Artificial Intelligence Review (2020) 53:475–499
https://doi.org/10.1007/s10462-018-9666-7

Independence test and canonical correlation analysis based on the alignment between kernel matrices for multivariate functional data

Tomasz Górecki[1] · **Mirosław Krzyśko**[1,2] · **Waldemar Wołyński**[1]

Published online: 10 November 2018

Abstract
In the case of vector data, Gretton et al. (Algorithmic learning theory. Springer, Berlin, pp 63–77, 2005) defined Hilbert–Schmidt independence criterion, and next Cortes et al. (J Mach Learn Res 13:795–828, 2012) introduced concept of the centered kernel target alignment (KTA). In this paper we generalize these measures of dependence to the case of multivariate functional data. In addition, based on these measures between two kernel matrices (we use the Gaussian kernel), we constructed independence test and nonlinear canonical variables for multivariate functional data. We show that it is enough to work only on the coefficients of a series expansion of the underlying processes. In order to provide a comprehensive comparison, we conducted a set of experiments, testing effectiveness on two real examples and artificial data. Our experiments show that using functional variants of the proposed measures, we obtain much better results in recognizing nonlinear dependence.

Keywords Multivariate functional data · Functional data analysis · Correlation analysis · Canonical correlation analysis

1 Introduction

The theory and practice of statistical methods in situations where the available data are functions (instead of real numbers or vectors) is often referred to as Functional Data Analysis (FDA). The term Functional Data Analysis was already used by Ramsay and Dalzell (1991)

✉ Tomasz Górecki
tomasz.gorecki@amu.edu.pl

Mirosław Krzyśko
mkrzysko@amu.edu.pl

Waldemar Wołyński
wolynski@amu.edu.pl

[1] Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Umultowska 87, 61-614 Poznań, Poland

[2] Faculty of Management, President Stanisław Wojciechowski Higher Vocational State School, Nowy Świat 4, 62-800 Kalisz, Poland

two decades ago. This subject has become increasingly popular from the end of the 1990s and is now a major research field in statistics (Cuevas 2014). Good access to the large literature in this field comes from the books by Ramsay and Silverman (2002, 2005), Ferraty and Vieu (2006), and Horváth and Kokoszka (2012). Special issues devoted to FDA topics have been published by different journals, including Statistica Sinica 14(3) (2004), Computational Statistics 22(3) (2007), Computational Statistics and Data Analysis 51(10) (2007), Journal of Multivariate Analysis 101(2) (2010), Advances in Data Analysis and Classification 8(3) (2014).

The range of real world applications, where the objects can be thought of as functions, is as diverse as speech recognition, spectrometry, meteorology, medicine or clients segmentation, to cite just a few (Ferraty and Vieu 2003; James et al. 2009; Martin-Baragan et al. 2014; Devijver 2017).

The uncentered kernel alignment originally was introduced by Cristianini et al. (2001). Gretton et al. (2005) defined Hilbert–Schmidt Independence Criterion (HSIC) and the empirical HSIC. Centered kernel target alignment (KTA) was introduced by Cortes et al. (2012). This measure is a normalized version of HSIC. Zhang et al. (2011) gave an interesting kernel-based independence test. This independence testing method is closely related to the one based on the Hilbert–Schmidt independence criterion (HSIC) proposed by Gretton et al. (2008). Gretton et al. (2005) described a permutation-based kernel independence test. There is a lot of work in the literature for kernel alignment and its applications (good overview can be found in Wang et al. 2015).

This work is devoted to a generalization of these measures of dependence to the case of multivariate functional data. In addition, based on these measures we constructed independence test and nonlinear canonical correlation variables for multivariate functional data. These results are based on the assumption that the applied kernel function is Gaussian. Functional HSIC and KTA canonical correlation analysis can be viewed as natural nonlinear extension of functional canonical correlation analysis (FCCA). So, we propose two nonlinear functional CCA extensions that capture nonlinear relationship. Moreover, both algorithms are capable of extracting also linear dependency. Additionally, we show that functional KTA approach is only a normalized variant of HSIC coefficient also for functional data. Finally, we propose some interpretation of module weighting functions for functional canonical correlations.

Section 2 provides an overview of centered measures alignment for random vectors. They are defined by such concepts as: kernel function alignment, kernel matrix alignment, and Hilbert–Schmidt Independence Criterion (HSIC) and associations between them have been shown. Functional data can be seen as values of random processes. In our paper, the multivariate random function $\boldsymbol{X}$ and $\boldsymbol{Y}$ have special representation (8) in finite dimensional subspaces of the spaces of square integrable functions on the given intervals. In Sect. 3 we present kernel-based independence test. Section 4 discusses the concept of alignment for multivariate functional data. The kernel function, the alignment between two kernels functions, the centered kernel alignment (KTA) between two kernel matrices and the empirical Hilbert–Schmidt Independence Criterion (HSIC) are defined. The HSIC was used as the basis for an independence test. In Sect. 5 we present kernel-based independence test for multivariate functional data. In Sect. 5, based on the concept of alignment between kernel matrices, nonlinear canonical variables were constructed. It is a generalization of the results of Chang et al. (2013) for random vectors. In Sect. 5 we present an one artificial and two real examples which confirm the usefulness of proposed coefficients in detection of nonlinear dependency for group of variables.

2 An overview of kernel alignment and its applications

We introduce the following notational convention. Throughout this section, $\boldsymbol{X}$ and $\boldsymbol{Y}$ are random vectors, with domains $\mathbb{R}^p$ and $\mathbb{R}^q$, respectively. Let $P_{\boldsymbol{X},\boldsymbol{Y}}$ be a joint probability measure on $(\mathbb{R}^p \times \mathbb{R}^q, \Gamma \times \Lambda)$ (here Γ and Λ are the Borel σ-algebras on $\mathbb{R}^p$ and $\mathbb{R}^q$, respectively), with associated marginal probability measures $P_{\boldsymbol{X}}$ and $P_{\boldsymbol{Y}}$.

Definition 1 (*Kernel functions*, Shawe-Taylor and Cristianini 2004) A kernel is a function k that for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^p$ satisfies

$$k(\boldsymbol{x}, \boldsymbol{x}') = \langle \boldsymbol{\varphi}(\boldsymbol{x}), \boldsymbol{\varphi}(\boldsymbol{x}') \rangle_{\mathcal{H}},$$

where φ is a mapping from $\mathbb{R}^p$ to an inner product feature space $\mathcal{H}$

$$\varphi : \boldsymbol{x} \to \varphi(\boldsymbol{x}) \in \mathcal{H}.$$

We call $\boldsymbol{\varphi}$ a feature map.

A kernel function can be interpreted as a kind of similarity measure between the vectors $\boldsymbol{x}$ and $\boldsymbol{x}'$.

Definition 2 (*Gram matrix*, Mercer 1909; Riesz 1909; Aronszajn 1950) Given a kernel k and inputs $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^p$, the $n \times n$ matrix $\boldsymbol{K}$ with entries $K_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is called the Gram matrix (kernel matrix) of k with respect to $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$.

Definition 3 (*Positive semi-definite matrix*, Hofmann et al. 2008) A real $n \times n$ symmetric matrix $\boldsymbol{K}$ with entries K_{ij} satisfying

$$\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j K_{ij} \geq 0$$

for all $c_i \in \mathbb{R}$ is called positive semi-definite.

Definition 4 (*Positive semi-definite kernel*, Mercer 1909; Hofmann et al. 2008) A function $k : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ which for all $n \in \mathbb{N}$, $\boldsymbol{x}_i \in \mathbb{R}^p$, $i = 1, \ldots, n$ gives rise to a positive semi-definite Gram matrix is called a positive semi-definite kernel.

This raises an interesting question: given a function of two variables $k(\boldsymbol{x}, \boldsymbol{x}')$, does there exist a function $\boldsymbol{\varphi}(\boldsymbol{x})$ such that $k(\boldsymbol{x}, \boldsymbol{x}') = \langle \boldsymbol{\varphi}(\boldsymbol{x}), \boldsymbol{\varphi}(\boldsymbol{x}') \rangle_{\mathcal{H}}$? The answer is provided by Mercer's theorem (1909) which says, roughly, that if k is positive semi-definite then such a φ exists.

Often, we will not known $\boldsymbol{\phi}$, but a kernel function $k : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ that encodes the inner product in $\mathcal{H}$, instead.

Popular positive semi-definite kernel functions on $\mathbb{R}^p$ include the polynomial kernel of degree $d > 0$, $k(\boldsymbol{x}, \boldsymbol{x}') = (1 + \boldsymbol{x}^\top \boldsymbol{x}')^d$, the Gaussian kernel $k(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\lambda \|\boldsymbol{x} - \boldsymbol{x}'\|^2)$, $\lambda > 0$, and the Laplace kernel $k(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\lambda \|\boldsymbol{x} - \boldsymbol{x}'\|)$, $\lambda > 0$. In this paper we use, the Gaussian kernel.

We start with the definition of centering and the analysis of its relevant properties.

2.1 Centered kernel functions

A feature mapping $\boldsymbol{\phi}\colon \mathbb{R}^p \to \mathcal{H}$ is centered by subtracting from it its expectation, that is transforming $\boldsymbol{\phi}(\boldsymbol{x})$ to $\tilde{\boldsymbol{\phi}}(\boldsymbol{x}) = \boldsymbol{\phi}(\boldsymbol{x}) - \mathrm{E}_{\boldsymbol{X}}[\boldsymbol{\phi}(\boldsymbol{X})]$, where $\mathrm{E}_{\boldsymbol{X}}$ denotes the expected value of $\boldsymbol{\phi}(\boldsymbol{X})$ when $\boldsymbol{X}$ is distributed according to $P_{\boldsymbol{X}}$. Centering a positive semi-definite kernel function $k\colon \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ consists centering in the feature mapping $\boldsymbol{\phi}$ associated to k. Thus, the the centered kernel $\tilde{k}$ associated to k is defined by

$$\begin{aligned}\tilde{k}(\boldsymbol{x}, \boldsymbol{x}') &= \langle \boldsymbol{\phi}(\boldsymbol{x}) - \mathrm{E}_{\boldsymbol{X}}[\boldsymbol{\phi}(\boldsymbol{X})], \boldsymbol{\phi}(\boldsymbol{x}') - \mathrm{E}_{\boldsymbol{X}'}[\boldsymbol{\phi}(\boldsymbol{X}')]\rangle \\ &= k(\boldsymbol{x}, \boldsymbol{x}') - \mathrm{E}_{\boldsymbol{X}}[k(\boldsymbol{X}, \boldsymbol{x}')] - \mathrm{E}_{\boldsymbol{X}'}[k(\boldsymbol{x}, \boldsymbol{X}')] + \mathrm{E}_{\boldsymbol{X},\boldsymbol{X}'}[k(\boldsymbol{X}, \boldsymbol{X}')],\end{aligned}$$

assuming the expectations exist. Here, the expectation is taken over independent copies $\boldsymbol{X}$, $\boldsymbol{X}'$ distributed according to $P_{\boldsymbol{X}}$. We see that, $\tilde{k}$ is also a positive semi-definite kernel. Note also that for a centered kernel $\tilde{k}$, $\mathrm{E}_{\boldsymbol{X},\boldsymbol{X}'}[\tilde{k}(\boldsymbol{X}, \boldsymbol{X}')] = 0$, that is, centering the feature mapping implies centering the kernel function.

2.2 Centered kernel matrices

Let $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ be a finite subset of $\mathbb{R}^p$. A feature mapping $\boldsymbol{\phi}(\boldsymbol{x}_i)$, $i = 1, \ldots, n$, is centered by subtracting from it its empirical expectation, i.e., leading to $\tilde{\boldsymbol{\phi}}(\boldsymbol{x}_i) = \boldsymbol{\phi}(\boldsymbol{x}_i) - \overline{\boldsymbol{\phi}}$, where $\overline{\boldsymbol{\phi}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{\phi}(\boldsymbol{x}_i)$. The kernel matrix $\boldsymbol{K} = (K_{ij})$ associated to the kernel function k and the set $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ is centered by replacing it with $\widetilde{\boldsymbol{K}} = (\widetilde{K}_{ij})$ defined for all $i, j = 1, 2, \ldots, n$ by

$$\tilde{K}_{ij} = K_{ij} - \frac{1}{n}\sum_{i=1}^n K_{ij} - \frac{1}{n}\sum_{j=1}^n K_{ij} + \frac{1}{n^2}\sum_{i,j=1}^n K_{ij}, \tag{1}$$

where $K_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$, $i, j = 1, \ldots, n$.

The centered kernel matrix $\widetilde{\boldsymbol{K}}$ is a positive semi-definite matrix. Also, as with the kernel function $\frac{1}{n^2}\sum_{i,j}^n \tilde{K}_{ij} = 0$.

Let $\langle \cdot, \cdot \rangle_F$ denote the Frobenius product and $\| \cdot \|_F$ the Frobenius norm defined for all $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times n}$ by

$$\begin{aligned}\langle \boldsymbol{A}, \boldsymbol{B} \rangle_F &= \mathrm{tr}(\boldsymbol{A}^\top \boldsymbol{B}), \\ \|\boldsymbol{A}\|_F &= (\langle \boldsymbol{A}, \boldsymbol{A} \rangle_F)^{1/2}.\end{aligned}$$

Then, for any kernel matrix $\boldsymbol{K} \in \mathbb{R}^{n\times n}$, the centered kernel matrix $\widetilde{\boldsymbol{K}}$ can be expressed as follows (Schölkopf et al. 1998):

$$\widetilde{\boldsymbol{K}} = \boldsymbol{H}\boldsymbol{K}\boldsymbol{H}, \tag{2}$$

where $\boldsymbol{H} = \boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$, $\mathbf{1}_n \in \mathbb{R}^{n\times 1}$ denote the vector with all entries equal to one, and $\boldsymbol{I}_n$ the identity matrix of order n. The matrix $\boldsymbol{H}$ is called "centering matrix".

Since $\boldsymbol{H}$ is the idempotent matrix ($\boldsymbol{H}^2 = \boldsymbol{H}$), then we get for any two kernel matrices $\boldsymbol{K}$ and $\boldsymbol{L}$ based on the subset $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ of $\mathbb{R}^p$ and the subset $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ of $\mathbb{R}^q$, respectively,

$$\langle \widetilde{\boldsymbol{K}}, \widetilde{\boldsymbol{L}} \rangle_F = \langle \boldsymbol{K}, \widetilde{\boldsymbol{L}} \rangle_F = \langle \widetilde{\boldsymbol{K}}, \boldsymbol{L} \rangle_F. \tag{3}$$

2.3 Centered kernel alignment

Definition 5 (*Kernel function alignment*, Cristianini et al. 2001; Cortes et al. 2012) Let k and l be two kernel functions defined over $\mathbb{R}^p \times \mathbb{R}^p$ and $\mathbb{R}^q \times \mathbb{R}^q$, respectively, such that $0 < \mathrm{E}_{\boldsymbol{X},\boldsymbol{X}'}[\tilde{k}^2(\boldsymbol{X},\boldsymbol{X}')] < \infty$ and $0 < \mathrm{E}_{\boldsymbol{Y},\boldsymbol{Y}'}[\tilde{l}^2(\boldsymbol{Y},\boldsymbol{Y}')] < \infty$, where $\boldsymbol{X}, \boldsymbol{X}'$ and $\boldsymbol{Y}, \boldsymbol{Y}'$ are independent copies distributed according to $P_{\boldsymbol{X}}$ and $P_{\boldsymbol{Y}}$, respectively. Then the alignment between k and l is defined by

$$\rho(k,l) = \frac{\mathrm{E}_{\boldsymbol{X},\boldsymbol{X}',\boldsymbol{Y},\boldsymbol{Y}'}[\tilde{k}(\boldsymbol{X},\boldsymbol{X}')\tilde{l}(\boldsymbol{Y},\boldsymbol{Y}')]}{\sqrt{\mathrm{E}_{\boldsymbol{X},\boldsymbol{X}'}[\tilde{k}^2(\boldsymbol{X},\boldsymbol{X}')]\,\mathrm{E}_{\boldsymbol{Y},\boldsymbol{Y}'}[\tilde{l}^2(\boldsymbol{Y},\boldsymbol{Y}')]}}.$$

We can define similarly the alignment between two kernel matrices $\boldsymbol{K}$ and $\boldsymbol{L}$ based on the finite subset $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ and $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$, respectively.

Definition 6 (*Kernel matrix alignment*, Cortes et al. 2012) Let $\boldsymbol{K} \in \mathbb{R}^{n\times n}$ and $\boldsymbol{L} \in \mathbb{R}^{n\times n}$ be two kernel matrices such that $\|\widetilde{\boldsymbol{K}}\|_F \neq 0$ and $\|\widetilde{\boldsymbol{L}}\|_F \neq 0$. Then, the centered kernel target alignment (KTA) between $\boldsymbol{K}$ and $\boldsymbol{L}$ is defined by

$$\hat{\rho}(\boldsymbol{K},\boldsymbol{L}) = \frac{\langle \widetilde{\boldsymbol{K}}, \widetilde{\boldsymbol{L}}\rangle_F}{\|\widetilde{\boldsymbol{K}}\|_F \|\widetilde{\boldsymbol{L}}\|_F}. \tag{4}$$

Here, by the Cauchy–Schwarz inequality, $\hat{\rho}(\boldsymbol{K},\boldsymbol{L}) \in [-1,1]$ and in fact $\hat{\rho}(\boldsymbol{K},\boldsymbol{L}) \in [0,1]$ when $\boldsymbol{K}$ and $\boldsymbol{L}$ are the kernel matrices of the positive semi-definite kernel $\tilde{k}$ and $\tilde{l}$.

Gretton et al. (2005) defined Hilbert–Schmidt Independence Criterion (HSIC) as a test statistic to distinguish between null hypothesis $H_0\colon P_{\boldsymbol{X},\boldsymbol{Y}} = P_{\boldsymbol{X}}P_{\boldsymbol{Y}}$ (equivalently we may write $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}$) and alternative hypothesis $H_1\colon P_{\boldsymbol{X},\boldsymbol{Y}} \neq P_{\boldsymbol{X}}P_{\boldsymbol{Y}}$.

Definition 7 (*Reproducing kernel Hilbert space*, Riesz 1909; Mercer 1909; Aronszajn 1950) Consider a Hilbert space $\mathcal{H}$ of functions from $\mathbb{R}^p$ to $\mathbb{R}$. Then $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS) if for each $\boldsymbol{x} \in \mathbb{R}^p$, the Dirac evaluation operator $\delta_{\boldsymbol{x}}\colon \mathcal{H} \to \mathbb{R}$, which maps $f \in \mathcal{H}$ to $f(\boldsymbol{x}) \in \mathbb{R}$, is a bounded linear functional.

Let $\varphi\colon \mathbb{R}^p \to \mathcal{H}$ be a map such that for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^p$ we have $\langle \boldsymbol{\phi}(\boldsymbol{x}), \boldsymbol{\phi}(\boldsymbol{x}')\rangle_{\mathcal{H}} = k(\boldsymbol{x},\boldsymbol{x}')$, where $k\colon \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ is a unique positive semi-definite kernel. We will require in particular that $\mathcal{H}$ be separable (it must have a complete, countable orthonormal system). We likewise define a second separable RKHS $\mathcal{G}$, with kernel $l(\cdot,\cdot)$ and feature map $\boldsymbol{\psi}$, on the separable space $\mathbb{R}^q$.

We may now define the mean elements $\mu_{\boldsymbol{X}}$ and $\mu_{\boldsymbol{Y}}$ with respect to the measures $P_{\boldsymbol{X}}$ and $P_{\boldsymbol{Y}}$ as those members of $\mathcal{H}$ and $\mathcal{G}$, respectively, for which

$$\langle \mu_{\boldsymbol{X}}, f\rangle_{\mathcal{H}} = \mathrm{E}_{\boldsymbol{X}}[\langle \boldsymbol{\phi}(\boldsymbol{X}), f\rangle_{\mathcal{H}}] = \mathrm{E}_{\boldsymbol{X}}[f(\boldsymbol{X})],$$
$$\langle \mu_{\boldsymbol{Y}}, g\rangle_{\mathcal{G}} = \mathrm{E}_{\boldsymbol{Y}}[\langle \boldsymbol{\psi}(\boldsymbol{Y}), g\rangle_{\mathcal{G}}] = \mathrm{E}_{\boldsymbol{Y}}[g(\boldsymbol{Y})],$$

for all functions $f \in \mathcal{H}$, $g \in \mathcal{G}$, where $\boldsymbol{\phi}$ is the feature map from $\mathbb{R}^p$ to the RKHS $\mathcal{H}$, and $\boldsymbol{\psi}$ maps from $\mathbb{R}^q$ to $\mathcal{G}$ and assuming the expectations exist.

Finally, $\|\mu_{\boldsymbol{X}}\|^2_{\mathcal{H}}$ can be computed by applying the expectation twice via

$$\|\mu_{\boldsymbol{X}}\|^2_{\mathcal{H}} = \mathrm{E}_{\boldsymbol{X},\boldsymbol{X}'}[\langle \boldsymbol{\phi}(\boldsymbol{X}), \boldsymbol{\phi}(\boldsymbol{X}')\rangle_{\mathcal{H}}] = \mathrm{E}_{\boldsymbol{X},\boldsymbol{X}'}[k(\boldsymbol{X},\boldsymbol{X}')],$$

assuming the expectations exist. The expectation is taken over independent copies $\boldsymbol{X}, \boldsymbol{X}'$ distributed according to $P_{\boldsymbol{X}}$. The means $\mu_{\boldsymbol{X}}$, $\mu_{\boldsymbol{Y}}$ exist when positive semi-definite kernels k and l are bounded. We are now in a position to define the cross-covariance operator.

Definition 8 (*Cross-covariance operator*, Gretton et al. 2005) The cross-covariance operator $\boldsymbol{C}_{\boldsymbol{X},\boldsymbol{Y}} : \mathcal{G} \to \mathcal{H}$ associated with the joint probability measure $P_{\boldsymbol{X},\boldsymbol{Y}}$ on $(\mathbb{R}^p \times \mathbb{R}^q, \Gamma \times \Lambda)$ is a linear operator $\boldsymbol{C}_{\boldsymbol{X},\boldsymbol{Y}} : \mathcal{G} \to \mathcal{H}$ defined as

$$\boldsymbol{C}_{\boldsymbol{X},\boldsymbol{Y}} = \mathrm{E}_{\boldsymbol{X},\boldsymbol{Y}}[\phi(\boldsymbol{X}) \otimes \psi(\boldsymbol{Y})] - \mu_{\boldsymbol{X}} \otimes \mu_{\boldsymbol{Y}},$$

for all $f \in \mathcal{H}$ and $g \in \mathcal{G}$, where the tensor product operator $f \otimes g : \mathcal{G} \to \mathcal{H}, f \in \mathcal{H}, g \in \mathcal{G}$, is defined as

$$(f \otimes g)h = f\langle g, h\rangle_{\mathcal{G}}, \text{ for all } h \in \mathcal{G}.$$

This is a generalization of the cross-covariance matrix between random vectors. Moreover, by the definition of the Hilbert–Schmidt (HS) norm, we can compute the HS norm of $f \otimes g$ via

$$\|f \otimes g\|_{HS}^2 = \|f\|_{\mathcal{H}}^2 \|g\|_{\mathcal{G}}^2.$$

Definition 9 (*Hilbert–Schmidt Independence Criterion*, Gretton et al. 2005) Hilbert–Schmidt Independence Criterion (HSIC) is the squared Hilbert–Schmidt norm (or Frobenius norm) of the cross-covariance operator associated with the probability measure $P_{\boldsymbol{X},\boldsymbol{Y}}$ on $(\mathbb{R}^p \times \mathbb{R}^q, \Gamma \times \Lambda)$:

$$\mathrm{HSIC}(P_{\boldsymbol{X},\boldsymbol{Y}}) = \|\boldsymbol{C}_{\boldsymbol{X},\boldsymbol{Y}}\|_F^2.$$

To compute it we need to express HSIC in terms of kernel functions (Gretton et al. 2005):

$$\begin{aligned}\mathrm{HSIC}(P_{\boldsymbol{X},\boldsymbol{Y}}) &= \mathrm{E}_{\boldsymbol{X},\boldsymbol{X}',\boldsymbol{Y},\boldsymbol{Y}'}[k(\boldsymbol{X},\boldsymbol{X}')l(\boldsymbol{Y},\boldsymbol{Y}')] \\ &+ \mathrm{E}_{\boldsymbol{X},\boldsymbol{X}'}[k(\boldsymbol{X},\boldsymbol{X}')]\,\mathrm{E}_{\boldsymbol{Y},\boldsymbol{Y}'}[l(\boldsymbol{Y},\boldsymbol{Y}')] \\ &- 2\,\mathrm{E}_{\boldsymbol{X},\boldsymbol{Y}}[\mathrm{E}_{\boldsymbol{X}'}[k(\boldsymbol{X},\boldsymbol{X}')]\,\mathrm{E}_{\boldsymbol{Y}'}[l(\boldsymbol{Y},\boldsymbol{Y}')]]. \end{aligned} \tag{5}$$

Here $\mathrm{E}_{\boldsymbol{X},\boldsymbol{X}',\boldsymbol{Y},\boldsymbol{Y}'}$ denotes the expectation over independent pairs $(\boldsymbol{X},\boldsymbol{Y})$ and $(\boldsymbol{X}',\boldsymbol{Y}')$ distributed according to $P_{\boldsymbol{X},\boldsymbol{Y}}$.

It follows from (5) that the Frobenius norm of $\boldsymbol{C}_{\boldsymbol{X},\boldsymbol{Y}}$ exists when the various expectations over the kernels are bounded, which is true as long as the kernels k and l are bounded.

Definition 10 (*Empirical HSIC*, Gretton et al. 2005) Let $S = \{(\boldsymbol{x}_1,\boldsymbol{y}_1),\ldots,(\boldsymbol{x}_n,\boldsymbol{y}_n)\} \subseteq \mathbb{R}^p \times \mathbb{R}^q$ be a series of n independent observations drawn from $P_{\boldsymbol{X},\boldsymbol{Y}}$. An estimator of HSIC, written HSIC(S), is given by

$$\mathrm{HSIC}(S) = \frac{1}{n^2}\langle \boldsymbol{K}, \widetilde{\boldsymbol{L}}\rangle_F, \tag{6}$$

where $\boldsymbol{K} = (k(\boldsymbol{x}_i,\boldsymbol{x}_j))$, $\boldsymbol{L} = (l(\boldsymbol{y}_i,\boldsymbol{y}_j)) \in \mathbb{R}^{n\times n}$.

Comparing (4) and (6) and using (3), we see that the centered kernel target alignment (KTA) is simply a normalized version of HSIC(S).

In two seminar papers on Székely et al. (2007) and Székely and Rizzo (2009) introduced the distance covariance (dCov) and distance correlation (dCor) as powerful measures of dependence.

For column vectors $\boldsymbol{s} \in \mathbb{R}^p$ and $\boldsymbol{t} \in \mathbb{R}^q$, denote by $\|\boldsymbol{s}\|_p$ and $\|\boldsymbol{t}\|_q$ the standard Euclidean norms on the corresponding spaces. For jointly distributed random vectors $\boldsymbol{X} \in \mathbb{R}^p$ and $\boldsymbol{Y} \in \mathbb{R}^q$, let

$$f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{s},\boldsymbol{t}) = \mathrm{E}_{\boldsymbol{X},\boldsymbol{Y}}\{\exp[i\langle \boldsymbol{s},\boldsymbol{X}\rangle_p + i\langle \boldsymbol{t},\boldsymbol{Y}\rangle_q]\},$$

be the joint characteristic function of $(\boldsymbol{X}, \boldsymbol{Y})$, and let $f_{\boldsymbol{X}}(\boldsymbol{s}) = f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{s}, \boldsymbol{0})$ and $f_{\boldsymbol{Y}}(\boldsymbol{t}) = \varphi_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{0}, \boldsymbol{t})$ be the marginal characteristic functions of $\boldsymbol{X}$ and $\boldsymbol{Y}$, where $\boldsymbol{s} \in \mathbb{R}^p$ and $\boldsymbol{t} \in \mathbb{R}^q$. The distance covariance between $\boldsymbol{X}$ and $\boldsymbol{Y}$ is the nonnegative number $\nu(\boldsymbol{X}, \boldsymbol{Y})$ defined by

$$\nu^2(\boldsymbol{X}, \boldsymbol{Y}) = \frac{1}{C_p C_q} \int_{\mathbb{R}^{p+q}} \frac{|f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{s}, \boldsymbol{t}) - f_{\boldsymbol{X}}(\boldsymbol{s}) f_{\boldsymbol{Y}}(\boldsymbol{t})|^2}{\|\boldsymbol{s}\|_p^{p+1} \|\boldsymbol{t}\|_q^{q+1}} d\boldsymbol{s} d\boldsymbol{t},$$

and $|z|$ denotes the modulus of $z \in \mathbb{C}$ and

$$C_p = \frac{\pi^{\frac{1}{2}(p+1)}}{\Gamma(\frac{1}{2}(p+1))}.$$

The distance correlation between $\boldsymbol{X}$ and $\boldsymbol{Y}$ is the nonnegative number defined by

$$\mathcal{R}(\boldsymbol{X}, \boldsymbol{Y}) = \frac{\nu(\boldsymbol{X}, \boldsymbol{Y})}{\sqrt{\nu(\boldsymbol{X}, \boldsymbol{X}) \nu(\boldsymbol{Y}, \boldsymbol{Y})}}$$

if both $\nu(\boldsymbol{X}, \boldsymbol{X})$ and $\nu(\boldsymbol{Y}, \boldsymbol{Y})$ are strictly positive, and defined to be zero otherwise. For distributions with finite first moments, the distance correlation characterizes independence in that $0 \leq \mathcal{R}(\boldsymbol{X}, \boldsymbol{Y}) \leq 1$ with $\mathcal{R}(\boldsymbol{X}, \boldsymbol{Y}) = 0$ if and only if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent.

Sejdinovic et al. (2013) demonstrated that distance covariance is an instance of the Hilbert–Schmidt Independence Criterion. Górecki et al. (2016, 2017) showed an extension of the distance covariance and distance correlation coefficients to the functional case.

2.4 Kernel-based independence test

Statistical tests of independence have been associated with a broad variety of dependence measures. Classical tests such as Spearman's ρ and Kendall's τ are widely applied, however they are not guaranteed to detect all modes of dependence between the random variables. Contingency table-based methods, and in particular the power-divergence family of test statistics (Read and Cressie 1988) are the best known general purpose tests of independence, but are limited to relatively low dimensions, since they require a partitioning of the space in which random variable resides. Characteristic function-based tests (Feuerverger 1993; Kankainen 1995) have also been proposed. They are more general than kernel-based tests, although to our knowledge they have been used only to compare univariate random variables.

Now, we describe how HSIC can be used as an independence measure, and as the basis for an independence test. We begin by demonstrating that the Hilbert–Schmidt norm can be used as a measure of independence, as long as the associated RKHSs are universal.

A continuous kernel k on a compact metric space is called universal if the corresponding RKHS $\mathcal{H}$ is dense in the class of continuous functions of the space.

Denote by $\mathcal{H}$, $\mathcal{G}$ RKHSs with universal kernels k, l on the compact domains $\mathcal{X}$ and $\mathcal{Y}$ respectively. We assume without loss of generality that $\|f\|_\infty \leq 1$ and $\|g\|_\infty \leq 1$ for all $f \in \mathcal{H}$ and $g \in \mathcal{G}$. Then Gretton et al. (2005) proved that $\|\boldsymbol{C}_{\boldsymbol{X},\boldsymbol{Y}}\|_{HS} = 0$ if and only if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent. Examples of universal kernels are Gaussian kernel and Laplacian kernel, while the linear kernel $k(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x}^\top \boldsymbol{x}'$ is not universal—the corresponding HSIC tests only linear relationships, and a zero cross-covariance matrix characterizes independence only for multivariate Gaussian distributions. Working with the infinite dimensional operator with universal kernels, allows us to identify any general nonlinear dependence (in the limit) between any pair of vectors, not just Gaussians.

We recall that in this paper we use the Gaussian kernel. We now consider the asymptotic distribution of statistics (6).

We introduce the null hypothesis H_0: $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}$ ($\boldsymbol{X}$ is independent of $\boldsymbol{Y}$, i.e., $P_{\boldsymbol{X},\boldsymbol{Y}} = P_{\boldsymbol{X}} P_{\boldsymbol{Y}}$). Suppose that we are given the i.i.d. samples $S_{\boldsymbol{x}} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ and $S_{\boldsymbol{y}} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ for $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. Let $\widetilde{\boldsymbol{K}}$ and $\widetilde{\boldsymbol{L}}$ be the centered kernel matrices associated to the kernel function k and the sets $S_{\boldsymbol{x}}$ and $S_{\boldsymbol{y}}$, respectively. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$ be the eigenvalues of the matrix $\widetilde{\boldsymbol{K}}$ and let $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ be a set of orthonormal eigenvectors corresponding to these eigenvalues. Let $\lambda'_1 \geq \lambda'_2 \geq \cdots \geq \lambda'_n \geq 0$ be the eigenvalues of the matrix $\widetilde{\boldsymbol{L}}$ and let $\boldsymbol{v}'_1, \ldots, \boldsymbol{v}'_n$ be a set of orthonormal eigenvectors corresponding to these eigenvalues. Let $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, $\Lambda' = \operatorname{diag}(\lambda'_1, \ldots, \lambda'_n)$, $\boldsymbol{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n)$ and $\boldsymbol{V}' = (\boldsymbol{v}'_1, \ldots, \boldsymbol{v}'_n)$. Suppose further that we have the eigenvalue decomposition (EVD) of the centered kernel matrices $\widetilde{\boldsymbol{K}}$ and $\widetilde{\boldsymbol{L}}$, i.e., $\widetilde{\boldsymbol{K}} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\top$ and $\widetilde{\boldsymbol{L}} = \boldsymbol{V}'\boldsymbol{\Lambda}'(\boldsymbol{V}')^\top$.

Let $\boldsymbol{\Psi} = (\boldsymbol{\Psi}_1, \ldots, \boldsymbol{\Psi}_n) = \boldsymbol{V}\boldsymbol{\Lambda}^{1/2}$ and $\boldsymbol{\Psi}' = (\boldsymbol{\Psi}'_1, \ldots, \boldsymbol{\Psi}'_n) = \boldsymbol{V}'(\boldsymbol{\Lambda}')^{1/2}$, i.e., $\boldsymbol{\Psi}_i = \sqrt{\lambda_i}\boldsymbol{v}_i$, $\boldsymbol{\Psi}'_i = \sqrt{\lambda'_i}\boldsymbol{v}'_i$, $i = 1, \ldots, n$.

The following result is true (Zhang et al. 2011): under the null hypothesis that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent, the statistic (6) has the same asymptotic distribution as

$$Z_n = \frac{1}{n^2} \sum_{i,j=1}^{n} \lambda_{i,n}\lambda'_{j,n} Z_{ij}^2, \tag{7}$$

where Z_{ij}^2 are i.i.d. χ_1^2-distributed variables, $n \to \infty$.

Note that the data-based test statistic HSIC (or its probabilistic counterpart) is sensible to dependence/independence and therefore can be used as a test statistic. Also important is the knowledge of its asymptotic distribution. These facts inspire the following dependence/independence testing procedure. Given the sample $S_{\boldsymbol{x}}$ and $S_{\boldsymbol{y}}$, one first calculates the centered kernel matrices $\widetilde{\boldsymbol{K}}$ and $\widetilde{\boldsymbol{L}}$ and their eigenvalues λ_i and λ'_i, and then evaluates the statistic HSIC(S) according to (6). Next, the empirical null distribution of Z under the null hypothesis can be simulated in the following way: one draws i.i.d. random samples from the χ_1^2-distributed variables Z_{ij}^2, and then generates samples for Z according to (7). Finally the p value can be found by locating HSIC(S) in the simulated null distribution.

A permutation-based test is described in Gretton et al. (2005). In the first step they propose to calculate the test statistic T (HSIC or KTA) for the given data. Next, keeping the order of the first sample we randomly permute the second sample a large number of times, and recompute the selected statistic each time. This destroy any dependence between samples simulating a draw from the product of marginals, making the empirical distribution of the permuted statistics behave like the null distribution of the test statistic. For a specified significance level α, we calculate threshold t_α in the right tail of the null distribution. We reject H_0 if $T > t_\alpha$. This test was proved to be consistent against any fixed alternative. It means that for any fixed significance level α, the power goes to 1 as the sample size tends to infinity.

2.5 Functional data

In recent years methods for representing data by functions or curves have received much attention. Such data are known in the literature as the functional data (Ramsay and Silverman 2005; Horváth and Kokoszka 2012; Hsing and Eubank 2015). Examples of functional data can be found in various application domains, such as medicine, economics, meteorology and many others. Functional data can be seen as the values of random process $X(t)$. In practice, the

values of the observed random process $X(t)$ are always recorded at discrete times $t_1, \ldots, t_J$, less frequently or more densely spaced in the range of variability of the argument t. So we have a time series $\{x(t_1), \ldots, y(t_J)\}$. However, there are many reasons to model these series as elements of functional space., because the functional data has many advantages over other ways of representing the time series.

1. They easily cope with the problem of missing observations, an inevitable problem in many areas of research. Unfortunately, most data analysis methods require complete time series. One solution is to delete a time series that has missing values from the data, but this can lead to , and generally leads to, loss of information. Another option is to use one of many statistical methods to predict the missing values, but then the results will depend on the interpolation method. In contrast to this type of solutions, in the case of functional data, the problem of missing observations is solved by expressing time series in the form of a set of continuous functions.
2. The functional data naturally preserve the structure of observations, i.e. they maintain the time dependence of the observations and take into account the information about each measurement.
3. The moments of observations do not have to be evenly spaced in individual time series.
4. Functional data avoids the curse of dimensionality. When the number of time points is greater than the number of time series considered, most statistical methods will not give satisfactory results due to overparametrization. In the case of functional data, this problem can be avoided because the time series are replaced with a set of continuous functions independent of the number of time points in which observations are measured.

In most of the papers on functional data analysis, objects are characterized by only one feature observed at many time points. In several applications there is a need to use statistical methods for objects characterized by many features observed at many time points (double multivariate data). In this case, such data are transformed into multivariate functional data.

Let us assume that $\boldsymbol{X} = (X_1, \ldots, X_p)^\top = \{\boldsymbol{X}(s), s \in I_1\} \in L_2^p(I_1)$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_q)^\top = \{\boldsymbol{Y}(t), t \in I_2\} \in L_2^q(I_2)$ are random processes, where $L_2(I)$ is a space of square integrable functions on the interval I. We also assume that

$$\mathrm{E}(\boldsymbol{X}(s)) = \boldsymbol{0},\ s \in I_1, \quad \mathrm{E}(\boldsymbol{Y}(t)) = \boldsymbol{0},\ t \in I_2.$$

We will further assume that each component X_g of the random process $\boldsymbol{X}$ and Y_h of the random process $\boldsymbol{Y}$ can be represented by a finite number of orthonormal basis functions $\{\varphi_e\}$ and $\{\varphi_f\}$ of space $L_2(I_1)$ and $L_2(I_2)$, respectively:

$$X_g(s) = \sum_{e=0}^{E_g} \alpha_{ge}\varphi_e(s),\ s \in I_1,\ g = 1, 2, \ldots, p,$$

$$Y_h(t) = \sum_{f=0}^{F_h} \beta_{hf}\varphi_f(t),\ t \in I_2,\ h = 1, 2, \ldots, q,$$

where α_{ge} and β_{hf} are the random coefficients. The degree of smoothness of processes X_g and Y_h depends on the values E_g and F_h respectively (small values imply more smoothing). The optimum values for E_g and F_h are selected using Bayesian Information Criterion (BIC) (see Górecki et al. 2018). As basis functions we can use e.g. the Fourier basis system or spline functions.

We introduce the following notation:

$$\boldsymbol{\alpha} = (\alpha_{10}, ..., \alpha_{1E_1}, ..., \alpha_{p0}, ..., \alpha_{pE_p})^\top,$$

$$\boldsymbol{\beta} = (\beta_{10}, ..., \beta_{1F_1}, ..., \beta_{q0}, ..., \beta_{qF_q})^\top,$$
$$\boldsymbol{\varphi}_{E_g}(s) = (\varphi_0(s), \ldots, \varphi_{E_g}(s))^\top, s \in I_1, g = 1, 2, ..., p,$$
$$\boldsymbol{\varphi}_{F_h}(t) = (\varphi_0(t), \ldots, \varphi_{F_h}(t))^\top,$$
$$t \in I_2, h = 1, 2, ..., q, \boldsymbol{\Phi}_1(s) = \begin{bmatrix} \boldsymbol{\varphi}_{E_1}^\top(s) & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\varphi}_{E_2}^\top(s) & \ldots & \mathbf{0} \\ \ldots & \ldots & \ldots & \ldots \\ \mathbf{0} & \mathbf{0} & \ldots & \boldsymbol{\varphi}_{E_p}^\top(s) \end{bmatrix},$$
$$\boldsymbol{\Phi}_2(t) = \begin{bmatrix} \boldsymbol{\varphi}_{F_1}^\top(t) & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\varphi}_{F_2}^\top(t) & \ldots & \mathbf{0} \\ \ldots & \ldots & \ldots & \ldots \\ \mathbf{0} & \mathbf{0} & \ldots & \boldsymbol{\varphi}_{F_q}^\top(t) \end{bmatrix},$$

where $\boldsymbol{\alpha} \in \mathbb{R}^{K_1+p}$, $\boldsymbol{\beta} \in \mathbb{R}^{K_2+q}$, $\boldsymbol{\Phi}_1 \in \mathbb{R}^{p+(K_1+p)}$, $\boldsymbol{\Phi}_2 \in \mathbb{R}^{q+(K_2+q)}$, $K_1 = E_1 + \cdots + E_p$, $K_2 = F_1 + \cdots + F_p$.

Using the above matrix notation the random processes $\boldsymbol{X}$ and $\boldsymbol{Y}$ can be represented as:

$$\boldsymbol{X}(s) = \boldsymbol{\Phi}_1(s)\boldsymbol{\alpha},\ s \in I_1,\ \boldsymbol{Y}(t) = \boldsymbol{\Phi}_2(t)\boldsymbol{\beta},\ t \in I_2, \tag{8}$$

where $\mathrm{E}(\boldsymbol{\alpha}) = \mathbf{0}$, $\mathrm{E}(\boldsymbol{\beta}) = \mathbf{0}$.

This means that the values of random processes $\boldsymbol{X}$ and $\boldsymbol{Y}$ are in finite dimensional subspaces of $L_2^p(I_1)$ and $L_2^q(I_2)$, respectively. We will denote these subspaces by $\mathcal{L}_2^p(I_1)$ and $\mathcal{L}_2^q(I_2)$.

Typically data are recorded at discrete moments in time. The process of transformation of discrete data to functional data is performed for each realization and each variable separately. Let x_{gj} denote an observed value of the feature X_g, $g = 1, 2, \ldots p$ at the jth time point s_j, where $j = 1, 2, ..., J$. Similarly, let y_{hj} denote an observed value of feature $Y_h, h = 1, 2, \ldots q$ at the jth time point t_j, where $j = 1, 2, ..., J$. Then our data consist of pJ pairs of (s_j, x_{gj}) and of qJ pairs of (t_j, y_{hj}). Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ and $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ be independent trajectories of random processes $\boldsymbol{X}$ and $\boldsymbol{Y}$ having the representation (8).

The coefficients $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}_i$ are estimated by the least squares method. Let us denote these estimates by $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$, $i = 1, 2, \ldots, n$.

As a result, we obtain functional data of the form:

$$\boldsymbol{X}_i(s) = \boldsymbol{\Phi}_1(s)\boldsymbol{a}_i,\ \boldsymbol{Y}_i(t) = \boldsymbol{\Phi}_2(t)\boldsymbol{b}_i, \tag{9}$$

where $s \in I_1, t \in I_2, \boldsymbol{a}_i \in \mathbb{R}^{K_1+p}, \boldsymbol{b}_i \in \mathbb{R}^{K_2+q}$, $K_1 = E_1 + \cdots + E_p$, $K_2 = F_1 + \cdots + F_q$, and $i = 1, 2, \ldots, n$.

Górecki and Smaga (2017) described a multivariate analysis of variance (MANOVA) for functional data. In the paper by Górecki et al. (2018), three basic methods of dimension reduction for multidimensional functional data are given: principal component analysis, canonical correlation analysis, and discriminant coordinates.

3 Alignment for multivariate functional data

3.1 The alignment between two kernel functions and two kernel matrices for multivariate functional data

Let $\boldsymbol{x}(s) \in \mathcal{L}_2^p(I_1)$, $s \in I_1$, where $\mathcal{L}_2^p(I_1)$ is a finite-dimensional space of continuous square-integrable vector functions over interval I_1.

Let

$$k^\star \colon \mathcal{L}_2^p(I_1) \times \mathcal{L}_2^p(I_1) \to \mathbb{R}$$

be a kernel function on $\mathcal{L}_2^p(I_1)$. As already mentioned, in this paper we use the Gaussian kernel. For the multivariate functional data this kernel has the form:

$$k^\star(\boldsymbol{x}(s), \boldsymbol{x}'(s)) = \exp(-\lambda_1 \|\boldsymbol{x}(s) - \boldsymbol{x}'(s)\|^2), \ \lambda_1 > 0.$$

But from (9), and by the orthonormality of the basis functions, we have:

$$\begin{aligned}\|\boldsymbol{x}(s) - \boldsymbol{x}'(s)\|^2 &= \int_{I_1} (\boldsymbol{x}(s) - \boldsymbol{x}'(s))^\top (\boldsymbol{x}(s) - \boldsymbol{x}'(s)) ds \\ &= \|\boldsymbol{a} - \boldsymbol{a}'\|^2.\end{aligned}$$

Hence

$$k^\star(\boldsymbol{x}(s), \boldsymbol{x}'(s)) = k(\boldsymbol{a}, \boldsymbol{a}')$$

and

$$k^\star(\boldsymbol{y}(t), \boldsymbol{y}'(t)) = k(\boldsymbol{b}, \boldsymbol{b}'),$$

where $\boldsymbol{a}$ and $\boldsymbol{b}$ are vectors occurring in the representation (9) of vector functions $\boldsymbol{x}(s)$, $s \in I_1$, $\boldsymbol{y}(t)$, $t \in I_2$.

For a given subset $\{\boldsymbol{x}_1(s), \ldots, \boldsymbol{x}_n(s)\}$ of $\mathcal{L}_2^p(I_1)$ and the given kernel function $k^\star$ on $\mathcal{L}_2^p(I_1) \times \mathcal{L}_2^p(I_1)$, the matrix $\boldsymbol{K}^\star$ of size $n \times n$, which has its (i, j)th element $K_{ij}^\star(s)$, given by $K_{ij}^\star(s) = k^\star(\boldsymbol{x}_i(s), \boldsymbol{x}_j(s))$, $s \in I_1$, is called the kernel matrix of the kernel function $k^\star$ with respect to the set $\{\boldsymbol{x}_1(s), \ldots, \boldsymbol{x}_n(s)\}$, $s \in I_1$.

Definition 11 (*Kernel function alignment for functional data*) Let $\tilde{k}^\star$ and $\tilde{l}^\star$ be two kernel functions defined over $\mathcal{L}_2^p(I_1) \times \mathcal{L}_2^p(I_1)$ and $\mathcal{L}_2^q(I_2) \times \mathcal{L}_2^q(I_2)$, respectively, such that $0 < \mathrm{E}_{\boldsymbol{X},\boldsymbol{X}'}[\tilde{k}^{\star 2}(\boldsymbol{X}, \boldsymbol{X}')] < \infty$ and $0 < \mathrm{E}_{\boldsymbol{Y},\boldsymbol{Y}'}[\tilde{l}^{\star 2}(\boldsymbol{Y}, \boldsymbol{Y}')] < \infty$, where $\boldsymbol{X}, \boldsymbol{X}'$ and $\boldsymbol{Y}, \boldsymbol{Y}'$ are independent copies distributed according to $P_{\boldsymbol{X}}$ and $P_{\boldsymbol{Y}}$, respectively. Then the alignment between $\tilde{k}^\star$ and $\tilde{l}^\star$ is defined by

$$\rho(\tilde{k}^\star, \tilde{l}^\star) = \frac{\mathrm{E}_{\boldsymbol{X},\boldsymbol{Y}}[\tilde{k}^\star(\boldsymbol{X}, \boldsymbol{X}')\tilde{l}^\star(\boldsymbol{Y}, \boldsymbol{Y}')]}{\sqrt{\mathrm{E}_{\boldsymbol{X}}[\tilde{k}^{\star 2}(\boldsymbol{X}, \boldsymbol{X}')]\,\mathrm{E}_{\boldsymbol{Y}}[\tilde{l}^{\star 2}(\boldsymbol{Y}, \boldsymbol{Y}')]}}. \tag{10}$$

We can define similarly the alignment between two kernel matrices $\widetilde{\boldsymbol{K}}^\star$ and $\widetilde{\boldsymbol{L}}^\star$ based on the subset $\{\boldsymbol{x}_1(s), \ldots, \boldsymbol{x}_n(s)\}$, $s \in I_1$, and $\{\boldsymbol{y}_1(t), \ldots, \boldsymbol{y}_n(t)\}$, $t \in I_2$, of $\mathcal{L}_2^p(I_1)$ and $\mathcal{L}_2^q(I_2)$, respectively.

Definition 12 (*Kernel matrix alignment for functional data*) Let $\widetilde{\boldsymbol{K}}^\star \in \mathbb{R}^{n \times n}$ and $\widetilde{\boldsymbol{L}}^\star \in \mathbb{R}^{n \times n}$ be two kernel matrices such that $\|\widetilde{\boldsymbol{K}}^\star\|_F \neq 0$ and $\|\widetilde{\boldsymbol{L}}^\star\|_F \neq 0$. Then, the centered kernel target alignment (KTA) between $\widetilde{\boldsymbol{K}}^\star$ and $\widetilde{\boldsymbol{L}}^\star$ is defined:

$$\hat{\rho}(\boldsymbol{K}^\star, \boldsymbol{L}^\star) = \frac{\langle \widetilde{\boldsymbol{K}}^\star, \widetilde{\boldsymbol{L}}^\star \rangle_F}{\|\widetilde{\boldsymbol{K}}^\star\|_F \|\widetilde{\boldsymbol{L}}^\star\|_F}. \tag{11}$$

If $\widetilde{\boldsymbol{K}}^\star$ and $\widetilde{\boldsymbol{L}}^\star$ are positive semi-definite matrices, then $\hat{\rho}(\boldsymbol{K}^\star, \boldsymbol{L}^\star) \in [0, 1]$. We have

$$\hat{\rho}(\boldsymbol{K}^\star, \boldsymbol{L}^\star) = \hat{\rho}(\boldsymbol{K}, \boldsymbol{L}),$$

where $\boldsymbol{K}$ is the matrix of size $n \times n$, which has its (i, j)th element K_{ij}, given by $K_{ij} = k(\boldsymbol{a}_i, \boldsymbol{a}_j)$.

3.2 Kernel-based independence test for multivariate functional data

Definition 13 (*Empirical HSIC for functional data*) The empirical HSIC for functional data is defined as

$$\mathrm{HSIC}(S^\star) = \frac{1}{n^2}\langle \boldsymbol{K}^\star, \boldsymbol{L}^\star \rangle_F,$$

where $S^\star = \{(\boldsymbol{x}_1(s), \boldsymbol{y}_1(t)), \ldots, (\boldsymbol{x}_n(s), \boldsymbol{y}_n(t))\},\ s \in I_1,\ t \in I_2$, $\boldsymbol{K}^\star$ and $\boldsymbol{L}^\star$ are kernel matrices based on the subsets $\{\boldsymbol{x}_1(s), \ldots, \boldsymbol{x}_n(s)\},\ s \in I_1$, and $\{\boldsymbol{y}_1(t), \ldots, \boldsymbol{y}_n(t)\},\ t \in I_2$ of $\mathcal{L}_2^p(I_1)$ and $\mathcal{L}_2^q(I_2)$, respectively.

But $\boldsymbol{K}^\star = \boldsymbol{K}$, where $\boldsymbol{K}$ is the kernel matrix of size $n \times n$, which has its (i, j)th element K_{ij} given by $K_{ij} = k(\boldsymbol{a}_i, \boldsymbol{a}_j)$, where $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ are vectors occurring in the representation (9) vector functions $\boldsymbol{X}(s),\ s \in I_1$. Analogously, $\boldsymbol{L}^\star = \boldsymbol{L}$, where $\boldsymbol{L}$ is the kernel matrix of size $n \times n$, which has its (i, j)th element L_{ij} given by $L_{ij} = l(\boldsymbol{b}_i, \boldsymbol{b}_j)$, where $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n$ are vectors occurring in the representation (9) vector functions $\boldsymbol{Y}(t),\ t \in I_2$. Hence

$$\mathrm{HSIC}(S^\star) = \mathrm{HSIC}(S_v),$$

where $S_v = \{(\boldsymbol{a}_1, \boldsymbol{b}_1), \ldots, (\boldsymbol{a}_n, \boldsymbol{b}_n)\}$.

Note also that the null hypothesis $H_0 \colon \boldsymbol{X} \perp \boldsymbol{Y}$ of independence of the random processes $\boldsymbol{X}$ and $\boldsymbol{Y}$ is equivalent to the null hypothesis $H_0 \colon \boldsymbol{\alpha} \perp \boldsymbol{\beta}$ of independence of random vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ occurring in the representation (8) random processes $\boldsymbol{X}$ and $\boldsymbol{Y}$. We can therefore use the tests described in Section 2.4, replacing $\boldsymbol{x}$ and $\boldsymbol{y}$ by $\boldsymbol{a}$ and $\boldsymbol{b}$.

3.3 Canonical correlation analysis based on the alignment between kernel matrices for multivariate functional data

In classical canonical correlation analysis (Hotelling 1936), we are interested in the relationship between two random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$. In the functional case we are interested in the relationship between two random functions $\boldsymbol{X}$ and $\boldsymbol{Y}$. Functional canonical variables U and V for random processes $\boldsymbol{X}$ and $\boldsymbol{Y}$ are defined as follows

$$U = \langle \boldsymbol{u}, \boldsymbol{X} \rangle = \int_{I_1} \boldsymbol{u}^\top(s)\boldsymbol{X}(s)ds, \tag{12}$$

$$V = \langle \boldsymbol{v}, \boldsymbol{Y} \rangle = \int_{I_2} \boldsymbol{v}^\top(t)\boldsymbol{Y}(t)dt, \tag{13}$$

where the vector functions $\boldsymbol{u}$ and $\boldsymbol{v}$ are called the vector weight functions and are of the form

$$\boldsymbol{u}(s) = \boldsymbol{\Phi}_1(s)\boldsymbol{u},\ \boldsymbol{v}(t) = \boldsymbol{\Phi}_2(t)\boldsymbol{v} \text{ where } \boldsymbol{u} \in \mathbb{R}^{K_1+p}, \boldsymbol{v} \in \mathbb{R}^{K_2+q}. \tag{14}$$

Classically the weight functions $\boldsymbol{u}$ and $\boldsymbol{v}$ are chosen to maximize the sample correlation coefficient (Górecki et al. 2018):

$$\rho = \frac{\mathrm{Cov}(U, V)}{\sqrt{\mathrm{Var}(U)\,\mathrm{Var}(V)}}. \tag{15}$$

The sample correlation coefficient between the variables U and V is now replaced by a centered kernel target alignment (KTA) between kernel matrices $\boldsymbol{K}$ and $\boldsymbol{L}$ based on the projected data $\langle \boldsymbol{u}(s), \boldsymbol{x}_i(s) \rangle_{\mathcal{H}}$ and $\langle \boldsymbol{v}(t), \boldsymbol{y}_i(t) \rangle_{\mathcal{H}}$, i.e. their (i, j)th entry are

$$K_{i,j} = k(\langle \boldsymbol{u}(s), \boldsymbol{x}_i(s)\rangle_{\mathcal{H}}, \langle \boldsymbol{u}(s), \boldsymbol{x}_j(s)\rangle_{\mathcal{H}}), s \in I_1,$$

and

$$L_{i,j} = l(\langle \boldsymbol{v}(t), \boldsymbol{y}_i(t)\rangle_{\mathcal{H}}, \langle \boldsymbol{v}(t), \boldsymbol{y}_j(t)\rangle_{\mathcal{H}}), t \in I_2,$$

respectively, $i, j = 1, \ldots, n$:

$$\hat{\rho}(\boldsymbol{u}(s), \boldsymbol{v}(t)) = \frac{\mathrm{tr}(\widetilde{\boldsymbol{K}}^\top \widetilde{\boldsymbol{L}})}{\sqrt{\mathrm{tr}(\widetilde{\boldsymbol{K}}^\top \widetilde{\boldsymbol{K}})\,\mathrm{tr}(\widetilde{\boldsymbol{L}}^\top \widetilde{\boldsymbol{L}})}} \tag{16}$$

subject to

$$\|\boldsymbol{u}(s)\| = \|\boldsymbol{v}(t)\| = 1. \tag{17}$$

But

$$\begin{aligned}\boldsymbol{K}_{i,j} &= k\left(\int_{I_1} \boldsymbol{u}^\top(s)\boldsymbol{x}_i(s)ds, \int_{I_1} \boldsymbol{u}^\top(s)\boldsymbol{x}_j(s)ds\right)\\ &= k(\boldsymbol{u}^\top \boldsymbol{a}_i, \boldsymbol{u}^\top \boldsymbol{a}_j) = \boldsymbol{K}^{(\boldsymbol{u})}_{i,j}\end{aligned}$$

and

$$\begin{aligned}\boldsymbol{L}_{i,j} &= l\left(\int_{I_2} \boldsymbol{v}^\top(t)\boldsymbol{y}_i(t)dt, \int_{I_2} \boldsymbol{v}^\top(t)\boldsymbol{y}_j(t)dt\right)\\ &= l(\boldsymbol{v}^\top \boldsymbol{b}_i, \boldsymbol{v}^\top \boldsymbol{b}_j) = \boldsymbol{L}^{(\boldsymbol{v})}_{i,j},\end{aligned}$$

where $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$ are vectors occuring in the representation (9) vectors functions $\boldsymbol{x}(s), s \in I_1$, $\boldsymbol{y}(t), t \in I_2, i = 1, \ldots, n, \boldsymbol{u} \in \mathbb{R}^{K_1+p}, \boldsymbol{v} \in \mathbb{R}^{K_2+q}$.

Thus, the choice of weighting functions $\boldsymbol{u}(s)$ and $\boldsymbol{v}(t)$ so that the coefficient (16) has a maximum value subject to (17) is equivalent to the choice of vectors $\boldsymbol{u} \in \mathbb{R}^{K_1+p}$ and $\boldsymbol{v} \in \mathbb{R}^{K_2+q}$ such that the coefficient

$$\hat{\rho}(\boldsymbol{u}, \boldsymbol{v}) = \frac{\mathrm{tr}(\widetilde{\boldsymbol{K}}_{\boldsymbol{u}}^\top \widetilde{\boldsymbol{L}}_{\boldsymbol{v}})}{\sqrt{\mathrm{tr}(\widetilde{\boldsymbol{K}}_{\boldsymbol{u}}^\top \widetilde{\boldsymbol{K}}_{\boldsymbol{u}})\,\mathrm{tr}(\widetilde{\boldsymbol{L}}_{\boldsymbol{v}}^\top \widetilde{\boldsymbol{L}}_{\boldsymbol{v}})}} \tag{18}$$

has a maximum value subject to

$$\|\boldsymbol{u}\| = \|\boldsymbol{v}\| = 1, \tag{19}$$

where $\boldsymbol{K}_{\boldsymbol{u}} = (K^{(\boldsymbol{u})}_{i,j})$, $\boldsymbol{L}_{\boldsymbol{v}} = (L^{(\boldsymbol{v})}_{i,j})$, $i, j = 1, \ldots, n$.

In order to maximize the coefficient of (18) we can use the result of Chang et al. (2013). Authors used a gradient descent algorithm, with modified gradient to ensure the unit length constraint is satisfied at each step (Edelman et al. 1998). Optimal step-sizes were found numerically using the Nelder-Mead method. This article employs the Gaussian kernel exclusively while other kernels are available. The bandwidth parameter λ of the Gaussian kernel was chosen using the "median trick" (Song et al. 2010), i.e. the median Euclidean distance between all pairs of points.

The coefficients of the projection of the ith value $\boldsymbol{x}_i(t)$ of random process $\boldsymbol{X}$ on the kth functional canonical variable are equal to

$$\hat{U}_{ik} = \langle \boldsymbol{u}_k, \boldsymbol{x}_i\rangle = \int_{I_1} \boldsymbol{u}_k^\top(s)\boldsymbol{x}_i(s)ds = \boldsymbol{a}_i^\top \boldsymbol{u}_k,$$

analogously the coefficients of the projection of the ith value $\boldsymbol{y}_i(t)$ of random process $\boldsymbol{Y}_t$ on the kth functional canonical variable are equal to

$$\hat{V}_{ik} = \boldsymbol{b}_i^\top \boldsymbol{v}_k,$$

where $i = 1, \ldots, n$, $k = 1, \ldots, \min(\text{rank}(\boldsymbol{A}), \text{rank}(\boldsymbol{B}))$, where $\boldsymbol{A} \in \mathbb{R}^{n \times p}$ and $\boldsymbol{B} \in \mathbb{R}^{n \times q}$, where the ith rows are $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$, respectively, which have column means of zero.

As we mentioned earlier KTA is a normalized variant of HSIC. Hence, we can repeat the above reasoning for HSIC criterion. However, we should remember that both approaches are not equivalent and we can obtain different results.

4 Experiments

Let us recall some and introduce another symbols:

- KTA—centered kernel target alignment,
- HSIC—Hilbert–Schmidt Independence Criterion,
- FCCA—classical functional canonical correlation analysis (Ramsay and Silverman 2005; Horváth and Kokoszka 2012),
- HSIC.FCCA—functional canonical correlation analysis based on HSIC,
- HSIC.KTA—functional canonical correlation analysis based on KTA.

4.1 Simulation

We generated random processes along with some noises to test the performance of the introduced measures. Random processes are specified by

$$\begin{aligned} X_t &= \varepsilon_t, \\ Y_t &= 3X_t + \eta_t, \\ Z_t &= X_t^2 + \xi_t, \end{aligned}$$

where ε_t, η_t and ξ_t are jointly independent random variables from Gaussian distribution with 0 mean and 0.25 variance. We generated processes of length 100. $N = 10000$ samples are generated for all processes. The objective is to examine how well functional variants (Fourier basis with 15 basis functions) of KTA and HSIC measures perform compared to measures used on raw data (artificially generated at discrete time stamps). Here, raw data are represented as vector data of generated trajectories, so raw data are three 10000 by 100 dimensional matrices(one for X_t, second for Y_t and third for Z_t). On the other hand, functional data are three 10000 by 15 dimensional matrices (coefficients of Fourier basis). Here, X_t and Y_t are linearly dependent, whereas X_t, Z_t and Y_t, Z_t are nonlinearly dependent (Fig. 1).

From Fig. 2 and Table 1, we see that the proposed extension of HSIC and KTA coefficients to functional data gives larger values of coefficients than the variants for raw time series. Unfortunately, it is not possible to perform inference based only on the values of coefficients. We have to apply tests. In Fig. 3 and in Table 2, we observe that when we use functional variants of the proposed measures, we obtain much better results in recognizing nonlinear dependence. Linear dependence between X_t and Y_t was easily recognized by each method (100% of correct decisions—p values below 5%). Results of functional KTA and HSIC are very similar. Non-functional measures HSIC and KTA give only 7.2% and 6.7% correct decisions (p values below 5%) for relationship X_t, Z_t and Y_t, Z_t, respectively. On the

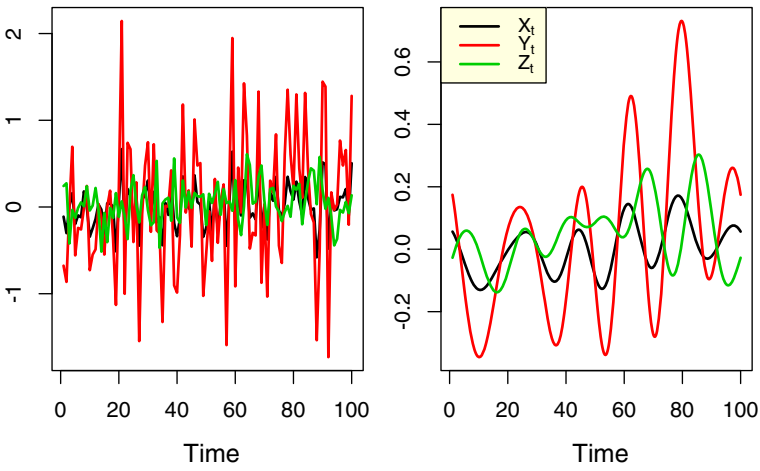

Fig. 1 Sample trajectories of X_t, Y_t and Z_t time series for raw (left plot) and functional (right plot) representation

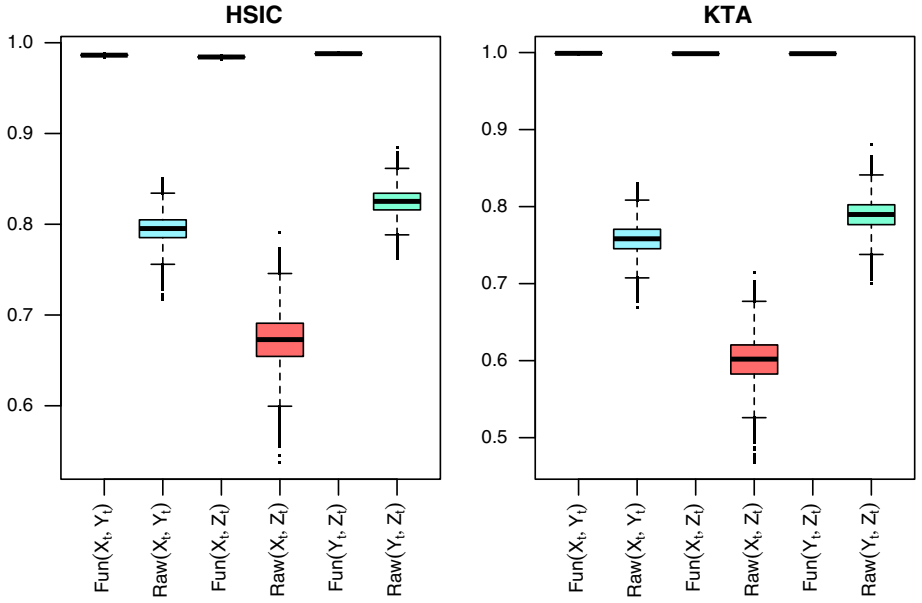

Fig. 2 Raw and functional HSIC and KTA coefficients for artificial time series

other hand, functional variants recognize dependency (p values below 5%) in 63.3% (both measures) for X_t, Y_t and in 47.8% (HSIC), 63.3% (KTA) for Y_t, Z_t.

4.2 Univariate example

As a first real example we used average daily temperature (in Celsius degrees) for each day of the year and average daily rainfall (in mm) for each day of the year rounded to 0.1 mm at

Table 1 Average raw and functional HSIC and KTA coefficients for artificial time series (number in brackets means standard deviation)

	(X_t, Y_t)	(X_t, Z_t)	(Y_t, Z_t)
Raw			
HSIC	0.795 (0.015)	0.672 (0.027)	0.825 (0.014)
KTA	0.758 (0.019)	0.601 (0.028)	0.789 (0.019)
Functional			
HSIC	0.986 (0.000)	0.984 (0.001)	0.988 (0.000)
KTA	0.999 (0.000)	0.999 (0.000)	0.999 (0.000)

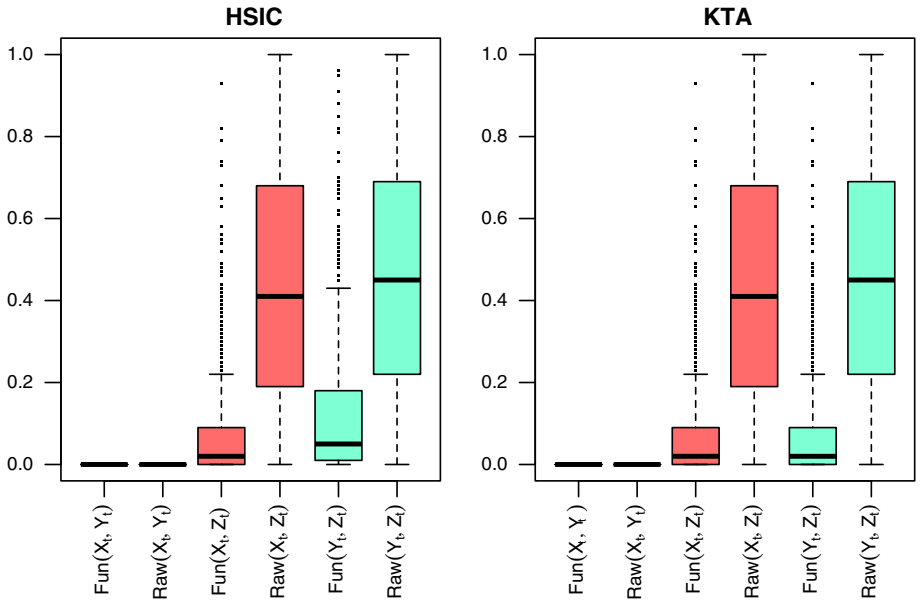

Fig. 3 p Values from permutation-based tests for raw and functional variants of HSIC and KTA coefficients

Table 2 Average p values from permutation-based tests for raw and functional variants of HSIC and KTA coefficients (number in brackets means standard deviation)

	(X_t, Y_t)	(X_t, Z_t)	(Y_t, Z_t)
Raw			
HSIC	0.000 (0.000)	0.445 (0.290)	0.458 (0.282)
KTA	0.000 (0.000)	0.445 (0.290)	0.458 (0.282)
Functional			
HSIC	0.000 (0.000)	0.077 (0.129)	0.125 (0.169)
KTA	0.000 (0.000)	0.077 (0.129)	0.077 (0.129)

35 different weather stations in Canada from 1960 to 1994. Each station belongs to one of four climate zone: Arctic (3 stations—blue color on plots), Atlantic (15—red color on plots), Continental (12— black color on plots) or Pacific (5—green color on plots) zone (Fig. 4). This data set comes from Ramsay and Silverman (2005).

In the first step, we smoothed data. We used the Fourier basis with various values of the smoothing parameter (number of basis functions) from 3 to 15. We can observe the effect

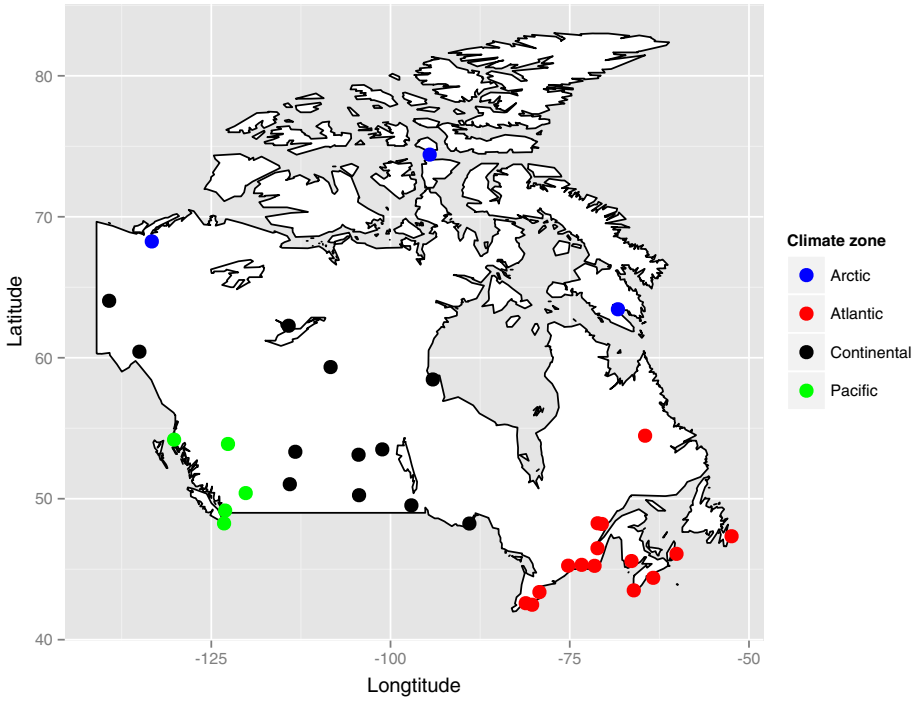

Fig. 4 Location of Canadian weather stations

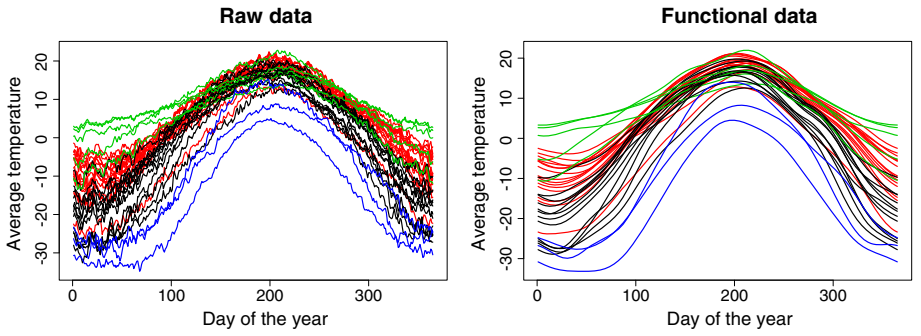

Fig. 5 Raw and functional temperature for Canadian weather stations

of smoothing in Figs. 5 and 6 (for Fourier basis with 15 basis functions). We decided to use the Fourier basis for two reasons: it has excellent computational properties, especially if the observations are equally spaced, and it is natural for describing periodic data, such as the annual weather cycles. Here, raw data are two 35 by 100 dimensional matrices (one for temperature and second for precipitation). On the other hand, functional data are two 35 by 15 dimensional matrices (coefficients of Fourier basis).

From the plots we can observe that the level of smoothness seems big enough. Additionally, we can observe some relationship between average temperature and precipitation. Namely, for weather stations with large average temperature, we observe relatively bigger average precipitation while for Arctic stations with lowest average temperatures we observe

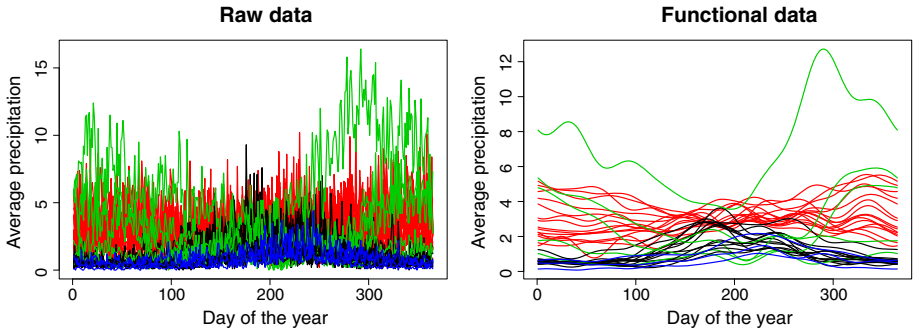

Fig. 6 Raw and functional precipitation for Canadian weather stations

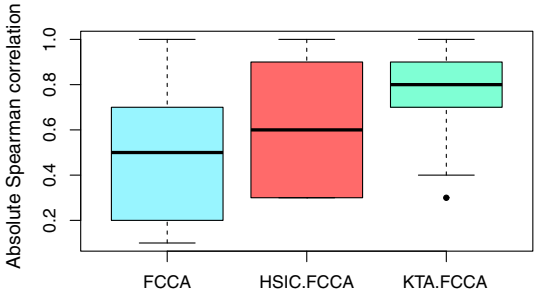

Fig. 7 Absolute Spearman correlation coefficient for the first set of functional canonical variables

the smallest average precipitation. So we can expect some relationship between average temperature and average precipitation for Canadian weather stations.

In the next step, we calculated the values of described earlier coefficients, the values of which are presented in Fig. 8. We observe quite big values of HSIC and KTA, but it is impossible to infer dependency from these values. We see that the values of HSIC and KTA coefficients are stable (both do not depend on basis size).

To statistically confirm the association between temperature and precipitation we performed some simulation study. This study based on Chang et al. (2013) simulation. Finding a good nonlinear dependency measure is not trivial. KTA and HSIC are not on the same scale. As Chang et al. (2013) we used Spearman correlation coefficient. We performed 50 random splits with the inclusion of 25 samples to identify models. The Spearman correlation coefficient was then calculated using the remaining 10 samples for each of 50 splits. As we know the strongest signal between the temperature and precipitation for Canadian weather stations is nonlinear (Chang et al. 2013). From Fig. 7 we can observe that HSIC.FCCA and KTA.FCCA have produced larger absolute Spearman coefficients than FCCA. Such results suggest that HSIC.FCCA & KTA.FCCA can be viewed as natural nonlinear extensions of CCA also in the case of multivariate functional data.

Finally, we performed permutation-based tests for HSIC and KTA coefficients. The results are presented in Fig. 8. All tests rejected H_0 (p values close to 0) for all basis sizes, so we can infer that we have some relationship between average temperature and average precipitation for Canadian weather stations. Unfortunately, we know nothing about the strength and direction of the dependency. Only a visual inspection of the plots suggests that there is a strong and positive relationship.

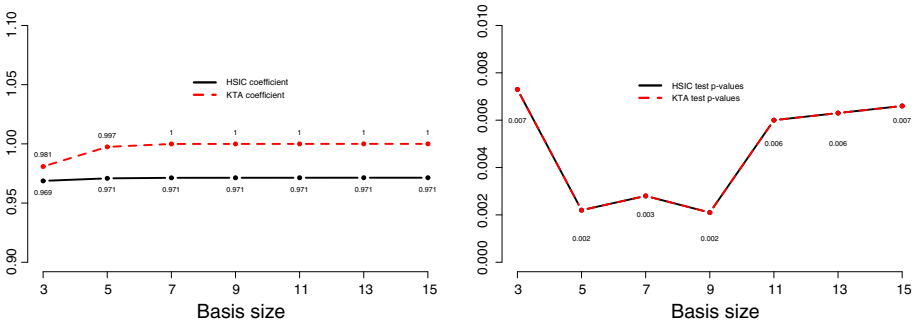

Fig. 8 HSIC and KTA coefficients and p values of permutation-based tests for Canadian weather data

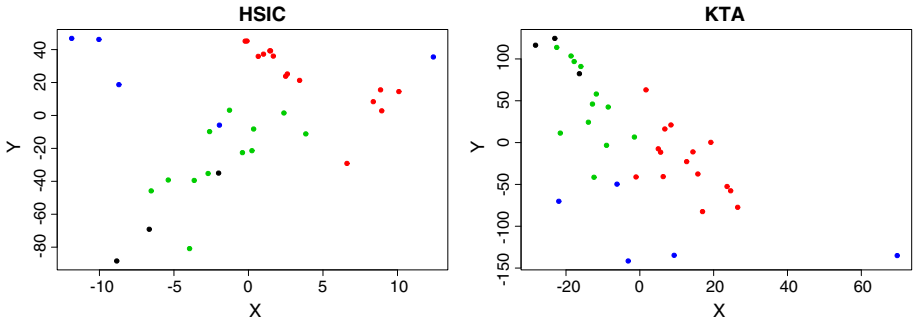

Fig. 9 Projection of the 35 Canadian weather stations on the plane $(\hat{U}_1, \hat{V}_1)$

The relative positions of the 35 Canadian weather stations in the system $(\hat{U}_1, \hat{V}_1)$ of functional canonical variables are shown in Fig. 9. It seems that for both coefficients the weather stations group reasonably.

4.3 Multivariate example

The described method was employed here to cluster the twelve groups (pillars) of variables of 38 European countries in the period 2008-2015. The list of countries used in the dependency analysis is contained in Table 3. Table 4 describes the pillars used in the analysis. For this purpose, use was made of data published by the World Economic Forum (WEF) in its annual reports (http://www.weforum.org). Those are comprehensive data, describing exhaustively various socio-economic conditions or spheres of individual states (Górecki et al. 2016). The data were transformed into functional data. Calculations were performed using the Fourier basis. In view of a small number of time periods ($J = 7$), for each variable the maximum number of basis components was taken to be equal to five. Here, raw data are twelve matrices (one for each pillar). Dimensions of matrices are different and depend on the number of variables in the pillar. Eg. for the first pillar we have 16 (number of variables) * 7 (number of time points) = 112 columns, hence dimensionality of the matrix for this pillar is 38 by 112. Similarly for the others. On the other hand, functional data are twelve matrices with 38 rows and appropriate number of columns (coefficients of Fourier basis). Number of columns for functional data eg. for the first pillar we calculate as 16 (number of variables) * 5 (number of basis elements) = 80.

Table 3 Countries used in analysis, 2008–2015

1	Albania (AL)	14	Greece (GR)	27	Poland (PL)
2	Austria (AT)	15	Hungary (HU)	28	Portugal (PT)
3	Belgium (BE)	16	Iceland (IS)	29	Romania (RO)
4	Bosnia and Herzegovina (BA)	17	Ireland (IE)	30	Russian Federation (RU)
5	Bulgaria (BG)	18	Italy (IT)	31	Serbia (XS)
6	Croatia (HR)	19	Latvia (LV)	32	Slovak Republic (SK)
7	Cyprus (CY)	20	Lithuania(LT)	33	Slovenia (SI)
8	Czech Republic (CZ)	21	Luxembourg (LU)	34	Spain (ES)
9	Denmark (DK)	22	Macedonia FYR (MK)	35	Sweden (SE)
10	Estonia (EE)	23	Malta (MT)	36	Switzerland (CH)
11	Finland (FI)	24	Montenegro (ME)	37	Ukraine (UA)
12	France (FR)	25	Netherlands (NL)	38	United Kingdom (GB)
13	Germany (DE)	26	Norway (NO)		

Table 4 Pillars used in analysis, 2008–2015

	Pillar	Number of variables
G1	Institutions	16
G2	Infrastructure	6
G3	Macroeconomic environment	2
G4	Health and primary education	7
G5	Higher education and training	6
G6	Goods market efficiency	10
G7	Labor market efficiency	6
G8	Financial market development	5
G9	Technological readiness	4
G10	Market size	4
G11	Business sophistication	9
G12	Innovation	5

Tables 5 and 6 contain the values of functional HSIC and KTA coefficients. As expected, they are all close to one. But high values of these coefficients do not necessarily mean that there is a significant relationship between the two groups of variables. We can expect association between groups of pillars. However, it is really hard to guess what groups are associated.

Similarly to the Canadian weather example we performed small simulation study for pillars G5 and G6. From Fig. 10 we can observe that HSIC.FCCA and KTA.FCCA have produced larger absolute Spearman coefficients than FCCA. This result suggest that proposed measures have better characteristic in discovering nonlinear relationship for this example.

We performed permutation-based tests for the HSIC and KTA coefficients discussed above. For most of tests, p values were close to zero, on the basis of which it can be inferred that there is some significant relationship between the groups (pillars) of variables. Table 7 contains the p values obtained for each test. We have exactly the same p values for both methods. Now,

Table 5 Functional HSIC coefficients

	1	2	3	4	5	6	7	8	9	10	11
2	0.9736										
3	0.9736	0.9737									
4	0.9736	0.9737	0.9737								
5	0.9708	0.9706	0.9706	0.9706							
6	0.9728	0.9727	0.9727	0.9727	0.9753						
7	0.9687	0.9683	0.9683	0.9683	0.9799	0.9780					
8	0.9730	0.9730	0.9730	0.9730	0.9725	0.9740	0.9721				
9	0.9736	0.9737	0.9737	0.9737	0.9706	0.9727	0.9683	0.9730			
10	0.9736	0.9737	0.9737	0.9737	0.9706	0.9727	0.9683	0.9730	0.9737		
11	0.9714	0.9711	0.9711	0.9711	0.9785	0.9755	0.9828	0.9726	0.9711	0.9711	
12	0.9688	0.9683	0.9683	0.9683	0.9778	0.9741	0.9897	0.9715	0.9783	0.9683	0.9830

Table 6 Functional KTA coefficients

	1	2	3	4	5	6	7	8	9	10	11
2	1.0000										
3	1.0000	1.0000									
4	1.0000	1.0000	1.0000								
5	0.9918	0.9916	0.9916	0.9916							
6	0.9980	0.9978	0.9978	0.9978	0.9951						
7	0.9741	0.9736	0.9736	0.9936	0.9801	0.9821					
8	0.9991	0.9990	0.9990	0.9990	0.9933	0.9989	0.9772				
9	1.0000	1.0000	1.0000	1.0000	0.9916	0.9978	0.9736	0.9990			
10	1.0000	1.0000	1.0000	1.0000	0.9916	0.9978	0.9736	0.9990	1.0000		
11	0.9927	0.9924	0.9924	0.9924	0.9947	0.9957	0.9833	0.9936	0.9924	0.9924	
12	0.9793	0.9788	0.9788	0.9788	0.9831	0.9834	0.9794	0.9917	0.9788	0.9788	0.9887

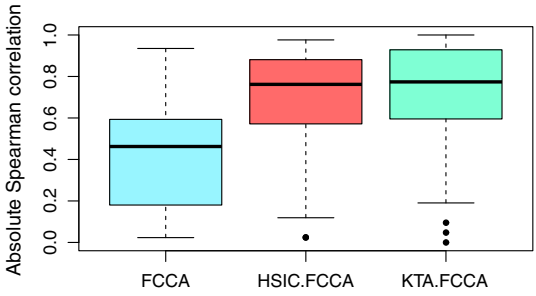

Fig. 10 Absolute Spearman correlation coefficient for the first set of functional canonical variables for pillars G5 & G6

we can observe that some groups are independent ($\alpha = 0.05$): G1 & G3, G3 & G6, G3 & G8, G3 & G11, G3 & G12, G4 & G9.

The graphs of the components of the vector weight function for the first functional canonical variables of the processes are shown in Fig. 11. From Fig. 11 (left) it can be seen that the greatest contribution in the structure of the first functional canonical correlation (U_1) comes

Table 7 Functional HSIC & KTA p values permutation-based tests (only non-zero)

	1	2	3	4	5	6	7	8	9	10	11
2	0.0142										
3	**0.0714**	0.0332									
4		0.0042	0.0343								
5		0.0001	0.0268								
6		0.0157	**0.0772**								
7		0.0009	0.0061								
8		0.0294	**0.0636**								
9	0.0030	0.0055	0.0198	**0.0640**	0.0002	0.0003	0.0009	0.0040			
10		0.0059	0.0294	0.0021					0.0055		
11		0.0039	**0.1034**						0.0008		
12		0.0008	**0.0563**						0.0044		

p-values greater than usual level of significance 5% are given in bold

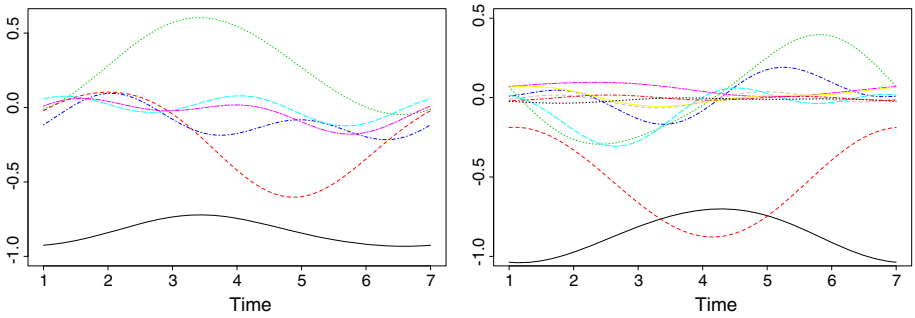

Fig. 11 Weight functions for first functional canonical variable U_1 (left) and V_1 (right)

from "black" process, and this holds for all of the observation years considered. Figure 11 (right) shows that, on specified time intervals, the greatest contribution in the structure of the first second functional canonical correlation (V_1 comes alternately from the processes "black" and "red dotted". The total contribution of a particular original process in the structure of a given functional canonical correlation is equal to the area under the module weighting function corresponding to this process. These contributions for the components are given in Table 8.

Figure 12 contains the relative positions of the 38 European countries in the system $(\hat{U}_1, \hat{V}_1)$ of functional canonical variables for selected groups of variables. The high correlation of the first two functional canonical variables can be seen in Fig. 12 for two pillars G5 and G6. For KTA criterion, the countries with the highest value of functional canonical variables U_1 and V_1 are: Finland (FI), France (FR), Hungary (HU), Greece (GR), Estonia (EE), Germany (DE), Iceland (IS), Czech Republic (CZ) and Denmark (DK). The countries with the lowest value of functional canonical variables U_1 and V_1 are: Romania (RO), Poland (PL), Norway (NO), Portugal (PT), Netherlands (NL) and Russian Federation (RU). Other countries belong to the intermediate group.

During the numerical calculation process we used R software (R Core Team 2018) and packages `fda` (Ramsay et al. 2018) and `hsicCCA` (Chang 2013).

Table 8 Sorted areas under module weighting functions

No.	Area	Proportion (in %)
First functional canonical variable (G5)		
1	5.008	51.74
2	1.724	17.81
3	1.567	16.19
4	0.713	7.36
5	0.351	3.63
6	0.317	3.27
First functional canonical variable (G6)		
1	5.187	44.77
2	3.194	27.56
3	1.287	11.11
4	0.580	5.00
5	0.511	4.41
6	0.323	2.79
7	0.206	1.77
8	0.152	1.31
9	0.091	0.78
10	0.057	0.49

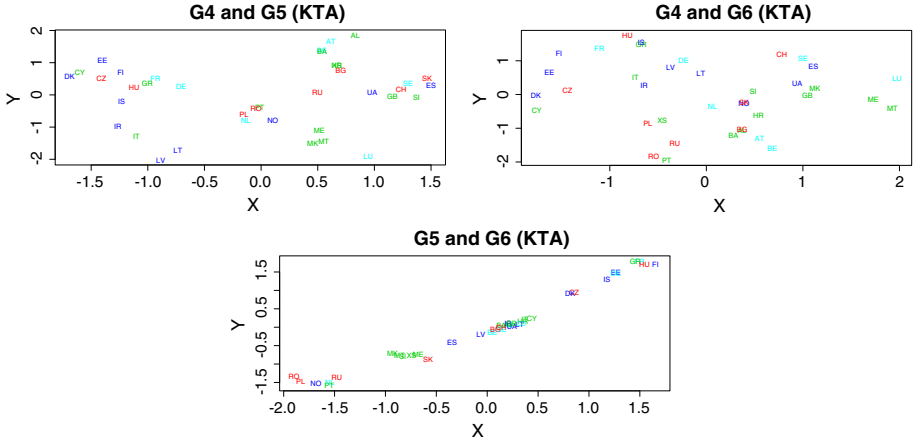

Fig. 12 Selected projections of the 38 European countries on the plane $(\hat{U}_1, \hat{V}_1)$. Regions used for statistical processing purposes by the United Nations Statistics Division: blue square—Northern Europe, cyan square—Western Europe, red square—Eastern Europe, green square—Southern Europe. (Color figure online)

5 Conclusions

We proposed an extension of two dependency measures for two sets of variables for multivariate functional data. We proposed to use tests to examine the significance of results because the values of proposed coefficients are rather hard to interpret. Additionally, we presented the methods of constructing nonlinear canonical variables for multivariate functional data using HSIC and KTA coefficients. Tested on two real examples, the proposed method has proven

useful in investigating the dependency between two sets of variables. Examples confirm usefulness of our approach in revealing the hidden structure of co-dependence between groups of variables.

During the study of proposed coefficients we discovered that the size of basis (smoothing parameter) is rather unimportant, the values (and p values for tests) do not depend on the basis size.

Of course, the performance of the methods needs to be further evaluated on additional real and artificial data sets. Moreover, we can examine the behavior of coefficients (and tests) for different bases like B-splines or wavelets (when data are not periodic, the Fourier basis could fail). This constitutes the direction of our future research.

Acknowledgements The authors are grateful to editor and two anonymous reviewers for giving many insightful and constructive comments and suggestions which led to the improvement of the earlier manuscript.

References

Aronszajn N (1950) Theory of reproducing kernels. Trans Am Math Soc 68:337–404

Chang B (2013) hsicCCA: Canonical Correlation Analysis based on Kernel Independence Measures. R package version 1.0. https://CRAN.R-project.org/package=hsicCCA

Chang B, Kruger U, Kustra R, Zhang J (2013) Canonical correlation analysis based on hilbert-schmidt independence criterion and centered kernel target alignment. In: Proceedings of the 30th international conference on machine learning, Atlanta, Georgia. JMLR: W and CP 28(2), 316–324

Cortes C, Mohri M, Rostamizadeh A (2012) Algorithms for learning kernels based on centered alignment. J Mach Learn Res 13:795–828

Cristianini N, Shawe-Taylor J, Elisseeff A, Kandola JS (2001) On kernel-target alignment. In: NIPS-2001, 367–373

Cuevas A (2014) A partial overview of the theory of statistics with functional data. J Stat Plan Inference 147:1–23

Devijver E (2017) Model-based regression clustering for high-dimensional data: application to functional data. Adv Data Anal Classif 11(2):243–279

Edelman A, Arias TA, Smith S (1998) The geometry of algorithms with orthogonality constraints. SIAM J Matrix Anal Appl 20(2):303–353

Ferraty F, Vieu P (2003) Curves discrimination: a nonparametric functional approach. Comput Stat Data Anal 44(1–2):161–173

Ferraty F, Vieu P (2006) Nonparametric functional data analysis: theory and practice. Springer, Berlin

Feuerverger A (1993) A consistent test for bivariate dependence. Int Stat Rev 61(3):419–433

Górecki T, Krzyśko M, Ratajczak W, Wołyński W (2016) An extension of the classical distance correlation coefficient for multivariate functional data with applications. Stat Transit 17(3):449–9466

Górecki T, Krzyśko M, Wołyński W (2017) Correlation analysis for multivariate functional data. In: Palumbo F, Montanari A, Montanari M (eds) Data science. Studies in classification, data analysis, and knowledge organization. Springer, Berlin, pp 243–258

Górecki T, Krzyśko M, Waszak Ł, Wołyński W (2018) Selected statistical methods of data analysis fir multivariate functional data. Stat Papers 59:153–182

Górecki T, Smaga Ł (2017) Multivariate analysis of variance for functional data. J Appl Stat 44:2172–2189

Gretton A., Bousquet O., Smola A., and Schölkopf B., (2005): Measuring statistical dependence with Hilbert–Schmidt norms. In: Jain S, Simon HU, Tomita E (eds) Algorithmic learning theory. Lecture notes in computer science 3734, 63–77. Springer

Gretton A, Fukumizu K, Teo CH, Song L, Schölkopf B, Smola AJ (2008) A kernel statistical test of independence. In: Platt JC, Koller D, Singer Y, Roweis S (eds) Advances in neural information processing systems. Curran, Red Hook, pp 585–592

Hofmann T, Schölkopf B, Smola AJ (2008) Kernel methods in machine learning. Ann Stat 36(3):1171–1220

Horváth L, Kokoszka P (2012) Inference for functional data with applications. Springer, Berlin
Hotelling H (1936) Relations between two sets of variates. Biometrika 28:321–377
Hsing T, Eubank R (2015) Theoretical foundations of functional data analysis, with an introduction to linear operators. Wiley, Hoboken
James GM, Wang JW, Zhu J (2009) Functional linear regression that's interpretable. Ann Stat 37(5):2083–2108
Kankainen A (1995) Consistent testing of total independence based on the empirical charecteristic function, Ph.D. thesis, University of Jyväskylä
Martin-Baragan B, Lillo R, Romo J (2014) Interpretable support vector machines for functional data. Eur J Oper Res 232:146–155
Mercer J (1909) Functions of positive and negative type and their connection with the theory of integral equations. Philos Trans R Soc Lond Ser A 209:415–446
R Core Team (2018) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/
Ramsay JO, Dalzell CJ (1991) Some tools for functional data analysis (with discission). J R Stat Soc Ser B 53(3):539–572
Ramsay JO, Silverman BW (2002) Applied functional data analysis. Springer, New York
Ramsay JO, Silverman BW (2005) Functional data analysis, 2nd edn. Springer, Berlin
Ramsay JO, Wickham H, Graves S, Hooker G (2018) fda: Functional data analysis. R package version 2.4.8. https://CRAN.R-project.org/package=fda
Read T, Cressie N (1988) Goodness-of-fit statistics for discrete multivariate analysis. Springer, Berlin
Riesz F (1909) Sur les opérations functionnelles linéaires. Comptes rendus hebdomadaires des séances de l'Académie des sciences 149:974–977
Sejdinovic D, Sriperumbudur B, Gretton A, Fukumizu K (2013) Equivalence of distance-based and RKHS-based statistics in hypothesis testing. Ann Stat 41(5):2263–2291
Schölkopf B, Smola AJ, Müller KR (1998) Nonlinear component analysis as a kernel eigenvalue problem. Neural Comput 10:1299–1319
Shawe-Taylor J, Cristianini N (2004) Kernel methods for pattern analysis. Cambridge University Press, Cambridge
Song L, Boots B, Siddiqi S, Gordon G, Somla A (2010) Hilbert space embeddings of hidden Markov models. In: Proceedings of the 26th international conference on machine learning (ICML2010)
Székely GJ, Rizzo ML, Bakirov NK (2007) Measuring and testing dependence by correlation of distances. Ann Stat 35(6):2769–2794
Székely GJ, Rizzo ML (2009) Brownian distance covariance. Ann Appl Stat 3(4):1236–1265
Wang T, Zhao D, Tian S (2015) An overview of kernel alignment and its applications. Artif Intell Rev 43(2):179–192
Zhang K, Peters J, Janzing D, Schölkopf B (2011) Kernel-based conditional independence test and application in causal discovery. In: Cozman FG, Pfeffer A (eds) Proceedings of the 27th conference on uncertainty in artificial intelligence, AUAI Press, Corvallis, OR, USA, 804–813