K-Sample Test for Equality of Copulas

Yves Ismaël Ngounou Bakam, Denys Pommmeret

Submitted to the Annals of Statistics

K-SAMPLE TEST FOR EQUALITY OF COPULAS

BY YVES I. NGOUNOU BAKAM [1] AND DENYS POMMMERET [1,2]

[1]*Aix-Marseille University, Ecole Centrale, CNRS, I2M, Campus de Luminy, 13288 Marseille cedex 9, France, yves-ismael.ngounou-bakam@univ-amu.fr*

[2]*ISFA, Univ Lyon, UCBL, LSAF EA2429, F-69007, Lyon, France, denys.pommeret@univ-amu.fr*

We propose a test procedure to compare simultaneously K copulas, with $K \geq 2$. The K observed populations can be paired. The test statistic is based on the differences between orthogonal projection coefficients associated to the density copulas, that we called *copula coefficients*. The procedure is data driven and we obtain a chi-square asymptotic distribution of the test statistic under the null. We illustrate our procedure via numerical studies and through two real datasets. Eventually, a clustering algorithm is deduced from the K-sample test and its performances are illustrated in a simulation experiment.

1. Introduction and motivations. Copulas have been extensively studied in the statistical literature and their field of application covers a very wide variety of areas (see for instance the book of Joe (2014) and references therein). The problem of goodness-of-fit for copulas is therefore an important topic and can deserve many situations as in insurance to compare the dependence between portfolios (see for instance Shi, Feng and Boucher (2016)), in finance to compare the dependence between indices (see for instance the book of Cherubini, Luciano and Vecchiato (2004)), in biology to compare dependence between genes (Kim et al., 2008), in medicine to compare diagnostics (see for instance Hoyer and Kuss (2018)), or more recently in ecology to compare dependence between species (see Ghosh et al. (2020)).

In the one-sample case, many testing methods have been proposed within the frame of parametric families of copulas (see for instance the review paper of Genest, Remillard and Beaudoin (2009), or more recently Omelka, Gijbels and Veraverbeke (2009), Can et al. (2015) and Can, Einmahl and Laeven (2020)).

Despite this attractiveness and the continuous increase of data, little work has been done in the K-sample case, for $K > 1$. When $K = 2$ an important reference is the work of Rémillard and Scaillet (2009). They proposed a non-parametric test based on the integrated square difference between the empirical copulas. Their approach requires the continuity of partial derivatives of copulas which permits to obtain an approximation of the distribution under the null. It is adapted to independent as well as paired populations and a R package is available in Remillard and Plante (2012).

When $K > 2$, there is no test procedure that exists to our knowledge. An extension of Rémillard and Scaillet (2009) is proposed in Bouzebda, Keziou and Zari (2011) when the K populations are observed independently but the test statistic proposed seems usable only to test the simultaneous independence of the K populations, that is to say to test if each copula is independent. Thus it seems that a direct extension of Rémillard and Scaillet (2009) is still a complex open problem. Eventually we can also report the recent work of Derumigny, Fermanian and Min (2021) considering the K-sample problem but in a different setting by restricting their study to conditional copulas.

In this paper we propose to tackle the problem of K copulas comparison with a new approach where a data driven procedure permits to reduce the complexity of the test statistic.

MSC2020 subject classifications: Primary 62H05; secondary 62H15.

Keywords and phrases: Clustering, copula coefficients, data driven, Legendre polynomials.

We do not directly compare the empirical copulas, but we compare their projections in a basis of Legendre polynomials. We restrict our study to continuous variables which populations can be paired. Then it makes possible to compare simultaneously the dependence structures of various populations, such as various portfolios in insurance, but also to compare the same population followed over several periods, such as a medical cohort.

More precisely, let $\mathbf{X} = (X_1, \cdots, X_p)$ be a p-dimensional continuous random variable with joint probability distribution function (pdf) $F_{\mathbf{X}}$ that can be expressed in terms of copula as

$$F_{\mathbf{X}}(x_1, \cdots, x_p) = C(F_1(x_1), \cdots, F_p(x_p)), \tag{1}$$

where F_j denotes the marginal pdf of X_j, and C denotes the copula associated to $\mathbf{X}$. Writing

$$U_j = F_j(X_j), \text{ for } j = 1, \cdots, p,$$

we have for all $u_j \in (0,1)$

$$C(u_1, \cdots, u_p) = F_{\mathbf{U}}(u_1, \cdots, u_p),$$

with $\mathbf{U} = (U_1, \cdots, U_p)$, and deriving this expression p times with respect to $u_1, \cdots, u_p$, we get an expression of the density copula

$$c(u_1, \cdots, u_p) = f_{\mathbf{U}}(u_1, \cdots, u_p), \tag{2}$$

where $f_{\mathbf{U}}$ denotes the joint density of the vector $\mathbf{U}$. Write $\mathcal{L} = \{L_n; n \in \mathbb{N}\}$ the set of orthogonal Legendre polynomials with first terms $L_0 = 1$ and $L_1(x) = \sqrt{3}(2x-1)$, such that L_n is of degree n and satisfies (see Appendix C for more detail):

$$\int_0^1 L_j(u) L_k(u) du = \delta_{jk},$$

where $\delta_{jk} = 1$ if $j = k$ and 0 otherwise. The random variables U_i are uniformly distributed and we have the following decomposition

$$f_{\mathbf{U}}(u_1, \cdots, u_p) = \sum_{j_1, \cdots, j_p \in \mathbb{N}} \rho_{j_1, \cdots, j_p} L_{j_1}(u_1) \cdots L_{j_p}(u_p), \tag{3}$$

where

$$\rho_{j_1, \cdots, j_p} = \mathbb{E}(L_{j_1}(U_1) \cdots L_{j_p}(U_p)), \tag{4}$$

as soon as $f_{\mathbf{U}}$ belongs to the space of all square-integrable functions with respect to the Lebesgue measure on $[0,1]^p$, that is, if

$$\int_0^1 \cdots \int_0^1 f_{\mathbf{U}}(u_1, \cdots, u_p)^2 du_1 \cdots du_p < \infty. \tag{5}$$

Write $\mathbf{j} = (j_1, \cdots, j_p)$ and $\mathbf{0} = (0, \cdots, 0)$. We can observe that $\rho_{\mathbf{0}} = 1$. Moreover, since by orthogonality we have $\mathbb{E}(L_{j_i}(U_i)) = 0$ for all $i = 1, \cdots, p$, we see that $\rho_{\mathbf{j}} = 0$ if only one element of $\mathbf{j}$ is non null. From (2) and (3) we deduce the expression of both copula and copula density, for all $u_1, \cdots, u_p \in (0,1)$ under condition (5):

$$c(u_1, \cdots, u_p) = 1 + \sum_{\mathbf{j} \in \mathbb{N}_*^p} \rho_{\mathbf{j}} L_{j_1}(u_1) \cdots L_{j_p}(u_p), \tag{6}$$

$$C(u_1, \cdots, u_p) = u_1\, u_2 \cdots u_p + \sum_{\mathbf{j} \in \mathbb{N}_*^p} \rho_{\mathbf{j}} I_{j_1}(u_1) \cdots I_{j_p}(u_p), \tag{7}$$

where

$$I_j(u) = \int_0^u L_j(x)dx,$$

and where $\mathbb{N}_*^p$ stands for the set $\{\mathbf{j} = (j_1, \cdots, j_p) \in \mathbb{N}^p; \mathbf{j} \neq \mathbf{0}\}$. Clearly the sequence $(\rho_\mathbf{j})_{\mathbf{j}\in\mathbb{N}_*^p}$ characterizes the copula and we call it the *copula coefficients*. Then under (5) the comparison of copulas coincides with the comparison of these coefficients. In this way assume that we observe K iid samples, possibly paired, with associated copulas denoted by $C_1, \cdots, C_K$. We consider the problem of testing the equality

$$H_0 : C_1 = \cdots = C_K \tag{8}$$

against H_1 : there exists $1 \leq k \neq k' \leq K$ such that $C_k \neq C_{k'}$.

By the previous expansions (7), testing the equality (8) remains to test the equality of all copula coefficients, that is

$$\tilde{H}_0 : \rho_\mathbf{j}^{(1)} = \cdots = \rho_\mathbf{j}^{(K)}, \quad \forall \mathbf{j} \in \mathbb{N}_*^p, \tag{9}$$

where $\rho^{(k)}$ stands for the copula coefficients associated to C_k. We propose to test $\tilde{H}_0$ with a statistic based on the estimation of these quantities.

Assumption (5) is often encountered in the literature and is discussed in Beare (2010). Is is satisfied for various parametric copulas as the Farlie-Gumbel-Morgenstern, Frank, and Gaussian copulas. It is also obviously satisfied when $f_\mathbf{U}$ is bounded. Moreover, Beare (2010) noted that in the bivariate case, copulas associated to Lancaster type distributions (see Lancaster (1958)) satisfied (5). This is the case for bivariate gamma, Poisson, binomial, and hypergeometric distributions, and for the compound correlated bivariate Poisson distribution (see for instance Hamdan and Al-Bayyati (1971)). However, copulas exhibiting lower or upper tail dependence (in the sense of McNeil, Frey and Embrechts (2015)) do not have square integrable density. In particular, the Gumbel, Clayton and t-copulas all have upper or lower tail dependence and then do not satisfy condition (5). But it is important to note that our framework is nonparametric and that the copulas to be compare are unknown. Moreover, since all components of $\mathbf{U}$ are bounded we know that all the copula coefficients exist and this is why we propose to test hypothesis $\tilde{H}_0$ rather than H_0. Then our procedure consists in comparing all the copula coefficients and can be used even if (5) is not verified.

Our method is a data driven smooth test derived from the Neyman's theory (see Neyman (1937)). These smooth tests are omnibus tests and detect any departure from the null. A penalized rule is introduced to select automatically an optimal number of coefficients to be compared. Under the null such a rule selects only one coefficient, leading to a chi-square asymptotic null distribution. We also prove that our test procedure detects alternatives such that $\rho_\mathbf{j}^{(k)} \neq \rho_\mathbf{j}^{(k')}$, for some integers $\mathbf{j} \in \mathbb{N}_*^p$, and $k \neq k'$; that is, there is at least one different copula coefficient.

Since this approach is data driven, we can deduce a clustering algorithm that permits to regroup automatically populations with similar dependence structure. For instance it can be useful in the case where many portfolios are compared in insurance and it yields a very easy way to construct similar groups with a given confidence level. Conversely, it can also be used to diversify portfolios and thus protect against excessively dependent risks.

A numerical study shows the very good behaviour of the test. We apply this approach on two datasets. The first one is the very well-known Iris dataset. While this dataset is very famous there was no simultaneous comparison between the 4-dimensional dependence structures of the three species involved. We therefore propose to apply our method to compare the dependence between sepals and petals, thus providing a new analysis. The second dataset is

a large medical insurance database with possibly paired data and concerns claims from three years: 1997, 1998 and 1999. We apply our method on several variables from this dataset. Finally we also illustrate the clustering algorithm on the two datasets.

The paper is organized as follows: in Section 2 we introduce the estimators of the copula coefficients and we set up notation. Section 3 presents our method in the two-sample case. In Section 4, we extend the result to the K ($K > 2$) sample case and in Section 5 we proceed with the study of the convergence of the test under alternatives. Section 6 establishes the relation between the K-sample procedure and a clustering algorithm. Section 7 is devoted to the numerical study and Section 8 contains two real-life illustrations. Section 9 discusses extensions and connections. Section 10 presents the proofs of the main results while the remaining proof and technical materials are deferred to the Supplementary Material.

2. Notation and estimation step. We consider K continuous random vectors, namely

$$\mathbf{X}^{(1)} = (X_1^{(1)}, \cdots, X_p^{(1)}), \cdots, \mathbf{X}^{(K)} = (X_1^{(K)}, \cdots, X_p^{(K)}),$$

with joint cumulative distribution function (cdf) $\mathbf{F}^{(1)}, \cdots, \mathbf{F}^{(K)}$, and with associated copulas $C_1, \cdots, C_K$, respectively. Assume that we observe K iid samples from $\mathbf{X}^{(1)}, \cdots, \mathbf{X}^{(K)}$, possibly paired, denoted by

$$(X_{i,1}^{(1)}, \cdots, X_{i,p}^{(1)})_{i=1,\cdots,n_1}, \cdots, (X_{i,1}^{(K)}, \cdots, X_{i,p}^{(K)})_{i=1,\cdots,n_K}.$$

We assume that

$$\text{for all } 1 \le k < \ell \le K,\ n_k/(n_k + n_\ell) \to a_{k\ell},\ \text{with } 0 < a_{k\ell} < \infty. \tag{10}$$

We will denote by $F_j^{(k)}$ the marginal cdf of the jth component of $\mathbf{X}^{(k)}$ and we write

$$U_{i,j}^{(k)} = F_j^{(k)}(X_{i,j}^{(k)}).$$

For testing (9) we first estimate the copula coefficients by

$$\widehat{\rho}_{j_1 \cdots j_p}^{(k)} = \frac{1}{n_k} \sum_{i=1}^{n_k} L_{j_1}(\widehat{U}_{i,1}^{(k)}) \cdots L_{j_p}(\widehat{U}_{i,p}^{(k)})), \tag{11}$$

where

$$\widehat{U}_{i,j}^{(k)} = \widehat{F}_j^{(k)}(X_{i,j}^{(k)}),$$

and where $\widehat{F}$ denotes the empirical distribution functions associated to F.

Considering the null hypothesis H_0 as expressed in (9), our test procedure is based on the sequences of differences

$$r_{\mathbf{j}}^{(\ell,m)} := \widehat{\rho}_{\mathbf{j}}^{(\ell)} - \widehat{\rho}_{\mathbf{j}}^{(m)},\ \text{for } 1 \le \ell \le m \le K,\ \text{and } \mathbf{j} \in \mathbb{N}_*^p,$$

with the convention that $r_{\mathbf{j}}^{(\ell,m)} = 0$ when only one element of $\mathbf{j}$ is different of zero, since in this case we have $\rho_{\mathbf{j}}^{(\ell)} = \rho_{\mathbf{j}}^{(m)} = 0$, from the orthogonality of the Legendre polynomials.

In order to select automatically the number of copula coefficients, for any vector $\mathbf{j} = (j_1, \cdots, j_p)$ we denote by

$$\|\mathbf{j}\|_1 = |j_1| + \cdots + |j_p|,$$

its L^1 norm and for any integer $d > 1$ we write

$$\mathcal{S}(d) = \{\mathbf{j} \in \mathbb{N}^p; \|\mathbf{j}\|_1 = d \text{ and there exists } k \neq k' \text{ such that } j_k > 0 \text{ and } j_{k'} > 0\}.$$

The set $\mathcal{S}(d)$ contains all non null positive integers $\mathbf{j} = (j_1, \cdots, j_p)$ with norm d and such that $j_k < d$ for all $k = 1, \cdots, p$. We will denote by $c(d) = \binom{d}{d+p-1} - p$ the cardinal of $\mathcal{S}(d)$ and we introduce a lexicographic order on $\mathbf{j} \in \mathcal{S}(d)$ as follows:

$$\begin{aligned}
&\mathbf{j} = (d-1, 1, 0, \cdots, 0) \Rightarrow ord(\mathbf{j}, d) = 1\\
&\mathbf{j} = (d-1, 0, 1, \cdots, 0) \Rightarrow ord(\mathbf{j}, d) = 2\\
&\cdots\\
&\mathbf{j} = (0, \cdots, 0, 2, d-2) \Rightarrow ord(\mathbf{j}, d) = c(d) - 1\\
&\mathbf{j} = (0, \cdots, 0, 1, d-1) \Rightarrow ord(\mathbf{j}, d) = c(d).
\end{aligned}$$

For instance, in the bivariate case, that is $p = 2$, we have

- if $d = 2$ there is only one possibility: $\mathbf{j} = (j_1, j_2) = (1, 1)$ with $ord(\mathbf{j}, 2) = 1$. The cases $(2, 0)$ or $(0, 2)$ are excluded.
- if $d = 3$ there are two possibilities: $\mathbf{j} = (2, 1)$ with $ord(\mathbf{j}, 3) = 1$ and $\mathbf{j} = (1, 2)$ with $ord(\mathbf{j}, 3) = 2$. The cases $\mathbf{j} = (0, 3)$ and $\mathbf{j} = (3, 0)$ are excluded.

3. The two-sample case. We first consider the two-sample case with $K = 2$ to detail the construction of our test statistics. Here we want to test

$$\tilde{H}_0 : \rho_{\mathbf{j}}^{(1)} = \rho_{\mathbf{j}}^{(2)}, \ \ \forall \mathbf{j} \in \mathbb{N}_*^p,$$

which is equivalently to $H_0 : C_1 = C_2$, when (5) is satisfied. We restrict our attention to the paired case and we write $n_1 = n_2 = n$. The independent case with $n_1 \neq n_2$ is briefly described in Appendix B. We then have iid observations $\{(X_{ik}^{(1)}, X_{ik}^{(2)}), k = 1, \cdots, p\}, i = 1, \cdots, n$ where $X_{ik}^{(1)}$ and $X_{ik}^{(2)}$ are dependent. To compare the copulas of $\mathbf{X}^1$ and $\mathbf{X}^2$ we introduce a series of statistics based on the differences between their copula coefficients as follows: for $1 \leq k \leq c(2)$ we define

$$T_{2,k}^{(1,2)} = n \sum_{\mathbf{j} \in \mathcal{S}(2); ord(\mathbf{j},2) \leq k} (r_{\mathbf{j}}^{(1,2)})^2, \tag{12}$$

and for $d > 2$ and $1 \leq k \leq c(d)$,

$$T_{d,k}^{(1,2)} = T_{d-1,c(d-1)}^{(1,2} + n \sum_{\mathbf{j} \in \mathcal{S}(d); ord(\mathbf{j},d) \leq k} (r_{\mathbf{j}}^{(1,2)})^2. \tag{13}$$

Clearly all these statistics are embedded since we have for $2 \leq k < c(d)$

$$\begin{aligned}
T_{d,k}^{(1,2)} &= T_{d,k-1}^{(1,2)} + n(r_{\mathbf{j}}^{(1,2)})^2 \mathbb{I}_{\mathbf{j} \in \mathcal{S}(d); ord(\mathbf{j},d) = k}\\
&= n \left(\sum_{u=2}^{d-1} \sum_{\mathbf{j} \in \mathcal{S}(u)} (r_{\mathbf{j}}^{(1,2)})^2 + \sum_{\mathbf{j} \in \mathcal{S}(d); ord(\mathbf{j},d) \leq k} (r_{\mathbf{j}}^{(1,2)})^2 \right),
\end{aligned}$$

where $\mathbb{I}$ denotes the indicator function. It follows that

$$T_{2,1}^{(1,2)} \leq T_{2,2}^{(1,2)} \leq T_{2,c(2)}^{(1,2)} \leq T_{3,1}^{(1,2)} \leq \cdots \leq T_{d,k}^{(1,2)} \leq \cdots \leq T_{d,c(d)}^{(1,2)} \leq T_{d+1,1}^{(1,2)} \leq \cdots .$$

Each statistic $T_{d,k}^{(1,2)}$ contains information permitting to compare the copula coefficients $\rho_{\mathbf{j}}^{(1)}$ and $\rho_{\mathbf{j}}^{(2)}$ up to the norm $\|\mathbf{j}\|_1 = d$ and $ord(\mathbf{j}, d) = k$. So when d is large it will make it possible

to compare high coefficient orders through $r_{\mathbf{j}}^{(1,2)}$, while k will permit to visit all the values of $\mathbf{j}$ for this given order. To simplify notation we write such a sequence of statistics as

$$V_1^{(1,2)} = T_{2,1}^{(1,2)};\ V_2^{(1,2)} = T_{2,2}^{(1,2)};\cdots\ V_{c(2)}^{(1,2)} = T_{2,c(2)}^{(1,2)};\ V_{c(2)+1}^{(1,2)} = T_{3,1}^{(1,2)}\cdots$$

By construction, for all integer $k > 0$ there exists a set $\mathcal{H}(k) \subset \mathbb{N}_*^p$, with $card(\mathcal{H}(k)) = k$, such that

$$V_k^{(1,2)} = n \sum_{\mathbf{j}\in\mathcal{H}(k)} (r_{\mathbf{j}}^{(1,2)})^2. \tag{14}$$

It can be observed that if $\mathbf{j}$ belongs to $\mathcal{H}(k)$ then $\|\mathbf{j}\|_1 \leq k$. Moreover, we have the following relation: for all $k \geq 1$ and $j = 1, \cdots, c(k+1)$

$$V_{c(1)+c(2)+\cdots+c(k)+j}^{(1,2)} = T_{k+1,j}^{(1,2)}, \quad \text{with the convention } c(1) = 0.$$

Notice that we need to compare all copula coefficients and then to let k tend to infinity to detect all possible alternatives. However, choosing a too large value tends to power dilution of the test. Following Kallenberg and Ledwina (1995), we suggest a data driven procedure to select automatically the number of coefficients to test the hypothesis H_0. Namely, we set

$$D(n) := \min\Big\{ \underset{1\leq k\leq d(n)}{\operatorname{argmax}} (V_k^{(1,2)} - kq_n)\Big\}, \tag{15}$$

where q_n and $d(n)$ tend to $+\infty$ as $n \to +\infty$, kq_n being a penalty term which penalizes the embedded statistics proportionally to the number of copula coefficients used. Finally, the data-driven test statistic that we use to compare C_1 and C_2 is $V_{D(n)}^{(1,2)}$ and we consider the following rate for the number of components in the statistic:

(A) $d(n)^{(p+4)} = o(q_n)$

A classical choice for q_n is $\log(n)$ initially used in Schwarz (1978) (see also the seminal work of Ledwina (1994)). This choice is convenient to detect smooth alternatives (see Section 5) and will be adopted in our simulation, up to a tuning factor. Our first result shows that under the null the least penalized statistic will be selected.

THEOREM 3.1. *Let assumption* **(A)** *holds. Then, under* $\tilde{H}_0$, $D(n)$ *converges in probability towards 1 as* $n \to +\infty$.

It is worth noting that under the null, the asymptotic distribution of the statistic $V_{D(n)}^{(1,2)}$ coincides with the asymptotic distribution of $V_1^{(1,2)} = T_{2,1}^{(1,2)} = n(r_{\mathbf{j}}^{(1,2)})^2$, with $\mathbf{j} = (1,1,0,\cdots,0)$. In that case we have

$$r_{\mathbf{j}}^{(1,2)} = \frac{1}{n}\sum_{i=1}^{n} \big(L_1(\widehat{U}_{i,1}^{(1)})L_1(\widehat{U}_{i,2}^{(1)}) - L_1(\widehat{U}_{i,1}^{(2)})L_1(\widehat{U}_{i,2}^{(2)})\big).$$

It follows that $T_{2,1}^{(1,2)}$ measures the discrepancy between $\mathbb{E}(L_1(U_1^{(1)})L_1(U_2^{(1)}))$ and $\mathbb{E}(L_1(U_1^{(2)})L_1(U_2^{(2)}))$. Asymptotically, the null distribution reduces to that of $V_1^{(1,2)}$ and is given below.

THEOREM 3.2. *Assume that* $\mathbf{j} = (1,1,0,\cdots,0)$. *Then, under* $\tilde{H}_0$, $\sqrt{n}r_{\mathbf{j}}^{(1,2)}$ *converges in law towards a central normal distribution with variance*

$$\sigma^2(1,2) = \mathbb{V}\Bigg(L_1(U_1^{(1)})L_1(U_2^{(1)}) - L_1(U_1^{(2)})L_1(U_2^{(2)})$$
$$+2\sqrt{3}\int\int \big(\mathbb{I}(X_1^{(1)} \le x) - F_1^{(1)}(x)\big)L_1(F_2^{(1)}(y))dF^{(1)}(x,y)$$
$$-2\sqrt{3}\int\int \big(\mathbb{I}(X_1^{(2)} \le x) - F_1^{(2)}(x)\big)L_1(F_2^{(2)}(y))dF^{(2)}(x,y)$$
$$+2\sqrt{3}\int\int \big(\mathbb{I}(X_2^{(1)} \le y) - F_2^{(1)}(y)\big)L_1(F_1^{(1)}(x))dF^{(1)}(x,y)$$
$$-2\sqrt{3}\int\int \big(\mathbb{I}(X_2^{(2)} \le y) - F_2^{(2)}(y)\big)L_1(F_1^{(2)}(x))dF^{(2)}(x,y)\Bigg).$$

In order to normalize the test, write

$$\widehat{\sigma}^2(1,2) = \frac{1}{n}\sum_{i=1}^{n}\Big(M_{i,1} - M_{i,2} - \overline{M}_1 + \overline{M}_2\Big)^2,$$

with

$$\overline{M}_s = \frac{1}{n}\sum_{i=1}^{n} M_{i,s}, \text{ for } s = 1,2,$$

where

$$M_{i,s} = L_1(\widehat{U}_{i,1}^{(s)})L_1(\widehat{U}_{i,2}^{(s)}) + \frac{2\sqrt{3}}{n}\sum_{k=1}^{n}\Big(\mathbb{I}\big(X_{i,1}^{(s)} \le X_{k,1}^{(s)}\big) - \widehat{U}_{k,1}^{(s)}\Big)L_1(\widehat{U}_{k,2}^{(s)})$$
$$+\frac{2\sqrt{3}}{n}\sum_{k=1}^{n}\Big(\mathbb{I}\big(X_{i,2}^{(s)} \le X_{k,2}^{(s)}\big) - \widehat{U}_{k,2}^{(s)}\Big)L_1(\widehat{U}_{k,1}^{(s)}).$$

PROPOSITION 1. *Under $\tilde{H}_0$ we have the following convergence in probability*

$$\widehat{\sigma}^2(1,2)) \xrightarrow{\mathbb{P}} \sigma^2(1,2).$$

We then deduce the limit distribution under the null.

COROLLARY 3.3. *Let assumption* **(A)** *holds. Then under $\tilde{H}_0$, $V_{D(n)}^{(1,2)}/\widehat{\sigma}^2(1,2)$ converges in law towards a chi-squared distribution χ_1^2 as $n \to +\infty$.*

4. The *K*-sample case. Write $\mathbf{n} = (n_1, \cdots, n_K)$. We restrict our attention to the paired case here, fixing then $n_1 = n_2 = \cdots = n_K := n$. The independent case is treated in Appendix B. Our aim is to generalize the two-sample case by considering a series of embedded statistics, each new of them including a new pair of populations to be compared. In this way we introduce the following set of indexes:

$$\mathcal{V}(K) = \{(\ell, m) \in \mathbb{N}^2; 1 \le \ell < m \le K\}.$$

Clearly $\mathcal{V}(K)$ contains $v(K) = K(K-1)/2$ elements which represent all the pairs of populations that we want to compare and that can be ordered as follows: we write $(\ell, m) <_{\mathcal{V}}$

(ℓ', m') if $\ell < \ell'$, or $\ell = \ell'$ and $m < m'$, and we denote by $rank_{\mathcal{V}}(\ell, m)$ the associated rank of (ℓ, m) in $\mathcal{V}(K)$. This can be seen as a natural order (left to right and bottom to top) of the elements of the upper triangle of a $(K-1) \times (K-1)$ matrix as represented below:

$$\begin{array}{ccccc} (1,2) & (1,3) & \cdots\cdots & & (1,K) \\ & (2,3) & \cdots\cdots & & (2,K) \\ & & \ddots & & \\ & & & & (K-1,K) \end{array}$$

We see at once that $rank_{\mathcal{V}}(1,2) = 1, rank_{\mathcal{V}}(1,3) = 2$ and more generally, for $\ell, m \in \mathcal{V}(K)$ we have

$$rank_{\mathcal{V}}(\ell, m) = K(l-1) - \frac{l(l+1)}{2} + m.$$

We construct an embedded series of statistics as follows

$$V_1 = V_{D(n)}^{(1,2)}, \;\; V_2 = V_{D(n)}^{(1,2)} + V_{D(n)}^{(1,3)}, \;\; \cdots, \;\; V_{v(K)} = V_{D(n)}^{(1,2)} + \cdots + V_{D(n)}^{(K-1,K)},$$

or equivalently,

$$V_k = \sum_{(\ell,m) \in \mathcal{V}(K); rank_{\mathcal{V}}(\ell,m) \leq k} V_{D(n)}^{(\ell,m)},$$

where $D(n)$ is given by (15) and $V_{D(n)}^{(\ell,m)}$ is defined as in (14). We have $V_1 < \cdots < V_{v(K)}$. The first statistic V_1 compares the first two populations 1 and 2. The second statistic V_2 compares the populations 1 and 2, and, in addition, the populations 1 and 3. And so on. For each $1 < k < v(K)$, there exists a unique pair (ℓ, m) such that $rank_{\mathcal{V}}(\ell, m) = k$. To choose automatically the appropriate number k we introduce the following penalization procedure, mimicking the Schwarz criteria procedure (Schwarz, 1978):

$$s(\mathbf{n}) = \min \left\{ \underset{1 \leq k \leq v(K)}{\operatorname{argmax}} \left(V_k - k \sum_{(\ell,m) \in \mathcal{V}(K)} p_{\mathbf{n}}(\ell, m) \mathbb{I}_{rank_{\mathcal{V}}(\ell,m)=k} \right) \right\},$$

where $p_{\mathbf{n}}(\ell, m)$ is a penalty term. In the sequel we consider the penalty term as a function of the sample sizes only, that is $p_{\mathbf{n}}(\ell, m) = p_{\mathbf{n}}$ for all $\ell, m = 1, \cdots, K$. And since $n_1 = \cdots = n_K = n$ we simply write $p_{\mathbf{n}} = p_n$. We then obtain

$$s(\mathbf{n}) = \min \left\{ \underset{1 \leq k \leq v(K)}{\operatorname{argmax}} \left(V_k - k p_n \right) \right\}. \tag{16}$$

We discuss this choice in Remark 1. We make the following assumption:

(A') $d(n)^{(p+4)} = o(p_n)$

The following result shows that under the null, the penalty chooses the first element of $\mathcal{V}(K)$ asymptotically.

THEOREM 4.1. *Assume that* **(A)** *and* **(A')** *hold. Then under* $\tilde{H}_0$, $s(\mathbf{n})$ *converges in probability towards 1 as* $n \to +\infty$.

COROLLARY 4.2. *Assume that* **(A)** *and* **(A')** *hold. Then under* $\tilde{H}_0$, $V_{s(\mathbf{n})}/\widehat{\sigma}^2(1,2))$ *converges in law towards a* χ_1^2 *distribution.*

Then our final data driven test statistic is given by

$$V = V_{s(\mathbf{n})}/\widehat{\sigma}^2(1,2)). \tag{17}$$

REMARK 1. In the classical smooth test approach (Ledwina, 1994) a standard penalty is $q_n = p_n = \alpha \log(n)$, which is related to the Schwarz criteria (Schwarz, 1978) as discussed in Kallenberg and Ledwina (1995). In practice, the factor α permits to stabilize the empirical level to be as close as possible to the asymptotic one. Note also that Inglot and Ledwina (2006) compared this type of Schwarz penalty to the Akaike one where they proposed p_n or q_n to be constant. In our simulation we consider the classical choice $q_n = p_n = \alpha \log(n)$, with an automatic choice of α described in Section 6 which makes it possible to calibrate the test very simply.

5. Alternative hypotheses. We consider the following series of alternative hypotheses:

$$H_1(1): \text{ the two first copulas } C_1 \text{ and } C_2 \text{ have at least one different copula coefficient}$$

and for $k > 1$:

$$H_1(k): \begin{cases} \text{if } rank_{\mathcal{V}}(k,\ell) < k, C_k \text{ and } C_\ell \text{ have the same copula coefficients} \\ \text{if } rank_{\mathcal{V}}(k,\ell) = k, C_k \text{ and } C_\ell \text{ have at least a different copula coefficient} \end{cases}$$

The hypothesis $H_1(k)$ means that the kth and ℓth populations such that $rank_{\mathcal{V}}(k,\ell) = k$ are the first (in the sense of the order in $\mathcal{V}(K)$) with at least one different copula coefficient.

We make the following assumption:

(B) $p_n = o(n)$.

THEOREM 5.1. *Assume that* **(A)-(A')-(B)** *hold. Then under* $H_1(k)$, $s(\mathbf{n})$ *converges in probability towards* k, *as* $\mathbf{n} \to +\infty$, *and* V *converges to* $+\infty$, *that is,* $\mathbb{P}(V < \epsilon) \to 0$, *for all* $\epsilon > 0$.

6. Clustering. In the sequel we propose to adapt the previous test procedure to obtain a data-driven method to cluster K populations into N subgroups characterized by a common dependence structure. The number N of clusters is unknown and will be automatically chosen by the previous procedure and validated by our testing method.

More precisely, assume that we observe K iid samples from K populations, possibly paired. The clustering algorithm starts by choosing the two populations that are the most similar in terms of dependence structure, through their copulas. In this way, it chooses the smaller two-sample statistic. If the equality of both associated copulas is accepted these two populations form the first cluster. Then the algorithm proposes the closer population of this cluster, that is the smaller statistic having a common population index. While the test accepts the simultaneous equality of the copulas, the cluster growths. If the last test is rejected then the cluster is closed and the last rejected population forms a new cluster. One can iterate this several times until every sample is associated with a cluster. We can summarize the clustering algorithm as follows:

Algorithm: K-sample copulas clustering

Initialization: $c=1$. By convention, $S=\{C_1,\cdots,C_K\}$ and $S_0=\emptyset$;
Select $\{\ell^\star,m^\star\}=\text{argmin}\{V_{D(n)}^{(\ell,m))};\ell\neq m\in S\setminus\bigcup_{k=1}^{c}S_{k-1}\}$;
Test $\tilde{H}_0$ between all $\rho_{\mathbf{j}}^{(\ell^\star)}$ and $\rho_{\mathbf{j}}^{(m^\star)}$;
if $\tilde{H}_0$ *is not rejected* **then**
    $S_1=\{C_{\ell^\star},C_{m^\star}\}$;
else
    STOP. There is no cluster.
end
while $S\setminus\bigcup_{k=1}^{c}S_k\neq\emptyset$ **do**
    Select $\{j^\star\}=\text{argmin}\{T_{D(n)}^{(i,j)};i\in S_c,j\in S\setminus\bigcup_{k=1}^{c}S_k\}$;
    Test $\tilde{H}_0$ the simultaneous equality of all the $\rho_{\mathbf{j}}^{(i)}$, $i\in S_c$ and $\rho_{\mathbf{j}}^{(j^\star)}$;
    if $\tilde{H}_0$ *not rejected* **then**
        $S_c=S_c\bigcup\{C_{j^\star}\}$;
    else
        $S_{c+1}=\{C_{j^\star}\}$;
        $c=c+1$;
    end
end

This clustering procedure can solve several complex problems in a very short time and is useful in practice, particularly in risk management and more generally in the world of actuarial science and finance markets by making it possible to detect mutualizable risks and not mutualizable; but also to build a well-diversified portfolio.

7. Numerical study of the test.

7.1. *Tuning the test statistic.* As evoked in Remark 1 we can choose the penalty $q_n=p_n=\alpha\log(n)$. We fix $\alpha=1$ in the proofs of this paper for simplicity. But in practice we can empirically improve this tuning factor by using the following data driven procedure:

- Assume we observe K populations.
- We merge all populations to get only one (larger) population.
- Split randomly this population into $K'>2$ sub-populations.
- Clearly these K' sub-populations have the same copula and then the null hypothesis $\tilde{H}_0$ is satisfied.
- We then approximate numerically the value of the factor $\alpha>0$ such that the selection rule retains the first component, that is $s(\mathbf{n})=1$. From Theorem 4.1 this is the asymptotic expected value under the null.
- We can repeat N times such a procedure to get N K'-sample under the null.

Finally we fix

$$\widehat{\alpha}=\min\{\alpha>0;\ \text{such that } s(\mathbf{n})=1 \text{ for the previous} N \text{selection rules}\}.$$

In our simulation we fixed arbitrarily $K'=3$, which seems to give a very correct empirical level.

Concerning the value of $d(n)$, the condition **(A)** is an asymptotic condition and from our experience choosing $d(n) = 3$ or 4 is enough to have a very fast procedure which detects alternatives such that copulas differ by a coefficient with a norm less or equal to $d(n)$.

7.2. *Simulation design.* In order to evaluate the performance of our test, we consider the following classical copulas families: the Gaussian copulas, the Student copulas, the Gumbel copulas, the Frank copulas, the Clayton copulas and the Joe copulas which we denote for hereafter *Gaus*, *Stud*, *Gumb*, *Fran*, *Clay* and *Joe* respectively. For the explicit functional forms and properties of these copulas we refer the reader to Nelsen (2007). For each copula C, the sample is generated with a given kendall's τ parameter, and we denote this model briefly by $C(\tau)$. When τ is close to zero the variables are close to the independence. Conversely, if τ is close to 1 the dependence becomes linear.

We consider two cases: i) a 5-sample case; ii) a 10-sample case; and for both we compute empirical levels and powers under null hypothesis and alternatives.

In Appendix F we also consider the two-sample case where we compare our test procedure to that proposed in Rémillard and Scaillet (2009) which is the only one competitor we found. Both methods give very similar results, with slightly higher power for our test procedure. Note that a large sample size n can increase significantly the computing time for the test proposed in the package *Twocop* of Rémillard and Scaillet (2009) and it can become too heavy to compare and less competitive.

7.3. *Five sample case.* In this case ($K = 5$), we fix $p = 3$ and we consider the same sizes for all sample, that is $n = n_1 = n_2 = n_3 = n_4 = n_5 \in \{50, 100, 200, \cdots, 900, 1000\}$. We fixed a theoretical level $\alpha = 5\%$.

Null hypotheses: we consider the following null hypotheses with with three levels of dependence: $\tau = 0.1$ (low dependence), $\tau = 0.5$ (middle dependence) and $\tau = 0.8$ (high dependence).

- **Null(Gaus)**: the same Gaussian copulas
- **Null(Stud)**: the same Student copulas
- **Null(Gumb)**: the same Gumbel copulas
- **Null(Fran)**: the same Frank copulas
- **Null(Clay)**: the same Clayton copulas
- **Null(Joe)**: the same Joe copulas

Alternatives: we consider the following alternatives hypotheses with $C_1, \cdots, C_5$ in the same copula family but with different τ as follows

- **Alt1**: $C_1(0.1)$ and $C_2(0.3) = C_3(0.3) = C_4(0.3) = C_5(0.3)$
- **Alt2**: $C_1(0.1)$ and $C_2(0.55) = C_3(0.55) = C_5(0.55)$, and $C_4(0.3)$
- **Alt3**: $C_1(0.1)$ and $C_2(0.8) = C_3(0.8) = C_5(0.8)$, and $C_4(0.3)$

Alt1 contains only one different population. Concerning **Alt2** and **Alt3**, they differ only from their Kendall coefficient and then allow to underline its effect. Figures 1-3 show the empirical levels (in $\%$) with respect to the sample sizes when $\tau = 0.1, 0.3$ and 0.8, respectively. For each case one can observe that the empirical level is close to the theoretical 5% as soon as n is greater than 200.

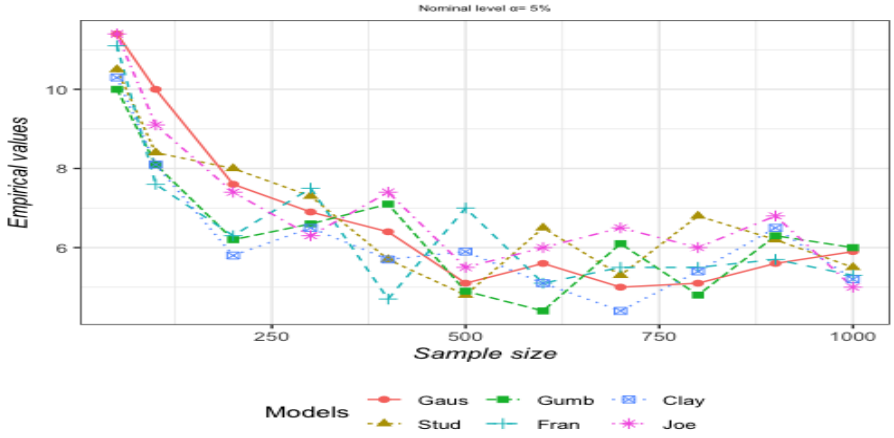

Fig 1: Five-sample case:empirical level for the null hypotheses with $\tau = 0.1$

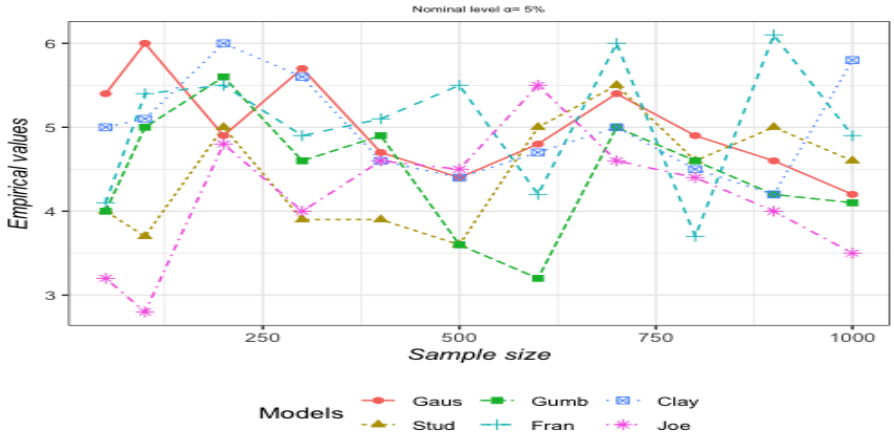

Fig 2: Five-sample case: empirical level for the null hypotheses with $\tau = 0.5$

Concerning the empirical power, Tables 1-3 contain all results under the alternatives. We omit some large sample size results where empirical powers are equal to 100%. It is important to note that even a sample size equal to 1000 the program runs very fast. It can be seen for the last two series of alternatives that the empirical powers are extremely high even for small sample sizes. The first series of alternatives yields good empirical powers but lower than in the two other series. This result was clearly expected because the tau are much closer and then dependence structure are more similar.

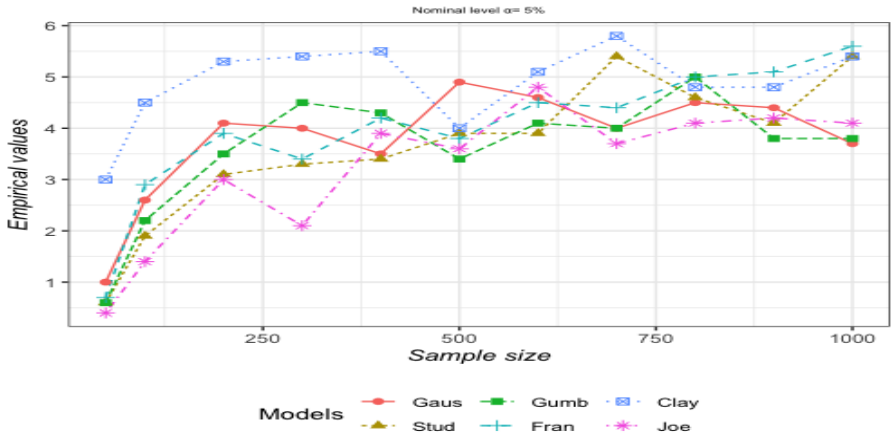

Fig 3: Five-sample case: empirical level for the null hypotheses with $\tau = 0.8$

TABLE 1
Five-sample test: Empirical powers for Alternative **Alt1**.

	Alternatives					
Size	Gaussian	Student	Gumbel	Frank	Clayton	Joe
50	39.9	35.7	35.6	36.6	35.9	35.5
100	64.1	61.8	60.3	64.0	61.1	60.7
200	91.5	88.4	87.5	91.1	89.9	87.7
300	97.9	98.0	97.7	98.2	97.3	97.2
400	99.8	99.7	99.6	99.8	99.7	99.8
500	100	100	100	100	100	99.9
600	100	100	100	100	100	100

TABLE 2
Five-sample test: Empirical powers for Alternative **Alt2**.

	Alternatives					
Size	Gaussian	Student	Gumbel	Frank	Clayton	Joe
50	97.8	97.6	96.3	98.6	97.4	95.6
100	100	100	99.9	100	100	100
200	100	100	100	100	100	100

TABLE 3
Five-sample test: Empirical powers for Alternative **Alt3**.

	Alternatives					
Size	Gaussian	Student	Gumbel	Frank	Clayton	Joe
50	100	100	100	100	100	100

7.4. *Ten sample case.* Analogously to the previous 5-sample case and wit the same notation, we considered six null hypotheses, as previously denoted by:

- **Null(Gaus), Null(Stud), Null(Gumb), Null(Fran), Null(Clay), Null(Joe)**,

and the following alternative where only one copula differs from the others.

- **Alt4** : $C_1(0.55)$ and $C_2(0.1) = C_3(0.1) = \cdots = C_{10}(0.1)$, with $\tau = 0.1$.

Empirical levels seem to tend fast to 0.5 and are relegated in Appendix F.

Table 4 shows empirical powers under alternatives. We only treat the case $n = 50$ and 100 since beyond all the powers are equal to 100%. We can observe a very good behavior of the test even for small sample sizes.

TABLE 4
Ten-sample test: Empirical powers for Alternative **Alt4**.

	Alternatives					
Size	Gaussian	Student	Gumbel	Frank	Clayton	Joe
50	98.0	96.7	96.2	97.9	97.1	97.3
100	100	100	100	100	100	100

7.5. *Clustering simulation.* We consider the following designs:

- **D1**: $n = 100$, $p = 3$, $K = 6$ populations with 3 groups $C_1 = Gumb(0.8)$ and $C_2 = C_3 = Gaus(0.2)$ and $C_4 = C_5 = C_6 = Clay(0.9)$
- **D2 =** D1 with $n = 500$
- **D3**: $n = 100$, $p = 5$, $K = 4$ different populations with 4 groups $C_1 = Gumb(0.8)$, $C_2 = Gaus(0.2)$, $C_3 = Clay(0.9)$, $C_4 = Gumb(1)$
- **D4**: $n = 100$, $p = 4$, $K = 5$ populations with one group $C^{(1)} = C^{(2)} = C^{(3)} = C^{(4)} = C^{(5)} = Clay(0.9)$
- **D5**: $n = 100$, $p = 2$, $K = 10$ populations with two unbalanced groups $C_1 = C_2 = \cdots = C_9 = Clay(0.9)$ and $C_{10} = Gumb(0.9)$

We applied the clustering algorithm described in Section 6. The results are summarized below:

- Results for **D1**
  - In 82.5 % of cases the algorithm found 3 groups. In such cases, 74 % of the time it was the 3 correct groups.
  - In 11.4 % of cases the algorithm found 4 groups
  - In 5 % of cases the algorithm found 2 groups
  - In 0.1 % of cases the algorithm found 5 groups.
  - Note that the first group (with the Gumbel copula) was well identified 99 % of the time.
- Results for **D2** The three groups were well identified in 93 % of cases. In other cases the algorithm obtained 4 groups (merging populations of the second and the third group).
- Results for **D3** In 78 % of cases the null hypothesis was rejected and we obtained 4 different groups. In other cases the algorithm merged two groups (Clayton with Normal or Clayton with Gumbel) and then proposed 3 clusters.
- Results for **D4** In 98 % of cases the algorithm found one group. In other cases it gave two groups.
- Results for **D5** More than 99% of cases the algorithm found the 2 correct groups. In other cases (less than 1%) the algorithm found 3 group obtained by a rejection of one of the 9 similar populations.

8. Real datasets applications.

8.1. *Biology data.* We analyse the well-known Fisher's Iris dataset. The data consists of fifty observations of four measures: Sepal Length (SL), Sepal Width (SW), Petal Length (PL), and Petal Width (PW), for each of three Species: Setosa, Versicolor, and Virginica. We then have $K = 3$ populations, and the dimension is $p = 4$. Figure 4 in Appendix D represents the lengths and widths for the three species. In Dhar, Chakraborty and Chaudhuri (2014) the authors shown that multivariate normal distributions seem to fit the data well for all three Iris species. Looking at their mean parameters the 4-dimensional joint distributions seem different but that does not tell us about their dependence structures.

We propose to test the equality of the dependence structure between the four variables (SL, SW, PL, PW) in the three-sample case, that is:

$$H_0 : C_{Setosa} = C_{Versicolor} = C_{Virginica}$$

This hypothesis implies that all their copula coefficients are equal, which is the hypothesis denoted by $\tilde{H}_0$ that we are testing and which is equivalent under (5). We obtain a p-value close to zero (10^{-11}), a selected rank equal to $D(n) = 2$ and a very large test statistic $V = 45.9$. We clearly reject the equality of the dependence structure here.

In case of reject we can process to an "ANOVA" type procedure as follows: we proceed to a series of 2-sample tests. Table 5 contains the associated p-values and we conclude to the equality of the dependence structure between Versicolor and Virginica.

TABLE 5
P-values for the two-sample tests

	Setosa	Versicolor	Virginica
Setosa	1	10^{-8}	0.0021
Versicolor	10^{-8}	1	0.68
Virginica	0.0021	0.68	1

8.2. *Insurance data.* Insurance is an area in which the knowledge of the dependence structure between several portfolios can be useful in pricing particularly for risk pooling or price segmentation. As an illustration, we consider the Society of Actuaries Group Medical Insurance Large Claims Database. It contains claims information of each claimant over the period 1997 to 1999 from seven insurers. Each row of the database presents a summary of claims for an individual claimant in 27 fields (columns) where the first five columns provide a general information about claimant, the next twelve quantify various types of medical charges and expenses and the last ten columns summarize details connected to the diagnosis. We refer to Grazier and G'Sell (2004) for detailed and thorough description of the data available online with the database at the web page of Society of Actuaries. Here we only consider $p = 3$ dimensional variables $\mathbf{X} = (X_1, X_2, X_3)$, where

- X_1 = paid hospital charges
- X_2 = paid physician charges
- X_3 = paid other charges,

for all claimant insured by a Preferred Provider Organization plan providing exposure for members. We apply our procedure with three scenarios where we study the dependence structure of $\mathbf{X}$ as follows:

Three-sample test, paired case. In this case, we consider the same claimants present over

the three periods $1997-1999$. At the end of the data processing, we obtain three samples of size $n = 6874$ observations. We analyse the dependence structure of the charges $\mathbf{X}$ between the three years ($C_{\mathbf{X}}^{1997} = C_{\mathbf{X}}^{1998} = C_{\mathbf{X}}^{1999}$). Here we have clearly a 3-sample test with paired data.
The test concluded to the non rejection of the equality of the three dependence structure with a p-value = 0.68 and a test statistic $V = 0.17$. Hence, the dependence structure of paid for insured over the three years seems to be similar.

Three-sample test, independent case. Here we restrict our attention to the female claimants. The three populations are composed by the relationship with the subscriber which can be "Employee" ($n_E = 18144$ observations), "Spouse" ($n_S = 10969$ observations) or "Dependent" ($n_D = 3555$ observations), for the year 1999. We want to test the equality of the dependence structure between charges $\mathbf{X}$. Here the $K = 3$ populations are assumed to be independent. Using our test procedure, we obtained a p-value close to zero. Therefore, the null hypothesis of equal dependence structure of those charges is rejected. The two-by-two equalities are rejected for "Dependent"/"Employee" and "Dependent"/"Spouse" with p-value in each case closing to 0. The p-value of "Employee"/"Spouse" is 0.0059. Thus the fact of being "Employee" or "Spouse" involves similar dependence structure of the charges and the two are different from "Dependent".

Ten-sample test, independent case. Here, we merely consider the data of the year 1999 where the relationship to subscriber is employee. We split the charges $\mathbf{X}$ by age range of three years and consider 10 groups as follows: Group1 = $[1936,\ 1938]$, ..., Group10 = $[1963,\ 1965]$.

The null hypothesis is H_0: the dependence structure of these 10 samples groups are identical. Applying our test procedure, we obtained a p-value equal to 0.156 and a test statistic equal to $V = 2.01$. So, we conclude that the null hypothesis of equal dependence structure by age is not rejected at a significant level $\alpha = 5\%$. There is no evidence to believe that the dependence structure of $\mathbf{X}$ changes over age. We proceeded to an Anova procedure and we present the results in Appendix G where Table 6 dresses the two-by-two comparisons. We can see that there are no significant differences between two successive years. But the difference increases with the gap between the years, as for example between the first age categories and the last ones.

Clustering Finally we applied the clustering algorihtm to the previous data.

- For the Iris dataset, as expected we obtain two groups: $\{Versicolor, Virginica\}$ and $\{Setosa\}$.
- For the Insurance dataset, i) in the three-sample paired case we obtain only one cluster which confirms the result of the test; ii) in the three sample independent case we obtain two clusters: {"Employee","Spouse"} and {"Dependent"} in accordance with the two-sample tests. iii) in the ten-sample case we obtained only one cluster which is concordant with the global testing procedure.

9. Conclusion. In this paper we used new quantities, called copulas coefficients, for testing the equality of copulas. A data driven procedure is developed in the two-sample case, for independent as well as paired populations. Its extension to the K-sample case is obtained by a second data driven method and then our test can be seen as an automatic comparison method. In this sense the test can also be used as an automatic clustering method permitting to regroup populations having the same dependence structure, whatever their distributions.

It can lead to various applications to bring together similar populations or on the contrary to have very diverse populations.

An important simulation study shows the behaviour of this approach and its practical implementation for more than two populations. The test is simple to use and can run for large dimensions. For the two-sample case it seems as efficient as his competitor proposed in Rémillard and Scaillet (2009). A R program of our procedure is available on Github-yvesngounou. Assumption (5) of square integrability can be bypassed since all copula coefficients exist and then the test can be applied to any copulas as a test of equality of all their coefficients.

Furthermore our approach can be extended in different directions and we mention two of them below:

- First, it can be used to compare Spearman's rho in the two-sample case. Let us recall that for any continuous bivariate random variable (X_1, X_2) with copula C, the Spearman's rho can be express as (see Nelsen (2007)):

$$\rho_C = 12 \int_0^1 \int_0^1 C(u,v) du dv - 3.$$

  Then the Spearman's rho coincides with the first copula coefficient, that is $\rho_C = \rho_{11}$. We can immediately deduce an estimator of the Spearman's rho as follows:

$$\hat{\rho}_C = \hat{\rho}_{11} = \frac{3}{n} \sum_{i=1}^{n} \left(2\widehat{U}_{i1} - 1\right) \left(2\widehat{U}_{i2} - 1\right).$$

  This estimator seems new from recent reviews on this topic (see for instance Pérez and Prieto-Alaiz (2016)) and it could be used to construct a goodness-of-fit test.
- Second, the proposed method in this paper is based on the copula coefficients. These quantities characterize the dependence structure and could be used for testing independence between vectors. This is a work in progress.

10. Proofs.

Proof of Theorem 3.1. We want to show that $\mathbb{P}(D(n) > 1) \to 0$ as n tends to infinity. We have

$$\begin{aligned}
\mathbb{P}_0\Big(D(n) > 1\Big) &= \mathbb{P}_0\Big(\exists k \in \{2, \cdots, d(n)\} : V_k^{(1,2)} - k\, q_n \geq V_1^{(1,2)} - q_n\Big) \\
&= \mathbb{P}_0\Big(\exists k \in \{2, \cdots, d(n)\} : V_k^{(1,2)} - V_1^{(1,2)} \geq (k-1) q_n\Big) \\
&= \mathbb{P}_0\Big(\exists k \in \{2, \cdots, d(n)\} : n \sum_{\mathbf{j} \in \mathcal{H}^*(k)} (r_{\mathbf{j}}^{(1,2)})^2 \geq (k-1) q_n\Big) \\
&\leq \mathbb{P}_0\Big(n \sum_{\mathbf{j} \in \mathcal{H}^*(d(n))} (r_{\mathbf{j}}^{(1,2)})^2 \geq q_n\Big), \qquad (18)
\end{aligned}$$

with $\mathcal{H}(k)$ satisfying (14) and where $\mathcal{H}^*(k) = \mathcal{H}(k) \backslash \mathcal{H}(1)$. The last inequality comes from the fact that if a sum of $(k-1)$ positive terms, say $\sum_{j=2}^{k} r_j$ is greater than a constant c, then necessarily there exists a term r_j such that $r_j > c/(k-1)$. The important point here is that $card(\mathcal{H}^*(k)) = k - 1$, which corresponds to the number of elements of the form $(r_{\mathbf{j}}^{(1,2)})^2$ in the difference $V_k^{(1,2)} - V_1^{(1,2)}$. For simplification of notation, we write $\mathcal{H}^*$ instead of $\mathcal{H}^*(d(n))$.

Under the null $\rho_{\mathbf{j}}^{(1)} = \rho_{\mathbf{j}}^{(2)}$ and we decompose $(r_{\mathbf{j}}^{(1,2)})^2$ as follows

$$\begin{aligned}\big(r_{\mathbf{j}}^{(1,2)}\big)^2 &= \big((\widehat{\rho}_{\mathbf{j}}^{(1)} - \rho_{\mathbf{j}}^{(1)}) - (\widehat{\rho}_{\mathbf{j}}^{(2)} - \rho_{\mathbf{j}}^{(2)})\big)^2 \\ &\leq 2(\widehat{\rho}_{\mathbf{j}}^{(1)} - \rho_{\mathbf{j}}^{(1)})^2 + 2(\widehat{\rho}_{\mathbf{j}}^{(2)} - \rho_{\mathbf{j}}^{(2)})^2,\end{aligned}$$

that we combine with the standard inequality for positive random variables: $\mathbb{P}(X+Y>z) \leq \mathbb{P}(X>z/2) + \mathbb{P}(Y>z/2)$, to get

$$\begin{aligned}\mathbb{P}_0\big(D(n)>1\big) &\leq \mathbb{P}_0\big(n \sum_{\mathbf{j}\in\mathcal{H}^*} (\widehat{\rho}_{\mathbf{j}}^{(1)} - \rho_{\mathbf{j}}^{(1)})^2 \geq q_n/4\big) + \mathbb{P}_0\big(n \sum_{\mathbf{j}\in\mathcal{H}^*} \widehat{\rho}_{\mathbf{j}}^{(2)} - \rho_{\mathbf{j}}^{(2)})^2 \geq q_n/4\big) \\ &:= \qquad A \qquad + \qquad B.\end{aligned}$$

We now study the first quantity A, the quantity B being similar. Writing

$$\tilde{\rho}_{\mathbf{j}}^{(1)} = \frac{1}{n}\sum_{s=1}^{n} L_{j_1}(U_{s,1}^{(1)}) \cdots L_{j_p}(U_{s,p}^{(1)})$$

we obtain

$$\begin{aligned}\widehat{\rho}_{\mathbf{j}}^{(1)} - \rho_{\mathbf{j}}^{(1)} &= (\widehat{\rho}_{\mathbf{j}}^{(1)} - \tilde{\rho}_{\mathbf{j}}^{(1)}) + (\tilde{\rho}_{\mathbf{j}}^{(1)} - \rho_{\mathbf{j}}^{(1)}) \\ &:= \qquad E_{\mathbf{j}} \qquad + \qquad G_{\mathbf{j}},\end{aligned} \tag{19}$$

where

$$E_{\mathbf{j}} = \frac{1}{n}\sum_{s=1}^{n}\Big(L_{j_1}(\widehat{U}_{s,1}^{(1)}) \cdots L_{j_p}(\widehat{U}_{s,p}^{(1)}) - L_{j_1}(U_{s,1}^{(1)}) \cdots L_{j_p}(U_{s,p}^{(1)})\Big),$$

$$G_{\mathbf{j}} = \frac{1}{n}\sum_{s=1}^{n}\Big(L_{j_1}(U_{s,1}^{(1)}) \cdots L_{j_p}(U_{s,p}^{(1)}) - \mathbb{E}\big(L_{j_1}(U_1^{(1)}) \cdots L_{j_p}(U_p^{(1)})\big)\Big).$$

Then we have

$$A \leq \mathbb{P}_0(n \sum_{\mathbf{j}\in\mathcal{H}^*} (E_{\mathbf{j}})^2 \geq q_n/16) + \mathbb{P}_0(n \sum_{\mathbf{j}\in\mathcal{H}^*} (G_{\mathbf{j}})^2 \geq q_n/16). \tag{20}$$

We first study the quantity involving $E_{\mathbf{j}}$ in (20). Write

$$S_i^{(1)} = \sup_x |\widehat{F}_i^{(1)}(x) - F_i^{(1)}(x)|, \quad i = 1, \cdots, p. \tag{21}$$

Applying the mean value theorem to $E_{\mathbf{j}}$ we obtain

$$|E_{\mathbf{j}}| \leq \frac{1}{n}\sum_{s=1}^{n}\sum_{i=1}^{p} S_i^{(1)} \sup_x |L'_{j_i}(x) \prod_{u\neq i} L_{j_u}(x)|.$$

From (43) and (44) (see Appendix C) there exists a constant $\tilde{c} > 0$ such that

$$|E_{\mathbf{j}}| \leq \tilde{c}\sum_{i=1}^{p} S_i^{(1)} (j_i{}^{5/2} \prod_{u\neq i} j_u{}^{1/2}). \tag{22}$$

When $\mathbf{j}$ belongs to $\mathcal{H}^* = \mathcal{H}^*(d(n))$ we necessarily have $\|\mathbf{j}\|_1 \leq d(n)$. It follows that

$$\mathbb{P}_0\big(n \sum_{\mathbf{j}\in\mathcal{H}^*} (E_{\mathbf{j}})^2 \geq q_n/16\big)$$

$$\leq \mathbb{P}_0\big(n \sum_{\mathbf{j}\in\mathcal{H}^*} \tilde{c} \sum_{i=1}^{p}\sum_{i'=1}^{p} S_i^{(1)} S_{i'}^{(1)} j_i^{5/2} j_{i'}^{5/2} \prod_{s\neq i} {j_s}^{1/2} \prod_{s'\neq i'} {j_{s'}}^{1/2} \geq q_n/16\big)$$

$$\leq \mathbb{P}_0\big(\tilde{c} \sum_{i=1}^{p}\sum_{i'=1}^{p} n S_i^{(1)} S_{i'}^{(1)} d(n)^{p+4} \geq q_n/16\big)$$

$$\to 0 \text{ as } n\to\infty, \tag{23}$$

since for all $i = 1, \cdots, p$, $\sqrt{n}S_i^{(1)}$ converges in law to a Kolmogorov distribution and $d(n)^{p+4} = o(q_n)$ by **(A)**.

Coming back to (19) we now study the quantity involving $G_{\mathbf{j}}$. First note that $\mathbb{E}(G_{\mathbf{j}}) = 0$. Moreover, $\mathbb{V}(G_{\mathbf{j}}) = \mathbb{V}(\prod_{i=1}^{p} L_{j_i}(U_i^{(1)}))/n$. Then, by Markov inequality we have

$$\mathbb{P}_0\big(n \sum_{\mathbf{j}\in\mathcal{H}^*} (G_{\mathbf{j}})^2 \geq q_n/16\big) \leq \frac{\sum_{\mathbf{j}\in\mathcal{H}^*} \mathbb{V}(\prod_{i=1}^{p} L_{j_i}(U_i^{(1)}))}{q_n/16}.$$

From (43) (see Appendix C) there exists a constant $c > 0$ such that

$$\mathbb{V}(\prod_{i=1}^{p} L_{j_i}(U_i^{(1)})) \leq c \prod_{i=1}^{p} j_i.$$

It follows that

$$\mathbb{P}_0\big(n \sum_{\mathbf{j}\in\mathcal{H}^*} (G_{\mathbf{j}})^2 \geq q_n/16\big) \leq \frac{c d(n)^p}{q_n/16} \to 0 \text{ as } n\to\infty, \tag{24}$$

We now combine (23) and (24) with (20) to conclude that

$$A \to 0, \text{ as } n\to\infty.$$

In the same manner we can show that $B \to 0$, as $n\to\infty$, which completes the proof. ■

Proof of Theorem 3.2. Let $\mathbf{j} = (1, 1, \cdots, 0, 0)$. We have $V_1^{(1,2)} = T_{2,1}^{(1,2)} = \left(\sqrt{n} r_{\mathbf{j}}^{(1,2)}\right)^2$ and we can decompose $\sqrt{n} r_{\mathbf{j}}^{(1,2)}$ under the null as follows:

$$\begin{aligned}
\sqrt{n} r_{\mathbf{j}}^{(1,2)} &= \sqrt{n}\left(\widehat{\rho}_{\mathbf{j}}^{(1)} - \widehat{\rho}_{\mathbf{j}}^{(2)}\right) \\
&= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} L_1(\widehat{U}_{i,1}^{(1)}) L_1(\widehat{U}_{i,2}^{(1)}) - \frac{1}{n}\sum_{i=1}^{n} L_1(\widehat{U}_{i,1}^{(2)}) L_1(\widehat{U}_{i,2}^{(2)})\right) \\
&= \sqrt{n}\left(\frac{1}{n}(\sum_{i=1}^{n} L_1(\widehat{U}_{i,1}^{(1)}) L_1(\widehat{U}_{i,2}^{(1)}) - m\right) - \sqrt{n}\left(\frac{1}{n}(\sum_{i=1}^{n} L_1(\widehat{U}_{i,1}^{(2)}) L_1(\widehat{U}_{i,2}^{(2)}) - m\right) \\
&:= \qquad R_n^{(1)} \qquad - \qquad R_n^{(2)}
\end{aligned}$$

where under the null

$$m = \mathbb{E}(L_1(\widehat{U}_{i,1}^{(1)}) L_1(\widehat{U}_{i,2}^{(1)})) = \mathbb{E}(L_1(\widehat{U}_{i,1}^{(2)}) L_1(\widehat{U}_{i,2}^{(2)})).$$

By Taylor expansion, using the fact that the Legendre polynomials satisfy $L_1' = 2\sqrt{3}$ and $L_1'' = 0$, we obtain

$$\begin{aligned}
R_n^{(1)} &= \sqrt{n}\Big(\int\int L_1(\widehat{F}_1^{(1)}(x)) L_1(\widehat{F}_2^{(1)}(y)) d\widehat{F}_n^{(1)}(x,y) - m \Big) \\
&= \sqrt{n}\Big(\int\int L_1(F_1^{(1)}(x)) L_1(F_2^{(1)}(y)) d\widehat{F}_n^{(1)}(x,y) - m \Big) \\
&\quad + \sqrt{n} \int\int (\widehat{F}_1^{(1)}(x) - F_1^{(1)}(x)) 2\sqrt{3} L_1(F_2^{(1)}(y)) dF^{(1)}(x,y) \\
&\quad + \sqrt{n} \int\int (\widehat{F}_2^{(1)}(y) - F_2^{(1)}(y)) 2\sqrt{3} L_1(F_1^{(1)}(x)) dF^{(1)}(x,y) \\
&\quad + \sqrt{n} \int\int (\widehat{F}_1^{(1)}(x) - F_1^{(1)}(x)) 2\sqrt{3} L_1(F_2^{(1)}(y)) d(\widehat{F}_n^{(1)}(x,y) - F^{(1)}(x,y)) \\
&\quad + \sqrt{n} \int\int (\widehat{F}_2^{(1)}(y) - F_2^{(1)}(y)) 2\sqrt{3} L_1(F_1^{(1)}(x)) d(\widehat{F}_n^{(1)}(x,y) - F^{(1)}(x,y)) \\
&:= \sqrt{n} \left(A_{1,n}^{(1)} + A_{2,n}^{(1)} + A_{3,n}^{(1)} + B_n^{(1)} + C_n^{(1)} \right).
\end{aligned}$$

By symmetry, the second term $R_n^{(2)}$ can be expressed as:

$$R_n^{(2)} = \sqrt{n} \left(A_{1,n}^{(2)} + A_{2,n}^{(2)} + A_{3,n}^{(2)} + B_n^{(2)} + C_n^{(2)} \right),$$

and finally

$$\sqrt{n} r_{\mathbf{j}}^{(1,2)} = \sqrt{n} \left(A_{1,n}^{(1)} + A_{2,n}^{(1)} + A_{3,n}^{(1)} - A_{1,n}^{(2)} - A_{2,n}^{(2)} - A_{3,n}^{(2)} + B_n^{(1)} + C_n^{(1)} - B_n^{(2)} - C_n^{(2)} \right).$$

Since $(\widehat{F} - F)(x) = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} \Big(\mathbb{I}(X_i \le x - F(x) \Big)$, we can rewrite

$$\begin{aligned}
A_{1,n}^{(1)} + A_{2,n}^{(1)} + A_{3,n}^{(1)} &= \frac{1}{n} \sum_{i=1}^{n} \Big\{ L_1(F_1^{(1)}(X_{1,i}^{(1)}) L_1(F_2^{(1)}(X_{2,i}^{(1)})) - m \\
&\qquad + 2\sqrt{3} \int\int (\mathbb{I}(X_{1,i}^{(1)} \le x) - F_1^{(1)}(x)) L_1(F_2^{(1)}(y)) dF^{(1)}(x,y) \\
&\qquad + 2\sqrt{3} \int\int (\mathbb{I}(X_{2,i}^{(1)} \le y) - F_2^{(1)}(y)) L_1(F_1^{(1)}(x)) dF^{(1)}(x,y) \Big\} \\
&:= \frac{1}{n} \sum_{i=1}^{n} (Z_{1,i}^{(1)} + Z_{2,i}^{(1)} + Z_{3,i}^{(1)})
\end{aligned}$$

and then

$$\begin{aligned}
A_{1,n}^{(1)} + A_{2,n}^{(1)} + A_{3,n}^{(1)} - A_{1,n}^{(2)} - A_{2,n}^{(2)} - A_{3,n}^{(2)} &= A_{1,n}^{(1)} - A_{1,n}^{(2)} + A_{2,n}^{(1)} - A_{2,n}^{(2)} + A_{3,n}^{(1)} - A_{3,n}^{(2)} \\
&:= \frac{1}{n} \sum_{i=1}^{n} \left((Z_{1,i}^{(1)} - Z_{1,i}^{(2)}) + (Z_{2,i}^{(1)} - Z_{2,i}^{(2)}) + (Z_{3,i}^{(1)} - Z_{3,i}^{(2)}) \right) \\
&:= \frac{1}{n} \sum_{i=1}^{n} Z_i
\end{aligned}$$

where Z_i are iid random variables. Clearly $\mathbb{E}(Z_{1,i}^{(1)} - Z_{1,i}^{(2)}) = 0$. Since $\mathbb{E}(\mathbb{I}(X_{1,i}^{(1)} \leq x)) = F_1^{(1)}(x)$ and $\mathbb{E}(\mathbb{I}(X_{1,i}^{(2)} \leq x)) = F_1^{(2)}(x)$, we also have $\mathbb{E}(Z_{2,i}^{(1)} - Z_{2,i}^{(2)}) = 0$ and similarly $\mathbb{E}(Z_{3,i}^{(1)} - Z_{3,i}^{(2)}) = 0$. Moreover, $\mathbb{V}(Z_i) \leq \infty$. By the Central Limit Theorem we have

$$\sqrt{n}\left(A_{1,n}^{(1)} + A_{2,n}^{(1)} + A_{3,n}^{(1)} - A_{1,n}^{(2)} - A_{2,n}^{(2)} - A_{3,n}^{(2)}\right) \to N(0, \sigma^2(1,2)),$$

where

$$\begin{aligned}
\sigma^2(1,2) = \mathbb{V}(Z_i) &= \mathbb{V}\Big(Z_{1,i}^{(1)} - Z_{1,i}^{(2)} + Z_{2,i}^{(1)} - Z_{2,i}^{(2)} + Z_{3,i}^{(1)} - Z_{3,i}^{(2)}\Big) \\
&= \mathbb{V}\Bigg(L_1(U_1^{(1)})L_1(U_2^{(1)}) - L_1(U_1^{(2)})L_1(U_2^{(2)}) \\
&\quad + 2\sqrt{3}\int\int \big(\mathbb{I}(X_1^{(1)} \leq x) - F_1^{(1)}(x)\big) L_1(F_2^{(1)}(y)) dF^{(1)}(x,y) \\
&\quad - 2\sqrt{3}\int\int \big(\mathbb{I}(X_1^{(2)} \leq x) - F_1^{(2)}(x)\big) L_1(F_2^{(2)}(y)) dF^{(2)}(x,y) \\
&\quad + 2\sqrt{3}\int\int \big(\mathbb{I}(X_2^{(1)} \leq y) - F_2^{(1)}(y)\big) L_1(F_1^{(1)}(x)) dF^{(1)}(x,y) \\
&\quad - 2\sqrt{3}\int\int \big(\mathbb{I}(X_2^{(2)} \leq y) - F_2^{(2)}(y)\big) L_1(F_1^{(2)}(x)) dF^{(2)}(x,y) \Bigg).
\end{aligned}$$

We proceed to show that $B_n^{(1)}$, $C_n^{(1)}$, $B_n^{(2)}$ and $C_n^{(2)}$ are $o_{\mathbb{P}}(n^{-1/2})$. We treat only the case of $B_n^{(1)}$, since the case of $C_n^{(1)}$ is similar and by symmetric the same reasoning applies to $B_n^{(2)}$ and $C_n^{(2)}$. We can rewrite

$$\begin{aligned}
\sqrt{n}B_n^{(1)} &= 2\sqrt{3}\int\int (\widehat{F}_1^{(1)}(x) - F_1^{(1)}(x)) L_1(F_2^{(1)}(y)) d(\widehat{F}_n^{(1)}(x,y) - F^{(1)}(x,y)) \\
&= \frac{2\sqrt{3}}{n}\sum_{k=1}^{n}\int\int \left((\mathbb{I}(X_{1,k}^{(1)} \leq x) - F_1^{(1)}(x)\right) L_1(F_2^{(1)}(y)) d(\widehat{F}_n^{(1)}(x,y) - F^{(1)}(x,y)) \\
&:= -\frac{2\sqrt{3}}{n}\sum_{k=1}^{n}\left(B_{1,k,n} + B_{2,k,n}\right),
\end{aligned}$$

where

$$B_{1,k,n} = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{1}_{X_{k,1}^{(1)} \leq X_{i,1}^{(1)}} - U_{i,1}^{(1)}\right) L_1(U_{i,2}^{(1)}) - \frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{1}_{X_{k,1}^{(1)} \leq X_{i,1}^{(1)}} - \widehat{U}_{i,1}^{(1)}\right) L_1(\widehat{U}_{i,2}^{(1)})$$

and

(25)

$$B_{2,k,n} = \iint \left(\mathbb{1}_{X_{k,1}^{(1)} \leq x} - F_1^{(1)}(x)\right) L_1\big(F_2^{(1)}(y)\big) dF_{1,2}^{(1)}(x,y) - \frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{1}_{X_{k,1}^{(1)} \leq X_{i,1}^{(1)}} - U_{i,1}^{(1)}\right) L_1(U_{i,2}^{(1)}).$$

For $B_{1,k,n}$, we have

$$B_{1,k,n} = \frac{2\sqrt{3}}{n}\sum_{i=1}^{n}\mathbb{1}_{X_{k,1}^{(1)} \leq X_{i,1}^{(1)}}\left(U_{i,2}^{(1)} - \widehat{U}_{i,2}^{(1)})\right) + \frac{2\sqrt{3}}{n}\sum_{i=1}^{n}\widehat{U}_{i,1}^{(1)}\left(\widehat{U}_{i,2}^{(1)} - U_{i,2}^{(1)}\right) + \frac{1}{n}\sum_{i=1}^{n}L_1(U_{i,2}^{(1)})\left(\widehat{U}_{i,1}^{(1)} - U_{i,1}^{(1)}\right).$$

By Glivenko-Cantelli's Theorem we obtain

$$|B_{1,k,n}| \leq 2\sqrt{3}S_2^{(1)} + 2\sqrt{3}S_1^{(1)} + \sqrt{3}S_1^{(1)}$$
$$= o_{\mathbb{P}}(1). \tag{26}$$

We can decompose $B_{2,k,n}$ as follows

$$\begin{aligned} B_{2,k,n} &= \left(\frac{1}{n}\sum_{i=1}^{n} U_{i,1}^{(1)} L_1(U_{i,2}^{(1)}) - \iint F_1^{(1)}(x) L_1\big(F_2^{(1)}(y)\big) dF_{1,2}^{(1)}(x,y) \right) \\ &\quad + \left(\iint \mathbb{1}_{X_{k,1}^{(1)} \leq x_1^{(1)}} L_1\big(F_2^{(1)}(y)\big) dF_{1,2}^{(1)}(x,y) - \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_{k,1}^{(1)} \leq X_{i,1}^{(1)}} L_1(U_{i,2}^{(1)}) \right) \\ &\equiv B_{2,k,n}^1 + B_{2,k,n}^2. \end{aligned}$$

To deal with $B_{2,k,n}^1$, we note that

$$\begin{aligned} B_{2,k,n}^1 &= \frac{1}{n}\sum_{s=1}^{n} U_{s,1}^{(1)} L_1(U_{s,2}^{(1)}) - \iint F_1^{(1)}(x) L_1\big(F_2^{(1)}(y)\big) dF_{1,2}^{(1)}(x,y) \\ &= \frac{1}{n}\sum_{s=1}^{n} U_{s,1}^{(1)} L_1(U_{s,2}^{(1)}) - \mathbb{E}\Big(U_1^{(1)} L_1(U_2^{(1)})\Big). \end{aligned}$$

Since $(U_{1,1}^{(1)}, U_{1,2}^{(1)}), (U_{2,1}^{(1)}, U_{2,2}^{(1)}), \cdots, (U_{n_1,1}^{(1)}, U_{n_1,2}^{(1)})$ are iid from $(U_1^{(1)}, U_2^{(1)})$, the Weak Law of Large Numbers and the Continuous Mapping Theorem show that

$$B_{2,k,n}^1 = o_{\mathbb{P}}(1). \tag{27}$$

For $B_{2,k,n}^2$, we have

$$\begin{aligned} B_{2,k,n}^2 &= \iint \mathbb{1}_{X_{k,1}^{(1)} \leq x} L_1\big(F_2^{(1)}(y)\big) dF_{1,2}^{(1)}(x,y) - \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_{k,1}^{(1)} \leq X_{i,1}^{(1)}} L_1(U_{i,2}^{(1)}) \\ &= \iint \mathbb{1}_{F_1^{(1)}(X_{k,1}^{(1)}) \leq F_1^{(1)}(x)} L_1\big(F_2^{(1)}(y)\big) dF_{1,2}^{(1)}(x,y) - \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{F_1^{(1)}(X_{k,1}^{(1)}) \leq F_1^{(1)}(X_{i,1}^{(1)})} L_1(U_{i,2}^{(1)}) \\ &= \int_0^1 \int_0^1 \mathbb{1}_{U_{k,1}^{(1)} \leq u} L_1(v) dC^{(1)}(u,v) - \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{U_{k,1}^{(1)} \leq U_{i,1}^{(1)}} L_1(U_{i,2}^{(1)}) \end{aligned}$$

and since $U_{i,1}^{(1)}$ has continuous uniform distribution it follows that

$$\begin{aligned} |B_{2,k,n}^2| &\leq \sup_{t\in[0,1]} \left| \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{t \leq U_{i,1}^{(1)}} L_1(U_{i,2}^{(1)}) - \int_0^1 \int_0^1 \mathbb{1}_{t \leq u_1^{(1)}} L_1(u_2^{(1)}) dC^{(1)}(u_1^{(1)}, u_2^{(1)}) \right| \\ &\leq \sup_{t\in[0,1]} \left| \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{t \leq U_{i,1}^{(1)}} L_1(U_{i,2}^{(1)}) - \mathbb{E}\left(\mathbb{1}_{t \leq U_1^{(1)}} L_1(U_2^{(1)}) \right) \right| \\ &\leq \sup_{t\in[0,1]} \left| g\left(t, (U_{1,1}^{(1)}, U_{1,2}^{(1)}), \cdots, (U_{n,1}^{(1)}, U_{n,2}^{(1)})\right) - \mathbb{E}\Big(g\big(t, (U_1^{(1)}, U_2^{(1)})\big)\Big) \right| \end{aligned}$$

where

$$g\left(t, z_1, \cdots, z_n\right) = \frac{1}{n}\sum_{k=1}^{n} \mathbb{1}_{t \leq u_k} L_1(v_k), \text{ with } z_k = (u_k, v_k) \text{ for } k = 1, \cdots, n.$$

Observe that for all $t \in [0,1]$,

$$\sup_{\substack{z_1,\cdots,z_n,\\ z_i'}} \left|g(t,z_1,\cdots,z_{n_1}) - g(t,z_1,\cdots,z_{i-1},z_i',z_{i+1},\cdots,z_n)\right| \leq \frac{2\|L_1\|_\infty}{n} = \frac{4\sqrt{3}}{n},$$

that is, if we change the ith variable z_i of g while keeping all the others fixed, then the value of the function does not change by more than $4\sqrt{3}/n$. Then, by McDiarmid's inequality, we get $\forall \epsilon > 0$

$$\mathbb{P}\left(\forall t, \left|g\left(t,(U_{1,1}^{(1)},U_{1,2}^{(1)}),\cdots,(U_{n,1}^{(1)},U_{n,2}^{(1)})\right) - \mathbb{E}\Big(g\big(t,(U_1^{(1)},U_2^{(1)})\big)\Big)\right| \geq \epsilon\right) \quad \leq 2e^{-n\epsilon^2/24} \underset{n\to\infty}{\longrightarrow} 0.$$

It implies that $B_{2,k,n}^2 = o_{\mathbb{P}}(1)$, and we conclude that $B_n^{(1)} = o_{\mathbb{P}}(n^{-1/2})$. The same result occurs for $C_n^{(1)}$. Finally, by symmetry we obtain $B_n^{(2)} = o_{\mathbb{P}}(n^{-1/2})$, and $C_n^{(2)} = o_{\mathbb{P}}(n^{-1/2})$, which proves the theorem.

■

Proof of Proposition 1. Let us define

$$\overline{W}_s = \frac{1}{n}\sum_{i=1}^n W_{i,s}, \text{ for } s = 1,2,$$

where

$$W_{i,s} = L_1(U_{i,1}^{(s)})L_1(U_{i,2}^{(s)}) + 2\sqrt{3}\int\int \big(\mathbb{I}(X_{i,1}^{(s)} \leq x) - F_1^{(s)}(x)\big)L_1(F_2^{(s)}(y))dF^{(1)}(x,y)$$
$$+2\sqrt{3}\int\int \big(\mathbb{I}(X_{i,2}^{(s)} \leq y) - F_2^{(s)}(y)\big)L_1(F_1^{(s)}(x))dF^{(1)}(x,y).$$

By construction $W_{1,1} - W_{1,2}, \cdots, , W_{n,1} - W_{n,2}$ are iid and we have

$$\frac{1}{n}\sum_{i=1}^n \Big(W_{i,1} - W_{i,2} - \overline{W}_1 + \overline{W}_2\Big)^2 \xrightarrow{\mathbb{P}} \sigma^2(1,2)). \tag{28}$$

According to Slusky's Lemma and (28), the proof is completed by showing that

$$\frac{1}{n}\sum_{i=1}^n \Big(W_{i,1} - W_{i,2} - \overline{W}_1 + \overline{W}_2\Big)^2 - \widehat{\sigma}^2(1,2)) \xrightarrow{\mathbb{P}} 0.$$

We have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \Big(W_{i,1} - W_{i,2} - \overline{W}_1 + \overline{W}_2\Big)^2 - \widehat{\sigma}^2(1,2)) \\
&= \frac{1}{n}\sum_{i=1}^n \Big(W_{i,1} - W_{i,2}\Big)^2 - \frac{1}{n}\sum_{i=1}^n \Big(M_{i,1} - M_{i,2}\Big)^2 + \Big(\overline{M}_1 - \overline{M}_2\Big)^2 - \Big(\overline{W}_1 - \overline{W}_2\Big)^2 \\
&= \frac{1}{n}\sum_{i=1}^n \Big(W_{i,1} - W_{i,2} - M_{i,1} + M_{i,2}\Big)\Big(W_{i,1} - W_{i,2} + M_{i,1} - M_{i,2}\Big) \\
&\quad -\frac{1}{n}\sum_{i=1}^n \Big(W_{i,1} - W_{i,2} - M_{i,1} + M_{i,2}\Big)\Big(\overline{M}_1 - \overline{M}_2 + \overline{W}_1 - \overline{W}_2\Big) \\
&= \frac{1}{n}\sum_{i=1}^n \Big(W_{i,1} - W_{i,2} - M_{i,1} + M_{i,2}\Big)\Big(W_{i,1} - W_{i,2} + M_{i,1} - M_{i,2} - \overline{M}_1 + \overline{M}_2 - \overline{W}_1 + \overline{W}_2\Big).
\end{aligned}$$

From (43), there exists a constant $\kappa > 0$ such that, for all $n > 0$ and for all $i = 1, \cdots, n$,

$$\max(|W_{i,1}|, |M_{i,1}|, |W_{i,2}|, |M_{i,2}|) \leq \kappa,$$

which implies that

$$\left| \frac{1}{n} \sum_{i=1}^{n} \left(W_{i,1} - W_{i,2} - \overline{W}_1 + \overline{W}_2 \right)^2 - \widehat{\sigma}^2(1,2)) \right| \leq \frac{8\kappa}{n} \sum_{i=1}^{n} |W_{i,1} - M_{i,1} + M_{i,2} - W_{i,2}| .$$

It remains to prove that $W_{i,1} - M_{i,1} + M_{i,2} - W_{i,2} \xrightarrow{\mathbb{P}} 0$. We have

$$W_{i,1} - M_{i,1} = I_{i,1} + 2\sqrt{3} I_{i,2} + 2\sqrt{3} I_{i,3}, \tag{29}$$

where

$$I_{i,1} = L_1(U_{i,1}^{(1)}) L_1(U_{i,2}^{(1)}) - L_1(\widehat{U}_{i,1}^{(1)}) L_1(\widehat{U}_{i,2}^{(1)})$$

$$I_{i,2} = \iint \left(\mathbb{I}(X_{i,1}^{(1)} \leq x) - F_1^{(1)}(x) \right) L_1\big(F_2^{(1)}(y)\big) dF^{(1)}(x,y), - \frac{1}{n} \sum_{k=1}^{n} \left(\mathbb{I}(X_{i,1}^{(1)} \leq X_{k,1}^{(1)}) - \widehat{U}_{k,1}^{(1)} \right) L_1(\widehat{U}_{k,2}^{(1)})$$

$$I_{i,3} = \iint \left(\mathbb{I}(X_{i,2}^{(1)} \leq x) - F_1^{(1)}(x) \right) L_1\big(F_2^{(1)}(y)\big) dF^{(1)}(x,y) - \frac{1}{n} \sum_{k=1}^{n} \left(\mathbb{I}(X_{i,2}^{(1)} \leq X_{k,2}^{(1)}) - \widehat{U}_{k,2}^{(1)} \right) L_1(\widehat{U}_{k,1}^{(1)}).$$

Since $L_1(t) = \sqrt{3}(2t-1)$, we get

$$\begin{aligned} |I_{i,1}| &= |2\sqrt{3} L_1(U_{i,1}^{(1)}) \left(U_{i,2}^{(1)} - \widehat{U}_2^{(1)} \right) + 2\sqrt{3} L_1(\widehat{U}_2^{(1)}) \left(U_{i,1}^{(1)} - \widehat{U}_1^{(1)} \right)| \\ &\leq 6(S_2^{(1)} + S_1^{(1)}) \quad = o_{\mathbb{P}}(1), \end{aligned}$$

where $S_2^{(1)}$ and $S_1^{(1)}$ are given by (21). Next, we remark that $I_{i,2} = B_{2,k,n}$, where $B_{2,k,n}$ is defined in (25). Then $I_{i,2} = o_{\mathbb{P}}(1)$ and similarly $I_{i,3} = o_{\mathbb{P}}(1)$. It follows that $W_{i,1} - M_{i,1} \xrightarrow{\mathbb{P}} 0$ and by symmetric we get $W_{i,2} - M_{i,2} \xrightarrow{\mathbb{P}} 0$ which completes the proof. ■

Proof of Theorem 4.1. Let us prove that $\mathbb{P}(s(\mathbf{n}) \geq 2)$ vanishes as $\mathbf{n} \to +\infty$. By definition of $s(\mathbf{n})$ we have:

$$\begin{aligned} &\mathbb{P}\big(s(\mathbf{n}) \geq 2\big) \\ &= \mathbb{P}\big(\text{there exists } 2 \leq k \leq v(K) : \; V_k - k p_{\mathbf{n}} \geq V_1 - p_{\mathbf{n}}\big) \\ &= \mathbb{P}\big(\text{there exists } 2 \leq k \leq v(K) : V_k - V_1 \geq (k-1) p_{\mathbf{n}}\big) \\ &= \mathbb{P}\big(\text{there exists } 2 \leq k \leq v(K) : \sum_{2 \leq ord_{\mathcal{V}}(\ell,m) \leq k} V_{D(n)}^{(\ell,m)} \geq (k-1) p_{\mathbf{n}}\big). \end{aligned}$$

Since the previous sum contains $(k-1)$ positive elements, there is at least one element greater than p_n. It follows that

$$\begin{aligned} \mathbb{P}\big(s(\mathbf{n}) \geq 2\big) &\leq \mathbb{P}\big(\text{there exists } (\ell,m) \text{ with } 2 \leq ord_{\mathcal{V}}(\ell,m) \leq v(K) : \; V_{D(n)}^{(\ell,m)} \geq p_{\mathbf{n}}\big) \\ &\leq \mathbb{P}\left(\sum_{2 \leq ord_{\mathcal{V}}(\ell,m) \leq v(K)} V_{D(n)}^{(\ell,m)} \geq p_n \right). \end{aligned}$$

First we can remark that $\mathcal{V}(K)$ is finite and then there is a finite number of terms in $\sum_{2\leq ord_{\mathcal{V}}(\ell,m)\leq v(K)} V_{D(n)}^{(\ell,m)}$. It follows that we simply have to show that the probability $\mathbb{P}(V_{D(n)}^{(\ell,m)} \geq p_n)$ vanishes as $n \to +\infty$ for any values of (ℓ, m) . Since $D(n) \leq d(n)$ have:

$$\begin{aligned}\mathbb{P}(V_{D(n)}^{(\ell,m)} \geq p_{\mathbf{n}}) &\leq \mathbb{P}(V_{d(n)}^{(\ell,m)} \geq p_{\mathbf{n}}) \\ &= \mathbb{P}_0\Big(n \sum_{\mathbf{j}\in\mathcal{H}(d(n))} (r_{\mathbf{j}}^{(\ell,m)})^2 \geq p_n\Big). \end{aligned} \tag{30}$$

Comparing (30) and (18) we can see that the study is now similar to the two-sample case and we can simply mimic the Proof of Theorem 3.1 to conclude.

■

REFERENCES

ABRAMOWITZ, M. and STEGUN, I. A. (1964). *Handbook of mathematical functions with formulas, graphs, and mathematical tables* **55**. US Government printing office.

BEARE, B. K. (2010). Copulas and temporal dependence. *Econometrica* **78** 395–410.

BOUZEBDA, S., KEZIOU, A. and ZARI, T. (2011). K-Sample Problem Using Strong Approximations of Empirical Copula Processes. *Mathematical Methods of Statistics* **20** 14–2.

CAN, S. U., EINMAHL, J. H. J. and LAEVEN, R. J. A. (2020). Goodness-of-fit testing for copulas: A distribution-free approach. *Bernoulli* **26** 3163 – 3190.

CAN, S. U., EINMAHL, J. H. J., KHMALADZE, E. V. and LAEVEN, R. J. A. (2015). Asymptotically distribution-free goodness-of-fit testing for tail copulas. *The Annals of Statistics* **43** 878 – 902.

CHERUBINI, U., LUCIANO, E. and VECCHIATO, W. (2004). *Copula methods in finance*. John Wiley & Sons.

DERUMIGNY, A., FERMANIAN, J. D. and MIN, A. (2021). Testing for equality between conditional copulas given discretized conditioning events. *arXiv:2008.09498*.

DHAR, S. S., CHAKRABORTY, B. and CHAUDHURI, P. (2014). Comparison of multivariate distributions using quantile–quantile plots and related tests. *Bernoulli* **20** 1484 – 1506.

GENEST, C., REMILLARD, B. and BEAUDOIN, D. (2009). Goodness-of-fit tests for copulas: A review and a power study. *Insurance: Mathematics and Economics* **44** 199-213.

GHOSH, S., SHEPPARD, L. W., HOLDER, M. T., LOECKE, T. D., REID, P. C., BEVER, J. D. and REUMAN, D. C. (2020). Copulas and their potential for ecology. *Advances in Ecological Research* **62** 409–468.

GRAZIER, K. L. and G'SELL, W. (2004). *Group Medical Insurance Claims Database Collection and Analysis. Report for public release*. Society of Actuaries.

HAMDAN, M. A. and AL-BAYYATI, H. A. (1971). Canonical Expansion of the Compound Correlated Bivariate Poisson Distribution. *Journal of the American Statistical Association* **66** 390–393.

HOYER, A. and KUSS, O. (2018). Meta-analysis for the comparison of two diagnostic tests - A new approach based on copulas. *Statistics in Medicine* **37** 739–748.

INGLOT, T. and LEDWINA, T. (2006). Towards data driven selection of a penalty function for data driven Neyman tests. *Linear Algebra and its Applications* **417** 124–133.

JOE, H. (2014). *Dependence modeling with copulas*. CRC press.

KALLENBERG, W. C. M. and LEDWINA, T. (1995). Consistency and Monte Carlo Simulation of a Data Driven Version of Smooth Goodness-of-Fit Tests. *The Annals of Statistics* **23** 1594 – 1608.

KIM, J.-M., JUNG, Y.-S., SUNGUR, E. A., HAN, K.-H., PARK, C. and SOHN, I. (2008). A copula method for modeling directional dependence of genes. *BMC bioinformatics* **9** 225.

LANCASTER, H. O. (1958). The Structure of Bivariate Distributions. *The Annals of Mathematical Statistics* **29** 719 – 736.

LEDWINA, T. (1994). Data-driven version of Neyman's smooth test of fit. *Journal of the American Statistical Association* **89** 1000–1005.

MASSART, P. (1990). The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The annals of Probability* 1269–1283.

MCNEIL, A. J., FREY, R. and EMBRECHTS, P. (2015). *Quantitative risk management: concepts, techniques and tools-revised edition*. Princeton university press.

NELSEN, R. B. (2007). *An introduction to copulas*. Springer Science & Business Media.

NEYMAN, J. (1937). » Smooth test» for goodness of fit. *Scandinavian Actuarial Journal* **1937** 149–199.

OMELKA, M., GIJBELS, I. and VERAVERBEKE, N. (2009). Improved kernel estimation of copulas: Weak convergence and goodness-of-fit testing. *The Annals of Statistics* **37** 3023 – 3058.

PÉREZ, A. and PRIETO-ALAIZ, M. (2016). A note on nonparametric estimation of copula-based multivariate extensions of Spearman's rho. *Statistics & Probability Letters* **112** 41–50.

REMILLARD, B. and PLANTE, J.-F. (2012). TwoCop: Nonparametric test of equality between two copulas R package version 1.0.

RÉMILLARD, B. and SCAILLET, O. (2009). Testing for equality between two copulas. *Journal of Multivariate Analysis* **100** 377-386.

SCHWARZ, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6** 461–464.

SHI, P., FENG, X. and BOUCHER, J.-P. (2016). Multilevel modeling of insurance claims using copulas. *The Annals of Applied Statistics* **10** 834 – 863.

SUPPLEMENTARY MATERIAL

This supplementary material document contains: A) The proof of Theorem 5.1; B) The rewritten of all results in the independent case, C) Further details about the Legendre polynomials; D) Representations of sepals and petals distributions for Iris dataset; E) Additional simulation and comparison in the two sample-case; F) Empirical levels for the ten sample case; G) The two-by-two comparison for Insurance dataset.

APPENDIX A: PROOF OF THEOREM 5.1

We give the proof for the case $k > 1$, the particular case $k = 1$ being similar. We first show that $\mathbb{P}(s(\mathbf{n}) \geq k)$ tends to 1. Under $H_1(k)$, we have for all $k' < k$:

$$\begin{aligned}
\mathbb{P}(s(\mathbf{n}) < k) &\leq \mathbb{P}\left(V_k - kp_{\mathbf{n}} \leq V_{k'} - k'p_{\mathbf{n}}\right) \\
&= 1 - \mathbb{P}\left((V_k - V_{k'}) \geq (k-k')p_{\mathbf{n}}\right) \\
&= 1 - \mathbb{P}\left(\sum_{k' < rank_{\mathcal{V}}(\ell,m) \leq k} V_{D(n)}^{(\ell,m)} \geq (k-k')p_{\mathbf{n}}\right) \\
&= 1 - \mathbb{P}\left(\sum_{k' < rank_{\mathcal{V}}(\ell,m) \leq k} n \sum_{\mathbf{j} \in \mathcal{H}(D(n))} (r_{\mathbf{j}}^{(\ell,m)})^2 \geq (k-k')p_{\mathbf{n}}\right) \\
&\leq 1 - \mathbb{P}\left(\mathbb{I}_{\{rank_{\mathcal{V}}(\ell,m)=k\}}\, n \sum_{\mathbf{j} \in \mathcal{H}(D(n_\ell, n_m))} (r_{\mathbf{j}}^{(\ell,m)})^2 \geq (k-k')p_{\mathbf{n}}\right).
\end{aligned}$$

When $rank_{\mathcal{V}}(\ell, m) = k$, under $H_1(k)$, since $C^{(\ell)} \neq C^{(m)}$, there exists $\mathbf{j}_0$ such that $\rho_{\mathbf{j}_0}^{(\ell)} \neq \rho_{\mathbf{j}_0}^{(m)}$, that is, $r_{\mathbf{j}_0}^{(\ell,m)} \neq 0$. We can write

$$\begin{aligned}
&\mathbb{P}\left(\mathbb{I}_{\{rank_{\mathcal{V}}(\ell,m)=k\}} n \sum_{\mathbf{j} \in \mathcal{H}(D(n))} (r_{\mathbf{j}}^{(\ell,m)})^2 \geq (k-k')p_{\mathbf{n}}\right) \\
&\geq \mathbb{P}\left(\mathbb{I}_{\{rank_{\mathcal{V}}(\ell,m)=k\}} n \mathbb{I}_{\mathbf{j}_0 \in \mathcal{H}(D(n))} (r_{\mathbf{j}_0}^{(\ell,m)})^2 \geq (k-k')p_{\mathbf{n}}\right), \qquad (31)
\end{aligned}$$

and we can decompose $r_{\mathbf{j}_0}^{(\ell,m)}$ as follows

$$\begin{aligned}
r_{\mathbf{j}_0}^{(\ell,m)} &= \left((\widehat{\rho}_{\mathbf{j}_0}^{(\ell)} - \rho_{\mathbf{j}_0}^{(\ell)}) - (\widehat{\rho}_{\mathbf{j}_0}^{(m)} - \rho_{\mathbf{j}_0}^{(m)})\right) + \left(\rho_{\mathbf{j}_0}^{(\ell)} - \rho_{\mathbf{j}_0}^{(m)})\right) \\
&:= \quad (\quad A \quad - \quad B \quad) \; + \quad D. \qquad (32)
\end{aligned}$$

We first decompose the quantities A and B. We only detail the calculus for A, since the case of B is similar. We have

$$\begin{aligned}
A &= (\widehat{\rho}_{\mathbf{j}_0}^{(\ell)} - \tilde{\rho}_{\mathbf{j}_0}^{(\ell)}) + (\tilde{\rho}_{\mathbf{j}_0}^{(\ell)} - \rho_{\mathbf{j}_0}^{(\ell)}) \\
&:= \quad E_{\mathbf{j}_0} \quad + \quad G_{\mathbf{j}_0}.
\end{aligned}$$

We can reuse (22) to get:

$$\begin{aligned}
|E_{\mathbf{j}_0}| &\leq \tilde{c} \sum_{i=1}^{p} S_i^{(\ell)} ({j_i}^{5/2} \prod_{u \neq i} {j_u}^{1/2}) \\
&\leq \tilde{c}' \|\mathbf{j}_0\|_1^{(p+4)/2} \sum_{i=1}^{p} S_i^{(\ell)},
\end{aligned}$$

for some constants $\tilde{c}$ and $\tilde{c}'$. Since $\sqrt{n}S_i^{(\ell)} = o_{\mathbb{P}}(1)$ (see for instance Massart (1990)) we have $nE_{\mathbf{j}_0}^2 = O_{\mathbb{P}}(1)$. As $G_{\mathbf{j}_0}$ is an empirical estimator we also have $nG_{\mathbf{j}_0}^2 = O_{\mathbb{P}}(1)$, which yields

$$nA^2 = O_{\mathbb{P}}(1). \tag{33}$$

We now consider the quantity D in (32). The inequality $\rho_{\mathbf{j}_0}^{(\ell)} \neq \rho_{\mathbf{j}_0}^{(m)}$ implies that

$$nD^2 = O_{\mathbb{P}}(n). \tag{34}$$

Finally, under $H_1(k)$, we combine (33) and (34) with (32) to get

$$n(r_{\mathbf{j}_0}^{(\ell,m)})^2 = O_{\mathbb{P}}(n).$$

If we prove that $\mathbb{I}_{\mathbf{j}_0 \in \mathcal{H}(D(n))} \to 1$ as $\mathbf{n}$ tends to infinity then (31) tends to 1, from assumption **(B)**. Mimicking the proof of Theorem 3.1 we can prove that $\mathbb{P}(D(n) < ord(\mathbf{j}_0, \|\mathbf{j}_0\|_1)) \to 0$ which gives the result.

Our next goal is to determine the limit of $\mathbb{P}(V < \epsilon)$ for $\epsilon > 0$. It is sufficient to prove that $\mathbb{P}(V_{s(n)} < \epsilon) \to 0$ as n tends to infinity. We have

$$\begin{aligned}
\mathbb{P}(V_{s(n)} < \epsilon) &= \sum_{s=1}^{v(K)} \mathbb{P}(V_s < \epsilon \cap s(n) = s) \\
&= \sum_{s=1}^{k-1} \mathbb{P}(V_s < \epsilon \cap s(n) = s) + \sum_{s=k}^{v(K)} \mathbb{P}(V_s < \epsilon \cap s(n) = s) \\
&\leq \sum_{s=1}^{k-1} \mathbb{P}(V_s < \epsilon \cap s(n) = s) + \sum_{s=k}^{v(K)} \mathbb{P}(V_s < \epsilon) \\
&:= E + F.
\end{aligned}$$

From what has already been proved, under $H_1(k)$

$$\lim_{n\to\infty} E = \sum_{s=1}^{k-1} \lim_{n\to\infty} \mathbb{P}(V_s < \epsilon)\mathbb{P}(s(n) = s) = 0.$$

For the second quantity F we obtain

$$\lim_{n\to\infty} F \leq \sum_{s=k}^{v(K)} \lim_{n\to\infty} \mathbb{P}(V_s < \epsilon) \leq (v(K) - k) \lim_{n\to\infty} \mathbb{P}(V_k < \epsilon),$$

which is due to the fact that the statistics are embedded. Let (ℓ, m) be such that $rank_{\mathcal{V}}(\ell, m) = k$. Since $V_k > V_{D(n)}^{(\ell,m)}$, we have

$$\lim_{n\to\infty} \mathbb{P}(V_k < \epsilon) \leq \lim_{n\to\infty} \mathbb{P}(V_{D(n)}^{(\ell,m)} < \epsilon).$$

Under $H_1(k)$, as in the proof of Theorem 3.1 we can see that the probability $\mathbb{P}(D(n) < k)$ tends to zero as n tends to infinity. It follows that

$$\lim_{n\to\infty} \mathbb{P}(V_{D(n)}^{(\ell,m)} < \epsilon) = \lim_{n\to\infty} \mathbb{P}(V_{D(n)}^{(\ell,m)} < \epsilon \cap D(n) \geq k)$$

and since the statistics are embedded we have $V_{k'}^{(\ell,m)} \geq n\left(r_{\mathbf{j}_0}^{(\ell,m)}\right)^2$ for all $k' \geq k$ which implies that

$$\lim_{n\to\infty} \mathbb{P}(V_{D(n)}^{(\ell,m)} < \epsilon) \leq \lim_{n\to\infty} \mathbb{P}(n\left(r_{\mathbf{j}_0}^{(\ell,m)}\right)^2 < \epsilon)$$

and finally

$$\lim_{n\to\infty} \mathbb{P}(V_{s(n)} < \epsilon) \leq \lim_{n\to\infty} (E+F) = 0.$$

■

APPENDIX B: THE INDEPENDENT CASE

We briefly describe the adaptation in the case of independent samples, rewriting the previous definitions and the main results.

The 2-sample independent case. The constructions (12) and (13) become

$$T_{2,k}^{(1,2)} = \frac{n_1 n_2}{n_1+n_2} \sum_{\mathbf{j}\in\mathcal{S}(2);ord(\mathbf{j},2)\leq k} (r_{\mathbf{j}}^{(1,2)})^2, \text{ for } 1 \leq k \leq c(2), \tag{35}$$

and, for $d > 2$ and $1 \leq k \leq c(d)$,

$$T_{d,k}^{(1,2)} = T_{d-1,c(d-1)}^{(1,2)} + \frac{n_1 n_2}{n_1+n_2} \sum_{\mathbf{j}\in\mathcal{S}(d);ord(\mathbf{j},d)\leq k} (r_{\mathbf{j}}^{(1,2)})^2. \tag{36}$$

Then (14) and (15) become

$$V_k^{(1,2)} = \frac{n_1 n_2}{n_1+n_2} \sum_{\mathbf{j}\in\mathcal{H}(k)} (r_{\mathbf{j}}^{(1,2)})^2, \tag{37}$$

$$D(n_1,n_2) := \min\Big\{ \operatorname*{argmax}_{1\leq k\leq d(n_1,n_2)} (V_k^{(1,2)} - k q_{\mathbf{n}}) \Big\}, \tag{38}$$

where $q_{\mathbf{n}}$ and $d(n_1,n_2)$ tend to $+\infty$ as $n_1, n_2 \to +\infty$. A classical choice for $q_{\mathbf{n}}$ is $\alpha \log(2n_1n_2/(n_1+n_2))$, where α can be simply equal to 1, or obtained by the tuning procedure described in Section 7.1. When $n_1 = n_2 = n$ it gives $\alpha\log(n)$.

Finally, the associated data-driven test statistic to compare C_1 and C_2 is

$$V^{(1,2)} = V_{D(n_1,n_2)}^{(1,2)}. \tag{39}$$

We consider the following rate for the number of components in the statistic :

(A') $d(n_1,n_2)^{(p+4)} = o(p_{\mathbf{n}})$

THEOREM B.1. *If* **(A')** *holds, then, under* $\tilde{H}_0$, $D(n_1,n_2)$ *converges in Probability towards 1 as* $n_1, n_2 \to +\infty$.

Asymptotically, the null distribution reduces to that of $V_1^{(1,2)}$ and is given below.

THEOREM B.2. *Let* $\boldsymbol{j} = (1,1,0\cdots,0)$. *Then Under* $\tilde{H}_0$,

$$(V^{(1,2)})^{1/2} \xrightarrow{D} \mathcal{N}\left(0, \sigma^2(1,2)\right) \text{ with } \sigma^2(1,2) = (1-a_{1,2})\sigma^2(1) + a_{1,2}\sigma^2(2)$$

where for $s = 1,2$

$$\sigma^2(s) = \mathbb{V}\Bigg(L_1(U_1^{(s)}) L_1(U_2^{(s)}) + 2\sqrt{3}\int\int \big(\mathbb{I}(X_1^{(s)} \leq x) - F_1^{(s)}(x)\big) L_1(F_2^{(s)}(y)) dF^{(s)}(x,y)$$
$$+2\sqrt{3}\int\int \big(\mathbb{I}(X_2^{(s)} \leq y) - F_2^{(s)}(y)\big) L_1(F_1^{(s)}(x)) dF^{(s)}(x,y) \Bigg).$$

To normalize the test, we consider the following estimator

$$\widehat{\sigma}^2(1,2)) = \frac{(1-a_{1,2})}{n_1}\sum_{i=1}^{n_1}(M_{i,1}-\overline{M}_1)^2 + \frac{a_{1,2}}{n_2}\sum_{i=1}^{n_2}(M_{i,2}-\overline{M}_2)^2,$$

with

$$\overline{M}_s = \frac{1}{n}\sum_{i=1}^{n} M_{i,s}, \text{ for } s=\ell,m$$

where

$$M_{i,s} = L_1(\widehat{U}_{i,1}^{(s)})L_1(\widehat{U}_{i,2}^{(s)}) + \frac{2\sqrt{3}}{n}\sum_{k=1}^{n}\Big(\mathbb{I}\big(X_{i,1}^{(s)} \le X_{k,1}^{(s)}\big) - \widehat{U}_{k,1}^{(s)}\Big)L_1(\widehat{U}_{k,2}^{(s)}) + \frac{2\sqrt{3}}{n}\sum_{k=1}^{n}\Big(\mathbb{I}\big(X_{i,2}^{(s)} \le X_{k,2}^{(s)}\big) - \widehat{U}_{k,2}^{(s)}\Big)L_1(\widehat{U}_{k,1}^{(s)}).$$

PROPOSITION 2. *Under $\tilde{H}_0$,*

$$\widehat{\sigma}^2(1,2) := \xrightarrow{\mathbb{P}} \sigma^2(1,2).$$

We then obtain the following result.

COROLLARY B.3. *Assume that* **(A')** *holds. Under $\tilde{H}_0$, $V^{(1,2)}/\widehat{\sigma}^2(1,2)$ converges in law towards a chi-squared distribution χ_1^2 as $n_1, n_2 \to +\infty$.*

The K-sample independent case. Write $\mathbf{n} = (n_1, \cdots, n_K)$. The rule (16) becomes

$$s(\mathbf{n}) = \min\Big\{ \underset{1\le k\le v(K)}{\operatorname{argmax}} \big(V_k - kp_{\mathbf{n}}\big)\Big\}. \tag{40}$$

where $p_{\mathbf{n}}$ satisfies

(A") $d(\mathbf{n})^{p+4} = o(p_{\mathbf{n}})$.

In practice we choose $p_{\mathbf{n}} = \alpha\log(K^{(K-1)}n_1\cdots n_K/(n_1+\cdots+n_K)^{K-1})$. The following result shows that under the null, the penalty chooses the first element of $\mathcal{V}(K)$ asymptotically.

THEOREM B.4. *Assume that* **(A")** *holds. Under $\tilde{H}_0$, $s(\mathbf{n})$ converges in probability towards 1 as $\mathbf{n} \to +\infty$.*

COROLLARY B.5. *Assume that* **(A")** *holds. Under $\tilde{H}_0$, $V_{s(\mathbf{n})}/\widehat{\sigma}^2(1,2))$ converges in law towards a χ_1^2 distribution.*

Then our final data driven test statistic is given by

$$V = V_{s(\mathbf{n})}/\widehat{\sigma}^2(1,2)). \tag{41}$$

Alternative hypotheses. We need the following assumption:

(B') $p_{\mathbf{n}} = o(\mathbf{n})$.

THEOREM B.6. *Assume that* **(B')** *holds. Under $H_1(k)$, $s(\mathbf{n})$ converges in probability towards k as $\mathbf{n} \to +\infty$, and V converges to $+\infty$, that is, $\mathbb{P}(V < \epsilon) \to 0$ for all $\epsilon > 0$.*

APPENDIX C: LEGENDRE POLYNOMIALS

The Legendre polynomials used in this paper are defined on $[0,1]$ by

$$L_0 = 1, L_1(x) = \sqrt{3}(2x-1), \text{ and for } n > 1:$$

$$(n+1)L_{n+1}(x) = \sqrt{(2n+1)(2n+3)}(2x-1)L_n(x) - \frac{n\sqrt{2n+3}}{\sqrt{2n-1}}L_{n-1}(x). \tag{42}$$

They satisfy

$$\int_0^1 L_j(x)L_k(x)dx = \delta_{jk},$$

where $\delta_{jk} = 1$ if $j = k$ and 0 otherwise.

Throughout the proofs we will use the following inequalities satisfied by Legendre polynomials (see Abramowitz and Stegun (1964))

$$L_j(x) \leq cj^{1/2}, \quad \forall x \in (0,1) \tag{43}$$

$$L_j'(x) \leq c'j^{5/2}, \quad \forall x \in (0,1) \tag{44}$$

$$L_j''(x) \leq c''j^{9/2}, \quad \forall x \in (0,1) \tag{45}$$

where $c > 0, c' > 0, c'' > 0$, are constant.

APPENDIX D: REPRESENTATIONS OF SEPALS AND PETALS DISTRIBUTIONS

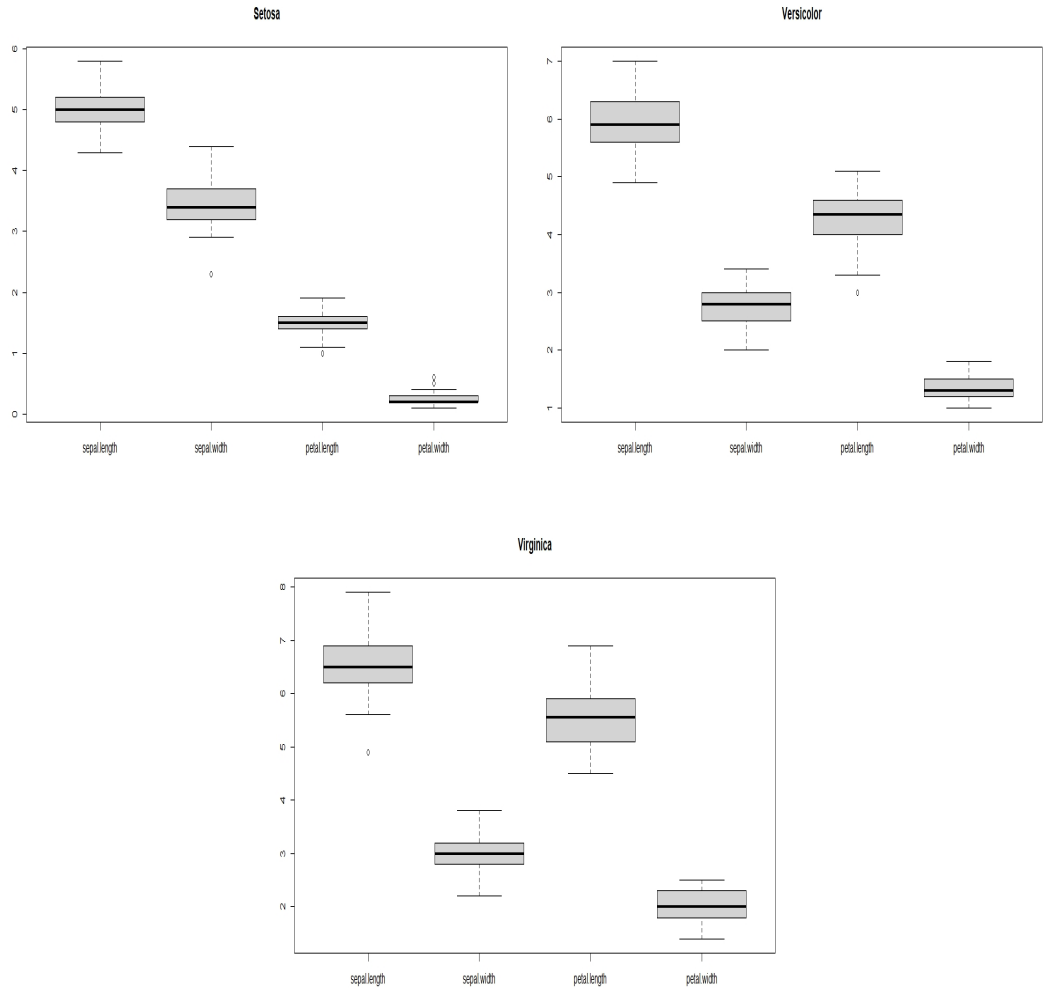

Fig 4: Lengths and widths for Setosa, Versicolor and Virginica.

APPENDIX E: SIMULATION RESULTS IN THE TWO-SAMPLE CASE

In this case ($K = 2$) we consider the procedure of Rémillard and Scaillet (2009) as a competitor. Let recall that this approach is based on the Cramer-von-Mises statistic between the two empirical copulas and an approximate p-value is obtained through multiplier technique with 1000 replications. We adopt the name of their R package and we call it the *Twocop* procedure. Similarly our procedure will be denoted by *Kcop*.

In simulation, we fix the dimension $p = 2$ and the nominal level $\alpha = 5\%$. The following groups of scenarios were considered:

1. **A25050**: group of 6 alternatives with size $n_1 = n_2 = 50$:
   - $A2norm : C_1 = Gaus(\tau_1 = 0.2)$ and $C_2 = Gaus(\tau_2 \in \{0.1, 0.2, \cdots, 0.9\})$
   - $A2stu : C_1 = Stud(df = 17, \tau_1 = 0.2)$ and $C_2 = Stud(df = 17, \tau_2 \in \{0.1, 0.2, \cdots, 0.9\})$ where df is a degree of freedom
   - $A2gum : C_1 = Gumb(\tau_1 = 0.2)$ and $C_2 = Gumb(\tau_2 \in \{0.1, 0.2, \cdots, 0.9\})$
   - $A2fran : C_1 = Fran(\tau_1 = 0.2)$ and $C_2 = Fran(\tau_2 \in \{0.1, 0.2, \cdots, 0.9\})$
   - $A2clay : C_1 = Clay(\tau_1 = 0.2)$ and $C_2 = Clay(\tau_2 \in \{0.1, 0.2, \cdots, 0.9\})$

- $A2joe$: $C_1 = Joe(\tau_1 = 0.2)$ and $C_2 = Joe(\tau_2 \in \{0.1, 0.2, \cdots, 0.9\})$

2. **A250100** $= A25050$ with $n_1 = 50$ and $n_2 = 100$
3. **A210050** $= A25050$ with $n_1 = 100$ and $n_2 = 50$
4. **A2100100** $= A25050$ with $n_1 = 100$ and $n_2 = 100$

Recall that this methodology to evaluate the finite sample performance was proposed in Rémillard and Scaillet (2009). We follow their designs with the same sample sizes $(n_1, n_2) \in \big\{(50, 50), (50, 100), (100, 50), (100, 100)\big\}$.

We note also that when $\tau_2 = 0.2$, the scenarios are under the null hypothesis.

Figures 5-8 show that both methods (*Twocop* and *Kcop*) give very similar performance. As expected, the more the Kendall's tau is different, the more the power increases. In our simulation, the first tau is fixed and equal to 0.2. The second varies and the power is maximum when it is equal to 0.9, and minimum (close to 5%) for 0.2 (the null case).

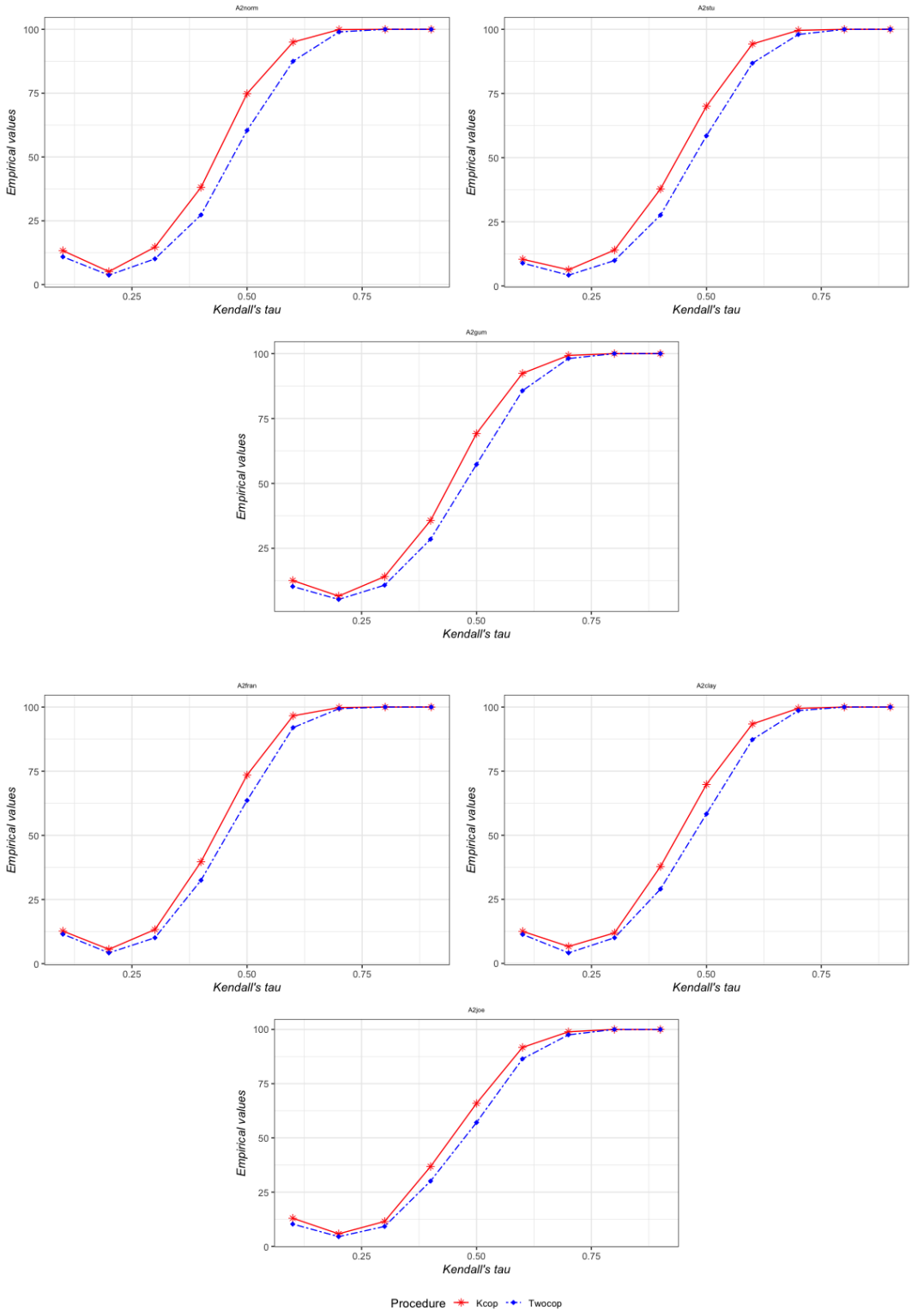

Fig 5: Two-sample case: Empirical power for **A25050**

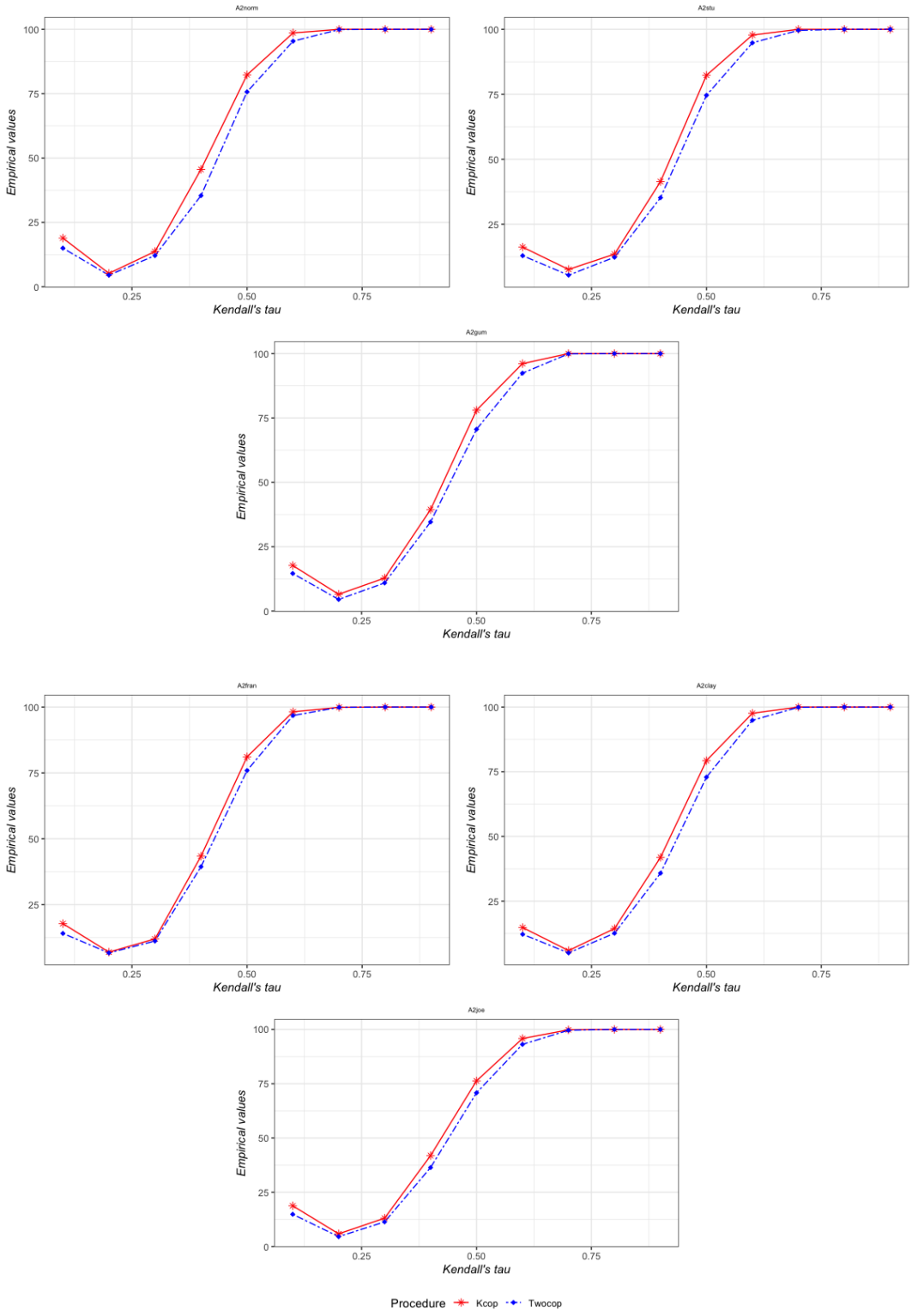

Fig 6: Two-sample case: Empirical power for **A250100**

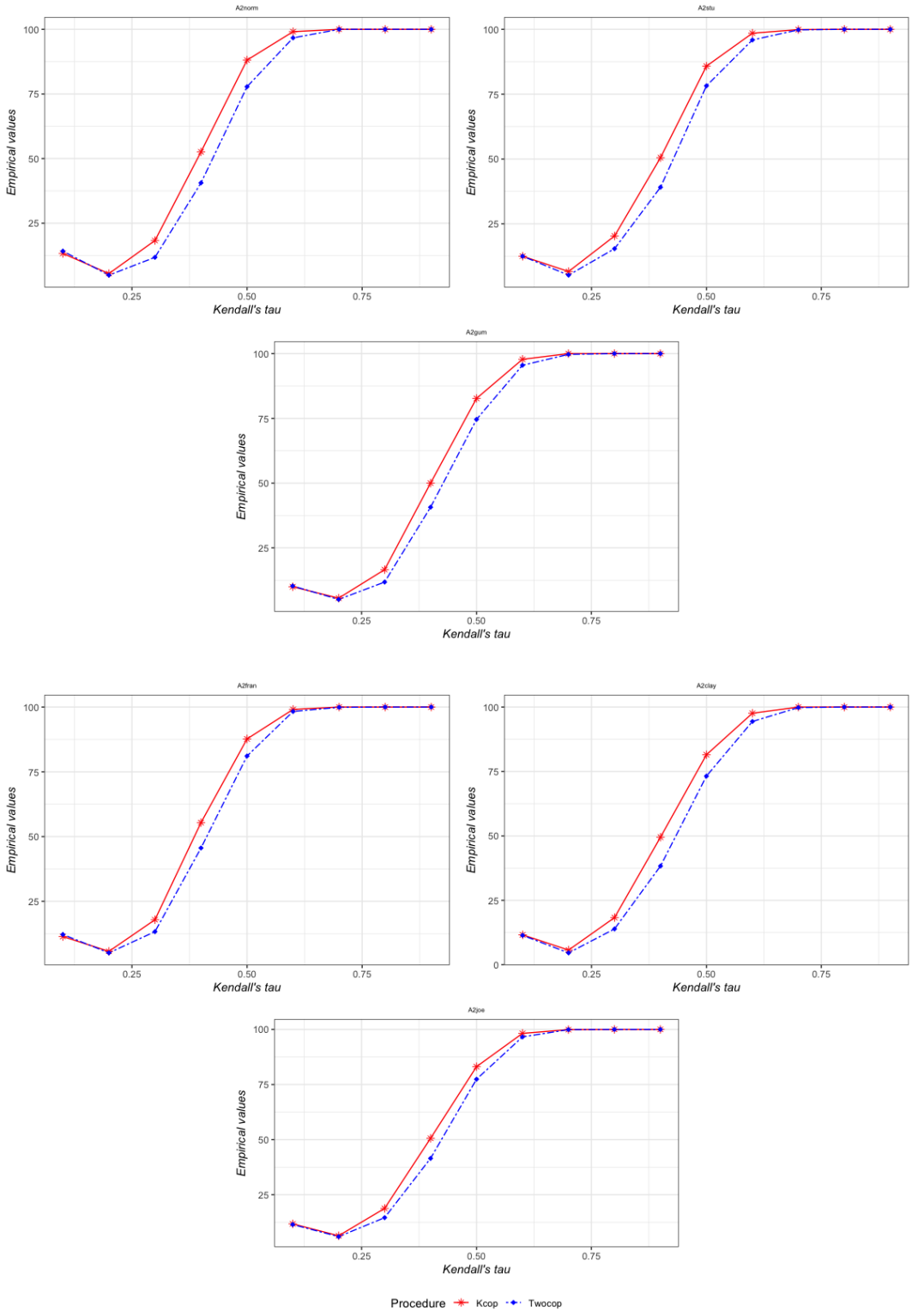

Fig 7: Two-sample case: Empirical power for **A210050**

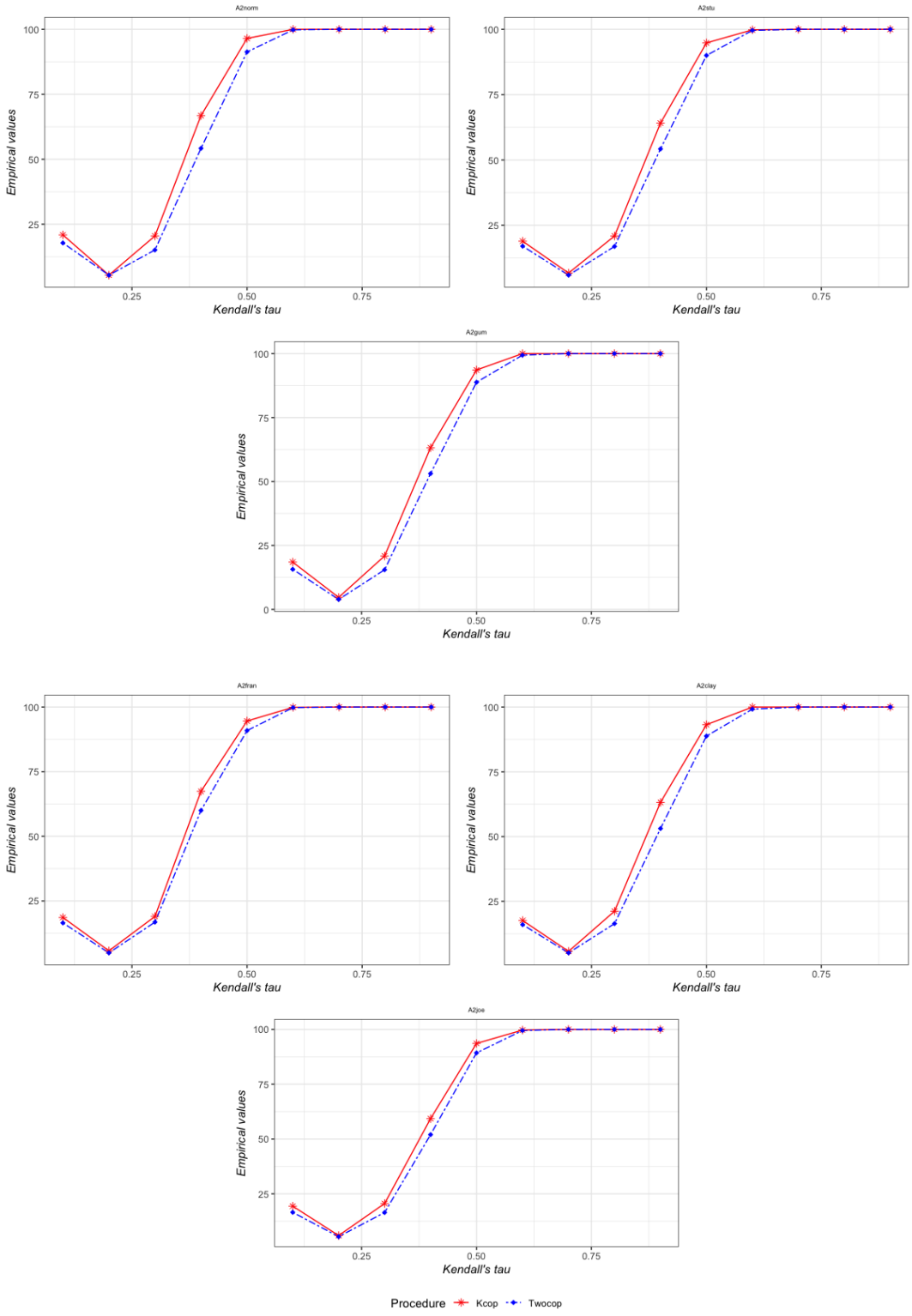

Fig 8: Two-sample case: Empirical power for **A2100100**

APPENDIX F: EMPIRICAL LEVELS FOR THE TEN-SAMPLE CASE

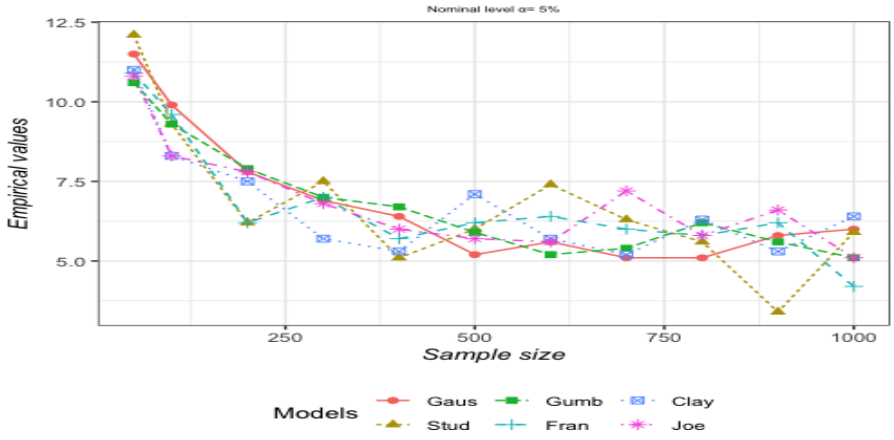

Fig 9: Ten-sample case: Empirical level for null hypotheses with $\tau = 0.1$

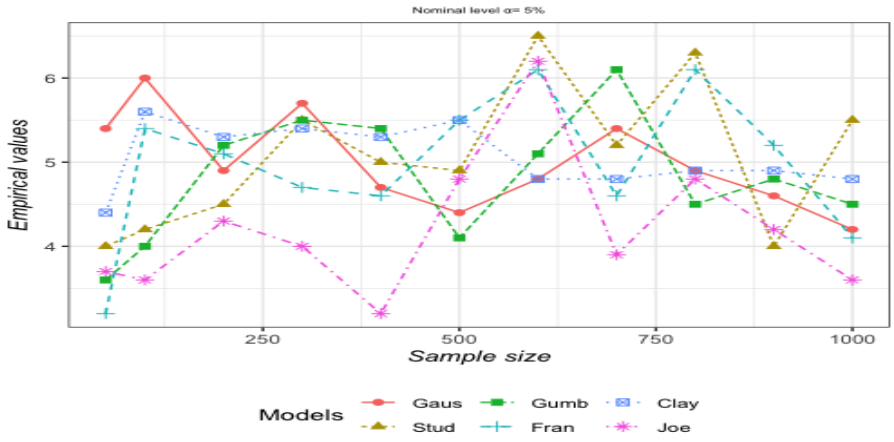

Fig 10: Ten-sample case: Empirical level for null hypotheses with $\tau = 0.5$

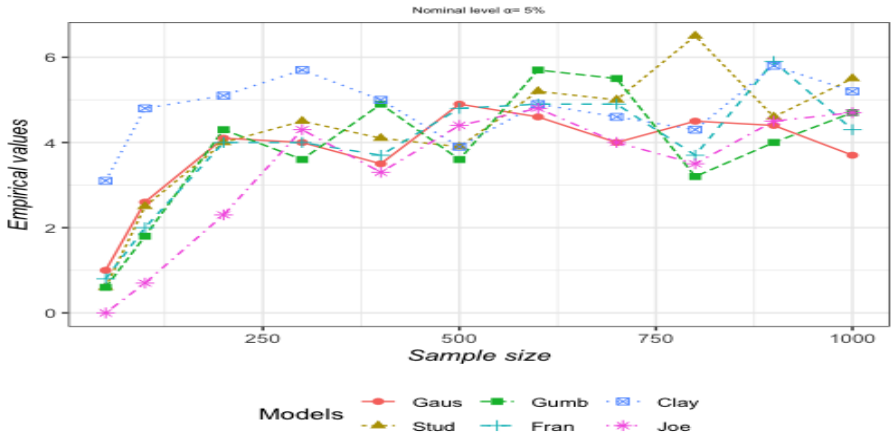

Fig 11: Ten-sample case: Empirical level for null hypotheses with $\tau = 0.8$

APPENDIX G: INSURANCE DATA: THE TWO-BY-TWO COMPARISON

TABLE 6
ANOVA test p-values (in bold the cases where the equality is not rejected

	group 1	group 2	group 3	group 4	group 5	group 6	group 7	group 8	group 9	group 10
group 1		**0.156**	**0.876**	0.035	2.88e-04	2.55e-04	3.26e-07	5.83e-06	2.49e-04	0.00
group 2			**0.217**	**0.585**	0.042	0.037	3.48e-04	2.94e-03	2.89e-02	2.67e-02
group 3				**0.059**	7.13e-04	6.29e-04	1.14e-06	1.79e-05	5.74e-04	5.05e-04
group 4					**0.111**	**0.100**	1.24e-03	9.54e-03	**0.076**	**0.071**
group 5						**0.951**	**0.089**	**0.301**	**0.789**	**0.767**
group 6							**0.103**	**0.333**	**0.836**	**0.814**
group 7								**0.502**	**0.171**	**0.179**
group 8									**0.469**	**0.485**
group 9										**0.979**
group 10										