Accepted Manuscript

Generalizing distance covariance to measure and test multivariate mutual dependence via complete and incomplete V-statistics

Ze Jin, David S. Matteson

PII: S0047-259X(17)30783-2
DOI: https://doi.org/10.1016/j.jmva.2018.08.006
Reference: YJMVA 4401

To appear in: *Journal of Multivariate Analysis*

Received date : 29 December 2017

Please cite this article as: Z. Jin, D.S. Matteson, Generalizing distance covariance to measure and test multivariate mutual dependence via complete and incomplete V-statistics, *Journal of Multivariate Analysis* (2018), https://doi.org/10.1016/j.jmva.2018.08.006

Generalizing distance covariance to measure and test multivariate mutual dependence via complete and incomplete V-statistics

Ze Jin[a,*], David S. Matteson[a]

[a]Department of Statistical Science, Cornell University, Ithaca, NY 14853, USA

Abstract

We propose three new measures of mutual dependence between multiple random vectors. Each measure is zero if and only if the random vectors are mutually independent. The first generalizes distance covariance from pairwise dependence to mutual dependence, while the other two measures are sums of squared distance covariances. The proposed measures share similar properties and asymptotic distributions with distance covariance, and capture non-linear and non-monotone mutual dependence between the random vectors. Inspired by complete and incomplete V-statistics, we define empirical and simplified empirical measures as a trade-off between the complexity and statistical power when testing mutual independence. The implementation of corresponding tests is demonstrated by both simulation results and real data examples.

Keywords: Characteristic functions, Distance covariance, Multivariate analysis, Mutual independence, V-statistics

1. Introduction

Let $X = (X_1, \dots, X_d)$ be a set of variables where each component is a random vector, and let $\mathbf{X} = \{X^1, \dots, X^n\}$ be a random sample from F_X, the joint distribution of X. We are interested in testing the hypotheses

$$\mathcal{H}_0 : X_1, \dots, X_d \text{ are mutually independent} \quad \text{vs.} \quad \mathcal{H}_A : X_1, \dots, X_d \text{ are dependent.}$$

This problem has many applications, including independent component analysis [16, 26], graphical models [8, 10, 22, 23], naive Bayes classifiers [38, 40], causal inference [5, 25], etc. It has been studied under different settings and assumptions, including pairwise ($d = 2$) and mutual ($d \geq 2$) independence, univariate ($X_1, \dots, X_d \in \mathbb{R}$) and multivariate ($X_1 \in \mathbb{R}^{p_1}, \dots, X_d \in \mathbb{R}^{p_d}$) components, and more. Here we consider the general case where $X_1, \dots, X_d$ are not assumed jointly normal.

The most extensively studied case is pairwise independence with univariate components ($X_1, X_2 \in \mathbb{R}$). Rank correlation is considered as a nonparametric counterpart to Pearson's product-moment correlation [28], including Kendall's tau [19], Spearman's rho [32], etc. Bergsma and Dassios [2] proposed a test based on an extension of Kendall's tau, testing an equivalent condition to $\mathcal{H}_0$. Additionally, Hoeffding [15] proposed a nonparametric test based on marginal and joint distribution functions, testing a necessary condition to investigate $\mathcal{H}_0$.

For pairwise independence with multivariate components ($X_1 \in \mathbb{R}^{p_1}, X_2 \in \mathbb{R}^{p_2}$), Székely et al. [37] and Székely and Rizzo [34] proposed a test based on distance covariance with fixed p_1, p_2 and $n \to \infty$ testing an equivalent condition to $\mathcal{H}_0$; this has been extended to martingale difference divergence in [31] with [17] testing conditional mean independence. Under the same setting, Gretton et al. [13] proposed a test based on Hilbert−Schmidt independence criterion (HSIC), which is 0 if and only if pairwise independence holds. Further, Székely and Rizzo [35] proposed a t-test based on a modified distance covariance for the setting in which n is finite and $p_1, p_2 \to \infty$, testing an equivalent condition to $\mathcal{H}_0$ as well.

For mutual independence with univariate components ($X_1, \dots, X_d \in \mathbb{R}$), one natural way to extend the pairwise rank correlation to multiple components is to collect the rank correlations between all pairs of components, and examine the norm ($\mathcal{L}_2, \mathcal{L}_\infty$) of this collection. Leung and Drton [21] proposed a test based on the $\mathcal{L}_2$ norm with $n, d \to \infty$, and $d/n \to \gamma \in (0, \infty)$, and Han et al. [14] proposed a test based on the $\mathcal{L}_\infty$ norm with $n, d \to \infty$, and $d/n \to \gamma \in [0, \infty]$. Each are testing a necessary condition to $\mathcal{H}_0$, in general.

*Corresponding author
Email addresses: zj58@cornell.edu (Ze Jin), matteson@cornell.edu (David S. Matteson)