Two-sample testing of high-dimensional linear regression coefficients via complementary sketching

Fengnan Gao* and Tengyao Wang†

(November 27, 2020)

Abstract

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.

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 $X_1, \ldots X_{n_1} \overset{\text{iid}}{\sim} N(\mu_1, \sigma^2)$ and $Y_1, \ldots, Y_{n_2} \overset{\text{iid}}{\sim} N(\mu_2, \sigma^2)$, 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 $X_1, \ldots, X_{n_1} \overset{\text{iid}}{\sim} F_{\theta_1,\gamma_1}$ and $Y_1, \ldots, Y_{n_2} \overset{\text{iid}}{\sim} F_{\theta_2,\gamma_2}$ 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

*Fudan University. Email: fngao@fudan.edu.cn
†University College London. Email: tengyao.wang@ucl.ac.uk

matrix $X_1 \in \mathbb{R}^{n_1 \times p}$ 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 generalised 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 generalised 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 or not 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.

As an example of applications where such assumptions come about natrually, consider the area of differential networks. Here, researchers are interested in whether two biological 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. 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 small, i.e., there are only a few different edges from one network to another. In the above case of dense coefficients, sparsity assumption makes no sense 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 much assumptions.

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 Bonferonni 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.

1. We propose a novel method to solve the testing problems formulated in (5) and (6) for model (2). Through complementary sketching, 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 optimisation problems. Moreover, when complementary sketching is combined with any methods designed for one-sample problems, our proposal substantially supplies a novel class of testing and estimation procedures for the corresponding two-sample problems.

2. 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, if $\rho \geq \sqrt{\frac{8k \log p}{n\kappa_1}}$, 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 $\rho \leq \sqrt{\frac{c_\alpha k \log p}{n\kappa_1}}$ for some c_α depending on α only, then almost surely no test has asymptotic size 0 and power 1 .

3. 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 $\rho^2 \asymp k \log p/(n\kappa_1)$, with the precise definition of κ_1 given in Lemma 12. 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, maximised 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 characterises the precise dependence of the difficulty of the two-sample problem on the sample size and dimensionality parameters.

4. We observe a phase transition phenomena of how the minimax detection limit depends on the sparsity parameter k. In addition to establishing 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, suited 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 $\rho \asymp \sqrt{\frac{k \log p}{n\kappa_1}}$ in the sparse regime, but of order $\rho \asymp p^{-1/4}$ 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).

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. Proofs of our main results are deferred until Section 5 with ancillary results in Section 6.

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 $\|v\|_q := \Big\{\sum_{i=1}^n (v_i)^q\Big\}^{1/q}$ 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., $(\mathrm{diag}(a))_{i,j} = \mathbb{1}_{\{i=j\}} a_i$. For $A \in \mathbb{R}^{p\times p}$, 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 $k \in [p]$, the k-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 $|a_n| \leq C b_n$ 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 reparametrise (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

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

Under this new parametrisation, 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, idealised 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 $Z := A_1^\top Y_1 + A_2^\top Y_2$ and $W := A_1^\top X_1 - A_2^\top X_2$. 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 synthesises 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 $(\tilde{A}X, AY)$, 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 summarising raw data with a smaller number of synthesised 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 summarise, 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 standardised 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 λ and τ as $\psi_{\lambda,\tau}^{\mathrm{sparse}}(X_1, X_2, Y_1, Y_2)$. As we will see in Section 3, the choice of $\lambda = 2\sigma\sqrt{\log p}$ and $\tau = k\sigma^2 \log p$ 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 $\lambda = 0$. 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 define our test as

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

Algorithm 1: Pseudo-code for complementary sketching-based test $\psi_{\lambda,\tau}^{\text{sparse}}$.

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 $\lambda \geq 0$, 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 $Q \leftarrow \operatorname{diag}(W^\top W)^{-1/2} W^\top Z$.

6 Compute the test statistic

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

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

for a suitable choice of threshold level η.

The computational complexity of both $\psi_{\lambda,\tau}^{\text{sparse}}$ and $\psi_{\eta}^{\text{dense}}$ 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 $\mathcal{O}(n^2p + nm^2)$. Finally, it is worth emphasising that while the matrix A generated this way is random, our test statistics $T = \sum_{j=1}^{p} Q_j^2 \mathbb{1}_{\{|Q_j| \geq \lambda\}}$ and $\|Z\|_2^2$, are in fact deterministic. To see this, we observe that both

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

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 9, $(\|W_j\|_2^2)_{j \in [p]}$, 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 $\psi_{\lambda,\tau}^{\text{sparse}}$ and $\psi_{\eta}^{\text{dense}}$. 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.

(C1) All entries of X_1 and X_2 are independent standard normals.

(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 main purpose for assuming (C1) is to ensure that after standardising columns to have unit length, the matrix W, as computed in Step 4 of Algorithm 1, satisfies a restricted isometry condition with high probability as in Proposition 2. In fact, as revealed in the proof of our theory, even if matrices X_1 and X_2 do not follow the Gaussian design, as long as W satisfies (9) and (10), all results in this section will still be true. 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. 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 what follows, we also make the simplifying assumption that the noise variance σ^2 is known. By replacing X_1, X_2, Y_1, Y_2 with X_1/σ, X_2/σ, Y_1/σ and Y_2/σ, we may further assume that $\sigma^2 = 1$ without loss of generality, which we assume for the rest of this section. In practice, if the noise variance is unknown, we can replace it with one of its consistent estimators $\hat{\sigma}^2$ (see e.g. Reid, Tibshirani and Friedman, 2016, and references therein).

3.1 Sparse case

We consider in this subsection the test $\psi_{\lambda,\tau}^{\text{sparse}}$, which is suitable for distinguishing β_1 and β_2 that differ in a small number of coordinates, the setting that has more subtle phenomena and hence is most of interest to us. Our first result below states that with a choice of hard-thresholding level λ of order $\sqrt{\log p}$, the test has asymptotic size 0.

Theorem 1. *If Condition (C2) holds and $\beta_1 = \beta_2$, then, with the choice of parameters $\tau > 0$ and $\lambda = \sqrt{(4+\varepsilon)\log p}$ for any $\varepsilon > 0$, we have*

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

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 $\psi_{\lambda,\tau}^{\mathrm{sparse}}$ is more involved. A key step in the argument is to show that $W^\top W$ is suitably close to a multiple of the identity matrix. More precisely, in Proposition 2 below, we derive entrywise and k-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 κ_1 is as defined in Lemma 12. 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:

$$\|\tilde{W}^\top\tilde{W} - I\|_{k,\mathrm{op}} \le C_{s,r}\sqrt{\frac{k\log(ep/k)}{n}}. \tag{11}$$

We note that condition (10) is relatively mild and would be satisfied if $k \le p^\alpha$ for any $\alpha \in [0, 1)$.

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 $\psi_{\lambda,\tau}^{\mathrm{sparse}}$, 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) and (C2), we further assume $k \in [p]$ and that (10) holds. If $\theta = (\beta_1 - \beta_2)/2 \in \Theta_{p,k}(\rho)$ with $\rho \ge \sqrt{\frac{8k\log p}{n\kappa_1}}$, and we set input parameters $\lambda = 2\sqrt{\log p}$ and $\tau \le 3k\log p$ in Algorithm 1, then*

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

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

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

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 $\mathcal{M}_X(k, \rho) \xrightarrow{\text{p}} 0$, there exists a test ψ that with asymptotic probability 1 correctly differentiates the null and the alternative. On the other hand, if $\mathcal{M}_X(k, \rho) \xrightarrow{\text{p}} 1$, 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. *Let $k \in [p]$ and assume Conditions (C1), (C2) and (10). If $\rho \geq \sqrt{\frac{8k \log p}{n\kappa_1}}$, and we set input parameters $\lambda = 2\sqrt{\log p}$ and $\tau \in (0, 3k \log p]$ in Algorithm 1, then*

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

Corollary 4 shows that the test $\psi^{\text{sparse}}_{\lambda,\tau}$ has an asymptotic detection limit, measured in $\|\beta_1 - \beta_2\|_2$, of at most $\sqrt{\frac{8k \log p}{n\kappa_1}}$ 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, the following theorem shows that when $\alpha < 1/2$, the detection limit of $\psi^{\text{sparse}}_{\lambda,\tau}$ is essentially optimal.

Theorem 5. *Under conditions (C1) and (C2), if further assume $k \leq p^\alpha$ for some $\alpha \in [0, 1/2)$ and $\rho \leq \sqrt{\frac{(1-2\alpha-\varepsilon)k \log p}{4n\kappa_1}}$ 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 if the signal ℓ_2 norm is a factor of $32/(1 - 2\alpha - \varepsilon)$ smaller than what can be detected by $\psi^{\text{sparse}}_{\lambda,\tau}$ 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 $\psi^{\text{sparse}}_{\lambda,\tau}$ has a minimax optimal detection limit measured in $\|\beta_1 - \beta_2\|_2$, up to constants depending on α 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, 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 $n\kappa_1$. By the explicit expression of κ_1 from Lemma 12, we have

$$n\kappa_1 = \frac{nr}{(1 + r)^2(1 + s)}, \tag{12}$$

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, (12) also gives a precise characterisation 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 maximised 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. *Let $\eta = m + 2^{3/2}\sqrt{m \log p} + 4 \log p$. Under Conditions (C1) and (C2), we further assume $k \in [p]$, $\rho^2 \geq \frac{2\sqrt{m \log p}}{n\kappa_1}$ 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 $k \in [p]$, 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 $\psi_{\lambda,\tau}^{\mathrm{sparse}}$. In our asymptotic regime (C2), $\frac{2\sqrt{m \log p}}{n\kappa_1}$ is, up to constants depending on s and r, of order $p^{-1/2} \log^{1/2} p$. The following theorem establishes that 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 $p^{1/2} \leq k \leq p^\alpha$ for some $\alpha \in [1/2, 1]$ and $\rho = o(p^{-1/4} \log^{-3/4} p)$, then $\mathcal{M}_X(k, \rho) \xrightarrow{\text{a.s.}} 1$.*

Theorem 7 points out that the lower bound on detectable signal size ρ^2 prescribed in Theorem 6 is necessary up to poly-logarithmic factors. The following proposition makes it explicit that the upper bound on sparsity imposed by (10) in Theorem 6 cannot be completely removed, i.e., the same result may not hold if we allow k to be a constant fraction of n.

Proposition 8. *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*

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

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

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 from 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). Specifically, after obtaining W and Z in Step 4 of Algorithm 1, we compute

$$\hat{M}_1 := \frac{1}{p}\mathrm{tr}\biggl(\frac{1}{m}W^\top W\biggr) \quad \text{and} \quad \hat{M}_2 := \frac{1}{p}\mathrm{tr}\biggl\{\biggl(\frac{1}{m}W^\top W\biggr)^2\biggr\} - \frac{1}{pm}\biggl\{\mathrm{tr}\biggl(\frac{1}{m}W^\top W\biggr)^2\biggr\}.$$

This allows us to estimate

$$\hat{\sigma}^2 := \biggl\{1 + \frac{p\hat{M}_1^2}{(m+1)\hat{M}_2}\biggr\}\frac{\|Z\|_2^2}{m} - \frac{\hat{M}_1}{m(m+1)\hat{M}_2}\|W^\top Z\|_2^2.$$

We implement our estimators $\psi_{\lambda,\tau}^{\mathrm{sparse}}$ and $\psi_{\eta}^{\mathrm{dense}}$ on standardised data $X_1/\hat{\sigma}$, $X_2/\hat{\sigma}$, $Y_1/\hat{\sigma}$ and $Y_2/\hat{\sigma}$ with the tuning parameters $\lambda = \sqrt{4\log p}$, $\tau = 3k\log p$ and $\eta = m + \sqrt{8m\log p} + 4\log p$ 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 $\psi_{\lambda,\tau}^{\mathrm{sparse}}$ relies on various problem parameters. In light of our results in Theorems 1 and 3, we define

$$\nu := \frac{rn\rho^2}{\sigma^2(1+s)(1+r)^2k\log p}, \tag{13}$$

where $s := p/m$ and $r := n_1/n_2$. As discussed after Theorem 3, $rn/\{(1+s)(1+r)^2\}$ in the definition of ν can be viewed as the effective sample size in the testing problem. In Figure 1, we plot the estimated test power of $\psi_{\lambda,\tau}^{\mathrm{sparse}}$ 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. In 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.

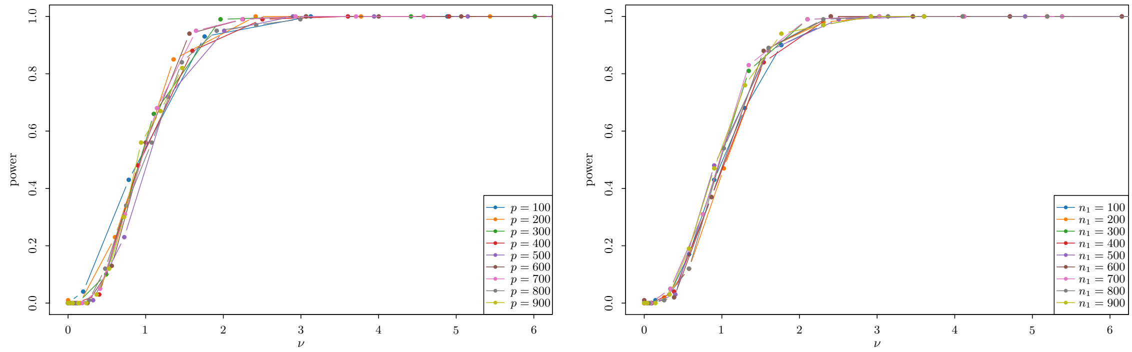

Figure 1: Power function of $\psi_{\lambda,\tau}^{\text{sparse}}$, estimated over 100 Monte Carlo repetitions, plotted against ν, as defined in (13), in various parameter settings. Left panel: $n_1 = n_2 = 500$, $p \in \{100, 200, \ldots, 900\}$, $k = 10$, $\rho \in \{0, 0.2, \ldots, 2\}$. Right panel: $n_1 \in \{100, 200, \ldots, 900\}$, $n_2 = 1000 - n_1$, $p = 400$, $k = 10$, $\rho \in \{0, 0.2, \ldots, 2\}$.

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 optimisation problems. Then based on the obtained estimators $\hat{\gamma}$ and $\hat{\Pi}$, ψ^{ZB} proceeds to compute a test statistic

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

Their test rejects the null if the test statistic exceeds an empirical upper-α-quantile (obtained 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 are then compared with Bonferonni-corrected thresholds at size α. For both the ψ^{LRT} and ψ^{CVV}, we set $\alpha = 0.05$.

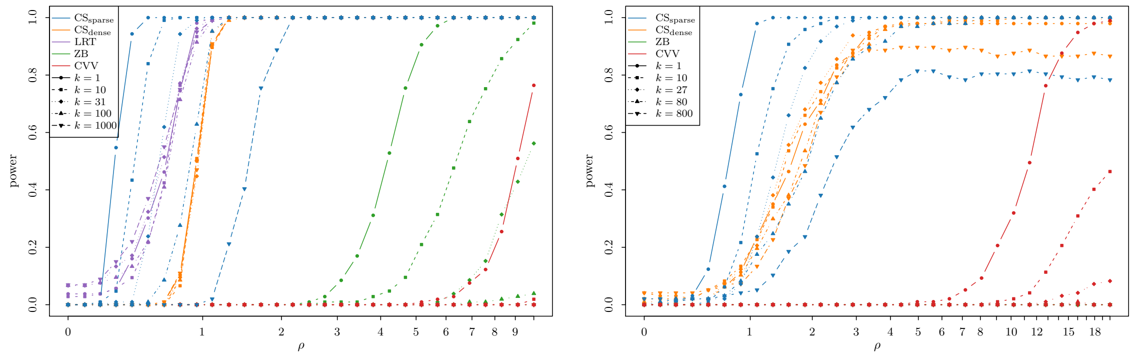

Figure 2: Power comparison of different methods at different sparsity levels $k \in \{1, 10, \lfloor p^{1/2} \rfloor, 0.1p, p\}$ and different signal ℓ_2 norm ρ on a logarithmic grid (noise variance $\sigma^2 = 1$). Left panel: $n_1 = n_2 = 500$, $p = 800$, $\rho \in [0, 10]$; right panel: $n_1 = n_2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

Figure 2 compares the estimated power, as a function of $\|\beta_1 - \beta_2\|_2$, of $\psi^{\mathrm{sparse}}_{\lambda,\tau}$ 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 maximise 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 $\psi^{\mathrm{sparse}}_{\lambda,\tau}$ to reach power of almost 1 in the sparsest case. Note though, in the densest case on the right panel ($k = 800$), $\psi^{\mathrm{sparse}}_{\lambda,\tau}$ and $\psi^{\mathrm{dense}}_{\eta}$ did not have saturated power curves, because noise variance is over-estiamted by $\hat{\sigma}^2$ in this setting.

We also observe that the power of $\psi^{\mathrm{sparse}}_{\lambda,\tau}$ has a stronger dependence on the level k than that of $\psi^{\mathrm{dense}}_{\eta}$. For $k \leq \sqrt{p}$, $\psi^{\mathrm{sparse}}_{\lambda,\tau}$ 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 $\psi_\eta^{\mathrm{dense}}$. 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.

4.3 Model misspecification

We have so far focused on the case of Gaussian random design X_1, X_2 with identity covariance and Gaussian regression noises ϵ_1, ϵ_2. This complies with these assumptions that have helped us gain theoretical insights into the two-sample testing problem. However, our proposed testing procedures can still be used even if these modelling assumptions were not satisfied. To evaluate the performance of our proposed procedure under model misspecification, 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

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

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 standardises 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. Figure 3 compares the performance of $\psi_{\lambda,\tau}^{\mathrm{sparse}}$, $\psi_\eta^{\mathrm{dense}}$ 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 $\psi_{\lambda,\tau}^{\mathrm{sparse}}$ is robust to model misspecification and exhibits good power in all settings. The test $\psi_\eta^{\mathrm{desne}}$ is robust to non-normal design and noise, but exhibits a slight reduction in power in a correlated design. The advantage of $\psi_{\lambda,\tau}^{\mathrm{sparse}}$ and $\psi_\eta^{\mathrm{dense}}$ 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, $\psi_{\lambda,\tau}^{\text{sparse}}$ and $\psi_{\eta}^{\text{dense}}$ start having nontrivial power at a signal ℓ_2 norm 10 to 20 times smaller than that of the competitors.

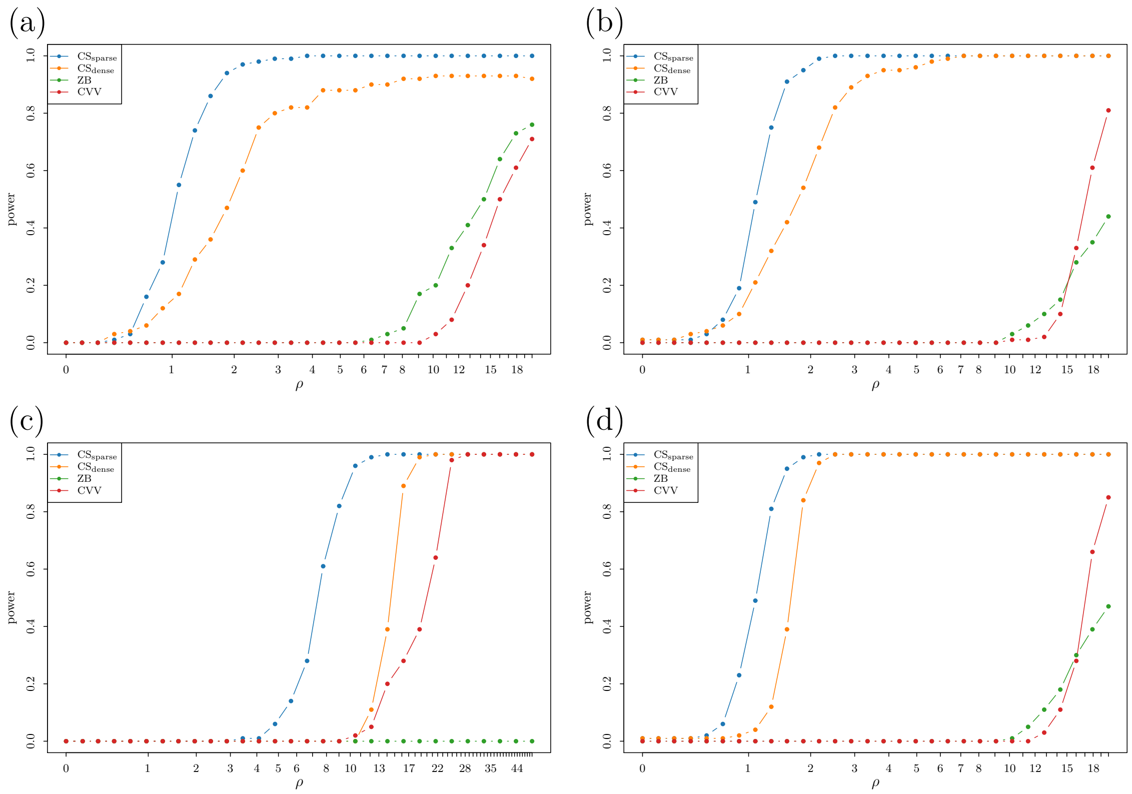

Figure 3: Power functions of different methods in models with non-Gaussian design or non-Gaussian noise, plotted against signal ℓ_2 norm ρ on a logarithmic grid. (a) correlated design matrix $\Sigma = (2^{-|i-j|})_{i,j\in[p]}$; (b) Rademacher design; (c) one-way balanced ANOVA design; (d) Gaussian design with $t_4/\sqrt{2}$-distributed noise. Details of the models are in Section 4.3.

5 Proof of main results

Proof of Theorem 1. Under the null hypothesis where $\beta_1 = \beta_2$, we have $\theta = 0$ and therefore, $Z = W\theta + \xi = \xi \sim N_m(0, I_m)$. In particular, $Q_j \sim N(0,1)$ for all $j \in [p]$. Thus, by a union bound, we have for $\lambda = \sqrt{(4+\epsilon)\log p}$ and any $\tau > 0$ that

$$\mathbb{P}(T \geq \tau) \leq \sum_{j=1}^{p} \mathbb{P}(|Q_j| \geq \lambda) \leq pe^{-\lambda^2/2} = p^{-1-\varepsilon/2},$$

The almost sure convergence is derived from the Borel–Cantelli lemma after noting that $p^{-1-\varepsilon/2}$ is summable for any $\varepsilon > 0$. $\square$

Proof of Proposition 2. Let $H \in \mathbb{O}^{p\times p}$ be any orthogonal matrix, then since $X \stackrel{\mathrm{d}}{=} XH$, by Lemma 9 we have

$$\begin{aligned} H^\top W^\top W H &= 4(H^\top X_1^\top X_1 H)(H^\top X^\top X H)^{-1}(H^\top X_2^\top X_2 H) \\ &\stackrel{\mathrm{d}}{=} 4(X_1^\top X_1)(X^\top X)^{-1}(X_2^\top X_2) = W^\top W. \end{aligned} \tag{14}$$

In particular, all diagonal entries of $W^\top W$ have the same distribution and all off-diagonal entries of $W^\top W$ have the same distribution. So it suffices to study $(W^\top W)_{1,1}$ and $(W^\top W)_{1,2}$.

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. By Lemma 13, 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 13, Q is uniformly distributed on $\mathbb{O}^{n\times p}$, which means $Q \stackrel{\mathrm{d}}{=} QH$ for any $H \in \mathbb{O}^{p\times p}$. Consequently $Q_1 \stackrel{\mathrm{d}}{=} Q_1 H$ and $B \stackrel{\mathrm{d}}{=} H^\top B H = (H^\top V)\Lambda(H^\top V)^\top$. 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.

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 9, we have

$$\begin{aligned} W^\top W &= 4X_1^\top A_1 A_1^\top X_1 = 4(X_1^\top X_1)(X_1^\top X_1 + X_2^\top X_2)^{-1}(X_2^\top X_2) \\ &= 4T^\top B(I_p - B)T = 4T^\top V\Lambda(I_p - \Lambda)V^\top T. \end{aligned} \tag{15}$$

Let $1 \ge \lambda_1 \ge \cdots \ge \lambda_p \ge 0$ 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 $s_1^2 \sim \chi^2_p$ and $r_1^2 \sim \chi^2_{n-p}$ such that $s_1 \ge 0$, $r_1 \ge 0$ are independent from each other and independent of everything else. By Lemma 13, 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 $\delta > 0$ 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 of the following inequalities hold:

$$\begin{gathered} \|a\|_1 - 2\|a\|_2\sqrt{\log(1/\delta)} \le \sum_{j=1}^p a_j G_{j,1}^2 \le \|a\|_1 + 2\|a\|_2\sqrt{\log(1/\delta)} + 2\|a\|_\infty \log(1/\delta), \\ p - 2\sqrt{p\log(1/\delta)} \le s_1^2 \le p + 2\sqrt{p\log(1/\delta)} + 2\log(1/\delta), \\ n - 2\sqrt{n\log(1/\delta)} \le t_1^2 \le n + 2\sqrt{n\log(1/\delta)} + 2\log(1/\delta). \end{gathered}$$

Using the fact that $\|a\|_\infty \leq 1/4$, we have with probability at least $1-6\delta$ that

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

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

$$\begin{aligned}\left|\frac{(W^\top W)_{1,1}}{4}-\frac{n}{p}\|a\|_1\right| &\leq \|a\|_1\frac{2\sqrt{n\log(1/\delta)}}{p}\bigl(1+\sqrt{n/p}\bigr)+\frac{n}{p}\left\{2\|a\|_2\sqrt{\log(1/\delta)}+\frac{\log(1/\delta)}{2}\right\}\\ &\quad+\mathcal{O}_s\left(\frac{\|a\|_1\log(1/\delta)}{p}+\frac{\|a\|_2\log(1/\delta)}{\sqrt{p}}+\frac{\log^{3/2}(1/\delta)}{p^{1/2}}\right).\end{aligned} \tag{16}$$

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

$$H^\top BH = H^\top T^{-\top}X_1^\top X_1T^{-1}H=(X^\top X)^{-1/2}(X_1^\top X_1)(X^\top X)^{-1/2},$$

which follows the matrix-variate Beta distribution $\mathrm{Beta}_p(n_1/2,n_2/2)$ as defined before Lemma 12. 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 Lemma 12, throughout the rest of this proof we may restrict ourselves to an almost sure event on which

$$\|a\|_1/p\to\kappa_1 \quad\text{and}\quad \|a\|_2/\sqrt{p}\to\kappa_2.$$

By (16), for each p, with probability at least $1-6\delta$, we have

$$\left|(W^\top W)_{1,1}-\frac{4n}{p}\|a\|_1\right| \leq 8\sqrt{n\log(1/\delta)}\Bigl((\kappa_1+\kappa_2)\sqrt{n/p}+\kappa_1\Bigr)+\mathcal{O}_s\left(\log(1/\delta)+\frac{\log^{3/2}(1/\delta)}{p^{1/2}}\right). \tag{17}$$

For the first claim in the proposition, we take $\delta=1/p^3$. Using (17), Lemma 12, a union bound over $j\in[p]$ and the Borel–Cantelli lemma (noting that $1/p^2$ is summable), we obtain that,

$$\max_{j\in[p]}\left|\frac{(W^\top W)_{j,j}}{4n\kappa_1}-1\right| \xrightarrow{\text{a.s.}} 0.$$

For the second part of the claim, define $\Delta := W^\top W-4np^{-1}\|a\|_1I_p$. By Lemma 11, there exists a $1/4$-net $\mathcal{N}$ of cardinality at most $\binom{p}{k}9^k$ such that

$$\|\Delta\|_{k,\mathrm{op}}\leq 2\sup_{u\in\mathcal{N}}u^\top\Delta u. \tag{18}$$

By (14), we have $u^\top\Delta u \stackrel{\mathrm{d}}{=} e_1^\top\Delta e_1$ for all $u\in\mathcal{S}^{p-1}$. Hence, by (17), (18) and a union bound over $u\in\mathcal{N}$, there exists $C_s>0$, depending only on s, such that for each p, with probability at least $1-6|\mathcal{N}|\delta$ that

$$\|\Delta\|_{k,\mathrm{op}}\leq 8\sqrt{n\log(1/\delta)}\Bigl\{(\kappa_1+\kappa_2)\sqrt{n/p}+\kappa_1\Bigr\}+\mathcal{O}_s\left(\log(1/\delta)+\frac{\log^{3/2}(1/\delta)}{p^{1/2}}\right). \tag{19}$$

If $\log(1/\delta) = o(p)$, as will be verified for our choice of δ later, then on the event where (19) is true, we have $\|W^\top W\|_{k,\mathrm{op}} = 4(1+o(1))np^{-1}\|a\|_1$ and $\|\mathrm{diag}(W^\top W)\|_{\mathrm{op}} = 4(1+o(1))np^{-1}\|a\|_1$. Define $D := \bigl(\frac{\mathrm{diag}(W^\top W)}{4np^{-1}\|a\|_1}\bigr)^{1/2}$, we have

$$\tilde{W}^\top\tilde{W} - \frac{W^\top W}{4n\|a\|_1/p} = (D^{-1}-I)\frac{W^\top W}{4n\|a\|_1/p}D^{-1} + \frac{W^\top W}{4n\|a\|_1/p}(D^{-1}-I).$$

Taking k-operator norms on both sides of the above identity, since D is a diagonal matrix, we have

$$\begin{aligned}\left\|\tilde{W}^\top\tilde{W} - \frac{W^\top W}{4n\|a\|_1/p}\right\|_{k,\mathrm{op}} &\le \|D^{-1}-I\|_{\mathrm{op}}\left\|\frac{W^\top W}{4n\|a\|_1/p}\right\|_{k,\mathrm{op}}(\|D^{-1}\|_{\mathrm{op}}+1)\\ &\le (2+o(1))\|D^{-1}-I_p\|_{\mathrm{op}} \le (1+o(1))\max_{j\in[p]}\left|\frac{(W^\top W)_{j,j}}{4np^{-1}\|a\|_1}-1\right|.\end{aligned}$$

Combining the above inequality with the observation that $|(W^\top W)_{j,j} - 4np^{-1}\|a\|_1| \le \|\Delta\|_{k,\mathrm{op}}$, we have by (19) and Lemma 12 that, asymptotically with probability at least $1-6|\mathcal{N}|\delta$,

$$\begin{aligned}\|\tilde{W}^\top\tilde{W}-I\|_{k,\mathrm{op}} &\le \left\|\tilde{W}^\top\tilde{W} - \frac{W^\top W}{4n\|a\|_1/p}\right\|_{k,\mathrm{op}} + \left\|\frac{W^\top W}{4n\|a\|_1/p} - I_p\right\|_{k,\mathrm{op}}\\ &\le (2+o(1))\frac{\|\Delta\|_{k,\mathrm{op}}}{4np^{-1}\|a\|_1} \le (4+o(1))\Bigl\{(1+\kappa_2/\kappa_1)\sqrt{n/p}+1\Bigr\}\sqrt{\frac{\log(1/\delta)}{n}}.\end{aligned}$$

Choosing $\delta := (10ep/k)^{-(k+4)}$. By (10), we indeed have $\log(1/\delta) = (k+4)\log(10ep/k) = o(p)$, as required in the above calculation. Also,

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

which is summable over p. Thus, by the Borel–Cantelli lemma, we see that for any $\varepsilon > 0$, with probability 1, the following sequence of events

$$\left\{\|\tilde{W}^\top\tilde{W}-I\|_{k,\mathrm{op}} > (4+\varepsilon)\Bigl\{(1+\kappa_2/\kappa_1)\sqrt{n/p}+1\Bigr\}\sqrt{\frac{(k+4)\log(10ep/k)}{n}}\right\}$$

happen finitely often. Hence the desired conclusion holds with, for instance, $C_{s,r} = 5(1+\sqrt{1+1/s}+\sqrt{s+r-1+1/s+1/r})$. □

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 $S := \mathrm{supp}(\theta) = \mathrm{supp}(\tilde{\theta})$, then

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

Our strategy will be to control $\|Q_S\|_2$. To this end, by (9), we have

$$\|\tilde{\theta}_S\|_2^2 = \sum_{j \in S} \theta_j^2 \|W_j\|_2^2 \geq \{4 - o(1)\} n\kappa_1 \sum_{j \in S} \theta_j^2 = (4 - o(1)) n\kappa_1 \rho^2. \tag{20}$$

Moreover, by (11) and (10), we have for sufficiently large p that

$$\|I_k - (\tilde{W}^\top \tilde{W})_{S,S}\|_{\mathrm{op}} \leq \|I_p - \tilde{W}^\top \tilde{W}\|_{k,\mathrm{op}} \leq C_{s,r}\sqrt{\frac{k\log(ep/k)}{n}} = o(1). \tag{21}$$

Define $R := \|(\tilde{W}^\top \tilde{W})_{S,S}\tilde{\theta}_S\|_2$. By the triangle inequality, (20) and (21), we have

$$R^2 \geq \Big[\|\tilde{\theta}_S\|_2 \Big\{1 - \|I_k - (\tilde{W}^\top \tilde{W})_{S,S}\|_{\mathrm{op}}\Big\}\Big]^2 \geq (4 - o(1)) n\kappa_1\rho^2 \geq (32 - o(1)) k \log p, \tag{22}$$

where we have evoked the condition $\rho^2 \geq 8k\log p/(n\kappa_1)$ in the final bound.

By Lemma 14 and (21), for sufficiently large p, we have with probability at least $1 - 2p^{-2}$ that

$$\begin{aligned}
\|Q_S\|_2^2 &\geq (1 - o(1))(k + R^2) - (2 + o(1))\Big\{\sqrt{2(k + 2R^2)\log p} + 2\log p\Big\} \\
&\geq (1 - o(1))(k + R^2) - (2 + o(1))\sqrt{(k + 2R^2)R^2/16} - (1/8 + o(1))R^2 \\
&\geq \bigg(1 - \frac{1}{4\sqrt{2}} - o(1)\bigg)k + \bigg(\frac{7}{8} - \frac{1}{\sqrt{2}} - o(1)\bigg)R^2 \geq 5k\log p, \qquad (23)
\end{aligned}$$

where both the second and the final inequalities hold because of (22).

From (23), using the tuning parameters $\lambda = 2\sqrt{\log p}$ and $\tau = k\log p$, we have for sufficiently large p that with probability at least $1 - 2p^{-2}$,

$$T = \sum_{j=1}^{p} Q_j^2 \mathbb{1}_{\{|Q_j| \geq \lambda\}} \geq \|Q_S\|_2^2 - k\lambda^2 \geq k\log p \geq \tau,$$

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

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 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 $\sup_{\beta \in \mathbb{R}^p} P^X_{\beta,\beta}(\psi^{\mathrm{sparse}}_{\lambda,\tau} \neq 0) \xrightarrow{\mathrm{a.s.}} 0$ and from Theorem 3 that $\sup_{\beta_1,\beta_2 \in \mathbb{R}^p : (\beta_1 - \beta_2)/2 \in \Theta_{p,k}(\rho)} P^X_{\beta,\beta}(\psi^{\mathrm{sparse}}_{\lambda,\tau} \neq 1) \xrightarrow{\mathrm{a.s.}} 0$. Combining the two completes the proof. □

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 $\mathcal{M} \ge 1 - o(1)$. Also, 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.

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 $P_0 := P_{0,0}^X$ and let $P_\pi := \int_{\theta\in\Theta_0} P_{L\theta,\theta}^X \, d\pi(\theta)$ denote the uniform mixture of $P_{\gamma,\theta}^X$ for $\{(\gamma,\theta) : \theta \in \Theta_0, \gamma = L\theta\}$. Let $\mathcal{L} := dP_\pi/dP_0$ be the likelihood ratio between the mixture alternative P_π and the simple null P_0. We have that

$$\begin{aligned}\mathcal{M} &\ge \inf_{\tilde{\psi}}\Big\{1 - (P_0 - P_\pi)\tilde{\psi}\Big\} = 1 - \frac{1}{2}\int\left|1 - \frac{dP_\pi}{dP_0}\right| dP_0 \\ &\ge 1 - \frac{1}{2}\left\{\int\left(1 - \frac{dP_\pi}{dP_0}\right)^2 dP_0\right\}^{1/2} \ge 1 - \frac{1}{2}\{P_0(\mathcal{L}^2) - 1\}^{1/2}.\end{aligned}$$

So it suffices to prove that $P_0(\mathcal{L}^2) \le 1 + o(1)$. Writing $\tilde{X}_1 = X_1 L + X_1$ and $\tilde{X}_2 = X_2 L - X_2$ and suppressing the dependence of P's on X in notations, by the definition of P_π, we compute that

$$\begin{aligned}\mathcal{L} = \int \frac{dP_{L\theta,\theta}}{dP_0}\, d\pi(\theta) &= \int \frac{e^{-\frac{1}{2}(\|Y_1 - X_1 L\theta - X_1\theta\|^2 + \|Y_2 - X_2 L\theta + X_2\theta\|^2)}}{e^{-\frac{1}{2}(\|Y_1\|^2 + \|Y_2\|^2)}}\, d\pi(\theta) \\ &= \int e^{\langle \tilde{X}_1\theta, Y_1\rangle - \frac{1}{2}\|\tilde{X}_1\theta\|^2 + \langle \tilde{X}_2\theta, Y_2\rangle - \frac{1}{2}\|\tilde{X}_2\theta\|^2}\, d\pi(\theta).\end{aligned}$$

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 $\mathrm{supp}(\theta) = J_0$. Let J, J' be independently and uniformly distributed on $\{J_0 \subseteq [p] : |J_0| = k\}$. By Fubini's theorem, we have

$$\begin{aligned}P_0(\mathcal{L}^2) &= \iint_{\theta,\theta'} e^{\frac{1}{2}\|\tilde{X}_1(\theta+\theta')\|^2 - \frac{1}{2}\|\tilde{X}_1\theta\|^2 - \frac{1}{2}\|\tilde{X}_1\theta'\| + \frac{1}{2}\|\tilde{X}_2(\theta+\theta')\|^2 - \frac{1}{2}\|\tilde{X}_2\theta\|^2 - \frac{1}{2}\|\tilde{X}_2\theta'\|^2}\, d\pi(\theta)\, d\pi(\theta') \\ &= \iint_{\theta,\theta'} e^{\theta^\top(\tilde{X}_1^\top\tilde{X}_1 + \tilde{X}_2^\top\tilde{X}_2)\theta'}\, d\pi(\theta)\, d\pi(\theta') \\ &= \iint_{\theta,\theta'} e^{\theta^\top W^\top W\theta'}\, d\pi(\theta)\, d\pi(\theta') \\ &= \mathbb{E}\bigg[\mathbb{E}\bigg\{e^{\tilde{\theta}_{J\cap J'}^\top\tilde{\theta}'_{J\cap J'}} \underbrace{\int_{\theta'_{J'}}\int_{\theta_J} e^{\tilde{\theta}_J^\top(\tilde{W}^\top\tilde{W} - I_p)_{J,J'}\tilde{\theta}'_{J'}}\, d\pi_J(\theta_J)\, d\pi_{J'}(\theta'_J)}_{\mathcal{I}} \,\bigg|\, J, J'\bigg\}\bigg], \end{aligned} \tag{24}$$

where we have employed Lemmas 10 and 9 in the penultimate equality and used the decomposition $\theta^\top W^\top W\theta' = \tilde{\theta}^\top\tilde{W}^\top\tilde{W}\tilde{\theta}' = \tilde{\theta}^\top(\tilde{W}^\top\tilde{W} - I_p)\tilde{\theta}' + \tilde{\theta}^\top\tilde{\theta}'$ in the final step.

We first control the integral $\mathcal{I}$. Define $\tilde{\theta} := (\tilde{\theta}_1, \ldots, \tilde{\theta}_p)^\top$ and $\tilde{\theta}' := (\tilde{\theta}'_1, \ldots, \tilde{\theta}'_p)^\top$ such that $\tilde{\theta}_j := \theta_j\|W_j\|_2$ and $\tilde{\theta}'_j := \theta'_j\|W_j\|_2$. By (9), we have

$$\vartheta := \max\Big\{\max_{j\in J}|\tilde{\theta}_j|, \max_{j\in J'}|\tilde{\theta}'_j|\Big\} \le (1 + o(1))\sqrt{\frac{4n\kappa_1\rho^2}{k}}. \tag{25}$$

By (25) and (11), we have for any $\rho = o(p^{-1/4}\log^{-3/4} p)$ (which includes the case $\rho \le \sqrt{\frac{(1-2\alpha-\varepsilon)k\log p}{4n\kappa_1}}$ for $k \le p^\alpha$ with $\alpha < 1/2$) that

$$\begin{aligned}\vartheta^2 \|(\tilde{W}^\top\tilde{W} - I_p)_{J,J'}\|_{\mathrm{F}} &\le \vartheta^2\|(\tilde{W}^\top\tilde{W} - I_p)_{J\cup J',J\cup J'}\|_{\mathrm{F}} \le \sqrt{2k}\vartheta^2\|\tilde{W}^\top\tilde{W} - I_p\|_{2k,\mathrm{op}} \\ &\le (1+o(1))\sqrt{2k}\frac{4n\kappa_1\rho^2}{k}\cdot C_{s,r}\sqrt{\frac{2k\log\{ep/(2k)\}}{n}} = o(\log^{-1}(p)).\end{aligned}$$

Consequently, by Arias-Castro, Candès and Plan (2011, Lemma 4), we have

$$\mathcal{I} \le e^{2\vartheta^2\|(\tilde{W}^\top\tilde{W}-I_p)_{J,J'}\|_{\mathrm{F}}\log(3k)} + \vartheta^2\|(\tilde{W}^\top\tilde{W} - I_p)_{J,J'}\|_{\mathrm{F}} \le e^{o(1)} + o(1) = 1 + o(1).$$

Plugging the above display in (24), we obtain

$$\begin{aligned}P_0(\mathcal{L}^2) &\le (1+o(1))\mathbb{E}\Big\{\mathbb{E}\Big(e^{\tilde{\theta}^\top_{J\cap J'}\tilde{\theta}'_{J\cap J'}} \,\Big|\, J\cap J'\Big)\Big\} \\ &= (1+o(1))\mathbb{E}\Big[\mathbb{E}\Big\{\prod_{j\in J\cap J'}\exp\Big(\tilde{\theta}_j\tilde{\theta}'_j\Big) \,\Big|\, J\cap J'\Big\}\Big] \\ &\le (1+o(1))\mathbb{E}\Big\{\prod_{j\in|J\cap J'|} 2\cosh(\vartheta^2)\Big\} = (1+o(1))\mathbb{E}\Big\{\cosh^{|J\cap J'|}(\vartheta^2)\Big\}. \qquad (26)\end{aligned}$$

Hence, it suffices to show that $\mathbb{E}\{\cosh^{|J\cap J'|}(\vartheta^2)\} = 1 + o(1)$. Note that $|J\cap J'| \sim \mathrm{HyperGeom}(k;k,p)$ is a hypergeometric random variable (defined as the number of black balls obtained from k draws without replacement from an urn containing p balls, k of which is black). Let $B \sim \mathrm{Bin}(k, k/p)$. By Hoeffding (1963, Theorem 4) and the fact that $\cosh(x) \le e^x$ for all $x \ge 0$, we have

$$\mathbb{E}\Big\{\cosh^{|J\cap J'|}(\vartheta^2)\Big\} \le \mathbb{E}e^{\vartheta^2|J\cap J'|} \le \mathbb{E}e^{\vartheta^2 B} = \Big\{1 + \frac{k}{p}\Big(e^{\vartheta^2} - 1\Big)\Big\}^k \le \exp\Big\{\frac{k^2}{p}(e^{\vartheta^2} - 1)\Big\}. \quad (27)$$

Since $\alpha \in [0, 1/2)$ and $\rho \le \sqrt{\frac{(1-2\alpha-\varepsilon)k\log p}{4n\kappa_1}}$, from (25), we can deduce that

$$\vartheta \le \sqrt{(1 - 2\alpha - \varepsilon + o(1))\log p}$$

and hence $e^{\vartheta^2}k^2/p = p^{-\varepsilon+o(1)} = o(1)$. So, from (27) we have $\mathbb{E}\{\cosh^{|J\cap J'|}(\vartheta^2)\} = 1+o(1)$, which 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. 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 $\xi \sim N_m(0, I_m)$. Hence, under the null hypothesis, we have $\|Z\|_2^2 \sim \chi^2_m$, which by Laurent and Massart (2000, Lemma 1) yields that

$$\mathbb{P}\Big\{\|Z\|_2^2 \ge m + 2\sqrt{m\log(1/\delta)} + 2\log(1/\delta)\Big\} \le \delta.$$

Setting $\delta = p^{-2}$, for $\eta = m + 2^{3/2}\sqrt{m\log p} + 4\log p$, we have by the Borel–Cantelli lemma that $\psi_\eta^{\mathrm{dense}}(X_1, X_2, Y_1, Y_2) \xrightarrow{\text{a.s.}} 0$.

On the other hand, under the alternative hypothesis, $\|Z\|_2^2 \sim \chi^2_m(\|W\theta\|_2^2)$, which by Birgé (2001, Lemma 8.1) implies that

$$\mathbb{P}\Big\{\|Z\|_2^2 \le m + \|W\theta\|_2^2 - 2\sqrt{(m + 2\|W\theta\|_2^2)\log(1/\delta)}\Big\} \le \delta.$$

Again, setting $\delta = p^{-2}$, observe from (11) and (10) that

$$\begin{aligned}\|W\theta\|_2^2 &= \|\tilde{W}\tilde{\theta}\|_2^2 = \|\tilde{\theta}\|_2^2 + \tilde{\theta}^\top(\tilde{W}^\top\tilde{W} - I_p)\tilde{\theta} \ge \|\tilde{\theta}\|_2^2\Big(1 - \|\tilde{W}^\top\tilde{W} - I_p\|_{k,\mathrm{op}}\Big)\\ &\ge (4 - o(1))n\kappa_1\rho^2,\end{aligned}$$

where the final bound comes from (20). Similarly we bound $\|W\theta\|_2^2 \le (4 + o(1))n\kappa_1\rho^2$. If $\rho^2 \ge \frac{2\sqrt{m\log p}}{n\kappa_1}$, then from the above display, we have $\|W\theta\|_2^2 \ge (8 - o(1))\sqrt{m\log p} \gg \log p$ and $\|W\theta\|_2^2 \le (8 + o(1))\sqrt{m\log p} \ll m$, and so

$$\begin{aligned}m + \|W\theta\|_2^2 &- 2\sqrt{(m + 2\|W\theta\|_2^2)\log(1/\delta)} - \eta\\ &\ge \|W\theta\|_2^2 - 2^{5/2}\sqrt{(m + 2\|W\theta\|_2^2)\log p} - 4\log p\\ &\ge (1 - o(1))\|W\theta\|_2^2 - 2^{5/2}(1 + o(1))\sqrt{m\log p} > 0\end{aligned}$$

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

Proof of Theorem 7. We follow the proof of Theorem 5 up to (26). Let $B \sim \mathrm{Bin}(k, k/p)$. By Hoeffding (1963, Theorem 4) and the fact that $\cosh(x) \le e^{x^2/2}$ for all $x \in \mathbb{R}$, we have

$$\mathbb{E}\Big\{\cosh^{|J\cap J'|}(\vartheta^2)\Big\} \le \mathbb{E}e^{|J\cap J'|\vartheta^4/2} \le \mathbb{E}e^{\vartheta^4 B/2} = \bigg\{1 + \frac{k}{p}\Big(e^{\vartheta^4/2} - 1\Big)\bigg\}^k \le \exp\bigg\{\frac{k^2}{p}(e^{\vartheta^4/2} - 1)\bigg\}.$$

If $\alpha \in [1/2, 1]$ and $\rho = o(p^{-1/4}\log^{-3/4} p)$, the by (25), $\vartheta^4 = o(p^{1-2\alpha}) = o(1)$, and hence $(e^{\vartheta^4/2} - 1)k^2/p = (1/2 + o(1))k^2\vartheta^4/p = o(1)$. By (26), $P_0(\mathcal{L}^2) = 1 + o(1)$ and we conclude as in the proof of Theorem 5. □

Proof of Proposition 8. 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 (24), we have

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

When $p \ge n_1$ or $p \ge n_2$, by Lemma 9, $W^\top W = (X_1^\top X_1)(X^\top X)^{-1}(X_2^\top X_2)$ 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 $p < \min\{n_1, n_2\}$, we work on the almost sure event where (9) and (11) hold. Let T and Λ be defined as in the proof of Proposition 2, then by (15), we have

$$\lambda_{\min}(W^\top W) \le 4\|T\|_{\mathrm{op}}^2\lambda_{\min}\Big(\Lambda(I - \Lambda)\Big) \le 4\|X^\top X\|_{\mathrm{op}}\min\{\lambda_{\min}(\Lambda), 1 - \lambda_{\max}(\Lambda)\}$$

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

$$\|X^\top X\|_{\mathrm{op}} \leq n\biggl(1 + \sqrt{\frac{p}{n}} + \sqrt{\frac{2\log p}{n}}\biggr)^2 \leq 5n$$

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 (28). 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 (28) with respect to $1-\xi$ and $1-\eta$ respectively, we obtain that

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

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

$$\rho^2\lambda_{\min}(W^\top W) = o\biggl(\max\biggl\{\frac{1}{(1-\xi)^2 p}, \frac{1}{(1-\eta)^2 p}\biggr\}\biggr)\mathcal{O}_\varepsilon(n\min\{(1-\xi)^2, (1-\eta)^2\}) \; = o(1),$$

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

6 Ancillary results

Lemma 9. *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 $X_2 \in \mathbb{R}^{n_2\times p}$. Let $A_1 \in \mathbb{R}^{n_1\times m}$ and $A_2 \in \mathbb{R}^{n_2\times m}$ be chosen 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 10. *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

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

The proof is complete by recalling the definitions of G_1 and G_2. $\square$

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 11. *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 $|\operatorname{supp}(v) \cup \operatorname{supp}(\tilde{v})| \leq k$ and $\|v - \tilde{v}\|_2 \leq 1/4$. We have

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

The desired inequality is obtained after rearranging terms in the above display. $\square$

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 $F^A(t) := n^{-1} \sum_{i=1}^{p} \mathbb{1}_{\{\lambda_i^A \leq t\}}$, 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 12. *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

$$t_\ell, t_r := \frac{(\xi + \eta)\eta + \xi\eta(\xi - \eta) \mp 2\xi\eta\sqrt{\xi - \xi\eta + \eta}}{(\xi + \eta)^2}. \tag{28}$$

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 13. *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 3.2.4) 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 lemma below provides concentration inequalities for the norm of a multivariate normal random vector with near-identity covariance.

Lemma 14. *Let $X \sim N_d(\mu, \Sigma)$. Suppose $\|\mu\|_2 = R$ and $\|\Sigma - I\|_{\mathrm{op}} \leq 1/2$. Then*

$$\mathbb{P}\Big[\|X\|_2^2 - (d+R^2)(1+2\|\Sigma-I\|_{\mathrm{op}}) \geq \Big\{2\sqrt{(d+2R^2)t}+2t\Big\}(1+2\|\Sigma-I\|_{\mathrm{op}})\Big] \leq 2e^{-t},$$
$$\mathbb{P}\Big[\|X\|_2^2 - (d+R^2)(1-2\|\Sigma-I\|_{\mathrm{op}}) \leq -\Big\{2\sqrt{(d+2R^2)t}+2t\Big\}(1+2\|\Sigma-I\|_{\mathrm{op}})\Big] \leq 2e^{-t}.$$

Proof. Define $Y = \Sigma^{-1/2}(X-\mu)$. Then

$$\|X\|_2^2 = Y^\top\Sigma Y + 2\mu^\top\Sigma^{1/2}Y + \|\mu\|_2^2 = \|Y+\mu\|_2^2 + Y^\top(\Sigma-I)Y + 2\mu^\top(\Sigma^{1/2}-I)Y.$$

Observe that $\|Y+\mu\|_2^2 \sim \chi_d^2(R^2)$. By Birgé (2001, Lemma 8.1), we have

$$\mathbb{P}\Big\{\|Y+\mu\|_2^2 \geq d + R^2 + 2\sqrt{(d+2R^2)t} + 2t\Big\} \leq e^{-t} \tag{29}$$

$$\mathbb{P}\Big\{\|Y+\mu\|_2^2 \leq d + R^2 - 2\sqrt{(d+2R^2)t}\Big\} \leq e^{-t} \tag{30}$$

On the other hand, we also have

$$\begin{aligned} |Y^\top(\Sigma - I)Y| + |2\mu^\top(\Sigma^{1/2} - I)Y| &\leq \|Y\|_2^2\|\Sigma - I\|_{\mathrm{op}} + 2R\|Y\|_2\|\Sigma^{1/2} - I\|_{\mathrm{op}} \\ &\leq (2\|Y\|_2^2 + R^2)\|\Sigma - I\|_{\mathrm{op}}, \end{aligned} \tag{31}$$

where the final inequality follows from a Taylor expansion of $\{I + (\Sigma - I)\}^{1/2}$ after noting $\|\Sigma - I\|_{\mathrm{op}} \leq 1/2$ and the elementary inequality $2ab \leq a^2 + b^2$ for $a, b \in \mathbb{R}$. By Laurent and Massart (2000, Lemma 1), we have with probability at least $1 - e^{-t}$ that

$$\|Y\|_2^2 \leq d + 2\sqrt{dt} + 2t. \tag{32}$$

Combining (29), (31) and (32), we have with probability at least $1 - 2e^{-t}$ that

$$\|X\|_2^2 \leq \Big\{d + R^2 + 2\sqrt{(d+2R^2)t} + 2t\Big\}(1 + 2\|\Sigma - I\|_{\mathrm{op}}).$$

Similarly, for the lower bound, we have again by (30), (31) and (32) that

$$\begin{aligned} \|X\|_2^2 &\geq d + R^2 - 2\sqrt{(d+2R^2)t} - \{2d + 4\sqrt{dt} + 4t + R^2\}\|\Sigma - I\|_{\mathrm{op}} \\ &\geq (d + R^2)(1 - 2\|\Sigma - I\|_{\mathrm{op}}) - \Big\{2\sqrt{(d+2R^2)t} + 2t\Big\}(1 + 2\|\Sigma - I\|_{\mathrm{op}}). \end{aligned}$$

holds with probability at least $1 - 2e^{-t}$. □

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, 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 Lepski's 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.

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.

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.

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

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.

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

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

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

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.

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.

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.