Journal of Multivariate Analysis 153 (2017) 189–210

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Multivariate nonparametric test of independence

Yanan Fan [a], Pierre Lafaye de Micheaux [a,b], Spiridon Penev [a,*], Donna Salopek [a]

[a] The University of New South Wales, Department of Statistics, Sydney, 2052 NSW, Australia

[b] CREST, ENSAI, Campus de Ker Lann, Rue Blaise Pascal-BP37203, 35172 BRUZ CEDEX, France

ARTICLE INFO

Article history:
Received 16 December 2015
Available online 12 October 2016

AMS 2000 subject classifications:
62G20
62P05

Keywords:
Central limit theorem
Empirical characteristic function
Multivariate K sample independence

ABSTRACT

The problem of testing mutual independence of p random vectors in a general setting where the dimensions of the vectors can be different and the distributions can be discrete, continuous or both is of great importance. We propose such a test which utilizes multivariate characteristic functions and is a generalization of known results. We characterize the limiting distribution of the test statistic under the null hypothesis. The limiting null distribution is approximated and the method is validated. Numerical results based on simulations are investigated and our methodology is implemented in the R package `IndependenceTests`. Power comparisons are also presented for some partial cases of our general test, where some competitive procedures exist.

1. Introduction

Very often, at certain stage of a statistical inference procedure, the question arises if a certain number p of random vectors (with any combination of component sub-vectors) can be assumed to be mutually independent. Such tests are of great importance to functional magnetic resonance imaging (fMRI) for example. In this situation, it is important to find which areas of the brain are involved in certain activities and if there are any statistical associations between these brain activities. Recovering the underlying components of these activities from a given set of linearly mixed observations can be an ill-defined problem. However, the assumption of independence among the sources provides a surprisingly powerful and effective technique for a wide range of problems in various practical domains. One of the techniques to deal with this is Independent Component Analysis (ICA). The main goal of ICA is to extract multivariate sources by using an explicit assumption about the independence of the sources. Therefore, the need for testing for multivariate independence arises naturally.

In the influential paper [1], the distance correlation is introduced as a new dependence measure between two random vectors. The empirical version of this measure is easily computable in their R package called `Energy`. This distance correlation is based on functionals of the characteristic functions so it can cover a wide variety of situations such as purely continuous, purely discrete, or mixed components of the random vectors. More details on the properties of this distance correlation, the issue of testing independence of two random vectors, and uniqueness are discussed in [29–31].

When the number of vectors p is equal to 2, the distance correlation and its latest extensions are indicative about independence and could be used to construct an independence test, too. In fact, [31, p. 2783] discusses such a test. However, when the number of vectors p is larger than 2, it is not sufficient to just look at pairs of vectors only when testing for mutual independence. Hence there is a need to generalize the test of multivariate independence for the case of $p > 2$.

* Corresponding author.
E-mail addresses: y.fan@unsw.edu.au (Y. Fan), pierre.lafaye-de-micheaux@ensai.fr (P.L. de Micheaux), s.penev@unsw.edu.au (S. Penev), dm.salopek@unsw.edu.au (D. Salopek).

http://dx.doi.org/10.1016/j.jmva.2016.09.014
0047-259X/

Such a test for $p > 2$, based on a Cramér–von Mises type functional of a process defined from the empirical characteristic functions, has been proposed earlier in [3]. However, in order to determine the asymptotic distribution of the resulting test-statistic, the authors resort to a simplifying assumption. Although they do not assume joint multivariate normality, they still assume that each of the sub-vectors is marginally normally distributed. [23] also deals with the case of arbitrary $p > 2$, but only tests for joint independence of all the *p univariate* components using only Monte Carlo approximations of the distribution of the test-statistic. The paper [19] also discusses testing independence based on empirical characteristic functions, but are only concerned with the joint independence of all the *p univariate* components. They also modify (by staying within the framework of testing joint independence of all the *p* components) their original test that uses values of the empirical characteristic functions on compact supports only. The assumption of compact support is restrictive since a test of such type would be inconsistent against virtually all possible deviations from the null hypothesis (there are counterexamples of two different distribution functions with identical characteristic functions on a compact interval (see [32])). However, considering characteristic functions or their empirical counterparts on the whole space (to avoid test inconsistency) introduces many technical difficulties (especially because of their periodicity).

The modified test of [32] has an asymptotically distribution free version for testing independence of all *p univariate* components only, but this modification can only be applied when all components of the joint distribution are *continuous*. In another development, [2] proposes a non-parametric test of independence between random *vectors* based on characterization of mutual independence defined from probabilities of half-spaces in a combinatorial formula of Möbius. Their paper, which generalizes [14], is related to our current project. They use the combinatorial formula expressed by using the cumulative distribution functions of (sub)vectors. Now our approach, by using the characteristic functions, allows us to easily accommodate continuous, discrete, and mixed components. In addition, our approach is supported by earlier works showing that goodness-of-fit and independence tests based on the empirical characteristic functions are very competitive for testing the real multinormal distribution (see, e.g., [16,22]).

As pointed out in [2], without the assumption that each sub-vector is one-dimensional with a continuous cumulative distribution function, any test of independence can no longer be distribution free. Hence both tests in [2] and in the current paper, are naturally not distribution free. The paper [2] deals with this issue by computing bootstrap approximations whereas we evaluate the asymptotic distribution of our test statistic.

With the recent advances of copulas in statistical applications, another focus of a stream of papers has been in deriving tests of independence among random vectors based on Cramér–von Mises functionals of the empirical copula process. The paper [13] studies the limiting distribution of such statistics under contiguous sequences of alternatives and analyzes asymptotic relative efficiencies in some classes of copula alternatives. A summary of the efforts in this direction, as well as further references can be seen in [20,25]. The copula approach is tempting to apply in this setting on ideological grounds. Indeed, the copula function is meant to precisely "extract" the dependence structure by "leaving aside" the marginals thus making a copula-based approach very suitable to testing multivariate independence. Both papers utilize the same empirical process based on the characterization of stochastic vectorial independence in terms of copulas to construct the test-statistic. However, this approach is limited only to *continuous* random vectors. Besides, their asymptotic distribution arising from the testing procedure, is even further restricted by the requirement that the copula has continuous partial derivatives. They also apply bootstrap approximations which may be more computationally intensive than our proposed procedures. For the above reasons we do not consider further comparison of our methods with copula-based approaches in this paper.

Our paper proposes a test in the most general form. That is, our test statistic is based on a Cramér–von Mises type functional of a process defined from the empirical characteristic functions, does not need any of the restrictive assumptions such as in [3] and treats the case when $p \geq 2$, where each of the p components can itself be a vector of arbitrary length $q_i, i = 1, \ldots, p$. We offer a theory about the asymptotic distribution of our characteristic functions-based test statistic as opposed to the use of the bootstrap approach as in [2]. This allows us to determine the p-values for our test and are able to reduce the computational time in some cases without compromising the ability of the test to keep the correct level of significance asymptotically (see Section 6). Our paper, by stating the asymptotic distribution of the test statistic, represents a generalization of [1] and of [28]. We note that in [1], (for the case of $p = 2$ only), a bound of the control of the first type error of the test is given. This bound is fast to compute but being a bound only, may not be accurate enough. Further, we solve the issues with numerical calculation of our test statistic via its decomposition into feasible components and calculate critical values numerically based on asymptotic approximations. We also have performed extensive investigation of the power of our test. We have included an example of the case $p \geq 3$ where, to the level of generality considered, the only competitor to our test is the bootstrap-based procedure from the program `dependogram` in [2], and have demonstrated the favourable performance of the new test in terms of computational time and ability to keep the size close to the nominal level. For the case of $p = 2$ we have compared our test to the tests of [2,15,28] and have demonstrated the favourable performance of our test in comparison to all competitors. We also offer a variety of weight functions thus further extending the applicability of our testing procedure.

The paper is organized as follows. In Section 2, we define our test statistic and in Section 3, we study its asymptotic distribution. We discuss the normalization of our test statistic in Section 4 and in Section 5, we propose and investigate several choices of weight functions in our construction of the test statistics. In Section 6, we discuss the numerical implementation. We also demonstrate the favourable performance of our test both under the null hypothesis and under the alternative, on specific simulated examples, by comparing it to the tests from [15,28]. Section 7 contains the proofs.