Journal of Econometrics 257 (2026) 106313

Contents lists available at ScienceDirect

Journal of Econometrics

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

Graph-based multisample comparison with application to feature selection for multi-category responses

Dan Pu [a], Haoming Shi [b,*], Wei Lan [c,*], Chih-Ling Tsai [d]

[a] School of Statistics and Joint Laboratory of Data Science and Business Intelligence, Southwestern University of Finance and Economics, Chengdu, China
[b] School of Finance, Southwestern University of Finance and Economics, Chengdu, China
[c] Laboratory for Data Intelligence and Innovation in Finance and Economics, Center of Statistical Research and Joint Lab of Data Science and Business Intelligence, Southwestern University of Finance and Economics, Chengdu, China
[d] Graduate School of Management, University of California, Davis, California, USA

ARTICLE INFO

Keywords:
Graph-based methods
Maximum pairwise difference tests
Multisample comparison
Sparse alternatives
Feature selection

ABSTRACT

This article proposes a graph-based maximum pairwise difference (MPD) test to compare K-sample distributions. This test can be utilized to identify the most informative variables contributing to the heterogeneity of multiple multivariate distributions. To construct the test, we first pool observations from multiple samples to create a similarity graph based on the distances between any two observations. For each observation, we next evaluate its connection probability to each sample. Then, we doubly maximize the squared differences of these connection probabilities via the within and between samples. Accordingly, the MPD test is a max-of-squares test and can effectively detect sparse alternatives. In addition, we show the asymptotic null distribution of the test statistic and establish its consistency under alternatives. To illustrate the finite sample performance of the proposed test, we conduct extensive simulation studies. The results indicate that the MPD test outperforms some commonly used multisample tests. Finally, we employ it to select relevant features related to corporate credit ratings.

1. Introduction

Testing the equality of K underlying distributions has been a widely studied problem in economics and statistics. This fundamental K-sample testing problem is frequently discussed in the research literature across various disciplines. For example, in data science, a standard assumption is that both the in-sample and out-of-sample data are drawn from an identical distribution; otherwise, the trained model may perform poorly in out-of-sample forecasts. Consequently, distribution shift detection emerges as a fundamental issue in data science. This task involves determining whether the distributions of two datasets are equivalent, as explored by Hur and Liang (2025). Additionally, in comparing predictive accuracy of various machine learning methods, an important question is whether a statistically significant difference exists between the distributions of the forecast errors induced by these methods (Hassani and Silva, 2015). Furthermore, in multiple-source data integration, testing distributional equality across different groups of data can simplify the model construction for the pooled dataset (Zhang et al., 2024). These examples highlight the practical significance for testing K-sample distributions.

During the past several decades, much effort in the literature has been devoted to addressing the K-sample problem via parametric and nonparametric methods see, e.g., Kruskal (1952), Zhang and Wu (2007), Cai et al. (2017), Chen (2020). In the parametric

* Corresponding authors.
E-mail addresses: shihm98@outlook.com (H. Shi), lanwei@swufe.edu.cn (W. Lan).

https://doi.org/10.1016/j.jeconom.2026.106313
Received 30 March 2025; Received in revised form 2 June 2026; Accepted 4 August 2026
Available online 18 August 2026
0304-4076/© 2026 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

approach, several tests have been developed to handle the K-sample problem see, e.g., Lopes et al. (2011), Cai et al. (2014). These tests can be applied to multivariate data under certain distributional assumptions, such as multivariate normal distributions. In such a setting, the K-sample comparison problem is equivalent to testing the equality of means or covariances across different samples. However, these tests are not applicable when the distribution of observations is mis-specified. To avoid this issue, numerous researchers have focused on nonparametric tests, such as the Kolmogorov-Smirnov test, the Wilcoxon test and the Wald-Wolfowitz runs test (Bickel, 1969; Oja and Randles, 2004). While these tests are useful and insensitive to distribution assumptions, they are not directly applicable to high-dimensional multivariate data. Here, the term "high-dimensional data" refers to the regime in which the data dimension is allowed to diverge with the sample size and may even exceed it. With the advent of the big data era, high-dimensional data are now commonly obtained and readily available in bioinformatics and genomics see, e.g., Bibby et al. (2022). In addition, non-Euclidean data (such as image data, network data and functional data) are commonly collected in many applications see, e.g., Kossinets and Watts (2006), Wang et al. (2016). Accordingly, how to effectively test the equality of multiple samples on such complex data types has attracted great interest and study.

Recently, graph-based nonparametric tests have received considerable attention due to their robustness, flexibility and computational convenience. This type of test reformulates the distribution comparison problem into a study of a specifically designed graph, which allows us to construct test statistics. For example, Friedman and Rafsky (1979) first constructed a k-minimum spanning tree (k-MST) based on the pairwise distances among the pooled observations, and then computed the number of edges that connect the nodes from different samples. However, this edge-count test usually suffers from low power when the dimension of the data is moderate to high. To overcome this problem, Chen and Friedman (2017) proposed the generalized edge-count test, and Chen et al. (2018) subsequently introduced the weighted edge-count test to address the sample imbalance problem. In addition to MST-type tests, Schilling (1986) and Henze (1988) constructed tests based on k-nearest neighbor (k-NN) graphs. Moreover, Rosenbaum (2005) employed the minimum distance nonbipartite pairing (MDP) and proposed the crossmatch test, which is exactly distribution-free in finite samples. Bhattacharya (2019) provided a general framework for analyzing commonly used graph-based two-sample tests and comparing their asymptotic properties. However, the abovementioned tests are only applicable for two-sample problems.

To address the K-sample problem, Nettleton and Banerjee (2001) proposed a graph-based multisample test based on the nearest-neighbor graph. Then, Petrie (2016) extended the aforementioned edge-count test to study multisample problems. In addition, Mukhopadhyay and Wang (2020) developed a K-sample testing procedure based on the LP graph kernel. Later, Song and Chen (2022) generalized a method from Chen and Friedman (2017) and utilized both within-sample edges and between-sample edges to construct the edge-count test. This test shows large power across a wide range of alternatives. Recently, Mukherjee et al. (2022) extended the work from Rosenbaum (2005) and proposed a distribution-free nonparametric test based on the minimum nonbipartite matching of pooled observations. This test can be efficiently applied to multivariate data as well as non-Euclidean data if the distance between observations is well-defined.

It is of interest to note that most of the above tests are based on the sum-of-squares statistics. For example, Mukherjee et al. (2022) proposed the multisample Mahalanobis crossmatch (MMCM) test, which is based on the Mahalanobis distance of the observed cross-counts. In addition, Song and Chen (2022) considered Mahalanobis-like tests based on the within-sample and between-sample edges. Based on the findings of Cai et al. (2014) and Feng et al. (2022), it is expected that the above methods will have high empirical power against non-sparse alternatives (i.e., most underlying distributions are different from each other). Under sparse alternatives, however, the sum-of-squares type tests may have low power. To get a better picture, we assume that the alternative hypothesis is sparse so that there are only a few heterogeneous samples and most of the samples are homogeneous. Then, under sparse alternatives, it is not surprising to find that the averaged signal is weak, although a few of the samples have strong signals. Consequently, the sum-of-squares type tests may not detect sparse signals. Similar studies can be found in Fan et al. (2015) and Feng et al. (2022).

In practice, sparse alternatives are frequently encountered when we employ the K-sample test to conduct feature selections for multi-category responses. Consider a scenario where we have multiple sets of multivariate data, each associated with distinct category labels. Our primary purpose is to identify the relevant features that cause significant differences in the distributions of multiple categories (Heller et al., 2016; Ghosh et al., 2021). In the context of feature selection, it is commonly found that certain features can only distinguish between two or three categories. That is to say, these informative features exhibit significant distributional differences among a small number of categories, while the distributions remain homogeneous across most of the other categories. This phenomenon has been verified in our real-data analysis in Section 4, which aims to select relevant features related to corporate credit ratings. For instance, the feature "PS" (Price to sales ratio; detailed illustrations are given in Table S.17 of the supplementary material) exhibits distributional homogeneity among four categories, with only one category being an exception. Consequently, a sparse alternative emerges in this problem. This practical challenge motivates us to develop a nonparametric multi-sample test. This test is designed to be powerful under sparse alternatives and can be used to select significant features for multi-category responses.

Our method. After exploring the K-sample testing problem under sparse alternatives, we propose a graph-based nonparametric test, namely the maximum pairwise difference (MPD) test. To compute the test statistic, we first construct a graph based on the pairwise distances among pooled observations. For each observation, we then calculate the probabilities of connections "within" the samples and "between" the samples, respectively. Subsequently, we doubly maximize the squared differences between the "within" and "between" probabilities across all possible pairs of samples for each observation. Hence, under the null hypothesis, the connection probabilities within the sample and between the samples are close to each other. By doubly maximizing the differences between the probabilities of connections within the samples and between the samples, the proposed test statistic is expected to approach zero under the null hypothesis, whereas the test statistic is expected to increase under the alternative hypothesis.

Our contribution. Compared with existing tests, the proposed test enjoys the following three appealing advantages.

(1) **Completely Nonparametric and Highly Versatile.** The proposed test does not impose any parametric assumption on data distributions or structural constraints on shift alternatives. Moreover, it is applicable to multivariate and non-Euclidean data, as long as the similarity graph is well-defined.
(2) **Suitable for Sparse Alternatives.** According to the construction of the test statistic, the proposed test maximizes the squared differences between two types of connection probabilities and belongs to the max-of-squares type test. Compared with the sum-of-squares tests, it is expected to exhibit higher power against sparse alternatives in multiple-sample scenarios. This finding is further validated through simulation studies in Section 3 of the main paper and Section S.4.1 of the supplementary material. Under sparse alternatives, the proposed test outperforms other existing tests in effectively detecting small differences among the K distributions.
(3) **Robustness to Non-Sparse Alternatives.** In addition to sparse alternatives, the proposed test performs well under non-sparse alternatives, as evidenced by simulation results in Section S.4.4 of the supplementary material. To further characterize the property of the MPD test under non-sparse alternatives, we establish its connection with the edge-count test proposed by Chen and Friedman (2017). The analysis in Section 2.4 of the main paper suggests that the proposed test exhibits superior power under non-sparse alternatives, particularly in scenarios involving covariance shifts and distribution shifts. This observation aligns with simulation results presented in Tables S.11 and S.12 of the supplementary material.

Technical challenges. Under some mild conditions, we theoretically derive the asymptotic null distribution of the test statistic and prove its consistency under alternatives. However, this theoretical analysis is nontrivial, since the proposed test statistic MPD is the doubly maximum value induced by a number of probability differences. Hence, it naturally gives rise to two main technical challenges. On the one hand, the covariance matrix resulting from these differences is singular because some differences can be represented as linear combinations of other differences. On the other hand, there exist complex correlations among these differences, rendering the test statistic a maximum of sums of correlated random vectors. Consequently, the classical asymptotic results for the maxima of independent summations cannot be directly applied. To overcome these difficulties, we first devise a procedure to remove redundant differences, thereby ensuring that the resulting covariance matrix is positive definite. To tackle the second challenge, we employ two expansion techniques-Slepian interpolation and Stein's leave-one-out method-to explicitly isolate the correlated components and demonstrate that the discrepancy between the extracted and original forms is asymptotically negligible. Through this approach, we rigorously establish a Gaussian approximation result for the proposed test when the number of samples K diverges.

Application novelty. Building on these advantages, we integrate the proposed MPD test with the graph-based feature selection algorithm of Ghosh et al. (2021) to identify features that drive significant distributional differences across multiple categories. Compared with conventional feature selection methods, this approach is completely nonparametric and imposes no additional assumptions on the data distribution or model specification. Furthermore, it is capable of detecting subtle differences among different categories, thereby identifying informative features that distinguish only a small subset of categories. This finding is supported by an empirical analysis of a corporate credit rating dataset.

Numerical findings. To evaluate the finite-sample performance of the MPD test, we conduct extensive simulation studies. Simulation results suggest that the proposed test not only controls the size well, but also is more powerful than existing tests. These results align with theoretical findings, reinforcing the strong empirical performance of the proposed test.

The rest of this paper is organized as follows. Section 2 introduces the proposed MPD test, establishes its theoretical properties, and investigates its relation to the edge-count test. Sections 3 and 4 present numerical studies and a real-data example, respectively. Finally, the paper is concluded with a brief discussion in Section 5.

2. Hypothesis tests

2.1. Notations and tests

Let $X_i^{(k)}$ $(i = 1, \dots, n_k)$ be the independent and identically distributed observations from F_k $(k = 1, \dots, K)$, and the total number of observations is $N = \sum_{k=1}^{K} n_k$. Here, $X_i^{(k)}$ can represent various types of data, such as multivariate data and non-Euclidean data. Without loss of generality, we focus on multivariate data, where the number of features is allowed to be either fixed or diverging with the sample size. For the sake of convenience, we rearrange the observations sample-by-sample, and denote $\boldsymbol{X} = (X_1, \dots, X_N) = (X_1^{(1)}, \dots, X_{n_1}^{(1)}, X_1^{(2)}, \dots, X_{n_2}^{(2)}, \dots, X_1^{(K)}, \dots, X_{n_K}^{(K)})$. Accordingly, if $\sum_{l=0}^{k-1} n_l < i \le \sum_{l=0}^{k} n_l$ with $n_0 = 0$, then the observation X_i belongs to the kth sample for $k = 1, \dots, K$. In addition, assume that the set $\mathcal{C}_k = \{\sum_{l=0}^{k-1} n_l + 1, \dots, \sum_{l=0}^{k} n_l\}$ contains the collection of indices associated with the observations coming from the kth sample. Moreover, for each $i = 1, \dots, N$, define g_i as the sample label of observation X_i. Thus, $i \in \mathcal{C}_k$ means that $g_i = k$, for some $k = 1, \dots, K$ and $i = n_0 + \cdots + n_{k-1} + 1, \dots, n_0 + \cdots + n_k$.

To compare the K-sample distributions, we consider the following hypotheses:

$$H_0 : F_1 = \cdots = F_K \quad \text{versus} \quad H_1 : F_{k_1} \neq F_{k_2} \text{ for some } 1 \le k_1 < k_2 \le K. \tag{1}$$

In this paper, we mainly focus on sparse alternatives, that is, there exist $k_1, \dots, k_m \in \{1, \dots, K\}$ and $k_{m+1}, \dots, k_K \in \{1, \dots, K\}$, such that $F_{k_1} \neq \cdots \neq F_{k_m}, F_{k_{m+1}} = \cdots = F_{k_K}$, where m is relatively small compared with K. We then construct the test statistic via the following three steps.

First, we regard every observation as a node of the graph, and create a graph based on the pairwise distances among pooled observations. Specifically, observations X_i and X_j are defined to be connected if $\|X_i - X_j\| < \tau$; otherwise, these two observations are disconnected. Here, $\tau > 0$ is a pre-specified tuning parameter. Accordingly, the adjacency matrix $A_\tau = (A_{ij,\tau})_{N\times N}$ is defined as

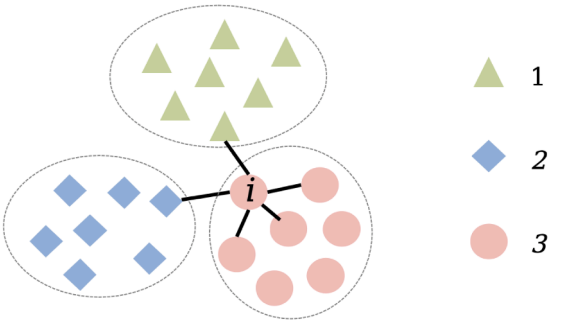

Fig. 1. The connectivity patterns of the observation i.

follows: $A_{ij,\tau} = 1$ if observations X_i and X_j are connected; otherwise, $A_{ij,\tau} = 0$. For the sake of completeness, we define $A_{ii,\tau} = 0$ for $i = 1, \ldots, N$. This adjacency matrix constitutes a similarity graph.

Second, we construct the connection probabilities of the "within" sample and the "between" samples. Specifically, for each observation, the connection probability within the sample is defined as

$$\hat{P}_{i,\tau}^{(g_i)} = \frac{1}{n_{g_i} - 1} \sum_{j \in C_{g_i}} A_{ij,\tau}.$$

Analogously, the connection probability between the sample g_i and the sample k ($g_i \neq k$) is defined as follows:

$$\hat{P}_{i,\tau}^{(k)} = \frac{1}{n_k} \sum_{j \in C_k} A_{ij,\tau}.$$

If the distributions of K samples are homogeneous, the connection probabilities of the "within" sample and the "between" samples are quantitatively similar for each observation, and hence, the difference between $\hat{P}_{i,\tau}^{(g_i)}$ and $\hat{P}_{i,\tau}^{(k)}$ approaches 0. This motivates us to propose the test statistic in the third step below.

Third, under the null hypothesis, we have $E\big(\hat{P}_{i,\tau}^{(g_i)} - \hat{P}_{i,\tau}^{(k)}\big) = 0$ and $Var\big(\hat{P}_{i,\tau}^{(g_i)} - \hat{P}_{i,\tau}^{(k)}\big) = \left(\frac{1}{n_{g_i}-1} + \frac{1}{n_k}\right)\big(p_{11,\tau} - p_{12,\tau}\big)$, where $p_{11,\tau} = p_{0,\tau}(1 - p_{0,\tau}), p_{0,\tau} = E\big\{I(\|X_i - X_j\| < \tau)\big\}, p_{12,\tau} = E(\delta_{ij,\tau}\delta_{im,\tau})$, and $\delta_{ij,\tau} = I(\|X_i - X_j\| < \tau) - p_{0,\tau}$ for any $i \neq j \neq m$. To improve the power of this test under sparse alternatives, we doubly maximize the squared difference between the two types of connection probabilities across all observations and samples. Accordingly, the test statistic is

$$T_\tau = \max_{1 \le i \le N} \max_{g_i \neq k} T_{i,\tau}^{(k)2} = \max_{1 \le i \le N} \max_{g_i \neq k} \frac{\left(\hat{P}_{i,\tau}^{(g_i)} - \hat{P}_{i,\tau}^{(k)}\right)^2}{\left(\frac{1}{n_{g_i}-1} + \frac{1}{n_k}\right)\left[\hat{p}_{11,\tau}^{(i)}(k) - \hat{p}_{12,\tau}^{(i)}(k)\right]}, \tag{2}$$

where $\hat{p}_{11,\tau}^{(i)}(k) = \hat{p}_{0,\tau}^{(i)}(k)\left[1 - \hat{p}_{0,\tau}^{(i)}(k)\right], \hat{p}_{0,\tau}^{(i)}(k) = \frac{1}{n_k + n_{g_i} - 1} \sum_{j \in \{C_{g_i}, C_k\}, i \neq j} A_{ij,\tau}$, and $\hat{p}_{12,\tau}^{(i)}(k) = \frac{1}{n_k + n_{g_i} - 1} \frac{1}{n_k + n_{g_i} - 2} \sum_{j \in \{C_{g_i}, C_k\}, i \neq j} \sum_{m \in \{C_{g_i}, C_k\}, j \neq m \neq i} (A_{ij,\tau} - \hat{p}_{0,\tau}^{(i)}(k))(A_{im,\tau} - \hat{p}_{0,\tau}^{(i)}(k))$. Note that $\hat{p}_{11,\tau}^{(i)}(k)$ and $\hat{p}_{12,\tau}^{(i)}(k)$ in the denominator of (2) are the estimators of $p_{11,\tau}$ and $p_{12,\tau}$, respectively, for each $i = 1, \ldots, N$ and $k = 1, \ldots, K$. Under the null hypothesis H_0, the test statistic T_τ is expected to be small. Conversely, the test statistic T_τ is expected to be large under the alternative hypothesis H_1.

To illustrate the construction of the proposed test statistic more clearly, we provide a straightforward computation based on a simple toy example. As depicted in Fig. 1, we first generate seven observations $X_i^{(k)}$ ($i = 1, \ldots, 7;\ k = 1, \ldots, 3$) from three different distributions, respectively. Then, for a given observation i that belongs to the third sample, the estimated probabilities of the "within" sample and the "between" samples are $\hat{P}_{i,\tau}^{(g_i)} = \frac{1}{2}$ and $\hat{P}_{i,\tau}^{(k)} = \frac{1}{7}$ for $k = 1, 2$. Furthermore, we obtain $\hat{p}_{0,\tau}^{(i)}(k) = \frac{4}{13}, \hat{p}_{11,\tau}^{(i)}(k) = \frac{36}{169}$, and $\hat{p}_{12,\tau}^{(i)}(k) = -\frac{3}{169}$ for $k = 1, 2$. Consequently, for this observation, we have $\max_{g_i \neq k} T_{i,\tau}^{(k)2} = \frac{25}{14}$. Repeating this calculation for all observations and taking the maximum across them yields the overall test statistic T_τ.

According to (2), the core of the proposed statistic T_τ is the maximum of the differences between all possible pairs of samples, rather than the sum of the differences between all possible pairs of samples. As a result, compared with other graph-based tests, the proposed test is expected to be powerful against sparse alternatives for multiple samples. It is worth noting that although we mainly focus on sparse alternatives, our theoretical results are applicable to non-sparse alternatives; see discussion in Section 2.4. Furthermore, the proposed test leverages the similarity graph to construct the test statistic. This implies that, in addition to Euclidean distance, alternative similarity metrics can be used to construct the graph among multivariate continuous observations. Besides, as long as pairwise distances between observations are well-defined, the proposed test is also applicable to various types of data, including non-Euclidean data and mixed continuous/discrete data. More details are provided in Section S.1 of the supplementary material.

2.2. Theoretical properties

Before establishing the theoretical properties of the proposed test statistic T_τ, we investigate the structure of the covariance matrix of $(T_{1,\tau},\dots,T_{L,\tau})^\top := (T_{1,\tau}^{(2)},\dots,T_{1,\tau}^{(K)},\dots,T_{N,\tau}^{(1)},\dots,T_{N,\tau}^{(K-1)})^\top$, where $L = N(K-1)$. To this end, we define

$$Q_{i,\tau}^{(k)} = \frac{\hat{P}_{i,\tau}^{(g_i)} - \hat{P}_{i,\tau}^{(k)}}{\left\{\left(\frac{1}{n_{g_i}-1}+\frac{1}{n_k}\right)(p_{11,\tau}-p_{12,\tau})\right\}^{1/2}} = \alpha_\tau(g_i,k)\left(\hat{P}_{i,\tau}^{(g_i)} - \hat{P}_{i,\tau}^{(k)}\right),$$

where $\alpha_\tau(g_i,k) = \left\{\left(\frac{1}{n_{g_i}-1}+\frac{1}{n_k}\right)(p_{11,\tau}-p_{12,\tau})\right\}^{-1/2}$. Then denote $(Q_{1,\tau},\dots,Q_{L,\tau})^\top := (Q_{1,\tau}^{(2)},\dots,Q_{1,\tau}^{(K)},\dots,Q_{N,\tau}^{(1)},\dots,Q_{N,\tau}^{(K-1)})^\top$, and let $\Sigma_\tau = (\sigma_{ij,\tau}) \in \mathbb{R}^{L\times L}$ be the covariance matrix of $(Q_{1,\tau},\dots,Q_{L,\tau})^\top$. Under the null hypothesis, the explicit expression of the covariance matrix Σ_τ is shown in Proposition 1.

Proposition 1. *Under H_0, the entries of the covariance matrix Σ_τ are given below.*

- *For any $1 \le i \le N$,*

$$cov(Q_{i,\tau}^{(k_1)}, Q_{i,\tau}^{(k_2)}) = \begin{cases} 1, & \text{if} \quad g_i \neq k_1 = k_2, \\ \alpha_\tau(g_i,k_1)\alpha_\tau(g_i,k_2)\frac{p_{11,\tau}-p_{12,\tau}}{n_{g_i}-1}, & \text{if} \quad g_i \neq k_1 \neq k_2. \end{cases}$$

- *For any $1 \le i,j \le N$ and $1 \le g_i = g_j \le K$,*

$$cov(Q_{i,\tau}^{(k_1)}, Q_{j,\tau}^{(k_2)}) = \begin{cases} \alpha_\tau(g_i,k_1)\alpha_\tau(g_i,k_2)\frac{p_{11,\tau}+(n_{g_i}-4)p_{12,\tau}}{(n_{g_i}-1)^2}, & \text{if} \quad g_i = g_j \neq k_1 \neq k_2, \\ \alpha_\tau(g_i,k_1)\alpha_\tau(g_i,k_1)\left[\frac{p_{11,\tau}+(n_{g_i}-4)p_{12,\tau}}{(n_{g_i}-1)^2}+\frac{p_{12,\tau}}{n_{k_1}}\right], & \text{if} \quad g_i = g_j \neq k_1 = k_2. \end{cases}$$

- *For any $1 \le i,j \le N$ and $1 \le g_i \neq g_j \le K$,*

$$cov(Q_{i,\tau}^{(k_1)}, Q_{j,\tau}^{(k_2)}) = \begin{cases} -\alpha_\tau(g_i,g_j)\alpha_\tau(g_j,k_2)\frac{p_{12,\tau}}{n_{g_j}}, & \text{if} \quad k_1 = g_j \neq g_i \neq k_2, \\ -\alpha_\tau(g_i,k_1)\alpha_\tau(g_j,g_i)\frac{p_{12,\tau}}{n_{g_i}}, & \text{if} \quad g_i = k_2 \neq k_1 \neq g_j, \\ \alpha_\tau(g_i,k_1)\alpha_\tau(g_j,k_1)\frac{p_{12,\tau}}{n_{k_1}}, & \text{if} \quad k_1 = k_2 \neq g_i \neq g_j, \\ \alpha_\tau(g_i,g_j)\alpha_\tau(g_j,g_i)\frac{p_{11,\tau}-(n_{g_i}+n_{g_j}+2)p_{12,\tau}}{n_{g_i}n_{g_j}}, & \text{if} \quad g_i = k_2 \neq k_1 = g_j, \\ 0, & \text{otherwise}. \end{cases}$$

Unfortunately, the covariance matrix Σ_τ is not positive definite, and the number of zero singular values is $(K-1)(K-2)/2$ for $K > 2$. For the sake of illustration, we consider the following example. Assume that $K = 3$ and $n_1 = n_2 = n_3 = N/3$. Then, the dimension of Σ_τ is $2N \times 2N$, and, for any $i \in C_1$, we have that $Q_{i,\tau}^{(2)} = \alpha_\tau\left(\hat{P}_{i,\tau}^{(1)} - \hat{P}_{i,\tau}^{(2)}\right)$ and $Q_{i,\tau}^{(3)} = \alpha_\tau\left(\hat{P}_{i,\tau}^{(1)} - \hat{P}_{i,\tau}^{(3)}\right)$, where $\alpha_\tau = \left\{\left(\frac{1}{N/3-1}+\frac{1}{N/3}\right)(p_{11,\tau}-p_{12,\tau})\right\}^{-1/2}$. Accordingly,

$$Q_{i,\tau}^{(2)} - Q_{i,\tau}^{(3)} = \alpha_\tau\left(\hat{P}_{i,\tau}^{(3)} - \hat{P}_{i,\tau}^{(2)}\right) = \frac{3\alpha_\tau}{N}\left(\sum_{j\in C_3} A_{ij,\tau} - \sum_{j\in C_2} A_{ij,\tau}\right)$$

and

$$\sum_{i\in C_1}\left(Q_{i,\tau}^{(2)} - Q_{i,\tau}^{(3)}\right) = \frac{3\alpha_\tau}{N}\left(\sum_{i\in C_1}\sum_{j\in C_3} A_{ij,\tau} - \sum_{i\in C_1}\sum_{j\in C_2} A_{ij,\tau}\right).$$

Analogously, we have

$$\sum_{i\in C_2}\left(Q_{i,\tau}^{(3)} - Q_{i,\tau}^{(1)}\right) = \frac{3\alpha_\tau}{N}\left(\sum_{i\in C_2}\sum_{j\in C_1} A_{ij,\tau} - \sum_{i\in C_2}\sum_{j\in C_3} A_{ij,\tau}\right)$$

and

$$\sum_{i\in C_3}\left(Q_{i,\tau}^{(1)} - Q_{i,\tau}^{(2)}\right) = \frac{3\alpha_\tau}{N}\left(\sum_{i\in C_3}\sum_{j\in C_2} A_{ij,\tau} - \sum_{i\in C_3}\sum_{j\in C_1} A_{ij,\tau}\right).$$

Since the adjacency matrix A_τ is symmetric, the above results imply that $\sum_{i\in C_1}\left(Q_{i,\tau}^{(2)} - Q_{i,\tau}^{(3)}\right) + \sum_{i\in C_2}\left(Q_{i,\tau}^{(3)} - Q_{i,\tau}^{(1)}\right) + \sum_{i\in C_3}\left(Q_{i,\tau}^{(1)} - Q_{i,\tau}^{(2)}\right) = 0$. Accordingly, there exists a collinearity problem among components $Q_{i,\tau}^{(k)}$ for any $1 \le k \neq g_i \le K$ and $i = 1,\dots,N$.

Suppose that the number of samples $K > 3$. Then, for any $1 \le k_1 \ne k_2 \ne k_3 \le K$, we can employ techniques similar to those used above to obtain the following result:

$$\sum_{i\in C_{k_1}} \left(Q_{i,\tau}^{(k_2)} - Q_{i,\tau}^{(k_3)}\right) + \sum_{i\in C_{k_2}} \left(Q_{i,\tau}^{(k_3)} - Q_{i,\tau}^{(k_1)}\right) + \sum_{i\in C_{k_3}} \left(Q_{i,\tau}^{(k_1)} - Q_{i,\tau}^{(k_2)}\right) = 0. \tag{3}$$

Thus, the collinearity problem exists. Note that if we set k_1 to 1, the number of equalities satisfying Eq. (3) is $(K-1)(K-2)/2$. This is because the number of combinations taken k_2 and k_3 from K is $(K-1)(K-2)/2$ when $1 = k_1 \ne k_2 \ne k_3 \le K$. This also explains why the number of zero singular values of matrix Σ_τ is $(K-1)(K-2)/2$.

To guarantee that the covariance matrix Σ_τ is positive definite, we adopt the following strategy of removing unnecessary components.

- For any $i_1 \in C_1$, we remove $Q_{i_1,\tau}^{(3)}, \ldots, Q_{i_1,\tau}^{(K)}$.
- For any $i_2 \in C_2$, we remove $Q_{i_2,\tau}^{(4)}, \ldots, Q_{i_2,\tau}^{(K)}$.
- ...
- For any $i_{K-2} \in C_{K-2}$, we remove $Q_{i_{K-2},\tau}^{(K)}$.

Although this strategy is not invariant to label permutations, it does not lead to any noticeable differences in practice, particularly when N is large. This observation is further supported by the simulation studies presented in Section S.4.6 of the supplementary material. Let $L^* = N(K-1) - (K-1)(K-2)/2$. Define $(\widetilde{Q}_{1,\tau}, \ldots, \widetilde{Q}_{L^*,\tau})^\top$ as the vector of remaining components, and denote $\Sigma_\tau^* = (\sigma_{ij,\tau}^*)$ as the covariance matrix of $(\widetilde{Q}_{1,\tau}, \ldots, \widetilde{Q}_{L^*,\tau})^\top$. After removing the linearly dependent components, we assume that the matrix Σ_τ^* is positive definite. Analogously, we obtain $(\widetilde{T}_{1,\tau}, \ldots, \widetilde{T}_{L^*,\tau})^\top$ after removing the corresponding redundant components. The resulting test is defined as

$$\widetilde{T}_\tau = \max_{1\le i\le L^*} \widetilde{T}_{i,\tau}^2, \tag{4}$$

and we name it the MPD test. To investigate the asymptotic distribution of MPD, we next introduce some technical conditions.

(C1) For any $1 \le k \le K$, there exists some constant γ_k such that $n_k/N \to \gamma_k$ as $N \to \infty$. In addition, there exist two constants $\gamma_{\min}$ and $\gamma_{\max}$ such that $0 < \gamma_{\min} \le \gamma_k \le \gamma_{\max} < 1$.

(C2) For any L^*, there exist some positive constants λ_1 and λ_2 such that $0 < \lambda_1 < \lambda_{\min}(\Sigma_\tau^*) < \lambda_{\max}(\Sigma_\tau^*) < \lambda_2$, where $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the minimum and maximum eigenvalues of the matrix A, respectively.

(C3) As $N \to \infty$, K is allowed to diverge as long as $\frac{K^3 \log^{28} N}{N} \to 0$.

Condition (C1) implies that the number of observations from different samples is approximately balanced. A similar condition is commonly used in the literature; see, e.g., Mukherjee et al. (2022) and Song and Chen (2022). Condition (C2) is mainly imposed to facilitate the theoretical analysis of the proposed test statistic. Under this condition, there does not exist redundant information among the remaining components $\widetilde{Q}_{l,\tau}$ ($l = 1, \ldots, L^*$). Similar to Kim (2021), Condition (C3) allows K to grow with N, but at a rate substantially slower than $N^{1/3}$. Such a restriction is required to control the Gaussian approximation error of the distribution of the proposed test statistic. Under these conditions, we have the following theorem.

Theorem 1. *Assume that Conditions (C1)–(C3) are satisfied, and the tuning parameter τ is selected such that $0 < p_{0,\tau} < 1$ and $p_{11,\tau} \ne p_{12,\tau}$. Under the null hypothesis H_0, we have that*

$$\sup_{x\in\mathbb{R}} \left| P\{\widetilde{T}_\tau / \log L^* < x\} - P\{\max_{1\le i\le L^*} Z_i^2 / \log L^* < x\} \right| \to 0,$$

as $N \to \infty$, where $Z = (Z_1, \ldots, Z_{L^*})^\top$ is a multivariate normally distributed random variable with mean zero and covariance matrix Σ_τ^*.

The proof of this theorem is given in Section S.3.4 of the supplementary material. To guarantee that $p_{0,\tau}$ remains non-degenerate, Theorem 1 imposes some mild conditions on the tuning parameter τ. These conditions are intuitive. For example, when $X_i, X_j \sim N(0, I_p)$, $\|X_i - X_j\|^2$ is approximately distributed as $N(2p, 8p)$ when p is large. Hence, to ensure that the probability $p_{0,\tau}$ remains non-degenerate, the tuning parameter τ should be chosen on the order of $\sqrt{p}$. In practice, τ can be selected based on empirical quantiles of the corresponding pairwise differences computed from the pooled observations. Further discussion on the selection of the optimal tuning parameter is provided in Section 2.3.

Based on Theorem 1, we can reject the null hypothesis at the significance level α if $\widetilde{T}_\tau \ge q_\alpha$, where q_α is the $(1-\alpha)$th quantile of the distribution $\max_{1\le i\le L^*} Z_i^2$. Since this distribution does not have a closed form, we can employ a numerical method to find the critical value. Specifically, we obtain the consistent estimator of the covariance matrix Σ_τ^*, by replacing $p_{0,\tau}, p_{11,\tau}$ and $p_{12,\tau}$ in Σ_τ^* with $\hat{p}_{0,\tau} = \{N(N-1)\}^{-1} \sum_{i=1}^N \sum_{j\ne i} A_{ij,\tau}$, $\hat{p}_{11,\tau} = \hat{p}_{0,\tau}(1-\hat{p}_{0,\tau})$, and $\hat{p}_{12,\tau} = \{N(N-1)(N-2)\}^{-1} \sum_{i=1}^N \sum_{j\ne i} \sum_{m\ne j\ne i} (A_{ij,\tau} - \hat{p}_{0,\tau})(A_{im,\tau} - \hat{p}_{0,\tau})$, respectively. Denote the resulting estimator by $\hat{\Sigma}_\tau^*$. We then generate M observations from the multivariate normal distribution with mean zero and covariance matrix $\hat{\Sigma}_\tau^*$. Denote the generated mth observation by $B^{(m)} = (B_1^{(m)}, \ldots, B_{L^*}^{(m)})^\top$ for $m = 1, \ldots, M$. Subsequently, we compute $S^{(m)} = \max_{1\le i\le L^*} B_i^{(m)2}$, and obtain the empirical critical value S_α, which is the $(1-\alpha)$th quantile of $S^{(1)}, \ldots, S^{(M)}$. Finally, we compare the test statistic $\widetilde{T}_\tau$ with the empirical critical value S_α, and draw the appropriate conclusion. Since this process is based on the asymptotic distribution of the MPD test in Theorem 1, we refer to it as the asymptotic MPD (aMPD) test hereafter.

It is worth noting that the convergence rate of the extreme value type tests may be slow (Lee et al., 2009). To boost the performance of the MPD test, we adopt a permutation procedure to determine the critical value. The detailed procedure is described as follows. At the bth step, we randomly permute the sample labels of all observations, and then calculate the test statistic $T^*_{b,\tau}$. We subsequently repeat the above procedure B times and denote the critical value t_α as the empirical $(1-\alpha)$ quantile of $T^*_{1,\tau}, \ldots, T^*_{B,\tau}$. Finally, we reject the null hypothesis if the test statistic $\widetilde{T}_\tau$ is greater than the critical value t_α.

This procedure is expected to control the size, as all observations are identically distributed under the null hypothesis and thus can be randomly relabeled. For convenience, we refer to this procedure as the permutation MPD (pMPD) test. Although the aMPD and pMPD tests adopt different strategies for determining the critical value, they are both fundamentally constructed upon the MPD statistic. For clarity and brevity, in the subsequent analysis we refer to them collectively as MPD, unless an explicit distinction is required.

Remark 1. To facilitate the derivation of the asymptotic distribution of the test statistic, we remove some redundant components when constructing the test statistics. Otherwise, the covariance matrix may become rank deficient, which would affect the implementation of the aMPD test. However, the validity of the permutation MPD test does not rely on whether these redundant components are removed. To examine whether this implementation choice affects finite-sample performance, we conducted additional simulation studies in Section S.4.7 of the supplementary material. The results indicate that removing the redundant components has little effect on the empirical size and power of the permutation-based test. Accordingly, for consistency, we remove the redundant components in both the aMPD and pMPD tests throughout the paper.

In addition to studying $\widetilde{T}_\tau$ under the null hypothesis, we next conduct a power analysis of the proposed test. To this end, we investigate the behavior of $\hat{p}^{(i)}_{0,\tau}(k)$ and $\hat{p}^{(i)}_{12,\tau}(k)$ included in the test statistic (4) under the alternative hypothesis. For any $i \in C_k, j \in C_l$ and $m \in C_s$, define $p_{kl,\tau} = E\{I(\|X_i - X_j\| < \tau)\}$ and $p_{kls,\tau} = E\{I(\|X_i - X_j\| < \tau) I(\|X_i - X_m\| < \tau)\}$. We then show the asymptotic properties of the means of $\hat{p}^{(i)}_{0,\tau}(k)$ and $\hat{p}^{(i)}_{12,\tau}(k)$ below.

Proposition 2. *For any $1 \le g_1, g_2, g_3 \le K$, define $\gamma^{g_3}_{g_1 g_2} = \frac{\gamma_{g_3}}{\gamma_{g_1} + \gamma_{g_2}}$. Under H_1, we have that*

$$
\begin{aligned}
E\{\hat{p}^{(i)}_{0,\tau}(k)\} \to \sigma^{(i)}_{0,\tau}(k) \;&:= \gamma^{g_i}_{g_i k} p_{g_i g_i,\tau} + \gamma^{k}_{g_i k} p_{g_i k,\tau},\\
\text{and } E\{\hat{p}^{(i)}_{12,\tau}(k)\} \to \sigma^{(i)}_{12,\tau}(k) \;&:= \frac{1}{n_{g_i} + n_k} \sum_{l \in \{g_i, k\}} \gamma^{l}_{g_i k} \Big\{ \sum_{s \in \{g_i, k\}} \gamma^{s}_{g_i k} p_{g_i l s,\tau} - p_{g_i l l,\tau} \\
&\quad + \sum_{s \in \{g_i, k\}} \gamma^{s}_{g_i k} p_{g_i l,\tau} p_{g_i s,\tau} - p^2_{g_i l,\tau} \Big\}.
\end{aligned}
$$

Under the alternative hypothesis, Proposition 2 demonstrates the asymptotic convergence of $E\{\hat{p}^{(i)}_{0,\tau}(k)\}$ and $E\{\hat{p}^{(i)}_{12,\tau}(k)\}$. To evaluate the differences between the "within" and the "between" probabilities for each observation i across samples g_i and k, we define $\sigma^{(i)}_{11,\tau}(k) = \sigma^{(i)}_{0,\tau}(k)(1 - \sigma^{(i)}_{0,\tau}(k))$ and

$$
\nu^{(i)}_\tau(k) = \frac{(p_{g_i g_i,\tau} - p_{g_i k,\tau})^2}{\left(\frac{1}{n_{g_i}-1} + \frac{1}{n_k}\right)\left[\sigma^{(i)}_{11,\tau}(k) - \sigma^{(i)}_{12,\tau}(k)\right]} \tag{5}
$$

It is worth noting that $\nu^{(i)}_\tau(k) = 0$ under the null hypothesis for any i and k. However, there exist some i and k such that $\nu^{(i)}_\tau(k) > 0$ under the alternative hypothesis. Accordingly, $\nu^{(i)}_\tau(k)$ can capture the discrepancy between the null and alternative hypotheses. Using Proposition 2 and (5), we are able to show the consistency of the MPD test below.

Theorem 2. *Assume that Conditions (C1)–(C3) are satisfied, and the tuning parameter τ is selected such that $0 < p_{0,\tau} < 1$ and $p_{11,\tau} \neq p_{12,\tau}$. Under the alternative hypothesis H_1, if $\max_{1 \le i \le N, 1 \le k \le K} \nu^{(i)}_\tau(k) > (4+\epsilon)\log L^*$ for some positive constant ϵ, then we have*

$$
P(\widetilde{T}_\tau > S_\alpha) \to 1,
$$

where S_α is the empirical critical value.

The proof of this theorem is given in Section S.3.5 of the supplementary material. Theorem 2 implies that the power of the proposed test converges to 1. To guarantee the consistency of the proposed test, Theorem 2 assumes that the maximum value of $\nu^{(i)}_\tau(k)$ is not too small. This assumption is sensible because it imposes a requirement on the strength of the difference between any two different samples. It is also important to note that Theorem 2 only imposes a mild condition on the maximum value of $\nu^{(i)}_\tau(k)$ for different i and k, rather than on the average value of $\nu^{(i)}_\tau(k)$. Based on the definition of $\nu^{(i)}_\tau(k)$, the value of $\max_{1 \le i \le N, 1 \le k \le K} \nu^{(i)}_\tau(k)$ can be relatively large even if most of the samples are homogeneous and only a few samples are heterogeneous. Accordingly, the MPD test is expected to perform well under sparse alternatives.

2.3. The choice of tuning parameter

Note that the result of Theorem 1 is valid for any τ. However, Theorem 2 shows that the power of the MPD test is associated with the tuning parameter τ. An inappropriate tuning parameter may lead T_τ to stay near zero even when the underlying distributions

of the K samples differ. This motivates us to select the optimal tuning parameter to maximize the power of the MPD test while maintaining the Type I error at a pre-specified significance level. Since it is a challenging task, this subsection explores a novel approach to choose the tuning parameter τ under some specific settings.

To simplify the analysis, we assume that all of the samples have the same number of observations, that is, $n_k = N/K \doteq n$ for every $1 \le k \le K$. For any $i \in C_k, j \in C_l$, let F_{kl} be the cumulative distribution function of $\|X_i - X_j\|^2$. Using the fact that $p_{kl,\tau} = E\{I(\|X_i - X_j\| < \tau)\}$ defined above Proposition 2, we have $F_{kl}(\tau^2) = p_{kl,\tau}$. In addition, denote the inverse function and the density function of F_{kl} by F_{kl}^{-1} and f_{kl}, respectively.

Under the null hypothesis, we have $F_{kk} = F_{kl}$ and $F_{kk}^{-1}F_{kl}(\tau^2) - \tau^2 = 0$ for any $1 \le k, l \le K$. In contrast, under the alternative hypothesis, there exists some l ($l \ne k$) such that $F_{kk} \ne F_{kl}$ and $F_{kk}^{-1}F_{kl}(\tau^2) - \tau^2 \ne 0$. To find the solution of the optimal tuning parameter, we consider a specific alternative hypothesis below.

$$H_1^* : F_{kk}^{-1}F_{kl}(\tau^2) - \tau^2 = \omega_{kl}\sqrt{K \log L^*/N},$$

for any τ and for some $l \ne k$, where ω_{kl} is a finite constant that does not depend on τ asymptotically. This hypothesis is quite common, as it encompasses a broad class of shift alternatives. Two illustrative scenarios are presented below, while the detailed analysis is provided in Section S.2 of the supplementary material.

Scenario 1: Mean Shift. Assume that $X_i^{(1)}$ $(i = 1, \dots, n)$ and $X_i^{(2)}$ $(i = 1, \dots, n)$ are independently generated from $N(0, I_p)$ and $N(\mu, I_p)$, respectively. Suppose that $p \to \infty$, $\tau^2 = 2p + O(\sqrt{p})$, and $\sqrt{p}\,\mu^2 \to 0$. A straightforward calculation shows that $F_{11}^{-1}\big(F_{12}(\tau^2)\big) - \tau^2 = -p\mu^2\{1 + o(1)\}$. Then the alternative hypothesis H_1^* holds if $\mu = O\Big(\frac{\log n}{np^2}\Big)^{1/4}$.

Scenario 2: Covariance Shift. Assume that $X_i^{(1)}$ $(i = 1, \dots, n)$ and $X_i^{(2)}$ $(i = 1, \dots, n)$ are randomly generated from the distributions $N(0, I_p)$ and $N(0, (1+\sigma^2)I_p)$, respectively. Suppose that $p \to \infty$, $\tau^2 = 2p + O(\sqrt{p})$, and $\sqrt{p}\,\sigma^2 \to 0$. Similarly, we obtain $F_{11}^{-1}\big(F_{12}(\tau^2)\big) - \tau^2 = -p\sigma^2\{1 + o(1)\}$. Therefore, the alternative hypothesis H_1^* holds if $\sigma^2 = O\Big(\sqrt{\frac{\log n}{np^2}}\Big)$.

Under the alternative hypothesis H_1^*, we have the following proposition.

Proposition 3. *Suppose that $n_k = n$ for all $1 \le k \le K$, and that f_{kl} is uniformly bounded for all $1 \le k, l \le K$. Under H_1^*, maximizing $\max_{1\le i\le N, 1\le k\le K} \nu_\tau^{(i)}(k)$ is asymptotically equivalent to maximizing the following function:*

$$\max_{1\le i\le N,\ 1\le k\le K} \frac{\omega_{g_i k}^2 f_{g_i g_i}^2(\tau^2)}{p_{g_i g_i,\tau}(1 - p_{g_i g_i,\tau}) - \bar{\sigma}_{12,\tau}^{(i)}(k)}, \tag{6}$$

where $\bar{\sigma}_{12,\tau}^{(i)}(k) = \frac{1}{8n}\Big(2p_{g_i g_i k,\tau} - p_{g_i g_i g_i,\tau} - p_{g_i k k,\tau}\Big)$.

The proof of this proposition is provided in Section S.3.3 of the supplementary material. To gain deeper insights into how the results of Proposition 3 inform the selection of the tuning parameter τ, we conduct a further analysis for Scenario 1. An analogous analysis can be conducted under the covariance shift alternatives.

Scenario 1: Mean Shift. Under this scenario, we have $F_{g_i g_i} = F$, $f_{g_i g_i} = f$ and $p_{g_i g_i,\tau} = F(\tau^2)$ for any i. Moreover, since the between-sample distance is symmetric, $\omega_{g_i k}$ is the same for all relevant pairs (g_i, k) and is asymptotically free of τ^2. In addition, we can show that $p_{g_i g_i k,\tau}, p_{g_i g_i g_i,\tau}, p_{g_i k k,\tau} = O(p^{-1})$, which implies that $\bar{\sigma}_{12,\tau}^{(i)}(k) \to 0$ as p diverges. Combining the above analysis, the function (6) can be simplified to

$$\mathcal{J}(\xi) = \frac{f^2(F^{-1}(\xi))}{\xi(1-\xi)},$$

where $\xi = F(\tau^2) \in (0, 1)$ and $\tau^2 = F^{-1}(\xi)$. Denote $\hat{\xi} = \operatorname{argmax}_{\xi\in(0,1)}\mathcal{J}(\xi)$. Under this scenario, we find that $\hat{\xi}$ is approximately 0.5. Accordingly, we can set the tuning parameter τ as the median of $\|X_i - X_j\|$ in this example.

Although this specific example only focuses on the comparison of two distributions, the above analyses can be straightforwardly extended to the K-sample ($K > 2$) problem by assuming that $X_i^{(k)}$ ($i = 1, \dots, n$ and $k = 1, \dots, K$) is generated from the multivariate normal distribution. In addition, this example leads us to recommend the median of $\|X_i - X_j\|$, computed from the pooled observations, as the tuning parameter τ in real practice. Simulation studies in Section 3 of the main paper and Section S.4 of the supplementary material support this recommendation. To assess the robustness of this choice, we also conduct simulation studies in Section S.4.5 of the supplementary material by considering three different τs (i.e., τ = 25%, 50% and 75% quantiles of $\|X_i - X_j\|$). Simulation results indicate that the three different τs yield comparable results, and the median of $\|X_i - X_j\|$ performs the best in most cases.

2.4. Comparison with the edge-count test

To further characterize the property of MPD, this subsection makes a power comparison between the MPD test and the edge-count test proposed by Chen and Friedman (2017) with $K = 2$. Such studies can be extended to $K > 2$ by comparing the MPD test and the generalized graph-based multisample test proposed by Song and Chen (2022).

The edge-count test statistic is defined as follows:

$$S_\tau = \big(R_1 - \mu_1, R_2 - \mu_2\big)\Sigma^{-1}\begin{pmatrix} R_1 - \mu_1 \\ R_2 - \mu_2 \end{pmatrix},$$

where $R_1 = \frac{1}{2}\sum_{i\in C_1}\sum_{j\in C_1} A_{ij,\tau}$, $R_2 = \frac{1}{2}\sum_{i\in C_2}\sum_{j\in C_2} A_{ij,\tau}$, $\mu_1 = E(R_1), \mu_2 = E(R_2)$ and Σ is the covariance matrix of the vector $(R_1, R_2)^\top$. For the sake of simplicity, we assume that $n_1 = n_2 = N/2 = n$. Denote G as the graph corresponding to adjacency matrix A and G_i as the subgraph in G that includes all edges that connect to observation X_i. Let $|G|$ and $|G_i|$ be the total number of edges in G and G_i, respectively. Based on Lemma 3.1 of Chen and Friedman (2017), we have $\lim_{N\to\infty}\mu_1 = \lim_{N\to\infty}\mu_2 = |G|/4$ and

$$\lim_{N\to\infty}\Sigma^{-1} = \frac{N}{4\Delta_2(N\Delta_1 - \Delta_2)}\begin{pmatrix} \Delta_1+\Delta_2 & -(\Delta_1-\Delta_2) \\ -(\Delta_1-\Delta_2) & \Delta_1+\Delta_2 \end{pmatrix},$$

where $\Delta_1 = |G| - \frac{2}{N^2}|G|^2$ and $\Delta_2 = \sum_{i=1}^N |G_i|^2 - \frac{4|G|^2}{N}$. For $i \in C_{g_i}$, define $p_{i,\tau}^{in} = \frac{1}{N}\sum_{j\in C_{g_i}} A_{ij,\tau}$ and $p_{i,\tau}^{out} = \frac{1}{N}\sum_{j\notin C_{g_i}} A_{ij,\tau}$. In order to assess the intrinsic difference between our proposed test and the edge-count test, we obtain the following property.

Lemma 1. *For the edge-count test S_τ, as $N \to \infty$, we have that*

$$\begin{aligned} lim_{N\to\infty} S_\tau = \frac{N^3}{64\Delta_2(N\Delta_1-\Delta_2)}\Big\{ &\Big(4\Delta_1+\Delta_2\Big)\sum_{i=1}^N (p_{i,\tau}^{in} - p_{i,\tau}^{out})^2 \\ &+ \Big(2\Delta_2 - 8\Delta_1\Big)\sum_{i\in C_1}\sum_{j\in C_2}(p_{i,\tau}^{in} - p_{i,\tau}^{out})(p_{j,\tau}^{in} - p_{j,\tau}^{out}) \\ &+ \Big(4\Delta_1+\Delta_2\Big)\sum_{i\in C_1}\sum_{i\neq j, j\in C_1}(p_{i,\tau}^{in} - p_{i,\tau}^{out})(p_{j,\tau}^{in} - p_{j,\tau}^{out}) \\ &+ \Big(4\Delta_1+\Delta_2\Big)\sum_{i\in C_2}\sum_{i\neq j, j\in C_2}(p_{i,\tau}^{in} - p_{i,\tau}^{out})(p_{j,\tau}^{in} - p_{j,\tau}^{out})\Big\}. \end{aligned}$$

The proof of the above lemma is given in Section S.3.1 of the supplementary material. According to Lemma 1, the test statistic S_τ can be regarded as the weighted summation of the elements $(p_{i,\tau}^{in} - p_{i,\tau}^{out})(p_{j,\tau}^{in} - p_{j,\tau}^{out})$ over $1 \le i, j \le N$. In particular, if the density of the graph G is 0.5, then $\Delta_1 = 0$ and we can obtain

$$S_\tau \propto \left[\sum_{i=1}^N (p_{i,\tau}^{in} - p_{i,\tau}^{out})\right]^2.$$

This indicates that the test statistic S is approximately proportional to the square of the summation of elements $p_{i,\tau}^{in} - p_{i,\tau}^{out}$ $(1 \le i \le N)$. In contrast, the MPD test can be regarded as the maximum of the elements $(p_{i,\tau}^{in} - p_{i,\tau}^{out})^2$ over $1 \le i \le N$. That is, the proposed test is based on the maximum difference between the "within" and the "between" probabilities across all observations. Hence, the proposed test is expected to exhibit high power when testing the small distributional difference between two samples.

To facilitate further analysis, we examine and compare the conditions necessary to guarantee the consistency of the MPD and S_τ tests across Scenarios 1 and 2.

Scenario 1: Mean Shift. Under this scenario, the MPD test requires that $\mu > C\{\log n/(pn)\}^{1/4}$, to satisfy the condition, $\max_{1\le i\le N, 1\le k\le K} \nu_\tau^{(i)}(k) > (4+\epsilon)\log L^*$, in Theorem 2. Conversely, the edge-count test statistic, S_τ, is consistent against all alternatives if $\mu > C/\sqrt{p}$ for some positive constant C. Hence, if $p \gg n$, the edge-count test enables us to detect subtle differences between two different means. However, if $p \ll n$, the MPD test only needs to impose a mild condition (i.e., $\mu > C\{\log n/(pn)\}^{1/4}$ given above) to detect mean differences. Accordingly, it can perform better than S_τ.

Scenario 2: Covariance Shift. Under this scenario, these two distributions differ in their covariances. If $\sigma^2 > C\{\log n/(pn)\}^{1/2}$, then the condition $\max_{1\le i\le N, 1\le k\le K} \nu_\tau^{(i)}(k) > (4+\epsilon)\log L^*$ holds in Theorem 2 for obtaining the consistency of MPD. To assure the S_τ test is consistent against all alternatives, it requires that $\sigma^2 > C/\sqrt{p}$ for some positive constant C. Accordingly, S_τ places a stricter requirement on the difference between two covariances. This finding is supported by simulation studies in Examples 5 and 6 of the supplementary material.

The detailed analysis is provided in Section S.2 of the supplementary material. In summary, under mean-shift alternatives, the MPD test tends to outperform the S_τ test when the sample size is large relative to the data dimension, whereas the edge-count test may be more powerful in the opposite regime. In addition, the proposed test is expected to exhibit higher power under the covariance shift alternatives and the distribution shift alternatives; see simulation results in Tables S.9 and S.10 in the supplementary material. As aforementioned, the above analyses can be extended to $K > 2$ by comparing the MPD test and the generalized graph-based multisample test.

3. Simulation studies

In this section, we conduct Monte Carlo experiments to illustrate the effectiveness of the proposed tests under various experiment settings. To make comparisons, we include our proposed tests (aMPD and pMPD) and the other four graph-based tests, which are the graph-based LP-nonparametric (GLP) test proposed by Mukhopadhyay and Wang (2020), the multisample crossmatch (MCM) and the multisample Mahalanobis crossmatch (MMCM) tests proposed by Mukherjee et al. (2022), and the generalized graph-based multisample (GGM) test proposed by Song and Chen (2022).

Note that the GLP test was implemented using the "LPKsample" R package, the MCM and MMCM tests were implemented by the "multicross" R package, and the GGM test was implemented by the "gTestsMulti" R package. Under the two-sample setting, the GGM

test is equivalent to the edge-count test proposed by Chen and Friedman (2017). According to Song and Chen (2022), the GGM test is based on a 5-MST. In the following experiments, the nominal level of the tests is set to 0.05, and each experiment is randomly replicated 500 times.

3.1. Experiment settings

This subsection employs the following example to compare the finite-sample performance of six tests under different dimensions. In this example, we consider three scenarios: mean shift, covariance shift, and distribution shift. Under the mean and covariance shifts, the data were generated from the multivariate normal distribution. Under the distribution shift, the data were simulated from multivariate t distributions. To demonstrate the effectiveness of the proposed test, sparse alternatives are designed to evaluate the empirical power of six tests. Specifically, for the K samples, we consider sparse alternatives $F_1 \neq (F_2 = \cdots = F_K)$ in our simulation studies. Additional experimental settings with sparse alternatives under different scenarios and dense alternatives are relegated to the supplementary material.

Example 1. In this example, there are $K = 5$ samples (groups) and five multivariate dimensions, $p = 5$, 25, 100, 200, and 400. In addition, three different alternatives are considered: mean shift, covariance shift, and distribution shift. In each setting, the sample sizes are $N = 250$ and 500.

- **Mean shift:** the mean varies among different groups. For $k = 1$, we randomly generate $X_i^{(k)}$ $(i = 1, \ldots, N/K)$ from a standard multivariate normal distribution $N(0, I_p)$. For $k = 2, \ldots, K$, $X_i^{(k)}$ $(i = 1, \ldots, N/K)$ is randomly generated from the distribution $N(\kappa_\mu, I_p)$. Here, the empirical sizes are evaluated at $\kappa_\mu = 0$, while the empirical power is calculated under $\kappa_\mu = 0.2$.
- **Covariance shift:** the covariance matrix varies among different groups. For $k = 1$, we randomly generate $X_i^{(k)}$ $(i = 1, \ldots, N/K)$ from a standard multivariate normal distribution $N(0, I_p)$. For $k = 2, \ldots, K$, $X_i^{(k)}$ $(i = 1, \ldots, N/K)$ is randomly simulated from the distribution $N(0, I_p + \kappa_\sigma 11^\top)$. The empirical sizes are evaluated at $\kappa_\sigma = 0$, while the empirical power is calculated under $\kappa_\sigma = 0.2$.
- **Distribution shift:** the distribution varies among different groups. For $k = 1$, we randomly generate $X_i^{(k)}$ $(i = 1, \ldots, N/K)$ from a standard multivariate normal distribution $N(0, I_p)$. For $k = 2, \ldots, K$, $X_i^{(k)}$ $(i = 1, \ldots, N/K)$ is randomly generated from a multivariate t-distribution, $t_{\kappa_f}(0, I_p)$, with mean 0, covariance I_p and degrees of freedom $\kappa_f = 20$. The empirical sizes are evaluated at $\kappa_f = \infty$, while the empirical power is calculated at $\kappa_f = 20$.

3.2. Experiment results

We next present the Monte Carlo experimental results of various tests based on 500 replications. The empirical sizes and power for Example 1 are summarized in Tables 1 to 2. Table 1 summarizes the empirical sizes of the six tests for Example 1. The results indicate that if the observations are generated from a multivariate normal distribution, the six tests generally control the empirical sizes well.

As for power analyses, Table 2 presents the empirical power of the six tests for Example 1 as the dimension p increases from 5 to 400. From Table 2, we can draw the following conclusions. First, regardless of the type of shift alternatives, almost all tests' power improves as the number of observations n increases. The results of our proposed tests are consistent with the theoretical finding in Theorem 2. Second, under the mean and covariance shift alternatives, the power of all six tests increases to 1 as the dimension p increases. This finding is not surprising since the differences among K groups become larger as the dimension p increases. Under the distribution shift alternatives, the proposed aMPD test, the pMPD test and the GLP test also show power improvement as the dimension p increases. However, the power of the MCM, MMCM and GGM tests exhibits either little improvement or reduction as the dimension p increases. This finding indicates that these three tests are not steady under distributional alternatives. Third, under the mean shift alternatives, the GLP and GGM tests perform better than the other four tests when p/n is not small. Otherwise, pMPD and aMPD are superior or comparable to their competitors; see, e.g., $p = 5$ and 25 under both $N = 250$ and 500. This finding is consistent with our discussion in Scenario 1 of Section 2.4. Note that, under the covariance shift alternatives, the pMPD and aMPD perform the best across all n and p, although GGM also exhibits good power. This finding supports our discussion in Scenario 2 of Section 2.4. As for the distribution shift scenario, the proposed pMPD and aMPD tests outperform other existing tests. Fourth, the power of the pMPD test is larger than that of the aMPD test when the number of observations is small regardless of the shift scenario. As the number of observations increases, the pMPD and aMPD tests yield similar power. This finding is expected since aMPD is evaluated under the asymptotic distribution of the MPD test in Theorem 1.

In sum, the simulation studies show that the pMPD and aMPD tests can control sizes well under all settings, and they have a power advantage under the covariance and distribution shift alternatives. We recommend using pMPD if the number of observations is not big. Otherwise, aMPD is recommended because it is computationally feasible and efficient. To save space, additional simulation studies for $K = 2$ are presented in Section S.4.3 of the supplementary material.

4. Real data analysis

Corporate credit ratings are valuable tools in the field of risk management because they allow investors and creditors to get a sense of a company's ability to pay its debts and then make risk-based lending decisions see, e.g., Huang et al. (2004), Wang and Ku (2021). Hence, it is critical to identify relevant features that influence credit rating scales see, e.g., Shin and Han (2001), Lee (2007),

Table 1
The empirical sizes for Example 1.

N	p	pMPD	aMPD	GLP	MCM	MMCM	GGM
250	5	0.038	0.026	0.058	0.062	0.046	0.054
	25	0.052	0.034	0.072	0.056	0.062	0.052
	100	0.076	0.046	0.063	0.066	0.044	0.052
	200	0.062	0.046	0.073	0.042	0.036	0.060
	400	0.054	0.038	0.067	0.058	0.028	0.042
500	5	0.062	0.040	0.037	0.066	0.046	0.050
	25	0.050	0.044	0.071	0.056	0.048	0.062
	100	0.044	0.038	0.048	0.056	0.044	0.066
	200	0.042	0.038	0.063	0.068	0.038	0.058
	400	0.052	0.032	0.055	0.046	0.046	0.050

Table 2
The empirical power of Example 1.

	N	p	pMPD	aMPD	GLP	MCM	MMCM	GGM
Mean Shift	250	5	0.130	0.092	0.105	0.104	0.044	0.096
		25	0.234	0.176	0.219	0.160	0.094	0.244
		100	0.392	0.308	0.629	0.298	0.216	0.700
		200	0.590	0.520	0.965	0.428	0.432	0.970
		400	0.872	0.798	1.000	0.640	0.792	1.000
	500	5	0.312	0.238	0.164	0.080	0.066	0.112
		25	0.642	0.588	0.432	0.200	0.124	0.470
		100	0.932	0.910	0.988	0.498	0.458	0.986
		200	0.996	0.990	1.000	0.772	0.862	1.000
		400	1.000	1.000	1.000	0.932	0.998	1.000
Covariance Shift	250	5	0.110	0.068	0.069	0.078	0.052	0.084
		25	0.464	0.356	0.219	0.144	0.090	0.306
		100	0.980	0.964	0.952	0.212	0.174	0.858
		200	1.000	1.000	0.966	0.336	0.262	0.980
		400	1.000	1.000	0.982	0.434	0.430	1.000
	500	5	0.200	0.146	0.092	0.094	0.052	0.122
		25	0.900	0.868	0.440	0.190	0.108	0.606
		100	1.000	1.000	1.000	0.422	0.394	0.994
		200	1.000	1.000	1.000	0.524	0.610	1.000
		400	1.000	1.000	1.000	0.722	0.832	1.000
Distribution Shift	250	5	0.050	0.024	0.074	0.066	0.052	0.056
		25	0.060	0.030	0.094	0.074	0.066	0.128
		100	0.534	0.378	0.160	0.080	0.048	0.186
		200	0.936	0.854	0.222	0.076	0.044	0.146
		400	1.000	1.000	0.341	0.074	0.056	0.084
	500	5	0.060	0.034	0.048	0.062	0.060	0.056
		25	0.142	0.104	0.136	0.058	0.036	0.296
		100	0.988	0.966	0.350	0.066	0.052	0.380
		200	1.000	1.000	0.595	0.096	0.042	0.208
		400	1.000	1.000	0.831	0.094	0.054	0.098

Hajek and Michalak (2013). However, this task is challenging due to the complex nature of the data. Some parametric methods, such as the classical multinomial logistic regression, face the risk of model misspecification. This necessitates the development of some robust non-parametric methods for feature selection. In this section, we follow the approach of Ghosh et al. (2021) and employ the proposed test to select relevant features that have the most influence on different credit rating scales.

To proceed with this empirical experiment, we collect the credit ratings of 258 companies listed on the New York Stock Exchange in 2022 from Standard and Poor's 500. The original rating of Standard and Poor's includes 22 scales, ranging from AAA to D. Figure S.1 in the supplementary material depicts the distribution of the credit rating levels of these 258 companies. As shown in Figure S.1, the credit ratings of most companies range from BBB to B. Few companies have credit ratings higher than A or lower than CCC. To avoid an imbalanced dataset, we reformulate the 22 rating levels into 5 groups so that the number of companies in each group is approximately balanced; see Table S.16 of the supplementary material. That is, the 258 companies are divided into five groups, which leads to $K = 5$ and $N = 258$ in our empirical analysis.

Motivated by Hajek and Michalak (2013), we collect 26 financial features (variables) of each company from Bloomberg, a global provider of business, financial, and news information. These features fall into six categories: size of a company, profitability ratios, activity ratios, liquidity ratios, leverage ratios, and market value ratios. The details of financial features are presented in Table S.17

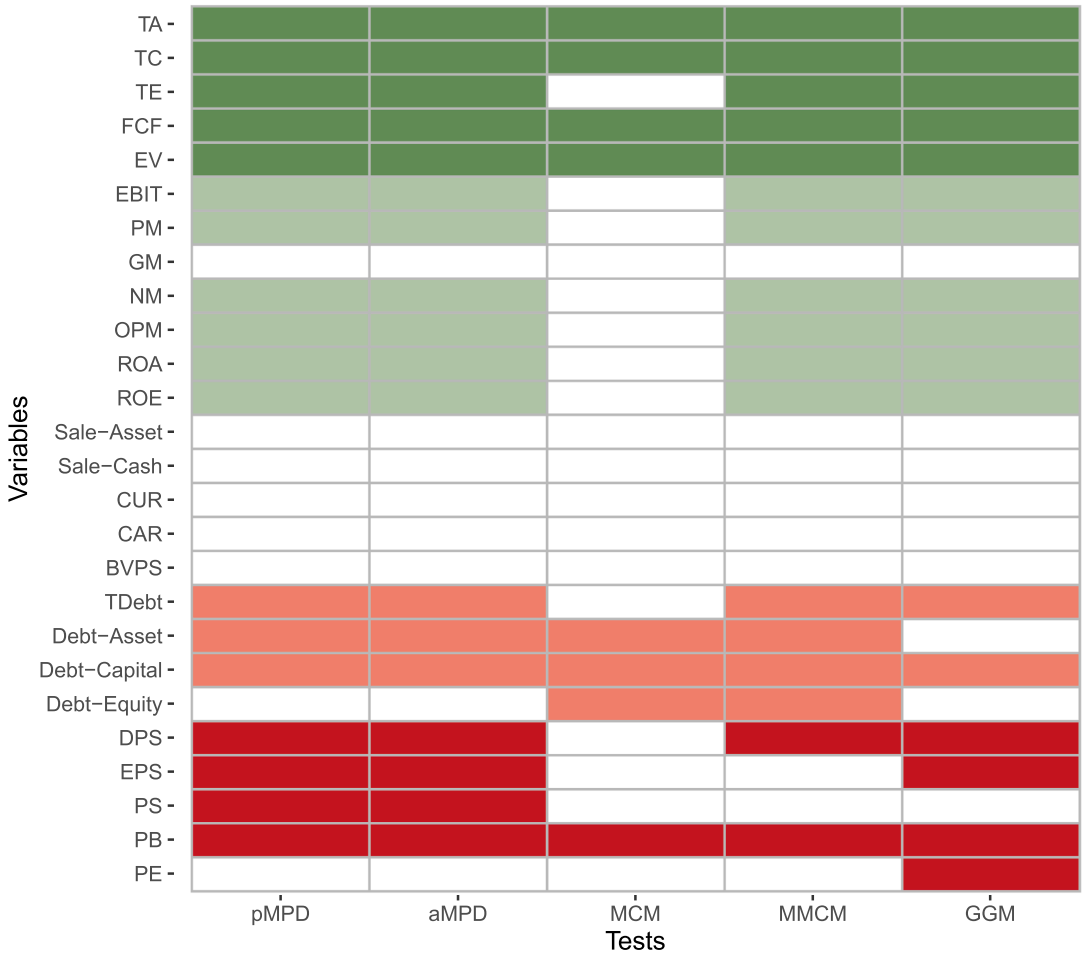

Fig. 2. The results of feature selection for various tests.

of the supplementary material. For each feature, the missing value is replaced by its mean value calculated from 2019 to 2021. In addition, each feature is centralized and standardized to have zero mean and unit variance. For brevity, let $X_i^{(k)} \in \mathbb{R}^{26}$ denote the financial feature vector associated with the ith company in the kth group.

We next apply the graph-based feature selection algorithm proposed by Ghosh et al. (2021), together with the proposed MPD test, to identify informative features that are related to corporate credit ratings. Specifically, we compute the 26×26 sample correlation matrix for the 26 features and apply single-linkage hierarchical agglomerative clustering to construct a dendrogram $\mathcal{T}$. The leaves of $\mathcal{T}$ correspond to individual features; ascending the tree sequentially merges them into progressively larger feature subsets until all features coalesce at the root.

Next, for each node of the tree $\mathcal{T}$, we apply the MPD test to the corresponding subset of features to assess whether the 258 companies in the five groups share the same distribution. If the null hypothesis is not rejected at a node, we treat the features in that node as uninformative and stop exploring that branch. If the null is rejected, we proceed to its child nodes and continue testing until we identify minimally significant subsets-nodes that are significant while their children are not. We carry out this top-down recursion over the tree and define the selected features as the union of those contained in all minimally significant nodes. By Theorem 1 of Ghosh et al. (2021), provided the underlying test is asymptotically consistent, this procedure recovers all truly contributing features with high probability and controls both the family-wise error rate and the false discovery rate at prespecified levels. Consequently, it allows us to select features that can accurately distinguish corporate credit ratings. The relevant features identified by the aMPD and pMPD tests are shown in Fig. 2.

For the sake of comparison, Fig. 2 also depicts the results selected by the MCM, MMCM and GGM tests. The GLP test is not included for comparison, since it does not control size well in some simulation studies; see Tables S.1 and S.2 of the supplementary material. From Fig. 2, we can draw the following two conclusions on feature selections. First, the features measuring the size of the company are selected by all of the tests. This finding implies that the credit rating of a company is closely related to its size. In addition to the size of the company, some financial indicators, which belong to the categories of profitability ratios, leverage ratios and market

Table 3
The p-values of the pMPD and aMPD tests in testing six different combinations of groups under the feature "PS".

Groups	{1,2,3,4,5}	{1,2,3,4}	{1,2,3,5}	{1,2,4,5}	{1,3,4,5}	{2,3,4,5}
pMPD	<0.001	0.002	<0.001	<0.001	<0.001	0.207
aMPD	<0.001	0.010	<0.001	<0.001	<0.001	0.648

value ratios, are also selected. However, none of the five tests select features in the categories of activity ratios and liquidity ratios. This suggests that the influence of those features on the credit rating of a company is weak and non-significant.

Second, the aMPD and pMPD tests yield the same results in terms of feature selections. Specifically, these two tests select 18 important features out of 26 features. Thus, they select more features than the other three tests; see Fig. 2. For example, the aMPD and pMPD tests select the feature "PS", while the other tests leave it out. To explore this finding, we apply the proposed test to investigate whether the feature "PS" exhibits homogeneity starting from the five groups. The results are shown in Table 3, which indicates that the "PS" exhibits distributional homogeneity among Groups {2, 3, 4, 5}, except for Group 1. Accordingly, the feature "PS" plays an important role in distinguishing Group 1 and the remaining groups. An analogous finding is on the feature "Debt-Asset". This is because, based on this feature, the aMPD and pMPD tests show that there exist significant differences among Groups {1,2,3,4,5}, but not among Groups {3,4,5}. As shown in Fig. 2, the proposed test can successfully select the relevant feature "Debt-Asset", while the GGM test cannot. The above two illustrations indicate that our proposed MPD test is powerful in detecting the small differences of features among different groups.

In addition, we compare the variable selection results obtained from our proposed procedures with those produced by the lasso and random forest; see Section S.5.3 of the supplementary material for additional details. The results show that the lasso identifies six predictors, all of which are also selected by the pMPD and aMPD tests. Similarly, the ten most important predictors identified by the random forest are likewise included in the sets selected by pMPD and aMPD. These findings suggest that the proposed methods are able to effectively identify the key features driving the differences across credit-rating categories.

5. Concluding remarks

This paper proposes a graph-based maximum pairwise difference test, MPD, to address the K-sample problem. Specifically, we construct a similarity graph among the pooled observations, and then doubly maximize the squared differences between the within and between probabilities of different samples for each observation. We not only derive the asymptotic distribution of the MPD test under the null hypothesis, but also establish its consistency under general alternatives. In addition, we demonstrate that the proposed MPD test is superior to other known multisample tests via extensive simulation studies and an empirical example.

Several potential topics can be considered for future research. First, extend the MPD test to handle complex datasets, such as panel count data see, e.g., Balakrishnan and Zhao (2009). Second, generalize the MPD test to change point detection see, e.g., Chen and Zhang (2015). We believe these efforts would expand the usefulness of our proposed test.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors are grateful to the editor, associate editor, and anonymous referees for their insightful comments and constructive suggestions. Dan Pu's research is supported by the National Natural Science Foundation of China (72403202). Wei Lan's research is supported by the National Key R&D Program of China (2022YFA1003702), the National Natural Science Foundation of China (72422020, 72333001, 12531011), and the Joint Lab of Data Science and Business Intelligence at Southwestern University of Finance and Economics.

Supplementary material

Supplementary material associated with this article can be found in the online version at 10.1016/j.jeconom.2026.106313.

References

Balakrishnan, N., Zhao, X., 2009. New multi-sample nonparametric tests for panel count data. Ann. Stat. 37 (3), 1112–1149.
Bhattacharya, B.B., 2019. A general asymptotic framework for distribution-free graph-based two-sample tests. J. R. Stat. Soc. Ser. B Stat. Methodol. 81 (3), 575–602.
Bibby, J.A., Agarwal, D., Freiwald, T., Kunz, N., Merle, N.S., West, E.E., Singh, P., Larochelle, A., Chinian, F., Mukherjee, S., et al., 2022. Systematic single-cell pathway analysis to characterize early t cell activation. Cell Rep. 41 (8), 111697.
Bickel, P.J., 1969. A distribution free version of the smirnov two sample test in the p-variate case. Ann. Math. Stat. 40 (1), 1–23.
Cai, S., Chen, J., Zidek, J.V., 2017. Hypothesis testing in the presence of multiple samples under density ratio models. Stat. Sin. 27 (2), 761–783.

Cai, T.T., Liu, W., Xia, Y., 2014. Two-sample test of high dimensional means under dependence. J. R. Stat. Soc. Ser. B Stat. Methodol. 76 (2), 349–372.
Chen, H., Chen, X., Su, Y., 2018. A weighted edge-count two-sample test for multivariate and object data. J. Am. Stat. Assoc. 113 (523), 1146–1155.
Chen, H., Friedman, J.H., 2017. A new graph-based two-sample test for multivariate and object data. J. Am. Stat. Assoc. 112 (517), 397–409.
Chen, H., Zhang, N., 2015. Graph-based change-point detection. Ann. Stat. 43 (1), 139–176.
Chen, S., 2020. A new distribution-free k-sample test: analysis of kernel density functionals. Can. J. Stat. 48 (2), 167–186.
Fan, J., Liao, Y., Yao, J., 2015. Power enhancement in high-dimensional cross-sectional tests. Econometrica 83 (4), 1497–1541.
Feng, L., Lan, W., Liu, B., Ma, Y., 2022. High-dimensional test for alpha in linear factor pricing models with sparse alternatives. J. Econom. 229 (1), 152–175.
Friedman, J.H., Rafsky, L.C., 1979. Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. Ann. Stat. 7 (4), 697–717.
Ghosh, S., Mukherjee, S., Agarwal, D., 2021. Feature selection in high-dimensional space using graph-based methods. arXiv preprint arXiv:2108.12682.
Hajek, P., Michalak, K., 2013. Feature selection in corporate credit rating prediction. Knowl. Based Syst. 51 (1), 72–84.
Hassani, H., Silva, E.S., 2015. A Kolmogorov-Smirnov based test for comparing the predictive accuracy of two sets of forecasts. Econometrics 3 (3), 590–609.
Heller, R., Heller, Y., Kaufman, S., Brill, B., Gorfine, M., 2016. Consistent distribution-free k-sample and independence tests for univariate random variables. J. Mach. Learn. Res. 17 (29), 1–54.
Henze, N., 1988. A multivariate two-sample test based on the number of nearest neighbor type coincidences. Ann. Stat. 16 (2), 772–783.
Huang, Z., Chen, H., Hsu, C.J., Chen, W.H., Wu, S., 2004. Credit rating analysis with support vector machines and neural networks: a market comparative study. Decis. Support Syst. 37 (4), 543–558.
Hur, Y., Liang, T., 2025. Detecting weak distribution shifts via displacement interpolation. J. Bus. Econ. Stat. 43 (1), 178–190.
Kim, I., 2021. Comparing a large number of multivariate distributions. Bernoulli 27 (1), 419–441.
Kossinets, G., Watts, D.J., 2006. Empirical analysis of an evolving social network. Science 311 (5757), 88–90.
Kruskal, W.H., 1952. A nonparametric test for the several sample problem. Ann. Math. Stat. 23 (4), 525–540.
Lee, S., Linton, O., Whang, Y.J., 2009. Testing for stochastic monotonicity. Econometrica 77 (2), 585–602.
Lee, Y.C., 2007. Application of support vector machines to corporate credit rating prediction. Expert Syst. Appl. 33 (1), 67–74.
Lopes, M., Jacob, L., Wainwright, M.J., 2011. A more powerful two-sample test in high dimensions using random projection. Adv. Neural Inf. Process. Syst. 24, 1206–1214.
Mukherjee, S., Agarwal, D., Zhang, N.R., Bhattacharya, B.B., 2022. Distribution-free multisample tests based on optimal matchings with applications to single cell genomics. J. Am. Stat. Assoc. 117 (538), 627–638.
Mukhopadhyay, S., Wang, K., 2020. A nonparametric approach to high-dimensional k-sample comparison problems. Biometrika 107 (3), 555–572.
Nettleton, D., Banerjee, T., 2001. Testing the equality of distributions of random vectors with categorical components. Comput. Stat. Data Anal. 37 (2), 195–208.
Oja, H., Randles, R.H., 2004. Multivariate nonparametric tests. Stat. Sci. 19 (4), 598–605.
Petrie, A., 2016. Graph-theoretic multisample tests of equality in distribution for high dimensional data. Comput. Stat. Data Anal. 96, 145–158.
Rosenbaum, P.R., 2005. An exact distribution-free test comparing two multivariate distributions based on adjacency. J. R. Stat. Soc. Ser. B Stat. Methodol. 67 (4), 515–530.
Schilling, M.F., 1986. Multivariate two-sample tests based on nearest neighbors. J. Am. Stat. Assoc. 81 (395), 799–806.
Shin, K.S., Han, I., 2001. A case-based approach using inductive indexing for corporate bond rating. Decis. Support Syst. 32 (1), 41–52.
Song, H., Chen, H., 2022. New graph-based multi-sample tests for high-dimensional and non-euclidean data. arXiv preprint arXiv:2205.13787.
Wang, J.L., Chiou, J.M., Müller, H.G., 2016. Functional data analysis. Annu. Rev. Stat. Appl. 3, 257–295.
Wang, M., Ku, H., 2021. Utilizing historical data for corporate credit rating assessment. Expert Syst. Appl. 165, 113925.
Zhang, J., Wu, Y., 2007. K-sample tests based on the likelihood ratio. Comput. Stat. Data Anal. 51 (9), 4682–4691.
Zhang, J.T., Guo, J., Zhou, B., 2024. Testing equality of several distributions in separable metric spaces: a maximum mean discrepancy based approach. J. Econom. 239 (2), 105286.