Statistica Sinica Preprint No: SS-2022-0141	
Title	Minimax Nonparametric Multi-sample Test under Smoothing
Manuscript ID	SS-2022-0141
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202022.0141
Complete List of Authors	Xin Xing, Zuofeng Shang, Pang Du, Ping Ma, Wenxuan Zhong and Jun S. Liu
Corresponding Authors	Xin Xing
E-mails	xinxing@vt.edu
Notice: Accepted version subject to English editing.	

Statistica Sinica

Minimax Nonparametric Multi-sample Test Under Smoothing

Xin Xing[1], Zuofeng Shang[2], Pang Du[1], Ping Ma[3], Wenxuan Zhong[3], Jun S. Liu[4]

[1]*Virginia Tech,* [2]*New Jersey Institute of Technology*

[3]*University of Georgia,* [4]*Harvard University*

We consider the problem of comparing probability densities among multiple groups. A new probabilistic tensor product smoothing spline framework is developed to model the joint density of two variables. Under such a framework, the probability density comparison is equivalent to testing the presence/absence of interactions. We propose a penalized likelihood ratio test for such interaction testing and show that the test statistic is asymptotically chi-square distributed under the null hypothesis. Furthermore, we derive a sharp minimax testing rate based on the Bernstein width for nonparametric multi-sample tests and show that our proposed test statistic is minimax optimal. In addition, a data-adaptive tuning criterion is developed to choose the penalty parameter. Simulations and real applications demonstrate that the proposed test outperforms the conventional approaches under various scenarios.

Key words: minimax optimality; nonparametric test; penalized likelihood ratio test; smoothing splines; multi-sample test; Wilks' phenomenon.

1. Introduction

A fundamental problem in statistics is to test whether the probability densities underlying U groups of observed data are the same, which is called the multi-sample test. It plays an essential role in different scientific fields ranging from modern biological sciences to deep learning. For instance, in metagenomics studies, comparing densities of specific microbial species (or strains) from different treatment groups helps researchers gain insights on the disease and treatments (Bilban et al., 2006; Turnbaugh et al., 2009; Qin et al., 2012); in genomics, identifying differentially expressed genes among multiple groups or conditions is fundamental to many downstream analyses; in machine learning, the multi-sample test is becoming an essential component in some deep learning algorithms (Li et al., 2017).

In these modern applications, the underlying distributions usually demonstrate complex patterns, including multi-modality and long-tails. Hence, it is often difficult to specify their distributional families. 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. Classical examples include distance-based tests such as the Kolmogorov-Smirnov (K-S) test (Darling, 1957), the Anderson-Darling test (Scholz and Stephens, 1987),

and their variants. An alternative direction is using discretization ("slicing") of continuous random variables (Miller and Siegmund, 1982). Jiang et al. (2015) proposed the dynamic slicing test (DSLICE), which penalizes the number of slices to regularize the test statistics. Gretton et al. (2007, 2012) proposed maximum mean discrepancy (MMD) two-sample tests via embedding the probability distribution into a reproducible kernel Hilbert space (RKHS). Eric et al. (2008) proposed the regularized MMD test by regularizing eigenvalues of the kernel matrix. Kim (2021) extended the MMD test to multi-sample test using the maximum of pair-wise MMDs. In addition, a class of approaches based on kernel density estimation was proposed (Anderson et al., 1994; Cao and Van Keilegom, 2006; Martínez-Camblor et al., 2008; Martínez-Camblor and de Uña-Álvarez, 2009; Zhan and Hart, 2014). One common challenge for MMD based and kernel density based testing approaches is the choice of tuning parameters, e.g., the kernel bandwidth or the roughness penalty parameter, since these parameters sensitively affect the methods' power. Furthermore, they have some drawbacks when applied to data of long-tailed distributions: since the kernel bandwidth is fixed across the entire sample (Silverman, 1986), they tend to have a low power in detecting changes at tails. In many applications such as gene expression analyses, metagenomics, and economics, long-tailed distributions are very common.

To overcome these limitations, we propose a likelihood-based test that can automatically adapt to densities with different shapes and develop a data-adaptive tuning method to automatically choose the penalization parameter. In this paper, we consider X as a continuous random vector and Z as a discrete random variable indicating the group information. Instead of directly comparing the multiple densities, we characterize the dependence between X and Z through its log-transformed joint density $\eta(x, z)$ within a space $\mathcal{H}$. The key idea is to uniquely decompose the log-transformed joint density $\eta \in \mathcal{H}$ into the main effects η_X, η_Z and the interaction effect η_{XZ} through a novel probabilistic decomposition of $\mathcal{H}$ so that the magnitude of the interaction exactly quantifies the density difference between multiple groups. The multi-sample test is thus equivalent to the interaction test

$$H_0 : \eta_{XZ}(x, z) = 0 \text{ vs. } H_1 : \eta_{XZ}(x, z) \neq 0. \tag{1.1}$$

We propose a penalized likelihood ratio (PLR) test by evaluating the penalized log-likelihood functional of η under H_0 and H_1, and establish its null distribution as a chi-square distribution. Compared with distance-based testing methods, which are not easily generalizable to handle multi-sample tests since the asymptotic distribution of the maximum pair-wise distance usually does not have an explicit form, the proposed PLR test can be directly applied to multi-sample tests by letting $Z \in \{1, \ldots, U\}$. We further propose a data-

adaptive rule to select the tuning parameter to guarantee testing optimality. The PLR test makes a full use of the distribution information and is sensitive to the density difference between the null and alternative hypotheses.

This work has main contributions sumarized in the following. **First**, without explicit expression of the function estimate, the classical technical tools used in Wald-type nonparamatric test in Xing et al. (2020) and Liu et al. (2021, 2020) can not be generalized to likelihood-based test. We propose a new probabilistic decomposition of the tensor product RKHS in Section 3. Existing references on functional decomposition without considering probabilistic measures (Gu, 2013; Wahba, 1990) mainly focus on estimation while leaving the hypothesis testing an open problem. Embedding the probability measures of X and Z into the tensor product decomposition of $\mathcal{H}$, we can transform the problem of density comparison to the problem of significance test of the interaction between X and Z, which provides a foundation to establish the minimax testing principle (see Section 4). This new probabilistic decomposition framework can be generalized to a broader class of dependence tests, including higher order independence tests and conditional independence tests, by using the magnitudes of the decomposed terms to measure the corresponding dependency. **Second**, we establish the minimax lower bound for density comparison problems based on the Bernstein width (Pinkus, 2012). Existing

minimax lower bounds of the testing rate are commonly derived based on Gaussian sequence models (Ingster, 1989, 1993; Wei and Wainwright, 2018; Xing et al., 2020) in a simple regression setting, and thus cannot be adapted to density comparison. In contrast, our result can be easily generalized to a wide range of dependence testing problems. We further prove the PLR based multi-sample test is minimax optimal. Compared with our proposed PLR test, the log-likelihood ratio without a penalty term does not enjoy the minimax optimality. Parallel to our work, Li and Yuan (2019) proposed a normalized MMD by choosing scaling parameters of the Gaussian kernel properly, and established its minimax property. Similar to the original MMD (Gretton et al., 2007), the approach in Li and Yuan (2019) is also based on a fixed kernel bandwidth, which can lead to low sensitivity when the underlying densities are long-tailed. However, our proposed approach is based on the penalized likelihood estimators, which can automatically adapt to long-tailed distributions. As shown in various simulation and real data studies in Sections 5 and 6, our proposed test shows a higher power when the underlying densities have complex features such as long-tails and multi-modality. In addition, we reveal an interesting connection between the PLR and MMD tests in our supplimenary. Also, we thank our referees for providing some helpful insights on the connections between MMD test and Hilbert-Schmidt independence criterion

(HSIC) test. We show that the MMD test (with a particularly selected kernel) is exactly the squared norm of the gradient of the log-likelihood ratio.

The rest of this paper is organized as follows. In Section 2, we construct our proposed penalized likelihood ratio test. Section 3 introduces the construction of the probabilistic decomposition of tensor product RHKS and main theoretical results, including the asymptotic distribution of the PLR test and its power performance. Section 4 established the minimax lower bound of density comparisons., we demonstrate the finite sample performance of our test through simulation studies. Section 6 is the analysis of two real-world examples using our test. Section 7 contains some discussion. In supplementary, we extend our PLR test to the case when the number of samples is divergent, and establish the minimax distinguishable rate and build the connection between our PLR test and the MMD test. Also, the the proofs of the main results are provided in Supplemenary.

2. Penalized likelihood ratio (PLR) for multi-sample test

The multi-sample problem can be stated as follows. Suppose that we have n independent d-dimensional observations, $X_i \in [0,1]^d, i = 1, \ldots, n$. Each X_i is associated with a label $Z_i \in \{1, \ldots, U\}$, which indicates that X_i is taken from the population indexed by Z_i with a probability density function f_{Z_i}. We aim

to test whether $f_1, \ldots, f_U$ are the same. Other than a smoothness constraint, we will not impose any other constraints on the probability density functions $f_1, \ldots, f_U$.

An equivalent formulation of the problem can be given in terms of the joint distribution of X and Z and their conditional independence. That is, consider n i.i.d. observations, $\mathbf{Y}_i = (X_i, Z_i)$, $i = 1, \ldots, n$, taken from a population $Y = (X, Z)$ with a joint probability density $f(x, z)$. Let

$$\eta(x, z) = \log(f(x, z)).$$

Let $f_{X|Z=z}(x)$ be the conditional density of X given $Z = z$ for $z = 1, \ldots, U$. The multi-sample problem is equivalent to testing whether X and Z are independent, i.e.,

$$H_0 : f_{X|Z=1}(\cdot) = \cdots = f_{X|Z=U}(\cdot)$$
$$v.s. \quad H_1 : \exists\ u_1 \neq u_2 \text{ such that } f_{X|Z=u_1}(\cdot) \neq f_{X|Z=u_2}(\cdot). \quad (2.1)$$

We denote $n_1 = |\{i : Z_i = 1\}|, \ldots, n_U = |\{i : Z_i = U\}|$, and assume that the n_j's are comparable, i.e., there exist constants $0 < c_1 \leq c_2$ such that $c_1 n_1 \leq n_u \leq c_2 n_1$, $\forall\ u = 1, \ldots, U$. We characterize the dependence between X and Z by their interaction with respect to their joint density, and show that testing the significance of such interaction is equivalent to the multi-sample test. We first consider the case when U is a fixed constant and then extend

the theory for diverging U.

In order to characterize the interaction between X and Z, we first define two averaging operators acting on the log-transformed joint density function $\eta(x,z)$. For any x, the operator $\mathcal{A}_x$ maps $\eta(x,z)$ to $\mathbb{E}_X\eta(X,z)$, a function in z; and for any z, the operator $\mathcal{A}_z$ maps $\eta(x,z)$ to $\mathbb{E}_Z\eta(x,Z)$. The interaction term is then characterized through the decomposition

$$\eta_{XZ}(x,z) = (\mathcal{I}-\mathcal{A}_x)(\mathcal{I}-\mathcal{A}_z)\eta(x,z) \equiv \eta(x,z)-(\mathcal{A}_x\eta)(z)-(\mathcal{A}_z\eta)(x)+\mathcal{A}_x\mathcal{A}_z\eta, \tag{2.2}$$

where $\mathcal{I}$ is the identity operator. Note that (2.2) is essentially derived from a functional ANOVA decomposition of $\eta(x,z)$ where $\mathcal{A}_x\mathcal{A}_z\eta$ is the constant, $(\mathcal{I}-\mathcal{A}_x)\mathcal{A}_z\eta$ and $(\mathcal{I}-\mathcal{A}_z)\mathcal{A}_x\eta$ are respectively the main effects of x and z, and $(\mathcal{I}-\mathcal{A}_x)(\mathcal{I}-\mathcal{A}_z)\eta$ is the interaction effect. A straightforward derivation shows that the multi-sample test is equivalent to testing whether η_{XZ} is zero or not; see Proposition S.4 in the Supplimentary.

We assume that η is in a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ and let $\mathcal{H}_0 = \{\eta \in \mathcal{H} \mid \eta_{XZ} = 0\}$ be the subspace of $\mathcal{H}$ containing all bivariate functions whose ANOVA decompositions have a zero interaction term. Based on Proposition S.4, the multi-sample test problem in (2.1) is equivalent to testing

$$H_0 : \eta \in \mathcal{H}_0 \quad v.s. \quad H_1 : \eta \in \mathcal{H}\backslash\mathcal{H}_0. \tag{2.3}$$

Consider estimating η by the minimizer of the penalized likelihood

$$\ell_{n,\lambda}(\eta) = -\frac{1}{n}\sum_{i=1}^{n}\eta(x_i, z_i) + \sum_{z\in\{1,\ldots,U\}}\int_{\mathcal{X}} e^{\eta(x,z)}dx + \frac{\lambda}{2}J(\eta), \tag{2.4}$$

where $\mathcal{X} = [0,1]^d$, the two sums form the negative log-likelihood representing the goodness-of-fit, $J(\cdot)$ is a quadratic functional enforcing a roughness penalty on η, and $\lambda > 0$ is a tuning parameter controlling the trade-off. We propose the following PLR test statistic

$$PLR = \inf_{\eta\in\mathcal{H}_0}\ell_{n,\lambda}(\eta) - \inf_{\eta\in\mathcal{H}}\ell_{n,\lambda}(\eta), \tag{2.5}$$

where the first and second terms are respectively the optimal penalized likelihoods under the reduced model $\mathcal{H}_0$ and the full model $\mathcal{H}$.

Note that the integrals in (2.4) are to guarantee the unitary constraint of a probability density function (see Theorem 3.1 in Silverman (1982)). We choose equation (2.4) instead of the logarithm of the integral in Gu and Qiu (1993) since the Fréchet derivative of the PLR will include an integral in the denominator, which makes the theoretical derivation more difficult.

2.1 Penalized likelihood functional under the full model

Under the full model, the minimization of (2.4) is performed in $\mathcal{H}$. Let $\mathcal{H}^{\langle X\rangle}$ be an RKHS of functions on the marginal domain $[0,1]^d$ and $\mathcal{H}^{\langle Z\rangle}$ be an RKHS of functions on $\{1,\ldots,U\}$. Then the full space $\mathcal{H} = \mathcal{H}^{\langle X\rangle}\otimes\mathcal{H}^{\langle Z\rangle}$ is their tensor

product and also an RKHS, where $\otimes$ denotes the tensor product of two linear spaces. Correspondingly, if $\mathcal{K}^{\langle X\rangle}$ and $\mathcal{K}^{\langle Z\rangle}$ are respectively the reproducing kernels (RKs) uniquely associated with the RKHS $\mathcal{H}^{\langle X\rangle}$ and $\mathcal{H}^{\langle Z\rangle}$, then the RK for $\mathcal{H}$ is simply the product of $\mathcal{K}^{\langle X\rangle}$ and $\mathcal{K}^{\langle Z\rangle}$, that is, $\mathcal{K}(\mathbf{Y}_i, \mathbf{Y}_j) = \mathcal{K}^{\langle X\rangle}(X_i, X_j)\mathcal{K}^{\langle Z\rangle}(Z_i, Z_j)$.

For the continuous domain $[0,1]^d$, we consider the mth order Sobolev space on $[0,1]^d$, i.e., $\mathcal{H}^{\langle X\rangle} = \{f \in L^2([0,1]^d) \mid f^{(\alpha)} \in L^2([0,1]^d), \quad \forall\, |\alpha| \le m\}$ where $|\alpha| = \sum_{l=1}^d \alpha_l$. When $d = 1$, the associated kernel $\mathcal{K}^{\langle X\rangle}(X_i, X_j) = 1 + (-1)^{m-1} k_{2m}(X_i - X_j)$, where $k_{2m}(x)$ is the $2m$-th order scaled Bernoulli polynomial (Abramowitz and Stegun, 1948). For $m = 2$, $k_4(x) = \frac{1}{24}((x - 0.5)^4 - 0.5(x - 0.5)^2 + \frac{7}{240})$ and the corresponding $\mathcal{K}^{\langle X\rangle}$ is known as the homogeneous cubic spline kernel. When $d > 2$, Novak et al. (2018) showed that the associated kernel is $\mathcal{K}^{\langle X\rangle}(X_i, X_j) = \int_{\mathbb{R}^d} [\prod_{l=1}^d \cos(2\pi(X_{il} - X_{jl})G_l)]/[1 + \sum_{0<|\alpha|\le m} \prod_{l=1}^d (2\pi G_l)^{2\alpha_l}] dG$ where $G \in \mathbb{R}^d$. An example for the discrete kernel is $\mathcal{K}(Z_i, Z_j) = \mathbb{1}_{\{Z_i = Z_j\}}$.

Let $\widehat{\eta}_{n,\lambda}$ be the penalized likelihood estimator of η under H_1, that is,

$$\widehat{\eta}_{n,\lambda} = \operatorname{argmin}_{\eta \in \mathcal{H}} \ell_{n,\lambda}(\eta). \tag{2.6}$$

Due to the integration in (2.4), the Representer Theorem (Wahba, 1990) does not apply here and the exact solution is not computable (Gu, 2013). We consider an efficient approximation in Gu (2013) by calculating the minimizer

of the penalized likelihood functional in $\mathcal{H}^\dagger = \text{span}\{\mathcal{K}(\mathbf{Y}_i, \cdot), i = 1, \ldots, n\}$. By the definition of $\mathcal{H}^\dagger$, the minimizer $\eta^\dagger(\cdot)$ of $\ell_{n,\lambda}(\eta)$ for $\eta^\dagger \in \mathcal{H}^\dagger$ has the form

$$\eta^\dagger(\cdot) = \sum_{i=1}^{n} \mathcal{K}(\mathbf{Y}_i, \cdot)c_i = \zeta^T \boldsymbol{c}, \quad \forall \eta^\dagger \in \mathcal{H}^\dagger \tag{2.7}$$

where $\zeta^T = (\mathcal{K}(\mathbf{Y}_1, \cdot), \cdots, \mathcal{K}(\mathbf{Y}_n, \cdot))$ is the vector of functions obtained from kernel $\mathcal{K}$ with its first argument fixed at $\mathbf{Y}_i$, and $\boldsymbol{c} = (c_1, \cdots, c_n)$ is the coefficient vector. Since $J(\eta)$ is $\langle \eta, \eta \rangle_{\mathcal{H}}$ where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is the inner product in $\mathcal{H}$ with reproducing kernel $\mathcal{K}$, we have $J(\eta^\dagger) = \boldsymbol{c}^T Q \boldsymbol{c}$ where $Q \in R^{n \times n}$ is the empirical kernel matrix with its (i, j)-th entry being $Q_{ij} = \mathcal{K}(\mathbf{Y}_i, \mathbf{Y}_j)$. This representation converts the infinite-dimensional minimization problem of (2.4) with respect to η to the finite-dimensional optimization problem with respect to the coefficient vector $\boldsymbol{c}$ by solving

$$\widehat{\boldsymbol{c}} = \underset{\boldsymbol{c}}{\text{argmin}} \left\{ -\frac{1}{n} \mathbf{1}_n^T Q \boldsymbol{c} + \int_{\mathcal{Y}} \exp\{\zeta^T \boldsymbol{c}\} dy + \frac{\lambda}{2} \boldsymbol{c}^T Q \boldsymbol{c} \right\}. \tag{2.8}$$

where $\mathbf{1}_n$ is an $n \times 1$ vector of ones, and the second term is the same as the second term in (2.4) with summation and integration over (x, z) replaced by integration over y for the convenience of presentation. The objective function in (2.8) is strictly convex (Tapia and Thompson, 1978). Its optimization with respect to $\boldsymbol{c}$ can be performed via a standard convex optimization procedure such as the Newton-Raphson algorithm; see, e.g., Gu (2013) and Wang (2011). The integrals in (2.8) can be calculated by numerical integration (see Section

7.4.2 in Gu (2013) for details). When n is large, the representation (2.7) involves a large number of coefficients, which may lead to numerical instability. To tackle this, one may consider only a subsample of $\{\mathbf{Y}_i : i = 1, \ldots, n\}$ to use in the presentation (Kim and Gu, 2004; Ma et al., 2015). For the nonparametric inference problem, subsampling method can maintain the minimax optimality through properly selected subsample size as shown in Liu et al. (2021). Practically, we follow the guide in Liu et al. (2021) to select the subsample sample size, which shows comparable power with the full data. In general, we denote by

$$\widehat{\eta}^{\dagger}_{n,\lambda} = \zeta^T \widehat{\boldsymbol{c}} \tag{2.9}$$

the penalized maximum likelihood estimate under the full model.

2.2 Penalized likelihood functional under the reduced model

Let $\widehat{\eta}^{0}_{n,\lambda}$ be the penalized likelihood estimator of η under H_0 in (2.3), that is,

$$\widehat{\eta}^{0}_{n,\lambda} = \text{argmin}_{\ \eta\in\mathcal{H}_0} \ell_{n,\lambda}(\eta). \tag{2.10}$$

In Section 3.1, we show that $\mathcal{H}_0$ is also an RKHS equipped with kernel function $\mathcal{K}^0(\cdot,\cdot)$, which enables us to use a similar reparameterization trick to solve the problem in (2.10). In the following, we show the kernel function $\mathcal{K}^0(\mathbf{Y}_i, \mathbf{Y}_j) = \mathcal{K}_0^{\langle X\rangle}(X_i, X_j)\mathcal{K}_0^{\langle Z\rangle}(Z_i, Z_j) + \mathcal{K}_1^{\langle X\rangle}(X_i, X_j)\mathcal{K}_0^{\langle X\rangle}(Z_i, Z_j) + \mathcal{K}_0^{\langle X\rangle}(X_i, X_j)\mathcal{K}_1^{\langle X\rangle}(Z_i, Z_j)$ where $\mathcal{K}_0^{\langle X\rangle}(X_i, X_j) = \mathbb{E}_X[\mathcal{K}^{\langle X\rangle}(X, X_j)] + \mathbb{E}_X[\mathcal{K}^{\langle X\rangle}(X_i, X)] - \mathbb{E}_{X,\widetilde{X}}\mathcal{K}^{\langle X\rangle}(X, \widetilde{X})$,

$\mathcal{K}_1^{\langle X\rangle} = \mathcal{K}^{\langle X\rangle} - \mathcal{K}_0^{\langle X\rangle}$, $\mathcal{K}_0^{\langle Z\rangle}(Z_i, Z_j) = \omega_{Z_i} + \omega_{Z_j} - \sum_{\ell=0}^{1} \omega_\ell^2$, $\mathcal{K}_1^{\langle Z\rangle} = \mathcal{K}^{\langle Z\rangle} - \mathcal{K}_1^{\langle Z\rangle}$, and $\omega_l = P(Z = l)$ for $l = 1, \ldots, U$. We plug the emprical estimate of $\widehat{\omega}_l = n_l/n$ for $l = 1, \ldots, U$ to calculate $\mathcal{K}^{\langle Z\rangle}$. The detailed derivation of $\mathcal{K}^0$ depends on our proposed probabilistic decomposition of $\mathcal{H}$, and is deferred to Section 3.1.

Similar to (2.7), we consider the efficient approximation in Gu (2013) by calculating the minimizer of the penalized likelihood functional in $\mathcal{H}^{0\dagger} = \text{span}\{\mathcal{K}^0(\mathbf{Y}_i, \cdot), i = 1, \ldots, n\}$, which has the form

$$\eta^{0\dagger}(\cdot) = \sum_{i=1}^{n} \mathcal{K}^0(\mathbf{Y}_i, \cdot) c_{0i} = \zeta_0^T \boldsymbol{c}_0, \quad \forall \eta^{0\dagger} \in \mathcal{H}^{0\dagger}. \tag{2.11}$$

To obtain the penalized likelihood estimators, we first solve the quadratic program

$$\widehat{\boldsymbol{c}}_0 = \underset{\boldsymbol{c}_0}{\text{argmin}} \left\{ -\frac{1}{n} \mathbf{1}_n^T Q_0 \boldsymbol{c}_0 + \int_{\mathcal{Y}} \exp\{\zeta_0^T \boldsymbol{c}_0\} + \frac{\lambda}{2} \boldsymbol{c}_0^T Q_0 \boldsymbol{c}_0 \right\} \tag{2.12}$$

where the (i, j)-th entry of Q_0 is $\mathcal{K}^0(\mathbf{Y}_i, \mathbf{Y}_j)$. Numerically, we could express

$$\begin{aligned} Q_0 &= [(I_n - H)Q^{\langle X\rangle}(I_n - H)] \circ [(I_n - H)Q^{\langle Z\rangle}(I_n - H)] \\ &+ [HQ^{\langle X\rangle}H] \circ [(I_n - H)Q^{\langle Z\rangle}(I_n - H)] + [(I_n - H)Q^{\langle X\rangle}(I_n - H)] \circ [HQ^{\langle Z\rangle}H] \end{aligned}$$

where $Q^{\langle X\rangle}$ is the empirical kernel matrix of $\mathcal{H}^{\langle X\rangle}$ with (i, j)-th entry $Q_{ij}^{\langle X\rangle} = \mathcal{K}^{\langle X\rangle}(X_i, X_j)$, $Q^{\langle Z\rangle}$ is the empirical kernel matrix of $\mathcal{H}^{\langle Z\rangle}$ with (i, j)-th entry $Q_{ij}^{\langle Z\rangle} = \mathcal{K}^{\langle Z\rangle}(Z_i, Z_j)$, and $H = I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$ with I_n being the $n \times n$ identity

matrix, $\mathbf{1}_n$ the $n \times 1$ vector of ones, and $\circ$ denotes the Hadamard product. Then we solve the quadratic optimization similar to (2.8) and output the function estimate

$$\widehat{\eta}^{0,\dagger}_{n,\lambda} = \zeta^{0T}\widehat{\boldsymbol{c}}^0. \tag{2.13}$$

2.3 Test statistics

Plugging the minimizers of the penalized likelihood functional under the full and reduced models into (2.5), we have the penalized likelihood ratio (PLR) statistic

$$PLR_{n,\lambda} = \ell_{n,\lambda}(\widehat{\eta}^0_{n,\lambda}) - \ell_{n,\lambda}(\widehat{\eta}_{n,\lambda}). \tag{2.14}$$

We will show in Section 3.2 that $PLR_{n,\lambda}$ is asymptotically χ^2 distributed under H_0 in the sense that $(2b_{n,\lambda})^{-1/2}(2PLR_{n,\lambda} - b_{n,\lambda}) \to N(0,1)$ with $b_{n,\lambda}$ diverges for a wide range of λ. Since $\widehat{\eta}_{n,\lambda}$ and $\widehat{\eta}^0_{n,\lambda}$ are not computable, we use their efficient approximations $\widehat{\eta}^{\dagger}_{n,\lambda}$ and $\widehat{\eta}^{0,\dagger}_{n,\lambda}$. Then an efficient approximation of the test statistic (2.14) is

$$PLR^{\dagger}_{n,\lambda} = \ell_{n,\lambda}(\widehat{\eta}^{0,\dagger}_{n,\lambda}) - \ell_{n,\lambda}(\widehat{\eta}^{\dagger}_{n,\lambda}).$$

We show that this efficient approximation has the same asymptotic distribution as $PLR_{n,\lambda}$. In practice, we use the gss package (Gu and Qiu, 1993) to obtatin the which implement the scalable computation via efficient approxi-

mation in Kim and Gu (2004) with compuation cost of the order $O(Nq^2)$ with $q = O(N^{2/(2m+1)})$ for the mth order Sobolev space.

For the nonparametric multi-sample test, the parameter space under H_0 is infinite-dimensional as $n \to \infty$. The assumptions of the Neyman-Pearson Lemma cannot be satisfied. Thus the uniformly most powerful test may not exist in general. We evaluate the power performance by the minimax rate of testing, which is defined as the minimal distance between the null and alternative hypotheses such that valid testing is possible (Ingster, 1989). For any generic 0-1 valued testing rule $\Phi = \Phi(\mathbf{Y}_1, \ldots, \mathbf{Y}_n)$ and a distinguishable rate $d_n > 0$ measuring the distance between the null and the alternative hypotheses, we define the *total error* $\mathrm{Err}(\Phi, d_n)$ of Φ under d_n as

$$\mathrm{Err}(\Phi, d_n) = \mathbb{E}_{H_0}\{\Phi\} + \sup_{\|\eta_{XZ}\|_2 \geq d_n} \mathbb{E}_{\eta}\{1 - \Phi\}, \tag{2.15}$$

where $\mathbb{E}_{H_0}\{\cdot\}$ denotes the expectation with respect to the truth η^* under H_0. The first and second terms on the right side of (2.15) represent type I and type II errors of Φ respectively. In Section 3, we show that the distinguishable rate of our proposed PLR test is related to the tuning parameter λ. We then derive the optimal distinguishable rate by carefully selecting λ. A data-adaptive tuning method is developed for practical use. In Section 4, we will use the information theory to establish the minimum distinguishable rate d_n for general testing rules, which extends the minimax testing principle pioneered

in Ingster (1989) to density comparison.

3. Theoretical Properties of PLR Test

In this section, we first introduce the probabilistic decomposition of a tensor product RKHS, enabling us to construct the kernel on the subspace $\mathcal{H}_0$. Such a decomposition is also of independent interest for studying different kinds of dependence among random variables. Compared with the function ANOVA decomposition in Wahba (1990) and Gu and Qiu (1993), the proposed probabilistic measure embedded decomposition makes the interaction term in (2.2) have zero expectation under null hypothesis which plays an essential role in deriving the limiting distribution of our test statistic. We then derive the asymptotic null distribution of our proposed test statistic and the optimal power of the test. Lastly, we develop a data-adaptive tuning procedure to choose the penalty parameter.

3.1 Probabilistic decomposition of the tensor product RKHS

We assume that the function $\eta(x, z)$ belongs to a tensor product RKHS $\mathcal{H} = \mathcal{H}^{\langle X\rangle} \otimes \mathcal{H}^{\langle Z\rangle}$, in which $\mathcal{H}^{\langle X\rangle}$ and $\mathcal{H}^{\langle Z\rangle}$ represent the marginal RKHS of X and Z respectively. We aim to decompose $\mathcal{H}$ into orthogonal subspaces with a hierarchical structure similar to the main effects and interactions in smooth-

ing spline ANOVA (Wahba, 1990; Gu, 2013; Lin, 2000; Wang, 2011), while embedding the probabilistic distributions of X and Z into the decomposition. Such decomposition enables us to convert the multi-sample test problem into testing the presence of the interaction. It includes two steps: decompose each marginal RKHS into mean and main effects; apply the distributive law to expand the tensor product of marginal RKHS into a series of subspaces.

We first introduce the probabilistic tensor decomposition of the discrete domain function space $\mathcal{H}^{\langle Z\rangle} := \{f(z) : z \in \{1,\ldots,U\}\}$ via a probabilistic averaging operator. Note that $\mathcal{H}^{\langle Z\rangle} = \mathbb{R}^U$ with the Euclidean inner product $(\langle\cdot,\cdot\rangle_2)$ and the kernel on $\mathcal{H}^{\langle Z\rangle}$ is $\mathcal{K}^{\langle Z\rangle}(z,\widetilde{z}) = \mathbb{1}_{\{z=\widetilde{z}\}}$. Consider a discrete probabilistic measure $\mathbb{P}_Z$ on $\mathcal{Z} = \{1,\ldots,U\}$ such that $\mathbb{P}_Z(Z=j) = \omega_j \geq 0$ with $\sum_{j=1}^U \omega_j = 1$. Let $\boldsymbol{\omega} = (\omega_1,\ldots,\omega_U)$, and define the probabilistic averaging operator as $\mathcal{A}_Z := f \to \mathbb{E}_Z f(Z) = \langle \boldsymbol{\omega}, f\rangle_{\mathcal{H}^{\langle z\rangle}}$. Since $\mathbb{E}_Z[\mathcal{K}_Z^{\langle Z\rangle}] = \boldsymbol{\omega}$, we can rewrite the probabilistic averaging operator as $\mathcal{A}_Z := f \to \mathbb{E}_Z f(Z) = \langle \mathbb{E}_Z[\mathcal{K}_Z^{\langle Z\rangle}], f\rangle_2$. Then $\mathbb{E}_Z[\mathcal{K}_Z^{\langle Z\rangle}]$ can be treated as a mean embedding of $\mathbb{P}_Z$ in $\mathcal{H}^{\langle Z\rangle}$. We further define the tensor sum decomposition of $\mathcal{H}^{\langle Z\rangle}$ as

$$\mathcal{H}^{\langle Z\rangle} = \mathcal{H}_0^{\langle Z\rangle} \oplus \mathcal{H}_1^{\langle Z\rangle} := span\{\mathbb{E}_Z \mathcal{K}_Z^{\langle Z\rangle}\} \oplus \{f \in \mathcal{H} : \mathbb{E}_Z\{f(Z)\} = 0\}, \quad (3.1)$$

where $\mathcal{H}_0^{\langle Z\rangle}$ is the grand mean space, $\mathcal{H}_1^{\langle Z\rangle}$ is the main effect space. Each subspace in (3.1) is an RKHS with their corresponding kernels stated in Lemma S.1 in Supplimentary. For fixed design of Z, we set $\omega_j = n_j / \sum_{j=1}^U n_j$.

Next, let us consider the continuous random variable $X \in \mathcal{X}$ and a probability measure $\mathbb{P}_X$ on $\mathcal{X}$. We suppose $\mathcal{H}^{\langle X \rangle}$ is the mth order Sobolev space with the corresponding inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}^{\langle X \rangle}}$. The results also hold for its homogeneous subspace. Let $\mathcal{K}^{\langle X \rangle}$ be the corresponding kernel satisfying $\langle f, \mathcal{K}_x^{\langle X \rangle} \rangle_{\mathcal{H}^{\langle X \rangle}} = f(x)$ for any $f \in \mathcal{H}^{\langle X \rangle}$. Similarly, the probabilistic averaging operator is $\mathcal{A}_X := f \to \mathbb{E}_X f(X) = \mathbb{E}_X \langle \mathcal{K}_X^{\langle X \rangle}, f \rangle_{\mathcal{H}^{\langle X \rangle}} = \langle \mathbb{E}_X \mathcal{K}_X^{\langle X \rangle}, f \rangle_{\mathcal{H}^{\langle X \rangle}}$. $\mathbb{E}_X \mathcal{K}_X^{\langle X \rangle}$ has the same role as $\boldsymbol{\omega}$ in the Euclidean space. Then, the tensor sum decomposition of a functional space is defined as

$$\mathcal{H}^{\langle X \rangle} = \mathcal{H}_0^{\langle X \rangle} \oplus \mathcal{H}_1^{\langle X \rangle} := span\{\mathbb{E}_X \mathcal{K}_X^{\langle X \rangle}\} \oplus \{f \in \mathcal{H}^{\langle X \rangle} : \mathcal{A}_X f = 0\}. \tag{3.2}$$

Analogously, we name $\mathcal{H}_0^{\langle X \rangle}$ as the grand mean space and $\mathcal{H}_1^{\langle X \rangle}$ as the main effect space. $\mathbb{E}_X \mathcal{K}_X^{\langle X \rangle}$ is known as the kernel mean embedding which is well established in the statistics literature (Berlinet and Thomas-Agnan, 2011). The construction of the kernel functions for $\mathcal{H}_0^{\langle X \rangle}$ and $\mathcal{H}_1^{\langle X \rangle}$ are included in Lemma S.2 in Supplementary.

We are now ready to consider the RKHS $\mathcal{H} = \mathcal{H}^{\langle X \rangle} \otimes \mathcal{H}^{\langle Z \rangle}$ on the product domain $\mathcal{Y} = \mathcal{X} \times \mathcal{Z}$. Applying the distributive rule, the decomposition of $\mathcal{H}$ is written as

$$\mathcal{H} = (\mathcal{H}_0^{\langle X \rangle} \oplus \mathcal{H}_1^{\langle X \rangle}) \otimes (\mathcal{H}_0^{\langle Z \rangle} \oplus \mathcal{H}_1^{\langle Z \rangle}) \equiv \mathcal{H}_{00} \oplus \mathcal{H}_{10} \oplus \mathcal{H}_{01} \oplus \mathcal{H}_{11}, \tag{3.3}$$

where $\mathcal{H}_{ij} = \mathcal{H}_i^{\langle X \rangle} \otimes \mathcal{H}_j^{\langle Z \rangle}$ for $i = 0, 1$ and $j = 0, 1$. Analogous to the classic

ANOVA, $\mathcal{H}_{10}$ and $\mathcal{H}_{01}$ are the RKHS's for the main effects, and $\mathcal{H}_{11}$ is the RKHS for the interaction. We call the decomposition of $\mathcal{H}$ in (3.3) as the *probabilistic decomposition* of the tensor product RKHS $\mathcal{H}$ since it embeds the probability measure of the random variable X and Z. Based on Theorems 2.6 in Gu (2013), we can construct the kernels $\mathcal{K}^{00}, \mathcal{K}^{10}, \mathcal{K}^{01}$ and $\mathcal{K}^{11}$ for the subspaces $\mathcal{H}_{00}, \mathcal{H}_{10}, \mathcal{H}_{01}$ and $\mathcal{H}_{11}$ accordingly; see Lemma S.3 in supplimentary for detailed construction.

3.2 Asymptotic distribution and Wilks' phenomenon

In this section, we present the asymptotic distribution of our PLR test statistic in Theorem 3.1. The proof relies on a technical lemma about the eigenstructures of $\mathcal{H}_0$ and $\mathcal{H}$; see Lemma 1 below. For any $\eta, \widetilde{\eta} \in \mathcal{H}$, define

$$\langle \eta, \widetilde{\eta} \rangle = V(\eta, \widetilde{\eta}) + \lambda J(\eta, \widetilde{\eta}), \tag{3.4}$$

where $V(\eta, \widetilde{\eta}) = \mathbb{E}_{\eta^*}\{\eta(\mathbf{Y})\widetilde{\eta}(\mathbf{Y})\}$ with the expectation taken under the true η^*, and J is a bilinear form corresponding to (2.4). It holds that $\mathcal{H}$ and $\mathcal{H}_0$, endowed with the inner product (3.4), are both RKHSs; see Lemma 2. In the following lemma, we characterize the eigenvalues and eigenvectors of the Rayleigh quotient V/J.

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

$V(\xi_p, \xi_{p'}) = \delta_{p,p'}$, $J(\xi_p, \xi_{p'}) = \rho_p \delta_{p,p'}$, *for all* $p, p' \geq 1$, *and that any* $\eta \in \mathcal{H}$ *can be written as* $\eta = \sum_{p=1}^{\infty} V(\eta, \xi_p)\xi_p$.

(b) Moreover, there exists a proper subset $\{\rho_p^0, \xi_p^0\}_{p=1}^{\infty}$ *of* $\{\rho_p, \xi_p\}_{p=1}^{\infty}$ *satisfying* $\{\xi_p^0\}_{p=1}^{\infty} \subset \mathcal{H}_0$ *and for any* $\eta \in \mathcal{H}_0$, $\eta = \sum_{p=1}^{\infty} V(\eta, \xi_p^0)\xi_p^0$. *Convergence of both series holds under (3.4).*

(c) $\rho_p^{\perp} \asymp p^{2m/d}$, *where* $\{\rho_p^{\perp}\}_{p=1}^{\infty} \subset \{\rho_p\}_{p=1}^{\infty}$ *is a subset of eigenvalues corresponding to* $\{\xi_p^{\perp}\}_{p=1}^{\infty} \equiv \{\xi_p\}_{p=1}^{\infty} \backslash \{\xi_p^0\}_{p=1}^{\infty}$. *The set* $\{\xi_p^{\perp}\}_{p=1}^{\infty}$ *generates the orthogonal complement of* $\mathcal{H}_0$ *under the inner product (3.4).*

Lemma 1 introduces an eigensystem that simultaneously diagonalizes the bilinear forms V and J. This eigensystem does not depend on the unknown null density, but only depends on the functional space $\mathcal{H}$. Moreover, $\mathcal{H}_0$ 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 $\mathcal{H}_0$. Specifically, for any $\eta, \widetilde{\eta} \in \mathcal{H}_0$, $\langle \eta, \widetilde{\eta} \rangle_0 = \langle \eta, \widetilde{\eta} \rangle$. Then $\mathcal{H}$ and $\mathcal{H}_0$ are both RKHS's endowed with these inner products.

Lemma 2. $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ *and* $(\mathcal{H}_0, \langle \cdot, \cdot \rangle_0)$ *are both RKHS's with the corresponding inner products.*

Following Lemma 2, there exist reproducing kernel functions $\widetilde{\mathcal{K}}(\cdot, \cdot)$ and

$\widetilde{\mathcal{K}}^0(\cdot,\cdot)$ defined on $\mathcal{Y}\times\mathcal{Y}$ satisfying, for any $\mathbf{y}\in\mathcal{Y}$, $\eta\in\mathcal{H}$, $\widetilde{\eta}\in\mathcal{H}_0$:

$$\widetilde{\mathcal{K}}_{\mathbf{y}}(\cdot)\equiv\widetilde{\mathcal{K}}(\mathbf{y},\cdot)\in\mathcal{H},\quad \widetilde{\mathcal{K}}^0_{\mathbf{y}}(\cdot)\equiv\widetilde{\mathcal{K}}^0(\mathbf{y},\cdot)\in\mathcal{H}_0,$$
$$\langle\widetilde{\mathcal{K}}_{\mathbf{y}},\eta\rangle=\eta(\mathbf{y}),\quad \langle\widetilde{\mathcal{K}}^0_{\mathbf{y}},\widetilde{\eta}\rangle_0=\widetilde{\eta}(\mathbf{y}). \tag{3.5}$$

We further introduce positive definite self-adjoint operators $W_\lambda:\mathcal{H}\to\mathcal{H}$ and $W^0_\lambda:\mathcal{H}_0\to\mathcal{H}_0$ such that

$$\langle W_\lambda\eta,\widetilde{\eta}\rangle=\lambda J(\eta,\widetilde{\eta})\text{ for all }\eta,\widetilde{\eta}\in\mathcal{H},$$
$$\langle W^0_\lambda\eta,\widetilde{\eta}\rangle_0=\lambda J_0(\eta,\widetilde{\eta})\text{ for all }\eta,\widetilde{\eta}\in\mathcal{H}_0, \tag{3.6}$$

where $J_0(\eta,\widetilde{\eta})=\theta_{01}^{-1}J_{01}(\eta,\widetilde{\eta})+\theta_{10}^{-1}J_{10}(\eta,\widetilde{\eta})$ is the restriction of J over $\mathcal{H}_0$. By (3.6) we get $\langle\eta,\widetilde{\eta}\rangle=V(\eta,\widetilde{\eta})+\langle W_\lambda\eta,\widetilde{\eta}\rangle$, $\langle\eta,\widetilde{\eta}\rangle_0=V(\eta,\widetilde{\eta})+\langle W^0_\lambda\eta,\widetilde{\eta}\rangle_0$. In the following, we give the explicit expression of $\widetilde{\mathcal{K}}_y(\cdot)$ and $W_\lambda\xi_p(\cdot)$.

Proposition 1. *For any $\mathbf{y}\in\mathcal{Y}$ and $\eta\in\mathcal{H}$, we have*

$$\|\eta\|^2=\sum_{p=1}^{\infty}|V(\eta,\xi_p)|^2(1+\lambda\rho_p),$$
$$\widetilde{\mathcal{K}}_{\mathbf{y}}(\cdot)=\sum_{p=1}^{\infty}\frac{\xi_p(\mathbf{y})}{1+\lambda\rho_p}\xi_p(\cdot),\quad \widetilde{\mathcal{K}}^0_{\mathbf{y}}(\cdot)=\sum_{p=1}^{\infty}\frac{\xi^0_p(\mathbf{y})}{1+\lambda\rho^0_p}\xi^0_p(\cdot),$$
$$W_\lambda\xi_p(\cdot)=\frac{\lambda\rho_p}{1+\lambda\rho_p}\xi_p(\cdot),\quad W^0_\lambda\xi^0_p(\cdot)=\frac{\lambda\rho^0_p}{1+\lambda\rho^0_p}\xi^0_p(\cdot).$$

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

As shown in Proposition 1, the eigenvalues for $\widetilde{\mathcal{K}}$ are $\{(1+\lambda\rho_p)^{-1}\}_{p=1}^{\infty}$, having a slower decay rate than the decay rate of eigenvalues for $\mathcal{K}$ due to

the scaling by λ. In particular, $\widetilde{\mathcal{K}}$ can be viewed as a scaled kernel comparing with the product kernel $\mathcal{K}^{\mathcal{H}} = \mathcal{K}^{00} + \mathcal{K}^{01} + \mathcal{K}^{10} + \mathcal{K}^{11}$ introduced in Lemma S.3 in supplimentary. Note that $\text{trace}(\widetilde{\mathcal{K}}) = \sum_{p=1}^{\infty}(1+\lambda\rho_p)^{-1} \asymp \lambda^{-d/(2m)}$ is the effective dimension that measures the complexity of $\mathcal{H}$; see Bartlett et al. (2005); Mendelson (2002).

Next, we will derive the null asymptotic distribution of the PLR statistics, which relies on the Taylor expansion of the PLR functional. First, we introduce the Frechét derivatives of the log-likelihood functional. Denote by D, D^2, D^3 the first-, second- and third-order Frechét derivatives of $\ell_{n,\lambda}(\eta)$. Let $S_{n,\lambda}(\eta)$ and $S^0_{n,\lambda}$ be respectively the score functions of the log-likelihood functionals $\ell_{n,\lambda}$ and $\ell^0_{n,\lambda}$. Define $\mathbf{y} = (x, z)$. Then these derivatives can be summarized as follows.

For any $\eta, \Delta\eta_1, \Delta\eta_2, \Delta\eta_3 \in \mathcal{H}$,

$$\begin{aligned} D\ell_{n,\lambda}(\eta)\Delta\eta_1 &= -\frac{1}{n}\sum_{i=1}^{n}\Delta\eta_1(\mathbf{Y}_i) + \int_{\mathcal{Y}}\Delta\eta_1(\mathbf{y})e^{\eta(\mathbf{y})}d\mathbf{y} + \lambda J(\eta, \Delta\eta_1) \\ &= \langle -\frac{1}{n}\sum_{i=1}^{n}\widetilde{\mathcal{K}}_{\mathbf{Y}_i} + \mathbb{E}_\eta\widetilde{\mathcal{K}}_{\mathbf{Y}} + W_\lambda\eta, \Delta\eta_1\rangle \\ &\equiv \langle S_{n,\lambda}(\eta), \Delta\eta_1\rangle, \end{aligned} \tag{3.7}$$

$$D^2\ell_{n,\lambda}(\eta)\Delta\eta_1\Delta\eta_2 = \int_{\mathcal{Y}}\Delta\eta_1(\mathbf{y})\Delta\eta_2(\mathbf{y})e^{\eta(\mathbf{y})}d\mathbf{y} + \lambda J(\Delta\eta_1, \Delta\eta_2), \tag{3.8}$$

$$D^3\ell_{n,\lambda}(\eta)\Delta\eta_1\Delta\eta_2\Delta\eta_3 = \int_{\mathcal{Y}}\Delta\eta_1(\mathbf{y})\Delta\eta_2(\mathbf{y})\Delta\eta_3(\mathbf{y})e^{\eta(\mathbf{y})}d\mathbf{y}. \tag{3.9}$$

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

$$\int_{\mathcal{Y}} \Delta\eta(\mathbf{y}) e^{\eta(\mathbf{y})} d\mathbf{y} = \mathbb{E}_\eta \Delta\eta_1(\mathbf{Y}) = \mathbb{E}_\eta \langle \widetilde{\mathcal{K}}_{\mathbf{Y}}, \Delta\eta_1 \rangle = \langle \mathbb{E}_\eta \widetilde{\mathcal{K}}_{\mathbf{Y}}, \Delta\eta_1 \rangle.$$

The Taylor expansion of the PLR functional gives

$$\begin{aligned} PLR_{n,\lambda} &= \ell_{n,\lambda}(\widehat{\eta}^0_{n,\lambda}) - \ell_{n,\lambda}(\widehat{\eta}_{n,\lambda}) \\ &= D\ell_{n,\lambda}(\widehat{\eta}_{n,\lambda})g + \int_0^1 \int_0^1 sD^2\ell_{n,\lambda}(\widehat{\eta}_{n,\lambda} + ss'g)ggdsds' \\ &= \int_0^1 \int_0^1 s\{D^2\ell_{n,\lambda}(\widehat{\eta}_{n,\lambda} + ss'g)gg - D^2\ell_{n,\lambda}(\eta^*)gg\}dsds' + \frac{1}{2}D^2\ell_{n,\lambda}(\eta^*)gg \\ &\equiv I_1 + I_2 \end{aligned} \tag{3.10}$$

where $g = \widehat{\eta}^0_{n,\lambda} - \widehat{\eta}_{n,\lambda}$ and η^* is the underlying truth. In the proof of Theorem 3.1, we will show that I_2 is a leading term compared with I_1. From (3.8), we have that $I_2 = \frac{1}{2}\|g\|^2 = \frac{1}{2}\|\widehat{\eta}^0_{n,\lambda} - \widehat{\eta}_{n,\lambda}\|^2$. As we will see, the asymptotic distribution of $\|\widehat{\eta}_{n,\lambda} - \widehat{\eta}^0_{n,\lambda}\|^2$ relies on the Bahadur representations of $\widehat{\eta}^0_{n,\lambda}$ and $\widehat{\eta}_{n,\lambda}$.

We further prove the following Bahadur representations for the difference of the two penalized likelihood estimators, by adapting an empirical processes technique in Shang and Cheng (2013). Lemma 3 is crucial for proving Theorem 3.1.

Lemma 3. *Suppose* $h = \lambda^{\frac{d}{2m}}$ *and* $nh^2 \to \infty$. *Then we have*

$$n^{1/2}\|\widehat{\eta}_{n,\lambda} - \widehat{\eta}^0_{n,\lambda}\| = n^{1/2}\|S^0_{n,\lambda}(\eta^*) - S_{n,\lambda}(\eta^*)\| + o_P(1).$$

where $S_{n,\lambda}(\eta^*)$ *and* $S^0_{n,\lambda}(\eta^*)$ *are the score functions for* $\ell_{n,\lambda}$ *and* $\ell^0_{n,\lambda}$*, respectively.*

This lemma shows that the main term I_2 in Taylor's expansion of the PLR functional is determined by the norm of the difference between the score function of $\ell_{n,\lambda}$ and the score function of $\ell^0_{n,\lambda}$. Since the score functions have explicit expressions through Proposition 1, we can characterize the asymptotic null distribution of I_2 by the eigensystem introduced in Lemma 1.

Before stating our main theorem, we introduce an assumption commonly used in literature for deriving the rates of density estimates; see, e.g., Theorem 9.3 of Gu (2013).

Assumption 1. There exists a convex set $B \subset \mathcal{H}$ around η^* and a constant $c_1 > 0$ such that, for any $\eta \in B$, $c\mathbb{E}_{\eta^*}\{\widetilde{\eta}^2(\mathbf{Y})\} \leq \mathbb{E}_{\eta}\{\widetilde{\eta}^2(\mathbf{Y})\}$. Furthermore, with probability approaching one, $\widehat{\eta}_{n,\lambda} \in B$; and under H_0, with probability approaching one, $\widehat{\eta}^0_{n,\lambda} \in B$.

This condition is satisfied when $\widehat{\eta}_{n,\lambda}$ and $\widehat{\eta}^0_{n,\lambda}$ are stochastically bounded and the members of B have uniform upper and lower bounds on the domain $\mathcal{Y}$. The following theorem provides the asymptotic distribution for the PLR test statistic under Assumption 1. The proof of Theorem 3.1 and Corollary 3.1.1 are in Supplimentary S.6.3.

Theorem 3.1. *Suppose $m \geq 1$ and Assumption 1 holds. Let $h = \lambda^{\frac{d}{2m}}$ and $nh^{2m+d} = O(1)$, $nh^2 \to \infty$ as $n \to \infty$. Under H_0, we have*

$$\frac{2n \cdot PLR_{n,\lambda} - \theta_\lambda}{\sqrt{2}\sigma_\lambda} \xrightarrow{d} N(0,1), \ n \to \infty, \tag{3.11}$$

where $\theta_\lambda = \sum_{p=1}^{\infty} \frac{1}{1+\lambda\rho_p^\perp}$, $\sigma_\lambda^2 = \sum_{p=1}^{\infty} \frac{1}{(1+\lambda\rho_p^\perp)^2}$.

We notice that $h \asymp n^{-c}$ with $\frac{1}{2m+d} \leq c \leq \frac{1}{2}$ satisfying the rate conditions in Theorem 3.1, so the asymptotic distribution (3.11) holds under a wide range of choices of h. The quantities θ_λ and σ_λ solely depend on the eigenvalues $\rho_p^\perp$'s and λ. Based on (3.11), we propose the following decision rule $\Phi_{n,\lambda}$ at the significance level α:

$$\Phi_{n,\lambda}(\alpha) = \mathbb{1}(|2n \cdot PLR_{n,\lambda} - \theta_\lambda| \geq z_{1-\alpha/2}\sqrt{2}\sigma_\lambda) \tag{3.12}$$

where $\mathbb{1}(\cdot)$ 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 α if $\Phi_{n,\lambda} = 1$. Wilks' phenomenon is also observed here similar to the nonparametric/semiparametric regression framework (Fan et al., 2001; Shang and Cheng, 2013). Specifically, let $r_\lambda = \frac{\theta_\lambda}{\sigma_\lambda^2}$, then (3.11) implies that, as $n \to \infty$,

$$\frac{2nr_\lambda \cdot PLR_{n,\lambda} - r_\lambda\theta_\lambda}{\sqrt{2r_\lambda\theta_\lambda}} \xrightarrow{d} N(0,1).$$

Therefore, $2nr_\lambda \cdot PLR_{n,\lambda}$ is asymptotically distributed as a χ^2 distribution with degrees of freedom $r_\lambda\theta_\lambda$. In the following corollary, we extend our asymptotic theory to the emiprical version of $\rho_p^\perp$'s.

Corollary 3.1.1. Assume that Assumption 1 holds. Let $h = \lambda^{\frac{d}{2m}}$ and $nh^{2m+d} = O(1)$, $nh^2 \to \infty$ as $n \to \infty$. Under H_0, we have

$$\frac{2n \cdot PLR^{\dagger}_{n,\lambda} - \theta_\lambda}{\sqrt{2}\sigma_\lambda} \xrightarrow{d} N(0,1), \ n \to \infty, \tag{3.13}$$

where $\widehat{\theta}_\lambda = \sum_{p=1}^{n} \frac{1}{1+\lambda\widehat{\rho}_p^{\perp}}$, $\widehat{\sigma}_\lambda^2 = \sum_{p=1}^{n} \frac{1}{(1+\lambda\widehat{\rho}_p^{\perp})^2}$, $\{\widehat{\rho}_p^{\perp}\}_{p=1}^{n}$ are empirical eigenvalues for $\mathcal{K}^{11}$.

In Corollary 3.1.1, we show the asymptotic distribution of the efficient approximation $PLR^{\dagger}_{n,\lambda}$. The proof of Corollary 3.1.1 is based on the local Radamacher complexity (Liu et al., 2021; Bartlett et al., 2005) to bound the tail sum of eigenvalues for $\mathcal{H}^{\dagger}$ and $\mathcal{H}^{0\dagger}$, and is also based on the accurate error bound for the eigenvalues of the kernel matrix in Braun (2006).

3.3 Power analysis and minimaxity

In this section, we investigate the power of PLR under local alternatives. Define the distinguishable rate as

$$d_n := \sqrt{\lambda + \sigma_\lambda/n}. \tag{3.14}$$

The distinguishable rate is used to measure the distance between the null and alternative hypotheses. Theorem 3.2 shows that the power of PLR approaches one, provided that the norm of η^*_{XZ}, the interaction term in the probabilistic

decomposition of η^*, has a norm bounded below by d_n. The squared distinguishable rate d_n^2 consists of two components: λ representing the squared bias of the estimator, and σ_λ/n with the order of $n^{-1}h^{-1/2}$ representing the standard derivation of $PLR_{n,\lambda}$. Since σ_λ decreases with λ, the minimal distinguishable rate for the PLR test is achieved by choosing an appropriate λ such that $\lambda \asymp \sigma_\lambda/n$. Our result owes much to the analytic expression of independence (in terms of interactions) based on the proposed probabilistic tensor product decomposition framework.

Let P_{η^*} denote the probability measure induced under η^*, $\|\eta\|_{\sup}$ the supremum norm over $\mathcal{Y}$, and $\|\eta\|_2 = \sqrt{V(\eta)}$.

Theorem 3.2. *Suppose Assumption 1 holds and let d_n be the distinguishable rate defined in (3.14), $m > 3/2$, $\eta^* \in \mathcal{H}$ with $\|\eta^*_{XZ}\|_{\sup} = o(1)$, $J(\eta^*_{XZ}) < \infty$, $\|\eta^*_{XZ}\|_2 \gtrsim d_n$. For any $\varepsilon \in (0,1)$, there exists a positive N_ε such that, for any $n \geq N_\varepsilon$, $\mathbb{P}_{\eta^*}(\Phi_{n,\lambda}(\alpha) = 1) \geq 1 - \varepsilon$. When $\lambda \asymp \lambda^* \equiv n^{-4m/(4m+d)}$, d_n is upper bounded by $d_n^* \equiv n^{-2m/(4m+d)}$.*

The proof of Theorem 3.2 is in Supplimentary S.6.3. Theorem 3.2 demonstrates that, when $\lambda \asymp \lambda^*$, PLR can successfully detect any local alternatives, provided that they separate from the null by at least d_n^*. In Section 4, we show that this upper bound is unimprovable by establishing the minimax lower bound for the distinguishable rate of a general multi-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^*. Therefore, we claim that our PLR test is minimax optimal.

For any $\varepsilon \in (0,1)$ and $\alpha \in (0,\varepsilon)$, Theorem 3.1 shows that $\mathbb{E}_{H_0}\{\Phi_{n,\lambda^*}(\alpha)\}$ tends to α; Theorem 3.2 shows that $\mathbb{E}_{\eta^*}\{1-\Phi_{n,\lambda^*}(\alpha)\} \le \varepsilon - \alpha$, provided that $\|\eta_{XZ}^*\|_2 \ge C_{\varepsilon-\alpha} d_n^*$ for a large constant $C_{\varepsilon-\alpha}$. That means, asymptotically,

$$\mathrm{Err}(\Phi_{n,\lambda^*}(\alpha), C_{\varepsilon-\alpha} d_n^*) \le \varepsilon. \tag{3.15}$$

In other words, the total error of PLR can be controlled by an arbitrary ε 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 distinguishable rate $d_n^\diamond(\varepsilon)$ as

$$d_n^\diamond(\varepsilon) = \inf\{d_n > 0 : \inf_{\Phi} \mathrm{Err}(\Phi, d_n) \le \varepsilon\}, \tag{4.1}$$

where the infimum in (4.1) is taken over all 0-1 valued testing rules based on the sample $\mathbf{Y}_i$'s. Note that $d_n^\diamond(\varepsilon)$ characterizes the smallest separation between the null and local alternatives such that there exists a testing approach with a total error of at most ε. Next we establish a lower bound for $d_n^\diamond$, i.e., if d_n is smaller than a certain lower bound, there exists no test that can distinguish the alternative from the null.

We first introduce a geometric interpretation of the hypothesis testing (2.3). Here we consider the local alternatives residing in $\mathcal{E} = \{\eta \in \mathcal{H} : \|\eta\|_{\mathcal{H}} < 1/2\}$. Geometrically, $\mathcal{E}$ is an ellipsoid with axis lengths equal to eigenvalues of $\mathcal{H}$. For any $\eta \in \mathcal{E}$, the projection of η on $\mathcal{E}_{11} := \mathcal{H}_{11} \cap \mathcal{E}$ is η_{XZ} where $\mathcal{H}_{11}$ is defined in (3.3). The magnitude of the interaction η_{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 $\mathcal{E}_{11}$.

Intuitively, the testing will be harder when the projection of η 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

$$b_{k,2}(C) := \operatorname*{argmax}_{r \geq 0} \{\mathbb{B}_2^{k+1}(r) \subset C \cap S \text{ for some subspace } S \in S_{k+1}\} \qquad (4.2)$$

where S_{k+1} denotes the set of all $k+1$ dimensional subspace, and $\mathbb{B}_2^{k+1}(r)$ is the $(k+1)$-dimensional L_2-ball with radius r and center 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 η projected in the ball with radius less than this bound, the total error is larger than $1/2$.

Lemma 4. *For any $\eta \in \mathcal{H}$, we have $Err(\Phi, d_n) \geq 1/2$ for all $d_n \ll r_B(\delta^*) := \sup\{\delta \,|\, \delta \leq \frac{1}{2\sqrt{n}}(k_B(\delta))^{1/4}\}$, where $k_B(\delta) := \operatorname{argmax}_k\{b_{k-1,2}^2(\mathcal{H}_{11}) \geq \delta^2\}$ is the*

Bernstein lower critical dimension and $r_B(\delta^*)$ *is called the Bernstein lower critical radius.*

In Lemma 4, we show that when d_n is less than $r_B(\delta^*)$, there is no test that can distinguish the alternative from the null. 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 $\mathcal{H}_{11}$. 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 a k-dimensional ball that can be embedded into $\mathcal{H}_{11}$ is determined by the kth largest eigenvalue. Lemma 5 characterizes the lower bound of $k_B(\delta)$ by the largest k such that the kth largest eigenvalue is larger than δ^2.

Lemma 5. *Let* γ_k *be the* k*th largest eigenvalue of* $\mathcal{H}_{11}$*. Then we have*

$$k_B(\delta) > \operatorname*{argmax}_k\{\sqrt{\gamma_k} \geq \delta\} \tag{4.3}$$

Note that $\gamma_k \asymp k^{-2m/d}$, then $\operatorname{argmax}_k\{\sqrt{\gamma_k} \geq \delta\} \asymp \delta^{-d/m}$. Plug in the lower bound of $k_B(\delta)$ to Lemma 4, we achieve $r_B(\delta^*)$, which is the minimax lower bound for the distinguishable rate in the following theorem.

Theorem 4.1. *Suppose* $\eta \in \mathcal{H}$*. For any* $\varepsilon \in (0,1)$*, the minimax distinguishable rate for the testing hypotheses (2.3) is* $d_n^\diamond(\varepsilon) \gtrsim n^{-2m/(4m+d)}$.

Theorem 4.1 provides a general guidance to justify a local minimax test for

testing $\eta_{XZ} = 0$. The proof of Theorem 4.1 is presented in the Supplimentary S.6.4. Comparing $d_n^\circ(\varepsilon)$ with d_n^* derived in Theorem 3.2, we see that the PLR test is minimax optimal.

5. Simulation Studies

In this section, we demonstrate the finite sample performance of the proposed test alongside its competitors through simulation studies. We choose the Kolmogorov-Smirnov (K-S) test and Anderson-Darling (AD) test as two representatives of the most popular CDF-based tests, the normalized MMD test (Li and Yuan, 2019) as a representative of kernel-based tests, the empirical likelihood tests (ELT) (Cao and Van Keilegom, 2006) and kernel density test (KDT) (Zhan and Hart, 2014) as representatives of density-based tests, and the dynamic slicing test (DSLICE) (Jiang et al., 2015) as a representative of discretization-based tests. We use the function *ad.test()* provided in the *kSamples* R package for the AD test, conduct the MMD test using the *dHSIC* R package with the default Gaussian kernel, use *dslice* R package for DSLICE test, and implement the ELT and KDT test using the code provided by the authors. For our proposed PLR test, we choose the roughness parameter based on the data-adaptive tuning parameter selection criteria in Section S.1 in supplimentary. Also, we have additional simulation studies for Beta,

Beta Mixtures, multivariate distribution ($d > 2$) and multiple distributions ($U > 2$) in Supplimentary S.4.

The samples $\mathbf{Y}_i = (X_i, Z_i)$, $i = 1, \ldots, n$, were generated as follows. We first generated $Z_i \overset{iid}{\sim} \text{Bernoulli}(0.5)$, 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. In each setting, we chose the averaged sample size n in each group as 125, 250, 375, 500, 625, 750, 875, 1000. Size and power were calculated as the proportions of rejection based on 1000 independent trials.

Setting 1: Gaussian distributions with mean zero and a group-specific variance: $X \mid Z = z \sim N\left(0, (1 + \delta_1 \mathbb{1}_{z=1})^2\right)$ where $\delta_1 = 0, 0.2, 0.3$.

Setting 2: Uni-modal Gaussian distribution versus bi-modal Gaussian distribution: $X \mid Z = z \sim 0.5N\left(-\delta_2 \mathbb{1}_{z=1}, (1 + \delta_2^2 \mathbb{1}_{z=0})\right) + 0.5N\left(\delta_2 \mathbb{1}_{z=1}, (1 + \delta_2^2 \mathbb{1}_{z=0})\right)$ where we set $\delta_2 = 0, 1, 1.2$.

Setting 3: Asymmetric mixture Gaussian distributions: $X \mid Z = z \sim 0.5N(2, 1) + 0.5N(-2, (1 - \delta_3 \mathbb{1}_{z=1})^2)$ where $\delta_3 = 0, 0.3, 0.45$.

Setting 4: Symmetric mixture distributions: $X \mid Z = z \sim 0.5N(2, (1 - \delta_4 \mathbb{1}_{z=1})^2) + 0.5N(-2, (1 - \delta_4 \mathbb{1}_{z=1})^2)$ where $\delta_4 = 0, 0.3, 0.6$.

Note that $\delta_1 = 0$, $\delta_2 = 0$, $\delta_3 = 0$ or $\delta_4 = 0$ corresponds to the true H_0 which will be used to examine the size of the test statistics. Nonzero δ's

represent different levels of heterogeneity between the two groups.

Figures S1 in supplimentary displays the powers of the six tests. For Setting 1, Figure S1(a)-(b) show that the powers of the PLR, MMD, ELT, AD, DSLICE, and KDT tests rapidly approach one when n or δ_1 increases. The power of the K-S test increases slightly slower than the other five tests. DSLICE appears to be slightly less powerful than the other four tests, maybe because of its discrete nature and its challenges in choosing a proper penalization parameter in their penalized slicing approach. For Setting 2, as shown in Figure S1(c)-(d), the MMD and PLR tests show comparable power. The PLR test has slightly higher power when the heterogeneity is higher. The power difference between these two tests increases as δ_2 increases. AD and K-S show significantly lower power. For Setting 3, Figure S1(e)-(f) show again that the PLR test has the highest power. DSLICE performs quite well here, maybe due to its flexibility in slicing. In contrast, the powers of K-S, MMD, ELT, AD, and KDT are significantly lower than both PLR and DSLICE. For Setting 4, PLR and DSLICE show similar power in Figure S1(g)-(h). The powers of MMD, K-S and AD tests are significantly lower than the others. The results demonstrate that both PLR and DSLICE are more adaptive to differently shaped distributions than the other four methods, while PLR enjoys additional advantages than DSLICE when the underlying distribution is

smooth.

Figure S2 in supplimentary displays the size of K-S, MMD, ELT, AD, DSLICE, KDT, and PLR tests. It can be seen that the sizes of the six tests are all around the nominal level 0.05 in Settings 1 and Setting 2, confirming that all tests are asymptotically valid. In Setting 3 and Setting 4, the size of the PLR test is still asymptotically correct, and that for DSLICE is reasonably close; while the sizes of K-S, MMD and ELT are way below 0.05, showing that these three tests are too conservative in handling bimodal distributions. We also test the performance under multivariate distribution ($d > 2$) and multiple distributions in Supplemenary, the proposed tests maintains highest power with controlled type-I error. In simulation studies with Beta and mixtrure of Beta distribution in Supplimentary, our proposed test also shows the highest power.

6. Real Data Analysis

In this section, application on metagenomic analysis of type II diabetes is provided to compare our PLR test with the Kolmogorov-Smirnov (K-S) and maximum mean discrepancy (MMD) tests. We also conduct another real example about gene expression analysis of chronic lymphocytic leukaemia in Supplementary S.5.2.

Recent studies have indicated that gut microbiota play an important role in many human diseases such as obesity and diabetes, and have observed significant association between diseases and gut microbial composition (Turnbaugh et al., 2009; Qin et al., 2012). Due to the rapid development of metagenomics, it is possible to study microbial DNA contents through environmental samples directly. Compared with traditional culture-based methods, metagenomics can study unculturable microorganisms and are much more scalable. Recently, several 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 differences since there are many environmental factors, such as age, dietary habits, and antibiotic usage, that can alter the composition of gut microbiota. A powerful test that can detect such distributional differences between different populations would be useful in metagenomic analysis.

This study aims 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 ith 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 that have an abundance larger than 1% of the mean abundance in more than 50% of the total samples. The 1005 p-values were calculated by K-S, 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 rate at 5%, we compared the identified species from the three methods in Figure S7 in supplimentary. The PLR, K-S, and MMD tests identified 101, 4, and 13 species, respectively. The species identified by PLR cover those by K-S or MMD.

Moreover, we highlighted two species that were identified only by the PLR test in Figure S7 (B-C). The densities of these two species are both bimodal in both the case and the control groups. Figure S7(B) plots the conditional density of the log-transformed abundance of *Roseburia intestinalis*. The majority of the case group has a significantly low abundance. In Figure S7(C), the other species, *Faecalibacterium prausnitzii* has a lower abundance for a

subgroup of patients in the case group. Both species are butyrate-producing bacteria which can 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 two species' concentrations are lower in T2D subjects. Also, we found several Lactobacillus species are increased in T2D patients which are also found in De La Vega-Monroy et al. (2013); Qin et al. (2012).

7. Discussion

In this paper, we proposed a probabilistic decomposition approach for probability densities based on the penalized likelihood ratio (PLR). As demonstrated in 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 that does not involve resampling.

Supplementary Materials Contain figures for simulation studies, figures real data analysis, additional simulated and real examples, the data-adaptive tuning parameter selection, extension to the case with a divergent number of samples, connection to maximum mean discrepancy, all technical proofs, and additional numerical reuslts.

References

Abramowitz, M. and I. A. Stegun (1948). *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, Volume 55. US Government printing office.

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, and S. Mendelson (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.

Bilban, M., D. Heintel, T. Scharl, T. Woelfel, M. M. Auer, E. Porpaczy, B. Kainz, A. Kröber, V. J. Carey, and M. Shehata (2006). Deregulated expression of fat and muscle genes in b-cell chronic lymphocytic leukemia with high lipoprotein lipase expression. *Leukemia 20*(6), 1080–1088.

Braun, M. L. (2006). Accurate error bounds for the eigenvalues of the kernel matrix. *Journal of Machine Learning Research 7*(Nov), 2303–2328.

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.

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

De La Vega-Monroy, M. L., E. Larrieta, M. German, A. Baez-Saldana, and C. Fernandez-Mejia (2013). Effects of biotin supplementation in the diet on insulin secretion, islet gene expression, glucose homeostasis and beta-cell proportion. *The Journal of nutritional biochemistry 24*(1), 169–177.

Eric, M., F. R. Bach, and Z. Harchaoui (2008). Testing for homogeneity with kernel fisher discriminant analysis. In *Advances in Neural Information Processing Systems*, pp. 609–616.

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

Gretton, A., K. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola (2007). A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems*, pp. 513–520.

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. (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.

Kim, I. (2021). Comparing a large number of multivariate distributions. *Bernoulli 27*(1), 419–441.

Kim, Y.-J. and C. Gu (2004). Smoothing spline gaussian regression: more scalable computation via efficient approximation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66*(2), 337–356.

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

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.

Liu, M., Z. Shang, and G. Cheng (2020). Nonparametric distributed learning under general designs. *Electronic Journal of Statistics 14*(2), 3070–3102.

Liu, M., Z. Shang, Y. Yang, and G. Cheng (2021). Nonparametric testing under randomized sketching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–1.

Ma, P., J. Z. Huang, and N. Zhang (2015). Efficient computation of smoothing splines via adaptive basis sampling. *Biometrika 102*(3), 631–645.

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.

Martínez-Camblor, P., J. De Una-Alvarez, and N. Corral (2008). k-sample test based on the common area of kernel density estimators. *Journal of Statistical Planning and Inference 138*(12), 4006–4020.

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.

Novak, E., M. Ullrich, H. Woźniakowski, and S. Zhang (2018). Reproducing kernels of sobolev spaces on $\mathbb{R}^d$ and applications to embedding constants and tractability. *Analysis and Applications 16*(05), 693–715.

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. and G. Cheng (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.

REFERENCES

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, Volume 26. CRC press.

Tapia, R. and J. Thompson (1978). *Nonparametric Probability Density Estimation*. Goucher College Series. Johns Hopkins University Press.

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.

Wang, Y. (2011). *Smoothing splines: methods and applications*. CRC Press.

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

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

Xing, X., M. Liu, P. Ma, and W. Zhong (2020). Minimax nonparametric parallelism test. *Journal of Machine Learning Research 21*(94), 1–47.

Zhan, D. and J. Hart (2014). Testing equality of a large number of densities. *Biometrika 101*(2), 449–464.