Submitted to the Annals of Statistics

TWO-SAMPLE TESTING OF HIGH-DIMENSIONAL LINEAR REGRESSION COEFFICIENTS VIA COMPLEMENTARY SKETCHING

BY FENGNAN GAO[*,†] AND TENGYAO WANG[‡,§]

Fudan University[*], SCMS[†], London School of Economics and Political Science[‡] and University College London[§]

School of Data Science
Shanghai Center for Mathematical Sciences
Fudan University
Handan Road 220
Shanghai 200433, China
fngao@fudan.edu.cn

Department of Statistics
London School of Economics
Columbia House
69 Aldwych
London WC2B 4RR, United Kingdom
t.wang59@lse.ac.uk

We introduce a new method for two-sample testing of high-dimensional linear regression coefficients without assuming that those coefficients are individually estimable. The procedure works by first projecting the matrices of covariates and response vectors along directions that are complementary in sign in a subset of the coordinates, a process which we call 'complementary sketching'. The resulting projected covariates and responses are aggregated to form two test statistics, which are shown to have essentially optimal asymptotic power under a Gaussian design when the difference between the two regression coefficients is sparse and dense respectively. Simulations confirm that our methods perform well in a broad class of settings and an application to a large single-cell RNA sequencing dataset demonstrates its utility in the real world.

1. Introduction. Two-sample testing problems are commonplace in statistical applications across different scientific fields, wherever researchers want to compare observations from different samples. In its most basic form, a two-sample Gaussian mean testing problem is formulated as follows: upon observing two samples $X_1, \dots X_{n_1} \overset{\text{iid}}{\sim} N(\mu_1, \sigma^2)$ and $Y_1, \dots, Y_{n_2} \overset{\text{iid}}{\sim} N(\mu_2, \sigma^2)$, we wish to test

$$H_0 : \mu_1 = \mu_2 \quad \text{versus} \quad H_1 : \mu_1 \neq \mu_2. \tag{1}$$

This leads to the introduction of the famous two-sample Student's t-test. In a slightly more involved form in the parametric setting, we observe $X_1, \dots, X_{n_1} \overset{\text{iid}}{\sim} F_{\theta_1,\gamma_1}$ and $Y_1, \dots, Y_{n_2} \overset{\text{iid}}{\sim} F_{\theta_2,\gamma_2}$ and would like to test $H_0 : \theta_1 = \theta_2$ versus $H_1 : \theta_1 \neq \theta_2$, where γ_1 and γ_2 are nuisance parameters.

Linear regression models have been one of the staples of statistics. A two-sample testing problem in linear regression arises in the following classical setting: fix $p \ll \min\{n_1, n_2\}$, we observe an n_1-dimensional response vector Y_1 with an associated design matrix $X_1 \in \mathbb{R}^{n_1 \times p}$

AMS 2000 subject classifications: 62H15, 62J05.

Keywords and phrases: two-sample hypotheses testing, high-dimensional data, linear model, sparsity, minimax detection.

in the first sample, and an n_2-dimensional response Y_2 with design matrix $X_2 \in \mathbb{R}^{n_2 \times p}$ in the second sample. We assume in both samples the responses are generated from standard linear models

$$\begin{cases} Y_1 = X_1\beta_1 + \epsilon_1, \\ Y_2 = X_2\beta_2 + \epsilon_2, \end{cases} \tag{2}$$

for some unknown regression coefficients $\beta_1, \beta_2 \in \mathbb{R}^p$ and independent homoscedastic noise vectors $\epsilon_1 \mid (X_1, X_2) \sim N_{n_1}(0, \sigma^2 I_{n_1})$ and $\epsilon_2 \mid (X_1, X_2) \sim N_{n_2}(0, \sigma^2 I_{n_2})$. The purpose is to test $H_0 : \beta_1 = \beta_2$ versus $H_1 : \beta_1 \neq \beta_2$. Suppose that $\hat{\beta}$ is the least square estimate of $\beta = \beta_1 = \beta_2$ under the null hypothesis and $\hat{\beta}_1$, $\hat{\beta}_2$ are the least square estimates of β_1 and β_2 respectively under the alternative hypothesis. Define the residual sum of squares as

$$\begin{aligned} \mathrm{RSS}_1 &= \|Y_1 - X_1\hat{\beta}_1\|_2^2 + \|Y_2 - X_2\hat{\beta}_2\|_2^2, \\ \mathrm{RSS}_0 &= \|Y_1 - X_1\hat{\beta}\|_2^2 + \|Y_2 - X_2\hat{\beta}\|_2^2. \end{aligned} \tag{3}$$

The classical generalized likelihood ratio test (Chow, 1960) compares the F-statistic

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS}_1)/p}{\mathrm{RSS}_1/(n_1 + n_2 - 2p)} \sim F_{p,\, n_1+n_2-2p} \tag{4}$$

against upper quantiles of the $F_{p,\, n_1+n_2-2p}$ distribution. It is well-known that in the classical asymptotic regime where p is fixed and $n_1, n_2 \to \infty$, the above generalized likelihood ratio test is asymptotically optimal.

High-dimensional datasets are ubiquitous in the contemporary era of Big Data. As dimensions of modern data p in genetics, signal processing, econometrics and other fields are often comparable to sample sizes n, the most significant challenge in high-dimensional data is that the fixed-p-large-n setup prevalent in classical statistical inference is no longer valid. Yet the philosophy remains true that statistical inference is only possible when the sample size relative to the *true* parameter size is sufficiently large. Most advances in high-dimensional statistical inference so far have been made under some 'sparsity' conditions, i.e., all but a small (often vanishing) fraction of the p-dimensional model parameters are zero. The assumption in effect reduces the parameter size to an estimable level, and it makes sense in many applications because often only few covariates are *really* responsible for the response, though identification of these few covariates is still a nontrivial task. In the high-dimensional regression setting $Y = X\beta + \epsilon$ where $Y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, $\beta \in \mathbb{R}^p$ with $p, n \to \infty$ simultaneously, a common assumption to make is $k \log p/n \to 0$ with $k = \|\beta\|_0 := \sum_{j=1}^p \mathbb{1}_{\{\beta_j \neq 0\}}$. Therefore, k is the true parameter size, which vanishes relative to the sample size n, and $\log p$ is understood as the penalty to pay for not knowing where the k true parameters are.

Aiming to take a step in studying the fundamental aspect of two-sample hypothesis testing in high dimensions, this paper is primarily concerned with the following testing problem: we need to decide whether the responses in the two samples have different linear dependencies on the covariates. More specifically, under the same regression setting as in (2) with $\min\{p, n\} \to \infty$, we wish to test the global null hypothesis

$$H_0 : \beta_1 = \beta_2 \tag{5}$$

against the composite alternative

$$H_1 : \|\beta_1 - \beta_2\|_2 \geq 2\rho,\ \|\beta_1 - \beta_2\|_0 \leq k. \tag{6}$$

In other words, we assume that under the alternative hypothesis, the difference between the two regression coefficients is a k-sparse vector with ℓ_2 norm at least 2ρ (the additional factor of 2 here exists to simplify relevant statements under the reparametrisation we will introduce

later in Section 2). Throughout this paper, we do not assume the sparsity of β_1 or β_2 under the alternative.

Classical F-tests no longer work well on the above testing problem, for the simple reason that it is not possible to get good estimates of β's through naive least square estimators, which are necessary in establishing RSS in (3) to measure the model's goodness of fit. A standard way out is to impose certain kinds of sparsity on both β_1 and β_2 to ensure that both quantities are estimable. To our best knowledge, this is the out-of-shelf approach taken by most literature, see, for instance, Städler and Mukherjee (2012); Xia, Cai and Cai (2015). Nevertheless, it is both more interesting and relevant in applications to study the testing problem where neither β_1 nor β_2 is estimable but only $\beta_1 - \beta_2$ is sparse.

Practically, the assumption that β_1 and β_2 are both dense, but their difference is sparse can be motivated by comparisons of paired high-dimensional datasets where the commonly seen sparsity assumption fails for each individual dataset. For instance, Kraft and Hunter (2009) pointed out that in some genetic studies, "many, rather than few, variant risk alleles are responsible for the majority of the inherited risk of each common disease". Hence, to compare the difference between two such populations, it may not be appropriate to assume that the number of responsible single nucleotide polymorphisms (SNPs) in each population is small. On the other hand, the difference between the two populations can still be accounted for by a few SNPs, as pointed out in the Framingham Offspring Study (Kannel and McGee, 1979; Xia, Cai and Cai, 2018). The area of differential networks provides further examples to motivate two-sample testing of regression coefficients assuming only sparsity in their difference. Here, researchers are interested in whether two networks formulated as Gaussian graphical models, such as 'brain connectivity network' and gene-gene interaction network (Xia, Cai and Cai, 2015; Charbonnier, Verzelen and Villers, 2015), are different in two subgroups of population. Such complex networks are mostly of high-dimensional nature, in the sense that the number of nodes or features in the networks are large, relative to the number of observations. Since the off-diagonal entries of the inverse covariance matrix in a graphical model can be equated to the node-wise regression coefficients, such differential network testing problems can be reduced to multiple two-sample high-dimensional regression coefficient testing problems. Such networks are often dense as interactions within different brain parts or genes are omnipresent, but because they are subject to the about same physiology, the differences between networks from two subpopulations are conceivably small, i.e., there are only a few different edges from one network to another. In the above case of dense coefficients, sparsity assumption may not be true, and it is impossible to obtain reasonable estimates of either regression coefficient β_1 or β_2 when p is of the same magnitude as n. For this reason, any approach to detect the difference between β_1 and β_2, which is built upon comparing estimates of β_1 and β_2 in some ways, fails. In fact, any inference on β_1 or β_2 is not possible unless we make some other stringent structural assumptions on the model. However, certain inference on the coefficient difference $\beta_1 - \beta_2$, such as testing the zero null with the sparse alternative, is feasible by exploiting sparse difference between different networks without many assumptions. See Section 5 for an application of our method to a real-world single cell RNA-sequencing dataset, which exemplifies the aforementioned two-sample differential network analysis.

1.1. *Related Works.* The two-sample testing problem in its most general form is not well-understood in high dimensions. Most of the existing literature has focused on testing the equality of means, namely the high-dimensional equivalence of (1), see, e.g. Cai, Liu and Xia (2014); Chen, Li and Zhong (2019). Similar to our setup, in the mean testing problems, we may allow for non-sparse means in each sample and test only for sparse differences between the two population means (Cai, Liu and Xia, 2014). The intuitive approach for testing equality

of means is to eliminate the dense nuisance parameter by taking the difference in means of the two samples and thus reducing it to a one-sample problem of testing a sparse mean against a global zero null, which is also known as the 'needle in the haystack' problem well studied previously by e.g. Ingster (1997); Donoho and Jin (2004). Such reduction, however, is more intricate in the regression problem, as a result of different design matrices for the two samples.

Literature is scarce for two-sample testing under high-dimensional regression setting. Städler and Mukherjee (2012), Xia, Cai and Cai (2018), Xia, Cai and Sun (2020) have proposed methods that work under the additional assumption so that both β_1 and β_2 can be consistently estimated. Charbonnier, Verzelen and Villers (2015) and Zhu and Bradic (2016) are the only existing works in the literature we are aware of that allow for non-sparse regression coefficients β_1 and β_2. Specifically, Charbonnier, Verzelen and Villers (2015) look at a sequence of possible supports of β_1 and β_2 on a Lasso-type solution path and then apply a variant of the classical F-tests to the lower-dimensional problems restricted on these supports, with the test p-values adjusted by a Bonferroni correction. Zhu and Bradic (2016) (after some elementary transformation) uses a Dantzig-type selector to obtain an estimate for $(\beta_1 + \beta_2)/2$ and then use it to construct a test statistic based on a specific moment condition satisfied under the null hypothesis. As both tests depend on the estimation of nuisance parameters, their power can be compromised if such nuisance parameters are dense.

1.2. *Our contributions.* Our contributions are four-fold. First, we propose a novel method to solve the testing problems formulated in (5) and (6) for model (2). Through 'complementary sketching', which is a delicate linear transformation on both the designs and responses, our method turns the testing problem with two different designs into one with the same design of dimension $m \times p$ where $m = n_1 + n_2 - p$. After taking the difference in two regression coefficients, the problem is reduced to testing whether the coefficient in the reduced one-sample regression is zero against sparse alternatives. The transformation is carefully chosen such that the error distribution in the reduced one-sample regression is homoscedastic. This paves the way for constructing test statistics using the transformed covariates and responses. Our method is easy to implement and does not involve any complications arising from solving computationally expensive optimization problems. Moreover, when complementary sketching is combined with any methods designed for one-sample global testing problems (e.g. Ingster, Tsybakov and Verzelen, 2010; Arias-Castro, Candès and Plan, 2011; Carpentier et al., 2019; Carpentier and Verzelen, 2021), our proposal substantially supplies a novel class of testing and estimation procedures for the corresponding two-sample problems. However, as the design matrices after the complementary sketching transformation possess complex dependence structure, theoretical results from the one-sample testing literature cannot be directly applied, and new techniques are required in the current work to analyse our two-stage procedure.

Our second contribution is that, in the sparse regime, where the sparsity parameter $k \sim p^\alpha$ in the alternative (6) for any fixed $\alpha \in (0, 1/2)$, we show that the detection limit of our procedure, defined as the minimal $\|\beta_1 - \beta_2\|_2$ necessary for asymptotic almost sure separation of the alternative from the null, is minimax optimal up to a multiplicative constant under a Gaussian design. More precisely, we show that in the asymptotic regime where n_1, n_2, p diverge at a fixed ratio, and for a large class of covariance matrices of the design, if $\rho^2 \gtrsim \frac{k \log p}{n \kappa_1}$, where κ_1 is a constant depending on n_1/n_2 and p/m only, then our test has asymptotic power 1 almost surely. On the other hand, in the same asymptotic regime, if $\rho^2 \leq \frac{c_\alpha k \log p}{n \kappa_1}$ for some c_α depending only on α, then almost surely no test has asymptotic size 0 and power 1.

Furthermore, our results reveal the effective sample size of the two-sample testing problem. Here, by effective sample size, we mean the sample size for a corresponding one-sample

testing problem (i.e. testing $\beta = 0$ in a linear model $Y = X\beta + \epsilon$ with rows of X following the same distribution as rows of X_1 and X_2) that has an asymptotically equal detection limit; see the discussion after Theorem 5 for a detailed definition. At first glance, one might think that the effective sample size is m, which is the number of rows in the reduced design. This hints that the reduction to the one-sample problem has made the original two-sample problem obsolete. However, on deeper thoughts, as an imbalance in the numbers of observations in X_1 and X_2 clearly makes testing more difficult, the effective sample size has to also incorporate this effect. We see from the previous point that uniformly for any α less than and bounded away from $1/2$, the detection boundary is of order $\rho^2 \asymp \frac{k \log p}{n\kappa_1}$, with the precise definition of κ_1 given in Proposition 2. Writing $n_1/n_2 = r$ and $p/m = s$, our results on the sparse case implies that the two-sample testing problem has the same order of detection limit as in a one-sample problem with sample size $n\kappa_1 = m(r^{-1} + r + 2)^{-1}$. We note that this effective sample size is proportional to m, and for each fixed m, maximized when $r = 1$ (i.e. $n_1 = n_2$) and approaches m/n in the most imbalanced design. This is in agreement with the intuition that testing is easiest when $n_1 = n_2$ and impossible when n_1 and n_2 are too imbalanced. Our study, thus, sheds light on the intrinsic difference between two-sample and one-sample testing problems and characterizes the precise dependence of the difficulty of the two-sample problem on the sample size and dimensionality parameters.

Finally, we observe a phase transition phenomena of how the minimax detection limit depends on the sparsity parameter k. On top of minimax rate optimal detection limit of our procedure in the sparse case when $k \asymp p^\alpha$ for $\alpha \in [0, 1/2)$, we also prove that a modified version of our procedure, designed for denser signals, is able to achieve minimax optimal detection limit up to logarithmic factors in the dense regime $k \asymp p^\alpha$ for $\alpha \in (1/2, 1)$. However, the detection limit is of order $\rho^2 \asymp \frac{k \log p}{n\kappa_1}$ in the sparse regime, but of order $\rho^2 \asymp p^{-1/2}$ up to logarithmic factors in the dense regime. Such a phase transition phenomenon is qualitatively similar to results previously reported in the one-sample testing problem (see, e.g. Ingster, Tsybakov and Verzelen, 2010; Arias-Castro, Candès and Plan, 2011; Carpentier et al., 2019; Carpentier and Verzelen, 2021).

1.3. *Organization of the paper.* We describe our methodology in detail in Section 2 and establish its theoretical properties in Section 3. Numerical results illustrate the finite sample performance of our proposed algorithm in Section 4. We present in Section 5 a real data example to compare gene regulatory networks in two close-related types of T cells. Proofs of our main results are deferred until Section 6 with the rest of proofs and ancillary results in Sections A and B of the online supplementary material.

1.4. *Notation.* For any positive integer n, we write $[n] := \{1, \ldots, n\}$. For a vector $v = (v_1, \ldots, v_n)^\top \in \mathbb{R}^n$, we define $\|v\|_0 := \sum_{i=1}^n \mathbb{1}_{\{v_i \neq 0\}}$, $\|v\|_\infty := \max_{i \in [n]} |v_i|$ and $\|v\|_q := \left\{\sum_{i=1}^n (v_i)^q\right\}^{1/q}$ for any positive integer q, and let $\mathcal{S}^{n-1} := \{v \in \mathbb{R}^n : \|v\|_2 = 1\}$. The support of vector v is defined by $\mathrm{supp}(v) := \{i \in [n] : v_i \neq 0\}$.

For $n \geq m$, $\mathbb{O}^{n \times m}$ denotes the space of $n \times m$ matrices with orthonormal columns. For $a \in \mathbb{R}^p$, we define $\mathrm{diag}(a)$ to the $p \times p$ diagonal matrix with diagonal entries filled with elements of a, i.e., $(\mathrm{diag}(a))_{i,j} = \mathbb{1}_{\{i=j\}} a_i$. Let $A \in \mathbb{R}^{p \times p}$, and we write $\|A\|_{\mathrm{op}}$, $\|A\|_{\mathrm{F}}$ and $\|A\|_{\max}$ for its operator, Frobenius and entrywise ℓ_∞ norm respectively. We define $\mathrm{diag}(A)$ to be the $p \times p$ diagonal matrix with diagonal entries coming from A, i.e., $(\mathrm{diag}(A))_{i,j} = \mathbb{1}_{\{i=j\}} A_{i,j}$. We also write $\mathrm{tr}(A) := \sum_{i \in [p]} A_{i,i}$. For a symmetric matrix $A \in \mathbb{R}^{p \times p}$ and $j \in [p]$, we write $\lambda_j(A)$ for its jth largest (real) eigenvalue. When $\lambda_p(A) \geq 0$, A is positive semidefinite, which we denote by $A \succeq 0$. For A symmetric and $k \in [p]$, the k-sparse operator norm of A is defined by

$$\|A\|_{k,\mathrm{op}} := \sup_{v \in \mathcal{S}^{p-1} : \|v\|_0 \leq k} |v^\top A v|.$$

For any $S \subseteq [n]$, we write v_S for the $|S|$-dimensional vector obtained by extracting coordinates of v in S and $A_{S,S}$ the matrix obtained by extracting rows and columns of A indexed by S.

Given two sequences $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ such that $b_n > 0$ for all n, we write $a_n = \mathcal{O}(b_n)$ if $|a_n| \le Cb_n$ for some constant C. If the constant C depends on some parameter x, we write $a_n = \mathcal{O}_x(b_n)$ instead. Also, $a_n = o(b_n)$ denotes $a_n/b_n \to 0$.

2. Testing via complementary sketching. In this section, we describe our testing strategy. Since we are only interested in the difference in regression coefficients in the two linear models, we reparametrize (2) with $\gamma := (\beta_1 + \beta_2)/2$ and $\theta := (\beta_1 - \beta_2)/2$ to separate the nuisance parameter from the parameter of interest. Define

$$\Theta_{p,k}(\rho) := \big\{\theta \in \mathbb{R}^p : \|\theta\|_2 \ge \rho \text{ and } \|\theta\|_0 \le k\big\}.$$

Under this new parametrization, the null and the alternative hypotheses can be equivalently formulated as

$$H_0 : \theta = 0 \quad \text{and} \quad H_1 : \theta \in \Theta_{p,k}(\rho).$$

The parameter of interest θ is now k-sparse under the alternative hypotheses. However, its inference is confounded by the possibly dense nuisance parameter $\gamma \in \mathbb{R}^p$. A natural idea, then, is to eliminate the nuisance parameter from the model. In the special design setting where $X_1 = X_2$ (in particular, $n_1 = n_2$), this can be achieved by considering the sparse regression model $Y_1 - Y_2 = X_1\theta + (\epsilon_1 - \epsilon_2)$. While the above example only works in a special, idealized setting, it nevertheless motivates our general testing procedure.

To introduce our test, we first concatenate the design matrices and response vectors to form

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \quad \text{and} \quad Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}.$$

A key idea of our method is to project X and Y respectively along $n - p$ pairs of directions that are complementary in sign in a subset of their coordinates, a process we call *complementary sketching*. Specifically, assume $n_1 + n_2 > p$ and define $n := n_1 + n_2$ and $m := n - p$ and let $A_1 \in \mathbb{R}^{n_1\times m}$ and $A_2 \in \mathbb{R}^{n_2\times m}$ be chosen such that

$$A_1^\top A_1 + A_2^\top A_2 = I_m \quad \text{and} \quad A_1^\top X_1 + A_2^\top X_2 = 0. \tag{7}$$

In other words, $A := (A_1^\top, A_2^\top)^\top$ is a matrix with orthonormal columns orthogonal to the column space of X. Such A_1 and A_2 exist since the null space of X has dimension at least m. Define $Z := A_1^\top Y_1 + A_2^\top Y_2 \in \mathbb{R}^m$, $W := A_1^\top X_1 - A_2^\top X_2 \in \mathbb{R}^{m\times p}$ and $\xi = A_1^\top \varepsilon_1 + A_2^\top \varepsilon_2 \in \mathbb{R}^m$. From the above construction, we have

$$Z = A_1^\top X_1\beta_1 + A_2^\top X_2\beta_2 + (A_1^\top \epsilon_1 + A_2^\top \epsilon_2) = W\theta + \xi, \tag{8}$$

where $\xi \mid W \sim N_m(0, A^\top A) = N_m(0, \sigma^2 I_m)$. We note that similar to conventional sketching (see, e.g. Mahoney, 2011), the complementary sketching operation above synthesizes m data points from the original n observations. However, unlike conventional sketching, where one projects the design X and response Y by the same sketching matrix $S \in \mathbb{R}^{m\times n}$ to obtain sketched data (SX, SY), here we project X and Y along different directions to obtain $(\tilde{A}^\top X, A^\top Y)$, where $\tilde{A} := (A_1^\top, -A_2^\top)^\top$ is complementary in sign to A in its second block. Moreover, the main purpose of the conventional sketching is to trade off statistical efficiency for computational speed by summarizing raw data with a smaller number of synthesized data points, whereas the main aim of our complementary sketching operation is to eliminate the nuisance parameter, and surprisingly, as we will see in Section 3, there is essentially no loss

of statistical efficiency introduced by our complementary sketching in this two-sample testing setting.

To summarize, after projecting X and Y via complementary sketching to obtain W and Z, we reduce the original two-sample testing problem to a one-sample problem with m observations, where we test the global null of $\theta = 0$ against sparse alternatives using data (W, Z). From here, we can construct test statistics as functions of W and Z, for which we describe two different tests. The first testing procedure, detailed in Algorithm 1, computes the sum of squares of hard-thresholded inner products between the response Z and standardized columns of the design matrix W in (8). We denote the output of Algorithm 1 with input X_1, X_2, Y_1 and Y_2 and tuning parameters ω and τ as $\psi^{\mathrm{sparse}}_{\omega,\tau}(X_1, X_2, Y_1, Y_2)$. As we will see in Section 3, if we have a 'good' estimator $\hat{\sigma}$ for the noise level σ, the choice of $\omega = 2\hat{\sigma}\sqrt{\log p}$ and $\tau = k\hat{\sigma}^2 \log p$ would be suitable for testing against sparse alternatives in the case of $k \leq p^{1/2}$. On the other hand, in the dense case when $k > p^{1/2}$, one option would be to choose $\omega = 0$. However, it turns out to be difficult to set the test threshold level τ in this dense case using the known problem parameters. Therefore, we decided to study instead the following as our second test. We apply steps 1 to 4 of Algorithm 1 to obtain the vector Z, and then for a suitable choice of threshold level η, define our test as

$$\psi^{\mathrm{dense}}_{\eta}(X_1, X_2, Y_1, Y_2) := \mathbb{1}\{\|Z\|_2^2 \geq \eta\}.$$

Algorithm 1: Pseudo-code for complementary sketching-based test $\psi^{\mathrm{sparse}}_{\omega,\tau}$.

Input: $X_1 \in \mathbb{R}^{n_1 \times p}, X_2 \in \mathbb{R}^{n_2 \times p}, Y_1 \in \mathbb{R}^{n_1}, Y_2 \in \mathbb{R}^{n_2}$ satisfying $n_1 + n_2 - p > 0$, a hard threshold level $\omega \geq 0$, and a test threshold level $\tau > 0$.

1 Set $m \leftarrow n_1 + n_2 - p$.

2 Form $A \in \mathbb{O}^{n \times m}$ with columns orthogonal to the column space of $(X_1^\top, X_2^\top)^\top$.

3 Let A_1 and A_2 be submatrices formed by the first n_1 and last n_2 rows of A.

4 Set $Z \leftarrow A_1^\top Y_1 + A_2^\top Y_2$ and $W \leftarrow A_1^\top X_1 - A_2^\top X_2$.

5 Compute $Q \leftarrow \{\mathrm{diag}(W^\top W)\}^{-1/2} W^\top Z$.

6 Compute the test statistic

$$T := \sum_{j=1}^{p} Q_j^2 \mathbb{1}_{\{|Q_j| \geq \omega\}}.$$

7 Reject the null hypothesis if $T \geq \tau$.

The computational complexity of both $\psi^{\mathrm{sparse}}_{\omega,\tau}$ and $\psi^{\mathrm{dense}}_{\eta}$ depends on Step 2 of Algorithm 1. In practice, we can form the projection matrix A as follows. We first generate an $n \times m$ matrix M with independent $N(0,1)$ entries, and then project columns of M to the orthogonal complement of the column space of X to obtain $\tilde{M} := (I_n - XX^\dagger)M$, where $X^\dagger$ is the Moore–Penrose pseudoinverse of X. Finally, we extract an orthonormal basis from the columns of $\tilde{M}$ via a QR decomposition $\tilde{M} = AR$, where R is upper triangular and A is a (random) $n \times m$ matrix with orthonormal columns that can be used in Step 2 of Algorithm 1. The overall computational complexity for our tests are therefore of order $\mathcal{O}(n^2 p + nm^2)$. Finally, it is worth emphasizing that while the matrix A generated this way is random, our test statistics $T = \sum_{j=1}^{p} Q_j^2 \mathbb{1}_{\{|Q_j| \geq \omega\}}$ and $\|Z\|_2^2$, are in fact deterministic. To see this, we observe that both

$$\begin{aligned} W^\top Z &= (A_1^\top X_1 - A_2^\top X_2)^\top (A_1^\top Y_1 + A_2^\top Y_2) \\ &= \begin{pmatrix} X_1^\top & -X_2^\top \end{pmatrix} \begin{pmatrix} A_1 \\ A_2 \end{pmatrix} \begin{pmatrix} A_1^\top & A_2^\top \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \end{aligned}$$

and $\|Z\|_2^2 = Y^\top AA^\top Y$ depend on A only through $AA^\top$, which is determined by the column space of A. Moreover, by Lemma 10, $(\|W_j\|_2^2)_{j\in[p]}$, being diagonal entries of $W^\top W = 4X_1^\top A_1 A_1^\top X_1$, are also functions of X alone. This attests that both test statistics, and consequently our two tests, are deterministic in nature.

3. Theoretical analysis. We now turn to the analysis of the theoretical performance of $\psi_{\omega,\tau}^{\mathrm{sparse}}$ and $\psi_{\eta}^{\mathrm{dense}}$. We consider both the size and power of each test, as well as the minimax lower bounds for smallest detectable signal strengths.

In addition to working under the regression model (2), we further assume the following conditions in our theoretical analysis. For some constants $0 < \underline{\lambda} \le \overline{\lambda} < \infty$, we write

$$\mathcal{C} := \{\Sigma \in \mathbb{R}^{p\times p} : \Sigma \succeq 0, \operatorname{diag}(\Sigma) = I_p \text{ and for all } S \subseteq [p] \text{ with } |S| = k \\ \underline{\lambda} \le \lambda_k(\Sigma_{S,S}) \le \lambda_1(\Sigma_{S,S}) \le \overline{\lambda}\}$$

(C1) All rows of X_1 and X_2 are independent and follows $N_p(0,\Sigma)$ distribution such that $\Sigma \in \mathcal{C}$.

(C2) Parameters n_1, n_2, p satisfy $m = n_1 + n_2 - p > 0$ and lie in the asymptotic regime where $n_1/n_2 \to r$ and $p/m \to s$ as $n_1, n_2, p \to \infty$.

The condition that $\operatorname{diag}(\Sigma) = I_p$ in (C1) means that all columns of the design matrix should have unit variance and that Σ is in fact a correlation matrix. This is assumed here both to simplify notation in our theoretical analysis and to reflect the common practice of column normalization in practical applications (especially when covariates are measured in different units). For a generic Σ, we remark that the testing boundary should be measured in terms of $\theta^\top \operatorname{diag}(\Sigma)\theta$ instead of $\|\theta\|_2^2$ and results similar to Theorems 1, 3, 5, 6, 7 can be derived via the reduction $X \mapsto X\{\operatorname{diag}(\Sigma)\}^{-1/2}$. The condition in (C1) that the spectrum of any $k \times k$ principal submatrix of Σ is contained in $[\underline{\lambda}, \overline{\lambda}]$ is relatively mild. It requires that any k covariates are not too collinear. We note that it is in particular implied if Σ itself has a bounded condition number, or alternatively if Σ satisfies the restricted isometry condition.

The condition $n_1 + n_2 - p > 0$ in (C2) is necessary in this two-sample problem, since otherwise, for any prescribed value of $\Delta := \beta_1 - \beta_2$, the equation system with β_1 as unknowns

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \beta_1 = \begin{pmatrix} Y_1 \\ Y_2 - X_2\Delta \end{pmatrix}$$

has at least one solution when $(X_1^\top, X_2^\top)^\top$ has rank n. As a result, except in some pathological cases, we can always find $\beta_1, \beta_2 \in \mathbb{R}^p$ that fit the data perfectly with $Y_1 = X_1\beta_1$ and $Y_2 = X_2\beta_2$, which makes the testing problem impossible. A more rigorous statement regarding the necessity of this condition in a minimax sense is proved in Proposition 9. Finally, we have carried out proofs of our theoretical results with finite sample arguments wherever possible. Nevertheless, due to a lack of finite-sample bounds on the empirical spectral density of matrix-variate Beta distributions, all results in this section are presented under the asymptotic regime set out in Condition (C2). Under this condition, we were able to exploit existing results in the random matrix theory to obtain a sharp dependence of the detection limit on s and r.

In practice, the noise variance σ^2 is typically unknown. The problem of noise variance estimation in high-dimensional linear models is a well-studied one itself that has received much attention recently (Fan, Guo and Hao, 2012; Sun and Zhang, 2012; Homrighausen and McDonald, 2013; Dicker, 2014; Reid, Tibshirani and Friedman, 2016). As the main focus of the current work is on two-sample testing of regression coefficients, we will make the simplifying assumption in our theoretical analysis that the noise variance σ^2 is either known or that good estimators exist. Specifically, we assume that one of the following two conditions about the noise variance is met:

(S1) There exists an independent estimator $\hat{\sigma}$ of σ such that $\hat{\sigma} \xrightarrow{\text{a.s.}} \sigma$.
(S2) There exists an independent estimator $\hat{\sigma}$ such that we have $|\hat{\sigma}/\sigma - 1| = o(p^{-1/2}\log^{1/2} p)$ almost surely.

For most of our theoretical analysis, the much weaker condition (S1) suffices. The stronger condition (S2) is needed only in Theorem 6, where we derive the upper bound for the test ψ_η^{dense}.

We remark that condition (S1) is very mild. It is for instance significantly weaker than most conditions where one requires the rate of convergence of $\hat{\sigma}$ of order at least a polynomial in p, i.e., for some $\alpha \le 1/2$ and any $\varepsilon > 0$, $\mathbb{P}(|\hat{\sigma}/\sigma - 1|p^\alpha > \varepsilon) = \mathcal{O}(1)$. The independence assumption on $\hat{\sigma}$ is often achieved in practice by sample splitting. For instance, using consistent estimators of σ proposed in Dicker (2014), we may use $\mathcal{O}(\log^2 n)$ data points from each sample to estimate σ, and use the remaining samples to perform the hypothesis test. However, we will assume that $\hat{\sigma}$ is available independent of (X_1, Y_1, X_2, Y_2) so that we can focus our theoretical analysis on the primary points of interest in this problem.

Condition (S2) is much stronger than (S1) but slightly weaker than the usual $\sqrt{n}$-consistency found in the parametric literature, albeit we need the convergence to take place almost surely. Similar to (S1), in reality we could possibly obtain such an estimator via the usual sample-splitting argument, where we take a fixed proportion of all data points to estimate σ and the rest for testing.

Finally, for ease of reference, we summarize all our theoretical findings in Table 1. Here, the lower bounds are proved in a subclass of covariance matrices $\mathcal{C}(D) \subseteq \mathcal{C}$ defined in (12).

	sparse	dense
upper bound	$\dfrac{7\sigma^2 k \log p}{\underline{\lambda}^2 n \kappa_1}$	$\dfrac{2\sigma^2\sqrt{m \log p}}{\underline{\lambda} n \kappa_1}$
lower bound	$\dfrac{(1-2\alpha-\epsilon)\sigma^2 k \log p}{8 D n \kappa_1}$	$\mathcal{O}(p^{-1/2}\min\{\log^{-1/2}(ep/k), D^{-3/2}\})$

TABLE 1

Comparison of upper and lower bounds of the testing boundary in terms of ρ^2 in both sparse and dense cases.

3.1. *Sparse case.* We consider in this subsection the test $\psi_{\omega,\tau}^{\text{sparse}}$, which is suitable for distinguishing β_1 and β_2 that differ in only a few coordinates, the setting that has more subtle phenomena and hence is most interesting to us. Our first result below states that with a choice of hard-thresholding level ω of order $\sigma\sqrt{\log p}$, the test has asymptotic size 0.

THEOREM 1. *If Conditions (C2) and (S1) hold and $\beta_1 = \beta_2$, then, with the choice of parameters $\tau > 0$ and $\omega = \hat{\sigma}\sqrt{(4+\varepsilon)\log p}$ for any $\varepsilon > 0$, we have*

$$\psi_{\omega,\tau}^{\text{sparse}}(X_1, X_2, Y_1, Y_2) \xrightarrow{\text{a.s.}} 0.$$

The almost sure statement in Theorem 1 and subsequent results in this section are with respect to both the randomness in $X = (X_1^\top, X_2^\top)^\top$ and in $\epsilon = (\epsilon_1^\top, \epsilon_2^\top)^\top$. However, a closer inspection of the proof of Theorem 1 tells us that the statement is still true if we allow an arbitrary sequence of matrices X (indexed by p) and only consider almost sure convergence with respect to the distribution of ϵ.

The control of the asymptotic power of $\psi_{\omega,\tau}^{\text{sparse}}$ is more involved. A key step in the argument is to show that $W^\top W$ is suitably close to a multiple of the population covariance matrix Σ. More precisely, in Proposition 2 below, we derive entrywise and k-sparse operator norm controls of the Gram matrix of the design matrix sketch W.

PROPOSITION 2. *Under Conditions (C1) and (C2), we further assume $k \in [p]$ and let W be defined as in Algorithm 1. Then with probability 1,*

$$\max_{j\in[p]}\left|\frac{(W^\top W)_{j,j}}{4n\kappa_1}-1\right|\to 0, \tag{9}$$

where $\kappa_1 := r/\{(1+r)^2(1+s)\}$. Moreover, define $\tilde{W} := W\{\mathrm{diag}(W^\top W)\}^{-1/2}$. If

$$\frac{k\log(ep/k)}{n}\to 0, \tag{10}$$

then there exists $C_{s,r} > 0$, depending only on s and r, such that with probability 1, the following holds for all but finitely many p:

$$\|\tilde{W}^\top\tilde{W}-\Sigma\|_{k,\mathrm{op}} \le C_{s,r}\left\{\bar{\lambda}\sqrt{\frac{k\log(ep/k)}{n}}+\bar{\lambda}^2\sqrt{\frac{\log p}{n}}\right\}. \tag{11}$$

We note that condition (10) is relatively mild and would be satisfied if $k \le p^\alpha$ for any $\alpha \in [0,1)$. As we will see later, the quantity $n\kappa_1$ in (9) can be viewed as the 'effective sample size' of the two-sample testing problem. A more detailed discussion about the intuition and interpretation of this quantity is provided after Theorem 5.

All theoretical results in this section except for Theorem 1 assume the random design Condition (C1) to hold. However, as revealed by the proofs, for any given (deterministic) sequence of X, these results remain true as long as (9) and (11) are satisfied. The asymptotic nature of Proposition 2 is a result of our application of Bai et al. (2015, Theorem 1.1), which guarantees an almost sure convergence of the empirical spectral distribution of Beta random matrices in the weak topology. This sets the tone for the asymptotic nature of our results, which depend on the aforementioned limiting spectral distribution.

The following theorem provides power control of our procedure $\psi^{\mathrm{sparse}}_{\omega,\tau}$, when the ℓ_2 norm of the scaled difference in regression coefficient $\theta = (\beta_1-\beta_2)/2$ exceeds an appropriate threshold.

THEOREM 3. *Under Conditions (C1), (C2) and (S1), we further assume $k \in [p]$ and that (10) holds. If $\theta = (\beta_1-\beta_2)/2 \in \Theta_{p,k}(\rho)$ with $\rho^2 \ge \frac{7\sigma^2 k\log p}{\underline{\lambda}^2 n\kappa_1}$, and we set input parameters $\omega = \hat{\sigma}\sqrt{(4+\varepsilon)\log p}$ for any $\varepsilon \in (0,1)$ and $\tau \le \hat{\sigma}^2 k\log p$ in Algorithm 1, then*

$$\psi^{\mathrm{sparse}}_{\omega,\tau}(X_1,X_2,Y_1,Y_2) \xrightarrow{\text{a.s.}} 1.$$

The size and power controls in Theorems 1 and 3 jointly provide an upper bound on the minimax detection threshold. Specifically, let $P^X_{\beta_1,\beta_2}$ be the conditional distribution of Y_1, Y_2 given X_1, X_2 under model (2). Conditionally on the design matrices X_1 and X_2 and given $k \in [p]$ and $\rho > 0$, the (conditional) *minimax risk* of testing $H_0 : \beta_1 = \beta_2$ against $H_1 : \theta = (\beta_1-\beta_2)/2 \in \Theta_{p,k}(\rho)$ is defined as

$$\mathcal{M}_X(k,\rho) := \inf_{\psi}\left\{\sup_{\beta\in\mathbb{R}^p} P^X_{\beta,\beta}(\psi\neq 0) + \sup_{\substack{\beta_1,\beta_2\in\mathbb{R}^p\\(\beta_1-\beta_2)/2\in\Theta_{p,k}(\rho)}} P^X_{\beta_1,\beta_2}(\psi\neq 1)\right\},$$

where we suppress all dependences on the dimension of data for notational simplicity and the infimum is taken over all $\psi : (X_1,Y_1,X_2,Y_2) \mapsto \{0,1\}$. If $\mathcal{M}_X(k,\rho) \xrightarrow{\mathrm{p}} 0$, there exists a test ψ that with asymptotic probability 1 correctly differentiates the null and the alternative. On the other hand, if $\mathcal{M}_X(k,\rho) \xrightarrow{\mathrm{p}} 1$, then asymptotically no test can do better than a random guess. The following corollary provides an upper bound on the signal size ρ for which the minimax risk is asymptotically zero.

COROLLARY 4. *Under conditions (C1), (C2) and (S1), we further assume* $k \in [p]$ *and that* (10) *holds. If* $\rho^2 \geq \frac{7\sigma^2 k \log p}{\underline{\lambda}^2 n\kappa_1}$, *and we set input parameters* $\omega = \hat{\sigma}\sqrt{(4+\varepsilon)\log p}$ *for any* $\varepsilon \in (0,1]$ *and* $\tau \in (0, \hat{\sigma}^2 k \log p]$ *in Algorithm 1, then*

$$\mathcal{M}_X(k,\rho) \leq \sup_{\beta\in\mathbb{R}^p} P^X_{\beta,\beta}(\psi^{\mathrm{sparse}}_{\omega,\tau} \neq 0) + \sup_{\substack{\beta_1,\beta_2\in\mathbb{R}^p \\ (\beta_1-\beta_2)/2\in\Theta_{p,k}(\rho)}} P^X_{\beta_1,\beta_2}(\psi^{\mathrm{sparse}}_{\omega,\tau} \neq 1) \xrightarrow{\text{a.s.}} 0$$

Corollary 4 shows that the test $\psi^{\mathrm{sparse}}_{\omega,\tau}$ has an asymptotic detection limit, measured in $\|\beta_1-\beta_2\|_2$, of at most $\{\frac{7\sigma^2 k\log p}{\underline{\lambda}^2 n\kappa_1}\}^{1/2}$ for all k satisfying (10). While (10) is satisfied for $k \leq p^\alpha$ with any $\alpha \in [0,1)$, the detection limit upper bound shown in Corollary 4 is suboptimal when $\alpha > 1/2$, as we will see later in Theorem 6. On the other hand, Theorem 5 below shows that when $\alpha < 1/2$, the detection limit of $\psi^{\mathrm{sparse}}_{\omega,\tau}$ is essentially optimal for a large subclass of covariance matrices. For some $D > 0$ (which we allow to diverge as $p \to \infty$), we write $\mathrm{RowSp}(D) \subseteq \mathbb{R}^{p\times p}$ for the subset of $p \times p$ matrices having at most D nonzero elements in each row and define

$$\mathcal{C}(D) := \left\{\Sigma \in \mathcal{C} : \Sigma = \Sigma_0 + \Gamma \text{ for } \Sigma_0 \in \mathrm{RowSp}(D) \text{ and } \|\Gamma\|_{\max} \leq \frac{D}{k\log^2 p}\right\}. \tag{12}$$

The class $\mathcal{C}(D)$ consists of matrices admitting a sparse plus noise decomposition, and contains many common covariance matrices for relatively small choice of D. For instance, if Σ is a banded matrix, we may take D to be its bandwidth and $\Gamma = 0$. When $\Sigma = (\Sigma_{j,\ell})_{j,\ell\in[p]} = (\varrho^{|j-\ell|})_{j-\ell\in[p]}$ has an auto-regressive structure, we may take $D = (\log k + \log\log^2 p)/\log(varrho)$. Another example is when $\Sigma = V\Lambda V^\top + \Xi$ has a spiked covariance structure such that V is uniformly sampled from $\mathbb{O}^{p\times r}$ and $\Lambda, \Xi \succeq 0$ are diagonal (this is commonly encountered in e.g. factor analysis). In this case, each row of V has ℓ_2 norm bounded by $\sqrt{(r\log p)/p}$ with high probability, so $\|V\Lambda V^\top\|_{\max} \leq (\overline{\lambda} r\log p)/p$ and hence $\Sigma \in \mathcal{C}(D)$ with $D = \max\{1, (\overline{\lambda} r\log^3 p)/p\}$.

THEOREM 5. *Under conditions (C1) and (C2), if further assume* $\Sigma \in \mathcal{C}(D)$ *for some* $D > 0$, $k \leq p^\alpha$ *for some* $\alpha \in [0, 1/2)$ *and* $\rho^2 \leq \frac{(1-2\alpha-\varepsilon)\sigma^2 k\log p}{8Dn\kappa_1}$ *for some* $\varepsilon \in (0, 1-2\alpha]$, *then* $\mathcal{M}_X(k,\rho) \xrightarrow{\text{a.s.}} 1$.

For any fixed $\alpha < 1/2$, Theorem 5 shows that for designs having covariance matrix in $\mathcal{C}(D)$, if the squared signal ℓ_2 norm is a factor of $56D/\{\underline{\lambda}^2(1-2\alpha-\varepsilon)\}$ smaller than what can be detected by $\psi^{\mathrm{sparse}}_{\omega,\tau}$ shown in Corollary 4, then all tests are asymptotically powerless in differentiating the null from the alternative. In other words, in the sparse regime where $k \leq p^\alpha$ for $\alpha < 1/2$, the test $\psi^{\mathrm{sparse}}_{\omega,\tau}$ has a minimax optimal detection limit measured in $\|\beta_1-\beta_2\|_2$, up to constants depending on $\alpha, \underline{\lambda}$ and D only.

It is illuminating to relate the above results with the corresponding ones in the one-sample problem in the sparse regime ($\alpha < 1/2$). Let X be an $n \times p$ matrix with independent $N(0,1)$ entries and $Y = X\beta + \epsilon$ for $\epsilon \mid X \sim N(0, I_n)$, and we consider the one-sample problem to test $H_0 : \beta = 0$ against $H_1 : \beta \in \Theta_{p,k}(\rho)$. Theorem 2 and 4 of Arias-Castro, Candès and Plan (2011) state that under the additional assumption that all nonzero entries of β have equal absolute values, the detection limit for the one-sample problem is at $\rho \asymp \sqrt{\frac{k\log p}{n}}$, up to constants depending on α. Thus, when $\underline{\lambda}$ and D are constants, Corollary 4 and Theorem 5 suggest that the two-sample problem with model (2) has up to multiplicative constants the same detection limit as the one-sample problem with sample size

$$n\kappa_1 = \frac{nr}{(1+r)^2(1+s)}, \tag{13}$$

which unveils how this 'effective sample size' depends on the relative proportions between sample sizes n_1, n_2 and the dimension p of the problem. It is to be expected that the effective sample size is proportional to m, which is the number of observations constructed from X_1 and X_2 in W. More intriguingly, (13) also gives a precise characterization of how the effective sample size depends on the imbalance between the number of observations in X_1 and X_2. For a fixed $n = n_1 + n_2$, $n\kappa_1$ is maximized when $n_1 = n_2$ and converges to $n_1 m/n$ (or $n_2 m/n$) if $n_1/n \to 0$ (or $n_2/n \to 0$).

3.2. *Dense case.* We now turn our attention to our second test, $\psi_\eta^{\mathrm{dense}}$. The following theorem states a sufficient signal ℓ_2 norm size for which $\psi_\eta^{\mathrm{dense}}$ is asymptotically powerful in distinguishing the null from the alternative.

THEOREM 6. *Under Conditions (C1), (C2) and (S2), we let* $\eta = \hat{\sigma}^2\big(m + 2\sqrt{(2+\varepsilon)m\log p} + 2(1+\varepsilon)\log p\big)$ *for any* $\varepsilon \in (0,5)$. *We further assume* $k \in [p]$, $\rho^2 \geq \frac{2\sigma^2\sqrt{m\log p}}{n\kappa_1 \underline{\lambda}}$ *and that* (10) *is satisfied.*

(a) *If* $\beta_1 = \beta_2$, *then* $\psi_\eta^{\mathrm{dense}}(X_1, X_2, Y_1, Y_2) \xrightarrow{\text{a.s.}} 0$.
(b) *If* $\theta = (\beta_1 - \beta_2)/2 \in \Theta_{p,k}(\rho)$, *then* $\psi_\eta^{\mathrm{dense}}(X_1, X_2, Y_1, Y_2) \xrightarrow{\text{a.s.}} 1$.

Consequently, $\mathcal{M}_X(k,\rho) \xrightarrow{\text{a.s.}} 0$.

Theorem 6 indicates that the sufficient signal ℓ_2 norm for asymptotic powerful testing via $\psi_\eta^{\mathrm{dense}}$ does not depend upon the sparsity level. While the above result is valid for all $k \in [p]$ such that (10) holds, it is more interesting in the dense regime where $k \geq p^{1/2}$. More precisely, by comparing Theorems 6 and 4, we see that if $k^2 \log p > m$ and $k\log(ep/k) \leq n/(2C_{s,r})$, the test $\psi_\eta^{\mathrm{dense}}$ has a smaller provable detection limit than $\psi_{\omega,\tau}^{\mathrm{sparse}}$. In our asymptotic regime (C2), $m \asymp n \asymp p$, so $\frac{2\sqrt{m\log p}}{n\kappa_1}$ is, up to constants depending on s and r, of order $p^{-1/2}\log^{1/2} p$. The following theorem points out that when $\Sigma \in \mathcal{C}(D)$ for some constant D, the detection limit of $\psi_\eta^{\mathrm{dense}}$ is minimax optimal up to poly-logarithmic factors in the dense regime.

THEOREM 7. *Under conditions (C1) and (C2), if we further assume* $\Sigma \in \mathcal{C}(D)$ *for some* $D > 0$, $p^{1/2} \leq k \leq p^\alpha$ *for some* $\alpha \in [1/2, 1)$ *and* $\rho^2 = o(p^{-1/2}\min\{\log^{-1/2}(ep/k), D^{-3/2}\})$, *then* $\mathcal{M}_X(k,\rho) \xrightarrow{\text{a.s.}} 1$.

4. Numerical studies. In this section, we study the finite sample performance of our proposed procedures via numerical experiments. Unless otherwise stated, the data generating mechanism for all simulations in this section is as follows. We first generate design matrices X_1 and X_2 with independent $N(0,1)$ entries. Then, for a given sparsity level k and a signal strength ρ, set $\Delta = (\Delta_j)_{j\in[p]}$ so that $(\Delta_1, \ldots, \Delta_k)^\top \sim \rho\,\mathrm{Unif}(\mathcal{S}^{k-1})$ and $\Delta_j = 0$ for $j > k$. We then draw $\beta_1 \sim N_p(0, I_p)$ and define $\beta_2 := \beta_1 + \Delta$. Finally, we generate Y_1 and Y_2 as in (2), with $\epsilon_1 \sim N_{n_1}(0, I_{n_1})$ and $\epsilon_2 \sim N_{n_2}(0, I_{n_2})$ independent of each other.

In Section 4.1, we supply the oracle value of $\hat{\sigma}^2 = 1$ to our procedures to check whether their finite sample performance is in accordance with our theory. In all subsequent subsections where we compare our methods against other procedures, we estimate the noise variance σ^2 with the method-of-moments estimator proposed by Dicker (2014). We implement our estimators $\psi_{\omega,\tau}^{\mathrm{sparse}}$ and $\psi_\eta^{\mathrm{dense}}$ on standardized data $X_1/\hat{\sigma}$, $X_2/\hat{\sigma}$, $Y_1/\hat{\sigma}$ and $Y_2/\hat{\sigma}$ with the tuning parameters $\omega = 2\hat{\sigma}\sqrt{\log p}$, $\tau = \hat{\sigma}^2 \log p$ and $\eta = \hat{\sigma}^2(m + \sqrt{8m\log p} + 4\log p)$ as suggested by Theorems 1, 3 and 6.

4.1. *Effective sample size in two-sample testing.* We first investigate how the empirical power of our test $\psi^{\mathrm{sparse}}_{\lambda,\tau}$ relies on various problem parameters. In light of our results in Theorems 1 and 3, we define

$$\nu := \frac{rn\rho^2}{\sigma^2(1+s)(1+r)^2 k \log p}, \tag{14}$$

where $s := p/m$ and $r := n_1/n_2$. Note that in the asymptotic regime (C2), we have $\nu \to n\kappa_1\rho^2/(\sigma^2 k \log p)$. As discussed after Theorem 3, $rn/\{(1+s)(1+r)^2\}$ in the definition of ν is asymptotically $n\kappa_1$ and can be viewed as the effective sample size in the testing problem. In Figure 1, we plot the estimated test power of $\psi^{\mathrm{sparse}}_{\lambda,\tau}$ against ν over 100 Monte Carlo repetitions for $n = 1000$, $k = 10$, $\rho \in \{0, 0.2, \ldots, 2\}$ and various values of p and n_1. In the left panel of Figure 1, p ranges from 100 to 900, which corresponds to s from $1/9$ to 9. As for the right panel, we vary n_1 from 100 to 900, which corresponds with an r varying between $1/9$ and 9. In both panels, the power curves for different s and r values overlap each other, with the phase transition all occurring at around $\nu \approx 1.5$. This conforms well with the effective sample size and the detection limit articulated in our theory.

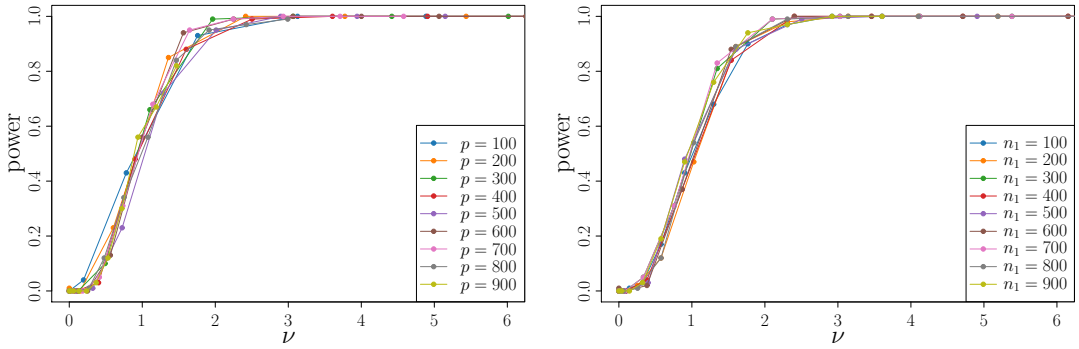


Fig 1: Power function of $\psi^{\mathrm{sparse}}_{\lambda,\tau}$, estimated over 100 Monte Carlo repetitions, plotted against ν, as defined in (14), in various parameter settings. Left panel: $n_1 = n_2 = 500$, $p \in \{100, 200, \ldots, 900\}$, $k = 10$, $\rho \in \{0, 0.2, \ldots, 2\}$. Right panel: $n_1 \in \{100, 200, \ldots, 900\}$, $n_2 = 1000 - n_1$, $p = 400$, $k = 10$, $\rho \in \{0, 0.2, \ldots, 2\}$.

4.2. *Comparison with other methods.* Next, we compare the performance of our procedures against competitors in the existing literature. The only methods we were aware of that could allow for dense regression coefficients β_1 and β_2 were those proposed by Zhu and Bradic (2016) and Charbonnier, Verzelen and Villers (2015). In addition, we also include in our comparisons the classical likelihood ratio test, denoted by ψ^{LRT}, which rejects the null when the F-statistic defined in (4) exceeds the upper α-quantile of an $F_{p,n-2p}$ distribution. Note that the likelihood ratio test is only well-defined if $p < \min\{n_1, n_2\}$. The test proposed by Zhu and Bradic (2016), which we denote by ψ^{ZB}, requires that $n_1 = n_2$ (when the two samples do not have equal sample size, a subset of the larger sample would be discarded for the test to apply). Specifically, writing $X_+ := X_1 + X_2$, $X_- := X_1 - X_2$ and $Y_+ := Y_1 + Y_2$, ψ^{ZB} first estimates $\gamma = (\beta_1 + \beta_2)/2$ and

$$\Pi := \{\mathbb{E}(X_+^\top X_+)\}^{-1}\mathbb{E}(X_+^\top X_-)$$

by solving Dantzig-Selector-type optimization problems. Then based on the obtained estimators $\hat{\gamma}$ and $\hat{\Pi}$, ψ^{ZB} proceeds to compute a test statistic

$$T_{\mathrm{ZB}} := \frac{\|\{X_- - X_+\hat{\Pi}\}^\top\{Y_+ - X_+\hat{\gamma}\}\|_\infty}{\|Y_+ - X_+\hat{\gamma}\|_2}.$$

Their test rejects the null if the test statistic exceeds an empirical upper-α-quantile (obtained

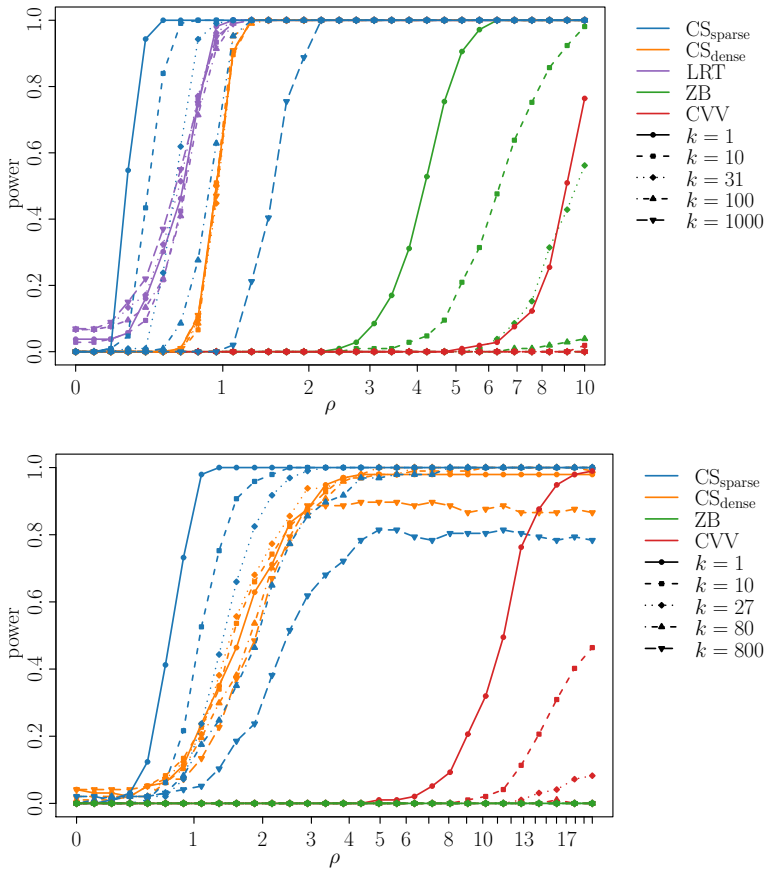


Fig 2: Power comparison of different methods at different sparsity levels $k \in \{1, 10, \lfloor p^{1/2} \rfloor, 0.1p, p\}$ and different signal ℓ_2 norm ρ on a logarithmic grid (noise variance $\sigma^2 = 1$). Top panel: $n_1 = n_2 = 1200$, $p = 1000$, $\rho \in [0, 10]$; bottom panel: $n_1 = n_2 = 500$, $p = 800$, $\rho \in [0, 20]$.

via Monte-Carlo simulation) of $\|\xi\|_\infty$ for $\xi \sim N(0, \{X_- - X_+\hat{\Pi}\}^\top\{X_- - X_+\hat{\Pi}\})$. As the estimation of Π involves solving a sequence of p Dantzig Selector problems, which is often time-consuming, we have implemented ψ^{ZB} with the oracle choice of $\hat{\Pi} = \Pi$, which is equal to I_p when covariates in the two design matrices X_1 and X_2 follow independent centred distribution with the same covariance matrix. The test proposed by Charbonnier, Verzelen and Villers (2015), denoted here by ψ^{CVV}, first performs a LARS regression (Efron et al., 2004) of concatenated response $Y = (Y_1^\top, Y_2^\top)^\top$ against the block design matrix

$$\begin{pmatrix} X_1 & X_1 \\ X_2 & -X_2 \end{pmatrix}$$

to obtain a sequence of regression coefficients $\hat{b} = (\hat{b}_1, \hat{b}_2) \in \mathbb{R}^{p+p}$. Then for every $\hat{b}$ on the LARS solution path with $\|\hat{b}\|_0 \leq \min\{n_1, n_2\}/2$, they restrict the original testing problem into the subset of coordinates where either $\hat{b}_1$ or $\hat{b}_2$ is non-zero, and form test statistics based on the Kullback–Leibler divergence between the two samples restricted to these coordinates. The sequence of test statistics is then compared with Bonferroni-corrected thresholds at size α. For both the ψ^{LRT} and ψ^{CVV}, we set $\alpha = 0.05$.

Figure 2 compares the estimated power, as a function of $\|\beta_1 - \beta_2\|_2$, of $\psi^{\mathrm{sparse}}_{\omega,\tau}$ and $\psi^{\mathrm{dense}}_{\eta}$ against that of ψ^{LRT}, ψ^{ZB} and ψ^{CVV}. We ran all methods on the same 100 datasets for each set of parameters. We performed numerical experiments in two high-dimensional settings with different sample-size-to-dimension ratio: $p = 1000$, $n_1 = n_2 = 1200$ in the left panel and $p = 800$, $n_1 = n_2 = 500$ in the right panel. Here, we took $n_1 = n_2$ to maximize the power of ψ^{ZB}. Also, since the likelihood ratio test requires $p < \min\{n_1, n_2\}$, it is only implemented in the left panel. For each experiment, we varied k in the set $\{1, 10, \lfloor p^{1/2} \rfloor, 0.1p, p\}$ to examine different sparsity levels.

We see in Figure 2 that both $\psi^{\mathrm{sparse}}_{\lambda,\tau}$ and $\psi^{\mathrm{dense}}_{\eta}$ showed promising finite sample performance. Both our tests did not produce any false positives under the null when $\rho = 0$, and showed better power compared to ψ^{ZB} and ψ^{CVV}. In the more challenging setting of the right panel with $p > \max\{n_1, n_2\}$, it takes a signal ℓ_2 norm more than 10 times smaller than that of the competitors for our test $\psi^{\mathrm{sparse}}_{\omega,\tau}$ to reach power of almost 1 in the sparsest case. Note though, in the densest case on the right panel ($k = 800$), $\psi^{\mathrm{sparse}}_{\omega,\tau}$ and $\psi^{\mathrm{dense}}_{\eta}$ did not have saturated power curves, because noise variance is over-estimated by $\hat{\sigma}^2$ in this setting.

We also observe that the power of $\psi^{\mathrm{sparse}}_{\omega,\tau}$ has a stronger dependence on the level k than that of $\psi^{\mathrm{dense}}_{\eta}$. For $k \leq \sqrt{p}$, $\psi^{\mathrm{sparse}}_{\omega,\tau}$ appears much more sensitive to the signal size. As k increases, $\psi^{\mathrm{dense}}_{\eta}$ eventually outperforms $\psi^{\mathrm{sparse}}_{\lambda,\tau}$, which is consistent with our observed phase transition behaviour as discussed after Theorem 6. It is interesting to note that when the likelihood ratio test is well-defined (left panel), it has better power than $\psi^{\mathrm{dense}}_{\eta}$. This is partly due to the fact that the theoretical choice of threshold η is relatively conservative to ensure asymptotic size of the test is 0 almost surely. In comparison, the rejecting threshold for the likelihood ratio test is chosen to have (p fixed and $n \to \infty$) asymptotic size of $\alpha = 0.05$, and the empirical size is sometimes observed to be larger than 0.08.

As remarked at the beginning of Section 1.2, the complementary sketching transforming can potentially be combined with other one-sample global testing procedure to obtain a two-sample test. Figure 3 illustrates this by comparing our methods with a two-stage procedure combining the complementary sketching transformation with the one-sample test proposed in Carpentier et al. (2019), which we call ψ^{CCCTW}. We remark that the test in Carpentier et al. (2019) requires the knowledge of the sparsity k and involves an unspecified parameter C_*. In our experience, the optimal choice of C_* seems to vary with different sparsity levels. As such, we have implemented ψ^{CCCTW} by choosing C_* in each simulation setting to maximize the power subject to a size constraint of $\alpha = 0.05$ (specifically, for $k = 1, 10, 31, 100, 1000$, we have chosen $C_* = 0.30, 0.67, 1.56, 3.37, 2.84$ respectively). We note that even granting ψ^{CCCTW} access to the additional sparsity parameter and this strong oracle parameter choice, $\psi^{\mathrm{sparse}}_{\omega,\tau}$ and $\psi^{\mathrm{dense}}_{\eta}$ are still competitive and in most cases outperforming ψ^{CCCTW} in the sparse and dense regimes respectively. We attribute this difference in performance to the fact that the matrix W after complementary sketching transformation does not satisfy typical design conditions (such as independent rows) assumed in most one-sample testing literature. As a result, two-stage methods such as the ψ^{CCCTW} test may suffer from power loss due to model misspecification.

4.3. *More general data generating mechanisms.* We have thus far focused on the case of Gaussian random design X_1, X_2 with identity covariance and Gaussian regression noises

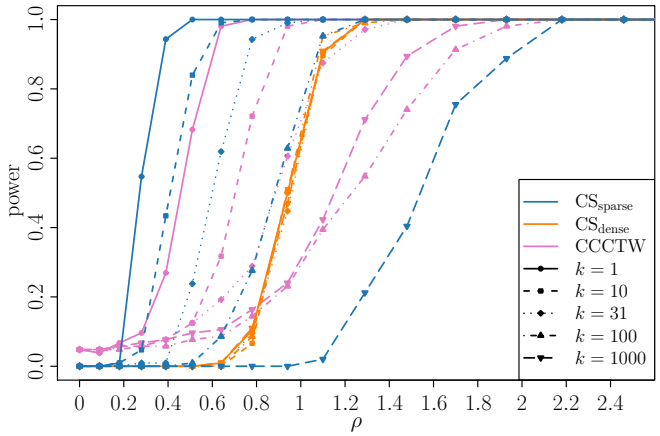


Fig 3: Power comparison of methods constructed by using different one-sample testing procedures after complementary sketching transformation. Parameters: $n_1 = n_2 = 1200$, $p = 1000$, $\rho \in [0, 2.5]$, $k \in \{1, 10, 31, 100, 1000\}$.

ϵ_1, ϵ_2. However, as our proposed testing procedures can still be used under more general data generating mechanisms. We consider the following four setups:

(a) Correlated design: assume rows of X_1 and X_2 are independently drawn from $N(0, \Sigma)$ with $\Sigma = (2^{-|j_1 - j_2|})_{j_1, j_2 \in [p]}$.

(b) Rademacher design: assume entries of X_1 and X_2 are independent Rademacher random variables.

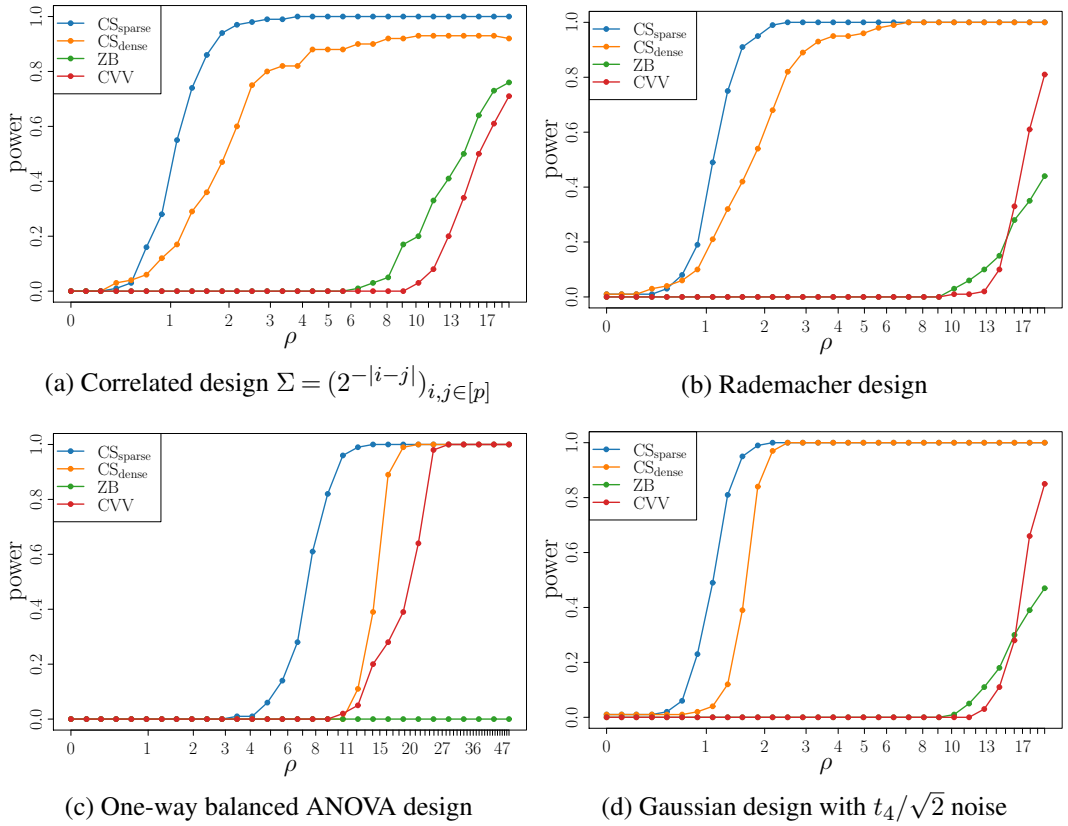
(c) One way balanced ANOVA design: assume $d_1 := n_1/p$ and $d_2 := n_2/p$ are integers and X_1 and X_2 are block diagonal matrices

$$X_1 = \begin{pmatrix} \mathbf{1}_{d_1} & & \\ & \ddots & \\ & & \mathbf{1}_{d_1} \end{pmatrix} \qquad X_2 = \begin{pmatrix} \mathbf{1}_{d_2} & & \\ & \ddots & \\ & & \mathbf{1}_{d_2} \end{pmatrix},$$

where $\mathbf{1}_d$ is an all-one vector in $\mathbb{R}^d$.

(d) Heavy tailed noise: we generate both ϵ_1 and ϵ_2 with independent $t_4/\sqrt{2}$ entries. Note that the $\sqrt{2}$ denominator standardizes the noise to have unit variance, to ensure easier comparison between settings.

In setups (a) to (c), we keep $\epsilon_1 \sim N_{n_1}(0, I_{n_1})$ and $\epsilon_2 \sim N_{n_2}(0, I_{n_2})$ and in setup (d), we keep X_1 and X_2 to have independent $N(0, 1)$ entries. Note that in setup (a) the covariance matrix belongs to $\mathcal{C}(D)$ with $D \asymp \log k + \log\log^2 p$. Figure 4 compares the performance of $\psi^{\text{sparse}}_{\lambda,\tau}$, $\psi^{\text{dense}}_{\eta}$ with that of ψ^{ZB} and ψ^{CVV}. In all settings, we set $n_1 = n_2 = 500$ and $k = 10$. In settings (a), (b) and (d), we choose $p = 800$ and ρ from 0 to 20. In setting (c), we choose $p = 250$ and ρ from 0 to 50. We see that $\psi^{\text{sparse}}_{\omega,\tau}$ is robust to model misspecification and exhibits good power in all settings. The test $\psi^{\text{desne}}_{\eta}$ is robust to non-normal design and noise, but exhibits a slight reduction in power in a correlated design. The advantage of $\psi^{\text{sparse}}_{\omega,\tau}$ and $\psi^{\text{dense}}_{\eta}$ over competing methods is least significant in the ANOVA design in setting (c), where each row vector of the design matrices has all mass concentrated in one coordinate. In all other settings where the rows of the design matrices are more 'incoherent' in the sense that all coordinate have similar magnitude, $\psi^{\text{sparse}}_{\omega,\tau}$ and $\psi^{\text{dense}}_{\eta}$ start having nontrivial power at a signal ℓ_2 norm 10 to 20 times smaller than that of the competitors.



(a) Correlated design $\Sigma = (2^{-|i-j|})_{i,j\in[p]}$

(b) Rademacher design

(c) One-way balanced ANOVA design

(d) Gaussian design with $t_4/\sqrt{2}$ noise

Fig 4: Power functions of different methods in models with non-Gaussian design or non-Gaussian noise, plotted against signal ℓ_2 norm ρ on a logarithmic grid. Details of the models are in Section 4.3.

5. Analysing a single-cell dataset. Here, we illustrate the applicability of our methodology on a single-cell RNA sequencing dataset from Suo et al. (2022). The dataset consists of the logarithmic normalized gene expression levels of 33538 genes measured in 91298 cells. For simplicity, we focus on the subset of $n = 7816$ cells that have been labelled as either CD4$^+$ T cells ($n_1 = 4852$) or T regulatory (TREG) cells ($n_2 = 2964$), two closely related T cell subpopulations, and only keep the $p = 4123$ genes whose log normalized expression variance is at least 1 in the two cell subpopulations. We are interested in testing for difference in the gene regulatory networks in the two cell subpopulations. This can be modelled by the difference in their respective Gaussian graphical model networks and tested by comparing the nodewise regression coefficients of each gene against the remaining genes in CD4$^+$ T cells and TREG cells. The left column of Table 2 summarizes the genes that report significant difference in their nodewise regression coefficients from our complementary sketching method, which we call 'master regulators'. Among the nine genes identified to have significant difference in their nodewise regression coefficients, FOXP3, CTLA4, IL2RA, IL7R, IKZF2, CD83, ANXA1 are all known to be important regulators, from several independent pathways, essential for the function of the TREG cell type (Bayer, Yu and Malek, 2007; Walker, 2013; Kim et al., 2015; Doebbeler et al., 2018; Toomer et al., 2019; Bai et al., 2020).

In addition to identifying the master regulator genes, a slight modification of our algorithm also allows us to identify their top interacting partners insofar as the two T cell subpopulations are concerned. Specifically, after performing complementary sketching to obtain sketched design W and response Z in Step 4 of Algorithm 1, we may compute the Lasso solution path

master regulators	top interacting partners
IKZF2	MT-ND4L, **HLA-B**, MT-ATP8, **ETS1**, **FYB1**, **JUNB**, RNF213, **HLA-C**
FOXP3	MT-ND4L, MT-ATP8, **S100A4**, **CD96**, **ISG20**, BIRC2, **SRSF4**, **GZMM**
CD83	**HSPA1A**, **NFKBIA**, **RGS2**, **MTRNR2L12**, **NR4A1**, **PSMC3**, BAG3, **SRP9**
IL2RA	**ENO1**, **ARID5B**, RPL23, CDC42, **CREM**, **CISH**, **GADD45B**, **PMAIP1**
ANXA1	**JUNB**, **JUN**, **TNFAIP3**, **FOSB**, **CALM2**, **ABLIM1**, **RGS2**, **CHI3L2**
CD8A	**FTL**, **SLC25A3**, **CD8B**, COTL1, **PTPRCAP**, PCBP1, **STMN1**, IGFBP2
CTLA4	**RGS1**, **GBP2**, **RPS10**, ZFP36L1, **TAGAP**, **STAT3**, **RPS4Y1**, **SRGN**
GNG8	RPL41, HSPB1, **OST4**, LTB, **TERF2IP**, **CUTA**, **PPDPF**, **IFITM1**
IL7R	RPL41, **RPL27A**, **VIM**, **TRBC2**, SLC25A6, **CORO1A**, **RPS26**, **TRAC**

TABLE 2

Genes with significant difference identified by the complementary sketching algorithm, together with their top eight interacting partners using graphical Lasso post complementary sketching. Genes that are identified to be significant by the Mann–Whitney–Wilcoxon test after Bonferroni correction are shown in bold.

(Tibshirani, 1996). The right column of Table 2 shows genes corresponding to the first eight nonzero coefficients entering the solution path, which can be interpreted as the top interacting partners of the master regulator genes.

Computationally, we remark that when applying Algorithm 1 to a differential network analysis setting, we can precompute an orthonormal basis spanning the orthogonal complement of the column span of $(X_1^\top, X_2^\top)^\top$, and obtain individual sketching matrices A for each nodewise regression by augmenting that basis with one additional vector. For example, on an 8-core 3.20 GHz desktop machine, our algorithm was able to test for all $p = 4123$ pairs of nodewise regressions in 1.6 hours (averaging 1.4 seconds per node). Our code and preprocessed dataset for the real data analysis are both available on GitHub[1].

It is interesting to contrast our analysis to the common differential-expression-based approach for identifying master regulator genes which determine the identity of different cell types. Differential expression analysis simply compares the expression levels of a gene in two different cell types, typically with the Mann–Whitney–Wilcoxon test. We have highlighted in bold in Table 2 all genes that are differentially expressed in $\mathrm{CD4}^+$ T cells and TREG cells (at 0.05 level after Bonferroni correction). It can be seen that all our master regulators are differentially expressed. However, differential expression analysis identifies a much larger set of genes, many potentially belonging to the same pathway and dependent on each other. Overall, our complementary sketching approach allows for more precise identification of the central players in gene regulatory networks.

6. Proof of main results.

PROOF OF THEOREM 1. By Condition (S1), we may work on the almost sure event $\Omega_\sigma := \{\hat{\sigma}/\sigma = 1 + \mathcal{O}(1)\}$. Under the null hypothesis where $\beta_1 = \beta_2$, we have $\theta = 0$ and therefore, $Z = W\theta + \xi = \xi \sim N_m(0, \sigma^2 I_m)$. In particular, $Q_j/\sigma \sim N(0,1)$ for all $j \in [p]$.

Thus, noting the independence of $\hat{\sigma}$ and the sample and employing a union bound, we have for $\omega = \hat{\sigma}\sqrt{(4+\varepsilon)\log p}$ and any $\tau > 0$ that

$$\mathbb{P}(T \geq \tau \mid \hat{\sigma}) \leq \sum_{j=1}^{p} \mathbb{P}(|Q_j|/\sigma \geq \omega/\sigma \mid \hat{\sigma}) \leq p \exp\bigl(-\omega^2/(2\sigma^2)\bigr).$$

[1] https://github.com/wangtengyao/compsket/

Keeping the preceding display in mind, by the independence of $\hat{\sigma}$ and the sample, we bound for p sufficiently large

$$\begin{aligned}\mathbb{P}(T \geq \tau \mid \Omega_\sigma) &\leq p \exp\bigl(-(1-\mathcal{O}(1))(2+\varepsilon/2)\log p\bigr) \\ &\leq p\exp(-(2+\varepsilon/4)\log p).\end{aligned}$$

Noting that Ω_σ is an almost sure event and that $p^{-1-\varepsilon/4}$ is summable for any $\varepsilon > 0$, the almost sure convergence in the theorem statement follows from the Borel–Cantelli lemma. □

PROOF OF THEOREM 3. By Proposition 2, it suffices to work with a deterministic sequence of W such that (9) and (11) holds, which we henceforth assume in this proof.

Define $\tilde{\theta} = (\tilde{\theta}_1, \ldots, \tilde{\theta}_p)^\top$ such that $\tilde{\theta}_j := \theta_j \|W_j\|_2$. Then, from (8), we have

$$Z = \tilde{W}\tilde{\theta} + \xi,$$

for $\xi \sim N_m(0, I_m)$. Write $Q := (Q_1, \ldots, Q_p)^\top$ and $S := \operatorname{supp}(\theta) = \operatorname{supp}(\tilde{\theta})$, then

$$Q_S = (\tilde{W}^\top Z)_S \sim N_k\bigl((\tilde{W}^\top \tilde{W})_{S,S}\tilde{\theta}_S,\, (\tilde{W}^\top \tilde{W})_{S,S}\bigr).$$

Our strategy will be to control $\|Q_S\|_2^2$. To this end, we first look at the quantity $\|(\tilde{W}^\top \tilde{W})_{S,S}^{-1/2} Q_S\|_2^2$, which has a noncentral chi-squared distribution $\chi_k^2(\|\tilde{\theta}_S^\top (\tilde{W}^\top \tilde{W})_{S,S}\tilde{\theta}_S\|_2^2)$. By (9) and (11), we have

$$\begin{aligned}\|\tilde{\theta}_S^\top (\tilde{W}^\top \tilde{W})_{S,S}\tilde{\theta}_S\|_2^2 \geq \|(\tilde{W}^\top \tilde{W})_{S,S}\|_{\mathrm{op}}\|\theta_S\|_2^2 &\geq \{\underline{\lambda} - \mathcal{O}(1)\} 4n\kappa_1\rho^2 \\ &\geq \frac{28\sigma^2 k\log p}{\underline{\lambda}},\end{aligned}$$

where we have used the fact that $\rho^2 \geq 7k\log p/(\underline{\lambda}^2 n\kappa_1)$ in the final bound. Thus, by Birgé (2001, Lemma 8.1), we have with probability at least $1 - p^{-2}$ that

$$\begin{aligned}&\|(\tilde{W}^\top \tilde{W})_{S,S}^{-1/2} Q_S\|_2^2 \\ &\qquad \geq k + \|\tilde{\theta}_S^\top (\tilde{W}^\top \tilde{W})_{S,S}\tilde{\theta}_S\|_2^2 - 2\sqrt{(2k + 4\|\tilde{\theta}_S^\top (\tilde{W}^\top \tilde{W})_{S,S}\tilde{\theta}_S\|_2^2)\log p} \\ &\qquad \geq \{1 - \mathcal{O}(1)\}\|\tilde{\theta}_S^\top (\tilde{W}^\top \tilde{W})_{S,S}\tilde{\theta}_S\|_2^2 - 4\|\tilde{\theta}_S^\top (\tilde{W}^\top \tilde{W})_{S,S}\tilde{\theta}_S\|_2\sqrt{\log p} \\ &\qquad \geq \frac{(6 - \mathcal{O}(1))\sigma^2 k\log p}{\underline{\lambda}}. \qquad (15)\end{aligned}$$

Consequently, by (15) and (11) again, we have with probability at least $1 - p^{-2}$ that

$$\begin{aligned}\|Q_S\|_2^2 &\geq \|(\tilde{W}^\top \tilde{W})_{S,S}\|_{\mathrm{op}}\|(\tilde{W}^\top \tilde{W})_{S,S}^{-1/2} Q_S\|_2^2 \\ &\geq \{\underline{\lambda} - \mathcal{O}(1)\}\frac{6\sigma^2 k\log p}{\underline{\lambda}} \geq (6 - \mathcal{O}(1))\sigma^2 k\log p. \qquad (16)\end{aligned}$$

By Condition (S1), we define the almost sure event $\Omega_\sigma := \{\hat{\sigma}/\sigma = 1 + \mathcal{O}(1)\}$ and $\omega_0 := \sigma\sqrt{(4+\varepsilon)\log p}$. We observe that on the event Ω_σ, $(\omega/\omega_0)^2 - 1 = (\hat{\sigma}/\sigma + 1)(\hat{\sigma}/\sigma - 1) = \mathcal{O}(1)$. From (16), using the tuning parameters $\omega = \hat{\sigma}\sqrt{(4+\varepsilon)\log p}$ and $\tau \leq \hat{\sigma}^2 k\log p$, we have, conditionally on Ω_σ, for sufficiently large p that with probability at least $1 - 2p^{-2}$,

$$T = \sum_{j=1}^p Q_j^2 \mathbb{1}_{\{|Q_j| \geq \omega\}} \geq \|Q_S\|_2^2 - k\omega_0^2\bigl(1 + (\omega/\omega_0)^2 - 1\bigr) \geq \hat{\sigma}^2 k\log p \geq \tau,$$

which would allow us to reject the null on the 'good' event Ω_σ. Keeping in mind that Ω_σ is an almost sure event, we proceed to bound

$$\mathbb{P}(T \le \tau) = \mathbb{P}(T \le \tau \mid \Omega_\sigma) \le 2p^{-2},$$

The desired almost sure convergence follows by the Borel–Cantelli lemma since $1/p^2$ is summable over $p \in \mathbb{N}$. □

PROOF OF THEOREM 5. By considering a trivial test $\tilde{\psi} \equiv 0$, we see that $\mathcal{M} \le 1$. Thus, it suffices to show that $\mathcal{M} \ge 1 - \mathcal{O}(1)$. Note that since $k \le p^{1/2}$, condition (10) is satisfied and hence by Proposition 2, it suffices to work with a deterministic sequence of X (and hence W) such that (9) and (11) holds, which we henceforth assume in this proof.

It is convenient to reparametrize the distributions in terms of $(\gamma, \theta) = ((\beta_1 + \beta_2)/2, (\beta_1 - \beta_2)/2)$ instead of (β_1, β_2). Define $Q^X_{\gamma,\theta} := P^X_{\beta_1,\beta_2}$. Let $L := (X_1^\top X_1 + X_2^\top X_2)^{-1}(X_2^\top X_2 - X_1^\top X_1)$ and π be the uniform distribution on

$$\Theta_0 := \{\theta \in \{k^{-1/2}\rho, -k^{-1/2}\rho, 0\}^p : \|\theta\|_0 = k\} \subseteq \Theta.$$

We write $Q_0 := Q^X_{0,0}$ and let $Q_\pi := \int_{\theta \in \Theta_0} Q^X_{L\theta,\theta}\, d\pi(\theta)$ denote the uniform mixture of $Q^X_{\gamma,\theta}$ for $\{(\gamma, \theta) : \theta \in \Theta_0, \gamma = L\theta\}$. Let $\mathcal{L} := dQ_\pi / dQ_0$ be the likelihood ratio between the mixture alternative Q_π and the simple null Q_0. We have that

$$\begin{aligned}
\mathcal{M} &\ge \inf_{\tilde{\psi}} \Big\{1 - (Q_0 - Q_\pi)\tilde{\psi}\Big\} = 1 - \frac{1}{2}\int \left|1 - \frac{dQ_\pi}{dQ_0}\right| dQ_0 \\
&\ge 1 - \frac{1}{2}\left\{\int \left(1 - \frac{dQ_\pi}{dQ_0}\right)^2 dQ_0\right\}^{1/2} \ge 1 - \frac{1}{2}\{Q_0(\mathcal{L}^2) - 1\}^{1/2}.
\end{aligned}$$

So it suffices to prove that $Q_0(\mathcal{L}^2) \le 1 + \mathcal{O}(1)$. Writing $\tilde{X}_1 = X_1 L + X_1$ and $\tilde{X}_2 = X_2 L - X_2$, by the definition of Q_π, we compute that

$$\begin{aligned}
\mathcal{L} = \int \frac{dQ^X_{L\theta,\theta}}{dQ_0}\, d\pi(\theta) &= \int \frac{e^{-\frac{1}{2}(\|Y_1 - X_1 L\theta - X_1\theta\|^2 + \|Y_2 - X_2 L\theta + X_2\theta\|^2)}}{e^{-\frac{1}{2}(\|Y_1\|^2 + \|Y_2\|^2)}}\, d\pi(\theta) \\
&= \int e^{\langle \tilde{X}_1\theta, Y_1\rangle - \frac{1}{2}\|\tilde{X}_1\theta\|^2 + \langle \tilde{X}_2\theta, Y_2\rangle - \frac{1}{2}\|\tilde{X}_2\theta\|^2}\, d\pi(\theta).
\end{aligned}$$

For $\theta \sim \pi$ and some fixed $J_0 \subseteq [p]$ with $|J_0| = k$, let π_{J_0} be the distribution of θ_{J_0} conditional on $\operatorname{supp}(\theta) = J_0$. Let J, J' be independently and uniformly distributed on $\{J_0 \subseteq [p] : |J_0| = k\}$. Define $\tilde{\theta} := (\tilde{\theta}_1, \dots, \tilde{\theta}_p)^\top$ and $\tilde{\theta}' := (\tilde{\theta}'_1, \dots, \tilde{\theta}'_p)^\top$ such that $\tilde{\theta}_j := \theta_j \|W_j\|_2$ and $\tilde{\theta}'_j := \theta'_j \|W_j\|_2$. Since $\Sigma \in \mathcal{C}(D)$, we can write $\Sigma = \Sigma_0 + \Gamma$ for $\Sigma_0 \in \operatorname{RowSp}(D)$ and $\|\Gamma\|_{\max} = \mathcal{O}(D/(k \log p))$. Also, since Σ is symmetric and $\|\Sigma\|_{\max} = \|\operatorname{diag}(\Sigma)\|_{\max} = 1$, we may assume without loss of generality that Σ_0 is symmetric and $\|\Sigma_0\|_{\max} \le 1$. By Fubini's theorem and Lemmas 11 and 10, we have

$$\begin{aligned}
Q_0(\mathcal{L}^2) &= \iint_{\theta,\theta'} e^{\frac{1}{2}\|\tilde{X}_1(\theta+\theta')\|^2 - \frac{1}{2}\|\tilde{X}_1\theta\|^2 - \frac{1}{2}\|\tilde{X}_1\theta'\|} \\
&\qquad\qquad \times e^{\frac{1}{2}\|\tilde{X}_2(\theta+\theta')\|^2 - \frac{1}{2}\|\tilde{X}_2\theta\|^2 - \frac{1}{2}\|\tilde{X}_2\theta'\|^2}\, d\pi(\theta)\, d\pi(\theta') \\
&= \iint_{\theta,\theta'} e^{\theta^\top(\tilde{X}_1^\top\tilde{X}_1 + \tilde{X}_2^\top\tilde{X}_2)\theta'}\, d\pi(\theta)\, d\pi(\theta') \\
&= \iint_{\theta,\theta'} e^{\theta^\top W^\top W\theta'}\, d\pi(\theta)\, d\pi(\theta') \\
&\le \{\mathbb{E}(e^{2\tilde{\theta}^\top(\tilde{W}^\top\tilde{W} - \Sigma_0)\tilde{\theta}'})\}^{1/2}\{\mathbb{E}(e^{2\tilde{\theta}^\top\Sigma_0\tilde{\theta}'})\}^{1/2}, \qquad (17)
\end{aligned}$$

where we apply the Cauchy–Schwarz inequality in the final inequality. We now bound the two factors in the final expression separately. By (9), we have

$$\vartheta := \max\bigl\{\max_{j\in J}|\tilde{\theta}_j|, \max_{j\in J'}|\tilde{\theta}'_j|\bigr\} \leq (1+\mathcal{o}(1))\sqrt{\frac{4n\kappa_1\rho^2}{k}}. \tag{18}$$

By (11), we have that

$$\begin{aligned}
\vartheta^2\|(\tilde{W}^\top\tilde{W}-\Sigma_0)_{J,J'}\|_{\mathrm{F}} &\leq \sqrt{k}\vartheta^2\|(\tilde{W}^\top\tilde{W}-\Sigma)_{J,J'}\|_{\mathrm{op}} + \vartheta^2\|\Gamma_{J,J'}\|_{\mathrm{F}} \\
&\leq \vartheta^2\bigl\{\sqrt{k}\|\tilde{W}^\top\tilde{W}-\Sigma\|_{2k,\mathrm{op}} + k\|\Gamma\|_{\max}\bigr\} \\
&\leq \frac{(4+\mathcal{o}(1))n\kappa_1\rho^2}{k}\biggl[C_{s,r}\biggl\{\overline{\lambda}\sqrt{\frac{k^2\log(ep)}{n}} + \overline{\lambda}^2\sqrt{\frac{k\log p}{n}}\biggr\} + \mathcal{o}\biggl(\frac{D}{\log p}\biggr)\biggr]. \qquad (19)
\end{aligned}$$

Since $\rho \leq \sqrt{\frac{(1-2\alpha-\varepsilon)k\log p}{4Dn\kappa_1}}$, we have from (19) that $\vartheta^2\|(\tilde{W}^\top\tilde{W}-\Sigma_0)_{J,J'}\|_{\mathrm{F}} = \mathcal{o}(1)$. Consequently, by Lemma 15, we have for sufficiently large p that

$$\begin{aligned}
\mathbb{E}(e^{2\tilde{\theta}^\top(\tilde{W}^\top\tilde{W}-\Sigma_0)\tilde{\theta}'}) &\leq 1 + C\vartheta^2\|(\tilde{W}^\top\tilde{W}-\Sigma_0)_{J,J'}\|_{\mathrm{F}}e^{4\vartheta^2\|(\tilde{W}^\top\tilde{W}-\Sigma_0)_{J,J'}\|_{\mathrm{F}}^2} \\
&= 1+\mathcal{o}(1). \qquad (20)
\end{aligned}$$

For the second factor on the right-hand side of (17), we have by Lemma 16 that

$$\mathbb{E}(e^{2\tilde{\theta}^\top\Sigma_0\tilde{\theta}'}) \leq \biggl\{1+\frac{Dk}{p}\bigl(\cosh(2\vartheta^2 D)-1\bigr)\biggr\}^k \leq \exp\biggl(\frac{Dk^2}{p}e^{2\vartheta^2 D}\biggr). \tag{21}$$

Since $\alpha \in [0,1/2)$ and $\rho^2 \leq \frac{(1-2\alpha-\varepsilon)k\log p}{8Dn\kappa_1}$, from (18), we deduce that

$$2\vartheta^2 D \leq (1-2\alpha-\varepsilon+\mathcal{o}(1))\log p$$

and hence $e^{2\vartheta^2 D}Dk^2/p = p^{-\varepsilon+\mathcal{o}(1)} = \mathcal{o}(1)$. So, from (17), (20) and (21) we have $Q_0(\mathcal{L}^2) \leq 1+\mathcal{o}(1)$, which completes the proof. □

Acknowledgements. This work was primarily supported by the Royal Society grant IEC/NSFC/170119. In addition, the research of FG was supported by NSFC grants 11701095 and 11690013 and that of TW was supported by EPSRC grant EP/T02772X/1. We would like to thank Chenqu Suo for suggesting the dataset and offering her valuable insights in the real data analysis. We thank the anonymous reviewers for helpful and constructive comments on an earlier draft.

SUPPLEMENTARY MATERIAL

Supplement: Supplementary material to 'Two-sample testing of high-dimensional linear regression coefficients via complementary sketching'
Additional proofs omitted in the main text and ancillary results.

REFERENCES

Arias-Castro, E., Candès, E. J. and Plan, Y. (2011) Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism. *Ann. Statist.*, **39**, 2533–2556.

Bai, F., Zhang, P., Fu, Y., Chen, H., Zhang, M., Huang, Q., Li, D., Li, B. and Wu, K. (2020) Targeting ANXA1 abrogates Treg-mediated immune suppression in triple-negative breast cancer. *Journal for immunotherapy of cancer*, **8**.

Bai, Z., Hu, J., Pan, G. and Zhou, W. (2015) Convergence of the empirical spectral distribution function of Beta matrices. *Bernoulli*, **21**, 1538–1574.

Bayer, A. L., Yu, A. and Malek, T. R. (2007) Function of the IL-2R for thymic and peripheral CD4+ CD25+ Foxp3+ T regulatory cells. *The Journal of Immunology*, **178**, 4062–4071.

Birgé, L. (2001) An alternative point of view on Lepskis method. In de Gunst, M and Klaassen, C. and van der Vaart, A. Eds,. *Lecture Notes-Monograph Series 36*, 113–133.

Cai, T. T., Liu, W. and Xia, Y. (2014) Two-sample test of high dimensional means under dependence. *J. Roy. Statist. Soc., Ser. B*, **76**, 349–372.

Carpentier, A., Collier, O., Comminges, L., Tsybakov, A. B. and Wang, Y. (2019) Minimax rate of testing in sparse linear regression. *Automation and Remote Control*, **80**, 1817–1834.

Carpentier, A. and Verzelen, N. (2021) Optimal sparsity testing in linear regression model. *Bernoulli*, **27**, 727–750.

Charbonnier, C., Verzelen, N. and Villers, F. (2015) A global homogeneity test for high-dimensional linear regression. *Electron. J. Statist.*, **9**, 318–382.

Chen, S. X., Li, J. and Zhong, P.-S. (2019) Two-sample and ANOVA tests for high dimensional means. *Ann. Statist.*, **47**, 1443–1474.

Chow, G. C. (1960) Tests of equality between sets of coefficients in two linear regressions. *Econometrica*, **28**, 591–605.

Dicker, L. H. (2014) Variance estimation in high-dimensional linear models. *Biometrika*, **101**, 269–284.

Doebbeler, M., Koenig, C., Krzyzak, L., Seitz, C., Wild, A., Ulas, T., Baßler, K., Kopelyanskiy, D., Butterhof, A., Kuhnt, C. et al. (2018) CD83 expression is essential for Treg cell differentiation and stability. *JCI insight*, **3**.

Donoho, D. and Jin, J. (2004) Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.*, **32**, 962–994.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004) Least angle regression. *Ann. Statist.*, **32**, 407–499.

Fan, J., Guo, S. and Hao, N. (2012) Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *J. Roy. Statist. Soc., Ser. B*, **74**, 37–65.

Homrighausen, D. and McDonald, D. (2013) The lasso, persistence, and cross-validation. In *International Conference on Machine Learning*, 1031–1039, PMLR.

Ingster, Y. I. (1997) Some problems of hypothesis testing leading to infinitely divisible distribution. *Math. Methods Statist.*, **6**, 47–69.

Ingster, Y. I., Tsybakov, A. and Verzelen, N. (2010) Detection boundary in sparse regression. *Electron. J. Statist.*, **4**, 1476–1526.

Kannel, W. B. and McGee, D. L. (1979) Diabetes and cardiovascular disease: the Framingham study. *J. Amer. Medical Assoc.*, **241**, 2035–2038.

Kim, H.-J., Barnitz, R. A., Kreslavsky, T., Brown, F. D., Moffett, H., Lemieux, M. E., Kaygusuz, Y., Meissner, T., Holderried, T. A., Chan, S. et al. (2015) Stable inhibitory activity of regulatory T cells requires the transcription factor Helios. *Science*, **350**, 334–339.

Kraft, P. and Hunter, D. J. (2009) Genetic risk prediction—are we there yet? *N. Engl. J. Medcine*, **360**, 1701–1703.

Mahoney, M. W. (2011) Randomized algorithms for matrices and data. *Found. Trends Mach. Learn.*, **3**, 123–224.

Reid, S., Tibshirani, R. and Friedman, J. (2016) A study of error variance estimation in lasso regression. *Statist. Sinica*, **26**, 35–67.

Städler, N. and Mukherjee, S. (2012) Two-sample testing in high-dimensional models. *arXiv preprint*, arxiv:1210.4584.

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

Suo, C., Dann, E., Goh, I., Jardine, L., Kleshchevnikov, V., Park, J.-E., Botting, R. A., Stephenson, E., Engelbert, J., Tuong, Z. K., Polanski, K., Yayon, N., Xu, C., Suchanek, O., Elmentaite, R., Conde, C. D., He, P., Pritchard, S., Miah, M., Moldovan, C., Steemers, A. S., Prete, M., Marioni, J. C., Clatworthy, M. R., Haniffa, M. and Teichmann, S. A. (2022) Mapping the developing human immune system across organs. *bioRxiv preprint*, doi.org/10.1101/2022.01.17.476665.

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

Toomer, K. H., Lui, J. B., Altman, N. H., Ban, Y., Chen, X. and Malek, T. R. (2019) Essential and non-overlapping IL-2Rα-dependent processes for thymic development and peripheral homeostasis of regulatory T cells. *Nature communications*, **10**, 1–16.

Walker, L. S. (2013) Treg and CTLA-4: two intertwining pathways to immune tolerance. *Journal of autoimmunity*, **45**, 49–57.

Xia, Y., Cai, T. and Cai, T. T. (2015) Testing differential networks with applications to the detection of gene-gene interactions. *Biometrika*, **102**, 247–266.

Xia, Y., Cai, T. and Cai, T. T. (2018) Two-sample tests for high-dimensional linear regression with an application to detecting interactions. *Statist. Sinica*, **28**, 63–92.

Xia, Y., Cai, T. T. and Sun, W. (2020) Gap: A general framework for information pooling in two-sample sparse inference. *J. Amer. Statist. Assoc.*, **115**, 1236–1250.

Zhu, Y. and Bradic, J. (2016) Two-sample testing in non-sparse high-dimensional linear models. *arXiv preprint*, arxiv:1610.04580.

Submitted to the Annals of Statistics

SUPPLEMENTARY MATERIAL TO 'TWO-SAMPLE TESTING OF HIGH-DIMENSIONAL LINEAR REGRESSION COEFFICIENTS VIA COMPLEMENTARY SKETCHING'

BY FENGNAN GAO[*,†] AND TENGYAO WANG[‡,§]

Fudan University[*], SCMS[†], London School of Economics and Political Science[‡] and University College London[§]

School of Data Science
Shanghai Center for Mathematical Sciences
Fudan University
Handan Road 220
Shanghai 200433, China
fngao@fudan.edu.cn

Department of Statistics
London School of Economics
Columbia House
69 Aldwych
London WC2B 4RR, United Kingdom
t.wang59@lse.ac.uk

APPENDIX A: PROOFS OF THE REST OF THEORETIC RESULTS

We collect here proofs of the remaining theoretical results omitted in the main text.

To prove Proposition 2, we need the following proposition, which considers tail bounds for $\|Wu\|$ for a fixed $u \in \mathcal{S}^{p-1}$ in the special case $\Sigma = I_p$.

PROPOSITION 8. *Under the conditions of Proposition 2, we further assume $\Sigma = I_p$. There exists a random sequence $(h_n)_n$ such that $h_n \xrightarrow{\text{a.s.}} 4\kappa_1$ for κ_1 defined in Lemma 13 and that for any sequence $(\delta_n)_n$ satisfying $\log(1/\delta_n) = o(p)$, we have for all large p that*

$$\mathbb{P}\left\{\left|\frac{1}{n}(W^\top W)_{1,1} - h_n\right| > (8+o(1))\sqrt{\frac{\log(1/\delta_n)}{n}}\left((\kappa_1+\kappa_2)\sqrt{n/p}+\kappa_1\right)\right\} \le \delta_n.$$

PROOF. Let $X = QT$ be the QR decomposition of X, which is almost surely unique if we require the upper-triangular matrix T to have non-negative entries on the diagonal.

Let Q_1 be the submatrix obtained from the first n_1 rows of Q. From Lemma 14, Q_1 and T are independent and T has independent entries distributed as $T_{j,j} = t_j > 0$ with $t_j^2 \sim \chi^2_{n-j+1}$ for $j \in [p]$ and $T_{j,k} = z_{j,k} \sim N(0,1)$ for $1 \le j < k \le p$.

Define $B := Q_1^\top Q_1$ and let $B = V\Lambda V^\top$ be its eigendecomposition, which is almost surely unique if we require the diagonal entries of Λ to be non-increasing and the diagonal entries of V to be nonnegative. By Lemma 14, Q is uniformly distributed on $\mathbb{O}^{n\times p}$, which means $Q \overset{\text{d}}{=} QH$ for any $H \in \mathbb{O}^{p\times p}$. Consequently, $Q_1 \overset{\text{d}}{=} Q_1 H$ and $B \overset{\text{d}}{=} H^\top B H = (H^\top V)\Lambda(H^\top V)^\top$. Since the group $\mathbb{O}^{p\times p}$ acts transitively on itself through left multiplication, the joint density of V and Λ must be a function of Λ only. In particular, V and Λ are independent.

AMS 2000 subject classifications: 62H15, 62J05.

Keywords and phrases: two-sample hypotheses testing, high-dimensional data, linear model, sparsity, minimax detection.

Note that $X_1 = Q_1 T$. Thus, $X_1^\top X_1 = T^\top B T$ and $X_2^\top X_2 = T^\top (I_p - B) T$. By Lemma 10, we have

$$W^\top W = 4X_1^\top A_1 A_1^\top X_1 = 4(X_1^\top X_1)(X_1^\top X_1 + X_2^\top X_2)^{-1}(X_2^\top X_2)$$
$$= 4T^\top B(I_p - B)T = 4T^\top V \Lambda (I_p - \Lambda) V^\top T. \tag{S.1}$$

Let $1 \geq \lambda_1 \geq \cdots \geq \lambda_p \geq 0$ be the diagonal entries of Λ. Define $a_j = \lambda_j(1-\lambda_j)$ for $j \in [p]$ and set $a := (a_1, \ldots, a_p)$. We can write $t_1^2 = s_1^2 + r_1^2$ with $s_1^2 \sim \chi_p^2$ and $r_1^2 \sim \chi_{n-p}^2$ such that $s_1 \geq 0$, $r_1 \geq 0$ are independent of each other and independent of everything else. By Lemma 14, we have that $G_{j,1} := s_1 V_{j,1}$ for $j \in [p]$ are independent $N(0,1)$ random variables. Note that

$$\frac{1}{4}(W^\top W)_{1,1} = \sum_{j=1}^p t_1^2 a_j V_{j,1}^2 = \frac{t_1^2}{s_1^2} \sum_{j=1}^p a_j G_{j,1}^2.$$

Let $\delta = \delta_n > 0$ be chosen later. By Laurent and Massart (2000, Lemma 1), applied conditionally on a, we have with probability at least $1 - 6\delta$ that all the following inequalities hold:

$$\|a\|_1 - 2\|a\|_2 \sqrt{\log \frac{1}{\delta}} \leq \sum_{j=1}^p a_j G_{j,1}^2 \leq \|a\|_1 + 2\|a\|_2 \sqrt{\log \frac{1}{\delta}} + 2\|a\|_\infty \log \frac{1}{\delta},$$

$$p - 2\sqrt{p \log \frac{1}{\delta}} \leq s_1^2 \leq p + 2\sqrt{p \log \frac{1}{\delta}} + 2\log \frac{1}{\delta},$$

$$n - 2\sqrt{n \log \frac{1}{\delta}} \leq t_1^2 \leq n + 2\sqrt{n \log \frac{1}{\delta}} + 2\log \frac{1}{\delta}.$$

Keeping in mind that $\|a\|_\infty \leq 1/4$, we have with probability at least $1 - 6\delta$ that

$$\frac{n - 2\sqrt{n\log(1/\delta)}}{p + 2\sqrt{p\log(1/\delta)} + 2\log(1/\delta)} \big\{ \|a\|_1 - 2\|a\|_2 \sqrt{\log(1/\delta)} \big\} \leq \frac{1}{4}(W^\top W)_{1,1}$$
$$\leq \frac{n + 2\sqrt{n\log(1/\delta)} + 2\log(1/\delta)}{p - 2\sqrt{p\log(1/\delta)}} \big\{ \|a\|_1 + 2\|a\|_2 \sqrt{\log(1/\delta)} + \frac{1}{2}\log(1/\delta) \big\}$$

If $\log(1/\delta) = o(p)$, then for each p with probability at least $1 - 6\delta$, we have that

$$\left| \frac{(W^\top W)_{1,1}}{4} - \frac{n}{p}\|a\|_1 \right| \leq \|a\|_1 \frac{2\sqrt{n\log(1/\delta)}}{p} \big(1 + \sqrt{n/p}\big)$$
$$+ \frac{n}{p} \left\{ 2\|a\|_2 \sqrt{\log(1/\delta)} + \frac{\log(1/\delta)}{2} \right\} \tag{S.2}$$
$$+ \mathcal{O}_s \left(\frac{\|a\|_1 \log(1/\delta)}{p} + \frac{\|a\|_2 \log(1/\delta)}{\sqrt{p}} + \frac{\log^{3/2}(1/\delta)}{p^{1/2}} \right).$$

By the definition of B, we have for $H := T(X^\top X)^{-1/2} \in \mathbb{O}^{p \times p}$ that

$$H^\top B H = H^\top T^{-\top} X_1^\top X_1 T^{-1} H = (X^\top X)^{-1/2}(X_1^\top X_1)(X^\top X)^{-1/2},$$

which follows the matrix-variate Beta distribution $\mathrm{Beta}_p(n_1/2, n_2/2)$ as defined before Lemma 13. Hence, the diagonal elements of Λ are the same as the eigenvalues of a $\mathrm{Beta}_p(n_1/2, n_2/2)$ random matrix. By (S.2), for each p, with probability at least $1 - 6\delta$,

we have

$$\left|(W^\top W)_{1,1} - \frac{4n}{p}\|a\|_1\right| \le 8\sqrt{n\log(1/\delta)}\big((\kappa_1+\kappa_2)\sqrt{n/p}+\kappa_1\big) + \mathcal{O}_s\bigg(\log(1/\delta) + \frac{\log^{3/2}(1/\delta)}{p^{1/2}}\bigg). \tag{S.3}$$

The desired result follows by taking $h_n = 4\|a\|_1/p$ and observing that by Lemma 13 $h_n \to 4\kappa_1$ almost surely. □

With Proposition 8, we are now in a position to prove Proposition 2 in its general form.

PROOF OF PROPOSITION 2. Let $V_i := X_i\Sigma^{-1/2}, i = 1, 2$ and set $V = (V_1^\top, V_2^\top)^\top$. Then each row of V follows $N(0, I_p)$ and is independent of each other. We have

$$\begin{aligned} W^\top W &= 4(X_1^\top X_1)(X^\top X)^{-1}(X_2^\top X_2) \\ &= 4(\Sigma^{1/2}V_1^\top V_1\Sigma^{1/2})(\Sigma^{-1/2}(V^\top V)^{-1}\Sigma^{-1/2})(\Sigma^{1/2}V_2^\top V_2\Sigma^{1/2}) \\ &= 4\Sigma^{1/2}V_1^\top V_1(V^\top V)^{-1}V_2^\top V_2\Sigma^{1/2} \overset{\mathrm{d}}{=} \Sigma^{1/2}(W_I^\top W_I)\Sigma^{1/2}, \end{aligned}$$

where W_I is the complementarily sketched design matrix when all entries of X are independent standard normals (i.e. $\Sigma = I_p$). Let h_n be the random sequence satisfying Proposition 8 and define $\Delta := W^\top W/n - h_n\Sigma$ and $\Delta_I := W_I^\top W_I/n - h_n I_p$. We start by controlling the ℓ-sparse operator norm of Δ for an arbitrary $\ell \in [p]$. By Lemma 12, there exists a $1/4$-net $\mathcal{N}_\ell$ of $\{v \in \mathcal{S}^{p-1} : \|v\|_0 \le \ell\}$ of cardinality at most $\binom{p}{\ell}9^\ell$, such that

$$\begin{aligned} \|\Delta\|_{\ell,\mathrm{op}} &= 2\sup_{u\in\mathcal{N}_\ell} u^\top\Delta u \overset{\mathrm{d}}{=} 2\sup_{u\in\mathcal{N}_\ell}(\Sigma^{1/2}u)^\top\Delta_I(\Sigma^{1/2}u) \\ &\le 2\|\Sigma\|_{k,\mathrm{op}}\sup_{u\in\mathcal{N}'_\ell} u^\top\Delta_I u \le 2\overline{\lambda}\sup_{u\in\mathcal{N}'_\ell} u^\top\Delta_I u, \end{aligned} \tag{S.4}$$

where $\mathcal{N}'_\ell := \{\Sigma^{1/2}u/\|\Sigma^{1/2}u\|_2 : u \in \mathcal{N}_\ell\}$. We claim that for any $u \in \mathcal{S}^{p-1}$, we have $u^\top\Delta_I u \overset{\mathrm{d}}{=} e_1^\top\Delta_I e_1$. This is because for any $H \in \mathbb{O}^{p\times p}$, we have $V \overset{\mathrm{d}}{=} VH$ and hence by Lemma 10 that

$$\begin{aligned} H^\top W_I^\top W_I H &= 4(H^\top V_1^\top V_1 H)(H^\top V^\top V H)^{-1}(H^\top V_2^\top V_2 H) \\ &\overset{\mathrm{d}}{=} 4(V_1^\top V_1)(V^\top V)^{-1}(V_2^\top V_2) = W_I^\top W_I, \end{aligned} \tag{S.5}$$

which in particular implies our claim. Consequently, by Proposition 8 and a union bound, when $\log(1/\delta) = o(p)$, we have with probability at least $1 - 6|\mathcal{N}_\ell|\delta$ that

$$\|\Delta\|_{\ell,\mathrm{op}} \le (16 + o(1))\overline{\lambda}\sqrt{\frac{\log(1/\delta)}{n}}\big\{(\kappa_1+\kappa_2)\sqrt{n/p}+\kappa_1\big\}. \tag{S.6}$$

To prove (9), we recall $\mathrm{diag}(\Sigma) = I_p$ and set $\ell = 1$ and $\delta = p^{-3}$ in (S.6) to obtain with probability at least $1 - 54p^{-2}$ that

$$\begin{aligned} \max_{j\in[p]}\left|\frac{(W^\top W)_{j,j}}{nh_n} - 1\right| &= \frac{1}{h_n}\|\Delta\|_{1,\mathrm{op}} \\ &\le \frac{(16+o(1))\overline{\lambda}}{h_n}\sqrt{\frac{3\log p}{n}}\big\{(\kappa_1+\kappa_2)\sqrt{n/p}+\kappa_1\big\}. \end{aligned} \tag{S.7}$$

The first conclusion follows by noting that $h_n \to 4\kappa_1$ and an application of the Borel–Cantelli lemma (since p^{-2} is summable).

To prove (11), we set $\ell = k$ and $\delta = (10ep/k)^{-(k+4)}$. By (10), we have $\log(1/\delta) = (k+4)\log(10ep/k) = \mathcal{O}(p)$ and

$$|\mathcal{N}|\delta \le 9^k \binom{p}{k}\left(\frac{10ep}{k}\right)^{-(k+4)} \le \left(\frac{9ep}{k}\right)^k \left(\frac{10ep}{k}\right)^{-(k+4)}$$
$$\le \frac{0.9^k}{(ep/k)^4} \le \max(p^{-2}, 0.9^{\sqrt{p}}).$$

By the Borel–Cantelli lemma,

$$\|\Delta\|_{k,\mathrm{op}} \le (16+\mathcal{O}(1))\overline{\lambda}\sqrt{\frac{(k+4)\log(10ep/k)}{n}}\big\{(\kappa_1+\kappa_2)\sqrt{n/p}+\kappa_1\big\}, \tag{S.8}$$

holds for all but finitely many p. We work on p sufficiently large such that (S.8) holds henceforth. Define $\hat{D} := \mathrm{diag}(W^\top W)/(nh_n)$. By (S.7) and a Taylor expansion, we have that $\|\hat{D}^{-1/2} - I\|_{\mathrm{op}} = (1+\mathcal{O}(1))(2h_n)^{-1}\|\Delta\|_{1,\mathrm{op}}$. Thus,

$$\|\hat{D}^{-1/2}W^\top W\hat{D}^{-1/2} - W^\top W\|_{k,\mathrm{op}}$$
$$\le \|\hat{D}^{-1/2} - I\|_{\mathrm{op}}\|W^\top W\|_{k,\mathrm{op}}(1+\|\hat{D}^{-1/2}\|_{\mathrm{op}})$$
$$\le (2+\mathcal{O}(1))\frac{\|\Delta\|_{1,\mathrm{op}}}{2h_n}\|n\Delta + nh_n\Sigma\|_{k,\mathrm{op}} \le (1+\mathcal{O}(1))n\overline{\lambda}\|\Delta\|_{1,\mathrm{op}}$$

where the final bound follows from (S.8) and the fact that $\|\Sigma\|_{k,\mathrm{op}} \le \overline{\lambda}$. Consequently, noting $h_n \to 4\kappa_1$, for all large p we have

$$\|\tilde{W}^\top\tilde{W} - \Sigma\|_{k,\mathrm{op}} = \frac{1}{h_n}\|\hat{D}^{-1/2}W^\top W\hat{D}^{-1/2}/n - h_n\Sigma\|_{k,\mathrm{op}}$$
$$\le \frac{1}{h_n}\|\Delta\|_{k,\mathrm{op}} + \frac{1}{nh_n}\|\hat{D}^{-1/2}W^\top W\hat{D}^{-1/2} - W^\top W\|_{k,\mathrm{op}}$$
$$\le \frac{1}{h_n}(\|\Delta\|_{k,\mathrm{op}} + \overline{\lambda}\|\Delta\|_{1,\mathrm{op}})$$
$$\le C_{s,r}\left\{\overline{\lambda}\sqrt{\frac{k\log(ep/k)}{n}} + \overline{\lambda}^2\sqrt{\frac{\log p}{n}}\right\},$$

for $C_{s,r} := 9(1+\sqrt{1+1/s}+\sqrt{s+r-1+1/s+1/r})$, which completes the proof. □

PROOF OF COROLLARY 4. The first inequality follows from the definition of $\mathcal{M}_X(k,\rho)$. An inspection of the proofs of Theorems 1 and 3 of the main text reveals that both results only depend on the complementary-sketched model $Z = W\theta + \xi$, and hence hold uniformly over (β_1, β_2). Thus, we have from Theorem 1 that $\sup_{\beta\in\mathbb{R}^p} P^X_{\beta,\beta}(\psi^{\mathrm{sparse}}_{\lambda,\tau} \ne 0) \xrightarrow{\text{a.s.}} 0$ and from Theorem 3 that $\sup_{\beta_1,\beta_2\in\mathbb{R}^p:(\beta_1-\beta_2)/2\in\Theta_{p,k}(\rho)} P^X_{\beta,\beta}(\psi^{\mathrm{sparse}}_{\lambda,\tau} \ne 1) \xrightarrow{\text{a.s.}} 0$. Combining the two completes the proof. □

PROOF OF THEOREM 6. As in the proof of Theorem 3, we work with a deterministic sequence of W such that (9) and (11) are satisfied. Furthermore, by Condition (S2), we henceforth work on the almost sure event $\Omega_\sigma = \{|\hat{\sigma}/\sigma - 1| = \mathcal{O}(p^{-1/2}\log^{1/2}p)\}$. For $\tilde{\theta} = (\tilde{\theta}_1, \ldots, \tilde{\theta}_p)^\top$ such that $\tilde{\theta}_j := \theta_j\|W_j\|_2$, we have from (8) that $Z = \tilde{W}\tilde{\theta} + \xi$, for

$\xi \sim N_m(0, \sigma^2 I_m)$. Hence, under the null hypothesis, $(\|Z\|_2^2/\sigma^2) \sim \chi^2_m$, which by Laurent and Massart (2000, Lemma 1) yields that

$$\mathbb{P}\big\{\|Z\|_2^2/\sigma^2 \geq m + 2\sqrt{m\log(1/\delta)} + 2\log(1/\delta)\big\} \leq \delta.$$

We set $\delta = p^{-(1+\varepsilon/2)}$ and have for $\eta_0 = \sigma^2(m + 2\sqrt{(1+\varepsilon/2)m\log p} + 2(1+\varepsilon/2)\log p)$, $\eta = \hat{\sigma}^2(m + 2\sqrt{m(1+\varepsilon)\log p} + 2(1+\varepsilon)\log p) \geq \eta_0$ on Ω_σ for all p sufficiently large. We bound, for all p sufficiently large,

$$\mathbb{P}(\|Z\|^2 \geq \eta) \leq \mathbb{P}(\|Z\|^2 \geq \eta_0) \leq p^{-(1+\varepsilon/2)},$$

whence, by the Borel–Cantelli lemma, we have $\psi_\eta^{\mathrm{dense}}(X_1, X_2, Y_1, Y_2) \xrightarrow{\text{a.s.}} 0$.

On the other hand, under the alternative, $(\|Z\|_2^2/\sigma^2) \sim \chi^2_m(\|W\theta\|_2^2)$. Observe from (9) and (11) that

$$\begin{aligned}\|W\theta\|_2^2 = \|\tilde{W}\tilde{\theta}\|_2^2 &= \|\Sigma^{1/2}\tilde{\theta}\|_2^2 + \tilde{\theta}^\top(\tilde{W}^\top\tilde{W} - \Sigma)\tilde{\theta} \\ &\geq \|\tilde{\theta}\|_2^2\big(\underline{\lambda} - \|\tilde{W}^\top\tilde{W} - \Sigma\|_{k,\mathrm{op}}\big) \\ &\geq (4\underline{\lambda} - \mathcal{O}(1))n\kappa_1\rho^2 \geq (8 - \mathcal{O}(1))\sigma^2\sqrt{m\log p}.\end{aligned}$$

By a similar argument, we also have $\|W\theta\|_2^2 \leq (4\overline{\lambda} + \mathcal{O}(1))n\kappa_1\rho^2 = \mathcal{O}(m)$. Thus, by Birgé (2001, Lemma 8.1), we have with probability at least $1 - p^{-2}$ that

$$\begin{aligned}\|Z\|_2^2/\sigma^2 &\geq m + \|W\theta\|_2^2 - 2\sqrt{(2m + 4\|W\theta\|_2^2)\log p} \\ &\geq m + (8 - 2\sqrt{2} - \mathcal{O}(1))\sqrt{m\log p}\end{aligned}$$

which is at least $\eta = \hat{\sigma}^2(m + 2\sqrt{2(1+\varepsilon)m\log p} + 2(1+\varepsilon)\log p) \leq$ for all p sufficiently large and any $\varepsilon \in (0, 5]$, conditionally on the almost sure event Ω_σ.

Consequently, we have $\mathbb{P}(\|Z\|_2^2 \leq \eta) \leq p^{-2}$ for all large p. As p^{-2} is summable, by the Borel–Cantelli lemma, $\psi_\eta^{\mathrm{dense}}(X_1, X_2, Y_1, Y_2) \xrightarrow{\text{a.s.}} 1$. □

PROOF OF THEOREM 7. Note that since $k \leq p^\alpha$ for $\alpha < 1$, condition (10) is satisfied and hence we can work with a deterministic sequence of W satisfying (9) and (11). Similar to the proof of Theorem 5, we write $\Sigma = \Sigma_0 + \Gamma$ for some $\Sigma_0 \in \mathrm{RowSp}(D)$ with $\|\Sigma_0\|_{\max} \leq 1$ and $\|\Gamma\|_{\max} = \mathcal{O}(D/(k\log p))$. We follow the proof of Theorem 5 up to (19).

Now, noting the assumption on ρ^2, we have by (18) that $\mathbb{E}(e^{2\tilde{\theta}^\top(\tilde{W}^\top\tilde{W} - \Sigma_0)\tilde{\theta}'}) = 1 + \mathcal{O}(1)$. It remains to show that $\mathbb{E}(e^{2\tilde{\theta}^\top\Sigma_0\tilde{\theta}'}) = 1 + \mathcal{O}(1)$. To this end, we have by Lemma 16 that

$$\begin{aligned}\mathbb{E}(e^{2\tilde{\theta}^\top\Sigma_0\tilde{\theta}'}) &\leq \left\{1 + \frac{Dk}{p}\big(\cosh(2\vartheta^2 D) - 1\big)\right\}^k \leq \left\{1 + \frac{Dk}{p}\big(2 + \mathcal{O}(1)\big)\vartheta^4 D^2\right\}^k \\ &\leq \exp\left\{\frac{(2 + \mathcal{O}(1))D^3k^2\vartheta^4}{p}\right\} = 1 + \mathcal{O}(1),\end{aligned}$$

where the second inequality follows by the Taylor expansion of $x \mapsto \cosh(x)$ and the fact that $\vartheta^2 D = (4 + \mathcal{O}(1))n\kappa_1\rho^2 D/k = \mathcal{O}(1)$, and the last equality holds by noting $D^3k^2\vartheta^4/p = (16 + \mathcal{O}(1))\kappa_1^2 D^3\rho^4 n^2/p = \mathcal{O}(1)$. □

APPENDIX B: ANCILLARY RESULTS

PROPOSITION 9. *If $k = p \geq \min\{n_1, n_2\}$, then $\mathcal{M}_X(k, \rho) = 1$. If $k = p$ and $p/n_1, p/n_2 \in [\varepsilon, 1)$ for any fixed $\varepsilon \in (0, 1)$, and $\theta = (\beta_1 - \beta_2)/2 \in \Theta_{p,k}(\rho)$ with*

$$\rho^2 = \mathcal{O}\left(\max\left\{\frac{p}{(n_1 - p)^2}, \frac{p}{(n_2 - p)^2}\right\}\right),$$

then $\mathcal{M}_X(k,\rho) \xrightarrow{\text{a.s.}} 1$.

PROOF. As in the proof of Theorem 5, it suffices to control $P_0(\mathcal{L}^2)$ for some choice of prior π. We write $\lambda_{\min}(W^\top W)$ for the minimum eigenvalue of $W^\top W$ and let θ be an associated eigenvector with ℓ_2 norm equal to ρ. We choose π to be the Dirac measure on θ. Then by (17), we have

$$P_0(\mathcal{L}^2) = e^{\theta^\top W^\top W\theta} = e^{\rho^2\lambda_{\min}(W^\top W)}.$$

When $p \geq n_1$ or $p \geq n_2$, we have by Lemma 10 that the Gram matrix $W^\top W = (X_1^\top X_1)(X^\top X)^{-1}(X_2^\top X_2)$ is singular. Hence, $\lambda_{\min}(W^\top W) = 0$ and $P_0(\mathcal{L}^2) = 1$, which implies that $\mathcal{M}_X(k,\rho) = 1$.

On the other hand, if $p < \min\{n_1, n_2\}$, let V_1, V_2, V, W_I be defined as in the proof of Proposition 2. Let T and Λ be defined as in the proof of Proposition 8, with V and W_I taking the roles of X and W therein respectively. Then, by (S.1), we have

$$\begin{aligned}\lambda_{\min}(W_I^\top W_I) &\leq 4\|T\|_{\text{op}}^2\lambda_{\min}\big(\Lambda(I-\Lambda)\big)\\ &\leq 4\|V^\top V\|_{\text{op}}\min\{\lambda_{\min}(\Lambda), 1-\lambda_{\max}(\Lambda)\}.\end{aligned} \tag{S.9}$$

Applying tail bounds for operator norm of a random Gaussian matrix (see, e.g. Wainwright, 2019, Theorem 6.1), we have

$$\|V^\top V\|_{\text{op}} \leq n\left(1 + \sqrt{\frac{p}{n}} + \sqrt{\frac{2\log p}{n}}\right)^2 \leq 5n$$

asymptotically with probability 1. Moreover, by Bai et al. (2015, Theorem 1.1), there is an almost sure event on which the empirical spectral distribution of Λ converges weakly to a distribution supported on $[t_\ell, t_r]$, for t_ℓ and t_r defined in (S.10). We will work on this almost sure event henceforth. For $p/n_1 \to \xi \in [\varepsilon, 1)$ and $p/n_2 \to \eta \in [\varepsilon, 1)$, we have $\limsup_{p\to\infty}\lambda_{\min}(\Lambda) \leq t_\ell$ and $\liminf_{p\to\infty}\lambda_{\max}(\Lambda) \geq t_r$. On the other hand, Taylor expanding the expression for t_ℓ and t_r in (S.10) with respect to $1-\xi$ and $1-\eta$ respectively, we obtain that

$$\begin{aligned}t_\ell &= \frac{1}{4}\eta(1-\xi)^2 + \mathcal{O}_\varepsilon\big((1-\xi)^3\big),\\ 1 - t_r &= \frac{1}{4}\xi(1-\eta)^2 + \mathcal{O}_\varepsilon\big((1-\eta)^3\big).\end{aligned}$$

Therefore, $\min\{\lambda_{\min}(\Lambda), 1-\lambda_{\max}(\Lambda)\} = \mathcal{O}_\varepsilon(\min\{(1-\xi)^2, (1-\eta)^2\})$. By the condition on ρ^2 and (S.9), we have

$$\rho^2\lambda_{\min}(W^\top W) \leq \overline{\lambda}\rho^2\lambda_{\min}(W_I^\top W_I) \ = o(1),$$

which implies that $P_0(\mathcal{L}^2) = 1 + o(1)$ and $\mathcal{M}_X \xrightarrow{\text{a.s.}} 1$. □

LEMMA 10. *Let* n_1, n_2, p, m *be positive integers such that* $n_1 + n_2 = p + m = n$. *Let* $X = (X_1^\top, X_2^\top)^\top \in \mathbb{R}^{n\times p}$ *be a non-singular matrix with block components* $X_1 \in \mathbb{R}^{n_1\times p}$ *and* $X_2 \in \mathbb{R}^{n_2\times p}$. *Choose* $A_1 \in \mathbb{R}^{n_1\times m}$ *and* $A_2 \in \mathbb{R}^{n_2\times m}$ *to satisfy* (7). *Then*

$$X_1^\top A_1 A_1^\top X_1 = -X_2^\top A_2 A_2^\top X_2 = (X_1^\top X_1)(X^\top X)^{-1}(X_2^\top X_2).$$

PROOF. The first equality follows immediately from (7). Define $\tilde{X}_1 := X_1(X^\top X)^{-1/2}$ and $\tilde{X}_2 := X_2(X^\top X)^{-1/2}$. Then $\tilde{X} := (\tilde{X}_1^\top, \tilde{X}_2^\top)^\top$ has orthonormal columns with the same column span as X, and so

$$\begin{pmatrix}\tilde{X}_1 & A_1\\ \tilde{X}_2 & A_2\end{pmatrix} \in \mathbb{O}^{n\times n}.$$

In particular, $\tilde{X}_1\tilde{X}_1^\top + A_1A_1^\top = I_{n_1}$. Therefore,

$$\begin{aligned} X_1^\top A_1 A_1^\top X_1 &= X_1^\top (I_{n_1} - \tilde{X}_1\tilde{X}_1^\top)X_1 = X_1^\top X_1 - X_1^\top X_1 (X^\top X)^{-1} X_1^\top X_1 \\ &= X_1^\top X_1 (X^\top X)^{-1}(X^\top X - X_1^\top X_1) \\ &= (X_1^\top X_1)(X^\top X)^{-1}(X_2^\top X_2), \end{aligned}$$

where the last equality holds by noting the block structure of X. □

LEMMA 11. *For $X_1 \in \mathbb{R}^{n_1\times p}$ and $X_2 \in \mathbb{R}^{n_2\times p}$, define $L := (X_1^\top X_1 + X_2^\top X_2)^{-1}(X_2^\top X_2 - X_1^\top X_1)$, $\tilde{X}_1 := X_1(L + I_p)$ and $\tilde{X}_2 := X_2(L - I_p)$. We have*

$$\tilde{X}_1^\top \tilde{X}_1 + \tilde{X}_2^\top \tilde{X}_2 = 4X_1^\top X_1 (X_1^\top X_1 + X_2^\top X_2)^{-1} X_2^\top X_2.$$

PROOF. Write $G_1 := X_1^\top X_1$, $G_2 := X_2^\top X_2$. It is clear that

$$\begin{aligned} L - I_p &= -2(X_1^\top X_1 + X_2^\top X_2)^{-1} X_1^\top X_1 = -2(G_1 + G_2)^{-1} G_1, \\ L + I_p &= 2(X_1^\top X_1 + X_2^\top X_2)^{-1} X_2^\top X_2 = 2(G_1 + G_2)^{-1} G_2. \end{aligned}$$

Therefore, we have

$$\begin{aligned} &\frac{1}{4}(\tilde{X}_1^\top \tilde{X}_1 + \tilde{X}_2^\top \tilde{X}_2) \\ &= \frac{1}{4}\{(L + I_p)^\top X_1^\top X_1 (L + I_p) + (L - I_p)^\top X_1^\top X_2 (L - I_p)\} \\ &= G_2(G_1 + G_2)^{-1} G_1 (G_1 + G_2)^{-1} G_2 \\ &\quad + G_1(G_1 + G_2)^{-1} G_2 (G_1 + G_2)^{-1} G_1 \\ &= -G_1(G_1 + G_2)^{-1} G_1 (G_1 + G_2)^{-1} G_2 \\ &\quad - G_1(G_1 + G_2)^{-1} G_2 (G_1 + G_2)^{-1} G_2 + 2G_1(G_1 + G_2)^{-1} G_2 \\ &= G_1(G_1 + G_2)^{-1} G_2. \end{aligned}$$

The proof is complete by recalling the definitions of G_1 and G_2. □

The following lemma concerns the control of the k-operator norm of a symmetric matrix. Similar results have been derived in previous works (see, e.g. Wang, Berthet and Samworth, 2016, Lemma 2). For completeness, we include a statement and proof of the specific version we use.

LEMMA 12. *For any symmetric matrix $M \in \mathbb{R}^{p\times p}$ and $k \in [p]$, there exists a subset $\mathcal{N} \subseteq \mathcal{S}^{p-1}$ such that $|\mathcal{N}| \leq \binom{p}{k} 9^k$ and*

$$\|M\|_{k,\mathrm{op}} \leq 2 \sup_{u\in\mathcal{N}} u^\top M u.$$

PROOF. Define $\mathcal{B}_0(k) := \cup_{J\subset[p],|J|=k} S_J$, where $S_J := \{v \in \mathcal{S}^{p-1} : v_i = 0, \forall i \notin J\}$. For each S_J, we find a $1/4$-net $\mathcal{N}_J$ of cardinality at most 9^k (Vershynin, 2012, Lemma 5.2). Define $\mathcal{N} := \cup_{J\subset[p],|J|=k} \mathcal{N}_J$, which has the desired upper bound on cardinality. By construction,

for $v \in \arg\max_{u \in \mathcal{B}_0(k)} u^\top M u$, there exists a $\tilde{v} \in \mathcal{N}$ such that $|\operatorname{supp}(v) \cup \operatorname{supp}(\tilde{v})| \le k$ and $\|v - \tilde{v}\|_2 \le 1/4$. We have

$$\begin{aligned}\|M\|_{k,\mathrm{op}} &= v^\top M v = v^\top M(v - \tilde{v}) + (v - \tilde{v})^\top M \tilde{v} + \tilde{v}^\top M \tilde{v} \\ &\le 2\|v - \tilde{v}\|_2 \|M\|_{k,\mathrm{op}} + \tilde{v}^\top M \tilde{v} \le \frac{1}{2}\|M\|_{k,\mathrm{op}} + \sup_{u \in \mathcal{N}} u^\top M u.\end{aligned}$$

The desired inequality is obtained after rearranging terms in the above display. □

The following lemma describes the asymptotic limit of the nuclear and Frobenius norms of the product of a matrix-variate Beta-distributed random matrix and its reflection. Recall that for $n_1 + n_2 > p$, we say that a $p \times p$ random matrix B follows a matrix-variate Beta distribution with parameters $n_1/2$ and $n_2/2$, written $B \sim \mathrm{Beta}_p(n_1/2, n_2/2)$, if $B = (S_1 + S_2)^{-1/2} S_1 (S_1 + S_2)^{-1/2}$, where $S_1 \sim W_p(n_1, I_p)$ and $S_2 \sim W_p(n_2, I_p)$ are independent Wishart matrices and $(S_1 + S_2)^{1/2}$ is the symmetric matrix square root of $S_1 + S_2$. Recall also that the spectral distribution function of any $p \times p$ matrix A is defined as $F^A(t) := n^{-1} \sum_{i=1}^p \mathbb{1}_{\{\lambda_i^A \le t\}}$, where λ_i^As are eigenvalues (counting multiplicities) of the matrix A. Further, given a sequence $(A_n)_{n \in \mathbb{N}}$ of matrices, their limiting spectral distribution function F is defined as the weak limit of the F^{A_n}, if it exists.

LEMMA 13. *Let $B \sim \mathrm{Beta}_p(n_1/2, n_2/2)$ and suppose that $\lambda_1, \ldots, \lambda_p$ are the eigenvalues of B. Define $a = (a_1, \ldots, a_p)^\top$, with $a_j = \lambda_j(1 - \lambda_j)$ for $j \in [p]$. In the asymptotic regime of (C2), we have*

$$\|a\|_1/p \xrightarrow{\text{a.s.}} \kappa_1,$$
$$\|a\|_2/\sqrt{p} \xrightarrow{\text{a.s.}} \kappa_2,$$

where

$$\kappa_1 = \frac{r}{(1+r)^2(1+s)} \quad \text{and} \quad \kappa_2^2 = \frac{r(r + s - rs + r^2 s + rs^2)}{(1+r)^4(1+s)^3}.$$

PROOF. We first look at the limiting spectral distribution of B. From the asymptotic relations between n_1, n_2 and p in (C2), we have that

$$p/n_1 \to \xi := \frac{s + sr}{r + sr} \quad \text{and} \quad p/n_2 \to \eta := \frac{s + sr}{1 + s}.$$

Define the left and right limits

$$t_\ell, t_r := \frac{(\xi + \eta)\eta + \xi\eta(\xi - \eta) \mp 2\xi\eta\sqrt{\xi - \xi\eta + \eta}}{(\xi + \eta)^2}. \tag{S.10}$$

By Bai et al. (2015, Theorem 1.1), almost surely, weak limit F of F^B exists and is of the form $\max\{1 - 1/\xi, 0\}\delta_0 + \max\{1 - 1/\eta, 0\}\delta_1 + \mu$, where δ_0 and δ_1 are point masses at 0 and 1 respectively, and μ has a density

$$\frac{(\xi + \eta)\sqrt{(t_r - t)(t - t_\ell)}}{2\pi\xi\eta t(1 - t)} \mathbb{1}_{[t_\ell, t_r]}$$

with respect to the Lebesgue measure on $\mathbb{R}$. Define $h_1 : t \mapsto t(1 - t)$. By the portmanteau lemma (see, e.g. van der Vaart, 2000, Lemma 2.2), we have almost surely that

$$\begin{aligned}\|a\|_1/p = F^B h_1 \to F h_1 &= \frac{\xi + \eta}{2\pi\xi\eta} \int_{t_\ell}^{t_r} \sqrt{(t_r - t)(t - t_\ell)}\, dt = \frac{\xi + \eta}{16\xi\eta}(t_r - t_\ell)^2 \\ &= \frac{r}{(1+r)^2(1+s)}.\end{aligned}$$

Similarly, for $h_2 : t \mapsto t^2(1-t)^2$, we have almost surely that

$$\begin{aligned}\|a\|_2^2/p \to Fh_2 &= \frac{\xi+\eta}{2\pi\xi\eta}\int_{t_\ell}^{t_r} t(1-t)\sqrt{(t_r-t)(t-t_\ell)}dt \\ &= \frac{\xi+\eta}{256\xi\eta}(t_r-t_\ell)^2(8t_\ell - 5t_\ell^2 + 8t_r - 6t_\ell t_r - 5t_r^2) \\ &= \frac{r(r+s-rs+r^2s+rs^2)}{(r+1)^4(s+1)^3}.\end{aligned}$$

Define $\kappa_1 := Fh_1$ and $\kappa_2 := (Fh_2)^{1/2}$, we arrive at the lemma. □

The following result concerning the QR decomposition of a Gaussian random matrix is probably well-known. However, since we did not find results in this exact form in the existing literature, we have included a proof here for completeness. Recall that for $n \geq p$, the set $\mathbb{O}^{n\times p}$ can be equipped with a uniform probability measure that is invariant under the action of left multiplication by $\mathbb{O}^{n\times n}$ (see, e.g. Stiefel manifold in Muirhead, 2009, Section 2.1.4).

LEMMA 14. *Suppose $n \geq p$ and X is an $n \times p$ random matrix with independent $N(0,1)$ entries. Write $X = HT$, with H taking values in $\mathbb{O}^{n\times p}$ and T an upper-triangular $p \times p$ matrix with non-negative diagonal entries. This decomposition is almost surely unique. Moreover, H and T are independent, with H uniformly distributed on $\mathbb{O}^{n\times p}$ with respect to the invariant measure and $T = (t_{j,k})_{j,k\in[p]}$ having independent entries satisfying $t_{j,j}^2 \sim \chi^2_{p-j+1}$ and $t_{j,k} \sim N(0,1)$ for $1 \leq j < k \leq p$.*

PROOF. The uniqueness of the QR decomposition follows since X has rank p almost surely. The marginal distribution of T then follows from the Bartlett decomposition of $X^\top X$ (Muirhead, 2009, Theorem 3.2.14) and the relationship between the QR decomposition of X and the Cholesky decomposition of $X^\top X$.

For any fixed $Q \in \mathbb{O}^{n\times n}$, we have $QX \overset{\mathrm{d}}{=} X$. Since $\mathbb{O}^{n\times n}$ acts transitively (by left multiplication) on $\mathbb{O}^{n\times p}$, the joint density of H and T must be constant in H for each value of T. In particular, we have that H and T are independent, and that H is uniformly distributed on $\mathbb{O}^{n\times p}$ with respect to the translation-invariant measure. □

The following two lemmas control the moment generation functions of (decoupled) quadratic Rademacher chaos random variables with respect to different matrices.

LEMMA 15. *Let $\xi = (\xi_1, \ldots, \xi_d)^\top$ and $\xi' = (\xi'_1, \ldots, \xi'_d)^\top$ be independent with independent Rademacher entries and fix $A \in \mathbb{R}^{d\times d}$. There exists a universal constant $C > 0$ such that for any $0 < \|A\|_{\mathrm{op}} \leq 1/32$, we have*

$$\mathbb{E}(e^{\xi^\top A\xi'}) \leq 1 + C\|A\|_{\mathrm{F}}e^{4\|A\|_{\mathrm{F}}^2}.$$

PROOF. By Hoeffding's inequality, we have

$$\mathbb{P}(\xi^\top A\xi' \geq t \mid \xi') \leq \exp\left\{-\frac{t^2}{2\|A\xi'\|_2^2}\right\}. \tag{S.11}$$

By Jensen's inequality, we have $\mathbb{E}(\|A\xi'\|_2) \leq \{\mathbb{E}(\xi'^\top A^\top A\xi')\}^{1/2} \leq \|A\|_{\mathrm{F}}$. Moreover, the map $x \mapsto \|Ax\|$ is Lipschitz with constant $\|A\|_{\mathrm{op}}$. Hence, from Boucheron, Lugosi and Massart (2013, Theorem 6.10), we have

$$\mathbb{P}(\|A\xi'\|_2 \geq \|A\|_{\mathrm{F}} + u) \leq \exp\left\{-\frac{u^2}{8\|A\|_{\mathrm{op}}^2}\right\}. \tag{S.12}$$

Combining (S.11) and (S.12), and setting $u = (2t\|A\|_{\mathrm{op}})^{1/2}$, we have

$$\begin{aligned}\mathbb{P}(\xi^\top A\xi' \geq t) &\leq \mathbb{P}(\|A\xi'\|_2 \geq \|A\|_{\mathrm{F}} + u) \\ &\quad + \mathbb{E}[\mathbb{P}(\xi^\top A\xi' \geq t \mid \xi')\mathbb{1}_{\{\|A\xi'\|_2 \leq \|A\|_{\mathrm{F}}+u\}}] \\ &\leq \exp\biggl\{-\frac{u^2}{8\|A\|_{\mathrm{op}}^2}\biggr\} + \exp\biggl\{-\frac{t^2}{2(\|A\|_{\mathrm{F}}+u)^2}\biggr\} \\ &\leq 2\exp\biggl\{-\frac{t^2}{4(\|A\|_{\mathrm{F}} + t^{1/2}\|A\|_{\mathrm{op}}^{1/2})^2}\biggr\} \\ &\leq 2\max\bigl\{e^{-t^2/(16\|A\|_{\mathrm{F}}^2)}, e^{-t/(16\|A\|_{\mathrm{op}})}\bigr\}.\end{aligned}$$

Consequently, if $32\lambda\|A\|_{\mathrm{op}} \leq 1$, we have

$$\begin{aligned}\mathbb{E}(e^{\xi^\top A\xi'}) &= \int_{u=0}^{1} \mathbb{P}(e^{\xi^\top A\xi'} \geq u)\,du + \int_{t=0}^{\infty} \mathbb{P}(\xi^\top A\xi' \geq t)e^t\,dt \\ &\leq 1 + 2\int_{t=0}^{\|A\|_{\mathrm{F}}^2/\|A\|_{\mathrm{op}}} e^{-t^2/(16\|A\|_{\mathrm{F}}^2)+t}\,dt \\ &\quad + 2\int_{\|A\|_{\mathrm{F}}^2/\|A\|_{\mathrm{op}}}^{\infty} e^{-t/(16\|A\|_{\mathrm{op}})+t}\,dt \\ &\leq 1 + 8\sqrt{2\pi}\|A\|_{\mathrm{F}}e^{4\|A\|_{\mathrm{F}}^2} + 64\|A\|_{\mathrm{op}}.\end{aligned}$$

Our claim follows since $\|A\|_{\mathrm{op}} \leq \|A\|_{\mathrm{F}}$. □

LEMMA 16. *Fix $A \in \mathrm{RowSp}(D) \subseteq \mathbb{R}^{p\times p}$ and let J and J' be independent and drawn uniformly at random from all subset of cardinality k of $[p]$. Let $\xi = (\xi_1, \ldots, \xi_d)^\top$ and $\xi' = (\xi_1', \ldots, \xi_d')^\top$ be independent (and independent of J and J') with independent Rademacher entries. Then*

$$\mathbb{E}(e^{\xi^\top A_{J,J'}\xi'}) \leq \biggl\{1 + \frac{Dk}{p}\bigl(\cosh(D\|A\|_{\max}) - 1\bigr)\biggr\}^k.$$

PROOF. Write $a := \|A\|_{\max}$. Also, for notational simplicity, we define $\theta, \theta' \in \mathbb{R}^p$ such that $\theta_J = \xi$, $\theta_{J^c} = 0$, $\theta'_{J'} = \xi'$, $\theta'_{J'^c} = 0$. So $\xi^\top A_{J,J'}\xi' = \theta^\top A\theta'$.

For each $j \in [p]$, we write $\mathrm{nb}(j) := \{j' \in [p] : A_{j,j'} \neq 0\}$. Note that by the definition of $\mathrm{RowSp}(D)$, $|\mathrm{nb}(j)| \leq D$ for all $j \in [p]$. Hence,

$$\theta^\top A\theta' = \sum_{j\in J}\sum_{j'\in \mathrm{nb}(j)\cap J'} A_{j,j'}\theta_j\theta'_{j'} = \sum_{j\in J} c_j\theta_j$$

where $c_j := \sum_{j'\in\mathrm{nb}(j)\cap J'} A_{j,j'}\theta'_{j'}$. We note that $|c_j| \leq Da$ and $c_j = 0$ unless $j \in \cup_{j'\in J'}\mathrm{nb}(j')$. Observe that $|\cup_{j'\in J'}\mathrm{nb}(j')| \leq Dk$, so $|\cup_{j'\in J'}\mathrm{nb}(j') \cap J|$ is stochastically dominated by the hypergeometric random variable $\mathrm{HyperGeom}(k; Dk, p)$ (defined as the number of black balls obtained from k draws without replacement from an urn containing p balls, Dk of which are black). Let $B \sim \mathrm{Bin}(k, Dk/p)$, we have by Hoeffding (1963, Theorem 4) that

$$\begin{aligned}\mathbb{E}(e^{\theta^\top A\theta'}) &= \mathbb{E}\Bigl\{\prod_{j\in[p]} \mathbb{E}(e^{c_j\theta_j} \mid J', \theta')\Bigr\} = \mathbb{E}\Bigl\{\prod_{j\in\cup_{j'\in J'}\mathrm{nb}(j')\cap J} \cosh(c_j)\Bigr\} \\ &\leq \mathbb{E}(e^{B\log\cosh(Da)}) = \biggl\{1 + \frac{Dk}{p}\bigl(\cosh(Da) - 1\bigr)\biggr\}^k.\end{aligned}$$

The proof is complete. □

REFERENCES

Bai, Z., Hu, J., Pan, G. and Zhou, W. (2015) Convergence of the empirical spectral distribution function of Beta matrices. *Bernoulli*, **21**, 1538–1574.

Birgé, L. (2001) An alternative point of view on Lepskis method. In de Gunst, M and Klaassen, C. and van der Vaart, A. Eds,. *Lecture Notes-Monograph Series 36*, 113–133.

Boucheron, S., Lugosi, G. and Massart, P. (2013) *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.

Hoeffding, W. (1963) Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, **58**, 13–30.

Laurent, B. and Massart, P. (2000) Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, **28**, 1302–1338.

Muirhead, R. J. (2009) *Aspects of Multivariate Statistical Theory*. John Wiley & Sons, Hoboken, New Jersey.

van der Vaart, A. W. (2000) *Asymptotic Statistics*. Cambridge University Press, Cambridge.

Vershynin, R. (2012) Introduction to the non-asymptotic analysis of random matrices. In Y. Eldar and G. Kutyniok (Eds.). *Compressed Sensing, Theory and Applications*, 210–268.

Wainwright, M. J. (2019) *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, Cambridge.

Wang, T., Berthet, Q. and Samworth, R. J. (2016) Statistical and computational trade-offs in estimation of sparse principal components. *Ann. Statist.*, **44**, 1896–1930.