calls for developing methodologies that can test whether such a condition holds true.

For a vector $\theta \in \mathbb{R}^p$, define the support of θ as $\text{supp}(\theta) = \{1 \leqslant i \leqslant p : \theta_i \neq 0\}$. In tests (3), letting $\Omega_0 = \{\theta \in \mathbb{R}^p : \min_{i \in \text{supp}(\theta)} |\theta_i| \geqslant c\}$, we can test the $\theta_{\min}$-condition for any given $c \geqslant 0$ and at a preassigned level of significance α.

1.1.2. Example 2 (confidence intervals for quadratic forms)

We can apply our method to test hypotheses of form

$$H_0 : \| Q\theta_0 \|_2 \in \Omega_0, \tag{5}$$

for some given set $\Omega_0 \subseteq [0, \infty)$ and a given matrix $Q \in \mathbb{R}^{m \times p}$. By duality between hypothesis testing and confidence intervals, we can also construct confidence intervals for quadratic forms $\| Q\theta_0 \|$.

In the case of $Q = \mathrm{I}$, this yields inference on the signal strength $\|\theta\|_2^2$. As noted in Janson *et al.* (2017), armed with such a testing method we can also provide confidence intervals for the estimation error, namely $\|\hat{\theta} - \theta_0\|_2^2$. Specifically, we split the collected samples into two independent groups $(y^{(0)}, X^{(0)})$ and $(y^{(1)}, X^{(1)})$, and construct an estimate $\hat{\theta}$ just by using the first group. Letting $\tilde{y} \equiv y^{(1)} - X^{(1)}\hat{\theta}$, we obtain a linear regression model $\tilde{y} = X^{(1)}(\theta_0 - \hat{\theta}) + w$. Further, if $\hat{\theta}$ is a sparse estimate, then $\theta_0 - \hat{\theta}$ is also sparse. Therefore, inference on the signal strength on the model obtained is similar to inference on the error size $\|\theta_0 - \hat{\theta}\|_2^2$.

Inference on quadratic forms turns out to be closely related to several well-studied problems, such as estimates of the noise level σ^2 and the proportion of explained variation (Fan *et al.*, 2012; Bayati *et al.*, 2013; Dicker, 2014; Janson *et al.*, 2017; Verzelen and Gassiat, 2018; Guo *et al.*, 2019). To expand on this point, suppose that attributes x_i are drawn IID from a Gaussian distribution with covariance Σ, and the noise level σ^2 is unknown. Then, $\text{var}(y_i) = \sigma^2 + \|\Sigma^{1/2}\theta_0\|_2^2$. Since $\|y\|_2^2/\text{var}(y_i)$ follows a χ^2-distribution with n degrees of freedom, we have $\|y\|_2^2/n = \text{var}(y_i)\{1 + O_P(n^{-1/2})\}$. Hence, the task of inference on the quadratic form $\|\Sigma^{1/2}\theta_0\|_2^2$ and the noise level σ^2 are intimately related. This is also related to the proportion of explained variation defined as

$$\eta(\theta_0, \sigma) = \frac{\mathbb{E}\{(x_i^{\mathrm{T}}\theta_0)^2\}}{\mathrm{var}(y_i)} = \frac{\mu}{1+\mu}, \tag{6}$$

with $\mu = (1/\sigma^2)\|\Sigma^{1/2}\theta_0\|_2^2$ the signal-to-noise ratio. This quantity is of crucial importance in genetic variability (Visscher *et al.*, 2008) as it somewhat quantifies the proportion of variance in a trait (the response) that is explained by genes (the design matrix) rather than the environment (the noise part).

1.1.3. Example 3 (testing individual parameters $\theta_{0,i}$)

Recently, there has been significant interest in testing individual hypothesis $H_{0,i} : \theta_i = 0$, in the high dimensional regime. This is a challenging problem because obtaining an exact characterization of the probability distribution of the parameter estimates in the high dimensional regime is notoriously difficult.

A successful approach is based on debiasing the regularized estimators. The resulting debiased estimator is amenable to distributional characterization which can be used for inference on individual parameters (Javanmard and Montanari, 2013, 2014a, b; Zhang and Zhang, 2014; Van de Geer *et al.*, 2014). Our methodology for testing hypotheses of form (3) is built on the debiasing idea. It also recovers the debiasing approach for $\Omega_0 = \{\theta \in \mathbb{R}^p : \theta_i = 0\}$.

1.1.4. Example 4 (confidence intervals for predictions)

For a new sample ξ, we can perform inference on the response value $\xi^{\mathrm{T}}\theta_0$ by letting $\Omega_0 = \{\theta : \xi^{\mathrm{T}}\theta_0 = c\}$ for a given value c. Further, by duality between hypothesis testing and confidence intervals, we can construct confidence intervals for $\xi^{\mathrm{T}}\theta_0$. We refer to Section 7 for further details.

1.1.5. Example 5 (confidence intervals for $f(\theta_0)$)

Let $f : \mathbb{R}^p \to \mathbb{R}$ be an arbitrary function. By letting $\Omega_0 = \{\theta : f(\theta_0) = c\}$ we can test different values of $f(\theta_0)$. Further, by employing the duality relationship between hypothesis testing and confidence intervals, we can construct confidence intervals for $f(\theta_0)$. Note that examples 3 and 4 are special cases of $f(\theta_0) = e_i^{\mathrm{T}}\theta_0$ and $f(\theta_0) = \xi^{\mathrm{T}}\theta_0$. Here e_i is the ith standard basis element with 1 at the ith entry and 0 everywhere else.

1.1.6. Example 6 (testing over convex cones)

For a given cone $\mathcal{C}$, our framework enables us to test whether θ_0 belongs to $\mathcal{C}$. Some examples that naturally arise in studying treatment effects are the non-negative cone $\mathcal{C}_{\geqslant 0} = \{\theta \in \mathbb{R}^p : \theta_i \geqslant 0 \text{ for all } 1 \leqslant i \leqslant p\}$ and the monotone cone $\mathcal{C}_{\mathrm{M}} = \{\theta \in \mathbb{R}^p : \theta_1 \leqslant \theta_2 \leqslant \ldots \leqslant \theta_p\}$. Letting θ_i denote the mean of treatment i, by testing $\theta_0 \in \mathcal{C}_{\geqslant 0}$, we can test whether all the treatments in the study are harmless. Another case is when treatments correspond to an ordered set of dosages of the same drug. Then, one might reason that, if the drug is of any effect, its effect should follow a monotone relationship with its dosage. This hypothesis can be cast as $\theta_0 \in \mathcal{C}_{\mathrm{M}}$. Such testing problems over cones have been studied for Gaussian sequence models by Kudo (1963), Robertson and Wegman (1978) and Raubertas *et al.* (1986), and very recently by Wei *et al.* (2019).

1.2. Other related work

Testing in the high dimensional linear model has experienced a resurgence in the past few years. Most closely related to us is the line of work on debiasing or desparsifying that was pioneered by Zhang and Zhang (2014), Van de Geer *et al.* (2014) and Javanmard and Montanari (2014b), who proposed a debiased estimator $\hat{\theta}^{\mathrm{d}}$ such that every co-ordinate $\hat{\theta}_i^{\mathrm{d}}$ is approximately Gaussian

under the condition that $s_0^2 \log(p)/n \to 0$, which is in turn used to test single co-ordinates of θ_0, $H_0 : \theta_{0,i} = 0$, and to construct confidence intervals for $\theta_{0,i}$. In a parallel line of work, Belloni *et al.* (2011, 2013, 2014, 2017) have also designed an asymptotically Gaussian pivot via the post-double-selection lasso, under the same sample size condition of $s_0^2 \log(p)/n \to 0$. Cai and Guo (2017) established that the sample size conditions that are required by debiasing and post-double-selection are minimax optimal, meaning that to construct a confidence interval of length $O(1/\sqrt{n})$ for a co-ordinate of θ_0 requires $s_0^2 \log(p)/n \to 0$ for general unknown population covariance. Javanmard and Montanari (2018) introduced a novel 'leave-one-out' technique by which they provided a sharper analysis of the debiasing approach for Gaussian designs and improved the required sample size to $s_0 = o\{n/\log(p)^2\}$, assuming that the population covariance can be estimated sufficiently well. Very recently, Deshpande *et al.* (2019) introduced 'on-line debiasing' for statistical inference in high dimensional models where the samples are collected adaptively and hence are correlated. The work further discusses applications of on-line debiasing to batched data settings and time series analysis.

The debiasing and post-double-selection approaches have also been applied to a wide variety of other models for testing $\theta_{0,i}$ including missing data linear regression (Wang *et al.*, 2019), quantile regression (Zhao *et al.*, 2014) and graphical models (Ren *et al.*, 2015; Chen *et al.*, 2016; Wang and Kolar, 2016; Barber and Kolar, 2018).

In the multiple-testing realm, the debiasing approach has been used to control directional false discovery rates (Javanmard and Javadi, 2019). Other methods such as false discovery rate thresholding and sorted l_1-penalized estimation ('SLOPE') procedures control the false discovery rate when the design matrix X is orthogonal (Su and Candes, 2016; Bogdan *et al.*, 2015; Abramovich *et al.*, 2006). In the non-orthogonal setting, the knockoff procedure (Barber and Candès, 2015) controls the false discovery rate whenever $n \geqslant 2p$, and the noise is isotropic. In Janson and Su (2016) the knockoff procedure was generalized to control also for the familywise error rate. More recently, Candès *et al.* (2018) developed the model-free knockoff which allows for $p > n$ when the distribution of X is known.

In parallel, there have been developments in selective inference, namely inference for the variables that the lasso selects. Lee *et al.* (2016) and Tibshirani *et al.* (2016) developed exact tests for the regression coefficients corresponding to variables that the lasso selects. This was further generalized to a wide variety of polyhedral model selection procedures including marginal screening and orthogonal matching pursuit in Lee and Taylor (2014). Tian and Taylor (2018), Fithian *et al.* (2014) and Harris *et al.* (2016) developed more powerful and general selective inference procedures by introducing noise in the selection procedure. To allow for selective inference in the high dimensional setting, Lee *et al.* (2016) combined the polyhedral selection procedure with the debiased lasso to construct selectively valid confidence intervals for $\theta_{0,i}$ when $s_0 \log(p)/\sqrt{n} \to 0$.

Much of the previous work has focused on testing co-ordinates or one-dimensional projections of θ_0. An exception is Nickl and van de Geer (2013), who studied the problem of constructing confidence sets for high dimensional linear models, so that, under Gaussian designs, the confidence sets are honest over the family of sparse parameters. In other words, the confidence sets should have the desired coverage over signals that are at least s sparse for a given sparsity level s. Our work increases the applicability of the debiasing approach by allowing for the general hypothesis, $\theta_0 \in \Omega_0$. The set Ω_0 can be non-convex or even disconnected. Our set-up encompasses a broad range of testing problems and it is shown to be minimax optimal for special cases such as $\Omega = \{\theta : \theta_i = 0\}$ and $\Omega_0 = \{\theta : \xi^{\mathrm{T}}\theta = c\}$.

Zhu and Bradic (2017) have studied problem (3) independently and indeed Zhu and Bradic (2017) was posted on line around the same time as the first draft of our paper was released. This work also leverages the idea of debiasing but greatly differs from this work, in both methodol-

ogy and theory, which we now discuss. In Zhu and Bradic (2017), the debiased estimator was constructed in the standard basis (compared with ours which is done in a lower dimensional subspace) and is followed by an l_1-projection to construct the test statistic. The test statistic involves a data-dependent vector and the method uses the bootstrap to approximate the distribution of the test statistic and to set the critical values. In terms of theory, Zhu and Bradic (2017) showed that the method proposed controls the type I error at the desired level assuming that $\log(p) = o(n^{1/8})$ and $s_0 = o\{n^{1/4}/\sqrt{\log(p)}\}$ (see theorem 1 therein), whereas we prove such a result for our test under $s_0 = o\{\sqrt{n}/\log(p)\}$. It was shown in Zhu and Bradic (2017) that the rule achieves asymptotic power 1 provided that the signal strength (measured in term of the l_∞-distance of θ_0 from Ω_0) asymptotically dominates $n^{-1/4}$. In comparison, in theorem 3 we establish a lower bound of the power for *all values* of the signal strength and as a corollary of that we show that the method achieves power 1 if the signal strength dominates $n^{-1/2}$ asymptotically.

1.3. Organization of the paper

In the remaining part of Section 1, we present the notation and a few preliminary definitions. The rest of the paper presents the following contributions.

(a) In Section 2, we explain our testing methodology. It consists of constructing a debiased estimator for the projections of the model parameters in a lower dimensional subspace. It is then followed by an l_∞-projection to form the test statistic.
(b) In Section 3, we present our main results. Specifically, we show that our method controls the false positive rate under a preassigned α-level. We also derive an analytical lower bound for the statistical power of our test. In the case $\Omega_0 = \{\theta \in \mathbb{R}^d : \theta_i = 0\}$ (example 3), it matches the bound that was proposed in Javanmard and Montanari (2014b), theorem 3.5, which is also shown to be minimax optimal.
(c) In Section 5, we explain the notion of approximate sparsity and discuss how our results can be extended to allow for approximately sparse models.
(d) In Section 6, we relax the Gaussianity assumption on the noise component and discuss how to address possibly non-Gaussian heteroscedastic noise under proper moment conditions.
(e) In Section 7, we provide applications of our framework for some special cases: inference on linear predictions, quadratic forms of the parameters and testing the $\theta_{\min}$-condition. In Section 7.1, we discuss the existing literature for these subproblems and compare it with our proposed methodology.
(f) In Section 8, we provide numerical experiments to corroborate our findings and evaluate the type I error and statistical power of our test under various settings.
(g) In Appendix A, proofs of the theorems are given whereas proofs of technical lemmas are deferred to the on-line appendices.

1.4. Notation

We start by adapting some simple notation that will be used throughout the paper, along with some basic definitions from the literature on high dimensional regression.

We use e_i to refer to the ith standard basis element, e.g. $e_1 = (1, 0, \ldots, 0)$. For a vector v, $\mathrm{supp}(v)$ represents the positions of non-zero entries of v. For a vector θ and a subset S, θ_S is the restriction of θ to indices in S. For an integer $p \geqslant 1$, we use the notation $[p] = \{1, \ldots, p\}$. We write $\|v\|_p$ for the standard l_p-norm of a vector v, i.e. $\|v\|_p = (\Sigma_i |v_i|^p)^{1/p}$ and $\|v\|_0$ for the number of non-zero entries of v. Whenever the subscript p is not mentioned it should be read as the l_2-norm. For a matrix A, we denote by $|A|_\infty \equiv \max_{i \leqslant m, j \leqslant n} |A_{ij}|$ the maximum absolute value of entries

of A. Further, its maximum and minimum singular values are respectively indicated by $\sigma_{\max}(A)$ and $\sigma_{\min}(A)$. Throughout, $\Phi(x) \equiv \int_{-\infty}^{x} \exp(-t^2/2)\mathrm{d}t/\sqrt{(2\pi)}$ denotes the cumulative distribution function of the standard normal distribution. We also denote the z-values $z_\alpha = \Phi^{-1}(1-\alpha)$.

The term 'with high probability' means with probability converging to 1 as $n \to \infty$ and, for two functions $f(n)$ and $g(n)$, the notation $f(n) = o\{g(n)\}$ means that g 'dominates' f asymptotically, namely, for every fixed positive C, there exists $n(C)$ such that $f(n) \leqslant Cg(n)$ for $n > n(C)$. Likewise, $f(n) = O\{g(n)\}$ indicates that f is 'bounded' above by g asymptotically, i.e. $f(n) \leqslant Cg(n)$ for some positive constant C. Analogously, we use the notation $o_P(\cdot)$ and $O_P(\cdot)$ to indicate asymptotic behaviour in probability as the sample size n grows to ∞.

Let $\hat{\Sigma} = (X^{\mathrm{T}}X)/n \in \mathbb{R}^{p\times p}$ be the sample covariance of the design $X \in \mathbb{R}^{n\times p}$. In the high dimensional setting, where p exceeds n, $\hat{\Sigma}$ is singular. As is common in high dimensional statistics, we assume the *compatibility condition* which requires $\hat{\Sigma}$ to be non-singular in a restricted set of directions.

We use the notation $\|\cdot\|_{\psi_2}$ to refer to the sub-Gaussian norm. Specifically, for a random variable X, we let

$$\|X\|_{\psi_2} = \sup_{q \geqslant 1} q^{-1/2}(\mathbb{E}|X|^q)^{1/q}. \tag{7}$$

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

$$\|X\|_{\psi_2} = \sup_{\|x\| \leqslant 1} \|\langle X, x\rangle\|_{\psi_2}.$$

Definition 1. For a symmetric matrix $J \in \mathbb{R}^{p\times p}$ and a set $S \subseteq [p]$, the compatibility condition is defined as

$$\phi^2(J, S) \equiv \min_{\theta\in\mathbb{R}^p} \left\{ \frac{|S|\langle\theta, J\theta\rangle}{\|\theta_S\|_1^2} : \theta \in \mathbb{R}^p, \|\theta_{S^c}\|_1 \leqslant 3\|\theta_S\|_1 \right\}. \tag{8}$$

Matrix J is said to satisfy the compatibility condition for a set $S \subseteq [p]$, with constant ϕ_0 if $\phi(J, S) \geqslant \phi_0$.

2. Projection statistic

Depending on the structure of Ω_0 it may be useful if, instead of testing the null hypothesis H_0: $\theta_0 \in \Omega_0$, we test it in a lower dimensional space. Consider a k-dimensional subspace represented by an orthonormal basis $\{u_1, \ldots, u_k\}$, with $u_i \in \mathbb{R}^p$. For this section, we assume that the basis $\{u_1, \ldots, u_k\}$ is predetermined and fixed. In Section 4, we discuss how to choose the subspace depending on Ω_0 to maximize the power of the test. The projection onto this subspace is given by

$$\mathcal{P}_U(\theta) = \sum_{i=1}^{k} \langle\theta, u_i\rangle u_i = UU^{\mathrm{T}}\theta,$$

where $U = (u_1, \ldots, u_k) \in \mathbb{R}^{p\times k}$. We also use the notation $\mathcal{P}_U(\Omega_0) = \{\mathcal{P}_U(\theta) : \theta \in \Omega_0\}$ to denote the projection of Ω_0 onto the subspace U. Define the hypothesis

$$\tilde{H}_0 : \mathcal{P}_U(\theta_0) \in \mathcal{P}_U(\Omega_0). \tag{9}$$

Under the null H_0, $\tilde{H}_0$ also holds, so controlling the type I error of $\tilde{H}_0$ also controls the type I

error of H_0. In what follows we propose a testing rule $R \in \{0,1\}$ for the null hypothesis $\tilde{H}_0$ and show that it controls the type I error below a preassigned level α. Consequently,

$$\sup_{\theta \in \Omega_0} \mathbb{P}_\theta(R=1) \leqslant \sup_{\mathcal{P}_U(\theta) \in \mathcal{P}_U(\Omega_0)} \mathbb{P}_\theta(R=1) \leqslant \alpha.$$

For now, we consider an arbitrary fixed subspace U, and then after we analyse the statistical power of our test we provide guidelines on how to choose U to increase the power.

To test $\tilde{H}_0$ we construct a test statistic based on the debiasing approach.

We first let $\{\hat{\theta}, \hat{\sigma}\}$ be the scaled lasso estimator (Sun and Zhang, 2012) given by

$$\{\hat{\theta}^n(\lambda), \hat{\sigma}(\lambda)\} = \underset{\theta \in \mathbb{R}^p, \sigma > 0}{\arg\min} \left\{ \frac{1}{2\sigma n} \|y - X\theta\|_2^2 + \frac{\sigma}{2} + \lambda \|\theta\|_1 \right\}. \tag{10}$$

This optimization simultaneously gives an estimate of θ_0 and σ. We use regularization parameter $\lambda = \sqrt{\{2.05 \log(p)/n\}}$. Because of the l_1-penalization, the lasso estimator $\hat{\theta}$ is biased towards small l_1-norm, and so is the projection $\mathcal{P}_U(\theta_0)$. We view $\mathcal{P}_U(\theta_0)$ in the basis U, namely $\gamma_0 = U^{\mathrm{T}}\theta_0$, and construct a debiased estimator for it in the following way:

$$\hat{\gamma}^{\mathrm{d}} = U^{\mathrm{T}}\hat{\theta} + \frac{1}{n} G^{\mathrm{T}} X^{\mathrm{T}}(y - X\hat{\theta}), \tag{11}$$

with the decorrelating matrix $G = [g_1 | \ldots | g_k] \in \mathbb{R}^{p \times k}$, where each g_i is obtained by solving the following optimization problems for each $1 \leqslant i \leqslant k$:

$$\begin{aligned} &\text{minimize} \quad g^{\mathrm{T}} \hat{\Sigma} g \\ &\text{subject to} \quad \|\hat{\Sigma} g - u_i\|_\infty \leqslant \mu. \end{aligned} \tag{12}$$

Note that the decorrelating matrix $G \in \mathbb{R}^{p \times p}$ is a function of X, but not of y. We next state a lemma that provides a a bias–variance decomposition for $\hat{\gamma}^{\mathrm{d}}$ and brings insight about the form of debiasing given by equation (11).

Lemma 1. Let $X \in \mathbb{R}^{n \times p}$ be any (deterministic) design matrix. Assuming that optimization problem (12) is feasible for $i \in [k]$, let $\hat{\gamma}^{\mathrm{d}} = \hat{\gamma}^{\mathrm{d}}(\lambda)$ be a general debiased estimator as per equation (11). Then, setting $Z = G^{\mathrm{T}} X^{\mathrm{T}} w / \sqrt{n}$, with w the noise vector in regression (2), we have

$$\sqrt{n}(\hat{\gamma}^{\mathrm{d}} - U^{\mathrm{T}}\theta_0) = Z + \Delta, \qquad Z \sim N(0, \sigma^2 G^{\mathrm{T}} \hat{\Sigma} G), \quad \Delta = \sqrt{n}(G^{\mathrm{T}}\hat{\Sigma} - U^{\mathrm{T}})(\theta_0 - \hat{\theta}). \tag{13}$$

Further, assume that X satisfies the compatibly condition for the set $S = \mathrm{supp}(\theta_0)$, $|S| \leqslant s_0$, with constant ϕ_0, and let $K \equiv \max_{i \in [p]} (X^{\mathrm{T}} X/n)_{ii}$. Then, choosing $\lambda = c\sqrt{\{\log(p)/n\}}$, we have

$$\mathbb{P}\left\{ \|\Delta\|_\infty \geqslant \frac{c\mu\sigma s_0}{\phi_0^2} \sqrt{\log(p)} \right\} \leqslant 2p^{-c_0} + 2\exp\left(-\frac{n}{16} \right), \qquad c_0 = \frac{c^2}{32K} - 1. \tag{14}$$

Lemma 1 can be proved in a similar way to theorem 2.3 of Javanmard and Montanari (2014b) and its proof is omitted here. The decomposition (13) explains the rationale behind optimization (12). Indeed the convex program (12) aims at optimizing two objectives. On one hand, the constraint controls the term $|G^{\mathrm{T}}\hat{\Sigma} - U^{\mathrm{T}}|_\infty$, which by lemma 1 controls the bias term $\|\Delta\|_\infty$. On the other hand, it minimizes the objective function $g^{\mathrm{T}}\hat{\Sigma} g$, which controls the variance of $\hat{\gamma}_i^{\mathrm{d}}$. Therefore, the parameter μ in optimization (12) controls the bias–variance trade-off and should be chosen sufficiently large to ensure that solving problem (12) is feasible. (See Section 3.1 for further discussion.)

Remark 1. In the special case of $k=1$ and $u=e_i$, the debiased estimator (11) reduces to the estimator that was introduced in Javanmard and Montari (2014b). For the special case of $k=1$, it becomes similar to the estimator that was proposed by Cai and Guo (2017) that is used to construct confidence intervals for linear functionals of θ_0. Note that the proposed debiasing procedure incurs small bias in the infinity norm with respect to the rotated basis, $\|\hat{\gamma}^{\mathrm{d}} - U^{\mathrm{T}}\theta_0\|_\infty$, as opposed to the standard debiasing procedure (Javanmard and Montanari, 2013, 2014a, b; Zhang and Zhang, 2014; Van de Geer *et al.*, 2014) which incurs small bias, in the infinity norm, with respect to the original basis, and not necessarily in the rotated basis.

The following assumption ensures that the entries of noise Z have non-vanishing variances.

Assumption 1. We have $\liminf_{n\to\infty} \min_{i\in[k]}(G^{\mathrm{T}}\hat{\Sigma}G)_{i,i} \geqslant c_0 > 0$, for some positive constant c_0.

Assumption 1 entails the decorrelating matrix G, where our proposal constructs via optimization (12). In the following lemma, we provide a sufficient condition for assumption 1 to hold.

Lemma 2. Suppose that

$$\lim_{n\to\infty} \sup \mu(\max_{i\in[k]} \|u_i\|_1) \leqslant c < 1$$

and

$$\lim_{n\to\infty} \sup \max_{i\in[k]} (u_i^{\mathrm{T}}\hat{\Sigma}u_i) < C < \infty,$$

for some constant c, C. Then, assumption 1 holds.

We refer to the on-line appendix A.1 for the proof of lemma 2.

Remark 2. The very recent work Cai *et al.* (2019) uses the debiasing approach for inference on individualized treatment effect (and for general linear function $u^{\mathrm{T}}\theta_0$). The mechanism proposed slightly differs from problem (12) in that it includes an extra constraint. By this trick, the proposed mechanism of Cai *et al.* (2019) can be used for inference on a broad family of loading vector u. We can follow the same idea and replace optimization problem (12) by

$$\begin{aligned} &\text{minimize} \quad g^{\mathrm{T}}\hat{\Sigma}g \\ &\text{subject to} \quad \|\hat{\Sigma}g - u_i\|_\infty \leqslant \mu, \\ &\qquad\qquad\ \ |u_i^{\mathrm{T}}\hat{\Sigma}g - 1| \leqslant \mu. \end{aligned} \tag{15}$$

This way assumption 1 is automatically satisfied (see Cai *et al.* (2019), lemma 1, for the details).

Define the shorthand

$$\begin{aligned} Q^{(n)} &\equiv \frac{\hat{\sigma}^2}{n}(G^{\mathrm{T}}\hat{\Sigma}G), \\ D^{(n)} &\equiv \mathrm{diag}(\{Q_{ii}^{(n)}\}^{-1/2}). \end{aligned} \tag{16}$$

To ease the notation, we hereafter drop the superscript '(n)'. We next construct a test statistic T_n so that the large values of T_n provide evidence against the null hypothesis. For this, consider the l_∞ projection estimator that is given by

$$\begin{aligned} \theta^{\mathrm{p}} = \arg\min_{\theta\in\mathbb{R}^p} &\|D(\hat{\gamma}^{\mathrm{d}} - U^{\mathrm{T}}\theta)\|_\infty \\ &\text{subject to } \theta \in \Omega_0. \end{aligned} \tag{17}$$

We then define the test statistic to be the optimal value of estimator (17), i.e.

$$T_n = \|D(\hat{\gamma}^{\mathrm{d}} - U^{\mathrm{T}}\theta^{\mathrm{p}})\|_\infty. \tag{18}$$

The reason for using the l_∞-norm in the projection is that the bias term of $\hat{\gamma}^{\mathrm{d}}$ is controlled in l_∞-norm. (See lemma 1.) The decision rule is then based on the test statistic

$$R_X(y) = \begin{cases} 1 & \text{if } T_n \geqslant z_{\alpha/(2k)} \text{ (reject } \tilde{H}_0), \\ 0 & \text{otherwise (fail to reject } \tilde{H}_0). \end{cases} \tag{19}$$

The above procedure generalizes the debiasing approach of Javanmard and Montanari (2014b). Specifically, for $\Omega_0 = \{\theta : \theta_1 = 0\} = \{0\} \times \mathbb{R}^{p-1}$ and $U = e_1 e_1^{\mathrm{T}}$, the test rule becomes the rule that was proposed by Javanmard and Montanari (2014b) for testing hypotheses of the form $H_0 : \theta_{0,1} = 0$ *versus* its alternative.

Remark 3. Using lemma 1, under the null hypothesis $H_0 : \theta_0 \in \Omega_0$, we have that $D(\hat{\gamma}^{\mathrm{d}} - U^{\mathrm{T}}\theta^{\mathrm{p}})$ is (asymptotically) stochastically dominated by DZ, whose entries are dependent and are distributed as standard normal. The choice of threshold $z_{\alpha/(2k)}$ in expression (19) comes from using this observation and union bounding to control the (two-sided) tail of $\|DZ\|_\infty$. Given that lemma 1 also characterizes the dependence structure of the entries of DZ, we can pursue another (less conservative) approach to choose the rejection threshold. As an implication of lemma 1, and since k (the dimension of Z) is fixed, we have that, for all $t \in \mathbb{R}$,

$$\mathbb{P}\{\|D(\hat{\gamma}^{\mathrm{d}} - U^{\mathrm{T}}\theta_0)\|_\infty \leqslant t\} - \mathbb{P}(\|DZ\|_\infty \leqslant t) = o_P(1). \tag{20}$$

Under the null hypothesis H_0, we have that $\|D(\hat{\gamma}^{\mathrm{d}} - U^{\mathrm{T}}\theta^{\mathrm{p}})\|_\infty \leqslant \|D(\hat{\gamma}^{\mathrm{d}} - U^{\mathrm{T}}\theta_0)\|_\infty$ and, by result (20), the distribution of $\|D(\hat{\gamma}^{\mathrm{d}} - U^{\mathrm{T}}\theta_0)\|_\infty$ is asymptotically equal to the maximum of dependent standard normal variables $\|DZ\|_\infty$, whose distribution can be easily simulated since the covariance of the multivariate Gaussian vector DZ is known.

In the next section, we prove that decision rule (19) controls the type I error below the target level α provided that the basis U is independent of the samples (y_i, x_i), $1 \leqslant i \leqslant n$. We also develop a lower bound on the statistical power of the testing rule and use that to choose the basis U.

3. Main results

3.1. Controlling false positive rate

Definition 2. Consider a given triple $(X; U; G)$ where $X \in \mathbb{R}^{n \times p}$, $U \in \mathbb{R}^{p \times k}$ with $U^{\mathrm{T}}U = I$ and $G \in \mathbb{R}^{p \times k}$. The generalized coherence parameter of $(X; U; G)$ denoted by $\mu_*(X; U; G)$ is given by

$$\mu_*(X; U; G) \equiv |\hat{\Sigma}G - U|_\infty, \tag{21}$$

where $\hat{\Sigma} = (X^{\mathrm{T}}X)/n$ is the sample covariance of X. The minimum generalized coherence of $(X; U)$ is $\mu_{\min}(X; U) = \min_{G \in \mathbb{R}^{p \times k}} \mu_*(X; U; G)$.

Note that, choosing $\mu \geqslant \mu_{\min}(X; U)$, solving the optimization problem (12) becomes feasible.

We take a minimax perspective and require that the probability of type I error (false positive) is controlled uniformly over s_0-sparse vectors.

For a testing rule $R \in \{0, 1\}$ and a set Ω_0, we define

$$\alpha_n(R) \equiv \sup\{\mathbb{P}_{\theta_0}(R = 1) : \theta_0 \in \Omega_0,\ \|\theta_0\|_0 \leqslant s_0(n)\}. \tag{22}$$

Our first result shows the validity of our test for general set Ω_0 under deterministic designs.

Theorem 1. Consider a sequence of design matrices $X \in \mathbb{R}^{n \times p}$, with dimensions $n \to \infty$, $p = p(n) \to \infty$ satisfying the following assumptions. For each n, the sample covariance $\hat{\Sigma} = (X^{\mathrm{T}}X)/n$ satisfies the compatibility condition for the set $S_0 = \mathrm{supp}(\theta_0)$, with a constant $\phi_0 > 0$. Also, assume that $K \geqslant \max_{i \in [p]} \hat{\Sigma}_{ii}$ for some constant $K > 0$. Also consider a sequence of matrices $U \in \mathbb{R}^{p \times k}$, with fixed k and $p = p(n) \to \infty$, such that $U^{\mathrm{T}}U = I_k$.

Consider the linear regression (2) and let $\hat{\theta}^n$ and $\hat{\sigma}$ be obtained by the scaled lasso, given by expression (10), with $\lambda = c\sqrt{\{\log(p)/n\}}$. Construct a debiased estimator $\hat{\gamma}^{\mathrm{d}}$ as in expression (11) by using $\mu \geqslant \mu_{\min}(X; U)$, where $\mu_{\min}(X; U)$ is the minimum generalized coherence parameter as per definition 2, and suppose that assumption 1 holds. Choose $c^2 > 32K$ and suppose that $s_0 = o(\min[1/\{\mu\sqrt{\log(p)}\}, n/\log(p)])$. For the test R_X that is defined in equation (19), and for any $\alpha \in [0, 1]$, we have

$$\limsup_{n \to \infty} \alpha_n(R_X) \leqslant \alpha. \tag{23}$$

We next prove validity of our test for general set Ω_0 under random designs.

Theorem 2. Let $\Sigma \in \mathbb{R}^{p \times p}$ such that $\sigma_{\min}(\Sigma) \geqslant C_{\min} > 0$ and $\sigma_{\max}(\Sigma) \leqslant C_{\max} < \infty$ and $\max_{i \in [p]} \Sigma_{ii} \leqslant 1$. Suppose that $X\Sigma^{-1/2}$ has independent sub-Gaussian rows, with mean 0 and sub-Gaussian norm $\|\Sigma^{-1/2}x_1\|_{\psi_2} = \kappa$, for some constant $\kappa > 0$.

Let $\hat{\theta}^n$ and $\hat{\sigma}$ be obtained by the scaled lasso, given by expression (10), with $\lambda = c\sqrt{\{\log(p)/n\}}$, and $c^2 > 48$. Consider an arbitrary $U \in \mathbb{R}^{p \times k}$, with $U^{\mathrm{T}}U = I$, that is independent of the samples $\{(x_i, y_i)\}_{i=1}^n$. Construct a debiased estimator $\hat{\gamma}^{\mathrm{d}}$ as in expression (11) with $\mu = a\sqrt{\{\log(p)/n\}}$ and $a^2 > 48e^2\kappa^4 C_{\max}/C_{\min}$. In addition, suppose that $\limsup_{n \to \infty} \mu(\max_{i \in [k]} \|u_i\|_1) \leqslant c'$, for some constant $0 < c' < 1$ and $s_0 = o\{\sqrt{n}/\log(p)\}$.

For the test R_X that is defined in equation (19), and for any $\alpha \in [0, 1]$, we have

$$\limsup_{n \to \infty} \alpha_n(R_X) \leqslant \alpha. \tag{24}$$

We refer to Appendix A for the proof of theorems 1 and 2.

3.2. Statistical power

We next analyse the statistical power of our test. Before proceeding, note that, without further assumption, we cannot achieve any non-trivial power, namely power of α which is obtained by a rule that randomly rejects the null hypothesis with probability α. Indeed, by choosing $\theta_0 \notin \Omega_0$ but arbitrarily close to Ω_0, one can make H_0 essentially indistinguishable from H_{A}. Taking this point into account, for a set $\Omega_0 \subseteq \mathbb{R}^p$ and $\theta_0 \in \mathbb{R}^p$, we define the distance $d(\theta_0, \Omega_0; U)$ as

$$d(\theta_0, \Omega_0; U) = \inf_{\theta \in \Omega_0} \|U^{\mathrm{T}}(\theta - \theta_0)\|_\infty. \tag{25}$$

We shall assume that, under the alternative hypothesis, $d(\theta_0, \Omega_0; U) \geqslant \eta$ as well. Define

$$\beta_n(R) \equiv \sup\{\mathbb{P}_{\theta_0}(R = 0) : \|\theta_0\|_0 \leqslant s_0(n), d(\theta_0, \Omega_0; U) \geqslant \eta\}. \tag{26}$$

Quantity β_n is the probability of type II error (false negative) and $1 - \beta_n$ is the statistical power of the test.

Theorem 3. Let R_X be the test that is defined in equation (19). Under the conditions of theorem 2, for all $\alpha \in [0, 1]$:

$$\liminf_{n \to \infty} \frac{1 - \beta_n(R_X)}{1 - \beta_n^*(\eta)} \geqslant 1, \qquad 1 - \beta_n^*(\eta) \equiv F\left(\alpha, \frac{\sqrt{n}\eta}{\hat{\sigma}m_0}, k\right)_+ \tag{27}$$

where we define m_0 as

$$m_0 \equiv \max_{i\in[k]} (u_i^{\mathrm{T}}\Sigma^{-1}u_i)^{1/2}. \tag{28}$$

Further, for $\alpha\in[0,1]$, $x\in\mathbb{R}_+$, and integer $k\geqslant 1$, the function $F(\alpha,x,k)$ is defined as follows:

$$F(\alpha,x,k) = 1-k\left[\Phi\left\{x+\Phi^{-1}\left(1-\frac{\alpha}{2k}\right)\right\}-\Phi\left\{x-\Phi^{-1}\left(1-\frac{\alpha}{2k}\right)\right\}\right]. \tag{29}$$

The proof of theorem 3 is given in Appendix A.3.

For any fixed $k\geqslant 1$ and $\alpha>0$, the function $x\mapsto F(\alpha,x,k)$ is continuous and monotone increasing, i.e. the larger $d(\theta_0,\Omega_0;U)$, the higher power is achieved. Also, to achieve a specific power $\beta>\alpha$, our scheme requires $\eta>c_\beta m_0(\sigma/\sqrt{n})$, for some constant c_β that depends on the desired power β. In addition, if $\eta\sqrt{n}\to\infty$, the rule achieves asymptotic power 1.

It is worth noting that, in the case of testing individual parameters $H_{0,i}:\theta_{0,i}=0$ (corresponding to $\Omega_0=\{\theta\in\mathbb{R}^p:\theta_i=0\}$ and $k=1$), we recover the power lower bound that was established in Javanmard and Montanari (2014b), which by comparing with the minimax trade-off that was studied in Javanmard and Montanari (2014a) is optimal up to a constant.

4. Choice of subspace U

Before we start this section, we stress again that, by theorems 1 and 2, the proposed testing rule controls the type I error below the desired level α, *for any choice of* $U\in\mathbb{R}^{p\times k}$, *with* $1\leqslant k\leqslant p$ *and* $U^{\mathrm{T}}U=I$ *that is independent of* X. Here, we provide guidelines for choosing U that yield high power. For this we use the result of theorem 3.

Note that

$$m_0 \leqslant \max_{i\in[k]}(C_{\min}^{-1}\|u_i\|^2)^{1/2} = C_{\min}^{-1/2},$$

where we recall that $\sigma_{\min}(\Sigma)>C_{\min}>0$ and $\|u_i\|=1$, for $i\in[k]$. Hence,

$$F\left\{\alpha,\frac{\sqrt{n}\,d(\theta_0,\Omega_0;U)}{\hat{\sigma}m_0},k\right\}\geqslant F\left\{\alpha,\frac{1}{\hat{\sigma}}\sqrt{(nC_{\min})}\,d(\theta_0,\Omega_0;U),k\right\}. \tag{30}$$

We propose to choose U by maximizing the right-hand side of inequality (30), which by theorem 3 serves as a lower bound for the power of the test. Nevertheless, the above optimization involves θ_0 which is unknown. To cope with this issue, we use the lasso estimate $\hat{\theta}$ via the following procedure.

Step 1: we randomly split the data (y,X) into two subsamples $(y^{(1)},X^{(1)})$ and $(y^{(2)},X^{(2)})$ each with sample size $n_0=n/2$. We let $\hat{\theta}^{(1)}$ be the optimizer of the scaled lasso applied to $(y^{(1)},X^{(1)})$.

Step 2: we choose $U\in\mathbb{R}^{p\times k}$ by solving the following optimization:

$$\underset{k\in[p],U\in\mathbb{R}^{p\times k},U^{\mathrm{T}}U=I}{\text{maximize}} \; F\left\{\alpha,\frac{1}{\hat{\sigma}}\sqrt{(nC_{\min})}\,d(\theta_0,\Omega_0;U),k\right\}. \tag{31}$$

Step 3: we construct the debiased estimator by using the data $(y^{(2)},X^{(2)})$. Specifically, set $\hat{\Sigma}^{(2)}\equiv(1/n_0)(X^{(2)})^{\mathrm{T}}(X^{(2)})$ and let g_i be the solution of the following optimization problems for each $1\leqslant i\leqslant k$:

$$\begin{aligned}&\text{minimize} \quad g^{\mathrm{T}}\hat{\Sigma}^{(2)}g\\ &\text{subject to} \quad \|\hat{\Sigma}^{(2)}g-u_i\|_\infty\leqslant\mu.\end{aligned} \tag{32}$$

Define the decorrelating matrix $G = [g_1|\ldots|g_k] \in \mathbb{R}^{p\times k}$ and let $\hat{\theta}^{(2)}$ be the optimizer of the scaled lasso applied to $(y^{(2)}, X^{(2)})$. Let

$$\hat{\gamma}^{\mathrm{d}} = U^{\mathrm{T}}\hat{\theta}^{(2)} + \frac{1}{n_0} G^{\mathrm{T}}(X^{(2)})^{\mathrm{T}}(y^{(2)} - X^{(2)}\hat{\theta}^{(2)}). \tag{33}$$

Step 4: set $Q \equiv (\hat{\sigma}^2/n)(G^{\mathrm{T}}\hat{\Sigma}^{(2)}G)$ and $D \equiv \mathrm{diag}(\{Q_{ii}\}^{-1/2})$. Find the l_∞-projection as

$$\theta^{\mathrm{p}} = \arg\min_{\theta\in\mathbb{R}^p} \|D(\hat{\gamma}^{\mathrm{d}} - U^{\mathrm{T}}\theta)\|_\infty \qquad \text{subject to } \theta \in \Omega_0. \tag{34}$$

Step 5: define the test statistics $T_n = \|D(\hat{\gamma}^{\mathrm{d}} - U^{\mathrm{T}}\theta^{\mathrm{p}})\|_\infty$. The testing rule is given by

$$R_X(y) = \begin{cases} 1 & \text{if } T_n \geqslant z_{\alpha/(2k)} \text{ (reject } H_0\text{)}, \\ 0 & \text{otherwise (fail to reject } H_0\text{)}. \end{cases} \tag{35}$$

Note that the data splitting above ensures that U is independent of $(y^{(2)}, X^{(2)})$, which is required for our analysis (see theorems 1–3.)

4.1. *Convex sets* Ω_0

When the set Ω_0 is convex, step 2 in the above procedure can be greatly simplified. Indeed, we can only focus on $k = 1$ in this case.

Lemma 3. Define the set $\mathcal{J}$ of matrices as

$$\mathcal{J} \equiv \arg\max_{U\in\mathbb{R}^{p\times k}} F\left\{\alpha, \frac{1}{\hat{\sigma}}\sqrt{(nC_{\min})}\, d(\hat{\theta}^{(1)}, \Omega_0; U), k\right\} \qquad \text{subject to } 1 \leqslant k \leqslant p, U^{\mathrm{T}}U = I_k. \tag{36}$$

If Ω_0 is convex then there is a unit norm $u^* \in \mathbb{R}^{p\times 1}$ such that $u^* \in \mathcal{J}$.

A proof of lemma 3 is given in the on-line appendix A.3.

Focusing on $k = 1$, optimization (31) reduces to the following optimization over $u \in \mathbb{R}^{p\times 1}$:

$$u \in \arg\max_{u\in\mathbb{R}^p, \|u\|_2=1} F\left\{\alpha, \frac{1}{\hat{\sigma}}\sqrt{(nC_{\min})}\, d(\hat{\theta}^{(1)}, \Omega_0; u), 1\right\}. \tag{37}$$

The function $x \mapsto F(\alpha, x, k)$ is monotone increasing in x and, by substituting for $d(\theta_0, \Omega_0; u)$, this becomes equivalent to the problem

$$\underset{u\in\mathbb{R}^p, \|u\|_2\leqslant 1}{\text{maximize}} \inf_{\theta\in\Omega_0} |u^{\mathrm{T}}(\theta - \hat{\theta}^{(1)})|. \tag{38}$$

Given that the objective is linear in u and θ, and the set Ω_0 is convex we can apply von Neumann's minimax theorem and change the order of max and min:

$$\inf_{\theta\in\Omega_0} \max_{u\in\mathbb{R}^p, \|u\|_2\leqslant 1} |u^{\mathrm{T}}(\theta - \hat{\theta}^{(1)})|. \tag{39}$$

Denote the orthogonal projection of $\hat{\theta}^{(1)}$ onto Ω_0 by $\mathcal{P}_{\Omega_0}(\hat{\theta}^{(1)}) = \arg\min_{\theta\in\Omega_0} \|\theta - \hat{\theta}^{(1)}\|_2$. Then it is straightforward to see that the optimal u is given by

$$u = \frac{\mathcal{P}^{\perp}_{\Omega_0}(\hat{\theta}^{(1)})}{\|\mathcal{P}^{\perp}_{\Omega_0}(\hat{\theta}^{(1)})\|}, \tag{40}$$

with $\mathcal{P}^{\perp}_{\Omega_0}(\hat{\theta}^{(1)}) = \hat{\theta}^{(1)} - \mathcal{P}_{\Omega_0}(\hat{\theta}^{(1)})$.

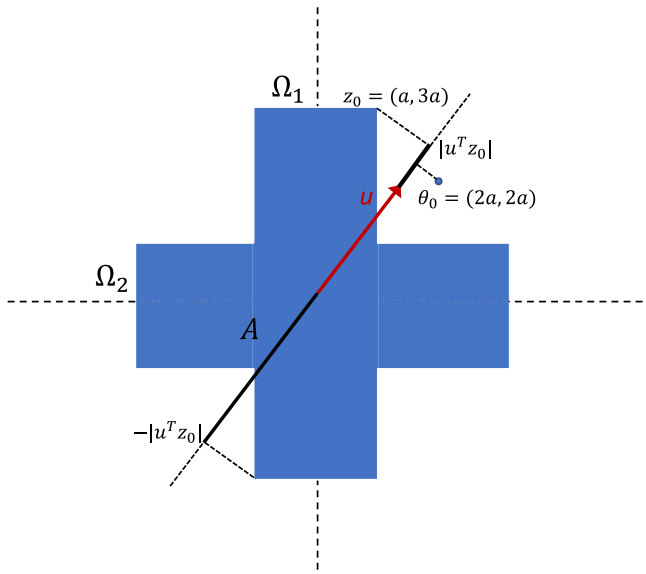

Fig. 1. Illustration of the example of non-convex Ω_0 discussed in remark 4 for $p = 2$

We remind again that the type I error is controlled at the desired level for any $U \in \mathbb{R}^{p\times k}$ with $U^{\mathrm{T}}U = I$ that is independent of (y, X). The choice of u in equation (40) is a guideline for increasing power in the case of convex sets Ω_0.

Remark 4. We stress again that the convexity assumption of set Ω_0 is crucial in deriving the recipe (40). To build further insight, we provide a concrete example of a non-convex Ω_0 and argue that $k = 1$ is not the right choice. Let $\Omega_0 = \Omega_1 \cup \Omega_2$, where $\Omega_i = \{x \in \mathbb{R}^p : |x_i| \leqslant a,\ |x_j| \leqslant 3a,\ \text{for } j \neq i\}$, for $i = 1, 2$ and a fixed constant $a > 0$. Let $\theta_0 = (2a, 2a, 0, \ldots, 0) \in \mathbb{R}^p$. Observe that Ω_0 is not convex and $\theta_0 \notin \Omega_0$. By choosing $k = p$ and $U = I_{p\times p}$, we have $d(\theta_0, \Omega_0; U) = a$ and hence our method achieves non-trivial power. However, we argue that, setting $k = 1$, our method cannot do better than random guessing. Specifically, we show that, for any vector $u \in \mathbb{R}^p$, we have $d(\theta_0, \Omega_0; u) = 0$. By symmetry, assume that $|u_1| \leqslant |u_2|$. Note that the point $z_0 = \pm(a\,\mathrm{sgn}(u_1), 3a\,\mathrm{sgn}(u_2), \ldots, 3a\,\mathrm{sgn}(u_p)) \in \Omega_1 \subset \Omega_0$. Further, $u^{\mathrm{T}}z_0 = \pm(a|u_1| + 3a|u_2| + \ldots + 3a|u_p|)$. By convexity of Ω_1, we have that $\mathcal{P}_u(\Omega_1) \supseteq A$ where $A = \{\alpha u : |\alpha| \leqslant |u^{\mathrm{T}}z_0|\}$. In addition, we have $u^{\mathrm{T}}\theta_0 = 2a(u_1 + u_2)$ and, using the assumption $|u_1| \leqslant |u_2|$, we obtain $|u^{\mathrm{T}}\theta_0| \leqslant |u^{\mathrm{T}}z_0|$. Therefore, $\mathcal{P}_u(\theta_0) \in A \subseteq \mathcal{P}_u(\Omega_1) \subset \mathcal{P}_u(\Omega_0)$. This implies that $d(\theta_0, \Omega_0; u) = 0$, meaning that we cannot do better than random guessing if the inference is done in the one-dimensional projected space. We refer to Fig. 1 for a schematic illustration of this example in $p = 2$.

5. Approximate sparsity

With the aim of broadening the application of our proposed method, we relax the sparsity assumption of the model to a so-called approximate sparsity structure. Consider the linear model

$$y = X\theta_* + w, \tag{41}$$

with $w \sim N(0, \sigma^2 I_{n\times n})$, and $\theta_* \in \mathbb{R}^p$ the unknown model parameters that are not necessarily sparse. However, we assume that there is at least one sparse linear combination of the covariates

that approaches close to the true signal. This is formally stated as the approximate sparsity that is stated below, which is similar to that introduced by Belloni *et al.* (2012).

Assumption 2 (approximately sparse model). The signal $X\theta_*$ is well approximated by a linear combination of unknown $s_0 \geqslant 1$ covariates:

$$X\theta_* = X\theta_0 + r, \qquad \|r\| = o_P(1). \tag{42}$$

The approximate sparsity assumption in Belloni *et al.* (2012) is weaker than the assumption that we are imposing here, as the former allows for $\|r\| = O_P(\sqrt{s_0})$. As shown in our analysis of the debiased estimator in theorem 4 below, assuming approximately sparse models instead of sparse models leads to an extra term $(1/\sqrt{n})G^{\mathrm{T}}X^{\mathrm{T}}r$ in the decomposition of $\sqrt{n}(\hat{\gamma}^{\mathrm{d}} - U^{\mathrm{T}}\theta_0)$. We control this term by a simple l_2–l_2-bound, namely

$$\left\| \frac{1}{\sqrt{n}} G^{\mathrm{T}} X^{\mathrm{T}} r \right\|_\infty \leqslant \left(\max_{i \in [k]} \left\| \frac{1}{\sqrt{n}} X g_i \right\| \right) \|r\|_2 = o_p(1),$$

where in the last step we used the approximate sparsity assumption. Although the l_2–l_2-bound could be conservative, it is difficult to improve it for general setting as the terms $G^{\mathrm{T}}X^{\mathrm{T}}$ and r both depend on X and hence are highly dependent.

The next assumption was also introduced by Belloni *et al.* (2012), under the name of the 'RF condition' for reduced form errors and regressors. This is basically an assumption on the moments of covariates and the noise component. In stating that we borrow the following empirical process notation from Belloni *et al.* (2012): $\mathbb{E}_n(f) \equiv \mathbb{E}_n\{f(z_i)\} \equiv \Sigma_{i=1}^n f(z_i)/n$ and $\bar{\mathbb{E}}(f) \equiv \mathbb{E}\{\mathbb{E}_n(f)\} = \mathbb{E}[\mathbb{E}_n\{f(z_i)\}] = \Sigma_{i=1}^n \mathbb{E}\{f(z_i)\}/n$.

Assumption 3 (moment condition). Suppose that the following moment conditions hold.

(a) For a constant $C_2 > 0$, $\bar{\mathbb{E}}(y_i^2) + \bar{\mathbb{E}}(X_{ij}^2 y_i^2) + 1/\bar{\mathbb{E}}(X_{ij}^2 w_i^2) \leqslant C_2$.
(b) We have $\max_{j \in [p]} \bar{\mathbb{E}}(|X_{ij}^3 w_i^3|) \leqslant o[\sqrt{\{n/\log(p)^3\}}]$, and also $s_0 \log(p) = o(n)$.
(c) $\max_{i \in [n], j \in [p]} X_{ij}^2 s_0 \log(p)/n \to 0$, in probability, and $\max_{j \in [p]} |(\mathbb{E}_n - \bar{\mathbb{E}})(X_{ij}^2 w_i^2)| + |(\mathbb{E}_n - \bar{\mathbb{E}})(X_{ij}^2 y_i^2)| \to 0$, in probability.

The above moment condition was proposed in Belloni *et al.* (2012) where they bounded the estimate error of selection methods such as the lasso under approximate sparsity conditions. Our lemma below provides a set of alternative conditions that, for sub-Gaussian designs, imply the moment condition 3.

Lemma 4. Suppose that the design X has independent sub-Gaussian centred rows with uniformly bounded sub-Gaussian norm ($\|x_i\|_{\psi_2} \leqslant C$). Assume that y_i and w_i have uniformly bounded conditional moments of order 4, i.e. $\mathbb{E}(y_i^4 | x_i) \leqslant C'$ and $\mathbb{E}(w_i^4 | x_i) \leqslant C''$, for $i \in [n]$. In addition, suppose that $s_0 = o\{n/\log^2(p)\}$ and $\log(p) = o(n^{1/3})$. Then the moment condition 3 holds.

We refer to the on-line appendix A.12 for the proof of lemma 4.

5.1. Iterated lasso

Following Belloni *et al.* (2012), we consider a weighted lasso estimator of θ_0. Formally, let $\hat{\theta}$ be given by

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^p} \left\{ \frac{1}{n} \|y - X\theta\|^2 + \lambda \sum_{j=1}^p |\tau_j \theta_j| \right\}, \tag{43}$$

where the regularization λ is chosen as

Table 1. Algorithm 1: choosing weights in the iterated lasso estimator

Input: response vector y, design matrix X, regularization parameter λ, number of iteration K
Output: estimator $\hat{\theta}$
1, (*initialization*) set $\tau_j = \sqrt{\mathbb{E}_n(X_{ij}^2 y_i^2)}$, for $j \in [p]$
2, *for* $k = 1, 2, \ldots, K$ *do*
3, compute estimator $\hat{\theta}$ given by estimator (43)
4, update the weights as $\tau_j = \sqrt{\mathbb{E}_n\{X_{ij}^2 (y_i - x_i^{\mathrm{T}}\hat{\theta})^2\}}$

$$\lambda = \frac{2.2}{\sqrt{n}} \Phi^{-1}\left\{1 - \frac{0.1}{2p\log(p)}\right\}. \tag{44}$$

The weights τ_j, for $j \in [p]$, are ideally chosen as $\tau_j = \sqrt{\mathbb{E}_n(X_{ij}^2 w_i^2)}$. But since the noise terms w_i are unobserved this ideal option is not realizable. Hence, we use an iterative method that was proposed in Belloni *et al.* (2012, 2014) to set the weights τ_j. (The resulting lasso estimator $\hat{\theta}$ is referred to as the 'iterated Lasso' in Belloni *et al.* (2012, 2014).) The details of the procedure are described in algorithm 1 (Table 1).

Our next theorem is analogous to theorem 2 and shows that our procedure controls the type I error for random designs under approximately sparse models.

Theorem 4. Let $\Sigma \in \mathbb{R}^{p\times p}$ such that $\sigma_{\min}(\Sigma) \geqslant C_{\min} > 0$ and $\sigma_{\max}(\Sigma) \leqslant C_{\max} < \infty$ and $\max_{i\in[p]} \Sigma_{ii} \leqslant 1$. Suppose that the regression model (41) is approximately sparse (assumption 2). Suppose that X has independent sub-Gaussian rows and the moment condition (assumption 3) or the alternative assumptions of lemma 4 hold.

Let $\hat{\theta}$ be the iterated lasso estimator using data (y, X), given by estimator (43). Consider an arbitrary $U \in \mathbb{R}^{p\times k}$, with $U^{\mathrm{T}}U = I$, that is independent of the samples $\{(x_i, y_i)\}_{i=1}^n$. Construct a debiased estimator $\hat{\gamma}^{\mathrm{d}}$ as in expression (11) with $\mu = a\sqrt{\{\log(p)/n\}}$, and $a^2 > 48e^2\kappa^4 C_{\max}/C_{\min}$. In addition, suppose that $\limsup_{n\to\infty} \mu(\max_{i\in[k]}\|u_i\|_1) \leqslant c'$, for some constant $0 < c' < 1$, $s_0 = o\{\sqrt{n}/\log(p)\}$ and $\log(p) = o(n^{1/3})$.

For the test R_X that is defined in equation (19), and for any $\alpha \in [0, 1]$, we have

$$\limsup_{n\to\infty} \alpha_n(R_X) \leqslant \alpha. \tag{45}$$

We refer to Appendix A.4 for the proof of theorem 4. We stress that the moment condition (assumption 3) or the alternative sufficient conditions that are stated in lemma 4 is needed only for the lasso l_1-estimation error bound developed in Belloni *et al.* (2012) under the approximate sparsity condition.

6. Extension to non-Gaussian heteroscedastic noise

Our analysis can be extended to the case of non-Gaussian heteroscedastic noise measurements. Specifically, suppose that the noise term w_i satisfies

$$\left.\begin{aligned} \mathbb{E}(w_i|X) &= 0, \\ \mathbb{E}(w_i^2|X) &= \sigma_i^2, \\ \mathbb{E}(|w_i|^{4+a}|X) &\leqslant B, \end{aligned}\right\} \tag{46}$$

for some constants $a, B > 0$, and $1 \leqslant i \leqslant n$.

Recall that our analysis is based on a bias–variance decomposition of the estimate $\hat{\gamma}^{\mathrm{d}}$ as in lemma 1. The bias term $\|\Delta\|_\infty$ can be bounded as

$$\|\Delta\|_\infty \leqslant \sqrt{n}\|G^{\mathrm{T}}\hat{\Sigma} - U\|_\infty \|\theta_0 - \hat{\theta}\|_1.$$

The first term does not involve the noise term w and can be treated as before. For bounding $\|\theta_0 - \hat{\theta}\|_1$, we used the result of Belloni *et al.* (2012), theorem 1 (see proposition 9.7 in the on-line appendix) that also applies to non-Gaussian heteroscedastic noise as long as the moment conditions (assumption 3) hold, which by lemma 4 for sub-Gaussian designs reduces to requiring that the noise variables w_i have bounded conditional moment of order 4.

So the remaining part is characterizing the limiting distribution of Z. For this, we shall show that the Lindeberg condition holds and hence Z admits an asymptotically normal distribution by virtue of the central limit theorem.

Similarly to the approach that was taken in Javanmard and Montanari (2014b), we slightly modify our construction of the decorrelating matrix G to ensure that the Lindeberg condition holds. Let $G = [g_1|\ldots|g_k] \in \mathbb{R}^{p\times k}$, where each g_i is obtained by solving the following optimization problems for each $1 \leqslant i \leqslant k$:

$$\begin{aligned} &\text{minimize} \quad g^{\mathrm{T}}\hat{\Sigma}g \\ &\text{subject to} \quad \|\hat{\Sigma}g - u_i\|_\infty \leqslant \mu \\ &\|Xg\|_\infty \leqslant n^\beta, \qquad \text{for arbitrary fixed } 0 < \beta < \tfrac{1}{2}. \end{aligned} \tag{47}$$

Our following proposition shows that Z admits an asymptotically normal distribution in the non-Gaussian setting.

Proposition 1. Suppose that the noise variables w_i are independent with $\mathbb{E}(w_i|X) = 0$, $\mathbb{E}(w_i^2|X) = \sigma_i^2$ and $\mathbb{E}(|w_i|^{4+a}|X) \leqslant B$ for some $a > 4\beta/(1-2\beta)$, such that $s_n^2 \equiv (1/n)\Sigma_{i=1}^n \sigma_i^2$ is bounded away from 0 and from above uniformly in n. Let $G = [g_1|\ldots|g_k] \in \mathbb{R}^{p\times k}$ be the matrix that is constructed by solving optimization problem (47). For $i \in [p]$, define

$$Z_i = \frac{1}{\sqrt{n}} \frac{g_i^{\mathrm{T}} X^{\mathrm{T}} w}{s_n (g_i^{\mathrm{T}} \hat{\Sigma} g_i)^{1/2}}. \tag{48}$$

Suppose that the assumptions of theorem 4 hold. Then, for any sequence $i = i(n) \in [p]$, and any $x \in \mathbb{R}$, we have

$$\lim_{n\to\infty} \mathbb{P}(Z_i \leqslant x|X) = \Phi(x), \tag{49}$$

with $\Phi(x)$ indicating the cumulative distribution function of a standard normal variable.

We refer to the on-line appendix A.4 for the proof of proposition 1.

The distributional characterization (49) involves the unknown quantity s_n. We next propose an estimator of s_n that can be used in lieu of s_n in equation (48) and still have result (49) in place. Denote by $\hat{\theta}$ the iterated lasso estimator given by expression (43) and define

$$\hat{s}_n^2 = \frac{1}{n}\sum_{i=1}^n (y_i - \langle x_i, \theta_* \rangle)^2. \tag{50}$$

Our next lemma shows that $\hat{s}_n$ is a consistent estimator of s_n.

Lemma 5. Consider regression model (41) along with assumption 2 (approximate sparsity)

and assumption 3 (moment conditions). Consider the estimator $\hat{s}_n$ that is given by equation (50). If $s_0 = o\{\sqrt{n}/\log(p)\}$, then $\lim_{n\to\infty}(\hat{s}_n - s_n) = 0$, with high probability.

The proof of lemma 5 is deferred to the on-line appendix A.5.

7. Discussion

It is useful to study the proposed methodology for some specific choices of Ω_0 and to discuss its optimality.

7.1. Example 1 (predictions)

Fix an arbitrary $c \in \mathbb{R}$ and consider the set $\Omega_0 = \{\theta : \xi^{\mathrm{T}}\theta = c\}$. This corresponds to the set where the (noiseless) unobserved response on the new feature vector ξ is c. We can use our methodology to test $H_0 : \theta_0 \in \Omega_0$ *versus* its alternative. Further, by duality of hypothesis testing and confidence intervals, our methodology provides confidence intervals for a linear functional of the form $\xi^{\mathrm{T}}\theta_0$.

Computing u from expression (40) in this case gives $u = \xi/\|\xi\|$. Since ξ is independent of (y, X), the data splitting step in the procedure becomes superfluous. By duality, we construct $1-\alpha$ confidence intervals for $\xi^{\mathrm{T}}\theta_0$ by finding the range of values c such that the rule fails to reject H_0 at level α. This is formalized in the next lemma.

Lemma 6. Consider a sequence of design matrices $X \in \mathbb{R}^{n\times p}$, with dimensions $n, p \to \infty$ and $p = p(n) \to \infty$ satisfying the assumptions of theorem 1. For given $\alpha \in (0, 1)$, define $C(\alpha) = [c_{\min}, c_{\max}]$ with

$$c_{\min} = \|\xi\|\hat{\gamma}^{\mathrm{d}} - \frac{\hat{\sigma}}{\sqrt{n}}\sqrt{(g^{\mathrm{T}}\hat{\Sigma}g)}\, z_{\alpha/2}\|\xi\|_2, \tag{51}$$

$$c_{\max} = \|\xi\|\hat{\gamma}^{\mathrm{d}} + \frac{\hat{\sigma}}{\sqrt{n}}\sqrt{(g^{\mathrm{T}}\hat{\Sigma}g)}\, z_{\alpha/2}\|\xi\|_2, \tag{52}$$

where $\hat{\gamma}^{\mathrm{d}}$ is the debiased estimator given by expression (33) with $u = \xi/\|\xi\|$. Then,

$$\liminf_{n\to\infty} \mathbb{P}\{\xi^{\mathrm{T}}\theta_0 \in C(\alpha)\} \geqslant 1-\alpha. \tag{53}$$

We refer to the on-line appendix A.6 for the proof of lemma 6. The confidence interval constructed has length of rate $\|\xi\|/\sqrt{n}$. In Cai and Guo (2017) it was shown that the minimax expected length of confidence intervals for $\xi^{\mathrm{T}}\theta_0$, with a sparse vector ξ (i.e. $\|\xi\|_0 = O(s_0)$) is $\|\xi\|\{1/\sqrt{n} + s_0\log(p)/n\}$. Therefore, in the regime $s_0 = o\{\sqrt{n}/\log(p)\}$, which is the focus of the current paper, the confidence intervals constructed are minimax rate optimal. It is worth noting that the confidence interval that is defined in lemma 6 is similar to that proposed by Cai and Guo (2017). For the case of non-sparse ξ, Cai and Guo (2017) established the minimax rate $\|\xi\|_\infty s_0\sqrt{\{\log(p)/n\}}$ for the expected length of confidence interval for $\xi^{\mathrm{T}}\theta_0$, and hence our construction (51) has an optimality gap in this case.

7.2. Example 2 (quadratic forms)

As another example we apply our framework to testing the squared l_2-norm of θ_0. Consider the set $\Omega_0(c) = \{\theta : \|\theta\|_2^2 = c\}$, where $c \geqslant 0$ is a fixed arbitrary constant. We use the framework proposed to test the null hypothesis $H_0 : \theta_0 \in \Omega_0(c)$. Computing u from expression (40) in this case gives $u = \hat{\theta}^{(1)}/\|\hat{\theta}^{(1)}\|$. We next use the duality between hypothesis testing and confidence intervals to construct confidence intervals for $\|\theta_0\|_2^2$.

Lemma 7. Consider a sequence of design matrices $X \in \mathbb{R}^{n\times p}$, with dimensions $n, p \to \infty$ and

$p = p(n) \to \infty$ satisfying the assumptions of theorem 2. For given $\alpha \in (0,1)$, define $C(\alpha) = [c_{\min}, c_{\max}]$ with

$$\begin{aligned} c_{\min} &= (2\hat{\gamma}^{\mathrm{d}}\|\hat{\theta}^{(1)}\| - \|\hat{\theta}^{(1)}\|^2 - L)_+, \\ c_{\max} &= (2\hat{\gamma}^{\mathrm{d}}\|\hat{\theta}^{(1)}\| - \|\hat{\theta}^{(1)}\|^2 + L), \end{aligned} \tag{54}$$

$$L = \|\hat{\theta}^{(1)}\|\sqrt{(g^{\mathrm{T}}\hat{\Sigma}g)}\{1+o(1)\}\frac{\hat{\sigma}z_{\alpha/2}}{\sqrt{n}}, \tag{55}$$

where $a_+ = \max(a, 0)$ and $\hat{\gamma}^{\mathrm{d}}$ is the debiased estimator given by expression (33) with $u = \hat{\theta}^{(1)}/\|\hat{\theta}^{(1)}\|$. Then,

$$\liminf_{n\to\infty} \mathbb{P}\{\|\theta_0\|_2^2 \in C(\alpha)\} \geqslant 1-\alpha. \tag{56}$$

We give the proof of lemma 7 in the on-line appendix A.7.

7.3. Example 3 (testing $\theta_{\min}$-condition)

For a given $c > 0$, define the set $\Omega_0 = \{\theta \in \mathbb{R}^p : \min_{j\in\mathrm{supp}(\theta)} |\theta_j| \geqslant c\}$. Apart from the importance of this example as discussed in Section 1, it differs from the previous examples in that the set Ω_0 is non-convex and disconnected. Recall that guideline (40) was provided for convex sets Ω_0, which is not true in this example.

Before proposing a choice of U for this example, we state a lemma.

Lemma 8. Let $v \in \mathbb{R}^p$ and define $\theta \in \mathbb{R}^p$ with $\theta_i = \mathcal{S}(v_i, c)$, where

$$\mathcal{S}(x,c) = \begin{cases} x & |x| \geqslant c, \\ c & x \in (c/2, c), \\ 0 & x \in [-c/2, c/2], \\ -c & x \in (-c, -c/2). \end{cases} \tag{57}$$

Then θ is a solution to $\min_{\theta\in\mathbb{R}^p} \|D(v-\theta)\|_\infty$, subject to $\theta \in \Omega_0$, for any diagonal matrix D.

A proof of lemma 8 is straightforward and so has been omitted.

In the numerical experiments, we apply our framework for this example with $k = 1$ and $U = u \in \mathbb{R}^p$ given by

$$u = e_{i^*}, \qquad i^* \equiv \arg\max_{i\in[p]} |\hat{\theta}_i^{(1)} - \mathcal{S}(\hat{\theta}_i^{(1)}, c)|. \tag{58}$$

We refer to the on-line appendix A.8 for a justification for this choice. By using lemma 8, the test statistic in this case amounts to $T_n = |d\{\hat{\gamma}^{\mathrm{d}} - \mathcal{S}(\hat{\gamma}^{\mathrm{d}}, c)\}|$ (see step 5 of the algorithm that was presented in Section 4).

7.4. Prior art

The inference problem (3) that is studied in this paper is very general and encompasses several important problems such as the examples that were discussed in Section 1.1. For specific choices of set Ω_0, we may use the structure of the set Ω_0 to come up with methods with higher statistical power. However, in what follows we argue that, for three classes of inferential problems, our proposed framework either recovers the previously proposed methods for that specific problem or has comparable performance. We also contrast the underlying assumptions of our framework and those of other methods that have been designed for these specialized problems.

7.4.1. Inference on prediction

As discussed in Section 7.1, for inference on linear functions $\gamma_0 = \xi^{\mathrm{T}}\theta_0$ (predictions), our framework proposes $u = \xi/\|\xi\|$ and we construct a debiased estimator of γ_0 taking the form

$$\hat{\gamma}^{\mathrm{d}} = \frac{\xi^{\mathrm{T}}}{\|\xi\|}\hat{\theta} + \frac{1}{n} g^{\mathrm{T}} X^{\mathrm{T}}(y - X\hat{\theta}), \tag{59}$$

with g obtained by solving optimization problem (12). As argued for the case of random designs with population covariance Σ, this implies that $g \approx \Sigma^{-1}\xi/\|\xi\|$. As also discussed earlier in Section 1 and the previous section, a similar approach has been used by Cai and Guo (2017) and they proved that the resulting confidence interval would be minimax rate optimal. It is indeed an appealing property of our method that, despite its generality, it recovers the method of Cai and Guo (2017) for this specific case and enjoys minimax optimality.

7.4.1.1. Assumptions. In terms of assumptions, Cai and Guo (2017) focused on high dimensional linear models with Gaussian designs (rows of design matrix are drawn IID from a multivariate normal distribution), sparse parameter vector and Gaussian measurement noise. Our analysis in Section 3 considers sub-Gaussian random designs (theorem 2) and coherent fixed designs (theorem 1). We also extended our analysis to *approximately* sparse models (Section 5) and non-Gaussian noise (Section 6).

7.4.1.2. Least favourable one-dimensional submodel. It is worth noting that the form of debiasing (59) for linear functionals of θ can also be derived from the perspective of least favourable scores that were discussed in Zhang and Zhang (2014). Akin to the semiparametric models, consider the one-dimensional submodel $\{\theta_0 + u\phi, |\phi| < \varepsilon_*\}$ with $\varepsilon_* \to 0$, ϕ scalar and $u \in \mathbb{R}^p$. By imposing the constraint $\xi^{\mathrm{T}}u = 1$, we have $\xi^{\mathrm{T}}(\theta_0 + u\phi) - \xi^{\mathrm{T}}\theta_0 = \phi$. The idea of Zhang and Zhang (2014) is to look for the least favourable submodels at θ_0, given by $\theta_0 + u\phi$ with u_0 the direction that minimizes Fisher's information. For the log-likelihood $l_i(\theta_0) = l(\theta_0|y_i, x_i)$, recall that the Fisher information operator at θ is defined as $F = -\mathbb{E}\{\ddot{l}_i(\theta)\}$ and, for linear regression with Gaussian errors, we have $F = (1/\sigma)^2\mathbb{E}(x_i x_i^{\mathrm{T}}) = (1/\sigma)^2\Sigma$. The least favourable direction in the submodel is then given by

$$u_0 = \arg\min_u \{u^{\mathrm{T}}\Sigma u : \xi^{\mathrm{T}}u = 1\} = \Sigma^{-1}\xi/(\xi^{\mathrm{T}}\Sigma^{-1}\xi).$$

Following Zhang and Zhang (2014), we can construct a low dimensional projection estimator as a one-step maximum likelihood correction of $\hat{\theta}$ in the direction of the least favourable submodel u as follows:

$$\begin{aligned}
\hat{\gamma}^{\mathrm{d}} &= \xi^{\mathrm{T}}\hat{\theta} + \arg\max_{\phi\in\mathbb{R}} \sum_{i=1}^{n} l_i(\hat{\theta} + u\phi) \\
&= \xi^{\mathrm{T}}\hat{\theta} + \frac{u^{\mathrm{T}}X^{\mathrm{T}}(y - X\hat{\theta})}{\|Xu\|^2} = \xi^{\mathrm{T}}\hat{\theta} + \frac{\xi^{\mathrm{T}}\Sigma^{-1}\xi}{\|X\Sigma^{-1}\xi\|^2}\,\xi^{\mathrm{T}}\Sigma^{-1}X^{\mathrm{T}}(y - X\hat{\theta}) \\
&\approx \xi^{\mathrm{T}}\hat{\theta} + \frac{1}{n}\xi^{\mathrm{T}}\Sigma^{-1}X^{\mathrm{T}}(y - X\hat{\theta}),
\end{aligned} \tag{60}$$

where in the last step we replaced the denominator by its expectation. Comparing expression (60) with equation (59) we see that (up to a normalization by $\|\xi\|$) they are the same if $g = \Sigma^{-1}\xi$. However, Σ is unknown in general and optimization problem (12) tries to find $g \approx \Sigma^{-1}\xi$ that also minimizes the variance of the debiased estimator obtained.

7.4.1.3. Choice of k and effect of sample splitting. Our procedure uses sample splitting to find the best subspace U for the sake of statistical power. On one side, the sample splitting incurs a loss in power as we are using only half of the data points. On the other side, the purpose of sample splitting was to choose U to increase the power. To understand this trade-off we consider the following inference problem. Consider a function $h : \mathbb{R}^p \mapsto \mathbb{R}^q$ defined as $h(\theta) = (\xi_1^{\mathrm{T}}\theta, \ldots, \xi_q^{\mathrm{T}}\theta)$,

for a linearly independent set $\{\xi_1, \ldots, \xi_q\}$. The goal is to do inference on the value of $h(\theta_0)$. We consider the following two methods of choosing U in constructing the debiased estimator.

(a) *Method 1*: we let $k = q$ and U be a basis for the space that is spanned by $\{\xi_1, \ldots, \xi_q\}$. This method does not require any sample splitting.
(b) *Method 2*: define $\Omega_0 = \{\theta : h(\theta) = c\}$, for a given $c > 0$. Since $\Omega_0(c)$ is convex, our methodology sets $k = 1$ and chooses u as in expression (40). Here we require sample splitting for $q \geqslant 2$ (see Section 4.1).

Note that the two methods become identical for $q = 1$. We next compare (the analytical lower bound on) the statistical power of these two methods for choosing U. Let $\eta_u = d(\hat{\theta}, \Omega_0; u)$ and $\eta_u = d(\hat{\theta}, \Omega_0; U)$, with u given by expression (40) and U a basis for the space $\{\xi_1, \ldots, \xi_q\}$. Using theorem 3 and equation (30), the lower bound for the power of method 1 and method 2 are respectively given by $F\{\alpha, 1/\hat{\sigma}\sqrt{(nC_{\min})}\eta_U, q\}$ and $F\{\alpha, 1/(\sqrt{2}\hat{\sigma})\sqrt{(nC_{\min})}\eta_u, 1\}$. Furthermore, by equation (92) in the on-line supplementary material we have $\eta_u \geqslant \eta_U$ and, since $F(\alpha, x, k)$ is increasing in x, we obtain

$$F\left\{\alpha, \frac{1}{\sqrt{2}\hat{\sigma}}\sqrt{(nC_{\min})}\eta_u, 1\right\} \geqslant F\left\{\alpha, \frac{1}{\sqrt{2}\hat{\sigma}}\sqrt{(nC_{\min})}\eta_U, 1\right\}.$$

In summary, we have

$$\begin{aligned} \liminf_{n\to\infty} \frac{\text{power}_1(n)}{F\{\alpha, 1/\hat{\sigma}\sqrt{(nC_{\min})}\eta_U, q\}} &\geqslant 1, \\ \liminf_{n\to\infty} \frac{\text{power}_2(n)}{F\{\alpha, 1/(\sqrt{2}\hat{\sigma})\sqrt{(nC_{\min})}\eta_U, 1\}} &\geqslant 1. \end{aligned} \tag{61}$$

These lower bounds nicely capture the trade-off between the choice of k and the sample splitting. The function $F(\alpha, x, k)$ is decreasing in k which supports the use of $k = 1$, but the function is increasing in x and hence decreases under sample splitting. To understand this trade-off we basically need to compare $F(\alpha, x, 1)$ and $F(\alpha, \sqrt{2}x, q)$, with $x = 1/(\sqrt{2}\hat{\sigma})\sqrt{(nC_{\min})}\eta_U$. In Fig. 2, we have plotted these curves for $\alpha = 0.05$ and several values of q. As we see for small values of signal strength x, method 2 ($k = 1$ and sample splitting) has higher statistical power, whereas, for larger signal strength x, method 1 ($k > 1$ and no sample splitting) prevails.

7.4.2. Inference on quadratic forms of parameters

Janson *et al.* (2017) proposed *EigenPrism*, which is a procedure to construct two-sided confidence intervals for the signal squared magnitude $\|\theta_0\|^2$. An appealing property of this procedure is that, albeit its applicability to the high dimensional setting ($p > n$), it does not make any assumption on the coefficient sparsity. However, it is theoretically justified only for standard Gaussian designs where $X_{ij} \sim N(0, 1)$, independently. As explained in Janson *et al.* (2017) this assumption is crucial because it ensures that, in the singular value decomposition of $X = UDV^{\mathrm{T}}$, the columns of V are uniformly distributed on the unit sphere, and hence enables computing the expectation and variance of inner products of columns of V with θ_0, which constitutes a main building component of Eigenprism. By contrast, our procedure (when specialized to inference on quadratic forms of parameters as discussed in Section 7, example 2) applies to a much broader family of sub-Gaussian random designs but assumes the coefficient sparsity $s_0 = o\{\sqrt{n}/\log(p)\}$.

In the limit $n, p \to \infty$ and $n/p \to \gamma \in (0, 1)$, the length of confidence intervals that are constructed by Eigenprism for $\|\theta_0\|^2$ works out at $C_\gamma(\|\theta_0\|^2 + \sigma^2)z_{\alpha/2}/\sqrt{n}$, with C_γ a numerical constant defined based on the Marcenko–Pastur distribution with parameter γ. By compari-

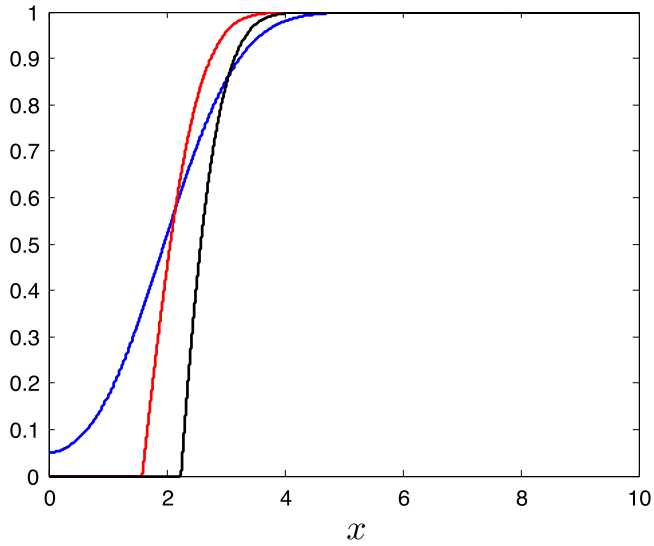

Fig. 2. Plot of $F(\alpha, x, 1)$ (——) and $F(\alpha, \sqrt{2}x, q)$ for $q = 2$ (——), 4 (——) and $\alpha = 0.05$

son, using lemma 7, the confidence interval that is obtained by our method is of length

$$2L < \frac{2z_{\alpha/2}}{\sqrt{C_{\min}}} \|\hat{\theta}^{(1)}\| \frac{\sigma}{\sqrt{n}}.$$

As we see, the length of confidence intervals for $\|\theta_0\|^2$ from both methods scale at rate $1/\sqrt{n}$.

7.4.3. Inference on individual parameters

As discussed in Section 1.1, for the special case of inference on an individual model parameter, our approach recovers the debiasing method of Javanmard and Montanari (2014b). A similar debiasing approach (with different construction of the decorrelating matrix, using nodewise regression) was proposed in Zhang and Zhang (2014) and Van de Geer *et al.* (2014) and its validity is proved under the assumption that the precision matrix Σ^{-1} is sparse. Belloni *et al.* (2014) proposed a significantly different approach for doing inference on an individual parameter, called 'post-double selection'. Suppose that we are interested in parameter θ_i. This method consists of two selection steps.

(a) Let I_1 be the covariates selected by the lasso in regressing columns i of the design matrix on the other columns.
(b) Let I_2 denote the covariates that are selected by the lasso in regressing y on the design X.

The estimation of parameter θ_i is then defined as the least squares estimator obtained by regression y on x_i and the selected features $I_1 \cup I_2$ (we may expand this set to include other features also that the statistician thinks are relevant). It is shown that the post-double estimator obeys an asymptotically normal distribution.

The limiting distribution of the post-double estimator is characterized under approximate sparsity structure and also applies to non-Gaussian noise as well, as far as some moment conditions (similar to assumption 3) hold. We stress that the approximate sparsity assumption in Belloni *et al.* (2014) is much weaker than ours in that it allows for $\|r\| = O_P(\sqrt{s_0})$, whereas we require $\|r\| = o_P(1)$.

8. Numerical illustration

In this section, we examine the performance of our inference framework in terms of coverage rate and length of confidence intervals, type I error and statistical power under various set-ups. We consider linear model (2) where the design matrix $X \in \mathbb{R}^{n\times p}$ has IID rows generated from $N(0,\Sigma)$, with $\Sigma \in \mathbb{R}^{p\times p}$ being the Toeplitz matrix $\Sigma_{i,j} = \rho^{|i-j|}$. For coefficient parameter θ_0, we consider a uniformly random support (set of non-zero parameters) $S \subseteq [p]$, with $|S| = s_0$. The measurement errors are $w_i \sim N(0,1)$.

8.1. Testing $\theta_{\min}$-condition

We consider the set $\Omega_0 = \{\theta : \min_{j\in \mathrm{supp}(\theta_0)} |\theta_{0,j}| \geqslant c\}$ and the null hypothesis $H_0 : \theta_0 \in \Omega_0$. As explained in Section 7 (example 3), the set Ω_0 is non-convex (indeed is disconnected) and we consider a one-dimensional projection of the problems along the direction u given by expression (58) for this example. For the scaled lasso estimator $\hat{\theta}^n$, given by expression (10), we set the regularization parameter $\lambda = \sqrt{\{2.05\log(p)/n\}}$. Further, the parameter μ in constructing the debiased estimator (see optimization problem (12)) is set to $\mu = 2\sqrt{\{\log(p)/n\}}$. We set $p = 1000$, $n = 600$ and $s_0 = 10$. The non-zero parameters $\theta_{0,i}$, $i \in S$, are chosen as $0.1, 0.2, \ldots, 1$. We set $\alpha = 0.05$ and vary the values of c and ρ. The rejection probabilities are computed based on 300 random samples for each value of pair (c,ρ). When $c \leqslant 0.1$, H_0 holds and thus the rejection probability corresponds to the type I error. When $c > 0.1$, the rejection probability corresponds to the power of the test. The results are reported in Table 2. As we see in Table 2, part (a), the type I error is controlled below the desired level $\alpha = 0.05$. Also, as evident in Table 2, part (b), the power of our test increases at a very fast rate as c increases.

8.2. Confidence intervals for linear functions

We use our methodology to construct 95% confidence intervals for functions of the form $\xi^{\mathrm{T}}\theta_0$. We set $p = 3000$ and $s_0 = 30$ and choose the correlation parameter $\rho = 0.5$. The model parameters are set as follows. We set $\theta_{0,j} = 0.5$ for $j = 1, \ldots, s_0$, and $\theta_{0,j} = 0.5/(j - s_0 + 1)$, for $j = s_0 + 1, \ldots, p$.

Table 2. Type I error and statistical power for H_0 : $\min_{j\in \mathrm{supp}(\theta_0)} |\theta_{0,j}| \geqslant c$, for level of significance $\alpha = 0.05$

c	*Results (%) for the following values of* ρ*:*			
	0.2	*0.4*	*0.6*	*0.8*
(a) Type I error				
0.02	0.00	0.004	1.33	2.33
0.04	0.33	1.66	2.33	3.00
0.06	1.66	2.00	3.00	3.66
0.08	3.33	4.33	3.66	4.66
0.1	3.00	4.00	4.66	4.33
(b) Statistical power				
0.2	8.00	10.66	18.66	14.33
0.3	17.33	24.66	28.66	35.33
0.4	86.00	93.33	92.66	84.66
0.5	90.00	88.00	97.33	86.66
0.6	100.00	88.33	100.00	100.00

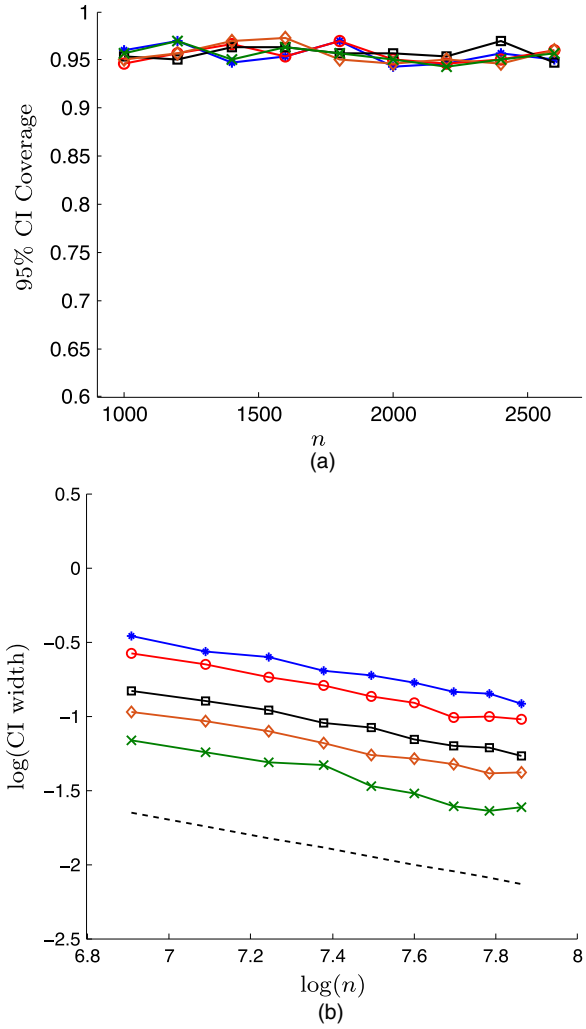

Fig. 3. (a) Coverage of 95% confidence intervals (51) for linear functions $\langle \xi, \theta_0 \rangle$ *versus* sample size n and (b) confidence interval widths *versus* sample size n (here $p = 3000$, $s_0 = 30$ and $\rho = 0.5$, and the model parameters are approximately sparse as described in Section 8.2): $\ast$, ξ_1; $\circ$, ξ_2; $\square$, ξ_3; $\diamond$, ξ_4; $\times$, ξ_5; - - - - -, $-0.5 \log(n)$

We construct confidence intervals according to lemma 6. We choose five vectors $\xi_1, \xi_2, \ldots, \xi_5$ as eigenvectors of Σ with well-separated eigenvalues. Specifically, sorting the eigenvalues of Σ as $\sigma_1 \geqslant \sigma_2 \geqslant \ldots \geqslant \sigma_{3000}$, we choose the eigenvectors corresponding to σ_1, σ_{750}, σ_{1500}, σ_{2250} and σ_{3000}. For each ξ_i, we vary n in $\{1000, 1200, 1400, \ldots, 2600\}$. For each configuration (ξ_i, n), we consider 300 independent realizations of measurement noise and, on each realization, we construct 95% confidence intervals for $\xi_i^{\mathrm{T}} \theta_0$ based on lemma 6.

In Fig. 3(a), we plot the average coverage probability of constructed confidence intervals for each configuration. Each curve corresponds to one of the vectors ξ_i. As we see, the coverage probability for all of them and across different values of n is close to the nominal value.

In Fig. 3(b), we plot the average length of confidence intervals as we vary the sample size n in the log–log-scale. As is evident from the Fig. 3(b), the length of confidence intervals scales as $1/\sqrt{n}$.

Table 3. Testing in the non-negative cone, $(n, s_0, p) = (600, 10, 1000)$†

b	*Results (%) for the following values of* ρ*:*			
	0.2	*0.4*	*0.6*	*0.8*
(a) Type I error				
1	2.00	2.00	2.00	3.33
0.8	0.66	2.33	2.33	2.66
0.6	3.00	3.66	1.00	2.66
0.4	2.66	2.33	1.33	2.00
0.2	2.33	1.66	2.33	3.66
(b) Statistical power				
−0.2	35.33	68.00	78.00	80.00
−0.4	99.33	100.00	100.00	100.00
−0.6	100.00	100.00	100.00	100.00
−0.8	100.00	100.00	100.00	100.00
−1	100.00	100.00	100.00	100.00

†The non-zero entries have magnitude b, and the covariance $\Sigma_{ij} = \rho^{|i-j|}$.

Table 4. Coverage rate of the confidence intervals for $\xi^{\mathrm{T}}\theta_0$ and $\|\theta_0\|_2^2$ computed as in equation (62) for the real data experiment and at various noise levels σ

σ	$\xi^{\mathrm{T}}\theta_0$	$\|\theta_0\|_2^2$
1	0.96	0.95
5	0.94	0.93
10	0.93	0.94

8.3. Testing for the non-negative cone

Define $\Omega_0 = \{\theta : \theta_i \geqslant 0 \text{ for all } i\}$ as the non-negative cone. In this section, we test $\theta_0 \in \Omega_0$ *versus* $\theta_0 \notin \Omega_0$. The null model is generated as follows. The non-zero entries in support S are chosen as $b, b/2, b/3, \ldots, b/s_0$, where $s_0 = |S|$ and $b > 0$. The entries outside S are set to 0. The alternative model is generated similarly where b is replaced by $-b$. As in the previous sections, the design matrix $X \in \mathbb{R}^{n \times p}$ has IID rows generated from $N(0, \Sigma)$, with $\Sigma \in \mathbb{R}^{p \times p}$ being the Toeplitz matrix $\Sigma_{i,j} = \rho^{|i-j|}$ and measurement errors $w_i \sim N(0, 1)$, with parameters $(n, s_0, p) = (600, 10, 1000)$. We set $\alpha = 0.05$ and vary the values of b and ρ. The rejection probabilities are computed based on 300 random samples for each value of the pair (b, ρ).

The simulation reported in Table 3 shows that the type I error is controlled below the target level $\alpha = 0.05$. Per statistical power, the method achieves power at least 99% for $|b| \geqslant 0.4$. Note that we have a very difficult alternative in the sense that only a small fraction of the co-ordinates, s_0/d, is negative with small magnitudes ranging in $[b/10, b]$, so it is a very mild violation of the null, yet our algorithm still has high power.

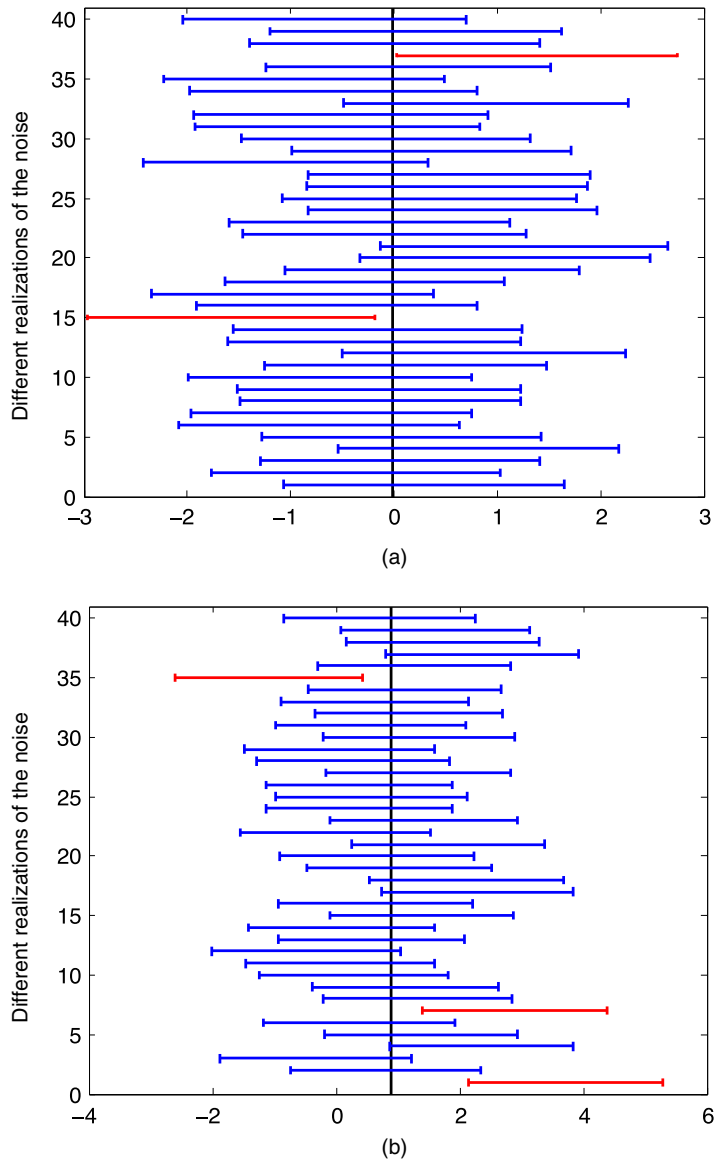

Fig. 4. 95% confidence intervals for (a) $\xi^{\mathsf{T}}\theta_0$ and (b) $\|\theta_0\|_2^2$ for the riboflavin data set: |, value of $\xi^{\mathsf{T}}\theta_0$ and $\|\theta_0\|_2^2$; ⊢—⊣, confidence interval covering the true value ⊢—⊣, confidence interval not covering the true value

8.4. Real data experiment

We measure the performance of our testing procedure on a riboflavin data set, which is publicly available from Bühlmann *et al.* (2014) and can be downloaded via the `hdi` R package. The data set includes $p = 4088$ predictors corresponding to the genes and $n = 71$ samples. The response variable indicates the logarithm of the riboflavin production rate and the covariates are the logarithm of the expression levels of the genes. We model the riboflavin production rate by a linear model. We first fit the lasso solution $\hat{\theta}$ by using the `glmnet` package (Friedman *et al.*, 2010)

and then generate $N = 100$ instances of the problem as $y^{(i)} = X\hat{\theta} + w^{(i)}$, where $w^{(i)} \sim N(0, \sigma^2 I_n)$. In other words, we treat $\hat{\theta}$ as the true parameter θ_0 and generate new data by resampling the noise.

We run two sets of experiments on these data.

8.4.1. Confidence interval for predictions

We fix a vector $\xi \in \mathbb{R}^p$ that is generated as $\xi_i \sim N(0, 1/\sqrt{p})$, independently for $i \in [p]$. On each problem instance (i), we construct confidence intervals $\mathrm{CI}^{(i)}$ for $\xi^{\mathrm{T}}\theta_0$, by using lemma 6. We compute the coverage rate as

$$\mathrm{Cov} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(\xi^{\mathrm{T}}\theta_0 \in \mathrm{CI}^{(i)}). \tag{62}$$

8.4.2. Confidence interval for squared norm

On each problem instance (i), we construct confidence intervals for $\|\theta_0\|_2^2$ by using lemma 7 and compute the coverage rate given by equation (62).

The results are reported in Table 4. As we see for various values of noise standard deviation σ, the coverage rates of the intervals constructed remain close to the nominal value. In Fig. 4, we depict the constructed confidence intervals for 40 random problem instances, in each experiment.

Acknowledgements

A. Javanmard was partially supported by an 'Outlier research in business' grant from the University of Southern California Marshall School of Business, a Google faculty research award and National Science Foundation Career award DMS-1844481. J. Lee acknowledges support of the Army Research Office under multidisciplinary university research initiative award W911NF-11-1-0303, a Sloan Research Fellowship and National Science Foundation grant CCF 2002272.

Appendix A: Proof of theorems

A.1. Proof of theorem 1

We first prove a lemma to bound the estimation error of $\hat{\sigma}$ returned by the scaled lasso. The following lemma uses the analysis of Sun and Zhang (2012) and its proof is given in the on-line appendix A.9 for readers' convenience.

Lemma 9. Under the assumptions of theorem 1, let $\hat{\sigma} = \hat{\sigma}(\lambda)$ be the scaled lasso estimator of the noise level, with $\lambda = c\sqrt{\{\log(p)/n\}}$ and define $\sigma_* = \|w\|/\sqrt{n}$. Then, $\hat{\sigma}$ satisfies

$$\mathbb{P}\left[\left|\frac{\hat{\sigma}}{\sigma^*} - 1\right| \geqslant \frac{2c}{\phi_0 \sigma^*}\sqrt{\left\{\frac{s_0 \log(p)}{n}\right\}}\right] \leqslant 2p^{-c_0} + 2\exp\left(-\frac{n}{16}\right), \qquad c_0 = \frac{c^2}{32K} - 1. \tag{63}$$

Armed with lemmas 9 and 1 we are ready to prove theorem 1. Under H_0, we have $\theta_0 \in \Omega_0$ and hence, by invoking lemma 1, we have

$$\begin{aligned} T_n &= \|D(\hat{\gamma}^{\mathrm{d}} - U^{\mathrm{T}}\theta^{\mathrm{p}})\|_\infty \leqslant \|D(\hat{\gamma}^{\mathrm{d}} - U^{\mathrm{T}}\theta_0)\|_\infty \\ &\leqslant \frac{1}{\sqrt{n}}\|DZ\|_\infty + \frac{1}{\sqrt{n}}\|D\Delta\|_\infty. \end{aligned} \tag{64}$$

Note that, for $\tilde{Z} \equiv \hat{\sigma} DZ/(\sigma\sqrt{n}) \in \mathbb{R}^k$, we have $\tilde{Z}_i \sim N(0, 1)$. The entries of $\tilde{Z}$ are correlated though.

Fix $\epsilon > 0$ and apply equation (64) to write

$$\begin{aligned}\mathbb{P}(T_n \geqslant x) &\leqslant \mathbb{P}\left(\frac{\sigma}{\hat{\sigma}}\|\tilde{Z}\|_\infty + \frac{1}{\sqrt{n}}\|D\Delta\|_\infty \geqslant x\right)\\ &\leqslant \mathbb{P}\left(\frac{\sigma}{\hat{\sigma}}\|\tilde{Z}\|_\infty \geqslant x-\epsilon\right) + \mathbb{P}\left(\frac{1}{\sqrt{n}}\|D\Delta\|_\infty \geqslant \epsilon\right)\\ &\leqslant \mathbb{P}\{\|\tilde{Z}\|_\infty \geqslant (1-\epsilon)(x-\epsilon)\} + \mathbb{P}\left(\left|\frac{\hat{\sigma}}{\sigma}-1\right| \geqslant \epsilon\right) + \mathbb{P}\left(\frac{1}{\sqrt{n}}\|D\Delta\|_\infty \geqslant \epsilon\right). \end{aligned} \tag{65}$$

For the second term, we proceed as follows:

$$\mathbb{P}\left(\left|\frac{\hat{\sigma}}{\sigma}-1\right| \geqslant \epsilon\right) \leqslant \mathbb{P}\left(\left|\frac{\hat{\sigma}}{\sigma^*}-1\right| \geqslant \frac{\epsilon}{2}\right) + \mathbb{P}\left(\left|\frac{\hat{\sigma}}{\sigma^*}-\frac{\hat{\sigma}}{\sigma}\right| \geqslant \frac{\epsilon}{2}\right). \tag{66}$$

Now, note that $\sigma^* \to \sigma$, in probability, as $n \to \infty$. Therefore, by applying lemma 9 and using the assumption $s_0 = o\{n/\log(p)\}$, we obtain

$$\limsup_{n\to\infty} \mathbb{P}\left(\left|\frac{\hat{\sigma}}{\sigma}-1\right| \geqslant \epsilon\right) = 0. \tag{67}$$

Using this in inequality (65), we have

$$\begin{aligned}\limsup_{n\to\infty}\ \mathbb{P}(T_n \geqslant x) \leqslant \limsup_{n\to\infty}\ &\mathbb{P}\{\|\tilde{Z}\|_\infty \geqslant (1-\epsilon)(x-\epsilon)\}\\ &+ \limsup_{n\to\infty}\ \mathbb{P}\left(\frac{1}{\sqrt{n}}\|D\Delta\|_\infty \geqslant \epsilon\right). \end{aligned} \tag{68}$$

We next note that by definition (16), and using the assumption $\liminf_{n\to\infty} \min_{i\in[k]}(G^{\mathrm{T}}\hat{\Sigma}G)_{ii} \geqslant c_0 > 0$, we have

$$\begin{aligned}\limsup_{n\to\infty} \mathbb{P}\left(\frac{1}{\sqrt{n}}\|D\Delta\|_\infty \geqslant \epsilon\right) &\leqslant \limsup_{n\to\infty} \mathbb{P}\left(\frac{1}{\hat{\sigma}\sqrt{c_0}}\|\Delta\|_\infty \geqslant \epsilon\right)\\ &\leqslant \limsup_{n\to\infty} \mathbb{P}\left(\frac{2}{\sigma\sqrt{c_0}}\|\Delta\|_\infty > \epsilon\right) + \mathbb{P}\left(\frac{\sigma}{\hat{\sigma}} \geqslant 2\right). \end{aligned} \tag{69}$$

By equation (67), we have $\mathbb{P}(\sigma/\hat{\sigma} \geqslant 2) \to 0$. In addition, since $s_0 = o[1/\{\mu\sqrt{\log(p)}\}]$, for n and p sufficiently large, we have $c\mu s_0\sqrt{\log(p)}/\phi_0^2 \leqslant \epsilon\sqrt{c_0}/2$. Hence, by expression (14),

$$\begin{aligned}\limsup_{n\to\infty}\ \mathbb{P}\left(\frac{1}{\sqrt{n}}\|D\Delta\|_\infty \geqslant \epsilon\right) &\leqslant \limsup_{n\to\infty}\ \mathbb{P}\left(\|\Delta\|_\infty > \frac{\epsilon\sigma\sqrt{c_0}}{2}\right)\\ &\leqslant \limsup_{n\to\infty} \left\{2p^{-c_0} + 2\exp\left(-\frac{n}{16}\right)\right\} = 0. \end{aligned} \tag{70}$$

By substituting inequality (70) into inequality (65), we obtain

$$\limsup_{n\to\infty}\ \mathbb{P}(T_n \geqslant x) \leqslant \limsup_{n\to\infty}\ \mathbb{P}(\|\tilde{Z}\|_\infty \geqslant x - \epsilon + \epsilon^2). \tag{71}$$

By union bounding over the entries of $\tilde{Z}$, we obtain

$$\mathbb{P}(\|\tilde{Z}\|_\infty \geqslant x - \epsilon x + \epsilon^2) \leqslant 2k\{1 - \Phi(x - \epsilon x + \epsilon^2)\}. \tag{72}$$

Observe that inequality (72) holds for any $\epsilon > 0$, and that the right-hand side is bounded pointwise for all

ϵ. Therefore, by applying the dominated convergence theorem, we obtain

$$\limsup_{n\to\infty} \mathbb{P}(T_n \geqslant x) \leqslant 2k\{1-\Phi(x)\}.$$

The result follows by choosing $x = \Phi^{-1}\{1-\alpha/(2k)\}$.

A.2. Proof of theorem 2

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

$$\mathcal{E}_n(\phi_0, s_0, K) \equiv \left\{ X \in \mathbb{R}^{n\times p} : \min_{S:|S|\leqslant s_0} \phi(\hat{\Sigma}, S) \geqslant \phi_0, \ \max_{i\in[p]} \hat{\Sigma}_{i,i} \leqslant K, \ \hat{\Sigma} = (X^{\mathrm{T}}X/n) \right\}. \tag{73}$$

Then, by the result of Rudelson and Zhou (2013), theorem 6 (see also Javanmard and Montari (2014b), theorem 2.4(a)), random designs satisfy the compatibility condition with constant $\phi_0 = \sqrt{C_{\min}}/2$, provided that $n \geqslant \nu s_0 \log(p/s_0)$, where $\nu = c\kappa^4 C_{\max}/C_{\min}$, for a constant $c > 0$. More precisely,

$$\mathbb{P}\{X \in \mathcal{E}_n(\sqrt{C_{\min}}/2, s_0, K)\} \geqslant 1 - 4\exp(-c_1 n/\kappa^4), \tag{74}$$

where $c_1 = c_1(c) > 0$ is a constant.

We next provide an explicit upper bound for the minimum generalized coherence $\mu_{\min}(X; U)$ (see definition 2) for random designs.

Proposition 2 (Javanmard and Montanari, 2014b). Let $\Sigma \in \mathbb{R}^{p\times p}$ be such that $\sigma_{\min}(\Sigma) \geqslant C_{\min} > 0$ and $\sigma_{\max}(\Sigma) \leqslant C_{\max} < \infty$ and $\max_{i\in[p]} \Sigma_{ii} \leqslant 1$. Suppose that $X\Sigma^{-1/2}$ has independent sub-Gaussian rows, with mean 0 and sub-Gaussian norm $\|\Sigma^{-1/2}x_1\|_{\psi_2} = \kappa$, for some constant $\kappa > 0$. For $U \in \mathbb{R}^{p\times k}$ independent of X satisfying $U^{\mathrm{T}}U = I$ and, for fixed constant $a > 0$, define

$$\mathcal{G}_n(a) \equiv \left[X \in \mathbb{R}^{n\times p} : \mu_{\min}(X; U) < a\sqrt{\left\{ \frac{\log(p)}{n} \right\}} \right]. \tag{75}$$

In other words, $\mathcal{G}_n(a)$ is the event that solving problem (12) is feasible for $\mu = a\sqrt{\{\log(p)/n\}}$. Then, for $n \geqslant a^2 C_{\min}\log(p)/(4e^2 C_{\max}\kappa^4)$, the following result holds true with high probability:

$$\mathbb{P}\{X \in \mathcal{G}_n(a)\} \geqslant 1 - 2p^{-c_2}, \qquad c_2 = \frac{a^2 C_{\min}}{24e^2\kappa^4 C_{\max}} - 2. \tag{76}$$

We refer to the on-line appendix A.10 for the proof of proposition 9.

The last step is to prove that assumption 1 holds. In doing that, we use lemma 2. Note that the first condition of lemma 2 holds by assumption of theorem 2. To prove the second condition, we use the following result.

Lemma 10. Let $\Sigma \in \mathbb{R}^{p\times p}$ such that $\sigma_{\min}(\Sigma) \geqslant C_{\min} > 0$ and $\sigma_{\max}(\Sigma) \leqslant C_{\max} < \infty$ and $\max_{i\in[p]} \Sigma_{ii} \leqslant 1$. Suppose that $X\Sigma^{-1/2}$ has independent sub-Gaussian rows, with mean 0 and sub-Gaussian norm $\|\Sigma^{-1/2}x_1\|_{\psi_2} = \kappa$, for some constant $\kappa > 0$. Let $\hat{\Sigma} \equiv (X^{\mathrm{T}}X)/n$. For $u_i \in \mathbb{R}^p$ independent of X, we have

$$\mathbb{P}\left[u_i^{\mathrm{T}}(\hat{\Sigma} - \Sigma)u_i \geqslant C\sqrt{\left\{ \frac{\log(p)}{n} \right\}} \right] \leqslant p^{-c}, \tag{77}$$

for a constant $C > 0$ depending on κ, $C_{\max}$, and $c > 2$ depending on C.

We refer to the on-line appendix A.2 for the proof of lemma 10. The second condition of lemma 2 follows from $u_i^{\mathrm{T}}\Sigma u_i \leqslant C_{\max}\|u_i\|^2 = C_{\max}$, union bounding over $i \in [k]$ and lemma 10 (along with the Borel–Cantelli lemma).

Putting the three probabilistic bounds (74), (76) and (77) together in theorem 1, we obtain that, for random designs with $s_0 = o\{\sqrt{n}/\log(p)\}$, $\limsup_{n\to\infty} \alpha_n(R_X) \leqslant \alpha$.

A.3. Proof of theorem 3

We start by stating a lemma that will be used later in the proof.

Lemma 11. Under the assumptions of theorem 2, for any $i \in [k]$ we have

$$\mathbb{P}\left[g_i^{\mathrm{T}} \hat{\Sigma} g_i \geqslant u_i^{\mathrm{T}} \Sigma^{-1} u_i + C \sqrt{\left\{ \frac{\log(p)}{n} \right\}} \right] \leqslant 2\, p^{-c},$$

where c is a constant depending on a and C and, by a suitable choice of them, we have $c \geqslant 2$.

We refer to the on-line appendix A.11 for the proof of lemma 11.

Corollary 1. Assuming the setting of theorem 2, by an application of the Borel–Cantelli lemma and using lemma 11, of any $i \in [k]$ we have almost surely

$$\limsup_{n \to \infty} (g_i^{\mathrm{T}} \hat{\Sigma} g_i - u_i^{\mathrm{T}} \Sigma^{-1} u_i) \leqslant 0. \tag{78}$$

Recalling the definition of m_0, given by expression (28), we have the following corollary.

Corollary 2. Recalling the definition of m_0 given by expression (28), for any $i \in [k]$, we have almost surely

$$\limsup_{n \to \infty} (g_i^{\mathrm{T}} \hat{\Sigma} g_i - m_0^2) \leqslant 0. \tag{79}$$

Let $z_* \equiv \Phi^{-1}\{1 - \alpha/(2k)\}$ and write

$$\begin{aligned} \liminf_{n \to \infty} \frac{1 - \beta_n(R_X)}{1 - \beta_n^*(\eta)} &= \liminf_{n \to \infty} \frac{1}{1 - \beta_n^*(\eta)} \inf_{\theta_0} \{ \mathbb{P}_{\theta_0}(R_X = 1) : \|\theta_0\|_0 \leqslant s_0,\, d(\theta_0, \Omega_0; U) \geqslant \eta \} \\ &= \liminf_{n \to \infty} \frac{1}{1 - \beta_n^*(\eta)} \inf_{\theta_0} [\mathbb{P}\{ \|D(\hat{\gamma}^{\mathrm{d}} - U^{\mathrm{T}} \theta^{\mathrm{p}})\|_\infty \geqslant z_* \} : \|\theta_0\|_0 \leqslant s_0,\, d(\theta_0, \Omega_0; U) \geqslant \eta]. \end{aligned} \tag{80}$$

We define the shorthands $v \equiv DU^{\mathrm{T}}(\theta^{\mathrm{p}} - \theta_0)$ and $\tilde{v} \equiv D(\hat{\gamma}^{\mathrm{d}} - U^{\mathrm{T}} \theta_0)$. Note that $v, \tilde{v} \in \mathbb{R}^k$. We further let $i^* \equiv \arg\max_{i \in [k]} |v_i|$. Then, we can write

$$\|D(\hat{\gamma}^{\mathrm{d}} - U^{\mathrm{T}} \theta^{\mathrm{p}})\|_\infty = |v - \tilde{v}|_\infty \geqslant |v_{i^*} - \tilde{v}_{i^*}|. \tag{81}$$

By an argument that is very similar to that used to derive equation (71), we can show that, for any fixed $i \in [k]$ and all $x \in \mathbb{R}$, we have

$$\limsup_{n \to \infty} \sup_{\|\theta_0\|_0 \leqslant s_0} |\mathbb{P}(\tilde{v}_i \leqslant x) \leqslant \Phi(x)| = 0. \tag{82}$$

In words, each co-ordinate of $\tilde{v}$ asymptotically admits a standard normal distribution.

The other remark that we want to make is about the quantity $\|v\|_\infty$, which will be a key factor in determining the power of the test. Because $\theta^{\mathrm{p}} \in \Omega_0$, we have

$$|v_{i^*}| = \|v\|_\infty \geqslant \min_{i \in [k]}(D_{ii})\, \|U^{\mathrm{T}}(\theta^{\mathrm{p}} - \theta_0)\|_\infty \geqslant \min_{i \in [k]}(D_{ii})\, d(\theta_0, \Omega_0; U) \geqslant \eta \min_{i \in [k]}(D_{ii}). \tag{83}$$

Continuing with equation (80), we write

$$\begin{aligned} \liminf_{n \to \infty} \frac{1 - \beta_n(R_X)}{1 - \beta_n^*(\eta)} &= \liminf_{n \to \infty} \frac{1}{1 - \beta_n^*(\eta)} \inf_{\theta_0} [\mathbb{P}\{ \|D(\hat{\gamma}^{\mathrm{d}} - U^{\mathrm{T}} \theta^{\mathrm{p}})\|_\infty \geqslant z_* \} : \|\theta_0\|_0 \leqslant s_0,\, d(\theta_0, \Omega_0; U) \geqslant \eta] \\ &\overset{(a)}{\geqslant} \liminf_{n \to \infty} \frac{1}{1 - \beta_n^*(\eta)} \inf_{\theta_0} \left\{ \mathbb{P}(|v_{i^*} - \tilde{v}_{i^*}| \geqslant z_*) : |v_{i^*}| \geqslant \eta \min_{i \in [k]}(D_{ii}) \right\} \end{aligned}$$

$$
\begin{aligned}
&= \liminf_{n\to\infty} \frac{1}{1-\beta_n^*(\eta)} \left[1 - \sup_{\theta_0}\{\mathbb{P}(|v_{i^*} - \tilde{v}_{i^*}| \leqslant z_*) : |v_{i^*}| \geqslant \eta \min_{i\in[k]}(D_{ii})\}\right] \\
&\geqslant \liminf_{n\to\infty} \frac{1}{1-\beta_n^*(\eta)} \left[1 - \sup_{\theta_0}\{\mathbb{P}(\exists j \in [k] : |v_{i^*} - \tilde{v}_j| \leqslant z_*) : |v_{i^*}| \geqslant \eta \min_{i\in[k]}(D_{ii})\}\right] \\
&\geqslant \liminf_{n\to\infty} \frac{1}{1-\beta_n^*(\eta)} \left[1 - k \sup_{\theta_0}\{\mathbb{P}(|v_{i^*} - \tilde{v}_1| \leqslant z_*) : |v_{i^*}| \geqslant \eta \min_{i\in[k]}(D_{ii})\}\right] \\
&\overset{\text{(b)}}{\geqslant} \liminf_{n\to\infty} \frac{1}{1-\beta_n^*(\eta)} \left\{1 - k\mathbb{P}\left(\left|\frac{\sqrt{n}\eta}{\hat{\sigma} m_0} - Z\right| \leqslant z_*\right)\right\} \\
&= \liminf_{n\to\infty} \frac{1}{1-\beta_n^*(\eta)} \left[1 - k\left\{\Phi\left(\frac{\sqrt{n}\eta}{\hat{\sigma} m_0} + z_*\right) - \Phi\left(\frac{\sqrt{n}\eta}{\hat{\sigma} m_0} - z_*\right)\right\}\right] \\
&\overset{\text{(c)}}{=} \liminf_{n\to\infty} \frac{1}{1-\beta_n^*(\eta)} F\left(\alpha, \frac{\sqrt{n}\eta}{\hat{\sigma} m_0}, k\right) = 1,
\end{aligned} \tag{84}
$$

where inequality (a) follows from equations (81) and (83) and inequality (b) holds because of corollary 2 and equation (82). Here Z is a standard normal variable; equation (c) follows by substituting for z_*.

A.4. Proof of theorem 4

The proof of theorem 4 goes along the same lines as the proof of theorems 1 and 2.

Defining $r = X\theta_* - X\theta_0$ and by plugging in for $y = X\theta_* + w = X\theta_0 + r + w$ in the definition (11), we obtain

$$
\begin{aligned}
\hat{\gamma}^{\mathrm{d}} &= U^{\mathrm{T}}\hat{\theta} + \frac{1}{n} G^{\mathrm{T}} X^{\mathrm{T}} X(\theta_0 - \hat{\theta}) + \frac{1}{n} G^{\mathrm{T}} X^{\mathrm{T}} r + \frac{1}{n} G^{\mathrm{T}} X^{\mathrm{T}} w \\
&= U^{\mathrm{T}}\theta_0 + (G^{\mathrm{T}}\hat{\Sigma} - U^{\mathrm{T}})(\theta_0 - \hat{\theta}) + \frac{1}{n} G^{\mathrm{T}} X^{\mathrm{T}} r + \frac{1}{n} G^{\mathrm{T}} X^{\mathrm{T}} w \\
&= U^{\mathrm{T}}\theta_0 + \frac{1}{\sqrt{n}}\Delta + \frac{1}{\sqrt{n}} Z,
\end{aligned} \tag{85}
$$

with

$$
\begin{aligned}
\Delta &\equiv \Delta_1 + \Delta_2, \\
\Delta_1 &\equiv \sqrt{n}(G^{\mathrm{T}}\hat{\Sigma} - U^{\mathrm{T}})(\theta_0 - \hat{\theta}), \\
\Delta_2 &\equiv \frac{1}{\sqrt{n}} G^{\mathrm{T}} X^{\mathrm{T}} r, \\
Z &\equiv \frac{1}{\sqrt{n}} G^{\mathrm{T}} X^{\mathrm{T}} w.
\end{aligned}
$$

Since $w \sim N(0, \sigma^2 I_{n\times n})$, we have $Z|X \sim N(0, \sigma^2 G^{\mathrm{T}}\hat{\Sigma}G)$. We next bound $\|\Delta\|_\infty$.

It is straightforward to see that the assumptions of theorem 4 imply the assumption of lemma 4 and hence, by the result of the lemma, the moment conditions (assumption 3) hold. To deal with Δ_1, we use the following result from Belloni *et al.* (2012) that bounds the l_1-error of the iterated lasso estimator under assumptions 2 and 3.

Proposition 3 (Belloni *et al.* (2012) theorem 1). Suppose that in the regression model (41), assumption 2 (approximate sparsity) and assumption 3 (moment conditions) hold. Let $\hat{\theta}$ be the iterated lasso estimator (43) with weights γ_j specified by algorithm (44). Then, $\hat{\theta}$ satisfies

$$
\|\hat{\theta} - \theta_0\|_1 \leqslant C C_{\min}^{-1} s_0 \sqrt{\left\{\frac{\log(p)}{n}\right\}}, \tag{86}
$$

with high probability, for some finite constant $C > 0$.

Now let $\mathcal{E}_n$ be the probability event that $\|\hat{\theta} - \theta_0\|_1 \leqslant C C_{\min}^{-1} s_0 \sqrt{\{\log(p)/n\}}$. Recall the event $\mathcal{G}_n(a)$ from

definition (75) and define $\mathcal{F}_n \equiv \mathcal{G}_n(a) \cap \mathcal{E}_n$. Then, by using propositions 2 and 3, we have that $\mathcal{F}_n$ happens with high probability. Further, on the event $\mathcal{F}_n$ we have

$$\|\Delta_1\| \leqslant \sqrt{n}a\sqrt{\left\{\frac{\log(p)}{n}\right\}}CC_{\min}^{-1}s_0\sqrt{\left\{\frac{\log(p)}{n}\right\}} = CC_{\min}^{-1}as_0\frac{\log(p)}{\sqrt{n}}. \tag{87}$$

We next bound Δ_2. Write

$$\|\Delta_2\|_\infty \leqslant \left(\max_{i\in[k]}\left\|\frac{1}{\sqrt{n}}Xg_j\right\|\right)\|r\|.$$

Using lemma 11, we have

$$\left\|\frac{1}{\sqrt{n}}Xg_i\right\|^2 = g_i^{\mathrm{T}}\hat{\Sigma}g_i \leqslant u_i^{\mathrm{T}}\Sigma^{-1}u_i + C\sqrt{\left\{\frac{\log(p)}{n}\right\}} \leqslant \frac{1}{C_{\min}} + C\sqrt{\left\{\frac{\log(p)}{n}\right\}} < C',$$

with $C' = 1/C_{\min} + C$, and with probability at least $1 - 2p^{-c}$, for $c \geqslant 2$. By union bounding over $i \in [k]$, we obtain

$$\max_{i\in[k]}\left\|\frac{1}{\sqrt{n}}Xg_i\right\| \leqslant C',$$

with probability at least $1 - 2kp^{-c} \geqslant 1 - 2p^{-c+1}$. Using assumption 2, $\|r\| = o_P(1)$, which gives us

$$\|\Delta_2\|_\infty = o_P(1). \tag{88}$$

Combining expressions (87) and (88), we have

$$\|\Delta\|_\infty = O_P\left\{s_0\frac{\log(p)}{\sqrt{n}}\right\} + o_P(1).$$

Hence $\|\Delta\|_\infty = o_p(1)$ and $Z|X$ is asymptotically normally distributed. Having this result, we can then follow the lines of the proof of theorem 2 to show that our procedure controls the type I error, i.e. $\limsup_{n\to\infty}\alpha_n(R_X) \leqslant \alpha$.

References

Abramovich, F., Benjamini, Y., Donoho, D. L. and Johnstone, I. M. (2006) Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.*, **34**, 584–653.

Barber, R. F. and Candès, E. J. (2015) Controlling the false discovery rate via knockoffs. *Ann. Statist.*, **43**, 2055–2085.

Barber, R. F. and Kolar, M. (2018) Rocket: robust confidence intervals via Kendall's tau for transelliptical graphical models. *Ann. Statist.*, **46**, no. 6B, 3422–3450.

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

Belloni, A., Chen, D., Chernozhukov, V. and Hansen, C. (2012) Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, **80**, 2369–2429.

Belloni, A., Chernozhukov, V., Fernández-Val, I. and Hansen, C. (2017) Program evaluation and causal inference with high-dimensional data. *Econometrica*, **85**, 233–298.

Belloni, A., Chernozhukov, V. and Hansen, C. (2011) Lasso methods for Gaussian instrumental variables models. *Preprint*. (Available from `http://arxiv.org/abs/1012.1297`.)

Belloni, A., Chernozhukov, V. and Hansen, C. (2014) Inference on treatment effects after selection among high-dimensional controls. *Rev. Econ. Stud.*, **81**, 608–650.

Belloni, A., Chernozhukov, V. and Hansen, C. B. (2013) *Inference for High-dimensional Sparse Econometric Models*, vol. 3, pp. 245–295. Cambridge: Cambridge University Press.

Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009) Simultaneous analysis of Lasso and Dantzig selector. *Am. J. Math.*, **37**, 1705–1732.

Bogdan, M., van den Berg, M., Sabatti, C., Su, W. and Candès, E. J. (2015) Slope—adaptive variable selection via convex optimization. *Ann. Appl. Statist.*, **9**, 1103–1140.

Bühlmann, P. and van de Geer, S. (2011) *Statistics for High-dimensional Data*. New York: Springer.

Bühlmann, P., Kalisch, M. and Meier, L. (2014) High-dimensional statistics with a view toward applications in biology. *A. Rev. Statist. Appl.*, **1**, 255–278.

Bunea, F., Tsybakov, A. and Wegkamp, M. (2007) Sparsity oracle inequalities for the lasso. *Electron. J. Statist.*, **1**, 169–194.
Cai, T., Cai, T. and Guo, Z. (2019) Individualized treatment selection: an optimal hypothesis testing approach in high-dimensional models. *Preprint arXiv:1904.12891*.
Cai, T. T. and Guo, Z. (2017) Confidence intervals for high-dimensional linear regression: minimax rates and adaptivity. *Ann. Statist.*, **45**, 615–646.
Candès, E., Fan, Y., Janson, L. and Lv, J. (2018) Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *J. R. Statist. Soc.* B, **80**, 551–577.
Candés, E. and Tao, T. (2007) The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, **35**, 2313–2351.
Chen, M., Ren, Z., Zhao, H. and Zhou, H. (2016) Asymptotically normal and efficient estimation of covariate-adjusted Gaussian graphical model. *J. Am. Statist. Ass.*, **111**, 394–406.
Chen, S. S. and Donoho, D. L. (1995) Examples of basis pursuit. In *Proc. Wavelet Applications in Signal and Image Processing III, San Diego*.
Deshpande, Y., Javanmard, A. and Mehrabi, M. (2019) Online debiasing for adaptively collected high-dimensional data. *Preprint arXiv:1911.01040*.
Dicker, L. H. (2014) Variance estimation in high-dimensional linear models. *Biometrika*, **101**, 269–284.
Fan, J., Guo, S. and Hao, N. (2012) Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *J. R. Statist. Soc.* B, **74**, 37–65.
Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Ass.*, **96**, 1348–1360.
Fan, J. and Lv, J. (2008) Sure independence screening for ultrahigh dimensional feature space (with discussion). *J. R. Statist. Soc.* B, **70**, 849–911.
Fithian, W., Sun, D. and Taylor, J. (2014) Optimal inference after model selection. *Preprint arXiv:1410.2597*. Department of Statistics, University of California at Berkeley, Berkeley.
Friedman, J., Hastie, T. and Tibshirani, R. (2010) Regularization paths for generalized linear models via coordinate descent. *J. Statist. Softwr.*, **33**, 1–22.
Greenshtein, E. and Ritov, Y. (2004) Persistence in high-dimensional predictor selection and the virtue of over-parametrization. *Bernoulli*, **10**, 971–988.
Guo, Z., Wang, W., Cai, T. T. and Li, H. (2019) Optimal estimation of genetic relatedness in high-dimensional linear models. *J. Am. Statist. Ass.*, **114**, 358–369.
Harris, X. T., Panigrahi, S., Markovic, J., Bi, N. and Taylor, J. (2016) Selective sampling after solving a convex problem. *Preprint arXiv:1609.05609*.
Janson, L., Barber, R. F. and Candes, E. (2017) Eigenprism: inference for high dimensional signal-to-noise ratios. *J. R. Statist. Soc.* B, **79**, 1037–1065.
Janson, L. and Su, W. (2016) Familywise error rate control via knockoffs. *Electron. J. Statist.*, **10**, 960–975.
Javanmard, A. and Javadi, H. (2019) False discovery rate control via debiased lasso. *Electron. J. Statist.*, **13**, 1212–1253.
Javanmard, A. and Montanari, A. (2013) Nearly optimal sample size in hypothesis testing for high-dimensional regression. In *Proc. 51st A. Allerton Conf., Monticello, June*, pp. 1427–1434. New York: Institute of Electrical and Electronics Engineers.
Javanmard, A. and Montanari, A. (2014a) Hypothesis testing in high-dimensional regression under the Gaussian random design model: asymptotic theory. *IEEE Trans. Inform. Theory*, **60**, 6522–6554.
Javanmard, A. and Montanari, A. (2014b) Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.*, **15**, 2869–2909.
Javanmard, A. and Montanari, A. (2018) Debiasing the lasso: optimal sample size for gaussian designs. *Ann. Statist.*, **46**, no. 6A, 2593–2622.
Kudo, A. (1963) A multivariate analogue of the one-sided test. *Biometrika*, **50**, 403–418.
Lee, J. D., Sun, D. L., Sun, Y. and Taylor, J. E. (2016) Exact post-selection inference, with application to the lasso. *Ann. Statist.*, **44**, 907–927.
Lee, J. D. and Taylor, J. E. (2014) Exact post model selection inference for marginal screening. In *Advances in Neural Information Processing Systems*, pp. 136–144. Cambridge: MIT Press.
Meinshausen, N. and Bühlmann, P. (2006) High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, **34**, 1436–1462.
Nickl, R. and van de Geer, S. (2013) Confidence sets in sparse regression. *Ann. Statist.*, **41**, 2852–2876.
Raskutti, G., Wainwright, M. J. and Yu, B. (2009) Minimax rates of estimation for high-dimensional linear regression over l_q-balls. In *Proc. 47th A. Allerton Conf., Monticello, Sept.* New York: Institute of Electrical and Electronics Engineers.
Raskutti, G., Wainwright, M. J. and Yu, B. (2011) Minimax rates of estimation for high-dimensional linear regression over l_q-balls. *IEEE Trans. Inform. Theory*, **57**, 6976–6994.
Raubertas, R. F., Lee, C.-I. C. and Nordheim, E. V. (1986) Hypothesis tests for normal means constrained by linear inequalities. *Communs Statist. Theory Meth.*, **15**, 2809–2833.
Ren, Z., Sun, T., Zhang, C.-H. and Zhou, H. H. (2015) Asymptotic normality and optimalities in estimation of large Gaussian graphical models. *Ann. Statist.*, **43**, 991–1026.

Robertson, T. and Wegman, E. J. (1978) Likelihood ratio tests for order restrictions in exponential families. *Ann. Statist.*, **6**, 485–505.
Rudelson, M. and Zhou, S. (2013) Reconstruction from anisotropic random measurements. *IEEE Trans. Inform. Theory*, **59**, 3434–3447.
Su, W. and Candes, E. (2016) Slope is adaptive to unknown sparsity and asymptotically minimax. *Ann. Statist.*, **44**, 1038–1068.
Sun, T. and Zhang, C.-H. (2012) Scaled sparse linear regression. *Biometrika*, **99**, 879–898.
Tian, X. and Taylor, J. (2018) Selective inference with a randomized response. *Ann. Statist.* **46**, 679–710.
Tibshirani, R. (1996) Regression shrinkage and selection with the lasso. *J. R. Statist. Soc.* B, **58**, 267–288.
Tibshirani, R. J., Taylor, J., Lockhart, R. and Tibshirani, R. (2016) Exact post-selection inference for sequential regression procedures. *J. Am. Statist. Ass.*, **111**, 600–620.
Van de Geer, S., Bühlmann, P., Ritov, Y. and Dezeure, R. (2014) On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.*, **42**, 1166–1202.
Verzelen, N. and Gassiat, E. (2018) Adaptive estimation of high-dimensional signal-to-noise ratios. *Bernoulli*, **24**, no. 4B, 3683–3710.
Visscher, P. M., Hill, W. G. and Wray, N. R. (2008) Heritability in the genomics era—concepts and misconceptions. *Nat. Rev. Genet.*, **9**, 255–266.
Wainwright, M. J. (2009) Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming. *IEEE Trans. Inform. Theory*, **55**, 2183–2202.
Wang, J. and Kolar, M. (2016) Inference for high-dimensional exponential family graphical models. *Proc. Int. Conf. Artificial Intelligence and Statistics*, pp. 751–760.
Wang, Y., Wang, J., Balakrishnan, S. and Singh, A. (2019) Rate optimal estimation and confidence intervals for high-dimensional regression with missing covariates. *J. Multiv. Anal.*, **174**, article 104526.
Wei, Y., Wainwright, M. J. and Guntuboyina, A. (2019) The geometry of hypothesis testing over convex cones: generalized likelihood ratio tests and minimax radii. *Ann. Statist.*, **47**, 994–1024.
Ye, F. and Zhang, C.-H. (2010) Rate minimaxity of the lasso and Dantzig selector for the ℓ_q loss in ℓ_r balls. *J. Mach. Learn. Res.*, **11**, 3519–3540.
Zhao, P. and Yu, B. (2006) On model selection consistency of Lasso. *J. Mach. Learn. Res.*, **7**, 2541–2563.
Zhao, T., Kolar, M. and Liu, H. (2014) A general framework for robust testing and confidence regions in high-dimensional quantile regression. *Preprint arXiv:1412.8724*.
Zhang, C.-H. and Zhang, S. S. (2014) Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Statist. Soc.* B, **76**, 217–242.
Zhu, Y. and Bradic, J. (2017) A projection pursuit framework for testing general high-dimensional hypothesis. *Preprint arXiv:1705.01024*.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Supplementary material to "A flexible framework for hypothesis testing in high-dimensions"'.