ORIGINAL ARTICLE

Kernel mean embedding of probability measures and its applications to functional data analysis

Saeed Hayati1 | Kenji Fukumizu2 | Afshin Parvardeh1

1Department of Statistics, Faculty of Mathematics and Statistics, University of Isfahan, Isfahan, Iran
2The Institute of Statistical Mathematics Tachikawa, Tokyo, Japan

Abstract
This study intends to introduce kernel mean embedding of probability measures over infinite-dimensional separable Hilbert spaces induced by functional response statistical models. The embedded function represents the concentration of probability measures in small open neighborhoods, which identifies a pseudo-likelihood and fosters a rich framework for statistical inference. Utilizing Maximum Mean Discrepancy, we devise new tests in functional response models. The performance of new derived tests is evaluated against competitors in three major problems in functional data analysis including Function-on-Scalar regression, functional one-way ANOVA, and equality of covariance operators.

KEYWORDS
equality of covariance operators, functional one-way ANOVA, functional regression, maximum mean discrepancy

1 | INTRODUCTION

Functional response models are among the major problems in the context of Functional Data Analysis. A fundamental issue in dealing with functional response statistical models arises due to the lack of practical frameworks on characterizing probability measure on function spaces. This is mainly a consequence of the tremendous gap on how we present probability measures in finite-dimensional and infinite-dimensional spaces.
A useful property of finite-dimensional spaces is the existence of a locally finite, strictly positive, and translation invariant measure like Lebesgue or counting measure, which makes it easy to take advantage of probability measures directly in the statistical inference. Fitting as a statistical model, and estimating parameters, hypothesis testing, deriving confidence regions and developing goodness of fit indices, all can be applied by integrating distribution or conditional distribution of response variables as a presumption into statistical procedures.

Sporadic efforts have been gone into approximating or representing probability measures on infinite-dimensional spaces. Let $\mathbb{H}$ be a separable infinite-dimensional Hilbert space and $X$ be a $\mathbb{H}$-valued random element with finite second moment and covariance operator $C$. Delaigle and Hall (2010) approximated probability of $B_r(x) = \{y \in \mathbb{H} : \|y - x\|_\mathbb{H} < r\}$ by the surrogate density of a finite-dimensional approximated version of $X$, obtained by projecting the random element $X$ into a space spanned by first few eigenfunctions of $C$ with largest eigenvalues. The approximated small-ball probability is on the basis of Karhunen–Loève expansion and putting an extra assumption that the component scores are independent. The precision of this approximation depends on the volume of ball and probability measure itself.

Let $I$ be a compact subset of $\mathbb{R}$ such as closed interval $[0, 1]$ and $X$ be a zero mean $L^2[I]$-valued random element with finite second moment and Karhunen–Loève expansion $X = \sum_{j \geq 1} \lambda_j^{1/2} X_j \psi_j$, in which $X_j = \lambda_j^{-1/2} \langle X, \psi_j \rangle$ and $\{ \lambda_j, \psi_j \}_{j \geq 1}$ is the eigensystem of covariance operator $C$. Suppose that the distribution of $X_j$ is absolutely continuous with respect to the Lebesgue measure with density $f_j$. Approximation of the logarithm of $p(x | r) = P(B_r(x)) = P(\{|X - x|_{L^2[I]} < r\})$ given by Delaigle and Hall (2010) is

$$\log p(x | r) = C_1(h, \{ \lambda_j \}_{j \geq 1}) + \sum_{j=1}^{h} \log f_j(x_j) + o(h),$$

in which $x_j = \langle x, \psi_j \rangle_{L^2[I]}$, and $h$ is the number of eigenvalues larger than or equal to $\sqrt{r}$ that tends to infinity as $r$ declines to zero. $C_1(\cdot)$ depends only on size of the ball and sequence of eigenvalues, though the quantity $o(h)$ as the precision of approximation depends on $P$.

The quantity $\sum_{j=1}^{h} \log f_j(x_j)$ is called log-density by Delaigle and Hall (2010). A serious concern with this approximation is its precision, which depends on the probability measure itself. Accordingly, it cannot be employed to compare small-ball probability in a family of probability measures. For example, in the case of estimating the parameters in a functional response regression model, the induced probability measure varies with different choices of parameters. Thus this approximation cannot be employed for parameter estimation and comparing the goodness of fit of different regression models.

Another work in representing probability measures on a general separable Hilbert space $\mathbb{H}$ presented by Lin et al. (2018). They constructed a dense subspace of $\mathbb{H}$ called Mixture Inner Product Space (MIPS), which is the union of a countable collection of finite-dimensional subspaces of $\mathbb{H}$. An approximating version of the given $\mathbb{H}$-valued random element lies in this subspace, which in consequence, lies in a finite-dimensional subspace of $\mathbb{H}$ according to a given discrete distribution. They defined a base measure on the MIPS, which is not translation-invariant, and introduced density functions for the MIPS-valued random elements.

The absence of a proper method in representing probability measures over infinite-dimensional spaces caused severe problems to statistical inference. To make it clear, as an example Greven et al. (2017) developed a general framework for functional additive mixed-effect regression models. They considered a log-likelihood function by summing up the log-likelihood
of response functions $Y_i$ at a grid of time-points $t_{id}$, $d = 1, \ldots, D_i$, assuming $Y_i(t_{id})$ to be independent within the grid of time-points. A simulation study by Kokoszka and Reimherr (2017) revealed the weak performance of the proposed framework in statistical hypothesis testing in a simple Gaussian Function-on-Scalar linear regression problem.

Currently, MLE and other density-based methods are out of reach in the context of functional response models. In this study, we follow a different path by identifying probability measures with their kernel mean functions and introduce a framework for statistical inference in infinite-dimensional spaces. A promising fact about the kernel mean functions, that is shown in this paper, is their ability to reflect the concentration of probability measures in small open neighborhoods, which unlike the approach of Delaigle and Hall (2010) is comparable among different probability measures. This property of kernel mean function motivates us to make use of it in fitting statistical models in the context of functional data analysis which consequently can be employed to build estimators for functional parameters. These estimators can then be used to introduce statistical tests for corresponding functional parameters.

This paper is organized as follows: In Section 2, kernel mean embedding of probability measures over infinite-dimensional separable Hilbert spaces is discussed. In Section 3 the problem of estimation by means of kernel mean embedding is discussed. The Maximum Kernel Mean (MKM) estimation method is introduced and estimators for Gaussian Response Regression models are derived. In Section 4, new statistical tests are developed employing Maximum Mean Discrepancy (MMD) metric and the estimators obtained in Section 3 for three major problems in functional data analysis and their performance evaluated using simulation studies. Section 5 has been devoted to discussion and conclusion. Major proofs are aggregated in the Appendix A.

## 2 | KERNEL MEAN EMBEDDING OF PROBABILITY MEASURES

We summarize the basics of kernel mean embedding. See Muandet et al. (2017) for a general reference. Let $(\mathbb{H}, B(\mathbb{H})), \mathcal{P}$ be a probability measure space. The space $\mathbb{H}$ is an infinite-dimensional separable Hilbert space equipped with inner product $\langle \cdot, \cdot \rangle_\mathbb{H}$, and $B(\mathbb{H})$ is the Borel $\sigma$-field. A function $k: \mathbb{H} \times \mathbb{H} \to \mathbb{R}$ is a positive definite kernel if it is symmetric, that is, $k(x, y) = k(y, x)$ and $\sum_{i,j=1}^n a_i a_j k(x_i, x_j) \geq 0$ for all $n \in \mathbb{N}$ and $a_i \in \mathbb{R}$ and $x_i \in \mathbb{H}$. $k$ is strictly positive definite if equality implies $a_1 = a_2 = \cdots = a_n = 0$. $k$ is said to be integrally strictly positive definite if $\int_{\mathbb{H}} k(x, y) \mu(dx) \mu(dy) > 0$ for any nonzero finite signed measure $\mu$ defined over the measure space $(\mathbb{H}, B(\mathbb{H}))$. Any integrally strictly positive definite kernel is strictly positive definite while the converse is not true (Striperumbudur et al., 2010). A positive definite kernel induces a Hilbert space of functions over $\mathbb{H}$, which is called Reproducing Kernel Hilbert Space (RKHS) and equals to $\mathcal{H}_k = \text{span}\{k(x, \cdot); x \in \mathbb{H}\}$ with inner product

$$\left\langle \sum_{i \geq 1} a_i k(x_i, \cdot), \sum_{j \geq 1} b_j k(y_j, \cdot) \right\rangle_{\mathcal{H}_k} = \sum_{i \geq 1} \sum_{j \geq 1} a_i b_j k(x_i, y_j).$$

For each $f \in \mathcal{H}_k$ and $x \in \mathbb{H}$ we have $f(x) = \langle f, k(\cdot, x) \rangle_{\mathcal{H}_k}$, which is the reproducing property of kernel $k$. A strictly positive definite kernel $k$ is said to be characteristic for a family of probability measures $\mathcal{P}$ if the map
\[ m : \mathcal{P} \rightarrow \mathcal{H}_k \quad \mathbb{P} \mapsto \int_{\mathbb{H}} k(x, \cdot) \mathbb{P}(dx), \]

which is defined as a Bochner integral, is injective. If \( \mathbb{E}_{\mathbb{P}}(\sqrt{k(X,X)}) < \infty \) then \( m_{\mathbb{P}}(\cdot) := (m(\mathbb{P}))(\cdot) \) exists in \( \mathcal{H}_k \) (Muandet et al., 2017), and the function \( m_{\mathbb{P}}(\cdot) = \int_{\mathbb{H}} k(x, \cdot) \mathbb{P}(dx) \) is called kernel mean function. Hereafter, for a given positive definite kernel \( \mathbb{E}_{\mathbb{P}} \) and for given two probability measures \( \mathbb{P} \) and \( \mathbb{Q} \), bounded and translation-invariant, then the kernel mean function \( m \) then exists. The next theorem and corollary. Proofs are provided in the Appendix A.

If the kernel \( k \) is characteristic then every probability measure \( \mathbb{P} \in \mathcal{P} \) is uniquely identified by an element \( m_{\mathbb{P}} \) of \( \mathcal{H}_k \) and MMD defined as

\[
\text{MMD}(\mathcal{H}_k, \mathbb{P}, \mathbb{Q}) = \sup_{f \in \mathcal{H}_k, \|f\|_{\mathcal{H}_k} \leq 1} \left\{ \int_{\mathbb{H}} f(x) \mathbb{P}(dx) - \int_{\mathbb{H}} f(x) \mathbb{Q}(dx) \right\}
= \sup_{f \in \mathcal{H}_k, \|f\|_{\mathcal{H}_k} \leq 1} \langle f, m_{\mathbb{P}} - m_{\mathbb{Q}} \rangle_{\mathcal{H}_k} = \|m_{\mathbb{P}} - m_{\mathbb{Q}}\|_{\mathcal{H}_k},
\]

is a metric on the family of measures \( \mathcal{P} \) (Muandet et al., 2017).

A similar quantity called Ball divergence is proposed by Pan et al. (2018) to distinguish probability measures defined over separable Banach spaces. For the case of infinite-dimensional spaces, Ball divergence distinguishes two probability measures if at least one of them possesses a full support, that is, \( \text{Supp}(\mathbb{P}) = \mathbb{H} \). They employed Ball divergence for a two-sample test, which according to their simulation results, the performance of both MMD and Ball divergence are close and superior to other tests.

Kernel mean functions can also be used to reflect the concentration of probability measures in small-balls, if the kernel function is translation-invariant. A positive definite kernel \( k \) is called translation-invariant if \( k(x,y) = \psi(x-y) \) for some positive definite function \( \psi \). Gaussian kernel \( e^{-\frac{1}{\sigma^2} \|x-y\|^2_{\mathbb{H}}} \) for \( \sigma > 0 \) is such an example. If we choose a continuous characteristic kernel that is bounded and translation-invariant, then the kernel mean function \( m_{\mathbb{P}} \) can be employed to represent the concentration of probability measure in different points of Hilbert space \( \mathbb{H} \). For example, consider

\[
m_{\mathbb{P}}(x) = \int_{\mathbb{H}} e^{-\frac{1}{\sigma^2} \|x-y\|^2_{\mathbb{H}}} \mathbb{P}(dy).
\]

If \( m_{\mathbb{P}}(\cdot) \) has an explicit form for a family of probability measures then \( m_{\mathbb{P}}(\cdot) \) can be employed to study and compare different probability measures. For example, if \( m_{\mathbb{P}_1}(x_1) > m_{\mathbb{P}_2}(x_2) \) then it could be concluded that the concentration of probability measure \( \mathbb{P} \) around the point \( x_1 \) is higher than \( x_2 \), and if for given two probability measures \( \mathbb{P}_1 \) and \( \mathbb{P}_2 \) we had \( m_{\mathbb{P}_1}(x) > m_{\mathbb{P}_2}(x) \) then we conclude that the concentration of probability measure \( \mathbb{P}_1 \) around the point \( x \) is higher than that of probability measure \( \mathbb{P}_2 \). This property of kernel mean functions makes them a good candidate to represent probability measures in infinite-dimensional spaces.

The representation property of probability measures by kernel mean functions is addressed in the next theorem and corollary. Proofs are provided in the Appendix A.

**Theorem 1.** Let \( \mathbb{P}_1 \) and \( \mathbb{P}_2 \) be two probability measures on a separable Hilbert space \( \mathbb{H} \) over the field \( \mathbb{R} \). Let \( k(x,y) = \psi(\|x-y\|_{\mathbb{H}}) \) be a translation-invariant characteristic kernel such that \( \psi : \mathbb{R}^+ \rightarrow [0,1] \), where \( \mathbb{R}^+ \) is the nonnegative reals, is a bounded continuous, strictly decreasing to zero and positive definite function, for example, \( \psi(t) = e^{-t^2} \).
Let \( m_{\mathbb{P}_1} (\cdot) \) and \( m_{\mathbb{P}_2} (\cdot) \) be the kernel mean embedding of \( \mathbb{P}_1 \) and \( \mathbb{P}_2 \), respectively, for the kernel \( k(\cdot, \cdot) \). If \( m_{\mathbb{P}_2} (y) > m_{\mathbb{P}_1} (y) \) for a given \( y \in \mathbb{H} \), then there exists an open ball \( B_r(y) \) such that \( \mathbb{P}_2(B_r(y)) > \mathbb{P}_1(B_r(y)) \), and \( r \) depends only on the difference \( m_{\mathbb{P}_2} (y) - m_{\mathbb{P}_1} (y) \) and the characteristic kernel itself.

**Corollary 1.** Let \( \mathbb{P} \) be a probability measure on a separable Hilbert space \( \mathbb{H} \) over the field \( \mathbb{R} \). Let \( k(x, y) = \psi(||x - y||_2) \) be a translation-invariant characteristic kernel such that \( \psi : \mathbb{R}^+ \to [0, 1] \) is a bounded continuous, strictly decreasing to zero and positive definite function, for example, \( \psi(t) = e^{-t^2} \), and let \( m_{\mathbb{P}} (\cdot) \) be the kernel mean embedding of \( \mathbb{P} \) for the kernel \( k(\cdot, \cdot) \). If \( m_{\mathbb{P}} (y_2) > m_{\mathbb{P}} (y_1) \) for some \( y_1, y_2 \in \mathbb{H} \), then there exist open balls of the same size \( B_r(y_1) \) and \( B_r(y_2) \) such that \( \mathbb{P}(B_r(y_2)) > \mathbb{P}(B_r(y_1)) \), and \( r \) depends only on the difference \( m_{\mathbb{P}} (y_2) - m_{\mathbb{P}} (y_1) \) and the characteristic kernel itself.

Kernel Mean Embedding of probability measures also has a connection with kernel scoring rules. Proper Scoring Rules are well-established instruments with applications in assessing probability models (Gneiting & Raftery, 2007). The following definition is borrowed from Steinwart and Ziegel (2019) and adapted to our context. In the following definition, \( c_{00} \) is the infinite-dimensional inner product space of sequences vanishing at infinity, which is a dense subspace of \( \ell_2 \).

**Definition 1.** Let \( \mathbb{X} \) be an arbitrary measurable space. Here it may be considered to be either the separable Hilbert space \( \ell_2 \) or the separable inner product space \( c_{00} \), and let \( \mathcal{M}_1 (\mathbb{X}) \) be the space of probability measures on \( \mathbb{X} \). For \( \mathcal{P} \subseteq \mathcal{M}_1 (\mathbb{X}) \), a scoring rule is defined as a function \( S : \mathcal{P} \times \mathbb{X} \to [-\infty, \infty] \) such that the integral \( \int_{\mathbb{X}} S(\mathbb{P}, x) \mathbb{Q}(dx) \) exists for all \( \mathbb{P}, \mathbb{Q} \in \mathcal{P} \). The scoring rule is proper if

\[
\int_{\mathbb{X}} S(\mathbb{P}, x) \mathbb{P}(dx) \leq \int_{\mathbb{X}} S(\mathbb{Q}, x) \mathbb{P}(dx), \quad \forall \mathbb{P}, \mathbb{Q} \in \mathcal{P},
\]

and is called strictly proper if the equality implies \( \mathbb{P} = \mathbb{Q} \).

**Kernel scores** are a general class of proper scoring rules, in which the scoring rule is generated by a symmetric positive definite kernel \( k : \mathbb{X} \times \mathbb{X} \to \mathbb{R} \) by

\[
S_k(\mathbb{P}, x) := -\int_{\mathbb{X}} k(x, \omega) d\mathbb{P}(d\omega) + \frac{1}{2} \int_{\mathbb{X}} \int_{\mathbb{X}} k(\omega, \omega') \mathbb{P}(d\omega) \mathbb{P}(d\omega')
= -m_\mathbb{P}(x) + \frac{1}{2} \norm{m_\mathbb{P}}^2.
\tag{2}
\]

The MMD distance between \( \mathbb{P}, \mathbb{Q} \in \mathcal{P} \) satisfies

\[
\norm{m_\mathbb{P} - m_\mathbb{Q}}_{\mathcal{H}_k}^2 = 2 \left( \int_{\mathbb{X}} S_k(\mathbb{Q}, x) \mathbb{P}(dx) - \int_{\mathbb{X}} S_k(\mathbb{P}, x) \mathbb{P}(dx) \right).
\tag{3}
\]

In effect, a kernel score rule \( S_k \) is a strictly proper scoring rule if and only if kernel mean embedding is injective or \( k \) to be characteristic.

There are a plethora of studies on the different class of characteristic kernels over locally compact spaces. For example, Steinwart (2001) proved that Gaussian kernel is characteristic on compact sets, Sriperumbudur et al. (2010, Thm. 9) showed that Gaussian kernel is characteristic on the whole space \( \mathbb{R}^d \), and Simon-Gabriel and Schölkopf (2018) studied the connection.
between various concepts of kernels such as universality, characteristic, and positive definiteness of kernels.

Given a separable Hilbert space \( H \), any integrally strictly positive definite kernel is characteristic (Sriperumbudur et al., 2010, Thm. 7), however, it is not clear which kernels are integrally strictly positive definite over infinite-dimensional separable Hilbert spaces. To the best of our knowledge, there is no study on existence and construction of characteristic kernels for infinite-dimensional spaces. The following two theorems, proofs of which are provided in Appendix A, try to tackle this problem. In Theorem 2, the result of Steinwart and Ziegel (2019, Thm. 3.14) is used to show the existence of a continuous characteristic kernel for infinite-dimensional separable Hilbert spaces, and Theorem 3 shows that Gaussian kernel is characteristic for \( c_{00} \), the infinite-dimensional inner product space of sequences vanishing at infinity, which is dense in \( \ell_2 \).

**Theorem 2.** Let \( H \) be an infinite-dimensional separable Hilbert space. There exists a continuous characteristic kernel on \( H \).

**Theorem 3.** Let \( c_{00} \) be the space of eventually zero sequences. The Gaussian kernel defined as \( k(x, y) = e^{-\frac{1}{2\sigma} \|x - y\|^2} \) for some \( \sigma > 0 \) is characteristic on \( c_{00} \).

Besides what are presented in Theorems 2 and 3, we show in Proposition 1 that Gaussian kernel is characteristic for the family of Gaussian probability measures over \( H \).

## 3 | MKM ESTIMATION

In the context of multivariate statistics, the density function is considered as one of the most ubiquitous tools in statistical inference. Density is a nonnegative function, which represents the amount of probability mass in a point or concentration of probability measure in a very small neighborhood. Typically a nominated family of probability measures is presented by the corresponding family of densities, and the aim is to choose a density from this family, which is the most likely one that generates a set of observations obtained by a probability-based survey sample. The aforementioned family of probability measures usually parameterized by a parameter \( \theta \) taking value either in a subset \( \Theta \) of a finite-dimensional or infinite-dimensional space.

Suppose that \( \{ P_\theta, \ \theta \in \Theta \} \) is a nominated family of probability measures indexed by \( \theta \). The idea behind Maximum Likelihood Estimation (MLE) is as follows: Suppose we randomly survey the population according to a sampling method and the result is an observation \( y \). If \( \theta \) is unknown, an estimation of \( \theta \) is one for which \( P_\theta \) is the most likely generator of \( y \). If the density function \( f_\theta = dP_\theta / d\lambda \) exists, we seek for a \( \theta \) for which \( f_\theta(y) \) is of maximum value.

What makes a density function suitable for this kind of inference is the base measure \( \lambda \) where the density is defined relative to it. A counting measure or Lebesgue measure are suitable options in finite-dimensional spaces. These base measures are positive, locally finite and translation invariant and a nontrivial measure with these properties does not exist in an infinite-dimensional separable Hilbert space (Gill et al., 2014).

Employing the kernel mean function, we can introduce a rather similar idea to likelihood-based estimation in infinite-dimensional spaces. Suppose \( k \) is a bounded, continuous, and translation-invariant characteristic kernel as described in Theorem 1, such as Gaussian kernel \( k(x, y) = e^{-\frac{1}{2\sigma} \|x - y\|^2} \) for a fixed \( \sigma > 0 \). The kernel mean function \( m_\varphi(\cdot) \) is a bounded function over \( H \) which reflects the concentration of \( P \) on a small neighborhood of \( y \in H \). Consider
the family of probability measures \( M = \{ \mathbb{P}_\theta, \ \theta \in \Theta \} \) and its counterpart family of kernel mean functions \( \{ m_{\mathbb{P}_\theta}(\cdot); \theta \in \Theta \} \). Assume that \( \theta \) is unknown and the endeavor is to estimate it through an observed random sample \( y \) from the population. Again the aim is to pick a \( \hat{\theta} \in \Theta \) such that \( \mathbb{P}_{\hat{\theta}} \) is the most likely generator of \( y \). Thus, it seems possible to estimate \( \theta \) by \( \hat{\theta} = \arg \max_{\theta \in \Theta} m_{\mathbb{P}_\theta}(y) \), which we call it MKM estimation of \( \theta \).

In this section, we derive the kernel mean embedding of probability measures induced by Functional response models, and we show how parameter estimation and hypothesis testing are capable in this framework.

3.1 Kernel mean embedding of Gaussian probability measure

In this section, we study the kernel mean embedding of functional response regressions, functional one-way ANOVA, and testing for homogeneity of covariance operators in the FDA under the assumption of Gaussian probability measures. In the context of FDA, the Gaussian assumption has been frequently used in many statistical problems, including the above category of models, while the random variables in a majority of statistical applications, either in FDA or non-FDA contexts, may not be normally distributed. By introducing the Gaussian assumption, the procedure of estimating the functional parameters is straightforward. By virtue of these estimators, in Section 4, statistical tests are developed with easy implementation of a permutational procedure of estimating the functional parameters.

Let \( \mathbb{H} \) be an arbitrary infinite-dimensional separable Hilbert space. An \( \mathbb{H} \)-valued random element is said to be a Gaussian random element with the mean function \( \mu \) and covariance operator \( C \), and is denoted by \( X \sim \mathcal{N}(\mu, C) \), if for any \( a \in \mathbb{H} \) we have \( \langle a, X \rangle \sim N(\langle a, \mu \rangle, \langle Ca, a \rangle) \). A Gaussian random element has a finite second moment, and its covariance operator is of trace class (Maniglia & Rhandi, 2004).

Denote by \(|\cdot|\) the determinant of a nonnegative symmetric operator on \( \mathbb{H} \), the kernel mean function of the Gaussian family of probability measures with mean \( \mu \) and covariance operator \( C \) and its uniqueness is given in the following proposition and is proved in the Appendix A.

**Proposition 1.** Let \( Y \sim \mathcal{N}(\mu, C) \) then for a Gaussian kernel,

\[
m_{\mathcal{N}(\mu, C)}(x) = \int_{\mathbb{H}} e^{-\frac{1}{2} \| x - y \|^2_{\mathbb{H}}} \mathcal{N}(\mu, C)(dy) = \left| I + 2 \frac{1}{\sigma} C \right|^{-1/2} e^{-\frac{1}{2} \left( (I + 2 \frac{1}{\sigma} C)^{-1} (x - \mu), (x - \mu) \right)}_{\mathbb{H}},
\]

the kernel mean embedding is injective and

\[
\left\| m_{\mathcal{N}(\mu, C)} \right\|^2_{L_{\mathbb{H}}} = \left| I + 4 \frac{1}{\sigma} C \right|^{-1/2}.
\]

If \( (\lambda_j, \psi_j)_{j \geq 1} \) is the eigensystem of \( C \), then \( (1 + \gamma \lambda_j, \psi_j)_{j \geq 1} \) is the eigensystem of \( I + \gamma C \) and the determinant and and inverse of \( I + \gamma C \) is given by \( |I + \gamma C| = \prod_{j \geq 1} (1 + \gamma \lambda_j) \) and \( (I + \gamma C)^{-1} = \sum_{j \geq 1} (1 + \gamma \lambda_j)^{-1} \psi_j \otimes_{\mathbb{H}} \psi_j \), for any \( \gamma > 0 \) and \( \psi_j \otimes_{\mathbb{H}} \psi_j \) is a rank-one operator with the action \( f \mapsto \langle \psi_j, f \rangle_{\mathbb{H}} \psi_j \).

Consider a random sample \( Y_i \in \mathbb{H}, \ i = 1, \ldots, n \) of independent and identically random elements with distribution \( \mathcal{N}(\mu, C) \). By choosing a suitable characteristic kernel for the product
space $\mathbb{H}^n$, kernel mean embedding of the induced probability measure by the random sample $\{Y_i\}$ that is, $\otimes_{i=1}^n \mathcal{N}(\mu, C)$ can be computed. Let $k$ be the Gaussian kernel, then according to the following theorem, which is proved in the Appendix A, $\prod_{i=1}^n k(\cdot, \cdot)$ is a characteristic kernel for the family of Gaussian distributions on $\mathbb{H}^n$ including $\otimes_{i=1}^n \mathcal{N}(\mu_i, C_i)$, and $\sum_{i=1}^n k(\cdot, \cdot)$ is a characteristic kernel for the family of product measures $\otimes_{i=1}^n \mathcal{N}(\mu_i, C_i)$ on $\mathbb{H}^n$.

**Proposition 2.** Let $k(\cdot, \cdot)$ be the Gaussian kernel defined over a separable Hilbert space $\mathbb{H}$, then product-kernel

$$k_\mathcal{P}(\cdot, \cdot) : \mathbb{H}^n \times \mathbb{H}^n \to \mathbb{R}$$

is a characteristic kernel for the family of Gaussian distributions on $\mathbb{H}^n$, and sum-kernel

$$k_\mathcal{S}(\cdot, \cdot) : \mathbb{H}^n \times \mathbb{H}^n \to \mathbb{R}$$

is a characteristic kernel for the family of product measures $\otimes_{i=1}^n \mathcal{N}(\mu_i, C_i)$ on $\mathbb{H}^n$.

Consider that according to the Proposition 2, product-kernel is characteristic for the broader family of multivariate Gaussian measures, while sum-kernel is characteristic for the family of product measures. For the case of Gaussian product-kernel, given a simple random sample $Y_1, \ldots, Y_n$ drawn from the Gaussian distribution, kernel mean function is

$$m_{\otimes_{i=1}^n \mathcal{N}(\mu, C)}(Y_1, \ldots, Y_n) = \int_{\mathbb{H}^n} e^{-\frac{1}{2} \sum_{i=1}^n \|y_i - \bar{z}\|_{\mathbb{H}}^2} \otimes_{i=1}^n \mathcal{N}(\mu, C)(dz_i),$$

whose logarithm equals to

$$\log m_{\otimes_{i=1}^n \mathcal{N}(\mu, C)}(Y_1, \ldots, Y_n) = \sum_{i=1, j \geq 1} \sum_{j \geq 1} \frac{-1}{1 + 2 \frac{1}{\sigma} \lambda_j} (y_i - \mu, \psi_j)^2 - \frac{n}{2} \sum_{j \geq 1} \log \left( 1 + 2 \frac{1}{\sigma} \lambda_j \right), \quad (4)$$

where $y_1, \ldots, y_n$ are the observation counterparts of $Y_1, \ldots, Y_n$, and $(\lambda_j, \psi_j)_{j \geq 1}$ is the eigensystem of $C$. By defining sample mean function and sample covariance operator as $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ and $\hat{C} = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y}) \otimes_H (y_i - \bar{y})$, respectively, where $\otimes_H$ denotes the tensor product defined as $(a \otimes_H b) c = (a \otimes_H b)$ for all $a, b, c \in \mathbb{H}$, then the logarithm of the kernel mean function for Gaussian product-kernel also is equal to

$$\log m_{\otimes_{i=1}^n \mathcal{N}(\mu, C)}(Y_1, \ldots, Y_n) = \sum_{j \geq 1} \frac{-n}{1 + 2 \frac{1}{\sigma} \lambda_j} \left[ (\hat{C} \psi_j, \psi_j)_H + (\bar{y} - \mu, \psi_j)_H^2 \right] - \frac{n}{2} \sum_{j \geq 1} \log \left( 1 + 2 \frac{1}{\sigma} \lambda_j \right). \quad (5)$$

We can see that the kernel mean function is dependent on $\{y_1, y_2, \ldots, y_n\}$ only through $\bar{y}$ and $\hat{C}$. Since Gaussian-product Kernel is characteristic, Equation (5) shows that for the family of
Gaussian probability measures, \( \bar{y}, \hat{C} \) is a typical joint sufficient statistic for parameters \((\mu, C)\). The possibility to identify sufficient statistics through kernel mean functions, alongside Theorem 1 and Corollary 2, reveals how a kernel mean embedding of probability measure behave akin to density function over finite-dimensional spaces.

The location and covariance parameters of the distribution can be estimated by maximizing \( \log m^{\otimes n}_{i=1} \mathcal{N}(\mu, C)(y_1, \ldots, y_n) \). The resulting estimator is slightly different in weights of components from the estimators one may obtain by the small ball probability approximation proposed by Delaigle and Hall (2010), or ordinary least squares (OLS) approach. As it is highlighted in Proposition 5, the MKM estimator of the location parameters converges to OLS, and the limiting estimator one may obtain by the small-ball probability approach, as \( \sigma \) tends to zero. It is also worth noting that although there is no estimation for covariance parameters by the small-ball probability approximation, there is an estimation for them by MKM.

In the context of functional regression, as it is addressed in Section 3.2, we may substitute a linear model for \( \mu \) in (4), and estimate the parameters of the model either by maximizing \( m^{\otimes n}_{i=1} \mathcal{N}(\mu, C)(y_1, \ldots, y_n) - \frac{1}{2}\|\| m^{\otimes n}_{i=1} \mathcal{N}(\mu, C) \|\|_H \) for parameters as in (2), or only choose to maximize \( \log m^{\otimes n}_{i=1} \mathcal{N}(\mu, C)(y_1, \ldots, y_n) \) for estimating the location parameters seeing as \( \|\| m^{\otimes n}_{i=1} \mathcal{N}(\mu, C) \|\|_H \) does not depend on the location parameters.

The Kernel Mean approach also provides a rich toolbox of kernel methods developed by the machine learning community, which can be used in statistical inference. To give just a few examples, we can name Kernel Bayes Rule for Bayesian inference and Latent variable modeling. The MMD for hypothesis testing and developing Goodness of Fit indices, and Hilbert Schmidt Independence Criterion (HSIC) for measuring dependency between random elements (see Gretton, Borgwardt, et al., 2012; Harchaoui et al., 2009; Muandet et al., 2017; Tang et al., 2017). In Section 4, MMD is used to derive and introduce new tests for three main problems in functional data analysis, including Function-on-Scalar regression, one-way ANOVA, and testing for homogeneity of covariance operators. The power of these tests is studied and compared with competitors by simulation.

### 3.2 MKM estimation of parameters in Function-on-Scalar regression

Let \( Y \) be a Gaussian random element taking value in \( \mathbb{H} \). Given a random sample of \( Y \), we can employ the kernel mean function to estimate the location and covariance parameters. Let \( y_1, y_2, \ldots, y_n \) be \( n \) independent realizations of \( Y \) according to the following Function-on-Scalar regression model:

\[
Y_i(t) = x_i^T \beta(t) + \epsilon_i(t), \quad i = 1, \ldots, n, \tag{6}
\]

where \( x_i \) is the vector of scalar covariates and \( \beta \) is the vector of \( p \) functional parameters. Residual functions \( \epsilon_i \) are \( n \) independent copies of a Gaussian random element with mean function zero and covariance operator \( C \). The following two propositions can be employed to obtain the MKM estimation of location and covariance parameters.

**Proposition 3.** Let \( y_1, y_2, \ldots, y_n \) be \( n \) independent realizations of model (6), where \( \epsilon_i \) is an \( \mathbb{H} \)-valued Gaussian random element with mean function zero and covariance
operator $C$. The MKM estimation of functional regression parameters coincide with the OLS estimation.

**Proof.** Defining $\mu_i := x_i^T \beta$, the logarithm of kernel mean function by (4) can be written as

$$
\log m_\beta(y_1, \ldots, y_n) := \log m_{\Phi_{i=1}^n \mathcal{N}(\mu, C)}(y_1, \ldots, y_n)
= \sum_{i=1}^n \left( I + 2 \frac{1}{\sigma} C \right)^{-1} (y_i - x_i^T \beta), x_i^T \beta \right)_H 
= \frac{n}{2} \sum_{j=1}^n \log \left( 1 + 2 \frac{1}{\sigma} \lambda_j \right).
$$

(7)

Fréchet derivation of (7) with respect to $\beta$ is an operator from $H^p$ to $\mathbb{R}$, i.e.

$$
\frac{\partial}{\partial \beta} \log m_\beta(y_1, \ldots, y_n) : H^p \to \mathbb{R}.
$$

$\hat{\beta}$ is a local extremum of $\log m_{\Phi_{i=1}^n \mathcal{N}(\mu, C)}(y_1, \ldots, y_n)$, if

$$
\left[ \frac{\partial}{\partial \beta} \log m_\beta(y_1, \ldots, y_n) \right](h) = 0 \quad \forall h \in H^p.
$$

Taking Fréchet derivation of (7) with respect to $\beta$, for an arbitrary $h \in H^p$ we have

$$
\left[ \frac{\partial}{\partial \beta} \log m_\beta(y_1, \ldots, y_n) \right](h) = \left[ \frac{\partial}{\partial \beta} \sum_{i=1}^n \left( I + 2 \frac{1}{\sigma} C \right)^{-1} (y_i - x_i^T \beta), x_i^T \beta \right)_H (h)
= 2 \frac{1}{\sigma} \sum_{i=1}^n \left( I + 2 \frac{1}{\sigma} C \right)^{-1} x_i^T h, y_i - x_i^T \beta \right)_H
= 2 \frac{1}{\sigma} \sum_{k=1}^p \left( h_k, (I + 2 \frac{1}{\sigma} C)^{-1} \sum_{i=1}^n x_{ik} (y_i - x_i^T \beta) \right)_{H^k}.
$$

So if $\hat{\beta}$ is a local extremum of (7), for each $1 \leq k \leq p$, we must have

$$
\sum_{i=1}^n x_{ik} (y_i - x_i^T \beta) = 0.
$$

(8)

Let $X$ be the $n \times p$ data matrix and $Y := [y_1, \ldots, y_n]^T$, then the Equation (8) implies $X^T (Y - X\hat{\beta}) = 0$, and consequently $\hat{\beta} = (X^T X)^{-1} X^T Y$. The remaining question arises here is that if $\hat{\beta}$ maximizes (7) or not. Let $\beta = \hat{\beta} + \nu$, then

$$
\log m_\beta(y_1, \ldots, y_n) = \log m_\beta(y_1, \ldots, y_n) - \frac{n}{\sigma} \sum_{i=1}^n \left( I + 2 \frac{1}{\sigma} C \right)^{-1} (x_i^T \nu), x_i^T \nu \right)_{H^k}
\leq \log m_\beta(y_1, \ldots, y_n),
$$

which completes the proof.
From the last proposition, \( \hat{\beta} = (X^TX)^{-1}X^TY \) is the MKM estimator of functional regression coefficients. It is also possible to derive a restricted MKM estimation of covariance operator with a similar approach to restricted MLE. Let \( A = [u_1, \ldots, u_{n-k}] \) be the first \( n-k \) eigenvectors of \( I - X(X^TX)^{-1}X^T \), then \( Y_i = u_i^TY \) is called the error contrast vector and is a sequence of \( n-k \) independent and identically distributed random elements with mean function zero and common covariance operator \( C \). We can then use the sequence of realizations \( y_1^*, \ldots, y_{n-k}^* \) and employ Proposition 4 to estimate the covariance operator by \( \hat{C} = \frac{1}{n-k} \sum_{i=1}^{n-k} y_i^* \otimes_{\mathbb{H}} y_i^* \).

**Proposition 4.** Let \( y_1, y_2, \ldots, y_n \) be \( n \) independent realizations of \( \mathbb{H} \)-valued Gaussian random element with mean function zero, and covariance operator \( C = \sum_{j \geq 1} \lambda_j \psi_j \otimes_{\mathbb{H}} \psi_j \). Let \( \hat{C} = \frac{1}{n} \sum_{i=1}^{n} y_i \otimes_{\mathbb{H}} y_i \), then as \( \frac{1}{n} \to \infty \) the MKM estimator of \( \{ \lambda_j, \psi_j \}_{j \geq 1} \) converges to

1. \( \hat{\psi}_k \) is the \( k \)th eigenfunction of \( \hat{C} \)
2. \( \hat{\lambda}_k = \frac{1}{n} \sum_{i=1}^{n} \langle y_i, \hat{\psi}_k \rangle_{\mathbb{H}}^2 \)

**Proof:** The logarithm of the kernel mean function of the product measure \( \mathbb{H} \)-valued Gaussian \( \mathcal{N}(\mu, C) \) is presented in (4), while we set \( \mu = 0 \). Parameter estimation is obtained by taking Fréchet derivation of kernel mean function with respect to \( \psi_k \) and usual derivation of kernel mean function with respect to \( \lambda_k \). In each case, it is shown that the local extremum is the global maximum of kernel mean function.

(1) \( \psi_k \): First we obtain the estimation of \( \psi_1 \). Taking Fréchet derivation of kernel mean function with respect to \( \psi_1 \), we have

\[
\left[ \frac{\partial}{\partial \psi_1} \log m_{\mathcal{N}(\mu, C)}(y_1, \ldots, y_n) \right](h) = \left[ \frac{\partial}{\partial \psi_1} \sum_{i=1}^{n} \sum_{j \geq 1} \frac{-1}{\sigma^2} \langle y_i, \psi_j \rangle_{\mathbb{H}}^2 \right](h) = \frac{-2}{\sigma^2} \sum_{i=1}^{n} \langle y_i, \psi_1 \rangle_{\mathbb{H}} \langle y_i, h \rangle_{\mathbb{H}} = \frac{-2}{\sigma^2} \langle \hat{C} \psi_1, h \rangle_{\mathbb{H}}.
\]

Consider that \( \psi_1 \) lies in a sphere of radius 1, thus \( \hat{\psi}_1 \) is an extremum point of \( \log m_{\mathcal{N}(\mu, C)}(y_1, \ldots, y_n) \), if \( \langle \hat{C} \psi_1, h \rangle_{\mathbb{H}} = 0 \) for any arbitrary \( h \) in the tangent space of unit sphere at point \( \hat{\psi}_1 \), that is,

\[
\forall h \in \{ \hat{\psi}_1 \}^\perp \Rightarrow \langle \hat{C} \hat{\psi}_1, h \rangle_{\mathbb{H}} = 0.
\]

In addition, for the case of identifiability \( \hat{\psi}_1 \) must be associates to the largest eigenvalue of \( \hat{C} \). This way, MKM estimation of \( \psi_1 \) is the solution to the following optimization problem:

\[
\hat{\psi}_1 = \arg \max_{\psi \in \mathbb{H}} \frac{1}{n} \sum_{i=1}^{n} \langle y_i, \psi \rangle_{\mathbb{H}}^2 \quad \text{s.t.} \quad \langle \hat{C} \hat{\psi}, h \rangle_{\mathbb{H}} = 0 \quad \forall h \in \{ \hat{\psi}_1 \}^\perp.
\]
which immediately follows that MKM estimation of $\psi_1$ is the first eigenfunction of $\hat{C}$, and is independent of kernel parameter $\frac{1}{\sigma}$. The remaining question to answer is that if $\hat{\psi}_1$ maximizes (4) or not. Consider that for any arbitrary $h \in \{\hat{\psi}_1\}^\perp$ and $\hat{\psi}_1 = \frac{1}{\|\hat{\psi}_1 + h\|_H^2}$$^\perp$,

$$
\log m_{\hat{\psi}_1}(y_1, \ldots, y_n) = \log m_{\hat{\psi}_1}(y_1, \ldots, y_n) - \frac{n}{2} \frac{1}{\sigma} \frac{1}{\|\hat{\psi}_1 + h\|_H^2} \left(1 + 2 \frac{1}{\sigma} \lambda_k\right) \langle \hat{C}h, h \rangle_H
$$

$$\leq \log m_{\hat{\psi}_1}(y_1, \ldots, y_n).$$

So $\hat{\psi}_1$ is the MKM estimation of $\psi_1$. For the MKM estimation of $\psi_k$, $k \geq 2$, the following constraint should be added to the optimization problem (10)

$$\langle \hat{\psi}_k, \hat{\psi}_j \rangle_H = 0 \quad \forall 1 \leq j < k,$$

which shows that the MKM estimation of all eigenfunctions is the same as the set of eigenfunctions of $\hat{C}$.

(2) $\lambda_k$: Taking derivation of kernel mean function with respect to $\lambda_k$, yields

$$
\frac{\partial}{\partial \lambda_k} \log m_{\Phi_{\psi_k}^\perp, \mathcal{N}(\mu, C)}(y_1, \ldots, y_n)
$$

$$= \frac{\partial}{\partial \lambda_k} \left[ \sum_{i=1}^n \sum_{j \geq 1} \frac{-1}{\sigma} \frac{1}{\|\hat{\psi}_1 + h\|_H^2} \left(1 + 2 \frac{1}{\sigma} \lambda_j\right) \langle y_i, \psi_j \rangle_H^2 - \frac{n}{2} \sum_{j \geq 1} \log \left(1 + 2 \frac{1}{\sigma} \lambda_j\right) \right]
$$

$$= \frac{n}{2} \frac{1}{\sigma} \frac{1}{\|\hat{\psi}_1 + h\|_H^2} \left(1 + 2 \frac{1}{\sigma} \lambda_k\right)^2 \langle y_i, \psi_k \rangle_H^2 - \frac{n}{2} \frac{1}{\sigma} \left(1 + 2 \frac{1}{\sigma} \lambda_k\right)
$$

$$= \frac{1}{\sigma} \frac{1}{\|\hat{\psi}_1 + h\|_H^2} \left(1 + 2 \frac{1}{\sigma} \lambda_k\right)^2 \sum_{i=1}^n \langle y_i, \psi_k \rangle_H^2 - n \left(1 + 2 \frac{1}{\sigma} \lambda_k\right).$$

Equating (11) to zero and given $\psi_k$, the value of $\lambda_k$ which maximizes (4) while we set $\mu = 0$ is given by $\hat{\lambda}_k = \frac{1}{n} \sum_{i=1}^n \langle y_i, \psi_k \rangle_H^2 - \frac{1}{n} \hat{\psi}_k^2$, in consequence by putting $\psi_k$ to be the $k$th eigenfunction of $\hat{C}$, we obtain $\hat{\lambda}_k = \frac{1}{n} \sum_{i=1}^n \langle y_i, \psi_k \rangle_H^2 - \frac{1}{n} \hat{\psi}_k^2$, which is a biased estimator of $\lambda_k$ and converges to $\frac{1}{n} \sum_{i=1}^n \langle y_i, \psi_k \rangle_H^2$ as $\sigma$ tends to zero.

In the following proposition, we obtained an estimation of the functional regression coefficients of model (6) by employing small-ball probability approximation.

**Proposition 5.** In the case of Function-on-Scalar regression with the assumption of normality as in the model (6), in estimating the location parameters, the MKM or OLS estimator is the same as the one obtained by the small-ball probability approximation proposed by Delaigle and Hall (2010).

**Proof.** Let $Y_1, \ldots, Y_n$ be a simple random sample generated by model (6), where $\epsilon_i$ is a sequence of $n$ independent copies of a $\mathbb{H}$-valued Gaussian random element with mean
function zero, and covariance operator \( C = \sum_{j \geq 1} \lambda_j \psi_j \otimes \psi_j \). Let functions \( \beta_k \) admit the Fourier decomposition \( \beta_k = \sum_{j \geq 1} \theta_{kj} \psi_j \) and define \( \theta_j = (\theta_{kj})_{k=1,\ldots,p}, x_i = (x_{ik})_{k=1,\ldots,p} \) and \( \beta = (\beta_k)_{k=1,\ldots,p} \).

The identity \( e_i = Y_i - x_i^T \beta \) i.i.d. \( \mathcal{N}(0, C) \) is equivalent to the situation where component scores \( (e, \psi_j) / \sqrt{\lambda_j} \) are independent and identically distributed according to the standard normal distribution for each \( j \in \mathbb{N} \). Fix \( r > 0 \) and let \( h = \arg \max_j r^2 \leq \lambda_j \). By the method of Delaigle and Hall (2010), the log-density with radius \( r \) equals to

\[
\log P(e_i | r) \propto \sum_{j=1}^{h} \log f_j(\sqrt{\lambda_j}(e_i, \psi_j)_\mathbb{H}) \propto -\sum_{j=1}^{h} \frac{\lambda_j}{2} (\langle Y_i, \psi_j \rangle_\mathbb{H} - x_i^T \theta_j)^2 ,
\]

(12)

and thus

\[
\sum_{i=1}^{n} \log P(e_i | r) \propto -\sum_{j=1}^{h} \sum_{i=1}^{n} \frac{\lambda_j}{2} (\langle Y_i, \psi_j \rangle_\mathbb{H} - x_i^T \theta_j)^2 = \sum_{j=1}^{h} \lambda_j \left(-\frac{1}{2} \theta_j^T (X^T X) \theta_j + B_j^T \theta_j \right),
\]

in which \( B_j = \sum_{i=1}^{n} x_i (Y_i, \psi_j)_\mathbb{H} \) and \( X \) is \( n \times p \) model matrix and \( Y \) is an \( n \times 1 \) column vector containing functions \( Y_i(\cdot) \). Estimate of \( \theta_j \) can be obtained by solving the equation \( \frac{\partial}{\partial \theta_j} \sum_{i=1}^{n} i.i.d. P(e_i | r) = 0 \) thus

\[
\hat{\theta}_j = (X^T X)^{-1} B_j.
\]

For a given \( r > 0 \) or its coupled quantity \( h \in \mathbb{N} \), estimation of \( \beta \) is

\[
\hat{\beta}^r = \left( \sum_{j=1}^{h} \hat{\theta}_j \psi_j \right)_{k=1,\ldots,p} = (X^T X)^{-1} \left( \sum_{i=1}^{n} x_{ik} \sum_{j=1}^{h} \langle Y_i, \psi_j \rangle_\mathbb{H} \psi_j \right)_{k=1,\ldots,p}
\]

\[
= (X^T X)^{-1} X^T \left( \sum_{j=1}^{h} \langle Y_i, \psi_j \rangle_\mathbb{H} \psi_j \right).
\]

Considering the limit of \( \hat{\beta}^r \) as \( r \) tends to zero, the limiting estimation ends up with

\[
\lim_{r \to 0} \hat{\beta}^r = (X^T X)^{-1} X^T Y ,
\]

which is the OLS estimation of \( \beta \).

4 | APPLICATIONS

MMD is a useful vehicle for hypothesis testing. If the kernel \( k \) is characteristic, then MMD is a metric on the space of probability measures. The induced distance by this metric can be employed to derive different statistical tests that can be hard to handle in the context of Functional data analysis, especially simultaneous hypothesis tests such as one-way ANOVA and testing for equality of covariance operators in more than two groups.

Kernel-based methods such as kernel mean and covariance embedding of probability measures have a wide range of applications in analyzing structured and nonstructured data and also developing non-parametric tests in finite-dimensional spaces such as testing for
homogeneity of location and variance parameters, change-point detection, and test of independence. See for example Harchaoui et al. (2009), Gretton, Borgwardt, et al. (2012), Tang et al. (2017), and Pfister et al. (2018) among others to get some insight. In this section, we employ MMD to develop new tests for three major problems in the context of Functional response models, including Function-on-Scalar regression, Functional one-way ANOVA, and testing for equality of covariance operators. The new tests are based on the estimators derived in Section 3. To develop new tests, the plug-in estimators of probability measures induced by null and alternative hypotheses are embedded in a Hilbert space and their distance is computed by MMD. The performance of new tests is compared to some state-of-the-art methods while the null distributions for all competitors are obtained by permutation algorithm for a homogeneous comparison of the competitors.

For choosing kernel bandwidth various methods have been proposed and employed in the literature. These methods include the widely used approach of median of distances which utilizes twice the median of pairwise distances as the kernel bandwidth (Sejdinovic et al., 2013). Another method is maximum-MMD (Fukumizu et al., 2009; Gretton, Sejdinovic, et al., 2012) that seeks to maximize the MMD distances over a family of characteristic kernels. Additionally, a double minimum $p$-value bandwidth is also introduced by Martínez-Camblor et al. (2008) which selects the kernel bandwidth attached to the smallest $p$-value from a set of bandwidths defined on a specific grid of values and controls the type-I error by a double bootstrap strategy.

In this paper, we employ two methods of median of pairwise distances and the double minimum $p$-value bandwidth to tackle with the problem of selecting kernel bandwidth in the developed tests. For the later approach, permutation tests are run with kernel bandwidths varies over a set of values, and the minimum $p$-value is calculated. The decision rule is to reject the null hypothesis if the minimum $p$-value is less than a critical value. The critical point is obtained by approximating the distribution of minimum $p$-value by a second round of permutation. Here we take the set of bandwidths of the form $\Sigma = \{10^\gamma S^2 : \gamma \in \Gamma\}$ where $S^2$ is the squared median of pairwise distances and $\gamma$ varies a set $\Gamma$. The following algorithm is a modified version of the algorithm introduced by Martínez-Camblor et al. (2008) to implement the double minimum $p$-value bandwidth:

- Step 1: For each value of kernel bandwidth $\sigma \in \Sigma$, compute $\mathrm{MMD}(\sigma)$ given data.
- Step 2: For each value of kernel bandwidth $\sigma \in \Sigma$, draw $K$ times $\mathrm{MMD}^*_{j}(\sigma)$ under null hypothesis by $B_1$ independent permutation of subjects, and let $p_\sigma = \frac{1}{B_1} \sum_{j=1}^{B_1} I \left( \mathrm{MMD}^*_{j}(\sigma) > \mathrm{MMD}(\sigma) \right)$ and $p = \min\{p_{\sigma} , \sigma \in \Sigma\}$, where $I(\cdot)$ is the indicator function.
- Step 3: For each value of kernel bandwidth $\sigma \in \Sigma$, draw $B_2$ values of $\mathrm{MMD}^*_{k}(\sigma)$ by permutation of subjects, and let $p^*_{k,\sigma} = \frac{1}{B_2} \sum_{j=1}^{B_2} I \left( \mathrm{MMD}^*_{j}(\sigma) > \mathrm{MMD}^*_{k}(\sigma) \right)$ and $p^*_k = \min\{p^*_{k,\sigma} , \sigma \in \Sigma\}$. Put $p^{**} = \frac{1}{B_1} \sum_{k=1}^{B_2} I \left( p^*_k \leq p \right)$.
- Step 4: Reject the null hypothesis whenever $p^{**} < \alpha$, where $\alpha$ is the nominal level of the test.

It is desirable to use newly generated values for $\mathrm{MMD}^*_{j}(\sigma)$ in step 3, however in this algorithm the generated values in step 2 are reused in step 3 to reduce the computational cost. The computational cost can further be reduced by putting $B_1 = B_2$ and set $\mathrm{MMD}^*_{k}(\sigma) := \mathrm{MMD}^*_{j}(\sigma)$. Here we put $B_1 = B_2 = 500$ and $\gamma$ varies in the range $-10$ to $+10$ with step $0.2$. In this section, in presenting the result the notations $\mathrm{MMD}(\sigma_0)$ and $\mathrm{MMD}(\sigma^2)$ are used to show the performance of
MMD test using the methods of median of pairwise distances and double permutation strategy, respectively.

Before proceeding, it seems indispensable to notice that in the methods developed in this section, we assumed that the sampled random functions are observed completely. However, in practice, a function is observed only in a sparse or dense subset of the domain. Accordingly, at the first stage of analysis, we may operate a smoothing procedure to construct functions. To scrutinize the effect of smoothing in the results, the first simulation in this section has been done with different number of points sampled per curve. The results seem to be acceptable despite using smoothed functions rather than complete functions.

### 4.1 Hypothesis testing in Function-on-Scalar regression model

Let $\mathbb{H}$ be the space of square-integrable functions $L^2[0, 1]$, and consider the following simple Function-on-Scalar regression problem:

$$y_i(t) = a(t) + x_i \beta(t) + \epsilon_i(t), \quad \epsilon_i \sim \mathcal{N}(0, C); \quad i = 1, \ldots, n \quad t \in [0, 1],$$

where $\mathcal{N}(0, C)$ is the Gaussian distribution over the space $L^2[0, 1]$, with mean function zero and covariance operator $C$. In Proposition 3 it is shown that the maximum kernel mean estimation of intercept and slope functions coincide with OLS estimation. To assess the uncertainty of estimation and testing for $H_0 : \beta = 0$ versus $H_1 : \beta \neq 0$, we run a simulation study proposed by Kokoszka and Reimherr (2017) to compare type-I error and power of new test devised from MMD with the current developed tests.

Let $a(t) = 2t$ and $\beta(t) = -c_0 \cos(\pi t)$, in which the parameter $c_0$ is used to switch between the null and alternative hypotheses. For the covariance operator $C$ in (13) we use the Matérn family of covariance operators, once with an infinitesimal smoothness parameter and once with the smoothness parameter set to 1/2,

$$C_\infty(s, t) = e^{-\frac{|s-t|^2}{\rho}} \quad \text{and} \quad C_{1/2}(s, t) = e^{-\frac{|s-t|^2}{\rho}}.$$

The kernel function $C_\infty$ referred to as squared-exponential covariance and $C_{1/2}$ referred to as exponential covariance function. To test $H_0 : \beta = 0$, we devise a new test using MMD statistic by employing Gaussian kernel as a characteristic kernel on $\mathbb{H}^n$. We use either the Gaussian sum-kernel

$$k(\cdot, \cdot) : \mathbb{H}^n \times \mathbb{H}^n \to \mathbb{R} \quad \left((x_i)_{i=1}^n, (y_i)_{i=1}^n\right) \mapsto \sum_{i=1}^n e^{-\frac{1}{\sigma} \|x_i - y_i\|_{\mathbb{H}}^2},$$

or Gaussian product-kernel

$$k(\cdot, \cdot) : \mathbb{H}^n \times \mathbb{H}^n \to \mathbb{R} \quad \left((x_i)_{i=1}^n, (y_i)_{i=1}^n\right) \mapsto \prod_{i=1}^n e^{-\frac{1}{\sigma} \|x_i - y_i\|_{\mathbb{H}}^2}.$$

Let $\mathbb{P}_0$ be the induced probability measure of model (13) under the null hypothesis, and $\mathbb{P}_1$ be the induced probability measure under the alternative hypothesis. Let $\hat{C}_0$ and $\hat{a}_0$ be the estimation of covariance and intercept function under the null hypothesis and $\hat{C}_1$, $\hat{a}_1$ and $\hat{\beta}_1$ be
the estimation of covariance, intercept and slope function under the alternative hypothesis as described in Proposition 3 and Proposition 4, then plugin estimators \( \hat{P}_0 \) and \( \hat{P}_1 \) are \( \otimes_{i=1}^{n} \mathcal{N}(\hat{a}_i, \hat{C}_0) \) and \( \otimes_{i=1}^{n} \mathcal{N}(\hat{a}_i + x_i \hat{\beta}_1, \hat{C}_1) \), respectively. The MMD statistic can then be defined as the MMD distance between \( \hat{P}_0 \) and \( \hat{P}_1 \). Using the Gaussian Sum-Kernel, MMD distance between \( \hat{P}_0 \) and \( \hat{P}_1 \) equals to:

\[
\begin{align*}
\text{MMD}(H_{k_1}, \hat{P}_0, \hat{P}_1) &= \left\| m_{\hat{p}_0} - m_{\hat{p}_1} \right\|_{H_k} = \left\| m_{\hat{p}_0} \right\|_{H_k}^2 + \left\| m_{\hat{p}_1} \right\|_{H_k}^2 - 2\left\{ m_{\hat{p}_0}, m_{\hat{p}_1} \right\}_{H_k} \\
&= \left[ n \left( I + \frac{4}{\sigma} \hat{C}_0 \right)^{-n/2} + n \left( I + \frac{4}{\sigma} \hat{C}_1 \right)^{-n/2} - 2 \left( I + \frac{4}{\sigma} \hat{C}_0 + \hat{C}_1 \right)^{-n/2} \right] \sum_{i=1}^{n} e^{-\frac{1}{n} \sum_{i=1}^{n} \left( (I + \frac{4}{\sigma} \hat{C}_0 + \hat{C}_1)^{-1} (\hat{a}_0 - \hat{a}_1 - x_i \hat{\beta}_1) (\hat{a}_0 - \hat{a}_1 - x_i \hat{\beta}_1) \right)}
\end{align*}
\]

and the Gaussian product-kernel yields:

\[
\begin{align*}
\text{MMD}(H_{k_1}, \hat{P}_0, \hat{P}_1) &= \left\| m_{\hat{p}_0} - m_{\hat{p}_1} \right\|_{H_k} = \left\| m_{\hat{p}_0} \right\|_{H_k}^2 + \left\| m_{\hat{p}_1} \right\|_{H_k}^2 - 2\left\{ m_{\hat{p}_0}, m_{\hat{p}_1} \right\}_{H_k} \\
&= \left[ \left( I + \frac{4}{\sigma} \hat{C}_0 \right)^{-n/2} + \left( I + \frac{4}{\sigma} \hat{C}_1 \right)^{-n/2} - 2 \left( I + \frac{4}{\sigma} \hat{C}_0 + \hat{C}_1 \right)^{-n/2} \right] \sum_{i=1}^{n} e^{-\frac{1}{n} \sum_{i=1}^{n} \left( (I + \frac{4}{\sigma} \hat{C}_0 + \hat{C}_1)^{-1} (\hat{a}_0 - \hat{a}_1 - x_i \hat{\beta}_1) (\hat{a}_0 - \hat{a}_1 - x_i \hat{\beta}_1) \right)}
\end{align*}
\]

The two test statistics are defined as

\[
\text{MMDS} := \text{MMD}(H_{k_1}, \hat{P}_0, \hat{P}_1), \quad \text{MMDP} := \text{MMD}(H_{k_1}, \hat{P}_0, \hat{P}_1).
\]

In a simulation study we will compare the MMD tests with few competitors including the test proposed by Greven et al. (2017) and implemented in the \texttt{pffr} function of \texttt{refund} package. Note that the distribution of MMD statistic under null hypothesis is approximated using the random permutation method. It is noted by Kokoszka and Reimherr (2017) that type-I error of \texttt{pffr} is inflated and is higher than the significance level. This problem is much worse with increasing number of sampling points per curve. Kokoszka and Reimherr (2017) proposed a partial fix to this problem. They suggest to ignore the uncertainty estimates from \texttt{pffr}. Instead, they use the estimated residual functions given by \texttt{pffr} and combine them with a classic estimate of uncertainty. To this end, let \( \hat{\beta}_1(t) \) and \( \hat{\epsilon}_1(t) \) be the slope function and residual functions estimated by \texttt{pffr}, respectively. Then an estimation of the uncertainty for \( \hat{\beta}_1(t) \) is

\[
\text{Cov}(\hat{\beta}_1(s), \hat{\beta}_1(t)) = \text{Cov}(\epsilon(s), \epsilon(t))X^TX_{(2,2)}^{-1} = \left(n \sum_{i=1}^{n} (x_i - \bar{x})^2 \right)^{-1} \sum_{i=1}^{n} \tilde{\epsilon}_i(s) \tilde{\epsilon}_i(t), \quad (14)
\]

in which \( X \) is an \( n \times 2 \) data matrix with vector of ones in the first column and vector of scalar covariate \( (x_i) \) in the second column. A variety of hypothesis tests then can be run by plugging \( \hat{\beta}_1(t) \) obtained from \texttt{pffr} and estimation of uncertainty obtained by (14) into the \texttt{fregion.test} function in \texttt{fregion} package.
Here we compare the proposed test with two existing ones; one is based on the $L^2[0,1]$ norm and the other test based on hyper-ellipsoid confidence regions proposed by Choi and Reimherr (2018).

Significance level is put at 0.05 and using a Monte Carlo simulation, the rate of rejection for different number of points sampled per curve ($m$) and two different covariance operators are computed. The null distribution for all of test statistics are approximated by a permutation strategy. For this purpose we permute the scalar covariate $x$ and will compute the parameter estimations and value of test statistics for large number of permutations. In the MMD tests B-spline basis has been used in the smoothing procedure, and the number of components for the smoothing procedure is considered to be fixed and equals 41. The kernel bandwidth is also chosen according to the median of pairwise distances approach and minimum $p$-value bandwidth method. The results are presented in Tables 1 and 2 are reported by 2000 iterations.

| $n$ | $c_0$ | 30 |  | 0.4 | 0.6 | 70 |  | 0.1 | 0.2 | 0.3 |
|-----|-------|----|---|-----|-----|----|---|-----|-----|-----|
| $m = 10$ | pffr | 5.0 | 4.0 | 15.8 | 55.4 | 5.3 | 3.6 | 3.4 | 17.4 |
| | Norm | 4.8 | 12.2 | 48.8 | 87.1 | 5.8 | 8.7 | 26.5 | 65.1 |
| | Ellipse | 4.6 | 20.7 | 70.7 | 96.4 | 5.6 | 13.2 | 47.6 | 86.2 |
| | MMDS($\sigma_0$) | 4.9 | 15.3 | 57.0 | 91.8 | 5.2 | 9.3 | 28.6 | 69.0 |
| | MMDP($\sigma_0$) | 4.8 | 26.2 | 78.5 | 97.7 | 5.0 | 17.9 | 58.4 | 91.7 |
| | MMDS($\pi^2$) | 4.5 | 20.4 | 72.7 | 97.0 | 5.6 | 14.6 | 54.2 | 92.0 |
| | MMDP($\pi^2$) | 4.9 | 22.2 | 75.7 | 97.6 | 5.1 | 15.7 | 57.0 | 93.3 |
| $m = 20$ | pffr | 4.2 | 4.0 | 16.6 | 57.1 | 4.8 | 2.6 | 4.2 | 18.4 |
| | Norm | 4.8 | 11.2 | 47.3 | 85.9 | 4.2 | 7.2 | 24.7 | 63.8 |
| | Ellipse | 5.0 | 20.1 | 71.2 | 96.4 | 4.3 | 12.2 | 47.5 | 85.8 |
| | MMDS($\sigma_0$) | 5.0 | 14.2 | 56.8 | 91.8 | 4.2 | 7.9 | 28.4 | 69.0 |
| | MMDP($\sigma_0$) | 4.0 | 24.8 | 76.8 | 97.4 | 4.3 | 16.1 | 55.9 | 90.2 |
| | MMDS($\pi^2$) | 4.8 | 20.1 | 72.0 | 97.1 | 5.0 | 14.7 | 55.0 | 92.0 |
| | MMDP($\pi^2$) | 5.0 | 22.3 | 75.8 | 97.8 | 5.0 | 15.7 | 58.2 | 93.0 |
| $m = 50$ | pffr | 5.8 | 4.6 | 20.3 | 60.8 | 5.8 | 4.0 | 6.8 | 26.6 |
| | Norm | 5.4 | 10.6 | 45.1 | 85.3 | 5.2 | 9.1 | 24.8 | 61.3 |
| | Ellipse | 4.8 | 20.2 | 71.0 | 96.2 | 5.7 | 14.6 | 47.4 | 86.6 |
| | MMDS($\sigma_0$) | 4.8 | 14.1 | 56.8 | 91.3 | 5.3 | 9.6 | 29.4 | 67.7 |
| | MMDP($\sigma_0$) | 4.8 | 24.9 | 77.0 | 97.6 | 5.2 | 18.8 | 58.4 | 90.4 |
| | MMDS($\pi^2$) | 5.4 | 20.4 | 73.6 | 97.1 | 5.7 | 16.0 | 55.0 | 93.0 |
| | MMDP($\pi^2$) | 5.0 | 22.8 | 76.0 | 97.2 | 5.8 | 17.0 | 58.9 | 93.9 |
From Table 1 and 2 it can be noticed that MMD test has higher power than \(pffr\). In all simulation scenarios, MMDP outperforms MMDS and its performance compete with the tests based on hyper-ellipsoid confidence regions.

### 4.2 Functional one-way ANOVA

The one-way ANOVA is a fundamental problem in statistical inference. Assume that \(\mathbb{H}\) is a separable Hilbert space, \(Y_{ij}\) for \(i = 1, \ldots, k\) and \(j = 1, \ldots, n_i\) are independent random samples taking values from \(\mathbb{H}\), and \(y_{ij}\) are their observations counterparts. For a typical functional one-way ANOVA problem with the assumption of homogeneity of covariance operators, the random elements \(Y_{ij}\) are assumed to be generated according to the following model:

\[
Y_{ij} = \mu_i + \epsilon_{ij}, \quad \epsilon_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, C); \quad i = 1, \ldots, k, j = 1, \ldots, n_i.
\]  

(15)
where \( \mu_i \) is the mean function of the \( i \)th group and the covariance operator is equal in all the \( k \) groups: \( C = E[\epsilon_{ij} \otimes \epsilon_{ij}] \) for all \( i = 1, \ldots, k \) and \( j = 1, \ldots, n_i \). It is of interest to test the equality of \( k \) mean functions, that is, \( H_0 : \mu_1 = \cdots = \mu_k \) versus alternative hypothesis \( H_1 \) which states at least two groups differs in their mean functions. Let \( P_0 \) be the probability measure of samples generated by model (15) under \( H_0 \) and \( P_1 \) be the probability measure induced by model (15), under the alternative hypothesis, that is, when parameters \( \mu_i \) are considered to be free. With the assumption of homogeneity of covariance operators, the squared MMD with Gaussian product-kernel equals to:

\[
\text{MMD}^2(H_{k_0}, P_0, P_1) = \| m_{P_0} - m_{P_1} \|^2_{H_k} \\
= \left| I + 4 \frac{1}{\sigma} C \right|^{-n/2} + \left| I + 4 \frac{1}{\sigma} C \right|^{-n/2} \\
- 2 \left| I + 4 \frac{1}{\sigma} C \right|^{-n/2} e^{-\frac{1}{2} \sum_{i=1}^{k} n_i \langle (I + 4 \frac{1}{\sigma} C)^{-1} (\mu_i - \mu)(\mu_i - \mu) \rangle_H} \\
= 2 \left| I + 4 \frac{1}{\sigma} C \right|^{-n/2} \left( 1 - e^{-\frac{1}{2} \sum_{i=1}^{k} n_i \langle (I + 4 \frac{1}{\sigma} C)^{-1} (\mu_i - \mu)(\mu_i - \mu) \rangle_H} \right).
\]

Covariance operator is assumed to be equal between the groups, so the new MMD test statistic can be simplified as

\[
\text{MMD}_0 = \sum_{i=1}^{k} n_i \left\langle \left( I + 4 \frac{1}{\sigma} C \right)^{-1} (\mu_i - \mu), (\mu_i - \mu) \right\rangle_H \\
= \sum_{i=1}^{k} n_i \left\| \left( I + 4 \frac{1}{\sigma} C \right)^{-1/2} (\mu_i - \mu) \right\|^2_H. \tag{16}
\]

Accordingly, the new test statistic is the weighted sum of the distance of group mean functions \( \mu_i \) from total mean function \( \mu \). By plugging the usual estimation of mean functions (which are also MKM estimations of mean functions) into (16), the new test statistic yields

\[
\text{\textit{\textbf{MMD}}}_0 = \sum_{i=1}^{k} n_i \left\langle \left( I + 4 \frac{1}{\sigma} C \right)^{-1} (\hat{\mu}_i - \hat{\mu}), (\hat{\mu}_i - \hat{\mu}) \right\rangle_H = \sum_{i=1}^{k} n_i \left\| \left( I + 4 \frac{1}{\sigma} C \right)^{-1/2} (\hat{\mu}_i - \hat{\mu}) \right\|^2_H.
\]

This test statistic is similar to the kernel Fisher discriminant analysis (KFDA) test statistic proposed by Harchaoui et al. (2013) for a two-sample kernel-based nonparametric test when the underlying space is finite-dimensional.

Let \( H \) again be the space of square-integrable functions \( L^2[0,1] \). Motivated by Zhang et al. (2019), a simulation study was run to evaluate the performance of the new MMD test against four other competitors developed for the space of square-integrable functions: an \( F \)-type test proposed by Shen and Faraway (2004), the Global Point-wise \( F \) (GPF) test and the \( F_{\text{max}} \) test proposed in Zhang and Liang (2014) and Zhang et al. (2019). We used the same setup as Zhang et al. (2019) for data generating procedure in our simulation study. Hence it is assumed that the functional samples in (15) are generated from the following one-way ANOVA model:

\[
y_{ij}(t) = \mu_i(t) + \epsilon_{ij}(t), \quad \mu_i(t) = c_i^T [1, t, t^2, t^3], \quad \epsilon_{ij}(t) = \sum_{r \geq 1} \sqrt{\lambda_r} \psi_{ij} \psi_r(t). \tag{17}
\]

\[i = 1, \ldots, k; \quad j = 1, \ldots, n_i; \quad t \in [0,1].\]
While our method works without the need to put any restriction on the number of components, we follow the same setup as in Zhang et al. (2019) and assume a finite number of $q$ nonzero eigenvalues in (17).

The parameter $n_i$ denotes the size of each group and the set of $\psi_r$ is the eigenfunctions. For all $i,j$, the design time points are considered to be balanced and equally spaced, thus all sampled curves are measured in the common grid of time points $t_j = j/(T + 1), j = 1, \ldots, T$.

The eigenvalues are assumed to follow the pattern $\lambda_r = a\rho^r - 1$ for fixed $a > 0$ and $\rho \in (0, 1)$. The tuning parameter $\rho$ determines the decay rate of eigenvalues. For $\rho$ close to zero (resp. close to one), eigenvalues decay fast (resp. slowly) and residual functions are more (resp. less) smooth.

We put $c_1 = [1, 2.3, 3.4, 1.5]^T$ and $u = [1, 2, 3, 4]^T / \sqrt{30}$. The vector $c_i = c_1 + (i - 1)\delta u$ for different values of $\delta$ represents the mean functions of the $k$ groups. The parameter $\delta$ switches between null and alternative hypotheses.

We fix $q = 100$, $a = 1.5$, $T = 80$ where $T$ is the number of time points where each curve is observed. For the eigenfunctions, we put $\psi_1(t) = 1$, $\psi_{2r}(t) = \sqrt{2}\sin(2\pi rt)$ and $\psi_{2r+1}(t) = \sqrt{2}\cos(2\pi rt)$ for $r = 1, \ldots, q$. Different setups for data generating procedure is a combination of the following set of parameters:

- $z_{ijr} \sim N(0, 1)$ or $z_{ijr} \sim t_4 / \sqrt{2}$.
- $\rho = 0.1, 0.5, 0.9$ for different level of smoothness of residual functions.
- $(n_i)^3 = (20, 30, 30)$ for the small sample and $(n_i)^3 = (70, 80, 100)$ for the large sample cases.

The parameter $\delta$ for each pair of $\rho$ and $n_i$ is selected in a way that the difference between the performance of tests can be distinguished. For the test statistics, $F$-type, GPF, and $F_{\text{max}}$, the authors proposed bootstrap methods to estimate the null distribution of the test statistic, however, in this section we used permutation strategy for estimating the null distribution. Consult Zhang (2014, p. 150) for the implementation of the $F$-type test, and Zhang et al. (2019) for the implementation of the GPF and $F_{\text{max}}$ tests.

Again for the MMD test B-spline basis has been used in the smoothing procedure, and the number of components for the smoothing procedure is considered to be fixed and equals 41. The kernel bandwidth is also chosen by median of pairwise distances and minimum p-value bandwidth method. The critical points for MMD and all other competitors are obtained by approximating the null distribution using random permutation. For this purpose the cases are iteratively shuffled within the groups, while ensuring that the random permutations maintain the original group sizes. By calculating the test statistic values for each permutation, an approximation of the null distribution for all test statistics is obtained.

Mention should be made that although in this paper the kernel mean embedding of probability measures and the MMD statistic are derived for the family of Gaussian probability distributions, the new MMD test, accompanied by double minimum p-value approach for bandwidth selection, is superior even in non-Gaussian scenarios. Simulation results are produced and reported by 2000 iterations in Tables 3 and 4.

### 4.3 Testing for equality of covariance operators

Let $\mathbb{H}$ be a separable Hilbert space. Assume that $Y_{ij}$ for $i = 1, \ldots, k$ and $j = 1, \ldots, n_i$ are independent $\mathbb{H}$-valued Gaussian random elements, and $y_{ij}$ are their observation counterparts generated from
TABLE 3 Type-I errors (in bold) and empirical powers of $L^2$, $F$, GPF, and $F_{\text{max}}$ and MMD$_0$ for one-way ANOVA problem when $z_{ijr} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$. All numbers are presented as percentages.

| $\rho \times 100$ | $(n_i) = (20, 30, 30)$ | $(n_i) = (70, 80, 100)$ |
|------------------|-----------------------|------------------------|
| $\delta \times 100$ | 0 | 0.5 | 20.0 | 40.0 | 60.0 | 0 | 0.1 | 10.0 | 20.0 | 30.0 |
| $\rho = 0.1$ | $F$ | 5.1 | 5.7 | 9.6 | 28.6 | 59.0 | 5.2 | 5.3 | 8.5 | 23.9 | 52.9 |
| | GPF | 5.1 | 5.6 | 9.6 | 28.5 | 59.0 | 5.3 | 5.3 | 8.5 | 24.0 | 52.8 |
| | $F_{\text{max}}$ | 5.3 | 5.8 | 25.1 | 81.2 | 99.1 | 5.3 | 5.1 | 23.8 | 75.2 | 98.8 |
| | MMD($\sigma_0$) | 5.1 | 5.6 | 10.8 | 41.1 | 88.6 | 5.3 | 5.8 | 9.7 | 35.6 | 84.6 |
| | MMD($\pi^2$) | 5.9 | 78.6 | 100.0 | 100.0 | 100.0 | 5.1 | 19.6 | 100.0 | 100.0 | 100.0 |
| $\rho = 0.5$ | $F$ | 5.0 | 5.6 | 9.5 | 26.3 | 56.9 | 5.9 | 4.7 | 9.2 | 23.1 | 50.4 |
| | GPF | 4.9 | 5.2 | 9.6 | 26.7 | 58.1 | 5.9 | 4.9 | 9.2 | 22.9 | 51.3 |
| | $F_{\text{max}}$ | 4.8 | 4.9 | 16.6 | 61.4 | 96.0 | 5.1 | 4.7 | 14.4 | 58.2 | 94.3 |
| | MMD($\sigma_0$) | 5.1 | 5.1 | 9.7 | 30.0 | 68.0 | 5.3 | 4.7 | 9.6 | 26.2 | 60.6 |
| | MMD($\pi^2$) | 4.5 | 84.6 | 100.0 | 100.0 | 100.0 | 5.9 | 86.2 | 100.0 | 100.0 | 100.0 |
| $\rho = 0.9$ | $F$ | 5.1 | 5.0 | 7.1 | 18.4 | 41.8 | 4.0 | 5.2 | 7.4 | 16.2 | 36.4 |
| | GPF | 4.9 | 4.8 | 7.3 | 18.2 | 41.4 | 4.0 | 5.2 | 7.4 | 16.4 | 36.0 |
| | $F_{\text{max}}$ | 4.8 | 4.9 | 5.8 | 13.8 | 33.2 | 4.8 | 5.7 | 6.4 | 13.9 | 27.8 |
| | MMD($\sigma_0$) | 4.9 | 4.8 | 6.7 | 18.0 | 40.0 | 4.9 | 5.2 | 7.3 | 14.9 | 34.2 |
| | MMD($\pi^2$) | 5.6 | 97.9 | 100.0 | 100.0 | 100.0 | 4.6 | 85.9 | 100.0 | 100.0 | 100.0 |

the following model:

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, C_i); \quad i = 1, \ldots, k, j = 1, \ldots, n_i, \quad (18)$$

where $\mu_i$ is the unknown mean function of group $i$, and $\varepsilon_{ij}$ accounts for subject-effect functions with mean zero and covariance operator $C_i = \mathbb{E} [\varepsilon_{ij} \otimes H \varepsilon_{ij}]$. It is of interest to test the equality of $k$ covariance operators, that is, $H_0 : C_1 = \cdots = C_k$ versus alternative hypothesis $H_1$ which states at least two groups differs in their covariance functions. Based on the proof of Proposition 1, the squared MMD for comparing two Gaussian probability measures $\mathcal{N}(\mu_1, C_1)$ and $\mathcal{N}(\mu_2, C_2)$ equal to

$$\left\| m_{\mathcal{N}(\mu_1, C_1)} - m_{\mathcal{N}(\mu_2, C_2)} \right\|^2_{\mathcal{H}_k}$$

$$= \left\| m_{\mathcal{N}(\mu_1, C_1)} \right\|^2_{\mathcal{H}_k} + \left\| m_{\mathcal{N}(\mu_2, C_2)} \right\|^2_{\mathcal{H}_k} - 2 \left\langle m_{\mathcal{N}(\mu_1, C_1)}, m_{\mathcal{N}(\mu_2, C_2)} \right\rangle_{\mathcal{H}_k}$$

$$= \left| I + 4 \frac{1}{\sigma} C_1 \right|^{-1/2} + \left| I + 4 \frac{1}{\sigma} C_2 \right|^{-1/2} - 2 \left| I + 2 \frac{1}{\sigma} (C_1 + C_2) \right|^{-1/2} e^{-\frac{1}{4} \left\langle (I+2 \frac{1}{\sigma} (C_1 + C_2))^{-1} (\mu_1 - \mu_2), (\mu_1 - \mu_2) \right\rangle_{\mathcal{H}}}.$$

To develop the MMD statistic based on a simple random sample from the model (18), first, we have to choose a proper characteristic kernel for the space $\mathbb{H}^n$ where $n = \sum_{i=1}^{k} n_i$. By Proposition 1,
### Table 4

Type-I errors (in bold) and empirical powers of $L^2$, $F$, GPF, and $F_{\text{max}}$ and MMD$_0$ for one-way ANOVA problem when $z_{ijr} \sim t_4/\sqrt{2}$. All numbers are presented as percentages.

| $\delta \times 100$ | $(n_i) = (20, 30, 30)$ | $(n_i) = (70, 80, 100)$ |
|---------------------|------------------------|------------------------|
|                     | 0  | 0.5 | 20.0 | 40.0 | 60.0 | 0  | 0.1 | 10.0 | 20.0 | 30.0 |
| $\rho = 0.1$        |    |     |      |      |      |    |     |      |      |      |
| $F$                 | **4.9** | 4.4 | 9.8  | 29.8 | 61.9 | **5.1** | 4.4 | 10.0 | 23.0 | 53.3 |
| GPF                 | **4.9** | 4.4 | 9.8  | 29.6 | 61.6 | **5.1** | 4.4 | 10.0 | 23.0 | 53.1 |
| $F_{\text{max}}$    | **4.8** | 4.2 | 28.4 | 83.2 | 98.6 | **4.7** | 4.4 | 25.3 | 78.1 | 98.2 |
| MMD($\sigma_0$)     | **4.9** | 4.0 | 12.6 | 50.3 | 94.6 | **5.0** | 4.2 | 11.8 | 41.3 | 92.9 |
| MMD($\pi^2$)        | **4.9** | 80.8 | 100.0 | 100.0 | 100.0 | **4.9** | 18.5 | 100.0 | 100.0 | 100.0 |
| $\rho = 0.5$        |    |     |      |      |      |    |     |      |      |      |
| $F$                 | **5.8** | 4.8 | 11.2 | 27.7 | 60.8 | **5.6** | 5.0 | 9.1  | 24.0 | 51.9 |
| GPF                 | **5.5** | 4.6 | 11.2 | 28.1 | 61.8 | **5.7** | 5.0 | 9.3  | 24.4 | 53.2 |
| $F_{\text{max}}$    | **5.3** | 4.3 | 18.4 | 65.7 | 96.8 | **5.3** | 5.2 | 14.6 | 59.3 | 93.8 |
| MMD($\sigma_0$)     | **5.6** | 4.6 | 12.2 | 33.5 | 72.6 | **5.1** | 4.8 | 9.8  | 29.3 | 65.3 |
| MMD($\pi^2$)        | **5.3** | 80.1 | 100.0 | 100.0 | 100.0 | **5.3** | 78.6 | 100.0 | 100.0 | 100.0 |
| $\rho = 0.9$        |    |     |      |      |      |    |     |      |      |      |
| $F$                 | **4.7** | 5.3 | 8.3  | 19.7 | 42.2 | **4.3** | 5.1 | 8.0  | 18.5 | 37.1 |
| GPF                 | **4.7** | 5.4 | 8.1  | 20.0 | 41.8 | **4.2** | 5.0 | 7.6  | 18.4 | 37.6 |
| $F_{\text{max}}$    | **4.6** | 5.1 | 6.9  | 14.9 | 34.0 | **4.2** | 5.1 | 7.3  | 14.8 | 31.1 |
| MMD($\sigma_0$)     | **4.6** | 5.1 | 7.6  | 19.6 | 41.2 | **4.3** | 4.9 | 7.8  | 18.0 | 36.5 |
| MMD($\pi^2$)        | **5.3** | 96.9 | 99.8 | 100.0 | 100.0 | **5.2** | 87.1 | 100.0 | 100.0 | 100.0 |

Gaussian kernel $k(x, y) = e^{-\frac{1}{\sigma} \|x - y\|^2}$ is characteristic for the family of Gaussian probability measures on $\mathbb{H}$, and from Proposition 2, $(x_i^n, y_i^n) \mapsto \sum_{i=1}^n k(x_i, y_i)$ is a characteristic kernel for the family of finite product of Gaussian probability measures on $\mathbb{H}^n$. Let $P_0$ be the probability measure of samples generated by the model (18) under $H_0$, that is, $C_1 = \ldots = C_k = C$ and $P_1$ be the probability measure induced by the model (18), when parameters $C_i$ considered to be free. For the centered version of the model (18), that is, $\mu_1 = \ldots = \mu_k = 0$, the squared MMD with Gaussian Sum-Kernel equals

$$MMD^2 = \|m_{P_0} - m_{P_1}\|_{H_k}^2$$

$$= n \left| \left| I + 4 \frac{1}{\sigma} C \right|^{-1/2} + \sum_{i=1}^k n_i \left| \left| I + 4 \frac{1}{\sigma} C_i \right|^{-1/2} - 2 \sum_{i=1}^k n_i \left| \left| I + 2 \frac{1}{\sigma} (C + C_i) \right|^{-1/2} \right. \right.$$  

$$\left. \right| - 2 \sum_{i=1}^k n_i \left( \left| \left| I + 4 \frac{1}{\sigma} C \right|^{-1/2} + \left| \left| I + 4 \frac{1}{\sigma} C_i \right|^{-1/2} - 2 \left| \left| I + 2 \frac{1}{\sigma} (C + C_i) \right|^{-1/2} \right) \right. \right.$$  

$$\left. \right|.$$  

There currently developed tests for the $k$-sample equality of covariance functions problem, if $\mathbb{H}$ considered being the space of square-integrable functions over a compact set like $L^2[0, 1]$. We address two recent successfully developed tests for homogeneity of covariance functions, namely quasi-GPF and quasi-$F_{\text{max}}$ introduced by Guo et al. (2019). There are other tests, which are shown to be less powerful than the currently mentioned tests in different settings. See Guo et al. (2019) and references therein for more information and simulation studies. We compared the new MMD test against quasi-GPF and quasi-$F_{\text{max}}$ in a simulation study. Our simulation study is motivated
by Guo et al. (2019), and we used the same setup for the data generating procedure. Assume that the mean function is zero and data is generated in a $k$-regime scheme according to the following model:

$$y_{ij}(t) = \varepsilon_{ij}(t),$$

$$\varepsilon_{ij}(t) = h(t) \sum_{r \geq 1} \sqrt{\lambda_r} z_{ijr} \psi_r(t),$$

$$i = 1, \ldots, k; \quad j = 1, \ldots, n_i; \quad t \in [0, 1].$$

where $h(t)$ is common for all the groups and $n_i$ denotes the size of each group. The set $\{\psi_r\}_{r \geq 1}$, for each $i$, is a set of basis functions, and we set the eigenvalues $\lambda_r = a \rho^{r-1}$ for fixed $a > 0$ and $\rho \in (0, 1)$. The tuning parameter $\rho$ determines the decay rate of eigenvalues. For a $\rho$ close to zero, eigenvalues decay fast and functional data is more correlated and more smooth, however, for a $\rho$ close to one, eigenvalues decay slowly and realization of functional data is less correlated across its domain and thus less smooth.

Although Guo et al. (2019) assumed a finite number of $q$ nonzero eigenvalues in the simulation process, our test works well without the need to put any restriction on the number of components. Here we follow Guo et al. (2019) and fix $q = 100, a = 1.5, T = 80, h(t) = \frac{T}{t+1}$ where $T$ is the number of time points that each curve is observed at. Different setups for data generating procedure is a combination of the following choice of parameters:

- $z_{ijr} \overset{i.i.d.}{\sim} N(0, 1)$ or $z_{ijr} \overset{i.i.d.}{\sim} \sqrt{3/5} t_5$.
- $\rho = 0.1, 0.5, 0.9$ for three class of high, moderate, and low correlations.
- $(n_i) = (20, 30, 30)$ for the small sample and $(n_i) = (70, 80, 100)$ for the large sample cases.
- $\psi_{ir}(t) = \phi_r(t)$ for $r = 1, 3, 4, \ldots, q$ and $\psi_{i2}(t) = \phi_2(t) + (i-1)\omega/h(t)$ for different choice of $\omega$ to reflect between group difference of covariance operators, and we can take $\{\phi_r\}_{r \geq 1}$ either the set of Fourier or B-spline basis for $L^2[0, 1]$.

Guo et al. (2019) compared quasi-GPF and quasi-$F_{\text{max}}$ with few other tests including two other tests $L^2$ and $T_{\text{max}}$ (Guo et al., 2018). According to their results, quasi-GPF is superior in low correlation schemes, and quasi-$F_{\text{max}}$ is superior in the high correlation schemes.

Let $\hat{C}_i$ be the usual estimation of the covariance operator under $H_0$ and $\hat{C}_i$ the usual covariance operator estimation of group $i$. Then our MMD test statistic equals

$$\overline{\text{MMD}} = \left[ \sum_{i=1}^{k} n_i \left( \left| I + 4 \frac{1}{\sigma} \hat{C}_i \right|^{-1/2} + \left| I + 4 \frac{1}{\sigma} \hat{C}_i \right|^{-1/2} - 2 \left| I + 2 \frac{1}{\sigma} (\hat{C} + \hat{C}_i) \right|^{-1/2} \right) \right]^{1/2}.$$

The critical points for MMD and all other of all competitors $L^2$, $T_{\text{max}}$, $F_{\text{max}}$, and GPF are obtained by approximating the null distribution using random permutation. This method involves iteratively assigning the cases into various groups, ensuring that the random permutations respect the original sizes of the groups. The null distribution for all the test statistics are then approximated by computing the values of the test statistics for each permutation. The empirical powers of the five test statistics are calculated in a simulation study with 2000 repetitions are reported in Tables 5 and 6 while setting $\alpha = 0.05$ and $\{\phi_r\}_{r \geq 1}$ selected to be the set of Fourier basis. In this simulation study, we used the B-spline basis for the smoothing procedure in MMD. The
TABLE 5  Type-I errors (in bold) and empirical powers of $L^2$, $T_{\text{max}}$, GPF, $F_{\text{max}}$ and MMD when $Z_{ij} \sim N(0,1)$. All numbers are presented as percentages.

| $\phi$ | $(n_i) = (20, 30, 30)$ | $(n_i) = (70, 80, 100)$ |
|-------|----------------------|------------------------|
|       | 0    | 0.25 | 0.50 | 0.75 | 1.00 | 0    | 0.15 | 0.30 | 0.45 | 0.60 |
| $\rho = 0.1$ |       |       |       |       |       |       |       |       |       |       |
| $L^2$  | 5.4  | 7.8  | 23.8 | 55.0 | 79.8 | 5.5  | 9.8  | 29.3 | 70.4 | 94.2 |
| $T_{\text{max}}$ | 5.4  | 9.5  | 24.3 | 55.1 | 82.9 | 4.2  | 13.2 | 37.2 | 75.3 | 95.8 |
| $F_{\text{max}}$ | 5.0  | 9.3  | 28.4 | 57.2 | 81.2 | 4.7  | 12.2 | 42.3 | 83.0 | 97.4 |
| GPF   | 5.7  | 7.4  | 21.3 | 53.3 | 79.1 | 5.2  | 8.4  | 23.6 | 64.6 | 92.7 |
| MMD($\sigma_0$) | 5.3  | 8.1  | 36.7 | 65.6 | 88.9 | 5.3  | 9.5  | 28.6 | 78.1 | 98.8 |
| MMD($\pi^2$) | 5.2  | 87.6 | 99.9 | 100.0| 100.0| 5.6  | 76.4 | 100.0| 100.0| 100.0|
| $\rho = 0.5$ |       |       |       |       |       |       |       |       |       |       |
| $L^2$  | 5.1  | 17.5 | 58.3 | 89.7 | 97.4 | 5.9  | 27.2 | 84.8 | 99.7 | 100.0|
| $T_{\text{max}}$ | 4.3  | 9.3  | 25.4 | 58.9 | 81.6 | 5.1  | 14.1 | 42.9 | 82.2 | 98.7 |
| $F_{\text{max}}$ | 4.6  | 12.8 | 39.6 | 78.1 | 93.4 | 5.1  | 17.5 | 62.9 | 95.3 | 99.9 |
| GPF   | 5.5  | 17.7 | 62.0 | 91.3 | 98.3 | 5.9  | 25.8 | 87.4 | 99.8 | 100.0|
| MMD($\sigma_0$) | 4.9  | 16.0 | 47.2 | 84.0 | 97.5 | 5.2  | 22.0 | 53.1 | 78.6 | 94.4 |
| MMD($\pi^2$) | 5.8  | 11.7 | 38.1 | 80.6 | 96.7 | 5.2  | 17.0 | 52.0 | 90.6 | 99.9 |
| $\rho = 0.9$ |       |       |       |       |       |       |       |       |       |       |
| $L^2$  | 4.8  | 9.2  | 43.2 | 83.8 | 97.0 | 6.0  | 11.8 | 58.1 | 97.4 | 100.0|
| $T_{\text{max}}$ | 5.3  | 4.7  | 6.5  | 10.8 | 26.1 | 6.3  | 5.1  | 6.8  | 13.2 | 31.4 |
| $F_{\text{max}}$ | 5.3  | 5.8  | 8.2  | 19.0 | 43.7 | 5.3  | 5.9  | 8.7  | 22.8 | 53.7 |
| GPF   | 4.8  | 9.4  | 42.2 | 84.2 | 97.7 | 5.8  | 12.5 | 57.6 | 97.1 | 100.0|
| MMD($\sigma_0$) | 4.7  | 11.3 | 46.1 | 85.7 | 98.2 | 5.1  | 12.9 | 63.5 | 95.0 | 99.9 |
| MMD($\pi^2$) | 5.4  | 9.8  | 36.6 | 75.6 | 96.0 | 5.6  | 9.7  | 50.4 | 92.5 | 100.0|

According to the results, the empirical powers of MMD test are higher than the other four tests in when the process is more smooth ($\rho = 0.1$). GPF outperforms MMD when the process is moderately smooth ($\rho = 0.5$).

4.3.1 Medfly data

In this section, we apply MMD and the other four tests introduced in Section 4.3, $L^2$, $T_{\text{max}}$, GPF and $F_{\text{max}}$, to test for homogeneity of covariance operators in a real data example, according to the model (18) with $k = 4$.

Medfly dataset is a functional data of mortality rate of medflies. Approximately, 7200 medflies of a given size were maintained in aluminum cages. Adults were given either a diet of sugar and water, or a diet of sugar, water, and ad libitum. Each day, dead flies were removed, counted, and
| $\delta$ | \( (n_i) = (20, 30, 30) \) | \( (n_i) = (70, 80, 100) \) |
|---|---|---|
| $\rho = 0.1$ | | |
| $L^2$ | 4.9 | 6.6 | 16.1 | 33.1 | 47.5 | 5.2 | 5.8 | 14.8 | 34.2 | 59.7 |
| $T_{\text{max}}$ | 5.1 | 7.6 | 14.8 | 28.1 | 49.6 | 5.1 | 8.5 | 19.5 | 38.6 | 62.4 |
| $F_{\text{max}}$ | 5.5 | 8.4 | 19.4 | 35.8 | 52.2 | 5.9 | 9.2 | 24.6 | 51.8 | 72.7 |
| GPF | 5.2 | 6.6 | 15.6 | 32.5 | 48.4 | 5.1 | 5.8 | 12.7 | 31.0 | 56.4 |
| MMD(\(\sigma_0\)) | 5.3 | 7.9 | 24.8 | 47.6 | 66.7 | 5.0 | 6.0 | 15.5 | 47.7 | 79.0 |
| MMD(\(\pi^2\)) | 5.5 | 58.0 | 91.9 | 99.1 | 99.9 | 4.8 | 75.2 | 100.0 | 100.0 | 100.0 |
| $\rho = 0.5$ | | |
| $L^2$ | 5.1 | 11.6 | 31.3 | 56.6 | 70.3 | 5.1 | 12.2 | 45.1 | 77.2 | 91.4 |
| $T_{\text{max}}$ | 5.0 | 7.3 | 14.6 | 29.9 | 47.8 | 5.3 | 8.9 | 20.0 | 43.8 | 71.9 |
| $F_{\text{max}}$ | 5.1 | 10.1 | 26.6 | 51.5 | 69.6 | 4.9 | 11.8 | 40.1 | 74.0 | 91.5 |
| GPF | 5.6 | 12.3 | 37.5 | 63.1 | 76.0 | 5.1 | 12.8 | 50.2 | 81.7 | 93.8 |
| MMD(\(\sigma_0\)) | 5.1 | 9.2 | 29.3 | 58.4 | 81.0 | 5.0 | 10.1 | 31.8 | 62.2 | 83.0 |
| MMD(\(\pi^2\)) | 5.7 | 7.3 | 20.6 | 48.8 | 74.8 | 5.0 | 7.6 | 24.8 | 60.8 | 88.9 |
| $\rho = 0.9$ | | |
| $L^2$ | 5.1 | 5.8 | 18.0 | 44.3 | 66.3 | 5.3 | 7.8 | 26.4 | 67.1 | 87.8 |
| $T_{\text{max}}$ | 5.2 | 4.7 | 5.2 | 8.1 | 13.3 | 5.3 | 5.1 | 6.9 | 9.6 | 17.3 |
| $F_{\text{max}}$ | 5.0 | 5.8 | 6.5 | 12.5 | 26.3 | 4.6 | 6.2 | 8.1 | 16.9 | 36.5 |
| GPF | 5.0 | 6.4 | 19.8 | 48.0 | 70.9 | 4.7 | 8.0 | 28.1 | 70.0 | 91.5 |
| MMD(\(\sigma_0\)) | 4.7 | 7.3 | 22.8 | 49.4 | 75.2 | 5.1 | 6.5 | 24.6 | 62.3 | 86.3 |
| MMD(\(\pi^2\)) | 5.3 | 6.0 | 17.6 | 38.2 | 66.4 | 5.9 | 5.9 | 16.0 | 48.2 | 79.7 |

**TABLE 6** Type-I errors (in bold) and empirical powers of $L^2$, $T_{\text{max}}$, GPF, $F_{\text{max}}$ and MMD when $z_{qr} \sim \sqrt{3/5} \, t_5$. All numbers are presented as percentages.

Their sex determined (Carey et al., 1992). The number and rate of alive medflies were recorded over a period of 101 days. In effect, the aim is to assess the effects of nutrition and gender on survival or mortality of medflies.

Cohorts of medflies consist of four groups, (a) Females on a sugar diet, (b) Females on a protein plus sugar diet, (c) Males on a sugar diet, and (d) Males on a protein plus sugar diet. The effect of gender and nutrition is studied before (see e.g., Chiou et al., 2003; Koenker & Geling, 2001), and it is known that there is an interaction between gender and nutrition on the survival of medflies (Müller & Wang, 1998). Survival functions of cohorts of medflies during a period of 30 days (days 2–31) are illustrated in Figure 1.

The panels in the first row demonstrate 33 sample functions in each group as well as the mean functions, and each sample function is the survival rate of medflies in one cage. The panels in the second row demonstrate deviation of samples from the group’s mean function as well as the first eigenfunction of the covariance operator. The first eigenfunction explains the major variation of functional samples within each group (89.3%, 94.5%, 96.4%, and 97.8% in each group, respectively). It could be noticed that there is a slight difference in eigenvalues and eigenfunctions of covariance operators between groups. The kernel functions of the covariance operators in the four
FIGURE 1 Survival functions of cohorts of medflies: (a) Females on a sugar diet, (b) Females on a protein plus sugar diet, (c) Males on a sugar diet, and (d) Males on a protein plus sugar diet; (Top row) Gray lines: survival functions of samples, Red line: mean function, (Bottom row) Gray lines: deviation of samples from the mean function, Blue line: first eigenfunction of covariance operators, which accounts for 89.3%, 94.5%, 96.4%, and 97.8% of the variation of survival functions in each of four groups, respectively.

FIGURE 2 Estimated covariance functions of the four groups of medflies: (a) Females on a sugar diet, (b) Females on a protein plus sugar diet, (c) Males on a sugar diet, and (d) Males on a protein plus sugar diet.

Groups are depicted in Figure 2, which magnifies the between-groups difference of covariance operators.

MMD and the four other tests including $L^2$, $T_{\text{max}}$, $F_{\text{max}}$, and GPF were employed to test the equality of covariance operators. The results are presented in Table 7.

It can be understood that the p-values of all pair-wise comparisons by MMD($\pi^2$) are generally smaller than of the other four tests. As described in Guo et al. (2019), it was expected that $F_{\text{max}}$ test to have higher power than GPF in this dataset. However, as it is shown in simulation studies, MMD test based on the Gaussian kernel has higher power than both $F_{\text{max}}$ and GPF when the underlying process is more smooth.
TABLE 7  
*p*-Values (in percent) of $L^2$, $T_{\text{max}}$, global point-wise $F$ (GPF), $F_{\text{max}}$ and maximum mean discrepancy (MMD) tests applied to compare covariance operators of survival functions of the four groups of medflies.

| Group  | $L^2$  | $T_{\text{max}}$ | $F_{\text{max}}$ | GPF | MMD($\sigma_0$) | MMD($\pi^2$) |
|--------|--------|-----------------|-----------------|-----|----------------|--------------|
| (a) vs (b) | 1.1    | 0.1             | 0.2             | 0.2 | 0.1            | 0.8          |
| (a) vs (c) | 0.1    | 0.1             | 0.1             | 0.2 | 4.2            | 0.8          |
| (a) vs (d) | 0.1    | 0.1             | 2.8             | 3.2 | 0.8            | 0.6          |
| (b) vs (c) | 0.1    | 0.2             | 0.8             | 12.0| 0.1            | 0.6          |
| (b) vs (d) | 8.6    | 12.0            | 4.2             | 38.0| 58.8           | 0.6          |
| (c) vs (d) | 14.2   | 5.6             | 16.0            | 19.2| 67.6           | 1.8          |
| All groups | 0.1    | 0.1             | 0.6             | 1.6 | 0.8            | 1.8          |

5 CONCLUSIONS AND DISCUSSION

This study explored kernel mean embedding of probability measures and its applications to functional data analysis. We derived a framework for introducing a pseudo-likelihood function over infinite-dimensional separable Hilbert spaces, and derived MKM estimators of location parameters and covariance operators by maximizing this pseudo-likelihood function. Plug-in estimators of probability measures are then employed to develop MMD-based tests for few common problems in functional data analysis including one-way ANOVA and homogeneity of covariance operators.

The null distribution of developed tests approximated by permutation strategy and power of tests studied and compared in multiple simulation studies. In the one-way ANOVA, the MMD test with the double permutation strategy generally had a better performance compared to the competitors. In the problem of homogeneity of covariance operators the performance of MMD test was higher compared to the competitors when the process was more smooth ($\rho = 0.1$). In the testing problem for location parameters of the function-on-scalar regression, the MMD test was generally better than $pffr$ and could compete with the tests based on hyper-ellipsoid confidence regions.

We employed two methods to tackle with the problem of kernel bandwidth selection, however, choosing the optimal kernel bandwidth is still an important open question as the performance of tests are dependent to the choose of kernel bandwidth. It is also crucial to acknowledge that a permutation strategy is utilized to approximate the null distribution of the new MMD test. However, its implementation is restricted in cases where such algorithm is not viable.

ACKNOWLEDGMENTS

The first author is grateful to the Graduate office of the University of Isfahan for their support. Part of this work was done while Saeed Hayati was visiting in the Institute of Statistical Mathematics under the support by the Research Organization of Information and Systems. Kenji Fukumizu has been supported in part by JSPS KAKENHI 18K19793. Afshin Parvardeh gratefully thanks Professor Victor Panaretos and EPFL in Switzerland for the kind hospitality that received during spending his sabbatical leave at EPFL, in which this work, in part, was prepared.
REFERENCES

Azzalini, A. (2013). *The skew-normal and related families* (Vol. 3). Cambridge University Press.

Carey, J. R., Liedo, P., Orozco, D., & Vaupel, J. W. (1992). Slowing of mortality rates at older ages in large medfly cohorts. *Science*, 258(5081), 457–461. https://doi.org/10.1126/science.1411540

Chiou, J.-M., Müller, H.-G., Wang, J.-L., & Carey, J. R. (2003). A functional multiplicative effects model for longitudinal data, with application to reproductive histories of female medflies. *Statistica Sinica*, 13(4), 1119–1133.

Choi, H., & Reimherr, M. (2018). A geometric approach to confidence regions and bands for functional parameters. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 80(1), 239–260. https://doi.org/10.1111/rssb.12239

Conway, J. B. (2014). *A course in point set topology*. Undergraduate texts in mathematics. Springer. https://doi.org/10.1007/978-3-319-02368-7

Delaigle, A., & Hall, P. (2010). Defining probability density for a distribution of random functions. *Annals of Statistics*, 38(2), 1171–1193. https://doi.org/10.1214/09-AOS741

Fukumizu, K., Gretton, A., Lanckriet, G., Schölkopf, B. & Sriperumbudur, B. K. (2009). *Kernel choice and classifiability for rkhs embeddings of probability distributions*. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, & A. Culotta (Eds.), *Advances in Neural Information Processing Systems* (Vol. 22). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2009/file/685ac8cad1be5ac98da9556bc1c8d9e-Paper.pdf

Gill, T., Kirtadze, A., Pantsulaia, G., & Plichko, A. (2014). Existence and uniqueness of translation invariant measures in separable Banach spaces. *Functiones et Approximatio Commentarii Mathematici*, 50(2), 401–419. https://doi.org/10.7169/facm/2014.50.2.12

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477), 359–378. https://doi.org/10.1198/016214506000001437

Gretton, A., Sejdinovic, D., Strathmann, H., Balakrishnan, S., Pontil, M., Fukumizu, K., & Sriperumbudur, B. K. (2012). Optimal kernel choice for large-scale two-sample tests. *Advances in Neural Information Processing Systems*, 25, 1–25.

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

Greven, S., Scheipl, F., Greven, S., & Scheipl, F. (2017). A general framework for functional regression modelling. *Statistical Modelling*, 17(1-2), 1–35. https://doi.org/10.1177/1471082X16681317

Guo, J., Zhou, B., & Zhang, J.-T. (2018). Testing the equality of several covariance functions for functional data: A supremum-norm based test. *Computational Statistics & Data Analysis*, 124, 15–26. https://doi.org/10.1016/j.csda.2018.02.002

Guo, J., Zhou, B., & Zhang, J.-T. (2019). New tests for equality of several covariance functions for functional data. *Journal of the American Statistical Association*, 114(527), 1251–1263. https://doi.org/10.1080/01621459.2018.1483827

Harchaoui, Z., Bach, F., Cappe, O., & Moulines, E. (2013). Kernel-based methods for hypothesis testing: A unified view. *IEEE Signal Processing Magazine*, 30(4), 87–97.

Harchaoui, Z., Moulines, E., & Bach, F. R. (2009). *Kernel change-point analysis*. In *Advances in neural information processing systems* (Vol. 21, pp. 609–616). Curran Associates, Inc.

Koenker, R., & Geling, O. (2001). Reappraising medfly longevity: A quantile regression survival analysis. *Journal of the American Statistical Association*, 96(454), 458–468. https://doi.org/10.1198/016214501753168172

Kokoszka, P., & Reimherr, M. (2017). Discussion of ‘a general framework for functional regression modelling’ by Greven and Scheipl. *Statistical Modelling*, 17(1-2), 45–49. https://doi.org/10.1177/1471082X16681331

Lin, Z., Müller, H.-G., & Yao, F. (2018). Mixture inner product spaces and their application to functional data analysis. *Annals of Statistics*, 46(1), 370–400. https://doi.org/10.1214/17-AOS1553

Maniglia, S., & Rhandi, A. (2004). Gaussian measures on separable hilbert spaces and applications. *Quaderni di Matematica*, 1, 2004.
Martínez-Camblor, P., De Una-Alvarez, J., & Corral, N. (2008). k-sample test based on the common area of kernel density estimators. *Journal of Statistical Planning and Inference, 138*(12), 4006–4020.

Minh, H. Q. (2017). Infinite-dimensional log-determinant divergences between positive definite trace class operators. *Linear Algebra and its Applications, 528*, 331–383. https://doi.org/10.1016/j.laa.2016.09.018

Muandet, K., Fukumizu, K., Sriperumbudur, B., & Schölkopf, B. (2017). Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning, 10*(1-2), 1–141.

Müller, H.-G., & Wang, J.-L. (1998). *Statistical tools for the analysis of nutrition effects on the survival of cohorts* (pp. 191–203). Springer US. https://doi.org/10.1007/978-1-4899-1959-5_12

Pan, W., Tian, Y., Wang, X., & Zhang, H. (2018). Ball divergence: Nonparametric two sample test. *Annals of Statistics, 46*(3), 1109–1137. https://doi.org/10.1214/17-AOS1579

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

Sejdinovic, D., Sriperumbudur, B., Gretton, A., & Fukumizu, K. (2013). Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The Annals of Statistics, 41*(5), 2263–2291.

Shen, Q., & Faraway, J. (2004). An F test for linear models with functional responses. *Statistica Sinica, 14*(4), 1239–1257.

Simon-Gabriel, C.-J., & Schölkopf, B. (2018). Kernel distribution embeddings: Universal kernels, characteristic kernels and kernel metrics on distributions. *Journal of Machine Learning Research, 19*(29), 44.

Smola, A., Gretton, A., Song, L., & Schölkopf, B. (2007). A hilbert space embedding for distributions. In *Algorithmic learning theory* (Vol. 4754, pp. 13–31). Lecture Notes in Computer Science. Springer.

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B., & Lanckriet, G. R. G. (2010). Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research, 11*, 1517–1561.

Steinwart, I. (2001). On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research, 2*, 67–93.

Steinwart, I., & Ziegel, J. F. (2021). Strictly proper kernel scores and characteristic kernels on compact spaces. *Applied and Computational Harmonic Analysis, 51*, 510–542. https://doi.org/10.1016/j.acha.2019.11.005

Tang, M., Athreya, A., Sussman, D. L., Lyzinski, V., & Priebe, C. E. (2017). A nonparametric two-sample hypothesis testing problem for random graphs. *Bernoulli, 23*(3), 1599–1630. https://doi.org/10.3150/15-BEJ789

Zhang, J.-T. (2014). *Analysis of variance for functional data Monographs on statistics and applied probability* (Vol. 127). CRC Press.

Zhang, J.-T., Cheng, M.-Y., Wu, H.-T., & Zhou, B. (2019). A new test for functional one-way ANOVA with applications to ischemic heart screening. *Computational Statistics & Data Analysis, 132*, 3–17. https://doi.org/10.1016/j.csda.2018.05.004

Zhang, J.-T., & Liang, X. (2014). One-way ANOVA for functional data via globalizing the pointwise F-test. *Scandinavian Journal of Statistics, 41*(1), 51–71. https://doi.org/10.1111/sjos.12025

---

**How to cite this article:** Hayati, S., Fukumizu, K., & Parvardeh, A. (2024). Kernel mean embedding of probability measures and its applications to functional data analysis. *Scandinavian Journal of Statistics, 51*(2), 447–484. https://doi.org/10.1111/sjos.12691

---

**APPENDIX A. PROOFS**

**A.1 Proof of Theorem 1 and Corollary 1**

To provide the proof of Theorem 1 we need the following lemma:

**Lemma 1.** Let \{b_j\}_{j\geq1} be a descending sequence of positive real numbers and \{a_j\}_{j\geq1} be a series of real numbers such that \(\sum_{j\geq1} |a_j| < \infty\) and \(\sum_{j\geq1} a_j b_j > 0\). Then there exists a finite \(N \in \mathbb{N}\) such that \(\sum_{j=1}^{N} a_j > 0\).
Proof. Let \( P = \{ n_1, n_2, \ldots \} \subseteq \mathbb{N} \) be the set of indices for which \( a_j > 0 \), define \( n_0 = 0 \) and for any \( n_i \in P \), let \( T_{n_i} = \mathbb{N} \cap (n_{i-1}, n_i) \). Then for any \( i \geq 1 \), we have \( b_{n_i} \sum_{j \in T_{n_i}} a_j \geq \sum_{j \in T_{n_i}} b_j a_j \). Let \( n_k \in P \) be the first index such that \( \sum_{j=1}^{n_k} a_j b_j > 0 \). If \( k = 1 \), the proof is straightforward. If \( k > 1 \), then

\[
\sum_{j=1}^{n_k} a_j = \sum_{i=1}^{k} \sum_{j \in T_{n_i}} a_j \geq \sum_{i=1}^{k-1} \sum_{j \in T_{n_i}} b_j a_j + \frac{1}{b_{n_{k-1}}} \sum_{j \in T_{n_k}} b_j a_j \\
\geq \frac{1}{b_{n_{k-1}}} \sum_{i=1}^{k} \sum_{j \in T_{n_i}} b_j a_j = \frac{1}{b_{n_{k-1}}} \sum_{j=1}^{n_k} b_j a_j > 0.
\]

\( \blacksquare \)

Proof of Theorem 1. Suppose \( m_{P_2}(y) - m_{P_1}(y) = \delta > 0 \). There exists \( r > 0 \) big enough such that \( \sup_{x \in B_r(y)} \psi(\|x - y\|_\mathbb{H}) \leq \delta / 2 \), in which \( B_r(y) = \{ x \in \mathbb{H} : \|x - y\|_\mathbb{H} < r \} \), and \( B_r(y)^c \) denotes the complement of \( B_r(y) \). We have

\[
0 < \delta = \int_{\mathbb{H}} \psi(\|x - y\|_\mathbb{H})(P_2 - P_1)(dx) \\
= \int_{B_r(y)} \psi(\|x - y\|_\mathbb{H})(P_2 - P_1)(dx) + \int_{B_r(y)^c} \psi(\|x - y\|_\mathbb{H})(P_2 - P_1)(dx) \\
\leq \int_{B_r(y)} \psi(\|x - y\|_\mathbb{H})(P_2 - P_1)(dx) + \delta / 2,
\]

and thus

\[
\int_{B_r(y)} \psi(\|x - y\|_\mathbb{H})(P_2 - P_1)(dx) \geq \frac{\delta}{2} > 0. \tag{A1}
\]

Let define

- \( r_{i,L} = (1 - L^i)r; i \geq 1, L \in (0, 1) \)
- \( b_{i,L} = B_{r_{i,L}}(y); i \geq 1 \)
- \( B_{L,L} = B_{1,L}, B_{i,L} = B_{i,L} \setminus B_{(i-1),L}; i \geq 2 \)

Thus from (A1) we have

\[
0 < \frac{\delta}{2} \leq \sum_{i \geq 1} \int_{B_{i,L}} \psi(\|x - y\|_\mathbb{H})(P_2 - P_1)(dx) \\
\leq \sum_{i \geq 1} m_{i,L}(P_2 - P_1)(B_{i,L}) + \sum_{i \geq 1} y_{i,L} P_2(B_{i,L}) \\
\leq \sum_{i \geq 1} m_{i,L}(P_2 - P_1)(B_{i,L}) + \sup_{i \geq 1} y_{i,L} P_2(B_r(y)),
\]

where \( m_{i,L} = \inf_{x \in B_{i,L}} \psi(\|x - y\|_\mathbb{H}) \) and \( M_{i,L} = \sup_{x \in B_{i,L}} \psi(\|x - y\|_\mathbb{H}) \) and \( y_{i,L} = M_{i,L} - m_{i,L} \). Because \( \psi \) is a bounded nonnegative continuous and strictly decreasing function,
we can choose $L \in (0, 1)$ such that $\sup_{\ell \geq 1} \gamma_{\ell L} \| \mathbb{P}_2(B_r(y)) \| < \frac{\delta}{4}$ or $\sup_{\ell \geq 1} \gamma_{\ell L} < \frac{\delta}{4P_2(B(y))}$ and thus $\sum_{\ell \geq 1} m_{\ell L} (\mathbb{P}_2 - \mathbb{P}_1)(B'_{\ell L}) > 0$. By employing Lemma 1, there exists $N < \infty$ such that $\sum_{\ell = 1}^N (\mathbb{P}_2 - \mathbb{P}_1) (B'_{\ell L}) > 0$, which immediately follows that $(\mathbb{P}_2 - \mathbb{P}_1)(B_r(y)) > 0$, where $r^* = (1 - L^N) r$.

**Proof of Corollary 1.** Let assume $\int_{\mathbb{H}} \psi(\|x - y\|_\mathbb{H}) \mathbb{P}(dx) > \int_{\mathbb{H}} \psi(\|x - y\|_\mathbb{H}) \mathbb{P}(dx)$ and let $\mathbb{P}_{-a}$ be the push forward of $\mathbb{P}$ by the map $x \mapsto x + a$, which translates $x$ to $x + a$, then we have $\int_{\mathbb{H}} \psi(\|x\|_\mathbb{H}) \mathbb{P}_{-y_2}(dx) > \int_{\mathbb{H}} \psi(\|x\|_\mathbb{H}) \mathbb{P}_{-y_1}(dx)$, and thus by the same argument as stated in the proof of Theorem 1, there exists $r > 0$ big enough such that $(\mathbb{P}_{-y_1} - \mathbb{P}_{-y_1})(B_r(0)) > 0$ and consequently $\mathbb{P}(B_r(y_2)) > \mathbb{P}(B_r(y_1))$.

**A.2 Proof of Theorem 2**

The existence proof of a continuous characteristic kernel for $\ell_2$ relies on the following theorem by Steinwart and Ziegel (2019). We also need Lemma 2 to complete the proof.

**Theorem 4** ((Steinwart & Ziegel, 2019, Thm. 3.14)). For a compact topological Hausdorff space $(X, \tau)$, the following statements are equivalent:

1. There exists a universal kernel $k$ on $X$.
2. There exists a continuous characteristic kernel $k$ on $X$.
3. $X$ is metrizable, that is, there exists a metric generating the topology $\tau$.

**Lemma 2.** Let $\mathbb{R}$ be the extended real line, and $\mathbb{R}^\infty$ and $\mathbb{R}^\infty$ be the countable products of $\mathbb{R}$ and $\mathbb{R}$ respectively, which are equipped with the product topologies, and let $\ell_2 \subset \mathbb{R}^\infty$ be the Hilbert space of square summable sequences. If the function $f : \mathbb{R}^\infty \rightarrow \mathbb{R}$ is continuous, then $f|_{\ell_2}$, which is restriction of $f$ to $\ell_2$, is continuous with respect to the norm of $\ell_2$.

**Proof.** Assume that $\varphi : \mathbb{R} \rightarrow [-1, 1]$ is defined as follows:

$$\varphi(x) = \frac{x}{1 + |x|}, \quad \forall x \in \mathbb{R} \quad \text{and} \quad \varphi(-\infty) = -1, \quad \varphi(+\infty) = 1.$$  

It is clear that $\varphi$ is a homeomorphic and order-preserving. Consider

$$\rho(x, y) := |\varphi(x) - \varphi(y)|, \quad \text{for all} \ x, y \in \mathbb{R}.$$  

Then $\rho$ is a metric and the topology induced by this metric is equivalent to the order topology. On the other hand, it is well known that the product topology on $\mathbb{R}^\infty$ can be generated by the following metric (Conway, 2014, Thm. 2.6.6),

$$d(x, y) := \sum_{k=1}^\infty \frac{\rho(x_k, y_k)}{2^k (1 + \rho(x_k, y_k))}, \quad \forall x = (x_k), y = (y_k) \in \mathbb{R}^\infty.$$  

Now, we show that the topology induced by the metric $d$ on $\ell_2$ is weaker than the norm topology. For this purpose, let $x_n = (x^k_n), x = (x^k) \in \ell_2$ and $|x_n - x| \rightarrow 0$. Therefore, $|x^k_n - x^k| \rightarrow 0$ for all $k \in \mathbb{N}$. Because $\varphi$ is a homeomorphic, $\rho(x^k_n, x^k) \rightarrow 0$ for any $k \in \mathbb{N}$ and hence $d(x_n, x) \rightarrow 0$. Thus, if $f$ is continuous with respect to the product topology, then $f|_{\ell_2}$ is continuous with respect to the norm topology.
Proof of Theorem 2. Without loss of generality assume $H$ to be the space of square summable sequences $\ell_2$, which is a subset of $\mathbb{R}^\infty$, and let $B(\ell_2)$ be Borel sigma-algebra generated by the open sets of $\ell_2$. $\mathbb{R}$ is a one-dimensional locally compact Hausdorff space, and the extended real line $\mathbb{R}$ equipped with order topology is a metrizable Hausdorff and compact topological space. Equip both $\mathbb{R}^\infty$ and $\mathbb{R}$ with the product topologies and let $B(\mathbb{R}^\infty)$ and $B\left(\mathbb{R}\right)$ be the Borel sigma-algebra generated by the open sets of these topologies. Consider that $B(\ell_2) = \{ A \cap \ell_2 : A \in B(\mathbb{R}^\infty) \}$, and we have $B(\ell_2) \subseteq B(\mathbb{R}^\infty) \subseteq B\left(\mathbb{R}\right)$. Note that $B(\mathbb{R}^\infty) \subseteq B\left(\mathbb{R}\right)$ since we equipped extended real line $\mathbb{R}$ with order topology, which includes the bases for the natural topology of $\mathbb{R}$. Let $\iota : \ell_2 \rightarrow \mathbb{R}^\infty$ be the usual inclusion map, then for every $A \in B\left(\mathbb{R}\right)$ we have $\iota^{-1}(A) = A \cap \ell_2 \in B(\ell_2)$, so $\iota$ is a $B(\ell_2) - B\left(\mathbb{R}\right)$ measurable map and hence every $\ell_2$-valued random element is an $\mathbb{R}^\infty$-valued random element and thus the space of Borel probability measures on $(\ell_2, B(\ell_2))$ is a subset of the space of Borel probability measures on $(\mathbb{R}^\infty, B(\mathbb{R}^\infty))$. $\mathbb{R}^\infty$ itself is a metrizable compact topological Hausdorff space, thus by invoking Theorem 4, there exists a continuous characteristic kernel $k(\cdot, \cdot)$ on $\mathbb{R}^\infty$, which by employing Lemma 2, its restriction to $\ell_2$ is also continuous with respect to the norm of $\ell_2$. □

A.3 Proof of Theorem 3

Let $\ell_2$ be the space of square summable sequences with inner product $\langle \cdot, \cdot \rangle_2$ and norm $\| \cdot \|_2$, and let $\Lambda_\theta$ be the infinite-dimensional Gaussian measure on the measurable space $(\mathbb{R}^\infty, B(\mathbb{R}^\infty))$ defined as the product of countably many copies of normal distribution with mean zero and variance $\theta$. The dual space of $\mathbb{R}^\infty$ is $c_{00}$, so the characteristic function of the Gaussian measure, for any $x \in c_{00}$ equals to

$$\psi(x) : = \int_{\mathbb{R}^\infty} e^{-i(\omega, x)} \Lambda_\theta^1 (d\omega) = e^{-\frac{1}{\theta} \|x\|_2^2}. \tag{A2}$$

Let $\mathbb{P}$ and $\mathbb{Q}$ be two arbitrary probability measures over $c_{00}$ such that

$$\text{MMD}(H_k, \mathbb{P}, \mathbb{Q}) = 0,$$

then

$$0 = \text{MMD}(H_k, \mathbb{P}, \mathbb{Q})^2 = \int_{c_{00}} \int_{c_{00}} e^{-\frac{1}{\theta} \|x-y\|_2^2} (\mathbb{P} - \mathbb{Q})(dx)(\mathbb{P} - \mathbb{Q})(dy)$$

$$= \int_{c_{00}} \int_{c_{00}} \left( \int_{\mathbb{R}^\infty} e^{-i(\omega, x-y)} \Lambda_\theta^1 (d\omega) \right) (\mathbb{P} - \mathbb{Q})(dx)(\mathbb{P} - \mathbb{Q})(dy)$$

$$\overset{(a)}{=} \int_{\mathbb{R}^\infty} \int_{c_{00}} e^{-i(\omega, x-y)} (\mathbb{P} - \mathbb{Q})(dx)(\mathbb{P} - \mathbb{Q})(dy) \Lambda_\theta^1 (d\omega)$$
\[\begin{align*}
&= \int_{\mathbb{R}^\infty} \left( \int_{\mathbb{R}^\infty} e^{-i(\omega x)}(P - Q)(dx) \right) \int_{\mathbb{R}^\infty} e^{i(\omega y)}(P - Q)(dy) \Lambda_2^{\perp}(d\omega) \\
&= \int_{\mathbb{R}^\infty} \left( \phi_P(\omega) - \phi_Q(\omega) \right) \left( \phi_P(\omega) - \phi_Q(\omega) \right) \Lambda_2^{\perp}(d\omega) \\
&= \int_{\mathbb{R}^\infty} |\phi_P(\omega) - \phi_Q(\omega)|^2 \Lambda_2^{\perp}(d\omega).
\end{align*}\] (A3)

In the above equation, Fubini–Tonelli’s theorem is invoked in (a). Dual of $c_{00}$ with norm $\| \cdot \|_2$, is the space of square summable sequences $\ell_2$. So to show that $P = Q$, it is enough to show that $\phi_P = \phi_Q$ agrees on $\ell_2$. By (A3) and by definition of the integral and the fact that $\text{supp} \left( \Lambda_2^{\perp} \right) = \mathbb{R}^\infty$, for any open set $B$ we have,

\[
\inf_{\omega \in B} |\phi_P(\omega) - \phi_Q(\omega)|^2 = 0.
\]

Fix $\omega_0 \in c_{00}$, and for any $m \in \mathbb{N}$ define

\[
B_m := \left\{ x \in \mathbb{R}^m : \sum_{i=1}^m (x_i - \omega_{0i})^2 < \frac{1}{m^2} \right\} \times \mathbb{R}^\infty,
\]

which is an open set in $\mathbb{R}^\infty$. Thus for each $m \in \mathbb{N}$, we have,

\[
\inf_{\omega \in B_m} |\phi_P(\omega) - \phi_Q(\omega)|^2 = 0,
\]

and so there exists $\omega_m \in B_m$ such that, $|\phi_P(\omega_m) - \phi_Q(\omega_m)|^2 < \frac{1}{m}$. Confirm that the sequence $\omega_m$ converges in the metric of $\mathbb{R}^\infty$ to $\omega_0$, since

\[
d(\omega_m, \omega_0) = \sum_{k \geq 1} 2^{-k} \frac{1}{1 + |\omega_{mk} - \omega_{0k}|} \leq \sum_{k \geq 1} 2^{-k} \frac{1}{1 + 1/m} + \sum_{k > m} 2^{-k} \leq \frac{1}{m + 1} (1 - 2^{-m}) + 2^{-m} \to 0.
\]

So $\langle \omega_m, x \rangle_2 \to \langle \omega_0, x \rangle_2$ for any $x \in c_{00}$. By a simple application of Bounded Convergence Theorem, we have

\[
\lim_{m \to \infty} |\phi_P(\omega_m) - \phi_Q(\omega_m)|^2 = \lim_{m \to \infty} \int_{c_{00}} \left| \int_{c_{00}} e^{-i(\omega_m x)} P(dx) - \int_{c_{00}} e^{-i(\omega_m x)} Q(dx) \right|^2
\]

\[
= \int_{c_{00}} \left| \lim_{m \to \infty} e^{-i(\omega_m x)} P(dx) - \lim_{m \to \infty} e^{-i(\omega_m x)} Q(dx) \right|^2
\]

\[
= \left| \phi_P(\omega_0) - \phi_Q(\omega_0) \right|^2,
\]
and thus
\[ |\phi_P(\omega_0) - \phi_Q(\omega_0)|^2 = \lim_{m \to \infty} |\phi_P(\omega_m) - \phi_Q(\omega_m)|^2 \leq \lim_{m \to \infty} \frac{1}{m} = 0. \]

So \( \phi_P = \phi_Q \) on \( c_{00} \). The space \( c_{00} \) is dense in \( \ell_2 \), so \( \phi_P = \phi_Q \) agrees on \( \ell_2 \) and thus \( P = Q \).

**A.4 Proof of Proposition 1**

Before providing the proof we need some tools, which are provided in the upcoming theorems and lemmas. The next theorem is a generalization of Ky Fan’s inequality, which is useful to show convexity of the map \( A \mapsto |I + A|^{-1/2} \) on the convex set of positive trace-class operators that is crucial to prove Gaussian kernel is characteristic for the family of Gaussian distributions. The following theorem is a special case of Minh (2017, Thm. 1) when \( \mu = \gamma = 1 \).

**Theorem 5.** Let \( \mathbb{H} \) be an infinite-dimensional separable Hilbert space, and \( A, B \) two arbitrary positive trace-class operators, for \( 0 \leq \alpha \leq 1 \)
\[ |\alpha(I + A) + (1 - \alpha)(I + B)| \geq |I + A|^\alpha |I + B|^{1-\alpha}. \]

For \( 0 < \alpha < 1 \), equality occurs if and only if \( A = B \).

**Lemma 3.** Let \( \mathbb{H} \) be a separable Hilbert space, and let \( |\cdot| \) be the determinant of a non-negative symmetric operator on \( \mathbb{H} \). \( A \mapsto |I + A|^{-1/2} \) is a convex function over the convex set of positive trace-class operators on \( \mathbb{H} \), and for any two arbitrary positive trace-class operators \( A \) and \( B \),
\[ 2 \left| I + \frac{A + B}{2} \right|^{-1/2} \leq |I + A|^{-1/2} + |I + B|^{-1/2}, \]
and \( 2 \left| I + \frac{A + B}{2} \right|^{-1/2} = |I + A|^{-1/2} + |I + B|^{-1/2} \) if and only if \( A = B \).

**Proof.** By Theorem 5 we have
\[ \log |I + (\alpha A + (1 - \alpha)B)| \geq \alpha \log |I + A| + (1 - \alpha) \log |I + B|, \]
so \( A \mapsto \log |I + A| \) is a concave function on the convex set of positive trace-class operators, and thus \( A \mapsto \log |I + A|^{-1/2} \) is a convex function and also is \( A \mapsto |I + A|^{-1/2} \), since \( x \mapsto e^x \) is a nondecreasing convex function. Consequently
\[ \left| I + \left( \frac{1}{2} A + \frac{1}{2} B \right) \right|^{-1/2} \leq \frac{1}{2} |I + A|^{-1/2} + \frac{1}{2} |I + B|^{-1/2}, \]
and thus
\[ 2 \left| I + \frac{A + B}{2} \right|^{-1/2} \leq |I + A|^{-1/2} + |I + B|^{-1/2}. \]

By invoking Theorem (5), equality occurs if and only if \( A = B \).

**Lemma 4** ((Maniglia & Rhandi, 2004, Prop. 1.2.8)). Let \( \mathbb{H} \) be a separable Hilbert space and \( \mathcal{N}(\mu, C) \) be a Gaussian probability measure on \( \mathbb{H} \) with mean function \( \mu \) and
covariance operator $C$. For any $\sigma > 0$

\[
\int_{\mathcal{H}} e^{-\frac{1}{\sigma} \|x\|_{\mathcal{H}}^2} \mathcal{N}(\mu, C)(dx) = \left| I + 2\frac{1}{\sigma} C \right|^{-1/2} e^{-\frac{1}{\sigma} \left( (I + 2\frac{1}{\sigma} C)^{-1} \mu, \mu \right)_{\mathcal{H}}}. \]

Proof of Proposition 1. If $Y \sim \mathcal{N}(\mu, C)$ then by Lemma 4 we have

\[
m_{\mathcal{N}(\mu,C)}(x) = \int_{\mathcal{H}} e^{-\frac{1}{\sigma} \|y-x\|_{\mathcal{H}}^2} \mathcal{N}(\mu, C)(dy) = \int_{\mathcal{H}} e^{-\frac{1}{\sigma} \|z\|_{\mathcal{H}}^2} \mathcal{N}(x - \mu, C)(dz)
= \left| I + 2\frac{1}{\sigma} C \right|^{-1/2} e^{-\frac{1}{\sigma} \left( (I + 2\frac{1}{\sigma} C)^{-1} (x-\mu), (x-\mu) \right)_{\mathcal{H}}}. \]

Let $T_1 = I + 2\frac{1}{\sigma} C_1$, then

\[
\left\langle m_{\mathcal{N}(\mu,C_1)}, m_{\mathcal{N}(\mu,C_2)} \right\rangle_{\mathcal{H}_k} = \int_{\mathcal{H}} \int_{\mathcal{H}} e^{-\frac{1}{\sigma} \|x-y\|_{\mathcal{H}}^2} \mathcal{N}(\mu_1, C_1)(dx) \mathcal{N}(\mu_2, C_2)(dy)
= \left| T_1 \right|^{-1/2} e^{-\frac{1}{\sigma} \left( T_1^{-1/2} (y-\mu_1), (y-\mu_1) \right)_{\mathcal{H}}} \mathcal{N}(\mu_2, C_2)(dy)
= \left| T_1 \right|^{-1/2} \int_{\mathcal{H}} e^{-\frac{1}{\sigma} \left( T_1^{-1/2} (y-\mu_1), T_1^{-1/2} (y-\mu_1) \right)_{\mathcal{H}}} \mathcal{N}(\mu_2, C_2)(dy)
= \left| T_1 \right|^{-1/2} \int_{\mathcal{H}} e^{-\frac{1}{\sigma} \|z\|_{\mathcal{H}}^2} \mathcal{N}(T_1^{-1/2} (\mu_2 - \mu_1), T_1^{-1/2} C_2 T_1^{-1/2}) (dz)
= \left| T_1 \right|^{-1/2} \left| I + 2\frac{1}{\sigma} T_1^{-1/2} C_2 T_1^{-1/2} \right|^{-1/2} e^{-\frac{1}{\sigma} \left( (I + 2\frac{1}{\sigma} C_1 T_1^{-1/2} C_1 T_1^{-1/2})^{-1} T_1^{-1/2} (\mu_2 - \mu_1), T_1^{-1/2} (\mu_2 - \mu_1) \right)_{\mathcal{H}}}
= \left| T_1 \right|^{-1/2} \left| I + 2\frac{1}{\sigma} T_1^{-1} C_2 \right|^{-1/2} e^{-\frac{1}{\sigma} \left( T_1^{-1/2} (I + 2\frac{1}{\sigma} C_1 C_2 T_1^{-1/2})^{-1} T_1^{-1/2} (\mu_2 - \mu_1), (\mu_2 - \mu_1) \right)_{\mathcal{H}}}
= \left| I + 2\frac{1}{\sigma} (C_1 + C_2) \right|^{-1/2} e^{-\frac{1}{\sigma} \left( (I + 2\frac{1}{\sigma} (C_1 + C_2))^{-1} (\mu_2 - \mu_1), (\mu_2 - \mu_1) \right)_{\mathcal{H}}},
\]

and thus

\[
\left\| m_{\mathcal{N}(\mu,C_1)} - m_{\mathcal{N}(\mu,C_2)} \right\|_{\mathcal{H}_k}^2
= \left\| m_{\mathcal{N}(\mu,C_1)} \right\|_{\mathcal{H}_k}^2 + \left\| m_{\mathcal{N}(\mu,C_2)} \right\|_{\mathcal{H}_k}^2 - 2 \left\langle m_{\mathcal{N}(\mu,C_1)}, m_{\mathcal{N}(\mu,C_2)} \right\rangle_{\mathcal{H}_k}
= \left| I + 4\frac{1}{\sigma} C_1 \right|^{-1/2} + \left| I + 4\frac{1}{\sigma} C_2 \right|^{-1/2}
- 2 \left| I + 2\frac{1}{\sigma} (C_1 + C_2) \right|^{-1/2} e^{-\frac{1}{\sigma} \left( (I + 2\frac{1}{\sigma} (C_1 + C_2))^{-1} (\mu_2 - \mu_1), (\mu_2 - \mu_1) \right)_{\mathcal{H}}}. \]
By invoking Lemma 3 we have
\[
\left|I + 4 \frac{1}{\sigma} C_1\right|^{-1/2} + \left|I + 4 \frac{1}{\sigma} C_2\right|^{-1/2} \geq 2 \left|I + 2 \frac{1}{\sigma} (C_1 + C_2)\right|^{-1/2}.
\]

The equality occurs if and only if \(C_1 = C_2\). So \(\left\|m_{\mathcal{N}(\mu_1, C_1)} - m_{\mathcal{N}(\mu_2, C_2)}\right\|_{H_k}^2 = 0\) if and only if \(\mu_1 = \mu_2\) and \(C_1 = C_2\). Hence, Gaussian kernel is characteristic for the family of Gaussian distributions.

### A.5 Proof of Proposition 2

**Proof.** We first give a proof for the product-kernel. Let \(\mathbb{H}^n\) be the Hilbert space of finite sequences \(\{(x_i)_{i=1}^n : x_i \in \mathbb{H}\}\) equipped with the inner product \(\langle (x_i)_{i=1}^n, (y_i)_{i=1}^n \rangle_{\mathbb{H}^n} = \sum_{i=1}^n \langle x_i, y_i \rangle_{\mathbb{H}}\), and the norm \(\left\| (x_i)_{i=1}^n - (y_i)_{i=1}^n \right\|_{\mathbb{H}^n}^2 = \sum_{i=1}^n \left\| x_i - y_i \right\|_{\mathbb{H}}^2\). Then according to Proposition 1, product-kernel
\[
k((x_i)_{i=1}^n, (y_i)_{i=1}^n) = \prod_{i=1}^n k(x_i, y_i) = e^{-\frac{1}{\sigma} \sum_{i=1}^n \langle x_i, y_i \rangle_{\mathbb{H}}},
\]
is characteristic for the family of Gaussian probability measures over \(\mathbb{H}^n\).

For the Gaussian sum-kernel
\[
k((x_i)_{i=1}^n, (y_i)_{i=1}^n) = \sum_{i=1}^n k(x_i, y_i)
\]
let \(\otimes_{i=1}^n \mathcal{N}(\mu_{1i}, C_{1i})\) and \(\otimes_{i=1}^n \mathcal{N}(\mu_{2i}, C_{2i})\) be two probability measures on \(\mathbb{H}^n\) then similar to the same approach in the proof of Proposition 1,
\[
\left\|m_{\otimes_{i=1}^n \mathcal{N}(\mu_{1i}, C_{1i})} - m_{\otimes_{i=1}^n \mathcal{N}(\mu_{2i}, C_{2i})}\right\|_{H_k}^2 = \sum_{i=1}^n \left|I + 4 \frac{1}{\sigma} C_{1i}\right|^{-1/2} + \left|I + 4 \frac{1}{\sigma} C_{2i}\right|^{-1/2} - 2 \left|I + 2 \frac{1}{\sigma} (C_{1i} + C_{2i})\right|^{-1/2} e^{-\frac{1}{\sigma} \left\langle (I + 2 \frac{1}{\sigma} (C_{1i} + C_{2i})^{-1}(\mu_{1i} - \mu_{2i})), (\mu_{1i} - \mu_{2i})\right\rangle_{\mathbb{H}}},
\]
\[
\geq 2 \left|I + 2 \frac{1}{\sigma} (C_{1i} + C_{2i})\right|^{-1/2} \left(1 - e^{-\frac{1}{\sigma} \left\langle (I + 2 \frac{1}{\sigma} (C_{1i} + C_{2i})^{-1}(\mu_{1i} - \mu_{2i})), (\mu_{1i} - \mu_{2i})\right\rangle_{\mathbb{H}}}\right).
\]

By invoking Lemma 3, equality occurs if and only if \(C_{1i} = C_{2i}\) for each \(i = 1, \ldots, n\), and thus \(\left\|m_{\otimes_{i=1}^n \mathcal{N}(\mu_{1i}, C_{1i})} - m_{\otimes_{i=1}^n \mathcal{N}(\mu_{2i}, C_{2i})}\right\|_{H_k}^2 = 0\) if and only if \(\mu_{1i} = \mu_{2i}\) and \(C_{1i} = C_{2i}\) for each \(i = 1, \ldots, n\).

### APPENDIX B. FURTHER NUMERICAL STUDIES

In this section we provide further simulation studies for the tests proposed in Sections 4.2 and 4.3 to investigate their performance for nonsymmetric distributions.
To investigate the performance of devised MMD tests in the case of nonsymmetric distributions, we have replicated the simulation studies in Sections 4.2 and 4.3 by considering Skew-t distribution for the component scores, that is, \( z_{ijr} \overset{i.i.d.}{\sim} \text{Skew-t}(\alpha = \pm 6, \nu = 4)/\sqrt{2} \). We consider the following form of density function for the Skew-t distribution with slant parameter \( \alpha \) and degrees of freedom \( \nu \) which is given by (Azzalini, 2013, p. 102) and implemented in the package \texttt{sn} in R language:

\[
f(x; \alpha, \nu) = \frac{2}{\sqrt{\nu + 1}} T(\alpha x; \nu + 1) \sqrt{1 + \frac{1}{\nu + 1}}
\]

where \( T(x; \nu) \) and \( t(x; \nu) \) are the density function and cumulative density function of Student’s \( t \) distribution. The density function is positively skewed if \( \alpha > 0 \) and negatively skewed if \( \alpha < 0 \) and reduces to the usual Student’s \( t \) distribution when \( \alpha = 0 \). The sign of the slant parameter \( \alpha \) for each component is selected randomly to reflect both positively and negatively skewed distributions.

In Table B1 a replication of the simulation study in Section 4.2 is presented with 2000 iterations and considering Skew-t distribution for the component scores to demonstrate the performance of Functional ANOVA tests in the case of nonsymmetric distributions for component scores.

A replication of the simulation study in Section 4.3 is also presented in Table B2 with 2000 iterations and considering Skew-t distribution for the component scores to demonstrate the performance of MMD in testing for homogeneity of covariance operators in the case of nonsymmetric distributions for component scores. In all of the simulation studies the null distributions for the MMD-based statistics and the competitors are approximated by permutation method.

**Table B1** Type-I errors (in bold) and empirical powers of \( L^2, T_{\text{max}}, \text{GPF}, \) and \( F_{\text{max}} \) and \( \text{MMD}_0 \) for one-way ANOVA problem when \( z_{ijr} \overset{i.i.d.}{\sim} \text{st}(\alpha = \pm 6, \nu = 4)/\sqrt{2} \). All numbers are presented as percentages.

| \( \rho \times 100 \) | \( (n_i) = (20, 30, 30) \) | \( (n_i) = (70, 80, 100) \) |
|------------------|------------------|------------------|
| \( \delta \times 100 \) | \( 0 \) | \( 0.5 \) | \( 20.0 \) | \( 40.0 \) | \( 60.0 \) | \( 0 \) | \( 0.1 \) | \( 10.0 \) | \( 20.0 \) | \( 30.0 \) |
| \( \rho = 0.1 \) | \( F \) | 6.3 | 5.1 | 17.5 | 59.9 | 92.0 | 4.9 | 5.1 | 12.8 | 49.1 | 86.6 |
| | GPF | 6.2 | 5.1 | 17.5 | 59.9 | 92.0 | 5.0 | 5.1 | 12.8 | 49.1 | 86.5 |
| | \( F_{\text{max}} \) | 6.0 | 5.0 | 56.0 | 97.4 | 99.8 | 5.0 | 5.1 | 45.6 | 97.4 | 99.8 |
| | \( \text{MMD}(\sigma_0) \) | 6.4 | 5.1 | 28.1 | 96.4 | 100.0 | 5.0 | 4.9 | 21.2 | 96.4 | 100.0 |
| | \( \text{MMD}(\pi^2) \) | 5.7 | 99.2 | 100.0 | 100.0 | 100.0 | 5.9 | 27.2 | 100.0 | 100.0 | 100.0 |
| \( \rho = 0.5 \) | \( F \) | 5.1 | 4.2 | 17.2 | 60.1 | 90.6 | 5.1 | 5.1 | 14.8 | 47.1 | 86.1 |
| | GPF | 5.1 | 4.1 | 17.8 | 60.9 | 91.2 | 4.8 | 5.1 | 15.0 | 48.4 | 87.3 |
| | \( F_{\text{max}} \) | 5.1 | 4.0 | 34.1 | 94.0 | 99.9 | 4.8 | 5.2 | 30.9 | 91.0 | 99.9 |
| | \( \text{MMD}(\sigma_0) \) | 4.8 | 4.3 | 19.6 | 73.2 | 97.9 | 5.0 | 5.3 | 17.2 | 60.5 | 96.4 |
| | \( \text{MMD}(\pi^2) \) | 5.4 | 81.2 | 100.0 | 100.0 | 100.0 | 5.5 | 76.8 | 100.0 | 100.0 | 100.0 |
| \( \rho = 0.9 \) | \( F \) | 4.5 | 4.7 | 12.0 | 39.8 | 74.4 | 5.1 | 5.1 | 10.3 | 32.0 | 67.2 |
| | GPF | 4.8 | 4.3 | 11.8 | 40.1 | 74.8 | 5.1 | 5.1 | 10.6 | 32.6 | 68.0 |
| | \( F_{\text{max}} \) | 5.1 | 4.7 | 9.8 | 31.8 | 68.7 | 5.2 | 6.0 | 8.0 | 26.0 | 61.6 |
| | \( \text{MMD}(\sigma_0) \) | 4.7 | 4.4 | 11.6 | 39.1 | 74.2 | 4.9 | 5.1 | 10.0 | 31.1 | 65.9 |
| | \( \text{MMD}(\pi^2) \) | 5.3 | 96.7 | 99.7 | 99.9 | 100.0 | 5.8 | 98.6 | 100.0 | 100.0 | 100.0 |
TABLE B2  Type-I errors (in bold) and empirical powers of $L^2$, $T_{\text{max}}$, GPF, $F_{\text{max}}$ and MMD when $z_{ijr} \sim \text{i.i.d.} \ N(\alpha = \pm 6, \nu = 4)/\sqrt{2}$. All numbers are presented as percentages.

| $\rho$ | $\omega$ | $(n_i) = (20, 30, 30)$ | $(n_i) = (70, 80, 100)$ |
|--------|----------|------------------------|------------------------|
| 0.1    | $L^2$    | 5.3  11.2  30.5  47.5  59.9 | 4.4  12.5  34.5  59.2  74.4 |
|        | $T_{\text{max}}$ | 4.8  13.2  30.0  49.6  64.9 | 3.6  16.0  44.4  68.7  82.8 |
|        | $F_{\text{max}}$ | 5.2  13.0  42.3  64.6  75.6 | 4.1  16.0  50.5  80.5  92.3 |
|        | GPF      | 5.6  10.2  31.6  49.1  60.7 | 4.6  10.5  31.2  56.1  73.2 |
|        | MMD$(\sigma_0)$ | 4.9  7.0  27.5  53.8  74.9 | 4.3  6.7  15.8  45.6  75.3 |
|        | MMD$(\sigma^2)$ | 5.9  56.6  87.6  97.6  99.8 | 5.1  74.9  100.0  100.0  100.0 |
| 0.5    | $L^2$    | 4.0  23.5  58.8  77.0  81.8 | 5.1  25.4  75.2  91.9  96.4 |
|        | $T_{\text{max}}$ | 4.8  16.0  37.0  57.4  72.2 | 5.1  17.7  52.8  82.2  93.2 |
|        | $F_{\text{max}}$ | 5.3  24.6  70.0  90.4  92.4 | 5.8  24.8  84.4  98.7  99.7 |
|        | GPF      | 4.2  25.9  68.0  84.4  85.8 | 4.6  23.5  78.7  95.2  97.8 |
|        | MMD$(\sigma_0)$ | 5.0  11.5  31.9  64.1  83.8 | 5.1  8.0  28.6  63.5  87.3 |
|        | MMD$(\sigma^2)$ | 5.3  7.8  25.8  60.9  84.2 | 5.3  6.4  19.7  58.7  90.2 |
| 0.9    | $L^2$    | 5.0  15.1  59.2  87.4  92.0 | 5.1  22.6  80.5  96.8  98.6 |
|        | $T_{\text{max}}$ | 5.2  8.4  16.7  33.0  52.8 | 4.0  10.4  22.6  50.4  80.0 |
|        | $F_{\text{max}}$ | 5.3  11.1  33.3  66.5  84.7 | 5.3  10.9  46.3  90.0  99.4 |
|        | GPF      | 5.2  15.6  62.8  90.2  94.0 | 5.1  21.9  82.1  98.2  99.4 |
|        | MMD$(\sigma_0)$ | 4.7  7.6  19.0  45.9  71.0 | 4.8  7.4  16.1  42.5  75.3 |
|        | MMD$(\sigma^2)$ | 5.1  7.1  16.7  41.8  70.0 | 5.9  6.0  11.2  33.0  72.0 |