MULTI-SAMPLE COMPARISON OF COPULAS AND CLUSTERING

Yves I. Ngounou Bakam[1] & Denys Pommeret[1,2]

[1] *CNRS, Ecole Centrale, I2M, Aix-Marseille Univ, Marseille, France*
yves-ismael.ngounou-bakam@univ-amu.fr
[2] *ISFA, Univ Lyon, UCBL, LSAF EA2429, F-69007, Lyon, France*
denys.pommeret@univ-amu.fr

Abstract. The comparison of copulas is a major challenge in copula modelling. Copulas are still an important topic with many applications as finance, actuarial science, environmental data, ecology, and so one. They can be used to model the dependence structures of multivariate observations.
In the two-sample case, Rémillard and Scaillet (2009) proposed a test to compare two nonparametric copulas, that is to test $H_0 : C1 = C2$, where C_1 and C_2 are two copulas observed on two iid samples, which may be paired.
To our knowledge, there is no extension to the K sample case. However, the increasing amount of data requires sometimes more comprehensive analyzes. It is in this sense that we propose an equality test of K copulas simultaneously when K populations are observed. We propose to test the following hypothesis:

$$H_0 : \quad C_1 = \cdots = C_K,$$

from K iid samples, possibly paired. It is therefore a generalization of Rémillard and Scaillet (2009). However, we obtain the exact asymptotic distribution of the test statistic and the convergence of the test. The idea of the test is to transform the observations to uniform laws, then to use the decomposition of the density of the copula in the Legendre polynomials orthogonal basis. Returning to the copula function we obtain what are called copula coefficients which characterize each copula. The test then amounts to simultaneously comparing these coefficients. The number of involved copula coefficients can growth with the sample size and is automatically selected by a data driven selection. We provide some illustrations of this method, in particular we suggest a clustering algorithm to classify populations with similar forms of dependence structure. A simulation study, analyse the level and the power of the test to show the good behavior of our test procedure and its performances to Rémillard and Scaillet approach within two-sample case. We provide an illustration through a real insurance dataset to demonstrate the method.

Keywords. Copula, Copula coefficients, Multi-sample, Clustering, Data driven.