Measuring and Testing Mutual Dependence for Functional Data

Tomasz Górecki, Mirosław Krzyśko, and Waldemar Wołyński

Abstract In this paper, measures of mutual independence of many-vector random processes were defined. Based on these measures, permutation tests of mutual independence of these random processes were also given. The properties of the described methods were presented using simulation studies for univariate and multivariate processes.

Keywords Functional data · Mutual correlation · Measures of multiple independence

1 Introduction

Many processes currently used in different fields of science and research lead to random observations that can be analyzed as curves. We can also find a large amount of data for which it would be more appropriate to use some interpolation techniques and consider them as functional data. This approach turns out to be essential when data have been observed at different time intervals.

Earlier, Górecki et al. (2017, 2020) showed how to use commonly known measures of correlation for two sets of variables: ρV coefficient (Escoufier 1973), distance

T. Górecki (✉)
Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Uniwersytetu Poznańskiego 4, Poznań 61-614, Poland
e-mail: tomasz.gorecki@amu.edu.pl

M. Krzyśko
Interfaculty Institute of Mathematics and Statistics, Calisia University, Nowy Świat 4, 62-800 Kalisz, Poland
e-mail: mkrzysko@amu.edu.pl

W. Wołyński
Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Uniwersytetu Poznańskiego 4, Poznań 61-614, Poland
e-mail: wolynski@amu.edu.pl

T. Chadjipadelis et al. (eds.), *Data Analysis and Rationality in a Complex World*, Studies in Classification, Data Analysis, and Knowledge Organization,
https://doi.org/10.1007/978-3-030-60104-1_8

correlation coefficient (dCorr) (Székely et al. 2007), and HSIC coefficient (Gretton et al. 2005) for multivariate functional data.

In this paper, using ρV and dCorr coefficients, we define measures of mutual independence of vector random processes whose realizations are multidimensional functional data. Based on these measures, permutation tests of mutual independence of vector random processes $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K$, $K \geq 2$, $\boldsymbol{X}_i \in L_2^{p_i}(I)$, where $L_2(I)$ is a Hilbert space of square-integrable functions on the interval I, $i = 1, \ldots, K$ are also considered.

The rest of this paper is organized as follows. We first review the concept of transformation of discrete data to multivariate functional data (Sect. 2). Section 3 contains the functional version of the ρV and dCorr coefficients. Section 4 is devoted to measures of mutual independence of vector random processes and permutation tests of mutual independence associated with these measures. Section 5 contains the results of our simulation experiments.

2 Functional Data

Let us assume that $\boldsymbol{X} = (X_1, X_2, ..., X_p)^\top \in L_2^p(I)$ is p-dimensional random process, where $L_2(I)$ is the Hilbert space of square-integrable functions on the interval I. Moreover, assume that the kth component of the vector $\boldsymbol{X}$ can be represented by a finite number of orthonormal basis functions $\{\varphi_b\}$ of space $L_2(I)$:

$$X_k(t) = \sum_{b=0}^{B_k} \alpha_{kb}\varphi_b(t),\ t \in I,\ k = 1, \ldots, p.$$

Let $\boldsymbol{\alpha} = (\alpha_{10}, \ldots, \alpha_{1B_1}, \ldots, \alpha_{p0}, \ldots, \alpha_{pB_p})^\top$ and

$$\boldsymbol{\Phi}(t) = \begin{bmatrix} \boldsymbol{\varphi}_1^\top(t) & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\varphi}_2^\top(t) & \ldots & \mathbf{0} \\ \ldots & \ldots & \ldots & \ldots \\ \mathbf{0} & \mathbf{0} & \ldots & \boldsymbol{\varphi}_p^\top(t) \end{bmatrix}, \tag{1}$$

where $\boldsymbol{\varphi}_k(t) = (\varphi_0(t), ..., \varphi_{B_k}(t))^\top$, $k = 1, ..., p$.

Using the above matrix notation, process $\boldsymbol{X}$ can be represented as

$$\boldsymbol{X}(t) = \boldsymbol{\Phi}(t)\boldsymbol{\alpha}.$$

This means that the realizations of a process $\boldsymbol{X}$ are in finite-dimensional subspace of $L_2^p(I)$.

We can estimate the vector $\boldsymbol{\alpha}$ on the basis of n independent realizations $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ of the random process $\boldsymbol{X}$ (functional data).

Typically data are recorded at discrete moments in time. Let x_{kj} denote an observed value of the feature $X_k, k = 1, 2, \ldots, p$ at the jth time point t_j, where $j = 1, 2, ..., J$. Then our data consist of the pJ pairs (t_j, x_{kj}). These discrete data can be smoothed by continuous functions x_k and I is a compact set such that $t_j \in I$, for $j = 1, ..., J$.

Details of the process of transformation of discrete data to functional data can be found in Ramsay and Silverman (2005), Horváth and Kokoszka (2012), or in Górecki et al. (2014).

3 $K = 2$ Case

For two random vectors $\boldsymbol{X} \in R^p$ and $\boldsymbol{Y} \in R^q$, Escoufier (1973) introduced correlation coefficient ρV as a nonnegative number given by

$$\rho V_{\boldsymbol{X},\boldsymbol{Y}} = \frac{\|\boldsymbol{\Sigma}_{XY}\|_F}{\sqrt{\|\boldsymbol{\Sigma}_{XX}\|_F \|\boldsymbol{\Sigma}_{YY}\|_F}},$$

where $\|\cdot\|_F$ denoted the Frobenius norm and

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{XX} & \boldsymbol{\Sigma}_{XY} \\ \boldsymbol{\Sigma}_{YX} & \boldsymbol{\Sigma}_{YY} \end{bmatrix}$$

is a covariance matrix of vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$.

Correlation coefficient ρV has the following properties: $\rho V_{\boldsymbol{X},\boldsymbol{Y}} = 0$ if and only if random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ are uncorrelated. Moreover, if the joint distribution of $\boldsymbol{X}$ and $\boldsymbol{Y}$ is $p+q$ dimensional normal distribution, random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent.

We may extend this coefficient to two random processes $\boldsymbol{X} \in L_2^p(I)$ and $\boldsymbol{Y} \in L_2^q(I)$ assuming that

$$\|\boldsymbol{\Sigma}_{XY}\|_F = \sqrt{\int_I \int_I \mathrm{tr}(\boldsymbol{\Sigma}_{XY}^\top(s,t)\boldsymbol{\Sigma}_{XY}(s,t))dsdt}.$$

Moreover, if processes $\boldsymbol{X}$ and $\boldsymbol{Y}$ have the form

$$\mathbf{X}(t) = \boldsymbol{\Phi}_1(t)\boldsymbol{\alpha}, \quad \mathbf{Y}(s) = \boldsymbol{\Phi}_2(s)\boldsymbol{\beta}, \quad t, s \in I, \tag{2}$$

then Górecki et al. (2017)

$$\rho V_{\boldsymbol{X},\boldsymbol{Y}} = \rho V_{\boldsymbol{\alpha},\boldsymbol{\beta}}.$$

In this case, the problem of testing the correlation of processes $\boldsymbol{X}$ and $\boldsymbol{Y}$ is equivalent to the problem of zeroing the coefficient $\rho V_{\boldsymbol{\alpha},\boldsymbol{\beta}}$.

Note, that the coefficient $\rho V_{\boldsymbol{X},\boldsymbol{Y}}$ is appropriate only for linear dependence. It is useless for more complicated situations. It “cannot see” nonlinear dependencies. In such a situation, we ought to use some other measures of dependence.

One such measure is proposed by Székely et al. (2007) distance correlation. Let us denote by $\phi_{X,Y}$ and ϕ_X, ϕ_Y the joint and the marginals characteristic functions of random vectors $\boldsymbol{X} \in R^p$ and $\boldsymbol{Y} \in R^q$, respectively. Distance correlation of random vectors $\boldsymbol{X} \in R^p$ and $\boldsymbol{Y} \in R^q$ is a nonnegative number given by

$$\mathrm{dCorr}(\boldsymbol{X}, \boldsymbol{Y}) = \frac{\mathrm{dCov}(\boldsymbol{X}, \boldsymbol{Y})}{\sqrt{\mathrm{dCov}(\boldsymbol{X}, \boldsymbol{X})\,\mathrm{dCov}(\boldsymbol{Y}, \boldsymbol{Y})}},$$

where

$$\mathrm{dCov}(\boldsymbol{X}, \boldsymbol{Y}) = \|\phi_{X,Y}(\boldsymbol{l}, \boldsymbol{m}) - \phi_X(\boldsymbol{l})\phi_Y(\boldsymbol{m})\|_{w},$$

and

$$\|f\|_w = \sqrt{\iint |f(\boldsymbol{l}, \boldsymbol{m})|^2 \boldsymbol{w}(\boldsymbol{l}, \boldsymbol{m}) d\boldsymbol{l} d\boldsymbol{m}}.$$

The weight function $\boldsymbol{w}$ is chosen to produce scale free and rotation invariant measure that does not go to zero for dependent random vectors.

Defining the joint characteristic function of processes $\boldsymbol{X} \in L_2^p(I)$ and $\boldsymbol{Y} \in L_2^q(I)$ as

$$\phi_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{l}, \boldsymbol{m}) = \mathrm{E}\{\exp[i < \boldsymbol{l}, \boldsymbol{X} >_p + i < \boldsymbol{m}, \boldsymbol{Y} >_q]\},$$

where

$$< \boldsymbol{l}, \boldsymbol{X} >_p = \int_{I_1} \boldsymbol{l}'(s)\boldsymbol{X}(s)ds, \quad < \boldsymbol{m}, \boldsymbol{Y} >_q = \int_{I_2} \boldsymbol{m}'(t)\boldsymbol{Y}(t)dt$$

and assuming that processes $\boldsymbol{X}$ and $\boldsymbol{Y}$ have the form (2) we have

$$\mathrm{dCorr}(\boldsymbol{X}, \boldsymbol{Y}) = \mathrm{dCorr}(\boldsymbol{\alpha}, \boldsymbol{\beta})$$

Górecki et al. (2017).

Thus, we can reduce the problem of testing the independence of random processes $\boldsymbol{X}$ and $\boldsymbol{Y}$ to the problem of testing the significance of their distance correlation $\mathrm{dCorr}(\boldsymbol{X}, \boldsymbol{Y})$.

4 $K > 2$ Case

Let us now discuss the problem of testing mutual independence for more than two vector processes.

Let $\boldsymbol{X}_1 \in L_2^{p_1}(I), \boldsymbol{X}_2 \in L_2^{p_2}(I), \ldots, \boldsymbol{X}_K \in L_2^{p_K}(I)$ be random processes with the following representation:

$$\boldsymbol{X}_1(t) = \boldsymbol{\Phi}_1(t)\boldsymbol{\alpha}_1, \boldsymbol{X}_2(t) = \boldsymbol{\Phi}_2(t)\boldsymbol{\alpha}_2, \ldots, \boldsymbol{X}_K(t) = \boldsymbol{\Phi}_K(t)\boldsymbol{\alpha}_K, \; t \in I. \tag{3}$$

Additionally, let the covariance matrix for vectors $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_K$ have the form:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} & \cdots & \boldsymbol{\Sigma}_{1K} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} & \cdots & \boldsymbol{\Sigma}_{2K} \\ \vdots & \vdots & & \vdots \\ \boldsymbol{\Sigma}_{K1} & \boldsymbol{\Sigma}_{K2} & \cdots & \boldsymbol{\Sigma}_{KK} \end{bmatrix}.$$

Assuming joint $p_1 + p_2 + \cdots + p_k$ dimensional normal distribution of vectors $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_K$, the problem of testing the null hypothesis

$$H_0 \colon \boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_K \text{ are independent}$$

is equivalent to the problem of testing the null hypothesis

$$H_0 \colon \sum_{i<j} \|\boldsymbol{\Sigma}_{ij}\|_F = 0.$$

Let us define coefficient of mutual correlation ρMV as a positive number given by

$$\rho^2 MV = \frac{2}{K(K-1)} \sum_{i<j} \rho^2 V(\boldsymbol{X}_i, \boldsymbol{X}_j).$$

Assuming that the processes meet the assumptions of model (3) and that the joint distribution of vectors $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_K$ is normal, the problem of testing the mutual independence is equivalent to the problem of testing the significance of coefficient ρMV.

Another way to test the mutual independence is to reduce this problem to a problem using two processes.

Let $\mathrm{Corr}(\boldsymbol{X}_i, \boldsymbol{X}_j)$ be some measure of dependence for vector processes $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ with property: $\mathrm{Corr}(\boldsymbol{X}_i, \boldsymbol{X}_j) = 0$ if and only if vector processes $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ are independent, $i, j = 1, 2, \ldots, K, i \neq j$.

Note that in the place of Corr we may put, e.g., dCorr.

Let

$$\boldsymbol{X}_{c+} = (\boldsymbol{X}_{c+1}^\top, \ldots, \boldsymbol{X}_K^\top)^\top, \; c = 1, \ldots, K-1,$$

$$\boldsymbol{X}_{c-} = (\boldsymbol{X}_1^\top, \ldots, \boldsymbol{X}_{c-1}^\top, \boldsymbol{X}_{c+1}^\top, \ldots, \boldsymbol{X}_K^\top)^\top, \; c = 1, \ldots, K.$$

Following the idea from Jin and Matteson (2018), we may define the coefficients of multiple independence as

$$\mathscr{R}(\boldsymbol{X}) = \frac{1}{K-1} \sum_{c=1}^{K-1} \mathrm{Corr}^2(\boldsymbol{X}_c, \boldsymbol{X}_{c+}),$$

and

$$\mathscr{S}(\boldsymbol{X}) = \frac{1}{K}\sum_{c=1}^{K} \mathrm{Corr}^2(\boldsymbol{X}_c, \boldsymbol{X}_{c-}).$$

Thus, the following theorem is true:

Theorem 1 *$\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_K$ are independent if and only if $\mathscr{R}(\boldsymbol{X}) = 0$ or $\mathscr{S}(\boldsymbol{X}) = 0$.*

Hence, the problem of testing the null hypothesis

$$H_0 \colon \boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_K \text{ are independent}$$

is equivalent to the problem of testing the null hypothesis

$$H_0 \colon \mathscr{R}(\boldsymbol{X}) = 0 \ (\mathscr{S}(\boldsymbol{X}) = 0).$$

To verify these hypotheses, we propose to use a permutation test.

5 Example

5.1 Univariate Case

Let

$$\begin{aligned} X_t &= \varepsilon_{1t}, \\ Y_t &= 3X_t + \varepsilon_{2t}, \\ Z_t &= X_t^2 + \varepsilon_{3t}, \end{aligned}$$

where ε_{1t}, ε_{2t} and ε_{3t} are independent random variables with $N(0, 0.25)$ distribution. We generated 1000 random realizations for each process with length 100 (Fig. 1). To smooth the data we used Fourier series with 15 elements. Clearly, processes X_t and Y_t are linearly dependent and processes X_t, Z_t and Y_t, Z_t are non-linearly dependent.

From Table 1 (third column), we see that all measures of correlation for functional data detect dependence (at significance level 5%) when at least one pair of linearly dependent processes exist. However, when we have nonlinear dependence only measures based on dCorr detect it.

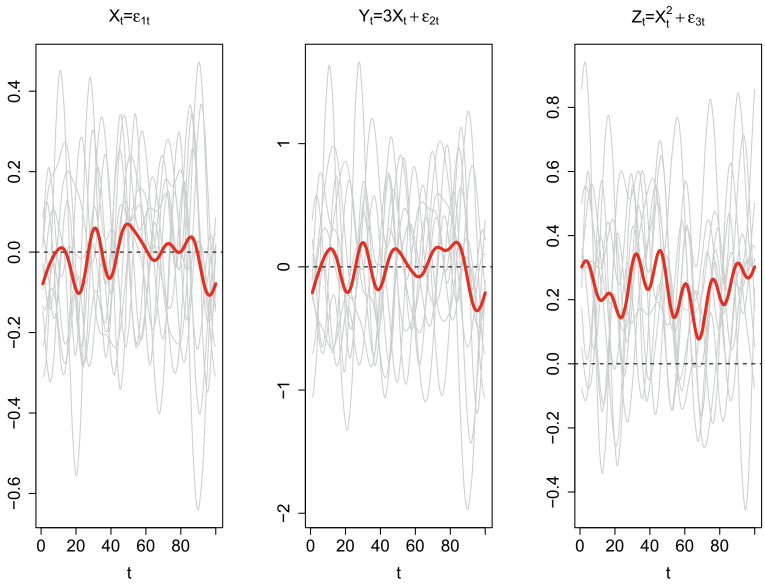

Fig. 1 10 realizations of univariate processes X_t, Y_t, and Z_t (functional means in red)

Table 1 Results of simulations (significant (5%) results are in bold)

Coefficient's name	Processes	p-value (Univ.)	p-value (Multi. – S1)	p-value (Multi. – S2)
ρMV	X_t, Y_t, Z_t	**0.014**	0.757	**0.036**
$\mathscr{R} - \rho V$	X_t, Y_t, Z_t	**0.006**	0.767	**0.015**
$\mathscr{S} - \rho V$	X_t, Y_t, Z_t	**0.027**	0.787	**0.024**
$\mathscr{R}$ – dCorr	X_t, Y_t, Z_t	**0.007**	0.581	**0.017**
$\mathscr{S}$ – dCorr	X_t, Y_t, Z_t	**0.016**	0.592	**0.006**
ρV	X_t vs Y_t	**0.001**	0.783	0.632
	X_t vs Z_t	0.367	0.568	0.203
	Y_t vs Z_t	0.481	0.566	0.526
dCorr	X_t vs Y_t	**0.001**	0.827	0.773
	X_t vs Z_t	**0.003**	0.486	0.094
	Y_t vs Z_t	**0.025**	0.457	0.470

5.2 *Multivariate Case*

Following Krzyśko and Smaga (2019) we consider the functional sample $\boldsymbol{x}_1(t), \ldots,$ $\boldsymbol{x}_n(t)$ of size $n = 1000$ containing realizations of the random process $\boldsymbol{X}(t) = (X(t), Y(t), Z(t))$, $t \in [0, 1]$ of dimension $p = 3$. These observations are generated in the following discretized way:

$$\boldsymbol{x}_i(t_j) = \boldsymbol{\Phi}(t_j)\boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_{ij},$$

where $i = 1, \ldots, n$, t_j, $j = 1, \ldots, 100$ are equally spaced design time points in [0, 1], the matrix $\boldsymbol{\Phi}(t)$ is as in (1) and contains the Fourier basis functions only and $B_k = 5$, $k = 1, \ldots, p$, $\boldsymbol{\alpha}_i$ are $5p$-dimensional random vectors, and $\boldsymbol{\varepsilon}_{ij} = (\varepsilon_{ij1}, \ldots, \varepsilon_{ijp})^\top$ are measurement errors such that $\varepsilon_{ijk} \sim N(0, 0.025 r_{ik})$ and r_{ik} is the range of the kth row of the matrix

$$\boldsymbol{\Phi}(t_1)\boldsymbol{\alpha}_i \ldots \boldsymbol{\Phi}(t_{100})\boldsymbol{\alpha}_i,$$

$k = 1, \ldots, p$. The random vectors $\boldsymbol{\alpha}_i$ are generated similarly to Todorov and Pires (2007) and Jin and Matteson (2018) in the following two setups:

S1 Normal distribution and equal covariance matrices: $\boldsymbol{\alpha}_i \sim N(\mathbf{0}_{5p}, \boldsymbol{I}_{5p})$.

S2 Part of $\boldsymbol{\alpha}_i$ for $(X(t), Y(t))$ is from $N(\mathbf{0}_{5(p-1)}, \boldsymbol{I}_{5(p-1)})$ and the first element of $\boldsymbol{\alpha}_i$ for $Z(t)$ is $\mathrm{sgn}(\alpha_1 \alpha_{5+1})W$, where $W \sim \mathrm{Exp}(1/\sqrt{2})$ and the remaining $p - 1$ elements are $N(\mathbf{0}_{5(p-1)-1}, \boldsymbol{I}_{5(p-1)-1})$. Clearly, $(X(t), Y(t), Z(t))$ is a pairwise independent but mutually dependent triplet.

Setup S1 is simple no dependence example. All tests correctly deal with this problem (Table 1, fourth column). Setup S2 is much harder to deal with. For whole triplet of data, all methods indicate dependence (Table 1, fifth column). For a pair of variables, all methods correctly detect independence for all pairs of processes.

6 Conclusions

We have considered the measuring and testing mutual dependence for multivariate functional data based on the basis functions representation of the data. We propose few measures of mutual dependence for multivariate functional data based on the equivalence to mutual independence through characteristic functions (Székely et al. 2007) and on ρV coefficient (Escoufier 1973). The performance of the proposed methods was studied in simulations. Their results have indicated that the proposed methods perform quite well. Finally, we can propose to use measures and tests based dCorr coefficient. Such methods correctly detect linear and nonlinear dependence structure both for univariate and multivariate processes.

References

Escoufier, Y.: Le traitement des variables vectorielles. Biometrics **29**(4), 751–760 (1973)

Górecki, T., Krzyśko, M., Waszak, Ł., Wołyński, W.: Methods of reducing dimension for functional data. Stat. Transit. New Ser. **15**(2), 231–242 (2014)

Górecki, T, Krzyśko, M., Wołyński, W.: Correlation analysis for multivariate functional data. In: Palumbo, F., Montanari, A., Vichi, M. (eds.) Data Science, Studies in Classification, Data Analysis, and Knowledge Organization, pp. 243–258. Springer International Publishing (2017)

Górecki, T., Krzyśko, M., Wołyński, W.: Independence test and canonical correlation analysis based on the alignment between kernel matrices for multivariate functional data. Artif. Intell. Rev. **53**, 475–499 (2020)

Gretton A., Bousquet O., Smola A., and Schölkopf B.: Measuring statistical dependence with Hilbert-Schmidt norms. In: Jain, S., Simon, H.U., Tomita, E. (eds.) Algorithmic Learning Theory. ALT 2005. Lecture Notes in Computer Science, vol. 3734, pp. 63–77. Springer, Berlin, Heidelberg (2005)

Horváth, L., Kokoszka, P.: Inference for Functional Data with Applications. Springer (2012)

Jin, Z., Matteson, D.S.: Generalizing distance covariance to measure and test multivariate mutual dependence via complete and incomplete V-statistics. J. Multivariate Anal. **168**, 304–322 (2018)

Krzyśko, M., Smaga, Ł.: A multivariate coefficient of variation for functional data. Stat. Interface **12**, 647–658 (2019)

Ramsay, J.O., Silverman, B.W.: Functional Data Analysis, 2nd edn. Springer (2005)

Székely, G.J., Rizzo, M.L., Bakirov, N.K.: Measuring and testing dependence by correlation of distances. Ann. Stat. **35**, 2769–2794 (2007)

Todorov, V., Pires, A.M.: Comparative performance of several robust linear discriminant analysis methods. Revstat. Stat. J. **5**, 63–83 (2007)