Bernoulli **30**(4), 2024, 2572–2597
https://doi.org/10.3150/23-BEJ1685

Kernel two-sample tests for manifold data

XIUYUAN CHENG[1,a] and YAO XIE[2,b]

[1]*Department of Mathematics, Duke University, Durham, NC,* [a]*xiuyuan.cheng@duke.edu*
[2]*H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA,* [b]*yao.xie@isye.gatech.edu*

We present a study of a kernel-based two-sample test statistic related to the Maximum Mean Discrepancy (MMD) in the manifold data setting, assuming that high-dimensional observations are close to a low-dimensional manifold. We characterize the test level and power in relation to the kernel bandwidth, the number of samples, and the intrinsic dimensionality of the manifold. Specifically, when data densities p and q are supported on a d-dimensional sub-manifold $\mathcal{M}$ embedded in an m-dimensional space and are Hölder with order β (up to 2) on $\mathcal{M}$, we prove a guarantee of the test power for finite sample size n that exceeds a threshold depending on d, β, and Δ_2 the squared L^2-divergence between p and q on the manifold, and with a properly chosen kernel bandwidth γ. For small density departures, we show that with large n they can be detected by the kernel test when Δ_2 is greater than $n^{-2\beta/(d+4\beta)}$ up to a certain constant and γ scales as $n^{-1/(d+4\beta)}$. The analysis extends to cases where the manifold has a boundary and the data samples contain high-dimensional additive noise. Our results indicate that the kernel two-sample test has no curse-of-dimensionality when the data lie on or near a low-dimensional manifold. We validate our theory and the properties of the kernel test for manifold data through a series of numerical experiments.

Keywords: Kernel methods; manifold data; Maximum Mean Discrepancy; two-sample test

1. Introduction

Two-sample testing aims to determine whether two sets of samples are drawn from the same distribution. In the classical setting, given two independent sets of data in $\mathbb{R}^m$,

$$x_i \sim p,\ i = 1, \cdots, n_X, \text{ i.i.d.}, \quad y_j \sim q,\ j = 1, \cdots, n_Y, \text{ i.i.d.}, \tag{1}$$

the two-sample problem seeks to accept or reject the null hypothesis $H_0 : p = q$. Here, we assume the data follow distributions with densities p and q, respectively. It is also of practical interest to identify where $p \neq q$ when the two distributions differ. The problem is fundamental in statistics and signal processing with broad applications in scientific discovery and machine learning. Exemplar applications include anomaly detection (Bhuyan, Bhattacharyya and Kalita, 2013, Chandola, Banerjee and Kumar, 2009, 2010), change-point detection (Cao et al., 2018, Xie and Siegmund, 2013, Xie and Xie, 2021), differential analysis of single-cell data (Zhao et al., 2021), model criticism (Bińkowski et al., 2018, Chwialkowski, Strathmann and Gretton, 2016, Lloyd and Ghahramani, 2015), general data analysis of biomedical data, audio and imaging data (Borgwardt et al., 2006, Cheng, Cloninger and Coifman, 2020, Chwialkowski et al., 2015, Jitkrittum et al., 2016), and machine learning applications (Chwialkowski, Strathmann and Gretton, 2016, Jitkrittum et al., 2017, Li et al., 2017, Lloyd and Ghahramani, 2015, Lopez-Paz and Oquab, 2017, Sutherland et al., 2017).

As an example of application in machine learning, suppose we are interested in performing an out-of-distribution (OOD) test (Ren et al., 2019) to determine whether or not the new incoming testing batch of data samples follows the same distribution as the training samples. If the distribution is significantly different, re-training the model to adapt to the new data distribution may be required, or the batch will be labeled as OOD. In performing such a task, we are to compare the two sets of samples from training

and the new arrival batch and determine whether (and how) their distributions differ. When data have low-dimensional structures, it is important to consider the data geometry in the OOD test.

In many applications, high-dimensional real data have intrinsically low-dimensional structures such as manifolds. For example, it is known that patches of natural images lie on sub-manifolds in the pixel space (Buades, Coll and Morel, 2005, Peyré, 2009), and so do image features extracted by deep neural networks (Sandler et al., 2018, Zhu et al., 2018). Another example is the single-cell RNA sequencing data where measurements lie near to curve-like structures due to the time development of cells, known as the "cell trajectory" (Saelens et al., 2019, Van den Berge et al., 2020). For natural images, a simple dataset is the MNIST hand-written digits (illustrated in Example 3.1 and Figure 2), which is one of the most commonly used datasets in statistics and machine learning research. Although the original MNIST data is not exactly on the manifold, they can be viewed as having approximately manifold-like structures. In Example 3.1, we provide a case where high dimensional image data lie exactly on a smooth manifold by simulating rotated copies of the same digit image for illustrative purposes. In this work, we consider the manifold data setting where distributions p and q are supported on (or near to) a d-dimensional manifold $\mathcal{M}$ embedded in $\mathbb{R}^m$, with $d \leq m$. We refer to $\mathbb{R}^m$ as the ambient space and d as the intrinsic dimensionality of the manifold data.

Traditional statistical methods for two-sample testing have focused on parametric or low-dimensional testing scenarios, such as Hotelling's two-sample test (Hotelling, 1931) and Student's t-test (Pfanzagl and Sheynin, 1996). When it is challenging to specify the exact parametric form of the distributions, non-parametric two-sample tests are more suitable. Earlier works on one-dimensional non-parametric two-sample tests are based on the Kolmogorov-Smirnov distance (Massey, 1951, Pratt and Gibbons, 1981), the total variation distance (Györfi and van der Meulen, 1991), among others. Extending these tests to high-dimensional data is non-trivial.

Modern non-parametric tests for high-dimensional data have been developed, many based on integral probability metrics (Sriperumbudur et al., 2012). A notable contribution is the Reproducing Kernel Hilbert Space (RKHS) kernel Maximum Mean Discrepancy (MMD) two-sample test (Gretton et al. 2009, 2012), which is related to U-statistics (Serfling, 2009). The asymptotic optimality of kernel MMD tests was recently studied in Balasubramanian, Li and Yuan (2021), Li and Yuan (2019). Wasserstein distance two-sample tests have been considered in del Barrio et al. (1999), Ramdas, García Trillos and Cuturi (2017), and graph-based statistics have been proposed for distribution-free tests in high dimensions (Bhattacharya, 2020, Chen and Friedman, 2017).

However, it is known that non-parametric two-sample tests face difficulties with high-dimensional data. For instance, Ramdas et al. (2015) provided a negative result for kernel MMD in high dimension that the test power decreases may decrease polynomially with increasing data dimension when applied to detect the mean shift of Gaussian distributions. However, the argument therein does not consider possible intrinsically low-dimensional structures of high-dimensional data. Furthermore, we also observe that the roles of the kernel bandwidth and the data dimensionality were not explicitly specified in the original kernel MMD test paper Gretton et al. (2012), both of which may play a crucial role in determining the performance of the kernel test in practice.

In this paper, we aim to answer the following fundamental questions about kernel tests applied to high-dimensional data with intrinsically low-dimensional structure:

Question 1. Will a decrease in test power be observed as data dimension increases when the data has intrinsic low-dimensionality such as lying on sub-manifolds?

Question 2. When using kernel tests on manifold data, how should one select the kernel bandwidth, given that it often significantly impacts the performance of kernel methods?

We provide a positive answer to Question 1 by providing a non-asymptotic result. Theoretically, we show that when data densities are supported on a d-dimensional sub-manifold $\mathcal{M}$ embedded in $\mathbb{R}^m$ (clean manifold data with no noise), the kernel two-sample test achieves a positive test power (at the specified test level) when the number of samples n exceeds a certain threshold depending on the manifold dimension d, the squared L^2-divergence $\Delta_2(p,q)$ between the two distributions on $\mathcal{M}$, the Hölder regularity β of densities defined with respect to the intrinsic manifold distance, among other intrinsically defined quantities and with a properly chosen kernel bandwidth γ (Theorem 3.4). This finite-sample result gives that, with large n, a small departure of q from p can be detected by the kernel test when Δ_2 exceeds $n^{-2\beta/(d+4\beta)}$ up to a certain constant (Corollary 3.5). In addition, to achieve test consistency under this regime, the kernel bandwidth γ needs to scale as $n^{-1/(d+4\beta)}$. This provides a theoretical answer to Question 2 for detecting a possibly small density departure given finite samples.

The above result holds for densities p and q in the Hölder class $\mathcal{H}^\beta(\mathcal{M})$, $0 < \beta \leq 2$. When higher order regularity of p and q presents, it no longer improves the theoretical rate (see Remark 3.3). Our finite-sample analysis shows that the properties of the kernel test are only affected by the intrinsic dimensionality d rather than the ambient dimensionality m. In our result, the definitions of the quantities d, Δ_2 and β are all intrinsic to the manifold geometry (see more in Section 2.3), while any characterization through kernel spectrum would be non-intrinsic at finite kernel bandwidth.

Our result indicates that kernel tests can avoid the curse of dimensionality for manifold data, which is consistent with a similar result for kernel density estimation in Ozakin and Gray (2009). When the kernel is positive semi-definite (PSD), the kernel test we study equals the RKHS kernel MMD statistic Gretton et al. (2012). However, our analysis also covers non-PSD kernels, where the technical requirement for the kernel function is regularity, decay, and positivity, as stated in Assumption 3. Our theory suggests that a larger class of kernel tests that is MMD-like but more general than MMD can have test power. This opens the possibility of constructing more general kernels for testing problems in practice. In Section 5.3, we provide experimental evidence demonstrating the testing power with non-PSD kernels.

Our result can also be connected to two-sample tests for Functional Data Analysis (Horváth and Kokoszka, 2012) where data samples are (discretized) functions. In fact, our Example 3.1 of image data lying on a manifold also happens to be a case of vector data having underlying functional limits (the image dimensionality increases with finer resolution). It was shown in Wynne and Duncan (2022) that when the kernel bandwidth is properly scaled, kernel MMD tests for functional data can retain power on high dimensional data by converging to a limiting kernel test over functions. This leads to the same positive answer of kernel tests in high dimension with our result but is from a different perspective. The underlying functional limit can be interpreted as effectively a low dimensionality of the data and a special case of data lying on (hidden) manifolds. The Riemannian manifold data considered in this work is a more general framework for the intrinsic low-dimensionality of vector data (for cases beyond those like Example 3.1), and the functional data setting extends to broader cases of non-vector data, e.g., functions evaluated on un-shared meshes. Notably, our result also indicates that a proper choice of kernel bandwidth is important for testing performance, where the optimal choice is not always the median distance heuristic.

We begin by proving the consistency of the kernel test when the data densities lie on a smooth manifold without a boundary. We then extend the theory to submanifolds with a smooth boundary. The manifold with boundary setting includes, as a special case, the Euclidean data case, where p and q are supported on a compact domain in $\mathbb{R}^m$ with a smooth boundary, and $d = m$. The theory also extends to the case where manifold data are corrupted by additive Gaussian noise in the ambient space $\mathbb{R}^m$. We show, theoretically, that as long as the coordinate-wise Gaussian noise level σ is less than $\gamma/\sqrt{m}$ up to an absolute constant (γ being the kernel bandwidth parameter), the kernel tests computed from noisy data have the same theoretical consistency rate as clean data lying on the manifold. In this case, the test consistency is determined only by the pair of two densities of the clean manifold data.

Our experiments demonstrate that the test power can be maintained (for fixed test level) as the ambient dimensionality m increases for low-dimensional manifold data embedded in high-dimensional space. Specifically, we construct an example of group-transformed images with increasingly refined resolution (i.e., increasing image size). We also conduct experiments on noise-corrupted data. In the theoretical regime of small additive noise, the performance of kernel tests on noisy data is similar to that on clean data, as predicted by the theory. Next, we apply kernel tests to the more complicated hand-written digits data set, which no longer lies exactly on manifolds. We demonstrate that kernel bandwidth much smaller than the median distance bandwidth can provide better performance. Finally, we numerically show that non-PSD kernels that may or may not satisfy the proposed theoretical conditions can provide a kernel test with power.

Our work adopts analytical techniques from the geometrical data analysis and manifold learning literature, particularly the analysis of local kernels on manifolds from Coifman and Lafon (2006). As a quick recap of related works: seminal works such as Belkin and Niyogi (2003, 2007), Coifman and Lafon (2006), Hein, Audibert and von Luxburg (2005) have demonstrated that the graph diffusion process on a kernelized affinity graph constructed from high-dimensional data vectors converges to a continuous diffusion process on the manifold as the sample size increases to infinity and the kernel bandwidth decreases to zero. The results in Singer (2006) and subsequent works demonstrate the approximation error to the manifold diffusion operator at a finite sample size, where the sample complexity only involves the intrinsic dimensionality. Another line of related works concerns the spectral convergence of kernel matrices constructed from manifold data. Note that the kernel function itself is computed from Euclidean coordinates of data in $\mathbb{R}^m$ and thus extrinsic. Therefore, any theoretical properties involving the kernel spectrum are also non-intrinsic to the manifold. In the limit of kernel bandwidth going to zero, the spectrum of kernelized graph Laplacian matrices has been shown to converge to that of the manifold Laplacian operator (Calder and García Trillos, 2022, Cheng and Wu, 2022b, Dunson, Wu and Wu, 2021, García Trillos et al., 2020). However, bounding the difference between the extrinsic kernel spectrum to the intrinsic limiting spectrum incurs more complicated analysis under additional assumptions. Our work addresses this limit by revealing the limiting population kernel MMD-like statistic as the squared L^2 divergence up to a constant scaling factor (Lemma 3.2), which is a simpler analysis.

In the rest of the paper, the necessary preliminaries and notations are provided in Section 2. In Section 3, we present the theory for kernel tests on manifold data and establish the consistency and power of the test. We then extend this theory to cover the case of a manifold with boundary and data containing high-dimensional noise in Section 4. Numerical experiments are presented in Section 5, and we discuss potential future research directions in Section 6. All proofs are provided in Section A of Cheng and Xie (2024).

2. Preliminaries

Following the setup in (1), we define $n := n_X + n_Y$. We also assume that n_X and n_Y are proportional, that is, as n increases, n_X/n approaches a constant $\rho_X \in (0,1)$. As our non-asymptotic analysis will consider a finite n, the constant proportion will be reflected in a "balancing" condition, see (11).

2.1. Classical RKHS kernel MMD statistic

The (biased) empirical estimate for the squared kernel MMD statistic Gretton et al. (2012) is defined as

$$\widehat{T} := \frac{1}{n_X^2}\sum_{i,i'=1}^{n_X} K_\gamma(x_i, x_{i'}) + \frac{1}{n_Y^2}\sum_{j,j'=1}^{n_Y} K_\gamma(y_j, y_{j'}) - \frac{2}{n_X n_Y}\sum_{i=1}^{n_X}\sum_{j=1}^{n_Y} K_\gamma(x_i, y_j), \tag{2}$$

where $K_\gamma(x, y)$ is a PSD kernel with a user-specified *bandwidth* parameter $\gamma > 0$. The corresponding population statistic T will be given in (7) below.

We consider a kernel with a fixed bandwidth, that is,

$$K_\gamma(x, y) = h\left(\frac{\|x - y\|^2}{\gamma^2}\right), \quad h : [0, \infty) \to \mathbb{R}, \tag{3}$$

where h usually is some non-negative function. A standard example is the Gaussian radial basis function (RBF) kernel, defined by $h(r) = \exp(-r/2)$. The classical theory of kernel MMD tests requires the kernel to be characteristic, ensuring that the MMD distance is a metric between distributions (Gretton et al., 2012). However, in this paper, we relax this assumption and only require h to be a non-negative, C^1 function that decays, which does not necessarily lead to a positive semi-definite (PSD) kernel K_γ. Please refer to Assumption 3 and the subsequent comments for further discussion.

The unbiased estimator of the kernel MMD removes the diagonal entries $K(x_i, x_i)$ and $K(y_j, y_j)$ in the summation in (2), and has a slightly different normalization (by $1/(N(N-1))$ rather than $1/N^2$, where $N = n_X$ and n_Y respectively). Since diagonal entries always equal $h(0)$, which is a constant, the biased and unbiased estimators give the same behavior in our setting qualitatively. In this paper, we focus on the biased estimator (2), and the analysis can be extended to the unbiased estimator.

2.2. Test level and power

We adopt the standard statistical definitions (Gretton et al., 2012) for the test level α_{level} and testing power. In the two-sample test setting, one computes the kernel test statistic $\widehat{T}$ from datasets X and Y and chooses a threshold t_{thres}. If $\widehat{T} > t_{\text{thres}}$, the test rejects the null hypothesis H_0.

The "level" of a test, denoted by α_{level}, is the target Type-I error. A test achieves a level α_{level} if

$$\mathbb{P}[\widehat{T} > t_{\text{thres}} | H_0] \le \alpha_{\text{level}}, \tag{4}$$

where $0 < \alpha_{\text{level}} < 1$ is typically set to a small constant, such as $\alpha_{\text{level}} = 0.05$. To control the Type-I error (4), the threshold t_{thres} needs to exceed the $(1 - \alpha_{\text{level}})$-quantile of the distribution of $\widehat{T}$ under H_0. Typical asymptotic theory determines t_{thres} by the limiting distribution of the detection statistic $\widehat{T}$ under H_0, which is a χ^2 distribution in many cases. However, the distribution of $\widehat{T}$ may significantly differ from the limiting distribution at a finite sample size. In practice, t_{thres} is usually estimated using a standard bootstrap procedure (Gretton et al., 2012, Higgins, 2004).

The Type-II error of the statistic $\widehat{T}$ and the threshold t_{thres} is given by $\mathbb{P}[\widehat{T} \le t_{\text{thres}} | H_1]$ under the alternative hypothesis. The *testing power* (at level α_{level}) corresponds to one minus the Type-II error. The test is said to be *asymptotically consistent* if the testing power can approach 1 as the sample size n increases. In this work, we will characterize the testing power of the kernel test at a finite sample size.

2.3. Riemannian manifold and intrinsic geometry

The differential geometric notations employed in this paper are standard and can be found in, for example, do Carmo (1992). We consider a smooth connected manifold $\mathcal{M}$ of dimension d equipped with a Riemannian metric tensor $g_{\mathcal{M}}$. The manifold $\mathcal{M}$ is isometrically embedded in the Euclidean space $\mathbb{R}^m$, where m is the ambient dimension and can be much larger than the intrinsic dimension d. Let $\iota : \mathcal{M} \to \mathbb{R}^m$ be the C^∞ isometric embedding, and let $\iota(x) \in \mathbb{R}^m$ denote the extrinsic coordinates. In this paper, we use the same notation x to represent both a point $x \in \mathcal{M}$ and its image $\iota(x) \in \mathbb{R}^m$, provided

that there is no ambiguity. Note that different embeddings in different spaces can be associated with the same Riemannian manifold $(\mathcal{M}, g_{\mathcal{M}})$. A quantity is called *intrinsic* if it solely depends on $g_{\mathcal{M}}$ and is independent of the embedding or extrinsic coordinates.

Given the Riemannian metric $g_{\mathcal{M}}$, the geodesic distance can be defined at least locally. We assume that the geodesic distance $d_{\mathcal{M}}(x, y)$ is globally defined on $\mathcal{M}$ and induces a metric on $\mathcal{M}$. The Euclidean distance in $\mathbb{R}^m$ is denoted by $\|x - y\|$. The manifold differential operators are defined intrinsically with respect to $g_{\mathcal{M}}$. For instance, for a C^1 function f on $\mathcal{M}$, $\nabla_{\mathcal{M}} f(x)$ denotes the manifold gradient of f at point x, which consists of partial derivatives with respect to the normal coordinates. The Hölder class $\mathcal{H}^\beta(\mathcal{M})$ is defined with respect to manifold geodesic distance, and in this work, we consider $0 < \beta \le 2$. Specifically,

(i) When $\beta \le 1$,

$$\mathcal{H}^\beta(\mathcal{M}) = \{f \in C^0(\mathcal{M}), \exists L > 0, |f(x) - f(y)| \le L d_{\mathcal{M}}(x, y)^\beta, \forall x, y \in \mathcal{M}\},$$

and we define the *Hölder constant* of f as $L_f := \sup_{x \ne y \in \mathcal{M}} |f(x) - f(y)| / d_{\mathcal{M}}(x, y)^\beta$.

(ii) When $1 < \beta \le 2$,

$$\mathcal{H}^\beta(\mathcal{M}) = \{f \in C^1(\mathcal{M}), \exists L > 0, \|\nabla_{\mathcal{M}} f(x) - \nabla_{\mathcal{M}} f(y)\| \le L d_{\mathcal{M}}(x, y)^{\beta-1}, \forall x, y \in \mathcal{M}\},$$

and then we define $L_f := \|\nabla_{\mathcal{M}} f\|_\infty + \sup_{x \ne y \in \mathcal{M}} \|\nabla_{\mathcal{M}} f(x) - \nabla_{\mathcal{M}} f(y)\| / d_{\mathcal{M}}(x, y)^{\beta-1}$.

Our notion of the Hölder constant L_f removes the C^0 norm $\|f\|_\infty$ from the usual definition of the Hölder norm. When $\beta = 1$, L_f is the Lipschitz constant of f (with respect to the manifold distance).

The Riemannian geometry also induces an intrinsic measure on $\mathcal{M}$. Let dV be the volume element on $\mathcal{M}$ associated with the local Riemann volume form. Then $(\mathcal{M}, dV)$ is a measure space. For any distribution $dP(x)$ on $\mathcal{M}$, it may have a density with respect to dV, that is, $dP(x) = p(x)dV(x)$, where p is the density function. In this paper, we consider densities that are Hölder continuous with respect to the metric $d_{\mathcal{M}}$ and square-integrable on $(\mathcal{M}, dV)$. Because $d_{\mathcal{M}}$ is intrinsic, the Hölder constants are intrinsically defined. Moreover, since the measure dV is intrinsic, dV-integrals such as the squared L^2 divergence $\int_{\mathcal{M}} (p(x) - q(x))^2 dV(x)$ between two distributions with densities p and q are also intrinsically defined.

2.4. Notations

Table 1 lists the default notations used in this paper. We may use abbreviated notation to omit the variable in an integral, e.g., $\int f dV = \int f(x) dV(x)$. The notation $\wedge$ stands for the minimum of two numbers, i.e., $a \wedge b = \min\{a, b\}$. The paper considers the joint limiting process of sample size $n \to \infty$ and kernel bandwidth $\gamma \to 0$, but the main result is non-asymptotic and holds for finite sample size n which is sufficiently large.

With respect to a limiting process, e.g., $\gamma \to 0$, the default asymptotic notations are as follows: $f = O(|g|)$ means that there is constant C such that $|f| \le C|g|$ eventually (meaning that there exists γ_0 s.t. when $\gamma < \gamma_0$ then $|g| \ge C|g|$). We use $O_{\mathrm{x}}(\cdot)$ to denote big-O notation with the constant depending on object x. In this work, we consider constants that depend on the manifold $\mathcal{M}$ and kernel function h as absolute ones and mainly focus on the constant dependence on data densities p and q. We will specify the constant dependence in the text, and we will also clarify the needed largeness of n or the smallness of γ for the bounds to hold. Additionally, $f \sim g$ means that $f, g \ge 0$ and there exist constants $C_1, C_2 > 0$ such that $C_1 g \le f \le C_2 g$ eventually; $f \gtrsim g$ means that $f \ge C_1 g$ eventually for some $C_1 > 0$; and $f \gg g$ means that for $f, g > 0$, $f/g \to \infty$ in the limit.

Table 1. List of default notations

m	dimensionality of the ambient space
d	intrinsic dimensionality of the manifold
$\mathcal{M}$	d-dimensional manifold in $\mathbb{R}^m$
dV	volume form on $\mathcal{M}$
$d_{\mathcal{M}}(x,y)$	manifold geodesic distance
$\|x-y\|$	Euclidean distance in $\mathbb{R}^m$
p, q	data sampling densities on $\mathcal{M}$
n_X, n_Y	number of samples in two-sample datasets X and Y respectively
n	$n = n_X + n_Y$
ρ_X	$n_X/n \to \rho_X$
$\widehat{T}$	empirical kernel statistic (2)
T	population kernel statistic (7)

β	Hölder class $\mathcal{H}^\beta(\mathcal{M})$
L_ρ	Upper bound of Hölder constants (defined in Section 3.1) of p and q on $\mathcal{M}$
$\rho_{\max}$	Uniform upper bound of p and q on $\mathcal{M}$
γ	kernel bandwidth parameter
K_γ	kernel applied to data, $K_\gamma(x,y) = h\left(\frac{\|x-y\|^2}{\gamma^2}\right)$
h	C^1 and decay function on $[0,\infty)$, $h \geq 0$
m_0	$m_0[h] := \int_{\mathbb{R}^d} h(\|u\|^2)du$

Asymptotic Notations	
$O(\cdot)$	$f = O(g)$: there exists $C > 0$ such that when $\|g\|$ is sufficiently small, $\|f\| \leq C\|g\|$.
$O_{\mathrm{x}}(\cdot)$	declaring the constant dependence on x.

3. Theoretical properties of kernel tests on manifold data

In this section, we study the property of the kernel MMD-like statistic in (2) for manifold data. Note that the kernel statistic $\widehat{T}$ can be computed from any two datasets $\{x_i\}_{i=1}^{n_X}$ and $\{y_j\}_{j=1}^{n_Y}$ as long as the bandwidth parameter γ is specified, and there is no need to estimate the intrinsic dimension d as an input parameter. The theory in this section studies the properties of the kernel test and the theoretical choice of γ when manifold structure is present in the high dimensional data. We begin by formulating the problem, introducing the local kernel, and stating the main result regarding the test size and power.

3.1. Manifold data in high-dimensional space

We state the necessary assumptions on the manifold data and sampling densities. An example of high dimensional image data satisfying our assumption is provided in Example 3.1, see Figure 2. In this section, we consider a compact manifold without a boundary:

Assumption 1 (Data manifold). $\mathcal{M}$ is a d-dimensional compact connected C^∞ manifold isometrically embedded in $\mathbb{R}^m$ without boundary.

An illustration of when $d = 1$ and $m = 3$ is shown in Figure 1(Left). Our theory extends when the manifold has a smooth boundary, which will be discussed in Section 4.1. This section assumes that the data densities p and q are supported on $\mathcal{M}$. In Section 4.2, we will discuss the extension of our analysis to the case where the data lie near the manifold and contain a certain type of additive Gaussian noise.

We introduce the following assumption on the Hölder regularity and boundedness of the data densities p and q. Recall the definition of $\mathcal{H}^\beta(\mathcal{M})$ in Section 2.3.

Assumption 2 (Data density). Data densities p and q are in $\mathcal{H}^\beta(\mathcal{M})$, $0 < \beta \leq 2$, and the Hölder constants of p and q are bounded by L_ρ, namely $L_\rho = \max\{L_p, L_q\}$. Since Hölder continuity implies continuity, due to compactness of $\mathcal{M}$, both densities are uniformly bounded, that is, there is constant $\rho_{\max}$ such that

$$0 \leq p(x), q(x) \leq \rho_{\max}, \quad \forall x \in \mathcal{M}.$$

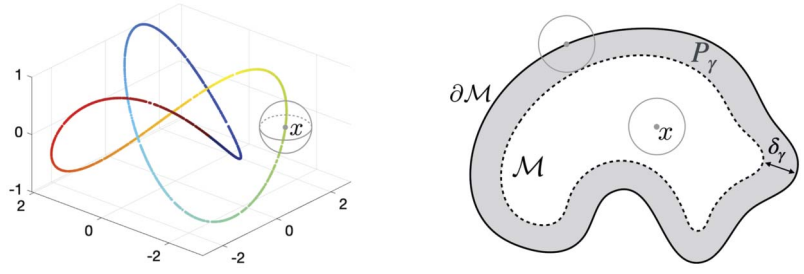

Figure 1. (Left) A one-dimensional manifold with no boundary (a closed curve) embedded in $\mathbb{R}^3$, and an Euclidean ball centered at a point x on the manifold. (Right) Illustration of a two-dimensional manifold with boundary, showing the near-boundary set P_γ (gray belt), and two Euclidean balls centered at a point away from the boundary and another point on the boundary, respectively.

To illustrate that the manifold structure naturally arises in real-world data, we provide an example of high-dimensional data lying on intrinsically low-dimensional manifolds. In this example, the change in data densities q from p is induced by the change in densities on a latent manifold independent of the ambient space $\mathbb{R}^m$.

Example 3.1 (Manifold data with increasing m). Consider data samples in the form of images I_i that have $W \times W$ pixels and thus can be represented as vectors in $\mathbb{R}^m$, where $m = W^2$. The image I_i is generated by evaluating a continuous function on an image grid given a latent variable z_i. Specifically,

$$I_i(j_1, j_2) = F\left(\left(\frac{j_1}{W}, \frac{j_2}{W}\right); z_i\right), \quad 1 \le j_1, j_2 \le W,$$

where $F(u; z_i)$ is a smooth mapping from $u \in [0,1] \times [0,1]$ to $\mathbb{R}$ that depends on a latent variable $z_i \in \mathcal{M}_z$. For instance, suppose $\mathcal{M}_z$ is a d-dimensional rotation group $SO(2)$, and the mapping $F(\cdot; z)$ corresponds to applying the rotation action $z \in SO(2)$ to the image, as illustrated in Figure 2. Under generic assumptions on F, the continuous functions $F(\cdot; z)$ for all z lie on a d-dimensional manifold in the function space. This construction defines the embedding map ι from the manifold $\mathcal{M}_z$ to $\mathbb{R}^{W \times W}$. In this example, when W increases, namely as the discretization gets finer, the image manifold in $\mathbb{R}^{W \times W}$ (up to a scalar normalization) also approaches a continuous limit determined by the latent manifold $\mathcal{M}_z$ (the rotation angle) and the mapping F on $[0,1]^2 \times \mathcal{M}_z$.

3.2. Local kernels on manifold and the population statistic

We consider local kernel $K_\gamma(x, y)$ defined as in (3) which is computed from Euclidean distances between data samples. In the term "local kernel", "local" means a small kernel bandwidth parameter γ, and typically γ decreases as the sample size increases. The following class of non-negative differential kernel function h contains K_γ being the Gaussian RBF kernel as a special case.

Assumption 3 (Differentiable kernel). We make the following assumptions about the function h, excluding the case where $h \equiv 0$:

(C1) *Regularity*. h is continuous on $[0, \infty)$, C^1 on $(0, \infty)$.

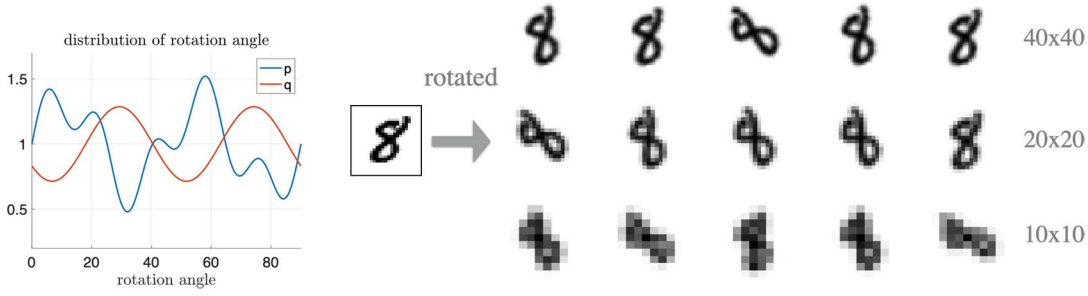

Figure 2. An example showing that the increase in the ambient dimension m does not affect the intrinsic dimensionality d nor the intrinsic geometry of the manifold data. An image of hand-written digit "8" is rotated by angles z and at different image sizes. The images for changing angle z lie on a one-dimensional manifold in the ambient space and approach a certain continuous limit as image resolution refines. The group element z has two distributions, which induce two distributions of data images in ambient space $\mathbb{R}^m$. When z changes from 0 to 2π the curve is closed and the data manifold has no boundary. When z changes from 0 to $\pi/2$ the curve has two endpoints and the data manifold has a boundary. The two-sample test results on this data are provided in Section 5.

(C2) *Decay condition*. h and h' are bounded on $(0,\infty)$ and have sub-exponential tail, specifically, $\exists a, a_k > 0$, s.t., $|h^{(k)}(\xi)| \le a_k e^{-a\xi}$ for all $\xi > 0$, $k = 0, 1$. Without loss of generality, assume that $a_0 = 1$.
(C3) *Non-negativity*. $h \ge 0$ on $[0,\infty)$.

Similar conditions on h have been used in Coifman and Lafon (2006) for kernelized graph Laplacian constructed from manifold data. For h that satisfies Assumption 3, we introduce the following moment constant of the kernel h,

$$m_0[h] := \int_{\mathbb{R}^d} h(\|u\|^2)du, \tag{5}$$

which is finite due to (C2). By (C1),(C2) and that h is not a zero function, $m_0[h] > 0$. We note that $0 \le K_\gamma(x,y) \le 1$ for any x, y, where $K_\gamma(x,y)$ is induced by h as defined in (3), due to (C2) and (C3). Note that the kernel $K_\gamma(x,y)$ is not necessarily PSD, but the theory herein remains valid in this case. (For the prototypical choice of the Gaussian RBF kernel, the kernel is indeed PSD.) The non-negativity condition (C3) may be relaxed, as it is only used to guarantee that $m_0[h] > 0$ and in the extension to the manifold with boundary in Section 4.1. We assume (C3) for simplicity.

The following lemma establishes the approximation of a Hölder function f by its kernel integral on a manifold when γ is small; this result is necessary for our subsequent analysis.

Lemma 3.1 (Kernel integral on manifold). *Suppose $\mathcal{M}$ satisfies Assumption 1, h satisfies Assumption 3, and f is in $\mathcal{H}^\beta(\mathcal{M})$, $0 < \beta \le 2$, with Hölder constant L_f. Then there is $\gamma_0 > 0$ which depends on $\mathcal{M}$ only, and constant C_1 that depends on $(\mathcal{M}, h)$, such that when $0 < \gamma < \min\{1, \gamma_0\}$, for any $x \in \mathcal{M}$,*

$$\left|\gamma^{-d}\int_{\mathcal{M}} h\left(\frac{\|x-y\|^2}{\gamma^2}\right) f(y)dV(y) - m_0[h]f(x)\right| \le C_1(L_f\gamma^\beta + \|f\|_\infty\gamma^2). \tag{6}$$

Specifically, γ_0 depends on manifold reach and curvature, and $C_1 > 0$ depends on manifold curvature and volume, the kernel function h (including the constants a, a_1 in Assumption 3(C2)), and the intrinsic dimensionality d.

In particular, if f is a constant function, then $L_f = 0$ and only the $O(\gamma^2)$ term remains in the bound in (6). The $O(\gamma^2)$ error is due to that the manifold has curvature while the local kernel function only accesses the Euclidean distance $\|x - y\|$ in the ambient space. When f is non-constant, the $O(\gamma^\beta)$ term results from the Taylor expansion (under manifold intrinsic coordinate) of f at x, and will be the leading term if $\beta < 2$. When $\beta = 2$, the bound in (6) becomes $O(\gamma^2)$, which echoes the $O(\gamma^2)$ error in Lemma 8 of Coifman and Lafon (2006) (the latter was proved for f with higher order regularity and under different technical assumptions). The constant γ_0 in Lemma 3.1 is for theoretical purposes, and, similar to other constant thresholds for the smallness of γ in later analysis, it is generally not to be computed in practice. We will clarify the choice of bandwidth γ in Remark 3.2. The proof of Lemma 3.1 follows the approach in Coifman and Lafon (2006) using standard techniques of differential geometry, and is included in Section A of Cheng and Xie (2024) for completeness.

The empirical test statistic $\widehat{T}$ is defined as in (2). Define the population kernel test statistic

$$\begin{aligned} T &:= \mathbb{E}_{x\sim p,\, y\sim p} K_\gamma(x,y) + \mathbb{E}_{x\sim q,\, y\sim q} K_\gamma(x,y) - 2\mathbb{E}_{x\sim p,\, y\sim q} K_\gamma(x,y) \\ &= \int_{\mathcal{M}} \int_{\mathcal{M}} K_\gamma(x,y)(p-q)(x)(p-q)(y) dV(x) dV(y), \end{aligned} \tag{7}$$

which equals the population (squared) kernel MMD when K_γ is PSD. Applying Lemma 3.1 gives the leading term in T as $\gamma \to 0$, as characterized in the following lemma. Define the squared L^2-divergence between p and q as

$$\Delta_2 := \int_{\mathcal{M}} (p-q)^2 dV = \Delta_2(p,q). \tag{8}$$

Lemma 3.2. *Under Assumptions 1, 2, 3, γ_0 and C_1 as in Lemma 3.1, when $0 < \gamma < \min\{1, \gamma_0\}$,*

$$\gamma^{-d} T = m_0[h]\Delta_2 + r_T, \quad |r_T| \le \tilde{C}_1 (L_\rho + \rho_{\max}) \gamma^\beta \Delta_2^{1/2}, \tag{9}$$

where $\tilde{C}_1 := 2C_1 \mathrm{Vol}(\mathcal{M})^{1/2}$ is a constant depending on $(\mathcal{M}, h)$.

We comment on the relationship between the population kernel statistic T and the L^2-divergence Δ_2 between the two densities p and q. Recall that T by definition depends on the kernel bandwidth parameter γ. By definition, if $\Delta_2 = 0$, then $p = q$ in the L^2 sense and this implies that $T = 0$ for any $\gamma > 0$; If $\Delta_2 > 0$, then Lemma 3.2 gives that

$$\gamma^{-d} T = \Delta_2^{1/2} \left(m_0[h] \Delta_2^{1/2} + O(\gamma^\beta) \right),$$

which means that the right-hand side will be strictly positive when γ is sufficiently small, and as a result, T is also strictly positive (the magnitude is up to a scaling factor of γ^{-d}). When γ is not small enough, then it is possible that T becomes zero even $\Delta_2 > 0$.

3.3. Control of the deviation of $\widehat{T}$ from mean

We now control the deviation of the empirical test statistic $\widehat{T}$ around T, where the latter equals $\mathbb{E}\widehat{T}$ up to an $O(1/n)$ bias. For the sample sizes of the two sets of samples, our analysis needs n_X and n_Y to grow proportionally to one another, namely, for some $\rho_X \in (0,1)$,

$$n = n_X + n_Y, \quad n_X/n \to \rho_X, \quad \text{as } n_X, n_Y \to \infty. \tag{10}$$

As our analysis considers sufficiently large samples, we introduce the following technical condition based on (10) (the constant 0.9 can be changed to any positive number less than 1)

$$0.9\rho_X \le \frac{n_X - 1}{n}, \quad 0.9(1-\rho_X) \le \frac{n_Y - 1}{n}, \quad 0 < \rho_X < 1. \tag{11}$$

The condition stands for the requirement of the largeness of n such that the balanced sizes of n_X and n_Y are achieved. We call (11) the *balancing condition* and assume it holds for all n. Since in our non-asymptotic result, we will derive the needed large n to guarantee the test level and power, the balancing condition (11) allows us to focus the characterization of the needed n on constants related to the manifold, the two densities, and the kernel, rather than the balancing of the two-sample sizes.

Proposition 3.3 proves a sample complexity result of the statistic $\widehat{T}$, which controls the deviation $\widehat{T} - T$ using the concentration of U-statistics. This estimation bound will be applied to control the upper tail of $\widehat{T}$ under H_0 and the lower tail of $\widehat{T}$ under H_1 respectively, and it can also be of independent interest. The U-statistic argument was used in (Gretton et al., 2012, Theorem 10) but the deviation bound therein was based on the point-wise boundedness of the kernel and the influence of kernel bandwidth was not explicit. Here we apply a Bernstein-type argument which allows to reveal the role of the bandwidth. The proof adopts the classical decoupling technique of the U-statistics (Hoeffding, 1963) and is included in Section A of Cheng and Xie (2024) for completeness.

Proposition 3.3 (Control of $|\widehat{T} - T|$). *Under Assumption 1, 2, 3, and the balancing condition* (11). *Define*

$$c := 0.9 \min\{\rho_X, 1-\rho_X\}, \quad \nu := (m_0[h^2] + 1)\rho_{\max}. \tag{12}$$

Then, there is a constant $C_1^{(2)} > 0$ depending on $(\mathcal{M}, h)$ such that when $0 < \gamma < \min\{1, \gamma_0, (C_1^{(2)})^{-1/2}\}$, for any $0 < \lambda < 3\sqrt{c\nu\gamma^d n}$, with probability $\ge 1 - 3e^{-\lambda^2/8}$,

$$\widehat{T} \le T + \frac{2}{cn} + 4\lambda\sqrt{\frac{\nu}{c}\frac{\gamma^d}{n}},$$

and with probability $\ge 1 - 3e^{-\lambda^2/8}$,

$$\widehat{T} \ge T - \frac{2}{cn} - 4\lambda\sqrt{\frac{\nu}{c}\frac{\gamma^d}{n}}.$$

The constant $C_1^{(2)}$ corresponds to the constant C_1 in Lemma 3.1 with the function h replaced by h^2.

Due to the fact that the proof of Proposition 3.3 reduces the concentration of the U-statistic to that of an $O(n)$-term independent sum, which is the same as the linear-time statistic (see Remark 3.4), an $O(n^{-1/2})$ fluctuation of the statistics $\widehat{T}$ around the mean is obtained (without considering γ in the big-O notation and up to the $O(n^{-1})$ bias). It is worth noting that, under H_0, the deviation is expected to scale as $O(n^{-1})$ (Cheng, Cloninger and Coifman, 2020, Gretton et al., 2012). In Section 6, we will discuss the possible influence on the asymptotic rate for detecting $q \neq p$. In practice, the testing threshold is usually estimated empirically using bootstrap methods rather than chosen according to theory, because the theoretical thresholds obtained by inequality can be over-conservative and those by approximation can be less accurate. See Section 5 for more details about the algorithm in practice.

3.4. Test level and power

We are ready to derive the main theorem which characterizes the kernel test's level and power when applied to manifold data at a finite sample size.

Theorem 3.4 (Power of kernel test). *Under Assumptions 1, 2, 3, and the balancing condition (11), let the constants γ_0 be as in Lemma 3.1, $\tilde{C}_1$ be as in Lemma 3.2, and c, ν, and $C_1^{(2)}$ be as in Proposition 3.3. Define $\lambda_1 := \sqrt{8\log(3/\alpha_{\text{level}})}$, and let the threshold for the test be $t_{\text{thres}} := 2/(cn) + 4\lambda_1\sqrt{\nu\gamma^d/(cn)}$. For $q \neq p$ under H_1, suppose $\Delta_2 = \int_{\mathcal{M}}(p-q)^2 dV > 0$. Then, when γ is small enough such that $0 < \gamma < \min\left\{1, \gamma_0, (C_1^{(2)})^{-1/2}\right\}$ and*

$$\tilde{C}_1(L_\rho + \rho_{\max})\gamma^\beta < 0.1 m_0[h]\Delta_2^{1/2}, \tag{13}$$

and meanwhile, for some constant $\lambda_2 > 0$, n is large enough such that

$$\gamma^d n > \max\left\{\frac{1}{c\nu}\left(\frac{\max\{\lambda_1,\lambda_2\}}{3}\right)^2, \frac{10}{cm_0[h]\Delta_2}, \frac{\nu}{c}\left(\frac{8(\lambda_1+\lambda_2)}{m_0[h]\Delta_2}\right)^2\right\}, \tag{14}$$

then

$$\mathbb{P}[\widehat{T} > t_{\text{thres}}|H_0] \leq \alpha_{\text{level}}, \quad \mathbb{P}[\widehat{T} \leq t_{\text{thres}}|H_1] \leq 3e^{-\lambda_2^2/8}. \tag{15}$$

We give a few comments to interpret the result in Theorem 3.4. First, the choice of the test threshold in the theorem is a theoretical one to facilitate our analysis, especially to obtain the dependence of test power on various factors like the dimensionality of data. Second, Theorem 3.4 considers a fixed alternative q, and the bound of testing power holds for finite samples and finite γ. To obtain a test power close to 1, namely a Type-II error in (15) as small as ϵ, one can make $\lambda_2 = \sqrt{8\log(3/\epsilon)}$, and then the theorem guarantees the test power when γ can be chosen to satisfy (13) and (14) simultaneously, which requires n to be large enough given Δ_2. This also leads to an argument for, with large n, what is the smallest Δ_2 (scales with a negative power of n) such that the H_1 can be correctly rejected using the kernel tests (with probability at least $1-\epsilon$). We call this the "rate-for-detection" and it is derived in Corollary 3.5. At last, in Theorem 3.4, only the intrinsic dimensionality d affects the testing power but not the ambient space dimensionality m. The constants Δ_2, $\rho_{\max}$, and L_ρ are determined by p and q as Hölder functions on $(\mathcal{M}, dV)$ and are intrinsically defined.

Remark 3.1 (Constant m_0). The constants $m_0[h^2]$ (appearing in the definition of ν) and $m_0[h]$ are integrals of the kernel function in $\mathbb{R}^d$ defined as in (5). The explicit values for the Gaussian RBF kernel are as follows:

Example 3.2 (Constants for Gaussian h). When $h(r) = e^{-r/2}$,

$$m_0[h] = \int_{\mathbb{R}^d} e^{-|u|^2/2}du = (2\pi)^{d/2}, \quad m_0[h^2] = \int_{\mathbb{R}^d} e^{-|u|^2}du = \pi^{d/2}.$$

For general h, both constants depend on d.

We consider the scenario where $\Delta_2(p,q)$ is allowed to decrease to zero as the sample size increases. The following corollary shows that the kernel test can achieve a positive test power (at the test level)

as long as $\Delta_2 \gtrsim n^{-2\beta/(d+4\beta)}$, and is asymptotically consistent (power approaches 1) when Δ_2 is greater than that order.

Corollary 3.5 (Rate-for-detection). *Under the same assumptions as in Theorem 3.4, suppose as n increases, $\gamma \sim n^{-1/(d+4\beta)}$, the densities p and q satisfy that their squared L^2-divergence Δ_2 is positive and is less than an $O(1)$ constant determined by $\rho_{\max}$, d and h, and, for $0 < \epsilon < 1$, with large n,*

$$\Delta_2 > c_3 \left(\log \frac{1}{\alpha_{\text{level}}} + \log \frac{1}{\epsilon} \right)^{1/2} n^{-2\beta/(d+4\beta)}, \tag{16}$$

where the constant c_3 depends on constants $\{L_\rho, \rho_{\max}, \rho_X, d, \beta\}$ and $(\mathcal{M}, h)$ and $\alpha_{\text{level}} < 1/2$. Then, for large enough n, the kernel test achieves a test level α_{level} and a test power at least $1 - \epsilon$. In particular, the test power $\to 1$ as $n \to \infty$ if $\Delta_2 \gg n^{-2\beta/(d+4\beta)}$.

Remark 3.2 (Choice of bandwidth). As shown in Corollary 3.5, when Δ_2 is small as in the regime therein, the bandwidth needs to scale with $n^{-1/(d+4\beta)}$ so that the test can have power. Such a kernel bandwidth $\gamma \to 0$ as n increases. The analysis suggests using small-bandwidth kernels for the test to detect small changes in distribution when large data samples are available. In contrast, the median distance choice of bandwidth Gretton et al. (2012) may lead to γ of order $O(1)$ in this case: on a manifold of diameter $O(1)$, suppose the data density is uniform, then the median of pairwise distance is generally $O(1)$. Thus the median distance γ may not be optimal for high-dimensional data, for example, when data lie on or near intrinsically low-dimensional manifolds or sub-manifolds, and there are sufficiently many samples in the dataset to detect a small departure of the density. We show in Section 5 that high-dimensional data kernel tests with a smaller bandwidth can outperform those with the median distance bandwidth in experiments. Theoretically, note that for kernel tests on data in Euclidean space, the optimal γ is shown to also scale with a negative power n to achieve minimax rate of detection (Li and Yuan, 2019). We further discuss the rate and the relation to this work in the discussion section.

In practice, kernel bandwidth is a hyper-parameter that can be determined by some cross-validation procedure at the cost of additional computation. The optimal choice of bandwidth depends on data distribution and sample size and would be difficult to predict theoretically. In particular, our theory (starting from Lemma 3.1 to Theorem 3.4) needs γ to be less than some $O(1)$ constant, and these theoretical constants can be difficult to obtain in practice especially when the data manifold is unknown. Our analysis does not suggest estimating these constants as a manner to gauge whether the kernel bandwidth is proper or not. Instead, the interpretation of our theory should be that, under the necessary conditions, there exists a γ such that the kernel test is guaranteed to distinguish the density departure. Such γ can be found, e.g., by cross-validation in practice. Of course, this only happens with sufficient data samples, and when the sample size is not large enough then the test power cannot be guaranteed – the selected γ in practice may still lead to a test with power that is not guaranteed by our theory here. Our rate for detection result provides a theoretical scaling of γ, which may provide guidance for the range of the value to search for in practice. For example, if the manifold intrinsic dimension is known *a priori* or can be estimated from data (Brito, Quiroz and Yukich, 2013, Farahmand, Szepesvári and Audibert, 2007, Levina and Bickel, 2004, Mordohai and Medioni, 2010, Pettis et al., 1979), our theoretical scaling would suggest how the bandwidth parameter should scale as the sample size increases. Generally, when more data samples are available, the theory suggests searching the smaller value range from the median distance, which can improve the detection ability of the kernel test for small density departures.

Remark 3.3 (Higher Hölder regularity). As shown in the proof of Corollary 3.5, the improvement of detection rate from higher regularity β is via analyzing how to fulfill (13) by setting γ sufficiently small. The condition (13) is based on the bound of $|r_T|$ in Lemma 3.2, and the latter is proved by

Lemma 3.1 which gives an $O(\gamma^{\beta\wedge 2})$ bound in (6). This means that when $\beta > 2$ (and $\|p-q\|_\infty > 0$) the bound of $|r_T|$ will remain $O(\gamma^2)$. As a result, higher Hölder regularity of the densities beyond two will not further improve the rate under the current analysis.

Remark 3.4 (Linear-time statistic). When $n_X = n_Y = 2m$, the linear-time test statistic, following (Gretton et al., 2012, Section 6), is defined as $\widehat{T}_{\text{lin}} = \frac{1}{m}\sum_{i=1}^{m} h(z_{2i-1}, z_{2i})$, where $z_i = (x_i, y_i)$ and $h(z_i, z_j) := K_\gamma(x_i, x_j) - K_\gamma(x_i, y_j) - K_\gamma(y_i, x_j) + K_\gamma(y_i, y_j)$. The construction in Gretton et al. (2012) is for kernel MMD test, but $\widehat{T}_{\text{lin}}$ is well-defined when the kernel K_γ is not PSD. The statistic $\widehat{T}_{\text{lin}}$ can be computed using $O(n)$ time and memory. The mean $\mathbb{E}\widehat{T}_{\text{lin}} = T$ as has been analyzed in Lemma 3.2; The deviation of $\widehat{T}_{\text{lin}}$ from mean observes bounds of the same order as in Proposition 3.3, since the finite-sample concentration of $\widehat{T}_{\text{lin}}$ is that of an independent sum of $n_X/2$ terms (technically the decoupling argument in the proof of Proposition 3.3 reduces the concentration of the U-statistic to that of the $(2i-1, 2i)$-indexed independent sum). As a result, the same test power analysis and rate of detection as proved in Theorem 3.4 and Corollary 3.5 hold for the linear-time statistic $\widehat{T}_{\text{lin}}$.

4. Theoretical extensions to manifold with boundary and noisy data

In this section, we extend the analysis in Section 3 to two important cases, namely when the data manifold has a boundary and data has additive noise in high dimensional ambient space.

4.1. Manifold with boundary

In many scenarios the data manifold has a boundary. For instance, when the range of rotation angle in Example 3.1 is less than $[0, 2\pi]$ the curve in the image space is not closed, and this is an example of manifold having boundaries, also see Figure 2. Another reason to consider boundary is the applicability of our theory to the Euclidean case (the manifold is "flat"), where after assuming compact support of the distributions the support domain will have a boundary, see more in Remark 4.1.

For our analysis, when a data point x approaches the manifold boundary the support of the local kernel will also intersect with the boundary, which makes the expression of local kernel integral in Lemma 3.1 not hold and voids the subsequent analysis. The current section is devoted to extending the theory in Section 3 to the case of the manifold with boundary by first extending Lemma 3.1. We assume

Assumption 4. $\mathcal{M}$ is a d-dimensional compact C^∞ sub-manifold isometrically embedded in $\mathbb{R}^m$, where the boundary $\partial\mathcal{M}$ is also C^∞.

The analysis proceeds using similar techniques and is based on the local kernel integral lemma (Lemma 4.1), which handles when x is on or near to $\partial\mathcal{M}$. Theorem 3.4 then extends under an additional Assumption 5 and certain modifications of the constants and condition (13), see the specifics in Theorem 4.3.

Remark 4.1 (Euclidean space). When data densities p and q are compactly supported on some domain Ω in $\mathbb{R}^m$ and Ω has a smooth boundary, this is a special case of the manifold-with-boundary setting where $d = m$. Our theoretical result thus covers such cases. When m is large, there is a curse-of-dimensionality revealed by the γ^d factor in the required lower bound of n in the condition (14).

We start by establishing the following lemma, which is the counterpart of Lemma 3.1.

Lemma 4.1. *Suppose $\mathcal{M}$ satisfies Assumption 4, h satisfies Assumption 3, and f is in $\mathcal{H}^\beta(\mathcal{M})$, $0 < \beta \le 2$, with Hölder constant L_f. Let $d_E(x,\partial\mathcal{M}) := \inf_{y\in\partial\mathcal{M}} \|x-y\|$, and define $\delta_\gamma := \sqrt{\frac{d+10}{a}\gamma^2 \log\frac{1}{\gamma}}$. Then, there is $\gamma_0' > 0$ which depends on $\mathcal{M}$ only, such that when $0 < \gamma < \min\{\gamma_0', 1\}$,*

(i) For any $x \in \mathcal{M}$ such that $d_E(x,\partial\mathcal{M}) > \delta_\gamma$, (6) holds.

(ii) There is constant C_1' that depends on $(\mathcal{M},h)$, such that for any x s.t. $d_E(x,\partial\mathcal{M}) \le \delta_\gamma$, there exists a function $m_0^{(\gamma)}[h](x)$ depending on γ s.t. $0 \le m_0^{(\gamma)}[h](x) \le m_0[h]$ for all x and

$$\left|\gamma^{-d}\int_{\mathcal{M}} h\left(\frac{\|x-y\|^2}{\gamma^2}\right) f(y)dV(y) - m_0^{(\gamma)}[h](x)f(x)\right| \le C_1'(L_f\gamma^{\beta\wedge 1} + \|f\|_\infty\gamma^2). \tag{17}$$

Similarly, as in Lemma 3.1, γ_0' depends on manifold reach and curvature, and the constant C_1' depends on manifold curvature and volume, and the kernel function h. The lemma shows that the error bound at x that is δ_γ away from $\mathcal{M}$ is $O(\gamma^{\beta\wedge 2})$ same as before, and at x that is within δ_γ distance from $\mathcal{M}$ is $O(\gamma^{\beta\wedge 1})$. This reflects the degeneracy of the kernel integral approximation at x, which is close to the manifold boundary.

When $\gamma < \min\{\gamma_0', 1\}$, we define $P_\gamma := \{x \in \mathcal{M}, d_E(x,\partial\mathcal{M}) \le \delta_\gamma\}$, which is the δ_γ-near-boundary set as shown in Figure 1(Right). To extend Lemma 3.2, we introduce the assumption that the major part of Δ_2 is not coming from the integral on P_γ.

Assumption 5. For $q \ne p$, there are positive constants γ_0'' and C_3 possibly depending on $\mathcal{M}$ (and independent from p and q), such that when $\gamma < \gamma_0''$,

$$\int_{P_\gamma} (p-q)^2 dV \le C_3\delta_\gamma \int_{\mathcal{M}} (p-q)^2 dV.$$

We then extend Lemma 3.2 which bounds the error between $\gamma^{-d}T$ and $m_0[h]\Delta_2(p,q)$ in the following lemma.

Lemma 4.2. *Under Assumptions 2, 3, 4, 5, γ_0' as in Lemma 4.1. Then, when $0 < \gamma < \min\{1, \gamma_0', \gamma_0''\}$, we have that*

$$\gamma^{-d}T = m_0[h]\Delta_2 + r_T, \quad |r_T| \le C_3\delta_\gamma m_0[h]\Delta_2 + (L_\rho + \rho_{\max})(\tilde{C}_1\gamma^\beta + \tilde{C}_2'\gamma^{\beta\wedge 1}\delta_\gamma)\Delta_2^{1/2}, \tag{18}$$

where the constants $\tilde{C}_1$ (as in Lemma 3.2) and $\tilde{C}_2'$ depend on $(\mathcal{M},h)$ only, including manifold curvature and volume, the regularity and volume of $\partial\mathcal{M}$, and the intrinsic dimensionality d.

Next, we extend Proposition 3.3 after replacing the constant $C_1^{(2)}$ with some $C_1'^{,(2)}$, and γ_0 with γ_0', in the statement (details in the proof), and this allows extending Theorem 3.4 to a data manifold with smooth boundary in the following theorem.

Theorem 4.3. *Under Assumptions 2, 3, 4, and 5, the same bound of test power as in Theorem 3.4 holds with the following changes: (i) replacing the constant $C_1^{(2)}$ with $C_1'^{,(2)}$ and requiring $0 < \gamma < \min\{1, \gamma_0', \gamma_0''\}$, (ii) condition (13) is replaced by*

$$C_3\delta_\gamma < 0.05, \quad (L_\rho + \rho_{\max})\left(\tilde{C}_1\gamma^\beta + \tilde{C}_2'\gamma^{\beta\wedge 1}\delta_\gamma\right) < 0.05 m_0[h]\Delta_2^{1/2}, \tag{19}$$

where the constants $\tilde{C}_1$ and $\tilde{C}_2'$ are as in Lemma 4.2 and depend on $(\mathcal{M},h)$ only.

Based on (19), which is again implied by the technical bound in (18), the above theorem induces a detection rate similar to in Corollary 3.5: Specifically, first note that $\delta_\gamma \sim \gamma\sqrt{\log(1/\gamma)}$ which is $o(1)$ as $\gamma \to 0$, thus $C_3\delta_\gamma < 0.05$ is satisfied when γ is less than some $O(1)$ threshold. In the second equation in (19), note that $\gamma^{\beta\wedge 1}\delta_\gamma \sim \gamma^{\beta\wedge 1+1}\sqrt{\log(1/\gamma)}$ which is dominated by γ^β (when $\beta = 2$, there is a factor of $\sqrt{\log(1/\gamma)}$). Thus the smallness of γ requirement is the same as in the proof of Corollary 3.5 (up to a factor of $\sqrt{\log(1/\gamma)}$). The largeness of n requirement is the same as before. As a result, the rate of detection for small enough Δ_2 in the order of n and the optimal scaling of γ are the same as in Corollary 3.5.

In case when Assumption 5 does not hold, one can derive upper bound of $|r_T|$ using similar techniques as in Lemma 4.2, and the rest of the analysis also generalizes. As the bound of $|r_T|$ will be worsen (due to that the kernel integral approximation error degenerates near the boundary as shown in Lemma 4.1(ii)), the resulting rate is also worse than in Theorem 4.3. Details are omitted.

In line with the theoretical results, the experiments in Section 5 are conducted on manifold data where $\mathcal{M}$ has a boundary. In Section 5.1, the data manifold is a continuous curve in the ambient space with endpoints. In Section 5.2, the original MNIST image data lie close to a collection of sub-manifolds in the ambient space, and it is also a case of a manifold with a boundary.

4.2. Near-manifold noisy data

In applications, data points may not lie exactly on the low-dimensional manifold but only near it. Since kernel $K_\gamma(x,y)$ is computed from Euclidean distances among data points, one can expect that if data samples are lying within a distance proportional to γ from the manifold $\mathcal{M}$, then the integration of kernel $K_\gamma(x,y)$ over such data distributions will preserve the magnitude to be of order γ^d and will not have a curse of dimensionality.

An important case is when near-manifold data are produced by adding Gaussian noise, which is distributed as $\mathcal{N}(0,\sigma^2 I_m)$, to data points that are lying on a manifold. In this case, to make the off-manifold perturbation to be of length up to constant times of γ (with high probability), it allows σ to be up to $c\gamma/\sqrt{m}$ for some $c > 0$. Here, we show that Theorem 3.4 can be extended under this noise regime for Gaussian kernel h. The analysis may also extend to other types of kernel functions.

Specifically, let $x_i = x_i^{(c)} + \xi_i^{(1)}$, $x_i^{(c)} \sim p_{\mathcal{M}}$, $\xi_i^{(1)} \sim \mathcal{N}(0,\sigma_{(1)}^2 I_m)$, and $y_i = y_i^{(c)} + \xi_i^{(2)}$, $y_i^{(c)} \sim q_{\mathcal{M}}$, $\xi_i^{(2)} \sim \mathcal{N}(0,\sigma_{(2)}^2 I_m)$, where the manifold clean data $x_i^{(c)}$ and $y_i^{(c)}$ are independent from the ambient space Gaussian noise $\xi_i^{(1)}$ and $\xi_i^{(2)}$. When $p_{\mathcal{M}}$ and $q_{\mathcal{M}}$ satisfies Assumption 2 and h is Gaussian kernel, Theorem 3.4 extends when, for some $c > 0$,

$$\sigma_{(1)}^2 + \sigma_{(2)}^2 \le \frac{c^2}{m}\gamma^2. \tag{20}$$

The argument is based on that the proof of Theorem 3.4 relies on the approximation of kernel integrals $\mathbb{E}_{x\sim p,y\sim p}K_\gamma(x,y)$ and the boundedness of $\mathbb{E}_{x\sim p,y\sim p}K_\gamma(x,y)^2$ at the order $O(\gamma^d)$, and similarly with $\mathbb{E}_{x\sim p,y\sim q}$, $\mathbb{E}_{x\sim q,y\sim q}$. Thus, when kernel h is Gaussian, and p (and q) equals $p_{\mathcal{M}}$ (and $q_{\mathcal{M}}$) convolved with a Gaussian with coordinate variance $\lesssim \gamma^2/m$ in $\mathbb{R}^m$, these integrals can be shown to be equivalent to those integrated over $p_{\mathcal{M}}$ and $q_{\mathcal{M}}$ with another Gaussian kernel having bandwidth $\tilde{\gamma}$, where $\tilde{\gamma}/\gamma$ is bounded between 1 and the absolute constant $\sqrt{1+c^2/m}$. As a result, the integrals of $K_\gamma(x,y)$ and $K_\gamma(x,y)^2$ can be computed same as before in Lemma 3.2 and Proposition 3.3, leading to a result of Theorem 3.4 after replacing the roles of p and q with $p_{\mathcal{M}}$ and $q_{\mathcal{M}}$. Details are left to Section A.3 of Cheng and Xie (2024).

This suggests that when the coordinate-wise noise level σ in $\mathbb{R}^m$ is bounded at the level of $\gamma/\sqrt{m}$, the behavior of the two-sample test with kernel $K_\gamma(x,y)$ applied to manifold-plus-noise data is essentially close to as if applied to the clean on-manifold data, and the testing power is determined by the on-manifold distributions $p_{\mathcal{M}}$ and $q_{\mathcal{M}}$. Experiments of data with additive Gaussian noise are given in Section 5, which verifies this theoretical prediction. In practice, we observe that the testing performances on clean and noisy data will stay close for small noise level σ, and start to show discrepancies when σ exceeds a certain level.

5. Numerical experiments

In this section, we present several numerical examples to demonstrate the validity of our theory. We first study a synthetic example of image data lying on a manifold, and then a density departure example using the MNIST dataset. The summary of the algorithm, including computation of the test threshold by bootstrap Arcones and Giné (1992), is provided in Section B of Cheng and Xie (2024). Code available at the public repository https://github.com/xycheng/manifold_mmd.

The notations are as follows: n_{run} is the number of replicas used to estimate the test power, and n_{boot} the number of bootstrap samples in computing the test threshold. We set the test level $\alpha_{\text{level}} = 0.05$ throughout. In our experiments, we test over a range of kernel bandwidth parameters γ. In practice, γ can also be chosen adaptively from data, e.g., the *median distance bandwidth* is set to be the median of all pairwise distances in the two sample datasets. Our theory in Section 3 suggests that the median distance γ may not always be the optimal choice: for manifold data of intrinsically low dimensionality, kernels with smaller bandwidth can achieve better testing power when there are sufficiently many samples. We verify this in experiments. In addition, we examine the Gaussian kernel and several possibly non-PSD kernels, and verify that the latter can also achieve high test power as suggested by our theory.

5.1. Images with differently distributed rotation angles

5.1.1. Clean data

We construct two datasets consisting of randomly rotated copies of an image of the handwritten digit '8', which are resized to be of different resolutions. The data-generating process was introduced in Example 3.1 and illustrated in Figure 2. This experiment is designed such that we specify the true distributions on the latent manifold (rotation angles), which induces the distributions of observed manifold data. The distributions p and q of the two datasets are induced by different rotation angle distributions, and the densities of rotation angles (between 0 to 90 degrees) are shown in the left of Figure 2. The image size changes from 10×10 to 40×40, and as a result, the data dimensionality increases from 100 to 1600. Note that since the rotation is only up to $\pi/2$, the corresponding manifold is a 1D curve with two endpoints, namely a manifold with a boundary.

Note that the image pixel values maintain the same magnitude as W increases, and then the value of $\frac{1}{m}\sum_{u=1}^{m} I_i(u)^2$ approaches an $O(1)$ limit, which is the squared integral of the underlying continuous function on $[0,1]^2$. This means that the image data vectors of size $W \times W$ need to divide by $\sqrt{m}$, $m = W^2$, so as to obtain isometric embedding of a manifold of diameter $O(1)$ in $\mathbb{R}^m$, In experiments we use bandwidth γ to resized images, can we call $\gamma/\sqrt{m}$ the "pixel-wise bandwidth". The pixel-wise bandwidth corresponds to the "γ" in theory in Section 3.

In computing the Gaussian kernel test statistics, we use bandwidth parameters over 5 values such that

$$\frac{\gamma}{\sqrt{m}} = \gamma_0 \left\{ \frac{1}{4}, \frac{1}{2}, 1, 2, 4 \right\}, \quad m = W^2, \quad W = 10, \cdots, 40, \tag{21}$$

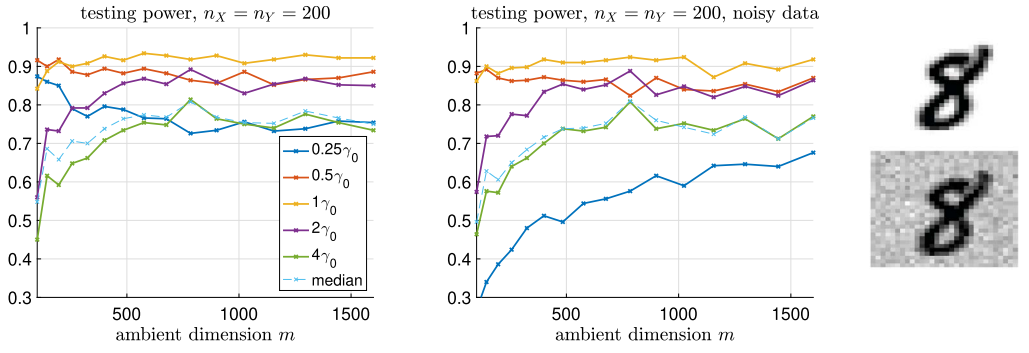

Figure 3. An example of simulated manifold data on which the kernel test power does not drop as the ambient dimension m increases, where the intrinsic dimension d remains constant. Gaussian kernel test statistics are computed on two datasets of rotated images with different distributions of rotation angles. Images are of sizes 10×10, $\cdots$, 40×40, and thus m increases from 100 to 1600. The test is computed with 5 values of kernel bandwidth as in (21), and that is chosen by the median distance from the data. The test power is estimated using $n_{\text{run}} = 500$. (Left) Results on clean images. (Middle) Results on images with additive Gaussian noise, where the noise level is chosen to be small and satisfies the condition in Section 4.2. (Right) Example clean and noisy images (size 30×30).

with $\gamma_0 = 20$, which we call the "baseline pixel-wise bandwidth". The median distance gives the pixel-wise bandwidth is about 70. Since the grayscale images take pixel values between 0 and 255, the pixel-wise bandwidth being 20 is relatively small, and is smaller than that chosen by the median distance. The estimated two-sample testing power on clean images is shown in Figure 3 (Left), which is computed using $n_{\text{boot}} = 400$ and $n_{\text{run}} = 500$. It can be seen that all the bandwidth choices give certain test power, which is consistent across m as m increases (showing a tendency of convergence after m exceeds 500 till 1600). The performance with pixel-wise bandwidth equal to 20 appears to be the best and is better than the bandwidth by median distance.

5.1.2. Noisy data

We add pixel-wise Gaussian noise of standard deviation $\sigma_0 = 20$ to the resized image data of dimension m, that is, $\sigma_0 = \gamma_0$ the baseline pixel-wise bandwidth in the previous clean data experiment. This falls under the scenario in Section 4.2: As was pointed out in the experiment with clean data, normalized clean image $I_i/\sqrt{m}$ lie on an $O(1)$ manifold, where I_i has size $W \times W$, $m = W^2$, and thus the pixel-wise bandwidth corresponds to the "γ" in the theory, If we add Gaussian noise $\mathcal{N}(0, \sigma_0^2 I_m)$ to the clean image I_i, it corresponds to adding noise $\mathcal{N}(0, (\sigma_0^2/m) I_m)$ to $I_i/\sqrt{m}$. Thus $\sigma_0/\sqrt{m}$ is the "σ" in Section 4.2. The small noise regime in Section 4.2 requires "$\sigma < c\gamma/\sqrt{m}$" for some constant c, and here, "γ" there is γ_0, and "σ" there is $\sigma_0/\sqrt{m}$, thus the condition translates into $\sigma_0/\sqrt{m} < c\gamma_0/\sqrt{m}$, which is satisfied if we set $\sigma_0 = \gamma_0$.

An example pair of clean and noise-corrupted images are shown in Figure 3 (Right). We conduct the two-sample testing experiments in the same way as in the experiment with clean data, and the the estimated testing power is shown in Figure 3 (Middle). The performance with the four pixel-wise bandwidth, which is greater than $\gamma_0/2$ are about the same as on the clean data; With the smallest pixel bandwidth $\gamma_0/4$, the test power degenerates and becomes worse than the choice by median distance, and the drop is more significant when dimensionality m is small. This suggests that this kernel bandwidth is too small for the amount of additive noise at the values of m and sample size n_X and n_Y.

5.2. Density departure in MNIST dataset

In this experiment, we compute the Gaussian kernel two-sample test on the original MNIST digit image dataset, where samples are of dimensionality 28×28. The data densities p and q are generated in the following way: p is uniformly subsample from the MNIST dataset, namely $p = p_{\text{data}}$. Though we only have finite samples (the MNIST dataset has 70000 images in 10 classes) of p_{data}, as we subsample $n_X = 6000$ from the whole dataset, it is approximate as if drawn from the population density p. q is constructed as a mixture of $q = 0.975p_{\text{data}} + 0.025p_{\text{cohort}}$, where p_{cohort} is the distribution of a local cohort within the samples of digit "1", having about 1700 samples. Since n_Y is set to be about 6000, and we subsample about 150 from the local cohort, the way we simulate samples in Y is approximately as if drawn from the density q. The local cohort corresponding to p_{cohort} is illustrated in Figure 4 (Bottom left), indicating the place where the density q departures from p. The experiment is conducted on one realization of dataset X and Y, where $n_X = 6000$, $n_Y = 5990$.

We apply the kernel test with two bandwidths, one using the median distance, which gives the pixel-wise bandwidth $\gamma/\sqrt{m} = 92.9$, and here $m = 28^2$; and the other takes $\gamma/\sqrt{m} = 25$. The test statistic $\widehat{T}$ under H_1 vs. the histogram under H_0 computed by bootstrap with $n_{\text{boot}} = 1000$ are shown in the top panel of Figure 4. It can be seen that with the smaller bandwidth, the test statistic shows a more clear rejection of H_0, indicating better testing power.

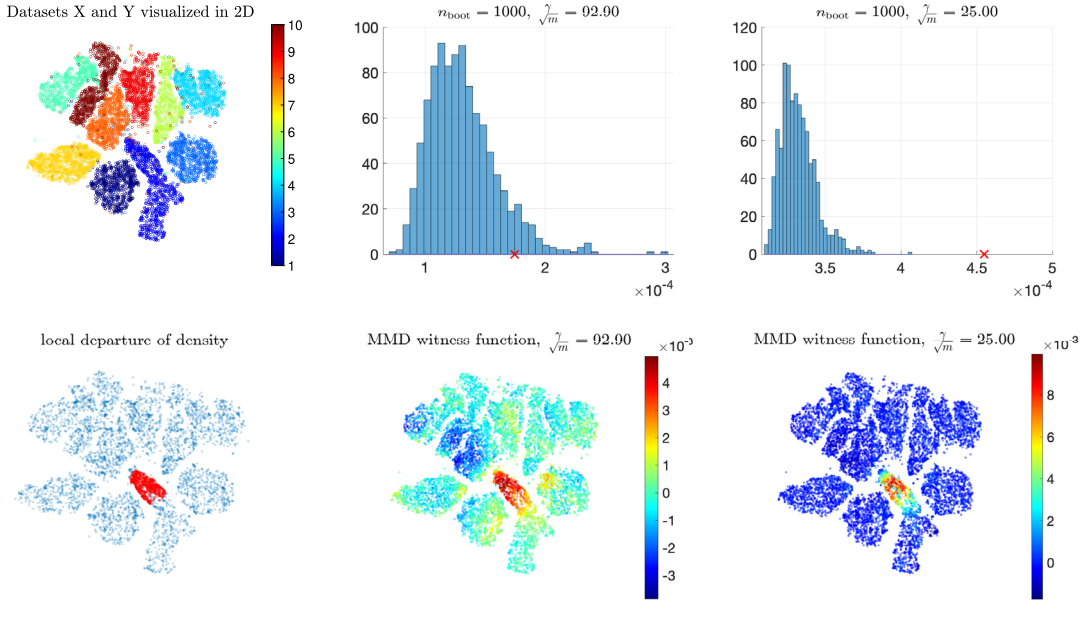

Figure 4. Kernel two-sample test to detect a local density departure of the MNIST image data distributions. (Top left) Datasets X and Y are visualized in 2D by tSNE, colored by 10-digit class labels. (Top middle and right) Kernel test statistic $\widehat{T}$ (red cross) plotted against the histogram of test statistic under H_0 computed by bootstrap Arcones and Giné (1992) (blue bar, see more in Section B of Cheng and Xie (2024)). The middle plot is for the Gaussian kernel test using median distance γ, and the right plot is by using a smaller γ. (Bottom left) The local cohort density p_{cohort} is illustrated by red dots. (Bottom middle and right) The witness function defined in (22) for kernel using median distance γ and a smaller bandwidth, respectively.

The *witness function* of kernel MMD (Gretton et al., 2012) is defined as

$$\hat{w}(x) = -\frac{1}{n_X}\sum_{i=1}^{n_X} K_\gamma(x, x_i) + \frac{1}{n_Y}\sum_{j=1}^{n_Y} K_\gamma(x, y_j), \tag{22}$$

and we visualize $\hat{w}$ with the Gaussian kernel as heat-map on the 2D embedding in the bottom panel of Figure 4. The witness function indicates where the two densities differ. Compared with the ground truth in the bottom left plot of the departure p_{cohort}, the witness function computed with the local kernel better detects the density departure than the median distance kernel, and this is consistent with the better test statistic separation in the top panel plots.

As a remark, unlike in Section 5.1, the MNIST image data do not lie on any constructed manifold induced by latent group action, but only lie near certain manifold-like structures in the ambient space – the latent manifold reveals all possible variations of images of the 10 digits, and since there are 10 classes, there are possibly 10 sub-manifolds (which may be of different intrinsic dimensionalities on each piece), as illustrated by the 2D embedding by tSNE (t-distributed Stochastic Neighbor Embedding (van der Maaten, 2014)) in Figure 4 (Top left). Thus, the case does not fall under the exact theoretical assumption of manifold data in Section 3, even though manifold-like structures are likely to be present in the dataset. The experimental results show that in this generalized case, there may still be a benefit to testing power by using a more localized kernel with a smaller bandwidth than the median distance bandwidth.

5.3. Two-sample tests with non-PSD kernels

Using the same data and experimental set-up as in Section 5.1, we examine different choices of the kernel function h, which are non-Gaussian and possibly non-PSD. The indicator kernel corresponds to the "epsilon-graph" construction frequently used in manifold learning, e.g. in ISOMAP (Tenenbaum, de Silva and Langford, 2000). The example kernels here are mainly designed to verify the theoretical prediction that the test with a non-PSD kernel can still have power, with no suggestion of any immediate practical advantage of these kernels for the testing problems. Specifically, we study

- Sigmoid kernel: $h(r) = \frac{\exp\{-10(r-2)\}}{1+\exp\{-10(r-2)\}}$, which is a translated and rescaled sigmoid function and satisfies Assumption 3, but generally gives a non-PSD kernel $K_\gamma(x, y)$ for data in $\mathbb{R}^m$.
- Sinc kernel: $h(r) = \sin(\frac{\pi}{2}r)/(\frac{\pi}{2}r)$, which takes both positive and negative values on $(0, \infty)$, and only has $1/r$ decay, violating both (C2) and (C3).
- Indicator kernel: $h(r) = \mathbf{1}_{[0,2)}(r)$, which is not continuous on $(0, \infty)$, violating (C1).

Plots of the kernel function $h(r)$ as univariate functions are shown in the left column of Figure 5. The testing power for clean and noisy data over a range of kernel bandwidth, including the median distance bandwidth, are shown in the right two columns (same plots as in Figure 3). The results show that these kernel tests, when the kernel is non-PSD and even violates the theoretical assumption, can obtain testing power if the kernel bandwidth is properly chosen. In addition, the optimal bandwidth may not be the median distance: the best performance achieved by the three kernels for clean data is obtained with γ_0 or $2\gamma_0$ in (21), which is smaller than the median distance as explained in Section 5.1. With noisy data, the testing powers worsen, but the power achieved by the best bandwidth is again comparable to that on the clean data for all three kernels; though the best bandwidth takes a different value from that under the clean data (which can be anticipated based on the analysis in Section A.3 of Cheng and Xie (2024)). Comparing the (translated and rescaled) sigmoid kernel and the indicator kernel in Figure 5, it can be seen that the smoothness of the kernel function leads to better noise robustness; On this data,

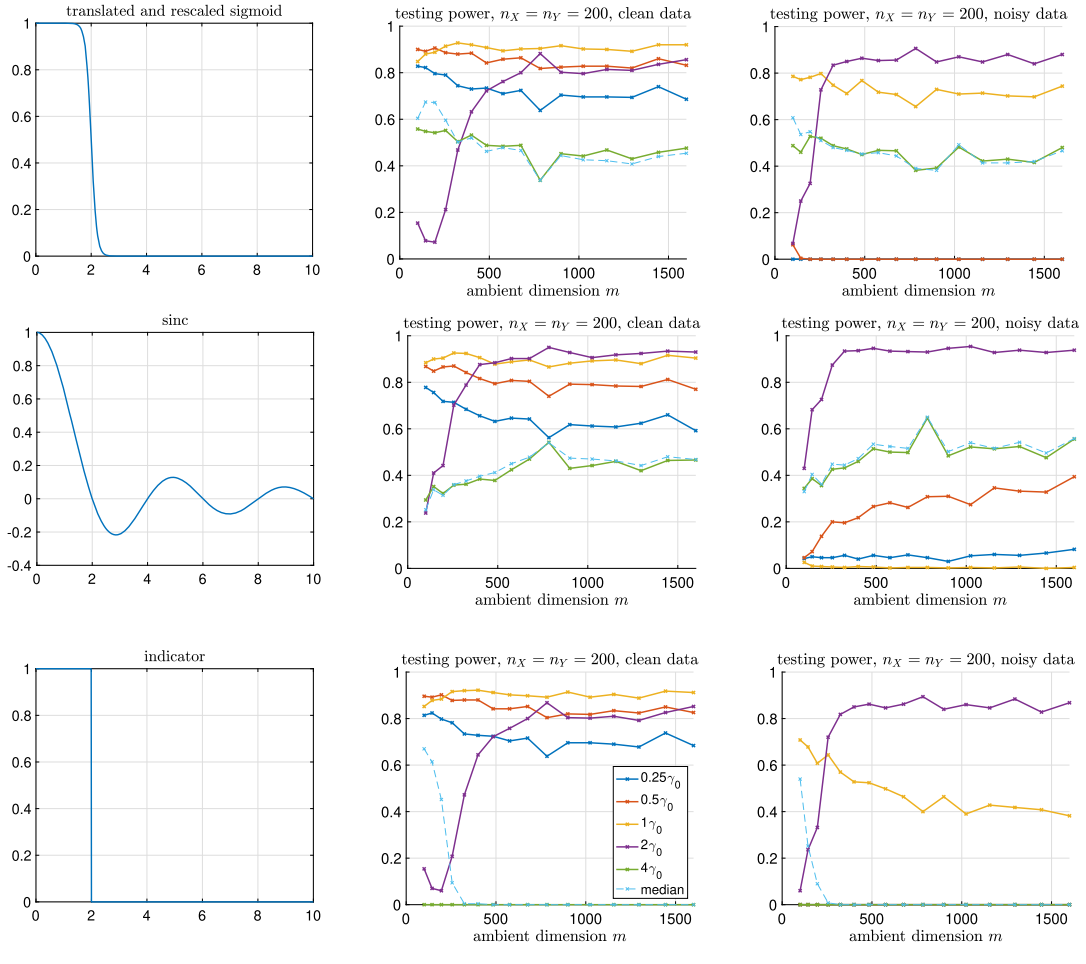

Figure 5. Two-sample test with non-Gaussian kernels that may not be PSD. (Left column) Three choices of kernel function h as in Section 5.3. (Right two columns) Same plots of testing powers on clean and noisy data of rotated images as in Figure 3, of the three kernel functions respectively.

the Gaussian kernel shows better noise robustness than the non-PSD kernels, comparing Figure 3 with Figure 5.

In all experimental trials presented in this paper, we observe that the testing power remains constant as the dimensionality of the data, denoted by m, increases and ultimately converges. Additionally, our results demonstrate that the curse-of-dimensionality does not affect manifold data and support theoretical proposals that suggest the kernel's consistency and power in two-sample tests do not depend on whether or not the kernel is PSD, which would result in the test statistic being an RKHS MMD.

6. Discussion

We provide a curse-of-dimensionality free result for kernel MMD-like statistics, which we expect to be tight for the linear-time version of the kernel statistic (see Remark 3.4) but not necessarily for the

full (quadratic-time) kernel statistic. For the full kernel statistic computed from data in d-dimensional Euclidean space, Li and Yuan (2019) showed that the Gaussian kernel statistic achieves a detection rate of $\Delta_2 \gg n^{-4\beta/(d+4\beta)}$, which is minimax optimal against smooth alternatives (c.f. Theorem 5 of Li and Yuan (2019), recall that Δ_2 is the squared L^2 divergence). One may conjecture the same minimax rate for data with intrinsic data dimension d. This rate of $n^{-4\beta/(d+4\beta)}$ is better than our proved rate of $\Delta_2 \gtrsim n^{-2\beta/(d+4\beta)}$ in Corollary 3.5. This gap may be due to the control of the fluctuation of the U-statistic by Proposition 3.3 not being tight (see the comment after the proposition). Obtaining a test power result at a finite sample size that matches the conjectured minimax rate under our manifold data setting is left to future work. However, the picture differs when considering computational complexity. Since the vanilla computation of the full statistic has $O(n^2)$ complexity, with the same computational and memory cost, the linear-time statistic can process $\tilde{n} \sim n^2$ samples. According to the proved rate by our result, it pushes the detection boundary to be $\Delta_2 \gtrsim \tilde{n}^{-2\beta/(d+4\beta)} \sim n^{-4\beta/(d+4\beta)}$, which is the same as the conjectured optimal rate. The linear-time statistic can be computed online (Flynn and Yoo, 2019), and thus can be viewed as trading the smaller variance by revisiting all the samples for faster computation on the fly. In summary, the statistical optimality of two-sample kernel statistics applied to intrinsically d-dimensional data remains to be further studied. It would be interesting to design kernel tests that achieve the theoretical detection rate with matched computational complexity.

Our work can be extended in several other directions. For instance, the current paper only considers isotropic kernels with fixed bandwidth. It would be of interest to generalize to other types of kernels used in practice, such as anisotropic kernels using local Mahalanobis distance, other non-Euclidean metrics, kernels with adaptive bandwidth (Cheng and Wu, 2022a, Zelnik-Manor and Perona, 2005), asymmetric kernels with a reference set (Cheng, Cloninger and Coifman, 2020, Jitkrittum et al., 2016), and so on. To expand the theoretical framework, it would be interesting to go beyond the compactness assumption of the manifold, which would allow extracting the low intrinsic dimensionality or low complexity of high dimensional data distributions that are unbounded or have long tails. One may extend the theory to more complicated manifold structures, like multiple sub-manifolds of different intrinsic dimensions or with complicated boundaries. More advanced analysis of the near-manifold setting, for example, by analyzing general high dimensional noise, would also be a useful extension. Algorithm-wise, while the theory in this work suggests using smaller bandwidth depending on data intrinsic dimensionality and sample size, providing a theoretical scaling of γ for large n, it remains to further develop efficient algorithms to choose kernel bandwidth in practice. Developing efficient kernel testing methods to reduce storage and computational costs would also be desirable. At last, it is natural to extend to other kernel-based testing problems, such as goodness-of-fit tests (Chwialkowski, Strathmann and Gretton, 2016, Jitkrittum, Kanagawa and Schölkopf, 2020, Shapiro, Xie and Zhang, 2021), and general hypothesis tests. Reducing sampling complexity by the intrinsic low-dimensionality of manifold data may also be beneficial therein.

Acknowledgments

The authors would like to thank the anonymous referees and the Associate Editor for their constructive comments that improved the quality of this paper.

Funding

The work was supported by NSF DMS-2134037. X.C. was also partially supported by NSF DMS-2237842 and DMS-2007040. Y.X. was also partially supported by an NSF CAREER CCF-1650913, NSF DMS-2134037, CMMI-2015787, CMMI-2112533, DMS-1938106, and DMS-1830210.

Supplementary Material

Supplement to "Kernel two-sample tests for manifold data". (DOI: 10.3150/23-BEJ1685SUPP; .pdf). This supplement contains proofs and details of the bootstrap estimation of test threshold.

References

Arcones, M.A. and Giné, E. (1992). On the bootstrap of U and V statistics. *Ann. Statist.* **20** 655–674. MR1165586 https://doi.org/10.1214/aos/1176348650

Balasubramanian, K., Li, T. and Yuan, M. (2021). On the optimality of kernel-embedding based goodness-of-fit tests. *J. Mach. Learn. Res.* **22** 1–45. MR4253694

Belkin, M. and Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.* **15** 1373–1396.

Belkin, M. and Niyogi, P. (2007). Convergence of Laplacian eigenmaps. In *Advances in Neural Information Processing Systems*. 129–136.

Bhattacharya, B.B. (2020). Asymptotic distribution and detection thresholds for two-sample tests based on geometric graphs. *Ann. Statist.* **48** 2879–2903. MR4152627 https://doi.org/10.1214/19-AOS1913

Bhuyan, M.H., Bhattacharyya, D.K. and Kalita, J.K. (2013). Network anomaly detection: Methods, systems and tools. *IEEE Commun. Surv. Tutor.* **16** 303–336.

Bińkowski, M., Sutherland, D.J., Arbel, M. and Gretton, A. (2018). Demystifying MMD GANs. In *International Conference on Learning Representations*.

Borgwardt, K.M., Gretton, A., Rasch, M.J., Kriegel, H.-P., Schölkopf, B. and Smola, A.J. (2006). Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics* **22** e49–e57.

Brito, M.R., Quiroz, A.J. and Yukich, J.E. (2013). Intrinsic dimension identification via graph-theoretic methods. *J. Multivariate Anal.* **116** 263–277. MR3049904 https://doi.org/10.1016/j.jmva.2012.12.007

Buades, A., Coll, B. and Morel, J.-M. (2005). A non-local algorithm for image denoising. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* **2** 60–65. IEEE.

Calder, J. and García Trillos, N. (2022). Improved spectral convergence rates for graph Laplacians on ε-graphs and k-NN graphs. *Appl. Comput. Harmon. Anal.* **60** 123–175. MR4393800 https://doi.org/10.1016/j.acha.2022.02.004

Cao, Y., Nemirovski, A., Xie, Y., Guigues, V. and Juditsky, A. (2018). Change detection via affine and quadratic detectors. *Electron. J. Stat.* **12** 1–57. MR3743736 https://doi.org/10.1214/17-EJS1373

Chandola, V., Banerjee, A. and Kumar, V. (2009). Anomaly detection: A survey. *ACM Comput. Surv.* **41**.

Chandola, V., Banerjee, A. and Kumar, V. (2010). Anomaly detection for discrete sequences: A survey. *IEEE Trans. Knowl. Data Eng.* **24** 823–839.

Chen, H. and Friedman, J.H. (2017). A new graph-based two-sample test for multivariate and object data. *J. Amer. Statist. Assoc.* **112** 397–409. MR3646580 https://doi.org/10.1080/01621459.2016.1147356

Cheng, X., Cloninger, A. and Coifman, R.R. (2020). Two-sample statistics based on anisotropic kernels. *Inf. Inference* **9** 677–719. MR4146351 https://doi.org/10.1093/imaiai/iaz018

Cheng, X. and Wu, H.-T. (2022a). Convergence of graph Laplacian with kNN self-tuned kernels. *Inf. Inference* **11** 889–957. MR4491976 https://doi.org/10.1093/imaiai/iaab019

Cheng, X. and Wu, N. (2022b). Eigen-convergence of Gaussian kernelized graph Laplacian by manifold heat interpolation. *Appl. Comput. Harmon. Anal.* **61** 132–190. MR4452681 https://doi.org/10.1016/j.acha.2022.06.003

Cheng, X. and Xie, Y. (2024). Supplement to "Kernel two-sample tests for manifold data." https://doi.org/10.3150/23-BEJ1685SUPP

Chwialkowski, K., Strathmann, H. and Gretton, A. (2016). A kernel test of goodness of fit. In *JMLR: Workshop and Conference Proceedings*.

Chwialkowski, K.P., Ramdas, A., Sejdinovic, D. and Gretton, A. (2015). Fast two-sample testing with analytic representations of probability measures. In *Advances in Neural Information Processing Systems* 1981–1989.

Coifman, R.R. and Lafon, S. (2006). Diffusion maps. *Appl. Comput. Harmon. Anal.* **21** 5–30. MR2238665 https://doi.org/10.1016/j.acha.2006.04.006

del Barrio, E., Cuesta-Albertos, J.A., Matrán, C. and Rodríguez-Rodríguez, J.M. (1999). Tests of goodness of fit based on the L_2-Wasserstein distance. *Ann. Statist.* **27** 1230–1239. MR1740113 https://doi.org/10.1214/aos/1017938923

do Carmo, M.P. (1992). *Riemannian Geometry*. *Mathematics: Theory & Applications*. Boston, MA: Birkhäuser, Inc. Translated from the second Portuguese edition by Francis Flaherty. MR1138207 https://doi.org/10.1007/978-1-4757-2201-7

Dunson, D.B., Wu, H.-T. and Wu, N. (2021). Spectral convergence of graph Laplacian and heat kernel reconstruction in L^∞ from random samples. *Appl. Comput. Harmon. Anal.* **55** 282–336. MR4279237 https://doi.org/10.1016/j.acha.2021.06.002

Farahmand, A.M. Szepesvári, C. and Audibert, J.-Y. (2007). Manifold-adaptive dimension estimation. In *Proceedings of the 24th International Conference on Machine Learning* 265–272.

Flynn, T. and Yoo, S. (2019). Change detection with the kernel cumulative sum algorithm. In *2019 IEEE 58th Conference on Decision and Control (CDC)*. 6092–6099.

García Trillos, N., Gerlach, M., Hein, M. and Slepčev, D. (2020). Error estimates for spectral convergence of the graph Laplacian on random geometric graphs toward the Laplace-Beltrami operator. *Found. Comput. Math.* **20** 827–887. MR4130541 https://doi.org/10.1007/s10208-019-09436-w

Gretton, A., Fukumizu, K., Harchaoui, Z. and Sriperumbudur, B.K. (2009). A fast, consistent kernel two-sample test. In *In Advances in Neural Information Processing Systems* **22** 673–681. Curran Associates.

Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B. and Smola, A. (2012). A kernel two-sample test. *J. Mach. Learn. Res.* **13** 723–773. MR2913716

Györfi, L. and van der Meulen, E.C. (1991). A consistent goodness of fit test based on the total variation distance. In *Nonparametric Functional Estimation and Related Topics (Spetses, 1990)*. *NATO Adv. Sci. Inst. Ser. C: Math. Phys. Sci.* **335** 631–645. Dordrecht: Kluwer Academic. MR1154355

Hein, M., Audibert, J.-Y. and von Luxburg, U. (2005). From graphs to manifolds—weak and strong pointwise consistency of graph Laplacians. In *Learning Theory*. *Lecture Notes in Computer Science* **3559** 470–485. Berlin: Springer. MR2203281 https://doi.org/10.1007/11503415_32

Higgins, J.J. (2004). *An Introduction to Modern Nonparametric Statistics*. Pacific Grove, CA: Brooks/Cole.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–30. MR0144363

Horváth, L. and Kokoszka, P. (2012). *Inference for Functional Data with Applications*. *Springer Series in Statistics*. New York: Springer. MR2920735 https://doi.org/10.1007/978-1-4614-3655-3

Hotelling, H. (1931). The generalization of student's ratio. *Ann. Math. Stat.* **2** 360–378.

Jitkrittum, W., Kanagawa, H. and Schölkopf, B. (2020). Testing goodness of fit of conditional density models with kernels. In *Conference on Uncertainty in Artificial Intelligence* 221–230. PMLR.

Jitkrittum, W., Szabó, Z., Chwialkowski, K.P. and Gretton, A. (2016). Interpretable distribution features with maximum testing power. In *Advances in Neural Information Processing Systems* 181–189.

Jitkrittum, W., Xu, W., Szabó, Z., Fukumizu, K. and Gretton, A. (2017). A linear-time kernel goodness-of-fit test. In *Advances in Neural Information Processing Systems* 262–271.

Levina, E. and Bickel, P. (2004). Maximum likelihood estimation of intrinsic dimension. *Adv. Neural Inf. Process. Syst.* **17**.

Li, T. and Yuan, M. (2019). On the optimality of Gaussian kernel based nonparametric tests against smooth alternatives. arXiv preprint. Available at arXiv:1909.03302.

Li, C.-L., Chang, W., Cheng, Y., Yang, Y. and Póczos, B. (2017). MMD GAN: Towards deeper understanding of moment matching network. In *Advances in Neural Information Processing Systems* 2203–2213.

Lloyd, J.R. and Ghahramani, Z. (2015). Statistical model criticism using kernel two sample tests. In *Advances in Neural Information Processing Systems*. 829–837.

Lopez-Paz, D. and Oquab, M. (2017). Revisiting classifier two-sample tests. In *International Conference on Learning Representations*.

Massey, F.J. (1951). The Kolmogorov-Smirnov test for goodness of fit. *J. Amer. Statist. Assoc.* **46** 68–78.

Mordohai, P. and Medioni, G. (2010). Dimensionality estimation, manifold learning and function approximation using tensor voting. *J. Mach. Learn. Res.* **11** 411–450. MR2591630

Ozakin, A. and Gray, A.G. (2009). Submanifold density estimation. In *Advances in Neural Information Processing Systems*. 1375–1382.

Pettis, K.W., Bailey, T.A., Jain, A.K. and Dubes, R.C. (1979). An intrinsic dimensionality estimator from near-neighbor information. *IEEE Trans. Pattern Anal. Mach. Intell.* **1** 25–37.

Peyré, G. (2009). Manifold models for signals and images. *Comput. Vis. Image Underst.* **113** 249–260.

Pfanzagl, J. and Sheynin, O. (1996). Studies in the history of probability and statistics. XLIV. A forerunner of the t-distribution. *Biometrika* **83** 891–898. MR1766040 https://doi.org/10.1093/biomet/83.4.891

Pratt, J.W. and Gibbons, J.D. (1981). Kolmogorov-Smirnov two-sample tests. In *Concepts of Nonparametric Theory*. 318–344. New York, NY: Springer.

Ramdas, A., García Trillos, N. and Cuturi, M. (2017). On Wasserstein two-sample testing and related families of nonparametric tests. *Entropy* **19** 47. MR3608466 https://doi.org/10.3390/e19020047

Ramdas, A., Reddi, S.J., Póczos, B., Singh, A. and Wasserman, L. (2015). On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Ren, J., Liu, P.J., Fertig, E., Snoek, J., Poplin, R., Depristo, M., Dillon, J. and Lakshminarayanan, B. (2019). Likelihood ratios for out-of-distribution detection. *Adv. Neural Inf. Process. Syst.* **32**.

Saelens, W., Cannoodt, R., Todorov, H. and Saeys, Y. (2019). A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37** 547–554. https://doi.org/10.1038/s41587-019-0071-9

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. and Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 4510–4520.

Serfling, R.J. (2009). *Approximation Theorems of Mathematical Statistics*. *Wiley Series in Probability and Mathematical Statistics*. New York: Wiley. MR0595165

Shapiro, A., Xie, Y. and Zhang, R. (2021). Goodness-of-fit tests on manifolds. *IEEE Trans. Inf. Theory* **67** 2539–2553. MR4282371 https://doi.org/10.1109/tit.2021.3050469

Singer, A. (2006). From graph to manifold Laplacian: The convergence rate. *Appl. Comput. Harmon. Anal.* **21** 128–134. MR2238670 https://doi.org/10.1016/j.acha.2006.03.004

Sriperumbudur, B.K., Fukumizu, K., Gretton, A., Schölkopf, B. and Lanckriet, G.R.G. (2012). On the empirical estimation of integral probability metrics. *Electron. J. Stat.* **6** 1550–1599. MR2988458 https://doi.org/10.1214/12-EJS722

Sutherland, D.J., Tung, H., Strathmann, H., De, S., Ramdas, A., Smola, A. and Gretton, A. (2017). Generative models and model criticism via optimized maximum mean discrepancy. In *International Conference on Learning Representations*.

Tenenbaum, J.B., de Silva, V. and Langford, J.C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science* **290** 2319–2323.

Van den Berge, K., Roux de Bézieux, H.R., Street, K., Saelens, W., Cannoodt, R., Saeys, Y., Dudoit, S. and Clement, L. (2020). Trajectory-based differential expression analysis for single-cell sequencing data. *Nat. Commun.* **11** 1201. https://doi.org/10.1038/s41467-020-14766-3

van der Maaten, L. (2014). Accelerating t-SNE using tree-based algorithms. *J. Mach. Learn. Res.* **15** 3221–3245. MR3277169

Wynne, G. and Duncan, A.B. (2022). A kernel two-sample test for functional data. *J. Mach. Learn. Res.* **23** 1–51. MR4576658

Xie, Y. and Siegmund, D. (2013). Sequential multi-sensor change-point detection. *Ann. Statist.* **41** 670–692. MR3099117 https://doi.org/10.1214/13-AOS1094

Xie, L. and Xie, Y. (2021). Sequential change detection by optimal weighted ℓ_2 divergence. *IEEE J. Sel. Areas Inf. Theory* **2** 747–761.

Zelnik-Manor, L. and Perona, P. (2005). Self-tuning spectral clustering. In *Advances in Neural Information Processing Systems* 1601–1608.

Zhao, J., Jaffe, A., Li, H., Lindenbaum, O., Sefik, E., Jackson, R., Cheng, X., Flavell, R.A. and Kluger, Y. (2021). Detection of differentially abundant cell subpopulations in scRNA-seq data. *Proc. Natl. Acad. Sci. USA* **118**. https://doi.org/10.1073/pnas.2100293118

Zhu, W., Qiu, Q., Huang, J., Calderbank, R., Sapiro, G. and Daubechies, I. (2018). LDMNet: Low dimensional manifold regularized neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2743–2751.

Received April 2023 and revised October 2023