Signal Processing 233 (2025) 109959

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Distance of mean embedding for testing independence of functional data

Mirosław Krzyśko [a], Łukasz Smaga [b], Jędrzej Wydra [c,d,*]

[a] Inter-Faculty Department of Mathematics and Statistics, University of Kalisz, ul. Nowy Świat 4,, Kalisz, Poland
[b] Faculty of Mathematics and Computer Science, Adam Mickiewicz University, ul. Uniwersytetu Poznańskiego 4, Poznań, Poland
[c] Criminalistics Research Unit, Adam Mickiewicz University, al. Niepodległości 53, Poznań, Poland
[d] Centre for Advanced Technologies, Adam Mickiewicz University, ul. Uniwersytetu Poznańskiego 10, Poznań, Poland

ARTICLE INFO

Keywords:
Basis expansion
Covariance-based tests
Distance of mean embedding
Functional data analysis
Independence testing

ABSTRACT

We investigate independence testing for functional data, which may be either univariate or multivariate. Broadly speaking, our approach involves first reducing the dimensionality of the functional data using basis expansion and then applying the distance of mean embedding - a flexible measure of independence. We enhance this method for pairwise independence by incorporating marginal aggregation, as well as asymmetric and symmetric aggregation measures, to improve test performance and adapt it to mutual independence testing. Our methods are compared with tests based on distance covariance and the Hilbert–Schmidt independence criterion. To evaluate their effectiveness, we present simulation studies and two real data examples using air pollution and chemometric data sets. The new testing procedures demonstrate favorable finite-sample properties, effectively controlling the type I error rate and exhibiting competitive power, making them viable alternatives to covariance-based tests.

1. Introduction

Functional data analysis (FDA) is a branch of statistics that deals with the analysis of data that takes the form of functions or curves. In traditional statistics, data is typically represented as a set of discrete points, such as a collection of numbers or vectors. However, in many fields, such as engineering, chemometrics, and biology, data is often collected in the form of continuous functions or curves, such as time series data, spectral data, or image data. Some examples of functional data include stock prices, weather patterns, audio signals, image spectra, medical images, and growth curves. The goal of the FDA is to develop methods and techniques for analyzing and modeling these types of data, taking into account their functional nature. This involves developing new statistical tools and methods that can handle the unique characteristics of functional data, such as infinite dimensionality, non-linear relationships and patterns, and complex correlation structures. The methods consider signal processing [1], regression analysis [2,3], classification [4], clustering [5], dimension reduction [6], and verifying statistical hypotheses [7–9], among others. There is a large literature about the FDA. Additionally to the above references, we just note the monographs by Ferraty and Vieu [10], Horváth and Kokoszka [11], Ramsay and Silverman [12], and Zhang [13]. There are also popular software packages for FDA in Matlab (fda), Python (pyfda), and R (fda).

In this paper, we address the problem of independence testing for functional data, which presents significant challenges. This is a critical issue in statistics, with applications in various domains. Independence measures and tests are commonly employed in areas such as independent component analysis [14], which involves finding a linear transformation that minimizes statistical dependence between the components of a random vector. For functional data, Vidal et al. [15] proposed an independent component analysis based on the spectral decomposition of a kurtosis operator of a smoothed principal component expansion, and applied them to mapping adverse artifactual events caused by body movements in electroencephalographic (EEG) signals. Additionally, the independence measures and tests are critical for variable and feature selection [16], a key process in regression and classification tasks. The problem of independence testing is not unfamiliar in the world of functional data analysis. In the field of testing pairwise independence for functional data, Lai et al. [17] extended a projection covariance method introduced by Zhu et al. [18]. In subsequent years, Djonguet et al. [19] undertook further development and introduced solutions based on kernel embedding. Research conducted by Górecki et al. [20] and Górecki et al. [21] focused on testing pairwise independence in the context of multivariate functional data. They utilized various tests based on distance covariance and Hilbert–Schmidt independence criterion. Krzyśko et al. [22] extended previous findings by introducing unbiased covariance estimators and

* Corresponding author.
E-mail addresses: mkrzysko@amu.edu.pl (M. Krzyśko), ls@amu.edu.pl (Ł. Smaga), jedrzej.wydra@amu.edu.pl (J. Wydra).

https://doi.org/10.1016/j.sigpro.2025.109959
Received 10 November 2024; Received in revised form 19 February 2025; Accepted 20 February 2025
Available online 1 March 2025

applying tests proposed by Zhu et al. [23], which are based on the aggregation of marginal sample covariances. Such an approach is significant for a simple reason: the limitation of traditional covariance-based tests is capturing only linear dependence in the high-dimensional scenario. Krzyśko et al. [22] also considered the mutual independence testing by applying techniques introduced by Jin and Matteson [24] in the context of functional data. These techniques aggregate the results of the pairwise independence and are powerful test procedures for mutual independence. On the other hand, Djonguet et al. [19] proposed the tests based on generalized Hilbert–Schmidt independence criterion.

In this paper, we extend the results of Krzyśko et al. [22] by applying in particular the recent findings of Zhang et al. [25]. The latter study introduced the distance of mean embedding (DIME) for testing pairwise independence of random vectors. Similar to HSIC, their independence measure was based on the maximum mean discrepancy between the joint distribution and the product of marginal distributions. However, instead of just utilizing the unit ball in a reproducing kernel Hilbert space, they employed the Hilbert space of square-integrable functions with a given measure. This results in a more flexible method compared to HSIC, potentially leading to greater statistical power. Therefore, its application to functional data can address the challenges of independence testing. Specifically, our approach involves a two-step process: first, the dimensionality of the functional data – whether univariate or multivariate – is reduced from infinite dimension to finite one through basis expansion; second, the distance of mean embedding is employed as an effective measure of independence. To increase the effectiveness for pairwise independence, we incorporate marginal aggregation introduced by Zhu et al. [23], which can substantially improve test performance. Furthermore, we extend the methodology to mutual independence testing using symmetric and asymmetric aggregation measures by Jin and Matteson [24], marking an advance over prior work by Zhang et al. [25] that focused on pairwise settings. To benchmark the effectiveness of these methods, we conduct comprehensive simulation studies and apply our procedures to two real-world datasets: air pollution and chemometric measurements. Compared to established methods based on distance covariance and the Hilbert–Schmidt independence criterion, our tests exhibit strong finite-sample properties, effectively control type I error rates, and deliver better power. These results underscore the potential of our methods for independence testing in functional data analysis.

The remainder of this paper is organized as follows. In Section 2, we present the methodology. First, we discuss the dimensionality reduction of functional data using basis expansion. Then, we explain its relevance to independence testing for functional data. Finally, we introduce the application of recent methods for multivariate data. Section 3 presents simulation studies that evaluate the finite-sample properties of the proposed approaches and their competitors. In Section 4, we provide two real data examples, analyzing air pollution and chemometric data sets. The paper concludes in Section 5.

2. Methodology

In this section, we present the methodology for constructing the independence tests for functional data.

2.1. Basis expansion of functional data

For theoretical consideration, one has to represent observed data in some general space. In the case of multivariate functional data, it is usually the Hilbert space $\mathcal{L}_2^d(\mathcal{T})$ of d-dimensional vectors of square integrable functions defined on the interval $\mathcal{T} = [a, b]$, where $a, b \in \mathbb{R}$ and $a < b$. This space is endowed with the following inner product:

$$\langle \mathbf{f}, \mathbf{g} \rangle = \int_a^b \mathbf{f}^\top(t)\mathbf{g}(t)dt = \sum_{i=1}^d \int_a^b f_i(t)g_i(t)dt$$

for $\mathbf{f} = (f_1, \dots, f_d)^\top \in \mathcal{L}_2^d(\mathcal{T})$ and $\mathbf{g} = (g_1, \dots, g_d)^\top \in \mathcal{L}_2^d(\mathcal{T})$, which designates the following norm:

$$\|\mathbf{f}\|_2 = \sqrt{\langle \mathbf{f}, \mathbf{f} \rangle}.$$

The space $\mathcal{L}_2^d(\mathcal{T})$ gives many possibilities to construct methods for functional data analysis. In the present paper, we use the basis expansion, which reduces the infinite dimension of functional data into finite dimension of random vectors of its coefficients. Let $\mathbf{X} = (X_1, \dots, X_d)^\top$ be a random process in the space $\mathcal{L}_2^d(\mathcal{T})$, and let $\mathbf{X}(t) = (X_1(t), \dots, X_d(t))^\top$ be a random vector of values of process $\mathbf{X}$ at $t \in \mathcal{T}$. In this paper, we distinguish between these two objects and generally use the process $\mathbf{X}$ notation, while the pointwise notation $\mathbf{X}(t)$ is used when needed. If $\{\varphi_{im}\}_{m=1}^\infty$ is a basis in $\mathcal{L}_2^1(\mathcal{T})$, then the basis expansion of elements of $\mathbf{X}$ is as follows:

$$X_i(t) = \sum_{m=1}^\infty \alpha_{im}\varphi_{im}(t), \ t \in \mathcal{T}.$$

Here, the coefficients $\alpha_{im} = \langle X_i, \varphi_{im} \rangle$ are random variables, $i = 1, \dots, d$, $m = 1, \dots$, which contain the information about the elements of vector $\mathbf{X}$, since the basis functions φ_{im} are fixed. Due to the form of α_{im}, the proposed method can be interpreted as a projection; however, in the FDA field, the so-called projection tests are constructed differently [26]. Therefore, we employ a test based on basis expansion, following established practices in the literature [27]. Note that the measurability of α_{im} is preserved and inherited from the original X_i, provided X_i is measurable in the Hilbert space, which we assume. Let us additionally note that when the basis $\{\varphi_{im}\}_{m=1}^\infty$ is non-orthonormal, as for example the B-spline basis, the projection involves solving a linear system using the Gram matrix, which corresponds to the inner products between basis functions.

In practice, the process X_i is not fully observed, and hence we cannot find all the coefficients α_{im}, $m = 1, \dots$. Thus, we approximate the process X_i using a truncated expansion with a number (say M_i) of the first components, which should contain most of the information about the process X_i in practice:

$$X_i(t) \approx \sum_{m=1}^{M_i} \alpha_{im}\varphi_{im}(t), \ t \in \mathcal{T}.$$

In fact, we take here the representation of X_i in some finite-dimensional subspace of $\mathcal{L}_2^1(\mathcal{T})$. Such representation is usually appropriate and effective, but its quality depends of the choice of the basis and the number M_i. Regarding the choice of a basis, there are some recommendations for selecting it, e.g., for periodical data, the Fourier basis is usually suggested [11]. The smaller M_i, we obtain the smoother representation, loosing more information, but having smaller dimension of the space of coefficients. On the other hand, the larger M_i, the basis expansion is closer to real X_i, but the mentioned dimension is larger. One usually tries to use small number of coefficients, which can achieve a good approximation. If this is not the case, the number of coefficients can be increased.

It is usually convenient to use the matrix notation. Let $\boldsymbol{\Phi}(t) = \operatorname{diag}(\boldsymbol{\varphi}_1^\top(t), \dots, \boldsymbol{\varphi}_d^\top(t))$ and $\boldsymbol{\varphi}_i(t) = (\varphi_{i1}(t), \dots, \varphi_{iM_i}(t))^\top$. Then,

$$\mathbf{X}(t) \approx \boldsymbol{\Phi}(t)\boldsymbol{\alpha},$$

where $\boldsymbol{\alpha} = (\alpha_{11}, \dots, \alpha_{1M_1}, \dots, \alpha_{d1}, \dots, \alpha_{dM_d})^\top \in \mathbb{R}^M$, $t \in \mathcal{T}$, and $M = M_1 + \cdots + M_d$. In practice, the coefficients in $\boldsymbol{\alpha}$ can be estimated by the least squares method or the roughness penalty approach based on a given sample of functional data [6]. The numbers M_i can be chosen deterministically based on for example visual inspection, but one can also use information criteria as Bayesian information criterion [28].

The information contained in the vector $\boldsymbol{\alpha}$ can be used to construct methods for functional data. There are many such procedures, for example, the analysis of variance with repeated measures [7], the canonical correlation analysis [6], and the variable selection for functional regression models [2]. In the next section, we will show how it can be used for verifying independence of functional data.

2.2. Independence testing for functional data

Let us now focus on important problem of testing independence for functional data, which can be univariate ($d = 1$) or multivariate ($d > 1$). Instead of just one random process $\mathbf{X}$, which we considered in the previous section, we have $L \geq 2$ random processes $\mathbf{X}_l = (X_{l1}, \ldots, X_{ld_l})^\top \in \mathcal{L}_2^{d_l}(\mathcal{T}_l)$, where $l = 1, \ldots, L$, $\mathcal{T}_l = [a_l, b_l]$, $a_l, b_l \in \mathbb{R}$, and $a_l < b_l$. Of interest is to verify the following statistical hypotheses:

$$H_0 : \mathbf{X}_1, \ldots, \mathbf{X}_L \text{ are mutually independent}, \quad H_1 : \neg H_0. \tag{1}$$

In case of $L = 2$, we say about the pairwise independence, while for $L \geq 3$, we have the mutual independence.

To construct a test for these hypotheses, we will use the basis expansion presented in Section 2.1. Thus, let $\{\varphi_{lim}\}_{m=1}^{\infty}$ be a basis in $\mathcal{L}_2^1(\mathcal{T}_l)$, and

$$\mathbf{X}_l(t_l) \approx \boldsymbol{\Phi}_l(t_l)\boldsymbol{\alpha}_l, \tag{2}$$

where $\boldsymbol{\Phi}_l(t_l) = \operatorname{diag}(\boldsymbol{\varphi}_{l1}^\top(t_l), \ldots, \boldsymbol{\varphi}_{ld_l}^\top(t_l))$, $\boldsymbol{\varphi}_{li}(t_l) = (\varphi_{li1}(t_l), \ldots, \varphi_{liM_{li}}(t_l))^\top$,

$\boldsymbol{\alpha}_l = (\alpha_{l11}, \ldots, \alpha_{l1M_{l1}}, \ldots, \alpha_{ld_l1}, \ldots, \alpha_{ld_lM_{ld_l}})^\top \in \mathbb{R}^{M_l}$,

$M_l = M_{l1} + \cdots + M_{ld_l}$, $t_l \in \mathcal{T}_l$, $l = 1, \ldots, L$, and $i = 1, \ldots, d_l$. Using (2), Krzyśko et al. [22] considered the practical verifying the hypotheses in (1) by testing the following hypotheses:

$$H_0^v : \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_L \text{ are mutually independent}, \quad H_1^v : \neg H_0^v, \tag{3}$$

which is caused by the fact that the vectors $\boldsymbol{\alpha}_l$ are random vectors and the basis functions in the matrices $\boldsymbol{\Phi}_l$ are fixed. However, note that the hypotheses (1) and (3) are not equivalent.

In practice, we assume that $\mathbf{X}_{l1}, \ldots, \mathbf{X}_{ln} \in \mathcal{L}_2^{d_l}(\mathcal{T}_l)$ is a sample of independent realizations of random process $\mathbf{X}_l$ for $l = 1, \ldots, L$ and $n \in \mathbb{N}$, $n > 3$. As in (2), we expand the observations in a basis as follows:

$$\mathbf{X}_{li}(t_l) \approx \boldsymbol{\Phi}_l(t_l)\boldsymbol{\alpha}_{li},$$

where $\boldsymbol{\alpha}_{li} = (\alpha_{l11i}, \ldots, \alpha_{l1M_{l1}i}, \ldots, \alpha_{ld_l1i}, \ldots, \alpha_{ld_lM_{ld_l}i})^\top \in \mathbb{R}^{M_l}$, $l = 1, \ldots, L$, $i = 1, \ldots, n$, and $t_l \in \mathcal{T}_l$.

To construct a test based on the relation between hypotheses in (1) and (3), one can use various tests for independence of random vectors. Here, we would like to focus on the methods proposed in the recent paper by Krzyśko et al. [22]. In that paper, the tests based on the distance covariance, the Hilbert–Schmidt independence criterion (HSIC), and their marginal versions by Zhu et al. [23] were considered, given good finite sample results. For the case of marginal independence testing, these test statistics were used with the asymmetric and symmetric aggregation methods proposed by Jin and Matteson [24]. In the present paper, we propose a method based on the distance between joint and marginal distributions proposed by Zhang et al. [25], which is presented in the following section. It can be seen as the extension of HSIC, which has a potential of being more powerful.

2.3. $\mathcal{L}^2$ distance of mean embedding (DIME)

Let us describe the test for (3) based on the $\mathcal{L}^2$ distance of mean embedding (DIME), which was considered in Zhang et al. [25]. We will use the notation from the previous sections for the functional data framework. Let $L = 2$ and $\boldsymbol{\alpha}_l \in \mathcal{A}_l \subset \mathbb{R}^{M_l}$ for $l = 1, 2$, where $\mathcal{A}_1$ and $\mathcal{A}_2$ are open sets. We denote the joint distribution of $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ by $P_{\boldsymbol{\alpha}_1\boldsymbol{\alpha}_2}$ and the marginal distributions of $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ by $P_{\boldsymbol{\alpha}_1}$ and $P_{\boldsymbol{\alpha}_2}$ respectively. Under the null (respectively alternative) hypothesis in (3), we have that $P_{\boldsymbol{\alpha}_1\boldsymbol{\alpha}_2} = P_{\boldsymbol{\alpha}_1} \times P_{\boldsymbol{\alpha}_2}$ (respectively $P_{\boldsymbol{\alpha}_1\boldsymbol{\alpha}_2} \neq P_{\boldsymbol{\alpha}_1} \times P_{\boldsymbol{\alpha}_2}$). Moreover, the well known fact is that $P_{\boldsymbol{\alpha}_1\boldsymbol{\alpha}_2} = P_{\boldsymbol{\alpha}_1} \times P_{\boldsymbol{\alpha}_2}$ if and only if $E_{P_{\boldsymbol{\alpha}_1\boldsymbol{\alpha}_2}}[f(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)] = E_{P_{\boldsymbol{\alpha}_1} \times P_{\boldsymbol{\alpha}_2}}[f(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)]$ for every f belonging to the space $C(\mathcal{A}_1 \times \mathcal{A}_2)$ of continuous bounded functions on $\mathcal{A}_1 \times \mathcal{A}_2$. Thus, the discrepancy between these two expected values can be used to verify the independence of $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$. Hence, the maximum mean discrepancy (MMD) is defined as follows:

$$\gamma_{\mathcal{F}}(P_{\boldsymbol{\alpha}_1\boldsymbol{\alpha}_2}, P_{\boldsymbol{\alpha}_1} \times P_{\boldsymbol{\alpha}_2}) = \sup_{f \in \mathcal{F}} |E_{P_{\boldsymbol{\alpha}_1\boldsymbol{\alpha}_2}} f - E_{P_{\boldsymbol{\alpha}_1} \times P_{\boldsymbol{\alpha}_2}} f|,$$

where $\mathcal{F} = \{f | f : \mathcal{A}_1 \times \mathcal{A}_2 \to \mathbb{R}\}$ should be a rich function class to identify the equality of distributions $P_{\boldsymbol{\alpha}_1\boldsymbol{\alpha}_2}$ and $P_{\boldsymbol{\alpha}_1} \times P_{\boldsymbol{\alpha}_2}$ and provide useful finite sample estimates. Very popular example of such a class is a unit ball in an reproducing kernel Hilbert space (RKHS) $\mathcal{H}_k = \mathcal{H}_{k_1} \otimes \mathcal{H}_{k_2}$ ($\otimes$ is the tensor product of RKHSs $\mathcal{H}_{k_1}$ and $\mathcal{H}_{k_2}$) of real-valued functions defined on $\mathcal{A}_1 \times \mathcal{A}_2$, where k is a reproducing kernel given by

$$k((\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2), (\boldsymbol{\alpha}_1', \boldsymbol{\alpha}_2')) = k_1(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_1')k_2(\boldsymbol{\alpha}_2, \boldsymbol{\alpha}_2'),$$

where $k_l : \mathcal{A}_l \times \mathcal{A}_l \to \mathbb{R}$ are the positive definite kernels associated with RKHS $\mathcal{H}_{k_l}$, $l = 1, 2$, and $(\boldsymbol{\alpha}_1', \boldsymbol{\alpha}_2')$ is independent and identically distributed copy of $(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)$. Then, under the assumption that $E_{P_{\boldsymbol{\alpha}_1\boldsymbol{\alpha}_2}}[k((\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2), (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2))]^{1/2} < \infty$, we have

$$\gamma_{\mathcal{F}}(P_{\boldsymbol{\alpha}_1\boldsymbol{\alpha}_2}, P_{\boldsymbol{\alpha}_1} \times P_{\boldsymbol{\alpha}_2}) = \|E_{P_{\boldsymbol{\alpha}_1\boldsymbol{\alpha}_2}} k((\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2), \cdot) - E_{P_{\boldsymbol{\alpha}_1} \times P_{\boldsymbol{\alpha}_2}} k((\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2), \cdot)\|_{\mathcal{H}_k}^2.$$

If k is characteristic kernel, i.e., $\gamma_{\mathcal{F}}(P_{\boldsymbol{\alpha}_1\boldsymbol{\alpha}_2}, P_{\boldsymbol{\alpha}_1} \times P_{\boldsymbol{\alpha}_2}) = 0$ if and only if $P_{\boldsymbol{\alpha}_1\boldsymbol{\alpha}_2} = P_{\boldsymbol{\alpha}_1} \times P_{\boldsymbol{\alpha}_2}$, then $\gamma_{\mathcal{F}}(P_{\boldsymbol{\alpha}_1\boldsymbol{\alpha}_2}, P_{\boldsymbol{\alpha}_1} \times P_{\boldsymbol{\alpha}_2})$ is a metric on the set of all Borel probability measures on $\mathcal{A}_1 \times \mathcal{A}_2$. Some examples of characteristic kernels are the Gaussian and Laplacian kernels. This is a way to obtain HSIC.

The choice of $\mathcal{F}$, presented in the above paragraph, is not the only one. Naturally, the class $C(\mathcal{A}_1 \times \mathcal{A}_2)$ of continuous bounded functions in principle allows us to uniquely identify the distributions. However, it is not practical to work with such a rich function class in the finite sample setting. Theorem 18 in Gretton et al. [29] shows the characterization of appropriate class $\mathcal{F}$, i.e., $\mathcal{F}$ being the subset of some vector space of functions from $\mathcal{A}_1 \times \mathcal{A}_2$ to $\mathbb{R}$ for which $S[\mathcal{F}] \cap C(\mathcal{A}_1 \times \mathcal{A}_2)$ is dense in $C(\mathcal{A}_1 \times \mathcal{A}_2)$ with respect to the supremum norm, where $S[\mathcal{F}] = \{af : f \in \mathcal{F} \text{ and } a \in [0, \infty)\}$. The examples of $\mathcal{F}$, other than these presented above, can be found in Section 7.2 of Gretton et al. [29]. If $\mathcal{F}$ is the class of functions of bounded variation 1, then the MMD is equal to the Kolmogorov metric, used in the famous Kolmogorov–Smirnov test.

Under the assumption that $E_{P_{\boldsymbol{\alpha}_1\boldsymbol{\alpha}_2}}[k((\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2), (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2))] < \infty$, we have $\mathcal{H}_k \subset \mathcal{L}_2(\mathcal{A}_1 \times \mathcal{A}_2)$, which was used by Zhang et al. [25] to construct DIME, a new distance between $P_{\boldsymbol{\alpha}_1\boldsymbol{\alpha}_2}$ and $P_{\boldsymbol{\alpha}_1} \times P_{\boldsymbol{\alpha}_2}$. Namely, the $\mathcal{L}^2$ distance of mean embedding is defined as

$$\begin{aligned}\mathcal{D}_\kappa^2(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2) =& \| E_{P_{\boldsymbol{\alpha}_1\boldsymbol{\alpha}_2}}[k_1(\boldsymbol{\alpha}_1, \cdot)k_2(\boldsymbol{\alpha}_2, \cdot)] \\ &- E_{P_{\boldsymbol{\alpha}_1}} k_1(\boldsymbol{\alpha}_1, \cdot) \cdot E_{P_{\boldsymbol{\alpha}_2}} k_2(\boldsymbol{\alpha}_2, \cdot) \|_{\mathcal{L}_2(\mathcal{A}_1 \times \mathcal{A}_2, \kappa)}^2 \\ =& \int_{\mathcal{A}_1 \times \mathcal{A}_2} \Big(E_{P_{\boldsymbol{\alpha}_1\boldsymbol{\alpha}_2}}[k_1(\boldsymbol{\alpha}_1, \mathbf{x})k_2(\boldsymbol{\alpha}_2, \mathbf{y})] \\ &- E_{P_{\boldsymbol{\alpha}_1}} k_1(\boldsymbol{\alpha}_1, \mathbf{x}) \cdot E_{P_{\boldsymbol{\alpha}_2}} k_2(\boldsymbol{\alpha}_2, \mathbf{y}) \Big)^2 d\kappa,\end{aligned}$$

where κ is a measure defined on $\mathcal{A}_1 \times \mathcal{A}_2$. Zhang et al. [25] proved that $\mathcal{D}_\kappa^2(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)$ is a metric on the space of Borel probability measures on $\mathcal{A}_1 \times \mathcal{A}_2$ under their Assumption 1 as follows:

(1) the kernel k is characteristic,
(2) κ is a Borel probability measure, and its support set covers the entire space $\mathcal{A}_1 \times \mathcal{A}_2$,
(3) k_1 and k_2 are bounded, i.e., there exist positive constants c_1 and c_2 such that

$$\sup_{\boldsymbol{\alpha}_l, \boldsymbol{\alpha}_l' \in \mathcal{A}_l} |k_l(\boldsymbol{\alpha}_l, \boldsymbol{\alpha}_l')| < c_l, \quad l = 1, 2.$$

The Assumption (1) ensures that the mean embedding map $E_P[k_1(\boldsymbol{\alpha}_1, \cdot)k_2(\boldsymbol{\alpha}_2, \cdot)]$, where P is a distribution on $\mathcal{A}_1 \times \mathcal{A}_2$, can identify the distributions. When k_1 and k_2 are the characteristic kernels, then k is also the characteristic kernel. It is worth noting that the characteristic kernel is less stringent assumption than the universal kernel that is required for HSIC. The simple example of κ satisfying Assumption (2)

is the standard normal distribution, which will be used in the practical part. The Assumption (3) is standard, and it is satisfied for the common kernels as the Gaussian and Laplacian kernels. Under Assumption (3), the kernel k is also bounded.

Therefore, under Assumption 1, $\mathcal{D}^2_\kappa(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2) = 0$ if and only if $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ are independent, and it is independence measure, which is flexible in choosing the characteristic kernels k_1 and k_2 as well as the metric κ. It is important that in DIME, one can choose the metric κ in contrast to HSIC.

In practice, $\mathcal{D}^2_\kappa(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)$ is not known as in particular, we do not assume any specific distribution of the data. Thus, we have to estimate it. Fortunately, it can be estimated using the methods of moments, since $\mathcal{D}^2_\kappa(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)$ can be expressed as follows:

$$\begin{aligned}\mathcal{D}^2_\kappa(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2) =& E_{\boldsymbol{\alpha}_1\boldsymbol{\alpha}_2} E_{\boldsymbol{\alpha}'_1\boldsymbol{\alpha}'_2} \Gamma(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}'_1, \boldsymbol{\alpha}_2, \boldsymbol{\alpha}'_2) \\ &- 2E_{\boldsymbol{\alpha}_1\boldsymbol{\alpha}_2} E_{\boldsymbol{\alpha}'_1} E_{\boldsymbol{\alpha}'_2} \Gamma(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}'_1, \boldsymbol{\alpha}_2, \boldsymbol{\alpha}'_2) \\ &+ E_{\boldsymbol{\alpha}_1} E_{\boldsymbol{\alpha}'_1} E_{\boldsymbol{\alpha}_2} E_{\boldsymbol{\alpha}'_2} \Gamma(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}'_1, \boldsymbol{\alpha}_2, \boldsymbol{\alpha}'_2),\end{aligned}$$

where $\Gamma(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}'_1, \boldsymbol{\alpha}_2, \boldsymbol{\alpha}'_2) = E_{\mathbf{XY}}[k_1(\boldsymbol{\alpha}_1, \mathbf{X}) k_1(\boldsymbol{\alpha}'_1, \mathbf{X}) k_2(\boldsymbol{\alpha}_2, \mathbf{Y}) k_2(\boldsymbol{\alpha}'_2, \mathbf{Y})]$ and $\mathbf{X}$ and $\mathbf{Y}$ are the random vectors defined on the same spaces as $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ respectively. Then, the estimator of $\mathcal{D}^2_\kappa(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)$ is given by the following formula:

$$\begin{aligned}\mathcal{D}^2_{n,\kappa}(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2) =& \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \Gamma(\boldsymbol{\alpha}_{1i}, \boldsymbol{\alpha}_{1j}, \boldsymbol{\alpha}_{2i}, \boldsymbol{\alpha}_{2j}) \\ &- \frac{2}{n^3} \sum_{i=1}^n \sum_{j=1}^n \sum_{s=1}^n \Gamma(\boldsymbol{\alpha}_{1i}, \boldsymbol{\alpha}_{1j}, \boldsymbol{\alpha}_{2i}, \boldsymbol{\alpha}_{2s}) \\ &+ \frac{1}{n^4} \sum_{i=1}^n \sum_{j=1}^n \sum_{s=1}^n \sum_{t=1}^n \Gamma(\boldsymbol{\alpha}_{1i}, \boldsymbol{\alpha}_{1j}, \boldsymbol{\alpha}_{2s}, \boldsymbol{\alpha}_{2t}).\end{aligned}$$

Let us consider the final computational detail for practical use (see also Section 3 in Zhang et al. [25]). When the vectors $\mathbf{X}$ and $\mathbf{Y}$ are independent, we have

$$\widetilde{\mathcal{D}}^2_{n,\kappa}(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2) = \frac{1}{n^2} \operatorname{tr}(\mathbf{K}_1 \mathbf{H} \mathbf{K}_2 \mathbf{H}), \tag{4}$$

where $\mathbf{K}_l$ are $n \times n$ matrices defined by their elements $(\mathbf{K}_l)_{ij} = \Gamma_l(\boldsymbol{\alpha}_{li}, \boldsymbol{\alpha}_{lj})$, $\Gamma_l(\boldsymbol{\alpha}_l, \boldsymbol{\alpha}'_l) = E_{\mathbf{Z}}[k_l(\boldsymbol{\alpha}_l, \mathbf{Z}) k_l(\boldsymbol{\alpha}'_l, \mathbf{Z})]$, $\mathbf{Z} = \mathbf{X}, \mathbf{Y}$ for $l = 1, 2$ respectively, and $\mathbf{H} = \mathbf{I}_n - (1/n)\mathbf{1}_n \mathbf{1}_n^\top$ is a centering matrix. To specify the form of matrices $\mathbf{K}_l$, Zhang et al. [25] took the Gaussian kernels k_l with parameters σ_l and the independent standard normal distributions for $\mathbf{X}$ and $\mathbf{Y}$. (We refer to the beginning of Section 3 for the Gaussian kernel and the practical choice of the parameters σ_l.) The resulting estimator was called by NDIME (N - normal), and it is based on the following $\mathbf{K}_l$ matrices:

$$(\mathbf{K}_l^N)_{ij} = \exp\left(- \frac{\|\boldsymbol{\alpha}_{li} - \boldsymbol{\alpha}_{lj}\|^2 + \frac{2\sigma_l^2}{1+\sigma_l^2} \boldsymbol{\alpha}_{li}^\top \boldsymbol{\alpha}_{lj}}{2\frac{\sigma_l^2(2+\sigma_l^2)}{1+\sigma_l^2}} \right),$$

where $l = 1, 2$, $i, j = 1, \ldots, n$, and $\|\cdot\|$ is the Euclidean norm in $\mathbb{R}^{M_l}$. Therefore, the formula for the estimator of the $\mathcal{L}^2$ distance of mean embedding given in (4) can be easily computed in practice. We provide the code implementing the methods considered in this paper in the R programming language [30] in our GitHub repository https://github.com/ls-git-17/functional_independence_testing_by_ndime. The HSIC has very similar form, but $\mathbf{K}_l = k_l(\boldsymbol{\alpha}_{li}, \boldsymbol{\alpha}_{lj})$ are the kernel matrices. Thus, NDIME reduces to HSIC, when the observations are orthogonal (i.e., $\boldsymbol{\alpha}_{li}^\top \boldsymbol{\alpha}_{lj} = 0$). To approximate the null distribution of $\widetilde{\mathcal{D}}^2_{n,\kappa}(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)$ and construct a test based on it as a test statistic, the permutation method was used (see for example Section 5 in [22]).

In addition to the use of permutation NDIME test, we also consider some modification and extension of it. First, we consider the marginal version of NDIME, denoted mNDIME, which is an application of aggregation method proposed by Zhu et al. [23]. More precisely, in the statistic proposed by Zhu et al. [23], we use NDIME instead of distance covariance or HSIC, i.e.,

$$m\text{NDIME} = \sqrt{\binom{n}{2}} \sum_{m_1=1}^{M_1} \sum_{m_2=1}^{M_2} \mathcal{D}^2_{n,\kappa}(\alpha_{1m_1}, \alpha_{2m_2}),$$

where α_{lm_l}, $l = 1, 2$, correspond to m_lth component of $\boldsymbol{\alpha}_l$. The values of $\mathcal{D}^2_{n,\kappa}(\alpha_{1m_1}, \alpha_{2m_2})$ are thus based on the samples $\alpha_{lm_l1}, \ldots, \alpha_{lm_ln}$, $l = 1, 2$, where $\boldsymbol{\alpha}_{li} = (\alpha_{l1i}, \ldots, \alpha_{lM_li})^\top$, $i = 1, \ldots, n$. As we will see in the next sections, this may help in improving the finite sample properties of NDIME permutation test. Second, we consider the case of mutual independence. To extend the tests based on NDIME and mNDIME for $L > 2$, we combine these tests with the R and S methods by Jin and Matteson [24], which also gives good results.

3. Simulation studies

In this section, we study the finite sample properties of the test procedures considered in Section 2. For pairwise independence ($L = 2$), we considered eight permutations tests based on:

1. distance covariance and its marginal version: dCov^2_n, $m\mathrm{dCov}^2_n$,
2. Hilbert–Schmidt covariance with Gaussian kernel $K(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x} - \mathbf{y}\|^2/(2\gamma^2))$ and its marginal version: ghCov^2_n, $m\mathrm{ghCov}^2_n$,
3. Hilbert–Schmidt covariance with Laplacian kernel $K(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x} - \mathbf{y}\|/\gamma)$ and its marginal version: lhCov^2_n, $m\mathrm{lhCov}^2_n$,
4. $\mathcal{L}^2$ distance of mean embedding with normal distribution and its marginal version: NDIME, mNDIME.

Here, $\|\cdot\|$ is the Euclidean norm in $\mathbb{R}^p$ for appropriate dimension p. The tests 1–3 were considered by Krzyśko et al. [22]. Following Zhu et al. [23], we set 200 permuted samples used in the permutation tests. For tests in 2–4, by Gretton et al. [31], the bandwidth parameter γ and the parameters σ_1 and σ_2 were taken separately for each sample as the median distance between points in a sample, i.e., $\gamma_l = median\{\|\boldsymbol{\alpha}_{lp} - \boldsymbol{\alpha}_{lq}\| : p, q = 1, \ldots, n; p \neq q\}$. In the case of mutual independence, we took $L = 3$ and had sixteen tests, which were the above tests combined with the R and S methods of Jin and Matteson [24]. We used the B-spline basis as the basis representation of the functional data with all numbers of basis functions M_{ij} set equal to five for simplicity. The coefficients of the basis representation were estimated by the least squares method. The implementation was prepared in the R programming language [30]. The code is available from the authors upon request.

We were interested in studying the type I error level and power of the tests, which are defined as probabilities that the test rejects the null hypothesis, when it is true and false respectively. Their empirical versions were estimated as the proportions of rejections of the null hypothesis on the basis of 500 simulation replications, when the data were generated under the null and alternative hypothesis respectively. The empirical type I error level is usually called the empirical size. The test controls the type I error level, when its empirical size is close to the significance level, which we set at $\alpha = 5\%$. On the other hand, the larger the power of the test, the better it is.

3.1. Simulation setup

To generate the artificial functional data, we used the basis representation (2) of them in the following way: Let $d_1 = \cdots = d_L = 3$. For each functional variable, the functional data were generated by their values in fifty design time points $t_1 = 0, t_2 = 1/49, t_3 = 2/49, \ldots, t_{50} = 1$ in the interval $[0, 1]$. For $i = 1, \ldots, n$ and $j = 1, \ldots, 50$, we have

$$\begin{bmatrix} \mathbf{X}_{1i}(t_j) \\ \mathbf{X}_{2i}(t_j) \\ \vdots \\ \mathbf{X}_{Li}(t_j) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Phi}_1(t_j) & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Phi}_2(t_j) & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \boldsymbol{\Phi}_L(t_j) \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha}_{1i} \\ \boldsymbol{\alpha}_{2i} \\ \vdots \\ \boldsymbol{\alpha}_{Li} \end{bmatrix} + \begin{bmatrix} e_{ij,1} \\ e_{ij,2} \\ \vdots \\ e_{ij,3L} \end{bmatrix}.$$

Here, the matrices $\boldsymbol{\Phi}_l$, defined in (2), contained the Fourier basis functions only, $l = 1, \ldots, L$. The random vectors $(\boldsymbol{\alpha}_{1i}^\top, \ldots, \boldsymbol{\alpha}_{Li}^\top)^\top$ were generated in various ways, as we describe below. Finally, the measurement errors $e_{ij,m}$ were independent random variables with the normal distribution $N(0, 0.025a_{im})$, where a_{im} was the range of the mth row of the following matrix:

$$\begin{bmatrix} \boldsymbol{\Phi}_1(t_1)\boldsymbol{\alpha}_{1i} & \ldots & \boldsymbol{\Phi}_1(t_{50})\boldsymbol{\alpha}_{1i} \\ \boldsymbol{\Phi}_2(t_1)\boldsymbol{\alpha}_{2i} & \ldots & \boldsymbol{\Phi}_2(t_{50})\boldsymbol{\alpha}_{2i} \\ \vdots & \ddots & \vdots \\ \boldsymbol{\Phi}_L(t_1)\boldsymbol{\alpha}_{Li} & \ldots & \boldsymbol{\Phi}_L(t_{50})\boldsymbol{\alpha}_{Li} \end{bmatrix}.$$

Set $n = 20$. We considered the following two scenarios for pairwise and mutual independence. Let $\boldsymbol{\alpha}_{li} = (\alpha_{li,1}, \ldots, \alpha_{li,15})^\top$ for $l = 1, \ldots, L$ and $i = 1, \ldots, n$.

Scenario 1. We considered pairwise independence ($L = 2$). We generated the random vectors $(\boldsymbol{\alpha}_{1i}^\top, \ldots, \boldsymbol{\alpha}_{Li}^\top)^\top$ in the following two settings under the null and alternative hypothesis respectively. The ideas of generating these vectors come from various earlier papers as Zhu et al. [23] and Zhang et al. [25].

Setting 1. In this setting, the null hypothesis is set to be true to investigate the type I error level of the tests. We generated i.i.d. vectors $\boldsymbol{\alpha}_{11}, \ldots, \boldsymbol{\alpha}_{1n}$ as independent of i.i.d. vectors $\boldsymbol{\alpha}_{21}, \ldots, \boldsymbol{\alpha}_{2n}$ in the following cases ($l = 1, 2$):

Case 1. $\boldsymbol{\alpha}_{li} \sim N_{15}(\mathbf{0}_{15}, \mathbf{I}_{15})$,
Case 2. $\boldsymbol{\alpha}_{1i} \sim AR_{0.5}(1)$ and $\boldsymbol{\alpha}_{2i} \sim AR_{-0.5}(1)$, where $AR_\rho(1)$ denotes the Gaussian autoregressive model of order 1 with parameter ρ,
Case 3. $\boldsymbol{\alpha}_{li} \sim N_{15}(\mathbf{0}_{15}, \boldsymbol{\Sigma}_{15})$, where $\boldsymbol{\Sigma}_a = (0.7^{|p-q|})_{p,q=1}^a$,
Case 4. $\boldsymbol{\alpha}_{li} \sim T_5(\mathbf{I}_{15})$, where $T_k(\boldsymbol{\Sigma})$ denotes the multivariate t-Student distribution with k degrees of freedom and covariance matrix $\boldsymbol{\Sigma}$,
Case 5. $\boldsymbol{\alpha}_{li} \sim T_5(\boldsymbol{\Sigma}_{15})$, where $\boldsymbol{\Sigma}_{15}$ is the same as in Case 3,
Case 6. the elements of $\boldsymbol{\alpha}_{li}$ were i.i.d. variables of uniform distribution $U(-1, 1)$,
Case 7. $\boldsymbol{\alpha}_{li} = \boldsymbol{\Sigma}_{15}^{1/2}\boldsymbol{a}_{li}$, where $\boldsymbol{a}_{li}$ are generated in the same way as in Case 6.

Setting 2. Here, we study the power of the tests, thus the data will be generated in the alternative hypothesis.

Case 1. $\boldsymbol{\alpha}_{1i} \sim N_{15}(\mathbf{0}_{15}, \mathbf{I}_{15})$ and $\alpha_{2i,m} = \alpha_{1i,m}^2$ for $m = 1, \ldots, 15$,
Case 2. $\boldsymbol{\alpha}_{1i} \sim N_{15}(\mathbf{0}_{15}, \boldsymbol{\Sigma}_{15})$ and $\alpha_{2i,m} = \alpha_{1i,m}^2$ for $m = 1, \ldots, 15$, $\boldsymbol{\Sigma}_a = (0.7^{|p-q|})_{p,q=1}^a$,
Case 3. $\boldsymbol{\alpha}_{1i} \sim N_{15}(\mathbf{0}_{15}, \mathbf{I}_{15})$ and $\alpha_{2i,m} = \log|\alpha_{1i,m}|$ for $m = 1, \ldots, 15$,
Case 4. $\boldsymbol{a}_i = (a_{i,1}, \ldots, a_{i,15})^\top \sim N_{15}(\mathbf{0}_{15}, \mathbf{I}_{15})$ and $\alpha_{1i,m} = \sin(a_{i,m})$ and $\alpha_{2i,m} = \cos(a_{i,m})$ for $m = 1, \ldots, 15$,
Case 5. $\boldsymbol{\alpha}_{1i} \sim N_{15}(\mathbf{0}_{15}, \mathbf{I}_{15})$, $\boldsymbol{\eta}_i \sim N_{15}(\mathbf{0}_{15}, \mathbf{I}_{15})$, and $\boldsymbol{\alpha}_{2i} = 0.9G(\boldsymbol{\alpha}_{1i}) + 0.1\boldsymbol{\eta}_i$, where $G(\boldsymbol{\alpha}_{1i}) = (\phi(\alpha_{li,1}), \ldots, \phi(\alpha_{li,15}))^\top$ and ϕ is the probability density function of standard normal distribution,
Case 6. $\boldsymbol{\alpha}_{1i} \sim N_{15}(\mathbf{0}_{15}, \mathbf{I}_{15})$, $\boldsymbol{\eta}_i \sim N_{15}(\mathbf{0}_{15}, \mathbf{I}_{15})$, $\boldsymbol{\tau}_i \sim N_{15}(\mathbf{0}_{15}, \mathbf{I}_{15})$, $\boldsymbol{\alpha}_{2i} = 0.9\boldsymbol{\alpha}_{1i} \circ \boldsymbol{\tau}_i + 0.1\boldsymbol{\eta}_i$, where $\circ$ denotes the component-wise multiplication.

Moreover, we consider Cases 7–12, which are the same as Cases 1–6, but the vectors $\boldsymbol{\alpha}_{1i}$ and $\boldsymbol{a}_{1i}$ are generated from T_5 distribution. Finally, in the last six Cases 13–18, the elements of these vectors were i.i.d. variables of uniform distribution $U(-1, 1)$.

Scenario 2. Now, we consider mutual independence with $L = 3$. The vectors $\boldsymbol{\alpha}_{1i}$ were generated in the same way as Scenario 1, except Setting 2 and Cases 4, 10, and 16, where it was generated from $N_{15}(\mathbf{0}_{15}, \mathbf{I}_{15})$, $T_5(\mathbf{I}_{15})$, and i.i.d. variables of uniform distribution $U(-1, 1)$ respectively. On the other hand, the vectors $\boldsymbol{\alpha}_{2i}$ and $\boldsymbol{\alpha}_{3i}$ were obtained in the same way as $\boldsymbol{\alpha}_{1i}$ and $\boldsymbol{\alpha}_{2i}$ in Scenario 1 respectively and independently from $\boldsymbol{\alpha}_{1i}$. Thus, under Setting 1, the null hypothesis is true, while in the case of Setting 2, the alternative hypothesis holds.

Table 1
Empirical sizes (as percentages) obtained in Scenario 1 and Setting 1.

Case	$d\text{Cov}_n^2$	$m d\text{Cov}_n^2$	$gh\text{Cov}_n^2$	$m gh\text{Cov}_n^2$	$lh\text{Cov}_n^2$	$m lh\text{Cov}_n^2$	NDIME	mNDIME
Normal distribution								
1	5.6	5.4	6.2	5.6	6.2	5.0	4.0	5.4
2	4.0	4.4	4.0	6.6	4.0	6.4	5.4	5.8
3	5.2	5.8	5.8	5.8	5.2	6.2	5.0	6.0
Student distribution								
4	4.2	4.6	4.0	4.8	5.4	4.2	5.7	5.9
5	5.8	5.8	6.0	5.2	5.8	5.0	5.2	4.2
Uniform distribution								
6	4.8	4.8	4.8	5.6	5.0	6.0	5.8	5.9
7	4.2	5.2	4.4	4.4	4.8	3.8	5.5	5.8

Table 2
Empirical sizes (as percentages) obtained in Scenario 2 and Setting 1. The column M denotes the use of R or S method.

Case	M	$d\text{Cov}_n^2$	$m d\text{Cov}_n^2$	$gh\text{Cov}_n^2$	$m gh\text{Cov}_n^2$	$lh\text{Cov}_n^2$	$m lh\text{Cov}_n^2$	NDIME	mNDIME
Normal distribution									
1	R	5.0	3.4	5.0	4.8	5.0	4.6	5.0	5.6
	S	4.8	4.2	4.6	4.0	3.4	4.6	6.2	5.2
2	R	6.4	5.8	5.8	5.2	6.2	5.0	6.0	6.1
	S	5.6	5.2	5.6	6.2	5.2	4.8	4.2	6.2
3	R	6.8	6.8	7.0	6.8	6.8	7.2	5.4	4.2
	S	4.8	6.0	4.4	5.4	4.6	5.2	5.0	4.8
Student distribution									
4	R	6.0	7.0	7.0	5.4	6.4	6.4	4.8	5.6
	S	5.8	6.6	6.0	5.4	5.8	4.4	4.8	5.2
5	R	6.6	6.2	6.0	5.4	5.8	5.8	5.2	5.8
	S	6.0	6.2	7.0	5.2	6.4	5.0	5.2	5.2
Uniform distribution									
6	R	4.2	4.0	4.6	5.0	3.8	5.2	5.4	5.6
	S	4.8	4.6	4.8	6.4	4.6	5.8	4.0	5.8
7	R	4.4	4.0	6.0	5.6	5.8	5.6	5.9	5.2
	S	4.2	4.2	4.6	5.4	5.2	4.6	6.2	4.4

3.2. Discussion about simulation results

In this section, we discuss the simulation results, which are presented in Tables 1–4. Let us first consider the control of the type I error level. In the case of pairwise independence ($L = 2$), the empirical sizes of all tests are presented in Table 1. We can easily observe that they are close to the significance level $\alpha = 5\%$ in all cases. This implies that all tests considered for pairwise independence control the type I error level at the given significance level. For mutual independence ($L = 3$, Table 2), the new tests NDIME and mNDIME still show this good property. However, some competing tests have empirical sizes greater than 6.9%, the upper limit of 95% binomial confidence interval for empirical size obtain under 500 simulation replications [32]. Thus, they may have too liberal character, which is unacceptable. Nevertheless, in most cases, the known tests present good behavior under the null hypothesis.

Now, we focus on the power. The higher the power the better the test procedure. Let us start with the empirical powers presented in Table 3 under the pairwise independence. In most cases, both new tests are more powerful than the competing tests. The exceptions are Cases 3 and 9 under the normal and Student distributions, where the NDIME test is one of the weakest test procedures. However, the mNDIME test is the most powerful among all tests considered in these cases. Similarly to the known tests, the marginal version of the new test mNDIME is usually more powerful than the NDIME test. However, this is not always the case, i.e., for Cases 16–18 under the uniform distribution, the NDIME test is better than the mNDIME test. Thus, the marginal test mNDIME does not overcome the standard test NDIME all the time. Finally, let us consider the mutual independence ($L = 3$), where the situation is more complicated. Similarly to the case of $L = 2$, all new tests are usually better than the known tests in terms of power. The exceptions are Cases 3, 9, and 16, but the most powerful test in these cases is still one of the new tests. Interestingly, the NDIME-R test is

Table 3
Empirical powers (as percentages) obtained in Scenario 1 and Setting 2.

Case	$dCov_n^2$	$mdCov_n^2$	$ghCov_n^2$	$mghCov_n^2$	$lhCov_n^2$	$mlhCov_n^2$	NDIME	*m*NDIME
Normal distribution								
1	27.2	41.6	28.8	75.4	47.8	78.0	99.6	100.0
2	35.0	52.8	43.4	91.4	71.8	97.0	100.0	100.0
3	12.8	21.0	13.6	42.4	20.4	43.8	13.4	58.8
4	9.6	11.4	8.6	21.2	13.8	26.6	64.0	67.0
5	10.8	15.6	12.0	30.2	18.4	31.0	37.2	88.2
6	7.6	10.2	6.8	16.6	8.8	17.2	61.8	78.0
Student distribution								
7	55.0	67.2	76.4	94.6	92.4	95.8	100.0	100.0
8	60.6	70.4	83.4	98.8	96.8	98.8	100.0	100.0
9	23.4	39.2	33.4	73.4	49.6	69.8	21.2	83.2
10	12.6	20.8	15.0	39.8	26.8	42.4	62.0	91.0
11	17.0	27.2	26.8	54.0	34.2	51.0	65.2	96.2
12	18.6	20.2	30.8	39.0	41.2	38.4	93.8	96.8
Uniform distribution								
13	12.8	21.0	13.0	45.0	23.8	49.6	95.2	98.4
14	31.4	47.0	32.8	79.2	56.0	89.0	100.0	100.0
15	7.0	14.8	7.8	26.4	14.2	30.2	38.2	61.0
16	9.6	19.2	10.0	34.8	17.4	39.4	91.6	74.8
17	6.6	6.8	7.0	9.2	7.8	9.8	24.2	17.6
18	7.0	5.8	7.4	7.0	7.2	7.4	34.2	31.0

Table 4
Empirical powers (as percentages) obtained in Scenario 2 and Setting 2. The column M denotes the use of R or S method.

Case	M	$dCov_n^2$	$mdCov_n^2$	$ghCov_n^2$	$mghCov_n^2$	$lhCov_n^2$	$mlhCov_n^2$	NDIME	*m*NDIME
Normal distribution									
1	R	17.6	25.6	24.8	51.4	40.6	56.2	98.6	97.6
	S	14.4	22.6	15.2	49.0	23.0	51.4	89.8	99.6
2	R	29.8	35.6	39.6	65.8	72.4	77.0	100.0	100.0
	S	22.6	34.4	26.2	61.6	39.2	73.4	96.6	100.0
3	R	10.0	10.4	13.0	22.8	20.2	22.6	11.4	24.4
	S	9.4	13.6	9.8	23.2	11.8	23.2	7.0	37.2
4	R	6.0	7.4	6.8	14.0	10.4	14.6	44.6	22.2
	S	5.4	5.2	5.2	12.6	6.2	13.8	7.4	33.8
5	R	5.6	4.6	10.0	15.0	12.8	15.6	28.6	59.4
	S	4.6	4.6	5.2	14.2	5.8	14.4	6.6	70.8
6	R	10.2	10.6	9.4	9.4	8.6	8.4	58.6	40.6
	S	8.2	8.2	6.2	8.6	6.8	8.0	36.8	58.0
Student distribution									
7	R	42.8	45.4	71.0	83.6	91.8	86.4	99.8	99.6
	S	32.4	44.6	48.6	78.8	62.4	82.4	94.8	100.0
8	R	52.6	52.8	77.4	89.0	94.6	94.6	100.0	100.0
	S	37.6	52.4	59.2	86.6	73.4	92.2	95.0	100.0
9	R	16.2	20.2	25.8	45.0	39.8	39.0	14.6	31.0
	S	10.6	20.8	13.6	43.2	23.2	41.6	9.6	50.6
10	R	6.4	7.0	11.6	19.4	20.2	21.2	25.2	28.0
	S	5.4	8.4	6.8	18.8	7.6	20.8	6.4	47.4
11	R	3.8	5.2	22.6	31.6	27.2	27.6	36.0	69.2
	S	4.0	6.0	6.8	30.4	6.8	26.2	9.0	83.4
12	R	13.4	10.0	24.8	19.8	37.2	22.4	92.8	67.2
	S	10.2	11.8	12.8	18.8	15.0	20.0	59.0	84.0
Uniform distribution									
13	R	8.8	8.2	11.2	21.8	20.8	26.0	94.8	68.6
	S	6.8	9.4	7.6	19.2	8.2	22.2	53.6	86.0
14	R	17.8	21.4	24.4	51.2	45.4	62.8	100.0	97.8
	S	12.2	19.6	14.0	47.0	16.0	59.2	84.6	99.8
15	R	9.0	10.6	9.8	14.4	12.4	13.8	34.6	25.8
	S	6.8	11.0	6.8	14.4	10.2	15.6	25.8	37.0
16	R	5.4	5.4	9.6	14.8	15.4	17.8	0.2	22.0
	S	5.0	5.8	4.6	15.2	5.2	17.8	13.2	39.6
17	R	5.0	5.0	5.2	5.0	5.8	5.6	20.2	7.8
	S	4.6	6.0	4.2	4.4	4.4	5.2	7.8	10.4
18	R	6.0	5.6	5.4	5.6	6.2	4.8	30.0	13.6
	S	6.4	5.6	5.8	7.2	6.2	8.2	19.8	19.8

usually more powerful than the NDIME-S test, while for the *m*NDIME-based tests, the opposite holds. This is perhaps the cause why it is difficult to indicate the most powerful test in most cases. Sometimes this is the NDIME-R test (Cases 4, 6, 12, 13, 14, 17, and 18), while the other time, it is the *m*NDIME-S test (Cases 1, 3, 5, 7, 9, 10, 11, 15, and 16).

To sum up, the new tests control the type I error level appropriately and at least one of them is the most powerful test procedure. Thus, they can be used as possible good statistical methods for verifying the independence of functional data.

4. Real data example

In this section, we present two real data examples of the use of the tests considered in Section 3. Unless otherwise stated, we perform

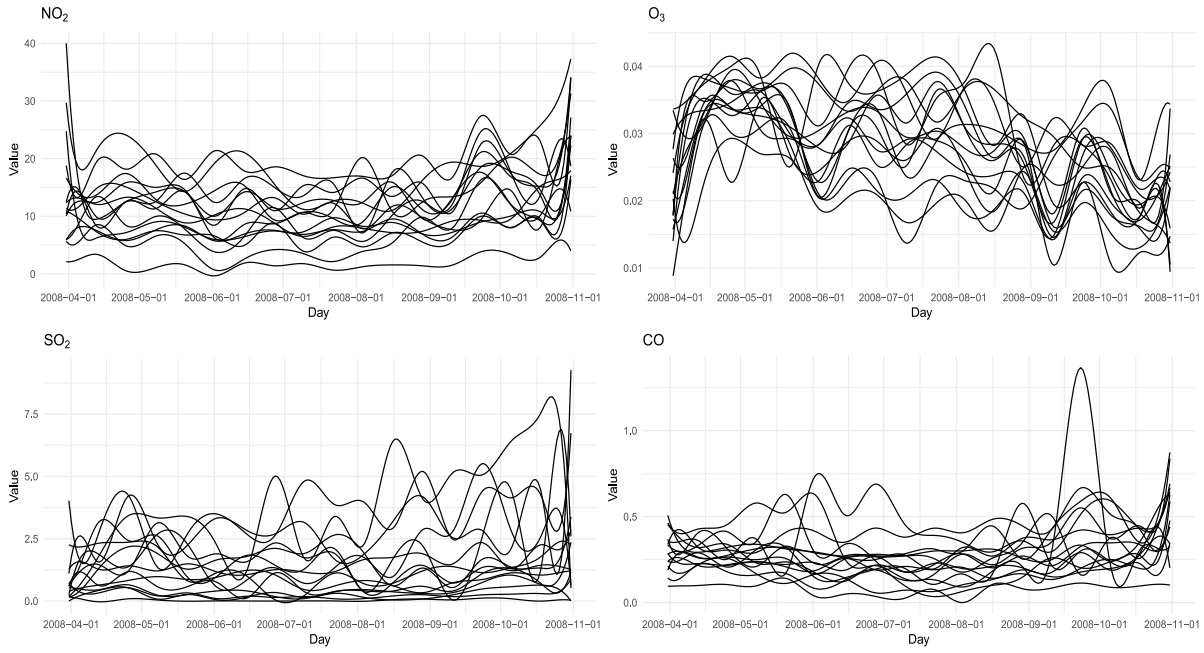

Fig. 1. Smoothed concentration measure curves of air pollution data.

Table 5
P-values (as percentages) for independence testing in the air pollution data set. The column C (respectively: M) denotes the comparison (respectively: the use of R or S method).

C	M	dCov_n^2	$m\mathrm{dCov}_n^2$	ghCov_n^2	$m\mathrm{ghCov}_n^2$	lhCov_n^2	$m\mathrm{lhCov}_n^2$	NDIME	*m*NDIME
Pairwise independence									
1		10.9	4.0	17.0	8.1	13.7	5.6	17.9	4.0
2		5.1	8.1	11.8	36.1	15.5	19.2	1.3	40.7
3		39.9	17.1	38.4	8.5	51.1	10.4	2.5	4.4
4		5.3	6.7	10.4	31.2	12.5	24.4	1.2	38.6
5		32.9	7.6	32.6	1.8	41.9	2.2	1.7	1.4
Mutual independence									
6	R	4.7	8.4	23.6	30.0	41.1	12.2	5.2	21.5
	S	4.7	8.4	16.7	54.4	21.4	29.9	0.6	23.3
7	R	97.9	31.0	26.8	9.1	47.3	6.8	4.8	12.9
	S	97.0	32.9	73.4	9.0	57.8	8.7	3.3	14.0
8	R	4.7	7.9	22.9	20.7	40.5	10.1	4.8	26.6
	S	4.7	8.1	17.2	34.7	21.6	21.4	0.5	25.3

them in the same way as in that section, but we use 1000 permutation samples.

4.1. Air pollution

The first data example is based on the U.S. air pollution data. They are available from https://www.kaggle.com/datasets/sogun3/uspollution. The data set contains daily concentration measurements of four major pollutants, namely, Nitrogen Dioxide (NO_2), Ozone (O_3), Sulfur Dioxide (SO_2), and Carbon Monoxide (CO) in U.S. cities for many years. As functional data, they were studied in Munko et al. [8] and Zhu et al. [9]. In [8], in particular, the data for $n = 15$ cities in the central states (Arizona, Arkansas, Colorado, Illinois, Iowa, Kansas, Louisiana, Missouri, and Texas) were considered. We take these data for illustrative purposes. The data consider the period from 31 March to 31 October 2008. As in [9], each daily concentration measurement raw curve was first smoothed using a B-spline basis of order 4 and with 20 basis functions. The resulting smoothed concentration measure curves are presented in Fig. 1.

In the literature (see for example [33,34]), the dependence between various air pollutants is studied. The results suggest a potential relation between them. We would like to check this using the tests that are considered for the independence of functional data. We consider the following comparisons between air pollutants: pairwise independence [1] NO_2 - O_3, [2] NO_2 - SO_2, [3] NO_2 - CO, [4] (NO_2, O_3) - (SO_2, CO), [5] (NO_2, SO_2) - (O_3, CO); mutual independence [6] NO_2 - O_3 - SO_2, [7] NO_2 - O_3 - CO, [8] NO_2 - O_3 - (SO_2, CO). The results are presented in Table 5. For pairwise independence, we can observe that at least one of the new tests detects the significant dependence between air pollutants. For two comparisons (3 and 5), both new tests agree to reject the null hypothesis about independence. On the other hand, for comparison 1, the *m*NDIME test rejects the null and the NDIME test does not do that, while for comparisons 2 and 4, the opposite is true. The known tests detect significant differences very rarely. In the case of mutual independence, the situation changes a little for new tests. Namely, the NDIME tests reject almost all null hypotheses (except the NDIME-R test for comparison 6 is on the boundary), while the *m*NDIME tests do not detect any significant differences. This can be a little surprising as in Section 3, the new *m*NDIME tests were the most powerful in most cases. The reason can be the distribution of the data in this particular data set. The *m*NDIME test was very powerful in Section 3, but not always the best test (see for example cases 17 and 18 in Table 4). It seems that the distribution of the air pollution data set is not in favor of the *m*NDIME test. The reason for this is difficult to find out. Here, we can see the potential for improvement of our methods, perhaps using functional PCA suggested by the Reviewer (see Section 5 for detail). For the NDIME tests, the S method seems to be more powerful than the R method in these cases as it gives smaller *p*-values. From the known tests,

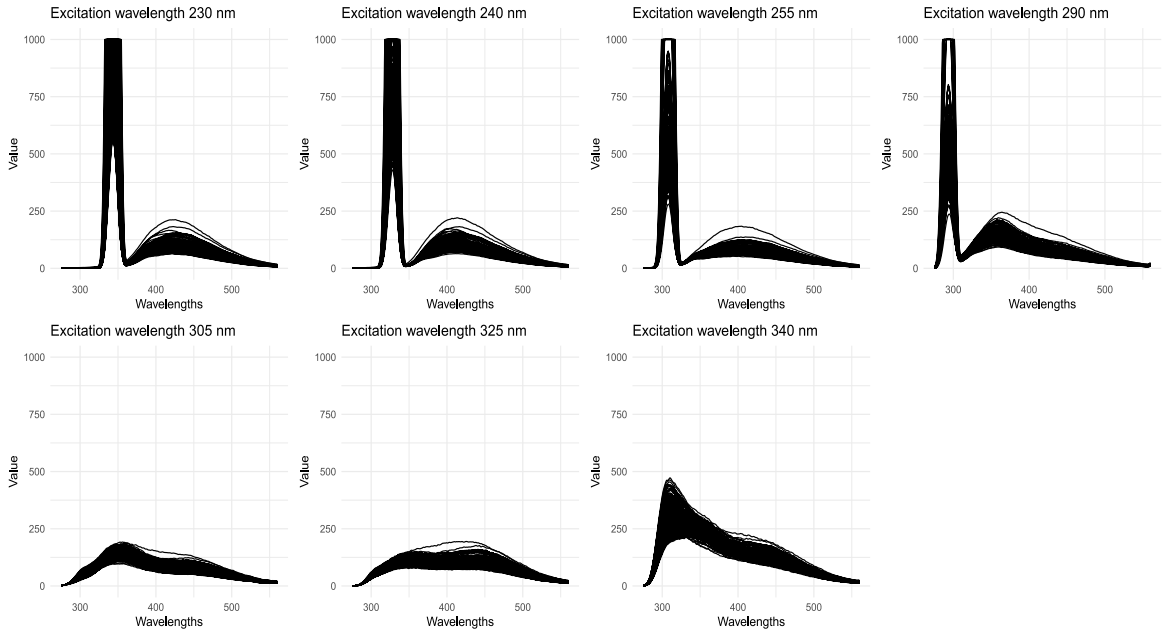

Fig. 2. The emission spectra measured at 571 wavelengths $275, 275.5, 276, \ldots, 560$ at seven excitation wavelengths 230, 240, 255, 290, 305, 325, and 340 nm.

the dCov_n^2-based tests have p-values slightly smaller than significant level only, while the other ones do not reject any null hypotheses.

4.2. Sugar spectra data

The second real data example concerns the chemometrical data. We consider the sugar spectra data set, which is described in Bro [35] and Munck et al. [36] and available from the following address: http://www.models.life.ku.dk/Sugar_Process. In this data set, the emission spectra from 275–560 nm were measured in 0.5 nm intervals for each sample of sugar, i.e., at 571 wavelengths $275, 275.5, 276, \ldots, 560$ nm. Moreover, the seven excitation wavelengths $230, 240, 255, 290, 305, 325,$ and 340 nm were studied. There are 268 samples of sugar. Then, for each excitation wavelength, the emission spectra can be treated as the functional data, which are given in discretized form, i.e., their values are provided at 571 design time points (wavelengths). Thus, we have seven functional variables (excitation wavelengths), presented in Fig. 2. We will verify the independence of some of these functional variables, but first, let us present a context for such a study.

In chemometrics, the content of ingredients is usually determined by certain functions, e.g., absorbance or emission spectra. Such a procedure is usually much cheaper than chemical analysis. In the sugar spectra data set, laboratory determinations of the sugar quality are also made. In particular, one of them is ash content, which measures the amount of inorganic impurities in refined sugar. In Gertheiss et al. [37] and Smaga and Matsui [3], the association between the ash content and the spectra measured at seven excitation wavelengths was considered using the scalar response functional regression model. The ash content was the scalar response variable, and the seven excitation wavelengths were the functional predictors. The aim of the study was to determine the most useful excitation wavelengths to predict the ash content. To obtain easier and cheaper analysis, the smallest number of excitation wavelengths was desirable to identify. By [3], the smallest set of excitation wavelengths with the lowest prediction error was $\{290, 325, 340\}$.

Coming back to the topic of the present paper, we would like to verify the independence of these three selected excitation wavelengths and the remaining ones. Thus, we have two groups of functional variables ($L = 2$), where in the first group, there are four variables $(230, 240, 255, 305)$, while in the second one, we have three variables $(290, 325, 340)$. The visual inspection from Fig. 2 indicates that the excitation wavelength 290 has a very similar shape to the excitation wavelengths 230,240, and 255. The same seems to be true for excitation wavelengths 305 and 325. Thus, we expect a strong dependence on these two groups of excitation wavelengths. To verify this, we applied all tests considered in this paper. For this purpose, we used 49 basis functions in the basis expansion of the data using Fourier basis for simplicity. Here, we used greater number of basis functions due to the complicated form of the data, especially for the excitation wavelengths $230, 240, 255,$ and 290. The number 49 was chosen based on graphical inspection, and the smoothed trajectories of the emission spectra are presented in Fig. 3. Nevertheless, we obtained the same results for five basis functions. The p-value of each of them was equal to zero. This consistency of the results of all tests is perhaps caused by the larger number of observations in this example, i.e., there are $n = 268$ functional observations. Therefore, we reject the null hypothesis and detect significant dependence between $230, 240, 255, 305$ and $290, 325, 340$ excitation wavelengths. This indicates the correctness of the proposed functional regression model and that not all excitation wavelengths have to be used in the analysis.

5. Conclusions

We have developed and analyzed new methods for independence testing in functional data, applicable to both univariate and multivariate cases. Our approach involves reducing the dimensionality of the functional data using basis expansion, followed by the application of the $\mathcal{L}^2$ distance of mean embedding by Zhang et al. [25] as a flexible measure of independence. To enhance test performance and applicability to mutual independence testing, we combined this measure for pairwise independence with marginal aggregation method by Zhu et al. [23], as well as asymmetric and symmetric aggregation measures by Jin and Matteson [24].

We conducted a comparative analysis with established tests based on distance covariance and the Hilbert–Schmidt covariance, including simulation studies and real data applications. The results demonstrate that our proposed methods exhibit robust finite-sample properties, with effective control of the type I error rate, whose performance proved competitive relative to other methods considered in this study. In most cases, the proposed methods exhibit higher statistical power than

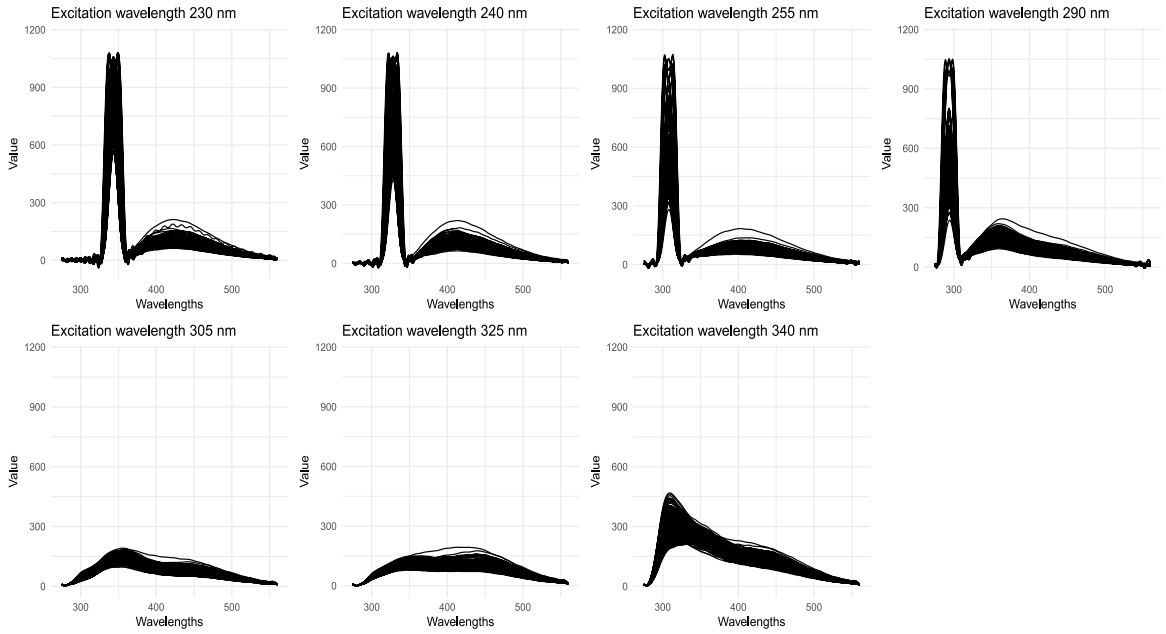

Fig. 3. The smoothed trajectories using the Fourier basis of the emission spectra measured at 571 wavelengths $275, 275.5, 276, \ldots, 560$ at seven excitation wavelengths 230, 240, 255, 290, 305, 325, and 340 nm.

covariance-based approaches. Empirical data tests further underscore the applicability of the proposed methods, as the results aligned with theoretical expectations.

These findings suggest that the new procedures are viable alternatives to traditional covariance-based tests, providing a promising tool for independence analysis in functional data settings. One issue that may be addressed in the future is the lack of a single test that is the most powerful, although this is often difficult to achieve. However, a compelling alternative may involve the combined application of both tests, as our analysis indicates that at least one consistently detects data dependency when present.

In a future study, we probably should consider use of the Karhunen–Loève expansion and the functional principal component analysis suggested by the Reviewer, which provides an optimal representation of a functional random variable in the sense of the mean squared error. It is also a popular analysis used for functional data (see, for example, Vidal et al. [38] for some recent application of it for classification). In our approach, the basis functions are fixed and independent of the data. Thus, independence testing is reduced to testing it for the random vectors of coefficients. On the other hand, the problem in using the Karhunen-Loève expansion and the functional principal component analysis in the independence testing is that the basis functions are unknown and have to be estimated from the data. Thus, they depend on all observations, and hence the information about the independence of functional data is not just in the coefficients as in our method. This problem for the portmanteau test of independence was also studied in Chapter 7 in Horváth and Kokoszka [11]. Therefore, using the functional principal component analysis in the testing for independence of functional data is more challenging than our approach.

CRediT authorship contribution statement

Mirosław Krzyśko: Writing – review & editing, Conceptualization. **Łukasz Smaga:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Jędrzej Wydra:** Writing – review & editing, Validation, Software, Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We would like to thank anonymous reviewers for their comments that helped us to improve the manuscript. A part of calculations for simulation study and real data example was made at the Poznań Supercomputing and Networking Center (grant pl0253-02).

Data availability

Data will be made available on request.

References

[1] J. Leśkow, A. Napolitano, Foundations of the functional approach for signal analysis, Signal Process. 86 (2006) 3796–3825.
[2] J.A.A. Collazos, R. Ronaldo Dias, A.Z. Zambom, Consistent variable selection for functional regression models, J. Multivariate Anal. 146 (2016) 63–71.
[3] Ł. Smaga, H. Matsui, A note on variable selection in functional regression via random subspace method, Stat. Methods Appl. 27 (2018) 455–477.
[4] R.H. Glendinning, S.L. Fleet, Classifying functional time series, Signal Process. 87 (2007) 79–100.
[5] M. Zhang, A. Parnell, Review of clustering methods for functional data, ACM Trans. Knowl. Discov. Data 17 (2023) 91.
[6] M. Krzyśko, Ł. Waszak, Canonical correlation analysis for functional data, Biom. Lett. 50 (2013) 95–105.
[7] C. Acal, A.M. Aguilera, Basis expansion approaches for functional analysis of variance with repeated measures, Adv. Data Anal. Classif. 17 (2023) 291–321.
[8] M. Munko, M. Ditzhaus, M. Pauly, Ł. Smaga, Multiple comparison procedures for simultaneous inference in functional MANOVA, 2024, preprint arXiv:2406.01242.
[9] T. Zhu, J.T. Zhang, M.Y. Cheng, One-way MANOVA for functional data via Lawley–Hotelling trace test, J. Multivariate Anal. 192 (2) (2022) 105095.
[10] F. Ferraty, P. Vieu, Nonparametric Functional Data Analysis: Theory and Practice, Springer, New York, 2006.
[11] L. Horváth, P. Kokoszka, Inference for Functional Data with Applications, Springer, New York, 2012.
[12] J. Ramsay, B.W. Silverman, Functional Data Analysis, Springer, 2005.
[13] J.T. Zhang, Analysis of Variance for Functional Data, Chapman & Hall, 2013.

[14] P. Comon, Independent component analysis, a new concept? Signal Process. 36 (1994) 287–314.
[15] M. Vidal, M. Rosso, A.M. Aguilera, Bi-smoothed functional independent component analysis for EEG artifact removal, Math. 9 (2021) 1243.
[16] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (2003) 1157–1182.
[17] T. Lai, Z. Zhang, Y. Wang, L. Kong, Testing independence of functional variables by angle covariance, J. Multivariate Anal. 182 (2021) 104711.
[18] L. Zhu, K. Xu, R. Li, W. Zhong, Projection correlation between two random vectors, Biom. 104 (2017) 829–843.
[19] T.K.M. Djonguet, A.M. Mbina, G.M. Nkiet, Testing independence of functional variables by an Hilbert-Schmidt independence criterion estimator, Statist. Probab. Lett. 207 (2024) 110016.
[20] T. Górecki, M. Krzyśko, W. Ratajczak, W. Wołyński, An extension of the classical distance correlation coefficient for multivariate functional data with applications, Stat. Transit. New Ser. 17 (2016) 449–466.
[21] T. Górecki, M. Krzyśko, W. Wołyński, Independence test and canonical correlation analysis based on the alignment between kernel matrices for multivariate functional data, Artif. Intell. Rev. 53 (2020) 475–499.
[22] M. Krzyśko, Ł. Smaga, P. Kokoszka, Marginal distance and Hilbert-Schmidt covariances-based independence tests for multivariate functional data, J. Artificial Intelligence Res. 73 (2022) 1355–1384.
[23] C. Zhu, X. Zhang, S. Yao, X. Shao, Distance-based and RKHS-based dependence metrics in high dimension, Ann. Statist. 48 (2020) 3366–3394.
[24] Z. Jin, D.S. Matteson, Generalizing distance covariance to measure and test multivariate mutual dependence via complete and incomplete V-statistics, J. Multivariate Anal. 168 (2018) 304–322.
[25] W. Zhang, W. Gao, H. Ng, Multivariate tests of independence based on a new class of measures of independence in Reproducing Kernel Hilbert Space, J. Multivariate Anal. 195 (2023) 105144.
[26] J.A. Cuesta-Albertos, M. Febrero-Bande, A simple multiway ANOVA for functional data, Test 19 (2010) 537–557.
[27] Ł. Smaga, Review of methods for functional one-way analysis of variance, REVSTAT - Stat. J. (2024) in press, URL https://revstat.ine.pt/index.php/REVSTAT/article/view/803/726.
[28] G. Shmueli, To explain or to predict? Statist. Sci. 25 (2010) 289–310.
[29] A. Gretton, K.M. Borgwardt, M.J. Rasch, B. Schölkopf, A. Smola, A kernel two-sample test, J. Mach. Learn. Res. 13 (2012) 723–773.
[30] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2024, URL www.R-project.org/.
[31] A. Gretton, K. Fukumizu, Z. Harchaoui, B.K. Sriperumbudur, A fast, consistent kernel two-sample test, in: Advances in Neural Information Processing Systems 22, Curran Associates, Inc., 2009, pp. 673–681.
[32] P. Duchesne, C. Francq, Multivariate hypothesis testing using generalized and {2}-inverses - with applications, Stat. 49 (2015) 475–496.
[33] S. Khedekar, S. Thakare, Correlation analysis of atmospheric pollutants and meteorological factors using statistical tools in Pune, Maharashtra, E3S Web Conf. 391 (2023) 01190.
[34] S. Rahaman, X. Tu, K. Ahmad, A. Qadeer, A real-time assessment of hazardous atmospheric pollutants across cities in China and India, J. Hazard. Mater. 479 (2024) 135711.
[35] R. Bro, Exploratory study of sugar production using fluorescence spectroscopy and multi-way analysis, Chemometr. Intell. Lab. Syst. 46 (1999) 133–147.
[36] L. Munck, L. Nørgaard, S.B. Engelsen, R. Bro, C.S. Andersson, Chemometrics in food science – A demonstration of the feasibility of a highly exploratory, inductive evaluation strategy of fundamental scientific significance, Chemometr. Intell. Lab. Syst. 44 (1998) 31–60.
[37] J. Gertheiss, A. Maity, A.M. Staicu, Variable selection in generalized functional linear models, Stat 2 (2013) 86–101.
[38] M. Vidal, M. Leman, A.M. Aguilera, Functional independent component analysis by choice of norm: a framework for near-perfect classification, Adv. Data Anal. Classif. (2025) http://dx.doi.org/10.1007/s11634-024-00622-5, in press.