Journal of Multivariate Analysis 188 (2022) 104813

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

# An overview of tests on high-dimensional means

Yuan Huang [a,1], Changcheng Li [b,1], Runze Li [b,1,*], Songshan Yang [c,1]

[a] Department of Biostatistics, Yale School of Public Health, New Haven, CT 06510, USA
[b] Department of Statistics, The Pennsylvania State University at University Park, PA 16802, USA
[c] Institute of Statistics and Big Data, Renmin University of China, Beijing 100872, China

ARTICLE INFO

*Article history:*
Received 1 August 2021
Received in revised form 17 August 2021
Accepted 17 August 2021
Available online 3 September 2021

*AMS 2020 subject classifications:*
62H15
62F03

*Keywords:*
Hotelling's $T^2$ test
Multiple comparison
Projection test
Regularization method

ABSTRACT

Testing high-dimensional means has many applications in scientific research. For instance, it is of great interest to test whether there is a difference of gene expressions between control and treatment groups in genetic studies. This can be formulated as a two-sample mean testing problem. However, the Hotelling $T^2$ test statistic for the two-sample mean problem is no longer well defined due to singularity of the sample covariance matrix when the sample size is less than the dimension of data. Over the last two decades, the high-dimensional mean testing problem has received considerable attentions in the literature. This paper provides a selective overview of existing testing procedures in the literature. We focus on the motivation of the testing procedures, the insights into how to construct the test statistics and the connections, and comparisons of different methods.



## 1. Introduction

With the rapid development of modern data collection and processing technologies, a vast amount of data with large dimensional features become increasingly popular and have been involved in many scientific areas such as biology, medicine, finance, and social science, calling for an advancement of classical methods to handle the high dimensionality. In recent years, considerable attention has been devoted to variable selection and feature screening ([17] and references therein). Statistical inference for high-dimensional means has been a very active research topic in the literature because of its important applications, such as those in genetic studies. For instance, many biological processes involve regulation of multiple genes, and such research suffers from low power for detecting important genetic markers and poor reproducibility if it focuses on the analysis of individual genes [47]. In other cases, genes are often analyzed in their functional groups to reduce the complexity of analysis [25]. Accordingly, the analysis of gene sets/pathways, which are groups of genes sharing common biological functions, chromosomal locations, or regulations, has become increasingly important in modern biological research. In many important applications, the problem of evaluating whether a group of genes are differentially expressed from another group can be formulated as a problem of testing two-sample means.

Hotelling's $T^2$ test [23] perhaps is the most well-known test on means in the multivariate analysis when the sample is from multivariate normal distributions. To implement the Hotelling $T^2$ test, the sample size $n$ should be greater than the dimension $p$ of data. Motivated by real-world applications, Dempster [13,14] proposed tests for a two-sample

* Corresponding author.
*E-mail address:* rzli@psu.edu (R. Li).

[1] The authors are listed in alphabetic order, and all authors have equally contributed to this work. All authors have made contributions to each sections.

https://doi.org/10.1016/j.jmva.2021.104813
0047-259X/

normal mean problem when $n < p$. Läuter [27] proposed exact $t$ and $F$ tests for normal mean problems based on left-spherical distribution theory [12] to improve the power of the Hotelling $T^2$ when $n < p$. See more detailed discussions in Section 2. Bai and Saranadasa [2] employed random matrix theory to prove that the power of $T^2$ test can be adversely affected even with $p < n$. Since the seminal work [2], testing hypotheses on high-dimensional means has become a very active topic.

This paper aims to provide a selective overview of research on testing high-dimensional mean problem. We will focus on the two-sample mean problem. Since the Hotelling $T^2$ test involves inverse of a sample covariance matrix and is not well defined when the inverse does not exist, Bai and Saranadasa [2] proposed a test statistic based on the $L_2$-distance between the sample mean and the population mean, and has inspired many follow-up works including, but not limited to, [8,10,40,41,48,54]. We review these methods in Section 3. Multiple comparison has been used to construct tests for high-dimensional means by considering tests for means of individual variables. This leads to $L_\infty$-type tests, which have been shown to be more powerful than the $L_2$-distance based tests in the presence of a few large sparse signals [4,52]. We review works on this topic in Section 4. Since the $L_2$-distance based tests may be more powerful than the $L_\infty$-type tests in the presence of dense signals, i.e., many small signals. These tests cannot dominate each other. Adaptive tests are $L_\gamma$-distance based tests with $\gamma$ being selected by data-driven methods. In other words, the adaptive tests essentially aim to achieve high power against various kinds of alternatives by adapting test statistics based on $p$-values calculated from statistics of different orders [22,51]. These tests are reviewed in Section 5. The $L_2$-distance based tests, $L_\infty$-type tests, and the adaptive tests do not take into account the correlation among variables. To utilize the correlation information in testing the high-dimensional means, researchers have considered projecting the high-dimensional samples to a low-dimensional space and then applying the classical Hotelling $T^2$ test on the projected data. Lopes et al. [34] constructed a random projection test, followed by Thulin [43] and Srivastava et al. [42] with permutation-based computation methods to handle multiple projections. Huang [24] derived the theoretical optimal direction with which the projection test possesses the best power under alternatives, and further proposed a sample-splitting strategy to construct an exact $t$ test. Li and Li [31] and Liu et al. [33] further studied how to implement the projection test using the optimal projection direction in practice. Section 6 provides a comprehensive review of these projection tests. We provide a numerical simulation comparison among these tests for the high-dimensional two-sample mean problem in Section 7, followed by discussions in Section 8.

## 2. The Hotelling $T^2$ and related tests

Suppose that $\boldsymbol{x}_i$, $i \in \{1, \ldots, N\}$, is an independent and identically distributed sample from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the $p$-dimensional normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Of interest is to test

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0 \text{ versus } H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0,$$

where $\boldsymbol{\mu}_0$ is a known constant. This test is referred to as the one-sample normal mean problem in the literature. The most well-known test for this hypothesis is the Hotelling $T^2$ test [23]. Let $\bar{\boldsymbol{x}}$ and $\boldsymbol{S}$ be the sample mean and sample covariance matrix, respectively. Based on the likelihood ratio criterion, one may derive the Hotelling $T^2$:

$$T^2 = N(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)^\top \boldsymbol{S}^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0).$$

The properties of $T^2$ have been well studied. See, for example, Chapter 5 of [1].

It has been observed that when the dimension $p$ is close to the sample size $N$, $T^2$ has low power [2,27]. In particular, when $N \leq p$, $\boldsymbol{S}$ is not invertible, and $T^2$ is not well defined. A natural question is how to construct an exact test when $N \leq p$ with fixed and finite $N$ and $p$. Let $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top$, which follows a matrix normal $N_{N\times p}(\mathbf{1}_N \boldsymbol{\mu}^\top, \boldsymbol{I}_N \otimes \boldsymbol{\Sigma})$, where $\mathbf{1}_N$ is an $N$-dimensional column vector with all elements being one, $\boldsymbol{I}_N$ is an $N \times N$ identity matrix, and $\otimes$ denotes the Kronecker product.

Dempster [13,14] dealt with the singularity issue of $\boldsymbol{S}$ by using an orthogonal transformation on data. Let $\boldsymbol{B}$ be an orthogonal matrix with the first row $\mathbf{1}_N/\sqrt{N}$, and let $\boldsymbol{Y} = \boldsymbol{B}\boldsymbol{X}$. Denote $\boldsymbol{y}_i^\top$ as the $i$th row of $\boldsymbol{Y}$. Using the property of matrix normal distribution, it follows that $\boldsymbol{y}_i$'s are independent, $\boldsymbol{y}_1 \sim N_p(\sqrt{N}\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and $\boldsymbol{y}_i \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$, $i \in \{2, \ldots, N\}$. The test proposed in [13,14] for the one-sample mean problem corresponds to

$$T_D = \frac{\|\boldsymbol{y}_1 - \boldsymbol{\mu}_0\|^2}{\sum_{i=2}^N \|\boldsymbol{y}_i\|^2/(N-1)}.$$

Under $H_0$, $\|\boldsymbol{y}_i\|^2$ is a quadratic form of $N_p(\mathbf{0}, \boldsymbol{\Sigma})$. Dempster [13,14] suggested approximating $\|\boldsymbol{y}_i\|^2$ by a scaled chi-square distribution and estimating its scale parameter and degrees of freedom by fitting the first two moment equations. Thus, under $H_0$, $T_D$ approximately follows an $F$-distribution.

Läuter [27] proposed a novel idea to construct $T^2$ by using the property of left-spherical distributions [12]. Without loss of generality, assume that $\boldsymbol{\mu}_0 = \mathbf{0}$. Then $\boldsymbol{X}$ follows $N_{N\times p}(\mathbf{0}, \boldsymbol{I}_N \otimes \boldsymbol{\Sigma})$, and therefore $\boldsymbol{X}$ follows a left-spherical distribution. That is, for any orthogonal $N \times N$ matrix $\boldsymbol{\Gamma}$, $\boldsymbol{X}$ and $\boldsymbol{\Gamma}\boldsymbol{X}$ have the same distribution. Using the invariance property of left-spherical distributions, Läuter [27] proposed projecting the data along the direction $\boldsymbol{D}(\boldsymbol{X}^\top\boldsymbol{X})$, a $p \times d$ matrix that depends on $\boldsymbol{X}$ only through $\boldsymbol{X}^\top\boldsymbol{X}$, and using the Hotelling $T^2$ based on $\boldsymbol{X}\boldsymbol{D}(\boldsymbol{X}^\top\boldsymbol{X})$ rather than $\boldsymbol{X}$. The authors showed that the resulting $T^2$ still follows the Hotelling $T^2$ distribution with the dimension $p$ replaced by $d$. Thus, the resulting $T^2$ test

is still an exact test for the one-sample mean problem. Frick [19] pointed out power insufficiency for the two special cases of Läuter's tests, one attains the highest power in the situation where all variables have nearly the same relative deviation and the same correlation to each other and the other works well when the covariance has a one-factor structure.

Chen et al. [9] proposed a regularized Hotelling $T^2$ test, referred to as the RHT test, by replacing $\boldsymbol{S}^{-1}$ in the definition of $T^2$ by $(\boldsymbol{S}+\lambda\boldsymbol{I})^{-1}$, a ridge-type estimate of the $\boldsymbol{\Sigma}^{-1}$, where $I$ is the identity matrix and $\lambda$ is a ridge tuning parameter. The authors further developed the theory of the RHT test and derived its limiting null distribution. Based on Chen et al. [9], Li et al. [30] further proposed a data-driven procedure to select the regularization parameter $\lambda$, and also proposed an adaptive test which combines the RHT statistics corresponding to a set of regularization parameters. Note that the RHT test can be viewed as an improvement of the BS and SD tests which are reviewed in Section 3 by incorporating the correlation between variables into the test to improve power. Similarly, the projection tests to be introduced in Section 6 give us an effective way to utilize the correlation information to improve power.

The Hotelling $T^2$ test has been further used for testing two-sample mean problems. Let $\boldsymbol{x}_{ij}, j\in\{1,\ldots,N_i\}$, be a random sample from $N_p(\boldsymbol{\mu}_i,\boldsymbol{\Sigma})$ for $i=1$ and 2, and $N=N_1+N_2$, the total sample size. The two-sample mean problem is to test

$$H_0:\boldsymbol{\mu}_1=\boldsymbol{\mu}_2 \text{ versus } H_1:\boldsymbol{\mu}_1\neq\boldsymbol{\mu}_2. \tag{1}$$

Let $\bar{\boldsymbol{x}}_i$ be the sample mean of $\boldsymbol{x}_{ij}$'s, and

$$\boldsymbol{S}=\frac{1}{n}\sum_{i=1}^{2}\sum_{j=1}^{N_i}(\boldsymbol{x}_{ij}-\bar{\boldsymbol{x}}_i)(\boldsymbol{x}_{ij}-\bar{\boldsymbol{x}}_i)^\top,$$

which is the pooled sample covariance matrix, where $n=N_1+N_2-2$.

The one-sample $T^2$ test can be naturally extended to the two-sample mean problem (1):

$$T^2=\frac{N_1N_2}{N_1+N_2}(\bar{\boldsymbol{x}}_1-\bar{\boldsymbol{x}}_2)^\top\boldsymbol{S}^{-1}(\bar{\boldsymbol{x}}_1-\bar{\boldsymbol{x}}_2), \tag{2}$$

which is well defined for an invertible $\boldsymbol{S}$. Furthermore under the null hypothesis,

$$\frac{n-p+1}{np}T^2\sim F_{p,n+1-p}.$$

As seen, the one-sample and two-sample mean problems essentially can be handled in the same strategy. Thus, we will focus on the two-sample mean problem, partly because the two-sample mean problem has many direct applications in high-dimensional genetic data analysis and other fields.

As analyzed in [2], the Hotelling $T^2$ test has low power when $\boldsymbol{S}$ is near singular. Of course, $\boldsymbol{S}$ is singular when $p>n$, and the Hotelling $T^2$ test is not well defined. To address the challenges, various tests for the one-sample and two-sample problems have been developed. In the following sections, we introduce the main testing procedures for the high-dimensional two-sample mean problem without normality assumption. As natural extensions of the multivariate normal distribution, independent component model and elliptical distributions are the two distribution classes mainly assumed in the literature of testing two-sample means.

**Definition 1.** A random vector $\boldsymbol{x}$ follows an independent component model (ICM) if $\boldsymbol{x}$ can be represented as $\boldsymbol{x}=\boldsymbol{\Gamma z}+\boldsymbol{\mu}$, where $\boldsymbol{\Gamma}$ is a $p\times m$ matrix for some $m\geq p$ such that $\boldsymbol{\Gamma\Gamma}^\top=\boldsymbol{\Sigma}$, and $\boldsymbol{z}=(z_1,\ldots,z_m)^\top$ is an $m$-dimensional random vector with independent and identically distributed elements $z_j$'s with $\mathrm{E}(z_j)=0$, $\mathrm{E}(z_j^2)=1$ and $\mathrm{E}(z_j^4)<\infty$.

**Definition 2** (*Elliptical Distribution*)**.** A random vector $\boldsymbol{x}$ follows an elliptically contoured distribution if its characteristic function $\mathrm{E}\{\exp(i\boldsymbol{x}^\top\boldsymbol{t})\}$ is of the form $\exp(i\boldsymbol{t}^\top\boldsymbol{\mu})\phi(\boldsymbol{t}^\top\boldsymbol{\Sigma t})$ for some function $\phi(\cdot)$. When $\boldsymbol{\mu}=\boldsymbol{0}$ and $\boldsymbol{\Sigma}=\boldsymbol{I}$, $\boldsymbol{x}$ follows a spherical distribution.

Both ICM and elliptical distributions are natural extensions of a multivariate normal distribution, which is the only distribution belonging to both ICM and elliptical distributions.

## 3. $L_2$-type tests

Consider the two-sample mean problem without normality assumption. Let $\boldsymbol{x}_{ij}, j\in\{1,\ldots,N_i\}$, $i\in\{1,2\}$, be a random sample from a population with mean $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}$. For $N_i$ large enough, $\bar{\boldsymbol{x}}_i$ asymptotically follows $N_p(\boldsymbol{\mu}_i,\boldsymbol{\Sigma})$, and $\bar{\boldsymbol{x}}_1-\bar{\boldsymbol{x}}_2$ asymptotically follows $N_p(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2,(N_1^{-1}+N_2^{-1})\boldsymbol{\Sigma})$. Then the $T^2$ defined in (2) asymptotically follows a $\chi^2_p$ when $p$ is fixed and finite. When $p>n$, $\boldsymbol{S}$ becomes singular. Thus, the Hotelling $T^2$ test cannot be used in the high-dimensional setting. Intuitively, a test statistic based on $\|\bar{\boldsymbol{x}}_1-\bar{\boldsymbol{x}}_2\|^2$, an estimate of $\|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\|^2$, may be used for testing the two-sample mean problem. We refer to such tests as $L_2$-type tests since they are based on the $L_2$-norm of the difference of the two sample means.

Since $\mathrm{E}\|\bar{\boldsymbol{x}}_1-\bar{\boldsymbol{x}}_2\|^2=(N_1^{-1}+N_2^{-1})\mathrm{tr}(\boldsymbol{\Sigma})$, Bai and Saranadasa [2] first considered the following test statistic for the two-sample problem in (1) as

$$T_n=(\bar{\boldsymbol{x}}_1-\bar{\boldsymbol{x}}_2)^\top(\bar{\boldsymbol{x}}_1-\bar{\boldsymbol{x}}_2)-(N_1{}^{-1}+N_2{}^{-1})\mathrm{tr}\boldsymbol{S}.$$

Under the null hypothesis, $\mathrm{E}(T_n) = 0$ and $\mathrm{Var}(T_n) = \sigma_{T_n}^2 = 2({N_1}^{-1}+{N_2}^{-1})^2(1+\frac{1}{n})\mathrm{tr}\boldsymbol{\Sigma}^2$ under the normality assumption, and $\mathrm{Var}(T_n) = \sigma_{T_n}^2\{1+o(1)\}$ under ICM.

Under the ICM assumption, $y_n = p/n \to y \in (0, 1)$, as $n \to \infty$, and $N_1/N \to \kappa > 0$. Bai and Saranadasa [2] showed that under the null hypothesis in (1), as $n \to \infty$,

$$\frac{T_n}{\sqrt{\mathrm{Var}(T_n)}} \sim N(0, 1).$$

Using large-dimensional random matrix theory, Bai and Saranadasa [2] showed that when $y \in (0, 1)$, the plug-in estimator of $\sigma_{T_n}^2$, i.e., substituting $\mathrm{tr}\boldsymbol{\Sigma}^2$ by $\mathrm{tr}\boldsymbol{S}^2$, is not a consistent estimator of $\mathrm{Var}(T_n)$ if $\lambda_{\max}(\boldsymbol{\Sigma}) = o(\mathrm{tr}\boldsymbol{\Sigma}^2)$, where $\lambda_{\max}(\boldsymbol{\Sigma})$ stands for the largest eigenvalue of $\boldsymbol{\Sigma}$. They further showed that $[n^2/\{(n+2)(n-1)\}][\mathrm{tr}\boldsymbol{S}^2 - (\mathrm{tr}\boldsymbol{S})^2/n]$ is an unbiased and ratio-consistent estimator of $\mathrm{tr}\boldsymbol{\Sigma}^2$, and proposed a test, referred to as the BS test, for the two-sample mean problem:

$$T_{\mathrm{BS}} = \frac{(\frac{1}{N_1}+\frac{1}{N_2})^{-1}(\bar{\boldsymbol{x}}_1-\bar{\boldsymbol{x}}_2)^\top(\bar{\boldsymbol{x}}_1-\bar{\boldsymbol{x}}_2) - \mathrm{tr}\boldsymbol{S}}{\sqrt{\frac{2(n+1)n}{(n+2)(n-1)}\{\mathrm{tr}\boldsymbol{S}^2 - (\mathrm{tr}\boldsymbol{S})^2/n\}}}.$$

Under some conditions, Bai and Saranadasa [2] derived the asymptotic power of the BS test as

$$\beta_{\mathrm{BS}}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) - \Phi\left\{-\xi_\alpha + \frac{n\kappa(1-\kappa)\|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\|^2}{\sqrt{2\mathrm{tr}\boldsymbol{\Sigma}^2}}\right\} \to 0,$$

where $\xi_\alpha$ is the $100(1-\alpha)$th percentile of the standard normal distribution for a given significance level $\alpha$. The authors further proved that the BS test may be more powerful than the Hotelling $T^2$ test when $p/n$ is close to one. They also noted that the BS test has the same asymptotic power with the test proposed in [13,14]. Notice that under the null hypothesis, the asymptotic distribution of the BS test and several $L_2$-type tests to be introduced are approximated using normal distributions. Instead of a normal approximation, Zhang et al. [54] proposed to use the Welch–Satterthwaite (W–S) $\chi^2$-approximation [37,50] to achieve adaptivity of the null distribution. Zhang et al. [54] further conducted a thorough analysis on theoretical properties and empirical analysis of the W–S $\chi^2$-approximation and concluded that the W–S $\chi^2$-approximation is at least comparable to and can be more accurate than the normal approximation under certain scenarios.

The Hotelling $T^2$ test is affine invariant. That is, the two-sample Hotelling $T^2$ test is invariant under a linear transformation $\boldsymbol{y}_{ij} = \boldsymbol{A}\boldsymbol{x}_{ij} + \boldsymbol{b}$ for a nonsingular constant square matrix $\boldsymbol{A}$ and a constant vector $\boldsymbol{b}$. The BS test does not possess this property. Indeed, the BS test is not invariant under $y_{ijk} = a_k x_{ijk} + b_k$ for $k \in \{1, \ldots, p\}$, where $x_{ijk}$ is the $k$th element of $\boldsymbol{x}_{ij}$. One way to deal with this issue is to scale each variable by dividing its sample standard deviation. Denote $\boldsymbol{D}_{\boldsymbol{S}} = \mathrm{diag}(\boldsymbol{S})$, the diagonalized matrix of $\boldsymbol{S}$, and consider $(\bar{\boldsymbol{x}}_1-\bar{\boldsymbol{x}}_2)^\top\boldsymbol{D}_{\boldsymbol{S}}^{-1}(\bar{\boldsymbol{x}}_1-\bar{\boldsymbol{x}}_2)$ instead of $\|\bar{\boldsymbol{x}}_1-\bar{\boldsymbol{x}}_2\|^2$. Srivastava and Du [41] proposed a test, referred to as the SD test, with the test statistic defined as

$$T_{\mathrm{SD}} = \frac{\frac{N_1N_2}{N_1+N_2}(\bar{\boldsymbol{x}}_1-\bar{\boldsymbol{x}}_2)^\top\boldsymbol{D}_{\boldsymbol{S}}^{-1}(\bar{\boldsymbol{x}}_1-\bar{\boldsymbol{x}}_2) - \frac{np}{n-2}}{\{2(\mathrm{tr}\boldsymbol{R}^2 - p^2/n)c_{p,n}\}^{\frac{1}{2}}},$$

where $c_{p,n}$ is an adjustment coefficient to improve convergence, and it should satisfy that $c_{p,n} \to 1$ in probability as $(n, p) \to \infty$. The authors suggested using $c_{p,n} = 1 + p^{-3/2}\mathrm{tr}R^2$, where $\boldsymbol{R} = \boldsymbol{D}_{\boldsymbol{S}}^{-\frac{1}{2}}\boldsymbol{S}\boldsymbol{D}_{\boldsymbol{S}}^{-\frac{1}{2}}$ is the sample correlation.

We denote $\boldsymbol{\mathcal{R}}$ to be the corresponding correlation matrix of the covariance matrix $\boldsymbol{\Sigma}$, and $\lambda_{\max}(\boldsymbol{\mathcal{R}})$ to be the largest eigenvalue of $\boldsymbol{\mathcal{R}}$. It is assumed that $n = O(p^\zeta)$ with $1/2 < \zeta \le 1$, $N_1/N \to \kappa \in (0, 1)$, $0 < \lim_{p\to\infty} p^{-1}\mathrm{tr}\boldsymbol{\mathcal{R}}^k < \infty$, $k \in \{1, \ldots, 4\}$, and $\lim_{p\to\infty}\lambda_{\max}(\boldsymbol{\mathcal{R}})/\sqrt{p} = 0$. Srivastava and Du [41] showed that $T_{\mathrm{SD}} \sim N(0, 1)$ under $H_0$, and further derived its asymptotic power function as

$$\beta_{\mathrm{SD}}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) - \Phi\left\{-\xi_\alpha + \frac{N_1N_2}{N_1+N_2}\frac{(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top\boldsymbol{D}_{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)}{\sqrt{2\mathrm{tr}\boldsymbol{\mathcal{R}}^2}}\right\} \to 0,$$

as $n, p \to \infty$.

Srivastava and Du [41] further compared the power function with that of the BS test and showed that the SD test may enjoy higher power than the BS test when the diagonal elements of $\boldsymbol{\Sigma}$ are not the same and some regularity conditions are satisfied.

Gregory et al. [21] proposed the generalized component test (referred to as GCT), which is a centered and scaled version of the statistic that takes the form of the mean of the squared two-sample $t$-statistics with unpooled variance over all $p$ components. The choices of centering quantity relate to the dimension and the formulation of scaling quantity rests on the assumption that the dependence among components is autocovarying and diminishing as components are further apart. Chakraborty and Chaudhuri [6] noted that the size of GCT is larger than the nominal level under the autoregressive model as well as spherical $t$ distributions for all values of $p$, which can be corrected using permutation-based critical values.

Chen and Qin [10] first noted that some strong moment conditions in [2] are due to the terms $\sum_{j=1}^{N_i}\boldsymbol{x}_{ij}^\top\boldsymbol{x}_{ij}$, $i \in \{1, 2\}$, in the expansion of $\|\bar{\boldsymbol{x}}_1-\bar{\boldsymbol{x}}_2\|^2$. However, these two terms are not useful in the sample mean testing problem. Chen and

Qin [10] proposed the following test, referred to as the CQ test, with the test statistic that does not involve these two unnecessary terms:

$$T_{\mathrm{CQ}} = \frac{1}{N_1(N_1-1)}\sum_{i\neq j}^{N_1}\boldsymbol{x}_{1i}^{\top}\boldsymbol{x}_{1j} + \frac{1}{N_2(N_2-1)}\sum_{i\neq j}^{N_2}\boldsymbol{x}_{2i}^{\top}\boldsymbol{x}_{2j} - \frac{2}{N_1N_2}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\boldsymbol{x}_{1i}^{\top}\boldsymbol{x}_{2j}.$$

Chen and Qin [10] considered the two-sample mean problem with unequal covariance matrix. Specifically, let $\boldsymbol{x}_{ij}, j \in \{1,\ldots,N_i\}$, be a random sample from a population with mean $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$. Chen and Qin [10] derived the mean and the asymptotic variance of $T_{\mathrm{CQ}}$. Under the null hypothesis, $\mathrm{E}(T_{\mathrm{CQ}}) = 0$, and

$$\mathrm{Var}(T_{\mathrm{CQ}}) - \left\{\frac{2}{N_1(N_1-1)}\mathrm{tr}(\boldsymbol{\Sigma}_1^2) + \frac{2}{N_2(N_2-1)}\mathrm{tr}(\boldsymbol{\Sigma}_2^2) + \frac{4}{N_1N_2}\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)\right\} \to 0.$$

Chen and Qin [10] established the asymptotic normality of $T_{\mathrm{CQ}}$ under the null and the local alternative hypothesis, and further derived the asymptotic power of their test under certain regularity conditions. Note that $T_{\mathrm{CQ}}$ is the same as that of the test proposed by [2], but Chen and Qin [10] studied the asymptotic properties of $T_{\mathrm{CQ}}$ and derived its asymptotic power under more general setting and weaker technical conditions than those given in [2].

Wang et al. [48] extended the CQ test to a nonparametric test, referred to as the WPL test, for high-dimensional one-sample mean problem $H_0 : \boldsymbol{\mu} = \mathbf{0}$. The WPL test shares the same form of the CQ test for the one-sample mean problem with $\boldsymbol{x}_i$ replaced by its spatial sign $\boldsymbol{x}_i/\|\boldsymbol{x}_i\|$. Under elliptical distribution assumption on the population and some other regularity conditions, Wang et al. [48] further studied the asymptotic properties of their proposed nonparametric test, and demonstrated that the nonparametric test can be more powerful than the CQ test when the sample is from heavy-tailed elliptical distributions. Li et al. [32] illustrated that classical spatial-sign-based procedures for a low-dimensional population are not robust for high-dimensional settings, and may lead to an inflated Type I error rate. Li et al. [32] further developed a correction to make the sign-based tests applicable for high-dimensional data, and proved that the corrected test statistic is asymptotically normal under elliptical distributions. Chakraborty and Chaudhuri [6] examined the CQ test and WPL test closely under the $\rho$-mixing and randomly scaled $\rho$-mixing assumptions. The two cover some commonly seen models, e.g., spherical Gaussian distributions is a special case of $\rho$-mixing models and multivariate spherical $t$ distribution is a special case of randomly scaled $\rho$-mixing models. Chakraborty and Chaudhuri [6] concluded that the power of CQ test and WPL test tend to be the same as $p \to \infty$ regardless of sample size given appropriate mixing conditions and some regularity conditions; in addition, the WPL test can be asymptotically more powerful than the CQ test under a stronger correlation and both $p, n \to \infty$. Other nonparametric tests including, but not limited to, [3,18,26,46] have also been developed for the two-sample mean problem.

Chen et al. [8] noted that the non-signal components only inflate the variance of $T_{\mathrm{CQ}}$ without any contribution to the power of the test when alternatives are sparse. As such, Chen et al. [8] proposed a hard thresholding method, referred to as the CLZ test, to remove the components with no signal before carrying out the $T_{\mathrm{CQ}}$ test. Although the actual statistic used by [8] is of a similar nature as $T_{\mathrm{CQ}}$, we represent the test statistic proposed in [8] similar to $T_{\mathrm{BS}}$ for notation simplicity. Let $\bar{X}_1^{(k)}$ and $\bar{X}_2^{(k)}$ are $k$th elements of $\bar{\boldsymbol{x}}_1$ and $\bar{\boldsymbol{x}}_2$, respectively, and $S_{1,kk}$ and $S_{2,kk}$ are the sample variance for the $k$th component in the first and second sample, respectively. Then,

$$T_{\mathrm{CLZ}}(s) = \sum_{k=1}^{p}\left\{\frac{(\bar{X}_1^{(k)} - \bar{X}_2^{(k)})^2}{S_{1,kk}/N_1 + S_{2,kk}/N_2} - 1\right\} I\left\{\frac{(\bar{X}_1^{(k)} - \bar{X}_2^{(k)})^2}{S_{1,kk}/N_1 + S_{2,kk}/N_2} > \lambda_p(s)\right\},$$

where $I\{\cdot\}$ is an indicator function, $\lambda_p(s) = 2s\log p$, $s \in (0,1)$, and the form of $\lambda_p(s)$ is based on the large deviation results [36]. Chen et al. [8] further established the asymptotic normality of $T_{\mathrm{CLZ}}(s)$ under certain regularity conditions and derived the asymptotic power of $T_{\mathrm{CLZ}}(s)$ under local alternatives.

One has to determine $s$ to implement $T_{\mathrm{CLZ}}(s)$. However, the optimal choice of the threshold $s$ depends on the difference between the true population means, which is unknown in general. If all of the signals in the population mean difference are strong enough, the threshold $s$ can be chosen very close to one to remove all the components with no signal while preserving all the components with signals. However, if some signals are weak, $s$ has to be chosen according to the strength of the weak signals, which is usually unknown in practice. To deal with this issue, Chen et al. [8] proposed to choose the most significant test statistic among possible choices of threshold values as their final test statistic:

$$T_{\mathrm{CLZ}} = \max_{s\in(0,1-\eta)}\{T_{\mathrm{CLZ}}(s) - \hat{\mu}_{T_{\mathrm{CLZ}}(s),0}\}/\hat{\sigma}_{T_{\mathrm{CLZ}}(s),0},$$

where $\eta$ is a parameter with a small positive value, and $\hat{\mu}_{T_{\mathrm{CLZ}}(s),0}$ and $\hat{\sigma}_{T_{\mathrm{CLZ}}(s),0}$ are estimates of the mean and standard deviation of $T_{\mathrm{CLZ}}(s)$ under the null hypothesis derived in [8]. Chen et al. [8] further derived the asymptotic null distribution for $T_{\mathrm{CLZ}}$ using the theory of extreme value distributions. Although the asymptotic null distribution can be derived using the extreme value theory, Chen et al. [8] found that the convergence rate of $T_{\mathrm{CLZ}}$ is slow. As a result, Chen et al. [8] proposed to use the bootstrap method to calculate the $p$-value of their test. Zhong et al. [56] developed a new test for high-dimensional means under sparsity, as an alternative to higher criticism (HC), which was introduced to determine whether there are any nonzero signals in the settings in which there is only a small fraction of significant signals against a predominantly null background. A comprehensive review on the basics of HC in both the testing and feature selection settings is given in Donoho and Jin [15].

## 4. $L_\infty$-type tests

Note that the two-sample mean problem is equivalent to testing simultaneously the following hypotheses:

$$H_{0k}: \mu_{1k} = \mu_{2k} \text{ versus } H_{1k}: \mu_{1k} \neq \mu_{2k},$$

for $k \in \{1, \ldots, p\}$. For each $k$, we may construct a $z$-test for each one-dimensional two-sample mean problem:

$$z_k = \frac{\bar{X}_{1k} - \bar{X}_{2k}}{\sqrt{S_{1,kk}/N_1 + S_{2,kk}/N_2}},$$

where the notation is the same as that in Section 3. Under mild conditions, $z_k$ asymptotically follows a normal distribution. Cai et al. [4] proposed a $L_\infty$-type test, referred to as the CLX test, with the test statistic accounting for sparse alternatives,

$$T_{\text{CLX}} = \max_{1\le k\le p} \frac{|\bar{X}_{1k} - \bar{X}_{2k}|^2}{S_{1,kk}/N_1 + S_{2,kk}/N_2},$$

which equals $\|\boldsymbol{u}\|_\infty$ with $\boldsymbol{u} = (z_1^2, \ldots, z_p^2)^\top$.

Using the extreme value theory and under some regularity conditions, Cai et al. [4] derived the asymptotic null distribution of $T_{\text{CLX}}$ and proposed an asymptotic test accordingly. The idea behind the construction of the $L_\infty$-type statistic $T_{\text{CLX}}$ is to pick up the strongest signals in the difference of means while ignoring other signals. Thus it will have advantages over the $L_2$-type testing methods when the signals are sparse, which has been shown in various simulation studies [2,10,41]. Chang et al. [7] advocated a data-driven approach to obtain critical values using Monte Carlo simulations based on the facts that convergence rate to the extreme value distribution for maximum-type statistics is usually slow and that the strong structural assumptions on the covariance matrices may be difficult to justify in applications. Chang et al. [7] also proposed a screening step to reduce the dimension and enhance power.

Xue and Yao [52] proposed a distribution and correlation-free two-sample mean test built upon an $L_\infty$-type test, referred to as the XY test, with the test statistic defined as

$$T_{\text{XY}} = \sqrt{N_1}\|\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2\|_\infty,$$

which only depends on the infinity norm of the sample mean difference. Xue and Yao [52] further derived theoretical properties of $T_{\text{XY}}$ based on a high-dimensional central limit theorem, and provided a data-driven critical value which can be easily computed via a multiplier bootstrap method. Notably, the result of [52] does not require samples to be independent and identically distributed and allows two samples to have highly unequal sizes. From the definition of $T_{\text{XY}}$, it is not invariant under scale transformation. In practice, one may have to scale it by using the trick of the SD test.

## 5. Adaptive tests

$L_2$-type tests like [2,10] and $L_\infty$-type tests like [4,52] present two extremes — the $L_2$-type tests use all the information in all the dimensions, while the $L_\infty$-type tests use only the dimension with the strongest signal as evidence against the null hypothesis. Typically, the $L_2$-type tests are powerful against dense alternatives, where the difference of two population means has a large proportion of non-zero elements; while the $L_\infty$-type tests are powerful against sparse alternatives, where the mean difference only has a small proportion of non-zero elements. In practice, it is unknown whether the alternative is dense or sparse. To deal with this issue, adaptive tests have been proposed to achieve high power against various kinds of alternatives simultaneously.

Xu et al. [51] developed an adaptive testing procedure, referred to as the XLWP test, which is powerful against both the sparse and dense alternatives or alternatives in-between in the high-dimensional setting. They incorporated the idea of $L_2$-type tests and $L_\infty$-type tests and proposed a family of sum-of-powers tests with a power index $\gamma$ as follows.

$$T_{\text{XLWP}}(\gamma) = \sum_{k=1}^{p} |\bar{X}_{1k} - \bar{X}_{2k}|^\gamma,$$

for $1 \le \gamma < \infty$ and

$$T_{\text{XLWP}}(\infty) = \max_{1\le k\le p} \frac{|\bar{X}_{1k} - \bar{X}_{2k}|^2}{S_{1,kk}/N_1 + S_{2,kk}/N_2}.$$

Note that $T_{\text{XLWP}}(\gamma)$ coincides with $T_{\text{BS}}$ if $\gamma = 2$ and $T_{\text{CLX}}$ if $\gamma = \infty$. Xu et al. [51] demonstrated that there are settings in which $T_{\text{XLWP}}(\gamma)$ with some $\gamma$ between 2 and $\infty$ is more powerful than $T_{\text{BS}}$ and $T_{\text{CLX}}$. Furthermore, Xu et al. [51] proposed an adaptive test to combine various sum-of-powers tests with different $\gamma$'s as follows.

$$T_{\text{XLWP}} = \min_{\gamma\in\mathcal{G}} P_{T_{\text{XLWP}}(\gamma)},$$

where $\mathcal{G}$ is a candidate set of $\gamma$ and $P_{T_{\text{XLWP}}(\gamma)}$ is the $p$-value calculated from $T_{\text{XLWP}}(\gamma)$. Note that since $T_{\text{XLWP}}$ is the minimum of some $p$-values, it is no longer a genuine $p$-value. In order to perform the proposed adaptive test, Xu et al. [51] derived the asymptotic null and alternative distributions for $T_{\text{XLWP}}$ under certain regularity conditions. Also note that to use the test $T_{\text{XLWP}}$, $\mathcal{G}$ needs to be pre-specified. Xu et al. [51] suggested using $\mathcal{G} = \{1, 2, \ldots, 6, \infty\}$, and more details can be found in [51].

He et al. [22] proposed an adaptive testing procedure which combines $p$-values computed from $U$-statistics of different orders. While He et al. [22] focused on a general framework of high-dimensional testing, their test can also be applied in the high-dimensional mean testing problem. For the two-sample mean testing problem, define

$$T_{\text{HXWP}}(a) = \sum_{j=1}^{p} \sum_{c=0}^{a} \binom{a}{c} \frac{(-1)^{(a-c)}}{P_c^{N_1} P_{a-c}^{N_2}} \sum_{\substack{(k_1,\ldots,k_c)\in A_c^{N_1} \\ (s_1,\ldots,s_{a-c})\in A_{a-c}^{N_2}}} \prod_{t=1}^{c} x_{1k_tj} \prod_{m=1}^{a-c} x_{2s_mj}, \tag{3}$$

for $1 \le a < \infty$, where $P_a^N = N!/(N-a)!$ is the arrangement number, $A_c^N = \{(a_1, \ldots, a_c) : 1 \le a_1 \ne \cdots \ne a_c \le N\}$ is the set of arrangements and $x_{ikj}$ is the $j$th element of $\boldsymbol{x}_{ik}$. He et al. [22] noted that $T_{\text{HXWP}}(a)$ is an unbiased $U$-statistic estimator for $\sum_{j=1}^{p}(\mu_{1j} - \mu_{2j})^a$, and also defined $T_{\text{HXWP}}(\infty)$ to be the same as $T_{\text{XLWP}}(\infty)$. Similar to $T_{\text{XLWP}}(\gamma)$, $T_{\text{HXWP}}(2)$ is powerful against dense alternatives, $T_{\text{HXWP}}(\infty)$ is powerful against sparse alternatives, and $T_{\text{HXWP}}(a)$ with an appropriate $a$ can be powerful for alternatives in-between.

He et al. [22] derived the asymptotic null distribution for $T_{\text{HXWP}}(a)$ with a finite integer $a$ and $a = \infty$, and they further showed that $T_{\text{HXWP}}(a)$ asymptotically follows a normal distribution for a finite integer $a$ and an extreme value distribution for $a = \infty$ under certain regularity conditions. From the property of $U$-statistics, He et al. [22] showed that $T_{\text{HXWP}}(a)$ with different $a$'s are asymptotically independent with each other. Similar to [51], He et al. [22] proposed to use an adaptive test to combine $p$-values from statistics of different orders as

$$T_{\text{HXWP}} = \min_{a\in\mathcal{A}} P_{T_{\text{HXWP}}(a)},$$

where $P_{T_{\text{HXWP}}(a)}$ is the $p$-value calculated from $T_{\text{HXWP}}(a)$ and $\mathcal{A}$ is some set of candidate $a$'s. Given the asymptotic independence among $T_{\text{HXWP}}(a)$ with different $a$'s, He et al. [22] derived the asymptotic $p$-value for $T_{\text{HXWP}}$ as $1 - (1 - T_{\text{HXWP}})^{|\mathcal{A}|}$, where $|\mathcal{A}|$ is the size of the candidate set $\mathcal{A}$. For implementation, the candidate set $\mathcal{A}$ needs to be pre-specified. He et al. [22] proposed to use $\mathcal{A} = \{1, 2, \ldots, 6, \infty\}$. Note that [51] and [22] are quite similar in the setting of the mean testing problem. A main difference is that He et al. [22] derived better theoretical properties such as asymptotic independence between testing statistics of different orders by using $U$-statistics instead of $V$-statistics as in [51]. However, the $U$-statistics in (3) is hard to compute directly when $a$ is large. To solve this problem, He et al. [22] also proposed a calculation scheme which can calculate (3) with time complexity $O(p^2(N_1+N_2))$ instead of $O(p^2(N_1+N_2)^a)$ as in the naive calculation approach. He et al. [22] also discussed other $p$-value combination methods such as Fisher's method beyond the minimum $p$-value combination method.

## 6. Projection tests

Test statistics introduced in Sections 3, 4, and 5 do not utilize correlation among the variables and therefore do not require an estimation of $\boldsymbol{\Sigma}^{-1}$, which may result in loss of power. Projection tests have been considered to achieve higher power by taking into account correlation. Earlier work on projection tests such as [27] target exact tests with finite $p$ and $n$. The exact tests proposed in [27] were further extended to linear multivariate tests on mean structures of matrix normal distribution in [28] with a correction in [29].

Lopes et al. [34] proposed a random projection test, referred to as the LJW test, that projects the sample to a randomly generated lower dimensional space such that the classical Hotelling $T^2$ test can be applied. Specifically, the LJW test is processed under the normality assumption when $p \ge n/2$. Let $\mathbf{P}_k^\top$ be a $k \times p$ projection matrix with independent and identically distributed $N(0, 1)$ entries, where $k$ is suggested to take $[n/2]$, where $[a]$ is the largest integer less than $a$. The projected samples $\{\mathbf{P}_k^\top \boldsymbol{x}_{11}, \ldots, \mathbf{P}_k^\top \boldsymbol{x}_{1N_1}\}$ and $\{\mathbf{P}_k^\top \boldsymbol{x}_{21}, \ldots, \mathbf{P}_k^\top \boldsymbol{x}_{2N_2}\}$ can be considered as independent and identically distributed samples from $N(\mathbf{P}_k^\top \boldsymbol{\mu}_1, \mathbf{P}_k^\top \boldsymbol{\Sigma}\mathbf{P}_k)$ and $N(\mathbf{P}_k^\top \boldsymbol{\mu}_2, \mathbf{P}_k^\top \boldsymbol{\Sigma}\mathbf{P}_k)$, respectively. The Hotelling $T^2$ test can be processed by testing the two projected samples with

$$H_{p0} : \mathbf{P}_k^\top \boldsymbol{\mu}_1 = \mathbf{P}_k^\top \boldsymbol{\mu}_2 \quad \text{versus} \quad H_{p1} : \mathbf{P}_k^\top \boldsymbol{\mu}_1 \ne \mathbf{P}_k^\top \boldsymbol{\mu}_2.$$

Suppose that $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = o(1)$ and $N_1/(N_1 + N_2) \to \kappa \in (0, 1)$, Lopes et al. [34] showed that under all sequences of projections $\mathbf{P}_k^\top$, the asymptotic power function of the LJW test satisfies, as $n \to \infty$,

$$\beta\{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}, \mathbf{P}_k^\top\} - \Phi\{-\xi_\alpha + \kappa(1-\kappa)\sqrt{n/2}\Delta_k^2\} \to 0,$$

where $\Delta_k^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \mathbf{P}_k (\mathbf{P}_k^\top \boldsymbol{\Sigma}\mathbf{P}_k)^{-1} \mathbf{P}_k^\top (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$.

The seemingly intuitive idea of the random projection test is motivated from the consideration that the test is designed to reduce the dimension to prevent accumulation of variance from the high-dimensional variables and meanwhile not to bring the distance too close so that it is harder to distinguish. Thulin [43] proposed a modification of the LJW test,

allowing the test statistics to be invariant under linear transformations of the marginal distributions. Multiple random projections are conducted and the test statistic averages the individual random projection Hotelling tests, whose power is then calculated by permutation. Thulin [43] demonstrated that its test offers a higher power when the variables are dependent. On a similar note, Srivastava et al. [42] proposed a test using multiple random projections with the test statistic averaging over the individual random projection Hotelling test p-values. These tests are referred to as the random projection (RP) tests.

A key question to projection tests is whether there exists an optimal projection so that the resulting projection test is the most powerful. To address this question, Huang [24] formulated this issue as follows. For $k \ll p$, let $\boldsymbol{A}$ be a $p \times k$ nonzero constant matrix with rank $k$. Based on the projected sample $\boldsymbol{y}_{ij} = \boldsymbol{A}^\top \boldsymbol{x}_{ij}$, the two-sample Hotelling $T^2$ test for $H_{0A} : \boldsymbol{A}^\top(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = 0$ can be written as

$$T_A^2 = \frac{N_1 N_2}{N_1 + N_2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{A}(\boldsymbol{A}^\top \boldsymbol{S}\boldsymbol{A})^{-1}\boldsymbol{A}^\top(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2),$$

provided that $\boldsymbol{A}^\top \boldsymbol{S}\boldsymbol{A}$ is invertible. Under the normality assumption,

$$\frac{N_1 + N_2 - k - 1}{kn} T_A^2 \sim F_{k, N_1+N_2-k-1},$$

under $H_0$, which implies $H_{0A}$ holds. Huang [24] proved that $T_A^2$ reaches its best power at $k = 1$ and $\boldsymbol{A} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$. Denote the optimal projection by $\boldsymbol{a} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$. As a result, $H_{0A}$ becomes $H_{0\boldsymbol{a}} : (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$. Suppose that $\boldsymbol{\Sigma}$ is positive definite, then $H_{0\boldsymbol{a}}$ is equivalent to $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$.

The estimation of the optimal projection $\boldsymbol{a}$ is challenging since it involves the estimation of $\boldsymbol{\Sigma}^{-1}$. To construct an exact projection test, Huang [24] proposed a sample-splitting strategy to estimate $\boldsymbol{a}$ where the data are partitioned into a subset for estimation and a subset for conducting the test. Let $\bar{\boldsymbol{x}}_1^{(1)} - \bar{\boldsymbol{x}}_2^{(1)}$ and $\boldsymbol{S}^{(1)}$ be the sample mean difference and pooled sample covariance matrix obtained from the subset for estimation, respectively. Since $\boldsymbol{S}^{(1)}$ is not invertible when $p > n$, Huang [24] proposed to estimate $\boldsymbol{a}$ by

$$\hat{\boldsymbol{a}} = \left(\boldsymbol{S}^{(1)} + \lambda \boldsymbol{D}_{\boldsymbol{S}^{(1)}}\right)^{-1}\left(\bar{\boldsymbol{x}}_1^{(1)} - \bar{\boldsymbol{x}}_2^{(1)}\right),$$

where $\boldsymbol{D}_{\boldsymbol{S}^{(1)}} = \mathrm{diag}(\boldsymbol{S}^{(1)})$ and $\lambda$ is a ridge tuning parameter. Thus, the projection test with the optimal direction, referred to as the OP test, is

$$T_{OP}^2 = \frac{N_1^{(2)} N_2^{(2)}}{N_1^{(2)} + N_2^{(2)}}\left(\bar{\boldsymbol{x}}_1^{(2)} - \bar{\boldsymbol{x}}_2^{(2)}\right)^\top \hat{\boldsymbol{a}}\left(\hat{\boldsymbol{a}}^\top \boldsymbol{S}^{(2)}\hat{\boldsymbol{a}}\right)^{-1}\hat{\boldsymbol{a}}^\top\left(\bar{\boldsymbol{x}}_1^{(2)} - \bar{\boldsymbol{x}}_2^{(2)}\right),$$

where $\bar{\boldsymbol{x}}_1^{(2)} - \bar{\boldsymbol{x}}_2^{(2)}$ and $\boldsymbol{S}^{(2)}$ are the sample mean difference and pooled sample covariance matrix obtained from the subset for conducting the test, and $N_1^{(2)}$ and $N_2^{(2)}$ are the sample sizes for the two samples in this subset. The authors also demonstrated that under the local alternative

$$H_1 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \boldsymbol{\delta}\sqrt{\frac{1}{N_1} + \frac{1}{N_2}},$$

where $\boldsymbol{\delta}$ is a constant vector, the asymptotic power of the OP test is no less than those of BS, DS, and CQ test under certain conditions.

Li and Li [31] investigated the projection tests for the linear hypothesis testing problem in linear models with high-dimensional responses, which includes the high-dimensional mean testing problem as a special case. In the setting of the two-sample mean problem, the test proposed by [31], referred to as the LL test, can be seen as a multiple data-splitting extension of [24] to solve the power loss problem of a single data splitting. Li and Li [31] derived the asymptotic normality of their test statistic under certain regularity conditions and proposed to use bootstrap methods to carry out the test. Li and Li [31] further showed that their test has similar asymptotic power with those of $T_{\mathrm{BS}}$ and $T_{\mathrm{CQ}}$ in the presence of low correlation among variables, and that their test can be much more powerful than some existing tests in the presence of high correlation.

In the construction of $T_{OP}^2$, a ridge-type estimator is used to estimate the optimal projection direction. However, the ridge type estimator is not consistent in high-dimensional settings in general. To deal with the problem of optimal projection direction estimation, Liu et al. [33] proposed to use nonconvex regularized quadratic programming to estimate the optimal projection direction. Although Liu et al. [33] mainly focused on the one-sample mean testing problem, we can easily modify it for the two-sample high-dimensional mean testing problem. Denote $\boldsymbol{w}^* = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$, the optimal projection direction to be estimated. Note that $\boldsymbol{w}^*$ is the solution to the following optimization problem $\boldsymbol{w}^* = \mathrm{argmin}_{\boldsymbol{w}}\left[\frac{1}{2}\boldsymbol{w}^\top \boldsymbol{\Sigma}\boldsymbol{w} - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{w}\right]$. Liu et al. [33] considered the following optimization problem to estimate the optimal projection direction

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\mathrm{argmin}}\left[\frac{1}{2}\boldsymbol{w}^\top \hat{\boldsymbol{\Sigma}}\boldsymbol{w} - (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{w} + \sum_{j=1}^{p} P_\lambda(w_j)\right], \tag{4}$$

where $\boldsymbol{w} = (w_1, \ldots, w_p)^\top$, $\hat{\boldsymbol{\Sigma}} = \boldsymbol{S}$ is the pooled sample covariance matrix, and $P_\lambda(w)$ is a penalty function with a tuning parameter $\lambda$ to promote the sparseness of the estimator. Commonly used penalty functions include the Lasso [44], the SCAD [16], the MCP [53], and others [17]. Liu et al. [33] further established that any stationary point $\hat{\boldsymbol{w}}$ of the problem (4) is a good estimator for optimal projection direction $\boldsymbol{w}^*$ under some regularity conditions.

To reduce the power loss from the data splitting, Liu et al. [33] further proposed a multiple-splitting projection test which repeats the single projection procedure $m$ times, obtaining $p$-values $p_k$, $k \in \{1, \ldots, m\}$ for some fixed integer $m$. Liu et al. [33] noted that these $p$-values are exchangeable in distribution. That is, $(p_1, \ldots, p_m) \overset{d}{=} (p_{\pi_1}, \ldots, p_{\pi_m})$ for any permutation $\pi$ on $\{1, \ldots, m\}$. They further proposed a $p$-value combination method which utilizes the exchangeability of the $p$-values. More specifically, let $Z_k = \Phi^{-1}(p_k)$, $k \in \{1, \ldots, m\}$. Under the null hypothesis, $Z_k$, $i \in \{1, \ldots, m\}$ are exchangeable standard normal random variables. Denote $\rho$ to be the correlation between $Z_i$ and $Z_j$, $1 \le i < j \le m$, and let $\hat{\rho}$ be some consistent estimator of $\rho$, Liu et al. [33] established that $M_{\hat{\rho}} = \bar{\boldsymbol{Z}}/\sqrt{\{1 + (m-1)\hat{\rho}\}/m}$ follows an asymptotic standard normal distribution under the null hypothesis. However, the asymptotic distribution needs $m$ to be large enough. Liu et al. [33] further proposed a critical value calculation method to control the finite-sample Type I error. Also, Liu et al. [33] proposed to choose $m \in [30, 60]$ for a trade-off between testing power and computational cost.

## 7. Numerical comparisons

In this section, we conduct intensive simulation to compare the performance of the tests introduced in the previous sections using R version 3.4.3. All simulations results are based on 5000 independent replicates. In our simulations, we set the dimension $p = 1000$, $n_1 = n_2 = n$ and the significance level 0.05.

We consider two types of alternatives: the sparse alternative where $\boldsymbol{\mu}_1 = \boldsymbol{0}$ and $\boldsymbol{\mu}_2 = c(\boldsymbol{1}_{10}^\top, \boldsymbol{0}_{p-10})^\top$ and dense local alternative where $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are generated from $N_p(\boldsymbol{0}, (c^2/n)\mathbf{I}_p)$. The sparse alternative is designed to challenge the $L_2$-type tests, while the dense local alternative is to challenge the $L_\infty$-type tests. We set $c = 0$, and 0.5 and 1 to examine the Type I error rate, and the power of the tests, respectively.

We consider two covariance structures: (1) compound symmetry (CS) with $\boldsymbol{\Sigma}_1 = (1-\rho)\boldsymbol{I}_p + \rho \boldsymbol{1}_p \boldsymbol{1}_p^\top$, where $\boldsymbol{I}_p$ is the $p \times p$ identity matrix; and (2) autoregressive (AR) correlation with $\boldsymbol{\Sigma}_2 = (\rho^{|i-j|})$, both with $\rho = 0.5$ for a moderate correlation and 0.8 for a high correlation. Denote $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ with $(i, j)$-element $\omega_{ij}$. Then for the CS correlation structure, $\omega_{ij}$ is a constant for $i \ne j$; for the AR correlation structure, $\omega_{ij} = 0$ for $|i-j| \ge 2$. The correlation among the variables is not utilized in the tests introduced in Sections 3, 4, and 5. Thus, the CS correlation structure is designed to challenge these tests, which may take advantage of the AR correlation structure whose inverse is very close to the identity matrix. The projection tests introduced in Section 6 may take advantage of the CS correlation structure since the correlation is taken into account.

We generate data from two multivariate distributions, multivariate normal and multivariate $t$ with degrees of freedom 6. A multivariate normal distribution belongs to the class of ICM, while a multivariate $t$ distribution is a special case of elliptical distributions. Using these two distributions enables us to examine how sensitive the performance of the tests is to the ICM assumption, and how the limiting null distributions are related to the ICM assumption.

In our simulations, we directly use the R package highmean version 3.0 to implement several tests, including [2] (aBS, eBS), [51] (aXLWP, eXLWP), [4] (aCLX, eCLX), [10] (aCQ, eCQ), [8] (aCLZ, eCLZ), and [41] (aSD, eSD). Here "a-" and "e-" represent asymptotic-based and permutation-based tests, respectively. The permutation parameter is set to 200 for permutation-based tests in the R package. For random projection tests, we conduct an asymptotic-based test with a single projection (aRP1) following [34], and a permutation-based test with a single projection (eRP1) and 30 projections (eRP30) following [43] and the codes provided in its supplementary material. Both permutation parameters are set to 100 for eRP30. We use the R package ARHT version 0.1.0 to implement [30] (ARHT). We also include [13] (DEM), [22] (HXWP), [27] (LAU), [24] (OP), [31] (LL), and [52] (XY) in this numerical comparison.

Due to the limited space, we present and discuss the results with $(n, \rho) = (40, 0.5)$. Results for $(n, \rho) = (40, 0.8)$, $(100, 0.5)$ and $(100, 0.8)$ are given in the supplementary material of this paper. It can be seen from the figures presented in the supplement that the overall patterns for $(n, \rho) = (40, 0.8), (100, 0.5)$ and $(100, 0.8)$ are similar to those for $(n, \rho) = (40, 0.5)$.

Fig. 1 depicts the Type I error and power for multivariate normal data. From Fig. 1(a) and (b), it can be seen that all tests retain the Type I error rate 0.05 very well except for aCLX and aCLZ. The aCLZ test inflates the Type I error rate for both correlation structures significantly, while the aCLX test inflates the Type I error rate only for the AR correlation structure. Fortunately, both eCLX and eCLZ retain the Type I error rate well. Thus, we should use the power of eCLX and eCLZ rather than that of aCLX and aCLZ for the power comparison.

Fig. 1(c) depicts the power for the AR correlation structure and dense local alternative, and implies that LL, eBS, eCQ, eSD tests have the highest power, followed by the adaptive tests and ARHT. The $L_\infty$-type tests, and both RP1 (aRP1 & eRP1) tests have low power. Fig. 1(d) depicts the power for CS correlation structure and dense local alternative, and indicates that the ARHT test, LL test, OP test, and eRP30 test have the highest power. For a larger signal with $c = 1$, the XY test and eCLX perform quite well, and the adaptive tests introduced in Section 5 have reasonable power. The $L_2$-type tests have the lowest power. This is expected since the $L_2$-tests ignore the correlation.

The power for the AR correlation structure and sparse constant alternative is depicted in Fig. 1(e), from which it can be seen that the adaptive tests perform the best, the $L_\infty$-type tests perform well. The eBS, eCQ, eSD and LL tests perform similarly. The ARHT, OP, aRP1, eRP1, and eRP30 tests have the lowest power.

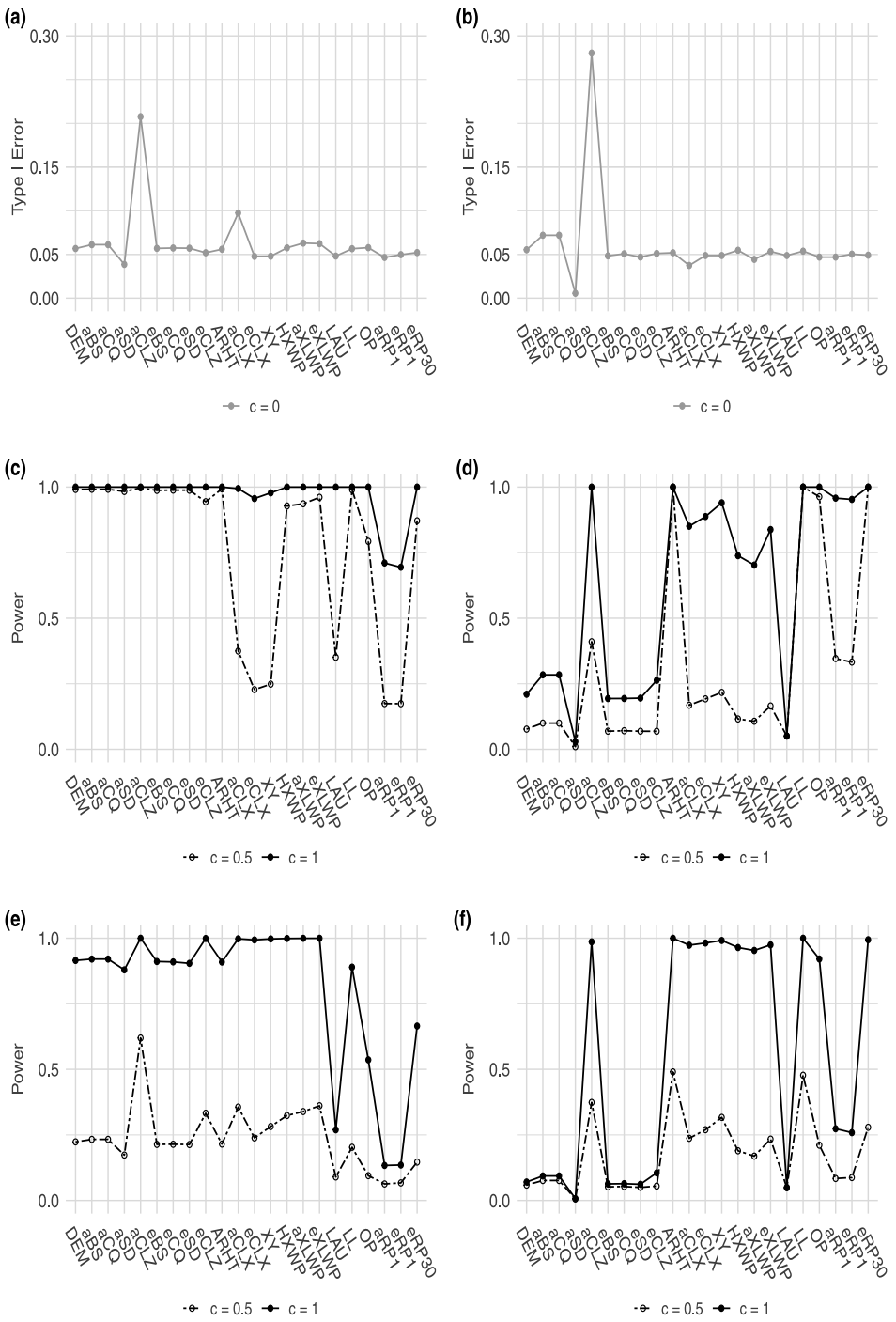


**Fig. 1.** Simulation results under the multivariate normal distribution with different values of $c$, the strength of signals. The null hypothesis corresponds to $c = 0$. The left and right panels are for the autoregressive (AR) and compound symmetric (CS) correlation structure, respectively. The top, middle, and bottom panels are for the Type I error, power for a dense local alternative, and power for a sparse constant alternative, respectively. Results are based on 5000 replications. Tests DEM [13], LAU [27] and ARHT [30] are introduced in Section 2. Tests BS (aBS and eBS, [2]), CQ (aCQ and eCQ, [10]), SD (aSD and eSD, [41]) and CLZ (aCLZ and eCLZ, [8]) can be found in Section 3. Tests CLX (aCLX and eCLX) [4] and XY [52] are defined in Section 4, tests HXWP [22] and XLWP (aXLWP, eXLWP, [51]) are introduced in Section 5. Tests OP [24], LL [31], eRP (eRP1, eRP30,[43]) are given in Section 6.

The power for the CS correlation structure and sparse constant alternative is depicted in Fig. 1(f), from which we can see that the LL and eRP30 tests perform the best, and followed by the XY test and ARHT test. The eCLX test and the adaptive tests have good performance too. The $L_2$-type tests, the aRP1, and eRP1 tests have the lowest power.

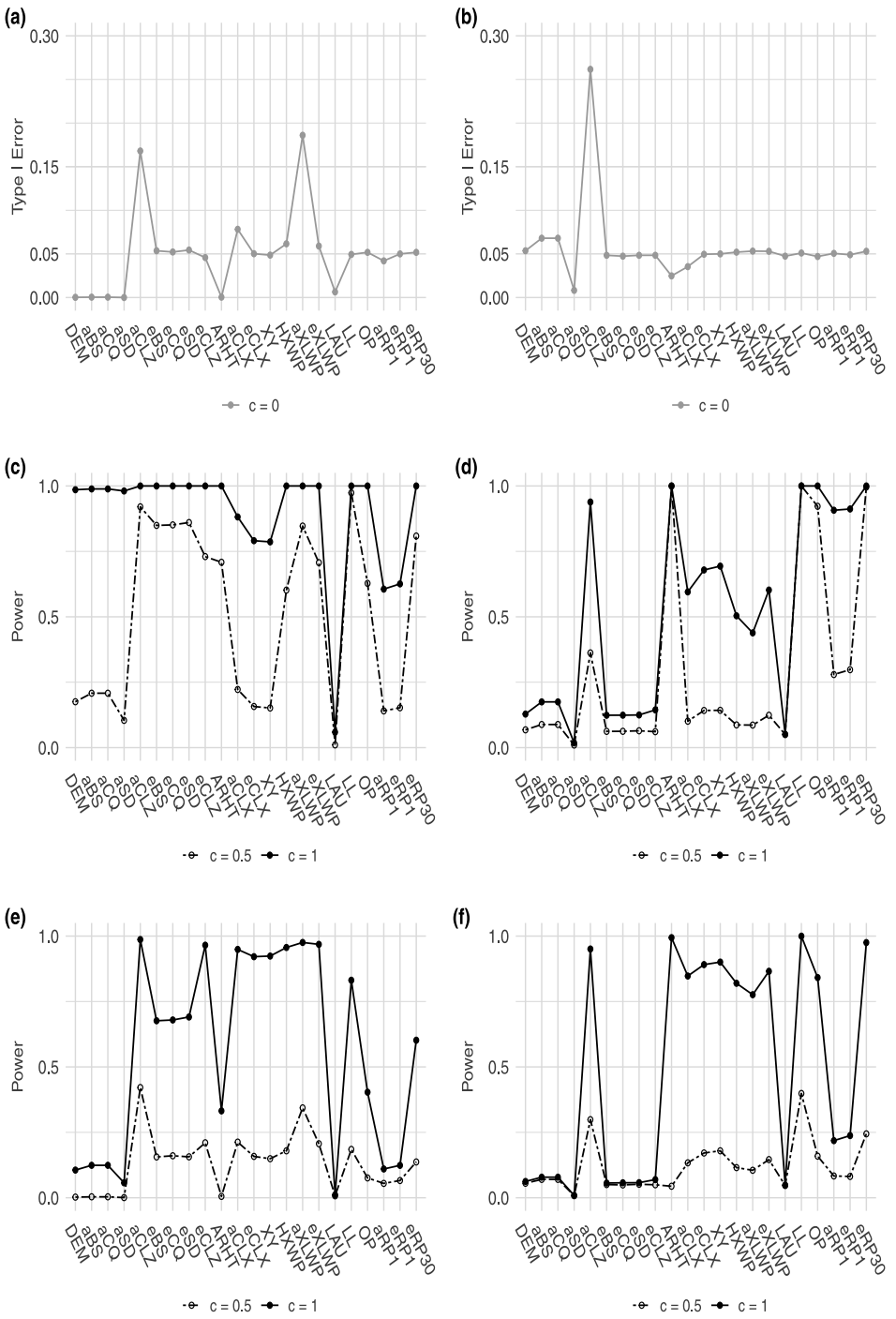


**Fig. 2.** Simulation results under the multivariate $t$ distribution with different values of $c$, the strength of signals. The null hypothesis corresponds to $c = 0$. The left and right panels are for the autoregressive (AR) and compound symmetric (CS) correlation structure, respectively. The top, middle, and bottom panels are for the Type I error, power for a dense local alternative, and power for a sparse constant alternative, respectively. Results are based on 5000 replications. Tests DEM [13], LAU [27] and ARHT [30] are introduced in Section 2. Tests BS (aBS and eBS, [2]), CQ (aCQ and eCQ, [10]), SD (aSD and eSD, [41]) and CLZ (aCLZ and eCLZ, [8]) can be found in Section 3. Tests CLX (aCLX and eCLX) [4] and XY [52] are defined in Section 4, tests HXWP [22] and XLWP (aXLWP, eXLWP, [51]) are introduced in Section 5. Tests OP [24], LL [31], eRP (eRP1, eRP30, [43]) are given in Section 6.

The Type I error and power for the multivariate $t$ distributions are depicted in Fig. 2, from which we can see that in addition to the aCLZ and aCLX tests, the aXLWP test cannot retain the Type I error rate, and the DEM, aBS, aCQ, aSD, and

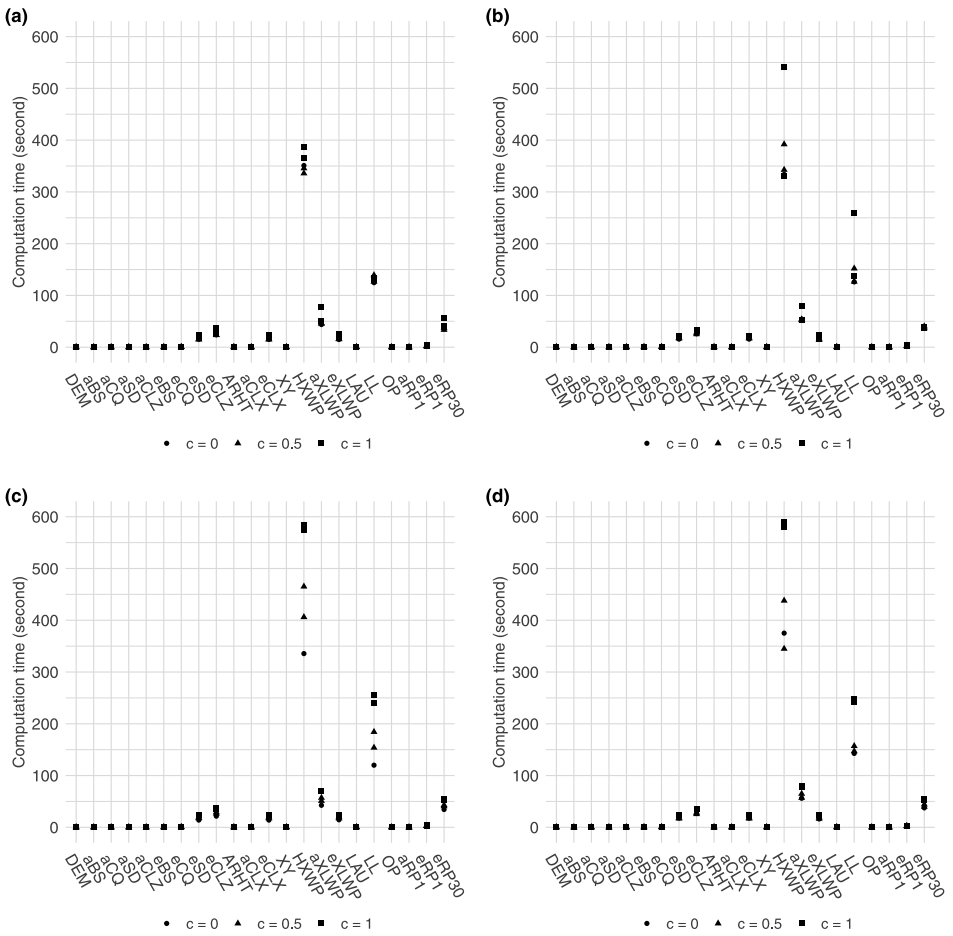


**Fig. 3.** Computation time (second) per replicate. The left and right panels are for the autoregressive (AR) and compound symmetric (CS) correlation structure, respectively. The upper and lower panels are for the multivariate normal and $t$ distributions with different values of $c$, the strength of signals, respectively. The null hypothesis corresponds to $c = 0$. Results are based on 5000 replications. Tests DEM [13], LAU [27] and ARHT [30] are introduced in Section 2. Tests BS (aBS and eBS, [2]), CQ (aCQ and eCQ, [10]), SD (aSD and eSD, [41]) and CLZ (aCLZ and eCLZ, [8]) can be found in Section 3. Tests CLX (aCLX and eCLX) [4] and XY [52] are defined in Section 4, tests HXWP [22] and XLWP (aXLWP, eXLWP, [51]) are introduced in Section 5. Tests OP [24], LL [31], eRP (eRP1, eRP30, [43]) are given in Section 6.

ARHT tests have much more conservative Type I error rates, when data are generated from the multivariate $t$ distribution with an AR correlation structure. As the result, it can be seen from Fig. 2(c) and (e) that these tests have much lower powers for multivariate $t$ distributions than for multivariate normal distributions. We also observe a conservative Type I error rate for ARHT under a CS correlation structure. The patterns of the Type I error and power of tests other than the DEM, aBS, aCQ and aSD tests are similar to those in Fig. 1.

In summary, there is no single test dominating all the other tests in all settings. The performance of the tests is related to the type of alternatives, correlation structures, and the population from which the data are generated. In general, we would recommend the LL and eXLWP tests since their performance is very good in all different settings.

Fig. 3 depicts the computing time for each test. The computing time may vary under settings. From Fig. 3, HXWP is the most costly one, followed by LL and eRP30.

## 8. Conclusions and discussion

This paper presents a selective overview on testing two-sample means of high dimensional data along with their motivations and properties. We classify these tests into several categories: the Hotelling $T^2$ related tests, $L_2$-type tests, $L_\infty$-type tests, adaptive-type tests, and projection tests. We conduct a comprehensive numerical comparison to demonstrate the strength and weakness of these tests. In general, the permutation-based test can retain the Type I error rate better than their asymptotic counterparts. As expected, there is no test which dominates all other tests in all scenarios. In general, we would recommend the LL and eXLWP tests since their performance is very good in all different settings.

There are many works on testing high-dimensional means. It is impossible to include all of them in a review article. For instance, this paper does not review tests for one-sample mean problem and two-sample mean problem based on empirical likelihood [11,49]. In addition, this paper does not include tests that are developed more specifically for data with special structures, such as compositional data [5] or genetic data incorporating pathway topology [25].

We conclude this paper by outlining a few future research directions. It has been common to impose sparsity in the high-dimensional data modeling. For two-sample mean problems, it is reasonable to assume that many variables have the same means. That is, many elements in $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ are 0. This implies that the vector $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ is sparse. How to construct a test that utilizes the sparsity to achieve better power would be an interesting topic for future research.

The challenge of testing high-dimensional means comes from the singularity of the sample covariance matrix or the estimation of the precision matrix high-dimensional data, i.e, the inverse of high-dimensional covariance matrix. There are some interesting works on estimation of the precision matrix of high-dimensional data such as high-dimensional Gaussian graphical models. How to incorporate the recent advances on high-dimensional precision matrix estimation to construct a high-dimensional mean test with better power and controlled Type I error is another interesting topic for future research.

Statistical inference for regression coefficient vectors in high-dimensional linear and generalized linear models have been a very active research topic. See, e.g., Zhang and Zhang [55], Van de Geer et al. [20], Ning and Liu [35], Tibshirani et al. [45], Shi et al. [38] and Shi et al. [39]. Extending the techniques developed in the context of testing high-dimensional means to testing linear hypothesis in high-dimensional regression models is also a great topic for future research.

## CRediT authorship contribution statement

**Yuan Huang:** Made contributions to each sections. **Changcheng Li:** Made contributions to each sections. **Runze Li:** Made contributions to each sections. **Songshan Yang:** Made contributions to each sections.

## Acknowledgments

The authors thank the Editor-in-Chief and the Executive editor for their comments and suggestions. All authors made equal contributions to this paper and are listed in the alphabetic order. This research was supported by National Science Foundation Grants DMS 1820702, DMS 1953196 and DMS 2015539.

## Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jmva.2021.104813.

## References

[1] T.W. Anderson, An Introduction To Multivariate Statistical Analysis, 3rd Edition, Wiley, New York, 2003.
[2] Z. Bai, H. Saranadasa, Effect of high dimension: by an example of a two sample problem, Statist. Sinica (1996) 311–329.
[3] M. Biswas, A.K. Ghosh, A nonparametric two-sample test applicable to high dimensional data, J. Multivariate Anal. 123 (2014) 160–171.
[4] T. Cai, W. Liu, Y. Xia, Two-sample test of high dimensional means under dependence, J. R. Stat. Soc. Ser. B Stat. Methodol. 76 (2) (2014) 349–372.
[5] Y. Cao, W. Lin, H. Li, Two-sample tests of high-dimensional means for compositional data, Biometrika 105 (1) (2018) 115–132.
[6] A. Chakraborty, P. Chaudhuri, Tests for high-dimensional data based on means, spatial signs and spatial ranks, Ann. Statist. 45 (2) (2017) 771–799.
[7] J. Chang, C. Zheng, W.-X. Zhou, W. Zhou, Simulation-based hypothesis testing of high dimensional means under covariance heterogeneity, Biometrics 73 (4) (2017) 1300–1310.
[8] S.X. Chen, J. Li, P.-S. Zhong, Two-sample and ANOVA tests for high dimensional means, Ann. Statist. 47 (3) (2019) 1443–1474.
[9] L.S. Chen, D. Paul, R.L. Prentice, P. Wang, A regularized Hotelling's $T^2$ test for pathway analysis in proteomic studies, J. Amer. Statist. Assoc. 106 (496) (2011) 1345–1360.
[10] S.X. Chen, Y.-L. Qin, A two-sample test for high-dimensional data with applications to gene-set testing, Ann. Statist. 38 (2) (2010) 808–835.
[11] X. Cui, R. Li, G. Yang, W. Zhou, Empirical likelihood test for large dimensional mean vector, Biometrika (2020) 591–607.
[12] A.P. Dawid, Spherical matrix distributions and multivariate model, J. R. Stat. Soc. Ser. B Stat. Methodol. 39 (1977) 254–261.
[13] A.P. Dempster, A high dimensional two sample significance test, Ann. Math. Stat. (1958) 995–1010.
[14] A.P. Dempster, A significance test for the separation of two highly multivariate small samples, Biometrics 16 (1) (1960) 41–50.
[15] D. Donoho, J. Jin, Higher criticism for large-scale inference, especially for rare and weak effects, Statist. Sci. 30 (1) (2015) 1–25.
[16] J. Fan, R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, J. Amer. Statist. Assoc. 96 (456) (2001) 1348–1360.
[17] J. Fan, R. Li, C.-H. Zhang, H. Zou, Statistical Foundations of Data Science, Chapman and Hall/CRC, 2020.
[18] L. Feng, C. Zou, Z. Wang, Multivariate-sign-based high-dimensional tests for the two-sample location problem, J. Amer. Statist. Assoc. 111 (514) (2016) 721–735.
[19] H. Frick, On the power behaviour of Läuter's exact multivariate one-sided tests, Biom. J. 38 (4) (1996) 405–414.
[20] S. Van de Geer, P. Bühlmann, Y. Ritov, R. Dezeure, et al., On asymptotically optimal confidence regions and tests for high-dimensional models, Ann. Statist. 42 (3) (2014) 1166–1202.
[21] K.B. Gregory, R.J. Carroll, V. Baladandayuthapani, S.N. Lahiri, A two-sample test for equality of means in high dimension, J. Amer. Statist. Assoc. 110 (510) (2015) 837–849.
[22] Y. He, G. Xu, C. Wu, W. Pan, Asymptotically independent U-statistics in high-dimensional testing, Ann. Statist. 49 (1) (2021) 154–181.
[23] H. Hotelling, The generalization of Student's ratio, Ann. Math. Stat. 2 (1931) 360–378.

[24] Y. Huang, Projection Test for High-Dimensional Mean Vectors with Optimal Direction Ph.D. dissertation, The Pennsylvania State University at University Park, 2015.
[25] P. Khatri, M. Sirota, A.J. Butte, Ten years of pathway analysis: current approaches and outstanding challenges, PLoS Comput. Biol. 8 (2) (2012) e1002375.
[26] I. Kim, S. Balakrishnan, L. Wasserman, Robust multivariate nonparametric tests via projection averaging, Ann. Statist. 48 (6) (2020) 3417–3441.
[27] J. Läuter, Exact t and f tests for analyzing studies with multiple endpoints, Biometrics 52 (1996) 964–970.
[28] J. Läuter, E. Glimm, S. Kropf, Multivariate tests based on left-spherically distributed linear scores, Ann. Statist. 26 (1998) 1972–1988.
[29] J. Läuter, E. Glimm, S. Kropf, Correction: Multivariate tests based on left-spherically distributed linear scores, Ann. Statist. 27 (1999) 1441–1441.
[30] H. Li, A. Aue, D. Paul, J. Peng, P. Wang, An adaptable generalization of Hotelling's $T^2$ test in high dimension, Ann. Statist. 48 (3) (2020) 1815–1847.
[31] C. Li, R. Li, Linear hypothesis testing in linear models with high dimensional responses, J. Amer. Statist. Assoc. (2021) 1–13, http://dx.doi.org/10.1080/01621459.2021.1884561.
[32] Y. Li, Z. Wang, C. Zou, A simpler spatial-sign-based two-sample test for high-dimensional data, J. Multivariate Anal. 149 (2016) 192–198.
[33] W. Liu, X. Yu, R. Li, Multiple-splitting projection test for high-dimensional mean vectors, 2021, in preparation.
[34] M. Lopes, L. Jacob, M.J. Wainwright, A more powerful two-sample test in high dimensions using random projection, in: Advances in Neural Information Processing Systems, 2011, pp. 1206–1214, Longer version: arXiv preprint arXiv:1108.2401.
[35] Y. Ning, H. Liu, A general theory of hypothesis tests and confidence regions for sparse high dimensional models, Ann. Statist. 45 (1) (2017) 158–195.
[36] V.V. Petrov, Limit Theorems of Probability Theory: Sequences of Independent Random Variables, Oxford Science Publications, Clarendon Press, Oxford, 1995.
[37] F.E. Satterthwaite, An approximate distribution of estimates of variance components, Biom. Bull. 2 (6) (1946) 110–114.
[38] C. Shi, R. Song, Z. Chen, R. Li, Linear hypothesis testing for high dimensional generalized linear models, Ann. Statist. 47 (5) (2019) 2671–2703.
[39] C. Shi, R. Song, W. Lu, R. Li, Statistical inference for high-dimensional models via recursive online-score estimation, J. Amer. Statist. Assoc. In press (2021).
[40] M.S. Srivastava, A test for the mean vector with fewer observations than the dimension under non-normality, J. Multivariate Anal. 100 (3) (2009) 518–532.
[41] M.S. Srivastava, M. Du, A test for the mean vector with fewer observations than the dimension, J. Multivariate Anal. 99 (3) (2008) 386–402.
[42] R. Srivastava, P. Li, D. Ruppert, Raptt: An exact two-sample test in high dimensions using random projections, J. Comput. Graph. Statist. 25 (3) (2016) 954–970.
[43] M. Thulin, A high-dimensional two-sample test for the mean using random subspaces, Comput. Statist. Data Anal. 74 (2014) 26–38.
[44] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. Ser. B Stat. Methodol. (1996) 267–288.
[45] R.J. Tibshirani, J. Taylor, R. Lockhart, R. Tibshirani, Exact post-selection inference for sequential regression procedures, J. Amer. Statist. Assoc. 111 (514) (2016) 600–620.
[46] S.-I. Tsukada, High dimensional two-sample test based on the inter-point distance, Comput. Statist. 34 (2) (2019) 599–615.
[47] G.B. Van der Voet, T.I. Todorov, J.A. Centeno, W. Jonas, J. Ives, F.G. Mullick, Metals and health: a clinical toxicological perspective on tungsten and review of the literature, Military Med. 172 (9) (2007) 1002–1005.
[48] L. Wang, B. Peng, R. Li, A high-dimensional nonparametric multivariate test for mean vector, J. Amer. Statist. Assoc. 110 (2015) 1658–1669.
[49] R. Wang, L. Peng, Y. Qi, Jackknife empirical likelihood test for equality of two high dimensional means, Statist. Sinica (2013) 667–690.
[50] B.L. Welch, The generalization of student's' problem when several different population variances are involved, Biometrika 34 (1/2) (1947) 28–35.
[51] G. Xu, L. Lin, P. Wei, W. Pan, An adaptive two-sample test for high-dimensional means, Biometrika 103 (3) (2016) 609–624.
[52] K. Xue, F. Yao, Distribution and correlation-free two-sample test of high-dimensional means, Ann. Statist. 48 (3) (2020) 1304–1328.
[53] C.-H. Zhang, Nearly unbiased variable selection under minimax concave penalty, Ann. Statist. (2010) 894–942.
[54] J.-T. Zhang, J. Guo, B. Zhou, M.-Y. Cheng, A simple two-sample test in high dimensions based on $L_2$-norm, J. Amer. Statist. Assoc. 115 (530) (2020) 1011–1027.
[55] C.-H. Zhang, S.S. Zhang, Confidence intervals for low dimensional parameters in high dimensional linear models, J. R. Stat. Soc. Ser. B Stat. Methodol. 76 (1) (2014) 217–242.
[56] P.-S. Zhong, S.X. Chen, M. Xu, Tests alternative to higher criticism for high-dimensional means under sparsity and column-wise dependence, Ann. Statist. 41 (6) (2013) 2820–2851.