RISE: Rank in Similarity Graph Edge-Count Two-Sample Test

Doudou Zhou and Hao Chen
Department of Statistics, University of California, Davis

Abstract

Two-sample hypothesis testing for high-dimensional data is ubiquitous nowadays. Rank-based tests are popular nonparametric methods for univariate data. However, they are difficult to be extended to high-dimensional data. In this paper, we propose a new non-parametric two-sample testing procedure, RISE: Rank in Similarity graph Edge-count two-sample test. The new test statistic is constructed on a rank-weighted similarity graph, such as the $k$-nearest neighbor graph. As a result, RISE can also be applied to non-Euclidean data. Theoretically, we prove that, under some mild conditions, the new test statistic converges to the $\chi^2$ distribution under the permutation null distribution, enabling a fast type-I error control. RISE exhibits good power under a wide range of alternatives compared to existing methods, as shown in extensive simulations. The new test is illustrated on the New York City taxi data for comparing travel patterns in consecutive months and a brain network dataset in comparing male and female subjects.

Keywords: Non-parametric two-sample test; rank-based method; similarity graph; high-dimensional data; non-Euclidean data
1 Introduction

For two independent random samples $X_1, \ldots, X_m \overset{i.i.d.}{\sim} F_X$ and $Y_1, \ldots, Y_n \overset{i.i.d.}{\sim} F_Y$, we consider the test

$$H_0 : F_X = F_Y \quad \text{against} \quad H_1 : F_X \neq F_Y.$$ 

Nowadays, it is common that the data is high-dimensional or non-Euclidean (Bullmore and Sporns, 2012; Tian et al., 2016; Menafoglio and Secchi, 2017; Jiang et al., 2020). In many of these problems, one has little information on $F_X$ and $F_Y$, which makes parametric approaches not applicable when the dimension is high. A number of nonparametric tests have been proposed for high-dimensional data such as the graph-based tests (Friedman and Rafsky, 1979; Schilling, 1986; Henze, 1988; Rosenbaum, 2005; Chen and Zhang, 2013; Chen and Friedman, 2017; Chen et al., 2018; Zhang and Chen, 2022), the classification-based tests (Hediger et al., 2019; Lopez-Paz and Oquab, 2016; Kim et al., 2021), the interpoint distances-based tests (Székely and Rizzo, 2013; Biswas and Ghosh, 2014; Li, 2018), and the kernel-based tests (Gretton et al., 2008; Harchaoui et al., 2007; Gretton et al., 2009, 2012; Song and Chen, 2020).

For non-parametric testing, rank-based tests are popular to approach given the success of the Wilcoxon’s rank-sum test (Wilcoxon, 1945) for univariate data. However, the rank for multivariate data is hard to define. There are some extended definitions of the rank to accommodate multivariate data, such as the spatial rank (Chaudhuri, 1996; Marden, 1999), the Mahalanobis rank (Hallin and Paindaveine, 2004, 2006), and the component-wise rank (Bickel, 1965; Puri and Sen, 2013). For instance, Oja (2010) proposed the multivariate spatial signs and ranks, which can be applied to construct a multivariate affine-invariant family of rank tests for the detection of the location difference. Based on the data depth rank, Liu and Singh (1993) proposed tests as a multivariate analog of Wilcoxon’s rank-sum
test, and Barale and Shirke (2021) proposed a test that worked both for location and scale difference. However, these tests are mainly for low-dimensional data.

Recently, Pan et al. (2018) introduced Ball Divergence (BD) to measure the the difference between two distributions and proposed a metric rank test procedure. Deb and Sen (2021) proposed to define the multivariate ranks through the theory of measure transportation (Hallin et al., 2021), based on which they built the multivariate rank-based distribution-free nonparametric testing. Both tests can be applied to high-dimensional data and achieved good performance for some useful settings. However, they also lose power under some common alternatives, which will be detailed in Section 4. Besides, they did not provide any analytic $p$-value approximations and relied on random permutations to obtain their $p$-values.

In this paper, we propose a new framework of two-sample testing procedure, Rank In Similarity graph Edge-Count two-sample test (RISE), which overcomes the curse of dimensionality (Chen and Friedman, 2017) and enables a fast type-I error control. Instead of dealing with the ranks of observations, we consider two types of ranks based on the similarity graph of the observations, the graph-induced rank defined by the inductive nature of the graph and the overall rank defined by the weight of edges in the graph. The similarity graph can be built from the pairwise similarity of observations, such as the $k$-nearest neighbor graph ($k$-NNG) (Henze, 1988) and the $k$-minimum spanning tree ($k$-MST)$^1$ (Friedman and Rafsky, 1979). As a result, our framework is applicable to non-Euclidean data as well.

Test statistics based on similarity graphs have attracted a lot of attention recently as $^1$The MST is a spanning tree connecting all observations while minimizing the sum of distances of edges in the tree. The $k$-MST is the union of the 1st, . . . , $k$th MSTs, where the $k$th MST is a spanning tree that connects all observations while minimizing the sum of distances across edges excluding edges in the $(k - 1)$-MST.
they can be applied to data with an arbitrary dimension and non-Euclidean data and perform well. The first test of this type was proposed in Friedman and Rafsky (1979) using the $k$-MST, later Schilling (1986) and Henze (1988) used the $k$-NNG, and Rosenbaum (2005) proposed to use the minimum distance non-bipartite pairing graph\(^2\) (MDP) to obtain an exact distribution-free test, which was extended to $k$-MDP\(^3\) in Chen and Friedman (2017). Recently, Chen and Friedman (2017) proposed a new test statistic, the generalized edge-count test (GET), on similarity graphs that utilizes a common pattern for high-dimensional data, and the test works well for a variety of alternatives.

The current graph-based tests treat each edge in the graph equally and ignore the differences on edges (Friedman and Rafsky, 1979; Henze, 1988; Chen and Friedman, 2017), which could lose power. There were attempts to use ranks in earlier studies (Schilling, 1986; Rosenbaum, 2005), but these tests lack power for high-dimensional data under some common alternatives. RISE solves the problems by incorporating weights on the edges of the similarity graphs and proposing a Mahalanobis-type statistic that works well for a variety of settings where existing methods work poorly.

The rest of the paper is organized as follows. In Section 2, we introduce in detail the new test statistic $T_R$ with its moment properties. The asymptotic property of $T_R$ is presented in Section 3. Extensive simulations are conducted in Section 4 with real data applications presented in Section 5. The details of proofs of the theorems are deferred to Supplementary Materials.

\(^2\)The MDP is constructed by dividing the $N$ observations into $N/2$ (assuming $N$ is even) non-overlapping pairs while minimizing the $N/2$ distances within pairs.

\(^3\)The $k$-MDP is the union of the 1st, . . . , $k$th MDPs, where the $k$th MDP is a minimum distance non-bipartite pairing while minimizing the sum of distances within pairs excluding the pairs in the $(k-1)$-MDP.
2 A new test statistic

2.1 Graph-based ranks

To simplify the notations, let

\[ Z_i = X_i, i = 1, \ldots, m; \quad Z_{m+j} = Y_j, j = 1, \ldots, n \]

be the pooled samples and \( N = m + n \). Let \( \{G_l\}_{l=1}^k \) be a sequence of similarity graphs with nodes \( \{Z_i\}_{i=1}^N \) and edges constructed by some optimization criteria in an inductive way such that

\[ G_{l+1} = G_l \cup G'_l \text{ with } G'_l = \arg \max_{G' \subset G_l} \sum_{(i,j) \in G'} S(Z_i, Z_j), \]

where \( G' \) subject to \( G' \cap G_l = \emptyset \) and some additional constraints. Here \( S(\cdot, \cdot) \) is some similarity measure, for example, \( S(Z_i, Z_j) = -\|Z_i - Z_j\| \) for Euclidean data, where \( \| \cdot \| \) is the Euclidean norm. For other choices of the similarity measures, see Chen and Zhang (2013); Sarkar and Ghosh (2018); Sarkar et al. (2020). By construction, we have \( G_1 \subset G_2 \ldots \subset G_k \). Many widely used similarity graphs can be constructed in this way with different constraints, for example,

- **k-NNG**: for each \( i \), there exists one and only one \( j \neq i \) such that \( (i,j) \in G' \);
- **k-MST**: \( G' \) should be a tree that connects all observations;
- **k-MDP**: \( G' \) should be a non-bipartite pairing;
- **k-SHP (Biswas et al., 2014)**: \( G' \) is a Hamiltonian path which visits each vertex exactly once, that is, connected and acyclic with \( N - 1 \) edges, where each node has degree at most two.

With \( \{G_l\}_{l=1}^k \), we define two types of graph-based rank matrix \( R = (R_{ij})_{i,j=1}^N \in \mathbb{R}^{N \times N} \) as follows:
• Graph-induced rank

\[ R_{ij} = \sum_{l=1}^{k} \mathbb{I}(i, j) \in G_l). \]

• Overall rank

\[ R_{ij} = \text{rank}(S(Z_i, Z_j), G_k), \]

where \( \text{rank}(S(Z_i, Z_j), G_k) \) is the rank of \( S(Z_i, Z_j) \) among \( \{S(Z_u, Z_v)\}_{(u,v) \in G_k} \) if \( (i, j) \in G_k \) and is zero if \( (i, j) \notin G_k \).

**Remark 2.1.** The two types of ranks are intuitive. For instance, the graph-induced rank of edges in the \( l \)th NNG or the \( l \)th MST will be \( k-l+1 \) for \( k \)-NNG and \( k \)-MST, respectively and the overall rank of edges will be the rank of the similarity of edges in the graph. Instead of rank, other choices of weights can also be considered. For instance, the interpoint distance or the kernel values. By incorporating different weights, the performance of the test can be different. In this work, we focus on ranks.

**Remark 2.2.** The rank used in BD (Pan et al., 2018) has some common grounds with our graph-induced rank on the \( k \)-NNG. They both utilize the rank of pairwise similarity. However, in BD, for each within-sample pairwise similarity \( S(Z_i, Z_j) \), they consider its rank among \( \{S(Z_i, Z_u)\}_{u=1}^{m} \) and \( \{S(Z_i, Z_u)\}_{u=m+1}^{N} \), respectively, and compare their difference. On the other hand, for each \( Z_i \), the graph-induced rank on \( k \)-NNG considers the ranks for \( \{S(Z_i, Z_u)\}_{u=1}^{N} \) and only the top \( k \) similarities are kept.

We first define two basic quantities based on the graph-based rank:

\[ U_x = \sum_{i=1}^{m} \sum_{j=1}^{m} R_{ij} \quad \text{and} \quad U_y = \sum_{i=m+1}^{N} \sum_{j=m+1}^{N} R_{ij}. \]

The proposed test statistic is defined as

\[ T_R = (U_x - \mu_x, U_y - \mu_y) \Sigma^{-1}(U_x - \mu_x, U_y - \mu_y)^\top, \]
where $\mu_x = \mathbb{E}(U_x)$, $\mu_y = \mathbb{E}(U_y)$ and $\Sigma = \text{Cov}((U_x, U_y)^T)$. We use $\mathbb{P}$, $\mathbb{E}$, Var, Cov to denote the probability, expectation, variance, and covariance under the permutation null distribution, respectively, which places $1/(N^m)$ probability on each of the $\binom{N}{m}$ permutations of the group labels where the first group has $m$ observations and the second group has $n$ observations.

Remark 2.3. The new test statistic $T_R$ is similar to GET in terms of its formula while GET treats each edge in the similarity graph equally. Actually, when the weights on the similarity graph are all set to be one, $T_R$ becomes GET when the similarity graph is undirected, and becomes the directed version of GET (Chu and Chen, 2018; Liu and Chen, 2022) when the similarity graph is directed. For GET, Chen and Friedman (2017) discussed that under the alternative hypothesis, there are two possible scenarios that (i) both samples tend to connect within each other within samples and (ii) one sample tends to connect within sample while the other sample tends to connect between sample. Similarly, for our rank quantities, under the alternative hypothesis, we also have two possible scenarios that (i) both $U_x$ and $U_y$ tend to be large and (ii) one of them tends to be large while the other one tends to be small. Hence, $T_R$ can capture the two different type of scenarios and is powerful for a wider range of alternatives.

2.2 Moment properties

We first symmetrize $R$ by using $\frac{1}{2}(R + R^T)$. With a slight notation abuse, in the following, $R$ is used for the symmetric version. This does not change the values of $U_x$ and $U_y$ by their definitions; while the derivation for their expectations and variances would be much simpler. Let $g_i = 1$ if the $i$th sample is from $F_X$ and $g_i = 0$ if from $F_Y$. Then $U_x$ and $U_y$
can be rewritten as
\[ U_x = \sum_{i=1}^{N} \sum_{j=1}^{N} g_i g_j R_{ij} \quad \text{and} \quad U_y = \sum_{i=1}^{N} \sum_{j=1}^{N} (1 - g_i)(1 - g_j) R_{ij}. \]

To simplify the notations, we further denote
\[ R_i = \sum_{j=1}^{N} R_{ij}, \quad \bar{R}_i = \frac{R_i}{N-1}, \quad r_0 = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} R_{ij}, \]
\[ r_1^2 = \frac{1}{N} \sum_{i=1}^{N} \bar{R}_{i}^2, \quad \text{and} \quad r_d^2 = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} R_{ij}^2. \]

**Theorem 2.1.** Under the permutation null distribution, we have that
\[
\mathbb{E}(U_x) = m(m-1)r_0, \quad \mathbb{E}(U_y) = n(n-1)r_0
\]
\[
\text{Var}(U_x) = \frac{2mn(m-1)}{(N-2)(N-3)} \left((n-1)(r_d^2 - r_0^2) + 2(m-2)(N-1)(r_1^2 - r_0^2)\right),
\]
\[
\text{Var}(U_y) = \frac{2mn(n-1)}{(N-2)(N-3)} \left((m-1)(r_d^2 - r_0^2) + 2(n-2)(N-1)(r_1^2 - r_0^2)\right),
\]
\[
\text{Cov}(U_x, U_y) = \frac{2m(m-1)n(n-1)}{(N-2)(N-3)} \left((r_d^2 - r_0^2) - 2(N-1)(r_1^2 - r_0^2)\right).
\]

The proof of Theorem 2.1 is in the Supplement. To assure that $T_R$ is well-defined, the covariance matrix $\Sigma$ should be invertible. Here we present the sufficient and necessary conditions.

**Theorem 2.2.** Given $m, n \geq 2$, $T_R$ is well-defined unless one of the following two cases happens:

(C1) $r_1^2 = r_0^2$;

(C2) $(N-2)(r_d^2 - r_0^2) = 2(N-1)(r_1^2 - r_0^2)$.

The proof of Theorem 2.2 is in the Supplement. In the following, we briefly discuss the two cases. By definition,
\[
r_0 = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} R_{ij} = \frac{1}{N} \sum_{i=1}^{N} \bar{R}_i.
\]
is the average of both \( \{ R_{ij} \}_{i,j=1}^N \) and \( \{ \bar{R}_i \}_{i=1}^N \), so \( r_d^2 - r_0^2 \) is the variance of \( \{ R_{ij} \}_{i,j=1}^N \) and \( r_1^2 - r_0^2 \) is the variance of \( \{ \bar{R}_i \}_{i=1}^N \). For instance, the graph-induced rank on the \( k \)-MDP satisfies (C1) as all nodes are required to have the exact same degree \( k \) for the \( k \)-MDP graph and thus \( \bar{R}_i = \frac{(1+k)k}{2(N-1)} \) for all \( i \). Except for such special graphs, it is rare to have graphs that satisfy (C1) or (C2). For example, the graph-induced rank on the \( k \)-NNG and the overall rank on the \( k \)-MDP would hardly ever run into either (C1) or (C2). We check it through Monte Carlo simulations by generating datasets from the standard multivariate multivariate Gaussian distribution with different \( N \)'s and \( d \)'s. For each dataset, we calculate the two ratios \( \frac{r_d^2}{r_0^2} \) and \( \frac{(N-2)(r_d^2-r_0^2)}{2(N-1)(r_1^2-r_0^2)} \). The procedure is repeated 1,000 times for each combination of \( N \) and \( d \). The details and results are in the Supplement. We find that neither (C1) nor (C2) happens in any of these simulation runs. In practice, when we apply the method, we could easily check whether the two cases happen. If it unfortunately happens, we could always use a different type of similarity graph to avoid the problem.

Define \( U_w = \frac{n-1}{N-2} U_x + \frac{m-1}{N-2} U_y \) and \( U_{\text{diff}} = U_x - U_y \), and their standardized statistics

\[
Z_w = \frac{U_w - \mathbb{E}(U_w)}{\sigma_w}, \quad \text{and} \quad Z_{\text{diff}} = \frac{U_{\text{diff}} - \mathbb{E}(U_{\text{diff}})}{\sigma_{\text{diff}}},
\]

where \( \sigma_w = \sqrt{\text{Var}(U_w)} \) and \( \sigma_{\text{diff}} = \sqrt{\text{Var}(U_{\text{diff}})} \).

**Theorem 2.3.** When \( T_R \) is well-defined, we have

\[
T_R = Z_w^2 + Z_{\text{diff}}^2 \quad \text{and} \quad \text{Cov}(Z_w, Z_{\text{diff}}) = 0. \quad (1)
\]

The proof of Theorem 2.3 is in the Supplement. From Theorem 2.1, it is easy to show that

\[
\sigma_w^2 = \frac{2m(m-1)n(n-1)}{(N-2)^2(N-3)} \left( (N-2)(r_d^2-r_0^2) - 2(N-1)(r_1^2-r_0^2) \right)
\]
\[ \sigma_{\text{diff}}^2 = 4(N - 1)mn(r_1^2 - r_0^2). \]

Hence, \( Z_{\text{diff}} \) or \( Z_w \) degenerates when (C1) or (C2) happens, respectively.

Remark 2.4. Some test statistics other than \( T_R \) can also be considered. For instance, the weighted rank sum statistic \( Z_w \) corresponding to the weighted edge-count test (Chen et al., 2018) that should work well for the location alternative and unbalanced sample sizes, and the max-rank test statistics \( R_{\text{max}} \equiv \max\{Z_w, |Z_{\text{diff}}|\} \) that corresponds to the max-type edge-count test statistic (Chu and Chen, 2019), which is preferred under the change-point setting.

3 Asymptotic properties

Obtaining the exact \( p \)-value of \( T_R \) by examining all permutations could be feasible for small sample sizes, but is time-prohibitive when the sample size is large. We thus work on the asymptotic distribution of \( T_R \).

3.1 Limiting distribution under the null hypothesis

Before stating the theorem, we define some notations. Let \( a_n = o(b_n) \) be that \( a_n \) is dominated by \( b_n \) asymptotically, i.e., \( \lim_{n \to \infty} \frac{a_n}{b_n} = 0 \), \( a_n = O(b_n) \) or \( a_n \asymp b_n \) be that \( a_n \) is bounded both above and below by \( b_n \) asymptotically, \( a_n \lesssim b_n \) be that \( a_n \) is bounded above by \( b_n \) (up to a constant factor) asymptotically, and ‘the usual limit regime’ be that \( m, n \to \infty \) and \( \frac{m}{n} \) does not go to 0 or \( \infty \) asymptotically. The following is a list of sufficient conditions for deriving the asymptotic distribution of \( T_R \).

Condition 3.1. \[ \sum_{i=1}^{N} \left( \sum_{j=1}^{N} R_{ij}^2 \right)^2 \lesssim N^3 r_d^4. \]
Condition 3.2. \[ \sum_{i=1}^{N} |\bar{R}_{i} - r_{0}|^{3} = o \left( (N(r_{1}^{2} - r_{0}^{2}))^{3} \right). \]

Condition 3.3. \[ \sum_{i=1}^{N} (\bar{R}_{i} - r_{0})^{3} = o(Nr_{d}(r_{1}^{2} - r_{0}^{2})). \]

Condition 3.4. \[ \sum_{i \neq j \neq k} N R_{ij} R_{ik} (\bar{R}_{i} - r_{0})(\bar{R}_{k} - r_{0}) = o(N^{3}r_{d}^{2}(r_{1}^{2} - r_{0}^{2})). \]

Condition 3.5. \[ \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k \neq i,j} \sum_{l \neq i,j} N R_{ij} R_{kl} (R_{ik} R_{jl} + R_{il} R_{jk}) = o(N^{4}r_{d}^{4}). \]

Condition 3.6. \[ r_{1} = o(r_{d}). \]

Theorem 3.1. Let \( R = (R_{ij})_{i,j \in [N]} \in \mathbb{R}^{N \times N} \) be the graph-induced rank or the overall rank matrix defined in Section 2. In the usual limit regime, under Conditions 3.1-3.6, we have that

\[ (Z_{w}, Z_{\text{diff}})^{T} \overset{D}{\longrightarrow} N_{2}(0_{2}, I_{2}) \quad \text{and} \quad T_{R} \overset{D}{\longrightarrow} \chi^{2}_{2} \]

under the permutation null distribution where \( \overset{D}{\longrightarrow} \) is convergence in distribution.

The proof of Theorem 3.1 is in the Supplement. Theorem 3.1 holds for general matrix \( R \) with some additional conditions.

Theorem 3.2. Let \( R = (R_{ij})_{i,j \in [N]} \in \mathbb{R}^{N \times N} \) be a symmetric matrix with non-negative entries and zero diagonal elements. Suppose further \( R_{ij} \geq 1 \) if \( R_{ij} > 0 \) and \( \max_{i,j} R_{ij} = o(N^{2}r_{d}^{2}) \). In the usual limit regime, under Conditions 3.1-3.6, we have that

\[ (Z_{w}^{P}, Z_{\text{diff}}^{P})^{T} \overset{D}{\longrightarrow} N_{2}(0_{2}, I_{2}) \quad \text{and} \quad T_{R} \overset{D}{\longrightarrow} \chi^{2}_{2} \]

under the permutation null distribution.

As a result, we can use different ways to weight the similarity graph. For Conditions 3.1-3.6, when \( R_{ij} \)'s only take values on 0 or 1 and \( R_{ij} = 1 \) if and only if the edge \((i,j)\) is in the similarity graph, then Conditions 3.1-3.5 degenerate to the conditions in Zhu and Chen (2021) and Condition 3.6 holds trivially by plugging in \( R_{ij} \)'s. The following remark discuss further on Conditions 3.1-3.6.
Remark 3.1. Although Conditions 3.1-3.6 seem complex, following the proof in Zhu and Chen (2021), we can have another version of conditions that are a bit stronger but easier to understand. Let \( G_i \) be the set of edges with one endpoint node \( i \), \( G_{i,2} \) be the set of edges with at least one endpoint in \( G_i \). In addition, we use \(| \cdot |\) to denote the cardinality of a set. So \(|G|\) is the number edges in \( G \). Let \( V_G = \sum_{i=1}^{N} |G_i|^2 - 4|G|^2/N \) which describes the variability of the sequence \(|G_i|\)'s, and \( V_R = \sum_{i=1}^{N} R_i^2 - (N(N-1)r_0)^2/N \). Then, Conditions 3.7-3.10 are sufficient conditions for obtaining the asymptotic distribution of \( T_R \).

**Condition 3.7.** \( \sum_{i=1}^{N} |G_i|^2 = o(|G|^2) \).

**Condition 3.8.** \( \sum_{i=1}^{N} |G_{i,2}|^2 = o(|G||V_G|) \).

**Condition 3.9.** \( \sum_{i=1}^{N} |G_i|^3 = o(\sqrt{|G| \land V_G}V_G) \).

**Condition 3.10.** \( \sum_{i=1}^{N} \sum_{j=1}^{N} R_{ij}^2 \simeq K^2|G|; \sum_{i=1}^{N} R_i^2 \simeq K^2 \sum_{i=1}^{N} |G_i|^2; \sum_{i=1}^{N} R_i^2 \simeq K^2V_G \), where \( K = \max_{i,j \in [N]} R_{ij} \).

This set of conditions tries to decouple the conditions on the graphs. Conditions 3.7-3.9 are borrowed from Zhu and Chen (2021), which are used to construct the limiting distributions of graph-based test statistics and shown to be satisfied easily by numerical experiments. Since we incorporate the weight (rank) information in the edges, we need an extra Condition 3.10, which is not hard to satisfy. For example, when \( R \) is the graph-induced rank on the \( k \)-NNG, we have \( K \simeq k \) and \( \sum_{i=1}^{N} \sum_{j=1}^{N} R_{ij}^2 \geq N \sum_{i=1}^{k} (l/2)^2 = Nk(k+1)/(2k+1)/24 \simeq K^2|G| \) and \( \sum_{i=1}^{N} \sum_{j=1}^{N} R_{ij}^2 \leq k^2|G| \), which implies \( \sum_{i=1}^{N} \sum_{j=1}^{N} R_{ij}^2 \simeq K^2|G| \).

We also know that \( R_i^2 \geq (\sum_{i=1}^{N} l/2)^2 = k^2(k+1)^2/16 \) and \( R_i^2 \leq (k|G_i|)^2 = k^2|G_i|^2 \). If \(|G_i| \simeq k \), we would have \( \sum_{i=1}^{N} R_i^2 \simeq K^2 \sum_{i=1}^{N} |G_i|^2 \).
3.2 Consistency

**Theorem 3.3.** For two continuous multivariate distributions $F_X$ and $F_Y$, if the graph-induced rank is used with the $k$-MST or $k$-NNG based on the Euclidean distance, where $k = O(1)$, then RISE is consistent against all alternatives in the usual limiting regime.

The proof is in the Supplement, which follows straightforwardly from Schilling (1986) and Henze and Penrose (1999).

4 Simulation studies

In this section, we conduct extensive simulations to examine the newly proposed test RISE. We mainly focus on the graph-induced rank on the $k$-NNG and the overall rank on the $k$-MDP as the representation of the two types of ranks.

Specifically, we consider a wide range of null and alternative distributions in moderate/high dimensions, including multivariate Gaussian distribution, Gaussian mixture distribution, multivariate log-normal distribution and multivariate $t_5$ distribution. These different distributions range from light-tails to heavy-tails, and the alternatives range from location difference, scale difference to mixed alternatives, with a hope that these simulation settings can cover real world scenarios.

Chen and Friedman (2017) suggested to use $k = 5$ for GET based on $k$-MST to achieve moderate power. For the $k$-NNG and $k$-MDP, the largest value of $k$ can be $N - 1$, while for the $k$-MST, the largest value of $k$ can only be $\frac{N}{2}$. So it is reasonable to choose $k$ for the $k$-NNG and $k$-MDP as twice of $k$ for the $k$-MST. Hence, we use $k = 10$ for simplicity in both simulation and real data analysis. We denote our methods as $R_{g}$-NN and $R_{o}$-MDP for RISE on the 10-NNG with the graph-based rank and on the 10-MDP graph with the
overall rank, respectively. More discussions on the choice of $k$ are in Section 6. Besides, a detailed comparison between RISE and GET including the results of RISE on the $k$-MST with the graph-induced rank and the overall rank is provided in Section 4.3.

We compare the type-I error and statistical power with seven state-of-art methods, including two graph-based methods: GET on 5-MST using the R package $g$Tests (Chen and Friedman, 2017), Rosenbaum’s cross matching test (CM) using the R package $crossmatch$ (Rosenbaum, 2005); two rank-based methods: multivariate rank-based test using measure transportation (MT) (Deb and Sen, 2021) and non-parametric two-sample test based on ball divergence (BD) using the R package $Ball$ (Pan et al., 2018); and three other tests: an LP-nonparametric test statistic (GLP) using the R package $LPKsample$ (Mukhopadhyay and Wang, 2020), a high-dimensional low sample size $k$-sample tests (HD) using the R package $HDLSSkST$ (Paul et al., 2021) and a kernel based two-sample test (MMD) using the R package $kerTests$ (Gretton et al., 2012). The tuning parameters of these comparable methods are set as their default values.

4.1 Settings

Throughout the simulation, we choose $m = n = 50$ and $m = 50, n = 100$, and set $d \in \{200, 500, 1000\}$. We consider diverse settings to examine the performance of these methods thoroughly. For each setting, we fix $F_X$, and choose different $F_Y$’s for the alternative hypothesis. We set the parameters of the distributions to make the tests have moderate power to be comparable. Each configuration is repeated 1000 times to estimate the power where the nominal significance level $\alpha$ is set as 0.05. We also check the empirical sizes of the tests under 0.01 and 0.05 nominal levels. The four settings are as follows:

I. $F_X = N_d(0_d, \Sigma_X)$ is the multivariate Gaussian distribution, where $\Sigma_{X,ij} = 6^{|i-j|}$. 

(a) Simple location: \( F_Y = N_d(\delta 1_d, \Sigma_X) \) where \( \delta = .5 \log d/\sqrt{d} \).

(b) Directed location: \( F_Y = N_d(\mu, \Sigma_X) \) where \( \mu = .5 \log d\mu'/||\mu'||_2 \) and \( \mu' \sim N_d(0_d, 1_d) \) is fixed.

(c) Simple scale: \( F_Y = N_d(0_d, \sigma^2 \Sigma_X) \) where \( \sigma = 1 + .12 \log d/\sqrt{d} \).

(d) Correlated scale: \( F_Y = N_d(0_d, \Sigma_Y) \) where \( \Sigma_{Y,ij} = .15|^{i-j}|. \)

(e) Location and scale mixed: \( F_Y = N_d(\mu, \Sigma_Y) \) where \( \mu = .2 \log d\mu'/||\mu'||_2 \) and \( \mu' \sim N_d(0_d, 1_d) \) is fixed.

II. \( F_X = WN_d(.31_d, 1_d) + (1 - W)N_d(-.31_d, 21_d) \) is the Gaussian mixture distribution, where \( W \sim \text{Bernoulli}(.5) \).

(a) Location: \( F_Y = WN_d((.3 + .75/\log d)1_d, 1_d) + (1 - W)N_d(-(.3 + .75/\log d)1_d, 21_d) \).

(b) Scale: \( F_Y = WN_d(.31_d, (1 + \sigma)^2 1_d) + (1 - W)N_d(-.31_d, (\sqrt{2} + \sigma)^2 1_d) \), where \( \sigma = .12\sqrt{50/d} \).

(c) Location and scale mixed: \( F_Y = WN_d(.351_d, \Sigma_Y) + (1 - W)N_d(-.351_d, 2\Sigma_Y) \), where \( \Sigma_{Y,ij} = .5|^{i-j}|. \)

III. \( F_X = \exp(N_d(0_d, \Sigma_X)) \) is the multivariate log-normal distribution, where \( \Sigma_{X,ij} = .6|^{i-j}|. \)

(a) Simple location: \( F_Y = \exp(N_d(\delta 1_d, \Sigma_X)) \) where \( \delta = .5 \log d/\sqrt{d} \).

(b) Sparse location: \( F_Y = \exp(N_d(\mu, \Sigma_X)) \) where \( \mu_j = (-1)^j 2.8 \log d/\sqrt{d}, j = 1, \ldots, [.05d], \mu_j = 0, j = [.05d] + 1, \ldots, d. \)

(c) Scale: \( F_Y = \exp(N_d(0_d, \sigma^2 \Sigma_X)), \) where \( \sigma = 1 + .15 \log d/\sqrt{d} \).

(d) Location and scale mixed: \( F_Y = \exp(N_d(\delta 1_d, \sigma \Sigma_X)) \) where \( \sigma = 1 + .1(50/d)^{25} \) and \( \delta = .25 \log d/\sqrt{d} \).
IV. $F_X = t_5(0_d, \Sigma_X)$ is the multivariate $t_5$ distribution, where $\Sigma_{X,ij} = .6^{|i-j|}$.

(a) Simple location: $F_Y = t_5(\delta 1_d, \Sigma_X)$ where $\delta = .5 \log d / \sqrt{d}$.

(b) Sparse location: $F_Y = t_5(\mu, \Sigma_X)$ where $\mu_j = (-1)^j 2.1 \log d / \sqrt{d}$, $j = 1, \ldots, \lfloor .05d \rfloor$, $\mu_j = 0$, $j = \lfloor .05d \rfloor + 1, \ldots, d$.

(c) Scale: $F_Y = t_5(0_d, \Sigma_Y)$, where $\Sigma_{Y,ij} = .7(1.1)^{|i-j|}$.

(d) Location and scale mixed: $F_Y = t_5(\delta 1_d, \Sigma_Y)$ where $\Sigma_{Y,ij} = (.8)^{|i-j|}$ and $\delta = .5 \log d / \sqrt{d}$.

4.2 Results

Here we present the results for $m = n = 50$ and $d \in \{200, 500, 1000\}$. The results for $m = 50, n = 100$ show similar patterns and are deferred to the Supplement.

The empirical sizes are presented in Table 1. RISE can control the type-I error well for different significant levels and settings, which validates the effectiveness of the asymptotic approximation even for relatively small sample sizes ($m = n = 50$). For other tests, MMD seems a little conservative and GLP has somewhat inflated type-I error for some settings, while all of the other tests can control the type-I error well.

The estimated power of these tests (in percent) is presented in Tables 2-4. The highest power for each setting and those with power higher than 95% of the highest one are highlighted in bold type.

Table 2 shows the results for the multivariate Gaussian distribution and the Gaussian mixture distribution settings. From Table 2, we see that for the multivariate Gaussian distribution, under the simple location alternative (a), MT performs the best, followed immediately by BD, Rg-NN and Ro-MDP. MMD is also good for $d = 200$ and 500. Under the directed location alternative (b), Rg-NN outperforms all of the other tests, followed
Table 1: Empirical sizes of the tests under the four settings when the nominal significance level $\alpha = .01$ and $0.05$, respectively, for $m = n = 50$ and $d = 200, 500, 1000$.

| $d$ | 200 | 500 | 1000 | 200 | 500 | 1000 | 200 | 500 | 1000 | 200 | 500 | 1000 |
|-----|-----|-----|------|-----|-----|------|-----|-----|------|-----|-----|------|
| $\alpha = 0.01$ | | | | | | | | | | | | |
| Setting I | Setting II | Setting III | Setting IV |
| R$_g$-NN | 0.01 | 0.00 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.00 | 0.01 | 0.00 | 0.01 |
| R$_o$-MDP | 0.01 | 0.01 | 0.01 | 0.01 | 0.02 | 0.01 | 0.01 | 0.02 | 0.01 | 0.00 | 0.01 | 0.01 |
| GET | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.00 | 0.01 | 0.00 | 0.00 | 0.01 | 0.00 | 0.01 |
| CM | 0.01 | 0.01 | 0.00 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.00 | 0.01 | 0.01 |
| MT | 0.01 | 0.01 | 0.02 | 0.01 | 0.01 | 0.00 | 0.01 | 0.01 | 0.00 | 0.01 | 0.01 | 0.01 |
| BD | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.00 | 0.01 | 0.01 | 0.01 |
| GLP | 0.01 | 0.01 | 0.01 | 0.02 | 0.03 | 0.03 | 0.06 | 0.07 | 0.06 | 0.01 | 0.01 | 0.01 |
| HD | 0.00 | 0.01 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 |
| MMD | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| $\alpha = 0.05$ | | | | | | | | | | | | |
| Setting I | Setting II | Setting III | Setting IV |
| R$_g$-NN | 0.05 | 0.05 | 0.04 | 0.05 | 0.05 | 0.04 | 0.04 | 0.04 | 0.03 | 0.06 | 0.04 | 0.05 |
| R$_o$-MDP | 0.06 | 0.05 | 0.04 | 0.04 | 0.06 | 0.04 | 0.05 | 0.06 | 0.04 | 0.05 | 0.04 | 0.05 |
| GET | 0.05 | 0.05 | 0.04 | 0.04 | 0.05 | 0.06 | 0.05 | 0.05 | 0.04 | 0.04 | 0.04 | 0.05 |
| CM | 0.04 | 0.04 | 0.03 | 0.04 | 0.03 | 0.04 | 0.03 | 0.03 | 0.04 | 0.04 | 0.03 | 0.03 |
| MT | 0.05 | 0.05 | 0.06 | 0.04 | 0.05 | 0.05 | 0.06 | 0.06 | 0.07 | 0.05 | 0.05 | 0.04 |
| BD | 0.04 | 0.05 | 0.06 | 0.04 | 0.06 | 0.04 | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 |
| GLP | 0.06 | 0.05 | 0.06 | 0.07 | 0.08 | 0.07 | 0.10 | 0.09 | 0.09 | 0.06 | 0.06 | 0.05 |
| HD | 0.03 | 0.04 | 0.03 | 0.03 | 0.04 | 0.03 | 0.02 | 0.03 | 0.02 | 0.02 | 0.02 | 0.02 |
| MMD | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.01 | 0.00 | 0.00 | 0.01 | 0.00 | 0.01 |
Table 2: Estimated power ($\alpha = 0.05$) under multivariate Gaussian I: (a) simple location, (b) directed location, (c) simple scale, (d) correlated scale, and (e) location and scale mixed and the Gaussian mixture II: (a) location, (b) scale, and (c) location and scale mixed.

|   | d  | m = n = 50 | Setting I (a) | Setting I (b) | Setting I (c) | Setting I (d) | Setting I (e) | Setting II (a) | Setting II (b) | Setting II (c) |
|---|---|------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|
|   | 200 | 500 | 1000 | 200 | 500 | 1000 | 200 | 500 | 1000 | 200 | 500 | 1000 |
| R_g-NN | 68 | 64 | 60 | 89 | **78** | **67** | 64 | 78 | 84 | **94** | **92** | **91** |
| R_o-MDP | 66 | 58 | 53 | 84 | 71 | 57 | 75 | 87 | 91 | **92** | **93** | **91** |
| GET | 62 | 56 | 50 | 81 | 68 | 56 | 59 | 71 | 80 | 81 | 78 | 75 |
| CM | 30 | 27 | 22 | 38 | 29 | 24 | 4 | 4 | 4 | 63 | 63 | 63 |
| MT | **98** | **96** | **93** | 7 | 6 | 7 | 5 | 5 | 4 | 13 | 14 | 14 |
| BD | 79 | 61 | 41 | 52 | 37 | 23 | **82** | **94** | **97** | 15 | 16 | 14 |
| GLP | 55 | 49 | 22 | 15 | 15 | 8 | 6 | 5 | 5 | 7 | 6 | 6 |
| HD | 4 | 4 | 3 | 3 | 3 | 4 | 55 | 71 | 84 | 8 | 9 | 7 |
| MMD | 90 | 54 | 6 | **98** | 54 | 3 | 0 | 0 | 0 | 0 | 0 | 0 |
|   | Setting I (e) | Setting II (a) | Setting II (b) | Setting II (c) |
| R_g-NN | **98** | **96** | **96** | **53** | **69** | **85** | **62** | **63** | **64** | **68** | **57** | **54** |
| R_o-MDP | **97** | **95** | **96** | 41 | 50 | 58 | 23 | 25 | 26 | 48 | 47 | 50 |
| GET | 91 | 87 | 86 | 44 | 59 | 75 | **63** | **65** | **66** | 51 | 40 | 38 |
| CM | 71 | 69 | 71 | 14 | 20 | 23 | 4 | 4 | 4 | 53 | **55** | **57** |
| MT | 16 | 14 | 11 | 49 | 54 | 56 | 4 | 5 | 5 | 7 | 11 | 12 |
| BD | 20 | 19 | 18 | 37 | 47 | 63 | 39 | 29 | 30 | 6 | 9 | 11 |
| GLP | 9 | 9 | 5 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 |
| HD | 8 | 8 | 7 | 2 | 4 | 2 | 3 | 4 | 3 | 2 | 4 | 2 |
| MMD | 1 | 0 | 0 | 1 | 2 | 1 | 0 | 1 | 0 | 1 | 1 | 0 |
immediately by $R_o$-MDP, then by GET. MMD is also good for $d = 200$, while all of other tests have low power. Under the simple sale alternative (c), BD performs the best and $R_o$-MDP performs the second best. $R_g$-NN, GET and HD also have satisfactory performance, while all of other tests have much lower power. Under the correlated scale alternative (d), $R_g$-NN and $R_o$-MDP exhibit the highest power and GET is also good enough. Under the location and scale mixed alternative (e), $R_g$-NN and $R_o$-MDP perform the best again, CM and GET have moderate power, and all other tests have low power. In these settings, $R_g$-NN, $R_o$-MDP and GET perform well in the multivariate Gaussian distribution setting, across a wide range of alternatives, while other tests can perform well in some alternatives, but have low power in other alternatives.

For the Gaussian mixture distribution setting II, we see that under the location alternative (a), $R_g$-NN performs the best. $R_o$-MDP, GET, MT and BD have moderate power while all of other tests have low power. Under the scale alternative (b), GET and $R_g$-NN outperform all other tests. Under the location and scale mixed alternative (c), $R_g$-NN and CM perform the best. So the overall performance of $R_g$-NN is the best in the Gaussian mixture setting.

Table 3 shows the result of the multivariate log-normal distribution. We see that under the simple location alternative (a), MT performs the best when $d$ is 200, and $R_o$-MDP performs the best when $d$ is 500 and 1000. $R_g$-NN, GET, GLP and BD also perform well. Under the sparse location alternative (b), $R_g$-NN outperforms all of the other tests, followed by $R_o$-MDP and GET, especially when $d$ is low ($d = 200$ or 500). MMD also performs well for $d = 200$ while other tests have low power. Under the scale alternative (c), BD performs the best and $R_o$-MDP performs the second best, followed immediately by $R_g$-NN and GET. Under the mixed alternative (d), $R_o$-MDP and BD perform the best, followed
Table 3: Estimated power ($\alpha = 0.05$) under the multivariate log-normal distribution III: (a) simple location, (b) sparse location, (c) scale, and (d) location and scale mixed.

|   | Setting III (a) | Setting III (b) | Setting III (c) | Setting III (d) |
|---|-----------------|-----------------|-----------------|-----------------|
| $d$ | 200 | 500 | 1000 | 200 | 500 | 1000 | 200 | 500 | 1000 |
| $m = n = 50$ | R$_g$-NN | 75 | 71 | 68 | 94 | 86 | 71 | 26 | 30 | 32 | 53 | 59 | 58 |
|          | R$_o$-MDP   | 94 | 95 | 95 | 85 | 80 | 68 | 46 | 58 | 63 | 80 | 88 | 93 |
|          | GET         | 68 | 61 | 56 | 85 | 69 | 49 | 24 | 26 | 27 | 49 | 51 | 50 |
|          | CM          | 18 | 17 | 15 | 32 | 30 | 25 | 6 | 6 | 6 | 9 | 10 | 12 |
|          | MT          | 97 | 94 | 88 | 11 | 25 | 43 | 17 | 19 | 13 | 68 | 65 | 60 |
|          | BD          | 91 | 93 | 94 | 17 | 14 | 10 | 56 | 68 | 72 | 82 | 91 | 94 |
|          | GLP         | 70 | 65 | 30 | 23 | 36 | 15 | 12 | 9 | 10 | 22 | 18 | 11 |
|          | HD          | 29 | 36 | 43 | 4 | 4 | 4 | 16 | 19 | 23 | 24 | 34 | 44 |
|          | MMD         | 83 | 57 | 20 | 98 | 79 | 8 | 19 | 7 | 0 | 54 | 32 | 10 |

immediately by MT, R$_g$-NN, and GET. So the overall performance of R$_o$-MDP is the best under the multivariate log-normal setting.

Finally, Table 4 shows the result of the multivariate $t_5$ distribution. MT performs the best under the simple location alternative (a), while R$_g$-NN and R$_o$-MDP are also good and outperform other tests. Under the sparse location alternative (b), R$_g$-NN performs the best. R$_o$-MDP performs the best in the scale alternative (c) and both R$_g$-NN and R$_o$-MDP perform the best in the mixed alternative (d). In these settings, R$_g$-NN and R$_o$-MDP are doing well consistently.

To summarize, we observe that RISE performs well in a wide range of alternatives under different distributions. Besides, MT performs well in the simple location alternative, e.g., Setting I (a), III (a), IV (a), but lacks power in directed or sparse location alternative and
Table 4: Estimated power ($\alpha = 0.05$) under the multivariate $t_5$ distribution IV: (a) simple location, (b) sparse location, (c) scale and (d) location and scale mixed.

| Setting | $d$ | 200 | 500 | 1000 | 200 | 500 | 1000 | 200 | 500 | 1000 | 200 | 500 | 1000 |
|---------|-----|-----|-----|------|-----|-----|------|-----|-----|------|-----|-----|------|
| (a) simple location | $m = n = 50$ | Setting IV (a) | Setting IV (b) | Setting IV (c) | Setting IV (d) |
| R$_g$-NN | 82 | 66 | 57 | 81 | 62 | 49 | 81 | 65 | 58 | 88 | 73 | 63 |
| R$_o$-MDP | 70 | 63 | 53 | 68 | 55 | 44 | 95 | 93 | 93 | 82 | 78 | 74 |
| GET | 66 | 44 | 33 | 58 | 36 | 24 | 70 | 46 | 39 | 76 | 56 | 43 |
| CM | 24 | 21 | 18 | 24 | 20 | 17 | 72 | 68 | 67 | 45 | 41 | 42 |
| MT | 95 | 92 | 88 | 10 | 9 | 6 | 17 | 19 | 19 | 75 | 72 | 67 |
| BD | 6 | 6 | 5 | 5 | 5 | 5 | 66 | 66 | 69 | 7 | 6 | 5 |
| GLP | 52 | 40 | 18 | 8 | 10 | 6 | 39 | 39 | 39 | 51 | 39 | 30 |
| HD | 2 | 2 | 2 | 3 | 2 | 2 | 13 | 11 | 11 | 2 | 3 | 1 |
| MMD | 62 | 17 | 4 | 42 | 8 | 3 | 30 | 29 | 35 | 60 | 20 | 5 |

scale alternatives, while BD performs well in the simple scale alternative but lacks power in the location alternatives. GET is doing a good job overall, but it is outperformed by RISE in most of the settings. Next, we compare RISE and GET in more details.

### 4.3 A detailed comparison between RISE and GET

Here, we compare the power of RISE and GET by varying $k$’s. We also explore the graph-induced rank (denoted by R$_g$-MST) and the overall rank (denoted by R$_o$-MST) on the $k$-MST. To compare different graphs in a more unified fashion, for the $k$-NNG and $k$-MDP, we set $k = 2\lceil N^\lambda \rceil$ while for the $k$-MST, we set $k = \lceil N^\lambda \rceil$, for $\lambda \in (0, 0.8)$, since for the $k$-NNG and $k$-MDP, the largest value of $k$ can be $N - 1$, while for the $k$-MST, the largest value of $k$ can only be $N^\lambda$. The results for different $n$’s and $d$’s show similar patterns, so we
only present the results for $m = n = 50$ and $d = 500$ here for Settings I-IV in Section 4.1 with $\alpha = 0.05$. Each configuration is repeated 1000 times to estimate the empirical size or power.

The empirical sizes of the five tests under Settings I-IV are presented in Figure 1. We see that all of these tests can control the type-I error well even for large $\lambda$ under all settings. The estimated power for Settings I and II are presented in Figure 2 and the estimated power for Settings III and IV are presented in Figure 3. From these figures, we see that RISE performs better than GET in most of the settings for a wide range of $k$’s.

Figure 1: Empirical sizes of RISE and GET for varying $\lambda$.

We notice that $R_g$-NN has the best performance in most of the settings for all $k$’s. The improvement of $R_g$-NN and $R_o$-MDP over GET is more significant under the heavy-tailed Setting III and IV. However, $R_o$-MDP is less powerful under the Gaussian mixed Setting II, which may be due to the intrinsic property of MDP. $R_o$-MST has a moderate performance
Figure 2: Estimated power of RISE and GET for varying $\lambda$ under Settings I and II.

Figure 3: Estimated power of RISE and GET for varying $\lambda$ under Settings III and IV.
such that it outperforms GET in the most of the settings but is dominated by R$_g$-NN in most instances. R$_g$-MST seems not very robust as it can achieve high power in some cases but is outperformed by GET sometimes.

5 Real data analysis

5.1 New York City taxi data

To illustrate the proposed tests, we here conduct an analysis on whether the travel patterns are different in consecutive months in the New York City. We use New York City taxi data from the NYC Taxi Limousine Commission (TLC) website\footnote{https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page}. The data contains rich information such as the taxi pickup and drop-off date/times, longitude and latitude coordinates of pickup and drop-off locations. Specifically, we are interested in the travel pattern from the John F. Kennedy International Airport of the year 2015. Similarly to Chu and Chen (2019), we set the boundary of JFK airport from 40.63 to 40.66 latitude and from $-73.80$ to $-73.77$ longitude. Additionally, we set the boundary of New York City from 40.577 to 41.5 latitude and from $-74.2$ to $-73.6$ longitude. We only consider those trips that began with a pickup at JFK and ended with a drop-off in New York City. The New York City is then split into a $30 \times 30$ grid with equal size and the numbers of taxi drop-offs that fall within each cell are counted for each day. Thus each day is represented by a $30 \times 30$ matrix and we use the negative Frobenius norm as the similarity measure.

We then conduct eleven comparisons over the consecutive months: January vs February, . . . , November vs December. With the aim for illustration, we treat them as eleven separate tests rather than a multiple testing problem. For simplicity, we only compare
our method with GET and two rank-based methods MT and BD. The five tests provide different conclusions for seven comparisons at 0.05 significance level, which is presented in Table 5. The p-values of the four comparisons with the same conclusion are presented in Table 6.

Table 5: The p-values of the tests showing inconsistent conclusion for the NYC taxi data.

| Method  | Jan/Feb | Feb/Mar | May/Jun | Jun/Jul | Jul/Aug | Aug/Sep | Sep/Oct |
|---------|---------|---------|---------|---------|---------|---------|---------|
| Rg-NN   | 0.007   | 0.005   | 0.021   | 0.000   | 0.008   | 0.000   | 0.011   |
| Ro-MDP  | 0.002   | 0.000   | 0.808   | 0.000   | 0.023   | 0.000   | 0.001   |
| GET     | 0.090   | 0.013   | 0.018   | 0.000   | 0.020   | 0.000   | 0.003   |
| MT      | 0.528   | 0.053   | 0.790   | 0.083   | 0.001   | 0.934   | 0.681   |
| BD      | 0.340   | **0.050** | 0.280   | **0.040** | 0.310   | 0.070   | **0.030** |

Table 6: The p-values of the tests showing consistent conclusion for the NYC taxi data.

| Method  | Mar/Apr | Apr/May | Oct/Nov | Nov/Dec |
|---------|---------|---------|---------|---------|
| Rg-NN   | 0.000   | 0.072   | 0.004   | 0.069   |
| Ro-MDP  | 0.008   | 0.076   | 0.007   | 0.316   |
| GET     | 0.000   | 0.367   | 0.008   | 0.211   |
| MT      | 0.030   | 0.093   | 0.001   | 0.371   |
| BD      | 0.020   | 0.190   | 0.010   | 0.270   |

We notice that for these inconsistent conclusions, our methods always have p-values smaller than 0.05 except for May vs June with Ro-MDP. GET also has p-values smaller than 0.05 except for January vs February. BD only rejects three of the comparisons while MT only rejects June vs July. It indicates that RISE and GET may be more powerful in this dataset.
Since RISE yields a different conclusion from all of the other tests in the comparison of January vs February, we take a closer look at it. We first examine each $k$th MST and $k$-MST separately for $k = 1, \ldots, 5$. The test statistic of GET depends on how far the two within-sample edge-counts deviate from their expectations under the null distribution, so we check how the two edge-counts statistics change when $k$ increases from 1 to 5.

Table 7: The edge-count statistics on the $k$th MST and the $p$-values of GET using the $k$th MST and the $k$-MST, respectively. The expected edges for each MST are 15.76 and 12.81 for Samples Jan and Feb, respectively.

| Edge-count | $k$ | 1    | 2    | 3    | 4    | 5    |
|------------|-----|------|------|------|------|------|
|            | Jan | 15   | 15   | 14   | 14   | 13   |
|            | Feb | 20   | 18   | 19   | 16   | 8    |
| $p$-values | $k$th MST | 0.034 | 0.112 | 0.105 | 0.540 | 0.109 |
|           | $k$-MST  | 0.034 | 0.007 | 0.002 | 0.003 | 0.090 |

Table 7 shows the within-sample edge-counts of each sample in each $k$th MST. The $p$-values of GET on the $k$th MST and the $k$-MST for different $k$’s are also presented. We notice that for most of the $k$th MSTs, at least one of the within-sample edge-counts somewhat deviates from their corresponding expectations. However, since GET treats all MSTs equally, there are two issues: (i) different MSTs can contain opposite information and (ii) a $k$th MST for a large $k$ can contain noisier information. The first issue is obvious from the edge-counts statistics. For example, the sample February has the within-sample edge-count above its expectation for the first to the forth MSTs, but below its expectation for the fifth MST. This makes the $p$-value increases from 0.003 on the 4-MST to 0.09 on the 5-MST. The second issue can be observed from the $p$-values of GET on the $k$th MST. The $p$-value of the comparison on the first MST is small, but it can be very large for other $k$th
MSTs. When the $k$th MST does not contain useful information but noise, the consequence for GET is to yield a larger $p$-value. On the other hand, RISE is less affected by the two issues by incorporating weights.

5.2 Brain network data

We here evaluate the performance of RISE in distinguishing differences in brain connectivity between male and female subjects using brain networks constructed from diffusion magnetic resonance imaging (dMRI). The data from the HNU1 study (Zuo et al., 2014) consists of dMRI records of fifteen male and fifteen female healthy subjects that were scanned ten times each over a period of one month. The processed data can be downloaded from [http://mrneurodata.s3-website-us-east-1.amazonaws.com/HNU1/ndmg_0-0-48/graphs/CPAC200/](http://mrneurodata.s3-website-us-east-1.amazonaws.com/HNU1/ndmg_0-0-48/graphs/CPAC200/), where each subject and scan is represented by a weighted network with 200 nodes registered to the CC200 atlas using the NeuroData’s MRI to Graphs pipeline (Kiar et al., 2018). Figure 4 plots four networks with two networks from male subjects and two from female subjects. The networks are then coded by $200 \times 200$ weighted adjacency matrices. For each subject, there are ten scans and we use the average of these ten matrices for the subject’s brain network representation. Then, we obtain fifteen networks for the male and female groups, respectively. Here, we also use the negative Frobenius norm as the similarity measure.

The results are presented in Table 8. Since the sample size is small ($N = 30$), to check the validity of the asymptotic $p$-value approximation, we also show the $p$-values of GET and RISE from 1000 permutations, which are showed in the brackets. We notice that for RISE, the approximate $p$-values are very close to the $p$-values from permutations even in such a small sample size. All of these tests have small $p$-values. BD shows some
Figure 4: The brain networks of two male subjects (blue) and two female subjects.

Table 8: The $p$-values of the tests for the brain network data.

| Method   | $R_x$-NN | $R_\omega$-MDP | GET       | MT     | BD     |
|----------|----------|-----------------|-----------|--------|--------|
| $p$-values | 0.003 (0.007) | 0.019 (0.019)   | 0.005 (0.011) | 0.095  | 0.057  |

evidence of difference with a $p$-value slightly larger than 0.05 while MT shows less evidence of difference, but RISE can provide a more confident conclusion with smaller $p$-values.

Besides, a heat map of the distance matrix of the 30 subjects is presented in Figure 5 where the first 15 subjects are male and the others female. We see an obvious difference between male and female subjects from the heat map, where the male subjects have larger within-sample distances but the female subjects have smaller within-sample distances. This is an evidence for scale difference.

We further plot the entrywise mean and standard deviation of the weighted networks for each sample as shown in Figure 6, which shows that the entrywise means of the two sample are close while significant differences exist for variances of some covariates. For example, several covariates have standard deviation near 3000 for male subjects but only near 2000 for female subjects. These results support the conclusion from RISE that the male and female brain networks are different.
Figure 5: The heatmap of the distance matrix of the 30 subjects, where the first 15 subjects are male and the others female.

6 Discussion and conclusion

We propose a new framework of asymptotically distribution-free rank-based test, which shows superior performance under a wide range of alternatives. Specifically, we suggest to use $R_g$-NN because of its robust performance and lower computational complexity. In most settings of the paper, we fix $k = 10$ for $R_g$-NN, which is already good enough in terms of power. For tests based on similarity graphs, the choice of graph is still an open question. Some previous works (Friedman and Rafsky, 1979; Zhang and Chen, 2022; Chen and Friedman, 2017; Chen et al., 2018) suggested to use the $k$-MST and set $k$ as a small constant number, e.g., $k = 3$ or $k = 5$. Recently, Zhu and Chen (2021) observed that a denser graph can improve the power of the tests such that $k = O(N^\lambda)$ for some $0 < \lambda < 1$ where $N$ is the total number of observations. Following this, Zhang and Chen (2021) compared the power for different $\lambda$’s under various simulation settings and suggested to
use $\lambda = 0.5$ for GET, where it showed adequate power across different simulation settings. Here we adopt a similar procedure to explore $k$ for RISE with details in the Supplement. Based on these numerical results as well as the results of Section 4.3, we found that if the sample size is large enough, it can be sufficient to use $k = 10$, otherwise, using $k = \lceil N^{0.65} \rceil$ for $k$-NNG or $k$-MDP could be a good choice when computation is not an issue.

References

Arroyo, J., A. Athreya, J. Cape, G. Chen, C. E. Priebe, and J. T. Vogelstein (2021). Inference for multiple heterogeneous networks with a common invariant subspace. *Journal of Machine Learning Research* 22(142), 1–49.

Barale, M. and D. Shirke (2021). A test based on data depth for testing location-scale of the two multivariate populations. *Journal of Statistical Computation and Simulation* 91(4), 768–785.
Bickel, P. J. (1965). On some asymptotically nonparametric competitors of Hotelling’s $t^2$

1. *The Annals of Mathematical Statistics*, 160–173.

Biswas, M. and A. K. Ghosh (2014). A nonparametric two-sample test applicable to high-dimensional data. *Journal of Multivariate Analysis* 123, 160–171.

Biswas, M., M. Mukhopadhyay, and A. K. Ghosh (2014). A distribution-free two-sample run test applicable to high-dimensional data. *Biometrika* 101(4), 913–926.

Bullmore, E. and O. Sporns (2012). The economy of brain network organization. *Nature Reviews Neuroscience* 13(5), 336–349.

Chaudhuri, P. (1996). On a geometric notion of quantiles for multivariate data. *Journal of the American Statistical Association* 91(434), 862–872.

Chen, H., X. Chen, and Y. Su (2018). A weighted edge-count two-sample test for multivariate and object data. *Journal of the American Statistical Association* 113(523), 1146–1155.

Chen, H. and J. H. Friedman (2017). A new graph-based two-sample test for multivariate and object data. *Journal of the American statistical association* 112(517), 397–409.

Chen, H. and N. R. Zhang (2013). Graph-based tests for two-sample comparisons of categorical data. *Statistica Sinica*, 1479–1503.

Chu, L. and H. Chen (2018). Sequential change-point detection for high-dimensional and non-Euclidean data. *arXiv preprint arXiv:1810.05973*.

Chu, L. and H. Chen (2019). Asymptotic distribution-free change-point detection for multivariate and non-Euclidean data. *The Annals of Statistics* 47(1), 382–414.
Deb, N. and B. Sen (2021). Multivariate rank-based distribution-free nonparametric testing using measure transportation. *Journal of the American Statistical Association*, 1–45.

Friedman, J. H. and L. C. Rafsky (1979). Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests. *The Annals of Statistics*, 697–717.

Gretton, A., K. Borgwardt, M. J. Rasch, B. Scholkopf, and A. J. Smola (2008). A kernel method for the two-sample problem. *arXiv preprint arXiv:0805.2368*.

Gretton, A., K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola (2012). A kernel two-sample test. *The Journal of Machine Learning Research 13*(1), 723–773.

Gretton, A., K. Fukumizu, Z. Harchaoui, and B. K. Sriperumbudur (2009). A fast, consistent kernel two-sample test. In *NIPS*, Volume 23, pp. 673–681.

Gretton, A., D. Sejdinovic, H. Strathmann, S. Balakrishnan, M. Pontil, K. Fukumizu, and B. K. Sriperumbudur (2012). Optimal kernel choice for large-scale two-sample tests. In *Advances in neural information processing systems*, pp. 1205–1213. Citeseer.

Hallin, M., E. Del Barrio, J. Cuesta-Albertos, and C. Matrán (2021). Distribution and quantile functions, ranks and signs in dimension d: A measure transportation approach. *The Annals of Statistics 49*(2), 1139–1165.

Hallin, M. and D. Paindaveine (2004). Rank-based optimal tests of the adequacy of an elliptic varma model. *The Annals of Statistics 32*(6), 2642–2678.

Hallin, M. and D. Paindaveine (2006). Parametric and semiparametric inference for shape: the role of the scale functional. *Statistics & Decisions 24*(3), 327–350.

Harchaoui, Z., F. R. Bach, and E. Moulines (2007). Testing for homogeneity with kernel fisher discriminant analysis. In *NIPS*, pp. 609–616. Citeseer.
Hediger, S., L. Michel, and J. Näf (2019). On the use of random forest for two-sample testing. *arXiv preprint arXiv:1903.06287*.

Henze, N. (1988). A multivariate two-sample test based on the number of nearest neighbor type coincidences. *The Annals of Statistics*, 772–783.

Henze, N. and M. D. Penrose (1999). On the multivariate runs test. *Annals of statistics*, 290–298.

Jiang, R., N. Zuo, J. M. Ford, S. Qi, D. Zhi, C. Zhuo, Y. Xu, Z. Fu, J. Bustillo, J. A. Turner, et al. (2020). Task-induced brain connectivity promotes the detection of individual differences in brain-behavior relationships. *NeuroImage 207*, 116370.

Kiar, G., E. W. Bridgeford, W. R. G. Roncal, V. Chandrashekhar, D. Mhembere, S. Ryan, X.-N. Zuo, D. S. Margulies, R. C. Craddock, C. E. Priebe, et al. (2018). A high-throughput pipeline identifies robust connectomes but troublesome variability. *bioRxiv*, 188706.

Kim, I., A. Ramdas, A. Singh, and L. Wasserman (2021). Classification accuracy as a proxy for two-sample testing. *The Annals of Statistics 49*(1), 411–434.

Li, J. (2018). Asymptotic normality of interpoint distances for high-dimensional data with applications to the two-sample problem. *Biometrika 105*(3), 529–546.

Liu, R. Y. and K. Singh (1993). A quality index based on data depth and multivariate rank tests. *Journal of the American Statistical Association 88*(421), 252–260.

Liu, Y.-W. and H. Chen (2022). A fast and efficient change-point detection framework based on approximate k-nearest neighbor graphs. *IEEE Transactions on Signal Processing*. 

33
Lopez-Paz, D. and M. Oquab (2016). Revisiting classifier two-sample tests. *arXiv preprint arXiv:1610.06545*.

Marden, J. I. (1999). Multivariate rank tests. *STATISTICS TEXTBOOKS AND MONOGRAPHS 159*, 401–432.

Menafoglio, A. and P. Secchi (2017). Statistical analysis of complex and spatially dependent data: a review of object oriented spatial statistics. *European journal of operational research 258*(2), 401–410.

Mukhopadhyay, S. and K. Wang (2020). A nonparametric approach to high-dimensional k-sample comparison problems. *Biometrika 107*(3), 555–572.

Oja, H. (2010). *Multivariate nonparametric methods with R: an approach based on spatial signs and ranks*. Springer Science & Business Media.

Pan, W., Y. Tian, X. Wang, and H. Zhang (2018). Ball divergence: nonparametric two sample test. *Annals of statistics 46*(3), 1109.

Paul, B., S. K. De, and A. K. Ghosh (2021). *HDLSSkST: Distribution-Free Exact High Dimensional Low Sample Size k-Sample Tests*. R package version 2.0.0.

Puri, M. L. and P. K. Sen (2013). On a class of multivariate multisample rank-order tests. In *Nonparametric Methods in Statistics and Related Topics*, pp. 659–682. De Gruyter.

Rosenbaum, P. R. (2005). An exact distribution-free test comparing two multivariate distributions based on adjacency. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67*(4), 515–530.

Sarkar, S., R. Biswas, and A. K. Ghosh (2020). On some graph-based two-sample tests for high dimension, low sample size data. *Machine Learning 109*(2), 279–306.
Sarkar, S. and A. K. Ghosh (2018). On some high-dimensional two-sample tests based on averages of inter-point distances. *Stat* 7(1), e187.

Schilling, M. F. (1986). Multivariate two-sample tests based on nearest neighbors. *Journal of the American Statistical Association* 81(395), 799–806.

Song, H. and H. Chen (2020). Generalized kernel two-sample tests. *arXiv preprint arXiv:2011.06127*.

Székely, G. J. and M. L. Rizzo (2013). Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference* 143(8), 1249–1272.

Tian, Z., L. Jia, H. Dong, F. Su, and Z. Zhang (2016). Analysis of urban road traffic network based on complex network. *Procedia engineering* 137, 537–546.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin* 1(6), 80–83.

Zhang, J. and H. Chen (2022). Graph-based two-sample tests for data with repeated observations. *Statistica Sinica*, DOI:10.5705/ss.202019.0116.

Zhang, Y. and H. Chen (2021). Graph-based multiple change-point detection. *arXiv preprint arXiv:2110.01170*.

Zhu, Y. and H. Chen (2021). Limiting distributions of graph-based test statistics. *arXiv preprint arXiv:2108.07446*.

Zuo, X.-N., J. S. Anderson, P. Bellec, R. M. Birn, B. B. Biswal, J. Blautzik, J. C. Breitner, R. L. Buckner, V. D. Calhoun, F. X. Castellanos, et al. (2014). An open science resource for establishing reliability and reproducibility in functional connectomics. *Scientific data* 1(1), 1–13.