Electronic Journal of Statistics
Vol. 13 (2019) 5253–5305
ISSN: 1935-7524
https://doi.org/10.1214/19-EJS1648

Global and local two-sample tests via regression

Ilmun Kim, Ann B. Lee, and Jing Lei

Carnegie Mellon University
Department of Statistics and Data Science
5000 Forbes Avenue
Pittsburgh, PA 15213
e-mail: ilmunk@stat.cmu.edu
e-mail: annlee@stat.cmu.edu
e-mail: jinglei@stat.cmu.edu

Abstract: Two-sample testing is a fundamental problem in statistics. Despite its long history, there has been renewed interest in this problem with the advent of high-dimensional and complex data. Specifically, in the machine learning literature, there have been recent methodological developments such as classification accuracy tests. The goal of this work is to present a regression approach to comparing multivariate distributions of complex data. Depending on the chosen regression model, our framework can efficiently handle different types of variables and various structures in the data, with competitive power under many practical scenarios. Whereas previous work has been largely limited to global tests which conceal much of the local information, our approach naturally leads to a local two-sample testing framework in which we identify local differences between multivariate distributions with statistical confidence. We demonstrate the efficacy of our approach both theoretically and empirically, under some well-known parametric and nonparametric regression methods. Our proposed methods are applied to simulated data as well as a challenging astronomy data set to assess their practical usefulness.

MSC 2010 subject classifications: Primary 62H15, 62G10; secondary 62G20.
Keywords and phrases: Galaxy morphology, intrinsic dimension, kernel regression, nearest neighbor regression, permutation test, random forests.

Received May 2019.

Contents

1. Introduction

Given two distributions P_0 and P_1 on $\mathbb{R}^D$, the global two-sample problem is concerned with testing $H_0 : P_0 = P_1$ versus $H_1 : P_0 \neq P_1$, based on independent random samples from each distribution. This fundamental problem has a long history in statistics and has been well-studied in a classical setting (see, e.g., Thas, 2010). Recently, however, there has been renewed interest in this field as modern data we encounter have become more complex and diverse. Traditional approaches, which focus on low-dimensional and Euclidean data, often fail or are not easily generalizable to high-dimensional and non-Euclidean data. Additionally, some recent developments in high-dimensional two-sample testing are limited to simple alternatives such as location and scale differences (see, Hu and Bai, 2016, for a recent review). In this context, there is a need to develop a new tool for the two-sample problem that can efficiently handle complex data and can detect differences beyond location and scale alternatives.

When the null hypothesis of the global two-sample test is rejected, it is often valuable (for e.g. scientific discovery, calibration of simulation models, and so on) to further explore *how* the two distributions are different. Specifically, as a follow-up study to the global test, one might wish to identify locally significant regions where the two distributions differ. This topic, which we refer to as the *local two-sample problem*, has been studied by Duong (2013) who uses kernel density estimators to identify local differences between two density functions. However, the kernel density approach may perform poorly when distributions are not in a low-dimensional Euclidean space, and hence another tool is needed for more challenging settings.

The goal of this work is to develop a general framework for both global and local two-sample problems that overcomes the aforementioned challenges. Specifically, we aim to design a two-sample test that can efficiently handle different types of variables (e.g. mixed data types) and various structure (e.g. manifold, irrelevant covariates) in the data. Consequently, the resulting test can have substantial power for a variety of challenging alternatives. We achieve our goal by connecting the two-sample problem to a regression problem as follows. Let f_0 and f_1 be density functions of P_0 and P_1 with respect to a common dominating measure. We view f_0 and f_1 as conditional densities $f(x|Y=0)$ and $f(x|Y=1)$ by introducing an indicator random variable $Y \in \{0,1\}$. Then by Bayes' theorem, the hypothesis $H_0 : f_0(x) = f_1(x)$ for all $x \in S = \{x \in \mathbb{R}^D : f_0(x)+f_1(x) > 0\}$ can be verified to be equivalent to the hypothesis that involves the regression function:

$$H_0 : \ \mathbb{P}(Y=1|X=x) = \mathbb{P}(Y=1), \quad \text{for all } x \in S. \tag{1}$$

We state the corresponding global and local alternative hypotheses as

$$H_1 : \ \mathbb{P}(Y=1|X=x) \neq \mathbb{P}(Y=1), \quad \text{for some } x \in S, \text{ and}$$

$$H_1(x) : \ \mathbb{P}(Y=1|X=x) \neq \mathbb{P}(Y=1), \quad \text{at fixed } x \in S,$$

respectively.

Motivated by the above reformulation, we propose a testing procedure that measures an empirical distance between the regression function $\mathbb{P}(Y=1|X=x)$ and the class probability $\mathbb{P}(Y=1)$. We refer to this approach as *the regression test*. Depending on the choice of regression method, the regression test can adapt to nontraditional data settings. As we shall see, the power of the test is closely related to the mean square error of the chosen regression estimator. In addition, by choosing a nonparametric regression method, the global regression test can be sensitive to general alternatives beyond location and scale differences. We will demonstrate the benefits of the regression test with both theoretical and empirical results.

1.1. Motivating example

We motivate our approach by comparing multivariate distributions of galaxy morphologies, but the proposed framework benefit other areas of science and technology as well (involving, e.g., outlier detection, calibration of simulation models, and comparison of cases and controls). A galaxy's morphology is the organization of a galaxy's light, as projected into our line of sight and observed at a particular wavelength as a pixelated image. Morphological studies are key to understanding the evolutionary history of galaxies and to constraining theories of the Universe; see e.g. Conselice (2014) for a review. So far astronomers have only been able to study one or two morphological statistics (or projections of these) at a time instead of an entire ensemble. The reason is a lack of tools for effectively comparing and jointly analyzing multivariate or high-dimensional

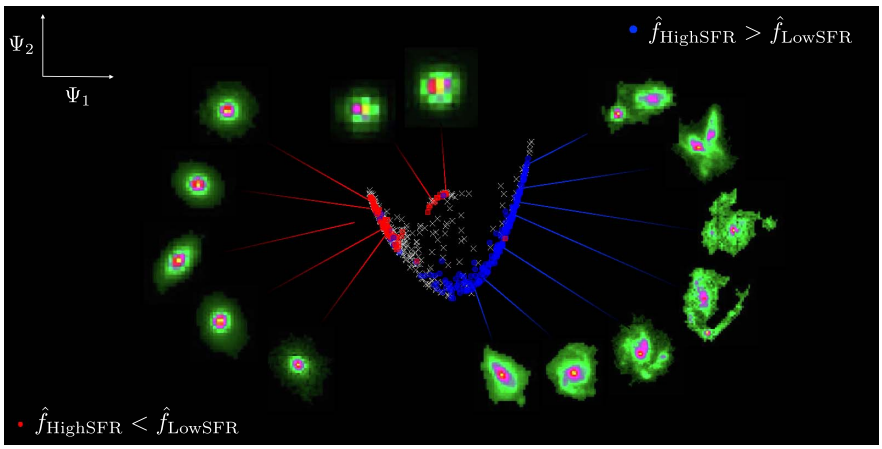

FIG 1. *Result of local two-sample test of differences between high- and low-SFR galaxies in a seven-dimensional morphology space. The red squares indicate regions where the density of low-star-forming galaxies are significantly higher, and the blue circles indicate regions in morphology space that are dominated by high-star-forming galaxies; the gray crosses represent insignificant test points. The galaxies are embedded in a two-dimensional diffusion space for visualization purposes only (see Appendix B for details);* Ψ_1 *and* Ψ_2 *here denote the first two coordinates.*

data in their native spaces. A global hypothesis test with a binary reject yes/no answer is also not informative enough to explain how two distributions are different in a multivariate feature space.

We illustrate the efficacy of the proposed global and local testing framework on the morphology statistics of two galaxy populations with high and low star-formation rate (SFR), respectively. The challenge here is not only that the problem involves multivariate data, but also that some of the morphological statistics are mixed discrete and continuous type with heavy outliers. We efficiently handle this issue by building on the success of random forests regression. The visualized local two-sample result is shown in Figure 1 and the details of the analysis can be found in Section 6.

1.2. Related work

In recent years, several attempts have been made to connect binary classification with two-sample testing. The main idea of this approach is to check whether the accuracy of a binary classifier is better than chance level and reject the null if the difference is significant. Such an approach, referred to as an accuracy or classification test, was conceptualized by Friedman (2003) and has since been investigated by several authors (Ojala and Garriga, 2010; Olivetti et al., 2015; Kim et al., 2019; Rosenblatt et al., 2016; Gagnon-Bartsch and Shem-Tov, 2016; Lopez-Paz and Oquab, 2016; Hediger et al., 2019). In the same manner as our regression framework, a key strength of the accuracy test is that it offers a flex-

ible way for the two-sample problem as it can utilize any existing classification procedure in the literature. However, the classification accuracy framework is not easily converted to a local two-sample test. In addition, many classifiers are estimated by dichotomizing regression estimators and the discrete nature of such classifiers may result in a less powerful test (see Section 5.2 and other simulation results).

For the global two-sample test, our framework can be viewed as an instance of goodness-of-fit testing for regression models (e.g. González-Manteiga and Crujeiras, 2013, for a review). There is a substantive literature on this topic including Hardle and Mammen (1993), Weihrather (1993), González-Manteiga and Cao (1993), Zheng (1996), Zhang and Dette (2004), Hart (2013) and among others. This line of work typically concentrates on comparing differences between parametric (e.g. linear regression) and nonparametric (e.g. kernel regression) fits from an asymptotic point of view. For example, Hardle and Mammen (1993) consider the squared deviation between a parametric regression estimator and a kernel estimator. They show that their test statistic converges to a normal distribution under the null hypothesis and justify the use of the wild bootstrap procedure. However, this type of asymptotic approach is challenging to analyze beyond kernel-type estimators and often requires strong technical assumptions. In contrast, our framework is designed to compare any type of regression estimators with a specific constant fit by building upon the permutation principle. Hence the resulting test is valid in any finite sample sizes. Moreover we present a unified framework of studying the power of the regression test by taking advantage of existing results on the estimation error.

For the local two-sample test, our approach has similarities to independent work by Cazáis and Lhéritier (2015) who estimate the Kullback-Leibler divergence between $\mathbb{P}(Y = 1|X = x)$ and $\mathbb{P}(Y = 1)$. Our procedure, however, identifies locally significant areas with statistical confidence whereas Cazáis and Lhéritier (2015) graphically decide a threshold for the significance.

1.3. Overview of this paper

We outline the paper as follows: In Section 2, we introduce the proposed metrics, test statistics and algorithms for the global and local regression tests. In Section 3, we study theoretical properties of the global regression test. We begin by considering a simple scenario where two populations only differ in their means in Section 3.1. In this scenario, we show that the regression test based on Fisher's linear discriminant analysis (LDA) achieves the same local optimality as the Hotelling's T^2 test. Moving on to general regression settings in Section 3.2, we establish a connection between the testing error of the global regression test and the mean integrated square error (MISE) of the regression estimator. In Section 4, we turn to the local two-sample problem and investigate general properties of the local regression tests. In Section 4.1, we describe the testing error of the local regression test in terms of the mean square error (MSE) of the regression estimator. We further establish an optimality of the

local regression tests over the Lipschitz class from a minimax point of view in Section 4.2. When data have intrinsic dimension, we show that the performance of the local regression tests based on kNN or kernel regression only depends on intrinsic dimension in Section 4.3. Section 4.4 studies the limiting distribution of the local permutation statistic to avoid a high computational cost from permutations for large sample size. In Section 5, simulation studies are provided to illustrate finite sample performance of the global and local regression tests. In Section 6, we apply the proposed approach to a problem in astronomy and demonstrate its efficacy. All the proofs are deferred to Appendix A.

Notation Throughout this paper, we denote the class probabilities $\mathbb{P}(Y = 0)$ and $\mathbb{P}(Y = 1)$ by π_0 and π_1, respectively, and write the joint distribution of (X, Y) by $\pi_0[P_0 \times \delta_0] + \pi_1[P_1 \times \delta_1]$ where δ_k denotes the point mass at k for $k = 0, 1$. We denote the corresponding conditional probability $\mathbb{P}(Y = 1|X = x)$ by $m(x)$, which can be explicitly written as

$$m(x) = \frac{\pi_1 f_1(x)}{\pi_1 f_1(x) + \pi_0 f_0(x)}.$$

We use $P_X(\cdot)$ to denote the marginal probability measure of X and $||Z||_2$ denotes the Euclidean norm of a vector $Z \in \mathbb{R}^D$. The symbols $\xrightarrow{p}$ and $\xrightarrow{d}$ stand for convergence in probability and in distribution, respectively. We use $a_n \lesssim b_n$ if there exists $C > 0$ such that $a_n \leq Cb_n$ for all n. Similarly, $a_n \asymp b_n$ if there exist constants $C, C' > 0$ such that $C \leq |a_n/b_n| \leq C'$ for all n. As convention, the acronym *i.i.d.* is used to represent independent and identically distributed.

2. Framework

2.1. Metrics

A common metric for comparing two distributions is the difference between two density functions $f_0(x)$ and $f_1(x)$; this metric has been used for global and local two-sample testing by Anderson et al. (1994) and Duong (2013). Another natural metric, suggested for global two-sample testing by Keziou and Leoni-Aubin (2005), Fokianos (2008) and Sugiyama et al. (2011), is the density ratio $f_1(x)/f_0(x)$. Although both the density difference and density ratio metrics are intuitive, there are several weaknesses associated with each of them. For example, the estimation of a density difference is largely limited to kernel density estimators, which are sensitive to the curse of dimensionality. The density ratio, on the other hand, could potentially be estimated using various regression methods thanks to the following reformulation:

$$\frac{f_1(x)}{f_0(x)} = \frac{\pi_0}{\pi_1} \frac{m(x)}{1 - m(x)},$$

(see, e.g., Qin and Zhang, 1997). The main weakness of the ratio approach, however, is that the ratio is highly sensitive to the tail behavior of distributions,

and it is not even well defined when $m(x) = 1$. To overcome these limitations, we propose an alternative approach which instead compares the regression function with the class probability. More specifically, we consider

$$\mathcal{T}_{global} = \int_S \{m(x) - \pi_1\}^2 dP_X(x), \quad \mathcal{T}_{local}(x) = \{m(x) - \pi_1\}^2 \tag{2}$$

as global and local measures of the discrepancy between two distributions where we assume that π_1 is a fixed constant within $0 < \pi_1 < 1$ throughout this paper. By construction, both $\mathcal{T}_{global}$ and $\mathcal{T}_{local}(x)$ are bounded between zero and one. More importantly, we can take advantage of numerous existing regression methods (see, e.g., Friedman et al., 2009, for popular methods and descriptions) when estimating $m(x)$. Hence, our approach maintains the flexibility of the density ratio approach while avoiding the problem of ill-defined quantities.

2.2. Test statistics and algorithms

Suppose we observe n pairs of samples $\{(X_i, Y_i)\}_{i=1}^n$, where $X_i \in \mathbb{R}^D$ and $Y_i \in \{0, 1\}$. Let $\widehat{m}(x)$ be an estimate of $m(x)$ based on the samples, and $\widehat{\pi}_1 = \frac{1}{n}\sum_{i=1}^n I(Y_i = 1)$. Then by plugging these statistics into (2), we define our global and local test statistics as

$$\widehat{\mathcal{T}}_{global} = \frac{1}{n}\sum_{i=1}^{n} \{\widehat{m}(X_i) - \widehat{\pi}_1\}^2, \quad \widehat{\mathcal{T}}_{local}(x) = \{\widehat{m}(x) - \widehat{\pi}_1\}^2. \tag{3}$$

The null distributions of the proposed test statistics are typically unknown, and they depend on the choice of regression method as well as the distribution of the data. Hence, to keep our framework as general as possible, we use a permutation procedure to set a critical value that yields a valid level α test for any given regression estimator under any sampling scheme given in Section 2.3. The proposed permutation framework for global and local two-sample testing are summarized in Algorithm 1 and Algorithm 2, respectively.

2.3. Sampling schemes

In the two-sample problem, there are two common sampling schemes for obtaining the paired data set $\{(X_i, Y_i)\}_{i=1}^n$, namely i) *i.i.d. sampling* and ii) *separate sampling* defined as follows:

- **i.i.d. sampling.** Under i.i.d. sampling, we observe n pairs of i.i.d. samples $\{(X_i, Y_i)\}_{i=1}^n$ from the joint distribution $\pi_1[P_1 \times \delta_1] + \pi_0[P_0 \times \delta_0]$. Here we note that n is fixed in advance. Then $n_1 = \sum_{i=1}^n I(Y_i = 1)$ and $n_0 = n - n_1$ are Binomial(n, π_1) and Binomial(n, π_0), respectively. This setting is common in applications of supervised learning where the goal is to build a model that can successfully predict the class label Y given the feature vector X (e.g. Friedman et al., 2009). Our goal, on the other hand, is to test whether the two distributions P_0 and P_1 are the same or not by leveraging existing methods in the regression literature.

Algorithm 1: Global Two-Sample Testing via Permutations

Require: samples $\{X_i, Y_i\}_{i=1}^n$, number of permutations B, significance level α, a regression method.

(1) Calculate the global test statistic $\widehat{\mathcal{T}}_{global}$.

(2) Randomly permute $\{Y_1, \ldots, Y_n\}$. Calculate the test statistic using the permuted data.

(3) Repeat the previous step B times to obtain $\{\widehat{\mathcal{T}}_{global}^{(1)}, \ldots, \widehat{\mathcal{T}}_{global}^{(B)}\}$.

(4) Approximate the permutation p-value by

$$p = \frac{1}{B+1}\left(1 + \sum_{b=1}^{B} I(\widehat{\mathcal{T}}_{global}^{(b)} > \widehat{\mathcal{T}}_{global})\right).$$

(5) Reject the null hypothesis when $p < \alpha$. Otherwise, accept the null hypothesis.

Algorithm 2: Local Two-Sample Testing via Permutations

Require: samples $\{X_i, Y_i\}_{i=1}^n$, test points $\{x_j\}_{j=1}^k$, number of permutations B, significance level α, a multiple testing procedure, a regression method.

(1) Calculate the test statistic $\widehat{\mathcal{T}}_{local}(x_j)$ at the k test points.

(2) Randomly permute $\{Y_1, \ldots, Y_n\}$. Calculate the test statistic using the permuted data.

(3) Repeat the previous step B times to obtain $\{\widehat{\mathcal{T}}_{local}^{(1)}(x_j)\}_{j=1}^k, \ldots, \{\widehat{\mathcal{T}}_{local}^{(B)}(x_j)\}_{j=1}^k$.

(4) Approximate the permutation p-value at each test point x_j by

$$p_j = \frac{1}{B+1}\left(1 + \sum_{b=1}^{B} I(\widehat{\mathcal{T}}_{local}^{(b)}(x_j) > \widehat{\mathcal{T}}_{local}(x_j))\right).$$

(5) Apply a multiple testing procedure for controlling the FWER or the FDR at α level.

(6) Return the significant local test points.

- **Separate sampling.** In the case of separate sampling, n_0 and n_1 are predetermined and they are not random. We then observe n_0 and n_1 independent sample points from P_0 and P_1 separately, which provides the data set $\{(X_i, Y_i)\}_{i=1}^n$ where $Y_i = 1$ if X_i was drawn from P_1 and $Y_i = 0$ otherwise.

We can link the separate sampling to the i.i.d. sampling scheme by randomly ordering the (X_i, Y_i) pairs, so that the data points are exchangeable and for each $i \in \{1, \ldots, n\}$, the conditional distribution of Y_i given $X_i = x$ is $m(x) = \pi_1 f_1(x)/\{\pi_1 f_1(x) + \pi_0 f_0(x)\}$ where the class probability is given by $\pi_1 = n_1/n$. Therefore, although the joint distributions of $\{(X_i, Y_i)\}_{i=1}^n$ are different under i.i.d. and separate sampling schemes, they share the same regression function.

Remark 2.1. *These two sampling schemes are also known as* prospective sampling *and* retrospective (or case-control) sampling*, respectively, and their relationships have been studied in different contexts. For example, it has been shown that the logistic slope estimates have similar behaviors under both sampling schemes (see, e.g. Anderson, 1972; Prentice and Pyke, 1979; Wang and Carroll, 1993, 1999; Bunea and Barbu, 2009). This result has been extended to general regression models by Scott and Wild (2001).*

3. Global two-sample tests via regression

The choice of regression method in our framework will ultimately decide whether we achieve competitive statistical power. In Section 3.1, we illustrate the point that the global regression test can be optimal if we choose a suitable regression method. For this theoretical purpose, we focus on the regression test based on Fisher's LDA and show its optimality. In Section 3.2, we turn our attention to more general regression settings and characterize the testing error of the global regression test in terms of the mean integrated square error (MISE) of the regression estimator.

3.1. Fisher's linear discriminant analysis

In this section, we consider a simple scenario of two sample normal mean to highlight the difference between our approach and the classification accuracy approach. In particular, we prove that the regression test based on Fisher's LDA achieves the same local power as Hotelling's T^2 test. This result has significance given that i) Hotelling's test is optimal under the considered scenario and ii) the classification accuracy test based on Fisher's LDA is usually underpowered (Kim et al., 2019; Rosenblatt et al., 2016). To facilitate comparison with the previous results, which are established under separate sampling, we also consider the case where n_0 and n_1 are predetermined throughout this subsection.

Suppose we observe $\{X_{i,0}\}_{i=1}^{n_0} \overset{i.i.d.}{\sim} N(\mu_0, \Sigma)$ and independently $\{X_{i,1}\}_{i=1}^{n_1} \overset{i.i.d.}{\sim} N(\mu_1, \Sigma)$. We denote the pooled samples by $\{X_i\}_{i=1}^{n} = \{X_{i,0}\}_{i=1}^{n_0} \cup \{X_{i,1}\}_{i=1}^{n_1}$ where $n = n_0 + n_1$. The two-sample problem then becomes the problem of testing for mean differences as

$$H_0 : \mu_0 = \mu_1 \quad \text{versus} \quad H_1 : \mu_0 \neq \mu_1. \tag{4}$$

For this particular problem, Fisher's LDA is a natural choice for regression, assuming normality and equal class covariances. Let $\widehat{\mu}_i$ be the sample mean vector for each group, $\mathcal{S}$ be the covariance matrix of the combined samples, i.e. $\mathcal{S} = n^{-1}\sum_{i=1}^{n}(X_i - \widehat{\mu})(X_i - \widehat{\mu})^\top$ where $\widehat{\mu} = n^{-1}\sum_{i=1}^{n} X_i$. Then, by putting $\pi_1 = n_1/n$, the regression estimator based on Fisher's LDA is given by

$$\widehat{m}_{\mathrm{LDA}}(x) \tag{5}$$
$$= \frac{\pi_1 \exp\left\{-\frac{1}{2}(x-\widehat{\mu}_1)^\top \mathcal{S}^{-1}(x-\widehat{\mu}_1)\right\}}{\pi_0 \exp\left\{-\frac{1}{2}(x-\widehat{\mu}_0)^\top \mathcal{S}^{-1}(x-\widehat{\mu}_0)\right\} + \pi_1 \exp\left\{-\frac{1}{2}(x-\widehat{\mu}_1)^\top \mathcal{S}^{-1}(x-\widehat{\mu}_1)\right\}}.$$

One of the most popular test statistics for testing (4) is Hotelling's T^2 statistic, which yields optimal power for the normal means problem (see, e.g. Anderson, 2003). For the two-sample problem, Hotelling's T^2 statistic is defined by

$$T^2_{\text{Hotelling}} = \frac{n_0 n_1}{n_0 + n_1}(\widehat{\mu}_0 - \widehat{\mu}_1)^\top \mathcal{S}_p^{-1}(\widehat{\mu}_0 - \widehat{\mu}_1),$$

where $\mathcal{S}_p$ is the pooled covariance matrix, that is

$$\mathcal{S}_p = \frac{1}{n_0 + n_1 - 2}\left(\sum_{i=1}^{n_0}(X_{i,0} - \widehat{\mu}_0)(X_{i,0} - \widehat{\mu}_0)^\top + \sum_{i=1}^{n_1}(X_{i,1} - \widehat{\mu}_1)(X_{i,1} - \widehat{\mu}_1)^\top\right).$$

On the other hand, the regression test statistic based on Fisher's LDA is given by

$$\widehat{\mathcal{T}}_{\text{LDA}} = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{m}_{\text{LDA}}(X_i) - \pi_1\right)^2.$$

The next theorem provides a connection between the seemingly unrelated $\widehat{\mathcal{T}}_{\text{LDA}}$ and $T^2_{\text{Hotelling}}$ statistics. Specifically, it shows that $n\pi_0^{-1}\pi_1^{-1}\widehat{\mathcal{T}}_{\text{LDA}}$ is asymptotically identical to Hotelling's T^2 statistic under the null. It is also worth pointing out that the theorem still holds without the normality assumption.

Theorem 3.1. *Let $\{X_{i,0}\}_{i=1}^{n_0}$ and $\{X_{i,1}\}_{i=1}^{n_1}$ be random samples under separate sampling from two multivariate distribution with the mean vectors μ_0 and μ_1, respectively, and the same covariance matrix Σ. Assume the pooled samples are mutually independent and the third moments of $X_{1,0}$ and $X_{1,1}$ are finite. Suppose that $\mathcal{S}_p$ and $\mathcal{S}$ satisfy $\mathcal{S}_p^{-1} = \Sigma^{-1}(1 + o_P(1))$ and $\mathcal{S}^{-1} = \Sigma^{-1}(1 + o_P(1))$. Then, under $H_0 : \mu_0 = \mu_1$, it holds that*

$$n\widehat{\mathcal{T}}_{\text{LDA}} = n\pi_0^2\pi_1^2(\widehat{\mu}_0 - \widehat{\mu}_1)^\top \mathcal{S}_p^{-1}(\widehat{\mu}_0 - \widehat{\mu}_1) + o_P(1). \tag{6}$$

Therefore,

$$n\pi_0^{-1}\pi_1^{-1}\widehat{\mathcal{T}}_{\text{LDA}} = T^2_{\text{Hotelling}} + o_P(1) \xrightarrow{d} \chi^2_D,$$

where χ^2_D is the chi-squared distribution with D degrees of freedom.

Let us now turn to the alternative hypothesis. To begin with, we consider a family of probability functions that satisfy the following smoothness condition.

Definition 3.1 (Definition 12.2.1 of Lehmann and Romano (2006))**.** *Let $\{P_\mu, \mu \in \Omega\}$ be a parametric model where Ω is an open subset of $\mathbb{R}^D$, and let $f_\mu(x) = dP_\mu(x)/d\nu(x)$ be the density function with respect to Lebesgue measure ν. The family $\{P_\mu, \mu \in \Omega\}$ is quadratic mean differentiable (q.m.d.) at μ_0 if there exists a vector of real-valued functions $\eta(\cdot, \mu_0) = (\eta_1(\cdot, \mu_0), \cdots, \eta_D(\cdot, \mu_0))^\top$ such that*

$$\int_{\mathbb{R}^D}\left[\sqrt{f_{\mu_0+h}(x)} - \sqrt{f_{\mu_0}(x)} - \langle\eta(x, \mu_0), h\rangle\right]^2 d\nu(x) = o(||h||_2^2) \tag{7}$$

as $||h||_2 \to 0$.

Such q.m.d. families include fairly large parametric models such as exponential families in natural form. For our purpose, we focus on location q.m.d. families, denoted by $\{\mathbb{P}_\mu, \mu \in \Omega\}$. Specifically, $\mathbb{P}_\mu$ is a member of $\{\mathbb{P}_\mu, \mu \in \Omega\}$ if its density satisfies $f_\mu(x) = f(x-\mu)$ for which $f(x)$ has zero mean and covariance matrix Σ. Next, for given $\mathbb{P}_{\mu_0}$ and $\mathbb{P}_{\mu_1}$ from $\{\mathbb{P}_\mu, \mu \in \Omega\}$, let us consider the local alternative

$$H_{1,n} : \mu_1 - \mu_0 = h/\sqrt{n}, \tag{8}$$

where $h = (h_1, \ldots, h_D)^\top$. Then, under $H_{1,n}$, $\widehat{\mathcal{T}}_{\mathrm{LDA}}$ has asymptotic behavior as follows.

Theorem 3.2. *Suppose under separate sampling that* $\{X_{i,0}\}_{i=1}^{n_0} \overset{i.i.d.}{\sim} \mathbb{P}_{\mu_0}$ *and independently* $\{X_{i,1}\}_{i=1}^{n_1} \overset{i.i.d.}{\sim} \mathbb{P}_{\mu_1}$ *where* $\mathbb{P}_{\mu_i}$ *is a member of the location q.m.d. family with the same covariance matrix* Σ *and finite third moments. Suppose that* $\mathcal{S}_p$ *and* $\mathcal{S}$ *satisfy* $\mathcal{S}_p^{-1} = \Sigma^{-1}(1 + o_P(1))$ *and* $\mathcal{S}^{-1} = \Sigma^{-1}(1 + o_P(1))$. *Under the sequence of local alternatives given in (8), we have*

$$n\pi_0^{-1}\pi_1^{-1}\widehat{\mathcal{T}}_{\mathrm{LDA}} = T^2_{\mathrm{Hotelling}} + o_P(1) \xrightarrow{d} \chi^2_D(\lambda),$$

where $\chi^2_D(\lambda)$ *denotes a noncentral chi-square distribution with* D *degrees of freedom and the noncentral parameter*

$$\lambda = \pi_0\pi_1 h^\top \Sigma^{-1} h.$$

The results from Theorem 3.1 and Theorem 3.2 imply that our regression test based on $\widehat{\mathcal{T}}_{\mathrm{LDA}}$ has the same asymptotic local power as Hotelling's T^2 test. As a result, the regression test based on $\widehat{\mathcal{T}}_{\mathrm{LDA}}$ is asymptotically optimal against the local alternatives as Hotelling's T^2 test.

To illustrate the main point of this section, we compare the performance of $\widehat{\mathcal{T}}_{\mathrm{LDA}}$ with Hotelling's T^2 test through Monte Carlo simulations. We randomly generate $n_0 = n_1 = 100$ samples from $N((0,\ldots,0)^\top, I_D)$ and $N((\mu,\ldots,\mu)^\top, I_D)$, respectively and set $\mu^2 = 0.05$ for $D = 5$ and $\mu^2 = 0.01$ for $D = 20$. We also consider two versions of the accuracy-based tests via Fisher's LDA: the in-sample (re-substitution) accuracy and the two-fold cross-validated accuracy. To calculate the cross-validated accuracy, we use the balanced sample splitting scheme in which the first part of data is used to train the LDA, and the second part is used to estimate the accuracy of the classifier (see, Definition 1 and 2 of Rosenblatt et al., 2016, for more details). To make a fair comparison, the critical values of the given tests were all decided by the permutation procedure. As shown in Figure 2, the regression test based on $\widehat{\mathcal{T}}_{\mathrm{LDA}}$ has comparable power to Hotelling's T^2 test that coincides with our theory. On the other hand, the accuracy tests have less power than Hotelling's T^2 test.

3.2. The MISE and testing error for global regression

We now turn to more general regression settings and investigate general properties of the global regression test in both separate and i.i.d. sampling cases. Let

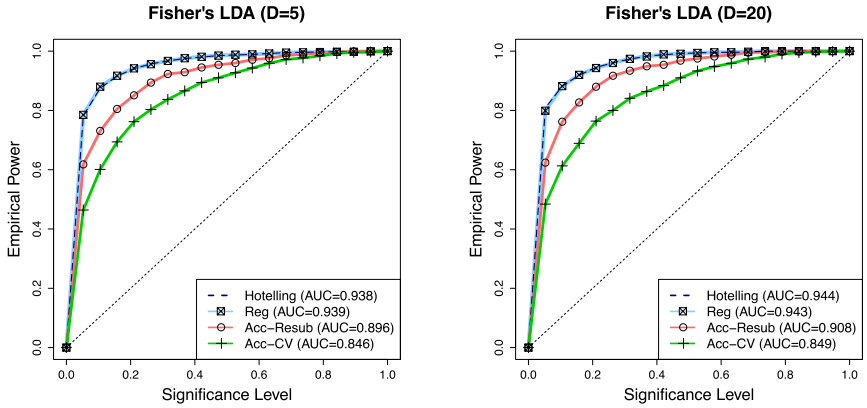

FIG 2. *Power comparisons between Hotelling's T^2 (Hotelling), $\widehat{T}_{LDA}$ (Reg), the in-sample accuracy (Acc-Resub), and the cross-validated accuracy (Acc-CV) via Fisher's LDA.*

$\mathcal{M}$ be a certain class of regression $m(x) : S \subseteq \mathbb{R}^D \mapsto [0,1]$ containing constant functions. Suppose that we have a regression estimator $\widehat{m}(x)$ that has the mean integrated square error as

$$\sup_{m \in \mathcal{M}} \mathbb{E} \int_S \left(\widehat{m}(x) - m(x)\right)^2 dP_X(x) \leq C_0 \delta_n, \tag{9}$$

where C_0 is a positive constant and $\delta_n = o(1)$. In the case of i.i.d. sampling, we further assume $\delta_n \geq n^{-1}$, which is typical for nonparametric regression estimators. Our main interest here is in employing the above MISE to characterize the testing error of the global regression test. Note that the plug-in global statistic in (3) is typically a biased estimator of the MISE and the bias differs from case to case. To simplify our analysis, we consider sample splitting where the half of data is used to estimate the regression function and the other is used to evaluate the empirical squared error. In detail, given samples $(X_1, Y_1), \ldots, (X_{2n}, Y_{2n})$, the regression test statistic based on (random) sample splitting is defined by

$$\widehat{T}'_{global} = \frac{1}{n} \sum_{i=n+1}^{2n} \left(\widehat{m}(X_i) - \widehat{\pi}_1\right)^2, \tag{10}$$

where $\widehat{m}(\cdot)$ and $\widehat{\pi}_1$ are calculated based on the first half of the data $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$. In the case of separate sampling, we assume a random ordering in the entire data set and similarly split it into two parts but with the additional restriction that class probabilities are the same in both parts. Based on $\widehat{T}'_{global}$, we argue that for sufficiently large $C_1 > 0$ and n, the testing error of the global regression test can be arbitrarily small against the class of global alternatives given by

$$\mathcal{M}(C_1 \delta_n) = \left\{ m \in \mathcal{M} : \int_S \left(m(x) - \pi_1\right)^2 dP_X(x) \geq C_1 \delta_n \right\}.$$

Note that since π_1 is assumed to be fixed, the regression function $m(x)$ is completely determined by f_0 and f_1. Thus in the following theorem and hereafter, we use the notation $f_0, f_1 \in \mathcal{M}$ to represent $m(x) = \pi_1 f_1(x)/\{\pi_0 f_0(x)+\pi_1 f_1(x)\} \in \mathcal{M}$. Similarly, we write $f_0, f_1 \in \mathcal{M}_0$ to signify that $\pi_1 f_1(x)/\{\pi_0 f_0(x)+\pi_1 f_1(x)\} = \pi_1$ for all $x \in S$. With this notation in hand, we state the main theorem of this subsection.

Theorem 3.3. *Consider the case of i.i.d. sampling or separate sampling. In each case, suppose that we have a regression estimator $\widehat{m}(\cdot)$ satisfying (9). Let t_α be the upper α quantile of the permutation distribution of $\widehat{T}'_{global}$ based on $\widehat{m}(\cdot)$ where we permute the first half of labels. For fixed $\alpha \in (0,1)$ and $\beta \in (0, 1-\alpha)$, we assume that there exists a positive constant $C'_{0,\alpha}$ such that $\sup_{f_0,f_1\in\mathcal{M}} \mathbb{P}_{f_0,f_1}(t_\alpha < C'_{0,\alpha}\delta_n) \geq 1-\beta/2$. Then there exist positive constants C_1 and N depending on $C_0, C'_{0,\alpha}, \alpha, \beta$ such that*

- *Type I error:* $\sup_{f_0,f_1\in\mathcal{M}_0} \mathbb{P}_{f_0,f_1}\left(\widehat{T}'_{global} > t_\alpha\right) \leq \alpha$ *and*
- *Type II error:* $\sup_{n\geq N} \sup_{f_0,f_1\in\mathcal{M}(C_1\delta_n)} \mathbb{P}_{f_0,f_1}\left(\widehat{T}'_{global} \leq t_\alpha\right) \leq \beta.$

Theorem 3.3 uses the assumption that the permutation critical value of the regression test is uniformly bounded by δ_n (up to some constant factor) with high probability. We end this subsection with a class of regression estimators, which satisfy this assumption. Let us consider a class of regression estimators with the following representation:

$$\widehat{m}(x) = \sum_{i=1}^{n} w_i(x) Y_i,$$

where $w_i(x) \geq 0$ and $\sum_{i=1}^n w_i(x) = 1$ for all x. In addition, we assume that $w_i(x)$ is a function of $\{X_1, \ldots, X_n\}$ but not $\{Y_1, \ldots, Y_n\}$. This class of estimators, often called linear smoothers, contains many popular regression methods such as k-nearest neighbor (kNN) regression, kernel regression and local polynomial regression. Focusing on linear smoothers, we provide the following corollary.

Corollary 3.1. *Consider the case of i.i.d. sampling or separate sampling. In each case, let $\widehat{T}'_{global}$ be the global regression test statistic in (10) based on a linear smoother $\widehat{m}(\cdot)$ with the property in (9). Let t_α be the upper α quantile of the permutation distribution of $\widehat{T}'_{global}$ where we permute the first half of labels. Then for fixed $\alpha \in (0,1)$ and $\beta \in (0, 1-\alpha)$, there exist positive constants C_1 and N depending on C_0, α, β such that*

- *Type I error:* $\sup_{f_0,f_1\in\mathcal{M}_0} \mathbb{P}_{f_0,f_1}\left(\widehat{T}'_{global} > t_\alpha\right) \leq \alpha$ *and*
- *Type II error:* $\sup_{n\geq N} \sup_{f_0,f_1\in\mathcal{M}(C_1\delta_n)} \mathbb{P}_{f_0,f_1}\left(\widehat{T}'_{global} \leq t_\alpha\right) \leq \beta.$

3.3. Examples

In the case of i.i.d. sampling, the convergence rate δ_n of commonly used regression estimators have been well-established and these results can be directly used to study the testing error of the global regression test. We list several known results here. More examples can be found in Györfi et al. (2002), Tsybakov (2009) and Devroye et al. (2013).

- **kNN regression.** When $\mathcal{M}$ is a class of Lipschitz continuous functions, the convergence rate of kNN estimators satisfies $\delta_n = n^{-2/(2+D)}$ (Györfi et al., 2002). This can be generalized to a Hölder space with smooth parameter β in which the rate becomes $\delta_n = n^{-2\beta/(2\beta+D)}$ (Györfi et al., 2002; Ayano, 2012) for $0 < \beta \leq 1.5$. Furthermore, Kpotufe (2011) shows that kNN estimators are adaptive to the intrinsic dimension $d \ll D$ under appropriate conditions. In this case, the convergence rate becomes much faster as $\delta_n = n^{-2/(2+d)} \ll n^{-2/(2+D)}$.
- **Kernel regression.** Kernel regression estimators also achieve the converge rate as $\delta_n = n^{-2/(2+D)}$ for Lipschitz continuous functions and more generally as $\delta_n = n^{-2\beta/(2\beta+D)}$ for a Hölder space with smooth parameter $0 < \beta \leq 1.5$ (Györfi et al., 2002). The adaptivity of kernel regression to the intrinsic dimension has been proved by Kpotufe and Garg (2013). Following their results, the convergence rate becomes $\delta_n = n^{-2/(2+d)} \ll n^{-2/(2+D)}$ when there exists a low-dimensional structure in the data.
- **Local polynomial regression.** Let $\mathcal{M}$ be a Sobolev space with smoothness α. Then local polynomial regression estimators has the convergence rate as $\delta_n = n^{-\alpha/(\alpha+d)}$ where d is manifold dimension smaller than the original dimension D (Bickel and Li, 2007).
- **Random forests regression.** For Lipschitz continuous functions, Biau (2012) shows that the random forest estimator converges at rate $\delta_n = n^{-\frac{0.75}{s \log 2 + 0.75}}$ where s is the number of the relevant features. Hence, the convergence rate of the random forests becomes faster than $n^{-2/(2+D)}$ when $s \leq D/2$ under certain conditions. Wager and Walther (2015) use the guess-and-check forest algorithm to show that the convergence rate of the random forest is $\delta_n = n^{-\log(\xi)/\log(2\xi)}$ where $\xi = 1/(1 - 3/4s)$.

To the best of our knowledge, there has been no detailed investigation of the regression estimation error under separate sampling. In this case, we cannot directly take advantage of existing results on regression. However, as the sample size becomes larger, the difference between i.i.d. sampling and separate sampling becomes minor. Hence we expect that a reasonable regression estimator behaves similarly under both sampling schemes in large sample sizes, while a detailed analysis is necessary in future work. It is also worth noting that for certain regression methods, consistency results are not significantly affected by sampling scheme. For example, the consistency theory for L_1 penalized regression relies mainly on the assumption about a design matrix, which can be fulfilled under both sampling schemes (Van de Geer, 2008; Bühlmann and Van De Geer,

2011). In such a case, the same convergence rate can be established under both sampling schemes.

4. Local two-sample tests via regression

The global two-sample test only answers the question whether two distributions are different, whereas in some applications, it would be more valuable to describe how these two distributions differ in a multivariate space. With this goal in mind, we now move on to the local two-sample problem and study general properties of the local regression test.

4.1. The MSE and testing error for local regression

We start by establishing similar results in Section 3.2 for local regression tests. Given a local point $x \in S$ of interest, suppose that a regression estimator has the mean square error such that

$$\sup_{m \in \mathcal{M}} \mathbb{E}\left[(\widehat{m}(x) - m(x))^2\right] \leq C_{0,x}\delta_{n,x}, \tag{11}$$

where $C_{0,x}$ is a positive constant and $\delta_{n,x} = o(1)$. In addition, we assume $\delta_{n,x} \geq n^{-1}$ for i.i.d. sampling. Then the next theorem shows that for sufficiently large $C_{1,x}$ and n, the local testing error based on the given regression estimator can be arbitrarily small against the class of local alternatives given by

$$\mathcal{M}(C_{1,x}\delta_{n,x}) = \left\{m \in \mathcal{M} : (m(x) - \pi_1)^2 \geq C_{1,x}\delta_{n,x}\right\}.$$

Theorem 4.1. *Consider the case of i.i.d. sampling or separate sampling. In each case, consider the local regression test statistic $\widehat{\mathcal{T}}_{local}(x)$ in (3) based on a linear smoother $\widehat{m}(x) = \sum_{i=1}^n w_i(x)Y_i$ with the property in (11). Let t_α be the upper α quantile of the permutation distribution of $\widehat{\mathcal{T}}_{local}(x)$. Then for fixed $\alpha \in (0,1)$ and $\beta \in (0, 1-\alpha)$, there exist positive constants $C_{1,x}$ and N_x such that*

- *Type I error:* $\sup_{f_0, f_1 \in \mathcal{M}_0} \mathbb{P}_{f_0,f_1}\left(\widehat{\mathcal{T}}_{local}(x) > t_\alpha\right) \leq \alpha$ *and*
- *Type II error:* $\sup_{n \geq N_x} \sup_{f_0, f_1 \in \mathcal{M}(C_{1,x}\delta_{n,x})} \mathbb{P}_{f_0,f_1}\left(\widehat{\mathcal{T}}_{local}(x) \leq t_\alpha\right) \leq \beta.$

Remark 4.1. *Although Theorem 4.1 focuses on a linear smoother, the same conclusion holds for other regression methods as long as there exists a positive constant $C_{0,x,\alpha}$ such that the permutation critical value t_α is bounded above by $C_{0,x,\alpha}\delta_n$ with high probability (see Theorem 3.3 for a more formal statement).*

In order to keep things as simple and concrete as possible, we next focus on the Lipschitz class and analyze the optimality of the local regression tests

from a minimax point of view. In the rest of this section (Section 4.2–4.4), we concentrate on *i.i.d. sampling scheme* to take full advantage of known regression results. However, as we discussed in Section 3.3, similar results are expected to hold under separate sampling as well.

4.2. Minimax optimality over the Lipschitz class

For a fixed constant $L > 0$, let us denote the Lipschitz function class by

$$\mathcal{M}_{Lip} = \Big\{ m : |m(x) - m(y)| \le L||x - y||_2 \text{ for all } x, y \in S \Big\}.$$

We also denote the collection of α level tests by $\Phi_{n,\alpha} = \{\phi : \sup_{f_0,f_1 \in \mathcal{M}_0} \mathbb{P}_{f_0,f_1}(\phi = 1) \le \alpha\}$ and denote the class of Lipschitz local alternatives by

$$\mathcal{M}_{Lip}(\delta_{n,x}) = \Big\{ m \in \mathcal{M}_{Lip} : (m(x) - \pi_1)^2 \ge \delta_{n,x} \Big\}. \tag{12}$$

With this notation and fixed $\alpha \in (0,1)$ and $\beta \in (0, 1-\alpha)$, the *minimum separation* is defined by

$$\delta^\star_{n,x} = \inf \Big\{ \delta_{n,x} : \inf_{\phi \in \Phi_{n,\alpha}} \sup_{f_0,f_1 \in \mathcal{M}_{Lip}(\delta_{n,x})} \mathbb{P}_{f_0,f_1}(\phi = 0) \le \beta \Big\}, \tag{13}$$

which is the smallest distance between $m(x)$ and π_1 such that the power becomes nontrivial. Then a test is called minimax rate optimal if it has power uniformly over $\mathcal{M}_{Lip}(\delta_{n,x})$ such that $\delta_{n,x} \asymp \delta^\star_{n,x}$.

In this section, we will investigate minimax rate optimality of local regression tests over the Lipschitz class under i.i.d. sampling. First we formally state an upper bound for the local estimation error based on kNN and kernel regression in Example 4.1 and Example 4.2, respectively. We then use these results to obtain the upper bound for the minimum separation in Corollary 4.1.

Example 4.1 (kNN regression)**.** *For a fixed point $x \in S$, list the data by*

$$(X_{1,n}(x), Y_{1,n}(x)), \ldots, (X_{n,n}(x), Y_{n,n}(x)),$$

where $X_{k,n}(x)$ is the kth nearest neighbor of x and $Y_{k,n}(x)$ is its pair. Consider the kNN regression estimator

$$\widehat{m}_{kNN}(x) = \frac{1}{k_n} \sum_{i=1}^{k_n} Y_{(i,n)}(x), \tag{14}$$

and assume that $\mathbb{P}(X \in B_{x,\epsilon}) > \tau_x \epsilon^D$ where $B_{x,\epsilon}$ is a ball of radius $\epsilon > 0$ centered at x and $\tau_x > 0$. Then

$$\sup_{m \in \mathcal{M}_{Lip}} \mathbb{E}\left[(\widehat{m}_{kNN}(x) - m(x))^2 \right] \le \frac{1}{4k_n} + L^2 \frac{2\Gamma(2/D)}{D\tau_x^{2/D}} \left(\frac{k_n}{n} \right)^{2/D},$$

and for $k_n = n^{2/(2+D)}$, *we have*

$$\sup_{m\in\mathcal{M}_{Lip}} \mathbb{E}\left[(\widehat{m}_{kNN}(x) - m(x))^2\right] \leq C_{0,x} n^{-\frac{2}{2+D}},$$

where $C_{0,x} = 1/4 + L^2\Gamma(2/D)D^{-1}\tau_x^{-2/D}$.

A similar result can be established for kernel regression estimators as follows.

Example 4.2 (Kernel regression)**.** *Given a kernel* $K : S \mapsto [0,\infty)$, *the kernel regression estimator at a fixed point* x *is given by*

$$\widehat{m}_{ker}(x) = \frac{\sum_{i=1}^n Y_i K\left(\frac{x-X_i}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right)}. \tag{15}$$

Assume there exists $0 < r < R$ *and* $0 < \lambda < 1$ *such that*

$$\lambda I(x \in B_{0,r}) \leq K(x) \leq I(x \in B_{0,R})$$

where $B_{0,\epsilon}$ *is a ball of radius* $\epsilon > 0$ *centered at the origin. Further assume that* $\mathbb{P}(X \in B_{x,\epsilon}) > \tau_x \epsilon^D$ *for some* $\tau_x > 0$. *Then*

$$\sup_{m\in\mathcal{M}_{Lip}} \mathbb{E}\left[(\widehat{m}_{ker}(x) - m(x))^2\right] \leq \left(\frac{1+\lambda}{4\lambda^2\tau_x r^D} + \frac{2e^{-1}}{\tau_x r^D}\right)\frac{1}{nh_n^D} + L^2R^2h_n^2$$

and for $h_n = n^{-2/(2+D)}$,

$$\sup_{f_0,f_1\in\mathcal{M}_{Lip}} \mathbb{E}\left[(\widehat{m}_{ker}(x) - m(x))^2\right] \leq C_{0,x} n^{-\frac{2}{2+D}}$$

where $C_{0,x} = (1+\lambda)/(4\lambda^2\tau_x r^D) + 2e^{-1}/(\tau_x r^D) + L^2R^2$.

Remark 4.2. *Example 4.1 and Example 4.2 are well-known and standard except that we keep track of the constant* $C_{0,x}$ *over the Lipschitz class. Similar results exist in the literature but for slightly different settings. Hence, in Appendix A, we present detailed proofs for these two examples heavily building on Györfi et al. (2002). The proofs will also be used to study the performance of the kNN and kernel local regression tests under the existence of intrinsic dimension in Proposition 4.1.*

From the previous examples together with Theorem 4.1, we conclude that the minimum separation in (13) satisfies $\delta^\star_{n,x} \lesssim n^{-2/(2+D)}$. We summarize this result in the following corollary.

Corollary 4.1 (Upper bound)**.** *Let us denote the local kNN and kernel regression test statistics by*

$$\widehat{\mathcal{T}}_{kNN}(x) = (\widehat{m}_{kNN}(x) - \widehat{\pi}_1)^2, \quad \widehat{\mathcal{T}}_{ker}(x) = (\widehat{m}_{ker}(x) - \widehat{\pi}_1)^2, \tag{16}$$

and the upper α quantile of the permutation distribution of each statistic by $t_{\alpha,kNN}$ and $t_{\alpha,ker}$ respectively. Suppose the conditions in Example 4.1 holds with $k_n = n^{2/(D+2)}$. Then for fixed $\alpha \in (0,1)$ and $\beta \in (0, 1-\alpha)$, there exist positive constants $C_{1,x}$ and N_x such that

- *Type I error:* $\sup_{f_0,f_1 \in \mathcal{M}_0} \mathbb{P}_{f_0,f_1}\left(\widehat{\mathcal{T}}_{kNN}(x) > t_{\alpha,kNN}\right) \leq \alpha$ *and*
- *Type II error:* $\sup_{n \geq N_x} \sup_{f_0,f_1 \in \mathcal{M}_{Lip}(C_{1,x} n^{-2/(2+D)})} \mathbb{P}_{f_0,f_1}\left(\widehat{\mathcal{T}}_{kNN}(x) \leq t_{\alpha,kNN}\right) \leq \beta.$

On the other hand, under the conditions in Example 4.2 with $h_n = n^{-2/(2+D)}$ and for fixed $\alpha \in (0,1)$ and $\beta \in (0, 1-\alpha)$, there exist positive constants $C_{1,x}$ and N_x such that

- *Type I error:* $\sup_{f_0,f_1 \in \mathcal{M}_0} \mathbb{P}_{f_0,f_1}\left(\widehat{\mathcal{T}}_{ker}(x) > t_{\alpha,ker}\right) \leq \alpha$ *and*
- *Type II error:* $\sup_{n \geq N_x} \sup_{f_0,f_1 \in \mathcal{M}_{Lip}(C_{1,x} n^{-2/(2+D)})} \mathbb{P}_{f_0,f_1}\left(\widehat{\mathcal{T}}_{ker}(x) \leq t_{\alpha,ker}\right) \leq \beta$

As a result, the minimum separation satisfies $\delta^\star_{n,x} \lesssim n^{-2/(2+D)}$.

Next based on the standard technique to lower bound the testing error (e.g., Ingster, 1987; Baraud, 2002), we establish a lower bound for the minimum separation by $n^{-2/(2+D)} \lesssim \delta^\star_{n,x}$. This results matches with the upper bound in Corollary 4.1. Therefore, the tests in Corollary 4.1 are minimax rate optimal and cannot be improved.

Theorem 4.2 (Lower bound)**.** *For any given $\alpha \in (0,1)$ and $\beta \in (1-\alpha)$, there exists a constant $C_{1,x} > 0$ such that*

$$\inf_{\phi \in \Phi_{n,\alpha}} \sup_{f_0,f_1 \in \mathcal{M}_{Lip}(C_{1,x} n^{-2/(2+D)})} \mathbb{P}_{f_0,f_1}(\phi = 0) \geq 1 - \alpha - \beta.$$

Remark 4.3. *In the context of two-sample testing, it is sometimes more natural to make smoothness assumptions on densities f_0 and f_1 rather than on the regression function. Here we briefly discuss how to translate the smoothness condition on f_0 and f_1 into a condition on the regression function. Suppose that density functions f_0 and f_1 are uniformly bounded below by $c > 0$ (see, e.g. Yang and Barron, 1999, for a similar assumption). Then some algebra shows that*

$$|m(x) - m(y)| \leq \pi_0 c^{-1} |f_0(x) - f_0(y)| + \pi_1 c^{-1} |f_1(x) - f_1(y)|.$$

In other words, if f_0 and f_1 are Lipschitz continuous (or more generally Hölder continuous), then the regression function is also Lipschitz continuous with a different Lipschitz constant. This means that our theoretical results will remain valid for the class of Lipschitz densities with the boundedness condition.

4.3. An approach to intrinsic dimension

The previous results show that no test is uniformly powerful when the square distance between $m(x)$ and π_1 is order of $n^{-2/(2+D)}$; therefore it demonstrates the typical curse of dimensionality. Suppose that data $X \in S \subseteq \mathbb{R}^D$ has low intrinsic dimension d which is smaller than the original dimension D (e.g. manifold data). In this case, we would like to have a test whose performance only depends on intrinsic dimension and thus avoids the curse of dimensionality. For this purpose, we consider the homogeneous measure which captures local dimension of data.

Definition 4.1. *(Definition 2 of Kpotufe, 2011) Fix $x \in S \subseteq \mathbb{R}^D$, and $r > 0$. Let $C > 0$ and $1 \leq d < D$. The probability measure $\mathbb{P}(\cdot)$ is (C, d)-homogeneous on $B_{x,r}$ if we have $\mathbb{P}(X \in B_{x,r'}) \leq C\epsilon^{-d}\mathbb{P}(X \in B_{x,\epsilon r'})$ for all $r' \leq r$ and $0 < \epsilon < 1$.*

Using Definition 4.1, we reproduce Corollary 4.1 and show that the performances of the local kNN and kernel regression tests depend on the intrinsic dimension instead of the original dimension.

Proposition 4.1. *Consider the same notations as in Corollary 4.1 and let $x \in S \subseteq \mathbb{R}^D$. Suppose the probability measure $\mathbb{P}(\cdot)$ is (C, d)-homogeneous on $B_{x,r}$. Then for the kNN regression test with $k_n = n^{2/(2+d)}$ and for any $\beta \in (0, 1-\alpha)$, there exist positive constants $C_{1,x}$ and N_x such that*

- *Type I error:* $\sup_{f_0,f_1 \in \mathcal{M}_0} \mathbb{P}_{f_0,f_1}\left(\widehat{\mathcal{T}}_{kNN}(x) > t_{\alpha,kNN}\right) \leq \alpha$ *and*
- *Type II error:* $\sup_{n \geq N_x} \sup_{f_0,f_1 \in \mathcal{M}_{Lip}(C_{1,x} n^{-2/(2+d)})} \mathbb{P}_{f_0,f_1}\left(\widehat{\mathcal{T}}_{kNN}(x) \leq t_{\alpha,kNN}\right) \leq \beta.$

On the other hand, for the kernel regression test with $h_n = n^{-2/(2+d)}$ and for any $\beta \in (0, 1-\alpha)$, there exist positive constants $C_{1,x}$ and N_x such that

- *Type I error:* $\sup_{f_0,f_1 \in \mathcal{M}_0} \mathbb{P}_{f_0,f_1}\left(\widehat{\mathcal{T}}_{ker}(x) > t_{\alpha,ker}\right) \leq \alpha$ *and*
- *Type II error:* $\sup_{n \geq N_x} \sup_{f_0,f_1 \in \mathcal{M}_{Lip}(C_{1,x} n^{-2/(2+d)})} \mathbb{P}_{f_0,f_1}\left(\widehat{\mathcal{T}}_{ker}(x) \leq t_{\alpha,ker}\right) \leq \beta.$

When the intrinsic dimension is unknown, one can employ a Bonferroni procedure to obtain the same results in Proposition 4.1. To illustrate the idea, let $k_n(i) = n^{-2/(i+2)}$ for $i = 1, \ldots, D$ and denote the resulting kNN tests by $\phi_i(\alpha) = I(\mathcal{T}^{(i)}_{kNN}(x) > t^{(i)}_{\alpha,kNN})$ where $\mathcal{T}^{(i)}_{kNN}(x)$ and $t^{(i)}_{\alpha,kNN}$ are the kNN test statistic calculated with $k_n(i)$ and the corresponding α level permutation critical value, respectively. Then the final test is defined by $\phi_{max} = \max_{1 \leq i \leq D} \phi_i(\alpha/D)$. By using the union bound, it is easy to see that $\sup_{f_0,f_1 \in \mathcal{M}_0} \mathbb{P}_{f_0,f_1}(\phi_{max} = 1) \leq \alpha$ and

$$\sup_{n \geq N_x} \sup_{f_0,f_1 \in \mathcal{M}_{Lip}(C_{1,x} n^{-2/(2+d)})} \mathbb{P}_{f_0,f_1}(\phi_{max} = 0) \leq \beta,$$

for certain $C_{1,x}$ and N_x. This shows that the Bonferroni test does not lose any power in terms of separation rate and it adapts to the unknown intrinsic dimension. Despite this theoretical guarantee, the Bonferroni approach should be used with caution in practice. Indeed the Bonferroni test might be too conservative since it does not take into account the dependency structure among $\phi_1, \ldots, \phi_D$.

Remark 4.4. *For simplicity, we illustrate our idea on the Lipschitz class which only requires a mild smoothness assumption. Nevertheless our results in Section 4.2–4.3 can be extended to a general function class such as Hölder class (e.g. Chapter 3.2 of Györfi et al., 2002) in a similar way. Indeed, all we need is a uniform bound for the MSE (11) over a general class, which can be found in the regression literature (see Section 3.3).*

4.4. Limiting distribution of local permutation test statistics

When the sample size is large, calculating the permutation distribution is time-consuming. Hence it would be useful to investigate the limiting distribution of the permutation statistic. Based on the combinatorial central limit theorem (e.g. Bolthausen, 1984), we show that the permutation distribution of our local test statistic converges to the chi-square distribution with one degree of freedom as the sample size tends to infinity.

Theorem 4.3. *Consider the local regression test statistic $\widehat{\mathcal{T}}_{local}(x)$ in (3) based on a linear smoother $\widehat{m}(x) = \sum_{i=1}^n w_i(x) Y_i$. Suppose that*

$$\frac{\max_{1\le i\le n} |w_i(x) - 1/n|}{\{\sum_{i=1}^n (w_i(x) - 1/n)^2\}^{1/2}} \xrightarrow{p} 0 \tag{17}$$

holds and let

$$\sigma_n^2 = \frac{n}{n-1} \widehat{\pi}_1 (1 - \widehat{\pi}_1) \sum_{i=1}^n \left(w_i(x) - \frac{1}{n} \right)^2. \tag{18}$$

Further let $\eta = (\eta_1, \ldots, \eta_n)$ be a permutation of $\{1, \ldots, n\}$. Then the permutation distribution of the one-side local regression statistic converges to the standard normal distribution as

$$\sup_{t\in\mathbb{R}} \left| \mathbb{P}_\eta \left(\sigma_n^{-1} (\widehat{m}_\eta(x) - \widehat{\pi}_1) \le t \middle| \mathcal{X}_n \right) - \mathbb{P}\left(N(0,1) \le t \right) \right| \xrightarrow{p} 0.$$

Here $\mathbb{P}_\eta(\cdot|\mathcal{X}_n)$ is the uniform probability measure over permutations conditioned on $(X_1, Y_1), \ldots, (X_n, Y_n)$ and $\widehat{m}_\eta(x) = \sum_{i=1}^n w_i(x) Y_{\eta_i}$. Thereby, $\sigma_n^{-2} \widehat{\mathcal{T}}_{local}(x)$ converges to the chi-square distribution with one degree of freedom as

$$\sup_{t\in\mathbb{R}} \left| \mathbb{P}_\eta \left(\sigma_n^{-2} \widehat{\mathcal{T}}_{local}(x) \le t \middle| \mathcal{X}_n \right) - \mathbb{P}\left(\chi_1^2 \le t \right) \right| \xrightarrow{p} 0.$$

We illustrate Theorem 4.3 using kNN and kernel regression and show that both $\sigma_n^{-2} \widehat{\mathcal{T}}_{kNN}(x)$ and $\sigma_n^{-2} \widehat{\mathcal{T}}_{ker}(x)$ converge to the chi-square distribution with one degree of freedom under appropriate conditions.

Corollary 4.2 (kNN regression)**.** *Consider the kNN estimator in (14) with*

$$\sigma_n^2 = \widehat{\pi}_1(1-\widehat{\pi}_1)\frac{(n-1)(n-k)}{n^2 k}.$$

Then the permutation distribution of $\sigma_n^{-2}\widehat{\mathcal{T}}_{kNN}(x)$ *converges to the chi-square distribution with one degree of freedom when* $n, k \to \infty$ *and* $2k < n$*.*

Corollary 4.3 (Kernel regression)**.** *Consider the kernel regression estimator in (15) and assume that* $\sup_t |K(t)| = \mathcal{K} < \infty$*,* $\int K^2(t)dt < \infty$ *and* $\int K_h(t)dx = 1$ *where* $K_h(t) = h^{-D}K(t/h)$*. Denote the density function of* X *by* $f(\cdot)$*. Assume that* $0 < f(x) < \infty$ *and* $f(\cdot)$ *is twice differentiable at* x*. Further assume that* $nh^D \to \infty$ *and* $h \to 0$*. Then the permutation distribution of* $\sigma_n^{-2}\widehat{\mathcal{T}}_{ker}(x)$ *converges to the chi-square distribution with one degree of freedom where* σ_n^2 *is given in (18).*

5. Simulations

In this section, we carry out simulation studies for global and local two-sample tests to examine the empirical performance of the proposed methods. Throughout our simulations, we focus on the separate sampling scenarios under which other existing two-sample tests are usually investigated. We begin by comparing the regression test based on random forests (Breiman, 2001) with other benchmark competitors in Section 5.1. Next in Section 5.2, we illustrate by an example that the classification accuracy tests can fail due to their discrete nature while the corresponding regression tests perform well. We also provide simulation results for the local regression test in Section 5.3 to validate our approach.

5.1. Random forests two-sample testing

Random forests have been proven to be a powerful tool for regression and classification problems in many application areas (see e.g., Hamza and Larocque, 2005; Díaz-Uriarte and De Andres, 2006; Cutler et al., 2007; Chen and Ishwaran, 2012). Despite the good performance of random forests in classification and regression problems, only a few works have applied these methods to statistical inference problems. To the best of our knowledge, only Gagnon-Bartsch and Shem-Tov (2016) and Hediger et al. (2019) use random forests for the two-sample problem. Now whereas Gagnon-Bartsch and Shem-Tov (2016) and Hediger et al. (2019) consider an accuracy test based on random forests, we propose a regression test based on random forests. The corresponding test statistic is given by

$$\widehat{\mathcal{T}}_{RF} = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{m}_{RF}(X_i) - \widehat{\pi}_1\right)^2, \tag{19}$$

where $\widehat{m}_{RF}$ is the regression estimator from the random forest algorithm. For our simulation study, we implement both the RF accuracy and regression tests

with the `randomForest` package (version 4.6-12) in `R` with default options for the parameters. We found in our simulation study that the in-sample classification accuracy of random forests is typically one even under the null case; therefore, the resulting test has no power against any alternative. For this reason, we instead estimate the classification accuracy from out-of-bag samples (which is a default option provided by the `randomForest` package). Throughout this section, we denote the accuracy test statistic based on random forests by $\widehat{\mathcal{A}}_{RF}$.

5.1.1. Simulation setting

Our simulations analyze two main settings. The first setting includes dense alternatives where the two distributions are different over a number of coordinates. The second setting, on the other hand, considers sparse alternatives where the two distributions differ in only a few coordinates. We carry out the simulations via the permutation procedure with 100 random permutations, repeated 300 times for all test statistics. The significance level is controlled at $\alpha = 0.05$.

Dense Alternatives. For the dense alternatives, we draw random samples of size $n_0 = n_1 = 20$ and dimension $D = 5, 20, 50, 100, 150$ and 200 from either multivariate normal distributions $N(\mu, \Sigma)$ or multivariate Cauchy distribution $\mathsf{C}(\mu, \Sigma)$ with different location μ and scale Σ parameters. We consider the following scenarios:

- **Dense Normal Location.** Test $N(0, I_D)$ versus $N(\mu, I_D)$, where $\mu = (0.2, 0.2, \ldots, 0.2)^\top$.
- **Dense Cauchy Location.** Test $\mathsf{C}(0, I_D)$ versus $\mathsf{C}(\mu, I_D)$, where $\mu = (0.3, 0.3, \ldots, 0.3)^\top$.
- **Dense Normal Scale.** Test $N(0, I_D)$ versus $N(0, J_D)$, where J_D is a diagonal matrix whose diagonal elements are $(0.6, 0.6, \ldots, 0.6)^\top$.
- **Dense Cauchy Scale.** Test $\mathsf{C}(0, I_D)$ versus $\mathsf{C}(0, J_D)$, where J_D is a diagonal matrix whose diagonal elements are $(0.5, 0.5, \ldots, 0.5)^\top$.

Sparse Alternatives. Similarly, we generate random samples with $n_0 = n_1 = 20$ and $D = 20, 50, 100, 200, 300$ and 400 from either multivariate normal distributions or multivariate Cauchy distributions. We consider the following problems:

- **Sparse Normal Location.** Test $N(0, I_D)$ versus $N(\mu, I_D)$, where $\mu = (2, 0, \ldots, 0)^\top$.
- **Sparse Cauchy Location.** Test $\mathsf{C}(0, I_D)$ versus $\mathsf{C}(\mu, I_D)$, where $\mu = (3, 0, \ldots, 0)^\top$.
- **Sparse Normal Scale.** Test $N(0, I_D)$ versus $N(0, J_D)$, where J_D is a diagonal matrix with diagonal elements $(0.01, 1, \ldots, 1)^\top$.
- **Sparse Cauchy Scale.** Test $\mathsf{C}(0, I_D)$ versus $\mathsf{C}(0, J_D)$, where J_D is a diagonal matrix with diagonal elements $(0.01, 1, \ldots, 1)^\top$.

As a benchmark competitor, we consider the maximum mean discrepancy (MMD) test (Gretton et al., 2012) based on

$$\begin{aligned} &\mathrm{MMD}_n^2 \\ &= -\frac{2}{n_0 n_1}\sum_{i,j=1}^{n_0,n_1} k(X_{i,0},X_{i,1}) + \frac{1}{n_0^2}\sum_{i,j=1}^{n_0} k(X_{i,0},X_{j,0}) + \frac{1}{n_1^2}\sum_{i,j=1}^{n_1} k(X_{i,0},X_{j,0}), \end{aligned} \tag{20}$$

where $k(x,y)$ is the Gaussian kernel with a bandwidth chosen by the median heuristic, i.e. $k(x,y) = \exp\left(-||x-y||_2^2/\sigma_{\text{median}}\right)$ (see, Gretton et al., 2012, for details). We also consider the Energy test (Székely and Rizzo, 2004; Baringhaus and Franz, 2004) based on

$$\begin{aligned} &\mathrm{Energy}_n \\ &= \frac{2}{n_0 n_1}\sum_{i,j=1}^{n_0,n_1} ||X_{i,0}-X_{j,1}||_2 - \frac{1}{n_0^2}\sum_{i,j=1}^{n_0} ||X_{i,0}-X_{j,0}||_2 - \frac{1}{n_1^2}\sum_{i,j=1}^{n_1} ||X_{i,1}-X_{j,1}||_2. \end{aligned} \tag{21}$$

5.1.2. Simulation results

Tables 1–4 summarize our simulation results. We see from Table 1 and 2 that MMD_n and Energy_n perform better than the regression test ($\widehat{\mathcal{T}}_{RF}$) and the accuracy test ($\widehat{\mathcal{A}}_{RF}$) against the dense normal location and scale alternatives. Indeed, MMD_n and Energy_n are known to be asymptotically optimal against the normal location alternative with the identity covariance matrix (Ramdas et al., 2015). However, they are both moment-based statistics, and hence sensitive to outliers. They are also based on the Euclidean metric. A major issue of the Euclidean and similar metrics is that they assign weights to the coordinates proportional to their scale without screening for irrelevant variables. Consequently, neither MMD_n nor Energy_n can properly deal with sparse alternatives, which explains their poor performance against the sparse location and scale alternatives. On the other hand, the base learner of the random forest algorithm is the decision tree. The usual splitting rule of decision trees is invariant to absolute values (see e.g., Chapter 9.2 of Friedman et al., 2009), which leads to robustness against outliers.

Random forests also have the ability to handle sparse alternatives by randomly selecting a few variables during the tree-growing process. By averaging each tree, random forests eventually put more weight on informative variables. In general, $\widehat{\mathcal{T}}_{RF}$ and $\widehat{\mathcal{A}}_{RF}$ are comparable to or more powerful than MMD_n and Energy_n under the sparse location and scale alternatives. Finally, we note from our simulations that the regression test $\widehat{\mathcal{T}}_{RF}$ exhibits higher power than the accuracy test $\widehat{\mathcal{A}}_{RF}$ for the dense as well as the sparse alternatives.

TABLE 1
Power analysis against dense location alternatives at level $\alpha = 0.05$

	Normal Dense Location						*Cauchy Dense Location*					
D	5	20	50	100	150	200	5	20	50	100	150	200
$\widehat{\mathcal{T}}_{\mathrm{RF}}$	0.123	0.187	0.303	0.417	0.573	0.633	**0.157**	**0.370**	**0.607**	**0.803**	**0.893**	**0.950**
$\widehat{\mathcal{A}}_{\mathrm{RF}}$	0.070	0.117	0.233	0.340	0.440	0.510	0.093	0.260	0.503	0.693	0.793	0.857
MMD_n	**0.143**	**0.290**	**0.520**	**0.723**	**0.880**	**0.937**	0.097	0.057	0.053	0.050	0.060	0.040
Energy_n	**0.156**	**0.283**	**0.530**	**0.720**	**0.877**	**0.940**	0.083	0.077	0.073	0.057	0.057	0.057

TABLE 2
Power analysis against dense scale alternatives at level $\alpha = 0.05$

	Normal Dense Scale						*Cauchy Dense Scale*					
D	5	20	50	100	150	200	5	20	50	100	150	200
$\widehat{\mathcal{T}}_{\mathrm{RF}}$	0.133	0.187	0.260	0.350	0.410	0.473	0.287	**0.557**	**0.790**	**0.937**	**0.953**	**0.970**
$\widehat{\mathcal{A}}_{\mathrm{RF}}$	0.097	0.150	0.200	0.277	0.277	0.290	0.230	0.407	0.663	0.783	0.840	0.877
MMD_n	**0.210**	**0.563**	**0.847**	**0.993**	**0.997**	**1.000**	**0.380**	0.380	0.407	0.407	0.400	0.400
Energy_n	0.080	0.263	0.397	0.657	0.847	0.913	0.283	0.293	0.310	0.310	0.313	0.297

TABLE 3
Power analysis against sparse location alternatives at level $\alpha = 0.05$

	Normal Sparse Location						*Cauchy Sparse Location*					
D	20	50	100	200	300	400	20	50	100	200	300	400
$\widehat{\mathcal{T}}_{\mathrm{RF}}$	0.953	0.880	**0.830**	**0.687**	**0.600**	**0.503**	**0.960**	**0.933**	**0.897**	**0.710**	**0.643**	**0.577**
$\widehat{\mathcal{A}}_{\mathrm{RF}}$	0.883	0.817	0.763	0.600	0.523	0.440	0.943	0.877	0.830	0.613	0.540	0.527
MMD_n	**0.977**	**0.943**	0.770	0.587	0.437	0.360	0.147	0.067	0.057	0.043	0.057	0.027
Energy_n	**0.977**	**0.943**	0.770	0.587	0.440	0.367	0.157	0.083	0.043	0.037	0.050	0.040

TABLE 4
Power analysis against sparse scale alternatives at level $\alpha = 0.05$

	Normal Sparse Scale						*Cauchy Sparse Scale*					
D	20	50	100	200	300	400	20	50	100	200	300	400
$\widehat{\mathcal{T}}_{\mathrm{RF}}$	**0.630**	**0.333**	**0.287**	**0.167**	**0.167**	**0.133**	**0.830**	**0.550**	**0.390**	**0.257**	**0.197**	**0.170**
$\widehat{\mathcal{A}}_{\mathrm{RF}}$	0.603	0.297	0.220	0.130	0.120	0.087	0.743	0.467	0.287	0.207	0.170	0.150
MMD_n	0.043	0.057	0.043	0.053	0.060	0.063	0.067	0.033	0.040	0.057	0.063	0.043
Energy_n	0.037	0.050	0.043	0.050	0.060	0.063	0.047	0.047	0.040	0.057	0.053	0.037

5.2. A comparison between regression and classification accuracy tests

As mentioned earlier, many classifiers are typically estimated by dichotomizing regression estimators. Depending on the alternative, this dichotomization can result in a less powerful accuracy test than the corresponding regression test. We specifically demonstrate this point by considering two commonly used nonparametric regression methods; namely, k-nearest neighbors regression and kernel regression.

5.2.1. Simulation setting

Recall the kNN estimator and the kernel regression estimator in (14) and (15), respectively. Using these estimators, the global regression test statistics are given

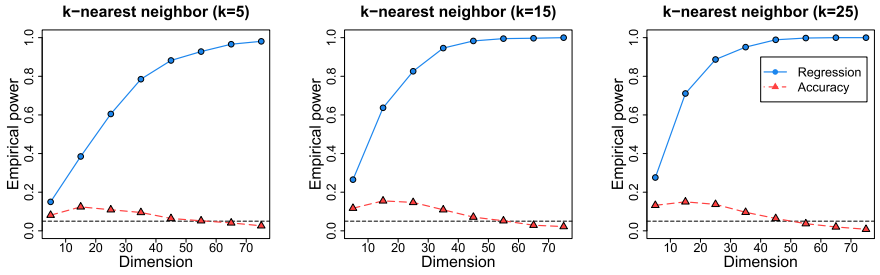

FIG 3. *Power comparison between the regression test and the classification accuracy test via k-NN regression at level* $\alpha = 0.05$ *for the toy example in Section 5.2.*

by

$$\widehat{\mathcal{T}}_{kNN} = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{m}_{kNN}(X_i) - \widehat{\pi}_1\right)^2 \quad \text{and} \quad \widehat{\mathcal{T}}_{ker} = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{m}_{ker}(X_i) - \widehat{\pi}_1\right)^2.$$

Here we use the Euclidean distance to measure the pairwise distance between observations for kNN. On the other hand, we consider the Gaussian kernel with a diagonal bandwidth matrix with identical components h for kernel regression. The corresponding accuracy test statistics are

$$\widehat{\mathcal{A}}_{kNN} = \frac{1}{n}\sum_{i=1}^{n} I\Big(I(\widehat{m}_{kNN}(X_i) > 1/2) = Y_i\Big) \quad \text{and}$$

$$\widehat{\mathcal{A}}_{ker} = \frac{1}{n}\sum_{i=1}^{n} I\Big(I(\widehat{m}_{ker}(X_i) > 1/2) = Y_i\Big),$$

respectively. For all tests, we reject the null hypothesis when the test statistic is larger than a permutation critical value.

For the simulation study, we let $\{X_{1,0}, \ldots, X_{n_0,0}\} \overset{i.i.d.}{\sim} N(\mu_0, \sigma_0^2 \times I_D)$ and $\{X_{1,1}, \ldots, X_{1,n_1}\} \overset{i.i.d.}{\sim} N(\mu_1, \sigma_1^2 \times I_D)$ where $\mu_0 = (0, \ldots, 0)^\top$, $\mu_1 = (0.2, \ldots, 0.2)^\top$, $\sigma_0^2 = 1$, and $\sigma_1^2 = 1.2$. Hence, there exist differences in both the location and scale parameters. We choose the sample sizes $n_0 = n_1 = 50$ and change the dimension from $D = 5$ to $D = 75$ by steps of 10. To compare the performance, we carry out the permutation test with 200 permutations, and the simulations are repeated 1,000 times to estimate the power of the test. We provide results for a range of different values of the tuning parameters: $k = 5, 15, 25$ for the k-NN regression, and $h = 5, 15, 25$ for the kernel regression.

5.2.2. Simulation results

Simulation results are presented in Figure 3 and 4. From the results, it is seen that the regression tests consistently outperform the corresponding classification

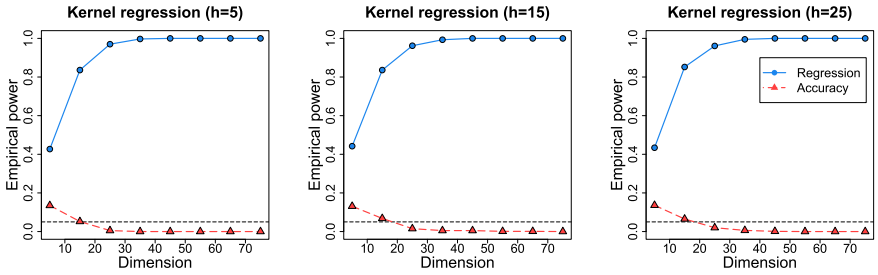

FIG 4. *Power comparison between the regression test and the classification accuracy test via kernel regression at level* $\alpha = 0.05$ *for the toy example in Section 5.2.*

accuracy tests under the given scenario. The power of the accuracy tests even decreases with dimension, whereas the power of the regression tests steadily increases with dimension. The increase in power with dimension is desirable under this scenario because each coordinate presents evidence towards the alternative. The counter-intuitive result for the accuracy tests is due to the fact that the tests employ a dichotomized regression estimator. To explain it more clearly, we borrow some results from Mondal et al. (2015). First, it can be shown by the weak law of large numbers that

$$\begin{aligned}
&1)\ D^{-1/2}||X_{i,0} - X_{j,0}||_2 \xrightarrow{p} \sigma_0\sqrt{2} \quad \text{for } 1 \le i < j \le n_0,\\
&2)\ D^{-1/2}||X_{i,1} - X_{j,1}||_2 \xrightarrow{p} \sigma_1\sqrt{2} \quad \text{for } 1 \le i < j \le n_1,\\
&3)\ D^{-1/2}||X_{i,0} - X_{j,1}||_2 \xrightarrow{p} \sqrt{\sigma_0^2 + \sigma_1^2 + (\mu_0 - \mu_1)^2}
\end{aligned}$$

for $1 \le i \le n_0$, $1 \le j \le n_1$, as $D \to \infty$ while n_0 and n_1 are fixed. For the given example, we have $\sigma_0\sqrt{2} < \sqrt{\sigma_0^2 + \sigma_0^2 + (\mu_0 - \mu_1)^2} < \sigma_1\sqrt{2}$, which implies that every instance is closer to an instance from the class $Y = 0$ than to other instances from the class $Y = 1$. In other words, the nearest neighbors of any observation are most likely to be from the class $Y = 0$. Note that both k-NN and kernel regression, explicitly or implicitly, use the Euclidean distance to calculate the proximity between two instances. Therefore, we observe with high probability that $\widehat{m}_{kNN}(X_i)$ and $\widehat{m}_{KerR}(X_i)$ are estimated as less than half and the dichotomized classifiers become

$$I\left(\widehat{m}_{kNN}(X_i) > 1/2\right) = \mathcal{I}\left(\widehat{m}_{KerR}(X_i) > 1/2\right) = 0, \quad \text{for all } i = 1, \ldots, n.$$

Due to this dichotomization, $\widehat{\mathcal{A}}_{kNN}$ and $\widehat{\mathcal{A}}_{KerR}$ converge to the empirical class probability n_0/n under the alternative, resulting in poor power performance. On the other hand, the regression tests based on $\widehat{\mathcal{T}}_{kNN}$ and $\widehat{\mathcal{T}}_{ker}$ can be powerful as long as $\widehat{m}_{kNN}(x)$ and $\widehat{m}_{ker}(x)$ significantly deviate from the class probability. This is indeed the case under the considered scenario and thus explains why the regression tests outperform the corresponding classification tests.

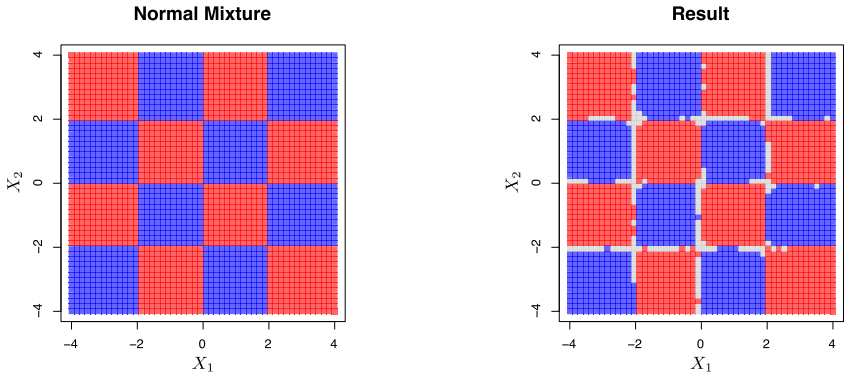

FIG 5. *Significant local regions for the normal mixture example. The left is the underlying true model and the right is the result of the local two-sample test. The difference regions are colored as follows — (a) red:* $f_1(x,y) > f_0(x,y)$*, (b) blue:* $f_1(x,y) < f_0(x,y)$ *and (c) gray: insignificant.*

5.3. Toy examples for local two-sample testing

Contrary to classification accuracy, our regression approach naturally leads to a local two-sample testing framework that provides further information on pointwise differences between two populations. We consider two toy examples to demonstrate the empirical performance of the local regression test. For the simulation study, we focus on the local kNN regression statistic in (16) with $k_n = n^{2/(2+D)}$ for the normal mixture example and $k_n = n^{2/(2+d)}$ for the manifold example. For both examples, we control the family-wise error rate (FWER) at $\alpha = 0.05$ via the Hochberg step up procedure (Hochberg, 1988).

5.3.1. Normal mixture example

In the first example, we consider two normal mixtures in $\mathbb{R}^2$:

$$f_0(x,y) = \frac{1}{8}\sum_{i=1}^{8}\phi_i(x,y) \quad \text{and} \quad f_1(x,y) = \frac{1}{8}\sum_{i=1}^{8}\phi_i'(x,y),$$

where ϕ_i is the bivariate normal density function with means $\mu_1 = (-3,-3)$, $\mu_2 = (-3,1)$, $\mu_3 = (-1,-1)$, $\mu_4 = (-1,3)$, $\mu_5 = (1,-3)$, $\mu_6 = (1,1)$, $\mu_7 = (3,-1)$, $\mu_8 = (3,3)$ and covariance matrix $\Sigma = 0.3^2 \times I_2$. ϕ_i' is similarly defined with means $\mu_1' = (-3,-1)$, $\mu_2' = (-3,3)$, $\mu_3' = (-1,-3)$, $\mu_4' = (-1,1)$, $\mu_5' = (1,-1)$, $\mu_6' = (1,3)$, $\mu_7' = (3,-3)$, $\mu_8' = (3,1)$ and the same covariance matrix. We generated $n_0 = n_1 = 2000$ samples from f_0 and f_1 and implemented Algorithm 2 to capture local significant points. The local tests were performed at a fixed uniform grid of 50×50 points over $(x,y) \in [-4,4] \times [-4,4]$ and the result is presented in Figure 5.

5.3.2. Manifold data example

In the second example, we create high-dimensional data with a low-dimensional manifold structure by generating edge images of size 16×16. Let x, y be integers on evenly spaced points between -30 and 30 that are 2 units apart. Hence the size of the domain of (x, y) becomes 16×16. Given two underlying parameters $\theta \in [-\pi, \pi]$ and $\rho \in [-5, 5]$, an edge image is defined by

$$\mathcal{I}(x, y) = I\left(x \cdot \cos(\theta) + y \cdot \sin(\theta) - \rho > 0\right).$$

For the simulation, we draw $n_0 = n_1 = 100$ samples from

$$(\theta_0, \rho_0) \sim \frac{1}{10}\text{Unif}([0, \pi] \times [0, 5]) + \frac{9}{10}\text{Unif}([-\pi, 0] \times [-5, 0]) \quad \text{and}$$
$$(\theta_1, \rho_1) \sim \frac{9}{10}\text{Unif}([0, \pi] \times [0, 5]) + \frac{1}{10}\text{Unif}([-\pi, 0] \times [-5, 0]),$$

and generate corresponding edge images. As a result, there are two sets of the edge images supported on $\mathbb{R}^{256}$. Using these image samples, we implemented Algorithm 2 to detect local significant points. The local tests were performed at fixed images whose parameters are defined on a uniform grid of 200×200 points over $(\theta, \rho) \in [-\pi, \pi] \times [-5, 5]$. For visualization purpose, we projected the testing points into the two-dimensional diffusion space (see Appendix B for details) and the final result is provided in Figure 6.

For both examples, the kNN local regression test performs reasonably well and detects most of the local differences between two distributions.

6. Application to astronomy data

Continuing our discussion from Section 1.1, we apply our two-sample framework to galaxies in the COSMOS, EGS, GOODS-North and UDS fields observed by the Hubble Space Telescope (HST) as part of the CANDELS program.[1] For the analysis, we compute seven morphological statistics that summarize galaxy images nonparametrically: *M, I, D* (Freeman et al., 2013), *Gini*, M_{20} (Lotz et al., 2004), C and A (Conselice, 2003). Each statistic (see the references for details) explains particular aspects of galaxy morphology. In brief, the *M, I, D* statistics capture galaxies with disturbed morphologies, *Gini* and M_{20} describe the variance of a galaxy's stellar light distribution, and the C and A statistics measure the concentration of light and asymmetry of a galaxy, respectively. We restrict our study to relatively nearby galaxy observations that have a redshift (proxy for distance) estimate between $0.56 < z < 1.12$. The final data set consists of 2736 so-called i-band-selected galaxy observations. For each galaxy, we have seven morphological image statistics along with an estimate of star-formation rate (SFR).

[1] http://candels.ucolick.org

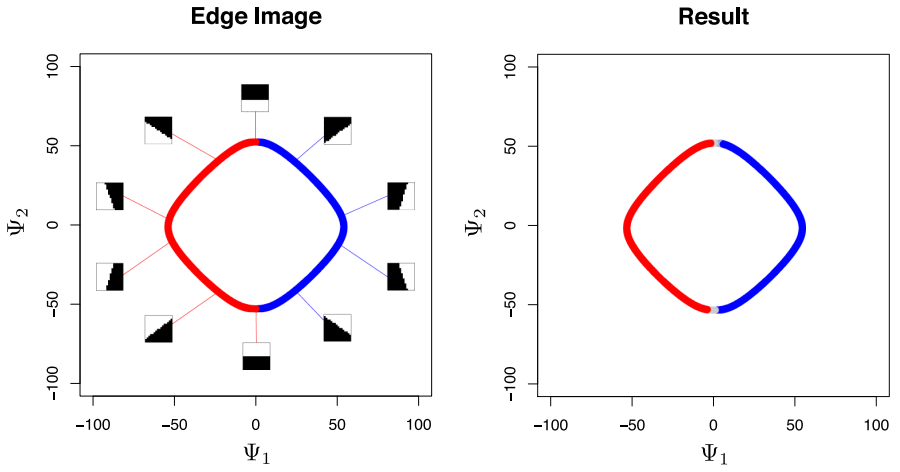

FIG 6. *Significant local regions for the manifold data example. The left is the underlying true model and the right is the result of the local two-sample test. The difference regions are colored as follows — (a) red:* $f_1(x_1, \ldots, x_{256}) > f_0(x_1, \ldots, x_{256})$*, (b) blue:* $f_1(x_1, \ldots, x_{256}) < f_0(x_1, \ldots, x_{256})$ *and (c) gray: insignificant. Here* Ψ_1 *and* Ψ_2 *denote the the first two coordinates of the diffusion map.*

Galaxy morphology is closely related to other physical properties such as star formation rate, mass and metallicity (see, e.g., Snyder et al., 2015). The aim of this study is to demonstrate that our local two-sample framework can be valuable in detecting and quantifying dependencies between variables of moderate or high dimension without resorting to low-dimensional projections of summary statistics. In particular, we demonstrate that local two-sample tests can identify galaxies that lie in regions of the feature space where the estimated proportion of a particular defined class of objects (such as star-forming galaxies) differs significantly from the global proportion. Hence, we start by defining two galaxy classes based on the SFR: we say that a galaxy belongs to the high-SFR group if its SFR is higher than the upper 25% quantile of the SFR distribution ($\log_{10}(\text{SFR}) > 1.201$), and that it belongs to the low-SFR group if its SFR is lower than the lower 25% quantile of the SFR distribution ($\log_{10}(\text{SFR}) < -0.915$). We further randomly divide the data into a training set ($n = 684$) and a test set ($n = 684$). We use the training data to construct the local test statistic in (3), and we perform the local-two sample tests at the points in the test set (that is, these are the evaluation points in Algorithm 2). Note that this particular application is especially challenging because the seven morphological statistics have very different properties, and some of the statistics (M and I) are essentially of mixed discrete and continuous type with heavy outliers; hence, any metric-based estimator is bound to perform poorly even after normalizing the variables. Our regression test, however, can by-pass this problem by leveraging the random forest algorithm. Another advantage of using random forests is that the algorithm returns variable importance measures

that can help us identify *which* morphology statistics are the most important in distinguishing the two populations (Figure 7).

6.1. Analysis and result

According to our global two-sample test ($\widehat{\mathcal{T}}_{RF} = .188$, $p < .001$), there is a significant difference between the low-SFR and the high-SFR populations in terms of galaxy morphology. We follow up on this result by implementing the local two-sample testing framework according to Algorithm 2 with FWER control at $\alpha = 0.05$ by the Hochberg step up procedure. To visualize locally significant points from the local test, we use diffusion maps with local scaling (Zelnik-Manor and Perona, 2005). For more information on our particular application of diffusion maps, see Appendix B. The main result of the local significance test is displayed in Figure 1. As we can see, the high-SFR and low-SFR dominated regions (that is, the regions where $f_{\text{LowSFR}} < f_{\text{HighSFR}}$ and $f_{\text{LowSFR}} > f_{\text{HighSFR}}$, respectively) are fairly well-separated in morphology space. Figure 1 also shows some examples of galaxy images at significant test points. By inspecting such images, we note that the "red" galaxies in the low-SFR dominated regions of the seven-dimensional space tend to be more concentrated and less disturbed than their "blue" counterparts in the high-SFR dominated regions — this result is consistent with previous astronomical studies about irregular galaxies displaying merger activities and high star-formation rates. Our test result is further supported by the variable importance measures in Figure 7: the two most important morphology statistics in distinguishing between high-SFR and low-SFR galaxies are the *Gini* (Lotz et al., 2004) and I (Freeman et al., 2013) morphology statistics. Indeed, by definition, the *Gini* statistic describes the variance of a galaxy's stellar light distribution, and the I statistic captures galaxies with disturbed morphologies.

7. Conclusions

In this work, we presented a new framework for both global and local two-sample testing via regression. Depending on the chosen regression model, our framework can efficiently deal with different types of variables and different structures in the data; thereby, providing tests with competitive power against many practical alternatives. Compared to other recent approaches in the two-sample literature (such as classification tests), our framework has the key advantage of being able to detect locally significant regions in multivariate spaces. Throughout this work, we studied theoretical properties of the regression tests by building on existing regression results. We established a connection between the power of the global and local tests to the MISE and MSE of the corresponding regression estimators, and we demonstrated practical usefulness of our methods via simulations.

By taking advantage of permutation tests under the global null hypothesis, the proposed local testing framework ensures that the type I error rate is less

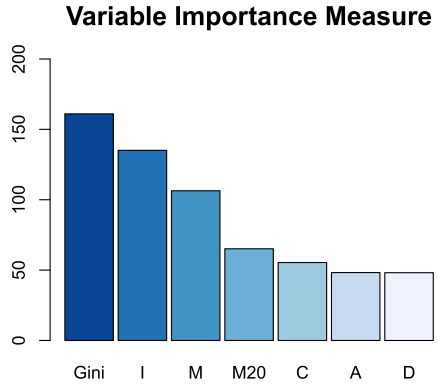

FIG 7. *Variable importance measures from random forest regression, as measured by the Mean Decrease Gini (MDG) metric when splitting the data along the indicated variables. For the morphology-SFR study, the Gini and I morphology statistics are the two most important features in distinguishing between high-star-forming and the low-star-forming galaxy populations.*

than or equal to the significance level. When the local null hypothesis $H_0(x) : m(x) = \pi$ is of interest, on the other hand, there is no such guarantee. In this case, it would be necessary to use an asymptotic framework and investigate the limiting behavior of a local test statistic. This topic is reserved for future work. Another direction for future work is to study the optimality of global regression tests. Contrary to the local regression test, a regression estimator with the optimal estimation error rate may not necessarily return minimax optimal global regression test. We hope that future studies will establish a lower bound and matching upper bound for the global regression test.

Appendix A: Proofs

A.1. Proof of Theorem 3.1

We start by simplifying $\widehat{m}_{\mathrm{LDA}}(x)$ as

$$\begin{aligned}
&\widehat{m}_{\mathrm{LDA}}(X_i) \\
&= \frac{\pi_1 \exp\left\{-\frac{1}{2}(X_i-\widehat{\mu}_1)^\top \mathcal{S}^{-1}(X_i-\widehat{\mu}_1)\right\}}{\pi_1 \exp\left\{-\frac{1}{2}(X_i-\widehat{\mu}_1)^\top \mathcal{S}^{-1}(X_i-\widehat{\mu}_1)\right\}+\pi_0 \exp\left\{-\frac{1}{2}(X_i-\widehat{\mu}_0)^\top \mathcal{S}^{-1}(X_i-\widehat{\mu}_0)\right\}} \\
&= \frac{\pi_1}{\pi_1+\pi_0 \exp\left\{-\frac{1}{2}(X_i-\widehat{\mu}_0)^\top \mathcal{S}^{-1}(X_i-\widehat{\mu}_0)+\frac{1}{2}(X_i-\widehat{\mu}_1)^\top \mathcal{S}^{-1}(X_i-\widehat{\mu}_1)\right\}} \\
&= \frac{\pi_1}{\pi_1+\pi_0 \exp\left\{(X_i-(\widehat{\mu}_0+\widehat{\mu}_1)/2)^\top \mathcal{S}^{-1}(\widehat{\mu}_0-\widehat{\mu}_1)\right\}}
\end{aligned}$$

and write

$$W_i = (X_i - (\widehat{\mu}_0 + \widehat{\mu}_1)/2)^\top \mathcal{S}^{-1} (\widehat{\mu}_0 - \widehat{\mu}_1).$$

For some $a \in (0,1)$, Taylor expansion of $f(x) = a/\{a + (1-a)e^x\}$ at $x = 0$ provides

$$\left|\left\{\widehat{m}_{\mathrm{LDA}}(X_i) - \pi_1\right\}^2 - \pi_0^2\pi_1^2 W_i^2\right| \le C|W_i|^3,$$

where C is a universal constant. This implies that

$$\left|\sum_{i=1}^n \left\{\widehat{m}_{\mathrm{LDA}}(X_i) - \pi_1\right\}^2 - \pi_0^2\pi_1^2 \sum_{i=1}^n W_i^2\right| \;\le\; C\sum_{i=1}^n |W_i|^3.$$

Now based on $|x+y|^3 \le 4|x|^3 + 4|y|^3$ and Cauchy-Schwarz inequality, it can be seen that

$$\sum_{i=1}^n |W_i|^3 \le 4n\left|((\widehat{\mu}_0+\widehat{\mu}_1)/2)^\top \mathcal{S}^{-1}(\widehat{\mu}_0-\widehat{\mu}_1)\right|^3 + 4\sum_{i=1}^n \left|X_i^\top \mathcal{S}^{-1}(\widehat{\mu}_0-\widehat{\mu}_1)\right|^3 = o_P(1).$$

As a result, $n\widehat{\mathcal{T}}_{\mathrm{LDA}}$ can be approximated by

$$n\widehat{\mathcal{T}}_{\mathrm{LDA}} = \sum_{i=1}^n \left\{\widehat{m}_{\mathrm{LDA}}(X_i) - \pi_1\right\}^2 = \pi_0^2\pi_1^2 \sum_{i=1}^n W_i^2 + o_P(1). \tag{22}$$

Let us denote $\delta_n = \mathcal{S}^{-1}(\widehat{\mu}_0 - \widehat{\mu}_1)$ and $\Delta_n = (\widehat{\mu}_0 + \widehat{\mu}_1)/2$, and recall $\mathcal{S} = n^{-1}\sum_{i=1}^n (X_i - \widehat{\mu})(X_i - \widehat{\mu})^\top$ where $\widehat{\mu} = n^{-1}\sum_{i=1}^n X_i$. Then we observe that

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n W_i^2 \;&=\; \frac{1}{n}\sum_{i=1}^n \left\{\delta_n^\top X_i - \delta_n^\top \Delta_n\right\}^2 \\
&=\; \delta_n^\top \left\{\frac{1}{n}\sum_{i=1}^n (X_i - \Delta_n)(X_i - \Delta_n)^\top\right\}\delta_n \\
&=\; \delta_n^\top \mathcal{S}\delta_n + \delta_n^\top (\widehat{\mu} - \Delta_n)(\widehat{\mu} - \Delta_n)^\top \delta_n \\
&=\; (\widehat{\mu}_0 - \widehat{\mu}_1)^\top \mathcal{S}^{-1}(\widehat{\mu}_0 - \widehat{\mu}_1) + R_n,
\end{aligned}$$

where $R_n = \delta_n^\top (\widehat{\mu} - \Delta_n)(\widehat{\mu} - \Delta_n)^\top \delta_n$. Hence, we have

$$n\widehat{\mathcal{T}}_{\mathrm{LDA}} = n\pi_0^2\pi_1^2\left\{(\widehat{\mu}_0 - \widehat{\mu}_1)^\top \mathcal{S}^{-1}(\widehat{\mu}_0 - \widehat{\mu}_1) + R_n\right\} + o_P(1).$$

We also note that the residual term is negligible under the null, i.e. $n\pi_0^2\pi_1^2 R_n = o_P(1)$, which results in

$$\begin{aligned}
n\pi_0^{-1}\pi_1^{-1}\widehat{\mathcal{T}}_{\mathrm{LDA}} &= \frac{n_0 n_1}{n_0 + n_1}(\widehat{\mu}_0 - \widehat{\mu}_1)^\top \mathcal{S}^{-1}(\widehat{\mu}_0 - \widehat{\mu}_1) + o_P(1) \\
&= T^2_{\mathrm{Hotelling}} + o_P(1).
\end{aligned}$$

The rest of the proof follows by the limiting property of Hotelling's T^2.

A.2. Proof of Theorem 3.2

Proof. First note that the likelihood ratio for testing (8) is given by

$$\mathcal{L}_n = \sum_{i=1}^{n_1} \log \frac{f_{\mu_0+h/\sqrt{n}}(X_{i,1})}{f_{\mu_0}(X_{i,1})}.$$

Since $\{\mathbb{P}_\mu, \mu \in \Omega\}$ is q.m.d. at μ_0, Theorem 12.2.3 of Lehmann and Romano (2006) under $n_1/(n_0+n_1) \to \pi_1$ yields that

$$\mathcal{L}_n \xrightarrow{d} N\left(-\frac{\pi_1}{2}\langle h, I(\mu_0)h\rangle,\ \pi_1\langle h, I(\mu_0)h\rangle\right),$$

where $I(\mu)$ is the Fisher information matrix. This implies by Corollary 12.3.1 of Lehmann and Romano (2006) that the joint distribution of $X_{1,0}$ and $X_{1,1}$ under the null and the alternative are mutually contiguous. Since contiguity implies

$$n\pi_0^{-1}\pi_1^{-1}\widehat{\mathcal{T}}_{\mathrm{LDA}} = \frac{n_0 n_1}{n_0+n_1}(\widehat{\mu}_0 - \widehat{\mu}_1)^\top \mathcal{S}^{-1}(\widehat{\mu}_0 - \widehat{\mu}_1) + o_P(1),$$

under $H_{1,n}$, the result follows by the limiting distribution of Hotelling's T^2 statistic. □

A.3. Proof of Theorem 3.3

Proof. The exact type I error control of the permutation test is well-known (see e.g. Chapter 15 of Lehmann and Romano, 2006). Strictly speaking, the considered test is not the usual permutation test since the only first half of labels are permuted to decide a critical value. However, it also controls the type I error under H_0 due to Theorem 15.2.1 of Lehmann and Romano (2006). Indeed, this result holds regardless of i.i.d. sampling or separate sampling. Hence we focus on the type II error control.

• Type II error control (i.i.d. sampling)

We start with the case of i.i.d. sampling. Based on the inequality $(x-y)^2 \le 2(x-z)^2 + 2(z-y)^2$, we lower bound the test statistic as

$$\begin{aligned}
\widehat{\mathcal{T}}'_{global} &= \frac{1}{n}\sum_{i=n+1}^{2n}(\widehat{m}(X_i) - \widehat{\pi}_1)^2 \\
&\ge \frac{1}{2n}\sum_{i=n+1}^{2n}(m(X_i) - \widehat{\pi}_1)^2 - \frac{1}{n}\sum_{i=n+1}^{2n}(\widehat{m}(X_i) - m(X_i))^2 \\
&\ge \frac{1}{4n}\sum_{i=n+1}^{2n}(m(X_i) - \pi_1)^2 - \frac{1}{2}(\pi_1 - \widehat{\pi}_1)^2 - \frac{1}{n}\sum_{i=n+1}^{2n}(\widehat{m}(X_i) - m(X_i))^2 .
\end{aligned} \tag{23}$$

Define the events $\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3, \mathcal{A}_4$ such that

$$\mathcal{A}_1 = \left\{(\pi_1 - \widehat{\pi}_1)^2 < C_2\delta_n\right\},$$

$$\mathcal{A}_2 = \left\{\frac{1}{n}\sum_{i=n+1}^{2n}(\widehat{m}(X_i) - m(X_i))^2 < C_3\delta_n\right\},$$

$$\mathcal{A}_3 = \left\{\left|\frac{1}{n}\sum_{i=n+1}^{2n}(m(X_i) - \pi_1)^2 - \mathbb{E}\left[(m(X_i) - \pi_1)^2\right]\right| < \frac{1}{2}\mathbb{E}\left[(m(X) - \pi_1)^2\right]\right\},$$

$$\mathcal{A}_4 = \left\{t_\alpha < C'_{0,\alpha}\delta_n\right\}.$$

Using Markov's inequality, we have

$$\mathbb{P}(\mathcal{A}_1^c) \le \frac{\pi_1(1-\pi_1)}{C_2 n\delta_n},$$

$$\mathbb{P}(\mathcal{A}_2^c) \le \frac{1}{C_3\delta_n}\mathbb{E}\left[\int_S (\widehat{m}(x) - m(x))^2 dP_X(x)\right] \le \frac{C_0}{C_3},$$

by the condition in (9). For the third event, denote $\Delta_n = \mathbb{E}\left[(m(X) - \pi_1)^2\right]$ and use Chebyshev's inequality to have

$$\begin{aligned}
\mathbb{P}(\mathcal{A}_3^c) &\le \frac{4}{n\Delta_n^2}\mathrm{Var}\left[(m(X) - \pi_1)^2\right] \\
&\le \frac{4}{n\Delta_n^2}\mathbb{E}\left[(m(X) - \pi_1)^4\right] \\
&\le \frac{4}{n\Delta_n^2}\mathbb{E}\left[(m(X) - \pi_1)^2\right] \quad \text{since } |m(X) - \pi_1| \le 1 \\
&\le \frac{4}{C_1 n\delta_n},
\end{aligned}$$

where the last inequality uses the assumption that $\Delta_n \ge C_1\delta_n$. Furthermore, under the assumption on the permutation critical value, $\mathbb{P}(\mathcal{A}_4^c) \le \beta/2$. Hence, we obtain

$$\mathbb{P}\left((\mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{A}_3 \cap \mathcal{A}_4)^c\right) \le \sum_{i=1}^{4}\mathbb{P}(\mathcal{A}_i^c) < \beta,$$

by choosing sufficiently large $C_1, C_2, C_3 > 0$ with the assumption that $\delta_n \ge n^{-1}$. Using (23), the type II error of the regression test is bounded by

$$\begin{aligned}
&\mathbb{P}(\widehat{\mathcal{T}}'_{global} \le t_\alpha) \\
&\le \mathbb{P}\left(\frac{1}{4n}\sum_{i=n+1}^{2n}(m(X_i) - \pi_1)^2 - \frac{1}{2}(\pi_1 - \widehat{\pi}_1)^2 - \frac{1}{n}\sum_{i=n+1}^{2n}(\widehat{m}(X_i) - m(X_i))^2 \le t_\alpha\right)
\end{aligned}$$

$$\leq \mathbb{P}\left(\left\{\frac{1}{4n}\sum_{i=n+1}^{2n}(m(X_i)-\pi_1)^2-\frac{1}{2}(\pi_1-\widehat{\pi}_1)^2-\frac{1}{n}\sum_{i=n+1}^{2n}(\widehat{m}(X_i)-m(X_i))^2 \leq t_\alpha\right\}\right.$$
$$\left.\bigcap\left\{\bigcap_{j=1}^{4}\mathcal{A}_j\right\}\right)+\mathbb{P}\left((\mathcal{A}_1\cap\mathcal{A}_2\cap\mathcal{A}_3\cap\mathcal{A}_4)^c\right)$$
$$\leq \mathbb{P}\left(\Delta_n < C_4\delta_n\right)+\beta,$$

where C_4 can be chosen by $C_4 = 2C'_{0,\alpha} + C_2 + 2C_3$. Now by choosing $C_1 > C_4$ for sufficiently large n, the type II error can be bounded by β. Hence the result follows.

- **Type II error control (Separate Sampling)**

The proof for separate sampling is almost the same as before except few details. First, we do not need to define $\mathcal{A}_1$ since π_1 is known. In terms of $\mathcal{A}_2$, apply Markov's inequality to obtain

$$\begin{aligned}\mathbb{P}\left(\mathcal{A}_2^c\right) &\leq \frac{1}{C_3\delta_n}\left\{\frac{n_0}{n}\mathbb{E}\left[\int_S(\widehat{m}(x)-m(x))^2dP_0(x)\right]\right.\\ &\qquad\left.+\frac{n_1}{n}\mathbb{E}\left[\int_S(\widehat{m}(x)-m(x))^2dP_1(x)\right]\right\}\\ &= \frac{C_0}{C_3}\mathbb{E}\left[\int_S(\widehat{m}(x)-m(x))^2dP_X(x)\right]\leq\frac{C_0}{C_3},\end{aligned}$$

where the last line uses the fact that $\frac{n_0}{n}P_0+\frac{n_1}{n}P_1 = P_X$. Similarly, for the event $\mathcal{A}_3$, we have by Chebyshev's inequality that

$$\begin{aligned}\mathbb{P}\left(\mathcal{A}_3^c\right) &\leq \frac{4}{\Delta_n^2}\frac{1}{n^2}\sum_{i=n+1}^{2n}\mathrm{Var}\left[(m(X_i)-\pi_1)^2\right]\\ &\leq \frac{4}{\Delta_n^2}\frac{1}{n^2}\sum_{i=n+1}^{2n}\mathbb{E}\left[(m(X_i)-\pi_1)^2\right]=\frac{4}{n\Delta_n^2}\mathbb{E}\left[(m(X)-\pi_1)^2\right]\\ &\leq \frac{4}{C_1 n\delta_n}.\end{aligned}$$

The rest follows exactly the same as before. Hence the proof is complete. □

A.4. Proof of Corollary 3.1

Proof. We prove the corollary by showing that the conditions in Theorem 3.3 are satisfied. In particular, it suffices to verify that for fixed $\alpha \in (0,1)$ and $\beta \in$

$(0, 1-\alpha)$, there exists a positive constant $C'_{0,\alpha}$ such that $\sup_{f_0,f_1\in\mathcal{M}} \mathbb{P}_{f_0,f_1}(t_\alpha < C'_{0,\alpha}\delta_n) \geq 1-\beta/2$. Then the rest of the proof proceeds the same as before.

• i.i.d. sampling

To start with the case of i.i.d. sampling, let $\eta = (\eta_1, \ldots, \eta_n)^\top$ be a permutation of $\{1, \ldots, n\}$. Now conditioned on the data $\mathcal{X}_{2n} = \{(X_1, Y_1), \ldots, (X_{2n}, Y_{2n})\}$, we denote the probability and expectation under permutations by $\mathbb{P}_\eta[\cdot] = \mathbb{P}_\eta[\cdot|\mathcal{X}_{2n}]$ and $\mathbb{E}_\eta[\cdot] = \mathbb{E}_\eta[\cdot|\mathcal{X}_{2n}]$ respectively. Then by Markov's inequality

$$\begin{aligned}\mathbb{P}_\eta\left(\widehat{\mathcal{T}}'_{global} \geq t\right) &= \mathbb{P}_\eta\left(\frac{1}{n}\sum_{i=n+1}^{2n}(\widehat{m}_\eta(X_i) - \widehat{\pi}_1)^2 \geq t\right) \\ &\leq \frac{1}{tn}\sum_{i=n+1}^{2n}\mathbb{E}_\eta\left[(\widehat{m}_\eta(X_i) - \widehat{\pi}_1)^2\right],\end{aligned}$$

where $\widehat{m}_\eta(x) = \sum_{i=1}^n w_i(x) Y_{\eta_i}$. Since $\sum_{i=1}^n w_i(x) = 1$ for any $x \in S$,

$$\mathbb{E}_\eta\left[\widehat{m}_\eta(x)\right] = \sum_{i=1}^n w_i(x)\mathbb{E}_\eta[Y_{\eta_i}] = \sum_{i=1}^n w_i(x)\widehat{\pi}_1 = \widehat{\pi}_1.$$

Further note that

$$\begin{aligned}\mathbb{E}_\eta\left[(\widehat{m}_\eta(x) - \widehat{\pi}_1)^2\right] &= \sum_{i_1=1}^n\sum_{i_2=1}^n w_{i_1}(x)w_{i_2}(x)\mathbb{E}_\eta\left[(Y_{\eta_{i_1}} - \widehat{\pi}_1)(Y_{\eta_{i_2}} - \widehat{\pi}_1)\right] \quad (24)\\ &\leq \sum_{i=1}^n w_i^2(x)\mathbb{E}_\eta\left[(Y_{\eta_i} - \widehat{\pi}_1)^2\right] \\ &= \widehat{\pi}_1(1-\widehat{\pi}_1)\sum_{i=1}^n w_i^2(x) \\ &\leq \frac{1}{4}\sum_{i=1}^n w_i^2(x), \quad (25)\end{aligned}$$

where the first inequality uses $\mathbb{E}_\eta\left[(Y_{\eta_{i_1}} - \widehat{\pi}_1)(Y_{\eta_{i_2}} - \widehat{\pi}_1)\right] \leq 0$ when $i_1 \neq i_2$.

Note that the permutation samples are not *i.i.d.* and thus in order to use the condition in (9) which holds for *i.i.d.* samples, we will associate the upper bound in (25) with *i.i.d.* samples. To do so, let $(Y_1^*, \ldots, Y_n^*)$ be *i.i.d.* Bernoulli random variables with parameter $p = 1/2$ independent of $\{X_1, \ldots, X_{2n}\}$. Then

$$\begin{aligned}&\mathbb{E}_{Y^*}\left[(\widehat{m}(x) - 1/2)^2 | X_1, \ldots, X_{2n}\right] \\ =\ &\mathbb{E}_{Y^*}\Big[\big(\sum_{i=1}^n w_i(x)Y_i^* - 1/2\big)^2 \big| X_1, \ldots, X_{2n}\Big]\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}_{Y^*}\Big[\big(\sum_{i=1}^{n} w_i(x)(Y_i^* - 1/2)\big)^2 \big| X_1, \ldots, X_{2n}\Big] \\
&= \sum_{i_1=1}^{n}\sum_{i_2=1}^{n} w_{i_1}(x) w_{i_2}(x) \mathbb{E}_{Y^*}[(Y_{i_1}^* - 1/2)(Y_{i_2}^* - 1/2)] \\
&= \frac{1}{4}\sum_{i=1}^{n} w_i^2(x).
\end{aligned}
$$

Therefore, we obtain

$$
\mathbb{E}_\eta\left[(\widehat{m}_\eta(x) - \widehat{\pi}_1)^2\right] \leq \mathbb{E}_{Y^*}\left[(\widehat{m}(x) - 1/2)^2 | X_1, \ldots, X_{2n}\right]
$$

which in turn implies that

$$
\mathbb{P}_\eta\left(\widehat{\mathcal{T}}'_{global} \geq t\right) \leq \frac{1}{tn}\sum_{i=n+1}^{2n} \mathbb{E}_{Y^*}\left[(\widehat{m}(X_i) - 1/2)^2 | X_1, \ldots, X_{2n}\right].
$$

So the critical value of the permutation distribution is bounded by

$$
t_\alpha^* \leq \frac{1}{\alpha n}\sum_{i=n+1}^{2n} \mathbb{E}_{Y^*}\left[(\widehat{m}(X_i) - 1/2)^2 | X_1, \ldots, X_{2n}\right]. \tag{26}
$$

Now choose $C'_{0,\alpha}$ such that $2C_0/(\alpha\beta) \leq C'_{0,\alpha}$. Then based on the assumption in (9) and Markov's inequality

$$
\begin{aligned}
&\sup_{f_0, f_1 \in \mathcal{M}} \mathbb{P}_{f_0,f_1}\left(t_\alpha^* \geq C'_{0,\alpha}\delta_n\right) \\
\leq &\sup_{f_0, f_1 \in \mathcal{M}} \mathbb{P}_{f_0,f_1}\left(\frac{1}{\alpha n}\sum_{i=n+1}^{2n} \mathbb{E}_{Y^*}\left[(\widehat{m}(X_i) - 1/2)^2 | X_1, \ldots, X_{2n}\right] \geq C'_{0,\alpha}\delta_n\right) \\
\leq &\frac{C_0}{C'_{0,\alpha}\alpha} \leq \beta/2.
\end{aligned}
$$

Hence the proof completes.

• Separate Sampling

Let $Y_1^{**}, \ldots, Y_n^{**}$ be Bernoulli random variables with parameter $\widehat{\pi}_1$ such that $\sum_{i=1}^n Y_i^{**} = n\widehat{\pi}_1$ and they are independent of $X_1, \ldots, X_{2n}$. In the case of separate sampling, the proof follows similarly by noting that the right-hand side of (24) is the same as

$$
\sum_{i_1=1}^{n}\sum_{i_2=1}^{n} w_{i_1}(x) w_{i_2}(x) \mathbb{E}_\eta\left[(Y_{\eta_{i_1}} - \widehat{\pi}_1)(Y_{\eta_{i_2}} - \widehat{\pi}_1)\right]
$$

$$= \sum_{i_1=1}^{n}\sum_{i_2=1}^{n} w_{i_1}(x)w_{i_2}(x)\mathbb{E}_{Y^{**}}\left[(Y_{i_1}^{**}-\widehat{\pi}_1)(Y_{i_2}^{**}-\widehat{\pi}_1)\right]$$

$$= \mathbb{E}_{Y^{**}}[(\widehat{m}(x)-\widehat{\pi}_1)^2|X_1,\ldots,X_n].$$

Now by putting the above quantity into the right-hand side of (26) and following the same lines afterwards, we complete the proof. □

A.5. Proof of Theorem 4.1

This result can be proved by following the same steps in the proof of Theorem 3.3. In fact, it is simpler than the previous proof since it does not involve sample splitting to estimate the integration error; hence we omit the proof.

A.6. Proof of Example 4.1

Proof. Let $\overline{m}_{kNN}(x) = \mathbb{E}[\widehat{m}_{kNN}(x)|X_1,\ldots,X_n]$. Then we have the following decomposition.

$$\mathbb{E}\left[(\widehat{m}_{kNN}(x)-m(x))^2\right]$$
$$= \underbrace{\mathbb{E}\left[(\widehat{m}_{kNN}(x)-\overline{m}_{kNN}(x))^2\right]}_{(I)} + \underbrace{\mathbb{E}\left[(\overline{m}_{kNN}(x)-m(x))^2\right]}_{(II)}.$$

For a fixed x, Proposition 8.1 of Biau and Devroye (2015) shows that conditioned on $\{X_1,\ldots,X_n\}$,

$$(X_{1,n}(x),Y_{1,n}(x)),\ldots,((X_{n,n}(x),Y_{n,n}(x)))$$

are independent. Using this independence property,

$$(I) = \mathbb{E}\left[\left(\frac{1}{k_n}\sum_{i=1}^{k_n}(Y_{i,n}(x)-m(X_{i,n}(x)))\right)^2\right] \le \frac{1}{4k_n}.$$

Next for (II),

$$(II) = \mathbb{E}\left[\left(\frac{1}{k_n}\sum_{i=1}^{k_n}(m(X_{i,n}(x))-m(x))\right)^2\right]$$
$$\le \mathbb{E}\left[\left(\frac{1}{k_n}\sum_{i=1}^{k_n}\left|m(X_{i,n}(x))-m(x)\right|\right)^2\right]$$
$$\le \mathbb{E}\left[\left(\frac{L}{k_n}\sum_{i=1}^{k_n}||X_{i,n}(x)-x||_2\right)^2\right]$$

where the last inequality uses the Lipschitz condition. Note that for fixed $\epsilon > 0$

$$\begin{aligned}\mathbb{P}\left(||X_{1,n}(x) - x||_2 > \epsilon\right) &= \left(1 - \mathbb{P}(X \in B_{x,\epsilon})\right)^n \\ &\leq (1 - \tau_x \epsilon^D)^n \leq \ e^{-\tau_x n \epsilon^D}\end{aligned} \tag{27}$$

by the assumption that $\mathbb{P}(X \in B_{x,\epsilon}) > \tau_x \epsilon^D$. Hence,

$$\begin{aligned}\mathbb{E}\left[||X_{1,n}(x) - x||^2\right] &= \int_0^\infty \mathbb{P}\left(||X_{1,n}(x) - x||_2 > \sqrt{\epsilon}\right) d\epsilon \\ &\leq \int_0^\infty e^{-\tau_x n \epsilon^{D/2}} d\epsilon \\ &= \frac{2\Gamma(2/D)}{D\tau_x^{2/D}} n^{-2/D}.\end{aligned} \tag{28}$$

Similarly to the proof of Theorem 6.2 of Györfi et al. (2002), divide the data into $k_n + 1$ parts where the first k_n parts have size $\lfloor n/k_n \rfloor$ and denote the first nearest neighbor of x from the jth partition by $\widetilde{X}_j^x$. This implies that

$$\sum_{i=1}^{k_n} ||X_{i,n}(x) - x||_2 \leq \sum_{i=1}^{k_n} ||\widetilde{X}_i^x - x||_2$$

and by Jensen's inequality,

$$\begin{aligned}(II) \leq \ \mathbb{E}\left[\left(\frac{L}{k_n}\sum_{i=1}^{k_n} ||\widetilde{X}_i^x - x||_2\right)^2\right] &\leq \ \frac{L^2}{k_n}\sum_{i=1}^{k_n}\mathbb{E}\left[||\widetilde{X}_i^x - x||_2^2\right] \\ &\leq \ L^2 \frac{2\Gamma(2/D)}{D\tau_x^{2/D}}\left(\frac{k_n}{n}\right)^{2/D}\end{aligned}$$

by the inequality (28). Combining the results, we have

$$\begin{aligned}\mathbb{E}\left[\left(\widehat{m}_{kNN}(x) - m(x)\right)^2\right] &= \ (I) + (II) \\ &\leq \ \frac{1}{4k_n} + L^2 \frac{2\Gamma(2/D)}{D\tau_x^{2/D}}\left(\frac{k_n}{n}\right)^{2/D}.\end{aligned}$$

This completes the proof. □

A.7. Proof of Example 4.2

Proof. Following the proof of Example 4.1, let

$$\overline{m}_{ker}(x) = \mathbb{E}\left[\widehat{m}_{ker}(x) | X_1, \ldots, X_n\right]$$

and thus

$$\mathbb{E}\left[(\widehat{m}_{ker}(x)-m(x))^2\right]=\underbrace{\mathbb{E}\left[(\widehat{m}_{ker}(x)-\overline{m}_{ker}(x))^2\right]}_{(I)}+\underbrace{\mathbb{E}\left[(\overline{m}_{ker}(x)-m(x))^2\right]}_{(II)}.$$

Define an event

$$\mathcal{A}_n=\left\{\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right)\geq\lambda\right\}.$$

Then

$$(I)=\underbrace{\mathbb{E}\left[(\widehat{m}_{ker}(x)-\overline{m}_{ker}(x))^2 I(\mathcal{A}_n)\right]}_{(I_1)}+\underbrace{\mathbb{E}\left[(\widehat{m}_{ker}(x)-\overline{m}_{ker}(x))^2 I(\mathcal{A}_n^c)\right]}_{(I_2)}.$$

For (I_1), we have

$$\begin{aligned}
&\mathbb{E}\left[(\widehat{m}_{ker}(x)-\overline{m}_{ker}(x))^2 I(\mathcal{A}_n)|X_1,\ldots,X_n\right]\\
=&\ \frac{\sum_{i=1}^n \mathrm{Var}(Y_i|X_i)K\left(\frac{x-X_i}{h_n}\right)}{\left(\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right)\right)^2}I(\mathcal{A}_n)\\
\leq&\ \frac{1}{4\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right)}I\left(\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right)\geq\lambda\right)\\
\leq&\ \frac{1+\lambda^{-1}}{4+4\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right)}\\
\leq&\ \frac{1+\lambda^{-1}}{4+4\lambda\sum_{i=1}^n I\left(||x-X_i||_2\leq rh_n\right)}\\
\leq&\ \frac{1+\lambda}{4\lambda^2}\frac{1}{1+\sum_{i=1}^n I\left(||x-X_i||_2\leq rh_n\right)}.
\end{aligned}$$

By Lemma 4.1 of Györfi et al. (2002),

$$\mathbb{E}\left[\frac{1}{1+B}\right]\leq\frac{1}{(n+1)p}\leq\frac{1}{np},$$

where $B\sim\mathrm{Binominal}(n,p)$. Using this result,

$$\begin{aligned}
(I_1)\leq&\ \frac{1+\lambda}{4\lambda^2}\frac{1}{n\mathbb{P}\left(X\in B_{x,rh_n}\right)}\\
\leq&\ \left(\frac{1+\lambda}{4\lambda^2\tau_x r^d}\right)\frac{1}{nh_n^d}.
\end{aligned}$$

For (I_2), note that $\left(\widehat{m}_{ker}(x) - \overline{m}_{ker}(x)\right)^2 \leq 1$ and thus

$$\begin{aligned}(I_2) \leq & \ \mathbb{P}\left(\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right) < \lambda\right) \\ \leq & \ \mathbb{P}\left(\sum_{i=1}^{n} I\left(||x-X_i||_2 \leq rh_n\right) = 0\right)\end{aligned}$$

where the second inequality is because if there exists X_i such that $||x-X_i||_2 \leq rh_n$, then $\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right) \geq \lambda$ by the assumption on the kernel. In addition,

$$\begin{aligned}\mathbb{P}\left(\sum_{i=1}^{n} I\left(||x-X_i||_2 \leq rh_n\right) = 0\right) = & \ \left(1 - \mathbb{P}\left(X \in B_{x,rh_n}\right)\right)^n \\ \overset{(i)}{\leq} & \ e^{-n\tau_x r^D h_n^D} \overset{(ii)}{\leq} \left(\frac{e^{-1}}{\tau_x r^D}\right)\frac{1}{nh_n^D},\end{aligned} \tag{29}$$

where (i) uses $1 + x \leq e^x$ with the assumption $\mathbb{P}\left(X \in B_{x,\epsilon}\right) \geq \tau_x \epsilon^D$ and (ii) uses $\sup_z ze^{-z} \leq e^{-1}$. As a result,

$$(I) = (I_1) + (I_2) \leq \left(\frac{1+\lambda}{4\lambda^2 \tau_x r^D} + \frac{e^{-1}}{\tau_x r^D}\right)\frac{1}{nh_n^D}.$$

For (II), we use Jensen's inequality and the Lipschitz condition to have

$$\begin{aligned}& \left(\overline{m}_{ker}(x) - m_{ker}(x)\right)^2 \\ = & \ \left(\frac{\sum_{i=1}^{n}\left(m(X_i) - m(x)\right)K\left(\frac{x-X_i}{h_n}\right)}{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right)}\right)^2 I\left(\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right) > 0\right) \\ & + m_{ker}(x)^2 I\left(\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right) = 0\right) \\ \leq & \ \frac{\sum_{i=1}^{n} L^2 ||X_i - x||_2^2 K\left(\frac{x-X_i}{h_n}\right)}{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right)} I\left(\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right) > 0\right) \\ & + I\left(\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right) = 0\right).\end{aligned}$$

Since $K(x) \leq I(x \in B_{0,R})$, we observe that

$$||X_i - x||_2^2 K\left(\frac{x-X_i}{h_n}\right) \leq R^2 h_n^2 K\left(\frac{x-X_i}{h_n}\right).$$

Consequently,

$$\begin{aligned}\left(\overline{m}_{ker}(x) - m_{ker}(x)\right)^2 \leq\ & L^2R^2h_n^2 + I\left(\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right) = 0\right) \\ \leq\ & L^2R^2h_n^2 + I\left(\sum_{i=1}^{n} I\left(||x - X_i||_2 \leq rh_n\right) = 0\right),\end{aligned}$$

where the second inequality is by the assumption $\lambda I(x \in B_{0,r}) \leq K(x)$. By taking the expectation,

$$\begin{aligned}(II) \leq\ & L^2R^2h_n^2 + \left(1 - \mathbb{P}\left(X \in B_{x,rh_n}\right)\right)^n \\ \leq\ & L^2R^2h_n^2 + \left(1 - \tau_x r^D h_n^D\right)^n \\ \leq\ & L^2R^2h_n^2 + \left(\frac{e^{-1}}{\tau_x r^D}\right)\frac{1}{nh_n^D}.\end{aligned} \tag{30}$$

Therefore, we conclude that

$$\begin{aligned}\mathbb{E}\left[\left(\widehat{m}_{ker}(x) - m(x)\right)^2\right] =\ & (I) + (II) \\ \leq\ & \left(\frac{1+\lambda}{4\lambda^2\tau_x r^D} + \frac{2e^{-1}}{\tau_x r^D}\right)\frac{1}{nh_n^D} + L^2R^2h_n^2,\end{aligned}$$

which completes the proof. □

A.8. Proof of Theorem 4.2

Proof. Suppose X has the uniform distribution over $[0, B]^D$ and $B > 0$. In addition, assume that for $0 < \epsilon < 1/2$, the regression function is given by

$$\begin{aligned}m(x) =\ & \epsilon\prod_{i=1}^{D}\left(1 - \frac{x_i}{B\epsilon}\right) I\left(0 \leq x_i \leq B\epsilon\right) \\ & + \epsilon\prod_{i=1}^{D}\left(\frac{B(1-\epsilon) - x_i}{B\epsilon}\right) I\{B(1-\epsilon) \leq x_i \leq B\} + \frac{1}{2}\end{aligned} \tag{31}$$

for $x = (x_1, \ldots, x_D) \in [0, B]^D$ and $m(x) = 0$ otherwise. Therefore, we have $\pi_1 = \pi_0 = 1/2$. Now for any $x, z \in [0, B]^D$, the telescoping argument gives

$$\begin{aligned}& |m(x_1, \ldots, x_D) - m(z_1, \ldots, z_D)| \\ \leq\ & |m(x_1, x_2, \ldots, x_D) - m(z_1, x_2, \ldots, x_D)| \\ & + \sum_{i=1}^{D-2} |m(z_1, \ldots, z_i, x_{i+1}, \ldots, x_D) - m(z_1, \ldots, z_i, z_{i+1}, x_{i+2}, \ldots, x_D)|\end{aligned}$$

$$+\ |m(z_1, z_2, \ldots, z_{D-1}, x_D) - m(z_1, z_2, \ldots, z_D)|.$$

For the first term,

$$\begin{aligned}
&|m(x_1, x_2, \ldots, x_D) - m(z_1, x_2, \ldots, x_D)| \\
\leq\ & \epsilon \left| \left(1 - \frac{x_1}{B\epsilon}\right) I\left(0 \leq x_1 \leq B\epsilon\right) - \left(1 - \frac{z_1}{B\epsilon}\right) I\left(0 \leq z_1 \leq B\epsilon\right) \right| \\
& \times \prod_{i=2}^{D} \left| \left(1 - \frac{x_i}{B\epsilon}\right) I\left(0 \leq x_i \leq B\epsilon\right) \right| \\
& + \epsilon \left| \left(\frac{B(1-\epsilon) - x_1}{B\epsilon}\right) I\Big\{B(1-\epsilon) \leq x_1 \leq B\Big\} \right. \\
& \qquad \left. - \left(\frac{B(1-\epsilon) - z_1}{B\epsilon}\right) I\Big\{B(1-\epsilon) \leq z_1 \leq B\Big\} \right| \\
& \times \prod_{i=2}^{D} \left| \left(\frac{B(1-\epsilon) - x_i}{B\epsilon}\right) I\Big\{B(1-\epsilon) \leq x_i \leq B\Big\} \right| \\
\leq\ & \epsilon \left| \left(1 - \frac{x_1}{B\epsilon}\right) I\left(0 \leq x_1 \leq B\epsilon\right) - \left(1 - \frac{z_1}{B\epsilon}\right) I\left(0 \leq z_1 \leq B\epsilon\right) \right| \\
& + \epsilon \left| \left(\frac{B(1-\epsilon) - x_1}{B\epsilon}\right) I\Big\{B(1-\epsilon) \leq x_1 \leq B\Big\} \right. \\
& \qquad \left. - \left(\frac{B(1-\epsilon) - z_1}{B\epsilon}\right) I\Big\{B(1-\epsilon) \leq z_1 \leq B\Big\} \right| \\
\leq\ & \frac{2}{B}|x_1 - z_1| \leq \frac{2}{B}||x - z||_2.
\end{aligned}$$

Applying the same logic to the other terms, we see that

$$|m(x) - m(z)| \leq \frac{2D}{B}||x - z||_2.$$

By choosing $B = 2D/L$, the regression function $m(x)$ becomes L-Lipschitz with

$$\delta_{n,x} = |m(x) - \pi_1|^2 = \epsilon^2 \quad \text{at } x = (0, \ldots, 0).$$

Next, we lower bound the testing error. Denote the product and joint measure of (X, Y) described above by P_0 and P_1 respectively. Using the standard approach to lower bound the testing error (e.g. Baraud, 2002), we obtain that for any α level test functions $\phi : \{(X_1, Y_1), \ldots, (X_n, Y_n)\} \mapsto \{0, 1\}$,

$$\inf_{\phi \in \Phi_{n,\alpha}} \sup_{f_0, f_1 \in \mathcal{M}_{Lip}(\delta_{n,x})} \mathbb{P}_{f_0,f_1}(\phi = 0) \geq\ 1 - \alpha - \mathrm{TV}(P_0^n, P_1^n)$$

where TV denotes total variation distance. Based on Pinsker's inequality, we get

$$\mathrm{TV}(P_0^n, P_1^n) \leq \sqrt{\frac{n}{2} D_{KL}(P_1||P_0)}$$

where D_{KL} is the Kullback-Leibler divergence and by the Jensen's inequality

$$\begin{aligned}
& D_{KL}(P_1||P_0) \\
= & \int \pi_1 f(x) \log \frac{f(x, Y=1)}{\pi_1 f(x)} dx + \int (1-\pi_1) f(x) \log \frac{f(x, Y=0)}{(1-\pi_1) f(x)} dx \\
= & \frac{1}{2} \int f(x) \log \frac{f(x|Y=1)}{f(x)} dx + \frac{1}{2} \int f(x) \log \frac{f(x|Y=0)}{f(x)} dx \\
\leq & \frac{1}{2} \int \frac{(f(x|Y=1) - f(x))^2}{f(x)} dx + \frac{1}{2} \int \frac{(f(x|Y=0) - f(x))^2}{f(x)} dx.
\end{aligned}$$

By the assumption on (X, Y), X has the marginal density $f(x) = B^{-D}$ and the conditional densities $f(x|Y=1) = 2B^{-D} m(x)$ and $f(x|Y=0) = 2B^{-D} - f(x|Y=1)$ for $x \in [0, B]^D$. Therefore,

$$\begin{aligned}
& \frac{1}{2} \int \frac{(f(x|Y=1) - f(x))^2}{f(x)} dx + \frac{1}{2} \int \frac{(f(x|Y=0) - f(x))^2}{f(x)} dx \\
= & \int \frac{(f(x|Y=1) - f(x))^2}{f(x)} dx \\
= & 4B^{-D} \int (m(x) - 1/2)^2 \, dx.
\end{aligned}$$

Using the definition of $m(x)$ in (31), the above integration is calculated by

$$4B^{-D} \int (m(x) - 1/2)^2 \, dx = \frac{8}{3^D} \epsilon^{2+D}.$$

Now by choosing $\epsilon = \beta^{2/(2+D)} 3^{D/(2+D)} 2^{-2/(2+D)} n^{-1/(2+D)}$, we have

$$\inf_{\phi \in \Phi_{n,\alpha}} \sup_{f_0, f_1 \in \mathcal{M}_{Lip}(C_{1,x} n^{-2/(2+D)})} \mathbb{P}_{f_0, f_1}(\phi = 0) \geq 1 - \alpha - \beta.$$

This completes the proof. □

A.9. Proof of Proposition 4.1

It is enough to show that there exist universal constants $C_0, C'_{0,\alpha}$ such that

$$\sup_{f_0, f_1 \in \mathcal{M}_{Lip}} \mathbb{E}\left[(\widehat{m}_{kNN}(x) - m(x))^2 \right] \leq C_0 n^{-\frac{2}{2+d}},$$

$$\sup_{f_0,f_1\in\mathcal{M}_{Lip}} \mathbb{E}\left[(\widehat{m}_{ker}(x)-m(x))^2\right] \leq C'_{0,\alpha} n^{-\frac{2}{2+d}}.$$

Then we can apply Theorem 4.1 to complete the proof. To start with kNN regression, we only need to modify (27) and follow the same steps in the proof of Example 4.1. From the definition of the (C,d)-homogeneous measure, we see that

$$\mathbb{P}\left(X\in B_{x,\epsilon}\right) \geq \frac{\epsilon^d}{C}\mathbb{P}\left(X\in B_{x,1}\right) = C'\epsilon^d.$$

As a result, (27) becomes

$$\begin{aligned}\mathbb{P}\left(||X_{1,n}(x)-x||_2>\epsilon\right) = &\ (1-\mathbb{P}(X\in B_{x,\epsilon}))^n \\ \leq &\ \left(1-C'\epsilon^d\right)^n \leq e^{-C'n\epsilon^d}.\end{aligned}$$

Then we end up having

$$\mathbb{E}\left[(\widehat{m}_{kNN}(x)-m(x))^2\right] \leq \frac{1}{4k_n} + L^2\frac{2\Gamma(2/d)}{dC'^{2/d}}\left(\frac{k_n}{n}\right)^{2/d}$$

and the result follows by setting $k_n = n^{\frac{2}{2+d}}$. Similarly, we only need to modify (29) and (30) in the proof of Example 4.2. By using the (C,d)-homogeneous measure,

$$\begin{aligned}(1-\mathbb{P}\left(X\in B_{x,rh_n}\right))^n \leq &\ \left(1-\frac{h_n^d}{C}\mathbb{P}\left(X\in B_{x,r}\right)\right)^n \\ = &\ \left(1-C'h_n^d\right)^n \\ \leq &\ e^{-C'nh_n^d}\end{aligned}$$

and apply this result to (29) and (30). We complete the proof by following the same steps in the proof of Example 4.2.

A.10. Proof of Theorem 4.3

Proof. We use a combinatorial central limit theorem in Bolthausen (1984) to prove the result. First denote $a_{ij} = w_i(x)Y_j$ for $1\leq i,j\leq n$ and

$$\mu = na_{..}, \quad \sigma_n^2 = \sum_{1\leq i,j\leq n}(a_{ij}-a_{i.}-a_{.j}+a_{..})^2/(n-1),$$

where

$$a_{i.} = \sum_{j=1}^n a_{ij}/n, \quad a_{.j} = \sum_{i=1}^n a_{ij}/n, \quad a_{..} = \sum_{1\leq i,j\leq n} a_{ij}/n^2.$$

In our case, $\mu = \widehat{\pi}_1$ and σ_n^2 is given in (18). Let $d_{ij} = a_{ij} - a_{i\cdot} - a_{\cdot j} + a_{\cdot\cdot} = (w_i(x) - 1/n)(Y_j - \widehat{\pi}_1)$. Then using the theorem in Bolthausen (1984), we obtain

$$\sup_{t\in\mathbb{R}} \left| \mathbb{P}\left(\frac{\widehat{m}(x) - \widehat{\pi}_1}{\sigma_n} \le t \middle| X_1, \ldots, X_n \right) - \Phi(t) \right| \le K \frac{1}{\sqrt{n}} \frac{\frac{1}{n^2}\sum_{i,j} |d_{i,j}|^3}{\left(\frac{1}{n^2}\sum_{i,j} d_{i,j}^2\right)^{3/2}},$$

where K is a universal constant. Note that

$$\frac{1}{n^2}\sum_{i,j} |d_{i,j}|^3 = \frac{1}{n}\sum_{i=1}^n \left| w_i(x) - \frac{1}{n} \right|^3 \cdot \frac{1}{n}\sum_{j=1}^n \left| Y_j - \widehat{\pi}_1 \right|^3$$

and

$$\frac{1}{n^2}\sum_{i,j} d_{i,j}^2 = \frac{1}{n}\sum_{i=1}^n \left(w_i(x) - \frac{1}{n} \right)^2 \cdot \frac{1}{n}\sum_{j=1}^n \left(Y_j - \widehat{\pi}_1 \right)^2.$$

As a result,

$$\frac{\frac{1}{n^2}\sum_{i,j} |d_{i,j}|^3}{\left(\frac{1}{n^2}\sum_{i,j} d_{i,j}^2\right)^{3/2}} = \frac{1}{\sqrt{n}} \frac{\frac{1}{n}\sum_{i=1}^n \left| w_i(x) - \frac{1}{n} \right|^3}{\left\{ \frac{1}{n}\sum_{i=1}^n \left(w_i(x) - \frac{1}{n} \right)^2 \right\}^{3/2}} \cdot \underbrace{\frac{\frac{1}{n}\sum_{j=1}^n \left| Y_j - \widehat{\pi}_1 \right|^3}{\left(\frac{1}{n}\sum_{j=1}^n \left(Y_j - \widehat{\pi}_1 \right)^2 \right)^{3/2}}}_{(II)}$$

$$\le \underbrace{\frac{\max_{1\le i\le n} (w_i(x) - 1/n)^2}{\sum_{i=1}^n (w_i(x) - 1/n)^2}}_{(I)} \cdot (II)$$

Note that $(I) = o_P(1)$ under the given assumption and (II) is stochastically bounded by the law of large number. Thus we conclude that

$$\sup_{t\in\mathbb{R}} \left| \mathbb{P}\left(\frac{\widehat{m}(x) - \widehat{\pi}_1}{\sigma_n} \le t \middle| X_1, \ldots, X_n \right) - \Phi(t) \right| = o_P(1),$$

which implies the desired result. □

A.11. Proof of Corollary 4.2

Proof. For kNN regression, there are k and $(n-k)$ number of k^{-1} and zero in $\{w_1(x), \ldots, w_n(x)\}$ respectively. Hence,

$$\sum_{i=1}^n \left(w_i(x) - \frac{1}{n} \right)^2 = k\left(\frac{1}{k} - \frac{1}{n} \right)^2 + \frac{n-k}{n^2}.$$

Furthermore, under the assumption that $2k < n$, we have

$$\max_{1\le i\le n} \left| w_i(x) - \frac{1}{n} \right| = \frac{1}{k} - \frac{1}{n}.$$

After direct calculations, one can show that

$$\frac{\max_{1\le i\le n}|w_i(x)-1/n|}{\{\sum_{i=1}^n (w_i(x)-1/n)^2\}^{1/2}}\to 0,$$

and thus the result follows. □

A.12. Proof of Corollary 4.3

Proof. Note that

$$\widehat{m}_{ker}(x)=\sum_{i=1}^n w_i(x)Y_i=\frac{\sum_{i=1}^n Y_i K\left(\frac{x-X_i}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right)}=\frac{\sum_{i=1}^n Y_i K_{h_n}(x-X_i)}{\sum_{i=1}^n K_{h_n}(x-X_i)}.$$

Hence it suffices to show that

$$\begin{aligned}&\frac{\max_{1\le i\le n}(w_i(x)-1/n)^2}{\sum_{i=1}^n (w_i(x)-1/n)^2}\\ =\;&\frac{\max_{1\le i\le n}\left(K_h(x-X_i)-\frac{1}{n}\sum_{j=1}^n K_h(x-X_j)\right)^2}{\sum_{i=1}^n\left(K_h(x-X_i)-\frac{1}{n}\sum_{j=1}^n K_h(x-X_j)\right)^2}\xrightarrow{p}0.\end{aligned}$$

Using the given condition, the numerator is bounded by

$$\max_{1\le i\le n}\left(K_h(x-X_i)-\frac{1}{n}\sum_{j=1}^n K_h(x-X_j)\right)^2\le 4h^{-D}\mathcal{K}^2.$$

Whereas the denominator can be decomposed into two parts:

$$\begin{aligned}&\sum_{i=1}^n\left(K_h(x-X_i)-\frac{1}{n}\sum_{j=1}^n K_h(x-X_j)\right)^2\\ =\;&\sum_{i=1}^n K_h^2(x-X_i)-2n\left(\frac{1}{n}\sum_{j=1}^n K_h(x-X_j)\right)^2\end{aligned}$$

Based on the usual bias-variance decomposition of the kernel density estimation (Wasserman, 2006), each part can be approximated as

$$\frac{1}{nh^D}\sum_{i=1}^n K^2\left(\frac{x-X_i}{h}\right)=f(x)\int K^2(u)du+O(h)+O_P\left(\frac{1}{\sqrt{nh^D}}\right)$$

$$\frac{1}{nh^D}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)=f(x)+O(h^2)+O_P\left(\frac{1}{\sqrt{nh^D}}\right).$$

Now, the sufficient condition can be further bounded by

$$\begin{aligned}
&\frac{\max_{1\le i\le n}\left(K_h(x-X_i)-\frac{1}{n}\sum_{j=1}^n K_h(x-X_j)\right)^2}{\sum_{i=1}^n\left(K_h(x-X_i)-\frac{1}{n}\sum_{j=1}^n K_h(x-X_j)\right)^2}\\
\le\ &\frac{4h^{-D}\mathcal{K}^2}{\frac{1}{h^{2D}}\sum_{i=1}^n K^2\left(\frac{x-X_i}{h}\right)-2n\left(\frac{1}{nh^D}\sum_{j=1}^n K\left(\frac{x-X_j}{h}\right)\right)^2}\\
=\ &\frac{4n^{-1}\mathcal{K}^2}{\frac{1}{nh^D}\sum_{i=1}^n K^2\left(\frac{x-X_i}{h}\right)-2h^D\left(\frac{1}{nh^D}\sum_{j=1}^n K\left(\frac{x-X_j}{h}\right)\right)^2}. \qquad (32)
\end{aligned}$$

Then using the previous approximations, the denominator becomes

$$\begin{aligned}
&\frac{1}{nh^D}\sum_{i=1}^n K^2\left(\frac{x-X_i}{h}\right)-2h^D\left(\frac{1}{nh^D}\sum_{j=1}^n K\left(\frac{x-X_j}{h}\right)\right)^2\\
=\ &f(x)\int K^2(u)du+O(h)+O_P\left(\frac{1}{\sqrt{nh^D}}\right)\\
&-2h^D\left(f(x)+O(h^2)+O_P\left(\frac{1}{\sqrt{nh^D}}\right)\right)^2\\
=\ &\underbrace{f(x)\int K^2(u)du}_{>0 \text{ by the assumption}}+o_P(1).
\end{aligned}$$

Hence (32) converges to zero in probability and the result follows. □

Appendix B: Diffusion maps

Dimensionality reduction methods can be useful for visualizing and describing low-dimensional structures that are embedded in higher-dimensional spaces. In this section, we briefly describe diffusion maps (Coifman et al., 2005; Coifman and Lafon, 2006) and the particular version that we use to visualize the results of our local two-sample test.

As a starting point for constructing a diffusion map, one first defines a weight that reflects the local similarity of two points x_i and x_j in $\mathcal{X}=\{x_1,\dots,x_n\}$. A common choice is the Gaussian kernel

$$w(x_i,x_j)=\exp\left(-\frac{s(x_i,x_j)^2}{\epsilon}\right), \qquad (33)$$

where $s(x_i,x_j)$ represents (for example, the Euclidean) distance between the points. These weights are used to build a Markov random walk on the data

with the transition probability from x_i to x_j defined as

$$p(x_i, x_j) = \frac{w(x_i, x_j)}{\sum_{k \in \Omega} w(x_i, x_k)}.$$

The one-step transition probabilities are stored in an $n \times n$ matrix denoted by $\mathbf{P}$, and then usually propagated by a t-step Markov random walk with transition probabilities $\mathbf{P}^t$. Instead of choosing a fixed time parameter t, however, we here combine diffusions at all times (Coifman et al., 2005) and define an averaged diffusion map[2] according to

$$\Psi_{\text{av}} : x \mapsto \left[\left(\frac{\lambda_1}{1-\lambda_1} \right) \psi_1(x), \left(\frac{\lambda_2}{1-\lambda_2} \right) \psi_2(x), \ldots, \left(\frac{\lambda_m}{1-\lambda_m} \right) \psi_m(x) \right],$$

where λ_i and ψ_i, respectively, represent the first mth eigenvalues and the corresponding right eigenvectors of $\mathbf{P}$.

In our application for galaxy morphologies, we also use a generalization of the weight in (33) proposed by Zelnik-Manor and Perona (2005) for spectral clustering. In their paper, the authors show that a data-driven varying bandwidth leads to more meaningful clustering results for data with multiple scales and propose the weight

$$\widehat{w}(x_i, x_j) = \exp\left(-\frac{s(x_i, x_j)^2}{\sigma_i \sigma_j} \right),$$

where $\sigma_{i(j)}$ is the distance between $x_{i(j)}$ and its kth neighbor. For our visualization purposes, we choose $m = 2$ and $k = 50$, but a range of other values give similar results.

Acknowledgements

ABL would like to thank Rafael Izbicki and Larry Wasserman for discussions that lead to the two-sample testing work, and Peter Freeman and Jeffrey Newman for acting as IK's co-advisors for the data analysis project on which Section 6 is based. The authors also thank the editor and the reviewers for their constructive comments and suggestions. This work was partially supported by NSF DMS-1520786.

References

Anderson, J. A. (1972). Separate sample logistic discrimination. *Biometrika*, 59(1):19–35. MR0345332

Anderson, N. H., Hall, P., and Titterington, D. M. (1994). Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *Journal of Multivariate Analysis*, 50(1):41–54. MR1292607

[2]This is also the default option of the function `diffuse()` in the `R` package `diffusionMap`.

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, volume 3. New York: Wiley-Interscience. MR1990662

Ayano, T. (2012). Rates of convergence for the k-nearest neighbor estimators with smoother regression functions. *Journal of Statistical Planning and Inference*, 142(9):2530–2536. MR2922003

Baraud, Y. (2002). Non-asymptotic minimax rates of testing in signal detection. *Bernoulli*, 8(5):577–606. MR1935648

Baringhaus, L. and Franz, C. (2004). On a new multivariate two-sample test. *Journal of Multivariate Analysis*, 88(1):190–206. MR2021870

Biau, G. (2012). Analysis of a random forests model. *Journal of Machine Learning Research*, 13(Apr):1063–1095. MR2930634

Biau, G. and Devroye, L. (2015). *Lectures on the Nearest Neighbor Method*. Springer. MR3445317

Bickel, P. J. and Li, B. (2007). Local polynomial regression on unknown manifolds. *Lecture Notes – Monograph Series*, pages 177–186. MR2459188

Bolthausen, E. (1984). An estimate of the remainder in a combinatorial central limit theorem. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 66(3):379–386. MR0751577

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32. MR3874153

Bühlmann, P. and Van De Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Science & Business Media. MR2807761

Bunea, F. and Barbu, A. (2009). Dimension reduction and variable selection in case control studies via regularized likelihood optimization. *Electronic Journal of Statistics*, 3:1257–1287. MR2566187

Cazáis, F. and Lhéritier, A. (2015). Beyond two-sample-tests: Localizing data discrepancies in high-dimensional spaces. In *IEEE International Conference on Data Science and Advanced Analytics (DSAA), 2015*, pages 1–10. IEEE.

Chen, X. and Ishwaran, H. (2012). Random forests for genomic data analysis. *Genomics*, 99(6):323–329.

Coifman, R. R. and Lafon, S. (2006). Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30. MR2238665

Coifman, R. R., Lafon, S., Lee, A. B., Maggioni, M., Nadler, B., Warner, F., and Zucker, S. W. (2005). Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the National Academy of Sciences of the United States of America*, 102(21):7426–7431. MR2238665

Conselice, C. J. (2003). The relationship between stellar light distributions of galaxies and their formation histories. *The Astrophysical Journal Supplement Series*, 147(1):1.

Conselice, C. J. (2014). The evolution of galaxy structure over cosmic time. *Annual Review of Astronomy and Astrophysics*, 52:291–337.

Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., and Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology*, 88(11):2783–2792.

Devroye, L., Györfi, L., and Lugosi, G. (2013). *A Probabilistic Theory of Pattern*

Recognition, volume 31. Springer Science & Business Media. MR1383093

Díaz-Uriarte, R. and De Andres, S. A. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7(1):3. MR2717520

Duong, T. (2013). Local significant differences from nonparametric two-sample tests. *Journal of Nonparametric Statistics*, 25(3):635–645. MR3174288

Fokianos, K. (2008). Comparing two samples by penalized logistic regression. *Electronic Journal of Statistics*, 2:564–580. MR2426102

Freeman, P., Izbicki, R., Lee, A., Newman, J., Conselice, C., Koekemoer, A., Lotz, J., and Mozena, M. (2013). New image statistics for detecting disturbed galaxy morphologies at high redshift. *Monthly Notices of the Royal Astronomical Society*, 434(1):282–295.

Friedman, J., Hastie, T., and Tibshirani, R. (2009). *The Elements of Statistical Learning*. Springer, New York. MR2722294

Friedman, J. H. (2003). On multivariate goodness of fit and two sample testing. *eConf*, 30908(SLAC-PUB-10325):311–313.

Gagnon-Bartsch, J. and Shem-Tov, Y. (2016). The classification permutation test: A nonparametric test for equality of multivariate distributions. *arXiv preprint* arXiv:1611.06408. MR4019146

González-Manteiga, W. and Cao, R. (1993). Testing the hypothesis of a general linear model using nonparametric regression estimation. *Test*, 2(1-2):161–188. MR1265489

González-Manteiga, W. and Crujeiras, R. M. (2013). An updated review of goodness-of-fit tests for regression models. *Test*, 22(3):361–411. MR3093195

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773. MR2913716

Györfi, L., Kohler, M., Krzyzak, A., and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer Science & Business Media. MR1920390

Hamza, M. and Larocque, D. (2005). An empirical comparison of ensemble methods based on classification trees. *Journal of Statistical Computation and Simulation*, 75(8):629–643.

Hardle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *The Annals of Statistics*, 21(4):1926–1947. MR1245774

Hart, J. (2013). *Nonparametric Smoothing and Lack-of-Fit Tests*. Springer Science & Business Media. MR1461272

Hediger, S., Michel, L., and Näf, J. (2019). On the use of random forest for two-sample testing. *arXiv preprint* arXiv:1903.06287.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802. MR0995126

Hu, J. and Bai, Z. (2016). A review of 20 years of naive tests of significance for high-dimensional mean vectors and covariance matrices. *Science China Mathematics*, 59(12):2281–2300. MR3578957

Ingster, Y. I. (1987). Minimax testing of nonparametric hypotheses on a distribution density in the L_p metrics. *Theory of Probability & Its Applications*,

31(2):333–337. MR0851000

Keziou, A. and Leoni-Aubin, S. (2005). Test of homogeneity in semiparametric two-sample density ratio models. *Comptes Rendus Mathématique*, 340(12):905–910. MR2152277

Kim, I., Ramdas, A., Singh, A., and Wasserman, L. (2019). Classification accuracy as a proxy for two sample testing. *arXiv preprint* arXiv:1602.02210v2.

Kpotufe, S. (2011). k-NN regression adapts to local intrinsic dimension. In *Advances in Neural Information Processing Systems*, pages 729–737.

Kpotufe, S. and Garg, V. (2013). Adaptivity to local smoothness and dimension in kernel regression. In *Advances in Neural Information Processing Systems*, pages 3075–3083.

Lehmann, E. L. and Romano, J. P. (2006). *Testing Statistical Hypotheses*. Springer Science & Business Media. MR2135927

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

Lotz, J. M., Primack, J., and Madau, P. (2004). A new nonparametric approach to galaxy morphological classification. *The Astronomical Journal*, 128(1):163.

Mondal, P. K., Biswas, M., and Ghosh, A. K. (2015). On high dimensional two-sample tests based on nearest neighbors. *Journal of Multivariate Analysis*, 141:168–178. MR3390065

Ojala, M. and Garriga, G. C. (2010). Permutation tests for studying classifier performance. *Journal of Machine Learning Research*, 11(Jun):1833–1863. MR2660654

Olivetti, E., Greiner, S., and Avesani, P. (2015). Statistical independence for the evaluation of classifier-based diagnosis. *Brain Informatics*, 2(1):13–19. MR3624321

Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, 66(3):403–411. MR0556730

Qin, J. and Zhang, B. (1997). A goodness-of-fit test for logistic regression models based on case-control data. *Biometrika*, 84(3):609–618. MR1603924

Ramdas, A., Reddi, S. J., Poczos, B., Singh, A., and Wasserman, L. (2015). Adaptivity and computation-statistics tradeoffs for kernel and distance based high dimensional two sample testing. *arXiv preprint* arXiv:1508.00655.

Rosenblatt, J., Gilron, R., and Mukamel, R. (2016). Better-than-chance classification for signal detection. *arXiv preprint* arXiv:1608.08873.

Scott, A. J. and Wild, C. (2001). Maximum likelihood for generalised case-control studies. *Journal of Statistical Planning and Inference*, 96(1):3–27. MR1843447

Snyder, G. F., Torrey, P., Lotz, J. M., Genel, S., McBride, C. K., Vogelsberger, M., Pillepich, A., Nelson, D., Sales, L. V., and Sijacki, D. (2015). Galaxy morphology and star formation in the illustris simulation at $z = 0$. *Monthly Notices of the Royal Astronomical Society*, 454(2):1886–1908.

Sugiyama, M., Suzuki, T., Itoh, Y., Kanamori, T., and Kimura, M. (2011). Least-squares two-sample test. *Neural Networks*, 24(7):735–751.

Székely, G. J. and Rizzo, M. L. (2004). Testing for equal distributions in high dimension. *InterStat*, 5:1–6.

Thas, O. (2010). *Comparing Distributions*. Springer. MR2547894

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation. Revised and Extended from the 2004 French Original. Translated by Vladimir Zaiats*. Springer Series in Statistics. New York: Springer. MR2724359

Van de Geer, S. A. (2008). High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2):614–645. MR2396809

Wager, S. and Walther, G. (2015). Adaptive concentration of regression trees, with application to random forests. *arXiv preprint* arXiv:1503.06388.

Wang, C. and Carroll, R. (1993). On robust estimation in logistic case-control studies. *Biometrika*, 80(1):237–241. MR1225228

Wang, S. and Carroll, R. J. (1999). High-order accurate methods for retrospective sampling problems. *Biometrika*, 86(4):881–897. MR1741984

Wasserman, L. (2006). *All of Nonparametric Statistics*. Springer Science & Business Media. MR2172729

Weihrather, G. (1993). Testing a linear regression model against nonparametric alternatives. *Metrika*, 40(1):367–379. MR1247139

Yang, Y. and Barron, A. (1999). Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, 27(5):1564–1599. MR1742500

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

Zhang, C. and Dette, H. (2004). A power comparison between nonparametric regression tests. *Statistics & Probability Letters*, 66(3):289–301. MR2045474

Zheng, J. X. (1996). A consistent test of functional form via nonparametric estimation techniques. *Journal of Econometrics*, 75(2):263–289. MR1413644