JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2023, VOL. 118, NO. 544, 2548–2561: Theory and Methods
https://doi.org/10.1080/01621459.2022.2061354

Power-Enhanced Simultaneous Test of High-Dimensional Mean Vectors and Covariance Matrices with Application to Gene-Set Testing

Xiufan Yu[a], Danning Li[b], Lingzhou Xue[c], and Runze Li[c]

[a]Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, Notre Dame, IN; [b]School of Mathematics and Statistics and KLAS, Northeast Normal University, Changchun, China; [c]Department of Statistics, The Pennsylvania State University, University Park, PA

ABSTRACT

Power-enhanced tests with high-dimensional data have received growing attention in theoretical and applied statistics in recent years. Existing tests possess their respective high-power regions, and we may lack prior knowledge about the alternatives when testing for a problem of interest in practice. There is a critical need of developing powerful testing procedures against more general alternatives. This article studies the joint test of two-sample mean vectors and covariance matrices for high-dimensional data. We first expand the high-power regions of high-dimensional mean tests or covariance tests to a wider alternative space and then combine their strengths together in the simultaneous test. We develop a new power-enhanced simultaneous test that is powerful to detect differences in either mean vectors or covariance matrices under either sparse or dense alternatives. We prove that the proposed testing procedures align with the power enhancement principles introduced by Fan, Liao, and Yao and achieve the accurate asymptotic size and consistent asymptotic power. We demonstrate the finite-sample performance using simulation studies and a real application to find differentially expressed gene-sets in cancer studies. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received July 2021
Accepted March 2022

KEYWORDS
Dense alternatives; Fisher's combination; Power-enhanced tests; Power enhancement components; Sparse alternatives

1. Introduction

Inferences on the equality of two distributions are of significant interest in a wide range of real applications. Genetic studies use the differential gene expression analysis to understand how genes are related to diseases (Wang, Peng, and Li 2015). Medical image analysis examines the differential structure of images to diagnose abnormal tissues (Ginestet et al. 2017). Pharmaceutical researchers rely on the analysis of comparative clinical trial outcomes for drug discovery and development (Cummings et al. 2019).

To make inferences on the discrepancies between two distributions, we usually consider their mean vectors and covariance matrices that characterize commonly used distributions, for example, the elliptical distributions (Anderson 2003). Over the past decade, there has been significant progress in testing the equality of two mean vectors (Chen and Qin 2010; Wang, Peng, and Li 2015; Wang and Yuan 2019; Chen, Li, and Zhong 2019) or covariance matrices (Li and Chen 2012; Zhu et al. 2017; Chen, Guo, and Qiu 2019) under the high-dimensional setting. Yet few works are capable of examining both mean vectors and covariance matrices simultaneously.

However, in practice, we often do not know whether the discrepancies reside in mean vectors or in covariance structure. It has been recognized that mean tests are powerful to detect the differences in mean vectors but cannot detect the different covariance structure. In contrast, covariance tests are powerful to identify the differences in covariance structure but are incompetent to distinguish the differential structure of two mean vectors. Thus, it is crucial to develop a new simultaneous testing procedure that is powerful to detect differences in either mean vectors or covariance matrices.

Let $\mathbf{X}$ and $\mathbf{Y}$ be two p-dimensional populations with mean vectors $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$ and covariance matrices $(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2)$, respectively. We consider the simultaneous test on the equality of mean vectors and covariance matrices of the two populations, that is,

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \text{ and } \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2. \tag{1.1}$$

In real-world applications such as genetic studies, the sample size is often less than a hundred, but the number of features can be thousands or even larger (Clarke et al. 2008). Throughout this article, we assume that the dimension p is much larger than the sample size n_1 or n_2. The challenge of high dimensionality leads to fundamental difficulties in understanding the asymptotic behavior of test statistics.

Two different classes of alternatives (i.e., dense alternatives and sparse alternatives) have been explored in the high-dimensional hypothesis testing. For dense alternatives, the parameter space of interest is defined using the squared entry-wise ℓ_2 norm, that is, $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2$ and $\mathrm{tr}\{(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2)^2\}$, and the distributions under H_0 and H_1 are hard to distinguish when the nonzero entries of $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2$ are of about the same size in the absolute value (Chen and Qin 2010; Li and Chen 2012). For sparse alternatives, the parameter space of interest is defined using the entry-wise maximum norm, that

CONTACT Danning Li lidn040@nenu.edu.cn School of Mathematics and Statistics and KLAS, Northeast Normal University, Changchun, China; Lingzhou Xue lzxue@psu.edu Department of Statistics, The Pennsylvania State University, University Park, PA.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

is, $\max_{1\le i\le p}(\mu_{1i}-\mu_{2i})^2$ and $\max_{1\le i,j\le p}(\sigma_{1,ij}-\sigma_{2,ij})^2$, and the distributions under H_0 and H_1 are hard to distinguish when there are only a few large nonzero entries of $\boldsymbol{\mu}_1-\boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}_1-\boldsymbol{\Sigma}_2$ (Arias-Castro, Candès, and Plan 2011; Cai, Liu, and Xia 2013). The mathematical definitions of dense alternatives and sparse alternatives will be presented in Section 3.

In the literature, there only exist a few works on jointly testing means and covariances. In the classical setting with a fixed dimension p, the likelihood ratio test (LRT) was extensively studied in the multivariate analysis (Anderson 2003) when the samples come from normal distributions. When p diverges proportionally as the sample size tends to infinity such that $p/\min\{n_1,n_2\}\to c$ for some $0<c\le 1$, Jiang and Yang (2013) studied the modified LRTs under the normal assumption and derived central limit theorems. The normal assumption was recently relaxed by Niu et al. (2019). To allow p to diverge at a comparable rate as the sample size tends to infinity, that is, $0<c<\infty$, Liu et al. (2017) proposed a new approach by replacing the entropy loss with the quadratic loss for covariance matrix estimation. Hyodo and Nishiyama (2018) proposed a new joint test using a weighted sum of multiple U-statistics to allow p to diverge faster than the sample size.

However, most existing testing procedures only allow for a moderately high dimension in the asymptotic regime such that the dimension diverges at a slower rate than the sample size. Also, these existing testing procedures are mainly based on the modified LRTs or the L_2-norm-based test. Like quadratic-form tests, they perform well against dense alternatives but perform poorly against sparse alternatives (Fan, Liao, and Yao 2015; Li and Xue 2015; Yu, Yao, and Xue 2019; Yu, Li, and Xue 2020). These tests suffer from power loss in detecting sparse signals, as the errors in estimating high-dimensional parameters accumulate (Fan, Liao, and Yao 2015). Moreover, these joint testing procedures are essentially based on a weighted sum of one test statistic related to the mean difference and another test statistic related to the covariance difference. The weighted sum is not an ideal combination due to potentially different scales of two test statistics. These tests could be driven by the test statistic of a larger scale, leading to undesired power loss in the corresponding alternative space (Xie, Singh, and Strawderman 2011).

This article aims to develop a new power-enhanced simultaneous testing procedure that is powerful to detect differences in either mean vectors or covariance structure against either sparse alternatives or dense alternatives under a high-dimensional setting. Fan, Liao, and Yao (2015) introduced the power enhancement framework for high-dimensional hypothesis testing, which consists of the following *power enhancement (PE) principles*: (a) no size distortion; (b) the power-enhanced test is at least as powerful as the original test; (c) the power is substantially enhanced under a more general alternative. In this work, we interpret the more general alternatives from the following two perspectives:

(a) expanding the high-power regions of mean tests or covariance tests to a wider alternative space, respectively. We aim to develop the power-enhanced tests against the union of their corresponding dense and sparse alternatives.

(b) extending the test capability to alternative spaces with respect to both mean vectors and covariance matrices. We aim to combine strengths of two power-enhanced tests and develop a joint test that is capable of detecting the difference from either mean vectors or covariance matrices.

To expand the high-power regions, we construct power-enhanced tests for mean vectors and covariance matrices separately. We revisit the test statistics of Chen and Qin (2010) and Li and Chen (2012) that are constructed based on the estimators of the squared Euclidean distance of two sample mean vectors and the squared Frobenius distance of two sample covariance matrices, respectively. It is known that they are powerful to detect dense signals but unable to detect sparse signals (Chen, Li, and Zhong 2019). We introduce their respective PE components to effectively enlarge the high-power regions to the union of sparse and dense alternatives. We show that the proposed power-enhanced tests satisfy three desired PE principles. It is worth pointing out that we need new ideas to deal with a more challenging setting than that in Fan, Liao, and Yao (2015). The mechanism of enhancing test power via PE components is to add a constructed component to an asymptotically pivotal statistic, so that the resultant testing power is strengthened upon the original test. The construction of PE components relies on a screening over the marginal test statistics. Fan, Liao, and Yao (2015) employs a quadratic-form OLS-based statistic, whose marginal distributions are asymptotically normal. However, Chen and Qin (2010) and Li and Chen (2012) use degenerate U-statistics, and the distributions of their marginal test statistics are no longer asymptotically normal but rather a χ^2 distribution under the null hypothesis. The asymmetrically distributed marginal statistics require additional attention in the design of PE components. To the best of our knowledge, this is the first work that constructs PE components based on degenerate U-statistics.

After expanding the high-power regions, we aim to combine their strengths to develop the power-enhanced simultaneous test to further enhance the test capability for jointly testing mean vectors and covariance matrices. We prove the asymptotic independence of two PE test statistics and then aggregate information from the two aspects via the combination of their respective p-values using Fisher's method (Fisher 1925). We also show that the proposed power-enhanced simultaneous test satisfies three PE principles. It is important to note that, unlike Fan, Liao, and Yao (2015), Li and Xue (2015); Yu, Yao, and Xue (2019), and Yu, Li, and Xue (2020), we do not require the stringent normal assumption or independent assumption when deriving the asymptotic independence result. Compared with Fan, Liao, and Yao (2015) and Li and Xue (2015), our proposed test is scale-invariant and computationally efficient.

We study the theoretical properties under an ultra-high dimensional setting where the dimension may grow at a nearly exponential rate of the sample size. Moreover, we conduct simulation studies to compare the proposed test's numerical performance against several benchmark tests under various alternatives. In a real application, we further demonstrate the power of the proposed test to find differentially expressed gene-sets using an acute lymphoblastic leukemia dataset. Our findings are supported by the biological literature.

The rest of this article is organized as follows. Section 2 presents the preliminaries, and Sections 3 and 4 include the complete methodological details. Theoretical properties, including the power enhancement properties, the asymptotic size and power analysis as well as the asymptotic optimality, are also established in these two sections. Section 5 conducts simulation studies to demonstrate the finite-sample properties under different alternative hypotheses. Section 6 presents an empirical study on identifying differentially expressed gene-sets among various types of cancers. Section 7 includes a few concluding remarks. All technical details are presented in the supplementary materials.

2. Preliminaries

Let $\mathbf{X}$ be a p-dimensional random vector with mean $\boldsymbol{\mu}_1 = (\mu_{11}, \ldots, \mu_{1p})'$ and covariance $\boldsymbol{\Sigma}_1 = \left(\sigma_{1,ij}\right)_{p\times p}$, and $\mathbf{Y}$ be a p-dimensional random vector with mean $\boldsymbol{\mu}_2 = (\mu_{21}, \ldots, \mu_{2p})'$ and covariance $\boldsymbol{\Sigma}_2 = \left(\sigma_{2,ij}\right)_{p\times p}$. Suppose that $\{\mathbf{X}_1, \ldots, \mathbf{X}_{n_1}\}$ are iid copies of $\mathbf{X}$, and $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_{n_2}\}$ are iid copies of $\mathbf{Y}$ that are independent of $\{\mathbf{X}_1, \ldots, \mathbf{X}_{n_1}\}$. Now, we consider the high-dimensional mean test

$$H_{0m} : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2, \tag{2.1}$$

and the high-dimensional covariance test

$$H_{0c} : \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2, \tag{2.2}$$

respectively. Chen and Qin (2010) proposed the following quadratic-form statistic M_{n_1,n_2} for testing if the two high-dimensional populations share the same mean vector in (2.1):

$$M_{n_1,n_2} = \frac{1}{n_1(n_1-1)}\sum_{u\neq v}^{n_1}\left(\mathbf{X}_u'\mathbf{X}_v\right) + \frac{1}{n_2(n_2-1)}\sum_{u\neq v}^{n_2}\left(\mathbf{Y}_u'\mathbf{Y}_v\right) - \frac{2}{n_1 n_2}\sum_{u}^{n_1}\sum_{v}^{n_2}\left(\mathbf{X}_u'\mathbf{Y}_v\right). \tag{2.3}$$

To test the equality of two covariance matrices in (2.2), Li and Chen (2012) constructed their test statistic based on the squared Frobenius norm of $\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2$. Since $\|\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2\|_F^2 = \mathrm{tr}\left((\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2)^2\right) = \mathrm{tr}(\boldsymbol{\Sigma}_1^2) + \mathrm{tr}(\boldsymbol{\Sigma}_2^2) - 2\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)$, they proposed a test statistic T_{n_1,n_2} in the form of linear combination of unbiased estimators for each term, that is,

$$T_{n_1,n_2} = A_{n_1} + B_{n_2} - 2C_{n_1,n_2}, \tag{2.4}$$

where A_{n_1}, B_{n_2}, and C_{n_1,n_2} are the unbiased estimators for $\mathrm{tr}(\boldsymbol{\Sigma}_1^2)$, $\mathrm{tr}(\boldsymbol{\Sigma}_2^2)$, and $\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)$, respectively. The following assumptions are discussed in Chen and Qin (2010) and Li and Chen (2012) to establish the asymptotic properties of two test statistics M_{n_1,n_2} and T_{n_1,n_2}.

Assumption 1. For any $i, j, k, l \in \{1, 2\}$, as $n_1, n_2, p \to \infty$,

$$\mathrm{tr}\left(\boldsymbol{\Sigma}_k\boldsymbol{\Sigma}_l\right) \to \infty, \quad \mathrm{tr}\left\{\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}_j\boldsymbol{\Sigma}_k\boldsymbol{\Sigma}_l\right\} = o\left\{\mathrm{tr}\left(\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}_j\right)\mathrm{tr}\left(\boldsymbol{\Sigma}_k\boldsymbol{\Sigma}_l\right)\right\}. \tag{2.5}$$

Theoretically, if we consider a simple case that $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$, the condition (2.5) reduces to $\mathrm{tr}(\boldsymbol{\Sigma}^4) = o(\mathrm{tr}^2(\boldsymbol{\Sigma}^2))$, which holds when $\lambda_{\max}^2 = o(\mathrm{tr}(\boldsymbol{\Sigma}^2))$. When the smallest eigenvalue is larger than 0, the condition allows $\lambda_{\max}$ to diverge as long as $\lambda_{\max}^2 = o(p)$.

Assumption 2. The random vectors $\{\mathbf{X}_u\}_{u=1}^{n_1}$, $\{\mathbf{Y}_v\}_{v=1}^{n_2}$ satisfy

$$\mathbf{X}_u = \boldsymbol{\Gamma}_1\mathbf{Z}_{1u} + \boldsymbol{\mu}_1,\ \mathbf{Y}_v = \boldsymbol{\Gamma}_2\mathbf{Z}_{2v} + \boldsymbol{\mu}_2 \quad 1 \le u \le n_1, 1 \le v \le n_2, \tag{2.6}$$

where $\boldsymbol{\Gamma}_i = (\boldsymbol{\gamma}_{i1}, \ldots, \boldsymbol{\gamma}_{ip})'$ is a $p \times m_i$ matrix for some $m_i \ge p$ such that $\boldsymbol{\Gamma}_i\boldsymbol{\Gamma}_i' = \boldsymbol{\Sigma}_i$ for $i = 1, 2$, and $\{\mathbf{Z}_{ij}\}_{j=1}^{n_i} = \{(z_{ij1}, \ldots, z_{ijm_i})'\}_{j=1}^{n_i} \in \mathbb{R}^{m_i}$ are iid random vectors such that for any positive integers q and α_l such that $\sum_{l=1}^{q}\alpha_l \le 8$, and for any $1 \le k_1 \ne k_2 \ne \cdots \ne k_q \le m_i$,

$$\begin{aligned} &E(z_{ijk}) = 0,\ \mathrm{var}(z_{ijk}) = 1,\ \mathrm{cov}(z_{ijk_1}, z_{ijk_2}) = 0, \\ &E(z_{ijk}^4) = 3 + \Delta_i,\ E(z_{ijk}^8) < \infty, \end{aligned} \tag{2.7}$$

and

$$E(z_{ijk_1}^{\alpha_1} z_{ijk_2}^{\alpha_2} \ldots z_{ijk_q}^{\alpha_q}) = E(z_{ijk_1}^{\alpha_1})E(z_{ijk_2}^{\alpha_2}) \ldots E(z_{ijk_q}^{\alpha_q}). \tag{2.8}$$

Note that (2.6) expresses the samples using a factor-model structure, and (2.7) spells the moment conditions needed for the factors z_{ijk}, in which the Δ_i measures the fourth-moment difference compared to a standard normal distribution. (2.8) depicts a pseudo-independent pattern among its components for each $\mathbf{Z}_{ij}$. The condition is satisfied if $\mathbf{Z}_{ij}$ does have independent structure.

Under the null hypothesis H_{0m}, Chen and Qin (2010) considered the standardized test statistic $M_{n_1,n_2}/\widehat{\sigma}_{01}$ and proved that,

$$\text{under } H_{0m}: \quad \frac{M_{n_1,n_2}}{\widehat{\sigma}_{01}} \xrightarrow{d} N(0,1) \quad \text{as } n_1, n_2, p \to \infty, \tag{2.9}$$

where $\widehat{\sigma}_{01}$ is a consistent estimator of $\sigma_{01} = (\frac{2}{n_1(n_1-1)}\mathrm{tr}(\boldsymbol{\Sigma}_1^2) + \frac{2}{n_2(n_2-1)}\mathrm{tr}(\boldsymbol{\Sigma}_2^2) + \frac{4}{n_1 n_2}\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2))^{\frac{1}{2}}$, which is the standard deviation of M_{n_1,n_2} under H_{0m}. The test rejects H_{0m} with significance level α if $M_{n_1,n_2} \ge \widehat{\sigma}_{01}z_\alpha$, where z_α is the upper α-quantile of standard normal distribution.

Under the null hypothesis H_{0c}, we note that the leading variance of T_{n_1,n_2} is $\sigma_{02}^2 = 4\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^2 \mathrm{tr}^2\left(\boldsymbol{\Sigma}^2\right)$. With $\widehat{\sigma}_{02}$ being a consistent estimator of σ_{02}, Li and Chen (2012) conducted the test for H_{0c} on the basis of the test statistic $T_{n_1,n_2}/\widehat{\sigma}_{02}$ and proved that,

$$\text{under } H_{0c}: \quad \frac{T_{n_1,n_2}}{\widehat{\sigma}_{02}} \xrightarrow{d} N(0,1) \quad \text{as } n_1, n_2, p \to \infty. \tag{2.10}$$

The test rejects H_{0c} with a nominal significance level α if $T_{n_1,n_2} \ge \widehat{\sigma}_{02}z_\alpha$.

In the sequel, we will present our proposed power-enhanced simultaneous test on jointly testing means and covariances in high dimensions. In Section 3, we propose power-enhanced tests for the mean test and the covariance test, respectively, to boost their respective power. In Section 4, anchored in these two power-enhanced test statistics, we study their asymptotic joint distribution and subsequently introduce our simultaneous test to expand the test capability for jointly testing high-dimensional mean vectors and covariance matrices.

3. Power-Enhanced Tests

Both M_{n_1,n_2} and T_{n_1,n_2} are quadratic-form statistics. It has been known that such type of statistics suffer from low power against sparse alternatives where the parameter of interest differs only in

a small proportion of coordinates. One predominant approach to achieve high testing power against sparse alternatives is to use extreme values to construct test statistics (Cai, Liu, and Xia 2013; Chernozhukov, Chetverikov, and Kato 2019), whereas another way continues with the quadratic-form statistics but rules out nonsignal bearing dimensions via thresholding (Fan 1996; Chen, Li, and Zhong 2019; Chen, Guo, and Qiu 2019). However, these tests generally require either stringent conditions or bootstrap to derive the limiting null distribution and are likely to suffer from size distortions due to slow convergence. Also, even though the extreme value tests and thresholding tests retain high power against sparse alternatives, they tend to lack the ability to detect dense and faint signals, in which circumstances the quadratic-form tests are favored.

To deal with the challenge mentioned above, we first explore power enhancement for testing high-dimensional mean vectors and covariance matrices based on $M_{n_1n_2}$ and T_{n_1,n_2}, respectively. Fan, Liao, and Yao (2015) provides a helpful insight for us to enhance testing power against sparse alternatives and preserve the merits of existing quadratic-form tests at the same time. We construct two PE components J_m and J_c, which are designed to take zero values under the null hypothesis but diverge quickly under sparse alternatives. The PE components are designed delicately following the guidance of the three PE principles. By adding the PE components to the original statistics, the resultant tests $\widehat{\sigma}_{01}^{-1}M_{n_1,n_2}+J_m$ and $\widehat{\sigma}_{02}^{-1}T_{n_1,n_2}+J_c$ acquire substantially enhanced power under sparse alternatives with little size distortion under the null hypothesis.

Different from Fan, Liao, and Yao (2015), the distributions of our marginal test statistics are no longer asymptotically normal under H_0. To be more specific, Fan, Liao, and Yao (2015) uses a quadratic-form OLS-based statistic to test the significance of the intercept in multi-factor pricing models. For each coordinate, the marginal test statistic asymptotically follows a standard normal distribution. Yet here, M_{n_1,n_2} and T_{n_1,n_2} are degenerate U-statistics. Under the null hypothesis, their marginal statistics are no longer asymptotically normal, causing difficulties in designing the PE components. In specific, the PE component is usually constructed using a screening technique. A properly chosen threshold is critical to capture the signal-bearing dimensions while exclude nonsignal-bearing dimensions effected by estimation noise. The choice of such threshold is straightforward for the well-known normal distribution but requires additional efforts for nonnormal distributions. After careful investigation, we prove that the marginal standardized statistics follow Chi-squared distributions. To overcome the challenge brought by these asymmetrically distributed marginal statistics, we control the tail probabilities using a generalized result (Petrov 1954) of Cramér's limiting theorem, and choose the thresholds accordingly.

Let $n=n_1+n_2$ and δ_p and η_p be the thresholds chosen for the mean statistics and covariances statistics, respectively. We choose J_m and J_c to be the sum of marginal standardized statistics whose values exceed δ_p and η_p. By construction, the screening procedure rules out all the noises under the null hypothesis. Still, it makes it capable of capturing nonzero signals under sparse alternatives, implying that J_m and J_c equal to zero under the null hypothesis but diverge quickly under the sparse alternatives.

3.1. Power-Enhanced Mean Tests

We use $\mathbf{X}=(X_1,\ldots,X_p)'$ and $\mathbf{Y}=(Y_1,\ldots,Y_p)'$ to denote the random vectors of interest. Let $\mathbf{X}_u=(X_{u1},\ldots,X_{up})'$ and $\mathbf{Y}_v=(Y_{v1},\ldots,Y_{vp})'$ be the corresponding random samples. We rewrite the statistic M_{n_1,n_2} into $M_{n_1,n_2}=\sum_{i=1}^p M_i$, where

$$M_i=\frac{1}{n_1(n_1-1)}\sum_{u\neq v}^{n_1}(X_{ui}X_{vi})+\frac{1}{n_2(n_2-1)}\sum_{u\neq v}^{n_2}(Y_{ui}Y_{vi})-\frac{2}{n_1n_2}\sum_{u}^{n_1}\sum_{v}^{n_2}(X_{ui}Y_{vi}).$$

For each $i=1,\ldots,p$, M_i consistently estimates $(\mu_{1i}-\mu_{2i})^2$ as $n_1,n_2\to\infty$. Under the null hypothesis $H_{0m}:\boldsymbol{\mu}_1=\boldsymbol{\mu}_2$, the variance of M_i is

$$v_i:=\frac{2}{n_1(n_1-1)}\sigma_{1,ii}^2+\frac{2}{n_2(n_2-1)}\sigma_{2,ii}^2+\frac{4}{n_1n_2}\sigma_{1,ii}\sigma_{2,ii},$$

which can be consistently estimated by $\widehat{v}_i:=\frac{2}{n_1(n_1-1)}\widehat{\sigma}_{1,ii}^2+\frac{2}{n_2(n_2-1)}\widehat{\sigma}_{2,ii}^2+\frac{4}{n_1n_2}\widehat{\sigma}_{1,ii}\widehat{\sigma}_{2,ii}$, with $\widehat{\sigma}_{1,ii}$ and $\widehat{\sigma}_{2,ii}$ being sample variances of X_i and Y_i, respectively. Define

$$J_m=\sqrt{p}\sum_{i=1}^{p}M_i\widehat{v}_i^{-1/2}\mathcal{I}\{\sqrt{2}M_i\widehat{v}_i^{-1/2}+1>\delta_p\} \tag{3.1}$$

with $\delta_p=2\log p$ as the power enhancement component for the mean test. The theoretical analysis regarding J_m is established upon $\delta_p=2\log p$. In practical implementations, we follow Fan, Liao, and Yao (2015) to choose a slightly larger thresholding value, specifically $\delta_{p,n}=2\log p\log\log n$, to mitigate finite-sample biases.

In what follows, we present some theoretical properties of the constructed PE component J_m as well as the proposed power-enhanced mean test. To ensure that adding the PE component does not bring in size distortion, Fan, Liao, and Yao (2015) assumes the errors in a regression model follow a normal distribution. Benefiting from the usage of concentration inequalities to analyze the tail probabilities for degenerate U-statistics, we only assume the distributions of both populations are sub-Gaussian.

Assumption 3. There exists a positive constant H such that for all $h\in[-H,H]$,

$$Ee^{h(X_{ui}-\mu_{1i})^2}<\infty,\ Ee^{h(Y_{vi}-\mu_{2i})^2}<\infty\quad\text{for } i=1,\ldots,p. \tag{3.2}$$

The sub-Gaussianity assumption is imposed to control the tail probability of marginal statistics, ensuring the PE components equal to zero under the null hypothesis. This condition has also been assumed in relevant literature such as Chen, Guo, and Qiu (2019) and Chen, Li, and Zhong (2019).

Theorem 1. Suppose $n_1/(n_1+n_2)\to\gamma$ for some constant $\gamma\in(0,1)$ as $\min\{n_1,n_2\}\to\infty$ and $\log p=o(n^{1/3})$. Given Assumptions 1–3, under the null hypothesis $H_{0m}:\boldsymbol{\mu}_1=\boldsymbol{\mu}_2$, as $n_1,n_2,p\to\infty$,

$$P(J_m=0|H_{0m})\to1,\quad M_{PE}=\frac{1}{\widehat{\sigma}_{01}}\sum_{i=1}^{p}M_i+J_m\xrightarrow{d}N(0,1). \tag{3.3}$$

Theorem 1 proves that $J_m = 0$ holds under H_{0m} with probability tending to 1. Thus, adding J_m to the mean statistic $\widehat{\sigma}_{01}^{-1} M_{n_1,n_2}$ will not affect its limiting null distribution. The proposed power-enhanced mean test rejects H_{0m} with the significance level α if $M_{\mathrm{PE}} \geq z_\alpha$.

3.2. Power-Enhanced Covariance Tests

As for the covariance test statistic T_{n_1,n_2}, we first decompose T_{n_1,n_2} into

$$T_{n_1,n_2} = \sum_{i=1}^{p}\sum_{j=1}^{p} T_{ij} = \sum_{i=1}^{p}\sum_{j=1}^{p}(A_{ij} + B_{ij} - 2C_{ij}),$$

where

$$\begin{aligned}
A_{ij} &= \frac{1}{n_1(n_1-1)} \sum_{u\neq v}^{n_1} X_{ui}X_{vi}X_{uj}X_{vj} \\
&\quad - \frac{2}{n_1(n_1-1)(n_1-2)} \sum_{u\neq v\neq k}^{n_1} X_{ui}X_{vi}X_{vj}X_{kj} \\
&\quad + \frac{1}{n_1(n_1-1)(n_1-2)(n_1-3)} \sum_{u\neq v\neq k\neq l}^{n_1} X_{ui}X_{vi}X_{kj}X_{lj}, \\
B_{ij} &= \frac{1}{n_2(n_2-1)} \sum_{u\neq v}^{n_2} Y_{ui}Y_{vi}Y_{uj}Y_{vj} \\
&\quad - \frac{2}{n_2(n_2-1)(n_2-2)} \sum_{u\neq v\neq k}^{n_2} Y_{ui}Y_{vi}Y_{vj}Y_{kj} \\
&\quad + \frac{1}{n_2(n_2-1)(n_2-2)(n_2-3)} \sum_{u\neq v\neq k\neq l}^{n_2} Y_{ui}Y_{vi}Y_{kj}Y_{lj}, \\
C_{ij} &= \frac{1}{n_1 n_2} \sum_{u}^{n_1}\sum_{v}^{n_2} X_{ui}Y_{vi}X_{uj}Y_{vj} \\
&\quad - \frac{1}{n_1 n_2(n_1-1)} \sum_{u\neq k}^{n_1}\sum_{v}^{n_2} X_{ui}Y_{vi}Y_{vj}X_{kj} \\
&\quad - \frac{1}{n_1 n_2(n_2-1)} \sum_{u\neq k}^{n_2}\sum_{v}^{n_1} Y_{ui}X_{vi}X_{vj}Y_{kj} \\
&\quad + \frac{1}{n_1 n_2(n_1-1)(n_2-1)} \sum_{u\neq k}^{n_1}\sum_{v\neq l}^{n_2} X_{ui}Y_{vi}X_{kj}Y_{lj}.
\end{aligned}$$

The decomposition is essential to derive the PE component. For each $i,j = 1,\ldots,p$, T_{ij} consistently estimates the element-wise difference in covariances, that is, $T_{ij} \xrightarrow{p} (\sigma_{1,ij} - \sigma_{2,ij})^2$ as $n_1, n_2 \to \infty$. Under the null hypothesis $H_{0c} : \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, the variance of T_{ij} is

$$\begin{aligned}
\xi_{ij} := 2\Big(&\frac{1}{n_1}\left(\sigma_{1,ij}^2 + \sigma_{1,ii}\sigma_{1,jj} + \Delta_1 \mathrm{tr}(\boldsymbol{\gamma}_{1i}\boldsymbol{\gamma}_{1j}^T \circ \boldsymbol{\gamma}_{1i}\boldsymbol{\gamma}_{1j}^T)\right) \\
&+ \frac{1}{n_2}\left(\sigma_{2,ij}^2 + \sigma_{2,ii}\sigma_{2,jj} + \Delta_2 \mathrm{tr}(\boldsymbol{\gamma}_{2i}\boldsymbol{\gamma}_{2j}^T \circ \boldsymbol{\gamma}_{2i}\boldsymbol{\gamma}_{2j}^T)\right)\Big)^2 (1 + o(1)),
\end{aligned}$$

where $\boldsymbol{\gamma}_{ki}$ and Δ_k, $k = 1,2$ are defined by (2.6) and (2.7) in Assumption 2. In addition, we know that $\mathrm{var}((X_i - \mu_{1i})(X_j - \mu_{1j})) = \sigma_{1,ij}^2 + \sigma_{1,ii}\sigma_{1,jj} + \Delta_1 \mathrm{tr}(\boldsymbol{\gamma}_{1i}\boldsymbol{\gamma}_{1j}^T \circ \boldsymbol{\gamma}_{1i}\boldsymbol{\gamma}_{1j}^T)$, and analogously, $\mathrm{var}((Y_i - \mu_{2i})(Y_j - \mu_{2j})) = \sigma_{2,ij}^2 + \sigma_{2,ii}\sigma_{2,jj} + \Delta_2 \mathrm{tr}(\boldsymbol{\gamma}_{2i}\boldsymbol{\gamma}_{2j}^T \circ \boldsymbol{\gamma}_{2i}\boldsymbol{\gamma}_{2j}^T)$. Therefore, ξ_{ij} can be consistently estimated by

$$\begin{aligned}
\widehat{\xi}_{ij} := 2\Bigg(&\frac{1}{n_1^2}\sum_{u=1}^{n_1}\{(X_{ui} - \bar{X}_i)(X_{uj} - \bar{X}_j) - \widehat{\sigma}_{1,ij}\}^2 \\
&+ \frac{1}{n_2^2}\sum_{v=1}^{n_2}\{(Y_{vi} - \bar{Y}_i)(Y_{vj} - \bar{Y}_j) - \widehat{\sigma}_{2,ij}\}^2\Bigg)^2,
\end{aligned}$$

where $\bar{X}_j$ and $\bar{Y}_j$ are the sample mean of X_j and Y_j, $\widehat{\sigma}_{1,ij}$ and $\widehat{\sigma}_{2,ij}$ are sample covariances of (X_i, X_j) and (Y_i, Y_j), respectively. Define

$$J_c = \sqrt{p}\sum_{i=1}^{p}\sum_{j=1}^{p} T_{ij}\widehat{\xi}_{ij}^{-1/2}\mathcal{I}\{\sqrt{2}T_{ij}\widehat{\xi}_{ij}^{-1/2} + 1 > \eta_p\} \tag{3.4}$$

as the power enhancement component for the covariance test, with $\eta_p = 4\log p$. Similar to the previous section, the theoretical analysis regarding J_c is established upon $\eta_p = 4\log p$. In practical implementations, we use a slightly larger thresholding value, specifically $\eta_{p,n} = 4\log p \log\log n$, for the purpose of mitigating finite-sample biases.

Theorem 2. Suppose $n_1/(n_1+n_2) \to \gamma$ for some constant $\gamma \in (0,1)$ as $\min\{n_1,n_2\} \to \infty$ and $\log p = o(n^{1/5})$. Given Assumptions 1–3, under the null hypothesis $H_{0c} : \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, as $n_1, n_2, p \to \infty$,

$$P(J_c = 0 | H_{0c}) \to 1, \quad T_{\mathrm{PE}} = \frac{1}{\widehat{\sigma}_{02}}\sum_{i=1}^{p}\sum_{j=1}^{p} T_{ij} + J_c \xrightarrow{d} N(0,1). \tag{3.5}$$

Theorem 2 proves that under the null hypothesis H_{0c}, $J_c = 0$ with probability approaching 1. The power-enhanced covariance test rejects H_{0m} with significance level α if $T_{\mathrm{PE}} \geq z_\alpha$.

Remark 3.1. The thresholding values δ_p and η_p are chosen such that the power enhancement principles (Fan, Liao, and Yao 2015) are guaranteed. Please see the theoretical analyses that are presented in Sections S.2.4 and S.2.5 of the supplement for the details. Intuitively, δ_p can be regarded as a threshold chosen to control the tail behavior of p marginal χ_1^2 random variables, whereas η_p is used to control the tail behavior of p^2 marginal χ_1^2 random variables. Given the fact that $\log p^2 = 2\log p$, we can specify $\eta_p = 4\log p$ and $\delta_p = 2\log p$.

3.3. Power Enhancement Properties

In this section, we study the power enhancement properties of our proposed power-enhanced tests M_{PE} and T_{PE}. Chen and Qin (2010) and Li and Chen (2012) provided power analysis of the mean test statistic M_{n_1,n_2} and the covariance test statistic T_{n_1,n_2}, respectively. Consider the following parameter spaces $\mathcal{G}_m^d$ and $\mathcal{G}_c^d$ for their alternative hypotheses:

$$\begin{aligned}
\mathcal{G}_m^d = \Big\{&(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) : \min\{n_1,n_2\}\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2 \\
&/\sqrt{\max\{\mathrm{tr}(\boldsymbol{\Sigma}_1^2), \mathrm{tr}(\boldsymbol{\Sigma}_2^2)\}} \to \infty\Big\},
\end{aligned}$$

$$\mathcal{G}_c^d = \Big\{(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2) : \boldsymbol{\Sigma}_1 > 0, \boldsymbol{\Sigma}_2 > 0, \frac{1}{n_1}\mathrm{tr}(\boldsymbol{\Sigma}_1^2) + \frac{1}{n_2}\mathrm{tr}(\boldsymbol{\Sigma}_2^2) = o\left(\mathrm{tr}\{(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2)^2\}\right)\Big\}.$$

Chen and Qin (2010) pointed out that as $n_1, n_2, p \to \infty$, the mean test statistic M_{n_1,n_2} would correctly reject the null hypothesis H_{0m} with probability approaching 1 if the mean differences $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ fall into the subspace $\mathcal{G}_m^d$. Li and Chen (2012) drew analogous conclusions in regard to the covariance alternative space $\mathcal{G}_c^d$ corresponding to the covariance test T_{n_1,n_2}. More specifically, as $n_1, n_2, p \to \infty$,

$$\inf_{(\boldsymbol{\mu}_1,\boldsymbol{\mu}_2)\in\mathcal{G}_m^d} P\left(M_{n_1,n_2} \geq \widehat{\sigma}_{01} z_\alpha\right) \to 1 \text{ and } \inf_{(\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2)\in\mathcal{G}_c^d} P\left(T_{n_1,n_2} \geq \widehat{\sigma}_{02} z_\alpha\right) \to 1. \tag{3.6}$$

Note that $\mathcal{G}_m^d$ and $\mathcal{G}_c^d$ use the squared Euclidean-norm $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2$ and the squared Frobenius-norm $\|\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2\|_F^2$ to specify a large magnitude of differences in mean vectors and covariance matrices in order for the tests to be powerful in detecting the discrepancies.

In what follows, we present the power enhancement properties of our proposed tests. We will show that adding the power enhancement components J_m and J_c enables the tests to observe sparse signals which only differ in a few coordinates.

Theorem 3. Suppose $n_1/(n_1 + n_2) \to \gamma$ for some constant $\gamma \in (0, 1)$ as $\min\{n_1, n_2\} \to \infty$ and $\log p = o(n^{1/5})$. Given Assumptions 1–3, as $n_1, n_2, p \to \infty$, we have

$$\inf_{(\boldsymbol{\mu}_1,\boldsymbol{\mu}_2)\in\mathcal{G}_m^d\cup\mathcal{G}_m^s} P\left(M_{\mathrm{PE}} \geq z_\alpha\right) \to 1, \quad \text{and} \quad \inf_{(\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2)\in\mathcal{G}_c^d\cup\mathcal{G}_c^s} P\left(T_{\mathrm{PE}} \geq z_\alpha\right) \to 1,$$

with

$$\mathcal{G}_m^s = \Big\{(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) : \max_{1\leq i\leq p} \frac{(\mu_{1i} - \mu_{2i})^2}{v_i^{1/2}} \geq C\delta_p\Big\}$$

$$\mathcal{G}_c^s = \Big\{(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2) : \boldsymbol{\Sigma}_1 > 0, \boldsymbol{\Sigma}_2 > 0, \max_{1\leq i,j\leq p} \frac{(\sigma_{1,ij} - \sigma_{2,ij})^2}{\xi_{ij}^{1/2}} \geq C\eta_p\Big\}$$

where C is an absolute constant that does not depend on n_1, n_2 and p.

Theorem 3 shows that the power-enhanced tests have the same rejection regions as those of the original tests, but the high power regions are substantially expanded from $\mathcal{G}_m^d$ and $\mathcal{G}_s^d$ to $\mathcal{G}_m^d \cup \mathcal{G}_m^s$ and $\mathcal{G}_c^d \cup \mathcal{G}_c^s$, respectively.

Remark 3.2. Theorems 1–3 demonstrate that δ_p and η_p dominate the maximum noise level under the null hypothesis, and select signals under the designated alternatives. As long as n and p are not too small such that $\delta_p, \eta_p > 1$, which coincides with the high-dimensional framework, the theorems confirms the resultant power-enhanced mean test M_{PE} and power-enhanced covariance test M_{PE} satisfy the three PE principles introduced by Fan, Liao, and Yao (2015).

4. Power-Enhanced Simultaneous Test

Given two power-enhanced tests, we have boosted the respective power of testing mean vectors and covariance matrices. Before heading to the aggregation of information from the two aspects, we study the joint limiting distribution of the two statistics M_{PE} and T_{PE}.

We begin with some insights on the joint distributions for statistics of the two aspects. Suppose we have a random sample iid drawn from a univariate normal distribution, then it is well-known that the sample mean and sample variance are independent. To a slightly more complex case, suppose we have a random sample iid drawn from a multivariate normal distribution $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ in the traditional statistical settings when p is fixed. We look into two likelihood ratio test (LRT) statistics. Let Λ_1 be the LRT statistic for testing $H_0 : \{\boldsymbol{\mu} = \mathbf{0}, \boldsymbol{\Sigma} = \mathbf{I}_p\}$ versus $H_a : \{\boldsymbol{\mu} \in \mathbb{R}^p, \boldsymbol{\Sigma} = \mathbf{I}_p\}$, and let Λ_2 be the LRT statistic for testing $H_0 : \{\boldsymbol{\mu} \in \mathbb{R}^p, \boldsymbol{\Sigma} = \mathbf{I}_p\}$ versus $H_a : \{\boldsymbol{\mu} \in \mathbb{R}^p, \boldsymbol{\Sigma} > \mathbf{0}\}$. In multivariate statistics, we know that Λ_1 and Λ_2 are independent (Anderson 2003, Lemma 10.3.1).

The above discussion inspires us to conjecture on analogous propositions regarding the joint distribution of M_{PE} and T_{PE}. As a matter of fact, in the following theorem, we prove that the two statistics are indeed asymptotically independent.

Theorem 4. Suppose $n_1/(n_1 + n_2) \to \gamma$ for some constant $\gamma \in (0, 1)$ as $\min\{n_1, n_2\} \to \infty$ and $\log p = o(n^{1/5})$. Given Assumptions 1–3, under H_0, for any $x_1, x_2 \in \mathbb{R}$, as $n_1, n_2, p \to \infty$,

$$P\left(M_{\mathrm{PE}} \leq x_1, T_{\mathrm{PE}} \leq x_2\right) \to \Phi(x_1)\Phi(x_2). \tag{4.1}$$

With the information of the two separate power-enhanced tests at hand, the next step is to reasonably aggregate the results for testing means and covariances simultaneously. Most existing works rely on the classical likelihood ratio test (Anderson 2003) and its variants (Jiang and Yang 2013; Liu et al. 2017; Niu et al. 2019). Their test statistics are in the form of a summation of two statistics, where one is designed for detecting discrepancies in covariance matrices, and the other is to catch signals of distinct mean vectors. We call this type of combined statistic as the weighted sum statistics (Li and Xue 2015; Li, Xue, and Zou 2018).

However, the weighted sum statistics bear some drawbacks. The two components are usually of different magnitudes. The combined test would be mostly driven by the statistic with a larger scale, but insensitive to the statistic with a smaller scale. Such inefficiency in combination would lead to power loss in certain alternative spaces. Also, the distribution of the weighted sum statistic depends on the convolution of two marginal distributions, which is usually computationally challenging, resulting in difficulty in choosing critical value.

We propose a scale-invariant statistic to simultaneously test the equality of mean vectors and covariance matrices, by combining their separate p-values via Fisher's method:

$$J_{n_1,n_2} = -2\log(p_m) - 2\log(p_c), \tag{4.2}$$

where $p_m = 1 - \Phi(M_{\mathrm{PE}})$ and $p_c = 1 - \Phi(T_{\mathrm{PE}})$ are the p-values acquired from the power-enhanced mean test and the covariance test, respectively, and $\Phi(\cdot)$ is the cdf of $N(0, 1)$.

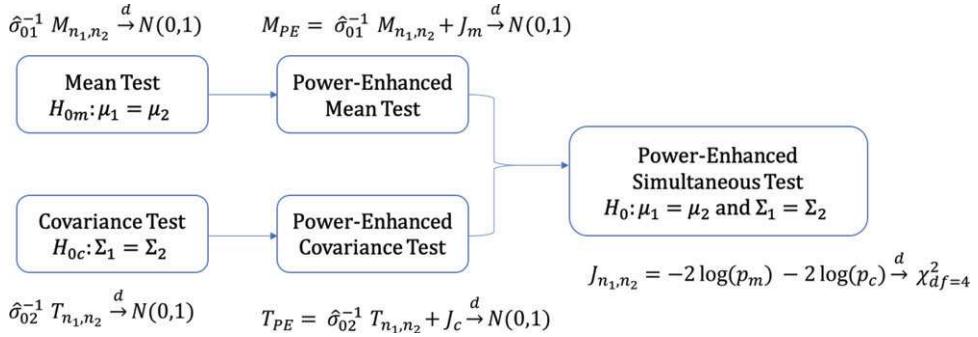

Figure 1. Power-enhanced simultaneous testing procedure.

As a matter of fact, the Fisher's method has been widely used in meta-analysis for combining the results of multiple scientific studies (Hedges and Olkin 2014). It is worth noticing that meta-analysis is designed for combining studies coming from independent sources. Yet combining two test statistics which are constructed from the same sample would be a different story, and therefore requires careful investigation on the independence assumption.

Theorem 4 proves that under the null hypothesis H_0, the two test statistics M_{PE} and T_{PE} are asymptotically independent. Hence, under H_0, p_m and p_c asymptotically independently follow a uniform distribution on the interval $[0, 1]$, and therefore $-2\log(p_m)$ and $-2\log(p_c)$ asymptotically independently follow a Chi-squared distribution with 2 degrees of freedom. As a result,

$$\text{under } H_0: \quad J_{n_1,n_2} \xrightarrow{d} \chi^2_4 \quad \text{as } n_1, n_2, p \to \infty. \tag{4.3}$$

Let q_α denote the upper-α quantile of χ^2_4 distribution, we reject the null hypothesis at the significance level α if

$$J_{n_1,n_2} \geq q_\alpha. \tag{4.4}$$

The procedures are summarized schematically in Figure 1. Equipped with the two key ingredients M_{PE} and T_{PE}, we proceed to investigate the size and power property of our proposed test J_{n_1,n_2} in Theorem 5. We show that our proposed test owns asymptotically accurate size approximation to the nominal significance level α and detects differences in either mean vectors or covariances over a wide range of alternatives.

Theorem 5 (Asymptotic Size and Power for Power-Enhanced Simultaneous Test). Suppose $n_1/(n_1+n_2) \to \gamma$ for some constant $\gamma \in (0,1)$ as $\min\{n_1, n_2\} \to \infty$ and $\log p = o(n^{1/5})$. Given Assumptions 1–3, as $n_1, n_2, p \to \infty$, the test J_{n_1,n_2} achieves (i) asymptotically accurate size, that is, under the null hypothesis $H_0: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, we have

$$P\left(J_{n_1,n_2} \geq q_\alpha\right) \to \alpha,$$

and (ii) asymptotically consistent power, specifically,

$$\inf_{\{(\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2)\in\mathcal{G}_c^d\cup\mathcal{G}_c^s\}\cup\{(\boldsymbol{\mu}_1,\boldsymbol{\mu}_2)\in\mathcal{G}_m^d\cup\mathcal{G}_m^s\}} P\left(J_{n_1,n_2} \geq q_\alpha\right) \to 1.$$

Remark 4.1. Theorem 5 confirms that our second PE procedure of expanding test capability from testing mean or covariances only to jointly testing mean vectors and covariance matrices satisfies the three PE principles.

There are other ways to aggregate information from the two aspects as the asymptotic independence permits the validity of many other combination methods. In what follows, we present two other tests using different methods to aggregate information to facilitate numerical comparison in the empirical studies. In Remark 4.2 and Section 5, we will show that the Fisher's combined test (4.4) outperforms other approaches as it is asymptotically optimal with respect to Bahadur efficiency.

One is a weighted statistics. Given the asymptotic independence, we may take the sum of squares of two statistics and transform the two asymptotic normal variables to an asymptotic χ^2_2 variable:

$$\text{under } H_0: \quad S_{n_1,n_2} = M^2_{\mathrm{PE}} + T^2_{\mathrm{PE}} \xrightarrow{d} \chi^2_2 \quad \text{as } n_1, n_2, p \to \infty. \tag{4.5}$$

The test rejects H_0 with a nominal significance level α if $S_{n_1,n_2} \geq c_\alpha$, where c_α is the upper-α quantile of χ^2_2 distribution. The other is an alternative p-value combination method. We consider the aggregation via Cauchy transformation (Liu and Xie 2020). The Cauchy combination is appealing for its insensitiveness of dependence between the statistics to be combined. In here, even though we obtain the asymptotic independence between M_{PE} and T_{PE}, we introduce the Cauchy combination test as a promising alternative. We define the Cauchy combination statistic as follows.

$$C_{n_1,n_2} = \frac{1}{2}\tan\left((0.5 - p_m)\pi\right) + \frac{1}{2}\tan\left((0.5 - p_c)\pi\right). \tag{4.6}$$

Under H_0, the asymptotic independence ensures that C_{n_1,n_2} converges to a standard Cauchy distribution as $n_1, n_2, p \to \infty$. The test rejects H_0 with a nominal significance level α if $C_{n_1,n_2} \geq k_\alpha$, where k_α is the upper-α quantile of standard Cauchy distribution.

Remark 4.2. Littell and Folks (1971, 1973) established the asymptotic optimality of Fisher's methods for combining independent tests in terms of the Bahadur slope. Singh, Xie, and Strawderman (2005) and Xie, Singh, and Strawderman (2011) further discussed such optimality within the framework of confidence distribution for meta-analysis. To combine the two p-values $p_m = 1 - \Phi(M_{\mathrm{PE}})$ and $p_c = 1 - \Phi(T_{\mathrm{PE}})$, the Fisher's method yields the largest exact Bahadur slope among all reasonable methods of combining independent tests, leading to the fastest decay rate of the p-values. Such results imply that to attain equal test power, the Fisher's combination test requires the smallest sample size. No other combining

method is superior to Fisher's method according to Bahadur relative efficiency. Therefore, J_{n_1,n_2} is asymptotically optimal with respect to Bahadur relative efficiency.

5. Simulation Studies

In this section, we conduct simulation studies to demonstrate the numerical performance of our proposed power-enhanced simultaneous test. To evaluate the power of the tests under different circumstances, we consider the following three types of alternative hypotheses: (a) H_m: $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$; (b) H_c: $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$, $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$; (c) H_b: $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$, $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$.

H_m describes the cases when the two populations share the same covariance matrix but have different means. H_c mimics the opposite situation in which the two populations have the same mean vector but differ in covariances. H_b considers the scenarios that there exist distinctions in both means and covariances among the two groups. For each alternative, we further consider two types of differences in the parameter of interest: the dense alternatives and the sparse alternatives. We use H_m^d and H_m^s to represent the existence of dense and sparse differences in $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$, and analogously, H_c^d and H_c^s to denote those in $\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2$.

We simulate our samples from the moving average structure shown below, so that we are able to accommodate the complex alternative hypotheses in a general data-generating process. For $i = 1, \ldots, p$, let

$$\begin{aligned} X_{u,i} &= \mu_{1,i} + Z_{u,i} + \theta_1 Z_{u,i+1}, \qquad u = 1, \ldots, n_1, \\ Y_{v,i} &= \mu_{2,i} + Z_{v+n_1,i} + \theta_2 Z_{v+n_1,i+1}, \quad v = 1, \ldots, n_2. \end{aligned} \tag{5.1}$$

In such a way, the parameters $\{\mu_{1,i}, \mu_{2,i}\}$ alter the mean vectors of our simulated samples $\{\mathbf{X}_u\}_{u=1}^{n_1}$ and $\{\mathbf{Y}_v\}_{v=1}^{n_2}$ to generate H_m^d and H_m^s, and $\{\theta_1, \theta_2\}$ control the covariance structure to account for H_c^d. By assigning different values to these parameters, we obtain simulated samples with various means and covariances. For the sparse alternatives with respect to covariance matrices H_c^s, we generate samples from a different approach by letting $\mathbf{X}_u = \boldsymbol{\Sigma}_1^{1/2}\mathbf{Z}_u + \boldsymbol{\mu}_1$, $\mathbf{Y}_v = \boldsymbol{\Sigma}_2^{1/2}\mathbf{Z}_{v+n_1} + \boldsymbol{\mu}_2$ for $u = 1, \ldots, n_1$, $v = 1, \ldots, n_2$, where $\mathbf{Z}_k = (Z_{k,1}, \ldots, Z_{k,p})'$, $k = 1, \ldots, n_1 + n_2$.

We first draw $\{Z_{k,i}\}_{1 \le k \le n_1+n_2, 1 \le i \le p+1}$ identically and independently from the standard normal $N(0,1)$. To check the robustness to nonnormally distributed data, we also generate the random data from the centralized Gamma(4, 2). We take the sample sizes as $n_1 = n_2 = N$ being 100 and 200, and let the dimension p take values in $\{100, 200, 500, 800, 1000\}$. For each setup, we compare our three proposed testing methods with four existing popular approaches: our proposed power-enhanced simultaneous test J_{n_1,n_2} as in (4.4), the proposed power-enhanced mean test M_{PE} as in (3.3), the proposed power-enhanced covariance test T_{PE} as in (3.5), the mean test M_{n_1,n_2} proposed by Chen and Qin (2010) as in (2.3), the covariance test T_{n_1,n_2} proposed by Li and Chen (2012) as in (2.4), the S_{n_1,n_2} approximation test as in (4.5), and the Cauchy combination test C_{n_1,n_2} as in (4.6). For each simulation setting, we report the frequencies of rejections over 5,000 replications with significance level $\alpha = 0.05$. We also compare the proposed power-enhanced tests M_{PE} and T_{PE} with those tests that are designed for sparse alternatives, specifically, the extreme-value-based tests proposed in Cai, Liu, and Xia (2013, 2014), see Section S.3.3 of the supplementary materials.

Table 1. Empirical size (%) with normal and gamma distributed $\{Z_{k,i}\}$ in the data generating process.

		Normal					Gamma				
N	Method	$p = 100$	200	500	800	1000	100	200	500	800	1000
100	M_{n_1,n_2}	5.24	5.24	5.12	5.06	5.32	5.10	4.72	5.00	5.04	5.10
	M_{PE}	5.96	5.84	5.36	5.46	5.48	5.64	5.18	5.32	5.36	5.34
	T_{n_1,n_2}	4.96	4.80	4.90	5.02	4.82	5.30	5.22	4.96	5.22	4.44
	T_{PE}	4.96	4.80	4.90	5.02	4.82	5.32	5.22	4.96	5.22	4.44
	S_{n_1,n_2}	5.70	5.84	5.98	5.12	5.40	5.92	5.60	5.34	5.10	4.84
	C_{n_1,n_2}	5.58	5.80	6.14	5.24	5.68	5.86	5.64	5.48	5.24	5.32
	J_{n_1,n_2}	5.60	5.56	5.22	5.42	5.12	5.58	5.56	5.16	5.54	5.06
200	M_{n_1,n_2}	5.48	5.30	5.46	5.16	5.22	4.94	5.06	5.06	5.24	5.26
	M_{PE}	5.68	5.56	5.62	5.18	5.30	5.34	5.32	5.18	5.32	5.34
	T_{n_1,n_2}	4.78	4.72	5.20	4.98	4.98	4.92	5.26	4.86	5.38	5.60
	T_{PE}	4.78	4.72	5.20	4.98	4.98	4.94	5.26	4.86	5.38	5.60
	S_{n_1,n_2}	5.14	4.80	5.46	5.22	5.46	5.64	5.22	5.18	5.56	5.54
	C_{n_1,n_2}	5.36	4.86	5.22	5.36	5.30	5.30	5.00	5.30	5.60	5.72
	J_{n_1,n_2}	5.50	5.12	5.24	5.30	5.08	5.50	5.22	5.56	5.54	5.54

NOTE: This table reports the frequencies of rejection by each method under the null hypothesis H_0 based on 5000 independent replications conducted at the significance level 5%.

To carry out $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, we set $\mu_{1i} = \mu_{2i} = 0$ for all $i = 1, \ldots, p$, and $\theta_1 = \theta_2 = 0$. Both samples are essentially iid from p-dimensional standard normal or multivariate gamma distribution. To evaluate the power, we fix $\{\mu_{1i}\}_{i=1}^p$ as zeros and $\theta_1 = 0$ for the first population, and vary $\{\mu_{2i}\}_{i=1}^p$ to set up the mean differences in H_m^d and H_m^s. As for the covariance alternatives, we change θ_2 to account for dense covariance differences in H_c^d and implement sparsely differed covariance matrix pair $(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2)$ to generate H_c^s.

As for H_m, we set $\theta_2 = 0$ to make sure the two samples share the same covariance matrix. In term of the mean vectors, for H_m^d, we follow Benjamini and Hochberg (1995) and consider a fixed percentage (pct) of violations in $\mu_{1,i} = \mu_{2,i}$ for $i = 1, \ldots, p$. The nonzero signal strength is determined in a similar fashion to Li and Chen (2012) as $\delta = \sqrt{\eta p^{-1/2}}$. To prevent trivial power of α and 1, we choose $\eta = 0.3$ and $pct = 15\%$. We set $\mu_{2,i} = \delta$ for $1 \le i \le [p \cdot pct]$ and zeros for the remaining ones. For sparse alternative H_m^s, we set the nonzero signal to be $\delta = 0.3\sqrt{\log p}$ and the number of nonzeros to be p^r with $r = 0.05$.

As for H_c, we ensure the two samples share equal means on every dimension. We set $\theta_2 = 0.2$ to create an MA(1) pattern of covariance as the dense alternative H_c^d. For the sparse alternative H_c^s, we follow Cai, Liu, and Xia (2013) to generate a symmetric sparse matrix $\mathbf{U}$ with eight random nonzero entries, each with a magnitude of $\delta = 0.3\sqrt{\log p^2}$. The locations of four nonzero entries are randomly selected from the upper triangle of $\mathbf{U}$ while the other four are specified by symmetry. Then we generate samples from $(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2)$ with $\boldsymbol{\Sigma}_1 = (1+\varepsilon)\mathbf{I}_p$ and $\boldsymbol{\Sigma}_2 = (1+\varepsilon)\mathbf{I}_p + \mathbf{U}$, where $\varepsilon = \left|\min\{\lambda_{\min}(\mathbf{U} + \mathbf{I}_p), 1\}\right| + 0.05$ is to make sure both $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are positive definite. Finally, with respect to H_b, we adopt the same idea as in H_m for the mean differences, and the same approach as in H_c for the covariance differences.

Table 1 presents the empirical size of the seven tests with Normal and Gamma distributed $\{Z_{k,i}\}$ in the data generating process (5.1). Tables 2–4 report the empirical power of the seven methods for testing H_m, H_c and H_b with normal distributed $\{Z_{k,i}\}$. We also carry out studies on the power analysis for Gamma distributed $\{Z_{k,i}\}$. The results show a similar pattern to the Gaus-

Table 2. Empirical power (%) against H_m and H_c with normal distributed $\{Z_{k,i}\}$ in the data-generating process.

		$N = 100$					$N = 200$				
H	Method	$p = 100$	200	500	800	1000	100	200	500	800	1000
H_m^d	M_{n_1,n_2}	47.30	44.94	47.00	46.64	46.52	87.04	88.92	90.76	91.54	91.32
	M_{PE}	48.64	45.92	47.52	46.88	46.88	87.34	89.10	90.86	91.58	91.34
	T_{n_1,n_2}	5.38	5.36	4.98	5.12	4.48	5.36	4.80	5.62	4.64	4.96
	T_{PE}	5.38	5.36	4.98	5.12	4.48	5.36	4.80	5.62	4.64	4.96
	S_{n_1,n_2}	34.06	30.16	30.32	29.32	28.44	75.18	76.74	77.74	79.46	79.26
	C_{n_1,n_2}	34.22	29.88	29.96	28.88	27.98	75.48	77.82	78.60	80.40	79.96
	J_{n_1,n_2}	39.36	37.48	36.78	36.80	36.72	80.32	82.20	83.32	84.46	84.64
H_m^s	M_{n_1,n_2}	42.24	33.38	54.22	44.68	17.74	82.24	72.48	94.46	88.60	40.12
	M_{PE}	79.00	79.54	96.30	95.58	78.98	99.22	99.68	100.00	99.98	99.74
	T_{n_1,n_2}	4.74	4.28	4.80	4.98	5.44	5.10	4.64	4.76	4.66	5.04
	T_{PE}	4.74	4.28	4.80	4.98	5.44	5.10	4.64	4.76	4.66	5.04
	S_{n_1,n_2}	76.80	78.10	95.66	95.22	77.88	99.18	99.66	100.00	99.98	99.74
	C_{n_1,n_2}	76.92	78.04	95.60	95.20	78.02	99.18	99.62	100.00	99.98	99.74
	J_{n_1,n_2}	77.68	78.92	95.86	95.50	78.50	99.18	99.64	100.00	99.98	99.74
H_c^d	M_{n_1,n_2}	4.80	4.92	5.12	4.84	5.10	5.10	5.00	5.26	5.18	5.26
	M_{PE}	5.50	5.60	5.38	5.26	5.48	5.48	5.26	5.36	5.32	5.40
	T_{n_1,n_2}	58.62	60.38	58.90	59.12	59.98	97.46	98.24	98.52	98.22	98.46
	T_{PE}	58.64	60.38	58.90	59.12	59.98	97.46	98.24	98.52	98.22	98.46
	S_{n_1,n_2}	37.22	37.52	36.12	36.52	38.16	91.80	92.62	93.94	92.90	93.16
	C_{n_1,n_2}	37.60	38.38	36.64	37.00	38.58	92.62	92.98	94.24	93.76	94.08
	J_{n_1,n_2}	45.20	46.52	46.02	45.94	47.24	94.34	94.68	95.62	95.50	95.64
H_c^s	M_{n_1,n_2}	5.12	4.98	5.14	5.28	5.22	5.08	5.38	5.40	4.90	5.16
	M_{PE}	5.76	5.44	5.38	5.54	5.44	5.56	5.60	5.46	5.02	5.18
	T_{n_1,n_2}	30.94	19.48	9.78	7.66	8.14	72.40	46.14	17.54	11.64	9.84
	T_{PE}	65.60	66.76	58.20	45.00	37.90	99.68	99.48	98.70	99.04	99.16
	S_{n_1,n_2}	60.00	63.74	57.06	44.02	36.72	99.60	99.44	98.80	98.98	99.16
	C_{n_1,n_2}	60.08	63.80	57.10	43.88	36.96	99.58	99.44	98.82	98.98	99.16
	J_{n_1,n_2}	62.46	65.58	57.92	44.82	37.54	99.62	99.44	98.74	98.98	99.18

NOTE: (a) This table reports the frequencies of rejection by each method under the alternative hypothesis based on 5000 independent replications conducted at the significance level 5%. (b) H_m^d and H_m^s stands for the type of alternative hypotheses (dense/sparse) in regards of the mean differences of the two populations. (c) H_c^d and H_c^s stands for the type of alternative hypotheses (dense/sparse) in regards of the covariance differences of the two populations.

Table 3. Empirical power (%) against H_b with normal distributed $\{Z_{k,i}\}$ in the data generating process.

H_b	N	Method	$p = 100$	200	500	800	1000
$H_m^d \cap H_c^d$	100	M_{n_1,n_2}	44.78	44.80	44.40	45.26	45.60
		M_{PE}	46.40	45.44	44.80	45.54	45.84
		T_{n_1,n_2}	57.80	58.70	58.94	59.16	59.78
		T_{PE}	57.80	58.70	58.94	59.16	59.78
		S_{n_1,n_2}	58.80	58.12	57.58	58.08	59.30
		C_{n_1,n_2}	57.22	56.56	55.40	55.76	56.72
		J_{n_1,n_2}	73.24	74.68	74.42	75.94	76.24
	200	M_{n_1,n_2}	84.44	85.58	87.92	89.20	89.22
		M_{PE}	84.86	85.74	88.02	89.26	89.22
		T_{n_1,n_2}	98.16	98.26	98.26	98.38	98.48
		T_{PE}	98.16	98.26	98.26	98.38	98.48
		S_{n_1,n_2}	98.84	99.12	99.38	99.24	99.30
		C_{n_1,n_2}	98.50	98.88	98.98	98.98	99.08
		J_{n_1,n_2}	99.64	99.88	99.88	99.88	99.90
$H_m^d \cap H_c^s$	100	M_{n_1,n_2}	38.90	42.42	41.68	39.08	38.40
		M_{PE}	40.52	43.12	42.16	39.42	38.70
		T_{n_1,n_2}	32.48	18.16	9.96	7.96	7.12
		T_{PE}	72.36	72.62	58.12	42.76	37.58
		S_{n_1,n_2}	75.42	77.76	65.86	52.62	48.62
		C_{n_1,n_2}	75.02	77.84	65.84	52.84	48.82
		J_{n_1,n_2}	81.34	81.98	70.56	59.42	54.56
	200	M_{n_1,n_2}	77.06	82.92	85.12	83.38	82.46
		M_{PE}	77.54	83.14	85.22	83.40	82.48
		T_{n_1,n_2}	74.24	46.52	17.68	11.54	10.06
		T_{PE}	99.82	99.46	98.84	98.94	99.00
		S_{n_1,n_2}	99.90	99.72	99.40	99.38	99.34
		C_{n_1,n_2}	99.90	99.72	99.42	99.36	99.34
		J_{n_1,n_2}	99.92	99.82	99.52	99.42	99.42

NOTE: (a) This table reports the frequencies of rejection by each method under the alternative hypothesis based on 5000 independent replications conducted at the significance level 5%. (b) H_b stands for the type of alternative hypotheses (dense/sparse) in regards of the mean and covariance differences of the two populations.

sian cases and are presented in Section 3.1 of the supplementary materials. Moreover, we examine the test performance in regard to data with two additional covariance structures, and the results are summarized in Section S.3.2 of the supplementary materials. These numerical comparisons provide us with the following findings:

1. Under H_0, all of the seven tests achieve reasonably accurate size approximation over a broad range of dimensionality. Besides, the empirical sizes with Gamma distribution illustrate that these tests are quite robust to non-Gaussianity.
2. The numerical results of the power-enhanced tests M_{PE} and T_{PE} echo with the power enhancement properties presented in Theorems 1 and 2. Table 1 reveals that adding power enhancement components does not inflate the testing size under the null hypothesis H_0. On the other hand, Tables 2–4 reflect that the testing power is substantially enhanced under sparse alternatives H_m^s and H_c^s.
3. As shown in Table 2, the mean tests (M_{n_1,n_2} and M_{PE}) are powerful in detecting mean differences as in H_m, but have almost no power in discovering the covariance differences under H_c. In contrast, the covariance tests (T_{n_1,n_2} and T_{PE}) perform well in declaring significance for covariance alternative H_c, however, it is powerless to identify the unequal means under H_m.
4. With respect to H_m and H_c, even though one of the M_{PE} test and T_{PE} test fails, the three combination tests remain powerful across all the experiments. This coincides with the power analysis shown in Theorem 5 that the combination of two tests makes use of their respective power under different alternatives, therefore, successfully discover the discrepancies in either mean vectors or covariance matrices.
5. Tables 3 and 4 illustrate that our proposed simultaneous test acquires additional gains when both mean differences and covariance differences exist. Under H_b, both M_{PE} test and and T_{PE} test successfully sense the differences in regard to the means and covariances, respectively. By combining the two tests together, our proposed approach yields to a higher testing power as it can simultaneously detect both types of differences.
6. What's more, for each simulation setting, the proposed test J_{n_1,n_2} prevails with higher power compared with the S_{n_1,n_2} and C_{n_1,n_2} tests. This finding resounds with the asymptotically optimal property discussed in Remark 4.2.

Additionally, Figure 2 provides a graphical representation of the testing power using seven approaches under H_b when both mean differences and covariances differences exist. We study four different hypotheses consisting of the combination of sparsely/densely differed means (H_m^s/H_m^d) and sparsely/densely differed covariances (H_c^s/H_c^d). The figure shows that the tests M_{n_1,n_2} and T_{n_1,n_2} favor dense alternatives H_m^d and H_c^d, respectively, because of its nature of quadratic forms, but lack the

Table 4. Empirical power (%) against H_b with normal distributed $\{Z_{k,i}\}$ in the data-generating process

H_b	N	Method	$p = 100$	200	500	800	1000
$H_m^s \cap H_c^d$	100	M_{n_1,n_2}	41.16	31.58	23.28	18.20	17.70
		M_{PE}	77.24	77.26	77.86	77.36	78.96
		T_{n_1,n_2}	58.18	59.96	58.80	58.26	58.78
		T_{PE}	58.20	59.96	58.80	58.26	58.78
		S_{n_1,n_2}	83.14	83.90	84.86	84.50	85.36
		C_{n_1,n_2}	83.14	84.06	84.98	84.64	85.72
		J_{n_1,n_2}	87.50	88.68	88.60	88.00	88.80
	200	M_{n_1,n_2}	81.52	69.82	51.78	41.72	38.26
		M_{PE}	99.10	99.46	99.60	99.74	99.74
		T_{n_1,n_2}	97.56	97.78	98.40	97.96	98.54
		T_{PE}	97.56	97.78	98.40	97.96	98.54
		S_{n_1,n_2}	99.96	99.96	100.00	99.98	99.98
		C_{n_1,n_2}	99.96	99.98	100.00	100.00	99.98
		J_{n_1,n_2}	99.98	100.00	100.00	100.00	99.98
$H_m^s \cap H_c^s$	100	M_{n_1,n_2}	34.84	29.34	20.34	16.94	15.04
		M_{PE}	66.12	71.62	70.74	67.76	68.60
		T_{n_1,n_2}	30.74	19.70	9.84	7.76	7.12
		T_{PE}	69.44	74.36	57.28	43.96	38.32
		S_{n_1,n_2}	85.56	89.86	84.80	78.64	78.14
		C_{n_1,n_2}	85.48	89.90	84.84	78.90	78.10
		J_{n_1,n_2}	87.30	91.14	86.12	79.82	79.44
	200	M_{n_1,n_2}	72.94	64.52	46.32	36.32	31.20
		M_{PE}	97.54	98.48	98.92	98.78	98.68
		T_{n_1,n_2}	74.22	46.46	17.76	11.72	10.94
		T_{PE}	99.72	99.36	99.00	98.86	98.88
		S_{n_1,n_2}	99.98	99.94	99.94	99.80	99.76
		C_{n_1,n_2}	99.98	99.94	99.94	99.84	99.76
		J_{n_1,n_2}	99.98	99.94	99.96	99.84	99.76

NOTE: (a) This table reports the frequencies of rejection by each method under the alternative hypothesis based on 5000 independent replications conducted at the significance level 5%. (b) H_b stands for the type of alternative hypotheses (dense/sparse) in regards of the mean and covariance differences of the two populations.

ability of detecting sparse alternatives such as H_m^s and H_c^s. Fortunately, the proposed power-enhanced tests M_{PE} and T_{PE} greatly promote the respective testing power under H_m^s and H_c^s. When it comes to jointly testing means and covariances, the plot clearly shows that our proposed Fisher's combined test J_{n_1,n_2} achieves the highest power among the three combination approaches (F_{n_1,n_2}, S_{n_1,n_2}, and C_{n_1,n_2}).

Moreover, we would like to discuss the effects of data characteristics (e.g., sample size n and data dimensionality p) on the test performance reflected by the empirical results. Using the M_{PE} test as an example, the convergence rate of M_{PE} is dominated by a leading term that depends on n, p, and the structure of $\mathbf{\Sigma}$. Hence, the size and power performance is related with data dimensionality in a complicated way. The theoretical results require both n and p go to infinity. The larger the p is, the better fit it is in regard to the asymptotic regime. It is likely the reason why when $n = 100$, $p = 500$ yields a more accurate size than that of $p = 100$ in Table 1. However, since the leading term also depends on the structure of $\mathbf{\Sigma}$, it is not always true that a larger p leads to a more accurate size or a more higher power. As for the sample size n, a larger n provides a better fit to the asymptotic regime. Yet we would like to point out that the main term of M_{PE} is a martingale and its convergence rate is slow with respect to n. When n is increased from 100 to 200, the convergence rate is not significantly improved. Together with the randomness in simulation studies, the empirical size with $n = 100$ is not necessarily closer to 5% compared with that of $n = 200$. In addition, the uncertainty due to randomness in simulation studies also has some impacts on the results of empirical size and empirical power.

In summary, the simulation results demonstrate the promising finite-sample performance of our proposed simultaneously tests, providing numerical evidence to verify the theoretical properties introduced in the previous section. Under the null hypothesis, the proposed test retains the desired nominal significance level. It is powerful in detecting either mean differences or covariance differences, and it remains high power for either sparse alternatives or dense alternatives. Moreover, the testing power is boosted against more general alternatives.

6. Application to Gene-Set Testing

This section demonstrates the power of our proposed tests through a real application on an Acute Lymphoblastic Leukemia (ALL) dataset from the Ritz Laboratory at the Dana-Farber Cancer Institute (DFCI). The data was originally published by Chiaretti et al. (2004) and is now available at the Bioconductor website. The ALL dataset contains gene expression levels of 12,625 probes on Affymetrix chip series HG-U95Av2 from 128 individuals with either T-cell ALL or B-cell ALL, depending on the type of lymphocyte for the leukemia cells. This study focuses on a subset of the ALL data for 79 patients with the B-cell ALL. We further divide the patients into two groups according to their B-cell tumors' subtypes: the BCR/ABL fusion and the cytogenetically normal NEG, whose sample sizes are 37 and 42, respectively.

Identifying differentially expressed gene-sets has received considerable attention in genetic studies (Efron and Tibshirani 2007; Goeman and Bühlmann 2007). Since each gene does not work individually but rather tend to function groups to achieve complex biological tasks, researchers look into gene expression profiles based on groups of genes depending on their functional characteristics. To make full use of prior biological knowledge, we group sets of genes according to their Gene Ontology (GO) annotations. The GO system describes the biological domains with respect to three aspects: biological process (BP), cellular component (CC), and molecular function (MF). We follow the same criteria to perform a prescreening procedure by excluding those probes with low fluorescence intensities and narrowly spread, characterized by small absolute values and small interquantile ranges. The filtering step retains 2391 probes, corresponding to 1849 unique GO terms in BP category, 306 in CC and 324 in MF.

Let $S_1, \ldots, S_K$ denote K gene-sets and $\{\boldsymbol{\mu}_{1S_k}, \mathbf{\Sigma}_{1S_k}\}$, $\{\boldsymbol{\mu}_{2S_k}, \mathbf{\Sigma}_{2S_k}\}$ be the mean vectors and covariance matrices of two types of tumors, respectively. We are interested in testing

$$H_{0,\mathrm{category}} : \boldsymbol{\mu}_{1S_k} = \boldsymbol{\mu}_{2S_k} \text{ and } \mathbf{\Sigma}_{1S_k} = \mathbf{\Sigma}_{2S_k}, \quad k = 1, \cdots, K$$

where category $\in \{BP, CC, MF\}$. We classify gene-sets into three different GO categories and shall test each GO category separately. Figure 3 plots the dimension of gene-sets contained in each category. The dimension of gene-sets in each category can be as large as two thousand, which is much larger than the sample sizes $n_1 = 37$ and $n_2 = 42$. To examine the test

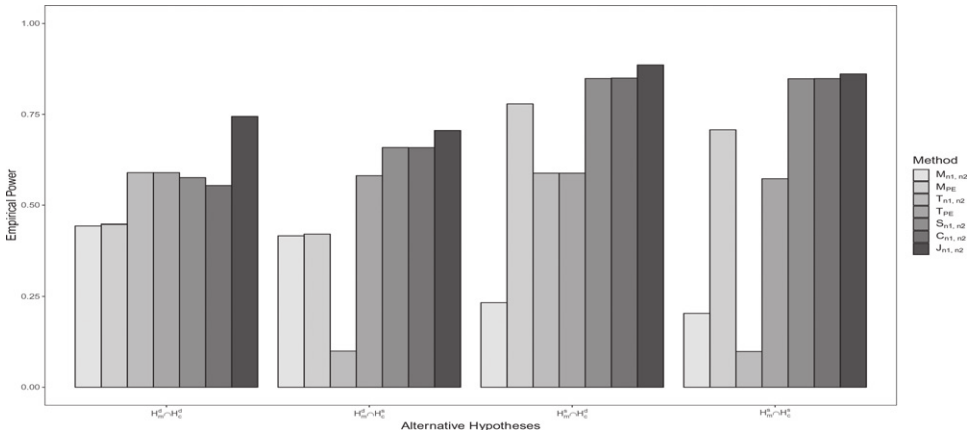

Figure 2. Empirical power comparison of the seven tests under H_b with normal distributed $\{Z_{k,i}\}$ and $N = 100, p = 500$.

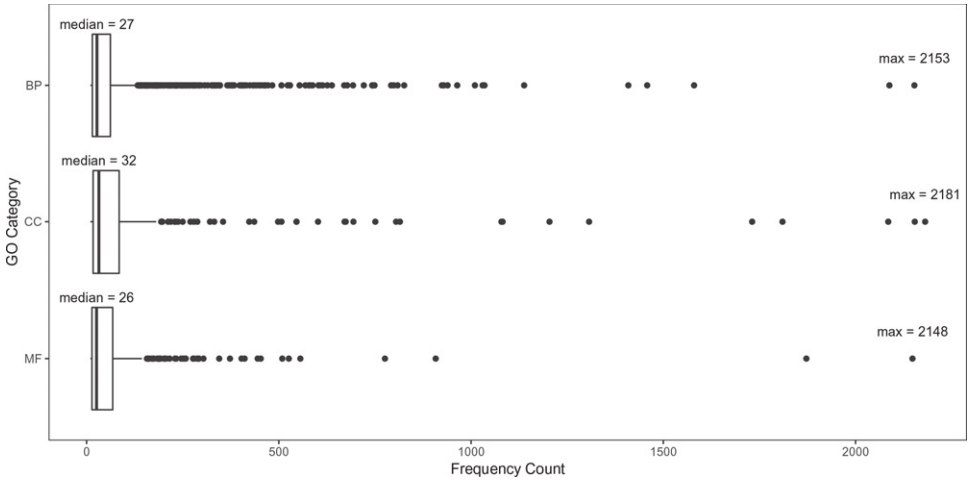

Figure 3. Boxplots of the dimension of gene-sets for three GO categories.

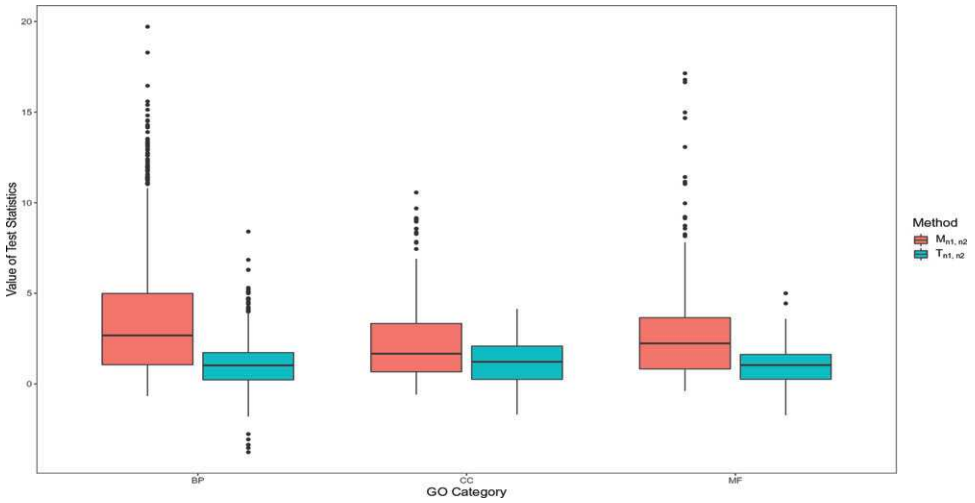

Figure 4. Boxplots of the $\widehat{\sigma}_{01}^{-1} M_{n_1,n_2}$ and $\widehat{\sigma}_{02}^{-1} T_{n_1,n_2}$ test statistics for three GO categories.

assumptions imposed on the covariance matrices, we compute the ratio of $\mathrm{tr}(\widehat{\mathbf{\Sigma}}_1^2\widehat{\mathbf{\Sigma}}_2^2)$ to $\mathrm{tr}(\widehat{\mathbf{\Sigma}}_1^2)\mathrm{tr}(\widehat{\mathbf{\Sigma}}_2^2)$ for the gene-sets, where $\widehat{\mathbf{\Sigma}}_1$ and $\widehat{\mathbf{\Sigma}}_2$ are sample covariances. We observe that the ratio is small for most gene-sets. The median of the ratio values is 0.240 and the 75th percentile is 0.335. There are 92% of gene-sets whose ratio is below 0.5.

Before proceeding, we explore the values of power-enhanced test statistics M_{n_1,n_2} and T_{n_1,n_2} for all gene-sets. Figure 4 presents boxplots of M_{n_1,n_2} and T_{n_1,n_2} within each GO category. The M_{n_1,n_2} statistics have relatively larger values compared with the T_{n_1,n_2} statistics. Recall that under the null hypothesis, both statistics converge to $N(0,1)$ in distribution. The finding that

Table 5. The number of significant gene-sets declared by different tests after BH control with nominal level $\alpha = 0.05$.

GO category		BP	CC	MF
Total number of gene-sets		1849	306	324
Number of Significant Gene-sets	M_{n_1,n_2}	1134	140	183
	M_{PE}	1469	216	236
	T_{n_1,n_2}	126	55	20
	T_{PE}	126	55	20
	S_{n_1,n_2}	1485	219	234
	C_{n_1,n_2}	1484	220	233
	J_{n_1,n_2}	1511	226	238

the M_{n_1,n_2} statistics have larger absolute values indicates that for these gene-sets, their mean vectors are more different compared to the covariance matrices between the two groups. Moreover, considering significance level $\alpha = 0.05$ and the upper α-quantile of $N(0,1)$ $z_\alpha = 1.645$, a large number of M_{n_1,n_2} statistics fall above the threshold z_α. Therefore, we would expect a lot of rejections for testing the equality of the mean vectors. The discussions in this paragraph give us an exploratory view of the dataset. Later on, we will present more precise comparisons among various test approaches.

We then apply our power-enhanced simultaneous test J_{n_1,n_2} to test the means and covariances simultaneously, together with the mean test M_{n_1,n_2}, the covariance test T_{n_1,n_2} and the two power-enhanced tests M_{PE} and T_{PE}. We compare our proposed approach J_{n_1,n_2} with the χ^2 approximation S_{n_1,n_2} as well as the Cauchy combination C_{n_1,n_2}. In order to control the false discovery rate (FDR), we apply the Benjamini–Hochberg (BH) procedure (Benjamini and Hochberg 1995) to each GO category. Table 5 reports the number of significant gene-sets declared by different tests with nominal level $\alpha = 0.05$ for every category.

As shown in Table 5, J_{n_1,n_2} identifies more significant gene-sets than the other methods. The M_{n_1,n_2} test declares a lot of significance whereas the T_{n_1,n_2} test only identifies a few. The M_{PE} identifies a few more differentially expressed gene-sets with respect to mean vectors, while the T_{PE} does not yield additional power in detecting the differ ences among covariances. This indicates there exist a large number of unequal means between the two types of tumors, but not much differences in their covariance patterns. This phenomenon emphasizes the importance of developing a powerful method for jointly testing the means and covariances, so that we have a better chance to detect differences between two distributions even though we are in lack of prior knowledge about whether the differences reside in means or covariances.

The χ^2 approximation S_{n_1,n_2} and the Cauchy combination C_{n_1,n_2} yield comparative performance. As shown in Table 5, the two methods identify more differences than the covariance test T_{PE}, yet potentially miss some differentially expressed gene-sets compared to the mean test M_{PE}. In contrast, our proposed J_{n_1,n_2} is able to identify more discrepancies between the two groups, compared to the other three combination approaches and also compared to the original means test as well as the covariance tests. In a short summary, our proposed Fisher's combined simultaneous test J_{n_1,n_2} benefits from incorporating the information from the mean tests and covariance tests, and outperforms other combination methods in detecting the significant differences among the gene-sets.

Next, we study those gene-sets which are declared significant only by J_{n_1,n_2} but not any other method. Among those, we pay special attention to the GO-term "GO:0005125" in the MF category. This gene-set contributes to cytokine activity, including interleukins which are a group of cytokines that regulate inflammatory and immune responses (Okada and Pollack 2004). Extensive scientific studies have revealed the close relationships between interleukins and leukemia (Touw et al. 1990; Paietta et al. 1997; Yoda et al. 2010; Canale et al. 2011). For another example, it is known microRNAs act complementarily to regulate disease-related mRNA modules in human diseases (Chavali et al. 2013). We observe that the expression levels of "GO:0006913" in the BP category are statistically different between the two groups. This GO-term refers to nucleocytoplasmic transport, whose association with leukemia has been validated by numerous cancer studies (Chavali et al. 2013; Gravina et al. 2014; Takeda and Yaseen 2014). The biological evidence suggests our power-enhanced simultaneous test J_{n_1,n_2} provides more useful information compared with other approaches, which further implies the importance of developing power-enhanced simultaneous tests.

7. Conclusion and Discussion

In this work, we study the problem of jointly testing the equality of two-sample mean vectors and covariance matrices of high-dimensional data. We introduce a new power-enhanced simultaneous test, and prove the test achieves accurate asymptotic size, enhanced and consistent asymptotic power under a more general alternative, and asymptotic optimality with respect to Bahadur efficiency. The proposed test is scale-invariant and computationally efficient. We demonstrate the finite-sample performance using simulation studies and a real application to gene-set testing. In our current setup, there are no structural assumptions imposed on the mean vectors and covariance matrices. In some applications, the mean vector or covariance matrix may admit some structure due to the nature of data, for example, the factor structure (Fan, Fan, and Lv 2008), and the banded structure (Bickel and Levina 2008). To boost the testing power of testing for parameters with certain structure, we may need first separate structural information and signals for the alternatives, and then construct the PE component upon the latter part. When the parameters naturally come with some kind of structure, developing power-enhanced tests which takes account of structural information would be an interesting and practically useful extension. We leave it for future work.

Supplementary Materials

The supplementary note consists of three sections to present lemmas, complete proofs of lemmas and theorems, and additional numerical results.

Acknowledgments

We would like to thank the Coeditor, Associate Editor and referees for their helpful comments and suggestions.

Funding

Xiufan Yu and Lingzhou Xue were partially supported by the National Science Foundation grants DMS-1811552 and DMS-1953189. Danning Li was partially supported by the National Natural Science Foundation of China grant 12101116. Runze Li was partially supported by the National Science Foundation grant DMS-1820702.

References

Anderson, T. (2003), *An Introduction to Multivariate Statistical Analysis*. Wiley Series in Probability and Statistics, New York: Wiley. [2548,2549,2553]

Arias-Castro, E., Candès, E. J., and Plan, Y. (2011), "Global Testing Under Sparse Alternatives: Anova, Multiple Comparisons and the Higher Criticism," *The Annals of Statistics*, 39, 2533–2556. [2549]

Benjamini, Y., and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society*, Series B, 57, 289–300. [2555,2559]

Bickel, P. J., and Levina, E. (2008), "Regularized Estimation of Large Covariance Matrices," *The Annals of Statistics*, 36, 199–227. [2559]

Cai, T., Liu, W., and Xia, Y. (2013), "Two-Sample Covariance Matrix Testing and Support Recovery in High-Dimensional and Sparse Settings," *Journal of the American Statistical Association*, 108, 265–277. [2549,2551,2555]

Cai, T. T., Liu, W., and Xia, Y. (2014), "Two-Sample Test of High Dimensional Means Under Dependence," *Journal of the Royal Statistical Society*, Series B, 76, 349–372. [2555]

Canale, S., Cocco, C., Frasson, C., Seganfreddo, E., Di Carlo, E., Ognio, E., Sorrentino, C., Ribatti, D., Zorzoli, A., and Basso, G. (2011), "Interleukin-27 Inhibits Pediatric B-acute Lymphoblastic Leukemia Cell Spreading in a Preclinical Model," *Leukemia*, 25, 1815–1824. [2559]

Chavali, S., Bruhn, S., Tiemann, K., Sætrom, P., Barrenäs, F., Saito, T., Kanduri, K., Wang, H., and Benson, M. (2013), "MicroRNAs Act Complementarily to Regulate Disease-Related mRNA Modules in Human Diseases," *RNA*, 19, 1552–1562. [2559]

Chen, S. X., Guo, B., and Qiu, Y. (2019), "Multi-Level Thresholding Test for High Dimensional Covariance Matrices," arXiv preprint arXiv:1910.13074. [2548,2551]

Chen, S. X., Li, J., and Zhong, P.-S. (2019), "Two-Sample and Anova Tests for High Dimensional Means," *The Annals of Statistics*, 47, 1443–1474. [2548,2549,2551]

Chen, S. X., and Qin, Y.-L. (2010), "A Two-Sample Test for High-Dimensional Data with Applications to Gene-Set Testing," *The Annals of Statistics*, 38, 808–835. [2548,2549,2550,2552,2553,2555]

Chernozhukov, V., Chetverikov, D., and Kato, K. (2019), "Inference on Causal and Structural Parameters Using Many Moment Inequalities," *The Review of Economic Studies*, 86, 1867–1900. [2551]

Chiaretti, S., Li, X., Gentleman, R., Vitale, A., Vignetti, M., Mandelli, F., Ritz, J., and Foa, R. (2004), "Gene Expression Profile of Adult t-cell Acute Lymphocytic Leukemia Identifies Distinct Subsets of Patients with Different Response to Therapy and Survival," *Blood*, 103, 2771–2778. [2557]

Clarke, R., Ressom, H. W., Wang, A., Xuan, J., Liu, M. C., Gehan, E. A., and Wang, Y. (2008), "The Properties of High-Dimensional Data Spaces: Implications for Exploring Gene and Protein Expression Data," *Nature Reviews Cancer*, 8, 37–49. [2548]

Cummings, J., Lee, G., Ritter, A., Sabbagh, M., and Zhong, K. (2019), "Alzheimer's Disease Drug Development Pipeline: 2019," *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 5, 272–293. [2548]

Efron, B., and Tibshirani, R. (2007), "On Testing the Significance of Sets of Genes," *The Annals of Applied Statistics*, 1, 107–129. [2557]

Fan, J. (1996), "Test of Significance Based on Wavelet Thresholding and Neyman's Truncation," *Journal of the American Statistical Association*, 91, 674–688. [2551]

Fan, J., Fan, Y., and Lv, J. (2008), "High Dimensional Covariance Matrix Estimation Using a Factor Model," *Journal of Econometrics*, 147, 186–197. [2559]

Fan, J., Liao, Y., and Yao, J. (2015), "Power Enhancement in High-Dimensional Cross-Sectional Tests," *Econometrica*, 83, 1497–1541. [2549,2551,2552,2553]

Fisher, R. A. (1925), *Statistical Methods for Research Workers* (Vol. 1), Edinburgh: Oliver and Boyd. [2549]

Ginestet, C. E., Li, J., Balachandran, P., Rosenberg, S., and Kolaczyk, E. D. (2017), "Hypothesis Testing for Network Data in Functional Neuroimaging," *The Annals of Applied Statistics*, 11, 725–750. [2548]

Goeman, J. J., and Bühlmann, P. (2007), "Analyzing Gene Expression Data in Terms of Gene Sets: Methodological Issues," *Bioinformatics*, 23, 980–987. [2557]

Gravina, G. L., Senapedis, W., McCauley, D., Baloglu, E., Shacham, S., and Festuccia, C. (2014), "Nucleo-Cytoplasmic Transport as a Therapeutic Target of Cancer," *Journal of Hematology & Oncology*, 7, 85. [2559]

Hedges, L. V., and Olkin, I. (2014), *Statistical Methods for Meta-Analysis*, Amsterdam: Academic Press. [2554]

Hyodo, M., and Nishiyama, T. (2018), "A Simultaneous Testing of the Mean Vector and the Covariance Matrix among Two Populations for High-Dimensional Data," *TEST*, 27, 680–699. [2549]

Jiang, T., and Yang, F. (2013), "Central Limit Theorems for Classical Likelihood Ratio Tests for High-Dimensional Normal Distributions," *The Annals of Statistics*, 41, 2029–2074. [2549,2553]

Li, D., and Xue, L. (2015), "Joint Limiting Laws for High-Dimensional Independence Tests," arXiv preprint arXiv:1512.08819. [2549,2553]

Li, D., Xue, L., and Zou, H. (2018), "Applications of Peter Hall's Martingale Limit Theory to Estimating and Testing High Dimensional Covariance Matrices," *Statistica Sinica*, 28, 2657–2670. [2553]

Li, J., and Chen, S. X. (2012), "Two Sample Tests for High-Dimensional Covariance Matrices," *The Annals of Statistics*, 40, 908–940. [2548,2549,2550,2552,2553,2555]

Littell, R. C., and Folks, J. L. (1971), "Asymptotic Optimality of Fisher's Method of Combining Independent Tests," *Journal of the American Statistical Association*, 66, 802–806. [2554]

——— (1973), "Asymptotic Optimality of Fisher's Method of Combining Independent Tests II," *Journal of the American Statistical Association*, 68, 193–194. [2554]

Liu, Y., and Xie, J. (2020), "Cauchy Combination Test: A Powerful Test with Analytic *p*-value Calculation Under Arbitrary Dependency Structures," *Journal of the American Statistical Association*, 115, 393–402. [2554]

Liu, Z., Liu, B., Zheng, S., and Shi, N.-Z. (2017), "Simultaneous Testing of Mean Vector and Covariance Matrix for High-Dimensional Data," *Journal of Statistical Planning and Inference*, 188, 82–93. [2549,2553]

Niu, Z., Hu, J., Bai, Z., and Gao, W. (2019), "On LR Simultaneous Test of High-dimensional Mean Vector and Covariance Matrix Under Non-normality," *Statistics & Probability Letters*, 145, 338–344. [2549,2553]

Okada, H., and Pollack, I. F. (2004), "Cytokine Gene Therapy for Malignant Glioma," *Expert Opinion on Biological Therapy*, 4, 1609–1620. [2559]

Paietta, E., Racevskis, J., Neuberg, D., Rowe, J., Goldstone, A., and Wiernik, P. (1997), "Expression of CD25 (Interleukin-2 Receptor α Chain) in Adult Acute Lymphoblastic Leukemia Predicts for the Presence of BCR/ABL Fusion Transcripts: Results of a Preliminary Laboratory Analysis of ECOG/MRC Intergroup Study E2993," *Leukemia*, 11, 1887–1890. [2559]

Petrov, V. V., (1954), "Generalization of Cramér's Limit Theorem," *Uspekhi Matematicheskikh Nauk*, 9, 195–202. [2551]

Singh, K., Xie, M., and Strawderman, W. E. (2005), "Combining Information from Independent Sources Through Confidence Distributions," *The Annals of Statistics*, 33, 159–183. [2554]

Takeda, A., and Yaseen, N. R. (2014), "Nucleoporins and Nucleocytoplasmic Transport in Hematologic Malignancies," *Seminars in Cancer Biology*, 27, 3–10. [2559]

Touw, I., Pouwels, K., Van Agthoven, T., Van Gurp, R., Budel, L., Hoogerbrugge, H., Delwel, R., Goodwin, R., Namen, A., and Lowenberg, B. (1990), "Interleukin-7 is a Growth Factor of Precursor B and T Acute Lymphoblastic Leukemia," *Blood*, 75, 2097–2101. [2559]

Wang, L., Peng, B., and Li, R. (2015), "A High-Dimensional Nonparametric Multivariate Test for Mean Vector," *Journal of the American Statistical Association*, 110, 1658–1669. [2548]

Wang, S., and Yuan, M. (2019), "Combined Hypothesis Testing on Graphs with Applications to Gene Set Enrichment Analysis," *Journal of the American Statistical Association*, 114, 1320–1338. [2548]

Xie, M., Singh, K., and Strawderman, W. E. (2011), "Confidence Distributions and a Unifying Framework for Meta-Analysis," *Journal of the American Statistical Association*, 106, 320–333. [2549,2554]

Yoda, A., Yoda, Y., Chiaretti, S., Bar-Natan, M., Mani, K., Rodig, S. J., West, N., Xiao, Y., Brown, J. R., and Mitsiades, C. (2010), "Functional Screening Identifies CRLF2 in Precursor B-Cell Acute Lymphoblastic Leukemia," *Proceedings of the National Academy of Sciences*, 107, 252–257. [2559]

Yu, X., Li, D., and Xue, L. (2020), "Fisher's Combined Probability Test for High-Dimensional Covariance Matrices," arXiv preprint arXiv:2006.00426. [2549]

Yu, X., Yao, J., and Xue, L. (2019), "Innovated Power Enhancement for Testing Multi-Factor Asset Pricing Models," available at SSRN 3809369. [2549]

Zhu, L., Lei, J., Devlin, B., and Roeder, K. (2017), "Testing High-Dimensional Covariance Matrices, with Application to Detecting Schizophrenia Risk Genes," *The Annals of Applied Statistics*, 11, 1810–1831. [2548]