Estimation of Distance Correlation: a Simulation-based Comparative Study

Blanca E. Monroy-Castillo, M.A. Jácome, and Ricardo Cao

Centro de Investigación CITIC, Universidade da Coruña, 15071 A Coruña, Spain
Research Group MODES, Faculty of Sciences, Universidade da Coruña, 15071 A Coruña, Spain
Research Group MODES, Faculty of Computer Science, Universidade da Coruña, 15071 A Coruña, Spain
Correspondence: b.mcastillo@udc.es

DOI: *https://doi.org/10.17979/spudc.000024.29*

Abstract: The notion of distance correlation was introduced to measure the dependence between two random vectors, not necessarily of equal dimensions, in a multivariate setting. In their work, Székely et al. (2007) proposed an estimator for the squared distance covariance, and they also proved that this estimator is a V-statistic. On the other hand, Székely and Rizzo (2014) introduced an unbiased version of the squared sample distance covariance, which was subsequently identified as a U-statistic in Huo and Székely (2016). In this study, a simulation is conducted to compare both distance correlation estimators: the U-estimator and the V-estimator. The analysis assesses their efficiency (mean squared error) and contrasts the computational times of both approaches across various dependence structures.

1 Introduction

Distance correlation is a novel measure of dependence between random vectors. The concept of distance correlation was introduced by Székely et al. (2007). They emphasize that distance covariance and distance correlation draw a parallel to product-moment covariance and correlation. Nevertheless, unlike the classical definition of correlation, distance correlation is zero only when the random vectors are independent. Essentially, for all distributions with finite first moments, distance correlation ($\mathcal{R}$) extends the concept of correlation in two fundamental ways:

(i) $\mathcal{R}(X, Y)$ is defined for X and Y in arbitrary dimensions;

(ii) $\mathcal{R}(X, Y) = 0$ characterizes independence of X and Y.

Distance correlation satisfies $0 \leqslant \mathcal{R} \leqslant 1$, and $\mathcal{R} = 0$ if and only if X and Y are independent.

Székely et al. (2007) introduced a sample distance covariance estimator, demonstrating that this estimator functions as a V-statistic. Furthermore, in the work of Székely and Rizzo (2014), intermediate findings are outlined, culminating in an unbiased estimator for squared distance covariance. The unbiased estimator is established as a U-statistic in Huo and Székely (2016), along with the introduction of a novel algorithm. This algorithm shows a computational complexity of $\mathcal{O}(n \log n)$, a noteworthy enhancement compared to the $\mathcal{O}\left(n^2\right)$ complexity associated with the direct implementation of the V-estimator put forward by Székely et al. (2007).

It is important to highlight that, to the best of our knowledge, the merits and drawbacks linked to each distance correlation estimator (U-estimator and V-estimator) have not been thoroughly investigated in the current body of literature. In this study, both estimators are comprehensively examined and compared.

2 Preliminaries

Consider two random vectors, $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$, where both p and q are positive integers. Let ϕ_X and ϕ_Y represent the characteristic functions of X and Y respectively, and denote the joint characteristic function of X and Y as $\phi_{X,Y}$. In this context, the distance covariance between these random vectors X and Y, assuming they possess finite first moments, is defined as a nonnegative scalar denoted as $\mathcal{V}(X,Y)$, given by the square root of:

$$\begin{aligned}\mathcal{V}^2(X,Y) &= ||\phi_{X,Y}(t,s) - \phi_X(t)\phi_Y(s)||^2 \\ &= \frac{1}{c_p c_q}\int_{\mathbb{R}^{p+q}} \frac{||\phi_{X,Y}(t,s) - \phi_X(t)\phi_Y(s)||^2}{|t|_p^{1+p}|s|_q^{1+q}}dt\,ds,\end{aligned}$$

where $c_d = \frac{\pi^{(1+d)/2}}{\Gamma((1+d)/2)}$. Similarly, distance variance ($\mathcal{V}(X)$) is defined as the square root of

$$\mathcal{V}^2(\mathrm{X}) = \mathcal{V}^2(\mathrm{X},\mathrm{X}) = ||\phi_{X,X}(t,s) - \phi_X(t)\phi_X(s)||^2.$$

And the distance correlation ($\mathcal{R}$) between random vectors X and Y, assuming they possess finite first moments, is the positive square root of a nonnegative quantity denoted as $\mathcal{R}^2(X,Y)$. This quantity is defined as follows:

$$\mathcal{R}^2(X,Y) = \begin{cases}\frac{\mathcal{V}^2(X,Y)}{\sqrt{\mathcal{V}^2(X)\mathcal{V}^2(Y)}}, & \mathcal{V}^2(X)\mathcal{V}^2(Y) > 0 \\ 0, & \mathcal{V}^2(X)\mathcal{V}^2(Y) = 0.\end{cases} \tag{29.1}$$

Alternatively, Székely et al. (2007) proposed an equivalent method for computing distance covariance through expectations. This is, if $E|X|_p^2 < \infty$ and $E|Y|_q^2 < \infty$, then $E[|X|_p|Y|_q] < \infty$, and

$$\begin{aligned}\mathcal{V}^2(X,Y) = E[|X_1 - X_2|_p|Y_1 - Y_2|_q] + E[|X_1 - X_2|_p]E[|Y_1 - Y_2|_q] \\ -2E[|X_1 - X_2|_p|Y_1 - Y_3|_q],\end{aligned} \tag{29.2}$$

where (X_1, Y_1), (X_2, Y_2) and (X_3, Y_3) are independent and identically distributed as (X, Y).

When dealing with an observed random sample $(\mathbf{X}, \mathbf{Y}) = \{(X_k, Y_k) : k = 1, \ldots, n\}$ drawn from the joint distribution of random vectors $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$, Székely et al. (2007) introduced the empirical distance covariance ($\mathcal{V}_n(\mathbf{X}, \mathbf{Y})$) as follows. The empirical distance covariance $\mathcal{V}_n(\mathbf{X}, \mathbf{Y})$ is a nonnegative value defined by the square root of:

$$\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y}) = \frac{1}{n^2}\sum_{k,l=1}^{n} A_{kl}B_{kl}, \tag{29.3}$$

where A_{kl} and B_{kl} denote the corresponding double-centered distance matrices defined as:

$$A_{kl} = \begin{cases} a_{kl} - \frac{1}{n}\sum_{j=1}^{n} a_{kj} - \frac{1}{n}\sum_{i=1}^{n} a_{il} + \frac{1}{n^2}\sum_{i,j=1}^{n} a_{ij}, & k \neq l \\ 0, & k = l, \end{cases}$$

where $a_{kl} = |X_k - X_l|$ the pairwise distances of the X observations, similarly for $b_{kl} = |Y_k - Y_l|$. In the same way

$$\mathcal{V}_n^2(\mathbf{X}) = \mathcal{V}_n^2(\mathbf{X}, \mathbf{X}) \;=\; \frac{1}{n^2}\sum_{k,l=1}^{n} A_{kl}^2. \tag{29.4}$$

Theorem 1 in Székely et al. (2007) establishes the nonnegativity of $\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y})$. Furthermore, it demonstrates that under independence, $\mathcal{V}_n^2$ behaves as a degenerate kernel V-statistic. The

computational complexity of this estimator is $\mathcal{O}(n^2)$. Now, the empirical distance correlation $\mathcal{R}_n(\mathbf{X}, \mathbf{Y})$ is defined as the square root of:

$$\mathrm{dCorV}^2(\mathbf{X}, \mathbf{Y}) \quad = \quad \mathcal{R}_n^2(\mathbf{X}, \mathbf{Y}) = \begin{cases} \frac{\mathcal{V}_n^2(\mathbf{X},\mathbf{Y})}{\sqrt{\mathcal{V}_n^2(\mathbf{X})\mathcal{V}_n^2(\mathbf{Y})}}, & \mathcal{V}_n^2(\mathbf{X})\mathcal{V}_n^2(\mathbf{Y}) > 0 \\ 0, & \mathcal{V}_n^2(\mathbf{X})\mathcal{V}_n^2(\mathbf{Y}) = 0, \end{cases} \tag{29.5}$$

which is always non-negative.

Similarly, in Székely and Rizzo (2014), the $\mathcal{U}$-centered matrix is introduced as follows. Consider a symmetric, real-valued $n \times n$ matrix $A = (a_{kl})$ with a zero diagonal, where $n > 2$. The entry in the (k, l)th position of the $\mathcal{U}$-centered matrix $\tilde{A}$ is defined as:

$$\tilde{A}_{kl} = \begin{cases} a_{kl} - \frac{1}{n-2}\sum_{j=1}^{n} a_{kj} - \frac{1}{n-2}\sum_{i=1}^{n} a_{il} + \frac{1}{(n-1)(n-2)}\sum_{i,j=1}^{n} a_{ij}, & k \neq l; \\ 0, & k = l. \end{cases}$$

Here "$\mathcal{U}-$centered" is so named because the inner product,

$$\mathcal{U}_n^2(\mathbf{X}, \mathbf{Y}) \quad = \quad \left(\tilde{A} \cdot \tilde{B}\right) = \frac{1}{n-3}\sum_{i \neq j} \tilde{A}_{kl}\tilde{B}_{kl}, \tag{29.6}$$

defines an unbiased estimator of the squared distance covariance. The work by Huo and Székely (2016) established that the estimator in Equation (29.6) is a U-statistic. This reevaluation paved the way for the creation of an efficient algorithm, which can be executed with a computational complexity of $\mathcal{O}(n \log n)$.
Thus, it is possible to define the empirical distance correlation through U-statistics (dCorU) which is the square root of

$$\mathrm{dCorU}^2(\mathbf{X}, \mathbf{Y}) = \begin{cases} \frac{\mathcal{U}_n^2(\mathbf{X},\mathbf{Y})}{\sqrt{\mathcal{U}_n^2(\mathbf{X})\mathcal{U}_n^2(\mathbf{Y})}}, & \mathcal{U}_n^2(\mathbf{X})\mathcal{U}_n^2(\mathbf{Y}) > 0 \\ 0, & \mathcal{U}_n^2(\mathbf{X})\mathcal{U}_n^2(\mathbf{Y}) = 0, \end{cases} \tag{29.7}$$

where $\mathcal{U}_n^2(\mathbf{X})$ represents the distance variance of $\mathbf{X}$, similarly $\mathcal{U}_n^2(\mathbf{Y})$ for $\mathbf{Y}$.

These results have spurred the development and advancement of numerous software packages, accessible for use in both the R software environment (R Core Team, 2022) and Python (Van Rossum and Drake Jr, 1995). In the realm of Python, libraries such as statsmodels (Seabold and Perktold, 2010), hyppo (Panda et al., 2021), dcor (Ramos-Carreño, 2022), and pingouin (Vallat, 2018) are available. In the R environment, notable packages include energy (Rizzo and Székely, 2022), dcortools (Edelmann and Fiedler, 2022), and the Rfast package (Papadakis et al., 2022).

3 Simulation study

A Monte Carlo simulation study was conducted in order to compare the efficiency of the dCorU and dCorV estimators across diverse dependence structures. For the simulation study, the dcortools package was utilized, specifically employing the distcor function. To calculate the distance correlation through dCorU, the code used is distcor(X, Y, bias.corr = TRUE). While, to compute dCorV, the code is distcor(X, Y, bias.corr = FALSE) or simply distcor(X, Y).

To facilitate a comprehensive comparison of each estimator's efficiency (MSE) and computational time, two different models are utilized. The first one, the bivariate normal model, which

is used because Székely et al. (2007) proved that the distance correlation, in terms of Pearson's correlation coefficient, is given by the square root of:

$$\mathcal{R}^2(X,Y) = \frac{\rho \arcsin \rho + \sqrt{1-\rho^2} - \rho \arcsin \rho/2 - \sqrt{4-\rho^2} + 1}{1 + \pi/3 - \sqrt{3}}.$$

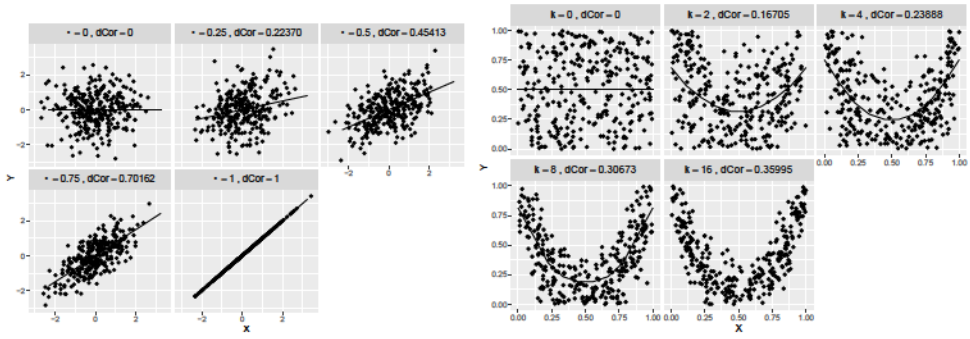
The second model corresponds to a nonlinear model defined as:

$$f_{X,Y}(x,y) = c\left[1 - \left(y - 4\left(x - \frac{1}{2}\right)^2\right)^2\right]^k I_{[0,1]}(x) I_{[0,1]}(y),$$

where $k \in \mathbb{N}$ and c is a constant that depends on the value of k.

These models encompass varying levels of dependence and are evaluated across two sample sizes, 100 and 10000. Each simulation is performed with 1000 Monte Carlo repetitions. The precise computation of the distance covariance for the second model is achieved using Equation (29.2). Similarly, the calculations for $\mathcal{V}(X)$ and $\mathcal{V}(Y)$ were performed. Finally, the value of $\mathcal{R}$ is obtained from Equation (29.1).

Five samples drawn from different values of the parameters of each model are depicted in Figure 1, accompanied by the respective parameter value and the corresponding distance correlation. The lines represent the conditional mean $E[Y|X = x]$ in each case.



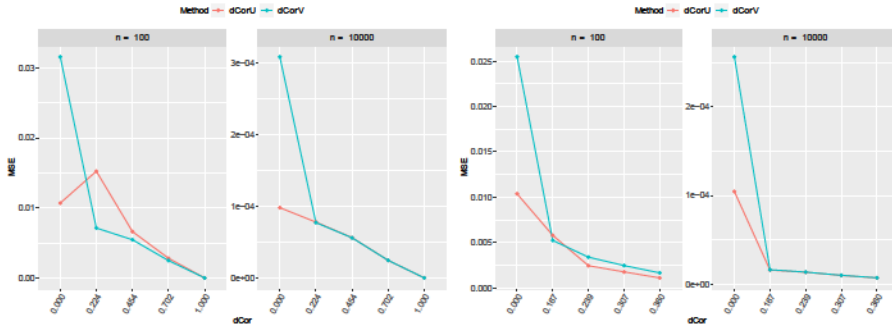
(a) Bivariate normal model samples for different values of ρ.

(b) Nonlinear model samples for different values of k.

Figure 1: Samples ($n = 300$) for different values of ρ, k and their corresponding distance correlation for the respective model.

In the bivariate normal model, complete dependence between X and Y results in $\mathcal{R} = 1$. However, the maximum value of the distance correlation for the nonlinear model, reached when X and Y are totally dependent (i.e. $k \to \infty$), is $\mathcal{R} = 0.41$.

The MSE results for each model are illustrated in Figure 2. Several observations can be made: under independence, dCorU outperforms dCorV in the two cases for both sample sizes. However, when it comes to cases of dependence, whether weak or strong, dCorV shows better performance than dCorU. These disparities are more noticeable with weak dependence and smaller sample sizes. Interestingly, this pattern deviates for the nonlinear model (depicted in Figure 2b), where, even with weak dependence, dCorU surpasses dCorV, although the differences are not substantial. Finally, for a larger sample size ($n = 10000$), both estimators exhibit remarkable similarity across both models.



(a) MSE under the bivariate normal model from $\mathcal{R} = 0$ ($\rho = 0$) to $\mathcal{R} = 1$ ($\rho = 1$).

(b) MSE under the nonlinear model from $\mathcal{R} = 0$ ($k = 0$) to $\mathcal{R} = 0.36$ ($k = 16$).

Figure 2: MSE of dCorU and dCorV under the two models for different distance correlation values.

Finally, the computational time for each estimator are in Table 1. The characteristics of the computer equipment used are the following ones: CPU 12th Gen Intel(R) Core(TM) i7-1280P 2.00 GHz and RAM 16 GB.

Note that there are no substantial differences in the computation times among each of the estimators. However, it is worth noting that these times were acquired using the dcortools package. If a different package, such as energy package, were employed, the timings would likely vary, potentially resulting in an increase.

Table 1: Computational time in secs for 1000 samples.

	$n = 100$		$n = 1000$		$n = 10000$	
	dCorU	dCorV	dCorU	dCorV	dCorU	dCorV
Time	0.36	0.33	0.63	0.58	2.22	2.51

4 Conclusions

This study focused on examining the performance of the dCorU and dCorV estimators for distance correlation through Monte Carlo simulations. The results presented here underscore the significance of the specific scenario. In cases of independence, dCorU demonstrates superiority over dCorV across all considered scenarios. However, when dependence is present, the outcomes diverge. The dCorV estimator aligns with superior results in terms of Mean Squared Error (MSE) for linear model (bivariate normal), as well as for the nonlinear model under weak dependence. Moreover, in the realm of computational efficiency, both estimators, dCorU and dCorV, stand on competitive ground.

Acknowledgements

CITIC is funded by the Xunta de Galicia through the collaboration agreement between the Consellería de Cultura, Educación, Formación Profesional e Universidades and the Galician universities for the reinforcement of the research centres of the Galician University System (CIGUS).

Bibliography

D. Edelmann and J. Fiedler. *dcortools: Providing Fast and Flexible Functions for Distance Correlation Analysis*, 2022. URL *https://CRAN.R-project.org/package=dcortools*. R package version 0.1.6.

X. Huo and G. J. Székely. Fast Computing for Distance Covariance. *Technometrics*, 58(4):435–447, 2016.

S. Panda, S. Palaniappan, J. Xiong, E. Bridgeford, R. Mehta, and C. Shen. *hyppo: A multivariate hypothesis testing Python package*, 2021. URL *https://github.com/neurodata/hyppo.*

M. Papadakis, M. Tsagris, M. Dimitriadis, S. Fafalios, I. Tsamardinos, M. Fasiolo, G. Borboudakis, J. Burkardt, C. Zou, K. Lakiotaki, and C. Chatzipantsiou. *Rfast: A Collection of Efficient and Extremely Fast R Functions*, 2022. URL *https://CRAN.R-project.org/package=Rfast*. R package version 2.0.6.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022. URL *https://www.R-project.org/*.

C. Ramos-Carreño. *dcor: distance correlation and energy statistics in Python*, 2022. URL *https://pypi.org/project/dcor/*.

M. L. Rizzo and G. J. Székely. *energy: E-Statistics: Multivariate Inference via the Energy of Data*, 2022. URL *https://CRAN.R-project.org/package=energy*. R package version 1.7-11.

S. Seabold and J. Perktold. *Statsmodels: Econometric and statistical modeling with Python*, 2010. URL *https://github.com/statsmodels/statsmodels/*.

G. J. Székely and M. L. Rizzo. Partial Distance Correlation with Methods for Dissimilarities. *Annals of Statistics*, 42(6):2382–2412, 2014.

G. J. Székely, M. L. Rizzo, and N. K. Bakirov. Measuring and testing dependence by correlation of distances. *Annals of Statistics*, 35(6):2769–2794, 2007.

R. Vallat. Pingouin: statistics in Python. *Journal of Open Source Software*, 3(31):1026, 2018.

G. Van Rossum and F. L. Drake Jr. *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam, 1995.