Some tests of independence between two random vectors in arbitrary dimensions

S. Sarkar[1*], M. Biswas[2] and A. K. Ghosh[1]

[1] *Theoretical Statistics and Mathematics Unit, Indian Statistical Institute, Kolkata.*
[2] *Department of Statistics, Brahmananda Keshab Chandra College, Kolkata.*
email : sohamsarkar1991@gmail.com, munmun.biswas08@gmail.com, akghosh@isical.ac.in.
[*] *Presenting author*

Keywords. *Distribution-free test; High-dimensional data; Inter-point distance; Minimal spanning tree; Multiscale approach*

Given a random sample from the distribution of $\mathbf{Z} = [\mathbf{X}'\,\mathbf{Y}']'$, we want to test whether $\mathbf{X} \in \mathbb{R}^p$ and $\mathbf{Y} \in \mathbb{R}^q$ are independent. There are several methods for this test in the literature, both in parametric and nonparametric regime. Most of these methods are not applicable when the dimension of the data is larger than the sample size. Recently some new methods have been proposed for this problem, which are based on inter-point distances of $\mathbf{X}$- and $\mathbf{Y}$-samples. Among them **?** proposed some tests using random traversal on minimal spanning trees, which are exactly distribution-free under the null hypothesis of independence. We identify some shortcomings of these tests and propose some modifications which rectify these problems, keeping the distribution-free property intact. Several simulated and real data sets are analyzed which demonstrate the superiority of our proposed methods.

However, all these above mentioned distribution-free tests sacrifice a lot of valuable information to achieve distribution-free property. Moreover, they may not yield the same result if the roles of $\mathbf{X}$ and $\mathbf{Y}$ are interchanged. We propose some new tests based on nearest neighbors, which use the unused information to come up with better performance, and are symmetric with respect to $\mathbf{X}$ and $\mathbf{Y}$. Multiscale versions of these tests are also developed. Performances of all these tests are evaluated using several simulated and real data sets.

Since our tests are based on ranks of inter-point distances, they are applicable to high-dimensional data and even for functional data taking values in infinite dimensional Banach spaces.

References

Heller, R., Gorfine, M., & Heller, Y. (2012). A class of multivariate distribution-free tests of independence based on graphs. *J. Statist. Plann. Inf.*, **142**, 3097-3106.