

Statistica Sinica Preprint No: SS-2025-0280	
Title	Delaunay Weighted Two-sample Test for High-dimensional Data by Incorporating Geometric Information
Manuscript ID	SS-2025-0280
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202025.0280
Complete List of Authors	Jiaqi Gu, Ruoxu Tan and Guosheng Yin
Corresponding Authors	Ruoxu Tan
E-mails	ruoxut@tongji.edu.cn
Notice: Accepted author version.	

Delaunay Weighted Two-sample Test for High-dimensional Data by Incorporating Geometric Information

Jiaqi Gu*

Department of Mathematics and Statistics, University of South Florida
and

Ruoxu Tan[†]

School of Mathematical Sciences and
School of Economics and Management, Tongji University
and

Guosheng Yin[‡]

Department of Statistics and Actuarial Science, University of Hong Kong

Abstract

Two-sample hypothesis testing is a fundamental problem with various applications, which faces new challenges in the high-dimensional context. To mitigate the issue of the curse of dimensionality, high-dimensional data are typically assumed to lie on a low-dimensional manifold. To incorporate geometric information in the data, we propose to apply the Delaunay triangulation and develop the Delaunay weight to measure the geometric proximity among data points. In contrast to existing similarity measures that only utilize pairwise distances, the Delaunay weight can take both the distance and direction information into account. A detailed computation procedure is developed to learn the unknown manifold and approximate the Delaunay weight. We further propose a novel nonparametric test statistic using the Delaunay weight matrix. Asymptotic normality under the null and consistency under the alternative of the test statistic are developed. Applied to simulated data, the new test shows robustness to the learning of the unknown manifold and exhibits substantial power gain if the distributions differ in the principal directions of covariance matrices. The proposed test also detects significant differences on a real dataset of mice protein expression levels.

Key words: Delaunay triangulation; geometric proximity; high dimension; manifold learning; permutation test.

*jiaqigu@usf.edu; ORCID: 0000-0002-8773-9558

[†]Corresponding author, ruoxut@tongji.edu.cn; ORCID: 0000-0002-2046-4377

[‡]gyin@hku.hk; ORCID: 0000-0003-3276-1392

1 Introduction

The two-sample test on whether the two underlying distributions are the same or not is a fundamental problem in statistics and machine learning with broad applications in change point detection (Cao et al., 2018; Chen, 2019), goodness-of-fit evaluation (Chwialkowski et al., 2016; Arias-Castro et al., 2018) and experimental design (Morgan and Rubin, 2012). In the era of big data, there are emerging challenges that the power of classical two-sample tests (e.g., Hotelling’s T -squared test, Wald test or likelihood ratio test) decreases or even vanishes as the data dimension increases. Under the high-dimensional setting, several methods have been developed to test a simpler hypothesis on whether two population means are the same. For example, under the sparse alternative that mean vectors of two samples only differ in a small number of dimensions, Cai et al. (2014) developed a maximum-type test statistic. To enhance the power and robustness to extreme values in detecting sparse mean differences, a parametric bootstrap approach with feature screening was proposed by Chang et al. (2017), while Wang et al. (2019) incorporated the trimmed mean approach to developing a robust test. Liu et al. (2024) developed a projection test by assuming the sparsity of the optimal projection direction. Yang et al. (2024) introduced a novel test whose estimation error does not affect asymptotic power while achieving superior performance. However, the aforementioned methods are not consistent against general alternatives, e.g., the two populations possess the same mean but differ in the covariance structures.

Nonparametric two-sample testing for general differences between distributions is a historical research topic with rich literature. For univariate data, the Kolmogorov–Smirnov test (Kolmogorov, 1933; Smirnov, 1948), the Wilcoxon–Mann–Whitney test (Wilcoxon, 1945; Mann and Whitney, 1947) and the Wald–Wolfowitz runs test (Wald and Wolfowitz, 1940) are early developments on this topic. Although such methods may be directly ap-

plicable to multivariate settings (Bickel, 1969), they would perform poorly unless the total sample size grows at an exponential rate with the dimension. To circumvent such difficulty, Friedman and Rafsky (1979) developed a two-sample test based on the minimum spanning tree (MST) that connects all data points with the minimal total length of edges. Given an MST of the pooled sample, the test statistic is constructed as the number of edges connecting data points from different samples, which is shown to be consistent against general alternatives. More recently, many nonparametric two-sample tests have been proposed for multivariate/high-dimensional data, including: (i) distance/energy-based tests based on pairwise distances of the pooled sample (Székely and Rizzo, 2004; Jurečková and Kalina, 2012; Marozzi, 2015); (ii) graph-based tests, e.g., the k -nearest neighbor (k -NN) graph (Schilling, 1986; Henze, 1988; Hall and Tajvidi, 2002) and the k -MST graph (Chen and Friedman, 2017; Chen et al., 2018); (iii) kernel tests via a decaying kernel of pairwise Euclidean distances of the pooled sample (Gretton et al., 2009, 2012; Cheng and Xie, 2024; Yan and Zhang, 2023); and (iv) regression tests by connecting the binary classification and two-sample testing (Kim et al., 2019; Hediger et al., 2022).

To mitigate the curse of dimensionality, high-dimensional data are often assumed to lie on an unknown low-dimensional manifold, which is reasonable in many applications, such as bioinformatics (Moon et al., 2018), image data analysis (Pless and Souvenir, 2009) and natural language processing (Zhao et al., 2021). Performances of the nonparametric tests under the low-dimensional manifold setting have been studied in the literature. For example, it is shown that the kernel test (Cheng and Xie, 2024) and the local regression test (Kim et al., 2019) are adaptive to the intrinsic dimension of the manifold by properly choosing the tuning parameter. Arias-Castro et al. (2018) also showed that the multivariate chi-squared test is adaptive to the intrinsic dimension of the manifold, while their work was

mainly focused on theoretical development rather than practical implementation. However, none of these tests directly take the manifold structure into account, while our work aims to fill this gap.

In particular, we develop a new nonparametric two-sample test, named the Delaunay weighted test, for high-dimensional data under the low-dimensional manifold assumption. Following the k -NN test by [Schilling \(1986\)](#), we propose the test statistic as the average of all within-group proximities of the pooled sample. The proximity is measured by our newly developed Delaunay weight instead of the k -NN. By incorporating the Delaunay triangulation on a Riemannian manifold ([Leibon and Letscher, 2000](#)), the Delaunay weight measures the geometric proximity among data points by taking both the geodesic distance and the relative direction into account. In practice, the Delaunay weight cannot be computed exactly because the manifold is usually unknown. As a remedy, we compute the Delaunay weight matrix of low-dimensional manifold representations for approximate inference. The manifold representations are obtained following the idea of Isomap ([Tenenbaum et al., 2000](#)). We develop the stereographic projected DELAUNAYSPARSE algorithm to find the Delaunay simplices for computing the Delaunay weight. The p -value of our test is obtained by a permutation procedure, and our Delaunay weighted test is applicable as long as the sample size is larger than the (estimated) intrinsic dimension of the underlying manifold. Compared with the kernel test and the local regression test based on local distances that are adaptive to the intrinsic dimension of the manifold, ours further takes the relative direction among data points into account, and thus it is more efficient in detecting the direction difference of two distributions. Numerical experiments on synthetic data reveal that our Delaunay weighted test outperforms existing approaches when the two distributions differ in the principal directions of covariance matrices, although the power of our method is not

the highest under location differences between distributions.

The rest of this article is organized as follows. Section 2 motivates and formulates the two-sample testing problem for high-dimensional data lying on an unknown Riemannian manifold and introduces preliminary geometric concepts. The Delaunay weighted test is proposed in Section 3, including the definition and advantages of Delaunay weight, the test statistic, computational strategies to approximate the Delaunay weight, and the permutation procedure for inference. Section 4 investigates the theoretical properties of the proposed test. In Section 5, we conduct various numerical experiments including a real data example to demonstrate the practical performance of our method, such as the statistical power, robustness to the manifold learning procedures, and computational cost. Section 6 concludes with discussions. The supplementary material contains details of our computational strategies, proofs of theorems and additional experimental results.

2 Manifold Hypothesis and Delaunay Triangulation

2.1 Manifold Hypothesis

Consider two samples $\mathbb{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_{n_1}\} \subset \mathcal{R}^D$ and $\mathbb{Y} = \{\mathbf{y}_1, \dots, \mathbf{y}_{n_0}\} \subset \mathcal{R}^D$ of independent and identically distributed (i.i.d.) data from two distributions F and G , respectively. Our goal is to test $H_0 : F = G$ versus $H_1 : F \neq G$. Let $\mathbb{Z}$ be the union of two samples, $\mathbb{Z} = \mathbb{X} \cup \mathbb{Y} = \{\mathbf{z}_1, \dots, \mathbf{z}_n\}$, and let $\boldsymbol{\delta} = \{\delta_1, \dots, \delta_n\}$ denote the group indicator of the pooled sample, where $n = n_1 + n_0$, $\delta_i = 1$ if $\mathbf{z}_i \in \mathbb{X}$ and $\delta_i = 0$ if $\mathbf{z}_i \in \mathbb{Y}$, for $i = 1, \dots, n$.

We assume that both F and G are supported on a d -dimensional geodesically convex Riemannian manifold $\mathcal{M}$ embedded in the ambient space $\mathcal{R}^D$ with $d \ll D$. This assumption, also known as the manifold hypothesis, is reasonable in many real data examples and

prevailing in high-dimensional data analysis (Pless and Souvenir, 2009; Meilă and Zhang, 2024). For example, each image often contain several thousand pixels, while Gong et al. (2019) showed that the intrinsic dimension of a face image dataset is around 10 to 20. In addition, the manifold hypothesis has been invoked in analyses of bioinformatics (Moon et al., 2018), natural language processing (Zhao et al., 2021), etc.

The manifold hypothesis is also investigated in the literature of two-sample testing (Arias-Castro et al., 2018; Kim et al., 2019; Cheng and Xie, 2024). To illustrate how the manifold hypothesis is helpful in two-sample testing applications, we consider an example of biological data analysis. In biological experiments, an essential task is to investigate whether the protein expression levels vary among cells of different conditions. Based on the experimental results, researchers are able to verify the disease mechanism in terms of protein expression and the efficacy of treatments. Because interactions among proteins only depend on a limited number of proteins' biochemical and structural properties (Terradot et al., 2004), the manifold hypothesis is usually imposed in the analysis of protein expression levels (Dorrity et al., 2020).

2.2 Delaunay Triangulation on Manifold

In Riemannian geometry, let $\ell_{\mathcal{M}}(\mathbf{a}, \mathbf{b})$ denote the shortest geodesic path between two points $\mathbf{a}, \mathbf{b} \in \mathcal{M}$, whose length is the geodesic distance, $d_{\mathcal{M}}(\mathbf{a}, \mathbf{b})$, induced by the ambient space $\mathcal{R}^D$. The manifold $\mathcal{M}$ is assumed to be geodesically convex meaning that the path $\ell_{\mathcal{M}}(\mathbf{a}, \mathbf{b})$ uniquely exists for any $\mathbf{a}, \mathbf{b} \in \mathcal{M}$. Let $\mathbf{T}_{\mathbf{a}}\mathcal{M}$ denote the tangent space of $\mathcal{M}$ at point $\mathbf{a} \in \mathcal{M}$. The logarithmic map $\log_{\mathbf{a}} : \mathcal{M} \rightarrow \mathbf{T}_{\mathbf{a}}\mathcal{M}$ is defined as the vector starting from $\mathbf{a}$ whose length equals $d_{\mathcal{M}}(\mathbf{a}, \mathbf{b})$ and direction is the derivative of $\ell_{\mathcal{M}}(\mathbf{a}, \mathbf{b})$ at $\mathbf{a}$. Since $\ell_{\mathcal{M}}(\mathbf{a}, \mathbf{b})$ always exists uniquely for any two points $\mathbf{a}, \mathbf{b} \in \mathcal{M}$, the domain of the logarithmic map

$\log_{\mathbf{a}}(\cdot)$ is the whole $\mathcal{M}$, for all $\mathbf{a} \in \mathcal{M}$. We can generalize the geometric concepts from the Euclidean space to the geodesically convex Riemannian manifold $\mathcal{M}$ as follows.

- *Convex set*: A subset $\mathcal{C} \subset \mathcal{M}$ is convex on $\mathcal{M}$ if for any two points $\mathbf{a}, \mathbf{b} \in \mathcal{C}$, $\ell_{\mathcal{M}}(\mathbf{a}, \mathbf{b}) \subset \mathcal{C}$.
- *Convex hull*: For a set of points $\mathbb{Z} \subset \mathcal{M}$, the convex hull of $\mathbb{Z}$, denoted by $\mathcal{H}_{\mathcal{M}}(\mathbb{Z}) \subset \mathcal{M}$, is the intersection of all convex sets on $\mathcal{M}$ that contain $\mathbb{Z}$.
- *Ball and sphere*: A ball on $\mathcal{M}$ with the center $\mathbf{z} \in \mathcal{M}$ and radius $r > 0$ is defined as $\mathcal{B}_{\mathcal{M}}(\mathbf{z}, r) = \{\mathbf{z}^* \in \mathcal{M}; d_{\mathcal{M}}(\mathbf{z}, \mathbf{z}^*) \leq r\}$, and the corresponding sphere is defined as $\partial\mathcal{B}_{\mathcal{M}}(\mathbf{z}, r) = \{\mathbf{z}^* \in \mathcal{M}; d_{\mathcal{M}}(\mathbf{z}, \mathbf{z}^*) = r\}$.
- *Simplex and circumscribing ball*: For $k = 0, \dots, d$, the k -simplex of $k + 1$ points $\mathbf{z}_1, \dots, \mathbf{z}_{k+1} \in \mathcal{M}$ is the convex hull $\mathcal{H}_{\mathcal{M}}(\{\mathbf{z}_1, \dots, \mathbf{z}_{k+1}\})$, and the circumscribing ball of $\mathbf{z}_1, \dots, \mathbf{z}_{k+1}$ is the smallest radius ball among all balls on $\mathcal{M}$ containing $\mathbf{z}_1, \dots, \mathbf{z}_{k+1}$. In particular, the 1-simplex of points $\{\mathbf{z}_1, \mathbf{z}_2\}$ is the shortest geodesic path $\ell_{\mathcal{M}}(\mathbf{z}_1, \mathbf{z}_2)$.
- *Projection*: For a set of points $\mathbb{Z} \subset \mathcal{M}$, the projection $\mathbf{p}^*$ of $\mathbf{p} \in \mathcal{M}$ on $\mathcal{H}_{\mathcal{M}}(\mathbb{Z})$ is defined as $\mathbf{p}^* = \arg \min_{\mathbf{q} \in \mathcal{H}_{\mathcal{M}}(\mathbb{Z})} d_{\mathcal{M}}(\mathbf{p}, \mathbf{q})$, in the case that the right hand side exists uniquely; otherwise, we first find $\mathbf{p}^\dagger = \arg \min_{\mathbf{q} \in \mathcal{H}_{\mathcal{R}^D}(\mathbb{Z})} d_{\mathcal{R}^D}(\mathbf{p}, \mathbf{q})$ and define $\mathbf{p}^*$ as the Euclidean projection of $\mathbf{p}^\dagger$ onto $\mathcal{H}_{\mathcal{M}}(\mathbb{Z})$, which exists uniquely almost everywhere with respect to the Lebesgue measure of $\mathcal{R}^D$ (Bhattacharya and Patrangenaru, 2003).

Following Leibon and Letscher (2000), the Delaunay triangulation can be defined on $\mathcal{M}$ as follows.

Definition 1. For a set of n points $\mathbb{Z} = \{\mathbf{z}_1, \dots, \mathbf{z}_n\} \subset \mathcal{M}$, the Delaunay triangulation of $\mathbb{Z}$, denoted by $\mathcal{DT}_{\mathcal{M}}(\mathbb{Z})$, is a mesh of m d -simplices $\{\mathcal{S}_1, \dots, \mathcal{S}_m\}$ satisfying:

1. For $j = 1, \dots, m$, the set of $d+1$ vertices of simplex $\mathcal{S}_j$, denoted by $\mathbb{V}(\mathcal{S}_j)$, is a subset of $\mathbb{Z}$ and is not contained in any k -simplex ($k = 0, \dots, d-1$) on $\mathcal{M}$.
2. Different simplices are disjoint except on their shared boundaries.
3. The union $\mathcal{S}_1 \cup \dots \cup \mathcal{S}_m$ is the convex hull of $\mathbb{Z}$ on $\mathcal{M}$, $\mathcal{H}_{\mathcal{M}}(\mathbb{Z})$.
4. The empty ball property: For $j = 1, \dots, m$, the circumscribing ball of $\mathcal{S}_j$ on $\mathcal{M}$ contains no points of $\mathbb{Z}$ except on its sphere.

We refer to simplices $\mathcal{S}_1, \dots, \mathcal{S}_m \in \mathcal{DT}_{\mathcal{M}}(\mathbb{Z})$ as the Delaunay simplices. According to [Leibon and Letscher \(2000\)](#), the Delaunay triangulation $\mathcal{DT}_{\mathcal{M}}(\mathbb{Z})$ is unique if and only if the pooled sample $\mathbb{Z}$ is generic, i.e., any $d+2$ points in $\mathbb{Z}$ do not lie on a sphere on $\mathcal{M}$ and any $k+2$ points do not lie in a k -simplex on $\mathcal{M}$ ($k = 1, \dots, d-1$). As long as both F and G have Lipschitz continuous densities f and g with respect to the measure on $\mathcal{M}$ induced by the Lebesgue measure of the ambient Euclidean space, the pooled sample $\mathbb{Z}$ is generic with probability one. Being the geometric dual of the Voronoi diagram, the Delaunay triangulation generates a mesh of simplices on $\mathcal{M}$ that are most regularized in shape. Figure 1 shows the empty ball property of the Delaunay triangulation $\mathcal{DT}_{\mathcal{M}}(\mathbb{Z})$ for $\mathcal{M} = \mathcal{R}^2$, and $\mathcal{DT}_{\mathcal{M}}(\mathbb{Z})$ maximizes the minimum angle in all the triangles (2-simplices) over all possible triangulations of $\mathbb{Z}$ ([Sibson, 1978](#)).

3 Delaunay Weighted Test

3.1 Delaunay Weight

Considering the simplest case $\mathcal{M} = \mathcal{R}^d$, for each internal point $\mathbf{p} \in \mathcal{H}_{\mathcal{R}^d}(\mathbb{Z}) = \mathcal{S}_1 \cup \dots \cup \mathcal{S}_m$, there exists a Delaunay simplex $\mathcal{S}(\mathbf{p}; \mathbb{Z}) \in \mathcal{DT}_{\mathcal{R}^d}(\mathbb{Z})$ such that $\mathbf{p} \in \mathcal{S}(\mathbf{p}; \mathbb{Z})$. We define

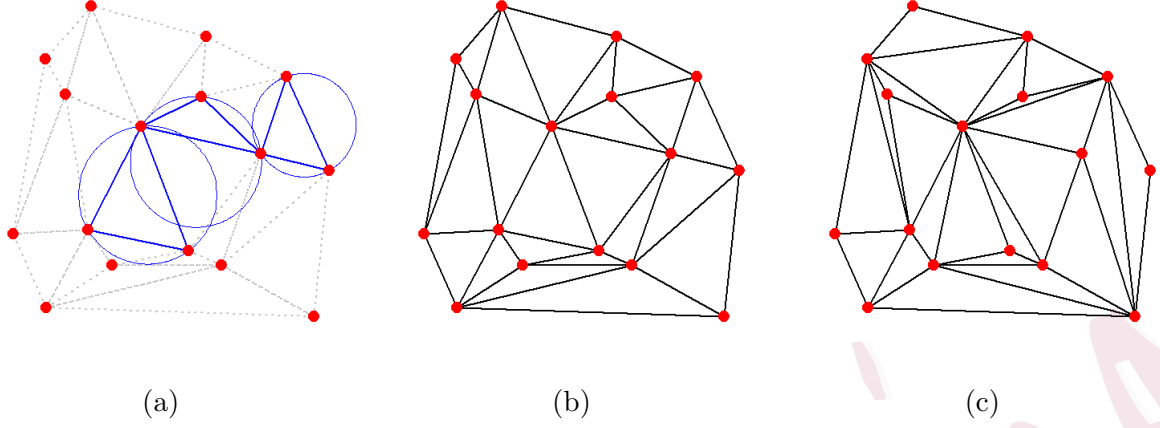


Figure 1: (a) Graphical illustration of the empty ball property of the Delaunay triangulation with $\mathcal{M} = \mathcal{R}^2$; (b)–(c) The Delaunay triangulation versus a random triangulation of the same $\mathbb{Z}$.

the Delaunay weight vector for $\mathbf{p} \in \mathcal{H}_{\mathcal{R}^d}(\mathbb{Z})$ with respect to $\mathbb{Z}$ as a vector $\mathbf{w}(\mathbf{p}; \mathbb{Z}) = (w_1(\mathbf{p}; \mathbb{Z}), \dots, w_n(\mathbf{p}; \mathbb{Z}))^\top$ such that

- (a) $w_i(\mathbf{p}; \mathbb{Z}) \in [0, 1]$ if $\mathbf{z}_i \in \mathbb{V}(\mathcal{S}(\mathbf{p}; \mathbb{Z}))$, and $w_i(\mathbf{p}; \mathbb{Z}) = 0$ if $\mathbf{z}_i \notin \mathbb{V}(\mathcal{S}(\mathbf{p}; \mathbb{Z}))$, where $\mathbb{V}(\mathcal{S}(\mathbf{p}; \mathbb{Z}))$ denotes the vertex set of the simplex $\mathcal{S}(\mathbf{p}; \mathbb{Z})$;
- (b) $\sum_{i=1}^n w_i(\mathbf{p}; \mathbb{Z}) = 1$;
- (c) $\sum_{i=1}^n w_i(\mathbf{p}; \mathbb{Z}) \cdot \mathbf{z}_i = \mathbf{p}$.

According to [Abadie and L'Hour \(2021\)](#), the Delaunay weight vector $\mathbf{w}(\mathbf{p}; \mathbb{Z})$ for $\mathbf{p} \in \mathcal{H}_{\mathcal{R}^d}(\mathbb{Z})$ corresponds to the convex combination of $\mathbb{Z}$ with the smallest compound discrepancy $\sum_{i=1}^n w_i \|\mathbf{p} - \mathbf{z}_i\|^2$ among all convex combinations satisfying $\mathbf{p} = \sum_{i=1}^n w_i \mathbf{z}_i$. That is, $\mathbf{w}(\mathbf{p}; \mathbb{Z})$ selects at most $d + 1$ points in $\mathbb{Z}$ (those with $w_i > 0$) that are convexly combined to form $\mathbf{p}$ and closest to $\mathbf{p}$ in the sense of the minimal compound discrepancy.

Conditions (a)–(c) only define the Delaunay weight vector for an internal point $\mathbf{p} \in \mathcal{H}_{\mathcal{R}^d}(\mathbb{Z})$. For generalization to an external point $\mathbf{p} \notin \mathcal{H}_{\mathcal{R}^d}(\mathbb{Z})$, we utilize the projection idea from Section 2.3 of [Abadie and L'Hour \(2021\)](#). Let $\mathbf{w}^{(\lambda)}(\mathbf{p}; \mathbb{Z}) = (w_1^{(\lambda)}(\mathbf{p}; \mathbb{Z}), \dots, w_n^{(\lambda)}(\mathbf{p}; \mathbb{Z}))^\top$

be the solution to the optimization problem,

$$\min \left\| \mathbf{p} - \sum_{i=1}^n w_i \mathbf{z}_i \right\|^2 + \lambda \sum_{i=1}^n w_i \|\mathbf{p} - \mathbf{z}_i\|^2,$$

subject to $\mathbf{w} = (w_1, \dots, w_n)^\top \in [0, 1]^n$ and $\sum_{i=1}^n w_i = 1$, where $\|\cdot\|$ denotes the Euclidean norm. The Delaunay weight vector $\mathbf{w}(\mathbf{p}; \mathbb{Z})$ for $\mathbf{p} \in \mathcal{H}_{\mathcal{R}^d}(\mathbb{Z})$ can be rewritten as $\mathbf{w}(\mathbf{p}; \mathbb{Z}) = \lim_{\lambda \rightarrow 0} \mathbf{w}^{(\lambda)}(\mathbf{p}; \mathbb{Z})$. For $\mathbf{p} \notin \mathcal{H}_{\mathcal{R}^d}(\mathbb{Z})$, [Abadie and L'Hour \(2021\)](#) showed that $\lim_{\lambda \rightarrow 0} \sum_{i=1}^n w_i^{(\lambda)}(\mathbf{p}; \mathbb{Z}) \mathbf{z}_i = \mathbf{p}^*$, where $\mathbf{p}^*$ is the projection of $\mathbf{p}$ on the convex hull $\mathcal{H}_{\mathcal{R}^d}(\mathbb{Z})$. Therefore, we have

$$\mathbf{w}(\mathbf{p}^*; \mathbb{Z}) = \lim_{\lambda \rightarrow 0} \mathbf{w}^{(\lambda)}(\mathbf{p}; \mathbb{Z}), \quad \forall \mathbf{p} \in \mathcal{R}^d. \quad (1)$$

Further, we can generalize the definition of the Delaunay weight for all $\mathbf{p} \in \mathcal{R}^d$ by substituting condition (c) with

$$(c^\dagger) \quad \sum_{i=1}^n w_i(\mathbf{p}; \mathbb{Z}) \cdot \mathbf{z}_i = \mathbf{p}^*.$$

However, when $\mathcal{M}$ is not Euclidean, condition $(c^\dagger)$ is no longer applicable because it relies on the linear operation of Cartesian coordinates, which is not well-defined on $\mathcal{M}$. To generalize the Delaunay weight vector to nonlinear manifolds, we make use of the logarithmic map.

Definition 2. For a set $\mathbb{Z} = \{\mathbf{z}_1, \dots, \mathbf{z}_n\} \subset \mathcal{M}$, the Delaunay weight vector of $\mathbf{p} \in \mathcal{M}$ is $\mathbf{w}(\mathbf{p}; \mathbb{Z}) = (w_1(\mathbf{p}; \mathbb{Z}), \dots, w_n(\mathbf{p}; \mathbb{Z}))^\top$ satisfying (a), (b) and

$$(c^\dagger) \quad \sum_{i=1}^n w_i(\mathbf{p}; \mathbb{Z}) \log_{\mathbf{p}^*}(\mathbf{z}_i) = \mathbf{0},$$

where $\mathbf{p}^*$ is the projection of $\mathbf{p}$ on the convex hull $\mathcal{H}_{\mathcal{M}}(\mathbb{Z})$ and $\mathcal{S}(\mathbf{p}; \mathbb{Z}) \in \mathcal{DT}_{\mathcal{M}}(\mathbb{Z})$ in (a) is the Delaunay simplex containing $\mathbf{p}^*$.

For the special case that $\mathcal{M} = \mathcal{R}^d$, $(c^\dagger)$ and $(c^\ddagger)$ are equivalent because

$$\sum_{i=1}^n w_i(\mathbf{p}; \mathbb{Z}) \cdot \overrightarrow{\mathbf{p}^* \mathbf{z}_i} = \overrightarrow{\mathbf{p}^* \mathbf{p}^*} \iff \sum_{i=1}^n w_i(\mathbf{p}; \mathbb{Z}) \log_{\mathbf{p}^*}(\mathbf{z}_i) = \log_{\mathbf{p}^*}(\mathbf{p}^*), \quad (2)$$

where $\overrightarrow{\mathbf{p}^* \mathbf{z}_i}$ denotes the vector from $\mathbf{p}^*$ to $\mathbf{z}_i$. Therefore, condition $(c^\ddagger)$ is indeed a generalization of $(c^\dagger)$ on a nonlinear manifold. Since the Delaunay triangulation $\mathcal{DT}_{\mathcal{M}}(\mathbb{Z})$ is unique for generic $\mathbb{Z}$ (Leibon and Letscher, 2000), the Delaunay weight vector $\mathbf{w}(\mathbf{p}; \mathbb{Z})$ is unique for $\mathbb{Z}$ and any point $\mathbf{p} \in \mathcal{M}$, as shown in Section S1 of the supplementary material. Based on Definition 2, the Delaunay weight matrix of $\mathbb{Z}$ is defined as $\mathbf{\Gamma}_{\mathbb{Z}} = (\gamma_{ij})_{n \times n}$, with

$$\gamma_{ii} = 0, \quad (\gamma_{i1}, \dots, \gamma_{i,i-1}, \gamma_{i,i+1}, \dots, \gamma_{in})^\top = \mathbf{w}(\mathbf{z}_i; \mathbb{Z} \setminus \{\mathbf{z}_i\}), \quad i = 1, \dots, n, \quad (3)$$

where γ_{ij} measures the geometric proximity of $\mathbf{z}_j$ to $\mathbf{z}_i$ and $\mathbb{Z} \setminus \{\mathbf{z}_i\}$ represents the remaining set of $\mathbb{Z}$ excluding $\mathbf{z}_i$. Because γ_{ij} and γ_{ji} are typically not the same, the Delaunay weight matrix $\mathbf{\Gamma}_{\mathbb{Z}}$ is a directed weighted adjacency matrix of $\mathbb{Z}$. Unlike the kernel, local regression and k -NN based approaches which only consider local distances to $\mathbf{z}_i$, the Delaunay weight $\mathbf{w}(\mathbf{z}_i; \mathbb{Z} \setminus \{\mathbf{z}_i\})$ evaluates the geometric proximity of the vertices of the Delaunay simplex $\mathcal{S}(\mathbf{z}_i; \mathbb{Z} \setminus \{\mathbf{z}_i\})$ to $\mathbf{z}_i$ using both the distance and direction information. In particular, γ_{ij} and γ_{ik} can be different even if $\|\log_{\mathbf{z}_i^*}(\mathbf{z}_j)\| = \|\log_{\mathbf{z}_i^*}(\mathbf{z}_k)\|$. Indeed, we see from Definition 2 that $\mathbf{w}(\mathbf{z}_i; \mathbb{Z} \setminus \{\mathbf{z}_i\})$ depends on the vectors $\log_{\mathbf{z}_i^*}(\mathbf{z}_j)$ rather than merely the distances $\|\log_{\mathbf{z}_i^*}(\mathbf{z}_j)\|$, for $j \neq i$.

To gain more insights on the Delaunay weight, we see that, when $\mathbf{z}_i$ is an inner point of the Delaunay simplex $\mathcal{S}(\mathbf{z}_i; \mathbb{Z} \setminus \{\mathbf{z}_i\})$, the non-zero elements of $\mathbf{w}(\mathbf{z}_i; \mathbb{Z} \setminus \{\mathbf{z}_i\})$ correspond to the points in $\mathbb{Z} \setminus \{\mathbf{z}_i\}$ from all directions surrounding $\mathbf{z}_i$. In contrast, under sparse or nonuniform sampling, nearest neighbors of $\mathbf{z}_i$ cannot cover all directions from $\mathbf{z}_i$ because

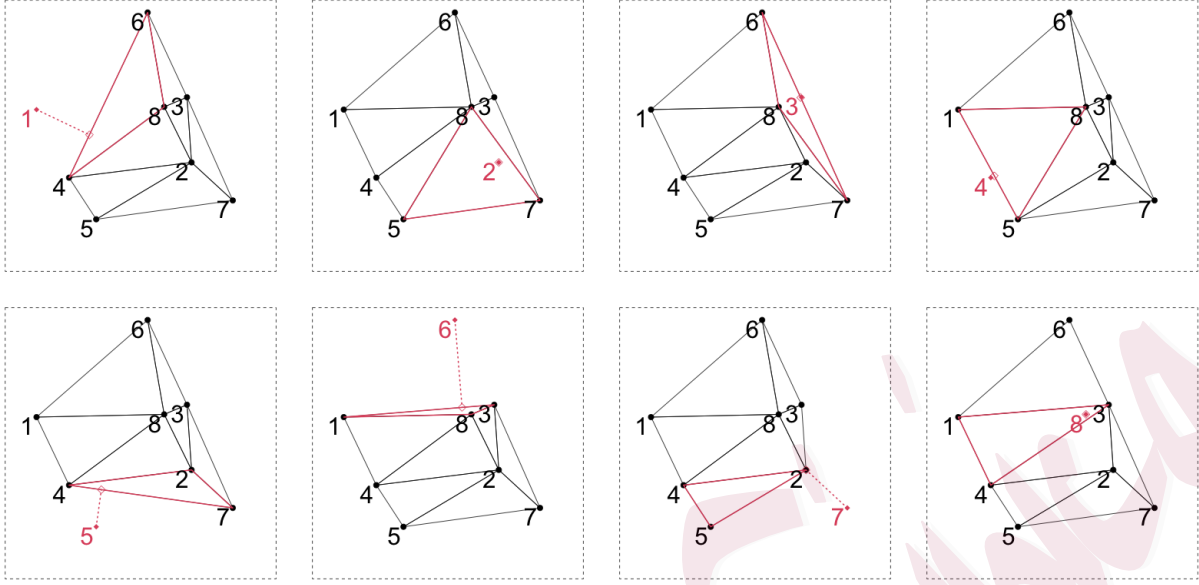


Figure 2: Graphical illustration of the Delaunay simplices (triangles) $\mathcal{S}(\mathbf{z}_i; \mathbb{Z} \setminus \{\mathbf{z}_i\})$ for $i = 1, \dots, 4$ (top panel) and for $i = 5, \dots, 8$ (bottom panel) computed from a random sample $\mathbb{Z} = \{\mathbf{z}_1, \dots, \mathbf{z}_8\}$ on $\mathcal{R}^2$.

they tend to cluster in certain directions. We illustrate this point using an example with a sample $\mathbb{Z} = \{\mathbf{z}_1, \dots, \mathbf{z}_8\}$ in $\mathcal{M} = \mathcal{R}^2$. Figure 2 shows the Delaunay simplex $\mathcal{S}(\mathbf{z}_i; \mathbb{Z} \setminus \{\mathbf{z}_i\})$ that contains the projection of $\mathbf{z}_i$ on the convex hull $\mathcal{H}_{\mathcal{M}}(\mathbb{Z} \setminus \{\mathbf{z}_i\})$. Based on $\mathcal{S}(\mathbf{z}_i; \mathbb{Z} \setminus \{\mathbf{z}_i\})$ ($i = 1, \dots, 8$), the weight vectors $\mathbf{w}(\mathbf{z}_i; \mathbb{Z} \setminus \{\mathbf{z}_i\})$ are concatenated to form the Delaunay weight matrix $\mathbf{\Gamma}_{\mathbb{Z}}$, as shown in Figure 3. For $\mathbf{z}_2 \in \mathcal{H}_{\mathcal{M}}(\mathbb{Z} \setminus \{\mathbf{z}_2\})$, the three nearest neighbors of $\mathbf{z}_2$ are $\mathbf{z}_3, \mathbf{z}_7$ and $\mathbf{z}_8$, which are all from the top-right direction of $\mathbf{z}_2$ and the information from the bottom-left direction is not captured. In contrast, the Delaunay simplex $\mathcal{S}(\mathbf{z}_2; \mathbb{Z} \setminus \{\mathbf{z}_2\})$ is formed by $\mathbf{z}_5, \mathbf{z}_7$ and $\mathbf{z}_8$, which covers all directions from $\mathbf{z}_2$, and the relative directions of $\overrightarrow{\mathbf{z}_2 \mathbf{z}_k}$ ($k = 5, 7, 8$) are used in constructing $\mathbf{w}(\mathbf{z}_2; \mathbb{Z} \setminus \{\mathbf{z}_2\})$. Therefore, the Delaunay weight utilizes the local geometric information in a more comprehensive way than the k -NN.

To better demonstrate the global advantage of the Delaunay weight matrix in capturing direction information, we generate a dataset $\mathbb{Z}$ in $\mathcal{R}^2$, where $\mathbb{Z} = \mathbb{X} \cup \mathbb{Y}$, $\mathbb{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_{20}\}$

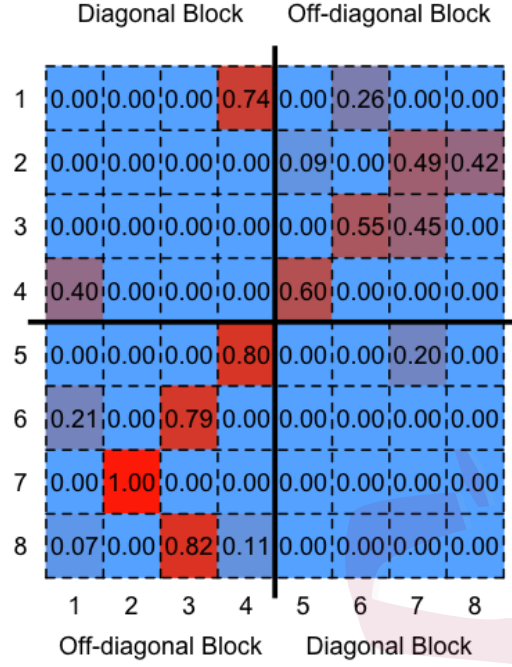


Figure 3: The Delaunay weight matrix $\Gamma_{\mathbb{Z}}$ computed from the $\mathbb{Z}$ in Figure 2, with the group indicators $\delta_i = 1$, for $i = 1, \dots, 4$, and $\delta_i = 0$ otherwise. The test statistic T_{DW} equals the average of all the Delaunay weights in the two diagonal blocks of $\Gamma_{\mathbb{Z}}$.

contains 20 points from $\text{MVN}(\mathbf{0}, \text{diag}(4, 1))$, and $\mathbb{Y} = \{\mathbf{y}_1, \dots, \mathbf{y}_{20}\}$ is obtained by rotating $\mathbb{X}$ by $\pi/2$ counterclockwise with respect to the origin. Here, we display in Figure 4 (a)–(c) the adjacency graphs of $\mathbb{Z}$ induced by k -NN, k -MST (Chen and Friedman, 2017; Chen et al., 2018) and $\Gamma_{\mathbb{Z}}$ respectively. Specifically, $k = 3$ for both k -NN and k -MST, and the adjacency graph induced by $\Gamma_{\mathbb{Z}}$ is directed with weight $\gamma_{ij} > 0$ for the edge $\mathbf{z}_i \rightarrow \mathbf{z}_j$. Unlike k -NN and k -MST graphs, the adjacency graph induced by the Delaunay weight matrix includes more edges connecting points in the principal directions of $\mathbb{X}$ and $\mathbb{Y}$. In this example, the principal directions of $\mathbb{X}$ is along the X -axis (0 and $\pm\pi$), and that of $\mathbb{Y}$ is along the Y -axis ($\pm\pi/2$). As a result, the Delaunay weight matrix better captures the principal directions of $\mathbb{X}$ and $\mathbb{Y}$. To visualize this more explicitly, we show the kernel density estimate (KDE) of directions among edges pointing $\mathbf{x}_i$'s and $\mathbf{y}_i$'s (see Section S2 of the supplementary material for computing the KDEs of direction) in Figure 4 (d)–(f). Specifically, Figure 4 (f) shows

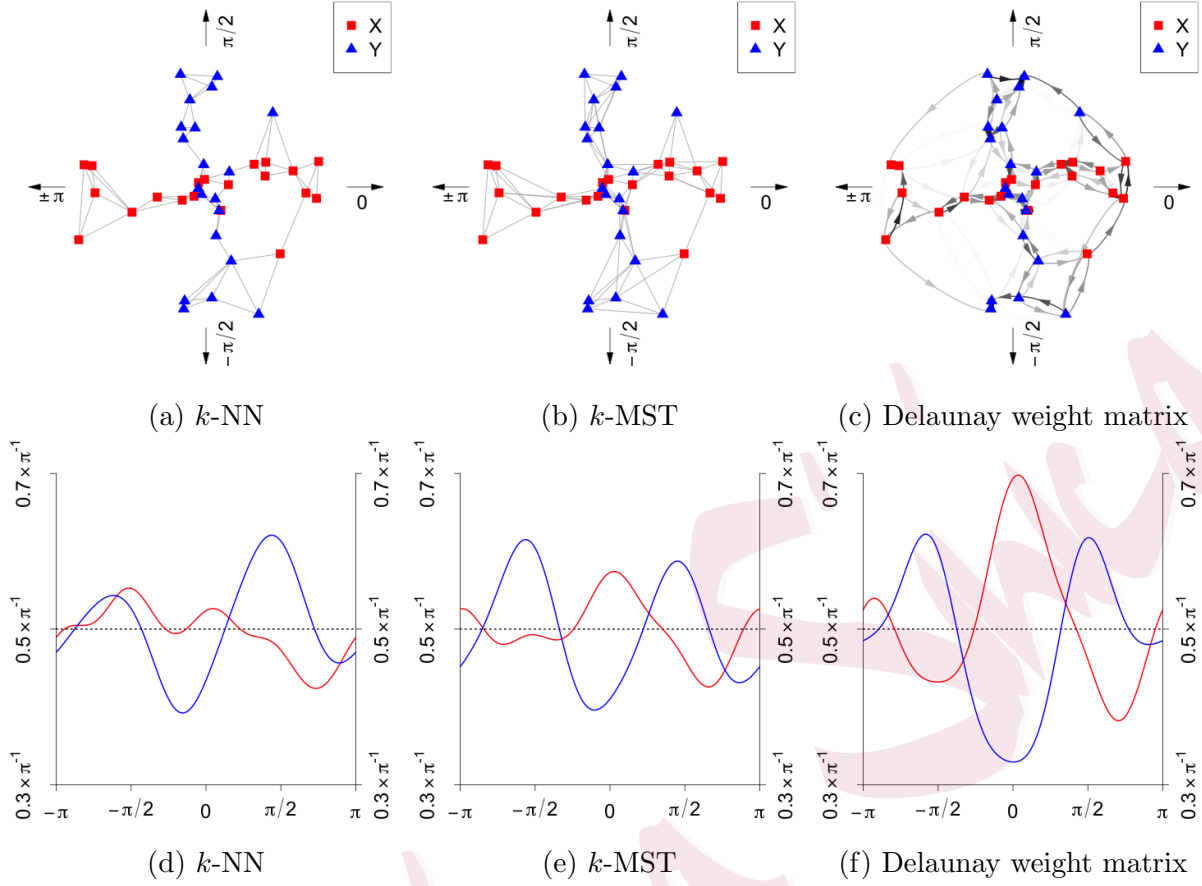


Figure 4: Graphical illustration of the global advantage of the Delaunay weight matrix in capturing direction information. (a)–(c): Adjacency graphs induced by k -NN, k -MST and the Delaunay weight matrix respectively. In (c), the shade of the edge $\mathbf{z}_i \rightarrow \mathbf{z}_j$ indicates the magnitude of γ_{ij} . (d)–(f): The kernel density estimates (KDEs) of directions among edges pointing to $\mathbf{x}_i$'s (red) and $\mathbf{y}_i$'s (blue) in neighborhood graphs induced by k -NN, k -MST and the Delaunay weight matrix, respectively.

that among edges pointing to $\mathbf{x}_i$'s, more weights are assigned to edges in directions close to the principal directions of $\mathbb{X}$. The same is true for $\mathbb{Y}$. In contrast, directions of edges pointing $\mathbf{x}_i$'s in both k -NN and k -MST are less prominent in the principal directions as shown in Figure 4 (d)–(e).

3.2 Test Statistic

For the pooled sample $\mathbb{Z} = \{\mathbf{z}_1, \dots, \mathbf{z}_n\}$ with the group indicator $\boldsymbol{\delta} = \{\delta_1, \dots, \delta_n\}$, the k -NN test statistic in Schilling (1986) is

$$\sum_{j=1}^n \sum_{i \neq j} \sum_{l=1}^k I(\mathbf{z}_j \text{ is the } l\text{-th nearest neighbor of } \mathbf{z}_i) \times I(\delta_i = \delta_j),$$

which has inspired a family of nonparametric two-sample tests for detecting distribution differences via a proximity graph of the pooled sample (Rosenbaum, 2005; Chen and Friedman, 2017; Chen et al., 2018). Based on the Delaunay weight γ_{ij} , we construct the test statistic,

$$T_{\text{DW}} = \frac{1}{n} \sum_{j=1}^n \sum_{i \neq j} \gamma_{ij} \times I(\delta_i = \delta_j), \quad (4)$$

where the new geometric proximity measure γ_{ij} accounts for the manifold structure and direction information.

3.3 Approximation of Delaunay Weight

In practice, the underlying manifold $\mathcal{M}$ is typically unknown, so that exact computation of the Delaunay weight matrix $\mathbf{\Gamma}_{\mathbb{Z}}$ via Definition 2 (and thus T_{DW}) is infeasible. Because $\mathbf{\Gamma}_{\mathbb{Z}}$ is constructed by Delaunay weight vectors $\mathbf{w}(\mathbf{z}_i; \mathbb{Z} \setminus \{\mathbf{z}_i\})$ which only measure the geometric proximity between $\mathbf{z}_i$ and the points locally surrounding $\mathbf{z}_i$ (i.e., the vertices of $\mathcal{S}(\mathbf{z}_i; \mathbb{Z} \setminus \{\mathbf{z}_i\})$), we can utilize manifold learning techniques to approximate $\mathbf{\Gamma}_{\mathbb{Z}}$ by $\mathbf{\Gamma}_{\tilde{\mathbb{Z}}}$, where $\tilde{\mathbb{Z}} = \{\tilde{\mathbf{z}}_1, \dots, \tilde{\mathbf{z}}_n\} \subset \mathcal{R}^d$ is a low-dimensional Euclidean representation of $\mathbb{Z}$ whose Delaunay weight matrix can be computed. Generally speaking, any nonlinear dimension reduction technique that aims to preserve the local geometric structure of $\mathbb{Z}$ can be applied to obtain $\tilde{\mathbb{Z}}$. We follow the Isomap algorithm (Tenenbaum et al., 2000) for its broad applicability,

which requires an adjacency graph. A k -NN or ϵ -ball graph is often used, but it cannot ensure connectivity. If the graph contains more than one connected components, then elements in $\mathbf{\Gamma}_{\tilde{\mathbb{Z}}}$ corresponding to all pairs of points from different components must be zero, which incurs information loss and power deterioration. To circumvent this issue, we propose using the union of a k -NN graph and the MST graph as our adjacency graph. We then estimate the geodesic distances $d_{\mathcal{M}}(\mathbf{z}_i, \mathbf{z}_j)$ by applying Dijkstra's algorithm to our adjacency graph and apply the classical MDS (Kruskal, 1964) to obtain $\tilde{\mathbb{Z}} \subset \mathcal{R}^d$; see Section S3 of the supplementary material for details. In case of unknown intrinsic dimension d , any existing method can be applied to estimate d , e.g., Facco et al. (2017) as summarized in Section S3 of the supplementary material. Our numerical experiments show that our method is robust to estimation of d .

To compute $\mathbf{\Gamma}_{\tilde{\mathbb{Z}}}$, we note the equivalence of conditions (c[†]) and (c[‡]) in (2) for the low-dimensional Euclidean representation $\tilde{\mathbb{Z}}$. Once we know the Delaunay simplex $\mathcal{S}(\tilde{\mathbf{z}}_i; \tilde{\mathbb{Z}} \setminus \{\tilde{\mathbf{z}}_i\})$ that contains $\tilde{\mathbf{z}}_i^*$, which is the projection of $\tilde{\mathbf{z}}_i$ on the convex hull $\mathcal{H}_{\mathcal{R}^d}(\tilde{\mathbb{Z}} \setminus \{\tilde{\mathbf{z}}_i\})$, we can easily compute $\mathbf{w}(\tilde{\mathbf{z}}_i; \tilde{\mathbb{Z}} \setminus \{\tilde{\mathbf{z}}_i\}) = (\tilde{\gamma}_{i1}, \dots, \tilde{\gamma}_{i,i-1}, \tilde{\gamma}_{i,i+1}, \dots, \tilde{\gamma}_{in})^\top$ by solving the linear equations,

$$\begin{cases} \tilde{\gamma}_{ij} = 0, & \forall \tilde{\mathbf{z}}_j \notin \mathbb{V}(\mathcal{S}(\tilde{\mathbf{z}}_i; \tilde{\mathbb{Z}} \setminus \{\tilde{\mathbf{z}}_i\})), \\ \sum_{\tilde{\mathbf{z}}_j \in \mathbb{V}(\mathcal{S}(\tilde{\mathbf{z}}_i; \tilde{\mathbb{Z}} \setminus \{\tilde{\mathbf{z}}_i\}))} \tilde{\gamma}_{ij} = 1, \\ \sum_{\tilde{\mathbf{z}}_j \in \mathbb{V}(\mathcal{S}(\tilde{\mathbf{z}}_i; \tilde{\mathbb{Z}} \setminus \{\tilde{\mathbf{z}}_i\}))} \tilde{\gamma}_{ij} \tilde{\mathbf{z}}_j = \tilde{\mathbf{z}}_i^*. \end{cases} \quad (5)$$

The remaining difficulty in computing $\mathbf{\Gamma}_{\tilde{\mathbb{Z}}}$ lies in finding the Delaunay simplex $\mathcal{S}(\tilde{\mathbf{z}}_i; \tilde{\mathbb{Z}} \setminus \{\tilde{\mathbf{z}}_i\})$, for $i = 1, \dots, n$.

Although the DELAUNAYSPARSE algorithm by Chang et al. (2020) can efficiently find $\mathcal{S}(\tilde{\mathbf{z}}_i; \tilde{\mathbb{Z}} \setminus \{\tilde{\mathbf{z}}_i\})$ for $\tilde{\mathbf{z}}_i \in \mathcal{H}_{\mathcal{R}^d}(\tilde{\mathbb{Z}} \setminus \{\tilde{\mathbf{z}}_i\})$, it does not work for those $\tilde{\mathbf{z}}_i \notin \mathcal{H}_{\mathcal{R}^d}(\tilde{\mathbb{Z}} \setminus \{\tilde{\mathbf{z}}_i\})$.

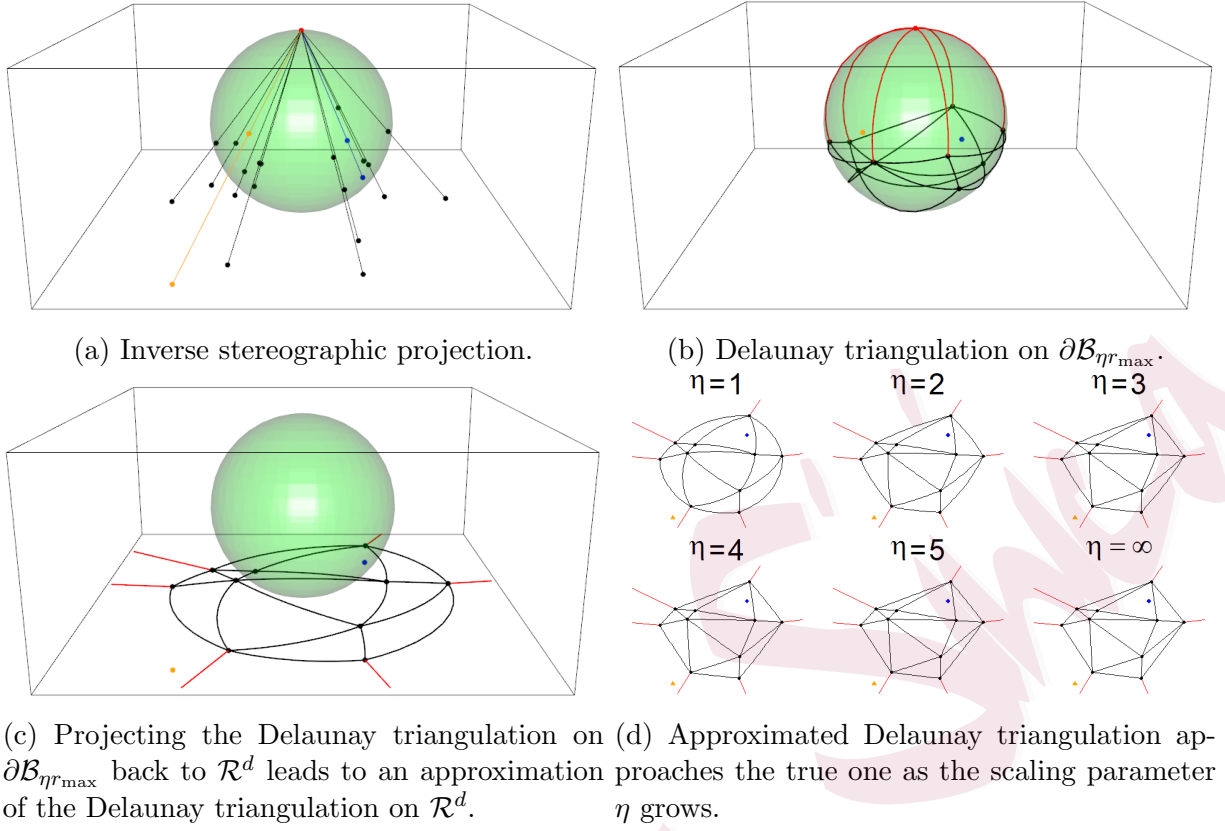


Figure 5: Graphical illustration of the stereographic projected DELAUNAYSPARSE algorithm.

To tackle this difficulty, we incorporate the inverse stereographic projection to the DELAUNAYSPARSE algorithm and develop the stereographic projected DELAUNAYSPARSE algorithm to find $\mathcal{S}(\tilde{\mathbf{z}}_i; \tilde{\mathbb{Z}} \setminus \{\tilde{\mathbf{z}}_i\})$, regardless of whether $\tilde{\mathbf{z}}_i \in \mathcal{H}_{\mathcal{R}^d}(\tilde{\mathbb{Z}} \setminus \{\tilde{\mathbf{z}}_i\})$ or not. In summary, the stereographic projected DELAUNAYSPARSE algorithm first projects $\tilde{\mathbb{Z}} \subset \mathcal{R}^d$ to $\tilde{\mathbb{Z}} \subset \partial\mathcal{B}_{\eta r_{\max}}$, then locates $\mathcal{S}(\tilde{\mathbf{z}}_i; \tilde{\mathbb{Z}} \setminus \{\tilde{\mathbf{z}}_i\})$ in a breadth-first search manner (similar to the crystallization search in Gu and Yin (2021)) and finally projects $\mathcal{S}(\tilde{\mathbf{z}}_i; \tilde{\mathbb{Z}} \setminus \{\tilde{\mathbf{z}}_i\})$ back to $\mathcal{R}^d$ for $\mathcal{S}(\tilde{\mathbf{z}}_i; \tilde{\mathbb{Z}} \setminus \{\tilde{\mathbf{z}}_i\})$. Such a procedure can be regarded as an extension of the original DELAUNAYSPARSE algorithm (Chang et al., 2020) to the case of $\tilde{\mathbf{z}}_i \notin \mathcal{H}_{\mathcal{R}^d}(\tilde{\mathbb{Z}} \setminus \{\tilde{\mathbf{z}}_i\})$. Without loss of generality, let the Euclidean representation $\tilde{\mathbb{Z}}$ be centered at $\mathbf{0}^d$. The stereographic projected DELAUNAYSPARSE algorithm is detailed as follows.

1. We first compute $r_{\max} = \max_{i=1, \dots, n} \|\tilde{\mathbf{z}}_i\|$, set a scaling parameter $\eta > 0$ and apply

the inverse stereographic projection,

$$\phi_{\eta r_{\max}}^{-1} : \mathcal{R}^d \rightarrow \partial \mathcal{B}_{\eta r_{\max}} \setminus \{(\mathbf{0}_d, \eta r_{\max})^\top\}, \quad \tilde{\mathbf{z}} \mapsto \left(\frac{\eta^2 r_{\max}^2}{\eta^2 r_{\max}^2 + \|\tilde{\mathbf{z}}\|^2} \cdot \tilde{\mathbf{z}}, \frac{\eta r_{\max} \|\tilde{\mathbf{z}}\|^2}{\eta^2 r_{\max}^2 + \|\tilde{\mathbf{z}}\|^2} \right)^\top, \quad (6)$$

to project $\tilde{\mathbb{Z}}$ to a d -sphere $\partial \mathcal{B}_{\eta r_{\max}} = \{\mathbf{z} \in \mathcal{R}^{d+1}; \|\mathbf{z} - (\mathbf{0}_d, \eta r_{\max}/2)^\top\| = \eta r_{\max}/2\}$.

Let $\check{\mathbf{z}}_i = \phi_{\eta r_{\max}}^{-1}(\tilde{\mathbf{z}}_i)$, for $i = 1, \dots, n$, and let the reference point $\check{\mathbf{z}}_0$ be $(\mathbf{0}_d, \eta r_{\max})$.

We define the augmented point set as $\check{\mathbb{Z}} = \{\check{\mathbf{z}}_0, \check{\mathbf{z}}_1, \dots, \check{\mathbf{z}}_n\} \subset \partial \mathcal{B}_{\eta r_{\max}}$. A graphical illustration of obtaining $\check{\mathbb{Z}}$ from $\tilde{\mathbb{Z}}$ with $d = 2$ is exhibited in Figure 5 (a).

2. As shown in Figure 5 (b), for any $i = 1, \dots, n$, the union of simplices in the Delaunay triangulation $\mathcal{DT}_{\partial \mathcal{B}_{\eta r_{\max}}}(\check{\mathbb{Z}} \setminus \{\check{\mathbf{z}}_i\})$ covers the whole sphere $\partial \mathcal{B}_{\eta r_{\max}}$. This is due to the augmentation of $\check{\mathbf{z}}_0 = (\mathbf{0}_d, \eta r_{\max})$, which is the image of points at infinity on $\mathcal{R}^d$ under the inverse stereographic projection $\phi_{\eta r_{\max}}^{-1}$. By projecting $\mathcal{DT}_{\partial \mathcal{B}_{\eta r_{\max}}}(\check{\mathbb{Z}} \setminus \{\check{\mathbf{z}}_i\})$ back to $\mathcal{R}^d$ under the convention that the geodesic path $\ell_{\partial \mathcal{B}_{\eta r_{\max}}}(\check{\mathbf{z}}_j, \check{\mathbf{z}}_0)$ (for any $j \neq i$) is mapped to the line in $\mathcal{R}^d$ from $\tilde{\mathbf{z}}_j$ in the opposite direction of $\mathbf{0}_d$, an approximation of $\mathcal{DT}_{\mathcal{R}^d}(\tilde{\mathbb{Z}} \setminus \{\tilde{\mathbf{z}}_i\})$ is obtained as shown in Figure 5 (c). Figure 5 (d) shows that such an approximation converges to $\mathcal{DT}_{\mathcal{R}^d}(\tilde{\mathbb{Z}} \setminus \{\tilde{\mathbf{z}}_i\})$ as the scaling parameter η tends to infinity.

3. Because $\mathcal{DT}_{\partial \mathcal{B}_{\eta r_{\max}}}(\check{\mathbb{Z}} \setminus \{\check{\mathbf{z}}_i\})$ covers $\partial \mathcal{B}_{\eta r_{\max}}$, for any $i = 1, \dots, n$, there must exist a Delaunay simplex $\mathcal{S}^* \in \mathcal{DT}_{\partial \mathcal{B}_{\eta r_{\max}}}(\check{\mathbb{Z}} \setminus \{\check{\mathbf{z}}_i\})$ that contains $\check{\mathbf{z}}_i$ on $\partial \mathcal{B}_{\eta r_{\max}}$. Thus, we use the breadth-first search procedure similar to the DELAUNAYSPARSE algorithm (Chang et al., 2020) to compute the Delaunay simplex $\mathcal{S}^*$. Finally, if $\check{\mathbf{z}}_0 \notin \mathbb{V}(\mathcal{S}^*)$, $\mathcal{S}(\tilde{\mathbf{z}}_i; \tilde{\mathbb{Z}} \setminus \{\tilde{\mathbf{z}}_i\})$ on $\mathcal{R}^d$ is obtained by the simplex formed by $\{\tilde{\mathbf{z}}_j : \check{\mathbf{z}}_j \in \mathbb{V}(\mathcal{S}^*)\}$; if $\check{\mathbf{z}}_0 \in \mathbb{V}(\mathcal{S}^*)$, $\mathcal{S}(\tilde{\mathbf{z}}_i; \tilde{\mathbb{Z}} \setminus \{\tilde{\mathbf{z}}_i\})$ is obtained by the simplex formed by $\{\tilde{\mathbf{z}}_j : \check{\mathbf{z}}_j \in \mathbb{V}(\mathcal{S}^\dagger)\}$, where $\mathcal{S}^\dagger$ is the neighbor simplex of $\mathcal{S}^*$ opposite to $\check{\mathbf{z}}_0$.

The complete pseudocode and other information of the stereographic projected DELAUNAYSPARSE algorithm are included in Section S4 of the supplementary material. We only need to generate the local (but not full) Delaunay triangulation to find the Delaunay simplex $\mathcal{S}(\tilde{\mathbf{z}}_i; \tilde{\mathbb{Z}} \setminus \{\tilde{\mathbf{z}}_i\})$, and thus the computation is not intensive. Specifically, Section S4 of the supplementary material shows that the computational complexity of the stereographic projected DELAUNAYSPARSE algorithm is $O(n^{2+1/d}d^2)$. As the computational complexity of solving (5) via Gaussian elimination is $O(nd^3)$ and the sample size n is usually much larger than the intrinsic dimension d , we conclude that the overall computational complexity for computing the approximation $\mathbf{\Gamma}_{\tilde{\mathbb{Z}}}$ is $O(n^{2+1/d}d^2)$.

The use of the inverse stereographic projection only has a small impact on the final Delaunay matrix $\mathbf{\Gamma}_{\tilde{\mathbb{Z}}}$ because of its local shape-preserving property. Specifically, the inverse stereographic projection preserves the angle between the logarithmic maps and rescales the length of the logarithmic maps by a locally constant factor. This rescaling factor converges to 1 as $\eta \rightarrow \infty$. As a result, any shape in the neighborhood of $\tilde{\mathbf{z}}_i$ would be preserved with tiny distortion under the inverse stereographic projection $\phi_{\eta r_{\max}}^{-1}$ when η is large, including $\mathcal{S}(\tilde{\mathbf{z}}_i; \tilde{\mathbb{Z}} \setminus \{\tilde{\mathbf{z}}_i\})$. Thus, the probability of returning an incorrect simplex is tiny (like in Figure 5 (d)). As long as the simplex $\mathcal{S}(\tilde{\mathbf{z}}_i; \tilde{\mathbb{Z}} \setminus \{\tilde{\mathbf{z}}_i\})$ is correctly found, the Delaunay weight can be correctly calculated using both $\tilde{\mathbf{z}}_i$ and $\mathcal{S}(\tilde{\mathbf{z}}_i; \tilde{\mathbb{Z}} \setminus \{\tilde{\mathbf{z}}_i\})$ on $\mathcal{R}^d$. To empirically evaluate the impact of the stereographic projection parameter η on the final Delaunay weight matrix, we perform an experiment as detailed in Section S4 of the supplementary material. It is shown that the Delaunay weight matrix $\mathbf{\Gamma}_{\tilde{\mathbb{Z}}, \eta}$ only differs due to unavoidable floating-point errors of the computer when $\eta \geq 15$. Thus, we suggest using $\eta = 15$ in practice.

3.4 Permutation Test

The Delaunay weighted test statistic in (4), $T_{\text{DW}} = \sum_{j=1}^n \sum_{i \neq j} \gamma_{ij} \times I(\delta_i = \delta_j)/n$, aims to detect differences between the underlying distributions F and G via the deviation of T_{DW} from its expectation under the null. Figure 3 gives an illustrative example of the Delaunay weight matrix $\Gamma_{\mathbb{Z}}$ with group indicators $\delta_i = 1$, for $i = 1, \dots, 4$, and $\delta_i = 0$ otherwise. If we partition $\Gamma_{\mathbb{Z}}$ at the 4-th row and column, T_{DW} equals the average of all Delaunay weights in the two diagonal blocks of $\Gamma_{\mathbb{Z}}$. Generally, off-diagonal elements in $\Gamma_{\mathbb{Z}}$ are identically distributed under the null. In contrast, under $H_1 : F \neq G$, more $\mathbf{z}_i$'s in $\mathbb{X}$ are located in the region where the density of F is higher than that of G . For these $\mathbf{z}_i$'s in $\mathbb{X}$, the expected proportion of data points from $\mathbb{X}$ in $\mathbb{V}(\mathcal{S}(\mathbf{z}_i; \mathbb{Z} \setminus \{\mathbf{z}_i\}))$ is larger than that in $\mathbb{Z} \setminus \{\mathbf{z}_i\}$. Similar arguments also hold for $\mathbb{Y}$. Thus, T_{DW} is stochastically larger under the alternative than under the null.

The exact distribution of T_{DW} under the null is complex and difficult to derive, and thus we implement a permutation procedure to compute the p -value as follows.

1. Compute the observed value of the test statistic T_{DW} via (4).
2. For $b = 1, \dots, B$, we randomly generate a permutation of group indicators $\boldsymbol{\delta}^{(b)} = \{\delta_1^{(b)}, \dots, \delta_n^{(b)}\}$, where $\sum_{i=1}^n \delta_i^{(b)} = n_1$ and $\sum_{i=1}^n (1 - \delta_i^{(b)}) = n_0$, and compute $T_{\text{DW}}^{(b)} = \sum_{j=1}^n \sum_{i \neq j} \gamma_{ij} \times I(\delta_i^{(b)} = \delta_j^{(b)})/n$ under $\boldsymbol{\delta}^{(b)}$.
3. Compute $p\text{-value} = \{\sum_{b=1}^B I(T_{\text{DW}} \leq T_{\text{DW}}^{(b)}) + 1\}/(B + 1)$, where 1 is added in both the numerator and denominator to calibrate that under H_0 , $\Pr(p\text{-value} \leq \alpha) \leq \alpha$ for all $\alpha \in [0, 1]$. Indeed, under H_0 , $T_{\text{DW}}, T_{\text{DW}}^{(1)}, \dots, T_{\text{DW}}^{(B)}$ are i.i.d. given the pooled sample $\mathbb{Z}$ and sample sizes (n_1, n_0) . Therefore, the p -value is uniformly distributed on $\{1/(B + 1), 2/(B + 1), \dots, 1\}$, which converges to $\text{Unif}[0, 1]$ as $B \rightarrow \infty$.

Under the unknown manifold setting where $\Gamma_{\mathbb{Z}}$ cannot be computed exactly, we replace it by $\Gamma_{\tilde{\mathbb{Z}}}$ for approximate inference. Unlike most of the existing tests based on pairwise distances only, the Delaunay weight takes the relative directions among data points into account, and thus our test gains more power in detecting the direction difference between F and G .

4 Theoretical Analysis

For convenience, we write $\check{\gamma}_{ij} = \gamma_{ij} + \gamma_{ji}$, for all $i \neq j$, and we use the subscript “ H_k ” in E_{H_k} and Var_{H_k} to denote the condition that H_k is true, for $k = 0, 1$. To derive asymptotic properties of the test statistic T_{DW} , it is common to deduce the distribution of the unconditional standardized statistic. However, since T_{DW} is a function of both δ and the Delaunay weight matrix $\Gamma_{\mathbb{Z}}$, it is difficult to obtain an explicit expression of its variance under the null due to the discrete nature of the Delaunay triangulation. Therefore, we first provide the expectation and variance of the test statistic conditional on $\mathbb{Z}$, where the proximity graph based on $\Gamma_{\mathbb{Z}}$ becomes nonrandom as in [Chen and Friedman \(2017\)](#) and [Chen et al. \(2018\)](#).

Theorem 1. *Let $\mathbb{Z} = \{\mathbf{z}_1, \dots, \mathbf{z}_n\}$ be a set of generic points on a d -dimensional geodesically convex Riemannian manifold $\mathcal{M} \subset \mathcal{R}^D$. When H_0 is true, we have*

$$E_{H_0}(T_{\text{DW}}|\mathbb{Z}) = \frac{n_1(n_1 - 1) + n_0(n_0 - 1)}{n(n - 1)}$$

and

$$\text{Var}_{H_0}(T_{\text{DW}}|\mathbb{Z}) = \frac{n_1(n_1 - 1)n_0(n_0 - 1)}{3n^2}V_{0,\mathbb{Z}} + \frac{n_1n_0(n - 2)}{3n^2}V_{1,\mathbb{Z}} + \frac{n_1n_0}{n^2}V_{2,\mathbb{Z}} - \frac{4n_1^2n_0^2}{(n - 1)^2n^2},$$

where

$$\begin{aligned} V_{0,\mathbb{Z}} &= \frac{\sum_{1 \leq i < j < k < l \leq n} (\check{\gamma}_{ij}\check{\gamma}_{kl} + \check{\gamma}_{ik}\check{\gamma}_{jl} + \check{\gamma}_{il}\check{\gamma}_{jk})}{n(n-1)(n-2)(n-3)/24}, \\ V_{1,\mathbb{Z}} &= \frac{\sum_{1 \leq i < j < k \leq n} (\check{\gamma}_{ij}\check{\gamma}_{ik} + \check{\gamma}_{ij}\check{\gamma}_{jk} + \check{\gamma}_{ik}\check{\gamma}_{jk})}{n(n-1)(n-2)/6}, \\ V_{2,\mathbb{Z}} &= \frac{\sum_{1 \leq i < j \leq n} \check{\gamma}_{ij}^2}{n(n-1)/2}, \end{aligned}$$

are average values of $\{(\check{\gamma}_{ij}\check{\gamma}_{kl} + \check{\gamma}_{ik}\check{\gamma}_{jl} + \check{\gamma}_{il}\check{\gamma}_{jk}) | 1 \leq i < j < k < l \leq n\}$, $\{(\check{\gamma}_{ij}\check{\gamma}_{ik} + \check{\gamma}_{ij}\check{\gamma}_{jk} + \check{\gamma}_{ik}\check{\gamma}_{jk}) | 1 \leq i < j < k \leq n\}$ and $\{\check{\gamma}_{ij}^2 | 1 \leq i < j \leq n\}$, respectively.

The proof of Theorem 1 is based on the fact that δ is uniformly distributed on all $\binom{n}{n_1}$ realizations containing n_1 ones and n_0 zeros conditional on $\mathbb{Z}$ under H_0 , as detailed in Section S5 of the supplementary material. Given the finite-sample conditional mean $E_{H_0}(T_{\text{DW}}|\mathbb{Z})$ and conditional variance $\text{Var}_{H_0}(T_{\text{DW}}|\mathbb{Z})$, we deduce the asymptotic null distribution of the standardized statistic of T_{DW} .

Theorem 2. Assume that (i) $\lim_{n \rightarrow \infty} n_1/n = \kappa \in (0, 1)$; (ii) $F = G$ and they have Lipschitz continuous densities $f = g$ with respect to the measure on $\mathcal{M}$ induced by the Lebesgue measure on the ambient Euclidean space; and (iii) there exists $\nu < 9/4$ such that $\sum_{i=1}^n |\mathcal{N}_i|^4 = o(n^\nu)$, with $\mathcal{N}_i = \cup_{S \in \mathcal{DT}_{\mathcal{M}}(\mathbb{Z}): \mathbf{z}_i \in S} \mathbb{V}(S)$. As $n \rightarrow \infty$, we have

$$\frac{T_{\text{DW}} - E_{H_0}(T_{\text{DW}}|\mathbb{Z})}{\sqrt{\text{Var}_{H_0}(T_{\text{DW}}|\mathbb{Z})}} \xrightarrow{D} N(0, 1), \quad (7)$$

where $\xrightarrow{D}$ denotes convergence in distribution.

The proof of Theorem 2 is given in Section S6 of the supplementary material. In Theorem 2, conditions (i) and (ii) are standard and mild. To gain insight into condition (iii), we define a hub $\mathbf{z}_i$ as a point with a degree much larger than the average degree in the $\mathcal{DT}_{\mathcal{M}}(\mathbb{Z})$, which only occurs under the rare case that many data points are located on the sphere $\partial\mathcal{B}_{\mathcal{M}}(\mathbf{z}_i, r)$ for some $r > 0$. Condition (iii) regularizes the Delaunay triangulation

$\mathcal{DT}_{\mathcal{M}}(\mathbb{Z})$, which is satisfied if no hub exists in $\mathcal{DT}_{\mathcal{M}}(\mathbb{Z})$. Empirical evidence of Theorem 2 is provided in Section S8 of the supplementary material.

The consistency of the Delaunay weighted test with the permutation procedure can be shown as follows.

Theorem 3. *Under the same conditions as in Theorem 2 except that $F = G$, the Delaunay weighted test using T_{DW} is consistent against all alternatives. That is, for any significance level $\alpha \in (0, 1)$ and $F \neq G$, the Delaunay weighted test with the permutation procedure rejects the null $H_0 : F = G$ with asymptotic probability one.*

The proof of Theorem 3 is given in Section S7 of the supplementary material, for which the main idea is adopted from Henze and Penrose (1999). In particular, we show that $T_{\text{DW}} - E_{H_0}(T_{\text{DW}}|\mathbb{Z})$ converges to a positive value and $\text{Var}_{H_0}(T_{\text{DW}}|\mathbb{Z})$ converges to zero a.s., as $n \rightarrow \infty$. Although the test statistic T_{DW} is constructed based on $\mathbf{\Gamma}_{\tilde{\mathbb{Z}}}$ instead of $\mathbf{\Gamma}_{\mathbb{Z}}$ in practice, the consistency of the test is preserved as long as the mapping from $\mathbb{Z}$ to $\tilde{\mathbb{Z}}$ is one-to-one, a standard assumption in manifold learning.

5 Experiments

To illustrate the empirical performance of the Delaunay weighted test, we conduct extensive experiments on both simulated and real data. For comparison, we also implement several existing nonparametric tests, including the k -NN test (Hall and Tajvidi, 2002), the e -distance test (Székely and Rizzo, 2004), the covariance test (Cai et al., 2013), the k -MST test (Chen and Friedman, 2017), the regression test (Kim et al., 2019) and the kernel test (Gretton et al., 2012; Cheng and Xie, 2024). Among these methods, the (local) regression test with the k -NN regression and the kernel test are adaptive to the intrinsic dimension of the underlying manifold, and thus we adopt the k -NN regression with $k = n^{2/(2+d)}$ as

Table 1: Rejection proportions (%) of the Delaunay weighted test on the Euclidean data under the null and different significance levels $\alpha = 0.01, 0.05, 0.10$.

d	$n_1 = 50, n_0 = 50$			$n_1 = 100, n_0 = 50$			$n_1 = 100, n_0 = 100$		
	$\alpha = 0.01$	0.05	0.10	0.01	0.05	0.10	0.01	0.05	0.10
20	1.4	4.6	8.7	1.0	4.7	10.3	0.5	4.8	10.5
50	1.0	4.2	9.5	0.6	4.5	10.1	1.0	4.7	9.9

suggested by Kim et al. (2019) in the regression test. Following the suggestion that the bandwidth should scale with $\sqrt{D}$ and $n^{-1/(d+2\beta)}$ (Cheng and Xie, 2024), we choose the bandwidth $s \cdot n^{-1/(d+2\beta)}$ for the kernel test, where s^2 is the sum of variances of $\mathbb{Z}$ and $\beta = 1$ is the Hölder exponent. As the Delaunay weight $\mathbf{w}(\mathbf{z}_i; \mathbb{Z} \setminus \{\mathbf{z}_i\})$, for each $\mathbf{z}_i$, contains at most $d + 1$ nonzero components, we use $k = d + 1$ in the k -NN tests and k -MST test for fair comparison. The covariance test is consistent only if the covariances of F and G differ, while it is of interest to compare its performance under general cases. For the standard permutation procedure, we set the number of permutations as $B = 1000$.

5.1 Scenario 1: Euclidean data

To validate the size of the Delaunay weighted test on Euclidean data, i.e., $\mathcal{M} = \mathcal{R}^D$, we generate 1000 Gaussian datasets under the null H_0 . In each dataset, $\mathbb{X}$ and $\mathbb{Y}$ respectively contain n_1 and n_0 i.i.d. samples from the multivariate normal distribution $\text{MVN}(\mathbf{0}_d, \mathbf{I}_{d \times d})$, where $\mathbf{I}_{d \times d}$ is the $d \times d$ identity matrix with $d = D$. We consider different dimensions $d \in \{20, 50\}$, sample sizes $(n_1, n_0) \in \{(50, 50), (100, 50), (100, 100)\}$ and significance levels $\alpha = 0.01, 0.05$ and 0.10 . Table 1 shows the rejection proportions of the Delaunay weighted test under different cases. In general, the Delaunay weighted test is able to control the type I error rate at the target level.

We also investigate the power of the Delaunay weighted test in distinguishing differences between distributions F and G on 1000 Gaussian datasets. We consider both the location

Table 2: Rejection proportions (%) of different two-sample tests on the Euclidean data under the location alternative and different significance levels $\alpha = 0.01, 0.05, 0.10$, where the best performance is highlighted in boldface.

d	Tests	$n_1 = 50, n_0 = 50$			$n_1 = 100, n_0 = 50$			$n_1 = 100, n_0 = 100$		
		$\alpha = 0.01$	0.05	0.10	0.01	0.05	0.10	0.01	0.05	0.10
20	Delaunay weighted	11.9	29.5	42.9	14.2	33.9	47.3	33.7	58.2	69.0
	k -NN	6.2	21.2	33.3	7.5	27.3	42.5	18.6	45.6	61.0
	k -MST	19.2	31.9	41.8	27.2	44.4	54.7	49.7	66.6	73.7
	Kernel	27.1	52.2	65.9	40.3	66.9	79.4	71.9	87.5	92.9
	e -distance	38.3	64.6	76.2	57.8	81.5	88.7	86.5	95.9	98.2
	Covariance	0.9	5.3	10.5	0.9	5.4	13.2	1.3	7.3	13.6
	Regression	5.5	17.4	28.1	6.2	18.4	33.9	15.8	36.8	50.6
50	Delaunay weighted	17.0	37.6	50.4	22.6	44.6	57.8	40.5	65.9	77.0
	k -NN	1.9	12.4	25.2	2.5	16.9	33.3	4.3	29.7	50.3
	k -MST	5.6	13.5	21.1	33.1	54.5	65.6	63.0	81.6	86.8
	Kernel	39.0	64.9	74.7	59.5	83.7	91.4	88.1	97.0	99.2
	e -distance	43.8	67.9	78.3	66.0	87.4	93.6	90.8	98.9	99.2
	Covariance	1.1	4.9	11.2	1.2	6.2	11.5	0.8	5.4	9.7
	Regression	2.5	13.0	21.8	1.9	10.5	20.2	6.5	22.4	36.5

alternative and the direction alternative, where the former imposes the difference in location parameters (mean or median) of distributions F and G , and the latter sets the same location parameter for distributions F and G but their covariance matrices have different principal directions (directions with large variance). Under the location alternative, $\mathbb{X}$ contains n_1 independent samples $\mathbf{x}_1, \dots, \mathbf{x}_{n_1}$ from $F = \text{MVN}(\mathbf{0}_d, \mathbf{I}_{d \times d})$, while $\mathbf{y}_1, \dots, \mathbf{y}_{n_0}$ in $\mathbb{Y}$ are generated from $G = \text{MVN}(\Delta_d, \mathbf{I}_{d \times d})$, where Δ_d is an independent variable uniformly distributed on $\partial \mathcal{B}_{\mathcal{R}^d}(\mathbf{0}, 0.8)$ for $d = 20$ and $\partial \mathcal{B}_{\mathcal{R}^d}(\mathbf{0}, 1)$ for $d = 50$. Under the direction alternative, both samples $\mathbb{X}$ and $\mathbb{Y}$ are generated from zero-mean Gaussian distributions $F = \text{MVN}(\mathbf{0}_d, \Sigma_{\mathbf{x}})$ and $G = \text{MVN}(\mathbf{0}_d, \Sigma_{\mathbf{y}})$, with covariance matrices,

$$\Sigma_{\mathbf{x}} = \begin{pmatrix} 1.25^2 \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \quad \text{and} \quad \Sigma_{\mathbf{y}} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & 1.25^2 \mathbf{I} \end{pmatrix},$$

where all submatrices $\mathbf{I}$ and $\mathbf{0}$ are of dimensions $(d/2) \times (d/2)$.

Table 2 presents the power of the Delaunay weighted test in comparison with other

Table 3: Rejection proportions (%) of different two-sample tests on the Euclidean data under the direction alternative and different significance levels $\alpha = 0.01, 0.05, 0.10$, where the best performance is highlighted in boldface.

d	Tests	$n_1 = 50, n_0 = 50$			$n_1 = 100, n_0 = 50$			$n_1 = 100, n_0 = 100$		
		$\alpha = 0.01$	0.05	0.10	0.01	0.05	0.10	0.01	0.05	0.10
20	Delaunay weighted	12.1	31.2	45.7	21.1	45.0	59.6	48.8	74.0	85.6
	k -NN	1.5	7.5	13.5	1.8	8.5	16.2	3.4	11.7	21.7
	k -MST	1.8	6.7	13.0	6.4	14.2	20.7	14.6	29.4	41.0
	Kernel	3.6	13.4	23.5	6.5	19.6	30.9	11.2	34.8	50.9
	e -distance	1.2	6.2	12.1	1.6	8.9	15.0	1.9	7.8	15.6
	Covariance	4.6	18.8	33.5	18.9	46.1	61.8	25.0	61.7	79.9
	Regression	1.7	7.5	14.2	1.3	7.7	14.7	3.5	13.0	22.7
50	Delaunay weighted	15.6	34.7	49.2	31.6	56.6	73.0	61.7	82.7	91.4
	k -NN	1.3	5.5	10.6	0.7	5.7	10.7	0.6	5.2	14.8
	k -MST	3.3	8.4	12.9	1.3	4.5	9.1	1.5	6.5	12.3
	Kernel	2.1	9.8	18.3	3.8	12.9	23.9	5.9	18.8	32.4
	e -distance	0.6	5.4	11.3	2.0	6.8	11.1	1.2	6.2	10.8
	Covariance	2.7	12.9	24.5	17.9	45.5	62.0	19.1	52.2	68.2
	Regression	0.8	5.9	10.4	1.3	5.0	10.7	3.1	8.6	16.7

methods under the location alternative. For both $d = 20$ and $d = 50$, the power of the Delaunay weighted test tends to 1 as the total sample size $n_1 + n_0$ grows, demonstrating its ability to distinguish the location difference of distributions. Compared with the existing ones, our test outperforms the k -NN test and the regression test and it yields comparable power with the k -MST test, while the kernel test and the e -distance test possess the highest power. However, the power difference between these two tests and our Delaunay weighted test becomes smaller for $d = 50$. As expected, the covariance test is powerless, because it is not consistent in distinguishing the location difference.

Table 3 presents the power of the Delaunay weighted test in comparison with other methods under the direction alternative. Our method outperforms all the tests in comparison, except that the covariance test has comparable power when $d = 20$. Under the direction alternative, our test, which takes the direction information into account, is more powerful than those tests that only use the distance information. The advantage of our Delaunay weighted test over the others further amplifies when the dimension d grows to

50, because the information contained in the Euclidean distance diminishes as d grows.

5.2 Scenario 2: Manifold Data

5.2.1 Size and Power

To validate the effectiveness of the Delaunay weighted test on high-dimensional data that lie on a low-dimensional manifold, we generate synthetic datasets on a d -dimensional manifold $\mathcal{M}$ embedded in $\mathcal{R}^D$ with $d = 10$ and $D = 100$. Specifically, the d -dimensional manifold $\mathcal{M}$ is obtained via the parametric transformation

$$\varphi : \mathcal{R}^d \mapsto \mathcal{R}^D, \quad \varphi(t_1, \dots, t_d) = \mathbf{Q} \begin{pmatrix} t_1 \cdot \cos(\pi t_1) \\ t_1 \cdot \sin(\pi t_1) \cdot \text{sign}(t_1) \\ \vdots \\ t_{d/2} \cdot \cos(\pi t_{d/2}) \\ t_{d/2} \cdot \sin(\pi t_{d/2}) \text{sign}(t_{d/2}) \\ t_{d/2+1} \\ \vdots \\ t_d \end{pmatrix},$$

where $\mathbf{Q}$ is a random $D \times (3d/2)$ matrix with orthonormal columns. Under this model, the geodesic distance between $\mathbf{z}_1 = \varphi(t_{11}, \dots, t_{1d})$ and $\mathbf{z}_2 = \varphi(t_{21}, \dots, t_{2d})$ is

$$d_{\mathcal{M}}(\mathbf{z}_1, \mathbf{z}_2) = \sqrt{\sum_{i=1}^{d/2} \left\{ \int_{t_{1i}}^{t_{2i}} \sqrt{1 + \pi^2 s^2} ds \right\}^2 + \sum_{i=d/2+1}^d (t_{2i} - t_{1i})^2}.$$

To validate the size and power of the Delaunay weighted test on manifold data, we randomly generate 1000 synthetic datasets with sample sizes $(n_1, n_0) \in \{(250, 250), (500, 500)\}$.

For $i = 1, \dots, n_1$, we generate $\mathbf{x}_i = \varphi(\zeta^{-1}(u_{i,1}), \dots, \zeta^{-1}(u_{i,d/2}), u_{i,d/2+1}, \dots, u_{i,d})$ where

$$\zeta : \mathcal{R} \mapsto \mathcal{R}, \quad \zeta(s) = \int_0^s \sqrt{1 + \pi^2 u^2} du,$$

and $(u_{i,1}, \dots, u_{i,d})^\top \sim \text{MVN}(\boldsymbol{\mu}_\mathbf{x}, \boldsymbol{\Sigma}_\mathbf{x})$. For $i = 1, \dots, n_0$, we generate $\mathbf{y}_i$ in the same way except that $(u_{i,1}, \dots, u_{i,d})^\top \sim \text{MVN}(\boldsymbol{\mu}_\mathbf{y}, \boldsymbol{\Sigma}_\mathbf{y})$. Values of $\boldsymbol{\mu}_\mathbf{x}, \boldsymbol{\Sigma}_\mathbf{x}$ and $\boldsymbol{\mu}_\mathbf{y}, \boldsymbol{\Sigma}_\mathbf{y}$ under different scenarios are summarized as follows.

- Null: $\boldsymbol{\mu}_\mathbf{x} = \boldsymbol{\mu}_\mathbf{y} = \mathbf{0}_d$ and $\boldsymbol{\Sigma}_\mathbf{x} = \boldsymbol{\Sigma}_\mathbf{y} = \mathbf{I}_{d \times d}$.
- Location alternative: $\boldsymbol{\mu}_\mathbf{x} = 0.06 \times \mathbf{1}_d$, $\boldsymbol{\mu}_\mathbf{y} = -0.06 \times \mathbf{1}_d$ and $\boldsymbol{\Sigma}_\mathbf{x} = \boldsymbol{\Sigma}_\mathbf{y} = \mathbf{I}_{d \times d}$.
- Direction alternative: $\boldsymbol{\mu}_\mathbf{x} = \boldsymbol{\mu}_\mathbf{y} = \mathbf{0}_d$,

$$\boldsymbol{\Sigma}_\mathbf{x} = \mathbf{I}_5 \otimes \begin{pmatrix} 1 & 0.12 \\ 0.12 & 1 \end{pmatrix} \text{ and } \boldsymbol{\Sigma}_\mathbf{y} = \mathbf{I}_{d/2} \otimes \begin{pmatrix} 1 & -0.12 \\ -0.12 & 1 \end{pmatrix},$$

where $\otimes$ stands for the Kronecker product.

Table 4 presents rejection proportions of all tests under different scenarios. Most of the estimated $\hat{d}$ are slightly overestimated as 11 or 12. Under the null, the rejection proportions of the Delaunay weighted test are close to the target significance levels, suggesting its ability to control the type I error rate. Although the Delaunay weighted test is not the most powerful under the location alternative, its power is comparable to most of the existing tests except the kernel test, which performs the best under the location alternative. Under the direction alternative, the Delaunay weighted test has the dominating power. Such an advantage even amplifies when sample sizes grow. The covariance test, designed to detect the directional difference, performs the second best under the direction alternative. In

Table 4: Rejection proportions (%) of different two-sample tests on the manifold data under the null and different alternatives (H_1) and different significance levels $\alpha = 0.01, 0.05, 0.10$, where the best performance under H_1 is highlighted in boldface.

Tests	Null (α)			Location H_1 (α)			Direction H_1 (α)		
	0.01	0.05	0.10	0.01	0.05	0.10	0.01	0.05	0.10
$n_1 = 250, n_0 = 250$									
Delaunay weighted	0.6	5.4	10.7	9.1	23.6	34.7	14.4	34.1	46.6
Kernel	0.7	4.9	9.7	19.8	42.9	57.8	3.5	15.9	28.5
Regression	0.4	4.1	9.3	9.9	27.5	36.0	2.3	11.8	21.8
k -NN	0.9	5.6	9.4	9.9	24.3	35.4	3.7	12.0	20.9
k -MST	1.7	5.7	11.6	9.2	18.1	27.9	3.5	12.1	18.2
e -distance	0.9	4.9	10.4	14.8	26.3	40.5	1.4	6.8	12.8
Covariance	1.1	3.6	6.7	0.2	1.5	4.4	11.5	31.0	42.5
$n_1 = 500, n_0 = 500$									
Delaunay weighted	0.6	5.7	10.8	29.2	44.0	58.1	46.7	69.7	81.4
Kernel	1.9	5.2	9.2	62.6	80.3	86.8	11.1	35.0	50.5
Regression	1.9	5.9	11.1	32.7	54.1	66.8	7.7	21.0	34.7
k -NN	1.9	5.7	10.0	27.5	47.2	60.2	5.7	22.1	34.7
k -MST	1.3	6.4	10.2	20.1	38.4	49.2	12.0	27.8	36.3
e -distance	1.6	6.0	10.2	30.2	47.8	59.0	1.7	7.5	15.5
Covariance	1.3	3.5	5.9	1.0	3.7	6.2	43.4	66.9	78.2

summary, when the data lie on a manifold, our Delaunay weighted test is able to detect general distribution differences with emphasis on the difference in the principal directions.

5.2.2 Sensitivity Analysis

As described in Section 3.3, a manifold learning procedure is required in our Delaunay weighted test on manifold data, including estimating the intrinsic dimension, estimating the geodesic distance, and obtaining the low-dimensional Euclidean representation $\tilde{\mathbf{Z}}$ via the MDS. It is of interest to see how the manifold learning steps affect the performance of our Delaunay weighted test. Specifically, we generate 1000 datasets under both the null and the direction alternative as in Section 5.2.1 with sample size $n_1 = n_0 = 500$. We consider three implementations of our test for comparison.

- Known Manifold: We implement the Delaunay weighted test based on the known manifold structure. That is to say, we directly calculate the Delaunay weights using

Definition 2 and perform the permutation test.

- Fixed $\hat{d}$: We implement the Delaunay weighted test by fixing $\hat{d}$ at several values including the true intrinsic dimension.
- Estimated $\hat{d}$: We implement the Delaunay weighted test with all the manifold learning steps.

Figure 6 displays the rejection proportions of different implementations of the Delaunay weighted test under different scenarios. In general, the Delaunay weighted test manages to control the type I error rate under the null and the power does not vary much with known or unknown manifold structures, indicating the Delaunay weighted test is robust to the manifold learning steps. This is corroborated by the similar results among the cases of “Known Manifold”, “Fixed $\hat{d} = 10 = d$ ”, and “Estimated $\hat{d}$ ”. Because our goal is to test the hypothesis rather than accurately estimate the manifold structure, the Delaunay weighted test only loses power if the estimated intrinsic dimension $\hat{d}$ is lower than the truth. Indeed, as shown in the results of “Fixed $\hat{d}$ ” with different values of $\hat{d}$, we see that the power keeps increasing as $\hat{d}$ grows from 6 to 10 and does not fluctuate drastically for a small range of $\hat{d} > 10$. Recall from Section 5.2.1 that the intrinsic dimension d tends to be slightly overestimated for this example. It is as expected that the implementation “Estimated $\hat{d}$ ” shows robust performance.

5.2.3 Computational Cost

To investigate the computational cost of the Delaunay weighted test compared with existing approaches, we generate 200 datasets under the null in a way similar to Section 5.2.1 with different sample sizes $n = n_1 + n_0$ ($n_1 = n_0$), intrinsic dimensions d and ambient dimensions D . Table 5 summarizes the computational time of different steps in the Delaunay weighted

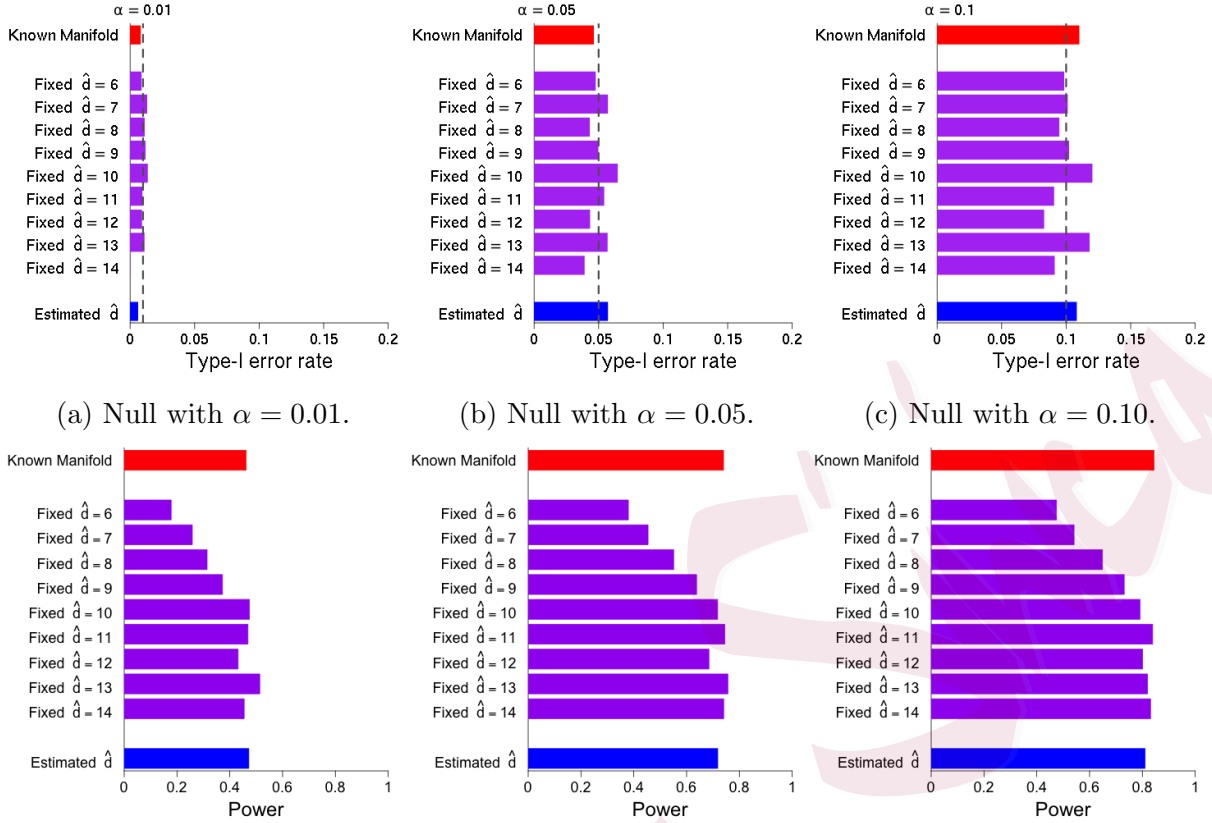


Figure 6: Rejection proportions of different implementations of the Delaunay weighted test under the null and the direction alternative as in Section 5.2.1 with sample size $n_1 = n_0 = 500$ and different significance levels (α).

test and existing approaches. In general, the computational time of the Delaunay weighted test is mainly dominated by the computational time in estimating the geodesic distance and calculating the Delaunay weight matrix. More specifically, the computational time in calculating the Delaunay weight matrix grows as the sample size and the intrinsic dimension increase, while the computational time in estimating the geodesic distance only depends on the sample size. Although the Delaunay weighted test is only more computationally efficient than the k -MST test, which requires an iterative procedure, its computational cost is still affordable.

Table 5: Mean (standard deviation) of the computational time (unit: s) of different steps in the Delaunay weighted test and existing approaches under different scenarios.

	$d = 10, D = 100$				$d = 10, n = 1000$			
	$n = 250$	$n = 500$	$n = 1000$		$D = 50$	$D = 100$	$D = 150$	$D = 200$
Algorithm S1	0.44(0.24)	1.18(0.29)	4.00(0.92)		3.51(0.81)	4.00(0.92)	4.36(0.87)	4.59(0.84)
Estimating $\tilde{d}_M(\mathbf{z}_i, \mathbf{z}_j)$	3.52(0.32)	17.92(0.67)	112.20(6.34)		110.64(5.47)	112.20(6.34)	112.27(6.61)	110.43(5.65)
MDS	0.01(0.00)	0.05(0.01)	0.30(0.04)		0.29(0.04)	0.30(0.04)	0.29(0.04)	0.28(0.01)
Calculating $\Gamma_{\tilde{z}}$	36.83(5.40)	140.27(11.66)	552.19(35.54)		550.21(40.17)	552.19(35.54)	560.09(31.87)	544.86(32.19)
Permutation test	0.39(0.04)	0.42(0.02)	0.52(0.02)		0.52(0.03)	0.52(0.02)	0.52(0.01)	0.51(0.01)
Total	41.20(5.99)	159.84(12.66)	669.20(42.86)		665.18(46.52)	669.20(42.86)	677.54(39.41)	660.67(38.70)
k -NN test	15.88(1.99)	17.31(1.12)	22.69(1.17)		22.17(1.19)	22.69(1.17)	23.45(1.00)	22.99(1.02)
k -MST test	36.80(4.30)	201.67(13.59)	1327.99(71.17)		1321.71(74.25)	1327.99(71.17)	1340.22(59.30)	1299.28(66.24)
Kernel test	0.40(0.05)	1.38(0.15)	4.79(0.48)		4.85(0.50)	4.79(0.48)	4.95(0.59)	4.77(0.50)
e-distance test	0.45(0.05)	1.74(0.16)	5.75(0.70)		5.52(0.61)	5.75(0.70)	5.84(0.73)	5.80(0.69)
Covariance test	0.19(0.28)	0.29(0.43)	0.48(0.31)		0.20(0.28)	0.48(0.31)	0.82(0.32)	1.20(0.33)
Regression test	0.51(0.02)	1.63(0.07)	5.25(0.29)		4.81(0.23)	5.25(0.29)	5.77(0.36)	6.09(0.38)

	$D = 100, n = 1000$			
	$d = 2$	$d = 4$	$d = 6$	$d = 10$
Algorithm S1	3.82(0.49)	3.80(0.57)	3.87(0.67)	4.00(0.92)
Estimating $\tilde{d}_M(\mathbf{z}_i, \mathbf{z}_j)$	110.42(3.94)	110.39(4.92)	110.16(5.08)	112.20(6.34)
MDS	0.31(0.04)	0.30(0.06)	0.27(0.02)	0.30(0.04)
Calculating $\Gamma_{\tilde{z}}$	46.25(1.42)	145.26(15.96)	296.68(17.11)	552.19(35.54)
Permutation test	0.47(0.03)	0.50(0.03)	0.51(0.01)	0.52(0.02)
Total	161.27(5.91)	260.25(21.54)	411.48(22.89)	669.2(42.86)
k -NN test	8.47(0.28)	11.40(0.59)	15.72(0.64)	22.69(1.17)
k -MST test	322.64(8.94)	544.54(38.21)	855.25(38.38)	1327.99(71.17)
Kernel test	5.10(0.57)	5.03(0.58)	4.69(0.31)	4.79(0.48)
e-distance test	5.76(1.13)	5.90(1.08)	5.67(0.94)	5.75(0.70)
Covariance test	0.44(0.30)	0.51(0.33)	0.62(1.36)	0.48(0.31)
Regression test	15.15(1.27)	7.14(0.48)	5.92(0.32)	5.25(0.29)

CPU: AMD EPYC 9754, 2.25GHz.

5.3 Real Data Analysis

As an illustration, we apply the Delaunay weighted test to the mice protein expression dataset¹ (Higuera et al., 2015). This dataset records the expression levels of 77 proteins in 1080 cells from the cortex nuclear fraction of 72 mice. These 72 mice belong to 8 groups according to their genotype (control or trisomic), behavior (stimulated to learn or not) and treatment (injected with the drug “memantine” or only saline). We remove 3 cells and 9 proteins with more than 10 missing values, resulting in a complete dataset of 1077 cells and $D = 68$ proteins.

To check whether or not the protein expression levels are likely supported on a low-dimensional manifold, we apply Algorithm S1 (Facco et al., 2017) to estimate the intrinsic dimension using all cells. This gives the estimated intrinsic dimension $\hat{d} = 3$, which is much smaller than $D = 68$. We further compute the low-dimensional Euclidean representations of all cells via the MDS, as visualized in Figure 7. The protein expression level difference is prominent between cells of different behaviors, while the difference is less obvious with respect to different genotypes and treatments. Thus, we focus on the two hypothesis tests:

- $H_{0,g}$: No difference in protein expression levels between cells of different genotypes.
- $H_{0,t}$: No difference in protein expression levels between cells of different treatments.

It is likely that protein expression levels are correlated for cells of the same mouse. Thus, a direct permutation test on all the cells may not be consistent. Instead, we perform the Delaunay weighted test with the mouse-level permutation, under the assumption that cells of different mice are independent. Due to possible correlation, bias may be introduced in computing the low-dimensional representations as well as in the Delaunay weights, which would incur power loss. However, the type I error is controlled by using the mouse-level

¹<https://archive.ics.uci.edu/dataset/342/mice+protein+expression>

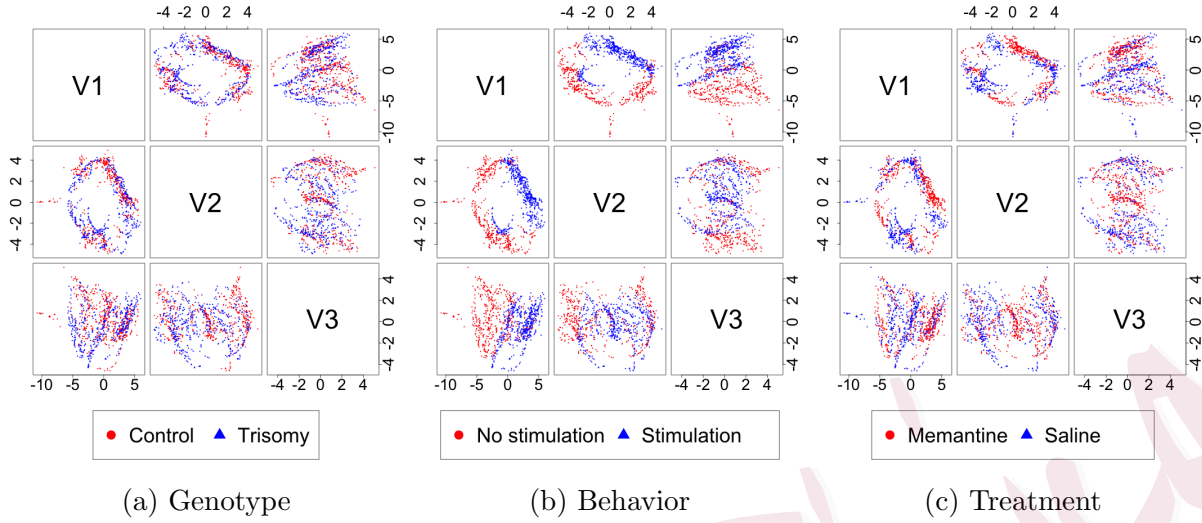


Figure 7: Low-dimensional Euclidean representations of protein expression levels of 1077 cells from the mouse cortex, where V_k denotes the k -th coordinate of the Euclidean representations, and cells are colored according to their (a) genotypes, (b) behaviors and (c) treatments.

permutation. Similarly, all existing methods under the comparison are also performed with the mouse-level permutation.

We compute p -values of our Delaunay weighted test and all other approaches in testing $H_{0,g}$ and $H_{0,t}$, as shown in Table 6. We find that all the tests except the e -distance test are significant with respect to $H_{0,g}$ under $\alpha = 0.05$. This suggests different protein expression levels between mice of different genotypes, as expected. However, in testing $H_{0,t}$, only our Delaunay weighted test yields a significant result under $\alpha = 0.05$, while all other tests fail to reject the null. This indicates that the drug “memantine” may cause changes in protein expression of mouse cortex cells, which can be detected by the tests that exploit geometric structures.

Based on the discovery by the Delaunay weighted and kernel tests that the drug “memantine” induces different protein expression, we further investigate which subgroups of mice the drug is effective on. Specifically, we perform a stratified analysis of $H_{0,t}$ over mice subgroups defined by genotypes and behaviors and report the results in Table 7. We

Table 6: The p -values of different two-sample tests in testing $H_{0,g}$ and $H_{0,t}$. Here, “***” denotes significance under $\alpha = 0.05$ and “*” denotes significance under $\alpha = 0.10$.

Tests	$H_{0,g}$	$H_{0,t}$
Delaunay weighted	0.028**	0.011**
Kernel	0.012**	0.076*
Regression	0.010**	0.776
k -NN	0.038**	0.501
k -MST	0.021**	0.880
e -distance	0.992	0.806

Table 7: The p -values of the Delaunay weighted test and the kernel test in testing $H_{0,t}$ over different subgroups. Here, “***” denotes significance under $\alpha = 0.05$.

Tests	Control & stimulated	Control & not stimulated	Trisomic & stimulated	Trisomic & not stimulated
Delaunay weighted	0.126	0.245	0.037**	0.005**
Kernel	0.252	0.171	0.135	0.002**

find that both the Delaunay weighted and kernel tests detect expression level changes of trisomic mice who are not stimulated. However, the Delaunay weighted test additionally detects a significant difference (under $\alpha = 0.05$) for the trisomic and stimulated mice. This is the main reason why the Delaunay weighted test has a smaller p -value than the kernel test does with respect to $H_{0,t}$ in Table 6.

6 Discussions

We propose a new two-sample test for high-dimensional data by incorporating geometric information on manifolds. Based on the Delaunay triangulation on manifolds, a new geometric proximity measure named the Delaunay weight is developed with a series of computational strategies for efficient calculation. The Delaunay weight distinguishes itself from others by taking the direction information into account, and thus the resulting test is more efficient in detecting the direction difference of two distributions. We define the test statistic as the average of within-group Delaunay weights, whose p -value is computed via

the permutation procedure. Theoretical derivation shows that under mild conditions, the test statistic is asymptotically normal under the null and the proposed test is consistent against general alternatives. Experiments demonstrate broad applicability and advantages of the Delaunay weighted test compared with existing ones when two distributions differ in their principal directions of covariance matrices.

There are several directions for future work. First, as the type of alternative is usually unknown in real applications and our Delaunay weighted test is particularly powerful in detecting directional differences, it is inspiring to combine our test with the kernel or e -distance test through a Cauchy combination test (Liu and Xie, 2020) for more robust performance across different types of distributional differences. Second, it is of interest to utilize the Delaunay weight matrix $\mathbf{\Gamma}_{\mathbb{Z}}$ that captures pairwise geometric information for more complicated statistical inferences other than hypothesis testing. Finally, it is interesting to generalize the Delaunay weight framework to the case of noise-contaminated data or data residing on a union of manifolds of different dimensions. Such extensions would substantially enhance the scope of our methodology, enabling its application to more complex real-world datasets.

Code Availability

The R code for reproducing all the experiments is available at <https://github.com/GuJQ5/Delaunay-weighted-test>.

Acknowledgments

We thank the Editor, the Associate Editor and two referees for their careful reviews and many insightful comments, which led to a much better exposition of our work. Tan's research was supported by the National Natural Science Foundation of China (No. 12401363 and 12471263), the Fundamental Research Funds for the Central Universities, and the Key Laboratory of Intelligent Computing and Applications (Ministry of Education). Yin's research was supported by the Patrick SC Poon endowment fund.

References

- Abadie, A. and L'Hour, J. (2021). A penalized synthetic control estimator for disaggregated data. *Journal of the American Statistical Association*, 116(536):1817–1834.
- Arias-Castro, E., Pelletier, B., and Saligrama, V. (2018). Remember the curse of dimensionality: The case of goodness-of-fit testing in arbitrary dimension. *Journal of Nonparametric Statistics*, 30(2):448–471.
- Bhattacharya, R. and Patrangenaru, V. (2003). Large sample theory of intrinsic and extrinsic sample means on manifolds. *The Annals of Statistics*, 31(1):1–29.
- Bickel, P. J. (1969). A distribution free version of the Smirnov two sample test in the p -variate case. *The Annals of Mathematical Statistics*, 40(1):1–23.
- Cai, T. T., Liu, W., and Xia, Y. (2013). Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. *Journal of the American Statistical Association*, 108(501):265–277.
- Cai, T. T., Liu, W., and Xia, Y. (2014). Two-sample test of high dimensional means under dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):349–372.
- Cao, Y., Nemirovski, A., Xie, Y., Guigues, V., and Juditsky, A. (2018). Change detection via affine and quadratic detectors. *Electronic Journal of Statistics*, 12(1):1–57.
- Chang, J., Zheng, C., Zhou, W.-X., and Zhou, W. (2017). Simulation-based hypothesis testing of high dimensional means under covariance heterogeneity. *Biometrics*, 73(4):1300–1310.
- Chang, T. H., Watson, L. T., Lux, T. C. H., Butt, A. R., Cameron, K. W., and Hong, Y. (2020). Algorithm 1012: DELAUNAYSPARSE: Interpolation via a sparse subset of the Delaunay triangulation in medium to high dimensions. *ACM Transactions on Mathematical Software*, 46(4):38:1–38:20.

- Chen, H. (2019). Sequential change-point detection based on nearest neighbors. *The Annals of Statistics*, 47(3):1381–1407.
- Chen, H., Chen, X., and Su, Y. (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 Friedman, J. H. (2017). A new graph-based two-sample test for multivariate and object data. *Journal of the American Statistical Association*, 112(517):397–409.
- Cheng, X. and Xie, Y. (2024). Kernel two-sample tests for manifold data. *Bernoulli*, 30(4):2572–2597.
- Chwialkowski, K., Strathmann, H., and Gretton, A. (2016). A kernel test of goodness of fit. In Balcan, M. F. and Weinberger, K. Q., editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2606–2615, New York, New York, USA. PMLR.
- Dorrity, M. W., Saunders, L. M., Queitsch, C., Fields, S., and Trapnell, C. (2020). Dimensionality reduction by UMAP to visualize physical and genetic interactions. *Nature Communications*, 11(1):1537.
- Facco, E., d’Errico, M., Rodriguez, A., and Laio, A. (2017). Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, 7(1):12140.
- Friedman, J. H. and Rafsky, L. C. (1979). Multivariate generalizations of the Wald–Wolfowitz and Smirnov two-sample tests. *The Annals of Statistics*, 7(4):697–717.
- Gong, S., Boddeti, V. N., and Jain, A. K. (2019). On the intrinsic dimensionality of image representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3987–3996.
- Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773.
- Gretton, A., Fukumizu, K., Harchaoui, Z., and Sriperumbudur, B. K. (2009). A fast, consistent kernel two-sample test. In *Advances in Neural Information Processing Systems*, volume 22, pages 673–681, Red Hook, New York, USA. Curran Associates, Inc.
- Gu, J. and Yin, G. (2021). Crystallization learning with the Delaunay triangulation. In Meilă, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 3854–3863. PMLR.
- Hall, P. and Tajvidi, N. (2002). Permutation tests for equality of distributions in high-dimensional settings. *Biometrika*, 89(2):359–374.
- Hediger, S., Michel, L., and Näf, J. (2022). On the use of random forest for two-sample testing. *Computational Statistics & Data Analysis*, 170:107435.

- Henze, N. (1988). A multivariate two-sample test based on the number of nearest neighbor type coincidences. *The Annals of Statistics*, 16(2):772–783.
- Henze, N. and Penrose, M. D. (1999). On the multivariate runs test. *The Annals of Statistics*, 27(1):290–298.
- Higuera, C., Gardiner, K. J., and Cios, K. J. (2015). Self-organizing feature maps identify proteins critical to learning in a mouse model of Down syndrome. *PLOS ONE*, 10(6):e0129126.
- Jurečková, J. and Kalina, J. (2012). Nonparametric multivariate rank tests and their unbiasedness. *Bernoulli*, 18(1):229–251.
- Kim, I., Lee, A. B., and Lei, J. (2019). Global and local two-sample tests via regression. *Electronic Journal of Statistics*, 13(2):5253–5305.
- Kolmogorov, A. N. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale dell’Istituto Italiano degli Attuari*, 4:83–91.
- Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27.
- Leibon, G. and Letscher, D. (2000). Delaunay triangulations and Voronoi diagrams for Riemannian manifolds. In *Proceedings of the sixteenth annual symposium on Computational geometry*, pages 341–349, New York, New York, USA. Association for Computing Machinery.
- Liu, W., Yu, X., Zhong, W., and Li, R. (2024). Projection test for mean vector in high dimensions. *Journal of the American Statistical Association*, 119(545):744–756.
- Liu, Y. and Xie, J. (2020). Cauchy combination test: A powerful test with analytic p -value calculation under arbitrary dependency structures. *Journal of the American Statistical Association*, 115(529):393–402.
- Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60.
- Marozzi, M. (2015). Multivariate multidistance tests for high-dimensional low sample size case-control studies. *Statistics in Medicine*, 34(9):1511–1526.
- Meilă, M. and Zhang, H. (2024). Manifold learning: What, how, and why. *Annual Review of Statistics and Its Application*, 11(1):393–417.
- Moon, K. R., Stanley, III, J. S., Burkhardt, D., van Dijk, D., Wolf, G., and Krishnaswamy, S. (2018). Manifold learning-based methods for analyzing single-cell RNA-sequencing data. *Current Opinion in Systems Biology*, 7:36–46.
- Morgan, K. L. and Rubin, D. B. (2012). Rerandomization to improve covariate balance in experiments. *The Annals of Statistics*, 40(2):1263–1282.

- Pless, R. and Souvenir, R. (2009). A survey of manifold learning for images. *IPSN Transactions on Computer Vision and Applications*, 1:83–94.
- 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.
- Schilling, M. F. (1986). Multivariate two-sample tests based on nearest neighbors. *Journal of the American Statistical Association*, 81(395):799–806.
- Sibson, R. (1978). Locally equiangular triangulations. *The Computer Journal*, 21(3):243–245.
- Smirnov, N. (1948). Table for estimating the goodness of fit of empirical distributions. *The Annals of Mathematical Statistics*, 19(2):279–281.
- Székely, G. J. and Rizzo, M. L. (2004). Testing for equal distributions in high dimension. *InterStat*, (5).
- Tenenbaum, J. B., de Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323.
- Terradot, L., Durnell, N., Li, M., Li, M., Ory, J., Labigne, A., Legrain, P., Colland, F., and Waksman, G. (2004). Biochemical characterization of protein complexes from the *Helicobacter pylori* protein interaction map: Strategies for complex formation and evidence for novel interactions within type IV secretion systems. *Molecular & Cellular Proteomics*, 3(8):809–819.
- Wald, A. and Wolfowitz, J. (1940). On a test whether two samples are from the same population. *The Annals of Mathematical Statistics*, 11(2):147–162.
- Wang, W., Lin, N., and Tang, X. (2019). Robust two-sample test of high-dimensional mean vectors under dependence. *Journal of Multivariate Analysis*, 169:312–329.
- Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.
- Yan, J. and Zhang, X. (2023). Kernel two-sample tests in high dimensions: Interplay between moment discrepancy and dimension-and-sample orders. *Biometrika*, 110(2):411–430.
- Yang, S., Zheng, S., and Li, R. (2024). A new test for high-dimensional two-sample mean problems with consideration of correlation structure. *The Annals of Statistics*, 52(5):2217–2240.
- Zhao, D., Wang, J., Lin, H., Chu, Y., Wang, Y., Zhang, Y., and Yang, Z. (2021). Sentence representation with manifold learning for biomedical texts. *Knowledge-Based Systems*, 218:106869.