sciendo

DOI: 10.2478/bile-2020-0009

Biometrical Letters
Vol. 57 (2020), No. 2, 119-129

Application of the ρV coefficient and distance correlation to the analysis of multivariate association

Malwina Janiszewska[1], Anna Szczepańska-Álvarez[1], Emilia Zawieja[2]

[1]Department of Mathematical and Statistical Methods, Poznań University of Life Sciences, Wojska Polskiego 28, PL-60-637 Poznań, Poland,
e-mail: malwina.janiszewska@up.poznan.pl, anna.szczepanska-alvarez@up.poznan.pl
[2] Institute of Human Nutrition and Dietetics, Poznań University of Life Sciences, Wojska Polskiego 31, PL-60-624 Poznań, Poland,
e-mail: emilia.zawieja@up.poznan.pl

Summary

The aim of this paper is to study the association between two random vectors related to two groups of characteristics. To analyze the multivariate association, the ρV coefficient and distance correlation are used. Two methods (classical and recent) are compared and illustrated with real data.

Key words: ρV coefficient, distance correlation, association between two random vectors

1. Introduction

In many experiments nowadays, the association between variables is studied. Depending on the nature of the data and the aim of the research, different measures are used in the analysis. For example, to describe the relationship between financial objects and the market, Tjøstheim and Hufthammer (2013) introduce Local Gaussian correlation. The authors give a mathematical description and an interpretation of the phenomena, where the correlation between financial objects becomes stronger as the market falls. Moreover, Reshef et al. (2011) present the maximal information coefficient (MIC) to detect a wide range of functional and non-functional relationships between variables for large data sets. MIC can be used in genomics, physics and political sciences, in research where hundreds of variables are observed. The above coefficients determine the relationship between two variables, but in some experiments, the association between groups of variables is studied.

The measured features can be divided into sets according to certain properties. For example, in medicine the relations between physical, biological and chemical characteristics are studied, in agriculture biological and chemical traits are often observed, etc. The problem of detecting associations between two random vectors is widely described in the literature. Among others, Escoufier (1973) proposes the RV coefficient as a multivariate generalization of the squared Pearson correlation coefficient. This approach is good for multivariate measure, but not for high-dimensional data. Heller et al. (2013) develop a powerful nonparametric multivariate test of association based on ranks of distances. Furthermore, Székely and Rizzo (2007) define the distance correlation, which provides a new approach to the problem of testing the joint independence of two random vectors. Székely and Rizzo (2013) notice that the bias of the coefficient increases with the dimension of the data, and the authors propose a modification of the coefficient and formulate an improved t-test for independence.
In our paper we compare the ρV coefficient and the distance correlation coefficient ($dCor$) and the corresponding tests. We focus on examining the association between two groups of characteristics measured on the same objects. The first set of variables is observed and measured on multiple occasions over the time of the experiment, while the second does not change during the time of the experiment.
The paper is organized as follows. In section 2 the RV coefficient and $dCor_n$ coefficient, and their properties, are described. An application of the considered methods to real data is presented in section 3. Results and conclusions are included in the final section.

2. Measures of association between two random vectors

Let us consider two random vectors $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$, and let the matrix $\mathbf{\Sigma}$ be the covariance matrix of $(X; Y)$ in the following form

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

where $\mathbf{\Sigma}_{XY}$ is the covariance matrix between X and Y, and $\mathbf{\Sigma}_{XX}$ and $\mathbf{\Sigma}_{YY}$ are appropriate variance-covariance matrices for X and Y respectively.
The association between two random vectors can be determined by the ρV coefficient or the distance correlation coefficient $dCor$. In this paper we

use the following notation: ρV and $dCor$ describe the relationship in the population, and RV and $dCor_n$ are calculated for the sample.

2.1. The ρV coefficient

Escoufier (1973) defined the population ρV correlation coefficient of two random vectors X and Y as:

$$\rho V(X,Y) = \frac{cov(X,Y)}{\sqrt{var(X)var(Y)}} = \frac{tr(\mathbf{\Sigma}_{YX}\mathbf{\Sigma}_{XY})}{\sqrt{tr(\mathbf{\Sigma}_{XX}^2)tr(\mathbf{\Sigma}_{YY}^2)}},$$

where $tr(\mathbf{\Sigma}_{YX}\mathbf{\Sigma}_{XY})$ is called the covariance between X and Y, and $tr(\mathbf{\Sigma}_{XX}^2)$ and $tr(\mathbf{\Sigma}_{YY}^2)$ are the so-called variances of X and Y respectively.

The main properties of the ρV coefficient are:

1. When $p = q = 1, \rho V = \rho^2$, where ρ^2 is the square of the simple correlation coefficient
2. $0 \leqslant \rho V(X,Y) \leqslant 1$
3. $\rho V(X,Y) = 0$ if and only if $\mathbf{\Sigma}_{YX} = 0$
4. $\rho V(X, a\mathbf{B}X + C) = 1$, where a is a constant, $\mathbf{B}$ is an orthogonal matrix and C is a constant vector. This means that ρV is invariant under shift, rotation and overall scaling.

We examine data matrices $\mathbf{X}_{n,p}$, $\mathbf{Y}_{n,q}$ representing n independent realizations of the random vectors, which are assumed column-centered.

The ρV coefficient can be estimated by:

$$RV(\mathbf{X},\mathbf{Y}) = \frac{\mathrm{tr}(\mathbf{S_{XY}S_{YX}})}{\sqrt{\mathrm{tr}(\mathbf{S_{XX}^2})\,\mathrm{tr}(\mathbf{S_{YY}^2})}},$$

where $\mathbf{S_{XY}} = \frac{1}{n-1}\mathbf{X'Y}$ is the sample covariance matrix between $\mathbf{X}$ and $\mathbf{Y}$, and $\mathbf{S_{XX}} = \frac{1}{n-1}\mathbf{X'X}$ and $\mathbf{S_{YY}} = \frac{1}{n-1}\mathbf{Y'Y}$ are sample variance-covariance matrices corresponding to $\mathbf{X}$ and $\mathbf{Y}$ (Josse and Holmes, 2016). Furthermore, the authors transformed the above equation and obtained the following form of the RV correlation coefficient:

$$RV(\mathbf{X},\mathbf{Y}) = \frac{\mathrm{tr}(\mathbf{XX'YY'})}{\sqrt{\mathrm{tr}[(\mathbf{XX'})^2]\,\mathrm{tr}[(\mathbf{YY'})^2]}},$$

where $\mathbf{XX'}$ and $\mathbf{YY'}$ are matrices representing the relative positions of the observations.

Two sets of variables are correlated if the relative position of the observations in one set is similar to the relative position of the samples in the other set. Let us write $\mathbf{W_X} = \mathbf{XX'}$ and $\mathbf{W_Y} = \mathbf{YY'}$ as the cross-product matrices. Since the two matrices $\mathbf{W_X}$ and $\mathbf{W_Y}$ may have different norms, a correlation coefficient, the RV coefficient, is computed by:

$$RV(\mathbf{X}, \mathbf{Y}) = \frac{\operatorname{tr}(\mathbf{XX'YY'})}{\sqrt{\operatorname{tr}[(\mathbf{XX'})^2]\operatorname{tr}[(\mathbf{YY'})^2]}} = \frac{< \mathbf{W_X}, \mathbf{W_Y} >_{HS}}{||\mathbf{W_X}||_{HS}||\mathbf{W_Y}||_{HS}},$$

where the nominator contains the Hilbert–Schmidt inner product between matrices $\mathbf{W_X}$ and $\mathbf{W_Y}$, and the denominator contains the product of the Hilbert–Schmidt norm matrices $\mathbf{W_X}$ and $\mathbf{W_Y}$. Moreover, to measure their proximity, the Hilbert–Schmidt inner product between matrices is computed:

$$< \mathbf{W_X}, \mathbf{W_Y} >_{HS} = \operatorname{tr}(\mathbf{XX'YY'}) = \sum_{l=1}^{p}\sum_{m=1}^{q} \operatorname{cov}^2(X_{.l}, Y_{.m}),$$

where cov is the sample covariance coefficient, $X_{.l}$ is column l of the matrix $\mathbf{X}$, and $Y_{.m}$ is column m of the matrix $\mathbf{Y}$. This coefficient can be viewed as the cosine between the two-dimensional vectors representing the inner product matrices. Furthermore, we can express the RV coefficient using distance matrices in the following form (Josse and Holmes, 2016):

$$RV(\mathbf{X}, \mathbf{Y}) = \frac{< \mathbf{C}\boldsymbol{\Delta}_{\mathbf{X}}^2\mathbf{C}, \mathbf{C}\boldsymbol{\Delta}_{\mathbf{Y}}^2\mathbf{C} >_{HS}}{||\mathbf{C}\boldsymbol{\Delta}_{\mathbf{X}}^2\mathbf{C}||_{HS}||\mathbf{C}\boldsymbol{\Delta}_{\mathbf{Y}}^2\mathbf{C}||_{HS}},$$

where $\boldsymbol{\Delta}_{\mathbf{X}} : n \times n$ is the distance matrix for the set $\mathbf{X}$, where elements d_{ij} represent the Euclidean distance between the samples i and j (similarly $\boldsymbol{\Delta}_{\mathbf{Y}} : n \times n$ is defined for the set $\mathbf{Y}$; for the definition of Euclidean distance see Gower, 1966), $\mathbf{C} = \mathbf{I}_n - \frac{\mathbf{1}_n\mathbf{1}_n'}{n}$ where $\mathbf{I}_n$ is the identity matrix of order n and $\mathbf{1}_n$ is a vector of ones of size n. The RV coefficient has the same properties as the ρV coefficient. To assess the significance of the association, the following hypothesis test based on the ρV coefficient is used:

$$\begin{cases} H_0 : \rho V = 0 & \text{there is no linear association between } X \text{ and } Y \\ H_1 : \rho V > 0 & \text{there is linear association between } X \text{ and } Y \end{cases}$$

The fact that $\rho V = 0$ means the absence of a linear relationship between groups of characteristics. Moreover, if the assumption of multivariate normality is satisfied, then $\rho V = 0$ implies that X and Y are independent. Furthermore, if the random variables have a multivariate normal joint distribution or if they belong to the class of elliptical distributions, it is possible to use the asymptotic test to verify the null hypothesis. It is common to make enhancements by the use of rank data

(Cléroux et al., 1995). However, Josse et al. (2008) show that asymptotic tests are misleading for small sample sizes, despite some improvements. These tests are good only for large sample sizes ($n > 300$). Since the asymptotic tests fail, it is possible to use the permutation test. When the number of individuals is sufficiently small, the exact RV permutation distribution can be obtained and an exact test can be performed; otherwise this method is computationally costly. The permutation test consists in simulating the null hypothesis by breaking the potential relationship between the two data sets, permuting the rows of one matrix. To obtain the RV permutation distribution under the null hypothesis, the rows of one matrix are permuted and the RV coefficient is computed. This step is repeated for each of the $n!$ permutations. Thus we obtain the RV permutation distribution under the null hypothesis.

Husson et al. (2013) implemented the RV coefficient in an R package, as coeffRV in FactoMineR. The function coeffRV provides the computation of the measure and the Pearson type III approximation to test its significance, and verifies hypotheses based on the permutation test.

2.2. The distance correlation $dCor$

Székely et al. (2007) defined a measure of the dependence between two random vectors with finite first moments called the distance correlation ($dCor$ or R). They showed that the distance correlation generalizes the idea of correlation for all random variables with finite first moments. It was proved that $dCor$ is defined for two random vectors X and Y of arbitrary dimensions, not necessarily equal. Moreover, $dCor = 0$ if and only if the random vectors are independent, and the coefficient $dCor$ does not require distributional assumptions. The distance correlation is defined as the non-negative number $R(X, Y)$ by

$$R^2(X,Y) = \begin{cases} \frac{V^2(X,Y)}{V^2(X)V^2(Y)}, & V^2(X)V^2(Y) > 0; \\ 0, & V^2(X)V^2(Y) = 0. \end{cases}$$

where $V^2(X, Y)$ is the distance covariance ($dCov$) between random vectors X and Y based on the distance between the joint characteristic function of X and Y denoted by $f_{X,Y}$ and the product of the marginal characteristic functions (f_X, f_Y), with a suitable weight function defined as:

$$V^2(X,Y;w) = \int_{\mathbb{R}^{p+q}} |f_{X,Y}(t,s) - f_X(t)f_Y(s)|^2 w(t,s)dtds$$

where $w(t,s) = (c_p c_q |t|_p^{1+p} |s|_q^{1+q})^{-1}$, $c_k = \frac{\pi^{(1+k)/2}}{\Gamma((1+k)/2}$ and $\Gamma(\cdot)$ is the complete gamma function. Similarly, the distance variance ($dVar$) is defined as

$$V^2(X;w) = V^2(X,X;w) = \int_{\mathbb{R}^{2p}} |f_{X,X}(t,s) - f_X(t) f_X(s)|^2 w(t,s) dt ds$$

and $V^2(Y;w)$ is defined analogously (Székely et al. (2007)).
The main properties of the $dCor$ coefficient given by Székely and Rizzo (2009) are as follows:

1. $0 \leqslant dCor(X,Y) \leqslant 1$
2. $dCor(X,Y) = 0$ if and only if X and Y are independent
3. $dCor(X, aX\mathbf{B} + C) = 1$ where a is a constant, $\mathbf{B}$ is an orthogonal matrix and C is a constant vector. This means that $dCor$ is invariant under shift, orthogonal transformation and scaling.

It is worth noting that the ρV coefficient has similar properties to the $dCor$ coefficient.
Moreover, for the examined data the empirical distance correlation $dCor_n$ is the square root of

$$dCor_n^2(\mathbf{X},\mathbf{Y}) = \begin{cases} \frac{V_n^2(\mathbf{X},\mathbf{Y})}{V_n^2(\mathbf{X}) V_n^2(\mathbf{Y})}, & V_n^2(\mathbf{X}) V_n^2(\mathbf{Y}) > 0; \\ 0, & V_n^2(\mathbf{X}) V_n^2(\mathbf{Y}) = 0. \end{cases} \tag{1}$$

and similarly $V_n^2(\mathbf{X},\mathbf{Y};w), V_n^2(\mathbf{X};w), V_n^2(\mathbf{Y};w)$ are defined by Székely et al. (2007). Josse and Holmes (2016) transformed the form of $dCor_n$ given in (1) to

$$dCor_n^2(\mathbf{X},\mathbf{Y}) = \frac{< \mathbf{C}\boldsymbol{\Delta}_\mathbf{X}\mathbf{C}, \mathbf{C}\boldsymbol{\Delta}_\mathbf{Y}\mathbf{C} >_{HS}}{||\mathbf{C}\boldsymbol{\Delta}_\mathbf{X}\mathbf{C}||_{HS} ||\mathbf{C}\boldsymbol{\Delta}_\mathbf{Y}\mathbf{C}||_{HS}}.$$

The difference in the forms of the $dCor_n^2$ and RV coefficients is that for the $dCor_n^2$ coefficient the distance matrices $\boldsymbol{\Delta}_\mathbf{X}$ and $\boldsymbol{\Delta}_\mathbf{Y}$ are not the squared Euclidean distances. This property implies that the $dCor_n$ coefficient detects non-linear relationships, whereas the RV coefficient is restricted to linear ones. Moreover, Josse and Holmes (2016) showed that the ρV test is more powerful for detecting a linear association between two groups of characteristics, but the $dCor$ test has the highest power in searching for

non-linear relationships. In both cases the power of the tests increases with the number of observations.

To assess independence between random vectors, the following hypothesis test is used (Székely et al., 2007):

$$\begin{cases} H_0 : f_{XY} = f_X f_Y & X \text{ and } Y \text{ are independent} \\ H_1 : f_{XY} \neq f_X f_Y & X \text{ and } Y \text{ are not independent} \end{cases}$$

The authors proposed a test of independence based on the statistic nV_n^2. They proved that under independence nV_n^2/S_2 ($S_2 = \frac{1}{n^2}\sum_{k,l=1}^{n} |X_k - X_l|_p \frac{1}{n^2}\sum_{k,l=1}^{n} |Y_k - Y_l|_q$, where $|X_k - X_l|_p$ is the Euclidean distance between the kth and lth observations for p characteristics, and $|Y_k - Y_l|_q$ is the Euclidean distance between the kth and lth observations for q characteristics) converges in the distribution to a quadratic form $Q = \sum_{j=1}^{\infty} \lambda_j Z_j^2$, where Z_j are independent standard normal random variables, λ_j are nonnegative constants that depend on the distribution of (X, Y), $E(Q) = 1$, and it tends to infinity otherwise. A test of independence that rejects independence for large nV_n^2/S_2 is statistically consistent against all alternatives with first moments, whereas some alternatives are ignored in the test based on the RV coefficient. For small sample sizes permutation tests are more often used to assess significance for the distance covariance coefficient.

Székely and Rizzo (2013) observed that the bias of the $dCor_n$ coefficient increases with the dimension of the data. The authors proposed a modification of this coefficient and formulated the t-test for independence (which is a transformation of the test given above). The modified coefficient is approximately normal for $n \geqslant 10$. Furthermore, the $dCor_n$ coefficient and the t-test of independence were implemented by them and developed as the R package energy with the function dcor.test.

3. Results of the experiment

In this section, the RV coefficient and the $dCor_n$ coefficient are illustrated on real data. The coefficients are used to assess the association between two groups of characteristics. The results obtained by two methods, the classical one given by Escoufier (1973) and the recent one given by Székely and Rizzo (2013), are compared.

The aim of the experiment is to measure the influence of a rapid weight loss diet on body mass and composition and blood cell numbers in athletes

competing in combat sports. Participants are provided with individually adjusted meal plans to follow for 11 days. Each participant is measured three times: before, during and after the diet, at rest and after the Ramp test.

We can distinguish two groups of characteristics measured on the same objects. One group of characteristics is observed over the time of the experiment, and the second does not change during the time of the experiment. In the study under consideration the diet is applied to 10 patients, and seven characteristics divided into two sets are measured. The first set contains five features: fat tissue, water level in the body, and the numbers of lymphocytes (LYM), erythrocytes (RBC) and thrombocytes (PLT) in the blood. The second set contains two characteristics that are invariable over the time of the experiment: the patient's age and height.

Now, let the matrix $\mathbf{X}$ with 10 rows and 5 columns and the matrix $\mathbf{Y}$ with 10 rows and 2 columns be the realizations of two random vectors. We test whether there is an association between the features from the first group $\mathbf{X} : 10 \times 5$ and the second group $\mathbf{Y} : 10 \times 2$ of features, at significance level 0.05.
For the given data set we compute the RV coefficient and the $dCor_n$ coefficient. The results are presented below.

It can be seen from Table 1 that the tests based on the ρV coefficient and the t-test of independence for the $dCor$ coefficient do not detect a significant multivariate relationship between the two considered groups of characteristics at each of the three time points. This implies that there is not a significant association (linear or non-linear) between the measures that vary over time and the characteristics that are constant in time.

Some of the characteristics, related to amounts of cells in blood (LYM, RBC, PLT), are measured in two states, at rest and immediately after the Ramp test. Now, we analyze the three characteristics that vary over time and two features that do not change in time (age and height of patients) which are measured in two states at three time points. We test whether there is an association between the features from the first group $\mathbf{X} : 10 \times 3$ and the second group $\mathbf{Y} : 10 \times 2$ of features, at significance level 0.05.

Table 2 shows results for characteristics measured at rest. Both measures of association do not identify a significant multivariate relationship between the two considered groups of characteristics at each of the three time points.

The results, collected in Table 3, for features measured after the Ramp test are similar to the results for characteristics measured at rest. Both

Table 1. Values of coefficients and p-values for testing association between two groups of characteristics at three time points

	Dependence measure	Coefficient	p-value
Before the diet			
	RV	0.1283	0.4464
	$dCor_n$	0.6577	0.2806
During the diet			
	RV	0.0294	0.8466
	$dCor_n$	0.5492	0.6527
After the diet			
	RV	0.1249	0.4801
	$dCor_n$	0.5715	0.5240

Table 2. Values of coefficients and p-values for testing association between two groups of characteristics measured at three time points at rest

	Dependence measure	Coefficient	p-value
Before the diet			
	RV	0.1142	0.4796
	$dCor_n$	0.6337	0.3110
During the diet			
	RV	0.0134	0.8814
	$dCor_n$	0.5314	0.5338
After the diet			
	RV	0.0935	0.5627
	$dCor_n$	0.5253	0.6266

measures of association do not identify a significant multivariate relationship between the two considered groups of characteristics at each of the three time points.

Next we analyze the relationship between two groups of features related to amounts of cells in the blood. The first group of characteristics consists of LYM, RBC and PLT measured at rest, and the second consists of LYM, RBC and PLT after the Ramp test. We test whether there is an association between the features from the first group $\mathbf{X} : 10 \times 3$ and the second group $\mathbf{Y} : 10 \times 3$ of features at three diet stages, at significance level 0.05.

The results in Table 4 show that in all cases a linear association between characteristics is detected (the test based on the ρV coefficient) and independence between features is rejected (the test based on $dCor$). We observe that the Ramp test influences the considered characteristics.

In the given experiment an association between the relevant characteris-

Table 3. Values of coefficients and p-values for testing association between two groups of characteristics measured after the Ramp test at three time points

	Dependence measure	Coefficient	p-value
Before the diet			
	RV	0.1784	0.3056
	$dCor_n$	0.6213	0.3298
During the diet			
	RV	0.1185	0.4711
	$dCor_n$	0.5433	0.6389
After the diet			
	RV	0.1605	0.3444
	$dCor_n$	0.6552	0.1908

Table 4. Values of coefficients and p-values for testing association between two groups of characteristics related to amounts of cells in the blood, measured at three time points

	Dependence measure	Coefficient	p-value
Before the diet			
	RV	0.7450	0.0051
	$dCor_n$	0.8341	0.0048
During the diet			
	RV	0.6319	0.0065
	$dCor_n$	0.7732	0.0193
After the diet			
	RV	0.8542	0.0010
	$dCor_n$	0.9295	0.0002

tics was not observed at each step of the experiment (before the diet, during the diet and after the diet); thus we may suspect that the diet does not affect the relationship between traits. Physical exercises (the Ramp test) do not change the conclusions. Moreover, we can conclude that the age and height of patients do not interfere with the influence of diet on the studied characteristics: fat tissue, water level in the body, and the numbers of lymphocytes (LYM), erythrocytes (RBC) and thrombocytes (PLT) in the blood. The results of the analysis may be impacted by the small number of observations and the nature of the data.

References

Cléroux R., Lazraq A., Lepage Y.(1995): Vector correlation based on ranks and a nonparametric test of no association between vectors. Communications in Statistics Theory and Methods 24: 713–733.

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

Gower J. (1966): Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53: 325–338.

Heller R., Heller Y., Gorfine M. (2013): A consistent multivariate test of association based on ranks of distances. Biometrika 100: 503–510.

Husson F., Josse J., Le S., Mazet J. (2013): FactoMineR: Multivariate Exploratory Data Analysis and Data Mining with R. URL http://cran.r-project.org/package=FactoMineR, R package version 1.24.

Josse J., Pages J., Husson F. (2008): Testing the significance of rv coefficient. Computational Statistics and Data Analysis 53: 82–91.

Josse J., Holmes S. (2016): Measuring multivariate association and beyond. Statistics Surveys 10: 132–167.

Tjøstheim D., Hufthammer K. (2013): Local Gaussian correlation: a new measure of dependence. Journal of Econometrics 172(1): 33–48.

Reshef D., Reshef Y., Finucane H., Grossman S., McVean G., Turnbaugh P., Lander E., Mitzenmacher M., Sabeti P.C. (2011): Detecting novel associations in large data sets. Science 334: 1518–1524.

Székely G.J., Rizzo M.L., Bakirov N.K. (2007): Measuring and testing independence by correlation of distances. Annals of Statistics 35: 2769–2794.

Székely G.J., Rizzo M.L., (2009): Brownian distance covariance. The Annals of Applied Statistics 3(4): 1236–1265.

Székely G.J., Rizzo M.L. (2013): energy: E-statistics (energy statistics). URL http://CRAN.R-project.org/package=energy. R package version 1.6.0.