| Statistica Sinica Preprint No: SS-2020-0339 | |
|---|---|
| **Title** | Unified Tests for Nonparametric Functions in RKHS With Kernel Selection and Regularization |
| **Manuscript ID** | SS-2020-0339 |
| **URL** | http://www.stat.sinica.edu.tw/statistica/ |
| **DOI** | 10.5705/ss.202020.0339 |
| **Complete List of Authors** | Tao He,<br>Ping-Shou Zhong,<br>Yuehua Cui and<br>Vidyadhar Mandrekar |
| **Corresponding Author** | Ping-Shou Zhong |
| **E-mail** | pszhong@stt.msu.edu |

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0339

Statistica Sinica

# Unified Tests for Nonparametric Functions in RKHS with Kernel Selection and Regularization

Tao He[1], Ping-Shou Zhong[2], Yuehua Cui[3] and Vidyadhar Mandrekar[3]

[1]*San Francisco State University,* [2]*University of Illinois at Chicago and* [3]*Michigan State University*

*Abstract:* This study develops a unified test procedure for nonparametric functions in a reproducing kernel Hilbert space of high-dimensional or functional covariates. The test procedure is simple, computationally efficient, and practical because we do not need to distinguish high-dimensional or functional covariates. We derive the asymptotic distributions of the proposed test statistic under the null and a series of local alternative hypotheses. The asymptotic distributions depend on the decay rate of the eigenvalues of the kernel function. This decay rate is determined by the kernel function and the types of covariates. We also develop a novel kernel selection procedure to maximize the power of the proposed test by maximizing the signal-to-noise ratio. The proposed kernel selection procedure is shown to be consistent in selecting the kernels that maximize the power function. Moreover, a test with a regularized kernel is constructed to further improve the power. It is shown that the proposed test nearly achieves the power of an oracle test if the regularization parameter is properly chosen. Extensive simulation studies evaluate the finite-sample performance of the proposed method. Finally, we apply the proposed method to a Yorkshire gilt data set to identify pathways that are associated with the triiodothyronine level. The proposed methods are included in an R package "KerUTest."

*Key words and phrases:* Gene set analysis; Kernel function; Nonparametric regression; Reproducing kernel Hilbert space

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0339

## 1. Introduction

High-dimensional or functional data arise in a wide range of areas, including biology, imaging, and climate. In genetic studies, millions of single nucleotide polymorphisms (SNPs) can be measured simultaneously using high-throughput technologies. The identification of genes that are associated with certain traits, such as blood pressure and grain yield, is becoming increasingly important in health and agriculture sciences. Although the traditional methods focus on a single gene-based analysis, this method has limitations (Manolio et al. (2009)). Gene-set based analysis (e.g., Subramanian et al. (2005)) holds great promise, because gene regulation is often complex and genes tend to work together in a nonlinear way (Liu et al. (2007); Li et al. (2012)) to achieve certain biological functions. To model the association between a certain trait $Y$ and a gene set $\mathbf{X}$, we consider the following nonparametric regression:

$$Y_i = \mu + h(\mathbf{X}_i) + \epsilon_i, \quad i = 1, \cdots, n, \tag{1.1}$$

where $\mathbf{X}_1, \cdots, \mathbf{X}_n$ are independent and identically distributed (i.i.d.) $p$-dimensional covariates generated from a probability measure on $R^p$, $h(\mathbf{X}_i)$ is an unknown nonparametric function of $\mathbf{X}_i = (X_{i1}, \cdots, X_{ip})^T$, and $\epsilon_i$ are i.i.d. random errors with mean zero and variance $\sigma^2$. For the purpose of model identification, without loss of generality, we assume $E\{h(\mathbf{X}_i)\} = 0$.

In a gene-set analysis, the number of genes $p$ in a gene-set can be in the order of thousands, but the sample size $n$ is limited and much smaller than $p$. If there is no

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0339

natural ordering among $\{X_{ij}\}_{j=1}^{p}$, $\mathbf{X}_i$ is a $p$-dim vector and $\mathbf{X}_i$ may be considered to be high-dimensional data (e.g., Bai and Saranadasa (1996)). A "large $p$, small $n$" setup can be used to model high-dimensional data when $p$ is much larger than $n$. If $\{X_{ij}\}_{j=1}^{p}$ can be indexed by a certain variable (e.g., chromosome locations), then $X_{ij}$ may be considered as a realization of a functional curve $X_i(\cdot)$ observed at $t_j$, where $t_1 < t_2 < \cdots < t_p$. Then, $\mathbf{X}_i = \{X_i(t_1), \cdots, X_i(t_p)\}^T$ is a collection of $p$ repeated measurements of $X_i(\cdot)$, a smooth curve in some underlying functional space (Ramsay and Silverman (2005)). When $p$ is much larger than $n$, $\mathbf{X}_i$ denotes dense functional data. An interesting procedure called "stringing" was developed by Chen et al. (2011) to transform high-dimensional data into functional data. However, in many real applications, considering $\mathbf{X}$ as high-dimensional or functional data is often subjective. To avoid this subjective choice, we use a general reproducing kernel Hilbert space (RKHS) for $h(\cdot)$, so that our approach is applicable to both high-dimensional and functional data.

This study aims to test the existence of a nonlinear association between a quantitative trait $Y$ and a gene set $\mathbf{X}$, which is equivalent to testing the following:

$$H_0 : h(\cdot) = 0 \quad \text{vs} \quad H_1 : h(\cdot) \neq 0. \tag{1.2}$$

Hypothesis testing for a nonparametric function of an explanatory variable in a finite-dimensional Euclidean space has been well studied in the literature. For example, Chen et al. (2003) and Gao and Gijbels (2008) considered inference for nonparametric functions based on kernel smoothing estimators. Shang and Cheng (2013) developed a general inference for nonparametric functions in a Sobolev space based on smoothing spline

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0339

estimators. Fan et al. (2001) developed generalized likelihood ratio tests for various nonparametric models with parametric distribution errors, and established Wilks theorems for a class of the generalized likelihood statistics using local polynomial estimators. Recently, Yang et al. (2020) developed a non-asymptotic test and Liu et al. (2018) developed a computationally efficient test for nonparametric functions. Most existing methods require an estimation of nonparametric functions, and suffer from the "curse of dimensionality" (Fan (2018)). Hence, they cannot be easily generalized to functions with explanatory variables in a high-dimensional space without a specific structure. In the high-dimensional linear regression with $h(\mathbf{X}) = \mathbf{X}^T\boldsymbol{\beta}$, testing $h(\cdot) = 0$ is equivalent to testing high-dimensional coefficients $\boldsymbol{\beta} = 0$ (e.g., Zhong and Chen (2011), Lan et al. (2014), Wang and Cui (2013)). However, these methods were designed for a linear model and do not apply to a general nonparametric function. When $\mathbf{X}_i$ is considered as functional data, extensive studies have been done for hypothesis testing under various model settings, for example, under the functional linear model (e.g., Kong et al. (2016); Su et al. (2017)), under generalized functional linear models (e.g., Shang and Cheng (2015); Li and Zhu (2020)), and considering nonparametric functions of functional covariates (e.g., Delsol et al. (2011); Delsol (2012)). See Tekbudak et al. (2019) for a recent review. Delsol et al. (2011) and Delsol (2012) constructed Cramér–von Mises-type test statistics based on a local smoothing estimator of the nonparametric function, and applied wild bootstrap procedures for practical implementation, which are computationally intensive.

An RKHS-based method is a popular approach for modeling nonparametric func-

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0339

tions. Most existing methods study RKHS for nonparametric functions of finite-dimensional covariates, where $p$ is a fixed constant and does not grow with the sample size (Wahba (1990), Liu et al. (2007), and Liu et al. (2008)). The estimation of the RKHS-based nonparametric function of functional data covariates (i.e., $h(\mathbf{X})$ is a function of functionals) was developed in Lian (2007) and Avery et al. (2014). However, there is no existing unified inference method for $h(\cdot)$ of high-dimensional or functional covariates.

The goal of this study is to develop a unified method for testing a nonparametric function in an RKHS of high-dimensional or functional covariates. The proposed method does not directly estimate the nonparametric function $h(\cdot)$ of the high-dimensional or functional covariates, and does not require a dimension-reduction method. Our key idea is to transform the hypothesis in (1.2) into an equivalent hypothesis. A U-statistic-based test statistic is then developed to test the equivalent hypothesis (see Section 2). The asymptotic distributions of the test statistic are obtained under the null hypothesis and a series of local alternatives, without a specific distribution assumption. The asymptotic distributions depend on the decay rate of the eigenvalues of a given kernel function. However, the decay rate is usually unknown, because it is determined both by the smoothness of the reproducing kernel $K$ and the distribution (hence, the types) of the covariates $\mathbf{X}$. As a result, the asymptotic distributions are not directly applicable. To address this challenge, we develop a unified and practical approximation method that does not require knowledge of the decay rate. Moreover, the proposed test procedure is computationally efficient without bootstrap procedures.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0339

An important finding in this study is that testing for the nonparametric function $h(\cdot)$ of high-dimensional covariates is feasible, even if no specific structure is imposed for $h(\cdot)$. The consistency of the test depends on the smoothness of the functional space and the data type of the covariates. If the functional space $\mathscr{H}_K$ generated by the kernel is sufficiently smooth (e.g., Gaussian kernel) or the covariates $\mathbf{X}$ are functional data, the proposed test is consistent without restrictions on the relationship between the dimension of the covariate $p$ and the sample size $n$. If the functional space $\mathscr{H}_K$ is not sufficiently smooth and the covariate $\mathbf{X}$ is high-dimensional data, some restrictions on $p$ and $n$ are needed to make the proposed test consistent.

In practice, the power of the proposed test depends on the choice of kernels. As a result, kernel selection is an important issue in a kernel machine-based testing procedure (Liu et al. (2007)). However, few studies have examined this area. We propose a new procedure for selecting kernels in the hypothesis testing context. By obtaining an explicit power function of the proposed test, we choose the kernel that maximizes the power function. Unlike the BIC proposed in Liu et al. (2007), our procedure is tailored to the hypothesis testing problem, and is particularly designed to improve the power of the proposed test. We show that the kernel selection procedure is consistent in the sense that it selects the kernels that maximize the power with probability one. Moreover, we can construct a regularized kernel to further improve the power of the test. A novel method for choosing the regularization parameter is introduced. We show that the proposed test with a regularized kernel achieves the power of an oracle test if the regularization

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0339

parameter is properly chosen.

The rest of the paper is organized as follows. In Section 2, we introduce the RKHS, functional space for $h(\cdot)$, and equivalent hypothesis. Section 3 proposes a new test statistic and establishes the main asymptotic distributions of the proposed test statistic under the null hypothesis and local alternatives. The kernel selection and regularization are discussed in Section 4. The finite-sample performance of the proposed test statistic is evaluated using extensive simulations in Section 5. In Section 6, we apply the proposed method to a Yorkshire gilt data set to identify gene sets associated with triiodothyronine levels. A brief discussion is given in Section 7. Some theoretical results, all the technical details, and additional simulation results are relegated to the Supplementary Material.

## 2. Functional space and equivalent hypothesis

Consider functions $h(\cdot)$ that belong to a functional space $\mathscr{H}_K$ generated by a kernel $K_{n,\theta_n}(\cdot,\cdot)$, where $\theta_n$ are tuning parameters that possibly depend on $n$. For notational convenience, we suppress $n$ in $\theta_n$ in the rest of this paper. The kernel $K_{n,\theta}(x_1, x_2) : R^p \times R^p \to R$ is any symmetric and positive semi-definite function defined on $R^p \times R^p$. Throughout the paper, we assume $p = p(n)$ is a function of $n$. A kernel $K_{n,\theta}(x_1, x_2)$ is said to be positive semi-definite if the associated kernel matrix $(K_{n,\theta}(x_i, x_j))_{i,j=1}^M$ is an $M \times M$ positive semi-definite matrix defined on any $M$ distinct points $x_1, \cdots, x_M \in R^p$. We use $K_{n,\theta}$ and bold font $\mathbf{K}$ to denote the kernel function and an $n \times n$ kernel matrix defined by $\mathbf{K} = \left\{K_{n,\theta}(\mathbf{X}_i, \mathbf{X}_j)\right\}_{i,j=1}^n$, respectively. Some commonly used kernel functions

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0339

include the linear kernel $K_{n,\theta}(z_1, z_2) = z_1^T z_2/\theta$ and the Gaussian kernel $K_{n,\theta}(z_1, z_2) = \exp(-\|z_1 - z_2\|^2/\theta)$. Additional examples of kernel functions can be found in Liu et al. (2007).

The functional space $\mathscr{H}_K$ is determined by the kernel function $K_{n,\theta}$. To define the functional space $\mathscr{H}_K$, we define the following normalized kernel

$$\mathcal{K}_{n,\theta}(x_1, x_2) = \frac{K_{n,\theta}(x_1, x_2)}{\sqrt{E\{K_{n,\theta}^2(\mathbf{X}_1, \mathbf{X}_2)\}}},$$

where $\mathbf{X_1}$ and $\mathbf{X_2}$ are two independent copies of $\mathbf{X}$ with probability measure $\boldsymbol{P}$. It is then obvious that $E\{\mathcal{K}_{n,\theta}^2(\mathbf{X}_1, \mathbf{X}_2)\} = 1$ and $\mathcal{K}_{n,\theta}(x_1, x_2)$ is still positive semi-definite and symmetric. The above normalization ensures $E\{\mathcal{K}_{n,\theta}^2(\mathbf{X}_1, \mathbf{X}_2)\} < \infty$, so that the eigen-decomposition of $\mathcal{K}_{n,\theta}$ can be properly defined according to Lemma 1 in the Supplemental Material. The normalization is needed because $E\{K_{n,\theta}^2(\mathbf{X}_1, \mathbf{X}_2)\}$ could diverge in the high-dimensional case. For instance, if $K_{n,\theta}(\mathbf{X}_1, \mathbf{X}_2) = \mathbf{X}_1^T\mathbf{X}_2$ and $\mathrm{Var}(\mathbf{X}) = \boldsymbol{\Sigma}$, then $E\{K_{n,\theta}^2(\mathbf{X}_1, \mathbf{X}_2)\} \geq \mathrm{tr}(\boldsymbol{\Sigma}^\mathbf{2})$, which implies that $E\{K_{n,\theta}^2(\mathbf{X}_1, \mathbf{X}_2)\}$ is at least of order $p$ if all the eigenvalues of $\boldsymbol{\Sigma}$ are bounded away from zero. Note that the normalization is mainly for theoretical analyses. Our standardized test statistic is invariant to the kernel normalization. Thus, the normalization is not needed in practice for the proposed test.

By Corollary 1 in the Supplemental Material, we can write

$$\mathcal{K}_{n,\theta}(x_1, x_2) = \sum_{m=1}^{\infty} \lambda_{\mathcal{K}_\theta,m}\psi_{n\theta,m}(x_1)\psi_{n\theta,m}(x_2),$$

where $\lambda_{\mathcal{K}_\theta,1} \geq \lambda_{\mathcal{K}_\theta,2} \geq \cdots$ are eigenvalues of $\mathcal{K}_{n,\theta}$, and $\{\psi_{n\theta,m}(\cdot)\}$ form a complete orthogonal normal system on $L^2(\boldsymbol{P})$. This representation extends the eigen-decomposition

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0339

of a kernel (or covariance) function from a one-dimensional space to a $p$-dimensional (or functional) space. Without causing confusion, we use $\lambda_{n\mathcal{K},m}$ and $\psi_{nm}(\cdot)$ to denote $\lambda_{n\mathcal{K}_\theta,m}$ and $\psi_{n\theta,m}(\cdot)$, respectively. Then, the space $\mathscr{H}_K$ is defined as (Cucker and Smale (2002))

$$\mathscr{H}_K = \{f(x) : f(x) = \sum_{m=1}^{\infty} \alpha_m \psi_{nm}(x) \text{ for } \alpha_m \text{ satisfying } \sum_{m=1}^{\infty} \alpha_m^2/\lambda_{n\mathcal{K},m} < \infty\}.$$

For example, if a centralized linear kernel $K_{n,\theta}(x_1, x_2) = (x_1 - \boldsymbol{\mu}_X)^T(x_2 - \boldsymbol{\mu}_X)$ with $\boldsymbol{\mu}_X = E(\mathbf{X})$ is used, the space $\mathscr{H}_K$ contains linear functions $f(x) = \boldsymbol{\beta}^T x$. If $\mathbf{X}$ is a high-dimensional vector, model (1.1) reduces to a linear model. If $\mathbf{X}$ is a functional data vector, model (1.1) becomes a functional linear model $h(x) = \int x(t)\beta(t)dt$. If nonlinear kernels such as polynomial and Gaussian kernels are given, the functional space $\mathscr{H}_K$ includes very general nonlinear models.

To distinguish $H_1$ from $H_0$, we define a measure to quantify the distance between $h(\cdot)$ and zero. Here, we define the norm $\|\cdot\|_K$ as a measure:

$$\|h\|_K^2 = \sum_{m=1}^{\infty} \lambda_{nm}\alpha_m^2, \tag{2.1}$$

where $\lambda_{nm} = \sqrt{E\{K_{n,\theta}^2(\mathbf{X}_1, \mathbf{X}_2)\}}\lambda_{n\mathcal{K},m}$, which may be considered as the eigenvalues of the kernel function $K_{n,\theta}(x, y)$. Obviously, the null hypothesis in (1.2) is true if and only if $\|h\|_K^2 = 0$, and $\|h\|_K^2 > 0$ under the alternative hypothesis. Therefore, the hypothesis considered in (1.2) is equivalent to

$$H_0 : \|h\|_K^2 = 0 \text{ vs } H_1 : \|h\|_K^2 > 0. \tag{2.2}$$

The connection between a nonparametric function and its eigen-decomposition has been used for statistical inference in the literature. For example, Fan (1996) developed Ney-

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0339

man's adaptive tests based on the Fourier transform of a nonparametric function.

For model identification, in the rest of this paper, we consider a centralized kernel $K_{n,\theta}$ that satisfies $\mu_K = E\{K_{n,\theta}(\mathbf{X}_1, \mathbf{X}_2)\} = 0$. Recall that $h(\cdot)$ needs to satisfy $\mathrm{E}\{h(\mathbf{X}_i)\} = 0$ for the purpose of identification. Note that the centralized kernel is equipped with the zero-mean eigenfunctions $\{\psi_{nm}(\cdot)\}_{m=1}^{\infty}$. As a result, the functions in the corresponding RKHS $\mathscr{H}_K$ have zero means, because $\mathrm{E}\{h(\mathbf{X}_i)\} = \mathrm{E}\{\sum_{m=1}^{\infty} \alpha_m \psi_{nm}(\mathbf{X}_i)\} = 0$. The centralized kernel $K_{n,\theta}$ can be constructed from any positive-definite kernel function $K^*_{n,\theta}$ by setting $K_{n,\theta}(\mathbf{x}_1, \mathbf{x}_2) = K^*_{n,\theta}(\mathbf{x}_1, \mathbf{x}_2) - K^*_{1,\theta}(\mathbf{x}_1) - K^*_{1,\theta}(\mathbf{x}_2) + \mu_{K^*}$, where $K^*_{1,\theta}(\mathbf{x}_1) = E\{K^*_{n,\theta}(\mathbf{x}_1, \mathbf{X}_2)\}$ is the first-order projection of $K^*_{n,\theta}$. By Lemma 3 in the Supplementary Material, $K_{n,\theta}$ is still semi-positive definite with only one zero eigenvalue $\lambda^*_{nm} = 0$ corresponding to the eigenfunction $\psi^*_{nm}(x) = 1$, if $K^*_\theta$ is positive definite. Some benefits of a centralized kernel are discussed in Lindsay et al. (2008). The practical construction of a centralized kernel is discussed in the next section.

## 3. Test statistics and asymptotic distributions

By the orthonormal expansion of $\mathcal{K}_{n,\theta}(x, y)$ in Section 2, we observe that $E\{(Y_i - \mu)(Y_j - \mu)K_{n,\theta}(\mathbf{X}_i, \mathbf{X}_j)\} = \|h\|_K^2$, for any $(i, j)$ pair such that $i \neq j$. Motivated by this observation, we consider the following test statistic:

$$T_n = \frac{1}{n(n-1)} \sum_{i \neq j} K_{n,\theta}(\mathbf{X}_i, \mathbf{X}_j)(Y_i - \bar{Y}_n)(Y_j - \bar{Y}_n)/\hat{\sigma}^2, \tag{3.1}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0339

where $\bar{Y}_n = n^{-1}\sum_{i=1}^n Y_i$ is the sample mean and $\hat{\sigma}^2 = (n-1)^{-1}\sum_{i=1}^n (Y_i - \bar{Y}_n)^2$ is the sample variance estimator of $\sigma^2$ under the null hypothesis (1.2). We can then check that $E(T_n) = o(1)$ under the null hypothesis and $E(T_n) = \|h\|_K^2/\sigma^2\{1+o(1)\}$ under the alternative. Therefore, the test statistic $T_n$ is able to distinguish the null and alternative hypotheses in (1.2).

Define $K_{2n,\theta}(x,y) = E\{K_{n,\theta}(x,\mathbf{X})K_{n,\theta}(\mathbf{X},y)\}$. Let $\lambda_{n1} \geq \lambda_{n2} \geq \lambda_{n3} \geq \cdots$ be eigenvalues of the kernel function $K_{n,\theta}$, and define $V_{kn} = \sum_{m=1}^{\infty} \lambda_{nm}^k$ for integers $k = 1, 2, \cdots$ The asymptotic framework considered here is $p(n) \to \infty$ as $n \to \infty$, where $p(n)$ diverges as $n$ diverges. However, we do not require an explicit relationship between $p(n)$ and $n$. To study the asymptotic distributions of the proposed test statistic $T_n$, we need the following technical assumptions:

(C1) Assume $\tau_8 < \infty$, where $\tau_k = E(\epsilon^k)$ is the $k$th moment of the random error $\epsilon$.

(C2) Assume $\sup_n V_{2n}^{-1-\delta/2} E|K_{n,\theta}(\mathbf{X}_1, \mathbf{X}_2)|^{2+\delta} < \infty$, for some $\delta > 0$, and $\sum_{m=M}^{\infty} \lambda_{n\mathcal{K},m}^2 \to 0$ uniformly for all $n > n_0$ as $n_0$ and $M \to \infty$.

(C3) Assume $E\{K_{2n,\theta}^4(\mathbf{X}_1, \mathbf{X}_2)\} = o(V_{2n}^4)$ and $E\{K_{2n,\theta}^2(\mathbf{X}_1, \mathbf{X}_1)\} = o(nV_{2n}^2)$.

The following theorem summarizes the asymptotic distribution of $T_n$ under $H_0$, and the proof can be found in the Supplemental Material.

**Theorem 1.** *Under the null hypothesis $H_0$ in (1.2) and (C1): (i) Assume (C2) holds. If $\lambda_{n\mathcal{K},m} \to \lambda_{\mathcal{K},m}$ as $n \to \infty$, then $nT_n/\sqrt{V_{2n}} \xrightarrow{d} \sum_{m=1}^{\infty} \lambda_{\mathcal{K},m}(\chi_m^2 - 1)$, where $\chi_m^2$ are independent chi-squared distributions with one degree of freedom; (ii) If condition (C3) holds, then $\sigma_{T_n}^{-1} nT_n \xrightarrow{d} N(0,1)$, where $\sigma_{T_n}^2 = 2V_{2n}$.*

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0339

**Remark 1**: Theorem 1 shows that the asymptotic distributions of $T_n$ depend on the decay rate of the eigenvalues $\lambda_{n\mathcal{K},m}$, which is determined by the kernel function and the dimension and distribution of the random vector $\mathbf{X}$. Consider a linear kernel given by $K_{n,\theta}(x_1, x_2) = (x_1 - \boldsymbol{\mu}_X)^T(x_2 - \boldsymbol{\mu}_X)$. The eigenvalues of the linear kernel are given by the eigenvalues of $\text{Var}(\mathbf{X}) = \boldsymbol{\Sigma}$. Part (i) of Theorem 1 provides an asymptotic distribution of $nT_n$ when the eigenvalues decay fast and $nT_n/\sqrt{V_{2n}}$ has the same distribution as the finite sum $\sum_{m=1}^{M} \lambda_{\mathcal{K},m}(\chi^2_m - 1)$, for some $M$ in (C2). If $\mathbf{X} = \{X_i(t_1), \cdots, X_i(t_p)\}^T$ is a functional data vector, the assumptions in (C2) are typical in a functional PCA type-based analysis, where the first few eigenvalues are dominant. The asymptotic distribution of $T_n$ is a weighted chi-squared distribution, not a chi-squared distribution, which differs from the Wilks' phenomena established for the nonparametric likelihood ratio test statistics (e.g., Fan et al. (2001)).

However, for high-dimensional data, the eigenvalues may not decay at a fast enough rate. Under this scenario, the asymptotic distribution is an asymptotic normal, as established in part (ii) of Theorem 1. For the above linear kernel, if we further assume that $\mathbf{X}_1$ and $\mathbf{X}_2$ are multivariate normal, then $E\{K^4_{2n,\theta}(\mathbf{X}_1, \mathbf{X}_2)\} = 3\text{tr}^2(\boldsymbol{\Sigma}^4) + 6\text{tr}(\boldsymbol{\Sigma}^8)$ and $E\{K^2_{2n,\theta}(\mathbf{X}_1, \mathbf{X}_1)\} = \text{tr}^2(\boldsymbol{\Sigma}^2) + 2\text{tr}(\boldsymbol{\Sigma}^4)$. If $\text{tr}(\boldsymbol{\Sigma}^4) = o\{\text{tr}^2(\boldsymbol{\Sigma}^2)\}$ (Zhong and Chen, 2011), then condition (C3) holds. The condition $\text{tr}(\boldsymbol{\Sigma}^4) = o\{\text{tr}^2(\boldsymbol{\Sigma}^2)\}$ is true for most scenarios when the eigenvalues of $\boldsymbol{\Sigma}$ decay slowly.

**Remark 2**: Because the decay rate of the eigenvalues $\lambda_{n\mathcal{K},m}$ is difficult to determine for a general kernel, and it relies on the distribution of $\mathbf{X}$, the asymptotic distributions

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0339

are not directly applicable. On the one hand, part (i) of Theorem 1 shows that the limiting distribution of $nT_n/\sqrt{V_{2n}}$ is $\sum_{m=1}^{\infty}\lambda_{\mathcal{K},m}(\chi^2_m-1)$. Because $\lambda_{nm}=\sqrt{V_{2n}}\lambda_{n\mathcal{K},m}$ and $\lambda_{n\mathcal{K},m}\to\lambda_{\mathcal{K},m}$, we may approximate the distribution of $nT_n$ by $\mathcal{T}_n=\sum_{m=1}^{\infty}\lambda_{nm}(\chi^2_m-1)$. On the other hand, if condition (C3) holds, then Lyapunov's condition $V_{4n}/V_{2n}^2\to 0$ is satisfied so that the central limit theorem holds for the sum of the weighted centralized chi-squared distributions $\mathcal{T}_n$; that is, $\sigma_{T_n}^{-1}\mathcal{T}_n\xrightarrow{d}N(0,1)$. This means that the asymptotic normality in Theorem 1 may be considered as the limiting distribution of $\mathcal{T}_n$. Thus, $\mathcal{T}_n$ is flexible enough to approximate the asymptotic distributions in both scenarios in Theorem 1, and $\mathcal{T}_n$ provides a unified inference approach for both high-dimensional and functional data.

In practice, obtaining accurate estimators for all the eigenvalues $\lambda_{nm}$ $(m=1,2,\cdots)$ simultaneously is difficult. Nevertheless, we apply a Satterthwaite approximation to the mixture of chi-squares $\sum_{m=1}^{\infty}\lambda_{nm}\chi^2_m$ using a scaled chi-squared distribution $\hat{a}_n\chi^2_{\hat{g}_n}$, where $\hat{g}_n=\hat{V}_{1n}/\hat{a}_n$, $\hat{a}_n=\hat{\sigma}^2_{T_n}/(2\hat{V}_{1n})$, and $\hat{V}_{1n}=n^{-1}\mathrm{tr}(\mathbf{HK})$ is an unbiased estimator of $V_{1n}$. Here, $\mathbf{H}=\mathbf{I}-n^{-1}\mathbf{J}$ is a projection matrix and $\mathbf{J}$ is an $n\times n$ matrix with all elements equal to one. Then, we approximate $\mathcal{T}_n$ by $\hat{a}_n\chi^2_{\hat{g}_n}-\hat{V}_{1n}$. The accuracy of the Satterthwaite approximation is at the order of $O\{(V_{3n}^2/V_{2n}^3)^{1/2}\}$. Taking the linear kernel as an example, if all the eigenvalues of $\mathbf{\Sigma}$ are finite, then $(V_{3n}^2/V_{2n}^3)^{1/2}$ is at the order of $p^{-1/2}$.

A unified asymptotic $\alpha$-level test rejects the null hypothesis if

$$(nT_n+\hat{V}_{1n})/\hat{a}_n>\chi^2_{\hat{g}_n,1-\alpha}, \tag{3.2}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0339

where $\chi^2_{g,1-\alpha}$ is the $1-\alpha$ quantile of a chi-squared distribution with $g$ degrees of freedom.

**Remark 3**: If conditions (C3) holds, then an $\alpha$-level test rejects the null if

$$\hat{\sigma}_{T_n}^{-1} n T_n > z_{1-\alpha}, \tag{3.3}$$

where $z_{1-\alpha}$ is the lower $1-\alpha$ quantile of the standard normal distribution, $\hat{\sigma}^2_{T_n} = 2(n-1)^{-2}\mathrm{tr}(\mathbf{H}\mathbf{K}^0\mathbf{H}\mathbf{K}^0)$ is a ratio-consistent estimator for $\sigma^2_{T_n}$ (see Proposition 1 in Section 4.1), where $\mathbf{A}^0 = (A^0_{ij})$ is a zero-diagonal matrix with $A^0_{ij} = A_{ij}$, for $i \neq j$ and $A^0_{ii} = 0$.

To achieve better accuracy in the size approximation, we adjust the variance estimator $\hat{\sigma}^2_{T_n}$ using the high-order moments of $\epsilon$ in (1.1). The adjusted variance estimator $\hat{\sigma}^2_{T_n,adj}$ replaces the estimator $\hat{\sigma}^2_{T_n}$ in the simulation study in Section 5 and the real-data analysis in Section 6. Assume the density function of $\epsilon$ is symmetric around zero. The adjusted variance estimator $\hat{\sigma}^2_{T_n,adj}$ is $\hat{\sigma}^2_{T_n,adj} = \big\{(2 - 12/(n-1) + 6\hat{\Delta}/n)\mathrm{tr}(\mathbf{H}\mathbf{K}^0\mathbf{H}\mathbf{K}^0) - (2/n + \hat{\Delta}/n)\mathrm{tr}^2(\mathbf{H}\mathbf{K}^0) + \hat{\Delta}\mathrm{tr}(\mathbf{A}\circ\mathbf{A})\big\}/(n-1)^2$, where $\circ$ denotes the Hadamard product, $\mathbf{A} = \mathbf{H}\mathbf{K}^0\mathbf{H}$, and $\hat{\Delta} = n^{-1}\sum_{i=1}^n[(Y_i - \bar{Y}_n)/\hat{\sigma}]^4 - 3$. The derivation of $\sigma^2_{T_n,adj}$ is provided in the Supplementary Material.

**Remark 4**: If the centralized kernel $K_{n,\theta}$ is unknown and is constructed from a kernel function $K^*_{n,\theta}$, it may contain unknown quantities $\mu_{K^*}$ and $K^*_{1,\theta}(\mathbf{X}_1)$. Thus, $T_n$ is not directly applicable. In this case, we can replace $K_{n,\theta}(\mathbf{X}_i, \mathbf{X}_j)$ with $\hat{K}_{n,\theta}(\mathbf{X}_i, \mathbf{X}_j)$, which is the $(i,j)$ element of $\mathbf{K} = \mathbf{K}^*_\theta - (n-1)^{-1}\mathbf{J}(\mathbf{K}^*_\theta)^0 - (n-1)^{-1}(\mathbf{K}^*_\theta)^0\mathbf{J} + n^{-1}(n-1)^{-1}\mathbf{J}(\mathbf{K}^*_\theta)^0\mathbf{J}$. Let $\hat{T}_n$ be the test statistic with corresponding kernel $\hat{K}_{n,\theta}$. It can be shown that $(nT_n - n\hat{T}_n)/\sqrt{V_{2n}} = o_p(1)$ (see the proof of Remark 4 in the Supplemental Material). This implies that $n\hat{T}_n/\sqrt{V_{2n}}$ has the same limiting distribution as $nT_n/\sqrt{V_{2n}}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0339

The next theorem studies the asymptotic distribution of the test statistic $T_n$ under a sequence of local alternative hypotheses,

$$H_{1n} : h(x) = d_n(x), \tag{3.4}$$

where $d_n(x)$ is any unknown function that possibly depends on $n$. For model identification, assume $\mathrm{E}\{d_n(\mathbf{X})\} = 0$. As usual, we consider local alternatives that are close to the null hypothesis because these are more challenging to detect than fixed alternatives. More specifically, assume that $d_n(\cdot)$ satisfies the following condition:

(C4) The local alternatives $d_n(x)$ satisfy $n\delta_K = O(V_{2n}^{1/2})$ and $n^2\mathrm{E}\{d_n^8(\mathbf{X})\} = o(1)$, where $\delta_K = \mathrm{E}\{K_{n,\theta}(\mathbf{X}_1, \mathbf{X}_2)d_n(\mathbf{X}_1)d_n(\mathbf{X}_2)\}$.

**Theorem 2.** *Under the local alternatives $H_{1n}$ in (3.4) satisfying (C4): (i) Assuming (C2) holds with $\delta = 2$, we have $V_{2n}^{-1/2}\{nT_n - \sigma_{T_n}\Psi(d_n)\} \xrightarrow{d} \sum_{m=1}^{\infty}\lambda_{\mathcal{K},m}(\chi_m^2 - 1)$, where $\Psi(d_n) = n\delta_K/(\sigma^2\sigma_{T_n})$ is the signal-to-noise ratio (SNR); (ii) If (C3) holds, then $\sigma_{T_n}^{-1}nT_n - \Psi(d_n) \xrightarrow{d} N(0,1)$.*

The proof of Theorem 2 can be found in the Supplementary Material. Applying Theorem 2, the power of an $\alpha$-level test for the rejection region in (3.3) under the local alternatives (3.4) is $\Omega(d_n) = 1 - \Phi\{z_{1-\alpha} - \Psi(d_n)\}$, where $\Phi(\cdot)$ is the CDF for the standard normal distribution. Therefore, the power of the proposed test is determined by the SNR $\Psi(d_n)$. If the $\alpha$-level rejection region in (3.2) is used, the power of the test is $\Omega(d_n) = P\big(\chi_{g_n}^2 > \chi_{g_n,1-\alpha}^2 - \sigma_{T_n}\Psi(d_n)/a_n\big)$, where $a_n = \sigma_{T_n}^2/(2V_{1n})$.

Let $d_n(\mathbf{x}) = b_n\Delta_n(\mathbf{x})$ such that $\mathrm{E}\{\lambda_{n1}^{-1}K_{n,\theta}(\mathbf{X}_1, \mathbf{X}_2)\Delta_n(\mathbf{X}_1)\Delta_n(\mathbf{X}_2)\}$ is a constant.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0339

Then, the proposed test has non-trivial power if $b_n = V_{2n}^{1/4}/\sqrt{n\lambda_{n1}}$. If $V_{2n}$ is a constant, which implies that $\lambda_{n1}$ is a constant, then the proposed test is able to detect alternatives of order $1/\sqrt{n}$. However, in high-dimensional cases, if $V_{2n}/\lambda_{n1}^2 \to \infty$ at a certain rate, the proposed test can detect alternatives of order $V_{2n}^{1/4}/\sqrt{n\lambda_{n1}}$, which is larger than $1/\sqrt{n}$. This reveals an adverse effect of dimensionality on the test. We observe that as long as $V_{2n} = o(n^2\lambda_{n1})$, the proposed test is consistent so that the power of the test converges to one. Depending on the chosen kernel, this condition might or might not impose conditions on $p(n)$ and $n$, because $V_{2n} = E\{K_{n,\theta}^2(\mathbf{X}_1, \mathbf{X}_2)\}$ depends on $p(n)$.

Assume that $\mathbf{X}$ is a $p$-dim random vector with mean $E(\mathbf{X}) = \mu_{\mathbf{X}}$ and covariance $\mathrm{Var}(\mathbf{X}) = \mathbf{\Sigma}$. Let $\eta_1 \geq \cdots \geq \eta_p$ be the eigenvalues of $\mathbf{\Sigma}$ and $r_m = \eta_m/\eta_1$ be the ratio of the eigenvalues. In the following, we discuss the implication of the condition $V_{2n} = o(n^2\lambda_{n1})$ on the relationship between $p$ and $n$ for four commonly used kernels: the linear, quadratic, polynomial, and Gaussian kernels.

**Example 1: (Linear Kernel)** If $K_{n,\theta}(\mathbf{X}_1, \mathbf{X}_2) = (\mathbf{X}_1 - \boldsymbol{\mu}_X)^T(\mathbf{X}_2 - \boldsymbol{\mu}_X)$ is a centralized linear kernel, then $V_{2n} = E\{K_{n,\theta}^2(\mathbf{X}_1, \mathbf{X}_2)\} = \mathrm{tr}(\mathbf{\Sigma}^2)$. Assume that $r_m \asymp m^{-\beta/2}$. The proposed test is consistent if $p = o\{n^{2/(1-\beta)}\}$ for $0 \leq \beta < 1$. If $\beta = 1$, the condition is $p = o\{\exp(n^2)\}$. If $\beta > 1$, then the proposed test is consistent for any relationship between $p$ and $n$.

**Example 2: (Quadratic Kernel)** Consider the quadratic kernel $K_{n,\theta}^*(\mathbf{X}_1, \mathbf{X}_2) = (\mathbf{X}_1^T\mathbf{X}_2 + 1)^2$. Then, the corresponding centralized kernel is $K_{n,\theta}(\mathbf{X}_1, \mathbf{X}_2) = 2(\mathbf{X}_1 - \boldsymbol{\mu}_X)^T(\mathbf{X}_2 - \boldsymbol{\mu}_X) + (\mathbf{X}_1^T\mathbf{X}_2)^2 - \mathbf{X}_1^T\mathbf{R}\mathbf{X}_1 - \mathbf{X}_2^T\mathbf{R}\mathbf{X}_2 + \mathrm{tr}(\mathbf{R}^2)$, where $\mathbf{R} = \mathbf{\Sigma} + \boldsymbol{\mu}_X\boldsymbol{\mu}_X^T$.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0339

If $\mathbf{X}_1$ and $\mathbf{X}_2$ are multivariate normally distributed with $\boldsymbol{\mu}_X = 0$, then $V_{2n} \asymp \mathrm{tr}^2(\boldsymbol{\Sigma}^2)$. Therefore, the proposed method is consistent if $\mathrm{tr}^2(\boldsymbol{\Sigma}^2) = o(n^2\lambda_{n1})$. If $r_m \asymp m^{-\beta/2}$, the proposed test is consistent if $p = o\{n^{1/(1-\beta)}\}$, for $0 \leq \beta < 1$. If $\beta = 1$, the condition is $p = o\{\exp(n)\}$. If $\beta > 1$, the proposed test is consistent for any relationship between $p$ and $n$.

**Example 3: (Polynomial Kernel)** Consider the polynomial kernel $K^*_{n,\theta}(\mathbf{X}_1, \mathbf{X}_2) = (\mathbf{X}_1^T\mathbf{X}_2)^d$ with a finite $d$. Assume $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent multivariate normally distributed with mean $\boldsymbol{\mu}_\mathbf{X}$ and variance $\boldsymbol{\Sigma}$. Let $\mathbf{X}_1 = \boldsymbol{\Sigma}^{1/2}\mathbf{Z}_1$ and $\boldsymbol{\Sigma} = \boldsymbol{Q}\Lambda\boldsymbol{Q}^T$ be the eigen-decomposition of $\boldsymbol{\Sigma}$, where $\Lambda = \mathrm{diag}(\eta_1, \cdots, \eta_p)$ is a diagonal matrix and $\boldsymbol{Q}$ is the corresponding eigenvector matrix. We then write $(\mathbf{X}_1^T\mathbf{X}_2)^d = (\mathbf{Z}_1^T\Lambda\mathbf{Z}_2)^d$, where $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are independent multivariate normally distributed vectors with mean $\boldsymbol{\mu}^* = \boldsymbol{Q}^T\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}_\mathbf{X}$ and identity covariance. As a result, we consider a polynomial kernel $K^*_{n,\theta}(\mathbf{Z}_1, \mathbf{Z}_2) = (\mathbf{Z}_1^T\Lambda\mathbf{Z}_2)^d$, where $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are independent multivariate distributed normal random vectors with mean $\boldsymbol{\mu}^*$ and covariance $\mathbf{I}_p$. In the Supplemental Material, we show that the centralized kernel of $K^*_{n,\theta}$ is

$$K_{n,\theta}(\mathbf{Z}_1, \mathbf{Z}_2) = \sum_{j_1+j_2+\cdots+j_p=d} \frac{d!}{j_1!\cdots j_p!} \prod_{l=1}^{S_J} \eta_{k_l}^{j_{k_l}} \{Z_{1k_l}^{j_{k_l}} - E(Z_{1k_l}^{j_{k_l}})\}\{Z_{2k_l}^{j_{k_l}} - E(Z_{2k_l}^{j_{k_l}})\},$$

where $j_1, \cdots, j_p$ are non-negative integers and $\{k_1, \cdots, k_{S_J}\}$ is a subset of $\{1, \cdots, p\}$, for which $j_{k_l} \neq 0$ and $l = 1, \cdots, S_J$ and $S_J$ is the number of nonzero integers in the set $J = \{j_1, \cdots, j_p\}$. Here, $\eta_j$ is the $j$th largest eigenvalue of $\Lambda$ and $\mathbf{Z}_1 = (Z_{11}, \cdots, Z_{1p})^T$. In the Supplemental Material, we also show that $V_{2n} \asymp \mathrm{tr}^d(\boldsymbol{\Sigma}^2)$. Therefore, the proposed

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0339

method is consistent if $\mathrm{tr}^d(\mathbf{\Sigma}^2) = o(n^2\lambda_{n1})$. If $r_m \asymp m^{-\beta/2}$, the proposed test is consistent if $p = o\{n^{2/\{d(1-\beta)\}}\}$, for $0 \leq \beta < 1$. If $\beta = 1$, the condition is $p = o\{\exp(n^{2/d})\}$. If $\beta > 1$, the proposed test is consistent for any relationship between $p$ and $n$.

**Example 4: (Gaussian Kernel)** Consider the Gaussian kernel $K^*_{n,\theta}(\mathbf{X}_1, \mathbf{X}_2) = \exp\{-(\mathbf{X}_1-\mathbf{X}_2)^T(\mathbf{X}_1-\mathbf{X}_2)/\theta\}$, with $\mathbf{X}_1$ and $\mathbf{X}_2$ following a normal distribution. The centralized kernel function $K_{n,\theta}(\mathbf{X}_1, \mathbf{X}_2)$ is $K_{n,\theta}(\mathbf{X}_1, \mathbf{X}_2) = \exp\{-(\mathbf{X}_1-\mathbf{X}_2)^T(\mathbf{X}_1-\mathbf{X}_2)/\theta\} - \kappa_1 \sum_{i=1}^2 \exp(-\mathbf{X}_i^T\mathbf{B}\mathbf{X}_i) + \kappa_2$, where $\mathbf{B} = \theta^{-1}\mathbf{I} - 2\theta^{-2}(2\theta^{-1}\mathbf{I}+\mathbf{\Sigma}^{-1})^{-1}$, $\kappa_1 = \Pi_{m=1}^p(2\theta^{-1}\eta_m+1)^{-1/2}$, $\kappa_2 = \Pi_{m=1}^p(4\eta_m/\theta+1)^{-1/2}$, and $\{\eta_m\}_{m=1}^p$ are the eigenvalues of $\mathbf{\Sigma}$. Moreover, $V_{2n} = \Pi_{m=1}^p(8\eta_m/\theta+1)^{-1/2} - 2\Pi_{m=1}^p(2\eta_m/\theta+1)^{-1/2}(6\eta_m/\theta+1)^{-1/2} + \Pi_{m=1}^p(4\eta_m/\theta+1)^{-1}$. When all the eigenvalues of $\mathbf{\Sigma}$ are bounded, we can see that $V_{2n}$ is a constant. Then, the condition $V_{2n} = o(n^2\lambda_{n1})$ is satisfied if $n^2\lambda_{n1}$ diverges. Under this condition, the proposed test is consistent regardless of the relationship between $p$ and $n$.

**Remark 5**: We observe some interesting phenomena from the above examples. If the eigenvalues of the kernel function $K_{n,\theta}$ decay slowly, some restrictions on the relationship between the data dimension and the sample size are needed. This corresponds to the case in which data should be considered as high-dimensional data. If the eigenvalues decay fast enough, we do not need any assumption on the data dimension and sample size. This is the case for functional data or kernel functions that generate sufficiently smooth functional spaces. For linear, quadratic, and polynomial kernels, the eigenvalues of the covariance of $\mathbf{X}$ need to decay fast enough that we can treat $\mathbf{X}$ as functional data. However, if the Gaussian kernel is used, the corresponding functional space is equipped

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0339

with smooth functional spaces, so that we do not need to worry about the data type of $\mathbf{X}$.

## 4. Kernel selection and regularization

To further improve the power of the proposed test, we consider the choice of kernel function and the construction of a regularized kernel in this section.

### 4.1 Kernel selection

In Sections 2–3, we assume that the kernel $K$ that generates the functional space $\mathscr{H}_K$ is known. However, the functional space $\mathscr{H}_K$ is typically unknown. Therefore, an important question in practice is how to select kernels to improve the power of the proposed test. The kernel selection problem has been studied for Fisher discriminant analysis (Kim et al. (2006)) and semi-supervised learning (Dai et al. (2007)). However, no kernel selection method is tailored to the hypothesis testing problem (Liu et al. (2007)).

We propose selecting kernels by maximizing the SNR of the proposed test. The motivation is to choose a kernel with a better SNR, so that the proposed test is more powerful. Because the SNR $\Psi_{\mathbb{K}_\theta}(d_n) = n\delta_{\mathbb{K}_\theta}/(\sigma^2\sigma_{T_n})$, it is equivalent to maximizing $\sigma_{T_n}^{-1}\delta_{\mathbb{K}_\theta}$, because $n$ and $\sigma^2$ do not depend on the kernel $K_\theta$. Therefore, given a family of candidate kernels $\mathscr{F}_{\mathbb{K}}$, the kernel $\mathbb{K}_\theta$ may be selected by maximizing the SNR, as follows:

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0339



$$\widehat{\mathbb{K}}_\theta = \arg\max_{\mathbb{K}_\theta \in \mathscr{F}_\mathbb{K}} \frac{\hat{\delta}_{\mathbb{K}_\theta}}{\hat{\sigma}_{T_n}}. \tag{4.1}$$

For a candidate kernel $\mathbb{K}_\theta \in \mathscr{F}_\mathbb{K}$, the unknown parameters $\delta_{\mathbb{K}_\theta}$ and $\sigma_{T_n}$ can be substituted using estimators, $\hat{\delta}_{\mathbb{K}_\theta} = \{n(n-1)\}^{-1} \sum_{i \neq j} \mathbb{K}_\theta(\mathbf{X}_i, \mathbf{X}_j)(Y_i - \bar{Y}_n)(Y_j - \bar{Y}_n)$ and $\hat{\sigma}^2_{T_n}$ defined in equation (3.3), respectively. These estimators are ratio consistent, as shown in Proposition 1.

Define $\widetilde{\mathbb{K}}_\theta = \arg\max_{\mathbb{K}_\theta \in \mathscr{F}_\mathbb{K}} \delta_{\mathbb{K}_\theta}/\sigma_{T_n}$ as the kernel with the largest SNR in the set $\mathscr{F}_\mathbb{K}$. Let $\mathscr{F}_{\mathbb{K},1}$ be the set of kernels with an SNR at the same order as the SNR of $\widetilde{\mathbb{K}}_\theta$, and $\mathscr{F}_{\mathbb{K},0} = \mathscr{F}_\mathbb{K} / \mathscr{F}_{\mathbb{K},1}$ be the set of kernels in $\mathscr{F}_\mathbb{K}$, but not in $\mathscr{F}_{\mathbb{K},1}$. Assume that all the kernels $K \in \mathscr{F}_{\mathbb{K},0}$ satisfy $|\sigma^{-1}_{T_n,\widetilde{\mathbb{K}}_\theta}\delta_{\widetilde{\mathbb{K}}_\theta} - \sigma^{-1}_{T_n,K}\delta_K| \asymp \sigma^{-1}_{T_n,\widetilde{\mathbb{K}}_\theta}\delta_{\widetilde{\mathbb{K}}_\theta}$. Here, $\sigma^2_{T_n,K}$ is the variance of $T_n$ constructed using kernel $K$. This means that the SNRs of the kernels in $\mathscr{F}_{\mathbb{K},0}$ and $\mathscr{F}_{\mathbb{K},1}$ have distinct orders. Moreover, let $R_{\min} = \min_{K \in \mathscr{F}_{\mathbb{K},0}} |\sigma^{-1}_{T_n,\widetilde{\mathbb{K}}_\theta}\delta_{\widetilde{\mathbb{K}}_\theta} - \sigma^{-1}_{T_n,K}\delta_K| / \sigma^{-1}_{T_n,K}\delta_K$ and $V_{\max,\mathbb{K}_\theta} = \max\big[n^{-1}V_{2n}, \mathrm{Var}\big\{\mathbb{K}_\theta(\mathbf{X}_1,\mathbf{X}_2)h(\mathbf{X}_1)h(\mathbf{X}_2)\big\}, \mathrm{Var}\big\{\mathbb{K}_\theta(\mathbf{X}_1,\mathbf{X}_2)h(\mathbf{X}_1)\big\}\big]$. Define $|\mathscr{F}_\mathbb{K}|$ as the cardinality of the set $\mathscr{F}_\mathbb{K}$. Assume the following condition:

(C5) The kernel $\widetilde{\mathbb{K}}_\theta$ satisfies $V_{\max,\widetilde{\mathbb{K}}_\theta} = o(n\delta^2_{\widetilde{\mathbb{K}}_\theta})$ and $|\mathscr{F}_{\mathbb{K},0}| = o\big\{\min(n\delta^2_{\widetilde{\mathbb{K}}_\theta}/V_{\max,\widetilde{\mathbb{K}}_\theta}, R_{\min})\big\}$.

The above condition (C5) is a mild condition on the SNR of the unknown function $h(\cdot)$ with respect to the kernel $\widetilde{\mathbb{K}}_\theta$. The signal is slightly stronger than those required in the local alternative condition (C4) so that the kernel selection consistency can be established. This is not surprising, because selection consistency typically requires a stronger signal than detection. In the first part of (C5), $V_{\max,\widetilde{\mathbb{K}}_\theta}$ quantifies the variation

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0339



of the estimator $\hat{\delta}_{\widetilde{\mathbb{K}}_\theta}$ whereas $\delta^2_{\widetilde{\mathbb{K}}_\theta}$ measures the signal strength of the projection of the underlying function $h(\cdot)$ to the kernel $\widetilde{\mathbb{K}}_\theta$, It requires that the signal strength is not too small when compared to the variation of its estimator so that the projection $\delta_{\widetilde{\mathbb{K}}_\theta}$ can be estimated consistently.

Note that the proposed kernel selection method is not designed to choose the underlying true kernel that generates the space $\mathscr{H}_K$. In the nonparametric function *estimation* context, if the kernel $\hat{\mathbb{K}}_\theta$ used for estimation is not the same as the underlying true kernel $K$ that generates the functional space $\mathscr{H}_K$, the functional space of the estimated functions $\mathscr{H}_{\hat{\mathbb{K}}_\theta}$ could be different from $\mathscr{H}_K$. However, in the *hypothesis testing* framework, the goal is to distinguish whether the true function $h(\mathbf{X})$ is in $H_0$ or in $H_1$. If $\hat{\mathbb{K}}_\theta \neq K$, the possible impact is that the decisions (reject $H_0$ or fail to reject $H_0$) based on the test statistics constructed using $K$ and $\hat{\mathbb{K}}_\theta$ could be different. The following Proposition 1 proves the ratio consistency of the SNRs by proving the ratio consistency of $\hat{\delta}_{\mathbb{K}_\theta}$ and $\hat{\sigma}^2_{T_n}$. Moreover, we show that a kernel with the same SNR order as $\widetilde{\mathbb{K}}_\theta$ will be selected with probability one, and the proposed kernel selection is consistent in the hypothesis testing context. The proof of Proposition 1 can be found in the Supplemental Material.

**Proposition 1.** *As $n \to \infty$, (i) $\hat{\sigma}^2_{T_n}/\sigma^2_{T_n} \xrightarrow{p} 1$; (ii) if condition (C5) holds, then $\hat{\delta}_{\widetilde{\mathbb{K}}_\theta}/\delta_{\widetilde{\mathbb{K}}_\theta} \xrightarrow{p} 1$ and $\widehat{\mathbb{K}}_\theta \in \mathscr{F}_{\mathbb{K},1}$ with probability one; and (iii) assuming $Var(Y) < \infty$ and the kernel $K$ that generates the RKHS $\mathscr{H}_K$ also satisfies condition (C5), then the proposed kernel selection is consistent in the sense that the decision rule (reject or fail to reject $H_0$) using $T_n$ built on the selected kernel $\hat{\mathbb{K}}_\theta$ is the same as that based on the true*

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0339



*kernel* $K$.

## 4.2 Kernel regularization

In this section, we show that the power of the proposed test can be further improved by using a regularized kernel. The power function is determined by the SNR $\Psi(d_n)$, which can be written as $\Psi(d_n) = n\sum_{m=1}^{\infty}\lambda_{nm}b_{nm}^2/(\sigma^2\sigma_{T_n})$, where $b_{nm} = \mathrm{E}\{d_n(\mathbf{X})\psi_{nm}(\mathbf{X})\}$ is the projection of $d_n(\mathbf{X})$ onto the $m$th eigenfunction $\psi_{nm}(\mathbf{X})$ of $K_{n,\theta}$. We observe that the numerator of $\Psi(d_n)$ (the signal part) is determined by the magnitude of the eigenvalues $\lambda_{nm}$ and the projections $b_{nm}$. For a given set of eigenfunctions $\{\psi_{nm}(x)\}_{m=1}^{\infty}$ and a function $d_n(x)$, the projections $b_{nm}$ are fixed. To increase the numerator of $\Psi(d_n)$, one could adjust the eigenvalues $\lambda_{nm}$ associated with the projection $b_{nm}$ so that larger nonzero projections receive higher weights than small projections do.

To adjust the eigenvalues of the kernel without changing the eigenfunctional space, we introduce a regularized kernel in the following. For any centralized kernel matrix $\mathbf{K}$, define the regularized kernel matrix $\mathbf{K}_{R,\gamma}$ as

$$\mathbf{K}_{R,\gamma} = \mathbf{K} - \mathbf{K}(n\gamma\mathbf{I} + \mathbf{K})^{-1}\mathbf{K}. \tag{4.2}$$

A similar version in a two-sample problem was discussed in Eric et al. (2008). Let $K_{R,\gamma}$ be the kernel function corresponding to the kernel matrix $\mathbf{K}_{R,\gamma}$. It can be proved (see Lemma 4 in the Supplementary Material) that the eigenfunctions of the kernel function $K_{R,\gamma}$ are still $\{\psi_{nm}(\mathbf{X})\}_{m=1}^{\infty}$, which are the same as those of $K_{n,\theta}$. However, the corresponding eigenvalues of $K_{R,\gamma}$ are $\{\gamma\lambda_{nm}/(\lambda_{nm}+\gamma)\}_{m=1}^{\infty}$. According to the definition

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0339



of the RKHS $\mathscr{H}_K$ in Section 2, the space $\mathscr{H}_K$ is mainly determined by the eigenfunctions and eigenvalues. As a result, the function smoothness in the RKHS defined by the regularized kernel could be different to that in the space defined by the unregularized kernel. However, similarly to the kernel selection in the last subsection, note that the regularization does not change the RKHS that generates the true function $h(\cdot)$. It is mainly designed to improve the power of the proposed test.

We now show how a regularized kernel $K_{R,\gamma}$ can improve the power of the proposed test. To see the point, we compare the SNRs $\Psi(d_n)$ and $\Psi_R(d_n,\gamma)$ corresponding to the kernels $K_{n,\theta}$ and $K_{R,\gamma}$, respectively. Let $C_n = n/(\sqrt{2}\sigma^2)$. Then, we have

$$\Psi(d_n) = C_n \frac{\sum_{m=1}^{\infty} \lambda_{nm} b_{nm}^2}{\sqrt{\sum_{m=1}^{\infty} \lambda_{nm}^2}} \quad \text{and} \quad \Psi_R(d_n,\gamma) = C_n \frac{\sum_{m=1}^{\infty} \lambda_{nm} b_{nm}^2/(\lambda_{nm}+\gamma)}{\sqrt{\sum_{m=1}^{\infty} \lambda_{nm}^2/(\lambda_{nm}+\gamma)^2}}. \tag{4.3}$$

By comparing the above two expressions, we see that $\sup_\gamma \Psi_R(d_n,\gamma) \geq \Psi(d_n)$. Because

$$\Psi_R(d_n,\gamma) = C_n \frac{\sum_{m=1}^{\infty} \lambda_{nm} b_{nm}^2/(\lambda_{nm}/\gamma+1)}{\sqrt{\sum_{m=1}^{\infty} \lambda_{nm}^2/(\lambda_{nm}/\gamma+1)^2}} \to \Psi(d_n) \quad \text{as} \quad \gamma \to \infty,$$

the regularized kernel $K_{R,\gamma}$ is the same as the unregularized kernel $K_{n,\theta}$ if $\gamma \to \infty$. Thus, the introduction of the regularization parameter $\gamma$ allows us to strike a balance between the numerator and the denominator so that $\Psi_R(d_n,\gamma)$ is larger than $\Psi(d_n)$ for some $\gamma$.

To select the best regularization parameter $\gamma$, it is natural to consider maximizing the SNR $\Psi_R(d_n,\gamma)$. That is, $\hat{\gamma} = \arg\max_{\gamma\in\mathbb{S}} \hat{\Psi}_R(d_n,\gamma)$, where $\mathbb{S} = \{s_1,\ldots,s_B\}$ is a set of positive candidate regularization parameters ordered in increasing order. Note that the denominator of $\Psi_R(d_n,\gamma)$ in (4.3) goes to infinity and the numerator of the SNR in (4.3) increases as $\gamma \to 0$. A reasonable estimate for the numerator of (4.3) should

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0339



be nondecreasing as $\gamma \to 0$. However, the numerator may not be well estimated if the sample size is small. We therefore propose a modification to the above approach. Let $s_l^* \in \mathbb{S}$ be the smallest regularization parameter in $\mathbb{S}$ such that $\hat{\delta}_{\mathbb{K},\gamma}(d_n)$, the numerator of $\Psi_R(d_n, \gamma)$, achieves its maximum value in $\mathbb{S}$. We then focus on the tuning parameters that are larger than $s_l^*$ in the set of $\mathbb{S}$. Given the samples, we can find the optimal tuning parameter by maximizing the following criterion:

$$\hat{\gamma} = \arg \max_{\gamma \in \{s_l^*, \ldots, s_B\}} \hat{\Psi}_R(d_n, \gamma). \tag{4.4}$$

For the stability selection consideration, we propose the following procedure to select the tuning parameter $\gamma$:

1. Randomly divide the sample $\{Y_i, \mathbf{X}_i\}_{i=1}^n$ into $L$ parts with equal sample sizes.

2. We drop the $l$th ($l = 1, 2, \ldots, L$) part of the sample, select the tuning parameter $\hat{\gamma}_l$ using the remaining $L - 1$ parts of the sample based on criterion (4.4).

3. Repeat step 2 for $l = 1, \ldots, L$. The stabilized tuning parameter is defined as $\tilde{\gamma} = \text{median}\{\hat{\gamma}_1, \ldots, \hat{\gamma}_L\}$.

The simulation studies in Section 5 demonstrate that the above tuning parameter selection method works well in practice. For the regularization parameter $\gamma$, we recommend choosing an interval that satisfies the conditions of Theorem 3 in the Supplementary Material, and then selecting a sequence of values that are discrete uniformly distributed within an appropriate interval to perform the stability selection procedure

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0339

described above. Based on our experience in the simulations, one could choose $L$ between four to eight. Please refer to Section 5.2 for more details. For a given candidate set $\mathbb{S} = \{s_1, \ldots, s_B\}$ for the regularization parameter $\gamma$, define $\mathbb{S}^* = \{s_l^*, \ldots, s_B\} \subset \mathbb{S}$ as the set of regularization parameters used in (4.4). Let $\tilde{\gamma} = \arg\max_{\gamma \in \mathbb{S}^*} \Psi_R(d_n, \gamma)$ and $|\mathbb{S}^*|$ be the cardinality of the set $\mathbb{S}^*$. If the regularized kernels corresponding to $\tilde{\gamma}$ and $|\mathbb{S}^*|$ satisfy the conditions in (C5), then the proposed kernel regularization method also has the consistency established in Proposition 1.

The regularization is most effective in the "sparse" case, in which the nonzero projections reside only in the first $N$ coordinates corresponding to the $N$ largest eigenvalues. *In Section S2 of the Supplementary Material, we show that the SNR $\Psi_R(d_n, \gamma^*)$ of the proposed test with a regularized kernel can attain the SNR of an oracle test within a factor of a slowly varying function* $\log(N)$.

## 5. Simulation study

The simulation studies were designed to evaluate the finite-sample performance of the proposed test for high-dimensional and functional covariates, kernel selection, and regularization methods. We simulated i.i.d. samples $\{\mathbf{X}_i, Y_i\}_{i=1}^n$ from the following model:

$$Y_i = \mu + h(\mathbf{X}_i) + \epsilon_i \quad i = 1, \cdots, n, \tag{5.1}$$

where the random error $\epsilon_i$ is simulated from $N(0, 1)$ or $Laplace(0, \sqrt{2}/2)$. We considered both high-dimensional and functional covariates $\mathbf{X}$. To generate high-dimensional $\mathbf{X}$, we first generated a $p$-dimensional normally distributed random vector $\mathbf{Z}$ with mean zero and

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0339

covariance $\mathbf{\Sigma} = (0.6^{|i-j|})_{i,j=1}^{p}$. Then, we obtained the covariates $\mathbf{X} = (X_1, \cdots, X_p)^T$ by setting the $j$th component using $X_j = F_{nj}(Z_j)$, for $j = 1, \cdots, p$. Here, $F_{nj}$ is the empirical cumulative distribution of the $j$th component given by $F_{nj}(z) = n^{-1}\sum_{i=1}^{n} I(Z_{ij} \leq z)$. To generate the functional covariates $\mathbf{X}$, we first generated a sequence of time points $0 < t_1 < \cdots < t_p < 1$ uniformly from (0,1), and then generated $X_j = X(t_j)$ using the stochastic process $X(t) = \sum_{k=1}^{100}(2\omega_{2k-1})^{1/2}\eta_{2k-1}\cos(2k\pi t) + \sum_{k=1}^{100}(2\omega_{2k})^{1/2}\eta_{2k}\sin(2k\pi t)$, where $\omega_k = 20(k+1.5)^{-3}$ and $\eta_k$s are i.i.d. $N(0,1)$. We considered two settings for the relationship between $n$ and $p$: (i) $p < n$ and (ii) $p >> n$, with $n = 40, 60$, and 100. Specifically, $p = (3, 5, 10)$ in setting (i), and $p = (1500, 3000, 4500)$ in setting (ii). All the results for evaluating the empirical power are based on 1000 simulation replicates and those for the empirical size are based on 5000 simulation replicates. To save space, the simulation results for setting (i) and the simulation studies for the Laplace errors are presented in Section S3 of the Supplementary Material. In all of our simulation and empirical studies, we used the scaled chi-squared approximation discussed in Remark 2 after Theorem 1. In particular, when the data dimension is low, we found that the chi-squared approximation was more accurate than the normal approximation.

We wish to test $H_0 : h(\cdot) = 0$. To assess the empirical size of the proposed test, we chose $h(\mathbf{x}) = 0$ under $H_0$. To evaluate the empirical power, we chose $h(\mathbf{x}) = h_H(\mathbf{x}) - E(h_H)$ in setting (ii), where $h_H(\mathbf{x}) = c_1\sum_{k=1}^{100}(-1)^k x_k + c_2\sum_{k=1}^{100}\{\exp(-x_k^2/p)H_2(x_k/p)\} + c_3\{x_1x_3 + \cos(x_3^2)\}$, where $H_k(\cdot)$ is the $k$th-order Hermite polynomial, and $c_1, c_2$, and $c_3$ are constants specified below. In setting (ii), we designed two scenarios with different

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0339

values of $c_1, c_2$, and $c_3$ for each setting: $\mathcal{S}_3 = \{c_1 = 0.1, c_2 = 100, c_3 = 0.1\}$ and $\mathcal{S}_4 = \{c_1 = 100u, c_2 = 0.1u, c_3 = 0.1u, u = 0.015\}$. In scenario $\mathcal{S}_3$, $c_2$ are chosen to be much larger than $c_1$ such that the nonlinear parts dominate the functions. In $\mathcal{S}_4$, $c_1$ are much larger than $c_2$ so that the linear parts dominate.

Three types of commonly used kernels were compared in all the simulations: the linear kernel $K_L(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T\mathbf{y}/\theta$, Gaussian kernel $K_G(\mathbf{x}, \mathbf{y}) = \exp\{-\|\mathbf{x} - \mathbf{y}\|^2/\theta\}$, and the exponential kernel $K_E(\mathbf{x}, \mathbf{y}) = \exp\{-(\|\mathbf{x}\|^2 + 3\|\mathbf{x} - \mathbf{y}\|^2 + \|\mathbf{y}\|^2)/\theta\}$. The tuning parameter $\theta$ was set to $p$ to make the computation more stable. This choice of $\theta$ is also closely related to the "median heuristic" used in the machine learning literature (see Schölkopf et al. (2002)). In practice, one might also apply the proposed kernel selection method to select the parameter $\theta_0 > 0$ in the tuning parameter $\theta$ with the form $\theta = p\theta_0$. Further discussion on the choice of $\theta$ can be found in Section S3.5 of the Supplementary Material.

Table 1 summarizes the empirical sizes of the proposed test and the test procedure (LLD) proposed by Liu et al. (2007) for high-dimensional and functional covariates. We see that both methods have similar empirical sizes and can control the type-I errors reasonably well. Table 2 contains the empirical power of the proposed test under setting (ii) with high-dimensional covariates. Several observations are given below: 1) There is a clear difference in power between the three types of kernels $K_E, K_G$, and $K_L$, especially when $p$ and $n$ are relatively small. The power difference is particularly striking in Table S2 in the Supplementary Material. The power based on the exponential kernel is higher

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0339

Table 1: Empirical size (in percentages) of the proposed test (Proposed) and the method of Liu et al. (2007) (LLD) for Gaussian errors with high-dimensional and functional covariates using different kernels.

| | | High-dimensional covariates | | | | | | | | | Functional covariates | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $n = 40$ | | | $n = 60$ | | | $n = 100$ | | | $n = 40$ | | | $n = 60$ | | | $n = 100$ | | |
| $p$ | method | $K_E$ | $K_L$ | $K_G$ | $K_E$ | $K_L$ | $K_G$ | $K_E$ | $K_L$ | $K_G$ | $K_E$ | $K_L$ | $K_G$ | $K_E$ | $K_L$ | $K_G$ | $K_E$ | $K_L$ | $K_G$ |
| 1500 | Proposed | 6.2 | 6.2 | 6.2 | 5.3 | 5.1 | 5.1 | 5.4 | 5.1 | 5.2 | 6.2 | 6.1 | 5.9 | 4.7 | 4.9 | 5.0 | 4.5 | 4.4 | 4.5 |
| | LLD | 4.9 | 4.9 | 4.9 | 3.9 | 4.0 | 4.0 | 4.7 | 5.1 | 5.3 | 4.5 | 5.2 | 4.9 | 5.5 | 5.1 | 5.6 | 4.4 | 4.2 | 4.1 |
| 3000 | Proposed | 6.4 | 6.3 | 6.3 | 5.2 | 5.1 | 5.2 | 5.9 | 5.6 | 5.3 | 5.6 | 5.0 | 5.2 | 5.5 | 5.2 | 5.5 | 5.1 | 5.1 | 5.0 |
| | LLD | 4.1 | 4.0 | 4.0 | 5.0 | 4.5 | 4.6 | 5.9 | 5.6 | 5.6 | 4.2 | 4.1 | 4.2 | 5.5 | 5.1 | 5.5 | 4.7 | 4.9 | 4.8 |
| 4500 | Proposed | 6.0 | 6.2 | 6.0 | 5.4 | 5.4 | 5.4 | 5.9 | 6.0 | 6.0 | 5.4 | 6.1 | 5.9 | 5.0 | 4.7 | 4.9 | 5.0 | 5.2 | 5.1 |
| | LLD | 4.7 | 4.4 | 4.3 | 5.4 | 5.3 | 5.4 | 6.1 | 5.9 | 6.0 | 4.7 | 5.1 | 5.0 | 4.4 | 4.1 | 4.1 | 5.1 | 4.8 | 5.0 |

than those using the other kernels. This is understandable, because the nonlinear parts dominate the function $h_L(\mathbf{x})$ (see Section S3.1 in the Supplementary Material) and the exponential and Gaussian kernels contain richer nonlinear eigenfunctions than that of the linear kernel, and can capture more information nonlinear functions; 2) The power increases as the sample size increases in all the cases; and 3) The proposed test is very robust to the change of error distributions. Because the power patterns for the functional and high-dimensional covariates are very similar, we omit the power results for the functional covariates. Additional simulation studies for $p > n$ cases can be found in the Supplementary Material.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0339

Table 2: Empirical power (in percentages) of the proposed test (Proposed) and the method of Liu et al. (2007) (LLD) for Gaussian errors with dependent covariates using different kernels under scenarios $\mathcal{S}_3$ and $\mathcal{S}_4$. The estimated theoretical power is given in parentheses, and the percentage of a kernel being selected among the three candidate kernels is displayed underneath.

| | | | $\mathcal{S}_3$ | | | $\mathcal{S}_4$ | | |
|---|---|---|---|---|---|---|---|---|
| $n$ | $p$ | method | $K_E$ | $K_L$ | $K_G$ | $K_E$ | $K_L$ | $K_G$ |
| 40 | 1500 | Proposed | 50.2(50.2) | 47.2(47.9) | 47.3(48.0) | 57.7(55.5) | 57.2(55.3) | 57.3(55.4) |
| | | | (83.9) | (11.0) | (5.1) | (33.9) | (33.0) | (33.1) |
| | | LLD | 43.6 | 42.7 | 42.8 | 50.7 | 51.5 | 51.7 |
| | 3000 | Proposed | 26.2(32.1) | 25.7(31.7) | 25.5(31.8) | 39.0(41.5) | 38.7(41.4) | 39.1(41.5) |
| | | | (52.1) | (29.2) | (18.7) | (35.2) | (36.9) | (27.9) |
| | | LLD | 20.8 | 21.1 | 21.1 | 32.8 | 34.8 | 34.8 |
| | 4500 | Proposed | 20.6(26.5) | 20.6(26.4) | 20.3(26.4) | 29.4(35.4) | 29.1(35.3) | 29.5(35.3) |
| | | | (39.2) | (41.3) | (19.5) | (38.1) | (39.7) | (22.2) |
| | | LLD | 16.1 | 15.4 | 15.5 | 25.4 | 26.0 | 26.0 |
| 60 | 1500 | Proposed | 76.3(71.1) | 74.1(68.6) | 74.2(68.7) | 84.4(78.5) | 84.3(78.4) | 84.2(78.4) |
| | | | (91.8) | (4.6) | (3.6) | (35.5) | (34.8) | (29.7) |
| | | LLD | 73.4 | 71.9 | 71.8 | 83.0 | 83.9 | 83.7 |
| | 3000 | Proposed | 41.0(43.1) | 40.0(42.5) | 39.9(42.6) | 62.1(59.8) | 61.7(59.7) | 62.1(59.7) |
| | | | (60.0) | (25.1) | (14.9) | (39.0) | (32.5) | (28.5) |
| | | LLD | 36.3 | 36.7 | 36.6 | 59.2 | 59.2 | 59.2 |
| | 4500 | Proposed | 32.2(36.5) | 31.8(36.2) | 32.0(26.3) | 51.3(50.6) | 51.4(50.5) | 50.9(50.6) |
| | | | (46.5) | (33.0) | (20.5) | (37.2) | (36.7) | (26.1) |
| | | LLD | 29.9 | 29.2 | 29.4 | 48.2 | 48.1 | 47.9 |
| 100 | 1500 | Proposed | 98.3(94.7) | 97.7(93.3) | 97.7(93.3) | 99.8(98.2) | 99.8(98.3) | 99.8(98.3) |
| | | | (98.2) | (1.2) | (0.6) | (37.2) | (34.9) | (27.9) |
| | | LLD | 98.3 | 97.6 | 97.6 | 99.7 | 99.7 | 99.7 |
| | 3000 | Proposed | 76.1(69.7) | 75.0(68.9) | 75.0(69.0) | 94.7(88.8) | 94.7(88.8) | 94.7(88.8) |
| | | | (70.3) | (16.7) | (13.0) | (39.9) | (31.1) | (29.0) |
| | | LLD | 74.2 | 74.3 | 74.5 | 93.9 | 94.3 | 94.3 |
| | 4500 | Proposed | 56.0(54.2) | 55.7(53.9) | 55.7(54.0) | 85.4(77.9) | 85.2(77.9) | 85.2(77.9) |
| | | | (52.2) | (26.9) | (20.9) | (42.3) | (30.7) | (27.0) |
| | | LLD | 53.6 | 53.5 | 53.6 | 83.0 | 83.4 | 83.4 |

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0339

### 5.1 Kernel selection

We observed from Table 2 and Tables S2, S3, S5, S6, S7, and S9 in the Supplementary Material that the empirical power of the test corresponding to different kernels can be very different. This naturally motivated us to select a kernel to improve the performance of the test. We applied the kernel selection method proposed in Section 4.1 to choose the optimal kernel among $K_E, K_G$, and $K_L$ for each simulation replicate.

We report the percentage of each kernel being selected in 1000 simulation replicates from among three candidate kernels $K_E, K_G$, and $K_L$. In almost all cases in Table 2 and Tables S2, S3 S5, S6, S7, and S9 in the Supplementary Material, the kernel selection method chooses the kernel with the highest power. This shows that the proposed kernel selection method works very well in selecting the optimal kernel. When the power of the different kernels was similar, the percentages were evenly distributed among the three kernels. To further confirm the validity of the proposed kernel selection method, for each simulation replicate, we estimated the theoretical power of the test using (4.1) for each kernel $K_E, K_L$, and $K_G$. In Table 2 and Tables S2, S3 S5, S6, S7, and S9 in the Supplementary Material, we report the mean of the estimated power for the three kernels based on 1000 simulation replicates. We observe that the estimated theoretical power is very close to the empirical power. In summary, the proposed kernel selection method is reliable for practical use.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0339

### 5.2 Regularization

To show the impact of the kernel regularization on the power improvement, we generated data according to model (5.1), where the random error $\epsilon$ follows a Laplace distribution, and the covariates $\mathbf{X}_i$ are i.i.d. random vectors with independently Uniform (0,1) components. The function $h(\mathbf{x})$ was chosen to be zero under $H_0$. Under the alternative, we chose $h(\mathbf{x}) = h_H(\mathbf{x})$, with the constants $c_1, c_2$, and $c_3$ set according to scenario $\mathcal{S}_3$. In this simulation, the sample size was $n = 60$ and the data dimension was $p = 200$. All the simulation results reported in this part are based on 1000 simulation replicates. To understand the computational cost for the proposed tests with and without regularized kernels, we also summarize the mean and standard deviation of the computation time in Section S3.6 in the Supplementary Material.

For each kernel $K_E$, $K_L$, and $K_G$, we constructed the regularized kernels with the regularization parameter $\gamma$ using (4.2). We selected a sequence of regularization parameters of different orders ($\gamma = 10^{-a}/n$, $a \in (-5, 2)$) to check their effects on the empirical power. For each regularization parameter value, we constructed the corresponding regularized test statistic and applied the test to data generated under $H_0$ and $H_1$. The simulation results for $K_L$ and $K_G$ are summarized in Section S3.4 in the Supplementary Material.

Figure 1 shows the empirical power and size of the proposed test using the regularized kernel $K_{R,\gamma}$. The x-axis represents the $-\log_{10}(\gamma)$, and the y-axis is the empirical power

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0339

or size. The power with large regularization parameters $\gamma$ is not displayed in the graph to enable a better view for small $\gamma$. When $\gamma$ is large $-\log_{10}(\gamma) \in (-3.222, 1.778)$, not shown in Figure 1, the power of the test was the same as that using non-regularized kernels (0.769 for $K_E$), and then started to grow slowly. For $-\log_{10}\gamma \in (1.778, 3.778)$, the power peak (0.810 for $K_E$) of the proposed test can be observed for all three kernels. It can be seen from Figure 1 that the empirical size of the regularized test is reasonably controlled.

To evaluate the method for selecting the regularization parameters proposed in Section 4.2, we also mark the regularization parameter selection results in Figure 1. The three vertical lines correspond to the first quantile ($Q_1$), median, and third quantile ($Q_3$), respectively, of the stabilized $\tilde{\gamma}$ obtained from the 1000 simulation replicates, where $L = 5$ was chosen in the stability selection. It can be seen from Figure 1 that the vertical lines are all very close to where the maximum power is achieved. This suggests that the proposed regularization selection method can locate the optimal regularization parameter to maximize the power of the proposed test.

## 6. An empirical study

We applied the proposed test to a Yorkshire gilt data set to find gene sets that are associated with triiodothyronine ($T_3$), which is an important thyroid hormone affecting growth and metabolism in the body. A total of 24,123 gene expressions were measured using liver tissues for 24 six-month-old Yorkshire gilts, whose $T_3$ levels in blood were

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0339

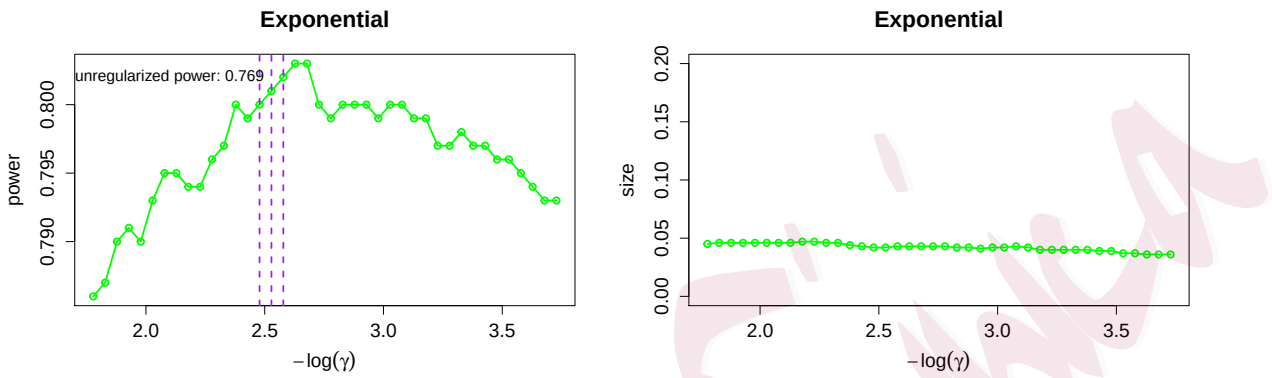


Figure 1: The empirical power (left panel) and size (right panel) for regularized kernels, where the vertical purple lines in the left panel denote the first, second, and third quantiles of the selected regularization parameters among 1000 simulation replicates. For each replicate, the regularization parameter was selected by the method introduced in Section 4.2.

also recorded. All the genes in the Yorkshire gilt data set were classified into 6176 gene ontology (GO) terms (gene sets), where each gene could be assigned to several GO terms according to its gene attributes in one of three domains: cellular component, molecular function, and biological process. More details about the data set can be found in Lkhagvadorj et al. (2009).

Let $Y_i$ and $\mathbf{X}_i^{(k)} = (\mathbf{X}_{i1}^{(k)}, \mathbf{X}_{i2}^{(k)}, \cdots, \mathbf{X}_{ip_k}^{(k)})^T$ be the measure of the $T_3$ level for the $i$th gilt and the standardized gene expression vector of the $k$th GO term for the $i$th gilt, respectively, where $p_k$ is the total number of genes in the $k$th GO term. Among the 6176 GO terms, 560 have $p_k$ larger than the sample size 24, with first quantile 36.75, median 60, third quantile 125.25, and maximum 5158. Our proposed methods work for both $p_k > n$ and $p_k < n$ cases. Simulation studies for $p_k > n$ cases are reported in Section 5 and Section S3 in the Supplementary Material, and simulation studies for

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0339

$p_k < n$ are included in Tables S1–S3 in Section S3 of the Supplemental Material. We consider the following nonparametric regression model $Y_i = \mu^{(k)} + h^{(k)}(\mathbf{X}_i^{(k)}) + \epsilon_i^{(k)}$, for $i = 1, \cdots, 24$ and $k = 1, \cdots, 6176$. For the $k$th GO term, we are interested in testing $H_0 : h^{(k)}(\cdot) = 0$ vs. $H_1 : h^{(k)}(\cdot) \neq 0$.

To apply our proposed kernel selection and regularization procedure, we applied the multiple splitting procedure in Meinshausen et al. (2009) to avoid double dipping. We randomly split the sample $B = 50$ times. For each split, the first half of the sample was used to search for the best combination of kernel function and regularization parameter $\gamma$ using our proposed methods in Section 4. The second half was used to perform the proposed hypothesis testing based on the selected regularized kernel from the first half. Specifically, we considered four different centralized kernels: the exponential kernel $K_E$, Gaussian kernel $K_G$, linear kernel $K_L$, and polynomial kernel $K_P$, where $K_E, K_G$, and $K_L$ are defined in Section 5, and $K_P(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^T \mathbf{x}_j/\theta)^2$ and $\theta$ was set as the dimension of $\mathbf{X}$ for each kernel. The regularization parameter $\gamma$ was set as $10^a$, where $a \in \{-3.00, -2.95, \ldots, 4.95, 5\}$. For each GO term, we obtained $B$ p-values from $B$ subsamples. These $B$ p-values were then aggregated into one p-value using the empirical quantile function of p-values (see Meinshausen et al. (2009)). For comparisons, we also applied LLD (Liu et al. (2007)) with the same centralized kernels. After controlling the false discovery rate at level 0.01 (Storey and Tibshirani (2003)), the proposed method declared 58 statistically significant GO terms, while the LLD test only identified 13 significant GO terms using the centralized Gaussian kernel. However, the LLD method with

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0339

the exponential, linear, and polynomial kernels did not find any significant GO terms. This indicates the advantages of the proposed approach. The two methods share five of the significant GO terms discovered.

## 7. Discussion

We have modeled the joint effect of high-dimensional or functional covariates in a set using a nonparametric function in an RKHS. We have addressed a fundamental question about testing nonparametric functions of high-dimensional data, without assuming any model structures. We proposed a nonparametric test for assessing the significance of a nonparametric function. In contrast to previous investigations, our method can be applied to both high-dimensional and functional data. We derived the asymptotic distributions of the test statistic under the null hypothesis and a sequence of local alternative hypotheses, and found the explicit effects of kernel functions and types of covariates on the asymptotic distributions.

Based on the obtained explicit power function, we proposed a kernel selection method designed to improve the power of the proposed test. Moreover, we introduced a test with the regularized kernel that can further improve the power and enhance the dimensionality the test can handle. It was shown that the regularized kernel plays a similar role to that of a re-weighting method that adds large weights to nonzero projections of the nonparametric function to the orthogonal bases of the RKHS. With a properly chosen regularization parameter, we demonstrated that the proposed test can achieve almost the

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0339



same power as the oracle test. A practical method for selecting regularization parameters was also introduced. Our method was motivated and further demonstrated by a genomic study. However, it can be broadly applied to other areas in which high-dimensional or functional data are routinely generated.

## Supplementary Material

Technical proofs, more details about the regularized kernel and its oracle property, and some additional simulation results are included in the Supplementary Material. An associated R package *"KerUTest"* is available on `https://github.com/hetao12/KerUTest`.

## Acknowledgments

We are grateful to the editor, associate editor, and two referees for their insightful, constructive, and careful comments. The authors acknowledge the support from NIH grants 1R21HG010073 and 1R01GM131398, and an NSF grant DMS-2137983.

## References

Avery, M., Y. Wu, H. H. Zhang, and J. Zhang (2014). Rkhs-based functional nonparametric regression for sparse and irregular longitudinal data. *Canadian Journal of Statistics 42*(2), 204–216.

Bai, Z. and H. Saranadasa (1996). Effect of high dimension: by an example of a two sample problem. *Statistica Sinica*, 311–329.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0339



Chen, K., K. Chen, H.-G. Müller, and J.-L. Wang (2011). Stringing high-dimensional data for functional analysis. *Journal of the American Statistical Association 106*(493), 275–284.

Chen, S. X., W. Härdle, and M. Li (2003). An empirical likelihood goodness-of-fit test for time series. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 65*(3), 663–678.

Cucker, F. and S. Smale (2002). On the mathematical foundations of learning. *Bulletin of the American mathematical society 39*(1), 1–49.

Dai, G., D.-Y. Yeung, and Y.-T. Qian (2007). Face recognition using a kernel fractional-step discriminant analysis algorithm. *Pattern recognition 40*(1), 229–243.

Delsol, L. (2012). No effect tests in regression on functional variable and some applications to spectrometric studies. *Computational Statistics 28*(4), 1–37.

Delsol, L., F. Ferraty, and P. Vieu (2011). Structural test in regression on functional variables. *Journal of Multivariate Analysis*, 422–447.

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. (1996). Test of significance based on wavelet thresholding and neyman's truncation. *Journal of the American Statistical Association 91*(434), 674–688.

Fan, J. (2018). *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*. Routledge.

Fan, J., C. Zhang, and J. Zhang (2001). Generalized likelihood ratio statistics and wilks phenomenon. *The Annals of Statistics 29*(1), 153–193.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0339



Gao, J. and I. Gijbels (2008). Bandwidth selection in nonparametric kernel testing. *Journal of the American Statistical Association 103*(484), 1584–1594.

Kim, S.-J., A. Magnani, and S. Boyd (2006). Optimal kernel selection in kernel fisher discriminant analysis. In *Proceedings of the 23rd international conference on Machine learning*, pp. 465–472. ACM.

Kong, D., A.-M. Staicu, and A. Maity (2016). Classical testing in functional linear models. *Journal of Nonparametric Statistics 28*(4), 813–838. PMID: 28955155.

Lan, W., H. Wang, and C.-L. Tsai (2014). Testing covariates in high-dimensional regression. *Annals of the Institute of Statistical Mathematics 66*(2), 279–301.

Li, S., Y. Cui, et al. (2012). Gene-centric gene–gene interaction: A model-based kernel machine method. *The Annals of Applied Statistics 6*(3), 1134–1161.

Li, T. and Z. Zhu (2020, jan). Inference for generalized partial functional linear regression. *Statistica Sinica*.

Lian, H. (2007). Nonlinear functional models for functional responses in reproducing kernel hilbert spaces. *Canadian Journal of Statistics 35*, 597–606.

Lindsay, B. G., M. Markatou, S. Ray, K. Yang, S.-C. Chen, et al. (2008). Quadratic distances on probabilities: A unified foundation. *The Annals of Statistics 36*(2), 983–1006.

Liu, D., D. Ghosh, and X. Lin (2008). Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. *BMC bioinformatics 9*(1), 292.

Liu, D., X. Lin, and D. Ghosh (2007). Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. *Biometrics 63*(4), 1079–1088.

Liu, M., Z. Shang, and G. Cheng (2018). Nonparametric testing under random projection.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0339

# REFERENCES

Lkhagvadorj, S., L. Qu, W. Cai, O. P. Couture, C. R. Barb, G. J. Hausman, D. Nettleton, L. L. Anderson, J. C. Dekkers, and C. K. Tuggle (2009). Microarray gene expression profiles of fasting induced changes in liver and adipose tissues of pigs expressing the melanocortin-4 receptor d298n variant. *Physiological genomics 38*(1), 98–111.

Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, et al. (2009). Finding the missing heritability of complex diseases. *Nature 461*(7265), 747.

Meinshausen, N., L. Meier, and P. Bühlmann (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association 104*(488), 1671–1681.

Ramsay, J. and B. Silverman (2005). *Functional data analysis*. Springer.

Schölkopf, B., A. J. Smola, F. Bach, et al. (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.

Shang, Z. and G. Cheng (2013, 10). Local and global asymptotic inference in smoothing spline models. *The Annals of Statistics 41*(5), 2608–2638.

Shang, Z. and G. Cheng (2015). Nonparametric inference in generalized functional linear models. *The Annals of Statistics 43*(4), 1742 – 1773.

Storey, J. D. and R. Tibshirani (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences 100*(16), 9440–9445.

Su, Y.-R., C.-Z. Di, and L. Hsu (2017). Hypothesis testing in functional linear models. *Biometrics 73*(2), 551–561.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0339

## REFERENCES

Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences 102*(43), 15545–15550.

Tekbudak, M., M. Alfaro-Córdoba, A. Maity, and A.-M. Staicu (2019, Semtember). A comparison of testing methods in scalar-on-function regression. *AStA Advances in Statistical Analysis 103*, 411–436.

Wahba, G. (1990). *Spline models for observational data*. SIAM Press.

Wang, S. and H. Cui (2013). Generalized f test for high dimensional linear regression coefficients. *Journal of Multivariate Analysis 117*, 134–149.

Yang, Y., Z. Shang, and G. Cheng (2020, 09–12 Jul). Non-asymptotic analysis for nonparametric testing. In J. Abernethy and S. Agarwal (Eds.), *Proceedings of Thirty Third Conference on Learning Theory*, Volume 125 of *Proceedings of Machine Learning Research*, pp. 3709–3755. PMLR.

Zhong, P.-S. and S. X. Chen (2011). Tests for high-dimensional regression coefficients with factorial designs. *Journal of the American Statistical Association 106*(493), 260–274.

Department of Mathematics, San Francisco State University

E-mail: hetao@sfsu.edu

Department of Mathematics, Statistics and Computer Science, University of Illinois at Chicago

E-mail: pszhong@uic.edu

Department of Statistics and Probability, Michigan State University

E-mail: cuiy@msu.edu and mandrek1@msu.edu