Journal of Multivariate Analysis 169 (2019) 312–329

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Robust two-sample test of high-dimensional mean vectors under dependence

Wei Wang, Nan Lin *, Xiang Tang

Department of Mathematics and Statistics, Washington University in St. Louis, St. Louis, MO 63130, USA

ARTICLE INFO

Article history:
Received 14 December 2017
Available online 9 October 2018

AMS subject classifications:
62H15

Keywords:
Cell-wise contamination
Robust precision matrix estimation
Sparse and strong alternatives
Two-sample mean test
Trimmed mean

ABSTRACT

A basic problem in modern multivariate analysis is testing the equality of two mean vectors in settings where the dimension p increases with the sample size n. This paper proposes a robust two-sample test for high-dimensional data against sparse and strong alternatives, in which the mean vectors of the populations differ in only a few dimensions, but the magnitude of the differences is large. The test is based on trimmed means and robust precision matrix estimators. The asymptotic joint distribution of the trimmed means is established, and the proposed test statistic is shown to have a Gumbel distribution in the limit. Simulation studies suggest that the numerical performance of the proposed test is comparable to that of non-robust tests for uncontaminated data. For cell-wise contaminated data, it outperforms non-robust tests. An illustration involves biomarker identification in an Alzheimer's disease dataset.

1. Introduction

In many contemporary applications, such as genomics and finance, high-dimensional data whose dimension p is comparable to, or larger than, the sample size n are ubiquitous. For example, gene expression data commonly measure tens of thousands of genes from a few hundred or fewer individuals. Many traditional statistical methods perform poorly or are even no longer applicable in such 'large p, small n' settings.

In this paper, we study the problem of testing the equality of p-dimensional mean vectors based on random samples independently drawn from two populations. To be specific, suppose that one collects random samples $\mathbf{X}_1, \ldots, \mathbf{X}_{n_1} \sim \mathcal{N}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $\mathbf{Y}_1, \ldots, \mathbf{Y}_{n_2} \sim \mathcal{N}_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ from which it is desired to test the hypotheses

$$\mathcal{H}_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \quad \text{vs.} \quad \mathcal{H}_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2. \tag{1}$$

The standard approach to testing (1) is through Hotelling's T^2 statistic [14], viz.

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} (\bar{\mathbf{X}} - \bar{\mathbf{Y}})^\top \mathbf{S}_n^{-1} (\bar{\mathbf{X}} - \bar{\mathbf{Y}}),$$

where $\bar{\mathbf{X}}$ and $\bar{\mathbf{Y}}$ are sample means, viz.

$$\bar{\mathbf{X}} = \frac{1}{n_1} \sum_{k=1}^{n_1} \mathbf{X}_k, \quad \bar{\mathbf{Y}} = \frac{1}{n_2} \sum_{\ell=1}^{n_2} \mathbf{Y}_\ell,$$

* Corresponding author.
E-mail address: nlin@wustl.edu (N. Lin).

https://doi.org/10.1016/j.jmva.2018.09.013
0047-259X/

and $\mathbf{S}_n$ is the pooled sample covariance matrix, viz.

$$\mathbf{S}_n = \frac{1}{n_1+n_2-2}\left\{\sum_{k=1}^{n_1}(\mathbf{X}_k-\bar{\mathbf{X}})(\mathbf{X}_k-\bar{\mathbf{X}})^\top + \sum_{\ell=1}^{n_2}(\mathbf{Y}_\ell-\bar{\mathbf{Y}})(\mathbf{Y}_\ell-\bar{\mathbf{Y}})^\top\right\}. \tag{2}$$

However, the power of Hotelling's T^2 test is low when the dimension p is high relative to the sample size n_1+n_2; see [15]. Furthermore, Hotelling's T^2 test becomes ill-defined for $p \geq n_1+n_2$ since $\boldsymbol{S}_n$ is singular.

In order to address these limitations in high-dimensional settings, various parametric and nonparametric methods have been proposed in recent years. Some nonparametric tests can be found, e.g., in [23,27]. As for parametric tests, Cai et al. [8] noted that they often rely on sum-of-squares-type statistics which aim to estimate the Euclidean norm $\|\mathbf{A}^{1/2}(\boldsymbol{\mu_1}-\boldsymbol{\mu_2})\|_2^2$ for some positive definite matrix $\mathbf{A}$; see, e.g., [3,9,22]. These tests all require $(n_1+n_2)\,\|\boldsymbol{\mu_1}-\boldsymbol{\mu_2}\|_2^2/\sqrt{p} \to \infty$ in order to distinguish between the null and alternative hypotheses with probability tending to 1. The sum-of-squares-type tests can have good power against so-called 'weak but dense' alternatives where the 'signals', i.e., non-zero entries in $\boldsymbol{\mu_1}-\boldsymbol{\mu_2}$, spread over a large number of dimensions, but the signal in each dimension is weak.

In applications such as genomics, however, the alternatives are 'sparse and strong', i.e., the mean vectors of two populations differ only in a few coordinates, but the magnitude of these rare signals is large. In such cases, the condition $(n_1+n_2)\,\|\boldsymbol{\mu_1}-\boldsymbol{\mu_2}\|_2^2/\sqrt{p} \to \infty$ does not hold, and hence Cai et al. [8] proposed a test based on the Kolmogorov distance. The test statistic is

$$M_{\boldsymbol{\Omega}} = \frac{n_1 n_2}{n_1+n_2}\max(\bar{Z}_1^2/v_{11},\ldots,\bar{Z}_p^2/v_{pp}),$$

where $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1} = (v_{ij})_{p\times p}$ is the true precision matrix, and $\bar{\mathbf{Z}} = (\bar{Z}_1,\ldots,\bar{Z}_p)^\top = \boldsymbol{\Omega}(\bar{\mathbf{X}}-\bar{\mathbf{Y}})$. The CLIME estimator in Cai et al. [7] can be used to estimate the precision matrix $\boldsymbol{\Omega}$.

High-dimensional data often contain a few outliers in each dimension, and hence there can be many outliers overall. A shared drawback of the aforementioned tests is that they are sensitive to outliers and tend to have unsatisfactory power when the multivariate distribution is heavy-tailed [25]. The goal of this paper is to develop a robust two-sample test that performs particularly well against sparse and strong alternatives for high-dimensional cell-wise contaminated data. Inspired by Cai et al. [8], we introduce a new test statistic based on the Kolmogorov distance, and adopt trimmed means and robust precision matrix estimators to achieve robustness. Under the null hypothesis, our test statistic has a limiting Gumbel distribution from which an asymptotic test can be designed for uncontaminated data. We show that in the absence of contamination, the performance of our robust test is comparable to that of Cai et al. [8]. More importantly, we show that our test has overwhelming power against the test in [8] for contaminated data, even if the contamination proportion is modest in each dimension.

In Section 2, we introduce our proposed test statistic. Section 3 gives the asymptotic properties of the test statistic under no contamination. In Section 4, we address the estimation of the precision matrix used in the test statistic when data are contaminated. Section 5 further discusses how to determine trimming levels. Simulation studies reported in Section 6 demonstrate the numerical performance of our method for both uncontaminated data and cell-wise contaminated data. In Section 7, our approach is applied to an Alzheimer's disease dataset. Discussions and other related work are given in Section 8. Proofs of the main results and theoretical details are in Appendix A.

2. Methodology

We deal with the testing problem (1) against sparse and strong alternatives. In Sections 2 and 3, we consider the uncontaminated scenario. Throughout the paper, our theoretical analysis is carried out under a normality assumption for convenience. However, this assumption is not indispensable as long as the tails of the distributions are not too heavy. We will first introduce our test procedure in the oracle setting where the precision matrix $\boldsymbol{\Omega}$ is known. We will then present a data-driven procedure for the general case of an unknown precision matrix $\boldsymbol{\Omega}$.

For a matrix $\mathbf{A} = (a_{ij})_{p\times p}$, we define the matrix L_1-norm, the matrix infinite norm and the element-wise ℓ_1-norm as

$$\|\mathbf{A}\|_{L_1} = \max_{1\leq j\leq p}\sum_{i=1}^{p}|a_{ij}|, \quad \|\mathbf{A}\|_\infty = \max_{i,j\in\{1,\ldots,p\}}|a_{ij}|, \quad \|\mathbf{A}\|_1 = \sum_{i=1}^{p}\sum_{j=1}^{p}|a_{ij}|,$$

respectively. We also define the operator norm as the largest eigenvalue $\|\mathbf{A}\|_{\text{op}} = \lambda_{\max}(\mathbf{A})$. For two sequences a_n and b_n, we write $a_n \asymp b_n$ if there exist two constants c, $C > 0$ such that $c \leq a_n/b_n \leq C$ for all $n \geq 1$; $a_n = O(b_n)$ if $|a_n| \leq C|b_n|$ holds for all sufficiently large n; and $a_n = o(b_n)$ if $a_n/b_n \to 0$ as $n \to \infty$.

2.1. Oracle procedure

Our statistic for testing (1) is

$$T_{\boldsymbol{\Omega}_{\boldsymbol{\eta}}} = \frac{n_1 n_2}{n_1+n_2}\max\{(\bar{Z}_{\eta_1,1})^2/w_{11},\ldots,(\bar{Z}_{\eta_p,p})^2/w_{pp}\}, \tag{3}$$

where $\bar{\mathbf{Z}}_{\boldsymbol{\eta}} = (\bar{Z}_{\eta_1,1}, \ldots, \bar{Z}_{\eta_p,p})^\top = \boldsymbol{\Omega}_{\boldsymbol{\eta}}(\bar{\mathbf{X}}_{\boldsymbol{\eta}} - \bar{\mathbf{Y}}_{\boldsymbol{\eta}})$, and $\boldsymbol{\Omega}_{\boldsymbol{\eta}} = (w_{ij})_{p\times p} = \boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{-1}$ is the common precision matrix for the $\boldsymbol{\eta}$-trimmed mean vectors $\sqrt{n_1}\,\bar{\mathbf{X}}_{\boldsymbol{\eta}}$ and $\sqrt{n_2}\,\bar{\mathbf{Y}}_{\boldsymbol{\eta}}$ with trimming level vector $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_p)^\top$.

Definition 1. For a random sample $X_1, \ldots, X_n$, let $X_{(i)}$ be the ith order statistic of the sample. For $\eta \in (0, 1/2)$, and with $\lfloor\cdot\rfloor$ denoting the floor function, the sample η-trimmed mean is given by

$$\bar{X}_\eta = \frac{1}{n - 2\lfloor n\eta\rfloor} \sum_{i=\lfloor n\eta\rfloor+1}^{n-\lfloor n\eta\rfloor} X_{(i)}.$$

Furthermore, for $\eta_1, \ldots, \eta_p \in (0, 1/2)$ and $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_p)^\top$, the $\boldsymbol{\eta}$-trimmed mean vectors of two p-variate samples are defined by $\bar{\mathbf{X}}_{\boldsymbol{\eta}} = (\bar{X}_{\eta_1,1}, \ldots, \bar{X}_{\eta_p,p})^\top$ and $\bar{\mathbf{Y}}_{\boldsymbol{\eta}} = (\bar{Y}_{\eta_1,1}, \ldots, \bar{Y}_{\eta_p,p})^\top$.

The following theorem gives the asymptotic distribution of the trimmed mean vector under normality in high-dimensional settings.

Theorem 1. *Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be a random sample of a p-variate random vector $\mathbf{X} = (X_1, \ldots, X_p)^\top \in \mathbb{R}^p$ from a multivariate normal distribution $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^\top$ and $\boldsymbol{\Sigma} = (\sigma_{ij})_{p\times p}$, where $K_0^{-1} \le \lambda_{\min}(\boldsymbol{\Sigma}) \le \lambda_{\max}(\boldsymbol{\Sigma}) \le K_0$ for some constant $K_0 > 0$. Suppose that for all $i \in \{1, \ldots, p\}$, $\eta_i \in [b_1, b_2]$ for some $0 \le b_1 < b_2 < 1/2$, and $\eta_i \to b_1$ as $p \to \infty$. If $\{\ln(pn)\}^7/n \to 0$ as $n, p \to \infty$, then $\sqrt{n}\,(\bar{\mathbf{X}}_{\boldsymbol{\eta}} - \boldsymbol{\mu}) \rightsquigarrow \mathcal{N}(0, \boldsymbol{\Sigma}_{\boldsymbol{\eta}})$, where $\boldsymbol{\Sigma}_{\boldsymbol{\eta}} = (\sigma^*_{ij})_{p\times p}$. Furthermore, if for each $i, j \in \{1, \ldots, p\}$ with $i \ne j$, F_i denotes the cdf of X_i and F_{ij} denotes the cdf of the pair (X_i, X_j), then the ith diagonal entry of $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$ is*

$$\sigma^*_{ii} = \begin{cases} \sigma_{ii} & \text{if } \eta_i = 0, \\ \dfrac{1}{(1-2\eta_i)^2}\left[\displaystyle\int (x-\mu_i)^2 \mathbf{1}\{F_i^{-1}(\eta_i) < x < F_i^{-1}(1-\eta_i)\}dF_i(x) + 2\eta_i\{F_i^{-1}(\eta_i) - \mu_i\}^2\right] & \text{if } \eta_i > 0, \end{cases} \tag{4}$$

where $\mathbf{1}(A)$ is the indicator function of the set A. Moreover, the (i, j)th off-diagonal entry of $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$ is

$$\sigma^*_{ij} = \begin{cases} \sigma_{ij} & \text{if } \eta_i = \eta_j = 0, \\ \Big[\{F_j^{-1}(\eta_j) - \mu_j\}\displaystyle\int x\mathbf{1}\{y < F_j^{-1}(\eta_j)\}dF_{ij}(x,y) & \\ \quad + \displaystyle\int (x-\mu_i)(y-\mu_j)\mathbf{1}\{F_j^{-1}(\eta_j) < y < F_j^{-1}(1-\eta_j)\}dF_{ij}(x,y) & \\ \quad + \{F_j^{-1}(1-\eta_j) - \mu_j\}\displaystyle\int x\mathbf{1}\{y > F_j^{-1}(1-\eta_j)\}dF_{ij}(x,y)\Big]/(1-2\eta_j) & \text{if } \eta_i = 0,\ \eta_j > 0, \\ \Big[2\{F_i^{-1}(\eta_i) - \mu_i\}\{F_j^{-1}(\eta_j) - \mu_j\}\Pr\{x < F_i^{-1}(\eta_i), y < F_j^{-1}(\eta_j)\} & \\ \quad + 2\{F_i^{-1}(\eta_i) - \mu_i\}\{F_j^{-1}(1-\eta_j) - \mu_j\}\Pr\{x < F_i^{-1}(\eta_i), y > F_j^{-1}(1-\eta_j)\} & \\ \quad + \tilde{\mu}_1 + \tilde{\mu}_2 + \tilde{\mu}_3 + \tilde{\mu}_4 + \tilde{\rho}\Big]/\{(1-2\eta_i)(1-2\eta_j)\} & \text{if } \eta_i,\ \eta_j > 0, \end{cases} \tag{5}$$

where $\tilde{\mu}_1$, $\tilde{\mu}_2$, $\tilde{\mu}_3$, $\tilde{\mu}_4$ contain integrals related to truncated means, viz.

$$\tilde{\mu}_1 = \{F_j^{-1}(\eta_j) - \mu_j\}\int x\mathbf{1}\{F_i^{-1}(\eta_i) < x < F_i^{-1}(1-\eta_i), y < F_j^{-1}(\eta_j)\}dF_{ij}(x,y),$$

$$\tilde{\mu}_2 = \{F_j^{-1}(1-\eta_j) - \mu_j\}\int x\mathbf{1}\{F_i^{-1}(\eta_i) < x < F_i^{-1}(1-\eta_i), y > F_j^{-1}(1-\eta_j)\}dF_{ij}(x,y),$$

$$\tilde{\mu}_3 = \{F_i^{-1}(\eta_i) - \mu_i\}\int y\mathbf{1}\{x < F_i^{-1}(\eta_i), F_j^{-1}(\eta_j) < y < F_j^{-1}(1-\eta_j)\}dF_{ij}(x,y),$$

$$\tilde{\mu}_4 = \{F_i^{-1}(1-\eta_i) - \mu_i\}\int y\mathbf{1}\{x > F_i^{-1}(1-\eta_i), F_j^{-1}(\eta_j) < y < F_j^{-1}(1-\eta_j)\}dF_{ij}(x,y),$$

and $\tilde{\rho}$ contains an integral related to truncated correlations,

$$\tilde{\rho} = \int (x-\mu_i)(y-\mu_j)\mathbf{1}\{F_i^{-1}(\eta_i) < x < F_i^{-1}(1-\eta_i), F_j^{-1}(\eta_j) < y < F_j^{-1}(1-\eta_j)\}dF_{ij}(x,y).$$

For Theorem 1, the bound $[b_1, b_2]$ on the trimming levels η_i guarantees, together with the condition that $\eta_i \to b_1$, that the high-dimensional trimmed mean vector admits a Bahadur type presentation based on the influence functions of trimmed means as $p \to \infty$. Given that the vector of influence functions of trimmed means belongs to the class of all hyperrectangles in $\mathbb{R}^p$ of the form $\{\boldsymbol{w} = (w_1, \ldots, w_p)^\top \in \mathbb{R}^p : a_i \le w_i \le b_i \text{ for all } i \in \{1, \ldots, p\}\}$, we can establish the asymptotic distribution

of the high-dimensional trimmed mean vector based on the Central Limit Theorem for hyperrectangles. The condition that $\{\ln(pn)\}^7/n \to 0$ as $n, p \to \infty$ is needed for that purpose.

Remark 1. Throughout the proof of Theorem 1 given in Appendix A, we only used specific properties of normal distributions in (A.8)–(A.9), i.e., moments of the standard normal and the moment generating function of the standard half-normal. Actually, it also works for distributions with a sub-exponential tail, i.e., the tail of X_{ki} satisfies $\Pr(|X_{ki}| \geq t) \leq 2\exp(-t/K_1)$ for all $t \geq 0$ and some constant $K_1 > 0$. If a distribution has a sub-exponential tail, its first four absolute moments are finite and the moment generating function of $|X_{ki}|$ is bounded; see Proposition 2.7.1 in [24]. Thus, normality is not indispensable. Essentially, the assumption of sub-exponential tails results from the requirement of the Central Limit Theorem for hyperrectangles in the high-dimensional setting (Lemma A.4), which is the core of the proof for Theorem 1. Conditions (M.2)–(M.3) in Lemma A.4 are related to finite moment requirements on the first four absolute moments and a bounded moment generating function of $|X_{ki}|$, respectively.

Note that $\bar{\mathbf{X}}_{\boldsymbol{\eta}}$ and $\bar{\mathbf{Y}}_{\boldsymbol{\eta}}$ in (3) share the same trimming level vector $\boldsymbol{\eta}$. If two populations share the common covariance matrix $\boldsymbol{\Sigma}$ and trimming level vector $\boldsymbol{\eta}$, then it follows directly from Theorem 1 that the trimmed mean vectors $\sqrt{n_1}\,\bar{\mathbf{X}}_{\boldsymbol{\eta}}$ and $\sqrt{n_2}\,\bar{\mathbf{Y}}_{\boldsymbol{\eta}}$ share a common covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$, and hence have the same precision matrix $\boldsymbol{\Omega}_{\boldsymbol{\eta}}$. The linear transformation of the data by the precision matrix $\boldsymbol{\Omega}_{\boldsymbol{\eta}}$ in (3) magnifies the signals, owing to the dependence within the data, and hence improves distinguishing the null and alternative hypotheses [8].

2.2. Data-driven procedure

In most applications, the precision matrix $\boldsymbol{\Omega}_{\boldsymbol{\eta}}$ is unknown and needs to be estimated. The CLIME estimator proposed by Cai et al. [7] is suitable for sparse precision matrix estimation. Their numerical analysis demonstrated that the CLIME estimator has desirable performance for estimating non-sparse precision matrices as well.

The CLIME estimator is a solution of the optimization problem

$$\min\|\boldsymbol{\Omega}\|_1, \text{ subject to } \|\mathbf{S}_n\boldsymbol{\Omega} - \mathbf{I}\|_\infty \leq \lambda_n, \tag{6}$$

where $\lambda_n = C\ln(p)/n$ for some sufficiently large constant C. In practice, the tuning parameter λ_n can be chosen through cross-validation. The matrix $\mathbf{S}_n$ is the pooled sample covariance matrix given by (2).

To estimate $\boldsymbol{\Omega}_{\boldsymbol{\eta}}$ in our test statistic (3), we replace $\mathbf{S}_n$ in (6) by $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\eta}}$, an estimator of the asymptotic covariance matrix for the trimmed mean which solves

$$\min\|\boldsymbol{\Omega}\|_1, \text{ subject to } \|\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\eta}}\boldsymbol{\Omega} - \mathbf{I}\|_\infty \leq \lambda_n. \tag{7}$$

Let $\hat{\boldsymbol{\Omega}}_1$ be a solution of the optimization problem in (7), and $\hat{w}_{ij}^1$ be the (i,j)th entry of $\hat{\boldsymbol{\Omega}}_1$. The CLIME estimator of the precision matrix $\boldsymbol{\Omega}_{\boldsymbol{\eta}}$ is defined to be $\hat{\boldsymbol{\Omega}}_{\boldsymbol{\eta}} = (\hat{w}_{ij})_{p\times p}$, where

$$\hat{w}_{ij} = \hat{w}_{ji} = \hat{w}_{ij}^1\mathbf{1}(|\hat{w}_{ij}^1| \leq |\hat{w}_{ji}^1|) + \hat{w}_{ji}^1\mathbf{1}(|\hat{w}_{ij}^1| > |\hat{w}_{ji}^1|).$$

To obtain $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\eta}}$, by Theorem 1, it suffices to solve the cdfs F_i and F_{ij}. Since each variable is normally distributed, given that the data are uncontaminated, we can use the sample mean and the sample covariance matrix to determine each F_i and F_{ij} in (4) and (5).

For testing the equality of two mean vectors, we apply the above procedure to $\boldsymbol{X}_k$ and $\boldsymbol{Y}_\ell$ separately, and then get the estimators of the asymptotic covariance matrices of the trimmed mean vectors $\sqrt{n_1}\,\bar{\mathbf{X}}_{\boldsymbol{\eta}}$ and $\sqrt{n_2}\,\bar{\mathbf{Y}}_{\boldsymbol{\eta}}$. Let $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\eta}}^X$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\eta}}^Y$ be these two estimators. The robust alternative $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\eta}}$ used in (7) is given by

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\eta}} = \frac{1}{n_1+n_2-2}\{(n_1-1)\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\eta}}^X + (n_2-1)\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\eta}}^Y\}.$$

By solving the optimization problem (7), we obtain an estimator $\hat{\boldsymbol{\Omega}}_{\boldsymbol{\eta}}$ of the precision matrix $\boldsymbol{\Omega}_{\boldsymbol{\eta}}$. For testing the hypothesis (1), our final test statistic $T_{\hat{\boldsymbol{\Omega}}_{\boldsymbol{\eta}}}$ is given by

$$T_{\hat{\boldsymbol{\Omega}}_{\boldsymbol{\eta}}} = \frac{n_1n_2}{n_1+n_2}\max\{(\hat{Z}_{\eta_1,1})^2/\hat{w}_{11}, \ldots, (\hat{Z}_{\eta_p,p})^2/\hat{w}_{pp}\},$$

where $\hat{\mathbf{Z}}_{\boldsymbol{\eta}} = (\hat{Z}_{\eta_1,1}, \ldots, \hat{Z}_{\eta_p,p})^\top = \hat{\boldsymbol{\Omega}}_{\boldsymbol{\eta}}(\bar{\mathbf{X}}_{\boldsymbol{\eta}} - \bar{\mathbf{Y}}_{\boldsymbol{\eta}})$, and $\hat{\boldsymbol{\Omega}}_{\boldsymbol{\eta}} = (\hat{w}_{ij})_{p\times p}$. Note that when $\boldsymbol{\eta} = \mathbf{0}$, this test boils down to the test proposed by Cai et al. [8].

The current estimator of $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\eta}}$ relies on the assumption of normality to guarantee the optimal convergence rate because the estimation is based on the sample covariance matrix $\mathbf{S}_n$. For non-normal distributions, we need to modify the estimation of $\boldsymbol{\Sigma}$. Avella-Medina et al. [2] presented three pilot covariance matrix estimators which perform better than the sample covariance matrix when the sub-Gaussian assumption fails. The adaptive Huber estimator therein only requires the existence of the fourth moments. As long as we have consistent estimators of means and variances, all the calculations involving F_i and F_{ij} in (4) and (5) can be carried out.

3. Theoretical analysis

3.1. Asymptotic distribution of the oracle test statistic

In the following sections, we assume that $n_1 \asymp n_2$, and $n = n_1 n_2/(n_1 + n_2)$. Let $\mathbf{D} = \mathrm{diag}(v_{11}, \ldots, v_{pp})$ and $\mathbf{D}_{\boldsymbol{\eta}} = \mathrm{diag}(w_{11}, \ldots, w_{pp})$, where v_{ii} is the diagonal entry of the precision matrix $\boldsymbol{\Omega}$, and w_{ii} is the diagonal entry of the precision matrix $\boldsymbol{\Omega}_{\boldsymbol{\eta}}$. Define $\mathbf{R} = (r_{ij})_{p\times p} = \mathbf{D}^{-1/2}\boldsymbol{\Omega}\mathbf{D}^{-1/2}$ and $\mathbf{R}_{\boldsymbol{\eta}} = (r^*_{ij})_{p\times p} = \mathbf{D}_{\boldsymbol{\eta}}^{-1/2}\boldsymbol{\Omega}_{\boldsymbol{\eta}}\mathbf{D}_{\boldsymbol{\eta}}^{-1/2}$. To obtain the null distributions of the test statistics, we assume the following conditions:

Condition 1: $K_0^{-1} \le \lambda_{\min}(\boldsymbol{\Sigma}) \le \lambda_{\max}(\boldsymbol{\Sigma}) \le K_0$ for some constant $K_0 > 0$.
Condition 2: $\max_{1\le i<j\le p}|r_{ij}| \le r_1 < 1$ for some constant $r_1 \in (0, 1)$.
Condition 3: For any $i \in \{1, \ldots, p\}$, the trimming proportion in the ith dimension $\eta_i \in [0, b_2]$, where $b_2 \in (0, 1/2)$, and as $p \to \infty$, the sequence η_i converges to 0.
Condition 4: $\{\ln(pn)\}^7/n \to 0$ as $n, p \to \infty$.
Condition 5: $\sup_{1\le i\le p} 1/|F_i^{-1}(\eta_i)| = o(1/\sqrt{\ln p})$, where for each $i \in \{1, \ldots, p\}$, F_i is the cdf of X_i or Y_i.

Remark 2. Condition 1 states that the eigenvalues of the covariance matrix $\boldsymbol{\Sigma}$ are bounded from below and above, which is a common assumption in high-dimensional settings [7,8]. Condition 2 is a mild requirement preventing $\boldsymbol{\Omega}$ from being singular. Conditions 3 and 4 are used in Theorem 1 to derive the asymptotic normality of the trimmed mean vector.

The result below, proved in Appendix A.3, shows that when constraints on the upper bound b_2 of trimming levels are imposed under Condition 5, Conditions 1–2 on $\boldsymbol{\Sigma}$ and $\boldsymbol{\Omega}$ imply parallel conditions on their counterparts $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$ and $\boldsymbol{\Omega}_{\boldsymbol{\eta}}$.

Lemma 1. *Suppose that Conditions 1, 3, 4 and 5 hold. Then,*

(a) $C_0^{-1} \le \lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{\eta}}) \le \lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{\eta}}) \le C_0$ *for some constant* $C_0 > 0$ *as* $n, p \to \infty$.
(b) *Furthermore, if Condition 2 holds, then* $\|\mathbf{R}_{\boldsymbol{\eta}}\|_\infty \le r_2 < 1$ *for some constant* $r_2 \in (0, 1)$ *as* $n, p \to \infty$.

The limiting null distribution of the statistic $T_{\boldsymbol{\Omega}_{\boldsymbol{\eta}}}$ in (3) is given next.

Theorem 2. *Suppose that Conditions 1–5 hold. Then under* $\mathcal{H}_0$ *in* (1)*, for any* $x \in \mathbb{R}$*, as* $n, p \to \infty$*,*

$$\Pr_{\mathcal{H}_0}[T_{\boldsymbol{\Omega}_{\boldsymbol{\eta}}} - 2\ln(p) + \ln\{\ln(p)\} \le x] \to \exp\{-\pi^{-1/2}\exp(-x/2)\} = \exp[-\exp[-\{x + \ln(\pi)\}/2]].$$

Remark 3. Based on the asymptotic normality in Theorem 1 for the trimmed means, the asymptotic properties of our test will be the same to those of the test in [8] so long as the conditions on $\boldsymbol{\Sigma}$ and $\boldsymbol{\Omega}$ for asymptotic properties of the test in [8] can imply similar conditions on $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$ and $\boldsymbol{\Omega}_{\boldsymbol{\eta}}$. It follows by Lemma 1 that the proof of Theorem 2 is very similar to Theorem 1(a) in [8]; thus it is omitted. For the subsequent asymptotic properties, we will only point out their counterparts in [8], and verify the congruence between conditions. One can refer to [8] for details of the proof.

Based on the null distribution, an asymptotically α-level test can be defined as

$$\Pi_\alpha(\boldsymbol{\Omega}_{\boldsymbol{\eta}}) = \mathbf{1}[T_{\boldsymbol{\Omega}_{\boldsymbol{\eta}}} \ge 2\ln(p) - \ln\{\ln(p)\} + q_\alpha],$$

where q_α is the $(1-\alpha)$th quantile of the Gumbel distribution, viz. $q_\alpha = -\ln(\pi) - 2\ln\{\ln(1-\alpha)^{-1}\}$. The null hypothesis $\mathcal{H}_0$ is rejected if $\Pi_\alpha(\boldsymbol{\Omega}_{\boldsymbol{\eta}}) = 1$. The critical value of the α-level test is

$$2\ln(p) - \ln\{\ln(p)\} + q_\alpha. \tag{8}$$

3.2. Asymptotic properties of the test

Let $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_p) = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$. Define the set of k_p-sparse vectors by

$$\mathcal{S}(k_p) = \{\boldsymbol{\delta} : \mathbf{1}(\delta_1 \ne 0) + \cdots + \mathbf{1}(\delta_p \ne 0) = k_p\}.$$

We analyze the power of our new test under the alternative hypothesis

$$\mathcal{H}_1 : \boldsymbol{\delta} \in \mathcal{S}(k_p) \text{ with } k_p = p^r \text{ for some } r \in [0, 1), \text{ and the non-zero locations are randomly drawn from } \{1, \ldots, p\}.$$

The following theorem shows that, the asymptotically α-level test $\Pi_\alpha(\boldsymbol{\Omega}_{\boldsymbol{\eta}})$ distinguishes $\mathcal{H}_0$ and $\mathcal{H}_1$ with probability tending to 1, which parallels Theorem 2 in [8].

Theorem 3. *Suppose that Conditions 1, 3, 4 and 5 hold. Under the alternative hypothesis* $\mathcal{H}_1$ *with* $r < 1/4$*, if* $\max_i|\delta_i/(\sigma_{ii})^{1/2}| \ge \sqrt{2\beta\ln(p)/n}$ *with* $\beta \ge 1/\min_i(\sigma_{ii}v_{ii}) + \varepsilon$ *for some constant* $\varepsilon > 0$*, then, as* $n, p \to \infty$*,* $\Pr_{\mathcal{H}_1}\{\Pi_\alpha(\boldsymbol{\Omega}_{\boldsymbol{\eta}}) = 1\} \to 1$.

Remark 4. It follows by Lemma 1(a) that $\sigma_{ii} \asymp \sigma_{ii}^*$ and $w_{ii} \asymp v_{ii}$. Therefore, the lower bound on the maximum normalized signal $\max_i|\delta_i/(\sigma_{ii})^{1/2}|$ implies the similar bound on $\max_i|\delta_i/(\sigma_{ii}^*)^{1/2}|$.

The next theorem shows that, under certain assumption on $\mathbf{\Omega}_{\boldsymbol{\eta}}$, $\Pi_\alpha(\hat{\mathbf{\Omega}}_{\boldsymbol{\eta}})$ based on the CLIME estimator $\hat{\mathbf{\Omega}}_{\boldsymbol{\eta}}$ performs equivalently well as $\Pi_\alpha(\mathbf{\Omega}_{\boldsymbol{\eta}})$ asymptotically.

Theorem 4. *Let $\hat{\mathbf{\Omega}}_{\boldsymbol{\eta}}$ be the* CLIME *estimator of $\mathbf{\Omega}_{\boldsymbol{\eta}}$. Suppose that Conditions 1–5 hold, and*

$$\mathbf{\Omega} \in \mathcal{U}(M_p) = \left\{ \mathbf{\Omega} \succ \mathbf{0} : \|\mathbf{\Omega}\|_{L_1} \le M_p, \max_{j\in\{1,\ldots,p\}} \sum_{i=1}^{p} \mathbf{1}(|v_{ij}| > 0) \le M_p \right\},$$

with $M_p^2 = o[\sqrt{n}/\{\ln(p)\}^{3/2}]$. Then, as $n, p \to \infty$,

$$\Pr_{\mathcal{H}_0}\{\Pi_\alpha(\hat{\mathbf{\Omega}}_{\boldsymbol{\eta}}) = 1\} \le \alpha + o(1). \tag{9}$$

Furthermore, if $\max_i|\delta_i/(\sigma_{ii})^{1/2}| \ge \sqrt{2\beta\ln(p)/n}$ with $\beta \ge 1/\min_i(\sigma_{ii}v_{ii}) + \varepsilon$ for some constant $\varepsilon > 0$, then, as $n, p \to \infty$,

$$\Pr_{\mathcal{H}_1}\{\Pi_\alpha(\hat{\mathbf{\Omega}}_{\boldsymbol{\eta}}) = 1\} \to 1. \tag{10}$$

Remark 5. Condition 4 implies that $\sqrt{n}/\{\ln(p)\}^{3/2} \to \infty$ as $n, p \to \infty$, so the sparsity assumption on $\mathbf{\Omega}$, i.e.,

$$\max_{j\in\{1,\ldots,p\}} \sum_{i=1}^{p} \mathbf{1}(|v_{ij}| > 0) \le M_p,$$

is a mild condition in the sense that it would not significantly narrow the class of the precision matrices.

4. Cell-wise contamination case

We have so far focused on the uncontaminated scenario. In this section, we turn to the cell-wise contamination case. We start with the definition of the cell-wise contamination model [1]. The model is defined, for all $k \in \{1, \ldots, n\}$, as

$$\mathbf{X}_k = (\mathbf{I} - \mathbf{B}_k)\mathbf{U}_k + \mathbf{B}_k\mathbf{V}_k,$$

where $\mathbf{X}_k \in \mathbb{R}^p$ is the observed contaminated random vector. The unobserved random vectors $\mathbf{U}_k$, $\mathbf{V}_k$ and $\mathbf{B}_k$ are independent. The majority or 'core' of the data is $\mathbf{U}_k$, and $\mathbf{V}_k$ represents the outlier that deviates from the core behavior of the data. Furthermore, $\mathbf{B}_k = \mathrm{diag}(B_{k1}, \ldots, B_{kp})$ is a diagonal matrix, where $B_{k1}, \ldots, B_{kp}$ are independent Bernoulli random variables with $\Pr(B_{ki} = 1) = \varepsilon_i$, for all $i \in \{1 \ldots, p\}$.

Suppose that the contamination proportion $\varepsilon_i = \varepsilon > 0$ for all $i \in \{1, \ldots, p\}$. That is, the expected percentage of outlying cells in each dimension is ε. The probability that any particular observation would be contaminated is $1 - (1 - \varepsilon)^p$. As the dimension $p \to \infty$, this probability goes to 1. Therefore, with high probability, all observations could be contaminated for high-dimensional data. This critical problem of cell-wise contaminated data highlights the appeal of our robust test based on the original data rather than the cleaned data after simply excluding outlying observations.

Throughout this section, we will work under the cell-wise contaminated models, and assume that, for all $k \in \{1, \ldots, n\}$,

$$\mathbf{X}_k = (\mathbf{I} - \mathbf{B}_k)\mathbf{U}_k + \mathbf{B}_k\mathbf{V}_k \quad \text{and} \quad \mathbf{Y}_k = (\mathbf{I} - \mathbf{B}_k)\mathbf{U}_k' + \mathbf{B}_k\mathbf{V}_k',$$

where $\mathbf{U}_k \sim \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{\Sigma})$, and $\mathbf{U}_k' \sim \mathcal{N}(\boldsymbol{\mu}_2, \mathbf{\Sigma})$. There is no distributional assumption about the contamination components $\mathbf{V}_k$ and $\mathbf{V}_k'$. Our goal is to test $\mathcal{H}_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ from the uncontaminated normal component.

To deal with cell-wise contamination in the high-dimensional setting, more strategies for robustification are added to the procedure of estimating the precision matrix $\mathbf{\Omega}_{\boldsymbol{\eta}}$ for the trimmed mean. We use a trimmed mean $\bar{\mathbf{X}}_{\boldsymbol{\eta}}$ and a robust covariance matrix estimator $\hat{\mathbf{\Sigma}} = (\hat{\sigma}_{ij})_{p\times p}$ to estimate the mean, variance or covariance of each F_i and F_{ij} in (4) and (5). To summarize, we follow the following procedure to obtain the estimated covariance matrix $\hat{\mathbf{\Sigma}}_{\boldsymbol{\eta}}$:

(a) We obtain the trimmed mean $\bar{\mathbf{X}}_{\boldsymbol{\eta}}$ and the robust covariance matrix estimator $\hat{\mathbf{\Sigma}}$ for the original data $\mathbf{X}$.
(b) We plug $\bar{\mathbf{X}}_{\boldsymbol{\eta}}$ and $\hat{\mathbf{\Sigma}}$ into (4) and (5) to get the covariance matrix $\hat{\mathbf{\Sigma}}_{\boldsymbol{\eta}}$.

To get a robust estimate $\hat{\mathbf{\Sigma}}$, we adopt the approach in Öllerer and Croux [21]. The robust covariance matrix estimation is based on pairwise correlations. For all $i, j \in \{1, \ldots, p\}$, let $\hat{\sigma}_{ij}$ be the entry of the covariance matrix $\hat{\mathbf{\Sigma}}$, and write it as $\hat{\sigma}_{ij} = \hat{\sigma}_i\hat{\sigma}_j\hat{\rho}_{ij}$, where $\hat{\sigma}_i, \hat{\sigma}_j$ are suitable estimators of standard deviations of ith and jth variables, respectively, and $\hat{\rho}_{ij}$ is a suitable estimator of their correlation. For each standard deviation, let, for all $i \in \{1, \ldots, p\}$,

$$\hat{\sigma}_i = \{\Phi^{-1}(0.75)\}^{-1}\hat{d}_i,$$

where $\hat{d}_i$ is the median absolute deviation from the median (MAD), given, for all $i \in \{1, \ldots, p\}$, by

$$\hat{d}_i = \text{median}_{1\le k\le n}\left\{|X_{ki} - \text{median}_{1\le \ell\le n}(X_{\ell i})|\right\}, \tag{11}$$

and the factor $\{\Phi^{-1}(0.75)\}^{-1}$ is chosen to make the estimator consistent under normality.

For each r_{ij}, let, for all $i, j \in \{1, \ldots, p\}$,

$$\hat{\rho}_{ij} = \sin(\pi r_{ij}/2),$$

where r_{ij} is Kendall's τ, defined, for all $i, j \in \{1, \ldots, p\}$ with $i \neq j$, by

$$r_{ij} = \frac{2}{n(n-1)} \sum_{1\le k<\ell\le n} \text{sign}(X_{ki} - X_{\ell i})\text{sign}(X_{kj} - X_{\ell j}).$$

Here, sign(x) is the signum function defined as sign$(x) = -1$, 0 or 1 according as $x < 0$, $= 0$ or > 0, respectively.

For testing the equality of two mean vectors, we get the estimators of asymptotic covariance matrices for the trimmed mean vectors $\bar{\mathbf{X}}_{\boldsymbol{\eta}}$ and $\bar{\mathbf{Y}}_{\boldsymbol{\eta}}$. Let $\hat{\boldsymbol{\Sigma}}^X_{\boldsymbol{\eta}}$ and $\hat{\boldsymbol{\Sigma}}^Y_{\boldsymbol{\eta}}$ be these two estimators. Define

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\eta}} = \frac{1}{n_1 + n_2 - 2}\{(n_1 - 1)\hat{\boldsymbol{\Sigma}}^X_{\boldsymbol{\eta}} + (n_2 - 1)\hat{\boldsymbol{\Sigma}}^Y_{\boldsymbol{\eta}}\}.$$

Then, $\hat{\boldsymbol{\Omega}}_{\boldsymbol{\eta}}$ is obtained by plugging $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\eta}}$ into (7).

Remark 6. Compared to the data-driven procedure in Section 2.2, the only difference here is the use of the trimmed means and robust covariance matrix estimators instead of sample means and pooled sample covariance matrices to determine each F_i and F_{ij} in (4) and (5).

Remark 7. Theorem 1 in Loh and Tan [19] gives the statistical error rate of the robust covariance matrix estimator $\hat{\boldsymbol{\Sigma}}$ based on Kendall's τ correlations. Although the estimator $\hat{\boldsymbol{\Sigma}}$ is not equal to the sample covariance estimator $\mathbf{S}_n$ in the uncontaminated case, it converges nonetheless to the true covariance matrix $\boldsymbol{\Sigma}$ at the optimal rate. For cell-wise contaminated data, if the contamination proportion $\varepsilon = O\{\sqrt{\ln(p)/n}\}$, then the robust covariance estimator $\hat{\boldsymbol{\Sigma}}$ still enjoys the same error rate as the optimal covariance estimator in the uncontaminated setting, i.e., $\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_\infty = O_P\{\sqrt{\ln(p)/n}\}$. Furthermore, we have $\text{E}(X_{ki}) - \text{E}(Y_{ki}) = \delta_i + O\{\sqrt{\ln(p)/n}\}$. Therefore, under the same conditions in Theorem 4, we have that (9) and (10) hold as $n, p \to \infty$.

5. Trimming level

Another practical issue is how to determine the trimming level vector $\boldsymbol{\eta}$. We consider two situations, one with outliers on both sides, the other with outliers on one side.

If outliers appear in both tails, we can use the modified Z-score to detect them. The modified Z-score of an observation X_{ki} for the ith feature is defined, for all $i \in \{1, \ldots, p\}$, by

$$\Phi^{-1}(0.75)\{X_{ki} - \text{median}_{1\le \ell\le n}(X_{\ell i})\}/\hat{d}_i,$$

where $\hat{d}_i$ is the median absolute deviation from the median (MAD) defined in (11). Observations that have modified Z-scores with an absolute value greater than 3.5 are labeled as potential outliers [16]. Letting $\eta_{i,L}$ and $\eta_{i,R}$ be the proportions of detected outliers in the ith dimension on two sides, the ith trimming level is defined as $\eta_i = \max(\eta_{i,L}, \eta_{i,R})$.

If outliers appear only in one tail, we may adopt the approach proposed by Dhar and Chaudhuri [12]. Let F be a distribution with density f. In the model $F(x) = (1 - \beta)H(x) + \beta G(x)$, where H is symmetric and G is stochastically larger than H, an estimation of the contamination proportion β is given by

$$\hat{\beta} = \underset{\eta\in[b_1,b_2]}{\text{argmax}}\, V_n(x),$$

where $0 < b_1 < b_2 < 1/2$, and

$$V_n(\eta) = \frac{2}{1-2\eta}\left[\bar{x}_\eta - \frac{1}{2}\{\hat{F}_n^{-1}(\eta) + \hat{F}_n^{-1}(1-\eta)\} - \frac{1}{2}\left[\frac{1}{\hat{f}\{\hat{F}_n^{-1}(\eta)\}} - \frac{1}{\hat{f}\{\hat{F}_n^{-1}(1-\eta)\}}\right]\right]$$

is a natural estimate of the second derivative of the η-trimmed mean. Here, $\bar{x}_\eta$ is the sample η-trimmed mean, $\hat{F}_n$ is the empirical distribution function, and $\hat{f}$ is some suitable estimate of the density f. We use the kernel density estimate based on the standard Gaussian kernel provided in the R package `ks` [13].

Note that in our new test statistic, $\bar{\mathbf{X}}_{\boldsymbol{\eta}}$ and $\bar{\mathbf{Y}}_{\boldsymbol{\eta}}$ share the same trimming level vector $\boldsymbol{\eta}$. Suppose that the trimming level vector determined by the samples from $\mathbf{X}$ is $\hat{\boldsymbol{\eta}}^X = (\hat{\eta}^X_1, \ldots, \hat{\eta}^X_p)^\top$, and the trimming level vector determined by the samples from $\mathbf{Y}$ is $\hat{\boldsymbol{\eta}}^Y = (\hat{\eta}^Y_1, \ldots, \hat{\eta}^Y_p)^\top$. Then the final trimming level $\hat{\boldsymbol{\eta}} = (\hat{\eta}_1, \ldots, \hat{\eta}_p)^\top$ is given, for all $i \in \{1, \ldots, p\}$, by $\hat{\eta}_i = \max(\hat{\eta}^X_i, \hat{\eta}^Y_i)$.

Table 1
Size and power of CLX and RCLX: non-contaminated scenario.

	m	Signal	Model 1		Model 2		Model 3	
			CLX	RCLX	CLX	RCLX	CLX	RCLX
$p = 50$								
Size			0.037	0.025	0.041	0.027	0.061	0.047
Power	$0.05p$	fixed	0.669	0.599	0.462	0.407	0.478	0.435
	$\sqrt{p}$	fixed	0.978	0.964	0.886	0.845	0.873	0.837
	$0.05p$	varied	0.720	0.698	0.625	0.596	0.638	0.612
	$\sqrt{p}$	varied	0.986	0.983	0.964	0.949	0.969	0.964
$p = 100$								
Size			0.042	0.023	0.034	0.022	0.056	0.040
Power	$0.05p$	fixed	0.954	0.930	0.839	0.774	0.728	0.676
	$\sqrt{p}$	fixed	0.995	0.992	0.965	0.946	0.924	0.896
	$0.05p$	varied	0.971	0.967	0.922	0.908	0.907	0.896
	$\sqrt{p}$	varied	0.998	0.996	0.993	0.990	0.985	0.980

6. Simulation study

The numerical performance of our test was investigated, and the simulation results are summarized in this section. We compare it with the test in Cai et al. [8]. For convenience, we name the test therein CLX, and our test RCLX.

Without loss of generality, we will always set $\boldsymbol{\mu}_2 = \mathbf{0}$ in the simulation. Under the null hypothesis, we set $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}$, while under the alternative hypothesis, $\boldsymbol{\mu}_1 = (\mu_{1,1}, \ldots, \mu_{1,p})$ has m non-zero entries with support $S = \{\ell_1, \ldots, \ell_m : 1 \leq \ell_1 < \cdots < \ell_m \leq p\}$ uniformly and randomly drawn from $\{1, \ldots, p\}$.

Take $\mu_{1,k} = 0$ for $k \notin S^{\mathrm{C}}$. This setup is the same as in Cai et al. [8].

Suppose that $n_1 = n_2 = n$. Two settings of the magnitude of the nonzero entry $\ell_j \in S$ are considered, namely

(a) Fixed magnitude: Take $\mu_{1,\ell_j} = \pm\sqrt{2\ln(p)/n}$ with equal probability;
(b) Varied magnitude: Draw μ_{1,ℓ_j} uniformly from the interval $[-\sqrt{8\ln(p)/n}, \sqrt{8\ln(p)/n}]$.

Data are generated from the following three models having precision matrix $\boldsymbol{\Omega}$ with different structures:

> Model 1 (block diagonal $\boldsymbol{\Omega}$): $\boldsymbol{\Sigma} = (\sigma_{i,j})$ where $\sigma_{i,i} = 1$, and $\sigma_{i,j} = 0.8$ for any $i, j \in \{2k-1, 2k\}$ with $i \neq j$, $k \in \{1, \ldots, [p/2]\}$, and $\sigma_{i,j} = 0$ otherwise.
> Model 2 (bandable $\boldsymbol{\Sigma}$): $\sigma_{i,j} = 0.6^{|i-j|}$.
> Model 3 (banded $\boldsymbol{\Omega}$): $\boldsymbol{\Omega} = (v_{i,j})$ with $v_{i,i} = 2$ for $i \in \{1, \ldots, p\}$, $v_{i,i+1} = 0.8$ for $i \in \{1, \ldots, p-1\}$, $v_{i,i+2} = 0.4$ for $i \in \{1, \ldots, p-2\}$, $v_{i,i+3} = 0.4$ for $i \in \{1, \ldots, p-3\}$, $v_{i,i+4} = 0.2$ for $i \in \{1, \ldots, p-4\}$, $v_{i,j} = v_{j,i}$ and $v_{i,j} = 0$ otherwise.

6.1. Comparison with the CLX test under normal distributions

Under each model, uncontaminated data $\mathbf{X}_k$ and $\mathbf{Y}_\ell$ are generated from $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, respectively. To generate cell-wise contaminated data, we randomly select 10% of the entries in each dimension, then draw 5% of them from a normal $\mathcal{N}(5, 1)$ and the other 5% from $\mathcal{N}(-5, 1)$. The sample size n is always 100. The dimension p takes value $p = 50$ and 100. The size and the power at significance level 5% are then calculated from 1000 replicates.

For uncontaminated data, the CLX test used in the simulations is exactly the same as that in [8]. Since the RCLX test reduces to the CLX test when $\boldsymbol{\eta} = \mathbf{0}$, the two tests perform in exactly the same way in this case. Therefore, in order to investigate how trimming would affect the performance, we carry out the RCLX test based on 5%-trimmed means. Table 1 presents the results for the non-contaminated scenario.

For cell-wise contaminated data, we replace sample means with trimmed means in the CLX test statistic to reduce the deviation from the true mean. The contamination is on both sides, so the modified Z-scores are used to detect outliers and determine the trimming levels used in the RCLX test. Table 2 presents the results for the 10% contamination scenario.

As we can see from Tables 1 and 2, the estimated sizes are close to the nominal level 0.05 for our RCLX test in all settings. Table 1 shows that the RCLX test has slightly lower but comparable power compared to the CLX test for the no contamination scenario. For the cell-wise contaminated data, Table 2 shows that the CLX test breaks down with almost zero power in all settings; in contrast, the RCLX test maintains its power.

6.2. Comparison with the CLX test under non-normal distributions

For Model 3, we generate two independent random samples $\mathbf{X}_k$ and $\mathbf{Y}_k$ multivariate models $\mathbf{X}_k = \Gamma\mathbf{Z}_k^{(1)} + \boldsymbol{\mu}_1$ and $\mathbf{Y}_k = \Gamma\mathbf{Z}_k^{(2)} + \boldsymbol{\mu}_2$, with $\Gamma\Gamma^\top = \Sigma$, where the components of $\mathbf{Z}_k^{(i)}$ are iid standard $\mathcal{G}(5, 1)$ random variables. The results

Table 2
Size and power of CLX and RCLX: 10% contamination scenario.

	m	Signal	Model 1		Model 2		Model 3	
			CLX	RCLX	CLX	RCLX	CLX	RCLX
$p = 50$								
Size			0	0.032	0	0.027	0	0.048
Power	$0.05p$	fixed	0	0.190	0	0.162	0	0.257
	$\sqrt{p}$	fixed	0	0.492	0	0.411	0	0.567
	$0.05p$	varied	0.003	0.396	0.002	0.340	0.003	0.457
	$\sqrt{p}$	varied	0.006	0.817	0.008	0.769	0.007	0.848
$p = 100$								
Size			0	0.043	0	0.053	0	0.066
Power	$0.05p$	fixed	0	0.421	0	0.387	0	0.505
	$\sqrt{p}$	fixed	0	0.705	0	0.603	0	0.756
	$0.05p$	varied	0.008	0.741	0.004	0.689	0.003	0.778
	$\sqrt{p}$	varied	0.013	0.941	0.023	0.899	0.014	0.950

Table 3
Size and power of CLX and RCLX for Model 3: $\mathcal{G}(5, 1)$ scenario.

	Size	Power			
		$0.05p$ fixed	$\sqrt{p}$ fixed	$0.05p$ varied	$\sqrt{p}$ varied
$p = 50$					
CLX	0.044	0.085	0.210	0.128	0.375
RCLX	0.064	0.105	0.236	0.156	0.396
$p = 100$					
CLX	0.043	0.178	0.265	0.273	0.488
RCLX	0.066	0.224	0.312	0.333	0.522

are summarized in Table 3. As we can see from this table, under the heavy tailed distribution $\mathcal{G}(5, 1)$, our test is still slightly and uniformly more powerful than the CLX test in all settings.

6.3. Comparison with the CLX test on cleaned data

Under cell-wise contamination, although removing the entire observation results in a huge loss of information, it seems appropriate to remove those outlying cells in observations, and then non-robust tests can be carried out on the cleaned data after excluding outlying cells. In other words, the CLX test based on the cleaned data after excluding outlying cells in observations is an alternative to our robust test procedure on the contaminated data. In [20], an unbiased estimator of the covariance matrix was proposed when the dataset contains missing cells in observations. Suppose that $\delta \in (0, 1]$ is the proportion of observed cells. Given observations $\mathbf{X}_1, \ldots, \mathbf{X}_n$ with underlying covariance matrix $\mathbf{\Sigma}$, the unbiased estimator of $\mathbf{\Sigma}$ is given by

$$\tilde{\mathbf{\Sigma}} = (\delta^{-1} - \delta^{-2})\mathrm{diag}(\mathbf{S}_n^{(\delta)}) + \delta^{-2}\mathbf{S}_n^{(\delta)},$$

where $\mathbf{S}_n^{(\delta)} = (X_1 \otimes X_1 + \cdots + X_n \otimes X_n)/n$, and $\otimes$ denotes the outer product. Based on this empirical covariance matrix estimator, we can carry out the CLX test on the cleaned data after deleting outlying cells in observations. For convenience, we name such a test procedure DCLX. Table 4 presents the size and power of the DCLX test for the 10% contamination scenario. Although the power of the DCLX test is higher compared to the RCLX test, the size of the DCLX test is much higher than the nominal level 0.05. That is, the type-I error is not controlled, because the asymptotic null distribution of the max-norm typed test statistic on the cleaned data after deleting those outlying cells can be more complicated, which deserves further exploration.

7. Real data

Alzheimer's disease (AD) is a cognitive impairment disorder characterized by memory loss and a decrease in functional abilities above and beyond what is typical for a given age. Potential low-cost fluid biomarkers can help with early diagnosis of AD. For instance, protein levels of particular forms of the Aβ and Tau proteins, and the Apolipoprotein E genotype. These biomarkers could be obtained from plasma or cerebrospinal fluid (CSF) [18].

The *AlzheimerDisease* data in the R package '`AppliedPredictiveModeling`' [17] contains CSF measurements from 333 subjects, including some with mild cognitive impairment (class 'Impaired') as well as healthy individuals (class 'Control'). The dataset used here is a subset that contains 106 high-risk patients with the Apolipoprotein E genotype E3E4. Among

Table 4
Size and power of DCLX: 10% contamination scenario.

	m	Signal	Model 1	Model 2	Model 3
$p = 50$					
Size			0.100	0.083	0.095
Power	$0.05p$	fixed	0.359	0.318	0.393
	$\sqrt{p}$	fixed	0.730	0.680	0.779
	$0.05p$	varied	0.505	0.466	0.551
	$\sqrt{p}$	varied	0.902	0.867	0.924
$p = 100$					
Size			0.122	0.130	0.166
Power	$0.05p$	fixed	0.670	0.603	0.393
	$\sqrt{p}$	fixed	0.886	0.832	0.919
	$0.05p$	varied	0.834	0.816	0.879
	$\sqrt{p}$	varied	0.969	0.955	0.925

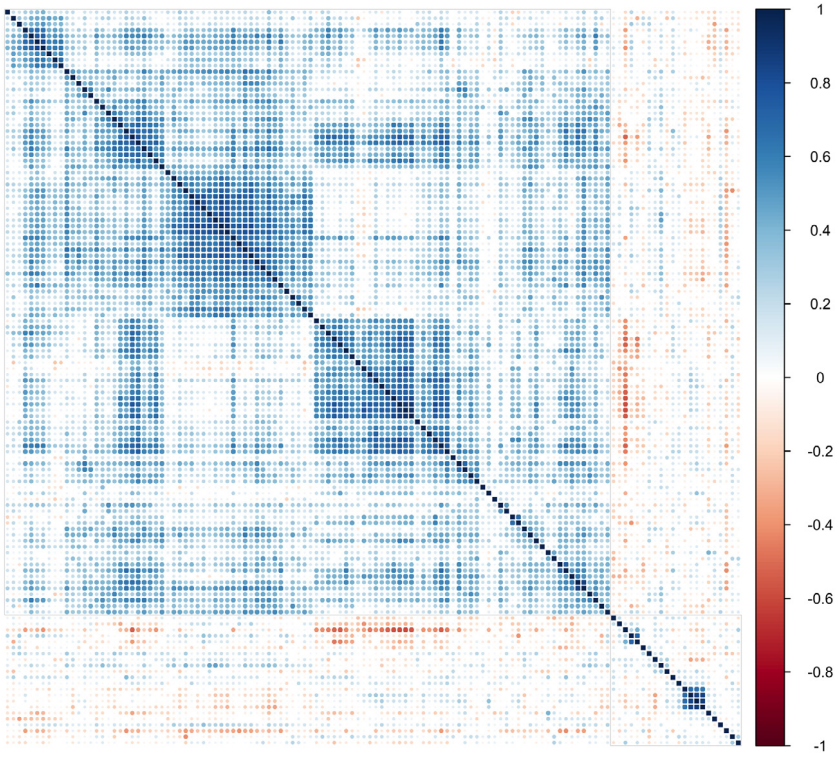

Fig. 1. Correlation matrix for the AD data. Each row and column represents a biomarker; they were sorted using clustering methods.

them, $n_1 = 41$ subjects are impaired, and $n_2 = 65$ subjects are healthy. Protein measurements of $p = 124$ exploratory biomarkers are collected on each subject, and standardized to the same scale. Fig. 1 shows the 124×124 correlation matrix of the exploratory biomarkers, which is quite sparse.

The goal of the experiment is to determine if subjects in the early states of impairment can be differentiated from cognitively healthy individuals via these biomarkers. One effective way is to compare the means of the biomarker measurements of the two groups. We want to investigate how the CLX test and our RCLX test behave on the contaminated data. To reveal the benchmark for the mean equality, we assume that the *AlzheimerDisease* data is a rather clean dataset, and apply the CLX test to the original data. The critical value calculated from (8), is 12.86 at the 5% significance level. The CLX test statistic is 22.16, so we should reject mean equality between the impaired group and the healthy group.

To generate cell-wise contaminated data, we randomly select 10% of the cells from each biomarker, and replace them by random values independently drawn from the normal distribution $\mathcal{N}(3, 1)$. Since the outliers are on one side, the trimming level η_i in each dimension $i \in \{1, \ldots, 124\}$ is obtained by solving the optimization problem in . Results are based on 1000 replicates where different sets of outliers are generated. Fig. 2 shows the distribution of test statistics with the 5% significance

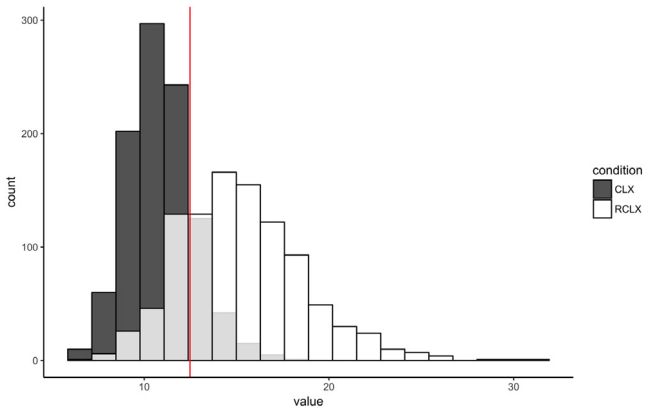

Fig. 2. Histogram of test statistics: CLX (dark gray), RCLX (white), overlapping (light gray) and critical value (vertical line).

level under contamination. The percentage of rejection among 1000 replicates based on our RCLX test is 74.3%, while the percentage of rejection based on the CLX test is only 11.8%.

8. Conclusion and discussion

In this paper, we proposed a robust two-sample test of high-dimensional means under dependence against sparse and strong alternatives based on trimmed means and robust precision matrix estimators, which robustifies the CLX test from Cai et al. [8]. Our test maintains comparable power compared to the CLX test for the uncontaminated scenario. Meanwhile, for cell-wise contaminated data, our RCLX test significantly outperforms the CLX test. Numerical studies show that while the CLX test is almost powerless under contamination, the RCLX test still achieves high power.

As mentioned in Remark 1, the result on the asymptotic distribution of a trimmed mean vector in Theorem 1 can be extended to non-normal distributions with sub-exponential tails. Therefore, the proposed test is applicable to data from heavy-tailed distributions. Numerical results show the adequacy of our method for such distributions.

The asymptotic properties in Section 3 rely on the assumption that $\{\ln(pn)\}^7/n \to 0$ as $n, p \to \infty$. It is a stronger assumption than that in Cai et al. [8], i.e., $\ln(p)/n \to 0$ as $n, p \to \infty$, and is merely resulted from the requirement for the Central Limit Theorem on hyperrectangles in high-dimensional settings [10], which to our knowledge, is the latest result available. This assumption might be weakened as high-dimensional probability theory develops.

Our proposed test could also be applied to datasets with missing values by treating missing data as outliers, if the missing data mechanism is completely at random.

Acknowledgments

The authors are grateful to the Editor-in-Chief, Christian Genest, to an Associate Editor and two referees for comments and suggestions that led to a much improved version of this paper. The work of Xiang Tang was supported by NSF, USA grant 1363250.

Appendix. Proof of main results

A.1. Lemmas

Lemma A.1 is Fact 9.9.59 in [4].

Lemma A.1. *Let* $\mathbf{A}$ *and* $\mathbf{B} \in \mathbb{R}^{n\times n}$*, assume that* $\mathbf{A}$ *and* $\mathbf{A}+\mathbf{B}$ *are nonsingular, and let* $\|\cdot\|$ *be a normalized submultiplicative norm on* $\mathbb{R}^{n\times n}$*. Then* $\|\mathbf{A}^{-1} - (\mathbf{A}+\mathbf{B})^{-1}\| \le \|\mathbf{A}^{-1}\| \times \|(\mathbf{A}+\mathbf{B})^{-1}\| \times \|\mathbf{B}\|$.

Lemma A.2, see DasGupta [11], shows that the univariate trimmed mean is asymptotically linear in the sense that it admits a Bahadur type linear representation.

Lemma A.2. *Let* $X_1, \ldots, X_n$ *be a random sample from a distribution* F *which is symmetric about* θ*. Let* $\eta \in (0, 1/2)$ *be fixed. Then the trimmed mean* $\bar{X}_\eta$ *admits the representation*

$$\bar{X}_\eta = \theta + \frac{1}{n}\sum_{i=1}^{n} \mathrm{IF}(X_i) + o_P(n^{-1/2}),$$

where for each $i \in \{1, \ldots, n\}$, $\mathrm{IF}(X_i)$ is the influence function of the trimmed mean given by

$$\mathrm{IF}(X_i) = \begin{cases} \{F^{-1}(\eta) - \theta\}/(1-2\eta) & \text{if } X_i < F^{-1}(\eta), \\ (X_i - \theta)/(1-2\eta) & \text{if } F^{-1}(\eta) \le X_i \le F^{-1}(1-\eta), \\ \{F^{-1}(1-\eta) - \theta\}/(1-2\eta) & \text{if } X_i > F^{-1}(1-\eta). \end{cases}$$

Lemma A.3. *Suppose that $\mathbf{X}_1, \ldots, \mathbf{X}_n$ form a random sample from a p-variate multivariate normal distribution $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ such that $K_0^{-1} \le \lambda_{\min}(\boldsymbol{\Sigma}) \le \lambda_{\max}(\boldsymbol{\Sigma}) \le K_0$ for some constant $K_0 > 0$. Let $\bar{\mathbf{X}}_{\boldsymbol{\eta}} = (\bar{X}_{\eta_1,1}, \ldots, \bar{X}_{\eta_p,p})^\top$ and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^\top$ be the trimmed mean vector and the true mean vector, respectively.*

(a) *If $0 < p < \infty$, then, as $n \to \infty$,*

$$\sqrt{n}\,(\bar{\mathbf{X}}_{\boldsymbol{\eta}} - \boldsymbol{\mu}) - \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \mathrm{IF}(\mathbf{X}_k) \xrightarrow{p} 0, \tag{A.1}$$

where for each $k \in \{1, \ldots, n\}$, $\mathrm{IF}(\mathbf{X}_k) = (\mathrm{IF}(X_{k1}), \ldots, \mathrm{IF}(X_{kp}))^\top$ with $\mathbf{X}_k = (X_{k1}, \ldots, X_{kp})^\top$.

(b) *Suppose that for any $i \in \{1, \ldots, p\}$, $\eta_i \in [b_1, b_2]$, where $0 \le b_1 < b_2 < 1/2$ and $\eta_i \to b_1$ as $p \to \infty$. If $\{\ln(pn)\}^7/n \to 0$ as $n, p \to \infty$, then (A.1) stills holds.*

Proof. It follows from Lemma A.2 that, for each $i \in \{1, \ldots, p\}$,

$$\sqrt{n}\,(\bar{X}_{\eta_i,i} - \mu_i) - \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \mathrm{IF}(X_{ki}) \xrightarrow{p} 0. \tag{A.2}$$

as $n \to \infty$. Furthermore, it is known that if $0 < p < \infty$, then as $n \to \infty$,

$$\sqrt{n}\,(\bar{\mathbf{X}}_{\boldsymbol{\eta}} - \boldsymbol{\mu}) - \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \mathrm{IF}(\mathbf{X}_k) \xrightarrow{p} 0.$$

Next, we shall show (A.1) still holds as $p \to \infty$. For all $i \in \{1, \ldots, p\}$, from (A.2), we have that, as $n \to \infty$,

$$\sqrt{n}\left[\frac{1}{n - 2\lfloor n\eta_i \rfloor} \sum_{k=\lfloor n\eta_i \rfloor + 1}^{n - \lfloor n\eta_i \rfloor} \frac{X_{(k)i} - \mu_i}{\sqrt{\sigma_{ii}}}\right] - \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \frac{\mathrm{IF}(X_{ki})}{\sqrt{\sigma_{ii}}} \xrightarrow{p} 0.$$

Set $\Phi^{-1}(\eta) = \inf_x\{x : \Phi(x) \ge \eta\}$ and $\Phi^{-1}(1-\eta) = \sup_x\{x : \Phi(x) < 1-\eta\}$. Note that

$$\mathrm{IF}(X_{ki})/\sqrt{\sigma_{ii}} = \begin{cases} \Phi^{-1}(\eta_i)/(1-2\eta_i) & \text{if } (X_{ki} - \mu_i)/\sqrt{\sigma_{ii}} < \Phi^{-1}(\eta_i), \\ (X_{ki} - \mu_i)/\{\sqrt{\sigma_{ii}}(1-2\eta_i)\} & \text{if } \Phi^{-1}(\eta_i) \le (X_{ki} - \mu_i)/\sqrt{\sigma_{ii}} < \Phi^{-1}(1-\eta_i), \\ \Phi^{-1}(1-\eta_i)/(1-2\eta_i) & \text{if } (X_{ki} - \mu_i)/\sqrt{\sigma_{ii}} \ge \Phi^{-1}(1-\eta_i). \end{cases}$$

Let $W_{ki} = (X_{ki} - \mu_i)/\sqrt{\sigma_{ii}} \sim \mathcal{N}(0, 1)$. Then we have, as $n \to \infty$,

$$\sqrt{n}\left\{\frac{1}{n - 2\lfloor n\eta_i \rfloor} \sum_{k=\lfloor n\eta_i \rfloor + 1}^{n - \lfloor n\eta_i \rfloor} W_{(k)i} - \frac{1}{n} \sum_{k=1}^{n} \mathrm{IF}(W_{ki})\right\} \xrightarrow{p} 0, \tag{A.3}$$

where

$$\mathrm{IF}(W_{ki}) = \begin{cases} \Phi^{-1}(\eta_i)/(1-2\eta_i) & \text{if } W_{ki} < \Phi^{-1}(\eta_i), \\ W_{ki}/(1-2\eta_i) & \text{if } \Phi^{-1}(\eta_i) \le W_{ki} < \Phi^{-1}(1-\eta_i), \\ \Phi^{-1}(1-\eta_i)/(1-2\eta_i) & \text{if } W_{ki} \ge \Phi^{-1}(1-\eta_i). \end{cases}$$

Therefore, (A.2) and (A.3) are equivalent. Now let

$$T_i = \sqrt{n}\,(\bar{X}_{\eta_i,i} - \mu_i) - \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \mathrm{IF}(X_{ki}), \quad Q_i = Q_n(\eta_i) = \sqrt{n}\left\{\frac{1}{n - 2\lfloor n\eta_i \rfloor} \sum_{k=\lfloor n\eta_i \rfloor + 1}^{n - \lfloor n\eta_i \rfloor} W_{(k)i} - \frac{1}{n} \sum_{k=1}^{n} \mathrm{IF}(W_{ki})\right\}.$$

Set $\mathbf{T} = (T_1, \ldots, T_p)^\top = \sqrt{n}\,(\bar{\mathbf{X}}_{\boldsymbol{\eta}} - \boldsymbol{\mu}) - \{\mathrm{IF}(\mathbf{X}_1) + \cdots + \mathrm{IF}(\mathbf{X}_n)\}/\sqrt{n}$. It follows from the definition of η-trimmed mean that the process $Q_n(\eta)$ for $\eta \in [b_1, b_2]$ lies in the space of real functions on $[b_1, b_2]$ that are right continuous, and have left-hand limits [6].

Since $\{\ln(pn)\}^7/n \to 0$ as $n, p \to \infty$, we must have $\{\ln(p)\}^7/n \to 0$ as $n, p \to \infty$. Therefore, there exist $M_1, N_1 > 0$, such that for all $p > M_1$, and for all $n \ge N_1$, $p/e^{n^{1/7}} < 1$. Also, since $\eta_i \to b_1$, it follows from the right continuity of the process $Q_n(\eta)$ that for all $\varepsilon > 0$, there exists $M_2 > M_1$, such that for all $i > M_2$, $|Q_n(\eta_i) - Q_n(b_1)| < \varepsilon/(2e^{N_1^{1/7}})$. Then, for all $n > N_1$,

and for all $p > M_2$,

$$\sum_{i=M_2+1}^{p} |Q_n(\eta_i) - Q_n(b_1)| < p\varepsilon/(2e^{N_1^{1/7}}) < \varepsilon/2.$$

Thus, as $n, p \to \infty$,

$$\Pr\left\{ \sum_{i=M_2+1}^{p} |Q_n(\eta_i) - Q_n(b_1)| > \varepsilon/2 \right\} \to 0. \tag{A.4}$$

Since for any i, $Q_n(\eta_i) \xrightarrow{p} 0$, as $n \to \infty$, we deduce that

$$\Pr\left\{ \sum_{i=1}^{M_2} |Q_n(\eta_i)| > \varepsilon/4 \right\} \to 0. \tag{A.5}$$

Also, there exists $N_2 > N_1$, such that for all $n > N_2$,

$$\Pr\{|Q_n(b_1)| > \varepsilon/(4e^{N_1^{1/7}})\} < \varepsilon/(4e^{N_1^{1/7}}).$$

Thus,

$$\Pr\left\{ \sum_{i=M_2+1}^{p} |Q_n(b_1)| > \varepsilon/4 \right\} \le p \times \Pr\{|Q_n(b_1)| > \varepsilon/(4e^{N_1^{1/7}})\} < p\varepsilon/(4e^{N_1^{1/7}}) < \varepsilon/4. \tag{A.6}$$

Let

$$\sigma_{\max} = \max(\sqrt{\sigma_{11}}, \ldots, \sqrt{\sigma_{pp}}) < \infty. \tag{A.7}$$

Combine (A.4)–(A.6). Given that we have

$$\sqrt{T_1^2 + \cdots + T_p^2} \le |T_1| + \cdots + |T_p|$$

and $|Q_i| \ge |T_i|/\sigma_{\max}$, one has

$$\begin{aligned}
\Pr\{\|\boldsymbol{T}\| > \varepsilon\sigma_{\max}\} &\le \Pr\left\{ \sum_{i=1}^{p} |T_i| > \varepsilon\sigma_{\max} \right\} \le \Pr\left\{ \sum_{i=1}^{p} |Q_i| > \varepsilon \right\} \\
&\le \Pr\left\{ \sum_{i=1}^{M_2} |Q_n(\eta_i)| > \varepsilon/2 \right\} + \Pr\left\{ \sum_{i=M_2+1}^{p} |Q_n(\eta_i) - Q_n(b_1)| > \varepsilon/4 \right\} \\
&\quad + \Pr\left\{ \sum_{i=M_2+1}^{p} |Q_n(b_1)| > \varepsilon/4 \right\} \to 0.
\end{aligned}$$

Thus, (A.1) still holds as $p \to \infty$. □

Let $\mathbf{U}_1, \ldots, \mathbf{U}_n$ be mutually independent random vectors in $\mathbb{R}^p$, where $p \ge 3$ may be larger, or even much larger, than n. Denote the ith coordinate of $\mathbf{U}_k$ by U_{ki}, so that $\mathbf{U}_k = (U_{k1}, \ldots, U_{kp})^\top$. We assume that each $\mathbf{U}_k$ is centered, i.e., $\mathrm{E}(U_{ki}) = 0$, and $\mathrm{E}(U_{ki}^2) < \infty$ for all $k \in \{1, \ldots, n\}$ and $i \in \{1, \ldots, p\}$. Let $\mathbf{V}_1, \ldots, \mathbf{V}_n$ be independent centered Gaussian random vectors in $\mathbb{R}^p$ such that each $\mathbf{V}_k$ has the same covariance matrix as $\mathbf{U}_k$, i.e., $\mathbf{V}_k \sim \mathcal{N}[0, \mathrm{E}(\mathbf{U}_k\mathbf{U}_k^\top)]$. Define the normalized sums

$$S_n^U = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \mathbf{U}_k \quad \text{and} \quad S_n^V = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \mathbf{V}_k.$$

Let $\mathcal{A}^{\mathrm{re}}$ be the class of all hyperrectangles in $\mathbb{R}^p$, which consists of all sets A of the form

$$A = \left\{ w \in \mathbb{R}^p : a_i \le w_i \le b_i \text{ for all } i \in \{1, \ldots, p\} \right\}$$

for some $-\infty \le a_i \le b_i \le \infty$ for all $i \in \{1, \ldots, p\}$. One's interest is to bound

$$\rho_n(\mathcal{A}) = \sup_{A \in \mathcal{A}^{\mathrm{re}}} |\Pr(S_n^U \in A) - \Pr(S_n^V \in A)|.$$

The following lemma is one of the main results from Chernozhukov et al. [10].

Lemma A.4. *Let $b > 0$ be a constant, and $B_n \geq 1$ be a sequence of constants, possibly growing to infinity as $n \to \infty$. Assume that the following conditions are satisfied:*

(M.1) $\sum_{k=1}^{n} \mathrm{E}(U_{ki}^2)/n \geq b$ *for all* $i \in \{1, \ldots, p\}$,
(M.2) $\sum_{k=1}^{n} \mathrm{E}(|U_{ki}|^{2+\gamma})/n \leq B_n^{\gamma}$ *for all* $i \in \{1, \ldots, p\}$ *and* $\gamma \in \{1, 2\}$.
(M.3) $\mathrm{E}\{\exp(|U_{ki}|/B_n)\} \leq 2$ *for all* $k \in \{1, \ldots, n\}$ *and* $i \in \{1, \ldots, p\}$.

Then we have $\rho_n(\mathcal{A}) \leq C[B_n^2\{\ln(pn)\}^7/n]^{1/6}$, *where the constant* C *depends only on* b.

The following lemma shows the relationship between the diagonal entry of $\boldsymbol{\Sigma}$ and the diagonal entry of $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$.

Lemma A.5. *There exist some constants* $C_1, C_2 > 0$ *such that* $C_1 \leq \sigma_{ii}^*/\sigma_{ii} \leq C_2$.

Proof. Obviously, for any $i \in \{1, \ldots, p\}$, we have $\sigma_{ii}^*/\sigma_{ii} \leq 1/(1-\eta_i)^2 \leq 1/(1-b_2)^2$. In fact by Theorem 4.1 in [5], the upper bound could be tighter, viz. $\sigma_{ii}^*/\sigma_{ii} \leq (1+4\eta_i) \leq (1+4b_2) < 3 = C_2$. For the lower bound, note that

$$\sigma_{ii}^* = \frac{1}{(1-2\eta_i)^2}\left[\sigma_{ii} + 2\eta_i\{F_i^{-1}(\eta_i) - \mu_i\}^2 - 2\int_{-\infty}^{F_i^{-1}(\eta_i)} (x-\mu_i)^2 dF_i(x)\right].$$

Let $h(\eta_i) = 2\eta_i\{F_i^{-1}(\eta_i) - \mu_i\}^2 - 2\int_{-\infty}^{F_i^{-1}(\eta_i)}(x-\mu_i)^2 dF_i(x)$. Then, $h(\eta_i) \leq 0$ and $h'(\eta_i) = 4\eta_i\{F_i^{-1}(\eta_i) - \mu_i\}/f\{F_i^{-1}(\eta_i) - \mu_i\}$. That is, $h(\eta_i)$ is strictly decreasing on $[b_1, b_2]$, and hence for any i, $\sigma_{ii}^* \geq \sigma_{ii} + h_i(b_2)$, where

$$h_i(b_2) = 2\sigma_{ii}\left[b_2\{\Phi^{-1}(b_2)\}^2 - \int_{-\infty}^{\Phi^{-1}(b_2)} x^2 d\Phi(x)\right] > h_i(0.5) = -\sigma_{ii}.$$

Therefore, $\sigma_{ii}^*/\sigma_{ii}$ is bounded below by some constant $C_1 > 0$ which only depends on b_2. □

A.2. Proof of Theorem 1

We first show that the three conditions in Lemma A.4 are satisfied.

(M.1): For all $i \in \{1, \ldots, p\}$,

$$\frac{1}{n}\sum_{k=1}^{n} \mathrm{E}\{|\mathrm{IF}(X_{ki})|^2\} = \mathrm{E}\{|\mathrm{IF}(X_{1i})|^2\} = \sigma_{ii}^*,$$

where σ_{ii}^* is defined in (4). If $K_0^{-1} \leq \lambda_{\min}(\boldsymbol{\Sigma}) \leq \lambda_{\max}(\boldsymbol{\Sigma}) \leq K_0$ for some constant $K_0 > 0$, by Lemma A.5, σ_{ii}^* must be uniformly bounded below by some positive constant. Let $b = \min_{1\leq i\leq p}\sigma_{ii}^* > 0$. Then, for all $i \in \{1, \ldots, p\}$,

$$\frac{1}{n}\sum_{k=1}^{n} \mathrm{E}\{|\mathrm{IF}(X_{ki})|^2\} \geq b.$$

(M.2): For all $i \in \{1, \ldots, p\}$ and $\gamma \in \{1, 2\}$,

$$\frac{1}{n}\sum_{k=1}^{n} \mathrm{E}\{|\mathrm{IF}(X_{ki})|^{2+\gamma}\} = \mathrm{E}\{|\mathrm{IF}(X_{1i})|^{2+\gamma}\}.$$

Note that

$$(1-2\eta_i)^{-1}\{F_i^{-1}(\eta_i) - \mu_i\} \leq \mathrm{IF}(X_{1i}) \leq (1-2\eta_i)^{-1}\{F_i^{-1}(1-\eta_i) - \mu_i\}.$$

Since $\eta_i \in [b_1, b_2]$, and $F_i(x)$ is a symmetric distribution about $x = \mu_i$, we have $|\mathrm{IF}(X_{1i})| \leq (1-2b_2)^{-1}|F_i^{-1}(b_1) - \mu_i|$. If $b_1 = 0$, then $(1-2b_2)^{-1}|F_i^{-1}(b_1) - \mu_i| = \infty$, and hence so far, we cannot find a finite constant to uniformly bound $|\mathrm{IF}(X_{1i})|$ for any i. Therefore, the following two cases are considered.

Case 1: $b_1 > 0$. Let $C_1(i) = 2(1-2b_2)^{-1}|F_i^{-1}(b_1) - \mu_i|$ and $C_2(i) = (1-2b_2)^{-(2+\gamma)/\gamma}|F_i^{-1}(b_1) - \mu_i|^{(2+\gamma)/\gamma}$. Let $B_n = \max\{C_1(1), \ldots, C_1(p), C_2(1), \ldots, C_2(p)\}$. Then, for all $i \in \{1, \ldots, p\}$,

$$\frac{1}{n}\sum_{k=1}^{n} \mathrm{E}\{|\mathrm{IF}(X_{ki})|^{2+\gamma}\} \leq B_n^{\gamma}.$$

Note that for any $i \in \{1, \ldots, p\}$, $\sigma_{ii} \leq \lambda_{\max}(\boldsymbol{\Sigma}) \leq K_0$, so $B_n < \infty$ and is independent of n.

Case 2: $b_1 = 0$. Since $\mathrm{IF}(W_{ki})$ is right continuous and $\eta_i \to 0$, then, for all $\varepsilon > 0$, there exist $\delta, N > 0$ such that, for all $i > N$, we have $0 \le \eta_i < \delta$ and $|\mathrm{IF}(W_{ki}) - W_{ki}| < \varepsilon$. Since $W_{ki} \sim \mathcal{N}(0, 1)$,

$$\begin{aligned}\mathrm{E}\{|\mathrm{IF}(W_{ki})|^3\} &= \mathrm{E}\{|\mathrm{IF}(W_{ki}) - W_{ki} + W_{ki}|^3\} \\ &\le \varepsilon^3 + 3\varepsilon\mathrm{E}(|W_{ki}|^2) + 3\varepsilon^2\mathrm{E}(|W_{ki}|) + \mathrm{E}(|W_{ki}|^3) = \varepsilon^3 + 3\varepsilon + 3\varepsilon^2\sqrt{2/\pi} + 2\sqrt{2/\pi},\end{aligned} \tag{A.8}$$

and hence $\mathrm{E}\{|\mathrm{IF}(W_{ki})|^3\}$ is bounded. Similarly, $\mathrm{E}\{|\mathrm{IF}(W_{ki})|^4\}$ is bounded. Thus, there exists $M > 0$ such that

$$\mathrm{E}\{|\mathrm{IF}(W_{ki})|^{2+\gamma}\} < M$$

for $\gamma \in \{1, 2\}$. Then, for any $i > N$, $\mathrm{E}\{|\mathrm{IF}(X_{ki})|^{2+\gamma}\} < M\sigma_{\max}^{2+\gamma}$, where $\sigma_{\max}$ is defined in (A.7). The behavior of the sequence η_i does not change when we rearrange its terms, so, without loss of generality, we can assume that for any $i \le N$, we have $\eta_i \ge \delta$ and hence $|F_i^{-1}(\eta_i) - \mu_i| \le |F_i^{-1}(\delta) - \mu_i|$.

Let

$$C_1'(i) = 2(1 - 2b_2)^{-1}|F_i^{-1}(\delta) - \mu_i| < \infty, \quad C_2'(i) = (1 - 2b_2)^{-(2+\gamma)/\gamma}|F_i^{-1}(\delta) - \mu_i|^{(2+\gamma)/\gamma} < \infty,$$

and

$$B_n = \max\{C_1'(1), C_1'(2), \ldots, C_1'(N), C_2'(1), C_2'(2), \ldots, C_2'(N), M^{1/\gamma}\sigma_{\max}^{(2+\gamma)/\gamma}\}.$$

Then, for all $i \in \{1, \ldots, p\}$,

$$\frac{1}{n}\sum_{k=1}^{n}\mathrm{E}\{|\mathrm{IF}(X_{ki})|^{2+\gamma}\} \le B_n^{\gamma}.$$

Note that $B_n < \infty$ and is independent of n.

(M.3): Note that in Case 1, for all $k \in \{1, \ldots, n\}$ and $i \in \{1, \ldots, p\}$, $2|\mathrm{IF}(X_{ki})| \le C_1(i) \le B_n$. Hence,

$$\mathrm{E}\{\exp(|\mathrm{IF}(X_{ki})|/B_n)\} \le \exp(1/2) < 2.$$

In Case 2, for all $k \in \{1, \ldots, n\}$ and $i \in \{1, \ldots, N\}$, $2|\mathrm{IF}(X_{ki})| \le C_1'(i) \le B_n$, so

$$\mathrm{E}\{\exp(|\mathrm{IF}(X_{ki})|/B_n)\} < 2.$$

For all $k \in \{1, \ldots, n\}$ and $i > N$, $B_n \ge M^{1/\gamma}\sigma_{\max}^{(2+\gamma)/\gamma}$. Then, choose sufficiently large M and sufficiently small ε, such that $M^{1/\gamma}\sigma_{\max}^{2/\gamma} \ge 1$ and $\exp(\varepsilon) \le 2/3$, then

$$\begin{aligned}\mathrm{E}\{\exp(|\mathrm{IF}(X_{ki})|/B_n)\} &\le \mathrm{E}\left\{\exp\left(\frac{|\mathrm{IF}(X_{ki})|}{M^{1/\gamma}\sigma_{\max}^{(2+\gamma)/\gamma}}\right)\right\} \\ &\le \mathrm{E}\left\{\exp\left(\frac{|\mathrm{IF}(W_{ki}) - W_{ki} + W_{ki}|}{M^{1/\gamma}\sigma_{\max}^{2/\gamma}}\right)\right\} \\ &\le \mathrm{E}\left\{\exp\left(\frac{|\mathrm{IF}(W_{ki}) - W_{ki}|}{M^{1/\gamma}\sigma_{\max}^{2/\gamma}}\right) \times \exp\left(\frac{|W_{ki}|}{M^{1/\gamma}\sigma_{\max}^{2/\gamma}}\right)\right\} \\ &\le \exp(\varepsilon) \times \mathrm{E}\{\exp(|W_{ki}|)\} \le 2.\end{aligned} \tag{A.9}$$

The last inequality follows from the fact that $|W_{ki}|$ is a standard half-normal distribution, and $\mathrm{E}\{\exp(|W_{ki}|)\} \approx 2.77 < 3$.

Let $\mathcal{A}^{\mathrm{re}}$ be the class of all hyperrectangles in $\mathbb{R}^p$, i.e., it consists of all sets A of the form

$$A = \left\{w \in \mathbb{R}^p : a_i \le w_i \le b_i \text{ for all } i \in \{1, \ldots, p\}\right\}.$$

Here, $a_i = \{F_i^{-1}(\eta_i) - \mu_i\}/(1 - 2\eta_i)$ and $b_i = \{F_i^{-1}(1 - \eta_i) - \mu_i\}/(1 - 2\eta_i)$.

Let $\mathbf{V}_1, \ldots, \mathbf{V}_n$ be iid observations drawn from $\mathcal{N}_p\left[0, \mathrm{E}\{\mathrm{IF}(\mathbf{X})\}\{\mathrm{IF}(\mathbf{X})\}^{\top}\right]$. Since all three conditions in Lemma A.4 are satisfied, we have

$$\sup_{A \in \mathcal{A}^{\mathrm{re}}}|\Pr(S_n^{\mathrm{IF}(\mathbf{X})} \in A) - \Pr(S_n^{V} \in A)| \le C[B_n^2\{\ln(pn)\}^7/n]^{1/6}.$$

That is, as $n, p \to \infty$, if $\{\ln(pn)\}^7/n \to 0$, then

$$\frac{1}{\sqrt{n}}\sum_{k=1}^{n}\mathrm{IF}(\mathbf{X}_k) \rightsquigarrow \mathcal{N}[0, \mathrm{E}_{\mathbf{X}}\{\mathrm{IF}(\mathbf{X})\}\{\mathrm{IF}(\mathbf{X})\}^{\top}].$$

Together with the result from Lemma A.3, by Slutsky's Theorem, $\sqrt{n}(\bar{\boldsymbol{X}}_{\boldsymbol{\eta}} - \boldsymbol{\mu}) \rightsquigarrow \mathcal{N}[0, \mathrm{E}_{\mathbf{X}}\{\mathrm{IF}(\mathbf{X})\}\{\mathrm{IF}(\mathbf{X})\}^{\top}]$. There remains for us to compute the asymptotic covariance matrix $\mathrm{E}_{\mathbf{X}}\{\mathrm{IF}(\mathbf{X})\}\{\mathrm{IF}(\mathbf{X})\}^{\top}$. □

A.3. Proof of Lemma 1

For convenience, we enumerate the four conditions as follows.

(a) $K_0^{-1} \le \lambda_{\min}(\mathbf{\Sigma}) \le \lambda_{\max}(\mathbf{\Sigma}) \le K_0$ for some constant $K_0 > 0$.
(b) $C_0^{-1} \le \lambda_{\min}(\mathbf{\Sigma}_{\boldsymbol{\eta}}) \le \lambda_{\max}(\mathbf{\Sigma}_{\boldsymbol{\eta}}) \le C_0$ for some constant $C_0 > 0$ as $n, p \to \infty$.
(c) $\|\mathbf{R}\|_\infty \le r_1$ for some constant $r_1 \in (0, 1)$.
(d) $\|\mathbf{R}_{\boldsymbol{\eta}}\|_\infty \le r_2$ for some constant $r_2 \in (0, 1)$ as $n, p \to \infty$.

First, we want to show that (a) implies (b). It suffices to show that there exists some constant $K_1 > 0$ such that for any $\lambda < K_1^{-1}$ or $\lambda > K_1$, $\lambda\mathbf{I} - \mathbf{\Sigma}_{\boldsymbol{\eta}}$ is invertible. Let $\lambda \in \mathbb{R}$ such that $\lambda > K_0^{-1} - \varepsilon_0$ or $\lambda < K_0 + \varepsilon_0$ given any $0 < \varepsilon_0 < K_0^{-1}$. Then (a) implies that λ is not the eigenvalue of $\mathbf{\Sigma}$. Equivalently, $\lambda\mathbf{I} - \mathbf{\Sigma}$ is invertible. Observe that

$$\lambda\mathbf{I} - \mathbf{\Sigma}_{\boldsymbol{\eta}} = \lambda\mathbf{I} - \mathbf{\Sigma} + \mathbf{\Sigma} - \mathbf{\Sigma}_{\boldsymbol{\eta}} = (\lambda\mathbf{I} - \mathbf{\Sigma})\{\mathbf{I} + (\lambda\mathbf{I} - \mathbf{\Sigma})^{-1}(\mathbf{\Sigma} - \mathbf{\Sigma}_{\boldsymbol{\eta}})\}.$$

Let $\|\cdot\|_{\text{op}}$ denote the operator norm. Then

$$\|(\lambda\mathbf{I} - \mathbf{\Sigma})^{-1}(\mathbf{\Sigma} - \mathbf{\Sigma}_{\boldsymbol{\eta}})\|_{\text{op}} < 1 \quad \Rightarrow \quad \mathbf{I} + (\lambda\mathbf{I} - \mathbf{\Sigma})^{-1}(\mathbf{\Sigma} - \mathbf{\Sigma}_{\boldsymbol{\eta}}) \text{ is invertible} \quad \Rightarrow \quad \lambda\mathbf{I} - \mathbf{\Sigma}_{\boldsymbol{\eta}} \text{ is invertible.}$$

Since

$$\|(\lambda\mathbf{I} - \mathbf{\Sigma})^{-1}\|_{\text{op}} = \sup_i |\lambda - \lambda_i(\mathbf{\Sigma})|^{-1} < 1/\varepsilon_0,$$

it suffices to show that $\|\mathbf{\Sigma} - \mathbf{\Sigma}_{\boldsymbol{\eta}}\|_{\text{op}} < \varepsilon_0$.

Without loss of generality, assume that the data are centered at $\mathbf{0}$, i.e., $\boldsymbol{\mu} = \mathbf{0}$. Let $\tilde{\mathbf{X}} = \text{IF}(\mathbf{X}) = (\text{IF}(X_1), \ldots, \text{IF}(X_p))^\top$. Then $\mathbf{\Sigma} = \text{E}(\mathbf{X}\mathbf{X}^\top)$ and $\mathbf{\Sigma}_{\boldsymbol{\eta}} = \text{E}(\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top)$. Note that

$$\tilde{X}_i = \min\{|X_i|, \sqrt{\sigma_{ii}}\,|\Phi^{-1}(\eta_i)|\}X_i/\{|X_i|(1 - 2\eta_i)\}.$$

Let $\mathbf{v} \in \mathcal{S}^{p-1}$, and $\mathcal{S}^{p-1}$ is $(p-1)$-dimensional Euclidean unit sphere. Let $\tilde{\boldsymbol{\eta}} = (1/(1-2\eta_1), \ldots, 1/(1-2\eta_p))^\top$, and $\odot$ denote the element-wise product. Then,

$$\begin{aligned}
\mathbf{v}^T(\mathbf{\Sigma} - \mathbf{\Sigma}_{\boldsymbol{\eta}})\mathbf{v} = \text{E}\{(\mathbf{v}^\top\mathbf{X})^2 - (\mathbf{v}^\top\tilde{\mathbf{X}})^2\} &\le \text{E}\left\{(\mathbf{v}^\top(\boldsymbol{\eta}\odot\mathbf{X}))^2 - (\mathbf{v}^T\tilde{\mathbf{X}})^2\right\} \\
&= \text{E}\left[\left\{(\mathbf{v}^\top(\boldsymbol{\eta}\odot\mathbf{X}))^2 - (\mathbf{v}^\top\tilde{\mathbf{X}})^2\right\}\mathbf{1}\{\forall_i\ |X_i| \le \sqrt{\sigma_{ii}}\,|\Phi^{-1}(\eta_i)|\}\right] \\
&\quad + \text{E}\left[\left\{(\mathbf{v}^\top(\boldsymbol{\eta}\odot\mathbf{X}))^2 - (\mathbf{v}^\top\tilde{\mathbf{X}})^2\right\}\mathbf{1}\{\exists_i\ |X_i| > \sqrt{\sigma_{ii}}\,|\Phi^{-1}(\eta_i)|\}\right] \\
&= \text{E}\left[\left\{(\mathbf{v}^\top(\boldsymbol{\eta}\odot\mathbf{X}))^2 - (\mathbf{v}^\top\tilde{\mathbf{X}})^2\right\}\mathbf{1}\{\exists_i\ |X_i| > \sqrt{\sigma_{ii}}\,|\Phi^{-1}(\eta_i)|\}\right] \\
&\le \text{E}\left[\{\mathbf{v}^\top(\boldsymbol{\eta}\odot\mathbf{X})\}^2\mathbf{1}\{\exists_i\ |X_i| > \sqrt{\sigma_{ii}}\,|\Phi^{-1}(\eta_i)|\}\right] \\
&\le \sqrt{\frac{1}{(1-2b_2)^4}\text{E}\{(\mathbf{v}^\top\mathbf{X})^4\}\,\text{Pr}\left\{\exists_i\ |X_i| > \sqrt{\sigma_{ii}}\,|\Phi^{-1}(\eta_i)|\right\}}
\end{aligned}$$

Since X_i has bounded fourth moment for any i, one has $\text{E}\{(\mathbf{v}^\top\mathbf{X})^4\} \le C'$ for some constant $C' > 0$. Moreover,

$$\text{Pr}\left\{\exists_i\ |X_i| > \sqrt{\sigma_{ii}}\,|\Phi^{-1}(\eta_i)|\right\} \le \sum_{i=1}^p \text{Pr}\left\{|X_i| > \sqrt{\sigma_{ii}}\,|\Phi^{-1}(\eta_i)|\right\} \le \sum_{i=1}^p \frac{\text{E}(X_i^2)}{\sigma_{ii}\,|\Phi^{-1}(\eta_i)|^2} = \sum_{i=1}^p \frac{1}{|\Phi^{-1}(\eta_i)|^2}.$$

Since the sequence η_i converges to 0, there exists $M > 0$, we can have for any $i > M$, $|\Phi^{-1}(\eta_i)| > e^{n^{1/7}}$, and hence when $p > e^M$, for any $i > \ln p > M$,

$$\sum_{i=1}^p \frac{1}{|\Phi^{-1}(\eta_i)|^2} = \sum_{i\le\ln p} \frac{1}{|\Phi^{-1}(\eta_i)|^2} + \sum_{i>\ln p} \frac{1}{|\Phi^{-1}(\eta_i)|^2} \le \frac{\ln p}{|\Phi^{-1}(b_2)|^2} + \frac{p}{e^{n^{1/7}}}.$$

Therefore,

$$\mathbf{v}^T(\mathbf{\Sigma} - \mathbf{\Sigma}_{\boldsymbol{\eta}})\mathbf{v} \le \sqrt{\frac{C'}{(1-2b_2)^4}\left\{\frac{\ln p}{|\Phi^{-1}(b_2)|^2} + \frac{p}{e^{n^{1/7}}}\right\}}.$$

Since $\{\ln(pn)\}^7/n \to 0$ as $n, p \to \infty$, we have $p/e^{n^{1/7}} = o(1)$. Furthermore, if $1/|\Phi^{-1}(b_2)| = o(1/\sqrt{\ln p})$, then, as $n, p \to \infty$, $\mathbf{v}^\top(\mathbf{\Sigma} - \mathbf{\Sigma}_{\boldsymbol{\eta}})\mathbf{v} = o(1)$. That is, $\|\mathbf{\Sigma} - \mathbf{\Sigma}_{\boldsymbol{\eta}}\|_{\text{op}} < \varepsilon_0$.

Next, we want to show that (a) and (c) imply (d). Let $\mathbf{A} = \mathbf{\Sigma}$ and $\mathbf{B} = \mathbf{\Sigma}_{\boldsymbol{\eta}} - \mathbf{\Sigma}$. Then by Lemma A.1,

$$\|\mathbf{\Omega}_{\boldsymbol{\eta}} - \mathbf{\Omega}\|_{\text{op}} \le \|\mathbf{\Omega}_{\boldsymbol{\eta}}\|_{\text{op}}\|\mathbf{\Omega}\|_{\text{op}}\|\mathbf{\Sigma}_{\boldsymbol{\eta}} - \mathbf{\Sigma}\|_{\text{op}} \le \lambda_{\min}^{-1}(\mathbf{\Sigma}_{\boldsymbol{\eta}})\lambda_{\min}^{-1}(\mathbf{\Sigma})\|\mathbf{\Sigma}_{\boldsymbol{\eta}} - \mathbf{\Sigma}\|_{\text{op}}, \tag{A.10}$$

and similarly,

$$\begin{aligned}\|\mathbf{D}_\eta^{-1}-\mathbf{D}^{-1}\|_{\mathrm{op}} &\le \|\mathbf{D}^{-1}\|_{\mathrm{op}}\|\mathbf{D}_\eta^{-1}\|_{\mathrm{op}}\|\mathbf{D}_\eta-\mathbf{D}\|_{\mathrm{op}} = \|\mathbf{D}^{-1}\|_{\mathrm{op}}\|\mathbf{D}_\eta^{-1}\|_{\mathrm{op}}\|\mathbf{D}_\eta-\mathbf{D}\|_\infty \\ &\le \|\mathbf{D}^{-1}\|_{\mathrm{op}}\|\mathbf{D}_\eta^{-1}\|_{\mathrm{op}}\|\boldsymbol{\Omega}_\eta-\boldsymbol{\Omega}\|_{\mathrm{op}} \le \|\mathbf{D}^{-1}\|_{\mathrm{op}}\|\mathbf{D}_\eta^{-1}\|_{\mathrm{op}}\lambda_{\min}^{-1}(\boldsymbol{\Sigma}_\eta)\lambda_{\min}^{-1}(\boldsymbol{\Sigma})\|\boldsymbol{\Sigma}_\eta-\boldsymbol{\Sigma}\|_{\mathrm{op}}. \end{aligned} \tag{A.11}$$

Note that

$$\begin{aligned}\|\mathbf{R}_\eta-\mathbf{R}\|_\infty &\le \|\mathbf{R}_\eta-\mathbf{R}\|_{\mathrm{op}} \\ &= \|(\mathbf{D}_\eta^{-1}\boldsymbol{\Omega}_\eta\mathbf{D}_\eta^{-1}-\mathbf{D}^{-1}\boldsymbol{\Omega}_\eta\mathbf{D}_\eta^{-1})+(\mathbf{D}^{-1}\boldsymbol{\Omega}_\eta\mathbf{D}_\eta^{-1}-\mathbf{D}^{-1}\boldsymbol{\Omega}\mathbf{D}_\eta^{-1})+(\mathbf{D}^{-1}\boldsymbol{\Omega}\mathbf{D}_\eta^{-1}-\mathbf{D}^{-1}\boldsymbol{\Omega}\mathbf{D}^{-1})\|_{\mathrm{op}} \\ &\le \|\mathbf{D}_\eta^{-1}-\mathbf{D}^{-1}\|_{\mathrm{op}}\|\boldsymbol{\Omega}_\eta\|_{\mathrm{op}}\|\mathbf{D}_\eta^{-1}\|_{\mathrm{op}}+\|\mathbf{D}^{-1}\|_{\mathrm{op}}\|\boldsymbol{\Omega}_\eta-\boldsymbol{\Omega}\|_{\mathrm{op}}\|\mathbf{D}_\eta^{-1}\|_{\mathrm{op}} \\ &\quad +\|\mathbf{D}^{-1}\|_{\mathrm{op}}\|\boldsymbol{\Omega}\|_{\mathrm{op}}\|\mathbf{D}_\eta^{-1}-\mathbf{D}^{-1}\|_{\mathrm{op}}. \end{aligned} \tag{A.12}$$

Combine (A.10)–(A.12). Since $\|\boldsymbol{\Sigma}_\eta-\boldsymbol{\Sigma}\|_{\mathrm{op}}<\varepsilon_0$, and all the remaining terms are uniformly bounded above by $(K_0^{-1}-\varepsilon_0)^{-1}$, we have $\|\mathbf{R}_\eta-\mathbf{R}\|_\infty \le 3(K_0^{-1}-\varepsilon_0)^{-6}\varepsilon_0$. So, we can make ε_0 sufficiently small such that $\|\mathbf{R}_\eta-\mathbf{R}\|_\infty \le (1-r_1)/2$. Therefore, $\|\mathbf{R}_\eta\|_\infty \le r_1+(1-r_1)/2=(1+r_1)/2$. Let $r_2=(1+r_1)/2<1$. Then, (d) holds as desired.

A.4. Proof of Theorem 4

First note that

$$\|\hat{\boldsymbol{\Sigma}}_\eta-\boldsymbol{\Sigma}_\eta\|_\infty = \|\hat{\boldsymbol{\Sigma}}_\eta-\mathbf{S}_n\|_\infty+\|\mathbf{S}_n-\boldsymbol{\Sigma}\|_\infty+\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_\eta\|_\infty.$$

It is known from [7] that $\|\mathbf{S}_n-\boldsymbol{\Sigma}\|_\infty = O_P\{\sqrt{\ln(p)/n}\}$ and from the proof of Lemma 1 that $\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_\eta\|_\infty = o(1)$. By the definition of $\hat{\boldsymbol{\Sigma}}_\eta$, $\|\hat{\boldsymbol{\Sigma}}_\eta-\mathbf{S}_n\|_\infty = o(1)$ given $\mathbf{S}_n$. Thus, we can have $\|\hat{\boldsymbol{\Sigma}}_\eta-\boldsymbol{\Sigma}_\eta\|_\infty = O_P\{\sqrt{\ln(p)/n}\}$. Then, by the proof of Theorem 5 in [8], it suffices to show $\|\boldsymbol{\Omega}_\eta\|_{L_1} \le M_p$ with $M_p^2 = o\{\sqrt{n}/(\ln p)^{3/2}\}$.

Observe that

$$\begin{aligned}\|\boldsymbol{\Omega}_\eta\|_{L_1} &= \max_{j\in\{1,\dots,p\}}\sum_{i=1}^p|w_{ij}| = \max_{j\in\{1,\dots,p\}}\sum_{i=1}^p|r_{ij}^*|\sqrt{\sigma_{ii}^*\sigma_{jj}^*} \\ &\le C\max_{j\in\{1,\dots,p\}}\sum_{i=1}^p|r_{ij}^*-r_{ij}+r_{ij}|\sqrt{\sigma_{ii}\sigma_{jj}} \qquad \text{because for any } i,\ \sigma_{ii}^* \asymp \sigma_{ii} \\ &\le C\max_{j\in\{1,\dots,p\}}\sum_{i=1}^p|r_{ij}|\sqrt{\sigma_{ii}\sigma_{jj}}+C\max_{j\in\{1,\dots,p\}}\sum_{i=1}^p|r_{ij}^*-r_{ij}|\sqrt{\sigma_{ii}\sigma_{jj}} \\ &\le C\|\boldsymbol{\Omega}\|_{L_1}+CK_0\max_{j\in\{1,\dots,p\}}\sum_{i=1}^p|r_{ij}^*-r_{ij}|. \end{aligned}$$

It is known from [26] that the population Winsorized covariance σ_{ij}^* is 0 if $\sigma_{ij}=0$. Since

$$\max_{j\in\{1,\dots,p\}}\sum_{i=1}^p\mathbf{1}(|v_{ij}|>0)\le M_p,$$

we have

$$\max_{j\in\{1,\dots,p\}}\sum_{i=1}^p|r_{ij}^*-r_{ij}| \le \|\mathbf{R}_\eta-\mathbf{R}\|_\infty\max_{j\in\{1,\dots,p\}}\sum_{i=1}^p\mathbf{1}(|v_{ij}|>0)\le M_p.$$

Together with $\|\boldsymbol{\Omega}\|_{L_1}\le M_p$, we have $\|\boldsymbol{\Omega}_\eta\|_{L_1}\le M_p$, as desired. □

Appendix B. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.jmva.2018.09.013.

References

[1] F.A. Alqallaf, K.P. Konis, R.D. Martin, R.H. Zamar, Scalable robust covariance and correlation estimates for data mining, in: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2002, pp. 14–23.
[2] M. Avella-Medina, H.S. Battey, J. Fan, Q. Li, Robust estimation of high-dimensional covariance and precision matrices, Biometrika 103 (2016) 1–14.
[3] Z. Bai, H. Saranadasa, Effect of high dimension: By an example of a two sample problem, Statist. Sinica 6 (1996) 311–329.
[4] D.S. Bernstein, Matrix Mathematics: Theory, Facts, and Formulas With Application to Linear Systems Theory, Princeton University Press, Princeton, NJ, 2005.

[5] P.J. Bickel, On some robust estimates of location, Ann. Math. Stat. (1965) 847–858.
[6] P. Billingsley, Convergence of Probability Measures, Wiley, New York, 1999.
[7] T. Cai, W. Liu, X. Luo, A constrained ℓ_1 minimization approach to sparse precision matrix estimation, J. Amer. Statist. Assoc. 106 (2011) 594–607.
[8] T. Cai, W. Liu, Y. Xia, Two-sample test of high dimensional means under dependence, J. R. Stat. Soc. Ser. B Stat. Methodol. 76 (2014) 349–372.
[9] S.X. Chen, Y.-L. Qin, A two-sample test for high-dimensional data with applications to gene-set testing, Ann. Statist. 38 (2010) 808–835.
[10] V. Chernozhukov, D. Chetverikov, K. Kato, Central limit theorems and bootstrap in high dimensions, Ann. Probab. 45 (2017) 2309–2352.
[11] A. DasGupta, Asymptotic Theory of Statistics and Probability, Springer, New York, 2008.
[12] S.S. Dhar, P. Chaudhuri, On the derivatives of the trimmed mean, Statist. Sinica 22 (2012) 655–679.
[13] T. Duong, ks: Kernel Smoothing, R Package version 1.10.7, 2017. URL https://CRAN.R-project.org/package=ks.
[14] H. Hotelling, The generalization of student's ratio, Ann. Math. Stat. 2 (1931) 360–378.
[15] J. Hu, Z. Bai, A review of 20 years of naive tests of significance for high-dimensional mean vectors and covariance matrices, Sci. China Math. 59 (2016) 2281–2300.
[16] K.S. Kannan, K. Manoj, S. Arumugam, Labeling methods for identifying outliers, Internat. J. Stat. Syst. 10 (2015) 231–238.
[17] M. Kuhn, K. Johnson, Functions and data sets for 'applied predictive modeling', R Package version 1.1-6, 2014. URL https://CRAN.R-project.org/package=AppliedPredictiveModeling.
[18] M. Kuhn, K. Johnson, Applied Predictive Modeling, Springer, New York, 2013.
[19] P.-L. Loh, X.L. Tan, High-dimensional robust precision matrix estimation: Cellwise corruption under ϵ-contamination, Electron. J. Stat. 12 (2018) 1429–1467.
[20] K. Lounici, High-dimensional covariance matrix estimation with missing observations, Bernoulli 20 (2014) 1029–1058.
[21] V. Öllerer, C. Croux, Robust high-dimensional precision matrix estimation, in: K. Nordhausen, S. Taskinen (Eds.), Modern Nonparametric, Robust and Multivariate Methods: Festschrift in Honour of Hannu Oja, Springer, 2015, pp. 325–350.
[22] M.S. Srivastava, M. Du, A test for the mean vector with fewer observations than the dimension, J. Multivariate Anal. 99 (2008) 386–402.
[23] M. Thulin, A high-dimensional two-sample test for the mean using random subspaces, Comput. Statist. Data Anal. 74 (2014) 26–38.
[24] R. Vershynin, High-Dimensional Probability: An Introduction with Applications in Data Science, Cambridge University Press, 2018.
[25] L. Wang, B. Peng, R. Li, A high-dimensional nonparametric multivariate test for mean vector, J. Amer. Statist. Assoc. 110 (2015) 1658–1669.
[26] R.R. Wilcox, Some results on a Winsorized correlation coefficient, British J. Math. Statist. Psych. 46 (1993) 339–349.
[27] J. Zhang, M. Pan, A high-dimension two-sample test for the mean using cluster subspaces, Comput. Statist. Data Anal. 97 (2016) 87–97.