Statistica Sinica Preprint No: SS-2022-0388	
Title	Cross Projection Test for High-Dimension Mean Vectors
Manuscript ID	SS-2022-0388
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202022.0388
Complete List of Authors	Guanpeng Wang and Hengjian Cui
Corresponding Authors	Hengjian Cui
E-mails	hjcui@bnu.edu.cn

Statistica Sinica

CROSS PROJECTION TEST FOR HIGH-DIMENSIONAL MEAN VECTORS

Guanpeng Wang[1,2] and Hengjian Cui[2*]

Weifang University[1] *and Capital Normal University*[2]

Abstract: A cross projection test (CPT) technique for a one-sample vector in a high-dimensional setting is introduced. To overcome the problems caused by the curse of dimensionality, we construct test statistics by employing a projection test to project high-dimensional samples into one-or multi-dimensional directions. First, we randomly split the sample into two groups. We then find the p projection directions from a sample covariance matrix of the first group of samples. The second group is used to construct a projection statistic and perform the test. Second, we find the projection directions by exchanging the order of the two groups of samples, and we perform the test again to obtain another test statistic. Finally, we construct the CPT statistic by adding the two asymptotically uncorrelated test statistics together using the cross projection technique, such that the information from the two independent split samples can be fully utilized. The simulation results show that our proposed cross projection test controls the type I error well, and it is more powerful than the existing mean tests for some covariance matrix structures. Meanwhile, after applying the power enhancement technique, the CPT method performs non-trivially

*Hengjian Cui is the corresponding author. Email: hjcui@bnu.edu.cn.

in general cases, especially for testing against sparse alternatives. A real gene-data analysis illustrates that the performance of our CPT is quite well.

Key words and phrases: Asymptotic distribution, cross projection test, mean test, high dimension, projection direction.

1. The First Section

It is well known that the hypothesis test of the mean vector is fundamental to multivariate statistical analysis (see Anderson (2003) and Muirhead (1982)), which in turn is instrumental in diverse fields of research and application domains. The rapid development of technology has introduced new types of data, such as internet portals, hyperspectral imagery, microarray analysis, and DNA, to many fields. Generally speaking, these are often high-dimensional data in which the dimensionality of variables p is much larger than the sample size (n). This brings about the "curse of dimensionality" in statistical data analysis, which renders classical test statistics invalid or no longer applicable. The past two decades have witnessed increasing interest in mean signals difference identification for high-dimensional settings, and the existing methods are generally classified into two categories. The first is the modified Hotelling's T^2 test statistic, and the second involves constructing projection test statistics in a lower dimensional space through the projection technique. The specific details of these methods are described as follows.

Let $\mathbf{x}_1, \mathbf{x}_2 \ldots, \mathbf{x}_n$ be an independent and identically distribution random sample drawn from p-variate distribution $F(x)$, with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. In a one-sample mean vector test, we are primarily concerned with the following hypothesis testing:

$$H_0 : \boldsymbol{\mu} = \mathbf{0} \quad \text{versus} \quad H_1 : \boldsymbol{\mu} \neq \mathbf{0}. \tag{1.1}$$

The classic Hotelling's T^2 test (Hotelling (1931)) works well for the above hypothesis with fixed dimension p, and it is still applicable when $p < n - 1$. However, when $p/n \to 1 - \varepsilon$ for $\varepsilon > 0$, the Hotelling's T^2 test suffers power loss, as demonstrated by Bai and Saranadasa (1996). It is a well-known fact that based on the theory of a large-dimensional random matrix, the sample covariance matrix has only $n-1$ non-zero eigenvalues for the case of $p > n$ (see Bai and Silverstein (2010) and Pan and Zhou (2011)). As a result, the sample covariance matrix becomes singular, and the Hotelling's T^2 test statistic is no longer applicable. To overcome this curse of dimensionality, modifications to Hotelling's T^2 test statistic have been proposed to allow the method work well in higher dimension settings. The common idea of these modifications is simply to replace the inverse of the sample covariance with the identity matrix in the Hotelling formula, namely, by removing the sample covariance matrix (see Bai and Saranadasa (1996), Chen and Qin (2010) and Aoshima

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0388

and Yata (2011)). However, removing the sample covariance matrix does not guarantee the scale invariance of many statistics in the literature for high-dimensional mean tests, see, for example, Bai and Saranadasa (1996), Chen and Qin (2010), Zhang et al. (2021), and the references therein. Therefore, many studies replace the sample covariance matrix with a diagonal matrix to construct a scalar-transformation-invariant test, see, for example, Srivastava and Du (2008), Srivastava (2009), Srivastava et al. (2013), Park and Ayyala (2013) and Srivastava and Kubokawa (2013). However, many studies seek to preserve more information from the covariance matrix by using a regularized method to estimate the inverse of the covariance matrix or by normalizing the diagonal matrix formed by the diagonal elements of the sample covariance matrix, see, Dong et al. (2016) and Feng et al. (2015), respectively. It is worth noting that the method of modifying Hotelling formula mentioned above do not make full use of the correlation information within the variables. To illustrate, if the pairwise variables are strongly correlated or the covariance matrix is not a diagonal matrix or a banded structure, the modified Hotelling's method will suffer substantial power loss. Regularization methods may also suffer from selection tuning parameter confusion and sparsity assumptions. Meanwhile, Wang et al. (2015) proposed a high-dimensional nonparametric multivariate test for mean vector based on spatial-signs. Subsequently, a mean test for high-dimensional data based on a covariance matrix with a spiked structure or strongly spiked eigenvalue model was proposed to improve the power of the test statistic, see Wang et al. (2015), Wang and Xu (2018), Aoshima and Yata (2018) and Ishii et al. (2019) for more discussion. At the same time, other scholars proposed the use of a projection test to map a high-dimensional sample onto a low-dimensional space, which can, to some extent, solve the mean test problem caused by the curse of dimensionality, see Lauter (1996), Lauter et al. (1998), Lopes et al. (2011), and Huang (2015). Finally, the optimal choice of projection direction $\mathbf{\Sigma}^{-1}\boldsymbol{\mu}$ for one- or two-sample mean tests was proved by Huang (2015).

In practical application, the optimal projection direction must be estimated using samples. Thus, estimating the inverse of the population covariance matrix or precision matrix ($\mathbf{\Sigma}^{-1}$) still confronts the same issues as the classical Hotelling's T^2 with a high-dimensional setting. The ridge-like estimator, $(\mathbf{S}_n + \lambda\mathbf{D})^{-1}$, is regarded as an estimator of $\mathbf{\Sigma}^{-1}$ in the optimal projection direction, as shown in Huang (2015) and Liu et al. (2021), where $\mathbf{S}_n$ is sample covariance matrix, λ is a tuning parameter that controls the degree of penalty, and $\mathbf{D}$ is a diagonal matrix of $\mathbf{S}_n$. Using this method of adding a penalty term to the sample covariance matrix will uniformly increase the sample eigenvalues to a positive value such that the smallest eigenvalue is a positive number.

This will make it inconsistent with the inverse of the population covariance matrix. Consequently, the estimation of optimal direction is not really the theoretical optimal projection direction. If parameter λ is poorly chosen or an estimator of the projection direction is far from the original optimal direction, it is no longer effective to perform the test using the single optimal projection direction approach. In this study, we project the test samples in p directions, and we propose a cross projection test (CPT) approach to the mean hypothesis test when dimensions p are comparable to, or even larger than, sample size n. This ensures that the information from the two groups of splitting samples is as fully utilized as possible. In addition, the CPT test statistic not only overcomes the problem of searching for the optimal direction, but it also has asymptotic normality, which allows it to carry out our test regardless of whether the random samples come from a normal or non-normal distribution. Moreover, the proposed CPT approach performs very well in various situations, including the iso-correlation covariance matrix, the factor model, and other compound structures, in which the modified Hotelling's method would not be efficient.

The remainder of this paper is organized as follows. In Section 2, the background of the projection test and the application of Hotelling statistics after projection mapping are introduced, and the specific implementation process for our proposed CPT statistic is described in detail. Theoretically, with some mild conditions, the standard CPT statistic follows asymptotically standard normal distribution under the null hypothesis, and the asymptotic power function under the local alternative is shown in Section 3. In Section 4, we use the power enhancement technique proposed by Fan et al. (2015) to improve the performance of the CPT when the mean vector has sparse settings. The results of numerical studies in Section 5 further show that the advantages of the CPT statistic coincide with our theoretical conclusion. This article concludes with a brief discussion outlining possible extensions of this work in Section 6. All technicalities and additional details are relegated to the Supplementary Material.

2. Cross projection test for a one-sample mean vector

In this section, we begin by explaining how to implement the projection test method for a few directions, such that high-dimensional samples are mapped onto low-dimensional space. This ensure that the classic Hotelling's T^2 test statistic is still feasible. Second, we present the detailed process of constructing the new test statistic (CPT) by combining two cross-projection statistics.

2.1 Background of projection tests for multiple directions

Assume that $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ are n independent and identically distributed random samples from a p-variate normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. To the best of our knowledge, the projection test essentially projects the p-dimensional vector, $\mathbf{x} \in \mathbb{R}^p$, onto low-dimensional space so that certain traditional test statistics remain applicable for the projected samples. The classic Hotelling T^2 test statistic is then implemented to test one- or two-sample mean vectors when the dimension of variable p is smaller than sample size n in multivariate statistical analysis (see Anderson (2003) and Muirhead (1982)). Let $\mathbf{P}$ be a $p \times k$ full column-rank projection matrix that satisfies $\mathbf{P}^T\mathbf{P} = \mathbf{I}_k$ for an integer $k \in \{1, \ldots, \min(n, p)\}$, and it is drawn independently of the data to be projected. After transformation using the projection technology, the hypothesis testing for a one-sample mean vector can be written as follows:

$$H_{0,\text{proj}} : \mathbf{P}^T\boldsymbol{\mu} = \mathbf{0}_k \quad \text{verus} \quad H_{1,\text{proj}} : \mathbf{P}^T\boldsymbol{\mu} \neq \mathbf{0}_k, \tag{2.1}$$

where $\mathbf{0}_k$ represents a zero-element vector of k dimension. In the case of projection testing (2.1), the Hotelling T^2 test statistic for k-dimensional projected sample $\{\mathbf{P}^T\mathbf{x}_1, \mathbf{P}^T\mathbf{x}_2, \ldots, \mathbf{P}^T\mathbf{x}_n\}$ takes the following form:

$$T^2_{k,\text{proj}} = n(\mathbf{P}^T\bar{\mathbf{x}})^T(\mathbf{P}^T\mathbf{S}_n\mathbf{P})^{-1}\mathbf{P}^T\bar{\mathbf{x}},$$

where $\bar{\mathbf{x}} = (\bar{x}_1, \ldots, \bar{x}_p)^T$ and $\mathbf{S}_n = (s_{ij})_{p\times p}$ denote the sample mean and sample covariance matrix, respectively. It is well known that when $p < n$, the test statistic $(\frac{n-p}{p(n-1)} T^2_{k,\mathrm{proj}})$ follows $F_{p,n-p}(\delta)$, with p and $n-p$ degrees of freedom with non-central parameter δ, where $\delta = n\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$, regardless of whether the hypothesis is $H_{0,\mathrm{proj}}$ or $H_{1,\mathrm{proj}}$. Meanwhile, under $H_{0,\mathrm{proj}}$, statistic $T^2_{k,\mathrm{proj}}$ asymptotically converges in distribution to χ^2_k as $n \to \infty$. Notably, when $k = 1$, the $t(n-1)$ distribution with $n-1$ degree of freedom can be applied to the aforementioned mean hypothesis testing under normal distribution.

It is worth noting that the above tests employ a projection matrix or vector $\mathbf{P}$ to operate the sample. Lopes et al. (2011) adopted a single random matrix $\mathbf{P} \in \mathbb{R}^{p\times k}$ with i.i.d. $N(0,1)$ entries to implement the projection test approach. Although all the information from the sample is used for the projection operation, the selection of the projection direction may not be optimal. Subsequently, a projection test for high-dimensional mean vectors with optimal direction using the random splitting sample was proposed by Huang (2015). However, the theoretically optimal projection direction in practical applications needs to be obtained through the ridge-like estimator of $(\mathbf{S}_n + \lambda\mathbf{D})^{-1}$, which is not a consistent estimator in high-dimensional settings. Thus, this strategy will result in power loss when a set of samples are used to perform the test. Therefore, motivated by a projection test from a single random (non-random) projection matrix or vector direction, we propose a more powerful cross projection test, which is displayed in the next subsection.

2.2 Implementation of CPT

In this subsection, we further elaborate on the specific implementation algorithm of the CPT. An easy-to-compute sample-splitting approach proposed by Wasserman and Roeder (2009) is adopted in our algorithm. To execute our algorithm, the sample is randomly partitioned into two independent sets with one splitting; that is, the index set ($\{1, 2, \ldots, n\}$) of a random sample is split into two disjointed subsets. These are defined as $\mathcal{H}_i$ with size $n_i = |\mathcal{H}_i|$ for $i = 1$ and 2, in which the two-sample size satisfies $n_1+n_2 = n$. Correspondingly, n independent and identically distribution samples $\{\mathbf{x}_i, i = 1, \ldots, n\}$ are randomly split into 2 disjointed batches: $\mathcal{D}_1$ and $\mathcal{D}_2$, with $\mathcal{D}_i = \{\mathbf{x}_j, j \in \mathcal{H}_i\}$ for i=1 and 2. Without loss of generality, the datasets are denoted by $\mathcal{D}_1 = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n_1}\}$ and $\mathcal{D}_2 = \{\mathbf{x}_{n_1+1}, \mathbf{x}_{n_1+2}, \ldots, \mathbf{x}_n\}$ for the two group samples. Data $\mathcal{D}_2$ is employed to find p projection directions, and data $\mathcal{D}_1$ is used to construct test statistic T_1^2. Subsequently, switching data $\mathcal{D}_1$ and $\mathcal{D}_2$, i.e., $\mathcal{D}_1$ is used to find projection directions and the another test statistic T_2^2 is constructed by using data $\mathcal{D}_2$. The detailed CPT algorithm is described below:

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0388

Step 1 The sample covariance matrices $\mathbf{S}_1$ and $\mathbf{S}_2$ are respectively used as estimators of the covariance matrix $\mathbf{\Sigma} = (\sigma_{ij})$ for datasets $\mathcal{D}_1$ and $\mathcal{D}_2$, and the spectral decomposition of sample covariance estimator $\mathbf{S}_1$ is written as $\mathbf{S}_1 = U_1 \Lambda_1 U_1^T$, where $U_1 = (\mathbf{u}_{11}, \mathbf{u}_{12}, \ldots, \mathbf{u}_{1p})$ is the eigenvector matrix and Λ_1 is a diagonal matrix consisting of the eigenvalues of $\mathbf{S}_1$. Similarly, the estimator $\mathbf{S}_2 = U_2 \Lambda_2 U_2^T$, where $U_2 = (\mathbf{u}_{21}, \mathbf{u}_{22}, \ldots, \mathbf{u}_{2p})$, is the eigenvector matrix of $\mathbf{S}_2$.

Step 2 Project data $\mathcal{D}_1$ onto p directions ($\mathbf{u}_{21}$, $\mathbf{u}_{22}$,..., $\mathbf{u}_{2p}$) that are from data $\mathcal{D}_2$. That is, n_1 samples are respectively directed to projection direction $\mathbf{u}_{2i}$ to obtain the following projection vector, which can be written as $\mathbf{y}_i^{(1)} = (y_{i1}, y_{i2}, \ldots, y_{in_1})$, where $y_{ij} = \mathbf{u}_{2i}^T \mathbf{x}_j$ for $i = 1, 2, \ldots, p$ and $j = 1, 2, \ldots, n_1$. The test statistic for data $\mathcal{D}_1$ is constructed as $T_1^2 = \sum_{i=1}^p T_{1i}^2$, where

$$T_{1i}^2 = \frac{n_1 \bar{\mathbf{y}}_{1i}^2}{\mathbf{u}_{2i}^T \mathbf{S}_1 \mathbf{u}_{2i}}, \tag{2.2}$$

and $\bar{\mathbf{y}}_{1i}$ is the sample mean of vector $\mathbf{y}_i^{(1)}$. That is, $\bar{\mathbf{y}}_{1i} = \mathbf{u}_{2i}^T \bar{\mathbf{x}}_1$ represents the projection on the mean level of the first group.

Step 3 Project $\mathcal{D}_2$ onto p directions ($\mathbf{u}_{11}$, $\mathbf{u}_{12}$,..., $\mathbf{u}_{1p}$) and obtain the projection vector on i-th projection direction $\mathbf{u}_{1i}$. The projection vector is defined as $\mathbf{y}_i^{(2)} = (y_{i(n_1+1)}, y_{i(n_1+2)}, \ldots, y_{in})$, where $y_{ij} = \mathbf{u}_{1i}^T \mathbf{x}_j$ for

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0388

$i = 1, 2, \ldots, p$ and $j = n_1 + 1, n_1 + 2, \ldots, n$. The test statistic for data $\mathcal{D}_2$ is constructed as $T_2^2 = \sum_{i=1}^{p} T_{2i}^2$, where

$$T_{2i}^2 = \frac{n_2 \bar{\mathbf{y}}_{2i}^2}{\mathbf{u}_{1i}^T \mathbf{S}_2 \mathbf{u}_{1i}}, \tag{2.3}$$

and $\bar{\mathbf{y}}_{2i}$ is the sample mean of vector $\mathbf{y}_i^{(2)}$.

Step 4 The CPT statistic is obtained by the summation of two asymptotically uncorrelated statistics, T_1^2 and T_2^2, which is defined as follows:

$$T_{\text{CP}}^2 =: T_1^2 + T_2^2 = \sum_{i=1}^{p} T_{1i}^2 + \sum_{i=1}^{p} T_{2i}^2. \tag{2.4}$$

To obtain the rejection region of the test statistic of the CPT, we further derive the asymptotic distribution of statistic T_{CP}^2 through our theoretical analysis under the null hypothesis (H_0).

3. Theoretical analysis

In this section, we derive the asymptotic distribution of test statistics T_{CP}^2 in equation (2.4) under the null hypothesis, $H_0 : \boldsymbol{\mu} = \mathbf{0}$ for the general covariance matrix forms when the suitable conditions hold. Before providing the content of the theorems, we present some mild assumptions, as used in Srivastava (2009), to obtain our theoretical results for the implementation of the CPT algorithm.

Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ be n independent and identically distributed p-variate random variables with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. They obey the independent component structure as follows:

$$\mathbf{x}_i = \boldsymbol{\mu} + \Gamma \mathbf{z}_i, \tag{3.1}$$

where Γ is a $p \times p$ non-singular matrix satisfying $\boldsymbol{\Sigma} = \Gamma\Gamma^T > 0$, $\mathbf{z}_i = (z_{i1}, z_{i2}, \ldots, z_{ip})^T$, $i = 1, 2, \ldots, n$, and z_{ij} are i.i.d. with a finite fourth moment, which is given by $\mathrm{E}(z_{ij}) = 0, \mathrm{E}(z_{ij}^2) = 1$, and $\mathrm{E}(z_{ij}^4) = \kappa < \infty$, for $i = 1, \ldots, n$, $j = 1, \ldots, p$. We further list some assumptions as follows:

Assumption 1. Sample size n and dimension p of random vector $\mathbf{x}$ satisfy $n = O(p^\tau)$, where $0 < \tau \leq 1$.

Assumption 2. Two groups of samples are formed by splitting one, and their sizes satisfy this relationship, as $n \to \infty$,

$$n_1/(n_1 + n_2) \to k \in (0, 1).$$

Assumption 3. Assume that the correlation matrix $\mathbf{R}$ of the random vector $\mathbf{x}$ and its dimensions satisfy the following limiting relation:

$$\lim_{p\to\infty} \varrho_i = \lim_{p\to\infty} \left(\frac{\mathrm{tr}\big((\mathbf{R})^i\big)}{p} \right) = \varrho_{i0} < \infty, \quad i = 1, \ldots, 4.$$

Assumption 4. Suppose that the smallest eigenvalue of the covariance matrix satisfies $\lambda_{\min}(\boldsymbol{\Sigma}) > c_0$ for some positive constant c_0.

Remark 1. Assumption 1 indicates that the sample size maintains a certain order relationship with the dimension. This requirement also appears in Chen and Qin (2010), Srivastava and Du (2008), and Srivastava (2009). The process of finding the projection directions and executing the test statistics through two split samples allows more efficiency to be exerted under the condition of Assumption 2. This is commonly used in the two-sample test; it requires that the sizes of two samples are comparable and not extremely imbalanced, see, for example, Chen and Qin (2010), Srivastava and Kubokawa (2013), and the references therein. It can be seen from the expression of Assumption 3 that the degree of correlation in the correlation matrix cannot be extremely heavy, or it can be overcome by dimension p and converge to a positive constant, ϱ_{i0}. Assumption 4 is a weak constraint on the covariance matrix, which requires the minimum eigenvalue of $\mathbf{\Sigma}$ to be far from zero. Readers can refer to Bickel and Levina (2008) for more details.

Notably, the covariance matrix satisfying the above assumptions has a very wide range of forms, such as the identity matrices, banded matrices, AR structures, and spiked eigenvalue models. Let λ_i's be the eigenvalues of the covariance matrix $\mathbf{\Sigma}$. We then discuss our Assumption 3 on the spiked eigenvalue model, such as

$$\lambda_i = a_i p^{\alpha_i} \ (i = 1, \ldots, k) \quad \text{and} \quad \lambda_i = c_i \ (i = k+1, \ldots, p) \tag{3.2}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0388

with positive and fixed constants, a_is, c_is and α_is, and a positive and fixed integer k. The requirement of $\lim_{p\to\infty} \varrho_i = \lim_{p\to\infty} \left(\frac{\mathrm{tr}\left((\mathbf{R})^i\right)}{p} \right) = \varrho_{i0} < \infty, \quad i = 1, \ldots, 4$ in Assumption 3 holds under the necessary condition of $\lambda_1(\mathbf{\Sigma}) = O(p^{\alpha})$ with $0 < \alpha \leq 1/4$ where λ_1 is the largest eigenvalue, that is $\alpha_1 \leq 1/4$ in (3.2). However, for the non-strongly spiked eigenvalue model in Aoshima and Yata (2018), $\lambda_1^2/\mathrm{tr}(\mathbf{\Sigma}^2) \to 0$ as $p \to \infty$, it is required that the largest eigenvalue satisfies $\alpha_1 < 1/2$ in the spiked model (3.2). For strongly spiked eigenvalue model $\liminf_{p\to\infty} \left(\lambda_1^2/\mathrm{tr}(\mathbf{\Sigma}^2)\right) > 0$, it is required that $\lambda_1 = O(p^{\alpha_1})$ with $\alpha_1 \geq 1/2$ in spiked model (3.2). In general, our Assumption 3 is slightly stronger than the non-strongly spiked eigenvalues mentioned in Aoshima and Yata (Aoshima and Yata (2018)). However, for some special cases, the spiked covariance matrix is $\mathbf{\Sigma} = \begin{pmatrix} \mathbf{D} & \mathbf{0}_{k\times(p-k)} \\ \mathbf{0}_{(p-k)\times k} & \mathbf{I}_{p-k} \end{pmatrix}$ or $\mathbf{\Sigma} = \begin{pmatrix} \mathbf{A}_{k\times k} + \mathbf{D} & \mathbf{0}_{k\times(p-k)} \\ \mathbf{0}_{(p-k)\times k} & \mathbf{I}_{p-k} \end{pmatrix}$, where $\mathbf{D} = \mathrm{diag}(a_1 p^{\alpha_1}, \ldots, a_k p^{\alpha_k})$ and the elements of the non-negative definite matrix $\mathbf{A}$ are constants independent of the dimension, then the correlation matrix is an identity matrix or an approximation of the identity matrix. Therefore, our Assumption 3 is satisfied for the special cases mentioned above, regardless of the values of α_1.

The test statistic, T^2_{CP}, is constructed as in equation (2.4). It can be expressed in matrix form as

$$T_{\mathrm{CP}}^2 = n_1 \bar{\mathbf{x}}_1^T U_2 \widehat{\mathbf{W}}_1^{-1} U_2^T \bar{\mathbf{x}}_1 + n_2 \bar{\mathbf{x}}_2^T U_1 \widehat{\mathbf{W}}_2^{-1} U_1^T \bar{\mathbf{x}}_2,$$

where $\widehat{\mathbf{W}}_1 = \mathrm{diag}(\mathbf{u}_{21}^T \mathbf{S}_1 \mathbf{u}_{21}, \ldots, \mathbf{u}_{2p}^T \mathbf{S}_1 \mathbf{u}_{2p})$ and $\widehat{\mathbf{W}}_2 = \mathrm{diag}(\mathbf{u}_{11}^T \mathbf{S}_2 \mathbf{u}_{11}, \ldots, \mathbf{u}_{1p}^T \mathbf{S}_2 \mathbf{u}_{1p})$. To obtain the theoretical properties of T_{CP}^2, we need to define

$$t_{\mathrm{o}}^2 = n_1 \bar{\mathbf{x}}_1^T U_2 \mathbf{W}_1^{-1} U_2^T \bar{\mathbf{x}}_1 + n_2 \bar{\mathbf{x}}_2^T U_1 \mathbf{W}_2^{-1} U_1^T \bar{\mathbf{x}}_2,$$

where $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_2$ represent the mean vectors of $\mathcal{D}_1$ and $\mathcal{D}_2$ with one splitting, respectively, and $\mathbf{W}_1 = \mathrm{diag}(\mathbf{u}_{21}^T \boldsymbol{\Sigma} \mathbf{u}_{21}, \ldots, \mathbf{u}_{2p}^T \boldsymbol{\Sigma} \mathbf{u}_{2p})$ and $\mathbf{W}_2 = \mathrm{diag}(\mathbf{u}_{11}^T \boldsymbol{\Sigma} \mathbf{u}_{11}, \ldots, \mathbf{u}_{1p}^T \boldsymbol{\Sigma} \mathbf{u}_{1p})$, and have to discuss its properties. The asymptotic normality of t_{o}^2 is given in Theorem 1.

Theorem 1. *Suppose that Assumptions 2–3 hold. Then, under the null hypothesis (H_0), as $n, p \to \infty$, we have*

$$\frac{t_{\mathrm{o}}^2 - 2p}{\left\{2\big(\mathrm{tr}(\mathbf{R}_1^2) + \mathrm{tr}(\mathbf{R}_2^2)\big)\right\}^{1/2}} \xrightarrow{d.} N(0,1),$$

where $\mathbf{R}_1 = \mathbf{D}_2^{-1/2}(U_2^T \boldsymbol{\Sigma} U_2)\mathbf{D}_2^{-1/2}$, with $\mathbf{D}_2 = \mathrm{diag}(\mathbf{u}_{21}^T \boldsymbol{\Sigma} \mathbf{u}_{21}, \ldots, \mathbf{u}_{2p}^T \boldsymbol{\Sigma} \mathbf{u}_{2p})$ for given projection matrix U_2, and $\mathbf{R}_2 = \mathbf{D}_1^{-1/2}(U_1^T \boldsymbol{\Sigma} U_1)\mathbf{D}_1^{-1/2}$, with $\mathbf{D}_1 = \mathrm{diag}(\mathbf{u}_{11}^T \boldsymbol{\Sigma} \mathbf{u}_{11}, \ldots, \mathbf{u}_{1p}^T \boldsymbol{\Sigma} \mathbf{u}_{1p})$ for given projection matrix U_1. "$\xrightarrow{d.}$" stands for the convergence in distribution.

Notably, Lemma 2 explains that the denominator of T_{CP}^2 in expression (2.4) enjoys a consistent property, and its proof is shown in Appendix A. This means that the quadratic form $\mathbf{u}_{2i}^T\mathbf{S}_1\mathbf{u}_{2i}$ converges to $\mathbf{u}_{2i}^T\mathbf{\Sigma}\mathbf{u}_{2i}$ and $\mathbf{u}_{1i}^T\mathbf{S}_2\mathbf{u}_{1i}$ converges to $\mathbf{u}_{1i}^T\mathbf{\Sigma}\mathbf{u}_{1i}$, with probability 1 for given the projection directions $\mathbf{u}_{1i}$ and $\mathbf{u}_{2i}$ $(1 \leq i \leq p)$. To obtain the rejection region for working our CPT statistic, we shall use $\frac{1}{p}\mathrm{tr}(\widehat{\mathbf{R}}_1 - p^2/(n_1-1))$ and $\frac{1}{p}\mathrm{tr}(\widehat{\mathbf{R}}_2 - p^2/(n_2-1))$ separately as estimators of $\frac{1}{p}\mathrm{tr}(\mathbf{R}_1)$ and $\frac{1}{p}(\mathbf{R}_2)$, where $\widehat{\mathbf{R}}_1$ and $\widehat{\mathbf{R}}_2$ are the sample correlation matrix of projection samples $U_2^T\boldsymbol{X}_1$ and $U_1^T\boldsymbol{X}_2$ with $\boldsymbol{X}_1 = (\mathbf{x}_1, \ldots, \mathbf{x}_{n_1})$, $\boldsymbol{X}_2 = (\mathbf{x}_{n_1+1}, \ldots, \mathbf{x}_n)$, respectively.

The asymptotic normality property of test statistic T_{CP}^2 under the null hypothesis (H_0) can be found in Theorem 2.

Theorem 2. *Suppose that Assumptions 1–4 hold, and given two independent projection directions U_1 and U_2 for their respective projection samples, then we have under the null hypothesis (H_0) that*

$$\frac{T_{\mathrm{CP}}^2 - p(\frac{n_1-1}{n_1-3}) - p(\frac{n_2-1}{n_2-3})}{\left\{2\left(\mathrm{tr}(\widehat{\mathbf{R}}_1^2) + \mathrm{tr}(\widehat{\mathbf{R}}_2^2) - \frac{p^2}{n_1-1} - \frac{p^2}{n_2-1}\right)\right\}^{1/2}} \xrightarrow{d.} N(0,1)$$

as $n, p \to \infty$.

It follows from the property of Theorem 2 that for a sufficiently large n, p, the rejection region of the new test, T_{CP}^2, at significance level α is $\{\mathbf{x} : (T_{\mathrm{CP}}^2 - p(\frac{n_1-1}{n_1-3}) - p(\frac{n_2-1}{n_2-3}))/\{2(\mathrm{tr}(\widehat{\mathbf{R}}_1^2) + \mathrm{tr}(\widehat{\mathbf{R}}_2^2) - \frac{p^2}{n_1-1} - \frac{p^2}{n_2-1})\}^{1/2} \geq z_\alpha\}$, where z_α denotes the upper α quantile of $N(0,1)$. Furthermore, we can get the power function of T^2_{CP} under the alternative hypothesis ($H_1 : \boldsymbol{\mu} \neq 0$). Define

$$T^2 =: \Big(T^2_{\mathrm{CP}} - p\big(\frac{n_1-1}{n_1-3}\big) - p\big(\frac{n_2-1}{n_2-3}\big)\big)\Big/\Big(2\big(\mathrm{tr}(\widehat{\mathbf{R}}_1^2) + \mathrm{tr}(\widehat{\mathbf{R}}_2^2) - \frac{p^2}{n_1-1} - \frac{p^2}{n_2-1}\big)\Big)^{1/2},$$

and it is convenient to derive the asymptotic normality of standardized CPT statistic T^2. We shall consider a local alternative under which

$$\boldsymbol{\mu} = \{1/(n(n-1))\}^{1/2}\boldsymbol{\delta}, \tag{3.3}$$

where $\boldsymbol{\delta}$ is a vector of constants. For every p, given two independent projection directions, U_1 and U_2, we shall assume that

$$\frac{1}{p}(\boldsymbol{\delta}^T U_2 \mathbf{W}_1^{-1} U_2^T \boldsymbol{\delta} + \boldsymbol{\delta}^T U_1 \mathbf{W}_2^{-1} U_1^T \boldsymbol{\delta}) \leq C, \tag{3.4}$$

where constant C is independent of p. The local alternative (3.3) and restricted condition (3.4) mentioned above are used in studies such as Srivastava and Du (2008) and Srivastava (2009).

Theorem 3. *Consider local alternative* $\boldsymbol{\mu} = 1/\{n(n-1)\}^{\frac{1}{2}}\boldsymbol{\delta}$*, assuming that* $\frac{1}{p}(\boldsymbol{\delta}^T U_2 \mathbf{W}_1^{-1} U_2^T \boldsymbol{\delta} + \boldsymbol{\delta}^T U_1 \mathbf{W}_2^{-1} U_1^T \boldsymbol{\delta}) \leq C$ *holds for the given projection directions,* U_1 *and* U_2*, then under the conditions of Theorem 2, we have*

$$\lim_{(n,p)\to\infty} \Big[P(T^2 > z_\alpha | U_1, U_2) - \Phi\Big(-z_\alpha + \frac{\Delta(\boldsymbol{\delta}; n, p)}{\sqrt{2(\mathrm{tr}(\mathbf{R}_1^2) + \mathrm{tr}(\mathbf{R}_2^2))}}\Big)\Big] = 0,$$

where $\Delta(\boldsymbol{\delta}; n, p) = \frac{1}{n-1}(k\boldsymbol{\delta}^T U_2 \mathbf{W}_1^{-1} U_2^T \boldsymbol{\delta} + (1-k)\boldsymbol{\delta}^T U_1 \mathbf{W}_2^{-1} U_1^T \boldsymbol{\delta})$.

It is easily seen that under the conditions of Theorem 3, the asymptotic power of standardized CPT statistic T^2 as $(n,p) \to \infty$ is given by

$$\boldsymbol{\beta}(T^2|\boldsymbol{\delta}) \simeq \mathrm{E}_{U_1,U_2}\Big(\Phi\Big(-z_\alpha + \frac{\Delta(\boldsymbol{\delta};n,p)}{\sqrt{2(\mathrm{tr}(\mathbf{R}_1^2)+\mathrm{tr}(\mathbf{R}_2^2))}}\Big)\Big).$$

Remark 2. Our proposed CPT procedure may be affected by the result of the single random splitting technique. To overcome this puzzle and improve the power of our test, a multi-splitting technique can be employed for mean testing in a high-dimensional setting, see Meinshausen et al. (2009). The CPT procedure, in which test statistic T^2_{CP} is obtained through the single splitting procedure, is repeated m times. Then, the multi-splitting technique yields m p-values. These p-values could be aggregated using equation (2.3) in Meinshausen et al. (2009) and the Cauchy combination test in Liu and Xie (2020). It is important to note that the multi-splitting approach not only eliminates the effect of random splitting, but it can also control the false discovery rate (FDR) well.

4. Power enhancement technique for sparse mean vector testing

Our proposed CPT statistic, T^2_{CP}, is more powerful for the dense mean vector $(H_1 : \boldsymbol{\mu} \neq 0)$ described above in Section 2. However, for the sparse alternative hypothesis $(H_{1s} : \boldsymbol{\mu} = \boldsymbol{\mu}_s)$, where $\boldsymbol{\mu}_s$ indicates there are many zero elements in $\boldsymbol{\mu}$, it is obvious that the cross projection test statistic (T^2_{CP}; sum-type test) may not capture significant signals in the margins very well, as discussed in Cai et al. (2014). Therefore, following Fan et al. (2015) and Guo and Cui (2019), a power enhancement technique is proposed to improve the test performance under the alternative hypothesis in our CPT statistic procedure. Similar to Assumption 3.1 in Fan et al. (2015), a power enhancement technique is proposed if the sample mean and sample covariance matrix satisfy the following assumption:

Assumption 5. As $n, p \to \infty$, the estimators of the sample mean and variance $\{\bar{x}_j, s_{jj}\}_{1\leq j\leq p}$ are satisfied for sequence $\delta_{n,p} = C_1 \log(\log(n))\sqrt{\log(p)}$ that

(a) $\inf_{\boldsymbol{\mu}\in\mathcal{U}} P(\max_{1\leq j\leq p} |\bar{x}_j - \mu_j|/s_{jj}^{1/2} < \delta_{n,p}/\sqrt{n})|\boldsymbol{\mu}) \to 1$,

(b) $\inf_{\boldsymbol{\mu}\in\mathcal{U}} P(4/9 < s_{jj}/\sigma_{jj} < 9/4, \forall j = 1, \ldots, p|\boldsymbol{\mu}) \to 1$,

where $\mathcal{U}$ is a collection of mean vectors of r.v. $\mathbf{x}$.

It is worth noting that the constants 4/9 and 9/4 in condition (b) are not optimally chosen, as this condition only requires $\{s_{jj}\}_{1\leq j\leq p}$ be not-too-bad estimators of their population counterparts. Let J_1 be a test statistic that has a correct asymptotic size (for example, the cross projected test statistic T^2_{CP} introduced in Section 2), which can become a composite test statistic by adding a non-negative power enhancement term, $J_0 \geq 0$. To satisfy the power enhancement property, define the sets $\mathcal{S}(\boldsymbol{\mu})$ and $\widehat{\mathcal{S}}$ as below

$$\mathcal{S}(\boldsymbol{\mu}) = \left\{ j \in \{1, \ldots, p\} : |\mu_j| > 3\sigma_{jj}^{1/2}\delta_{n,p}/\sqrt{n} \right\},$$

and

$$\widehat{\mathcal{S}} = \Big\{ j \in \{1, \ldots, p\} : |\bar{x}_j| > s_{jj}^{1/2} \delta_{n,p} / \sqrt{n} \Big\}.$$

Then, a power enhancement test statistic is constructed as

$$J_0 = n \cdot \mathbf{I}\{\max_j (|\bar{x}_j| / s_{jj}^{1/2}) > \delta_{n,p} / \sqrt{n}\}, \tag{4.1}$$

where n is a strengthened coefficient, $\mathbf{I}\{\cdot\}$ stands for the indication function, and formula (4.1) is simply J_0, as in Fan et al. (2015). The following theorem states the asymptotic behavior of statistic J_0 under both the null and alternative hypotheses.

Theorem 4. *Suppose Assumption 5 holds. As $n, p \to \infty$, $P(\widehat{\mathcal{S}} = \emptyset | H_0) \to 1$ under null hypothesis (H_0). Hence*

$$P(J_0 = 0 | H_0) \to 1 \quad \textit{and} \quad \inf_{\{\boldsymbol{\mu} \in \mathcal{U} : \mathcal{S}(\boldsymbol{\mu}) \neq \emptyset\}} P(J_0 > n | \boldsymbol{\mu}) \to 1.$$

Theorem 4 not only gives the asymptotic behavior of J_0, but it also describes the "sure screening property" of $\widehat{\mathcal{S}}$, which means that it selects all significant components whose indices are in set $\mathcal{S}(\boldsymbol{\mu})$. Obviously, this result is performed uniformly in $\boldsymbol{\mu}$ under both the null and alternative hypotheses, and the term J_0 can identify more significant signals from alternative $\boldsymbol{\mu} \neq \mathbf{0}$, significantly improving the corresponding power. From the asymptotic normality of Theorem 2, it can be easily to know that

$$J_{\text{CPT}} =: \Big(T_{\text{CP}}^2 - p\big(\frac{n_1 - 1}{n_1 - 3}\big) - p\big(\frac{n_2 - 1}{n_2 - 3}\big)\Big) \Big/ \Big(2\big(\text{tr}(\widehat{\mathbf{R}}_1^2) + \text{tr}(\widehat{\mathbf{R}}_2^2) - \frac{p^2}{n_1 - 1} - \frac{p^2}{n_2 - 1}\big)\Big)^{1/2}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0388

follows a standard normal distribution under the null hypothesis. Similar to the statistic constructed by Fan et al. (2015) to enhance the power, test statistic J_{CPT} has an asymptotic null distribution, $N(0,1)$. Hence, the critical region also takes the form $\{\mathbf{x} : J_{\mathrm{CPT}} + J_0 > z_\alpha\}$ at significance level $\alpha \in (0,1)$ via Theorem 5.

Theorem 5. *Suppose Assumption 4 holds. The test statistics* J_{CPT} *and* $J_{\mathrm{CPT}} + J_0$ *enjoy the same asymptotic null distribution,* $N(0,1)$*. Furthermore, as* $n, p \to \infty$*, the power enhancement test on the set* $\mathcal{U}_s = \{\boldsymbol{\mu} : \boldsymbol{\mu} \in \mathcal{U}, \mathcal{S}(\boldsymbol{\mu}) \neq \emptyset\}$ *has high power,*

$$\inf_{\boldsymbol{\mu} \in \mathcal{U}_s} P(J_{\mathrm{CPT}} + J_0 \geq z_\alpha | \boldsymbol{\mu}) \to 1.$$

It can be found that the test statistic can obtain high power for mean vector $\boldsymbol{\mu} \in \mathcal{U}_s$ in Theorem 5. Moreover, it is worth noting that some projection tests can be used in combination with the power enhancement technique. Our CPT method henceforth can also achieve a more powerful performance.

5. Numerical studies

5.1 Simulation results

In this section, we conduct some simulation studies to compare our proposed test statistic, T^2_{CP}, with existing tests for high-dimensional one-sample data by repeating each experiment 10000 times. We mainly evaluate the advantages of the CPT method over Huang (2015), which used a single optimal projection method to test the mean vector of the same data in terms of empirical size and power. Therefore, the performance level and empirical power of test statistic T^2_{CP} will be compared with the test statistics proposed by Bai and Saranadasa (1996)(abbreviated as T^2_{BS}), Chen and Qin (2010) (abbreviated as T^2_{CQ}), S-rivastava (2009) (abbreviated as T^2_{S}), and Huang (2015) (optimal projection direction, abbreviated as T^2_{OP}). To find the optimal projection in T^2_{OP}, we set the splitting percentage to 50% and the tuning parameter to $\lambda = (n/2)^{-0.5}$. These settings are reasonable ranges, and they are the parameter choices used in the simulations in Huang (2015). When the null hypothesis holds, the optimal test statistic, T^2_{OP}, converges to the student t-distribution for normal data, while the asymptotic chi-square property of $(T^2_{\text{OP}})^2$ is required for non-normal data. Random samples $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)^T \in \mathbb{R}^{n\times p}$ are still generated by considering the structure of (3.1). They follow multivariate model $\mathbf{x}_i = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\mathbf{z}_i$, where $\mathbf{z}_i$'s are independent and identically distribution random variables. In the multivariate model, $\mathbf{z}_i = (z_{i1}, z_{i2}, \ldots, z_{ip})$ is generated through three distributions: the multivariate normal distribution, student t distribution, and chi-square distribution. For specific parameter settings, see Examples (a)–(c) below.

Example (a): The random variables, z_{ij}'s, follow a standard normal distribution; that is, $z_{ij} \sim N(0,1)$.

Example (b): Random variable z_{ij} has a distribution $t(6)/\sqrt{3/2}$, where $t(6)/\sqrt{3/2}$ is a standardized t distribution with 6 degrees of freedom.

Example (c): Random variable z_{ij} has a distribution $(\chi_4^2 - 4)/(2\sqrt{2})$, where $(\chi_4^2 - 4)/(2\sqrt{2})$ is a standardized chi-square distribution with 4 degrees of freedom, which is a non-symmetric distribution.

To setup the covariance matrix model, we consider the following three structure types:

(I). Factor model structure $\mathbf{\Sigma}_1$, for which $\mathbf{\Sigma}_1 = \mathbf{I}_p + (\sigma^2/p^{3/4})\mathbf{A}_{p\times 5}\mathbf{A}_{p\times 5}^T$, where $\mathbf{I}_p$ is a p-dimensional identity matrix and $\sigma^2 = 18$, and a deterministic matrix $\mathbf{A}$ is factor loading with elements from standard normal $N(0,1)$ distribution.

(II). Factor model structure $\mathbf{\Sigma}_2$, for which $\mathbf{\Sigma}_2 = \mathbf{I}_p + \sigma^2\mathbf{A}_{p\times 5}\mathbf{A}_{p\times 5}^T$, where $\sigma^2 = 1.5$ and a deterministic matrix $\mathbf{A}$ is generated similarly as $\mathbf{\Sigma}_1$. ($\mathbf{\Sigma}_2$ does not satisfy the necessary condition of Assumption 3)

(III). Diagonal covariance matrix $\mathbf{\Sigma}_3$, where $\mathbf{\Sigma}_3 = \mathbf{I}_p$.

(IV). Autoregressive structure $\mathbf{\Sigma}_4$, in which $\mathbf{\Sigma}_{ij} = \rho^{|i-j|}$, where $\rho = 0.5$.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0388

Considering the projection direction and the estimation of the sample mean vector in the implementation of CPT, the segmentation percentage can be set to 50% for the best power and stable performance, which is shown in the Supplementary Material. In the following simulation studies, the splitting percentage is set to 50% to show all the simulation results.

Dense mean case: The p-dimension mean vector $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_p)^T$ is set to $\mu_i = (w-1)/20$ with $w = 1, 2, 3, 4$, for $i = 1, \ldots, p$. It is worth noting that when $w = 1$, the mean vector is exactly the null vector, and w is not equal to one, which is the alternative hypothesis.

Table 1: Empirical size and power (%) of test statistics (Example(a), nominal size $\alpha = 0.05$)

		Size					Dense mean $w = 2$					Dense mean $w = 3$				
	Type	T^2_{CP}	T^2_{OP}	T^2_{BS}	T^2_{S}	T^2_{CQ}	T^2_{CP}	T^2_{OP}	T^2_{BS}	T^2_{S}	T^2_{CQ}	T^2_{CP}	T^2_{OP}	T^2_{BS}	T^2_{S}	T^2_{CQ}
	$\boldsymbol{\Sigma}_1$	6.5	4.4	6.9	7.2	4.6	**92.6**	79.8	13.0	15.8	9.5	**100.0**	100.0	58.5	83.9	48.4
$n = 200$	$\boldsymbol{\Sigma}_2$	6.1	4.8	6.6	6.9	4.8	**92.4**	79.7	7.8	8.6	4.9	**100.0**	100.0	11.0	11.8	8.6
$p = 250$	$\boldsymbol{\Sigma}_3$	5.9	5.0	5.2	5.6	5.0	97.3	86.0	99.8	99.8	99.4	**100.0**	100.0	100.0	100.0	100.0
	$\boldsymbol{\Sigma}_4$	6.4	5.2	6.1	6.2	5.2	78.1	45.7	95.3	93.3	74.3	**100.0**	99.6	100.0	100.0	100.0
	$\boldsymbol{\Sigma}_1$	5.8	5.7	6.3	6.7	4.5	**98.2**	92.0	15.2	18.4	9.7	**100.0**	100.0	88.0	98.6	79.1
$n = 250$	$\boldsymbol{\Sigma}_2$	5.8	5.2	6.5	6.5	4.4	**98.0**	89.6	8.2	9.0	4.8	**100.0**	100.0	11.9	15.6	9.5
$p = 350$	$\boldsymbol{\Sigma}_3$	6.1	5.4	5.7	6.0	5.7	99.4	96.1	100.0	99.9	99.9	**100.0**	100.0	100.0	100.0	100.0
	$\boldsymbol{\Sigma}_4$	6.0	5.3	5.3	5.7	5.3	89.6	64.6	98.8	98.9	98.1	**100.0**	100.0	100.0	100.0	100.0

Table 1 shows a comparison of our cross projection test statistic (T^2_{CP}), the optimal projection method (T^2_{OP}), and the modified Hotelling method for testing the dense mean vector in terms of empirical size and power. Random samples were drawn from multivariate the normal distribution, student-t distribution, and chi-square distribution, respectively. The dense mean settings show that the mean signal gradually increases as w increases. Once the eigenvalues of the covariance matrix have one or more spikes (such as the factor model in types (I)$\sim$(II), where the correlations are distributed across all elements of the covariance matrix), the modified Hotelling method performs badly. However, our CPT method is still better than the optimal projection method. The reason for this is that in the spiked models, the sample eigen-subspace of sample covariance matrix $\mathbf{S}_n$, the one which is corresponded to the eigenspace of significant eigenvalues of $\mathbf{\Sigma}$, converges to the eigen-subspace of $\mathbf{\Sigma}$ with a high probability. Therefore, the alternative $\boldsymbol{\mu}$ is that as long as the high probability falls towards the spanning space of the eigenvector(s) corresponding to the mostly non-significant eigenvalue(s), we propose that the CPT test statistic is significantly better than the Hotelling type statistic.

As we all know, for example, test statistics T^2_{BS}, T^2_{S}, and T^2_{CQ} perform well by improving Hotelling's formula when the non-zero elements of the covariance matrix fall uniformly near the diagonal region. However, when the covariance matrix is diagonal, our proposed CPT method is very close to this in terms of power, but the optimal projection method (T^2_{OP}) proposed by Huang (2015)

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0388

performs very badly in when $w = 2$. When the sample size is fixed and the dimension increases, our CPT method is still able to overcome the dimensional problem, and the empirical power improves. It is worth noting that our CPT approach works well for both the heavy-tailed student-t distribution and chi-square distribution (skewed distribution) in the Supplementary Material.

Sparse mean case: For this part of the simulation studies, we employ by Fan et al. (2015) power enhancement technique introduced in Section 4 to improve the empirical power when the empirical size is controlled. The sparse mean vector is set to $\boldsymbol{\mu}_s = (0.35 * \mathbf{1}_4^T, \mathbf{0}_{p-4}^T)^T$. Under the settings of covariance model $\boldsymbol{\Sigma}_1$, the empirical size of the three distributions (Example(a)–Example(c)) can be well controlled around the nominal level of 0.05, as shown in Table 2. Table 3 shows that the enhancement power technique has improved the performance of T^2_{CP} in terms of the empirical power for the multi-normal distribution, under three covariance matrix structures. More simulation results for the student t and chi-square distributions are reported in Tables S5 and S6 in the Supplementary Material.

Table 2 shows the empirical size of some tests, when C_1 is set to 1.3 in the expression of $\delta_{n,p}$ in formula (4.1). It follows straightforwardly that when $n, p \to \infty$, probability $P(\widehat{\mathcal{S}} = \emptyset) \to 1$, so the empirical size of T^2_{CP} is equal to $J_{\mathrm{CPT}} + J_0$ with probability 1. This coincides with the screening mechanism in

Table 2: Empirical size (%) of tests with $\alpha = 0.05$ for $\mathbf{\Sigma}_1$

n	p	T^2_{BS}	T^2_{S}	T^2_{CQ}	T^2_{OP}	T^2_{CP}	$J_{\mathrm{OP}} + J_0$	$J_{\mathrm{CPT}} + J_0$	$P(\widehat{\mathcal{S}} = \emptyset)$
Example(a): Multivariate Gaussian									
150	150	6.77	6.89	5.30	5.15	6.34	5.18	6.35	99.96
	200	7.50	7.35	5.45	4.67	6.22	4.81	6.23	99.92
	300	6.60	6.68	4.87	4.87	5.76	4.88	5.76	99.98
200	150	5.90	6.40	3.90	4.68	5.86	4.70	5.89	99.96
	200	6.85	7.25	4.75	5.06	6.10	5.09	6.11	99.94
	300	7.10	7.15	4.90	5.02	5.97	5.03	5.97	100.00
Example(b): Multivariate Student t									
150	150	7.53	7.60	4.97	5.10	6.63	5.13	6.65	99.93
	200	6.30	6.47	4.10	5.67	6.43	5.70	6.47	99.97
	300	7.23	7.60	4.83	5.23	6.07	5.25	6.08	99.98
200	150	7.22	7.48	5.53	5.32	6.10	5.35	6.11	99.97
	200	7.12	7.45	4.80	5.28	5.63	5.28	5.63	100.00
	300	6.84	7.12	5.03	5.28	5.51	5.28	5.51	100.00
Example(c): Multivariate Chi-square									
150	150	7.03	7.29	4.83	5.83	6.22	5.97	6.29	99.83
	200	6.65	6.99	5.43	5.15	6.20	5.28	6.30	99.85
	300	7.27	7.85	4.87	5.24	6.31	5.38	6.43	99.84
200	150	6.42	6.62	4.83	5.55	5.70	5.60	5.73	99.94
	200	7.03	7.24	5.43	5.51	5.40	5.53	5.43	99.95
	300	7.16	7.30	4.87	5.32	5.84	5.33	5.85	99.99

Theorem 4, which does not affect the control of Type I errors. Therefore, as long as the value of C_1 is set slightly larger, or n, p tends to infinity, T^2_{CP} and T^2_{OP} will be equal to $J_{\mathrm{CPT}} + J_0$ and $J_{\mathrm{OP}} + J_0$, respectively, with probability 1.

We can see from the empirical power of the test statistics in Table 3 that the CPT approach performs better than the optimal projection method proposed by Huang (2015) in various situations. Essentially, in terms of empirical power, the performance advantage of the CPT in the sparse mean vector test is similar to that in the dense case. Regarding the covariance matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ with more dominant eigenvalues, our cross projection method still works very well, whereas the traditional method is almost infeasible. Further-

Table 3: Empirical power (%) of tests with $\alpha = 0.05$ for Example(a)

Type	n	p	T^2_{BS}	T^2_{S}	T^2_{CQ}	T^2_{OP}	T^2_{CP}	$J_{\text{OP}} + J_0$	$J_{\text{CPT}} + J_0$	$P(\widehat{\mathcal{S}} = \emptyset)$
$\boldsymbol{\Sigma}_1$	150	150	10.83	13.43	7.43	63.96	75.56	69.37	78.15	79.22
		200	10.87	11.02	7.33	41.30	63.80	43.89	64.19	93.04
		300	10.22	11.25	7.47	42.70	49.79	48.23	54.25	86.40
	200	150	12.21	13.51	8.97	71.89	91.06	73.79	91.31	88.77
		200	11.34	12.47	9.03	67.96	86.72	71.25	87.57	82.76
		300	11.31	12.73	7.23	56.54	73.96	63.21	77.37	77.62
$\boldsymbol{\Sigma}_2$	150	150	7.53	7.82	5.70	43.60	74.74	43.67	74.76	99.64
		200	7.43	7.50	5.50	31.74	61.02	31.80	61.02	99.76
		300	7.47	8.21	4.63	45.22	48.94	45.53	49.10	98.87
	200	150	8.30	8.64	5.80	72.13	93.34	72.19	93.34	99.30
		200	7.37	7.95	5.77	74.07	86.25	74.20	86.27	98.63
		300	7.80	8.13	5.73	48.95	71.29	48.97	71.29	99.89
$\boldsymbol{\Sigma}_3$	150	150	97.79	97.11	95.83	63.95	86.21	90.34	94.20	17.21
		200	94.39	94.19	91.33	60.98	79.46	86.78	90.68	22.97
		300	86.60	86.65	79.23	52.80	65.32	80.20	84.24	32.00
	200	150	99.75	99.76	99.77	76.64	97.18	97.65	99.23	4.70
		200	99.38	99.36	98.80	74.13	93.64	96.58	98.28	6.87
		300	97.20	97.19	95.00	68.29	84.72	94.16	96.03	11.58
$\boldsymbol{\Sigma}_4$	150	150	86.85	87.21	81.07	43.96	67.18	79.16	83.61	28.31
		200	80.54	81.02	72.17	40.78	59.69	74.66	78.77	34.70
		300	69.27	69.90	59.13	36.63	47.16	67.55	71.02	43.59
	200	150	96.41	96.37	94.27	52.93	83.52	90.88	94.00	13.11
		200	93.34	93.43	89.60	50.59	76.71	88.64	91.79	16.22
		300	86.64	86.74	79.30	47.18	64.74	83.79	86.99	25.22

more, when the elements of the covariance matrix are uniformly located near the diagonal in the cases of $\boldsymbol{\Sigma}_3$ and $\boldsymbol{\Sigma}_4$, our CPT approach is much better than the optimal projection method. However, the CPT approach is still the most effective after the power enhancement technique, the projection method (see tests T^2_{CP} and T^2_{OP}) is only slightly better than the method of modifying Hotelling's tests (e.g., the T^2_{BS}, T^2_{S}, and T^2_{CQ} methods) for random samples from a multivariate normal distribution when $P(\widehat{\mathcal{S}} = \emptyset)$ becomes larger, such as the case of $p = 300$ for $\boldsymbol{\Sigma}_4$. Now that the sample size is set to $n = 200$ and the covariance matrices are $\boldsymbol{\Sigma}_3$ and $\boldsymbol{\Sigma}_4$, it is obviously found that the probability $P(\widehat{\mathcal{S}} = \emptyset)$ becomes very small, and the empirical power of $J_{\text{OP}} + J_0$ and $J_{\text{CPT}} + J_0$ are both close to 1, which is exactly consistent with Theorem 5.

5.2 Real data analysis

In this section, we employ a pig gene dataset from the Department of Animal Science at Iowa State University. The dataset is significant at different levels with respect to certain treatments and was previously analyzed by Lkhagvadorj et al. (2009), Chen et al. (2010), and Guo and Cui (2019). To implement our cross projection test, we equally divided a dataset from an experiment with 24 six-month-old Yorkshire gilts into four groups. The gilts are genotyped by the MC4R (melanocortin-4 receptor) gene. Half with D298, and the other half with N298. However, two diet treatments were assigned randomly for each genotyped gilt, in which six of them were fed without restriction, and the others fasted. A more detailed description of the experiment can be found in Lkhagvadorj et al. (2009). There are 24,123 genes in the liver tissues and 6176 gene sets, which fluctuate in dimension from 1 to 5158. The original goal of this research analysis was to identify treatment effects on the gene expression structure levels, and our interest is testing the difference within each gene set for different treatments.

To meet our theoretical analysis requirements, we only focus on 302 gene sets, with a dimension greater than 50 and smaller than 800 to meet the requirement that p be greater than sample size n. Assume that $\mathcal{S}_1, \ldots, \mathcal{S}_{307}$ are used to quantify 302 sets of genes, where gene-set $\mathcal{S}_g$ consists of p_g genes. As the number of gilts in the different treatment groups was the same, 12 gilts without food restriction and 12 gilts with the fasting treatment, we used them as pairwise matching data. Let $\boldsymbol{x}_{1,i}^{(g)}$ and $\boldsymbol{x}_{2,i}^{(g)}$ be the i-th gilt for the fasting and unrestricted groups, respectively, both with the p_g dimensional vector for the g-th gene set. Let $\boldsymbol{y}_i^{(g)}$ be the difference between $\boldsymbol{x}_{1,i}^{(g)}$ and $\boldsymbol{x}_{2,i}^{(g)}$. If no significant difference exists between the treatment and control groups, the mean of $\left\{\boldsymbol{y}_i^{(g)}\right\}$ is equal to $\mathbf{0}_{pg}$. Therefore, the null and alternative hypotheses within each gene set are described as follows

$$H_{0g} : \boldsymbol{\mu}^{(g)} = \mathbf{0}_{p_g} \quad \text{verus} \quad H_{1g} : \boldsymbol{\mu}^{(g)} \neq \mathbf{0}_{p_g},$$

where the $\boldsymbol{\mu}^{(g)}$ is the p_g-dimensional mean vector of the g-th gene set $\{\boldsymbol{y}^{(g)}\}$. Under significance level $\alpha = 0.05$, Figure 1 shows a histogram of the p-values for the mean test for the 302 gene sets using the CPT method.

From Figure 1, we can see that most gene sets show significant differences between the treatment group with fasting and the control group without restriction. From our specific analysis, about 74.2% of the gene sets have significant differences in the mean level, which is basically consistent with the biological conclusion of Lkhagvadorj et al. (2009).

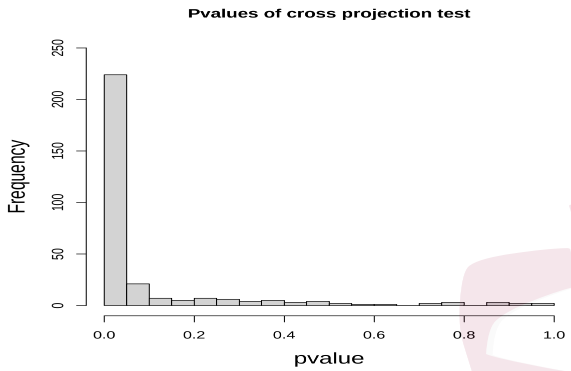

Figure 1: Histogram of p-values for cross projection tests.

6. Conclusion and discussion

In this study, we proposed a new cross projection test approach to the widely studied high-dimensional mean vector hypothesis tests. The main goal was to improve the performance of the optimal projection direction with only one splitting proposed by Huang (2015). In our approach, test statistic T^2_{CP} integrates p projection directions and makes full use of the information from the two segmented samples obtained through the splitting technique to construct our cross projection test. Simultaneously, the selection of tuning parameter λ in the estimation of the inverse covariance matrix and the splitting of a percentage of samples, as performed in the test in Huang (2015), are also avoided in our application. Instead, the CPT only uses the first half of the samples to select the projection direction. The other half of the samples are used to perform the test. Regarding the sample splitting technique, the test statistic loses information at the intersection of the two split samples, so the simulation showed some disadvantages compared to the modified Hotelling's test statistic. However, when the correlation between two variables is strong or the eigenvalues of the matrix have spikes, our CPT still works well, while the modified Hotelling's method fails. For the sparse mean test, the T^2_{CP} statistic becomes $J_{\text{CPT}} + J_0$ after the power enhancement technique added. Regardless of what type of structure the covariance matrix has, our new CPT approach performs the best, to a certain extent.

In our extended research on the high-dimensional two-sample mean vector test when the population covariance matrices $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$, the specific hypothesis testing was $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ versus $H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$. We then outlined the steps of this requirement. For instance, for two samples, $X_1 \in \mathbb{R}^{n_1 \times p}$ and $X_2 \in \mathbb{R}^{n_2 \times p}$, we first divide X_1 and X_2 into $X_{11} \in \mathbb{R}^{n_{11} \times p}$ and $X_{12} \in \mathbb{R}^{n_{12} \times p}$, and $X_{21} \in \mathbb{R}^{n_{21} \times p}$ and $X_{22} \in \mathbb{R}^{n_{22} \times p}$, respectively. We then combine samples X_{11} and X_{21} to find p projection directions with size $n_{11} + n_{21}$. The total size of the two split samples, X_{12} and X_{22}, is equal to $n_{12} + n_{22}$, and these are merged to perform the test statistic in the projection directions. The remaining steps of the two-sample test are shown in the CPT implementation process proposed for the one-sample case in Section 2.

Supplementary Material

The proofs of the theorems 1–5 are given in the Supplementary Material.

Acknowledgments

The authors thank the editor, an associate editor, and two referees for their many helpful comments that have resulted in significant improvements in the article. This work is supported by the State Key Program of the National Natural Science Foundation of China (12031016), National Natural Science Foundation of China (NNSFC) grants (11971324, 11901406), the Interdisciplinary Construction of Bioinformatics and Statistics, and the Academy for Multidisciplinary Studies, Capital Normal University.

References

Anderson, T. (2003). An introduction of multivariate statistical analysis, 3rd ed. *New York: Wiley*.

Aoshima, M. and K. Yata (2011). Two-stage procedures for high-dimensional data. *Sequential Analysis (Editors special invited paper) 30*(4), 356–399.

Aoshima, M. and K. Yata (2018). Two-sample tests for high-dimension, strongly spiked eigenvalue models. *Statistica Sinica 28*, 43–62.

Bai, Z. and H. Saranadasa (1996). Effect of high dimension: By an example of a two sample problem. *Statistica Sinica 6*(2), 311–329.

Bai, Z. and J. Silverstein (2010). Spectral analysis of large dimensional random matrices. *Springer, Second Edition*.

Bickel, P. and E. Levina (2008). Regularized estimation of large covariance matrices. *Annals of Statistics 36*, 199–227.

Cai, T., W. Liu, and Y. Xia (2014). Two-sample test of high dimensional means under dependence. *Journal of the Royal Statistical Socity. Series B 76*(2), 349–372.

Chen, S. and Y. Qin (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *Annals of Statistics 38*(2), 808–835.

Chen, S., L. Zhang, and P. Zhong (2010). Tests for high dimensional covariance matrices. *Journal of the American Statistical Association 105*(490), 810–819.

Dong, K., H. Pang, T. Tong, and M. G. Genton (2016). Shrinkage-based diagonal Hotelling's tests for high-dimensional small sample size data. *Journal of Multivariate Analysis 143*, 127–142.

Fan, J., Y. Liao, and J. Yao (2015). Power enhancement in high-dimensional cross-sectional tests. *Econometrica 83*(4), 1497–1541.

Feng, L., C. Zou, Z. Wang, and L. Zhu (2015). Two-sample Beheren-Fisher problem for high-dimensional data. *Statistica Sinica 25*(4), 1297–1312.

Guo, W. and H. Cui (2019). Projection tests for high-dimensional spiked covariance matrices. *Journal of Multivariate Analysis 169*, 21–32.

Hotelling, T. (1931). The generalization of student's ratio. *Annals of Mathematics and Statistics 2*,

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0388

360–378.

Huang, Y. (2015). Projection test for high-dimensional mean vector with optimal direction Ph.D. dissertation. *Department of Statistics, The Pennsylvania State University at University Park*.

Ishii, A., K. Yata, and M. Aoshima (2019). Inference on high-dimensional mean vectors under the strongly spiked eigenvalue model. *Japanese Journal of Statistics and Data Science, 2*, 105–128.

Lauter, J. (1996). Exact t and F tests for analyzing studies with multiple endpoints. *Biometrics 52*(3), 964–970.

Lauter, J., E. Glimm, and S. Kropf (1998). Multivariate tests based on left-spherically distributed linear scores. *Annals of Statistics 26*(5), 1972–1988.

Liu, W., X. Yu, and R. Li (2021). Multiple-splitting projection test for high-dimensional mean vectors. *arXiv:2110.15480v1*.

Liu, Y. and J. Xie (2020). Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. *Journal of the American statistical association 115*(529), 393–402.

Lkhagvadorj, S. et al. (2009). Microarray gene expression profiles of fasting induced changes in liver and adipose tissues of pigs expressing the melanocortin-4 receptor d298n variant. *Physiological Genomics 38*(1), 98–111.

Lopes, M., L. Jacob, and M. Wainwright (2011). A more powerful two-sample test in high di-

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0388

mensions using random projection. *Advances in Neural Information Processing Systems 24*, 1206–1214.

Meinshausen, N., L. Meier, and P. Buehlmann (2009). P-values for high-dimensional regression. *Journal of the American statistical association 104*(488), 1671–1681.

Muirhead, R. (1982). Aspects of multivariate statistical theory. *New York: Wiley*.

Pan, G. and W. Zhou (2011). Central limit theorem for Hotelling's t^2 statistic under large dimension. *Annals of Applied Probability 21*(5), 1860–1910.

Park, J. and D. Ayyala (2013). A test for the mean vector in large dimension and small samples. *Journal of Statistical Planning & Inference 143*(5), 929–943.

Srivastava, M. (2009). A test for the mean vector with fewer observations than the dimension under non-normality. *Journal of Multivariate Analysis 100*, 518–532.

Srivastava, M. and M. Du (2008). A test for the mean vector with fewer observations than the dimension. *Journal of Multivariate Analysis 99*(3), 386–402.

Srivastava, M., S. Katayama, and Y. Kano (2013). A two sample test in high dimensional data. *Journal of Multivariate Analysis 114*, 349–358.

Srivastava, M. and T. Kubokawa (2013). Tests for multivariate analysis of variance in high dimension under non-normality. *Journal of Multivariate Analysis 115*, 204–216.

Wang, L., B. Peng, and R. Li (2015). A high-dimensional nonparametric multivariate test for mean vector. *Journal of the American Statistical Association 110*, 1658–1669.

Wang, R. and X. Xu (2018). On two-sample mean tests under spiked covariances. *Journal of Multivariate Analysis 167*, 225–249.

Wang, Y., W. Lan, and H. Wang (2015). A high dimensional two-sample test under a low dimensional factor structure. *Journal of Multivariate Analysis 140*, 162–170.

Wasserman, L. and K. Roeder (2009). High dimesnional variable selection. *Annals of Statistics 37*, 2178–2201.

Zhang, L., T. Zhu, and J. Zhang (2021). Two-sample Behrens–Fisher problems for high-dimensional data: a normal reference scale-invariant test. *Journal of Applied Statistics 213*, 142–161.

School of Mathematics and Statistics, Weifang University, Weifang, 261061, China.

E-mail: (wguanpeng@163.com)

School of Mathematical Sciences, Capital Normal University, Beijing 100048, China.

E-mail: (hjcui@bnu.edu.cn)