Minimax Nonparametric Two-sample Test

Xin Xing\textsuperscript{*}, Zuofeng Shang\textsuperscript{†}, Pang Du \textsuperscript{‡}, Ping Ma\textsuperscript{§}, Wenxuan Zhong, \textsuperscript{§} Jun S. Liu \textsuperscript{*}

November 12, 2019

Abstract

We consider the problem of comparing probability densities between two groups. To model the complex pattern of the underlying densities, we formulate the problem as a nonparametric density hypothesis testing problem. The major difficulty is that conventional tests may fail to distinguish the alternative from the null hypothesis under the controlled type I error. In this paper, we model log-transformed densities in a tensor product reproducing kernel Hilbert space (RKHS) and propose a probabilistic decomposition of this space. Under such a decomposition, we quantify the difference of the densities between two groups by the component norm in the probabilistic decomposition. Based on the Bernstein width, a sharp minimax lower bound of the distinguishable rate is established for the nonparametric two-sample test. We then propose a penalized likelihood ratio (PLR) test possessing the Wilks’ phenomenon with an asymptotically Chi-square distributed test statistic and achieving the established minimax testing rate. Simulations and real applications demonstrate that the proposed test outperforms the conventional approaches under various scenarios.

Keywords: two-sample test, smoothing spline, penalized likelihood ratio testing, minimax optimality, nonparametric testing, Wilks’ phenomenon.

1 Introduction

Testing the equality of two probability densities based on observed samples is a fundamental problem in statistics. Formally speaking, let $X \in \mathcal{X}$ be

\textsuperscript{*}Department of Statistics, Harvard University
\textsuperscript{†}Department of Mathematical Sciences, New Jersey Institute of Technology
\textsuperscript{‡}Department of Statistics, Virginia Tech
\textsuperscript{§}Department of Statistics, University of Georgia
a random variable and $Z \in \{0, 1\}$ be a binary random variable indicating the group membership of $X$. Let $f_{X|Z=0}(\cdot)$ and $f_{X|Z=1}(\cdot)$ be the conditional probability densities of $X$ given $Z = 0$ and $Z = 1$, respectively. The problem of interest is to test the following hypothesis

$$H_0 : f_{X|Z=0}(\cdot) = f_{X|Z=1}(\cdot) \text{ vs. } H_1 : f_{X|Z=0}(\cdot) \neq f_{X|Z=1}(\cdot).$$

The above two-sample testing problem arises from many applications, ranging from modern biological sciences to deep learning.

In many applications, the underlying distributions usually demonstrate complex patterns and cannot be fitted well by a pre-determined family of parametric distributions. Classical normality-based tests such as the two-sample t-test (Anderson, 1958) and the Shapiro-Wilk test (Shapiro and Wilk, 1965) are generally inappropriate. Nonparametric approaches are more appealing due to their distribution-free feature, and have received increasing attention recently. Examples include distance-based tests such as the Kolmogorov-Smirnov (KS) test (Darling, 1957), distance correlation (Székely et al., 2007), and the maximum mean discrepancy (MMD) (Gretton et al., 2012; Pfister et al., 2018). However, these popular methods often suffer from insufficient power as demonstrated in the following toy example. Figure 1(a) displays two normal densities with zero means and different variances such that the Kullback-Leibler (KL) distance between the densities is 0.14. Figure 1(c) plots the empirical power of three competing tests. Overall the powers of KS and MMD are satisfactory and increase rapidly along with the sample size $n$. However, when the densities in comparison are multimodal, such as mixtures of normal densities shown in Figure 1(b) where their KL distance is still at 0.14, the powers of KS and MMD stay at zero even when the sample size $n$ is as large as 1000 (Figure 1(d)). This indicates that distance-based tests may not be sensitive to the distribution change especially when the densities demonstrate complex shapes. See Mason et al. (1983) for an earlier similar discovery.

### 1.1 A brief literature review

There is an extensive literature on nonparametric density comparisons, e.g., the Kolmogorov-Smirnov (KS) test (Darling, 1957), distance correlation (DC) (Székely et al., 2007), Hilbert-Schmidt independence criteria (HSIC) (Pfister et al., 2018), minimum mean discrepancy (MMD) (Gretton et al., 2012), empirical likelihood tests (ELT) (Cao and Van Keilegom, 2006), and kernel density test (Martínez-Camblor and de Uña-Álvarez, 2009). An alternative direction is using discretization (“slicing”) of continuous random
variables (Miller and Siegmund, 1982). Recently, (Jiang et al., 2015) proposed the dynamic slicing test (DSLICE), which penalizes the number of slices to regularize the test statistics. However, all these methods have complicated asymptotic distributions, and consequently Monte Carlo simulation techniques, such as resampling and bootstrap, are usually needed for determining the $p$-values. Moreover, it is nontrivial to analyze their power due to the complexity of the resampling procedure. Parallel to our work, Li and Yuan (2019) proposed a normalized MMD test for comparing two densities based on Gaussian kernels with appropriately chosen scaling parameters, and established its minimax testing optimality. Nonparametric minimax testing principle is pioneered by Ingster (1989) and Ingster (1993). Recently, Lepski et al. (1999) and Fromont et al. (2012) establish the minimax testing principle for nonparametric goodness-of-fit test, Wei et al. (2017) introduced minimax nonparametric hypothesis testing over convex cones and Wei and Wainwright (2018) derived sharp upper and lower bounds on the localized

Figure 1: (a) Two uni-modal densities with KL-distance 0.14. (b) Two bi-modal densities with KL-distance 0.14. (c) Powers of the KL, MMD and the proposed PLR test for uni-model case for sample size ranging from 125 to 1000. (d) Powers of the KL, MMD and the proposed PLR test for bi-model case for sample size ranging from 125 to 1000.
minimax testing radius in Gaussian sequence models using an information-theoretic view. Our minimax lower bound for separation rate can be viewed as an extension of Wei and Wainwright (2018) to the two-sample testing scenario. However, such an extension is nontrivial in that in general it is impossible to transform a two-sample testing problem into a Gaussian sequence model. Instead, we modify the original technique from Ingster (1989) and Ingster (1993), which efficiently treats the signal in time-domain, to fit in the information-theoretic framework of Wei and Wainwright (2018).

1.2 Main contributions of this work

In this paper, we propose a new test called the penalized likelihood ratio (PLR) to improve upon existing approaches. The intuition is to incorporate the likelihood ratio principle into the nonparametric two-sample test. Specifically, we characterize the log-transformed joint density $\eta(x,z)$ of $(X,Z)$ by a tensor product reproducing kernel Hilbert space (RKHS) $\mathcal{H}$, and establish a penalized log-likelihood function of $\eta \in \mathcal{H}$. Then the PLR test is defined as the difference between the maximum values of the penalized log-likelihoods under $H_0$ and $H_1$. The tensor product RKHS framework has following advantages. First, the maximum values of the penalized log-likelihoods are easy to compute, and numerous existing algorithms and software packages are publicly available (Silverman, 1982; Gu and Qiu, 1993; Gu, 2013). Second, the choice of RKHS is flexible. Though the current work focuses on Sobolev spaces, the results are readily extendable to general settings such as a Gaussian RKHS. Third, theoretical tools for justifying the performance of the proposed PLR test are mature given the recent advances in RKHS-based nonparametric inference (Shang et al., 2010, 2013; Shang and Cheng, 2015). In fact, the PLR test is proven to be minimax optimal and the asymptotic null distribution of its test statistic is a chi-square distribution. Comparing with distance-based tests whose asymptotic distributions are mostly unavailable analytically, the proposed PLR test is user-friendly.

The main contribution of the current work has three parts. First, our proposed PLR test is built upon functional ANOVA decomposition in tensor product RKHS proposed by Gu (2013) and Wahba (1990). However, the existing references on functional ANOVA mainly focus on estimation while leaving hypothesis testing an open problem. The current work fills this gap. Second, the proposed PLR test is proven explicitly related to the popular MMD test (Gretton et al., 2012), which is also a surprise to us (see Section 5). Specifically, we show that the MMD test (with a particularly selected kernel) is the squared norm of the gradient of our log-likelihood ratio. We note that
the log-likelihood ratio is slightly different from our PLR test since it does not involve a penalty term. This likelihood-based viewpoint may partly explain the success of these two seemingly unrelated approaches. Third, we establish a minimax lower bound for the separation rate between $H_0$ and $H_1$ to guarantee the existence of a successful test. This result is useful for proving the minimax optimality of the PLR test and is of independent interest. Comparing with the existing minimax testing rate established by Ingster (1989), Ingster (1993), and Wei and Wainwright (2018) in the simple regression setting, we generalize the minimax hypothesis testing framework to handle the density comparison problems. We use the Bernstein width proposed by Pinkus (2012) for deriving the order of the smallest separation rate between the hypotheses to guarantee a successful test.

The rest of the paper is organized as follows. Section 2 introduces some background on tensor product RKHS. Section 4 describes some theoretical results for the minimax lower bound of two-sample tests based on an information-theoretic view. Section 3 derives the asymptotic distribution of the PLR test and analyzes its asymptotic power. The PLR test is also shown to achieve the minimax lower bound established in Section 4. Section 5 provides an interpretation of the MMD as a likelihood ratio test and compares it with the proposed PLR test. Section 6 examines finite sample performances of the PLR test in comparison with a competitors through simulations. Section 7 contains two real-world examples using the PLR test, and Section 8 concludes with a short discussion. Additional proofs for the lemmas are deferred to the Supplementary Material.

2 Probabilistic decomposition for tensor product RKHS

Let $\mathcal{H}$ be an RKHS endowed with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and let $\mathcal{D}$ be a general domain for functions in $\mathcal{H}$. There always exists a symmetric and square integrable function $K(\cdot, \cdot) : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$, such that

$$\langle f, K(\alpha, \cdot) \rangle_{\mathcal{H}} = f(\alpha), \text{ for all } f \in \mathcal{H} \text{ and } \alpha \in \mathcal{D}. \quad (2.1)$$

Here $K(\cdot, \cdot)$ is called the reproducing kernel of $\mathcal{H}$. By Mercer’s theorem, $K$ has the following decomposition:

$$K(\alpha_1, \alpha_2) = \sum_{\nu=0}^{\infty} \mu_\nu \phi_\nu(\alpha_1) \phi_\nu(\alpha_2), \quad (2.2)$$
where \( \mu_\nu \)'s are non-negative descending eigenvalues and \( \phi_\nu \)'s are the corresponding normalized eigen-functions. For a discrete domain, we select the kernel as \( K(\alpha_1, \alpha_2) = 1\{\alpha_1 = \alpha_2\} \) for \( \alpha_1, \alpha_2 \in D \). For a continuous domain, there are different choices of kernels such as the Gaussian and Sobolev kernels. In this paper, we consider the \( m \)-th order Sobolev kernel whose eigenvalues satisfy

\[
\mu_\nu \approx \nu - 2m \quad \text{for} \quad \nu \geq 1 \quad (\text{Gu, 2013}).
\]

Suppose that the log-transformed joint probability density function \( \eta(x, z) \), for \((x, z) \in Y \equiv X \times Z \), belongs to a tensor product RKHS \( H = H^{(X)} \otimes H^{(Z)} \), in which \( H^{(Z)} \) and \( H^{(X)} \) represent the marginal RKHS of \( X \) and \( Z \), respectively. Specifically, we consider \( X \) as a continuous domain and \( Z \) as a discrete domain with values \( \{0, 1, \ldots, d-1\} \). We start with the Euclidean space as a simple example to illustrate the basic idea of tensor sum decomposition., which is often called the ANOVA decomposition in linear models. Consider a slightly more general setting with \( H^{(Z)} = \mathbb{R}^d \) being a \( d \)-dimensional Euclidean space, and let \( f \in \mathbb{R}^d \) be a vector, with \( f(z) \) being the \( z \)-th entry of the vector for \( z = 0, \ldots, d-1 \). We then have \( H^{(Z)} = \{ f : f \in \mathbb{R}^d \} \). Let \( A \) be the average operator defined as \( Af = (\frac{1}{d}1, f) \), where \( 1 = (1, \cdots, 1)^T \in \mathbb{R}^d \), and \( \langle \cdot, \cdot \rangle_{H^{(Z)}} \) is the Euclidean inner product. The tensor sum decomposition of the Euclidean space \( \mathbb{R}^d \) is

\[
\mathbb{R}^d = \mathbb{R}^d_0 \oplus \mathbb{R}^d_1 := \{ \frac{1}{d}1 \} \oplus \{ f \in \mathbb{R}^d | \sum_{z=0}^{d-1} f(z) = 0 \}
\]

where \( \mathbb{R}^d_0 \) represents the grand mean and \( \mathbb{R}^d_1 \) represents the main effect space. Lemma 2.1 provides the kernel for \( \mathbb{R}^d_0 \) and \( \mathbb{R}^d_1 \), respectively.

**Lemma 2.1.** For the RKHS \( H^{(Z)} \) on the discrete domain \( \{0, \ldots, d-1\} \) equipped with the Euclidean inner product, there exists a unique non-negative definite reproducing kernel \( K^{(Z)} \). Based on the tensor sum decomposition \( H^{(Z)} = H^{(Z)}_0 \oplus H^{(Z)}_1 \) where \( H^{(Z)}_0 = \{ \frac{1}{d}1 \} \) and \( H^{(Z)}_1 = \{ f \in H : \sum_{x=0}^{d-1} f(z) = 0 \} \), we have that the kernel for \( H^{(Z)}_0 \) is

\[
K^{(Z)}_{0, \text{init}}(z_1, z_2) = 1/d,
\]

and the kernel for \( H^{(Z)}_1 \) is

\[
K^{(Z)}_{1, \text{init}}(z_1, z_2) = 1\{z_1 = z_2\} - 1/d,
\]

where \( 1 \) denotes the indicator function.
Lemma 2.1 is related to the classic ANOVA decomposition which assumes a uniform distribution on the random variable, i.e., \( P(Z = 0) = \cdots = P(Z = d - 1) = 1/d \) without considering the true probability measure of \( Z \). Instead, we embed the probability measure of \( Z \) into the tensor sum decomposition of \( \mathcal{H}^{(Z)} \). Consider a discrete probabilistic measure \( \mathbb{P} \) on \( Z = \{0, \ldots, d - 1\} \) such that \( \mathbb{P}(Z = j) = \omega_j \geq 0 \), with \( \sum_{j=0}^{d-1} \omega_j = 1 \). Let \((\omega_0, \ldots, \omega_{d-1})\). The average operator \( \mathcal{A} \) is modified to \( \mathcal{A} := \mathcal{E}_z f(z) = \langle \omega, f \rangle_{\mathcal{H}(Z)} \). Since the kernel on \( \mathcal{H}(Z) \) is defined as \( \mathcal{H}(Z) = \{ f : f \in \mathbb{R}^d \} \) with the Euclidean inner product is \( K^{(Z)}(z_1, z_2) = 1_{\{z_1 = z_2\}}, \) we have \( \mathbb{E}_z [K^{(Z)}(z, z)] = \omega \) and the probabilistic averaging operator can be rewritten as \( \mathcal{A} := f \rightarrow \mathbb{E}_z f(z) = \langle \mathbb{E}_z [K^{(Z)}(z, \cdot)], f \rangle_{\mathcal{H}(Z)} \). Then \( \mathbb{E}_z [K^{(Z)}(z, \cdot)] \) can be treated as a mean embedding of the \( \mathbb{P} \) in \( \mathcal{H}(Z) \). Based on this probabilistic averaging operator, we introduce the tensor sum decomposition of \( \mathcal{H}(Z) \) in the following lemma.

Lemma 2.2. For the RKHS \( \mathcal{H}(Z) \) on the discrete domain \( \{0, \ldots, d-1\} \) with probability measure \( \mathbb{P}(Z = z) = \omega_z \) for \( z = 0, \ldots, d-1 \), there corresponds a unique non-negative definite reproducing kernel \( K^{(Z)} \). Based on the tensor sum decomposition \( \mathcal{H}(Z) = \mathcal{H}_0^{(Z)} \oplus \mathcal{H}_1^{(Z)} \) where \( \mathcal{H}_0^{(Z)} = \{ \mathbb{E}_z [K^{(Z)}(\cdot, \cdot)] \} \) and \( \mathcal{H}_1^{(Z)} = \{ f \in \mathcal{H} : \mathbb{E}_z (f(z)) = 0 \} \), we have that the kernel for \( \mathcal{H}_0^{(Z)} \) is

\[
K_0^{(Z)}(z_1, z_2) = \omega_{z_1} + \omega_{z_2} - \sum_{\ell=0}^{d-1} \omega_\ell^2,
\]

and the kernel for \( \mathcal{H}_1^{(Z)} \) is

\[
K_1^{(Z)}(z_1, z_2) = 1_{\{z_1 = z_2\}} - \omega_{z_1} - \omega_{z_2} + \sum_{\ell=0}^{d-1} \omega_\ell^2,
\]

where \( 1 \) is the indicator function.

Next we consider the continuous random variable \( X \in \mathcal{X} \) and \( \mathbb{P} \) be a probability measure on \( \mathcal{X} \). Suppose \( \mathcal{H}^{(X)} \) is a subspace in \( L^2(\mathbb{P}) \), and let \( K^{(X)} \) be the corresponding kernel satisfying \( \langle f, K^{(X)}(x, \cdot) \rangle_{\mathcal{H}^{(X)}} = f(x) \) for any \( f \in \mathcal{H}^{(X)} \). To conduct the tensor sum decomposition on \( \mathcal{H}^{(X)} \), we introduce the probabilistic averaging operator \( \mathcal{A} \) as \( \mathcal{A} := f \rightarrow \mathbb{E}_x f(x) = \mathbb{E}_x \langle K^{(X)}(x, \cdot), f \rangle_{\mathcal{H}^{(X)}} = \langle \mathbb{E}_x K^{(X)}(x, \cdot), f \rangle_{\mathcal{H}^{(X)}} \). \( \mathbb{E}_x K^{(X)}(x) \) has the same role as \( \omega \) in the Euclidean space. Then, the tensor sum decomposition of a functional space is defined as

\[
\mathcal{H}^{(X)} = \mathcal{H}_0^{(X)} \oplus \mathcal{H}_1^{(X)} := \{ \mathbb{E}_x K^{(X)}(x) \} \oplus \{ f \in \mathcal{H}^{(X)} : \mathcal{A} f = 0 \}. \tag{2.4}
\]
Analogously, we name \( \mathcal{H}_0^{(X)} \) as the grand mean space and \( \mathcal{H}_1^{(X)} \) as the main effect space. \( \mathbb{E}_x K_x^{(X)} \) is known as the kernel mean embedding which is well established in the statistics literature Berlien and Thomas-Agnan (2011). Next, we introduce Lemma 2.3 to construct the kernel function for \( \mathcal{H}_0^{(X)} \) and \( \mathcal{H}_1^{(X)} \).

**Lemma 2.3.** For the RKHS \( \mathcal{H}^{(X)} \) on a continuous domain \( X \) with probability measure \( \mathbb{P} \) equipped with inner product \( \langle \cdot, \cdot \rangle_{\mathcal{H}^{(X)}} \), there corresponds a unique nonnegative definite reproducing kernel \( K^{(X)} \). Based on the tensor sum decomposition \( \mathcal{H}^{(X)} = \mathcal{H}_0^{(X)} \oplus \mathcal{H}_1^{(X)} \) where \( \mathcal{H}_0^{(X)} = \{ \mathbb{E}_x K_x^{(X)} \} \) and \( \mathcal{H}_1^{(X)} = \{ f \in \mathcal{H} : \mathbb{E}_x (f(x)) = 0 \} \), we have that the kernel for \( \mathcal{H}_0^{(X)} \) is

\[
K_0^{(X)}(x, y) = \mathbb{E}_x [K(x, y)] + \mathbb{E}_y [K(x, y)] - \mathbb{E}_{x, y} K(x, y),
\]

and the kernel for \( \mathcal{H}_1^{(X)} \) is

\[
K_1^{(X)}(x, y) = \langle K_x^{(X)} - \mathbb{E}_x K_x^{(X)}, K_y^{(X)} - \mathbb{E}_y K_y^{(X)} \rangle_{\mathcal{H}^{(X)}} = K^{(X)}(x, y) - \mathbb{E}_x [K^{(X)}(x, y)] - \mathbb{E}_y [K^{(X)}(x, y)] + \mathbb{E}_{x, y} K^{(X)}(x, y).
\]

We are now ready to consider the RKHS \( \mathcal{H} = \mathcal{H}^{(X)} \otimes \mathcal{H}^{(Z)} \) on the product domain \( Y = X \times Z \) for the purpose of two-sample hypothesis testing, where \( X \) is a continuous domain and \( Z = \{0, 1\} \) is a discrete domain containing the group information. By Lemma 2.2, \( \mathcal{H}^{(Z)} \) has the probabilistic decomposition as tensor sums of subspaces \( \mathcal{H}^{(Z)} = \mathcal{H}_0^{(Z)} \oplus \mathcal{H}_1^{(Z)} \); and by Lemma 2.3, \( \mathcal{H}^{(X)} \) has the probabilistic decomposition as \( \mathcal{H}^{(X)} = \mathcal{H}_0^{(X)} \oplus \mathcal{H}_1^{(X)} \). Based on the distributive law, we have the decomposition of \( \mathcal{H} \) as

\[
\mathcal{H} = (\mathcal{H}_0^{(X)} \oplus \mathcal{H}_1^{(X)}) \otimes (\mathcal{H}_0^{(Z)} \oplus \mathcal{H}_1^{(Z)}) \equiv \mathcal{H}_{00} \oplus \mathcal{H}_{10} \oplus \mathcal{H}_{01} \oplus \mathcal{H}_{11},
\]

where \( \mathcal{H}_{ij} = \mathcal{H}_i^{(X)} \otimes \mathcal{H}_j^{(Z)} \) for \( i, j = 0, 1 \). We call this decomposition the *probabilistic decomposition* of the tensor product RKHS \( \mathcal{H} \) since it embeds the probability measure of the random variable \( X \) and \( G \). With the decomposition in (2.6), the uniqueness of the decomposition in (2.8) is guaranteed. In the following Lemma, we construct the kernel functions on each subspace \( \mathcal{H}_{ij} \).

**Lemma 2.4.** Suppose \( K_i^{(X)} \) is the reproducing kernel of \( \mathcal{H}_i^{(X)} \) on \( X \), and \( K_j^{(Z)} \) is the reproducing kernel of \( \mathcal{H}_j^{(Z)} \) on \( Z \) for \( i, j = 0, 1 \). Then the reproducing kernels of \( \mathcal{H}_i^{(X)} \otimes \mathcal{H}_j^{(Z)} \) on \( Y = X \times Z \) is \( K^{ij}((x_1, z_1), (x_2, z_2)) = K_i^{(X)}(x_1, x_2) K_j^{(Z)}(z_1, z_2) \) with \( x_1, x_2 \in X \) and \( z_1, z_2 \in Z \).
Lemma 2.4 states that the reproducing kernels of the tensor product space is the product of the reproducing kernels. Lemmas 2.4 can be easily proved with Theorems 2.6 in Gu (2013). Based on Lemmas 2.2, 2.3 and 2.4, we can construct the kernels \( K_{00}, K_{10}, K_{01} \) and \( K_{11} \) for the subspaces \( \mathcal{H}_{00}, \mathcal{H}_{10}, \mathcal{H}_{01} \) and \( \mathcal{H}_{11} \) accordingly.

2.1 Eigensystem for tensor product RKHS and hypothesis testing

Given \( \mathcal{H}^{(X)} \) and \( \mathcal{H}^{(Z)} \), \( \mathcal{H}^{(X)} \otimes \mathcal{H}^{(Z)} \) is defined as
\[
\mathcal{H}^{(X)} \otimes \mathcal{H}^{(Z)} = \left\{ \sum_{i=1}^{k} f_i(\cdot)g_i(\cdot) : f_i \in \mathcal{H}^{(X)}, g_i \in \mathcal{H}^{(Z)}, k \text{ is any positive integer.} \right\}
\]

By the decomposition in (2.2), we have the eigenvalue and eigenfunction pair for \( \mathcal{H}_0^{(X)} \) as \( \{\mu_0, \phi_0\} \) and the eigen basis for \( \mathcal{H}_1^{(X)} \) as \( \{\mu_i, \phi_i\}_{i=1}^{\infty} \). The eigenvalue and eigenfunction pair for \( \mathcal{H}_0^{(Z)} \) is \( \{\nu_0, \psi_0\} \) and eigen basis for \( \mathcal{H}_1^{(Z)} \) is \( \{\nu_1, \psi_1\} \). Then we have the basis for the tensor product RKHS \( \mathcal{H} \)
\[
\{\mu_0\nu_0, \phi_0\psi_0\}, \{\mu_0\nu_1, \phi_0\psi_1\}, \{\mu_i\nu_0, \phi_i\psi_0\}_{i=1}^{\infty}, \text{ and } \{\mu_i\nu_1, \phi_i\psi_1\}_{i=1}^{\infty} \quad (2.7)
\]
are the pairs of eigenvalues and eigenfunctions for \( \mathcal{H}_{00}, \mathcal{H}_{01}, \mathcal{H}_{10}, \) and \( \mathcal{H}_{11} \), respectively. In the following sections, we refer to (2.7) as the eigensystem for \( \mathcal{H} \) under the probabilistic decomposition.

Given the probabilistic decomposition tensor product RKHS (2.6), the log-transformed joint density function has a unique decomposition demonstrated in below:
\[
\eta(x, z) = \mu + \eta_X(x) + \eta_Z(z) + \eta_{XZ}(x, z), (x, z) \in \mathcal{Y}, \quad (2.8)
\]
where \( \mu \in \mathcal{H}_{00} \) is the ground mean, \( \eta_X(\cdot) \in \mathcal{H}_{10} \) and \( \eta_Z(\cdot) \in \mathcal{H}_{01} \) are the main effects, and \( \eta_{XZ} \in \mathcal{H}_{11} \) is the interaction. In the following lemma, we show that the distribution of \( X \) given group \( Z = 0 \) and the distribution of \( X \) given group \( Z = 1 \) are the same if and only if the interaction term \( \eta_{XZ} \) is zero.

Lemma 2.5. \( f_{X|Z=0}(\cdot) = f_{X|Z=1}(\cdot) \) if and only if \( \eta_{XZ} = 0 \).

Following this lemma and the probabilistic decomposition of \( \mathcal{H} \) in (2.6), \( f_{X|Z=0}(\cdot) = f_{X|Z=1}(\cdot) \) is equivalent to \( \eta \in \mathcal{H}_0 := \mathcal{H}_{00} \oplus \mathcal{H}_{10} \oplus \mathcal{H}_{01} \). Hence, we rewrite the two-sample test in (1.1) as the following hypothesis testing
\[
H_0 : \eta \in \mathcal{H}_0 \text{ vs. } H_1 : \eta \in \mathcal{H} \setminus \mathcal{H}_0, \quad (2.9)
\]
where \( \mathcal{H} \setminus \mathcal{H}_0 \) denotes the set difference of \( \mathcal{H} \) and \( \mathcal{H}_0 \).
3 Penalized likelihood ratio test

Suppose that \( y_i = (x_i, z_i), \ i = 1, \ldots, n, \) are iid observations generated from \((X, Z)\) with the log-transformed joint density \( \eta(x, z) \). Let \( \ell_{n, \lambda}\) be the negative penalized likelihood function defined as

\[
\ell_{n, \lambda}(\eta) = -\frac{1}{n} \sum_{i=1}^{n} \{\eta(y_i) + \sum_{z=0}^{1} \int_{X} e^{\eta(x, z)} dx\} + \frac{\lambda}{2} J(\eta), \quad \eta \in H, \tag{3.1}
\]

with a penalty parameter \( \lambda > 0 \) and a penalty function \( J(\cdot) \). Following Gu (2013), we define \( J(\cdot) \) as follows,

\[
J(\eta) = \theta_{10}^{-1} J_{10}(\eta) + \theta_{01}^{-1} J_{01}(\eta) + \theta_{11}^{-1} J_{11}(\eta), \tag{3.2}
\]

where \( \theta_{10}, \theta_{01}, \theta_{11} > 0 \) are tunable parameters, and \( J_{10}, J_{01}, J_{11} \) are norms induced by the inner products on \( H_{10}, H_{01}, H_{11} \) respectively. Specifically, \( J_{10} \) penalizes the roughness of the mean function \( \mathbb{E}_z \eta(x, z) \), \( J_{01} \) penalizes the variance of the marginal means \( \mathbb{E}_x \{\eta(x, z)\} \), and \( J_{11} \) penalizes the roughness of the deviations \( \eta - \mathbb{E}_z \eta(\cdot, z) - \mathbb{E}_x \eta(x, \cdot) \).

Let \( \hat{\eta}_{n, \lambda}^0 \) and \( \hat{\eta}_{n, \lambda} \) be the penalized likelihood estimators of \( \eta \) respectively under \( H_0 \) and \( H_1 \) in (2.9),

\[
\hat{\eta}_{n, \lambda}^0 = \arg\min_{\eta \in H_0} \ell_{n, \lambda}(\eta) \quad \text{and} \quad \hat{\eta}_{n, \lambda} = \arg\min_{\eta \in H} \ell_{n, \lambda}(\eta).
\]

The integral in (3.1) guarantees that such estimators fulfill the unitary constraint as shown by Silverman (1982). Numerically, we apply the reparametrization trick by expressing \( \eta \) in a finite-dimensional space

\[
\eta(y) = \sum_{i=1}^{n} \sum_{\beta=00,01,10} K^\beta(y_i, y) c_i := \xi_0^T c \quad \text{for} \ \eta \in H_0,
\]

\[
\eta(y) = \sum_{i=1}^{n} \sum_{\beta=00,01,10,11} K^\beta(y_i, y) c_i := \xi^T c \quad \text{for} \ \eta \in H.
\]

Plugging into (3.1), the calculation of \( \hat{\eta}_{n, \lambda}^0 \) and \( \hat{\eta}_{n, \lambda} \) reduces to the minimization of

\[
\frac{1}{n} Q_0 c + \int_y \exp{\xi_0^T c} + \frac{\lambda}{2} c^T Q_0 c \quad \text{and} \quad \frac{1}{n} Q c + \int_y \exp{\xi^T c} + \frac{\lambda}{2} c^T Q c
\]

with respect to \( c \), where the \( ij \)th entry of \( Q_0 \) is \( \sum_{\beta=00,01,10} K^\beta(y_i, y_j) \) and the \( ij \)th entry of \( Q_1 \) is \( \sum_{\beta=00,01,10,11} K^\beta(y_i, y_j) \).
We now propose the following penalized likelihood ratio (PLR) test statistic for testing the hypothesis (2.9):

\[ \text{PLR}_{n, \lambda} = \ell_{n, \lambda}(\eta_{0}) - \ell_{n, \lambda}(\hat{\eta}_{n, \lambda}). \]

In Section 3.1, we show that \( \text{PLR}_{n, \lambda} \) is asymptotically \( \chi^2 \) distributed, fulfilling the Wilks’ phenomenon, based on which an asymptotically valid testing rule will be proposed. We further show that \( \text{PLR}_{n, \lambda} \) is minimax optimal in the sense of Theorem 4.3.

3.1 Asymptotic distribution and Wilks’ Phenomenon

In this subsection, we present the asymptotic distribution of our PLR test (see Theorem 3.4). The proof relies on a technical lemma about the eigen-structures of \( H_0 \) and \( H \); see Lemma 3.1 below. For any \( \eta, \tilde{\eta} \in H \), define

\[ \langle \eta, \tilde{\eta} \rangle = V(\eta, \tilde{\eta}) + \lambda J(\eta, \tilde{\eta}), \tag{3.3} \]

where \( V(\eta, \tilde{\eta}) = \mathbb{E}_{\eta^*}\{\eta(Y)^\top \tilde{\eta}(Y)\} \) with expectation taken under the true \( \eta^* \), and \( J \) is a bilinear form corresponding to (3.2). It holds that \( H \) and \( H_0 \), endowed with the inner product (3.3), are both RKHSs; see Lemma 3.2 in the Appendix. In the following lemma, we characterize the eigenvalues and eigenvectors of the Rayleigh quotient \( V/J \).

Lemma 3.1. (a) There exist a sequence of functions \( \{\xi_p\}_{p=1}^{\infty} \subset H \) and a sequence of nonnegative eigenvalues \( \{\rho_p\}_{p=1}^{\infty} \) with \( \rho_p \asymp p^{2m} \) such that

\[ V(\xi_p, \xi_{p'}) = \delta_{p, p'}, \quad J(\xi_p, \xi_{p'}) = \rho_p \delta_{p, p'}, \quad \text{for all } p, p' \geq 1, \tag{3.4} \]

and that any \( \eta \in H \) can be written as \( \eta = \sum_{p=1}^{\infty} V(\eta, \xi_p)\xi_p \).

(b) Moreover, there is a proper subset \( \{\rho_p^0\}_{p=1}^{\infty} \) of \( \{\rho_p\}_{p=1}^{\infty} \) satisfying that \( \{\xi_p^0\}_{p=1}^{\infty} \subset H_0 \) and for any \( \eta \in H_0 \), \( \eta = \sum_{p=1}^{\infty} V(\eta, \xi_p^0)\xi_p^0 \). Convergence of both series holds under (3.3).

(c) \( \rho_p^0 \asymp p^{2m} \), where \( \{\rho_p^0\}_{p=1}^{\infty} \subset \{\rho_p\}_{p=1}^{\infty} \) is a subset of eigenvalues corresponding to \( \{\xi_p^0\}_{p=1}^{\infty} \equiv \{\xi_p\}_{p=1}^{\infty} \setminus \{\xi_p^0\}_{p=1}^{\infty} \). The set \( \{\xi_p^0\}_{p=1}^{\infty} \) generate the orthogonal complement of \( H_0 \) under the inner product (3.3).

Lemma 3.1 introduces an eigensystem that simultaneously diagonalizes the bilinear forms \( V \) and \( J \). This eigensystem does not depend on the unknown null density, and only depends on the functional space \( H \). Moreover,
$H_0$ can be generated by a proper subset of the eigenfunctions, which is crucial for analyzing the likelihood ratios.

Let $\langle \cdot, \cdot \rangle_0$ denote the restriction of $\langle \cdot, \cdot \rangle$ on the subspace $H_0$. Specifically, for any $\eta, \tilde{\eta} \in H_0$, $\langle \eta, \tilde{\eta} \rangle_0 = \langle \eta, \tilde{\eta} \rangle$. Then $H$ and $H_0$ are both RKHS’s endowed with these inner products.

**Lemma 3.2.** $(H, \langle \cdot, \cdot \rangle)$ and $(H_0, \langle \cdot, \cdot \rangle_0)$ are both RKHS’s with the corresponding inner products.

Following Lemma 3.2, there exist reproducing kernel functions $\tilde{K}(\cdot, \cdot)$ and $\tilde{K}^0(\cdot, \cdot)$ defined on $Y \times Y$, satisfying, for any $y \in Y$, $\eta \in H$, $\tilde{\eta} \in H_0$:

$$\tilde{K}_y(\cdot) \equiv \tilde{K}(y, \cdot) \in H, \quad \tilde{K}^0_y(\cdot) \equiv \tilde{K}^0(y, \cdot) \in H_0,$$

$$\langle \tilde{K}_y, \eta \rangle = \eta(y), \quad \langle \tilde{K}^0_y, \tilde{\eta} \rangle_0 = \tilde{\eta}(y). \quad (3.5)$$

Following Shang et al. (2013), we further introduce positive definite self-adjoint operators $W_\lambda : H \rightarrow H$ and $W_\lambda^0 : H_0 \rightarrow H_0$ such that

$$\langle W_\lambda \eta, \tilde{\eta} \rangle = \lambda J(\eta, \tilde{\eta}) \quad \text{for all } \eta, \tilde{\eta} \in H,$$

$$\langle W_\lambda^0 \eta, \tilde{\eta} \rangle_0 = \lambda J_0(\eta, \tilde{\eta}) \quad \text{for all } \eta, \tilde{\eta} \in H_0, \quad (3.6)$$

where $J_0(\eta, \tilde{\eta}) = \theta^{-1}_0 J_{01}(\eta, \tilde{\eta}) + \theta^{-1}_0 J_{10}(\eta, \tilde{\eta})$ is the restriction of $J$ over $H_0$. By (3.6) we get $\langle \eta, \tilde{\eta} \rangle = V(\eta, \tilde{\eta}) + \langle W_\lambda \eta, \tilde{\eta} \rangle, \quad \langle \eta, \tilde{\eta} \rangle_0 = V(\eta, \tilde{\eta}) + \langle W_\lambda^0 \eta, \tilde{\eta} \rangle_0$.

**Proposition 3.3.** For any $y \in Y$ and $\eta \in H$, we have

$$\|\eta\|^2 = \sum_{p=1}^{\infty} |V(\eta, \xi_p)|^2 (1 + \lambda \rho_p),$$

$$\tilde{K}_y(\cdot) = \sum_{p=1}^{\infty} \frac{\xi_p(y)}{1 + \lambda \rho_p} \xi_p(\cdot), \quad \tilde{K}^0_y(\cdot) = \sum_{p=1}^{\infty} \frac{\xi^0_p(y)}{1 + \lambda \rho^0_p} \xi^0_p(\cdot),$$

$$W_\lambda \xi_p(\cdot) = \frac{\lambda \rho_p}{1 + \lambda \rho_p} \xi_p(\cdot), \quad W_\lambda^0 \xi^0_p(\cdot) = \frac{\lambda \rho^0_p}{1 + \lambda \rho^0_p} \xi^0_p(\cdot),$$

where $\{\rho_p, \xi_p\}_{p=1}^{\infty}$ and $\{\rho^0_p, \xi^0_p\}_{p=1}^{\infty}$ are eigensystem defined in Lemma 3.1.

The eigenvalue for $\tilde{K}$ are $\{(1 + \lambda \rho_p)^{-1}\}_{p=1}^{\infty}$, having a slower decay rate due to scaling by $\lambda$. $\tilde{K}$ can be viewed as a scaled kernel comparing with the product kernel $K^H = K^{00} + K^{01} + K^{10} + K^{11}$ introduced in Lemma 2.4. Note that $\text{trace}(\tilde{K}) = \sum_{p=1}^{\infty} (1 + \lambda \rho_p)^{-1} \sim \lambda^{-1/(2m)}$ is the effective dimension that measures the complexity of $H$; see Bartlett et al. (2005); Mendelson (2002).

12
We remark that trace($\tilde{K}$) plays the same role as the Bernstein lower critical dimension in Lemma 4.2 by letting $\delta^2 = \lambda$.

Let $D, D^2, D^3$ be the first-, second- and third-order Frechét derivatives of $l_{n,\lambda}(\eta)$. Based on the above notation, these derivatives can be summarized as follows. Let $y = (x, z)$. For any $\eta, \Delta \eta_1, \Delta \eta_2, \Delta \eta_3 \in \mathcal{H}$,

\[
Dl_{n,\lambda}(\eta)\Delta \eta_1 = -\frac{1}{n}\sum_{i=1}^{n} \Delta \eta_1(y_i) + \int_{Y} \Delta \eta_1(y)e^{\eta(y)}dy + \lambda j(\eta, \Delta \eta_1)
\]

\[
= \langle -\frac{1}{n}\sum_{i=1}^{n} \tilde{K}_{y_i} + \mathbb{E}_{\eta}\tilde{K}_{y} + W_{\lambda\eta}, \Delta \eta_1 \rangle
\]

\[
\equiv \langle S_{n,\lambda}(\eta), \Delta \eta_1 \rangle, \tag{3.7}
\]

\[
D^2l_{n,\lambda}(\eta)\Delta \eta_1 \Delta \eta_2 = \int_{Y} \Delta \eta_1(y)\Delta \eta_2(y)e^{\eta(y)}dy + \lambda J(\Delta \eta_1, \Delta \eta_2), \tag{3.8}
\]

\[
D^3l_{n,\lambda}(\eta)\Delta \eta_1 \Delta \eta_2 \Delta \eta_3 = \int_{Y} \Delta \eta_1(y)\Delta \eta_2(y)\Delta \eta_3(y)e^{\eta(y)}dy. \tag{3.9}
\]

The second equality of (3.7) is due to the reproducing property (3.5) and that

\[
\int_{Y} \Delta \eta(y)e^{\eta(y)}dy = \mathbb{E}_{\eta}\Delta \eta_1(y) = \mathbb{E}_{\eta}\langle \tilde{K}_{y}, \Delta \eta_1 \rangle = \langle \mathbb{E}_{\eta}\tilde{K}_{y}, \Delta \eta_1 \rangle.
\]

Let $h = \lambda \frac{1}{\sqrt{n}}$ and define $V_{\eta}(\tilde{\eta}) = \mathbb{E}_{\eta}\{\tilde{\eta}^2(y)\}$ for any $\eta, \tilde{\eta} \in \mathcal{H}$. In particular, $V(\cdot) = V_{\eta}(\cdot)$. We need the following condition to regularize the behavior of $V(\cdot)$ and $V_{\eta}(\cdot)$.

**Assumption 1.** There exists a convex set $B \subset \mathcal{H}$ around $\eta^*$ and a constant $c > 0$ such that, for any $\eta \in B$, $cV(\eta) \leq V_{\eta}(\eta)$. Furthermore, with the probability approaching one, $\tilde{\eta}_{n,\lambda} \in B$; and under $H_0$, with the probability approaching one, $\tilde{\eta}_{0,\lambda}^0 \in B$.

Assumption 1 is commonly used in literature for deriving the rates of density estimates; see Theorem 9.3 of Gu (2013). This condition is satisfied when the members of $B$ have uniform upper and lower bounds on the domain $\mathcal{Y}$, as well as that $\tilde{\eta}_{n,\lambda}$ and $\tilde{\eta}_{0,\lambda}^0$ are stochastically bounded. The following theorem provides the asymptotic distribution for the PLR test statistic under Assumption 1.
Theorem 3.4. Suppose \( m \geq 1 \) and Assumption 1 holds, \( nh^{2m+1} = O(1) \), \( nh^2 \to \infty \) as \( n \to \infty \). Under \( H_0 \), we have
\[
\frac{2n \cdot PLR_{n,\lambda} - \theta_\lambda}{\sqrt{2}\sigma_\lambda} \overset{d}{\to} N(0,1), \quad n \to \infty,
\] (3.10)
where \( \theta_\lambda = \sum_{p=1}^{\infty} \frac{1}{1+\lambda \rho_p}, \quad \sigma_\lambda^2 = \sum_{p=1}^{\infty} \frac{1}{(1+\lambda \rho_p)^2} \).

We notice that \( h \approx n^{-c} \) with \( \frac{1}{2m+1} \leq c \leq \frac{1}{2} \) satisfies the rate conditions in Theorem 3.4, so the asymptotic distribution (3.10) holds under a wide-ranging choice of \( h \). The quantities \( \theta_\lambda \) and \( \sigma_\lambda \) solely depend on the eigenvalues \( \rho_p \)'s and \( \lambda \). Based on (3.10), we propose the following decision rule \( \Phi_{n,\lambda} \) at the significance level \( \alpha \):
\[
\Phi_{n,\lambda}(\alpha) = 1(\left| 2n \cdot PLR_{n,\lambda} - \theta_\lambda \right| \geq z_{1-\alpha/2} \sqrt{2}\sigma_\lambda)
\] (3.11)
where \( 1(\cdot) \) is the indicator function, \( z_{1-\alpha/2} \) is the \( 1-\alpha/2 \) quantile of the standard normal distribution. Hence, we reject \( H_0 \) at the significance level \( \alpha \) if \( \Phi_{n,\lambda} = 1 \). Theorem 3.4 is closely related to the Wilks' phenomenon demonstrated in the classical nonparametric/semiparametric regression framework (Fan et al., 2001; Shang et al., 2013; Cheng and Shang, 2015; Shang and Cheng, 2015). Specifically, let \( r_\lambda = \frac{\theta_\lambda}{\sigma_\lambda^2} \), then (3.10) implies that, as \( n \to \infty \),
\[
\frac{2nr_\lambda \cdot PLR_{n,\lambda} - r_\lambda \theta_\lambda}{\sqrt{2r_\lambda \theta_\lambda}} \overset{d}{\to} N(0,1).
\]
Therefore, \( 2nr_\lambda \cdot PLR_{n,\lambda} \) is asymptotically distributed as a \( \chi^2 \) distribution with degrees of freedom \( r_\lambda \theta_\lambda \). In practice, \( \rho_p^\perp \)'s can be estimated by the sample eigenvalues of the empirical kernel matrix, from which the quantities \( r_\lambda \) and \( \theta_\lambda \) can be accurately approximated. Our numerical study in Sections 6 and 7 adopt such an approximation and the performance is satisfactory.

3.2 Power Analysis and Minimaxity

We first introduce some notation and terminology, before we introduce the minimax principle pioneered in Ingster (1989) to characterize the level of difficulty for testing the hypothesis (2.9). Denote the observed samples by \( (y_1, \ldots, y_n) \). For a generic 0-1 valued testing rule \( \Phi = \Phi(y_1, \ldots, y_n) \) and a separation rate \( d_n > 0 \), define the total error \( \text{Err}(\Phi, d_n) \) of \( \Phi \) under \( d_n \) as follows:
\[
\text{Err}(\Phi, d_n) = \mathbb{E}_{H_0} \{ \Phi \} + \sup_{\|\eta_X\| \geq d_n} \mathbb{E}_{\eta} \{ 1 - \Phi \},
\] (3.12)
where \( \mathbb{E}_{H_0} \{ \cdot \} \) denotes the expectation under \( H_0 \). The first and second terms on the right side of (3.12) represent type I and type II errors of \( \Phi \) respectively. Specifically, we define \( d_{n,h} \) as the distinguishable rate for testing rule \( \Phi_{n,\lambda} \).

In this section, we investigate the power of PLR under local alternatives. Theorem 3.5 shows that the power of PLR approaches one, provided that the norm of \( \eta^{*}_{XZ} \) is bounded away from zero by an order \( d_{n,h} := \sqrt{h^{2m} + (nh^{1/2})^{-1}} \). Our result owes much to the analytic expression of independence (in terms of interactions) based on the proposed probabilistic tensor product decomposition framework.

The following theorem shows that our PLR can achieve a high power provided that \( \eta^{*}_{XZ} \), the interaction term in the probabilistic decomposition of \( \eta^{*} \), has a norm bounded below by \( d_{n,h} \). Let \( P_{\eta^{*}} \) denote the probability measure induced under \( \eta^{*} \), \( \| \eta^{*} \|_{\sup} \) the supremum norm over \( Y \), and \( \| \eta^{*} \|_2 = \sqrt{\mathbb{V}(\eta^{*})} \).

**Theorem 3.5.** Suppose Assumption 1 holds, \( m > 3/2 \), \( \eta^{*} \in \mathcal{H} \) with \( \| \eta^{*}_{XZ} \|_{\sup} = o(1) \), \( J(\eta^{*}_{XZ}) < \infty \), \( \| \eta^{*}_{XZ} \|_2 \geq d_{n,h} \). For any \( \varepsilon \in (0,1) \), there exists a positive \( N_\varepsilon \) such that, for any \( n \geq N_\varepsilon \), \( \mathbb{P}_{\eta^{*}}(\Phi_{n,\lambda}(\alpha) = 1) \geq 1 - \varepsilon \). When \( h \asymp h^{*} \equiv n^{-2/(4m+1)} \), \( d_{n,h} \) is upper bounded by \( d^{*}_{n} \equiv n^{-2m/(4m+1)} \).

Theorem 3.5 demonstrates that, when \( h \asymp h^{*} \), PLR can successfully detect any local alternatives, provided that they separate from the null at least by \( d^{*}_{n} \). In the next section, we shown that this upper bound is unimprovable by establishing the minimax lower bound of distinguishable rate for general two-sample test. It means that no test can successfully detect the local alternatives if they separate from the null by a rate faster than \( d^{*}_{n} \), we claim that our PLR test is minimax optimal.

Let \( \lambda^{*} = n^{-4m/(4m+1)} \). For any \( \varepsilon \in (0,1) \) and \( \alpha \in (0,\varepsilon) \), Theorem 3.4 shows that \( \mathbb{E}_{H_0}(\Phi_{n,\lambda^{*}}(\alpha)) \) tends to \( \alpha \); Theorem 3.5 shows that \( \mathbb{E}_{\eta^{*}}\{1 - \Phi_{n,\lambda^{*}}(\alpha)\} \leq \varepsilon - \alpha \), provided that \( \| \eta^{*}_{XZ} \|_2 \geq C_{\varepsilon - \alpha}d^{*}_{n} \) for a large constant \( C_{\varepsilon - \alpha} \). This implies that, asymptotically,

\[
\text{Err}(\Phi_{n,\lambda^{*}}(\alpha), C_{\varepsilon - \alpha}d^{*}_{n}) \leq \varepsilon. \tag{3.13}
\]

In other words, the total error of PLR is controlled by an arbitrary \( \varepsilon \) provided that the null and local alternatives are separated by \( d^{*}_{n} \).

### 4 Minimax lower bound of the distinguishable rate

For any \( \varepsilon \in (0,1) \), define the minimax separation rate \( d^{\dagger}_{n}(\varepsilon) \) as

\[
d^{\dagger}_{n}(\varepsilon) = \inf\{d_n > 0 : \inf_{\Phi} \text{Err}(\Phi, d_n) \leq \varepsilon\}, \tag{4.1}
\]
where the infimum in (4.1) is taken over all 0-1 valued testing rules based on samples $y_i$'s. Here we consider the local alternative by assuming $\|\eta\|_\mathcal{H} < 1/2$. And $d_n^\dagger(\varepsilon)$ characterizes the smallest separation between the null and local alternatives such that there exists a testing approach with a total error of at most $\varepsilon$. Next we establish a lower bound for $d_n^\dagger$, i.e., if $d_n$ is smaller than a certain lower bound, there exists no test that can distinguish the alternative and null.

We first introduce a geometric interpretation of the hypothesis testing (2.9). Geometrically, $\mathcal{E} = \{\eta \in \mathcal{H} : \|\eta\|_H < 1/2\}$ is an ellipse with axis lengths equal to eigenvalues defined in (2.7) as shown in Figure 2. For any $\eta \in \mathcal{E}$, the projection of $\eta$ on $\mathcal{E}_{11} := \mathcal{H}_{11} \cap \mathcal{E}$ is $\eta_{XZ}$. The magnitude of the interaction $\eta_{XZ}$ can be qualified by $\|\eta_{XZ}\|_2$. The distinguishable rate $d_n$ is the radius of the sphere centered at $\eta_{XZ} = 0$ in $\mathcal{E}_{11}$.

![Figure 2: Geometric interpretation of the distinguishable rate of the testing $H_0^\ast$.](image)

Intuitively, the testing will be harder when the projection of $\eta$ on $\mathcal{H}_{11}$ is closer to the original point $\eta_{XZ} = 0$. We then introduce the Bernstein width in Pinkus (2012) to characterize the testing difficulty. For a compact set $C$, the Bernstein $k$-width is defined as

$$b_{k,2}(C) := \arg\max_{r \geq 0} \{B_{k+1}^2(\mathbb{R}) \subset C \cap S \text{ for some subspace } S \in S_{k+1}\}$$

where $S_{k+1}$ denotes the set of all $k + 1$ dimensional subspaces, $B_{k+1}^2(\mathbb{R})$ is a $(k+1)$-dimensional $L_2$-ball with radius $r$ centered at $\eta_{XZ} = 0$ in $\mathcal{H}_{11}$. Based on the Bernstein width, we give an upper bound of the testing radius, i.e.,
for any $\eta$ projected in the ball with radius less than the certain bound, the total error is larger than $1/2$.

**Lemma 4.1.** For any $\eta \in \mathcal{H}$, we have

$$Err(\Phi, d_n) \geq 1/2$$

for all

$$d_n \ll r_B(\delta^*) := \sup\{\delta \mid \delta \leq \frac{1}{2\sqrt{n}}(k_B(\delta))^{1/4}\}$$

where $k_B(\delta) := \arg\max_k \{b_{k-1,2}(\mathcal{H}_{11}) \geq \delta^2\}$ is the Bernstein lower critical dimension and $r_B(\delta)$ is called the Bernstein lower critical radius.

In Lemma 4.1, we show that when $d_n$ is less than $r_B(\delta)$, there is no test can distinguish the alternative from the null hypothesis. In order to achieve a non-trivial power, we need $d_n$ to be larger than the Bernstein lower critical radius $r_B(\delta^*)$. The critical radius $r_B(\delta)$ depends on the shape of the space $\mathcal{H}_{11}$ which is characterized by its eigenvalues defined in (2.7). The lower bound of $k_B(\delta)$ depends on the decay rate of the eigenvalues for $\mathcal{H}_{11}$. According to the Liebig’s law, the radius of $k$-dimensional ball that can be embedded into $\mathcal{H}_{11}$ is determined by $k$th largest eigenvalue. In the next lemma, we characterize the lower bound of $k_B(\delta)$ by the largest $k$ such that the $k$th largest eigenvalue is larger than $\delta^2$.

**Lemma 4.2.** Let $\gamma_k$ be the $k$th largest eigenvalue of $\mathcal{H}_{11}$. Then we have

$$k_B(\delta) > \arg\max_k \{\sqrt{\gamma_k} \geq \delta\}$$

(4.3)

Note that $\gamma_k \propto k^{-2m}$, then $\arg\max_k \{\sqrt{\gamma_k} \geq \delta\} \propto \delta^{1/m}$. Plug in the lower bound of $k_B(\delta)$ to Lemma 4.1, we achieve $r_B(\delta^*)$, which is the minimax lower bound for the distinguishable rate in the following theorem.

**Theorem 4.3.** Suppose $\eta \in \mathcal{H}$. For any $\varepsilon \in (0, 1)$, the minimax distinguishable rate for the testing hypotheses (2.9) is $d_n(\varepsilon) \gtrsim n^{-2m/(4m+1)}$.

Theorem 4.3 provides a general guidance to justify a local minimax test for testing $\eta_{XZ} = 0$. The proof of Theorem 4.3 is presented in the Appendix. Combining with the upper bound derived in Theorem 3.5, we show that this lower bound is sharp.
5 Connection to Maximum Mean Discrepancy (MMD)

In this section, we revisit the MMD in Gretton et al. (2012) for the hypothesis testing (1.1) from the viewpoint of the likelihood principle based on our proposed probabilistic decomposition of $\mathcal{H}$.

We first briefly summarize the MMD. Given the kernel function $K^{(X)}$ on $\mathcal{H}^{(X)}$, denote the embedding that maps a probability distribution $f_{X|Z=z}$ to $\mathcal{H}^{(X)}$ by $\mu_z(\cdot) = \int_{\mathcal{X}} K^{(X)}(x, \cdot) f_{X|Z=z}(x) dx$, then the squared MMD between $f_{X|Z=0}$ and $f_{X|Z=1}$ is defined as the squared distance between embeddings of distributions to reproducing kernel Hilbert spaces (RKHS):

$$MMD^2(\mathcal{H}^{(X)}; f_{X|Z=0}, f_{X|Z=1}) := \|\mu_0 - \mu_1\|_{\mathcal{H}^{(X)}}$$

$$= \langle \mu_0, \mu_0 \rangle_{\mathcal{H}^{(X)}} + \langle \mu_1, \mu_1 \rangle_{\mathcal{H}^{(X)}} - 2 \langle \mu_0, \mu_1 \rangle_{\mathcal{H}^{(X)}}$$

$$= \mathbb{E}_{x_1, x_2}[K^{(X)}(x_1, x_2)] + \mathbb{E}_{x_1', x_2'}[K^{(X)}(x_1', x_2')] - 2 \mathbb{E}_{x_1, x_1'}[K^{(X)}(x_1, x_1')]$$

where $x_1, x_2 \sim f_{X|Z=0}$, and $x_1', x_2' \sim f_{X|Z=1}$. An estimate of the squared MMD is then provided by

$$MMD_b^2(\mathcal{H}^{(X)}; f_{X|Z=0}, f_{X|Z=1}) = \frac{1}{n^2_0} \sum_{i, j | z_i = z_j = 0} K^{(X)}(x_i, x_j)$$

$$- \frac{2}{n_0 n_1} \sum_{i, j | z_i \neq z_j} K^{(X)}(x_i, x_j) + \frac{1}{n^2_1} \sum_{i, j | z_i = z_j = 1} K^{(X)}(x_i, x_j); \quad (5.1)$$

We replace each instance of $K^{(X)}(x_i, x_j)$ in the sum of (5.1) by the centralized kernel $K_1^{(X)}(x_i, x_j)$ introduced in Lemma 2.3, and the $MMD_b^2$ remains the same since the mean term is canceled out under $H_0$ that $\mu_0 = \mu_1$. Therefore, we have

$$MMD_b^2(\mathcal{H}^{(X)}; f_{X|Z=0}, f_{X|Z=1}) = \frac{1}{n^2_0} \sum_{i, j | z_i = z_j = 0} K_1^{(X)}(x_i, x_j)$$

$$- \frac{2}{n_0 n_1} \sum_{i, j | z_i \neq z_j} K_1^{(X)}(x_i, x_j) + \frac{1}{n^2_1} \sum_{i, j | z_i = z_j = 1} K_1^{(X)}(x_i, x_j);$$

We next show that the MMD estimate is equivalent to the squared score function based on the likelihood functional without penalty. Let $\ell_n$ be the negative likelihood functional defined as $\ell_n(\eta) = -\frac{1}{n} \sum_{i=1}^n \eta(y_i)$, and $LR_n$ be the likelihood ratio functional defined as

$$LR_n(\eta) = \ell_n(\eta) - \ell_n(P_{H_0} \eta) = -\frac{1}{n} \sum_{i=1}^n \{\eta(y_i) - P_{H_0} \eta(y_i)\}, \quad \eta \in \mathcal{H}, \quad (5.2)$$
where \( P_{\mathcal{H}_0} \) is the projection operator from \( \mathcal{H} \) to \( \mathcal{H}_0 \). Using the reproducing property, we rewrite (5.2) as

\[
LR_n(\eta) = -\frac{1}{n} \sum_{i=1}^{n} \{(K^{\mathcal{H}}_{y_i}, \eta)_{\mathcal{H}} - (K^{\mathcal{H}_0}_{y_i}, \eta)_{\mathcal{H}}\}, \tag{5.3}
\]

where \( K^{\mathcal{H}} = K^{00} + K^{01} + K^{10} + K^{11} \) is the kernel for \( \mathcal{H} \) and \( K^{\mathcal{H}_0} = K^{00} + K^{01} + K^{10} \) is the kernel for \( \mathcal{H}_0 \).

Now we calculate the Fréchet derivative of the likelihood ratio functional as the score function, i.e.,

\[
DLR_n(\eta) \Delta \eta = \left\langle -\frac{1}{n} \sum_{i=1}^{n} (K^{\mathcal{H}}_{y_i} - K^{\mathcal{H}_0}_{y_i}), \Delta \eta \right\rangle_{\mathcal{H}} = \left\langle \frac{1}{n} \sum_{i=1}^{n} K^{11}_{y_i}, \Delta \eta \right\rangle_{\mathcal{H}},
\]

where \( K^{11} \) is the kernel for \( \mathcal{H}_{11} \). We further define a score test statistics as the squared \( \| \cdot \|_{\mathcal{H}} \) norm of the score function as follows

\[
S_n^2 = \| \frac{1}{n} \sum_{i=1}^{n} K^{11}_{y_i} \|^2_{\mathcal{H}} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} K^{11}(y_i, y_j), \tag{5.4}
\]

where the second equality holds by the reproducing property. Recall that by Lemma 2.2 the kernel on \( \mathcal{H}_1^{(Z)} \) is \( K_1^{(Z)}(z_i, z_j) = 1\{z_i = z_j\} - \omega_{z_i} - \omega_{z_j} + \sum_{l=1}^{2} \omega_l^2 \), and by Lemma 2.3, the kernel on \( \mathcal{H}_1^{(X)} \) is \( K_1^{(X)}(x_i, x_j) = K(x_i, x_j) - \mathbb{E}_x[K(x, x_j)] - \mathbb{E}_y[K(x_i, y)] + \mathbb{E}_{x,y}K(x, y) \). Then we have \( K^{11}(y_i, y_j) = K_1^{(Z)}(z_i, z_j)K_1^{(X)}(x_i, x_j) \) based on Lemma 2.4. Let \( \omega_0 = n_0/(n_0 + n_1) \) and \( \omega_1 = n_1/(n_0 + n_1) \) where \( n_0 \) is the number of observations in group 0 and \( n_1 \) is the number of observations in group 1. The score test statistics in (5.4) can be rewritten as

\[
\frac{4n_0n_1}{(n_0 + n_1)^2} S_n^2 = \frac{1}{n_0^2} \sum_{\{i,j \mid z_i = z_j = 0\}} K_1^{(X)}(x_i, x_j) - \frac{2}{n_0n_1} \sum_{\{i,j \mid z_i \neq z_j\}} K_1^{(X)}(x_i, x_j) + \frac{1}{n_1^2} \sum_{\{i,j \mid z_i = z_j = 1\}} K_1^{(X)}(x_i, x_j). \]

Thus, the scaled score test statistic is equivalent to the MMD test statistic, i.e.,

\[
\frac{4n_0n_1}{(n_0 + n_1)^2} S_n^2 = \text{MMD}^2_{b}(\mathcal{H}^{(X)}; f_{X|Z=0}, f_{X|Z=1}) \tag{5.5}
\]

under the null hypothesis. When \( n_0 = n_1 \), i.e. the number of observations are equal in two groups, we have \( S_n^2 = \text{MMD}^2_{b}(\mathcal{H}^{(X)}; f_{X|Z=0}, f_{X|Z=1}) \).

19
However, the minimax optimality of the score test statistics $S_n^2$ based on the likelihood ratio cannot be guaranteed, implying the suboptimal of the MMD. Instead, in previous Section 3, we establish the testing optimality for penalized likelihood ratio test. We further show the difference between the MMD and our proposed PLR statistic. As shown in the proof of Theorem 3.4, the PLR test statistic is asymptotically equivalent to

$$PLR_{n, \lambda} \sim \|S_{n, \lambda}^0(\eta) - S_{n, \lambda}(\eta)\|^2 \sim \frac{1}{n} \left\| \sum_{i=1}^{n} \tilde{K}^1_{y_i} \right\|^2,$$

where $S_{n, \lambda}$ and $S_{n, \lambda}^0$ are the score functions defined in (3.7) based on the penalized likelihood ratio functional, and $\tilde{K}^1_{y_i} = K_{y_i} - K^0_{y_i} = \sum_{p=1}^{\infty} \frac{\xi_{\perp}^{+}(y_i)\xi_{\perp}^{-}(y_i)}{1 + \lambda \rho_{\perp}^p}$. Notice that $\tilde{K}^1$ can be viewed as a scaled version of the product kernel $K^{11}$ by replacing the eigenvalues $\{\rho_{\perp}^p\}$ with $\{1 + \lambda \rho_{\perp}^p\}$. By choosing $\lambda = \lambda^*$, $\text{trace}(\tilde{K}^1) = \sum_{p=1}^{\infty} \frac{1}{1 + \lambda \rho_{\perp}^p} \asymp n^{2/(4m+1)}$ matches the lower bound of $k_B(d_n^*)$ with $d_n = n^{-2m/(4m+1)}$ as the minimax lower bound for the distinguishable rate in Lemma 4.2. In contrast, the MMD is based on kernel $K^{11}$ without regularization, thus the optimality of the power performance cannot be guaranteed.

6 Simulation Study

In this section, we demonstrate the finite sample performance of the proposed test alongside its competitors through a simulation study. We choose the KS test as one of the most popular CDF-based test, the MMD test (Gretton et al., 2012) as a kernel-based test, and the ELT (Cao and Van Keilegom, 2006) as a density-based test. In the supplementary material, we compare with the Anderson-Darling (AD) test (Scholz and Stephens, 1987), density equality test (DET) (Anderson et al., 1994), Mann-Whitney U-Test (MW) (Mann and Whitney, 1947), and dynamic slicing test (DSLICE) (Jiang et al., 2015). We summarize the simulation results in Supplementary Table 1-4.

6.1 Comparison on Gaussian Densities and Their Mixtures

The samples $y_i = (X_i, Z_i), i = 1, \ldots, n,$ were generated as follows. We first generated $Z_i \sim \text{Bernoulli}(0.5)$, with 0/1 representing the control/treatment group. Then $X_i$’s were independently generated from the conditional distribution $f_{X|Z}(x)$ in the following Settings 1 and 2. In each setting, we
chose the averaged sample size $n$ in each group as 125, 250, 375, 500, 625, 750, 875, 1000. We chose the roughness parameter by directly plugging in $\lambda = n^{-4m/(4m+1)}$, which is theoretically guaranteed by Theorem 4.3, and estimate $\theta$s by using the trace of the corresponding kernel matrix which is also suggested by Gu (2013). Size and power were calculated as the proportions of rejection based on 1000 independent trials.

In Setting 1, we consider the densities of $X$ in two group are unimodal Gaussian distribution, i.e.,

$$X \mid Z = z \sim N(0, (1 + \delta_1 1_{z=1})^2),$$

where $\delta_1 = 0, 0.1, 0.2, 0.3$. In Setting 2, the densities of $X$ in two group are multimodal Gaussian distribution, i.e.,

$$X \mid Z = z \sim 0.5N(2, 1) + 0.5N(-2, (1 - \delta_2 1_{z=1})^2)$$

where $\delta_2 = 0, 0.15, 0.3, 0.45$. The coefficients $\delta_1, \delta_2$ determine the variance of the conditional distributions. In particular, $\delta_1 = 0$ or $\delta_2 = 0$ corresponds to true $H_0$ which will be used to examine the size of the test statistics.

Results are summarized in Figures 3, 4 and 5. Figure 3 displays the size of PLR, KS and MMD. It can be seen that the sizes of the three tests are close to the nominal level 0.05 in Setting 1, confirming that all tests are asymptotically valid. In Setting 2, the size of the PLR test is still asymptotic correct, while the sizes of KS and MMD are below 0.05 showing that the latter two tests are more conservative in handling bimodal distributions.

Figures 4 and 5 display the powers of the four tests. In Setting 1, i.e., Figure 3, we can see that the powers of the PLR, MMD and ELT tests
rapidly approach one when $n$ or $\delta_1$ increases. The power of the KS test increases slightly slower than the other three tests. In Setting 2, as shown in Figure 5, the power of the PLR test rapidly approaches one when $n$ or $\delta_2$ increases, whereas the powers of KS, MMD and ELT are not as high as PLR. The highest power of KS, MMD and ELT is lower than 0.55 when the averaged sample size in each group reaches 1000 and $\delta_2 = 0.45$. The results clearly demonstrate that KS, MMD and ELT are only powerful for unimodal distribution, while PLR is powerful for both unimodal and bimodal distributions.

### 6.2 Comparison on Non-Gaussian Densities and Their Mixtures

In this section, we generated data from non-Gaussian distributions. We considered distribution including a unimodal Beta distribution (Setting 3):

$$X \mid Z = z \sim \text{Beta}(2(1 + \delta_3 1_{z=1}), 2(1 + \delta_3 1_{z=1}))$$
where \( \delta_3 = 0, 0.1, 0.2, 0.3 \). Also, we consider a mixture of Beta distributions (Setting 4):

\[
X \mid Z = z \sim 0.5 \text{Beta} \left( 2(1 + \delta_4 \mathbb{1}_{z=1}), 6(1 + \delta_4 \mathbb{1}_{z=1}) \right) + 0.5 \text{Beta} \left( 6(1 + \delta_4 \mathbb{1}_{z=1}), 2(1 + \delta_4 \mathbb{1}_{z=1}) \right)
\]

where \( \delta_4 = 0, 0.15, 0.3, 0.45 \). Similar as Section 6.1, we calculated the size and power based on 1000 independent trials.

Setting 3 corresponds to a Beta distribution, while Setting 4 corresponds to a mixture of Beta distributions. With \( \delta_3 = 0 \) and \( \delta_4 = 0 \), we intended to examine the size of the test under the null hypothesis \( H_0 \). The power of the testing methods were examined with positive \( \delta_3 \)'s and \( \delta_4 \)'s.

As shown in Figure 6(a), the empirical sizes of Setting 3 were all around 0.05 for the four test procedures when the density is a unimodal Beta distribution. Whereas Figure 6(b) shows that the empirical sizes of KS, MMD and ELT tests were significantly lower than 0.05, while the sizes of PLR test...
Figure 6: Size vs. sample size in Section 6.2 for PLR, KS and MMD tests. Results were obtained under $\delta_3 = 0$ for Setting 3 and $\delta_4 = 0$ for Setting 4.

Figure 7: Power vs. sample size in Section 6.2 for PLR (a), KS (b), MMD (c) and ELT (d). Results were obtained under nonzero $\delta_3$ in Settings 3.

were still around 0.05. This demonstrates that our PLR test are asymptotically correct for both unimodal and bimodal distributions.

Figure 7 examines the power of the three tests under the Setting 3. In
Setting 3, when $\delta_3 = 0.2, 0.3$, the empirical powers of the MMD and PLR test approached 1 as $n$ increased. In contrast, the power of KS and ELT test were lower than 0.5 even when the averaged sample size in each group reaches 1000. In Setting 4, as shown in Figure 8 the power of KS, MMD and ELT test were below 0.2 even when the averaged sample size in each group is 1000. In contrast, the power of PLR test approached 1 rapidly when $\delta_4$ was 0.30 or 0.45. We conclude that the PLR test is still the most powerful among the four tests in all the considered settings even when the data distribution is multimodal and non-Gaussian.

7 Real Data Analysis

In this section, two real-world applications are provided to compare our PLR test with KS and MMD tests.
7.1 Metagenomic Analysis of Type II Diabetes

The gut microbiota influences numerous biological functions throughout the body. Recent studies have indicated that gut microbiota plays an important role in many human diseases such as obesity and diabetes. The association between disease and gut microbial composition has been reported in many studies (Turnbaugh et al., 2009; Qin et al., 2012). Due to the rapid development of metagenomics, it is possible to directly study the DNA through environmental samples. Compared with traditional culture-based methods, metagenomics can study unculturable microorganisms. Recently, severall metagenomic binning algorithms such as MetaGen (Xing et al., 2017) were proposed to estimate the abundance of microbial species with high accuracy.

As observed in Turnbaugh et al. (2009), the microbial distributions demonstrate large cross-individual difference since there are many environment factors, such as age and antibiotic usage, that could alter the distribution of gut microbiota. A powerful test that can detect such distributional differences would be very useful in metagenomic analysis.

The aim of this study is to detect whether the microbial species have different distributions between case and control groups. For a particular microbial species, let $X_i$ be the log-transformed abundance for the $i$th individual, and let $Z_i = 1/0$ represent the case/control group. We applied the proposed PLR test to a metagenomic data set with 145 sequenced gut microbial DNA samples from 71 T2D patients (case group) and 74 individuals unaffected by T2D (control group) using Illumina Genome Analyzer and obtained 378.4 gigabase paired-end reads. We used MetaGen (Xing et al., 2017) to do the metagenomic binning in which DNA fragments were clustered into species-level bins, and estimated the abundance of 2450 identified species bins. We applied the KS, MMD and PLR tests on 1005 species clusters with abundance larger than 1% of the averaged abundance in more than 50% of the total samples. The 1005 p-values were then calculated by KS, MMD and PLR for each species. We adjusted the p-values by the Benjamini-Hochberg method (Benjamini and Hochberg, 1995). Through controlling the false discovery at 5%, we compared the identified species from the three methods in Figure 9(A). The PLR test identified 101 species, the KS test identified 4 species, and the MMD test identified 13 species. The species identified by PLR cover those by KS or MMD.

Moreover, we highlighted two species that were only identified by the PLR test in Figure 9(B-C). The densities of these two species are both bimodal in the case and control groups. Figure 9(B) plots the conditional density of the log-transformed abundance of *Roseburia intestinalis*. The
majority of the case group has significantly low abundance. In Figure 9(C), the other species, *Faecalibacterium prausnitzii* has lower abundance for a subgroup of patients in the case group. Both species are butyrate-producing bacteria which are able to exert profound immunometabolic effects, and thus are probiotic less abundant in T2D patients. Our finding is consistent with Tilg and Moschen (2014) who also observed that the concentrations of the two species are lower in T2D subjects.

### 7.2 Gene Expression of Chronic Lymphocytic Leukaemia

Chronic lymphocytic leukaemia (CLL), the most common leukaemia among adults in Western countries, is a heterogeneous disease with variable clinical presentation and evolution. Studies have shown that CLL patients with a mutated Immunoglobulin Heavy Chain Variable (IGHV) gene have a much more favorable outcome and a low probability of developing progressive disease, whereas those with the unmutated IGHV gene are much more likely to develop progressive disease and have a shorter survival. The molecular changes leading to the pathogenesis of the disease are still poorly understood. To further investigate the role of the mutation status in IGHV gene,
we aimed to test whether the distributions of the gene expressions are the same between the IGHV mutated and the IGHV unmutated patients.

In this study, we considered a data set of 225 CLL patients in which 131 were IGHV mutated and 85 were IGHV unmutated. The gene expressions were measured by the Affymetrix technique in which proper quality control and normalization methods were performed (Maura et al., 2015). We used the Log2-transformed expression value extracted from the CEL files as the measurement of the expression level. For the \( i \)th subject, let \( X_i \) denote the expression level and \( Z_i \) denote the IGHV mutation status. In particular, \( Z_i = 0 \) denotes the unmutated status and \( Z_i = 1 \) denotes the mutated status. We aimed to test \( H_0: f_{X|Z=0}(x) = f_{X|Z=1}(x) \), i.e. whether the conditional densities of the gene expression level are the same between the two IGHV mutation status. Rejection of \( H_0 \) implies that the gene expression level distribution varies significantly across the mutation status.

We applied the PLR, KS and MMD tests to the 18863 genes. Considering the overall lower p-values in this example, we performed the Bonferroni correction on the p-values, i.e., we rejected \( H_0 \) at a significance level of \( 0.05/18863 = 2.65 \times 10^{-6} \). Such correction was used to reduce the family-wise error rate. The three methods selected 1071 genes, 275 genes and 412 genes respectively. Results are summarized in a Venn diagram (Figure 10(A)) which clearly demonstrates that the genes selected by PLR cover those selected by KS and MMD. There were 272 genes selected by all methods and 412 genes selected by both PLR and MMD. For instance, TGFB2 was missed by KS but discovered by PLR and MMD. In literature, it has been verified by real-time quantitative PCR (Bomben et al., 2007) that TGFB2 is down-regulated in IGHV mutated CLL cases compared with IGHV unmutated cases; see Figure 10(B) for a comparison of the conditional densities from both groups. So PLR and MMD made the correct selection. There were 597 genes, including DTX1, uniquely selected by PLR. DTX1 is a well-established direct target of NOTCH1 which plays a significant role in a variety of developmental processes as well as in pathogenesis of certain human cancers and genetic disorders Yamamoto et al. (2001); Fabbri et al. (2017); see Figure 10(C) for a comparison of the conditional densities. The proposed PLR test correctly selected such a gene.

8 Discussion

We proposed a probabilistic decomposition approach for probability densities based on the penalized likelihood ratio (PLR). As demonstrated in
Figure 10: (A). A Venn diagram showing the numbers of genes selected by PLR, KS and MMD. (B). Densities of gene expression levels from TGFB2 in mutated/unmutated status. (C). Densities of gene expression levels from DTX1 in mutated/unmutated status. Both (B) and (C) demonstrate that the densities of the two expression levels from mutated and unmutated groups are different.

simulation studies, our method performs well under various families of density functions of different modalities. Notably, our test possesses the Wilks’ phenomenon and testing minimaxity. Such results are not easy to derive for distance-based methods. Furthermore, the Wilks’ phenomenon leads to an easy-to-execute testing rule which does not involve resampling.

In many real applications, the underlying densities are complex, usually neither unimodal nor Gaussian. The simulation results demonstrate the superior performance of the PLR test under many different situations. We applied the proposed test to identifying the microbial species with altered distribution in case and control groups. The discovered species with bimodal distributions in both groups were only discovered by the PLT test but omitted by the KS and MMD tests. In another application of comparing the conditional densities of gene expressions under different IGHV mutation status, we made discoveries that are widely supported by existing biological studies.

The proposed test can easily be extended to multidimensional case and k-sample test. In addition, a natural extension is to test the independence or conditional independence between random variables. This can be carried
out through a higher-order probabilistic decomposition of tensor product RKHS. A challenge for such an extension is to characterize the properties of the eigenvalues of the functional space spanned by interactions. We will explore this in a sequel work.

A Appendix: Proofs of the Main Results

This section contains the proofs of the main results Theorems 4.3 3.4, 3.5. Proofs of Lemmas 2.5, 4.1, 4.2, 2.5, 3.1, Proposition 3.3 as well as some auxiliary results, are also included. Proofs of Lemmas A.2-7 are included in supplementary.

- Section A.1 includes the proof of Lemma 2.5.
- Section A.2 includes preliminaries for the minimax lower bound.
- Section A.3 includes the proof of Lemma 4.1.
- Section A.4 includes the proof of Lemma 4.2.
- Section A.5 includes the proof of Theorem 4.3.
- Section A.6 includes the proof of Theorem 3.4.
- Section A.7 includes the proof of Theorem 3.5.
- Section A.8 includes the proof of Lemma 3.2.
- Section A.9 includes the proof of Lemma 3.1.
- Section A.10 includes the proof of Proposition 3.3.

A.1 Proof of Lemma 2.5

Proof. Write \( \eta^*(x, z) = \eta_0^* + \eta_X^*(x) + \eta_Z^*(z) + \eta_{XZ}^*(x, z) \) according to (2.8). If \( X \) and \( Z \) are independent, then \( f^*(x, z) = f_X(x)f_Z(z) \), where \( f_X, f_Z \) are the marginal densities of \( X \) and \( Z \). Take log-transformations on both sides, i.e., \( \eta^*(x, z) = \log(f^*(x, z)) = \log(f_X(x)) + \log(f_Z(z)) \), and hence, \( \eta_{XZ}^* = 0 \). On the other hand, if \( \eta_{XZ}^* = 0 \), then \( f^*(x, z) \propto e^{\eta_X^*(x)}e^{\eta_Z^*(z)} \), and hence, \( X, Z \) are independent. \( \square \)
A.2 Preliminaries for the minimax lower bound

Lemma A.1. Let $P_0$ be the probability measure under the null, and $P_1$ be the probability with density in $\{\eta | \|\eta_{X,G}\|_H < d_n\}$. We have
\[
\inf_{\phi_n} \text{Err}(\phi_n, d_n) \geq 1 - \delta(\sqrt{\delta + 4} - \delta),
\]
where $\delta^2 = \mathbb{E}_{P_0}(dP_1/dP_0 - 1)^2$.

Proof. The test is bounded below by $1 - \|P_0 - P_1\|_{TV}$, where $\|\cdot\|_{TV}$ is the total variation distance between $P_0$ and $P_1$. By the theorem in Ingster (1987), we have
\[
\frac{1}{2}\|P_0 - P_1\|_{TV} \leq \delta (1 - \frac{1}{2}\|P_0 - P_1\|_{TV})^{1/2},
\]
which directly implies the result. \hfill \Box

A.3 Proof of Lemma 4.1

Proof. As show in Lemma A.1, we have
\[
\inf_{\phi_n} \text{Err}(\phi_n, d_n) \geq 1 - \delta(\sqrt{\delta + 4} - \delta), \tag{A.1}
\]
Next we show that if $d_n^2 \leq \frac{\sqrt{k_B(d_n)}}{4n}$, we have that the last term in (A.1) is larger than $1/2$. For simplicity, denote $k = k_B(d_n)$. For any $b = (b_1, \ldots, b_k) \in \{-1, 1\}^k$, let $\theta_b = \frac{d_n}{\sqrt{k}} \sum_{i=1}^k b_i e_i \in \mathbb{R}^N$, where $e_i$ is the standard basis vector with ith coordinate as one. We assume $b$ is uniformly distributed over $\{-1, 1\}^k$ so that $\theta_b$ is uniformly distributed over $Q := \{\theta_b : b \in \{-1, 1\}^k\}$. Since $\mathbb{E}_{P_0} e^{\theta_b \phi_1} - 1 = 0$, we have $e^{\theta_b \phi_1} - 1 \in \mathcal{H}_{11}$. Define
\[
e^{\theta_b \phi_1} - 1 = \frac{d_n}{\sqrt{k}} \sum_{i=1}^k b_i \psi_i \phi_1, \tag{A.2}
\]
where $\{\psi_i \phi_1\}_{i=1}^k$ are basis function for $\mathcal{H}_{11}$. Consider $P_1^{(n)}$ as a mixture distribution defined as
\[
dP_1^{(n)}(y_i) = \mathbb{E}_{\theta_b} \prod_{i=1}^n \{\exp(\eta^{\theta_b}(y_i))\} = 2^{-k} \sum_{b \in \{-1\}^k} \prod_{i=1}^n \exp(\eta^{\theta_b}(y_i)).
\]
Then we have \( \frac{d\mathbb{P}_n^{(n)}}{d\mathbb{P}_0^{(n)}}(y) = \mathbb{E}_{\theta_b} \prod_{i=1}^{n} \exp(\eta_{XG}(y_i)) \). We rewrite \( \delta \) as

\[
\delta^2 = E_{\mathbb{P}_0^{(n)}}[d\mathbb{P}_1^{(n)}/d\mathbb{P}_0^{(n)} - 1]^2
\]

\[
= E_{\mathbb{P}_0^{(n)}}[\mathbb{E}_{\theta_b} \prod_{i=1}^{n} \exp(\eta_{XG}(y_i))]^2 - 1
\]

\[
= E_{\mathbb{P}_0^{(n)}}[\mathbb{E}_{\theta_b} \prod_{i=1}^{n} \exp(\eta_{XG}(y_i))][\mathbb{E}_{\theta_b'} \prod_{i=1}^{n} \exp(\eta_{XG}(y_i))] - 1
\]

\[
= \mathbb{E}_{\theta_b, \theta_b'} \prod_{i=1}^{n} E_{\mathbb{P}_0(y_i)} \exp(\eta_{XG}(y_i)) \exp(\eta_{XG}(y_i)) - 1
\]

\[
= \mathbb{E}_{\theta_b, \theta_b'} [E_{\mathbb{P}_0(y)} \exp(\eta_{XG}(y_i)) \exp(\eta_{XG}(y_i))]^n - 1.
\]

Plugging the above result into A.2, we have

\[
\delta_n + 1 = \mathbb{E}_{\theta_b, \theta'_b} [E_{\mathbb{P}_0(y)} (1 + \frac{d_n}{\sqrt{k}} \sum_{l=1}^{k} b_l \psi_l \phi_1)(1 + \frac{d_n}{\sqrt{k}} \sum_{l=1}^{k} b'_l \psi_l \phi_1)]^n
\]

\[
= \frac{1}{2^k} \sum_{b,b'} (1 + \frac{d_n^2 b^T b'}{k})^n
\]

\[
\leq \frac{1}{2^k} \sum_{b} \exp\left\{\frac{n d_n^2 b^T 1_k}{k}\right\}
\]

\[
= \frac{1}{2^k} \sum_{i=0}^{k} \binom{k}{i} \exp\left\{n(k-2i)d_n^2/k\right\}
\]

\[
= \frac{1}{2^k} (\exp(\frac{n d_n^2}{k}) + \exp(-\frac{n d_n^2}{k}))^k
\]

\[
\leq (1 + \frac{n^2 d_n^4}{k^2})^k
\]

\[
\leq \exp(\frac{n^2 d_n^4}{k}),
\]

where (i) is due to the fact \( \frac{1}{2} (\exp(x) + \exp(-x)) \leq 1 + x^2 \) for \(|x| \leq 1/2\) and (ii) is due to the fact \( 1 + x \leq e^x \). Thus for any \( d_n^4 \leq \frac{k}{\ln^2}, \) we have

\[
\inf_{\phi_n} \text{Err}(\phi_n, d_n) \geq 1 - \delta_n (\sqrt{\delta_n} + 4 - \delta_n) \geq 1 - e^{1/16} (\sqrt{e^{1/16} + 4}) \geq 1/2.
\]
For $d_n \lesssim k^{1/4}/\sqrt{n}$, we have

$$|\exp \eta XG - 1| = \frac{d_n}{\sqrt{k}} \left| \sum_{l=1}^{k} b_l \psi_l \phi_1 \right| \lesssim \frac{k^{3/4}}{\sqrt{n}}.$$ 

Thus, there exists $c_1, c_2 > 0$ such that

$$c_1 |\eta XG(y)| < |\exp \eta XG - 1| < c_2 |\eta XG(y)|,$$

which indicates that $\| \exp \eta XG - 1 \|_2 \lesssim \| \eta XG \|_2$. By the definition of $r_B(\epsilon)$, we have $\text{Err}(\phi_n, d_n) > 1/2$ for all $d_n \leq r_B(\epsilon)$. \hfill \Box

### A.4 Proof of Lemma 4.2

**Proof.** We show that $b_{k,2}(\mathcal{E}_{1,1})$ is bounded below by $\sqrt{\rho_{k+1}}$. It is sufficient to show that $\mathcal{E}_{1,1}$ contains a $l_2$ ball centered at $\eta XG = 0$ with radius $\sqrt{\rho_{k+1}}$. For any $v \in \mathcal{E}_{11}$ with $\| v \|_2 \leq \sqrt{\rho_{k+1}}$, we have

$$b_{2,k} \leq \sum_{i=1}^{k+1} \frac{v_i^2}{\rho_i} \leq \frac{1}{\mu_{k+1}} \sum_{i=1}^{k+1} v_i^2$$

where inequality (i) holds by set the $(k+1)$-dimensional subspace spaned by the eigenvectors corresponding to the first $(k+1)$ largest eigenvalues; inequality (ii) holds by the decreasing order of the eigenvalues, i.e., $\rho_1 \geq \rho_2 \geq \ldots \rho_{k+1}$.

Recall that the definition of the Bernstein lower critical dimension is $k_B(\epsilon) = \text{argmax}_k \{ b_{k-1,2}^2(\mathcal{E}_{1,2}) \geq \epsilon^2 \}$, we have

$$k_B(\epsilon) \geq \text{argmax}_k \{ \sqrt{\rho_k} \geq \epsilon \}$$

\hfill \Box

### A.5 Proof of Theorem 4.3

**Proof.** By Lemma 4.1, we have

$$d_n \leq \sup \{ \epsilon : k_B \geq 16n^2 \epsilon^4 \}.$$
Then we plug in the lower bound of $k_B$ in Lemma 4.2 and we have
\[ d_n \leq \sup \{ \epsilon : \arg \max_k \{ \sqrt{\rho_k} \geq \epsilon \} \geq 16n^2 \epsilon^4 \} \tag{A.4} \]
The eigenvalues have polynomial decay rate i.e., $\rho_p \propto p^{-2m}$, and consequently, $\arg \max_k \{ \sqrt{\rho_k} \propto \epsilon^{-1/m} \}$. Plugging this into (A.4), it is easy to see that the supremum on the right hand side has an order $n^{-2m/4m+1}$. Proof is thus completed. \qed

\section*{A.6 Proof of Theorem 3.4}

The proof of Theorem 3.4 relies on the following Taylor expansion of PLR test. Let $g = \hat{\eta}_0^n - \hat{\eta}_n^{\lambda}$, we have
\[ PLR_{n,\lambda} = \ell_{n,\lambda}(\hat{\eta}_0^n) - \ell_{n,\lambda}(\hat{\eta}_n^{\lambda}) \]
\[ = D\ell_{n,\lambda}(\hat{\eta}_n^{\lambda})g + \int_0^1 \int_0^1 sD^2\ell_{n,\lambda}(\hat{\eta}_n^{\lambda} + ss'g)ggdsds' \]
\[ = \int_0^1 \int_0^1 s\{ D^2\ell_{n,\lambda}(\hat{\eta}_n^{\lambda} + ss'g)gg - D^2\ell_{n,\lambda}(\eta^*)gg \}dsds' + \frac{1}{2} D^2\ell_{n,\lambda}(\eta^*)gg \]
\[ \equiv I_1 + I_2 \tag{A.5} \]
where $\eta^*$ is the underlying truth. We will show that $I_2$ is a leading term compared with $I_1$. From (3.8), we have that $I_2 = \frac{1}{2}\|g\|^2 = \frac{1}{2}\|\hat{\eta}_0^n - \hat{\eta}_n^{\lambda}\|^2$. As we will see, the asymptotic distribution of $\|\hat{\eta}_n^{\lambda} - \hat{\eta}_n^{\lambda}\|^2$ relies on Bahadur representations of $\hat{\eta}_0^n$ and $\hat{\eta}_n^{\lambda}$.

We first prove several lemmas. Define
\[ h^{-1} = \sum_{p=1}^{\infty} \frac{1}{(1+\lambda \rho_p)^2}, \quad h_0^{-1} = \sum_{p=1}^{\infty} \frac{1}{(1+\lambda \rho_0^p)^2}. \]
From Lemma 3.1, we have $\rho_p \propto p^{2m}$ and $\rho_0^p \propto p^{2m}$. The following lemma provides an relation between $h$ (or $h_0$) and $\lambda$.

\begin{lemma}
$h \asymp \lambda^{1/2m}$ and $h_0 \asymp \lambda^{1/2m}$.
\end{lemma}

The following Lemma presents a relationship between the two norms $\| \cdot \|_{\sup}$ and $\| \cdot \|$.

\begin{lemma}
There exists an absolute constant $c_m > 0$ s.t. $\|\eta\|_{\sup} \leq c_m h^{-1/2}\|\eta\|$.
\end{lemma}
Proofs of Lemmas A.1 and A.2 can be executed similar to Shang et al. (2013).

The following two lemmas characterize the convergence rates of $\hat{\eta}_{n,\lambda}$ and $\hat{\eta}_{0,\lambda}$ under $H_0$.

**Lemma A.3.** Assume $\lambda \to 0$ and $H_0$. Then $\|\hat{\eta}_{n,\lambda} - \eta^*\|_0 = O_P((nh_0)^{-1/2} + \lambda^{1/2})$ and $\|\hat{\eta}_{n,\lambda} - \eta^*\| = O_P((nh)^{-1/2} + \lambda^{1/2})$.

Lemma A.3 can be proved based on a quadratic approximation method proposed by Gu (2013), i.e., apply (Gu, 2013, Section 9.2.2) to both $(\hat{\eta}_{n,\lambda}, H)$ and $(\hat{\eta}_{0,\lambda}, H_0)$. The optimal rates for both estimators achieve at $h \asymp n^{-1/(2m+1)}$, $h_0 \asymp n^{-1/(2m+1)}$. Notice that $\| \cdot \|$ and $\| \cdot \|_0$ are equivalent under the null hypothesis for any $\eta \in H_0$. Thus, in what follows, we will not distinguish the two norms for notation convenience. We also do not distinguish $h$ and $h_0$ since they have the same order for achieving optimality.

Based on an empirical processes technique by Shang et al. (2013), one can further prove the following Bahadur representations for the two MLEs which will be crucial for proving Theorem 3.4.

**Lemma A.4.** Suppose $nh^2 \to \infty$. Then we have
\[
\|\tilde{\eta}_{n,\lambda} - \eta^* - S_{n,\lambda}(\eta^*)\| = O_p((nh)^{-3/2} + h^{2m}),
\]
\[
\|\tilde{\eta}_{0,\lambda} - \eta^* - S^0_{n,\lambda}(\eta^*)\| = O_p((nh)^{-3/2} + h^{2m}),
\]
where $S_{n,\lambda}(\eta^*)$ and $S^0_{n,\lambda}(\eta^*)$ are defined in (A.6) and (A.7).

The proof of Theorem 3.4 is sketched as follows. By Lemma A.4, $n^{1/2}\|\hat{\eta}_{n,\lambda} - \hat{\eta}_{n,\lambda} - S_{n,\lambda}(\eta^*) + S_{n,\lambda}(\eta^*)\| = o_P(1)$. So we have the following
\[
n^{1/2}\|\tilde{\eta}_{n,\lambda} - S^0_{n,\lambda}(\eta^*) - S_{n,\lambda}(\eta^*)\| = o_P(1).
\]
Thus we only focus on $n^{1/2}\|S^0_{n,\lambda}(\eta^*) - S_{n,\lambda}(\eta^*)\|$. Moreover, the following expressions of $S^0_{n,\lambda}(\eta^*)$ and $S_{n,\lambda}(\eta^*)$ are reserved for future use:
\[
S_{n,\lambda}(\eta^*) = -\frac{1}{n} \sum_{i=1}^n K_{Y_i} + E_{\eta^*} K_Y + W_\lambda \eta^*, \quad (A.6)
\]
\[
S^0_{n,\lambda}(\eta^*) = -\frac{1}{n} \sum_{i=1}^n K^0_{Y_i} + E_{\eta^*} K^0_Y + W^0_\lambda \eta^*. \quad (A.7)
\]
Proof of Theorem 3.4. Let us first analyze $I_1$. Let $\tilde{g} = \tilde{\eta}_{n,\lambda} + ss'g - \eta^*$, for any $0 \leq s, s' \leq 1$. By Lemma A.3, we have $\|g\| = O_P((nh)^{-1/2} + h^m) = o_P(1)$. Notice that

$$D^2 \ell_{n,\lambda}(\tilde{\eta}_{n,\lambda} + ss'g)gg = D^2 \ell_{n,\lambda}(\tilde{g} + \eta^*)gg = \int_{\mathcal{Y}} g^2(y)e^{\tilde{g}(y) + \eta^*(y)}dy + \lambda J(g, g),$$

(A.8)

and

$$D^2 \ell_{n,\lambda}(\eta^*)gg = \int_{\mathcal{Y}} g^2(y)e^{\eta^*(y)}dy + \lambda J(g, g).$$

(A.9)

Combining (A.8) and (A.9), we have

$$|D^2 \ell_{n,\lambda}(\tilde{\eta}_{n,\lambda} + ss'g)gg - D^2 \ell_{n,\lambda}(\eta^*)gg| \leq \int_{\mathcal{Y}} g^2(y)e^{\eta^*(y)}|e^{\tilde{g}(y)} - 1|dy.$$

By Taylor expansion of $e^{\tilde{g}(y) + \eta^*(y)}$ at $\eta^*(y)$ for any $y \in \mathcal{Y}$, it trivially holds that $e^{\eta^*(y)}|e^{\tilde{g}(y)} - 1| = e^{\eta^*(y)}O(\|\tilde{g}(y)\|)$ (Lemma A.2), and $h^{-1/2}((nh)^{-1} + \lambda)^{1/2} = o(1)$, we have

$$|I_1| = O_P(h^{-1/2}((\|\tilde{\eta}_{n,\lambda} - \eta^*\| + \|g\|) \cdot \|g\|^2) = o_P(\|g\|^2).$$

(A.10)

Let us then analyze $I_2$. From (3.8), we have $D^2 \ell_{n,\lambda}(\eta^*)gg = \|g\|^2 = \|\tilde{\eta}_{n,\lambda} - \tilde{\eta}_{n,\lambda}^0\|^2$, which dominates $I_1$, since $h^{-1/2}((\|\tilde{\eta}_{n,\lambda} - \eta^*\| + \|g\|) = o_P(1)$.

Next let us analyze $\|\tilde{\eta}_{n,\lambda} - \tilde{\eta}_{n,\lambda}^0\|^2$. By Lemma A.4, we have

$$n^{1/2}\|\tilde{\eta}_{n,\lambda} - \tilde{\eta}_{n,\lambda}^0 - S_{n,\lambda}^0(\eta^*) + S_{n,\lambda}(\eta^*)\| = O_P(n^{1/2}h^{-3/2}((nh)^{-1} + h^m)) = o_P(1).$$

Thus we only need to focus on $n^{1/2}\|S_{n,\lambda}^0(\eta^*) - S_{n,\lambda}(\eta^*)\|$. Recall $S_{n,\lambda}(\tilde{\eta}_{n,\lambda}) = 0$ and $S_{n,\lambda}(\eta^*)$, $S_{n,\lambda}^0(\eta^*)$ have expressions (A.6), (A.7). Define $K_{\mathcal{Y}} = K_{\mathcal{Y}} - K_0$ and $W_{\lambda}^1 = W_{\lambda} - W_0$, then $S_{n,\lambda}(\eta^*) - S_{n,\lambda}(\eta^*) = -\frac{1}{n}\sum_{i=1}^n K_{\mathcal{Y}}^1 + EK_{\mathcal{Y}} + W_{\lambda}^1\eta^*$.

By Proposition 3.3, $K_{\mathcal{Y}}^1$ can be expressed as a series of $\xi_{p,\lambda}^1(Y)$. Since $\xi_{p,\lambda}^1 \in \mathcal{H}_1$ and $\phi_0 = 1 \in \mathcal{H}_0$, we have

$$E_\eta^*\{\xi_{p,\lambda}^1(Y)\} = E_\eta^*\{\xi_{p,\lambda}^1(Y)\phi_0(X)\} = V(\xi_{p,\lambda}^1, \phi_0) = 0.$$

And so $E_\eta^*\{K_{\mathcal{Y}}^1\} = 0$. Therefore, $S_{n,\lambda}(\eta^*) - S_{n,\lambda}(\eta^*) = -\frac{1}{n}\sum_{i=1}^n K_{\mathcal{Y}}^1_{\lambda} +$
\( W_1^1 \eta^* \). Then

\[
n\| S_{n, \lambda}^0(\eta^*) - S_{n, \lambda}(\eta^*) \|^2 = n^{-1} \sum_{i=1}^{n} K_{\lambda}^1_{\eta} \|^2 - 2 \sum_{i=1}^{n} \langle K_{\lambda}^1_{\eta}, W_1^1 \eta^* \rangle + n\| W_1^1 \eta^* \|^2
\]

\[\equiv W_1 - 2W_2 + W_3.\]

Since \( \eta^* \in \mathcal{H}_0 \), it follows by Lemma 3.1 that \( \eta^* \) is expanded by a series of \( \xi^0 \). By Proposition 3.3, \( W_{\lambda} \xi^0_p \propto \xi^0_p \) which implies \( W_{\lambda} \eta^* = W_{\lambda}^0 \eta^* \). And hence, \( W_{\lambda}^1 \eta^* = W_{\lambda} \eta^* - W_{\lambda}^0 \eta^* \) which yields that \( W_2 = W_3 = 0 \). Write

\[
W_1 = n^{-1} \sum_{i=1}^{n} K_{\lambda}^1_{\eta} \|^2 = n^{-1} \sum_{i=1}^{n} \| K_{\lambda}^1_{\eta} \|^2 + n^{-1} W(n), \quad \text{where} \quad W(n) = \sum_{i \neq j} K^1(Y_i, Y_j).
\]

Next let us consider the term \( \sum_{i=1}^{n} K^1(Y_i, Y_i) \). Let \( E \) denote \( E_{\eta^*} \) unless otherwise indicated. Let \( \theta(n) = E\{ K^1(Y_i, Y_i) \} \). By Lemma A.2 we have

\[
E\{ \sum_{i=1}^{n} \{ K^1(Y_i, Y_i) - \theta(n) \}^2 \} \leq n E\{ K^1(Y_i, Y_i)^2 \} = O(nh^{-2}), \quad \text{so}
\]

\[
\sum_{i=1}^{n} [K(Y_i, Y_i) - \theta(n)] = O_p(n^{1/2}h^{-1}). \quad \text{(A.11)}
\]

Next, we derive the asymptotic distribution of \( W(n) \). Define \( W_{ij} = 2K^1(Y_i, Y_j) \), then \( W(n) = \sum_{1 \leq i < j \leq n} W_{ij} \). Let \( \sigma(n)^2 = \text{Var}(W(n)) \) and

\[
G_I = \sum_{i<j} E\{ W_{ij}^4 \},
\]

\[
G_{II} = \sum_{i<j<k<l} (E\{ W_{ij}^2 W_{ik}^2 \} + E\{ W_{ji}^2 W_{jk}^2 \} + E\{ W_{kl}^2 W_{kj}^2 \}), \quad \text{and}
\]

\[
G_{IV} = \sum_{i<j<k<l} (E\{ W_{ij} W_{ik} W_{il} W_{kl} \} + E\{ W_{ij} W_{il} W_{kj} W_{kl} \} + E\{ W_{ik} W_{il} W_{jk} W_{jl} \}).
\]

By \( E\{ K^1_{Y_i} \} = 0 \) and direct examinations we have

\[
\sigma^2(n) = \text{Var}(W(n)) = \sum_{1 \leq i < j \leq n} \mathbb{E}\{ (K^1(Y_i, Y_j) - \mathbb{E}[K^1(Y_i, Y_j)])^2 \}
\]

\[= \sum_{1 \leq i < j \leq n} \mathbb{E}\{ K^1(Y_i, Y_j)^2 \} \asymp n^2 h^{-1}.
\]

Since \( \mathbb{E}\{ W_{ij}^4 \} = 16E\{ K^1(Y_i, Y_j)^4 \} = O(h^{-4}) \), we have \( G_I = O(n^2 h^{-4}) \). Obviously, \( \mathbb{E}\{ W_{ij}^2 W_{lk}^2 \} \leq \mathbb{E}\{ W_{ij}^4 \} = O(h^{-4}) \), implying \( G_{II} = O(n^3 h^{-4}) \).
For pairwise different $i, j, k, l$, we have

$$
\mathbb{E}\{W_{ij}W_{ik}W_{lj}W_{lk}\} = 16 \mathbb{E}\{K^1(Y_i, Y_j)K^1(Y_i, Y_k)K^1(Y_l, Y_j)K^1(Y_l, Y_k)\}
$$

$$
= \sum_{p=1}^{\infty} \frac{1}{(1 + \lambda \rho_p^2)^4} = O(h^{-1}),
$$

which leads to $G_{IV} = O(n^4 h^{-1})$.

It follows by $h = o(1)$ and $(nh^2)^{-1} = o(1)$ that $G_I$, $G_{II}$ and $G_{IV}$ are of lower order than $\sigma(n)^4$. By Proposition 3.2 of de Jong (1987) we get that

$$
\frac{W(n)}{\sigma(n)} \xrightarrow{d} N(0, 1).
$$

(A.12)

From (A.11) and (A.12), we get $\frac{1}{n} \sum_{i=1}^{n} K^1(Y_i, Y_i)^2 = \theta(n) + o_P(1)$, which implies $n\|S_{n,\lambda}^0(\eta^*) - S_{n,\lambda}(\eta^*)\|^2 = O_P(h^{-1} + n\lambda + h^{-1/2}) = O_P(h^{-1})$, and hence $n^{1/2}\|S_{n,\lambda}^0(\eta^*) - S_{n,\lambda}(\eta^*)\| = O_P(h^{-1/2})$. Thus,

$$
2n \cdot PLR_{n,\lambda} = n\|\hat{\eta}_{n,\lambda} - \eta^*\|^2 + o_P(h^{-1/2})
$$

$$
= \left( n^{1/2}\|S_{n,\lambda}^0(\eta^*) - S_{n,\lambda}(\eta^*)\| + o_P(1) \right)^2 + o_P(h^{-1/2})
$$

$$
= n\|S_{n,\lambda}^0(\eta^*) - S_{n,\lambda}(\eta^*)\|^2 + 2n^{1/2}\|S_{n,\lambda}^0(\eta^*) - S_{n,\lambda}(\eta^*)\| \cdot o_P(1) + o_P(h^{-1/2})
$$

$$
= n^{-1}\|\sum_{i=1}^{n} K^1_{Y_i}\|^2 + o_P(h^{-1/2}).
$$

(A.13)

By (A.12), (A.13) and Slutsky’s theorem, $\frac{2n \cdot PLR_{n,\lambda} - \theta(n)}{\sigma(n)/n} \xrightarrow{d} N(0, 1)$. Since $\theta_\lambda = \sum_{p=1}^{\infty} \frac{1}{1 + \lambda \rho_p^2}$, $\sigma_\lambda^2 = \sum_{p=1}^{\infty} \frac{1}{(1 + \lambda \rho_p^2)^2}$, we have $\theta(n) = \theta_\lambda$ and $\frac{\sigma(n)}{n} = \sqrt{\frac{\sigma_\lambda^2}{2}} \mathbb{E}(W_{ij}^2)/n = \sqrt{2}\sigma_\lambda$.

A.7 Proof of Theorem 3.5

Before proving Theorem 3.5, we provide some preliminary lemmas. For $\eta^* \in \mathcal{H}$, consider decomposition $\eta^* = \eta_0^* + \eta^*_{XZ}$ where $\eta_0^*$ is the projection of $\eta^*$ on $\mathcal{H}_0$. The following lemma says that, for general $\eta^* \in \mathcal{H}$, the restricted MLE $\hat{\eta}_{n,\lambda}^0$ converges to $\eta_0^*$ with rate of convergence provided.

Lemma A.5. Suppose that Assumption 1 is satisfied. We have $\|\hat{\eta}_{n,\lambda}^0 - \eta_0^*\|_0 = O_P((nh)^{-1/2} + \lambda^{1/2})$. 

38
Parallel to Lemma A.4, when \( \eta^* \in \mathcal{H} \), we have the following result characterizing the higher order expansion of \( \hat{\eta}^0_{n,\lambda} \).

**Lemma A.6.** Suppose that \( nh^2 \to \infty \). We have

\[
\| \hat{\eta}^0_{n,\lambda} - \eta^*_{0} - S^0_{n,\lambda}(\eta^*_0) \|_0 = O_P(h^{-3/2}((nh)^{-1} + h^{2m})).
\]

**Proof of Theorem 3.5.** Let \( g = \hat{\eta}^0_{n,\lambda} - \hat{\eta}^0_{n,\lambda} \). Recall the Taylor expansion (A.5):

\[
PLR_{n,\lambda} = \ell_{n,\lambda}(\hat{\eta}^0_{n,\lambda}) - \ell_{n,\lambda}(\hat{\eta}_{n,\lambda})
\]

\[
= \int_0^1 \int_0^1 s\{D^2 f(\hat{\eta}_{n,\lambda} + ss'g)gg - D^2 f(\eta^*)gg\}dsds' + \frac{1}{2}D^2 f(\eta^*)gg
\]

\[
= O_P(\| \hat{\eta}_{n,\lambda} - \eta^* \|_{\text{sup}} + \| g \|_{\text{sup}} \cdot \| g \|^2) + \frac{1}{2} \| g \|^2,
\]

where the \( O_P \) term in the last equation follows from (A.10). By Lemmas A.2 and A.3, \( \| \hat{\eta}_{n,\lambda} - \eta^* \|_{\text{sup}} = o_P(1) \). By assumption \( \| \eta^*_X \|_{\text{sup}} \leq (\log n)^{-1} = o(1) \) and Lemma A.5, we have \( \| g \|_{\text{sup}} = \| \hat{\eta}^0_{n,\lambda} - \eta^*_0 + \eta^* - \hat{\eta}_{n,\lambda} - \eta^*_X \|_{\text{sup}} = o_P(1) \). Hence, the \( O_P \) term in (A.14) is dominated by \( \frac{1}{2} \| g \|^2 \), for which we only focus on the latter. Combining the results of Lemmas A.4 and A.6, we have

\[
\| \hat{\eta}_{n,\lambda} - \eta^* - S_{n,\lambda}(\eta^*) \| = O_P(h^{-2}((nh)^{-1} + h^{2m})),
\]

\[
\| \hat{\eta}^0_{n,\lambda} - \eta^*_0 - S^0_{n,\lambda}(\eta^*_0) \|_0 = O_P(h^{-2}((nh)^{-1} + h^{2m})).
\]

Recalling \( \eta^* - \eta^*_0 = \eta^*_X \), we have \( \| g \| = \| \eta^*_X + S_{n,\lambda}(\eta^*) - S^0_{n,\lambda}(\eta^*_0) \| + O_P(h^{-2}((nh)^{-1} + h^{2m})) \). In what follows, we focus on \( \| \eta^*_X + S_{n,\lambda}(\eta^*) - S^0_{n,\lambda}(\eta^*_0) \| \). By definition of \( S_{n,\lambda}(\eta^*) \), \( S^0_{n,\lambda}(\eta^*_0) \) (see (3.7)) and direct calculations, it can be shown that

\[
\| \eta^*_X + S_{n,\lambda}(\eta^*) - S^0_{n,\lambda}(\eta^*_0) \|^2
\]

\[
= \frac{1}{n} \sum_{i=1}^n K^1 \eta^*_X(Y_i)^2 + \| \eta^*_X \|^2 + \| EK^0_Y - EK^0 \|^2 + \| W^1 \eta^*_X \|^2
\]

\[
- \frac{2}{n} \sum_{i=1}^n \eta^*_X(Y_i) + 2E \eta^*_X(Y) + 2\langle W^1 \eta^*_X, \eta^*_X \rangle - \frac{2}{n} \sum_{i=1}^n EK^1(Y_i, Y)
\]

\[
- \frac{2}{n} (W^1 \eta^*_X)(Y_i) + 2E(W^1 \eta^*_X)(Y).
\]
where $E$ denotes $E_{\eta}$. Since $E\{K_{Y} - K_{\bar{Y}}^{0}\} = EK_{Y}^{1} = 0$, we have

$$
\|\eta_{XZ}^{*} + S_{n,\lambda}(\eta_{\bar{X}}^{*}) - S_{n,\lambda}(\eta_{0}^{*})\|^2 \\
\geq \|\frac{1}{n} \sum_{i=1}^{n} K_{Y_{i}}^{1}\|^2 + \|\eta_{XZ}^{*}\|^2 + \left[ -\frac{2}{n} \sum_{i=1}^{n} \eta_{XZ}^{*}(Y_{i}) + 2E_{\eta_{\bar{X}0}}(\eta_{XZ}^{*}(Y)) \right] + 2W_{\lambda}^{1}\eta_{XZ}^{*}, \eta_{XZ}^{*} \\
+ \left[ -\frac{2}{n} (W_{\lambda}^{1}\eta_{XZ}(Y_{i}) + 2E_{\eta_{\bar{X}}}W_{\lambda}(\eta_{XZ})(Y)) \right] = V_{1} + V_{2} + V_{3} + V_{4} + V_{5}.
$$

Since $\text{Var}(V_{3}) \leq \frac{2}{n} E(\eta_{XZ}^{*}(Y))^{2} \leq \frac{2}{n} \|\eta_{XZ}^{*}\|^{2}$,

$$
V_{3} = O_{P}(n^{-1/2}) \|\eta_{XZ}^{*}\|. \quad \text{(A.14)}
$$

By assumption $J(\eta_{XZ}^{*}, \eta_{XZ}^{*}) \leq C$, we have

$$
V_{4} = \lambda J(\eta_{XZ}^{*}, \eta_{XZ}^{*}) \leq C\lambda. \quad \text{(A.15)}
$$

Since $\text{Var}(V_{5}) \leq E(W_{\lambda}\eta_{XZ}^{*})^{2} = V(W_{\lambda}\eta_{XZ}^{*}, W_{\lambda}\eta_{XZ}^{*})$. By Proposition 3.3, we have

$$
V(W_{\lambda}\eta_{XZ}^{*}, W_{\lambda}\eta_{XZ}^{*}) = \sum_{p=1}^{\infty} |V(\eta_{XZ}^{*}, \xi_{p})|^{2} \left( \frac{\lambda \rho_{p}}{1 + \lambda \rho_{p}} \right)^{2} = o(\lambda),
$$

where the last equality follows by $\sum_{p=1}^{\infty} |V(\eta_{XZ}^{*}, \xi_{p})|^{2} \rho_{p} < \infty$ and the dominated convergence theorem. Thus we have

$$
V_{5} = o_{p}(n^{-1/2} \lambda^{1/2}) \quad \text{(A.16)}
$$

Combining (A.14), (A.15) and (A.16) we have

$$
\frac{2n \cdot PLR_{n,\lambda} - \theta(n)}{\sigma(n)} \\
\geq \frac{2n \cdot V_{1} - \theta(n)}{\sigma(n)} + \frac{2n \cdot (V_{2} + V_{3} + V_{4} + V_{5})}{\sigma(n)} \\
\geq O_{P}(1) + 2n \sigma^{-1}(n)(\|\eta_{XZ}^{*}\|^{2} + O_{P}(n^{-1/2}\|\eta_{XZ}^{*}\|) + O(\lambda) + o_{p}(n^{-1/2}\lambda^{1/2})).
$$

For $C_{\varepsilon} > 0$ sufficiently large, let $\eta_{XZ}^{*}$ satisfy $\|\eta_{XZ}^{*}\|^{2} \geq C_{\varepsilon} n^{-1/2}\|\eta_{XZ}^{*}\|$, $\|\eta_{XZ}^{*}\|^{2} \geq C_{\varepsilon}\lambda$, $n\varepsilon^{1/2}\|\eta_{XZ}^{*}\|^{2} \geq C_{\varepsilon}$, $n\|\eta_{XZ}^{*}\|^{2} / \sigma(n) \geq C_{\varepsilon}$, which implies that with probability greater than $1 - \varepsilon$, $\frac{2n \cdot PLR_{n,\lambda} - \theta(n)}{\sigma(n)} \geq c_{\alpha}$ (i.e., $\Phi_{n,\lambda}(\alpha) = 1$), where $c_{\alpha}$ is the $1 - \alpha$ percentile of standard normal distribution. It can be seen that the above conditions on $\eta_{XZ}^{*}$ are satisfied if $\|\eta_{XZ}^{*}\|^{2} \geq C_{\varepsilon}(\lambda + (nh^{1/2})^{-1})$. The result follows immediately by the fact $\|\eta_{XZ}^{*}\|_{2} \leq \|\eta_{XZ}^{*}\|$. Proof is completed. \qed

40
A.8 Proof of Lemma 3.2

Proof. Following Gu (2013), $J(\cdot)$ is the roughness penalty, hence it is standard in the sense of Lin (2000). Following Lin (2000), the norm based on $\int_Y \eta(x,z)^2 dx dz + J(\eta)$ is equivalent to $\|\cdot\|_{\mathcal{H}(X) \otimes \mathcal{H}(Z)}$, where $\|\cdot\|_{\mathcal{H}(X) \otimes \mathcal{H}(Z)}$ is the tensor product norm induced by the Sobolev norm $V_X(g_1,g_2) + J_X(g_1,g_2)$ on $\mathcal{H}(X)$ and the Euclidean norm on $\mathcal{H}(Z)$. Since $f^*(x,z)$ is bounded away from zero and infinity, there exist constants $0 < c_1 \leq c_2 < \infty$ such that, for any $\eta \in \mathcal{H}$,

$$c_1 \int_Y \eta(x,z)^2 dx dz \leq V(\eta,\eta) \leq c_2 \int_Y \eta(x,z)^2 dx dz. \quad (A.17)$$

Therefore, $\|\cdot\|$ and $\|\cdot\|_{\mathcal{H}(X) \otimes \mathcal{H}(Z)}$ are equivalent norms. Since $\mathcal{H}$ endowed with $\|\cdot\|_{\mathcal{H}(X) \otimes \mathcal{H}(Z)}$ is an RKHS, $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ is an RKHS. Since $\mathcal{H}_0$ is a closed subset of $\mathcal{H}$, and $\langle \cdot, \cdot \rangle_0$ is inherited from $\langle \cdot, \cdot \rangle$, we have that $(\mathcal{H}_0, \langle \cdot, \cdot \rangle_0)$ is also an RKHS.

A.9 Proof of Lemma 3.1

Proof. We first construct the eigensystems on the marginal domain $\mathcal{H}(X)$ and $\mathcal{H}(Z)$, based on which the eigensystem on $\mathcal{H}$ will be constructed. Recall the Sobolev norm $V_X(g_1,g_2) + J_X(g_1,g_2)$ on $\mathcal{H}(X)$. Following Shang et al. (2013), we choose the eigenvalues and eigenfunctions of $\mathcal{H}(X)$ as the solution to the following ordinary differential equations: for $k = 0, 1, \ldots$,

$$(-1)^m \phi_k^{(2m)}(\cdot) = \mu_k f_X(\cdot) \phi_k(\cdot),$$

$$\phi_k^{(j)}(0) = \phi_k^{(j)}(1) = 0, \text{ for } j = m, \ldots, 2m - 1 \quad (A.18)$$

where $f_X$ is the marginal density of $X$ and $0 = \mu_0 = \mu_1 = \cdots = \mu_{m-1} \leq \mu_m \leq \mu_{m+1} \leq \cdots$ are ordered eigenvalues. The solution to (A.18) satisfies $V_X(\phi_k, \phi_{k'}) = \delta_{kk'}$ and $J_X(\phi_k, \phi_{k'}) = \mu_k \delta_{kk'}$. The existence of such solution is guaranteed by Proposition 2.2 in Shang et al. (2013). Furthermore, one can choose $\phi_0 \equiv 1$. To see this, note that $\phi_1, \ldots, \phi_m$ are basis of the space of $m - 1$ order polynomials on $[0,1]$. Let $\phi_k(x) = \sum_{t=0}^{m-1} a_t,k x^t$ and $M_{tt'} = \int_0^1 a_t,t' f_X(x) dx$ for $k = 0, \ldots, m - 1$ and $0 \leq t, t' \leq m - 1$. Let $A_k = (a_{0,k}, \ldots, a_{m-1,k})^T$ and $M = [M_{tt'}]_{t,t'=0}^{m-1}$. Since $V_X(\phi_k, \phi_{k'}) = \delta_{kk'}$ for $k, k' =
0, \ldots, m - 1, we have $A_k^T MA_{k'} = \delta_{kk'}$. Purposely choose $A_0 = (1, 0, \ldots, 0)^T$ and treat the rest $A_1, \ldots, A_{m-1}$ as unknowns to be determined. This leaves us $m^2 - m$ unknown coefficients and $\frac{m^2 + m}{2} - 1$ equations. Since $m^2 - m \geq \frac{m^2 + m}{2} - 1$ for any positive integer $m$, there always exist $A_k$’s for $k = 1, \ldots, m - 1$ that satisfy $A_k^T MA_{k'} = \delta_{kk'}$. This shows that we can choose $\phi_0 \equiv 1$ while maintaining the simultaneous diagonalization.

The space $\mathcal{H}^{(Z)}$ is an $a$-dimensional Euclidean space endowed with Euclidean norm. Let $\{\phi_l\}_{l=0}^{a-1}$ denote the orthonormal eigenvectors. The corresponding eigenvalues are $\nu_0 = \cdots = \nu_{a-1} = 1$. To see this, note that the reproducing kernel is $R(z, z') = 1(z = z')$, hence, $\langle R_z, \phi_l \rangle_Z = \psi_l(z)$. On the other hand, $R_z(z, z') = \sum_{l=0}^{a-1} \nu_l \psi_l(z) \psi_l(z')$, hence, $\langle R_z, \phi_l \rangle_Z = \psi_l(z) \nu_l$, leading to $\nu_l = 1$. For convenience, we choose $\psi_0$ as constant function, i.e., $\psi_0(z) \equiv 1/\sqrt{a}$ for $z = 1, \ldots, a$.

Let $\| \cdot \|_{\mathcal{H}^{(X)} \otimes \mathcal{H}^{(Z)}}$ denote the tensor product norm induced by $V_X(g_1, g_2) + J_X(g_1, g_2)$ on $\mathcal{H}^{(X)}$ and the Euclidean norm on $\mathcal{H}^{(Z)}$. The marginal basis for $\mathcal{H}^{(X)}$ and $\mathcal{H}^{(Z)}$ naturally provide a basis for the tensor space, i.e., $\{\phi_k \phi_l : k \geq 0, 0 \leq l \leq a - 1\}$, that satisfy

$$
\langle \phi_k \phi_l, \phi_{k'} \phi_{l'} \rangle_{\mathcal{H}^{(X)} \otimes \mathcal{H}^{(Z)}} = (1 + \mu_k \nu_l) \delta_{kk'} \delta_{ll'}.
$$

The right hand side $\mu_k \nu_l$ of (A.19) is the eigenvalue correspondence to basis $\phi_k \phi_l$. Indeed, they form the eigenvalues of the Rayleigh quotient $\| \cdot \|_{\mathcal{H}^{(X)} \otimes \mathcal{H}^{(Z)}}^2 / \| \cdot \|_{\mathcal{H}^{(X)} \otimes \mathcal{H}^{(Z)}}^2$ since $\phi_k$ and $\psi_l$ are eigenvalues of the marginal Rayleigh quotients; see (Lin, 2000, Section 2.3). Represent these eigenvalues in an increasing order: $\pi_1 \leq \pi_2 \leq \cdots$, i.e., $\pi_{r_0+s} = \mu_r$ for $r \geq 0$ and $1 \leq s \leq a$.

Consider orthogonal decomposition $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$ in (3.3). By Weinberger (1974), we can use the Rayleigh quotient $V/(V + J)$ to produce $\xi_0 \in \mathcal{H}_0$ and $\xi_p \in \mathcal{H}_1$ with corresponding eigenvalues $\rho_p^0$ and $\rho_p^1$ that satisfy:

$V(\xi_0, \xi_0) = \delta_{pp'}, J(\xi_0, \xi_0) = \rho_p^0 \delta_{pp'},$ for $j = 0, \perp$. Let $\{\xi_p\}_{p=1}^{\infty} = \{\xi_p^0, \xi_p^1\}_{p=1}^{\infty}$ and $\{\rho_p\}_{p=1}^{\infty} = \{\rho_p^0, \rho_p^1\}_{p=1}^{\infty}$, where $\rho_p$ are arranged in an increasing order. It is easy to verify that $\xi_p$’s are Rayleigh quotient eigenvalues of $V/(V + J)$ over $\mathcal{H}$ as defined in (Weinberger, 1974, Section 2). We also have

$V(\xi_p, \xi_p) = \delta_{pp'}, \quad J(\xi_p, \xi_p) = \rho_p \delta_{pp'}$. 

42
By (A.17), the Rayleigh quotients corresponding to \( \| \cdot \|_{L^2(X) \otimes L^2(Z)} \) and \( (V, V + J) \) are equivalent. By Mapping theorem ([Weinberger, 1974, Section 3.3]), there exist constants \( c_1, c_2 > 0 \) s.t.

\[
\frac{c_1}{1 + \pi_p} \leq \frac{1}{1 + \rho_p} \leq \frac{c_2}{1 + \pi_p}, \quad p \geq 1.
\]

(A.20)

Following (A.20) we have \( \rho_p \asymp \pi_p \asymp p^{2m} \). By Fourier expansion, we have \( \eta = \sum_{p=1}^{\infty} V(\eta, \xi_p) \xi_p \).

When restricted on \( \mathcal{H}_0 \), the Rayleigh quotients corresponding to \( (V, V + J) \) and \( \| \cdot \|_{L^2(X) \otimes L^2(Z)}, \| \cdot \|_{\mathcal{H}(X) \otimes \mathcal{H}(Z)} \) are still equivalent. Similar to (A.20), by Mapping theorem,

\[
\frac{c_1}{1 + \pi_p} \leq \frac{1}{1 + \rho_p} \leq \frac{c_2}{1 + \pi_p}, \quad p \geq 1.
\]

(A.21)

where \( \{\pi^0_p\}_{p=1}^{\infty} = \{\mu_k, \nu_l : l = 0, \ldots, a - 1, k \geq 0\} \) are eigenvalues (with increasing order) corresponding to \( \{\phi_k, \psi_l : l = 0, \ldots, a - 1, k \geq 0\} \). Specifically, \( \pi^0_p = \pi_p \) for \( p = 1, \ldots, a \), and \( \pi^0_{a+s} = \pi_{sa+1} \) for \( s \geq 1 \). Now remove \( \{\pi^0_p\}_{p \geq 1} \) from \( \{\pi_p\}_{p \geq 1} \) and denote the rest as \( \{\pi^\perp_p\}_{p \geq 1} \). From (A.20) and (A.21), we have

\[
\frac{c_1}{1 + \pi^\perp_p} \leq \frac{1}{1 + \rho^\perp_p} \leq \frac{c_2}{1 + \pi^\perp_p}, \quad p \geq 1.
\]

Since \( \nu_1 = \cdots = \nu_{a-1} = 1 \) which leads to \( \pi^\perp_{r(a-1)+s} = \mu_{r+1} \) for \( r \geq 0 \) and \( s = 1, \ldots, a-1 \), we have \( \rho^\perp_p \asymp \pi^\perp_p \asymp \mu_{\lfloor p/(a-1) \rfloor} \asymp p^{2m} \).

\[\Box\]

A.10 Proof of Proposition 3.3

Proof. The proof of \( \| \eta \|^2 = \sum_{p=1}^{\infty} |V(\eta, \xi_p)|^2 (1 + \lambda \rho_p) \) follows by (3.3) and the Fourier expansion of \( \eta: \eta = \sum_{p=1}^{\infty} V(\eta, \xi_p) \xi_p \). For any \( p' \geq 1 \),

\[
\langle \eta, \xi_{p'} \rangle = \left( \sum_{p=1}^{\infty} V(\eta, \xi_p) \xi_{p'} \right) = V(\eta, \xi_{p'})(1 + \lambda \rho_{p'}).
\]

(A.22)

By (A.22), \( V(K_y, \xi_p) = \frac{(K_y \xi_p)}{1 + \lambda \rho_p} = \xi_p(y) \). Hence \( K_y(\cdot) = \sum_{p=1}^{\infty} \frac{\xi_p(y)}{1 + \lambda \rho_p} \xi_p(\cdot) \) follows. Meanwhile, (A.22) implies that \( V(W_\lambda \xi_p, \xi_{p'}) = \frac{(W_\lambda \xi_p \xi_{p'})}{1 + \lambda \rho_{p'}} = \frac{\lambda \rho_p}{1 + \lambda \rho_p} \xi_p(\cdot) \).

Thus we have \( W_\lambda \xi_p(\cdot) = \frac{\lambda \rho_p}{1 + \lambda \rho_p} \xi_p(\cdot) \).

43
By Lemma 3.1, any $\eta \in \mathcal{H}_0$ satisfies $\eta = \sum_{p=1}^{\infty} V(\eta, \xi^0_p) \xi^0_p$. Therefore, $V(K^0_{\xi^0_p}, \xi^0_p) = \langle K^0_{\xi^0_p}, \xi^0_p \rangle_0 / (1 + \lambda \rho^0_p)$. Hence, $K^0_{\xi}(\cdot) = \sum_{p=1}^{\infty} \frac{\xi^0_p(\cdot)}{1 + \lambda \rho^0_p} \xi^0_p(\cdot)$, and likewise, $W^0_{\lambda \rho^0_p}(\cdot) = \frac{\lambda \rho^0_p}{1 + \lambda \rho^0_p} \xi^0_p(\cdot)$.

References

Anderson, N. H., P. Hall, D. M. Titterington, et al. (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.

Anderson, T. W. (1958). *An introduction to multivariate statistical analysis*. New York: Wiley.

Bartlett, P. L., O. Bousquet, S. Mendelson, et al. (2005). Local rademacher complexities. The Annals of Statistics 33(4), 1497–1537.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the royal statistical society. Series B (Methodological), 289–300.

Berlinet, A. and C. Thomas-Agnan (2011). *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media.

Bomben, R., M. Dal Bo, D. Capello, D. Benedetti, D. Marconi, A. Zucchetto, F. Forconi, R. Maffei, E. M. Ghia, L. Laurenti, et al. (2007). Comprehensive characterization of ighv3-21–expressing b-cell chronic lymphocytic leukemia: an italian multicenter study. Blood 109(7), 2989–2998.

Cao, R. and I. Van Keilegom (2006). Empirical likelihood tests for two-sample problems via nonparametric density estimation. Canadian Journal of Statistics 34(1), 61–77.

Cheng, G. and Z. Shang (2015). Joint asymptotics for semi-nonparametric regression models with partially linear structure. The Annals of Statistics 43(3), 1351–1390.

Darling, D. A. (1957). The kolmogorov-smirnov, cramér-von mises tests. The Annals of Mathematical Statistics 28(4), 823–838.

de Jong, P. (1987). A central limit theorem for generalized quadratic forms. Probability Theory and Related Fields 75(2), 261–277.
Fabbri, G., A. B. Holmes, M. Viganotti, C. Scuoppo, L. Belver, D. Herranz, X.-J. Yan, Y. Kieso, D. Rossi, G. Gaidano, et al. (2017). Common nonmutational notch1 activation in chronic lymphocytic leukemia. *Proceedings of the National Academy of Sciences* 114(14), E2911–E2919.

Fan, J., C. Zhang, and J. Zhang (2001). Generalized likelihood ratio statistics and wilks phenomenon. *Annals of statistics* 29, 153–193.

Fromont, M., M. Lerasle, P. Reynaud-Bouret, et al. (2012). Kernels based tests with non-asymptotic bootstrap approaches for two-sample problems. In *Conference on Learning Theory*, pp. 23–1.

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

Gu, C. (2013). *Smoothing spline ANOVA models*, Volume 297. Springer Science & Business Media.

Gu, C. and C. Qiu (1993). Smoothing spline density estimation: Theory. *The Annals of Statistics*, 217–234.

Ingster, Y. I. (1987). Minimax testing of nonparametric hypotheses on a distribution density in the Lp metrics. *Theory of Probability & Its Applications* 31(2), 333–337.

Ingster, Y. I. (1989). Asymptotic minimax testing of independence hypothesis. *Journal of Soviet Mathematics* 44(4), 466–476.

Ingster, Y. I. (1993). Asymptotically minimax hypothesis testing for nonparametric alternatives. i, ii, iii. *Math. Methods Statist* 2(2), 85–114.

Jiang, B., C. Ye, and J. S. Liu (2015). Nonparametric k-sample tests via dynamic slicing. *Journal of the American Statistical Association* 110(510), 642–653.

Lepski, O. V., V. G. Spokoiny, et al. (1999). Minimax nonparametric hypothesis testing: the case of an inhomogeneous alternative. *Bernoulli* 5(2), 333–358.

Li, T. and M. Yuan (2019). On the optimality of gaussian kernel based nonparametric tests against smooth alternatives. *arXiv preprint arXiv:1909.03302*. 
Lin, Y. (2000). Tensor product space anova models. *Annals of Statistics*, 734–755.

Mann, H. B. and D. R. Whitney (1947). On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, 50–60.

Martínez-Camblor, P. and J. de Uña-Álvarez (2009). Non-parametric k-sample tests: density functions vs distribution functions. *Computational Statistics & Data Analysis* 53(9), 3344–3357.

Mason, D. M., J. H. Schuenemeyer, et al. (1983). A modified kolmogorov-smirnov test sensitive to tail alternatives. *The Annals of Statistics* 11(3), 933–946.

Maura, F., G. Cutrona, L. Mosca, S. Matis, S. Fabris, L. Agnelli, M. Colombo, C. Massucco, M. Ferracin, et al. (2015). Association between gene and mirna expression profiles and stereotyped subset# 4 b-cell receptor in chronic lymphocytic leukemia. *Leukemia & lymphoma* 56(11), 3150–3158.

Mendelson, S. (2002). Geometric parameters of kernel machines. In *International Conference on Computational Learning Theory*, pp. 29–43. Springer.

Miller, R. and D. Siegmund (1982). Maximally selected chi square statistics. *Biometrics*, 1011–1016.

Pfister, N., P. Bühlmann, B. Schölkopf, and J. Peters (2018). Kernel-based tests for joint independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80(1), 5–31.

Pinkus, A. (2012). *N-widths in Approximation Theory*, Volume 7. Springer Science & Business Media.

Qin, J., Y. Li, Z. Cai, S. Li, J. Zhu, F. Zhang, S. Liang, W. Zhang, Y. Guan, D. Shen, et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* 490(7418), 55.

Scholz, F. W. and M. A. Stephens (1987). K-sample anderson–darling tests. *Journal of the American Statistical Association* 82(399), 918–924.

Shang, Z. et al. (2010). Convergence rate and bahadur type representation of general smoothing spline m-estimates. *Electronic Journal of Statistics* 4, 1411–1442.
Shang, Z. and G. Cheng (2015). Nonparametric inference in generalized functional linear models. *The Annals of Statistics* 43(4), 1742–1773.

Shang, Z., G. Cheng, et al. (2013). Local and global asymptotic inference in smoothing spline models. *The Annals of Statistics* 41(5), 2608–2638.

Shapiro, S. S. and M. B. Wilk (1965). An analysis of variance test for normality (complete samples). *Biometrika* 52(3/4), 591–611.

Silverman, B. W. (1982). On the estimation of a probability density function by the maximum penalized likelihood method. *The Annals of Statistics*, 795–810.

Székely, G. J., M. L. Rizzo, N. K. Bakirov, et al. (2007). Measuring and testing dependence by correlation of distances. *The annals of statistics* 35(6), 2769–2794.

Tilg, H. and A. R. Moschen (2014). Microbiota and diabetes: an evolving relationship. *Gut* 63(9), 1513–1521.

Turnbaugh, P. J., M. Hamady, T. Yatsunenko, B. L. Cantarel, A. Duncan, R. E. Ley, M. L. Sogin, W. J. Jones, B. A. Roe, J. P. Affourtit, et al. (2009). A core gut microbiome in obese and lean twins. *nature* 457(7228), 480.

Wahba, G. (1990). *Spline models for observational data*, Volume 59. Siam.

Wei, Y. and M. J. Wainwright (2018). The local geometry of testing in ellipses: Tight control via localized kolmogorov widths. *arXiv:1712.00711*.

Wei, Y., M. J. Wainwright, and A. Guntuboyina (2017). The geometry of hypothesis testing over convex cones: Generalized likelihood tests and minimax radii. *arXiv preprint arXiv:1703.06810*.

Weinberger, H. F. (1974). *Variational methods for eigenvalue approximation*. SIAM.

Xing, X., J. S. Liu, and W. Zhong (2017). Metagen: reference-free learning with multiple metagenomic samples. *Genome biology* 18(1), 187.

Yamamoto, N., S.-i. Yamamoto, F. Inagaki, M. Kawaichi, A. Fukamizu, N. Kishi, K. Matsuno, K. Nakamura, G. Weinmaster, H. Okano, et al. (2001). Role of deltex-1 as a transcriptional regulator downstream of the notch receptor. *Journal of Biological Chemistry* 276(48), 45031–45040.
Supplement to Minimax Nonparametric Testing for Density Comparison

In this document, additional proofs are included.

- Section S.1 includes the proof of Lemma A.2.
- Section S.2 includes the proof of Lemma A.3.
- Section S.3 includes the proof of Lemma A.4.
- Section S.4 includes the proof of Lemma A.5.
- Section S.5 includes the proof of Lemma A.6.
- Section S.6 includes the proof of Lemma A.7.
- Section S.7 includes the supplementary simulation results

### S.1 Proof of Lemma A.2

Since $\rho_p \asymp p^{2m}$, we have

$$h^{-1} = \sum_{p=0}^{\infty} \frac{1}{(1 + \lambda \rho_p)^2} \times \int_{1}^{\infty} \frac{1}{(1 + \lambda p^{2m})^2} dx = \int_{\lambda^{1/2m}}^{\infty} \frac{1}{(1 + x^{2m})^2} dx = O(\lambda^{-1/2m})$$

Thus we have $h \asymp \lambda^{1/2m}$. Similarly, $h_0 \asymp \lambda^{1/2m}$.

### S.2 Proof of Lemma A.3

For any $y \in \mathcal{Y}$ and $\eta \in \mathcal{H}$, we have $|\eta(y)| = |\langle K_y, \eta \rangle| \leq \|K_y\| \cdot \|\eta\|$. So it is sufficient to find the upper bound for $\|K_y\|$. By Proposition A.1 and the boundedness of $\xi_p$’s, we have

$$\|K_y\|^2 = K(y, y) = \sum_{p=1}^{\infty} |\xi_p(y)|^2 \leq \frac{c_m}{1 + \lambda \rho_p} \leq c_m h^{-1}$$  \hspace{1cm} (S.1)

where $c_m > 0$ is a constant free of $y$ and $\eta$. 

1
S.3 Proof of Lemma A.4

The proof is rooted in Gu (2013). Consider the quadratic approximation of the integral \( \int_y e^{\eta(y)} dy \):

\[
\int_y e^{\eta(y)} dy \approx \int_y e^{\eta^*(y)} dy + \int_y (\eta - \eta^*) e^{\eta^*(y)} dy + \frac{1}{2} V(\eta - \eta^*, \eta - \eta^*). \tag{S.2}
\]

Dropping the terms that do not involve \( \eta \), and plugging (S.2) into (4), \( \ell_{n,\lambda}(\eta) \) has a quadratic approximation \( q_{n,\lambda}(\eta) \):

\[
q_{n,\lambda}(\eta) = -\frac{1}{n} \sum_{i=1}^n \eta(y_i) + \int_y \eta e^{\eta^*} dy + \frac{1}{2} V(\eta - \eta^*, \eta - \eta^*) + \frac{1}{2} J(\eta, \eta). \tag{S.3}
\]

Consider the Fourier expansions of \( \eta \) and \( \eta^* \):

\[
\eta(x, z) = \sum_{k=1}^{\infty} \sum_{l=1}^{a} \beta_{kl} \phi_k(x) \psi_l(z), \quad \eta^*(x, z) = \sum_{k=1}^{\infty} \sum_{l=1}^{a} \beta_{kl}^* \phi_k(x) \psi_l(z).
\]

Then, we have

\[
q_{n,\lambda}(\eta) = \sum_{k=1}^{\infty} \sum_{l=1}^{a} \left\{ -\beta_{kl}\left( \frac{1}{n} \sum_{i=1}^n \phi_k(x_i) \psi_l(z_i) - \mathbb{E}\{\phi_k(X)\psi_l(Z)\} \right) 
+ \frac{1}{2}(\beta_{kl} - \beta_{kl}^*)^2 + \frac{\lambda}{2} \mu_k \nu_l \beta_{kl}^2 \right\}. \tag{S.4}
\]

Write \( \gamma_{kl} = n^{-1} \sum_{i=1}^n \phi_k(X_i) \psi_l(Z_i) - \mathbb{E}\{\phi_k(X)\psi_l(Z)\} \). Minimizing (S.4) with respect to \( \beta_{kl} \)'s, we get the optimizer:

\[
\beta_{kl} = (\gamma_{kl} + \beta_{kl}^*)/(1 + \lambda \mu_k \nu_l), \quad k \geq 1, l = 1, \ldots, a.
\]

Then \( \bar{\eta} = \sum_{k=1}^{\infty} \sum_{l=1}^{a} \beta_{kl} \phi_k \psi_l \) becomes a linear approximation of \( \bar{\eta}_{n,\lambda} \). By direct calculations we get that

\[
V(\bar{\eta} - \eta^*) = \sum_{k=1}^{\infty} \sum_{l=1}^{a} (\beta_{kl} - \beta_{kl}^*)^2, \quad \lambda J(\bar{\eta} - \eta^*) = \sum_{i=1}^{\infty} \sum_{j=1}^{a} \lambda \mu_k \nu_l (\beta_{kl} - \beta_{kl}^*)^2.
\]

Since \( \mathbb{E} \gamma_{kl} = 0 \) and \( \mathbb{E} \gamma_{ij}^2 = 1/n \), we have

\[
\mathbb{E}\{V(\bar{\eta} - \eta^*)\} = \sum_{i=1}^{\infty} \sum_{j=1}^{a} \frac{1}{(1 + \lambda \mu_k \nu_l)^2} + \lambda \sum_{i=1}^{\infty} \sum_{j=1}^{a} \frac{\lambda \mu_k \nu_l}{(1 + \lambda \mu_k \nu_l)^2} \beta_{kl}^* \beta_{kl}^* 
\]

\[
\mathbb{E}\{\lambda J(\bar{\eta} - \eta^*)\} = \sum_{i=1}^{\infty} \sum_{j=1}^{a} \frac{1}{(1 + \lambda \mu_k \nu_l)^2} + \lambda \sum_{i=1}^{\infty} \sum_{j=1}^{a} \frac{\lambda \mu_k \nu_l}{(1 + \lambda \mu_k \nu_l)^2} \beta_{kl}^* \beta_{kl}^* \beta_{kl}^*
\]

\[\tag{S.5}\]
By similar derivations in Lemma A.2, it can be verified that
\[
\sum_{i=1}^{\infty} \sum_{j=1}^{a} \frac{1}{(1 + \lambda \mu_k \mu_l)^2} = O(\lambda^{-1/2m}),
\]
\[
\sum_{i=1}^{\infty} \sum_{j=1}^{a} \frac{\lambda \mu_k \mu_l}{(1 + \lambda \mu_k \mu_l)^2} = O(\lambda^{-1/2m}),
\]
\[
\sum_{i=1}^{\infty} \sum_{j=1}^{a} \frac{1}{(1 + \lambda \mu_k \mu_l)} = O(\lambda^{-1/2m}).
\]
Plugging into (S.5), we obtain that
\[
||\hat{\eta} - \eta^*||^2 = (V + \lambda J)(\hat{\eta} - \eta^*) = O_p(n^{-1} \lambda^{-1/2m} + \lambda). \tag{S.6}
\]

We now turn to the approximation error \(\hat{\eta} - \eta\). We calculate the Fréchet derivative of the quadratic approximation in (S.3) as
\[
Dq_{n,\lambda}(\eta)\Delta\eta = -\frac{1}{n} \sum_{i=1}^{n} \Delta\eta(y_i) + \int_{Y} \Delta\eta e^{\eta^*} d\eta + \lambda V(\eta - \eta^*, \Delta\eta) + \lambda J(\eta, \Delta\eta). \tag{S.7}
\]

Since \(Dq_{n,\lambda}(\eta) = 0\), setting \(\Delta\eta = \hat{\eta}_{n,\lambda} - \eta\), (S.7) is equal to
\[
-\frac{1}{n} \sum_{i=1}^{n} (\hat{\eta}_{n,\lambda} - \eta)(y_i) \int_{Y} (\hat{\eta}_{n,\lambda} - \eta)(y)e^{\eta^*} d\eta + \int_{Y} (\hat{\eta}_{n,\lambda} - \eta)(y) e^{\hat{\eta}_{n,\lambda}(y)} d\eta + \lambda J(\hat{\eta}_{n,\lambda}, \hat{\eta}_{n,\lambda} - \eta) \tag{S.8}
\]

Since \(D\ell_{n,\lambda}(\hat{\eta}_{n,\lambda}) = 0\), setting \(\Delta\eta = \hat{\eta}_{n,\lambda} - \eta\) yields
\[
D\ell_{n,\lambda}(\eta)\Delta\eta = -\frac{1}{n} \sum_{i=1}^{n} (\hat{\eta}_{n,\lambda} - \eta)(y_i) + \int_{Y} (\hat{\eta}_{n,\lambda} - \eta)(y) e^{\hat{\eta}_{n,\lambda}(y)} d\eta + \lambda J(\hat{\eta}_{n,\lambda}, \hat{\eta}_{n,\lambda} - \eta). \tag{S.9}
\]

Combining (S.8) and (S.9), we have
\[
\int_{Y} (\hat{\eta}_{n,\lambda} - \eta)(y) e^{\hat{\eta}_{n,\lambda}(y)} d\eta - \int_{Y} (\hat{\eta}_{n,\lambda} - \eta)(y) e^{\hat{\eta}(y)} d\eta + \lambda J(\hat{\eta}_{n,\lambda} - \eta) = \int_{Y} (\hat{\eta}_{n,\lambda} - \eta)(y) e^{\eta^*} d\eta - \int_{Y} (\hat{\eta}_{n,\lambda} - \eta)(y) e^{\hat{\eta}(y)} d\eta.
\]

By Taylor expansion,
\[
\int_{Y} (\hat{\eta}_{n,\lambda} - \eta)(y) e^{\hat{\eta}(y)} d\eta - \int_{Y} (\hat{\eta}_{n,\lambda} - \eta)(y) e^{\eta^*} d\eta = V(\hat{\eta}_{n,\lambda} - \eta, \hat{\eta} - \eta^*)(1 + o_p(1)),
\]
where the $o_p$ term holds as $\lambda \to 0$ and $n\lambda^{1/2m} \to \infty$. Define

$$D(\alpha) = \int_Y (\hat{\eta}_{n,\lambda} - \eta)(y)e^{\hat{\eta}_{n,\lambda}(y)} + o(\hat{\eta}_{n,\lambda} - \eta)(y)dy.$$  

It can be shown that $\dot{D}(\alpha) = V_{\hat{\eta} + \alpha(\hat{\eta}_{n,\lambda} - \eta)}(\hat{\eta}_{n,\lambda} - \eta)$. By the mean value theorem,

$$\int_Y (\hat{\eta}_{n,\lambda} - \eta)(y)e^{\hat{\eta}_{n,\lambda}(y)}dy - \int_Y (\hat{\eta}_{n,\lambda} - \eta)(y)e^\eta(y)dy = D(1) - D(0) = \dot{D}(\alpha) = V_{\hat{\eta} + \alpha(\hat{\eta}_{n,\lambda} - \eta)}(\hat{\eta}_{n,\lambda} - \eta),$$  

for some $\alpha \in [0,1]$. Then by Assumption 1, we have

$$c_1V(\hat{\eta}_{n,\lambda} - \eta) + \lambda J(\hat{\eta}_{n,\lambda} - \eta) \leq o_p(V(\eta - \eta^*, \hat{\eta} - \eta)) = o_p(V(\hat{\eta}_{n,\lambda} - \eta)V(\eta - \eta^*))^{1/2}$$

Combine with the estimation error (S.6), we have

$$\|\hat{\eta}_{n,\lambda} - \eta^*\|^2 = V(\hat{\eta}_{n,\lambda} - \eta^*) + \lambda J(\hat{\eta}_{n,\lambda} - \eta^*) = O_p(n^{-1}\lambda^{1/2m} + \lambda).$$

### S.4 Proof of Lemma A.5

Let $g = \hat{\eta}_{n,\lambda} - \eta^*$. By Taylor’s expansion we have

$$S_{n,\lambda}(\hat{\eta}_{n,\lambda}) = S_{n,\lambda}(\eta^*) + DS_{n,\lambda}(\eta^*)g + \int_0^1 \int_0^1 sD^2S_{n,\lambda}(\eta^* + ss'g)gdsds'.$$

By (A.16) and (6), one can check that $\langle DS_{n,\lambda}(\eta^*)g_1, g_2 \rangle = \langle g_1, g_2 \rangle$, and thus, $DS_{n,\lambda} = id$ is an identity operator. By the fact $S_{n,\lambda}(\hat{\eta}_{n,\lambda}) = 0$, we have

$$\|\hat{\eta}_{n,\lambda} - \eta^* - S_{n,\lambda}(\eta^*)\| = \| \int_0^1 \int_0^1 sD^2S_{n,\lambda}(\eta^* + ss'g)gdsds'\|. \tag{S.9}$$

By (A.17) we have $D^2S_{n,\lambda}(\eta^* + ss'g)g = \int_Y g(y)^2K_{\hat{\eta}}e^n(y) + ss'g(y)dy$. By Proposition A.1 and Lemma A.3, we have

$$\sup_{y \in Y} |g(y)|^2 \leq c_m h^{-1}\|g\|^2 = c_m h^{-1}O_P((nh)^{-1} + h^{2m}).$$

By (S.1), we have $\|E_{\eta^*} \{K_{\hat{\eta}}\} \| \leq c^{1/2}_m h^{-1/2}$. Thus, we have

$$\|D^2S_{n,\lambda}(\eta^* + ss'g)g\| = O(h^{-3/2}((nh)^{-1} + h^{2m})). \tag{S.10}$$

Plugging (S.10) into (S.9), we finish the proof.
S.5 Proof of Lemma A.6

Suppose the $\eta^\ast_0$ is the projection of $\eta^\ast$ on $\mathcal{H}_0$. Define an index set $I_0 = \{(k,l)|k = 1 \text{ or } l = 1\}$ corresponding to the basis, $\{\phi_k \psi_l|k = 1 \text{ or } l = 1\}$, of $\mathcal{H}_0$. When restricted to $\mathcal{H}_0$, the Fourier expansion of $\eta^\ast$ is

$$\eta^\ast_0(x, z) = \sum_{(k,l) \in I_0} \beta^0_{k,l} \phi_k(x) \psi_l(z).$$

Substituting the above $\eta^\ast_0$ as well as its Fourier expansion into the proof of Lemma A.4, all results remain valid, provided the following truth:

$$\mathbb{E}\left\{\frac{1}{n} \sum_{i=1}^{n} \phi_k(X_i) \psi_l(Z_i) - \mathbb{E}_{\eta^\ast}(\phi_k \psi_l)\right\}^2 = \frac{1}{n}$$

$$\mathbb{E}\left\{\frac{1}{n} \sum_{i=1}^{n} \phi_k(X_i) \psi_l(Z_i) \phi_{k'}(X_i) \psi_{l'}(Z_i) - \mathbb{E}_{\eta^\ast}(\phi_k \psi_l \phi_{k'} \psi_{l'})\right\}^2 \leq \frac{c}{n},$$

where $c$ is a positive constant. The existence of such $c$ is guaranteed by the uniform boundedness of $\phi_k(x)$’s as proved by Shang et al. (2013). Let $\tilde{\eta}_0^\ast$ be the projection of $\eta^\ast$ on the subspace $\mathcal{H}_0$ and $g = \tilde{\eta}_0^\ast - \eta^\ast_0$. Substituting $\eta^\ast_0$ and $\tilde{\eta}_0^\ast$ into the proof of Lemma A.4, the results would follow.

S.6 Proof of Lemma A.7

Let $\eta^\ast_0$ be the projection of $\eta^\ast$ on the subspace $\mathcal{H}_0$ and $g = \tilde{\eta}_0^\ast - \eta^\ast_0$. Substituting $\eta^\ast_0$ and $\tilde{\eta}_0^\ast$ into the proof of Lemma A.5, one can show the desired results.

S.7 Supplementary Simulation Results

In this section, we compare the proposed method with KS test, MMD test (Gretton et al., 2012), ELT (Cao and Van Keilegom, 2006), Anderson-Darling test (Scholz and Stephens, 1987), density equality test (Anderson et al., 1994) and Mann-Whitney U-Test (Mann and Whitney, 1947). We summarize the simulation results for settings 1-4. In each setting, we chose the averaged sample size $n$ in each group as 125, 250, 375, 500, 625, 750, 875, 1000. We chose the roughness parameter by directly plugging in $\lambda = n^{-4m/(4m+1)}$, which is theoretically guaranteed by Theorem 4.3, and estimated $\theta$s by using the trace of the corresponding kernel matrix which is also suggested by Gu (2013). Size and power were calculated as the proportions of rejection based on 1000 independent trials.
| $\delta$ | n  | PLR  | KS  | MMD  | ELT  | DET  | AD  | WM  | DSLICE |
|----------|-----|------|-----|------|------|------|-----|-----|---------|
|          | 125 | 0.059 | 0.020 | 0.035 | 0.065 | 0.065 | 0.030 | 0.025 | 0.040    |
|          | 250 | 0.037 | 0.050 | 0.050 | 0.060 | 0.065 | 0.075 | 0.085 | 0.030    |
|          | 375 | 0.041 | 0.065 | 0.050 | 0.050 | 0.045 | 0.060 | 0.045 | 0.035    |
|          | 500 | 0.039 | 0.070 | 0.055 | 0.050 | 0.040 | 0.050 | 0.035 | 0.035    |
|          | 625 | 0.041 | 0.045 | 0.030 | 0.055 | 0.055 | 0.040 | 0.050 | 0.030    |
| $\delta_2 = 0$ | 750 | 0.049 | 0.045 | 0.070 | 0.060 | 0.065 | 0.050 | 0.060 | 0.020    |
|          | 865 | 0.053 | 0.085 | 0.095 | 0.030 | 0.040 | 0.070 | 0.075 | 0.005    |
|          | 1000 | 0.041 | 0.040 | 0.035 | 0.070 | 0.075 | 0.035 | 0.035 | 0.020    |
|          | 125 | 0.120 | 0.040 | 0.060 | 0.105 | 0.095 | 0.055 | 0.030 | 0.090    |
|          | 250 | 0.123 | 0.050 | 0.115 | 0.150 | 0.155 | 0.075 | 0.050 | 0.090    |
|          | 375 | 0.225 | 0.060 | 0.150 | 0.150 | 0.140 | 0.120 | 0.030 | 0.110    |
|          | 500 | 0.309 | 0.080 | 0.225 | 0.160 | 0.135 | 0.145 | 0.035 | 0.095    |
|          | 625 | 0.328 | 0.125 | 0.340 | 0.250 | 0.210 | 0.240 | 0.060 | 0.145    |
|          | 750 | 0.391 | 0.105 | 0.385 | 0.330 | 0.290 | 0.260 | 0.060 | 0.140    |
| I        | 865 | 0.490 | 0.175 | 0.425 | 0.355 | 0.305 | 0.320 | 0.065 | 0.165    |
|          | 1000 | 0.522 | 0.140 | 0.505 | 0.500 | 0.425 | 0.360 | 0.055 | 0.210    |
| $\delta_2 = 0.15$ | 125 | 0.291 | 0.070 | 0.205 | 0.250 | 0.205 | 0.115 | 0.060 | 0.180    |
|          | 250 | 0.468 | 0.095 | 0.410 | 0.430 | 0.410 | 0.210 | 0.030 | 0.335    |
|          | 375 | 0.635 | 0.265 | 0.615 | 0.545 | 0.525 | 0.515 | 0.075 | 0.475    |
|          | 500 | 0.853 | 0.370 | 0.745 | 0.810 | 0.770 | 0.645 | 0.040 | 0.555    |
|          | 625 | 0.905 | 0.485 | 0.880 | 0.905 | 0.900 | 0.805 | 0.055 | 0.695    |
|          | 750 | 0.949 | 0.455 | 0.950 | 0.945 | 0.945 | 0.925 | 0.075 | 0.765    |
|          | 865 | 0.983 | 0.570 | 0.940 | 0.965 | 0.975 | 0.930 | 0.095 | 0.800    |
|          | 1000 | 0.995 | 0.655 | 0.990 | 0.980 | 0.980 | 0.970 | 0.040 | 0.900    |
| $\delta_2 = 0.30$ | 125 | 0.464 | 0.160 | 0.380 | 0.440 | 0.400 | 0.265 | 0.060 | 0.345    |
|          | 250 | 0.821 | 0.275 | 0.770 | 0.820 | 0.805 | 0.625 | 0.040 | 0.600    |
|          | 375 | 0.956 | 0.455 | 0.920 | 0.960 | 0.955 | 0.900 | 0.050 | 0.855    |
|          | 500 | 0.989 | 0.700 | 0.990 | 1.000 | 1.000 | 0.990 | 0.040 | 0.930    |
|          | 625 | 1.000 | 0.805 | 0.995 | 1.000 | 1.000 | 0.995 | 0.040 | 0.990    |
|          | 750 | 1.000 | 0.870 | 1.000 | 1.000 | 1.000 | 0.990 | 0.050 | 0.985    |
|          | 865 | 1.000 | 0.985 | 1.000 | 1.000 | 1.000 | 1.000 | 0.045 | 0.995    |
|          | 1000 | 1.000 | 0.960 | 1.000 | 1.000 | 1.000 | 1.000 | 0.035 | 1.000    |

Table 1: Simulation results for setting 1.
| $\delta$        | Sample size | PLR | KS | MMD | ELT | DET | AD | WM | DSLICE |
|-----------------|-------------|-----|----|-----|-----|-----|----|----|--------|
| $\delta_2 = 0$  | 125         | 0.066 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.045  |
|                 | 250         | 0.064 | 0.000 | 0.005 | 0.000 | 0.000 | 0.000 | 0.000 | 0.015  |
|                 | 375         | 0.044 | 0.000 | 0.000 | 0.005 | 0.005 | 0.005 | 0.000 | 0.005  |
|                 | 500         | 0.060 | 0.000 | 0.000 | 0.000 | 0.000 | 0.005 | 0.000 | 0.000  |
|                 | 625         | 0.038 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000  |
|                 | 750         | 0.044 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.010  |
|                 | 865         | 0.043 | 0.000 | 0.005 | 0.000 | 0.005 | 0.000 | 0.000 | 0.000  |
|                 | 1000        | 0.046 | 0.000 | 0.000 | 0.005 | 0.000 | 0.000 | 0.000 | 0.010  |
| $\delta_2 = 0.15$ | 125         | 0.078 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.025  |
|                 | 250         | 0.082 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.020  |
|                 | 375         | 0.098 | 0.000 | 0.000 | 0.015 | 0.005 | 0.005 | 0.000 | 0.015  |
|                 | 500         | 0.116 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.010  |
|                 | 625         | 0.123 | 0.000 | 0.000 | 0.015 | 0.005 | 0.005 | 0.000 | 0.020  |
|                 | 750         | 0.134 | 0.005 | 0.000 | 0.005 | 0.000 | 0.000 | 0.000 | 0.025  |
|                 | 865         | 0.150 | 0.000 | 0.000 | 0.015 | 0.005 | 0.005 | 0.000 | 0.025  |
|                 | 1000        | 0.164 | 0.000 | 0.000 | 0.015 | 0.005 | 0.005 | 0.000 | 0.015  |
| $\delta_2 = 0.30$ | 125         | 0.093 | 0.005 | 0.005 | 0.010 | 0.005 | 0.000 | 0.000 | 0.045  |
|                 | 250         | 0.184 | 0.005 | 0.000 | 0.010 | 0.005 | 0.010 | 0.000 | 0.110  |
|                 | 375         | 0.216 | 0.005 | 0.005 | 0.010 | 0.010 | 0.005 | 0.000 | 0.055  |
|                 | 500         | 0.296 | 0.000 | 0.000 | 0.010 | 0.000 | 0.000 | 0.000 | 0.100  |
|                 | 625         | 0.366 | 0.000 | 0.000 | 0.040 | 0.035 | 0.005 | 0.000 | 0.085  |
|                 | 750         | 0.428 | 0.025 | 0.010 | 0.045 | 0.025 | 0.010 | 0.000 | 0.155  |
|                 | 865         | 0.554 | 0.025 | 0.015 | 0.075 | 0.050 | 0.025 | 0.000 | 0.160  |
|                 | 1000        | 0.600 | 0.020 | 0.015 | 0.080 | 0.045 | 0.035 | 0.000 | 0.235  |
| $\delta_2 = 0.45$ | 125         | 0.144 | 0.000 | 0.000 | 0.030 | 0.000 | 0.000 | 0.000 | 0.110  |
|                 | 250         | 0.448 | 0.010 | 0.005 | 0.030 | 0.005 | 0.005 | 0.000 | 0.185  |
|                 | 375         | 0.544 | 0.005 | 0.010 | 0.060 | 0.030 | 0.015 | 0.000 | 0.270  |
|                 | 500         | 0.758 | 0.040 | 0.020 | 0.085 | 0.070 | 0.080 | 0.000 | 0.380  |
|                 | 625         | 0.842 | 0.040 | 0.015 | 0.235 | 0.170 | 0.095 | 0.000 | 0.525  |
|                 | 750         | 0.922 | 0.080 | 0.055 | 0.360 | 0.285 | 0.165 | 0.000 | 0.580  |
|                 | 865         | 0.942 | 0.135 | 0.085 | 0.465 | 0.400 | 0.290 | 0.000 | 0.715  |
|                 | 1000        | 0.966 | 0.195 | 0.125 | 0.530 | 0.535 | 0.370 | 0.000 | 0.870  |

Table 2: Simulation results for setting 2.
| $\delta$ | Sample size | PLR | KS | MMD | ELT | DET | AD | WM | DSLICE |
|---------|-------------|-----|----|-----|-----|-----|----|-----|--------|
| $\delta_2 = 0$ | 125 | 0.040 | 0.055 | 0.060 | 0.020 | 0.035 | 0.045 | 0.045 | 0.035 |
|         | 250 | 0.068 | 0.055 | 0.065 | 0.050 | 0.060 | 0.070 | 0.055 | 0.030 |
|         | 375 | 0.060 | 0.055 | 0.065 | 0.050 | 0.060 | 0.070 | 0.065 | 0.030 |
|         | 500 | 0.064 | 0.050 | 0.060 | 0.045 | 0.050 | 0.060 | 0.060 | 0.030 |
|         | 625 | 0.054 | 0.090 | 0.090 | 0.040 | 0.040 | 0.095 | 0.085 | 0.015 |
|         | 750 | 0.044 | 0.040 | 0.025 | 0.050 | 0.070 | 0.030 | 0.035 | 0.005 |
|         | 865 | 0.058 | 0.030 | 0.020 | 0.040 | 0.045 | 0.040 | 0.040 | 0.015 |
|         | 1000 | 0.056 | 0.030 | 0.030 | 0.040 | 0.040 | 0.030 | 0.030 | 0.010 |
| $\delta_2 = 0.15$ | 125 | 0.068 | 0.040 | 0.045 | 0.025 | 0.040 | 0.045 | 0.035 | 0.060 |
|         | 250 | 0.096 | 0.070 | 0.080 | 0.060 | 0.065 | 0.065 | 0.045 | 0.025 |
|         | 375 | 0.124 | 0.075 | 0.085 | 0.055 | 0.065 | 0.065 | 0.055 | 0.035 |
|         | 500 | 0.132 | 0.080 | 0.115 | 0.040 | 0.060 | 0.065 | 0.035 | 0.025 |
|         | 625 | 0.146 | 0.045 | 0.110 | 0.045 | 0.045 | 0.060 | 0.050 | 0.020 |
|         | 750 | 0.174 | 0.070 | 0.120 | 0.075 | 0.080 | 0.090 | 0.045 | 0.025 |
|         | 865 | 0.210 | 0.075 | 0.155 | 0.045 | 0.035 | 0.120 | 0.040 | 0.050 |
|         | 1000 | 0.236 | 0.065 | 0.150 | 0.045 | 0.055 | 0.130 | 0.055 | 0.075 |
| $\delta_2 = 0.30$ | 125 | 0.092 | 0.045 | 0.090 | 0.030 | 0.045 | 0.075 | 0.035 | 0.045 |
|         | 250 | 0.170 | 0.030 | 0.120 | 0.065 | 0.070 | 0.075 | 0.030 | 0.055 |
|         | 375 | 0.240 | 0.090 | 0.190 | 0.060 | 0.075 | 0.155 | 0.055 | 0.090 |
|         | 500 | 0.328 | 0.110 | 0.260 | 0.065 | 0.055 | 0.145 | 0.025 | 0.155 |
|         | 625 | 0.368 | 0.120 | 0.315 | 0.050 | 0.050 | 0.240 | 0.060 | 0.145 |
|         | 750 | 0.438 | 0.140 | 0.345 | 0.090 | 0.090 | 0.270 | 0.065 | 0.155 |
|         | 865 | 0.606 | 0.170 | 0.435 | 0.095 | 0.060 | 0.280 | 0.025 | 0.175 |
|         | 1000 | 0.622 | 0.170 | 0.455 | 0.110 | 0.060 | 0.380 | 0.035 | 0.240 |
| $\delta_2 = 0.45$ | 125 | 0.130 | 0.075 | 0.135 | 0.040 | 0.045 | 0.115 | 0.065 | 0.115 |
|         | 250 | 0.350 | 0.105 | 0.270 | 0.085 | 0.080 | 0.160 | 0.060 | 0.160 |
|         | 375 | 0.424 | 0.140 | 0.405 | 0.075 | 0.090 | 0.275 | 0.045 | 0.220 |
|         | 500 | 0.674 | 0.185 | 0.480 | 0.105 | 0.070 | 0.365 | 0.030 | 0.315 |
|         | 625 | 0.802 | 0.200 | 0.610 | 0.110 | 0.055 | 0.475 | 0.035 | 0.460 |
|         | 750 | 0.856 | 0.280 | 0.695 | 0.190 | 0.090 | 0.630 | 0.050 | 0.495 |
|         | 865 | 0.932 | 0.355 | 0.755 | 0.205 | 0.075 | 0.680 | 0.070 | 0.465 |
|         | 1000 | 0.952 | 0.495 | 0.870 | 0.260 | 0.085 | 0.840 | 0.095 | 0.610 |

Table 3: Simulation results for setting 3.
| $\delta$ | Sample size | PLR | KS | MMD | ELT | DET | AD | WM | DSLICE |
|---|---|---|---|---|---|---|---|---|---|
| $\delta_2 = 0$ | 125 | 0.060 | 0.000 | 0.000 | 0.000 | 0.005 | 0.000 | 0.000 | 0.020 |
|  | 250 | 0.060 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.010 |
|  | 375 | 0.056 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.020 |
|  | 500 | 0.040 | 0.005 | 0.000 | 0.005 | 0.000 | 0.000 | 0.000 | 0.020 |
|  | 625 | 0.044 | 0.000 | 0.000 | 0.010 | 0.000 | 0.000 | 0.000 | 0.020 |
|  | 750 | 0.059 | 0.000 | 0.000 | 0.005 | 0.000 | 0.000 | 0.000 | 0.015 |
|  | 865 | 0.055 | 0.005 | 0.000 | 0.010 | 0.000 | 0.000 | 0.000 | 0.010 |
| $\delta_2 = 0.15$ | 125 | 0.042 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.041 |
|  | 250 | 0.092 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.040 |
|  | 375 | 0.108 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.050 |
|  | 500 | 0.152 | 0.005 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.040 |
|  | 625 | 0.182 | 0.020 | 0.000 | 0.010 | 0.000 | 0.010 | 0.000 | 0.030 |
|  | 750 | 0.196 | 0.005 | 0.000 | 0.015 | 0.000 | 0.025 | 0.000 | 0.045 |
|  | 865 | 0.232 | 0.005 | 0.000 | 0.015 | 0.000 | 0.010 | 0.000 | 0.050 |
|  | 1000 | 0.286 | 0.010 | 0.000 | 0.015 | 0.000 | 0.020 | 0.000 | 0.025 |
| $\delta_2 = 0.30$ | 125 | 0.068 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.080 |
|  | 250 | 0.156 | 0.005 | 0.000 | 0.000 | 0.005 | 0.000 | 0.000 | 0.100 |
|  | 375 | 0.336 | 0.020 | 0.000 | 0.000 | 0.025 | 0.000 | 0.000 | 0.105 |
|  | 500 | 0.568 | 0.005 | 0.000 | 0.005 | 0.000 | 0.015 | 0.000 | 0.170 |
|  | 625 | 0.596 | 0.025 | 0.000 | 0.020 | 0.000 | 0.065 | 0.000 | 0.205 |
|  | 750 | 0.684 | 0.010 | 0.000 | 0.020 | 0.000 | 0.085 | 0.000 | 0.270 |
|  | 865 | 0.756 | 0.040 | 0.000 | 0.030 | 0.000 | 0.170 | 0.000 | 0.275 |
|  | 1000 | 0.822 | 0.045 | 0.000 | 0.050 | 0.000 | 0.135 | 0.000 | 0.295 |
| $\delta_2 = 0.45$ | 125 | 0.120 | 0.005 | 0.000 | 0.000 | 0.000 | 0.005 | 0.000 | 0.100 |
|  | 250 | 0.372 | 0.005 | 0.000 | 0.000 | 0.000 | 0.040 | 0.000 | 0.215 |
|  | 375 | 0.616 | 0.010 | 0.000 | 0.000 | 0.005 | 0.000 | 0.000 | 0.210 |
|  | 500 | 0.854 | 0.035 | 0.000 | 0.015 | 0.000 | 0.115 | 0.000 | 0.395 |
|  | 625 | 0.906 | 0.025 | 0.000 | 0.045 | 0.000 | 0.180 | 0.000 | 0.475 |
|  | 750 | 0.960 | 0.060 | 0.000 | 0.070 | 0.000 | 0.325 | 0.000 | 0.570 |
|  | 865 | 0.982 | 0.090 | 0.000 | 0.080 | 0.000 | 0.440 | 0.000 | 0.650 |
|  | 1000 | 0.998 | 0.105 | 0.000 | 0.095 | 0.000 | 0.590 | 0.000 | 0.700 |

Table 4: Simulation results for setting 4.