Journal of Multivariate Analysis 154 (2017) 19–39

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Distance correlation coefficients for Lancaster distributions

Johannes Dueck [a], Dominic Edelmann [b], Donald Richards [c,*]

[a] Institute of Applied Mathematics, University of Heidelberg, Im Neuenheimer Feld 294, 69120 Heidelberg, Germany
[b] Division of Biostatistics, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, 69120 Heidelberg, Germany
[c] Department of Statistics, Pennsylvania State University, University Park, PA 16802, USA

ARTICLE INFO

Article history:
Received 2 July 2015
Available online 31 October 2016

AMS 2010 subject classifications:
primary 60E05
62H20
secondary 33C05
42C05
60E10

Keywords:
Affine invariance
Bivariate gamma distribution
Bivariate normal distribution
Bivariate negative binomial distribution
Bivariate Poisson distribution
Characteristic function
Distance correlation coefficient
Lancaster distributions
Multivariate normal distribution

ABSTRACT

We consider the problem of calculating distance correlation coefficients between random vectors whose joint distributions belong to the class of Lancaster distributions. We derive under mild convergence conditions a general series representation for the distance covariance for these distributions. To illustrate the general theory, we apply the series representation to derive explicit expressions for the distance covariance and distance correlation coefficients for the bivariate normal distribution and its generalizations of Lancaster type, the multivariate normal distributions, and the bivariate gamma, Poisson, and negative binomial distributions which are of Lancaster type.

1. Introduction

The concepts of distance covariance and distance correlation, introduced by Székely, et al. [27,31], have been shown to be widely applicable for measuring dependence between collections of random variables. As examples of the ubiquity of distance correlation methods, we note the results on distance correlation given recently by: Székely, et al. [21,28–31], on statistical inference; Sejdinovic, et al. [26], on machine learning; Kong, et al. [10], on familial relationships and mortality; Zhou [33], on nonlinear time series; Lyons [17], on abstract metric spaces; Martínez-Gómez, et al. [18] and Richards, et al. [20], on large astrophysical databases; Dueck, et al. [5], on high-dimensional inference and the analysis of wind data; and Dueck, et al. [6], on a connection with singular integrals on Euclidean spaces.

A result which is of fundamental importance in distance correlation theory is the explicit formula for the empirical distance correlation coefficient [31, pp. 2773–2774]. By combining that explicit formula with the fast algorithm of Huo and Székely [9], it becomes straightforward to apply distance correlation methods to real-world data sets.

On the other hand, the calculation of population distance correlation coefficients remains an intractable problem generally. Székely, et al. [31, pp. 2785–2786] calculated the distance correlation coefficient for the bivariate normal

* Corresponding author.
E-mail address: richards@stat.psu.edu (D. Richards).

http://dx.doi.org/10.1016/j.jmva.2016.10.012
0047-259X/

distribution; Dueck, et al. [4, Appendix] extended that result to the general multivariate normal distribution; and Dueck, et al. [5] calculated the affinely invariant distance correlation coefficient for the multivariate normal distribution. Otherwise, no such results are yet available for any other distribution. Hence, the state of distance correlation theory hitherto is that the empirical coefficients can be calculated readily but the opposite holds for their population counterparts, generally. Consequently, it was not possible to calculate distance correlation coefficients explicitly for given nonnormal distributions in terms of the usual parameters that parametrize these distributions, or to ascertain for nonnormal distributions any analogs of the limit theorems derived by Dueck, et al. [5, Section 4].

We describe in detail the difficulties arising in attempts to calculate the population distance correlation coefficients. Let p and q be positive integers. For column vectors $s \in \mathbb{R}^p$ and $t \in \mathbb{R}^q$, denote by $\|s\|$ and $\|t\|$ the standard Euclidean norms on the corresponding spaces; thus, if $s = (s_1, \ldots, s_p)^\top$ then $\|s\| = (s_1^2 + \cdots + s_p^2)^{1/2}$, and similarly for $\|t\|$. Given vectors u and v of the same dimension, we let $\langle u, v\rangle$ be the standard Euclidean scalar product of u and v. For jointly distributed random vectors $(X, Y) \in \mathbb{R}^p \times \mathbb{R}^q$ and non-random vectors $(s, t) \in \mathbb{R}^p \times \mathbb{R}^q$, let

$$\psi_{X,Y}(s, t) = \mathrm{E}\exp\big[\,\mathrm{i}\langle s, X\rangle + \mathrm{i}\langle t, Y\rangle\big],$$

$\mathrm{i} = \sqrt{-1}$, be the joint characteristic function of (X, Y), and let $\psi_X(s) = \psi_{X,Y}(s, 0)$ and $\psi_Y(t) = \psi_{X,Y}(0, t)$ be the corresponding marginal characteristic functions. For any $z \in \mathbb{C}$, let $|z|^2$ denote the squared modulus of z; also, we use the notation

$$\gamma_p = \frac{\pi^{(p+1)/2}}{\Gamma\big((p+1)/2\big)}. \tag{1.1}$$

In the case of distributions with finite first moments, Székely, et al. [31, p. 2772] defined $\mathcal{V}(X, Y)$, the *distance covariance* between X and Y, to be the positive square-root of

$$\mathcal{V}^2(X, Y) = \frac{1}{\gamma_p\gamma_q}\int_{\mathbb{R}^{p+q}} \frac{|\psi_{X,Y}(s, t) - \psi_X(s)\psi_Y(t)|^2}{\|s\|^{p+1}\,\|t\|^{q+1}}\,\mathrm{d}s\,\mathrm{d}t \tag{1.2}$$

and they defined the *distance correlation coefficient* between X and Y as

$$\mathcal{R}(X, Y) = \frac{\mathcal{V}(X, Y)}{\sqrt{\mathcal{V}(X, X)\mathcal{V}(Y, Y)}} \tag{1.3}$$

if both $\mathcal{V}(X, X)$ and $\mathcal{V}(Y, Y)$ are strictly positive, and otherwise to be zero [31, p. 2773]. For distributions with finite first moments we have $0 \le \mathcal{R}(X, Y) \le 1$, and $\mathcal{R}(X, Y) = 0$ if and only if X and Y are mutually independent.

For given random vectors X and Y, the fundamental obstacle in calculating the population distance correlation coefficient (1.3) is the computation of the singular integral (1.2). In particular, the singular nature of the integrand precludes evaluation of the integral by expanding the numerator, $|\psi_{X,Y}(s, t) - \psi_X(s)\psi_Y(t)|^2$, and subsequent term-by-term integration of each of the resulting three terms.

In this paper, we calculate the distance correlation coefficients for pairs (X, Y) of random vectors whose joint distributions are in the class of *Lancaster distributions*, a class of probability distributions made prominent by Lancaster [15,16] and Sarmanov [24]. The distribution functions of the Lancaster family are well-known to have attractive expansions in terms of certain orthogonal functions (Koudou [14]; Diaconis, et al. [3]). By applying those expansions, we obtain explicit expressions for the distance covariance and distance correlation coefficients.

Consequently, we derive under mild convergence conditions a general formula for the distance covariance for the Lancaster distributions. We apply the general formula to obtain explicit expressions for the distance covariance and distance correlation for the bivariate normal distributions and some of its generalizations, for the multivariate normal distributions, and for bivariate gamma, Poisson, and negative binomial distributions. We remark that explicit results can also be obtained for other Lancaster-type expansions obtained by Bar-Lev, et al. [2]; however, we will omit the details for other cases because the formulas derived here are entirely representative of other cases.

2. The Lancaster distributions

To recapitulate the class of Lancaster distributions we generally follow the standard notation in that area, as given by Koudou [13,14]; cf., Lancaster [16], Pommeret [19], or Diaconis, et al. [3, Section 6].

Let $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$ be locally compact, separable probability spaces, such that $L^2(\mu)$ and $L^2(\nu)$ are separable. Let σ, a probability measure on $\mathcal{X} \times \mathcal{Y}$, have marginal distributions μ and ν; then there exist functions K_σ and L_σ such that

$$\sigma(\mathrm{d}x, \mathrm{d}y) = K_\sigma(x, \mathrm{d}y)\mu(\mathrm{d}x) = L_\sigma(\mathrm{d}x, y)\nu(\mathrm{d}y).$$

We note that K_σ and L_σ represent the conditional distributions of Y given $X = x$, and X given $Y = y$, respectively.

Let $\mathcal{C}$ denote a countable index set with a zero element, denoted by 0. Let $\{P_n : n \in \mathcal{C}\}$ and $\{Q_n : n \in \mathcal{C}\}$ be sequences of functions on $\mathcal{X}$ and $\mathcal{Y}$ which form orthonormal bases for the separable Hilbert spaces $L^2(\mu)$ and $L^2(\nu)$, respectively. We assume, by convention, that $P_0 \equiv 1$ and $Q_0 \equiv 1$.