Electronic Journal of Statistics
Vol. 18 (2024) 895–941
ISSN: 1935-7524
https://doi.org/10.1214/24-EJS2220

Smooth test for equality of copulas

Yves Ismaël Ngounou Bakam[1]

[1] *Univ Rennes, Ensai, CNRS, CREST - UMR 9194, F-35000 Rennes, France*
e-mail: yves.ngounou@ensai.fr

and

Denys Pommeret[2]

[2] *Aix-Marseille University, CNRS, Centrale Marseille, I2M, Campus de Luminy, 13288 Marseille cedex 9, France*
e-mail: denys.pommeret@univ-amu.fr

Abstract: A smooth test to simultaneously compare K copulas, where $K \geq 2$, is proposed. The K observed populations can be paired. The test statistic is based on the differences between moment sequences, called copula coefficients. These coefficients characterize the copulas, even in cases where the copula densities may not exist. The procedure involves a two-step data-driven procedure. In the initial step, the most significantly different coefficients are selected for all pairs of populations. The subsequent step utilizes these coefficients to identify populations that exhibit significant differences. To illustrate the efficacy of our method, we present numerical studies that demonstrate its performance. Furthermore, we apply our methodology, implemented in the "Kcop" R package, to two real datasets.

Keywords and phrases: Copula coefficients, data-driven smooth test, K-sample, Legendre polynomials.

Received August 2022.

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 [14] and references therein). The problem of goodness-of-fit for copulas is, therefore, an important topic and can be relevant to a wide range of situations, as in insurance to compare the dependence between portfolios (see for instance [31]), in finance to compare the dependence between indices (see for instance the book of [7]), in biology to compare the dependence between genes (see [16]), in medicine to compare diagnosis (see for instance [12]), or more recently in ecology to compare dependence between species (see [10]).

In the one-sample case, numerous testing methods have been proposed within parametric copula families (see for instance the review paper of [9], or more recently [23], [6], and [5]).

In the two-sample case, a notable reference is the nonparametric test proposed by [27], based on integrated square differences between empirical copulas.

Their test is convergent and requires the continuity of partial derivatives of copulas which allows to obtain an approximation of the distribution under the null. Their approach, adaptable to both independent and paired populations, is implemented in the "TwoCop" R package [26].

For $K > 2$, [24] introduced an innovative method to compare K copulas. [25] developed a second test statistic based on a generalized Szekely–Rizzo inequality. While these tests are consistent and can assess radial symmetry and exchangeability, they are limited to samples of the same size. More precisely both procedures consist of dividing the sample into sub-samples and testing the equality of the associated sub-copulas. Therefore, testing the equality of copulas from independent samples cannot be achieved by these works. Furthermore, in both cases the null distribution is intractable and the author needs a multiplier bootstrap method to implement these tests. Such bootstrap approach for copulas was initiated in [28]. Another extension of [27] is proposed in [4] when the K populations are observed independently, but the proposed test statistic seems to work only for testing the simultaneous independence of the K populations.

Recently, [22] conducted a study on a nonparametric copula estimator, demonstrating excellent numerical results. In this paper, we propose a novel approach to addressing the comparison of K-copulas based on such estimators. Instead of directly comparing empirical copulas, we focus on their projections onto the basis of Legendre polynomials. We restrict our study to continuous variables whose populations can be paired. This approach allows us to simultaneously compare the dependence structures of diverse populations, such as various insurance portfolios, and to compare the same population observed over multiple periods, as seen in medical cohorts. Importantly, our method is applicable not only to the paired case but also to scenarios involving several independent samples with varying sizes. This versatility is crucial for practical applications and represents a novel contribution compared to the previously mentioned works, even though the approaches of [24, 25] could potentially be extended in this direction.

Our approach is a data-driven procedure derived from Neyman's smooth tests theory (see [21]). These smooth tests serve as omnibus tests capable of detecting any departure from the null hypothesis. In our study, we consider the orthogonal projections of copula densities onto the basis of Legendre polynomials, and subsequently, we compare their coefficients. For each pair of populations, a penalized rule is introduced to select automatically the coefficients that are the most significantly different. A second penalized rule determines the number of populations to be compared. Thus, the procedure operates as a data-driven method with two selection steps. Under the null hypothesis, the penalties lead the rules to select only one pair of populations and one coefficient, resulting in a chi-square asymptotic null distribution. This simplicity distinguishes our test from the works of [24, 25], where the null distribution lacks an explicit form, requiring a multiplier bootstrap for p-value calculation. Furthermore, we demonstrate that our test procedure can detect any fixed alternative and provides insights into the rejection decision. Specifically, the second penalized rule is calibrated to identify the populations that differ most significantly. In case of rejection, we can pinpoint the pairs of populations that contributed the most to the test statistic value. Additionally, a two-by-two test can be conducted to identify similar populations. In practice, we have developed an R package "Kcop", which is accessible on the Comprehensive R Archive Network (CRAN) for implementing the K-sample procedure.

A numerical study validates the robust performance of the test. We apply this methodology to two datasets in the fields of biology and insurance. The first dataset, the well-known Iris dataset, lacks simultaneous comparison of the four-dimensional dependence structures of the three species involved. Consequently, we propose applying the smooth test to assess the dependence between sepals and petals, offering a new analysis. The second dataset is a substantial medical insurance database with possibly paired data, covering claims from three years: 1997, 1998, and 1999. We apply the smooth test to several variables from this dataset, illustrating the concepts of risk pooling and price segmentation.

All these results can be reproduced using the "Kcop" package.

The paper is organized as follows: in Section 2 we specify the null hypothesis considered in this paper and we set up the notation. Section 3 presents the 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 is devoted to the numerical study and Section 7 contains real-life illustrations. Section 8 discusses extensions and connections.

All proofs are located in Appendix A. The adaptation to the dependent case is straightforward and is summarized in Appendix B, where all results are rewritten in this context. A method for automating test parameters is available in Appendix C. Additionally, Appendices D to I contain supplementary materials, including various complements, additional simulations, and comparisons.

2. Notation and null hypotheses

Let $\mathbf{X} = (X_1, \ldots, X_p)$ be a p-dimensional continuous random vector with joint cumulative distribution function (cdf) $F_{\mathbf{X}}$, and with unique copula defined by

$$C(F_1(x_1), \ldots, F_p(x_p)) \quad = \quad F_{\mathbf{X}}(x_1, \ldots, x_p),$$

where F_j denotes the marginal cdf of X_j. Writing

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

we have for all $u_j \in [0, 1]$

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

with $\mathbf{U} = (U_1, \ldots, U_p)$. The copula density (if it exists) defined by

$$c(u_1, \ldots, u_p) \quad := \quad \frac{\partial^p C(u_1, \ldots, u_p)}{\partial u_1, \ldots, \partial u_p},$$

coincides with the probability density function (pdf) $f_{\mathbf{U}}$ 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 D for more details):

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

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

$$c(u_1,\ldots,u_p) = \sum_{j_1,\ldots,j_p \in \mathbb{N}} \rho_{j_1,\ldots,j_p} L_{j_1}(u_1)\ldots L_{j_p}(u_p), \tag{1}$$

where

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

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

$$\int_0^1 \ldots \int_0^1 c(u_1,\ldots,u_p)^2 du_1 \ldots du_p < \infty. \tag{2}$$

Write $\mathbf{j} = (j_1,\ldots,j_p)$ and $\mathbf{0} = (0,\ldots,0)$. We can observe that $\rho_{\mathbf{0}} = 1$. Moreover, since by orthogonality we have $\mathbb{E}(L_{j_i}(U_i)) = 0$, for all $i = 1,\ldots,p$, we see that $\rho_{\mathbf{j}} = 0$ if only one element of $\mathbf{j}$ is non null. When the copula density exists and is square integrable, we deduce from (1) that, for all $u_1,\ldots,u_p \in [0,1]$,

$$\begin{aligned} c(u_1,\ldots,u_p) &= 1 + \sum_{\mathbf{j}\in\mathbb{N}_*^p} \rho_{\mathbf{j}} L_{j_1}(u_1)\ldots L_{j_p}(u_p), \\ C(u_1,\ldots,u_p) &= u_1\, u_2 \ldots u_p + \sum_{\mathbf{j}\in\mathbb{N}_*^p} \rho_{\mathbf{j}} I_{j_1}(u_1)\ldots I_{j_p}(u_p), \end{aligned} \tag{3}$$

where $I_j(u) = \displaystyle\int_0^u L_j(x)dx$, and $\mathbb{N}_*^p$ stands for the set $\{\mathbf{j} = (j_1,\ldots,j_p) \in \mathbb{N}^p; \mathbf{j} \neq \mathbf{0}\}$. The sequence $(\rho_{\mathbf{j}})_{\mathbf{j}\in\mathbb{N}_*^p}$ will be referred to as the *copula coefficients* (as in [22]). Since $\mathbf{U}$ is bounded, all copula coefficients exist. The following result, due to [29] or [17], shows that such a sequence characterizes the copula. Moreover, it shows that assumption (2) is unnecessary.

Proposition 1. *Let $(\rho_{\mathbf{j}})_{\mathbf{j}\in\mathbb{N}^p}$ and $(\rho'_{\mathbf{j}})_{\mathbf{j}\in\mathbb{N}^p}$ be two sequence of copula coefficients associated to copulas C and C', respectively. Then*

$$\rho_{\mathbf{j}} = \rho'_{\mathbf{j}},\ \forall \mathbf{j} \in \mathbb{N}^p \Longleftrightarrow C = C'.$$

Thereby, the copula is determined by its sequence of copula coefficients, a property that holds even when condition (2) is not satisfied, and the copula density may not exist. Consequently, for any continuous random vectors, the comparison of their copulas coincides with the comparison of their copula coefficients. This equivalence holds true even when the random vectors lack a density or possess densities that are not square-integrable. We will use this characterization to construct the test statistic.

We consider K continuous random vectors, namely

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

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

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

The following assumption will be needed throughout the paper: we assume that for all $1 \leq \ell < m \leq K$, $\min(n_\ell, n_m) \to \infty$, and

$$n_\ell/(n_\ell + n_m) \to a_{\ell,m}, \text{ with } 0 < a_{\ell,m} < \infty. \tag{4}$$

Write $\mathbf{n} = (n_1, \ldots, n_K)$. Hence, it will cause no confusion if we write $\mathbf{n} \to +\infty$ when all $n_i \to +\infty$, and for a series of univariate random variable $(Q_n)_{n\in\mathbb{N}}$ the notation $Q_n = o_{\mathbb{P}}(\mathbf{n})$ means that $Q_n = o_{\mathbb{P}}(n_i)$, for all $i = 1, \ldots, K$.

We consider the problem of testing the equality

$$H_0: \qquad C_1 = \cdots = C_K, \tag{5}$$

against H_1: there exist $1 \leq k \neq k' \leq K$ such that $C_k \neq C_{k'}$. From Proposition 1, testing the equality (5) is equivalent to test the equality of all copula coefficients, that is

$$H_0: \rho_{\mathbf{j}}^{(1)} = \quad \cdots \quad = \rho_{\mathbf{j}}^{(K)}, \quad \forall \mathbf{j} \in \mathbb{N}_*^p, \tag{6}$$

against H_1: there exist $1 \leq k \neq k' \leq K$ and $\mathbf{j} \neq \mathbf{j}'$ such that $\rho_{\mathbf{j}}^{(k)} \neq \rho_{\mathbf{j}'}^{(k')}$, where $\rho^{(k)}$ stands for the copula coefficients associated to C_k.

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

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

For testing (6), we estimate the copula coefficients by

$$\widehat{\rho}_{j_1\ldots j_p}^{(\ell)} \quad = \quad \frac{1}{n_\ell} \sum_{i=1}^{n_\ell} L_{j_1}(\widehat{U}_{i,1}^{(\ell)}) \ldots L_{j_p}(\widehat{U}_{i,p}^{(\ell)}),$$

where $\widehat{U}_{i,j}^{(\ell)} = \widehat{F}_j^{(\ell)}(X_{i,j}^{(\ell)})$, and $\widehat{F}$ denotes the empirical distribution function associated to F. Such estimators $\widehat{\rho}_{j_1\ldots j_p}^{(\ell)}$ have been extensively studied in [22] where it is shown their excellent behavior. Considering the null hypothesis H_0 as expressed in (6), our test procedure is based on the sequences of differences

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

with the convention that $r_{\mathbf{j}}^{(\ell,m)} = 0$ when only one component of $\mathbf{j}$ is different from zero. This is due to the orthogonality of the Legendre polynomials, leading $\rho_{\mathbf{j}}^{(\ell)} = \rho_{\mathbf{j}}^{(m)} = 0$ in such cases.

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

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

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

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

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

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

This order will be used to compare successively the copula coefficients.

3. Two-sample case

We first consider the two-sample case when $K = 2$ to detail the construction of the test statistics. We want to test

$$H_0 : \rho_{\mathbf{j}}^{(1)} = \rho_{\mathbf{j}}^{(2)}, \quad \forall \mathbf{j} \in \mathbb{N}_*^p.$$

We restrict our attention to the iid case, the paired case with $n_1 = n_2$ being briefly described in Appendix B. To compare the copulas associated with $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$, we introduce a series of statistics derived from the differences between their copula coefficients. Specifically, for $1 \le k \le c(2)$, we define

$$T_{2,k}^{(1,2)} \quad := \quad \frac{n_1 n_2}{n_1 + n_2} \sum_{\mathbf{j} \in \mathcal{S}(2); ord(\mathbf{j},2) \le k} (r_{\mathbf{j}}^{(1,2)})^2, \tag{7}$$

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

$$T_{d,k}^{(1,2)} \quad := \quad T_{d-1,c(d-1)}^{(1,2)} + \frac{n_1 n_2}{n_1 + n_2} \sum_{\mathbf{j} \in \mathcal{S}(d); ord(\mathbf{j},d) \le k} (r_{\mathbf{j}}^{(1,2)})^2. \tag{8}$$

These statistics are embedded and we have for $2 \leq k < c(d)$,

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

It follows that

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

Each statistic $T_{d,k}^{(1,2)}$ contains information enabling the comparison of the copula coefficients $\rho_{\mathbf{j}}^{(1)}$ and $\rho_{\mathbf{j}}^{(2)}$ up to the norm $\|\mathbf{j}\|_1 = d$ and $ord(\mathbf{j}, d) = k$. Consequently, for a large value of d, it will be feasible to compare coefficients of high orders using $r_{\mathbf{j}}^{(1,2)}$, and the parameter k enables the exploration of all $\mathbf{j}$ values for the given order. To simplify notation, we write such a sequence of statistics as

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

By construction, for all integer $k > 0$, each statistic $V_k^{(1,2)}$ is a sum of k elements. More precisely there exists a set $\mathcal{H}(k) \subset \mathbb{N}_*^p$, with $card(\mathcal{H}(k)) = k$, such that

$$V_k^{(1,2)} = \frac{n_1 n_2}{n_1 + n_2} \sum_{\mathbf{j} \in \mathcal{H}(k)} (r_{\mathbf{j}}^{(1,2)})^2. \tag{9}$$

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 $j = 1, \ldots, c(k+1)$

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

Notice that we need to compare all copula coefficients and then let k tend to infinity to detect all potential alternatives. However, choosing a too large value for k can lead to a dilution of the test's power. Following [15], we suggest a data-driven procedure to automatically select the number of coefficients to test the hypothesis H_0. For this purpose, we set

$$D(\mathbf{n}) := \min \big\{ \operatorname*{argmax}_{1 \leq k \leq d(\mathbf{n})} (V_k^{(1,2)} - k p_{\mathbf{n}}) \big\}, \tag{10}$$

where $p_{\mathbf{n}}$ and $d(\mathbf{n})$ tend to $+\infty$, as $n_1, n_2 \to +\infty$, $k p_{\mathbf{n}}$ being a penalty term which penalizes the embedded statistics proportionally to the number of copula coefficients used. Roughly speaking, $D(\mathbf{n})$ automatically selects the coefficients that exhibit the most significant differences.

Therefore, the data-driven test statistic that we use to compare C_1 and C_2 is $V_{D(\mathbf{n})}^{(1,2)}$. We consider the following rate for penalty term:

(A) $d(n_i)^{(p+5)} = o(p_{\mathbf{n}})$, for $i = 1, 2$.

Our first result shows that under the null the least penalized statistic will be selected, specifically, the first one.

Theorem 1. *Let assumptions **(A)** and (4) hold. Then under H_0, $D(\mathbf{n})$ converges in probability towards 1 as $n_1, n_2 \to +\infty$.*

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 $V_1^{(1,2)} = T_{2,1}^{(1,2)} = \dfrac{n_1 n_2}{n_1+n_2}(r_{\mathbf{j}}^{(1,2)})^2$, with $\mathbf{j} = (1,1,0,\dots,0)$. In that case, we simply have

$$r_{\mathbf{j}}^{(1,2)} = \frac{1}{n_1}\sum_{i=1}^{n_1} L_1(\widehat{U}_{i,1}^{(1)})L_1(\widehat{U}_{i,2}^{(1)}) - \frac{1}{n_2}\sum_{i=1}^{n_2} L_1(\widehat{U}_{i,1}^{(2)})L_1(\widehat{U}_{i,2}^{(2)}).$$

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 $\mathbb{E}(L_1(U_1^{(2)})L_1(U_2^{(2)}))$. This simply means that all other copula coefficients are not significant under the null and are therefore not selected. Asymptotically, the null distribution reduces to that of $V_1^{(1,2)}$ and is given below.

Theorem 2. *Let $\boldsymbol{j} = (1,1,0\dots,0)$. If (4) holds, then under H_0,*

$$V_1^{(1,2)}/\sigma^2(1,2) \quad \xrightarrow{D} \quad \chi_1^2,$$

with $\sigma^2(1,2) = (1-a_{1,2})\sigma^2(1) + a_{1,2}\sigma^2(2)$, where $a_{1,2}$ is defined in (4), and where, for $s=1,2$,

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

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

$$\overline{M}^{(s)} = \frac{1}{n_s}\sum_{i=1}^{n_s} M_i^{(s)}, \quad \text{for} \quad s = 1,2,$$

where

$$\begin{aligned}M_i^{(s)} = L_1(\widehat{U}_{i,1}^{(s)})L_1(\widehat{U}_{i,2}^{(s)}) \quad & + \quad \frac{2\sqrt{3}}{n_s}\sum_{k=1}^{n_s}\Big(\mathbb{1}\big(X_{i,1}^{(s)} \le X_{k,1}^{(s)}\big) - \widehat{U}_{k,1}^{(s)}\Big)L_1(\widehat{U}_{k,2}^{(s)}) \\ & + \quad \frac{2\sqrt{3}}{n_s}\sum_{k=1}^{n_s}\Big(\mathbb{1}\big(X_{i,2}^{(s)} \le X_{k,2}^{(s)}\big) - \widehat{U}_{k,2}^{(s)}\Big)L_1(\widehat{U}_{k,1}^{(s)}).\end{aligned}$$

Proposition 2. *If (4) holds, then under H_0,*

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

We then deduce the limit distribution under the null.

Corollary 1. *Let assumptions **(A)** and (4) hold. Then under H_0, $V_{D(n)}^{(1,2)}/\widehat{\sigma}^2(1,2)$ converges in law towards a chi-squared distribution χ_1^2 as $n_1, n_2 \to +\infty$.*

4. *K*-sample case

In this section, we focus on the iid case, with treatment of the paired case provided in Appendix B.

Our objective is to extend the two-sample case by introducing a series of embedded statistics. Each new statistic will include a new pair of populations to be compared. We will use the first rule (10) to select a potentially different copula coefficient between each pair. A second rule will then be considered to select a possibly different pair between all populations. To select the pairs of populations we introduce the following set of indices:

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

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

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

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

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

We construct an embedded series of statistics as follows:

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

or equivalently,

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

where $D(\mathbf{n})$ is given by (10) and $V_{D(\mathbf{n})}^{(\ell,m)}$ is defined as in (9), replacing the pair index $(1,2)$ by (ℓ,m). We have $V_1 < V_2 < \ldots < V_{v(K)}$. 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 more generally, the statistic V_k compares k pairs of populations. For each $1 < k < v(K)$, there exists a unique pair (ℓ,m) such that $r_{\mathcal{V}}(\ell,m) = k$. To choose automatically the appropriate number of pairs k we introduce the following penalization procedure, mimicking the Schwarz criterion procedure [30]:

$$s(\mathbf{n}) = \min\Big\{ \underset{1\le k\le v(K)}{\operatorname{argmax}} \big(V_k - kq_{\mathbf{n}}\big)\Big\}, \tag{11}$$

where $q_{\mathbf{n}}$ is a penalty term. The choice of $q_{\mathbf{n}}$ is discussed in Remark 1. We will need the following assumption:

(A') $d(n_i)^{(p+5)} = o(q_{\mathbf{n}}),$ for $i = 1, \ldots, K$.

The following result shows that, under the null, the penalty will choose the first element of $\mathcal{V}(K)$ asymptotically. This means that all other pairs are not significantly different under the null and do not contribute to the statistic.

Theorem 3. *Let assumptions **(A)**, **(A')** and (4) hold. Then under H_0, $s(\mathbf{n})$ converges in probability towards 1 as $n_1, \ldots, n_K \to +\infty$.*

Corollary 2. *Let assumptions **(A)**, **(A')** and (4) hold. Then under H_0, $V_{s(\mathbf{n})}/\widehat{\sigma}^2(1,2)$ converges in law towards a χ_1^2 distribution as $n_1, \ldots, n_K \to +\infty$.*

Then the final data-driven test statistic is given by

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

Remark 1. In the classical smooth test approach (see [18]), the standard penalty in the univariate case is $q_n = p_n = \log(n)$, a choice closely linked to the Schwarz criteria [30] as detailed in [15]. Here, we extend this approach to the multivariate case with the following generalization:

$$q_{\boldsymbol{n}} = p_{\boldsymbol{n}} = \alpha \log\left(\frac{K^{(K-1)} n_1 \cdots n_K}{(n_1 + \cdots + n_K)^{K-1}}\right). \tag{12}$$

Proposition 5 demonstrates that this choice is sufficient for detecting alternatives. In practical applications, the introduction of the factor α serves to stabilize the empirical level, bringing it closer to the asymptotic one. Details on the automatic selection of α can be found in Appendix C, offering a straightforward calibration of the test.

It's worth noting that in [13], a comparison between this Schwarz penalty and the Akaike penalty was conducted. The latter proposes a constant value for $p_{\boldsymbol{n}}$ or $q_{\boldsymbol{n}}$, providing an alternative approach to calibrating the test.

Finally, in the paired case where $n := n_1 = \ldots = n_K$, we opt for $q_n = p_n = \alpha \log(n)$.

5. Alternative hypotheses

We consider the following series of alternative hypotheses: for $k \in \{1, \dots, v(K)\}$

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

The hypothesis $H_1(k)$ asserts that for a given k, the populations indexed by ℓ and m with $r_{\mathcal{V}}(\ell, m) = k$ are the first to exhibit a difference, as per the order defined on $\mathcal{V}(K)$). If $k = 1$, it means that the two first copulas C_1 and C_2 have at least one different copula coefficient. We will need the following assumption:

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

Proposition 3. *Let assumptions **(A)**, **(A')**, **(B)** and (4) hold. Then under $H_1(k)$, $s(\mathbf{n})$ converges in probability towards k, as $n_1, \dots, n_K \to +\infty$, and V converges to $+\infty$, that is, $\mathbb{P}(V < \epsilon) \to 0$, for all $\epsilon > 0$.*

Thus a value of $s(\mathbf{n})$ equal to k indicates that the first pairs of populations are equal and that a difference appears from the kth pair (following the order on $\mathcal{V}(K)$).

6. Numerical study of the test

We choose the penalty $q_{\mathbf{n}} = p_{\mathbf{n}} = \alpha \log(K^{(K-1)} n_1 \dots n_K / (n_1 + \dots + n_K)^{K-1})$, as indicated in Remark 1. In our proofs, we set $\alpha = 1$ for simplicity. However, in practice, we enhance this tuning factor empirically using the data-driven procedure outlined in Appendix C.

Concerning the value of $d(\mathbf{n})$, conditions **(A)** and **(A')** are asymptotic conditions and from our experience setting $d(\mathbf{n}) = 3$ or 4 is enough to have a very fast procedure which detects alternatives where copulas differ by a coefficient with a norm less than or equal to $d(\mathbf{n})$. This parameter can be modified in the package 'Kcop'. In our simulation, we fixed $d(\mathbf{n}) = 3$. The nominal level is equal to $\alpha = 5\%$.

6.1. Simulation design

We consider the following copula families: Gaussian, Student, Gumbel, Frank, Clayton, and Joe Copulas (briefly denoted by Gaus, Stud, Gumb, Fran, Clay and Joe). For the explicit forms and properties of these copulas, we refer the reader to [20]. For each copula C, the sample is generated with a given Kendall's τ parameter, and we denote it 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.

In our simulation, we compute empirical levels and empirical powers as the percentage of rejections under the null and alternative hypotheses based on 1000 replicates. We consider the following scenarios:

- We first consider the two-sample case where we compare our test procedure to that proposed in [27] which is the competitor we found for dependent as well as independent bivariate observations. Both methods give very similar results.
- Then, we consider two cases: a 5-sample case and a 10-sample case. In both situations, alternatives are constructed by modifying τ.
- We also compare the performance of the smooth test to the approach developed in [24] in the K-sample case, with $K = 2, 3, 4$, restricting our study to sub-samples from the observations as done in [24, 25].
- A 6-population case is studied where we change copulas, keeping the same τ.
- Finally an additional simulation study is proposed in Appendix H. We compared three Student copulas with $df = 5$ and with $\tau = 0.4$ or 0.6.

6.2. Simulation results in the two-sample case

In this case ($K = 2$) we consider the procedure of [27] as a competitor. Let us 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 the multiplier technique with 1000 replications. They also proposed a R package denoted by *Twocop*. By extension, we call our R package *Kcop*.

Here we fix the dimension $p = 2$. The following groups of scenarios are considered:

1. $\mathcal{A}2$: **50-50**: it includes six alternatives of size $n_1 = n_2 = 50$ which are:
   - $A2norm$: $C_1 = Gaus(\tau_1 = 0.2)$ and $C_2 = Gaus(\tau_2 \in \{0.1, 0.2, \ldots, 0.9\})$
   - $A2stu$: $C_1 = Stud(df = 17, \tau_1 = 0.2)$ and $C_2 = Stud(df = 17, \tau_2 \in \{0.1, 0.2, \ldots, 0.9\})$ where df is a degree of freedom
   - $A2gum$: $C_1 = Gumb(\tau_1 = 0.2)$ and $C_2 = Gumb(\tau_2 \in \{0.1, 0.2, \ldots, 0.9\})$
   - $A2fran$: $C_1 = Fran(\tau_1 = 0.2)$ and $C_2 = Fran(\tau_2 \in \{0.1, 0.2, \ldots, 0.9\})$
   - $A2clay$: $C_1 = Clay(\tau_1 = 0.2)$ and $C_2 = Clay(\tau_2 \in \{0.1, 0.2, \ldots, 0.9\})$
   - $A2joe$: $C_1 = Joe(\tau_1 = 0.2)$ and $C_2 = Joe(\tau_2 \in \{0.1, 0.2, \ldots, 0.9\})$
2. $\mathcal{A}2$: **50-100** = $\mathcal{A}2$: **50-50** with $n_1 = 50$ and $n_2 = 100$
3. $\mathcal{A}2$: **100-50** = $\mathcal{A}2$: **50-50** with $n_1 = 100$ and $n_2 = 50$
4. $\mathcal{A}2$: **100-100** = $\mathcal{A}2$: **50-50** with $n_1 = 100$ and $n_2 = 100$

Recall that this methodology to evaluate the finite sample performance was proposed in [27]. We follow their designs with the same sample sizes $(n_1, n_2) \in \big\{(50, 50), (50, 100), (100, 50), (100, 100)\big\}$. Such scenarios coincide with the null hypothesis when $\tau_2 = 0.2$.

The results are very similar for all scenarios and we present the *A*2*norm* alternatives in this section, reserving the remaining results for Appendix F. Figures 1–2 illustrate that both methods (`Twocop` and `Kcop`) exhibit highly comparable performance. As expected, the more different the Kendall tau, the greater the power. In our simulation, the tau associated with C_1 is fixed and equal to 0.2. The tau associated with C_2 varies and the power is maximal (100%) when it is greater than or equal to 0.7. Conversely, the power is minimal (approaching 5%) when the tau is set at 0.2, corresponding to the null hypothesis.

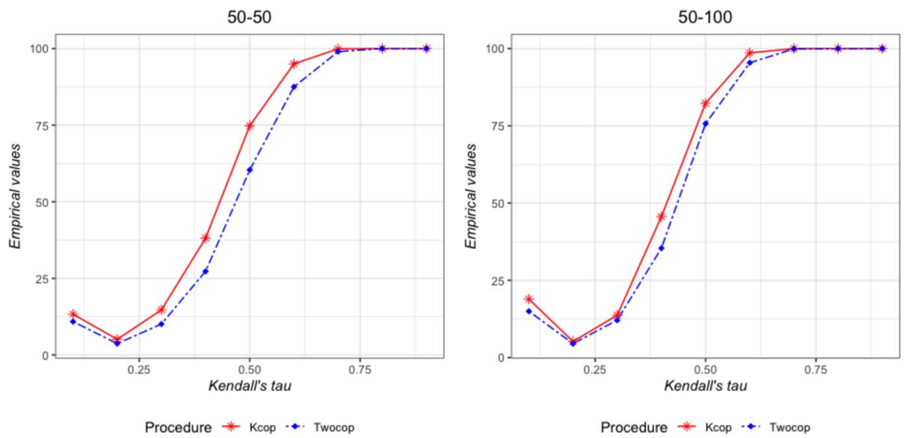

FIG 1. *Two-sample case: % of rejections under* $\mathcal{A}2$: ***50-50*** *(left) and* ***50-100*** *(right).*

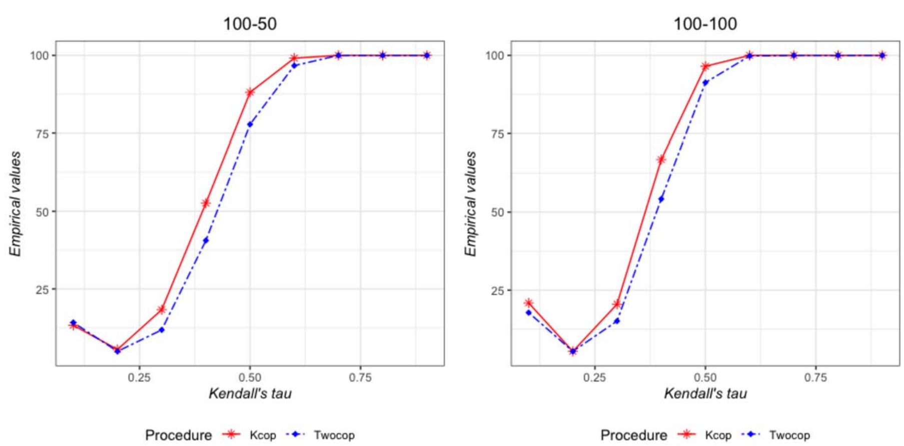

FIG 2. *Two-sample case: % of rejections under* $\mathcal{A}2$: ***100-50*** *(left) and* ***100-100*** *(right).*

6.3. Five-sample case

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

Null hypotheses: under the null hypothesis we consider the same copulas (Gaussian, Student with degree of freedom = 17, Gumbel, Frank, Clayton, Joe) with three levels of dependence: $\tau = 0.1$ (low dependence), $\tau = 0.5$ (middle dependence) and $\tau = 0.8$ (high dependence).

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

- **Alt1**: $C_1(0.3) = C_2(0.3) = C_3(0.3) = C_4(0.3)$ and $C_5(0.1)$
- **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 solely in their Kendall's tau, allowing us to highlight its effect.

Table 1 presents empirical levels (in %) with respect to sample sizes when $\tau = 0.1, 0.5$ and 0.8, respectively. In each case, one can observe that the empirical level is close to the theoretical 5% as soon as n is greater than 200. For $n = 50$ or 100, two phenomena emerge: the empirical level appears larger than the theoretical level when τ is small and smaller than the theoretical level when τ is large. Hence, with fewer observations, the procedure more readily identifies identical copulas when their dependence structure is stronger. This leads to the following recommendations: for a small size ($n < 200$) if the estimation of τ is close to 0.1, it is advisable to adopt a more conservative approach (choosing a larger theoretical level, e.g., around 0.09). Conversely, if the estimation of τ is close to 0.9, it is preferable to be anticonservative (choosing a lower theoretical level around 0.02). This implies a slight reduction in power in the first case, while power increases in the second case. A tuning procedure could be considered, incorporating a data-driven criterion based on the estimation of τ.

Concerning the empirical power, Tables 2–4 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 for a sample size equal to 1000, the program runs very fast. It can be seen for alternatives **Alt2** and **Alt3** that the empirical powers are extremely high even for small sample sizes. The first series of alternatives yields lower empirical powers since only one copula differs with a slight change in τ.

6.4. Ten-sample case

Analogously to the previous 5-sample case, we consider null hypotheses with Gaussian, Student, Gumbel, Frank, Clayton, and Joe copulas. We fixed $p = 2$.

Table 1
Empirical levels (in %) for the five-sample test.

n	Gaussian	Student	Gumbel	Frank	Clayton	Joe
	Kendall tau $\tau = 0.1$					
50	11.4	10.5	10.0	11.1	10.3	11.4
100	10.0	8.4	8.1	7.6	8.1	9.1
200	7.6	8.0	6.2	6.3	5.8	7.4
300	6.9	7.3	6.6	7.5	6.5	6.3
400	6.4	5.7	7.1	4.7	5.7	7.4
500	5.1	4.8	4.9	7.0	5.9	5.5
600	5.6	6.5	4.4	5.1	5.1	6.0
700	5.0	5.3	6.1	5.5	4.4	6.5
800	5.1	6.8	4.8	5.5	5.4	6.0
900	5.6	6.2	6.3	5.7	6.5	6.8
1000	5.9	5.5	6.0	5.3	5.2	5.0
	Kendall tau $\tau = 0.5$					
50	5.4	4.0	4.0	4.1	5.0	3.2
100	6.0	3.7	5.0	5.4	5.1	2.8
200	4.9	5.0	5.6	5.5	6.0	4.8
300	5.7	3.9	4.6	4.9	5.6	4.0
400	4.7	3.9	4.9	5.1	4.6	4.6
500	4.4	3.6	3.6	5.5	4.4	4.5
600	4.8	5.0	3.2	4.2	4.7	5.5
700	5.4	5.5	5.0	6.0	5.0	4.6
800	4.9	4.6	4.6	3.7	4.5	4.4
900	4.6	5.0	4.2	6.1	4.2	4.0
1000	4.2	4.6	4.1	4.9	5.8	3.5
	Kendall tau $\tau = 0.8$					
50	1.0	0.6	0.6	0.7	3.0	0.4
100	2.6	1.9	2.2	2.9	4.5	1.4
200	4.1	3.1	3.5	3.9	5.3	3.0
300	4.0	3.3	4.5	3.4	5.4	2.1
400	3.5	3.4	4.3	4.2	5.5	3.9
500	4.9	3.9	3.4	3.8	4.0	3.6
600	4.6	3.9	4.1	4.5	5.1	4.8
700	4.0	5.4	4.0	4.4	5.8	3.7
800	4.5	4.6	5.0	5.0	4.8	4.1
900	4.4	4.1	3.8	5.1	4.8	4.2
1000	3.7	5.4	3.8	5.6	5.4	4.1

Table 2
*Empirical powers (in %) under alternative **Alt1** (five-sample case).*

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

TABLE 3
*Empirical powers (in %) under alternative **Alt2** (five-sample case).*

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

TABLE 4
*Empirical powers (in %) under alternative **Alt3** (five-sample case).*

	Alternatives					
	Gaussian	Student	Gumbel	Frank	Clayton	Joe
$n = 50$	100	100	100	100	100	100

We consider the following alternatives where only one copula differs from the others.

- **Alt4**: $C_1(0.1) = C_2(0.1) = \cdots = C_9(0.1)$ and $C_{10}(0.55)$

Empirical levels seem to tend fast to 0.5 and are relegated in Appendix I. Table 5 shows empirical powers under alternatives **Alt4**. We only treat the cases where $n = 50$ and 100, as beyond these values, all empirical powers are equal to 100%. Remarkably, even for such small sample sizes, we observe very good behavior of the test even with small sample sizes.

TABLE 5
*Percentage of rejection under alternative **Alt4** (ten-sample case).*

	Alternatives					
	Gaussian	Student	Gumbel	Frank	Clayton	Joe
$n = 50$	98.0	96.7	96.2	97.9	97.1	97.3
$n = 100$	100	100	100	100	100	100

6.5. Alternatives with the same Kendall's tau

We consider a last alternative hypothesis with $C_1, \ldots, C_6$ which are the six copulas defined in the null hypothesis models above all with the same $\tau = 0.55$ and with a dimension p up to 5 as follows

- **Alt5**: $\tau = 0.55$; $C_1 = Gauss$, $C_2 = Student$, $C_3 = Gumbel$, $C_4 = Frank$, $C_5 = Clayton$, $C_6 = Joe$, and the dimension $p \in \{2, 3, 4, 5\}$.

Empirical powers are presented in Table 6. It can be seen that the power increases with the dimension p when the sample size is less than $n = 300$: it is then easier to detect differences between the dependence structures of the vectors. When $n \geq 300$, the empirical power is stable and equal to 100% in all scenarios.

Table 6
Empirical powers (in %) under alternative ***Alt5*** *(6-sample case).*

Dimension	$p = 2$	$p = 3$	$p = 4$	$p = 5$
$n = 50$	1.2	3.1	14.8	20.0
$n = 100$	2.0	27.3	73.6	79.1
$n = 200$	19.8	89.9	99.8	100
$n = 300$	60.3	100.0	100.0	100
$n = 400$	90.9	100.0	100.0	100
$n = 500$	98.3	100.0	100.0	100

6.6. Testing the equality of all the bivariate sub-copulas of copulas

The purpose of this section is to compare the performance of our test with that obtained by [24]. We follow the same design (see Tables 3 in [24]) and we adopt the same notation. More precisely, we simulated data $\boldsymbol{U} = (U_1, \ldots, U_{2K}) \sim C$, where C is a $2K$-dimensional copula and we examine the equality of all the bivariate sub-copulas of $\boldsymbol{U}$, that is

$$C^{(U_1,U_2)} = C^{(U_3,U_4)} = \cdots = C^{(U_{2K-1},U_{2K})}.$$

We denote by $N(\theta)$ the model where $\boldsymbol{U}$ is generated by the $2K$-variate normal copula and by $T(\theta)$ the model where $\boldsymbol{U}$ is generated by the Student copula with $\nu = 3$ degrees of freedom, where the correlation matrix Σ is such that $\theta = \Sigma_{1,2} = \Sigma_{2,1}$ and $\Sigma_{i,j} = 0.2$ for all $(i,j) \neq \{(1,2);(2,1)\}$

We compare our procedure ($Kcop$) to the following quadratic functional procedures proposed in [24]:

- Cramér-von Mises (CvM) statistic,
- Two characteristic function statistics, denoted as (Cf_1, Cf_2), correspond to the weights functions of normal and double-exponential distributions, respectively
- Diagonal statistics (Dia).

We refer the reader to [24] for more detail and to code for the program.

The results are provided in Tables 7 and 8. There is no overarching conclusion that allows determining a superior method. The various statistics seem to yield fairly similar results, except in the case of $K = 4$, where the emprical powers associated with our test statistic appear to be generally superior.

7. Real datasets applications

7.1. Biology data

We analyze Fisher's well-known 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, Virginica, and Versicolor. We then have $K = 3$ populations, and the dimension is $p = 4$. The lengths and widths for the three species are represented in Appendix E. In [8] the authors show that multivariate normal distributions seem to fit the data well for

Table 7
Empirical levels for different models studied in [24] with sample size $n = 50$ and $n = 100$.

		$n=50$					$n=100$				
K	Approaches	CvM	Dia	Cf_1	Cf_2	$Kcop$	CvM	Dia	Cf_1	Cf_2	$Kcop$
2	$N(.2)$	4.7	5.2	6.2	6.1	6.0	4.7	3.9	4.7	4.7	4.0
	$T(.2)$	4.1	3.9	4.8	4.6	6.0	4.4	4.6	5.2	5.3	5.0
3	$N(.2)$	3.3	4.8	5.7	4.8	4.0	2.9	4.1	3.3	3.8	4.0
	$T(.2)$	3.0	4.4	3.9	3.7	6.0	4.0	5.1	4.3	4.5	6.0
4	$N(.2)$	3.4	4.1	5.7	4.9	5.0	2.3	3.0	3.5	3.3	6.0
	$T(.2)$	1.7	4.9	3.5	3.0	6.0	4.4	4.5	6.2	6.1	6.0

Table 8
Empirical powers for different alternatives studied in [24] with $n = 50$ and $n = 100$.

		$n=50$					$n=100$				
K	Approaches	CvM	Dia	Cf_1	Cf_2	$Kcop$	CvM	Dia	Cf_1	Cf_2	$Kcop$
2	$N(.4)$	16.8	15.5	21.3	19.8	**32.0**	22.5	21.8	28.5	26.4	**43.0**
	$T(.4)$	44.2	48.1	**48.9**	48.1	35.0	78.0	75.4	**82.2**	81.2	40.0
	$N(.6)$	51.3	48.9	62.2	58.1	**68.0**	84.1	81.7	90.1	87.7	**94.0**
	$T(.6)$	88.6	92.2	**90.8**	90.7	60.0	**99.9**	**99.9**	**99.9**	**99.9**	84.0
3	$N(.4)$	12.0	12.7	17.8	16.1	**64.0**	20.3	21.1	25.2	23.6	**74.0**
	$T(.4)$	40.7	47.7	47.3	45.7	**70.0**	76.7	76.6	**81.2**	79.4	73.0
	$N(.6)$	47.3	48.2	63.0	56.8	**92.0**	85.8	84.7	91.5	88.6	**99.0**
	$T(.6)$	90.1	**93.5**	93.2	92.4	89.0	100.0	99.8	99.9	**100.0**	97.0
4	$N(.4)$	9.8	12.0	16.1	14.1	**78.0**	19.6	20.6	27.1	25.0	**84.0**
	$T(.4)$	34.6	41.7	43.9	42.3	**84.0**	74.4	75.5	78.8	77.6	**86.0**
	$N(.6)$	43.4	45.8	57.2	53.5	**95.0**	81.5	80.3	88.9	86.4	**100.0**
	$T(.6)$	86.3	91.7	91.0	90.5	**96.0**	99.8	99.8	**100.0**	**100.0**	99.0

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:

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

Since the observations between different species are not connected, we then apply the test for independent populations. We obtain a p-value close to zero (9.93^{-07}) and a very large test statistic $V = 23.94$. We reject the equality of the dependence structure here. The selected rank $s(n)$ is equal to 3. It means that the most significant difference is obtained when considering the statistics associated with population 1 versus 2 (Setosa and Virginica) and population 1 versus 3 (Setosa and Versicolor).

In case of rejection, we can proceed to an "ANOVA" type procedure, applying a series of two-sample tests. Table 9 contains the associated p-values and we conclude with the equality of the dependence structure between Versicolor and Virginica.

Table 9
P-values for the two-sample test (Iris dataset).

	Setosa	Virginica	Versicolor
Setosa	1	8.28×10^{-6}	9.87×10^{-5}
Virginica	8.28×10^{-6}	1	0.58
Versicolor	9.87×10^{-5}	0.58	1

7.2. Insurance data

Insurance is an area in which understanding the dependence structure among multiple portfolios is crucial for pricing, especially for risk pooling or price segmentation. To illustrate, we examine the Society of Actuaries Group Medical Insurance Large Claims Database, which contains claims information for each claimant from seven insurers over the period 1997 to 1999. Each row in the database presents a summary of claims for an individual claimant in 27 fields (columns). The first five columns provide general information about the claimant, the next twelve quantify various types of medical charges and expenses, and the last ten columns summarize details related to the diagnosis. For a detailed and thorough description of the data available online, refer to [11], accessible on the web page of the Society of Actuaries. In this context, we focus on $p = 3$ 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 claimants insured by a Preferred Provider Organization plan providing exposure for members. This consideration becomes pertinent for risk pooling if the objective is to group together similar charge scenarios or for price segmentation to provide similar guarantees for the charges. We employ a procedure with three scenarios to study the dependence structure of $\mathbf{X}$ as follows:

Three-sample test, paired case. In this case, we consider the same claimants (paired situation) present over the three periods $1997 - 1999$. At the end of the data processing, we obtained three samples of size $n = 6874$ observations. We analyse the dependence structure of the charges $\mathbf{X}$ between the three years, that is, we test $H_0 : C_{\mathbf{X}}^{1997} = C_{\mathbf{X}}^{1998} = C_{\mathbf{X}}^{1999}$. The test concluded with the non-rejection of the equality of the three dependence structures, as evidenced by a p-value $= 0.788$, a test statistic of $V = 0.072$ and a selected rank equal to $s(n) = 1$. Hence, the dependence structure of paid for insured over the three years seems to be similar. It can be an argument for keeping the same distribution of risks on the different charges X_1, X_2 and X_3.

Three-sample test, independent case. Here, we narrow our focus to female claimants. The three populations consist of individuals classified by their relationship with the subscriber, which can be “Employee” ($n_E = 18144$ observations), “Spouse” ($n_S = 10969$ observations), or “Dependent” ($n_D = 10969$ observations), all for the year 1999.

Our objective is to test the equality of the dependence structure among the charges $\mathbf{X}$. In this context, we assume independence among the $K = 3$ populations. Through our testing procedure, we obtain a p-value close to zero. Consequently, we reject the null hypothesis of equal dependence structure for these charges.

Subsequently, applying an ANOVA procedure reveals that the two-by-two equalities are rejected for “Dependent” vs “Employee” and “Employee” vs “Spouse”, with a p-value close to zero in each case. The p-value for “Dependent” vs “Spouse” is close to one.

Therefore, the status of being a “Dependent” or “Spouse” implies a similar dependence structure for the charges, distinct from the status of being an “Employee”. In the context of risk pooling, differentiating charges between these two groups becomes relevant.

Ten-sample test, independent case. Here, we analyze data from the year 1999 where the relationship to the subscriber is “Employee”. We categorize the charges $\mathbf{X}$ based on age ranges of three years, creating 10 groups as follows: $G_1 = [1936,\ 1938], \ldots, G_{10} = [1963,\ 1965]$.

The null hypothesis is H_0: the dependence structures of these 10-sample groups are identical. Applying our test procedure, we obtain a p-value close to 0 and a test statistic of $V = 16.20$. Thus, we reject the null hypothesis of equal dependence structure by age at a significant level of $\alpha = 5\%$.

There is evidence to suggest that the dependence structure of $\mathbf{X}$ changes over age. We further apply an ANOVA procedure, and the results are presented in Appendix G, Table 10, where a two-by-two comparison is proposed. Notably, there are no significant differences between two successive years. Additionally, Group 6 exhibits a similar dependence structure to the other groups, except for Group 3. The disparity increases with the gap between the years, especially between the first age categories and the last ones.

Observing the age range, we identify two clusters: $\{\text{Group } 1, \ldots, \text{Group } 5\}$ and $\{\text{Group } 6, \ldots, \text{Group } 10\}$. In terms of price segmentation, this allows the formation of two groups with similar dependencies.

8. Other similar tests

Some extensions of the K-sample test to various null hypotheses have been studied in [24, 2, 25]. Following this approach we indicate how to adapt the previous test procedure to answer the following hypotheses:

$$H_0^{RS} : C^{(\mathbf{U}^{(1)},\ldots,\mathbf{U}^{(K)})} = C^{(1-\mathbf{U}^{(1)},\ldots,1-\mathbf{U}^{(K)})}$$
$$H_0^{Exc} : C^{(\mathbf{U}^{(\ell)},\mathbf{U}^{(m)})} = C^{\mathbf{U}^{(m)},\mathbf{U}^{(\ell)})}\ ,\forall \ell \neq m$$
$$H_0^{ES} : C^{(\mathbf{U}^{(1)},\ldots,\mathbf{U}^{(K)})} = C^{\mathbf{U}^{(j_1)},\ldots\mathbf{U}^{(j_K)})},\ \text{for all permutations } \mathbf{j} \text{ of } \{1,\ldots,K\}$$

Clearly, H_0^{RS} coincides with the radial symmetry, indicating that $(\mathbf{U}^{(1)},\ldots,\mathbf{U}^{(K)})$ and $(1-\mathbf{U}^{(1)},\ldots,1-\mathbf{U}^{(K)})$ have the same joint distribution. H_0^{Exc} implies pairwise exchangeable copulas, and H_0^{ES} represents exchangeable symmetry. These three hypotheses have been elegantly grouped together and tested in

[24, 25]. We can also adapt our procedure to such hypotheses naturally by considering the density representation given by (3). For instance, in the two-sample case, testing H_0^{RS} involves comparing the coefficients $\mathbb{E}\left(L_{\mathbf{j}_1}(\mathbf{U}^{(1)})L_{\mathbf{j}_2}(\mathbf{U}^{(2)})\right)$ to the coefficients $\mathbb{E}\left(L_{\mathbf{j}_1}(1-\mathbf{U}^{(1)})L_{\mathbf{j}_2}(1-\mathbf{U}^{(2)})\right)$ for all $\mathbf{j}_1, \mathbf{j}_2$ in $\mathbb{N}^p$. Asymptotically, under H_0^{RS} the test statistic coincides with the comparison of $\mathbb{E}\left(L_1(U_1^{(1)})L_1(U_1^{(2)})\right)$ to $\mathbb{E}\left(L_1(1-U_1^{(1)})L_1(1-U_1^{(2)})\right)$ and the selected test statistic is

$$\frac{1}{n}\sum_{i=1}^{n}\left(L_1(\widehat{U}_{i,1}^{(1)})L_1(\widehat{U}_{i,1}^{(2)}) - L_1(1-\widehat{U}_{i,2}^{(1)})L_1(1-\widehat{U}_{i,2}^{(2)})\right),$$

which has an asymptotic centred normal distribution under H_0^{RS} with variance similar to that studied in Proposition 2 of the paper.

Similarly, H_0^{Exc} consists in comparing $\mathbb{E}\left(L_{\mathbf{j}_1}(\mathbf{U}^{(\ell)})L_{\mathbf{j}_2}(\mathbf{U}^{(m)})\right)$ to $\mathbb{E}\left(L_{\mathbf{j}_1}(\mathbf{U}^{(m)})L_{\mathbf{j}_2}(\mathbf{U}^{(\ell)})\right)$ for all $\ell \neq m$. Under the null hypothesis, the test statistic coincides simply with the comparison of the first coefficients (the least penalized) $\mathbb{E}\left(L_1(U_1^{(\ell)})L_1(U_2^{(\ell)})\right)$ and $\mathbb{E}\left(L_1(U_1^{(m)})L_1(U_2^{(m)})\right)$, asymptotically. Then the selected statistic under the null is

$$\frac{1}{n}\sum_{i=1}^{n}\left(L_1(\widehat{U}_{i,1}^{(\ell)})L_1(\widehat{U}_{i,2}^{(\ell)}) - L_1(\widehat{U}_{i,1}^{(m)})L_1(\widehat{U}_{i,2}^{(m)})\right),$$

which has asymptotically a centered normal null distribution.

Finally, the same reasoning applies to H_0^{ES} where the test statistic is asymptotically the same as the previous one.

9. Conclusion

In this paper, we introduced characteristic sequences, referred to as copula coefficients, for testing the equality of copulas. We developed a data-driven procedure in the two-sample case, accommodating both independent and paired populations. The extension to the K-sample case involves a second data-driven method, resulting in a two-step automatic comparison method. Our approach is applicable to all continuous random vectors, even in cases where the copula density does not exist.

Our method differs from the two-sample test proposed by [27] and complements the K-sample test developed by [24, 25], enabling the comparison of separate samples. The simulation study demonstrates the effectiveness of our approach, even for more than two populations. The test is user-friendly and performs efficiently. We have limited our simulations to the case of ten samples, but larger dimensions are conceivable with this method. For future exploration, studying high dimensions within limited computation time may require dimension reduction by selecting a limited number of copula coefficients and vector components, which extends beyond the scope of this paper.

Comparing our method to existing approaches in the two-sample case, it appears as efficient as the competitor proposed by [27]. In the K-sample case with $K > 2$, numerical results suggest performance at least as good as those obtained by [24, 25]. In both cases of comparison, we used the previous models proposed by the authors. An R package of our procedure, named "Kcop," is available on CRAN.

Following the seminal work of [24] we can adapt our procedure to test radial symmetry or exchangeability with a very similar statistic. This idea is already nicely developed in [24, 2, 25] with a general approach.

Eventually, our approach can be extended in various directions. Two potential directions include:

- Copula coefficients can be used to obtain a simplified and unified expression for some measures of association. Let us recall that for any continuous d-dimensional random variable $\boldsymbol{X} = (X_1, \ldots, X_d)$ with copula C, one of the well-known popular multivariate versions of Spearman's rho $\boldsymbol{\rho}_{\boldsymbol{X}}(C)$ can be expressed as (see [20]):

$$\boldsymbol{\rho}_{\boldsymbol{X}}(C) = h_\rho(d).\left\{2^d \int_{[0,1]^d} \pi(\boldsymbol{u}) dC(\boldsymbol{u}) - 1\right\} \text{ with } \pi(\boldsymbol{u}) = \prod_{j=1}^{d} u_j$$

  and where $h_\rho(d) = \dfrac{d+1}{2^d - (d+1)}$. Then Spearman's rho coincides with the first copula coefficients, that is

$$\boldsymbol{\rho}_{\boldsymbol{X}}(C) = h_\rho(d) \sum_{\mathbf{j} \in \{0,1\}^d / \{0\}^d} \rho_{\mathbf{j}} \prod_{k=1}^{d} \left(\delta_{0,j_k} + \frac{\sqrt{3}}{3} \delta_{1,j_k}\right).$$

  For instance, for $d = 3$, we have

$$\boldsymbol{\rho}_{\boldsymbol{X}}(C) = \frac{3}{4}\rho_{110} + \frac{3}{4}\rho_{101} + \frac{3}{4}\rho_{011} + \frac{3\sqrt{3}}{8}\rho_{111},$$

  and we deduce a novel estimator of the multivariate Spearman's rho as follows:

$$\hat{\boldsymbol{\rho}}_{\boldsymbol{X}}(C) = h_\rho(d) \sum_{\mathbf{j} \in \{0,1\}^d / \{0\}^d} \widehat{\rho}_{\mathbf{j}} \prod_{k=1}^{d} \left(\delta_{0,j_k} + \frac{\sqrt{3}}{3} \delta_{1,j_k}\right).$$

  This estimator opens up possibilities for constructing tests comparing Spearman's rho. However, this requires the calculation of the asymptotic distributions of copula coefficients as proposed in [32].
- Secondly, since the copula coefficients characterize the dependence structure, we could use such coefficients for testing independence between random vectors in the same spirit as the penalized smooth tests proposed here.

Appendix A: Proofs

We detail the proof in the independent case. The dependent case with $n_1 = \cdots = n_K := n$ is similar and will be indicated briefly in Appendix B. Throughout the proofs, we used the equality $L_1(x) = \sqrt{3}(2x-1)$ and the following inequalities are satisfied by Legendre polynomials (see [1]):

$$L_j(x) \leq cj^{1/2}, \quad \forall x \in [0,1] \tag{13}$$
$$L'_j(x) \leq c'j^{5/2}, \quad \forall x \in [0,1], \tag{14}$$

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

Proof of Proposition 1

From Corollary 6.7 of [29], if μ is a Radon measure on $\mathbb{R}^p$ for which all moments are finite and if there exists $\epsilon > 0$ such that

$$\int_{\mathbb{R}^p} e^{\epsilon\|x\|}\mu(dx) < +\infty, \tag{15}$$

then μ is said *determinate*, that is: if ν is a Radon measure with the same moments then $\nu = \mu$. Since U is bounded on $[0,1]^p$, all its moments are finite and (15) is satisfied for all $\epsilon > 0$. It follows that its distribution is *determinate*. ■

Proof of Theorem 1

We want to show that $\mathbb{P}_0(D(\mathbf{n}) > 1) \to 0$ as $\mathbf{n}$ tends to infinity. We have

$$\begin{aligned}
\mathbb{P}_0\Big(D(\mathbf{n}) > 1\Big) &= \mathbb{P}_0\Big(\exists k \in \{2,\dots,d(\mathbf{n})\} : V_k^{(1,2)} - k\, p_{\mathbf{n}} \geq V_1^{(1,2)} - p_{\mathbf{n}}\Big) \\
&= \mathbb{P}_0\Big(\exists k \in \{2,\dots,d(\mathbf{n})\} : V_k^{(1,2)} - V_1^{(1,2)} \geq (k-1)p_{\mathbf{n}}\Big) \\
&= \mathbb{P}_0\Big(\exists k \in \{2,\dots,d(\mathbf{n})\} : \frac{n_1 n_2}{n_1+n_2} \sum_{\mathbf{j}\in\mathcal{H}^*(k)} (r_{\mathbf{j}}^{(1,2)})^2 \geq (k-1)p_{\mathbf{n}}\Big) \\
&\leq \mathbb{P}_0\Big(\frac{n_1 n_2}{n_1+n_2} \sum_{\mathbf{j}\in\mathcal{H}^*(d(\mathbf{n}))} (r_{\mathbf{j}}^{(1,2)})^2 \geq p_{\mathbf{n}}\Big),
\end{aligned} \tag{16}$$

with $\mathcal{H}(k)$ satisfying (9) 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 $\mathcal{H}^*(d(\mathbf{n}))$.

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

$$\big(r_{\mathbf{j}}^{(1,2)}\big)^2 = \big((\widehat{\rho}_{\mathbf{j}}^{(1)} - \rho_{\mathbf{j}}^{(1)}) - (\widehat{\rho}_{\mathbf{j}}^{(2)} - \rho_{\mathbf{j}}^{(2)})\big)^2 \tag{17}$$

$$\leq \quad 2(\widehat{\rho}_{\mathbf{j}}^{(1)} - \rho_{\mathbf{j}}^{(1)})^2 + 2(\widehat{\rho}_{\mathbf{j}}^{(2)} - \rho_{\mathbf{j}}^{(2)})^2, \tag{18}$$

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

$$\begin{aligned}\mathbb{P}_0\big(D(\mathbf{n})>1\big) &\leq \mathbb{P}_0\big(\frac{n_1 n_2}{n_1+n_2}\sum_{\mathbf{j}\in\mathcal{H}^*}(\widehat{\rho}_{\mathbf{j}}^{(1)} - \rho_{\mathbf{j}}^{(1)})^2 \geq p_{\mathbf{n}}/4\big) \\ &\quad + \mathbb{P}_0\big(\frac{n_1 n_2}{n_1+n_2}\sum_{\mathbf{j}\in\mathcal{H}^*}(\widehat{\rho}_{\mathbf{j}}^{(2)} - \rho_{\mathbf{j}}^{(2)})^2 \geq p_{\mathbf{n}}/4\big) \qquad (19)\\ &:= A + B. \qquad (20)\end{aligned}$$

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

$$\tilde{\rho}_{\mathbf{j}}^{(1)} \quad = \quad \frac{1}{n_1}\sum_{s=1}^{n_1} L_{j_1}(U_{s,1}^{(1)})\cdots L_{j_p}(U_{s,p}^{(1)}),$$

we obtain

$$\widehat{\rho}_{\mathbf{j}}^{(1)} - \rho_{\mathbf{j}}^{(1)} \quad = \quad (\widehat{\rho}_{\mathbf{j}}^{(1)} - \tilde{\rho}_{\mathbf{j}}^{(1)}) + (\tilde{\rho}_{\mathbf{j}}^{(1)} - \rho_{\mathbf{j}}^{(1)}) \quad := \quad E_{\mathbf{j}} + G_{\mathbf{j}}, \tag{21}$$

where

$$\begin{aligned}E_{\mathbf{j}} &= \frac{1}{n_1}\sum_{s=1}^{n_1}\Big(L_{j_1}(\widehat{U}_{s,1}^{(1)})\cdots L_{j_p}(\widehat{U}_{s,p}^{(1)}) - L_{j_1}(U_{s,1}^{(1)})\cdots L_{j_p}(U_{s,p}^{(1)})\Big)\\ G_{\mathbf{j}} &= \frac{1}{n_1}\sum_{s=1}^{n_1}\Big(L_{j_1}(U_{s,1}^{(1)})\cdots L_{j_p}(U_{s,p}^{(1)}) - \mathbb{E}\big(L_{j_1}(U_1^{(1)})\cdots L_{j_p}(U_p^{(1)})\big)\Big).\end{aligned}$$

Then we have

$$A \leq \mathbb{P}_0\big(\frac{n_1 n_2}{n_1+n_2}\sum_{\mathbf{j}\in\mathcal{H}^*}(E_{\mathbf{j}})^2 \geq p_{\mathbf{n}}/16\big) + \mathbb{P}_0\big(\frac{n_1 n_2}{n_1+n_2}\sum_{\mathbf{j}\in\mathcal{H}^*}(G_{\mathbf{j}})^2 \geq p_{\mathbf{n}}/16\big). \tag{22}$$

We first study the quantity involving $E_{\mathbf{j}}$ in (22). Write

$$S_\ell^{(1)} \quad = \quad \sup_x |\widehat{F}_\ell^{(1)}(x) - F_\ell^{(1)}(x)|, \quad \ell = 1,\ldots,p. \tag{23}$$

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

$$|E_{\mathbf{j}}| \quad \leq \quad \frac{1}{n_1}\sum_{s=1}^{n_1}\sum_{i=1}^{p} S_i^{(1)} \sup_x |L'_{j_i}(x)\prod_{u\neq i} L_{j_u}(x)|.$$

From (13) and (14) there exists a constant $\tilde{c} > 0$ such that

$$|E_{\mathbf{j}}| \quad \leq \quad \tilde{c}\sum_{i=1}^{p} S_i^{(1)}({j_i}^{5/2}\prod_{u\neq i}{j_u}^{1/2}). \tag{24}$$

When $\mathbf{j}$ belongs to $\mathcal{H}^* = \mathcal{H}^*(d(\mathbf{n}))$ we necessarily have $\|\mathbf{j}\|_1 \leq d(n)$. Moreover $card(\mathcal{H}^*) = d(\mathbf{n}) - 1$. It follows that

$$\begin{aligned}
&\mathbb{P}_0\big(\frac{n_1 n_2}{n_1+n_2}\sum_{\mathbf{j}\in\mathcal{H}^*}(E_{\mathbf{j}})^2 \geq p_{\mathbf{n}}/16\big) \\
&\leq \mathbb{P}_0\big(\frac{n_1 n_2}{n_1+n_2}\sum_{\mathbf{j}\in\mathcal{H}^*}\tilde{c}^2\sum_{i=1}^{p}\sum_{i'=1}^{p} S_i^{(1)} S_{i'}^{(1)} j_i^{5/2} j_{i'}^{5/2}\prod_{s\neq i} {j_s}^{1/2}\prod_{s'\neq i'} {j_{s'}}^{1/2} \geq p_{\mathbf{n}}/16\big) \\
&\leq \mathbb{P}_0\big(\tilde{c}^2\sum_{i=1}^{p}\sum_{i'=1}^{p}\frac{n_1 n_2}{n_1+n_2} S_i^{(1)} S_{i'}^{(1)}\sum_{\mathbf{j}\in\mathcal{H}^*} d(\mathbf{n})^{p+4} \geq p_{\mathbf{n}}/16\big) \\
&\leq \mathbb{P}_0\big(\tilde{c}^2\sum_{i=1}^{p}\sum_{i'=1}^{p}\frac{n_1 n_2}{n_1+n_2} S_i^{(1)} S_{i'}^{(1)} d(\mathbf{n})^{p+5} \geq p_{\mathbf{n}}/16\big) \\
&\to 0 \;\text{ as } \mathbf{n}\to\infty, \qquad (25)
\end{aligned}$$

since for all $i = 1,\dots,p$, $\sqrt{\frac{n_1 n_2}{n_1+n_2}} S_i^{(1)}$ converges in law to a Kolmogorov distribution and $d(\mathbf{n})^{p+5} = o(p_{\mathbf{n}})$ by **(A)**.

Coming back to (21), we now study the quantity involving $G_{\mathbf{j}}$. First note that $\mathbb{E}(G_{\mathbf{j}}) = 0$. Moreover, $\mathbb{V}(G_{\mathbf{j}}) = \mathbb{E}((G_{\mathbf{j}})^2) = \mathbb{V}(\prod_{i=1}^{p} L_{j_i}(U_i^{(1)}))/n_1$. Then, by Markov inequality we have

$$\mathbb{P}_0\big(\frac{n_1 n_2}{n_1+n_2}\sum_{\mathbf{j}\in\mathcal{H}^*}(G_{\mathbf{j}})^2 \geq p_{\mathbf{n}}/16\big) \quad\leq\quad \frac{n_2}{n_1+n_2}\frac{\sum_{\mathbf{j}\in\mathcal{H}^*}\mathbb{V}(\prod_{i=1}^{p} L_{j_i}(U_i^{(1)}))}{p_{\mathbf{n}}/16}$$

and from (13) there exists a constant $c > 0$ such that

$$\mathbb{V}\big(\prod_{i=1}^{p} L_{j_i}(U_i^{(1)})\big) \quad\leq\quad c^2\prod_{i=1}^{p} j_i.$$

It follows that

$$\mathbb{P}_0\big(\frac{n_1 n_2}{n_1+n_2}\sum_{\mathbf{j}\in\mathcal{H}^*}(G_{\mathbf{j}})^2 \geq p_{\mathbf{n}}/16\big) \leq \frac{n_2}{n_1+n_2}\frac{c^2 d(\mathbf{n})^p}{p_{\mathbf{n}}/16} \quad \to 0, \text{ as } \mathbf{n}\to\infty. \qquad (26)$$

We now combine (25) and (26) with (22) to conclude that $A \to 0$, as $\mathbf{n}\to\infty$.

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

Proof of Theorem 2

Let $\mathbf{j} = (1, 1, \ldots, 0, 0)$. We have $V_1^{(1,2)} = T_{2,1}^{(1,2)} = \left(\sqrt{\frac{n_1 n_2}{n_1+n_2}} r_{\mathbf{j}}^{(1,2)}\right)^2$ and we can decompose $\sqrt{\frac{n_1 n_2}{n_1+n_2}} r_{\mathbf{j}}^{(1,2)}$ under the null as follows:

$$\begin{aligned}
\sqrt{\frac{n_1 n_2}{n_1+n_2}} r_{\mathbf{j}}^{(1,2)} &= \sqrt{\frac{n_1 n_2}{n_1+n_2}} \left(\widehat{\rho}_{\mathbf{j}}^{(1)} - \widehat{\rho}_{\mathbf{j}}^{(2)}\right) \\
&= \sqrt{\frac{n_1 n_2}{n_1+n_2}} \left(\frac{1}{n_1} \sum_{i=1}^{n_1} L_1(\widehat{U}_{i,1}^{(1)}) L_1(\widehat{U}_{i,2}^{(1)}) \right. \\
&\qquad \left. - \frac{1}{n_2} \sum_{i=1}^{n_2} L_1(\widehat{U}_{i,1}^{(2)}) L_1(\widehat{U}_{i,2}^{(2)})\right) \\
&= \sqrt{\frac{n_1 n_2}{n_1+n_2}} \left(\frac{1}{n_1} (\sum_{i=1}^{n_1} L_1(\widehat{U}_{i,1}^{(1)}) L_1(\widehat{U}_{i,2}^{(1)}) - m\right) \\
&\qquad - \sqrt{\frac{n_1 n_2}{n_1+n_2}} \left(\frac{1}{n_2} (\sum_{i=1}^{n_2} L_1(\widehat{U}_{i,1}^{(2)}) L_1(\widehat{U}_{i,2}^{(2)}) - m\right) \quad (27) \\
&:= R_{\mathbf{n}}^{(1)} - R_{\mathbf{n}}^{(2)}, \quad (28)
\end{aligned}$$

where, under the null

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

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

$$\begin{aligned}
R_{\mathbf{n}}^{(1)} &= \sqrt{\frac{n_1 n_2}{n_1+n_2}} \left\{ \sum_{i=1}^{n_1} \left(L_1(F_1^{(1)}(x_{i,1}^{(1)})) L_1(F_2^{(1)}(X_{i,2}^{(1)})) d\widehat{\mathbf{F}}^{(1)}(x,y) - m\right) \right. \\
&\quad + \int\int (\widehat{F}_1^{(1)}(x) - F_1^{(1)}(x)) 2\sqrt{3} L_1(F_2^{(1)}(y)) d\mathbf{F}^{(1)}(x,y) \\
&\quad + \int\int (\widehat{F}_2^{(1)}(y) - F_2^{(1)}(y)) 2\sqrt{3} L_1(F_1^{(1)}(x)) d\mathbf{F}^{(1)}(x,y) \\
&\quad + \int\int (\widehat{F}_1^{(1)}(x) - F_1^{(1)}(x)) 2\sqrt{3} L_1(F_2^{(1)}(y)) d\big(\widehat{\mathbf{F}}^{(1)} - \mathbf{F}^{(1)}\big)(x,y) \\
&\quad \left. + \int\int (\widehat{F}_2^{(1)}(y) - F_2^{(1)}(y)) 2\sqrt{3} L_1(F_1^{(1)}(x)) d\big(\widehat{\mathbf{F}}^{(1)} - \mathbf{F}^{(1)}\big(x,y\big) \right\} \\
&:= \sqrt{\frac{n_1 n_2}{n_1+n_2}} \left(A_{1,n_1}^{(1)} + A_{2,n_1}^{(1)} + A_{3,n_1}^{(1)} + B_{n_1}^{(1)} + C_{n_1}^{(1)}\right).
\end{aligned}$$

By symmetry, the second term $R_{\mathbf{n}}^{(2)}$ can be expressed as:

$$R_{\mathbf{n}}^{(2)} = \sqrt{\frac{n_1 n_2}{n_1+n_2}} \left(A_{1,n_2}^{(2)} + A_{2,n_2}^{(2)} + A_{3,n_2}^{(2)} + B_{n_2}^{(2)} + C_{n_2}^{(2)}\right)$$

and finally

$$\sqrt{\frac{n_1 n_2}{n_1+n_2}}\, r_{\mathbf{j}}^{(1,2)} = \sqrt{\frac{n_1 n_2}{n_1+n_2}}\Big(A_{1,n_1}^{(1)} + A_{2,n_1}^{(1)} + A_{3,n_1}^{(1)} - A_{1,n_2}^{(2)} - A_{2,n_2}^{(2)} - A_{3,n_2}^{(2)} + B_{n_1}^{(1)} + C_{n_1}^{(1)} - B_{n_2}^{(2)} - C_{n_2}^{(2)}\Big).$$

This expression is very similar to the expansion used in [32] (see his proof of Theorem 1) and [3] (see his equation (3.4)). We imitate their approach here. Therefore, we will show that $\sqrt{n_1}\sum_{i=1}^{3} A_{i,n_1}^{(1)}$ and $\sqrt{n_2}\sum_{i=1}^{3} A_{i,n_2}^{(2)}$ have a limiting normal distribution and the rest of the terms are all $o_{\mathbb{P}}(1)$. Using the expression of the empirical cdf we can rewrite

$$\begin{aligned} A_{1,\mathbf{n}}^{(1)} + A_{2,\mathbf{n}}^{(1)} + A_{3,\mathbf{n}}^{(1)} &= \frac{1}{n_1}\sum_{i=1}^{n_1}\Big\{L_1(F_1^{(1)}(X_{1,i}^{(1)})L_1(F_2^{(1)}(X_{2,i}^{(1)})) - m \\ &\quad + 2\sqrt{3}\int\int(\mathbb{1}(X_{1,i}^{(1)} \le x) - F_1^{(1)}(x))L_1(F_2^{(1)}(y))d\mathbf{F}^{(1)}(x,y) \\ &\quad + 2\sqrt{3}\int\int(\mathbb{1}(X_{2,i}^{(1)} \le y) - F_2^{(1)}(y))L_1(F_1^{(1)}(x))d\mathbf{F}^{(1)}(x,y)\Big\} \\ &:= \frac{1}{n_1}\sum_{i=1}^{n_1}(Z_{1,i}^{(1)} + Z_{2,i}^{(1)} + Z_{3,i}^{(1)}) := \frac{1}{n_1}\sum_{i=1}^{n_1} Z_i^{(1)}, \end{aligned}$$

where $Z_i^{(1)}$ are iid random variables. By symmetry we get

$$A_{1,\mathbf{n}}^{(2)} + A_{2,\mathbf{n}}^{(2)} + A_{3,\mathbf{n}}^{(2)} = \frac{1}{n_2}\sum_{i=1}^{n_2}(Z_{1,i}^{(2)} + Z_{2,i}^{(2)} + Z_{3,i}^{(2)}) := \frac{1}{n_2}\sum_{i=1}^{n_2} Z_i^{(2)}.$$

Clearly $\mathbb{E}(Z_{1,i}^{(1)} - Z_{1,i}^{(2)}) = 0$. Since $\mathbb{E}(\mathbb{1}(X_{1,i}^{(1)} \le x)) = F_1^{(1)}(x)$ and $\mathbb{E}(\mathbb{1}(X_{1,i}^{(2)} \le x)) = F_1^{(2)}(x)$, 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, $Z_i^{(1)}$ and $Z_i^{(2)}$ have finite variances. Applying the Central Limit Theorem to the independent iid series $Z_i^{(1)}$ and $Z_i^{(2)}$ we obtain

$$\sqrt{\frac{n_1 n_2}{n_1+n_2}}\left(\frac{1}{n_1}\sum_{i=1}^{n_1} Z_i^{(1)} + \frac{1}{n_2}\sum_{i=1}^{n_2} Z_i^{(2)}\right) \to N(0, \sigma^2(1,2)),$$

with

$$\sigma^2(1,2) = (1-a_{1,2})\mathbb{V}(Z_i^{(1)}) + a_{1,2}\mathbb{V}(Z_i^{(2)})$$

where $a_{1,2}$ is given by (4), and where, for $j = 1, 2$,

$$\begin{aligned} \mathbb{V}(Z_i^{(j)}) &= \mathbb{V}\Big(L_1(U_1^{(j)})L_1(U_2^{(j)}) \\ &\quad + 2\sqrt{3}\int\int\big(\mathbb{1}(X_1^{(j)} \le x) - F_1^{(j)}(x)\big)L_1(F_2^{(j)}(y))d\mathbf{F}^{(j)}(x,y) \end{aligned}$$

$$+2\sqrt{3}\int\int\big(\mathbb{1}(X_2^{(j)}\le y)-F_2^{(j)}(y)\big)L_1(F_1^{(j)}(x))d\mathbf{F}^{(j)}(x,y)\Bigg).$$

We now proceed to check that $B_{n_s}^{(s)}$, $C_{n_s}^{(s)}$ are $o_{\mathbb{P}}(n_s^{-1/2})$ for $s=1,2$. The asymptotic negligibility of $B_{n_s}^{(s)}, s=1,2$ and $C_{n_s}^{(s)}, s=1,2$ follows directly from those of $\boldsymbol{B_{1N}}$ and $\boldsymbol{B_{2N}}$ in [3]. The arguments are exactly similar to those of [3] (see his proof of Theorem 1) and we therefore omit the details. ■

Proof of Proposition 2

Let us define

$$\overline{W}^{(s)}=\frac{1}{n_s}\sum_{i=1}^{n_s}W_i^{(s)},\quad \text{for } s=1,2,$$

where

$$\begin{aligned}W_i^{(s)} &= L_1(U_{i,1}^{(s)})L_1(U_{i,2}^{(s)})\\ &\quad +2\sqrt{3}\int\int\big(\mathbb{1}(X_{i,1}^{(s)}\le x)-F_1^{(s)}(x)\big)L_1(F_2^{(s)}(y))dF^{(1)}(x,y)\\ &\quad +2\sqrt{3}\int\int\big(\mathbb{1}(X_{i,2}^{(s)}\le y)-F_2^{(s)}(y)\big)L_1(F_1^{(s)}(x))dF^{(1)}(x,y).\end{aligned}$$

We focus on $s=1$, the case $s=2$ being similar. We have

$$\frac{1}{n_1}\sum_{i=1}^{n_1}\left(W_i^{(1)}-\overline{W}^{(1)}\right)^2\xrightarrow{\mathbb{P}}\sigma^2(1). \tag{29}$$

According to Slusky's Lemma and (29), the proof is completed by showing that

$$\frac{1}{n_1}\sum_{i=1}^{n_1}\left(W_i^{(1)}-\overline{W}^{(1)}\right)^2-\widehat{\sigma}^2(1)\xrightarrow{\mathbb{P}}0.$$

We have

$$\begin{aligned}&\frac{1}{n_1}\sum_{i=1}^{n_1}\left(W_i^{(1)}-\overline{W}^{(1)}\right)^2-\widehat{\sigma}^2(1)\\ =&\ \frac{1}{n_1}\sum_{i=1}^{n}\left(W_i^{(1)}\right)^2-\left(\overline{W}^{(1)}\right)^2-\frac{1}{n_1}\sum_{i=1}^{n_1}\left(M_i^{(1)}\right)^2+\left(\overline{M}^{(1)}\right)^2\\ =&\ \frac{1}{n_1}\sum_{i=1}^{n_1}\left(W_i^{(1)}-M_i^{(1)}\right)\left(W_i^{(1)}+M_i^{(1)}-\overline{M}^{(1)}-\overline{W}^{(1)}\right).\end{aligned}$$

From (13), there exists a constant $\kappa>0$ such that, for all $n_1>0$ and for all $i=1,\ldots,n_1$,

$$\max(|W_i^{(1)}|,|M_i^{(1)}|)\le\kappa,$$

which implies that

$$\left|\frac{1}{n_1}\sum_{i=1}^{n_1}\left(W_i^{(1)}-\overline{W}^(1)\right)^2-\widehat{\sigma}^2(1)\right|\leq\frac{8\kappa}{n_1}\sum_{i=1}^{n_1}\left|W_i^{(1)}-M_i^{(1)}\right|.$$

It remains to prove that $W_i^{(1)}-M_i^{(1)}\xrightarrow{\mathbb{P}}0$. We have $W_i^{(1)}-M_i^{(1)}=I_{i,1}+2\sqrt{3}I_{i,2}+2\sqrt{3}I_{i,3}$, where

$$\begin{aligned}
I_{i,1} &= L_1(U_{i,1}^{(1)})L_1(U_{i,2}^{(1)})-L_1(\widehat{U}_{i,1}^{(1)})L_1(\widehat{U}_{i,2}^{(1)})\\
I_{i,2} &= \int\left(\mathbb{1}(X_{i,1}^{(1)}\leq x)-F_1^{(1)}(x)\right)L_1\big(F_2^{(1)}(y)\big)dF^{(1)}(x,y)\\
&\qquad -\frac{1}{n_1}\sum_{k=1}^{n_1}\left(\mathbb{1}(X_{i,1}^{(1)}\leq X_{k,1}^{(1)})-\widehat{U}_{k,1}^{(1)}\right)L_1(\widehat{U}_{k,2}^{(1)})\\
I_{i,3} &= \int\left(\mathbb{1}(X_{i,2}^{(1)}\leq x)-F_1^{(1)}(x)\right)L_1\big(F_2^{(1)}(y)\big)dF^{(1)}(x,y)\\
&\qquad -\frac{1}{n_1}\sum_{k=1}^{n_1}\left(\mathbb{1}(X_{i,2}^{(1)}\leq X_{k,2}^{(1)})-\widehat{U}_{k,2}^{(1)}\right)L_1(\widehat{U}_{k,1}^{(1)}).
\end{aligned}$$

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

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

where $S_2^{(1)}$ and $S_1^{(1)}$ are given by (23). We next show that $I_{i,2}=o_{\mathbb{P}}(1)$. We have

$$\begin{aligned}
I_{i,2} &= \frac{1}{n_1}\sum_{i=1}^{n_1}U_{i,1}^{(1)}L_1(U_{i,2}^{(1)})-\iint F_1^{(1)}(x)L_1\big(F_2^{(1)}(y)\big)dF_{1,2}^{(1)}(x,y)\\
&\quad+\iint\mathbb{1}_{X_{k,1}^{(1)}\leq x_1^{(1)}}L_1\big(F_2^{(1)}(y)\big)dF_{1,2}^{(1)}(x,y)-\frac{1}{n_1}\sum_{i=1}^{n_1}\mathbb{1}_{X_{k,1}^{(1)}\leq X_{i,1}^{(1)}}L_1(U_{i,2}^{(1)})\\
&:= I_{2,k}^1+I_{2,k}^2.
\end{aligned}$$

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

$$I_{2,k}^1 = \frac{1}{n_1}\sum_{s=1}^{n_1}U_{s,1}^{(1)}L_1(U_{s,2}^{(1)})-\mathbb{E}\left(U_1^{(1)}L_1(U_2^{(1)})\right).$$

Since the random vectors $(\mathbf{U}_i^{(1)}:=(U_{i,1}^{(1)},U_{i,2}^{(1)}))_{i=1,\ldots,n_1}$ are iid, the weak law of large numbers and the continuous mapping theorem show that

$$I_{2,k}^1=o_{\mathbb{P}}(1).$$

For $I^2_{2,k}$, we can write

$$\begin{aligned} I^2_{2,k} &= \iint \mathbb{1}_{F_1^{(1)}(X_{k,1}^{(1)})\le F_1^{(1)}(x)} L_1\big(F_2^{(1)}(y)\big) dF^{(1)}_{1,2}(x,y) \\ &\qquad - \frac{1}{n_1}\sum_{i=1}^{n_1} \mathbb{1}_{F_1^{(1)}(X_{k,1}^{(1)})\le F_1^{(1)}(X_{i,1}^{(1)})} L_1(U_{i,2}^{(1)}) \\ &= \int_0^1\int_0^1 \mathbb{1}_{U_{k,1}^{(1)}\le u} L_1(v) dC^{(1)}(u,v) - \frac{1}{n_1}\sum_{i=1}^{n_1} \mathbb{1}_{U_{k,1}^{(1)}\le U_{i,1}^{(1)}} L_1(U_{i,2}^{(1)}) \end{aligned}$$

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

$$\begin{aligned} |I^2_{2,k}| &\le \sup_{t\in[0,1]} \Bigg| \frac{1}{n_1}\sum_{i=1}^{n_1} \mathbb{1}_{t\le U_{i,1}^{(1)}} L_1(U_{i,2}^{(1)}) \\ &\qquad - \int_0^1\int_0^1 \mathbb{1}_{t\le u_1^{(1)}} L_1(u_2^{(1)}) dC^{(1)}(u_1^{(1)},u_2^{(1)}) \Bigg| \\ &\le \sup_{t\in[0,1]} \left| \frac{1}{n_1}\sum_{i=1}^{n_1} \mathbb{1}_{t\le U_{i,1}^{(1)}} L_1(U_{i,2}^{(1)}) - \mathbb{E}\left(\mathbb{1}_{t\le U_1^{(1)}} L_1(U_2^{(1)})\right)\right| \\ &\le \sup_{t\in[0,1]} \left| g\left(t, \mathbf{U}_1^{(1)},\dots,\mathbf{U}_{n_1}^{(1)}\right) - \mathbb{E}\Big(g\big(t,\mathbf{U}^{(1)}\big)\Big)\right|, \end{aligned}$$

where

$$g(t,z_1,\dots,z_{n_1}) = \frac{1}{n_1}\sum_{k=1}^{n_1} \mathbb{1}_{t\le u_k} L_1(v_k), \text{and } z_k=(u_k,v_k) \text{ for } k=1,\dots,n_1.$$

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

$$\sup_{\substack{z_1,\dots,z_{n_1},\\ z_i'}} |g(t,z_1,\dots,z_{n_1}) - g(t,z_1,\dots,z_{i-1},z_i',z_{i+1},\dots,z_{n_1})| \le \frac{2\|L_1\|_\infty}{n_1},$$

with $\frac{2\|L_1\|_\infty}{n_1} = \frac{4\sqrt{3}}{n_1}$, 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 $4\sqrt{3}/n_1$. Then, by McDiarmid's inequality, we get $\forall \epsilon > 0$

$$\mathbb{P}\left(\forall t, \left| g\left(t,\mathbf{U}_1^{(1)},\dots,\mathbf{U}_{n_1}^{(1)}\right) - \mathbb{E}\Big(g\big(t,\mathbf{U}^{(1)}\big)\Big)\right| \ge \epsilon\right) \le 2e^{-n_1\epsilon^2/24} \underset{n_1\to\infty}{\longrightarrow} 0.$$

It implies that $I^2_{2,k} = o_{\mathbb{P}}(1)$, and we conclude that $I_{i,2} = o_{\mathbb{P}}(1)$ and similarly that $I_{i,3} = o_{\mathbb{P}}(1)$. It follows that $W_i^{(1)} - M_i^{(1)} \xrightarrow{\mathbb{P}} 0$ which completes the proof. ∎

Proof of Theorem 3

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:

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

Since the previous sum contains $(k-1)$ positive elements, there is at least one element greater than $p_{\mathbf{n}}$. It follows that

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

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

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

Comparing (30) and (16) we can see that the study is now similar in spirit to the two-sample case and we can simply mimic the proof of Theorem 1 to conclude. ■

Proof of Proposition 3

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

$$\begin{aligned}\mathbb{P}(s(\mathbf{n}) < k) &\leq \mathbb{P}\left(V_k - kp_{\mathbf{n}} \leq V_{k'} - k'p_{\mathbf{n}}\right) \\ &= 1 - \mathbb{P}\left((V_k - V_{k'}) \geq (k-k')p_{\mathbf{n}}\right) \\ &= 1 - \mathbb{P}\left(\sum_{k' < r_{\mathcal{V}}(\ell,m) \leq k} V_{D(\mathbf{n})}^{(\ell,m)} \geq (k-k')p_{\mathbf{n}}\right)\end{aligned}$$

$$= 1 - \mathbb{P}\left(\sum_{k'<r_{\mathcal{V}}(\ell,m)\leq k} \frac{n_\ell n_m}{n_\ell + n_m} \sum_{\mathbf{j}\in\mathcal{H}} (r_{\mathbf{j}}^{(\ell,m)})^2 \geq (k-k')p_{\mathbf{n}}\right)$$
$$\leq 1 - \mathbb{P}\left(\mathbb{1}_{\{r_{\mathcal{V}}(\ell,m)=k\}} \frac{n_\ell n_m}{n_\ell + n_m} \sum_{\mathbf{j}\in\mathcal{H}} (r_{\mathbf{j}}^{(\ell,m)})^2 \geq (k-k')p_{\mathbf{n}}\right). \quad (31)$$

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

$$\mathbb{P}\left(\mathbb{1}_{\{r_{\mathcal{V}}(\ell,m)=k\}} \frac{n_\ell n_m}{n_\ell + n_m} \sum_{\mathbf{j}\in\mathcal{H}} (r_{\mathbf{j}}^{(\ell,m)})^2 \geq (k-k')p_{\mathbf{n}}\right)$$
$$\geq \mathbb{P}\left(\mathbb{1}_{\{r_{\mathcal{V}}(\ell,m)=k\}} \frac{n_\ell n_m}{n_\ell + n_m} \mathbb{1}_{\mathbf{j}_0\in\mathcal{H}} (r_{\mathbf{j}_0}^{(\ell,m)})^2 \geq (k-k')p_{\mathbf{n}}\right), \quad (32)$$

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

$$r_{\mathbf{j}_0}^{(\ell,m)} = \left((\widehat{\rho}_{\mathbf{j}_0}^{(\ell)} - \rho_{\mathbf{j}_0}^{(\ell)}) - (\widehat{\rho}_{\mathbf{j}_0}^{(m)} - \rho_{\mathbf{j}_0}^{(m)})\right) + \left(\rho_{\mathbf{j}_0}^{(\ell)} - \rho_{\mathbf{j}_0}^{(m)}\right) \quad := \quad (A-B) + D. \quad (33)$$

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

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

We can reuse (24) to get:

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

for some constants $\tilde{c}$ and $\tilde{c}'$. Since $\sqrt{n_\ell} S_i^{(\ell)} = O_{\mathbb{P}}(1)$ (see for instance [19]) we have $n_\ell E_{\mathbf{j}_0}^2 = O_{\mathbb{P}}(1)$. As $G_{\mathbf{j}_0}$ is an empirical estimator we also have $n_\ell G_{\mathbf{j}_0}^2 = O_{\mathbb{P}}(1)$, which yields

$$n_\ell A^2 = O_{\mathbb{P}}(1). \quad (34)$$

We now consider the quantity D in (33). The inequality $\rho_{\mathbf{j}_0}^{(\ell)} \neq \rho_{\mathbf{j}_0}^{(m)}$ implies that

$$\frac{n_\ell n_m}{n_\ell + n_m} D^2 = O(\mathbf{n}). \quad (35)$$

Finally, under $H_1(k)$, we combine (34) and (35) with (33) to get $\frac{n_\ell n_m}{n_\ell + n_m} (r_{\mathbf{j}_0}^{(\ell,m)})^2 = O_{\mathbb{P}}(\mathbf{n})$. If we prove that $\mathbb{1}_{\mathbf{j}_0 \in \mathcal{H}(D(\mathbf{n}))} \to 1$ as n tends to infinity then (32) tends to 1, from assumption **(B)**. Mimicking the proof of

Theorem 1 we can prove that $\mathbb{P}(D(\mathbf{n}) < ord(\mathbf{j}_0, \|\mathbf{j}_0\|_1)) \to 0$ which gives the result.

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

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

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

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

For the second quantity F, we obtain

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

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

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

Under $H_1(k)$, as in the proof of Theorem 1, we can see that the probability $\mathbb{P}(D(\mathbf{n}) < k)$ tends to zero as $\mathbf{n}$ tends to infinity. It follows that

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

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

$$\lim_{n\to\infty} \mathbb{P}(V_{D(\mathbf{n})}^{(\ell,m)} < \epsilon) \quad \leq \quad \lim_{\mathbf{n}\to\infty} \mathbb{P}\left(\frac{n_\ell n_m}{n_\ell + n_m}\left(r_{\mathbf{j}_0}^{(\ell,m)}\right)^2 < \epsilon\right) \quad = \quad 0, \qquad (36)$$

since by (33) $\dfrac{n_\ell n_m}{n_\ell + n_m}(r_{\mathbf{j}_0}^{(\ell,m)})^2 = O_{\mathbb{P}}(\mathbf{n})$, and finally

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

Appendix B: Paired case

We briefly describe the adaptation in the case of dependent samples, rewriting the previous definitions and the main results. In the following, we express $n := n_1 = \cdots = n_K$.

B.1. Two-sample paired case

The constructions (7) and (8) become

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

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

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

Then (9) and (10) become

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

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

where q_n and $d(n)$ tend to $+\infty$ as $n \to +\infty$. A classical choice for q_n is $\alpha \log(n)$, where α can be simply equal to 1, or obtained by the tuning procedure described in Appendix C.

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

$$V^{(1,2)} = V_{D(n)}^{(1,2)}.$$

We consider the following rate for the penalty:

(A") $d(n)^{(p+5)} = o(p_n)$.

Theorem 4. *Let assumption **(A")** holds. Then under H_0, $D(n)$ converges in Probability towards 1 as $n \to +\infty$.*

Asymptotically, the null distribution will reduce to that of $V_1^{(1,2)} = T_{2,1}^{(1,2)} = n(r_{\mathbf{j}}^{(1,2)})^2$, with $\mathbf{j} = (1, 1, 0, \ldots, 0)$ and

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

Theorem 5. *Let $\boldsymbol{j} = (1, 1, 0 \ldots, 0)$. Under H_0,*

$$(V^{(1,2)})^{1/2} \xrightarrow{D} \mathcal{N}\left(0, \sigma^2(1,2)\right),$$

where

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

To normalize the test, we consider the following estimator

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

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

where

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

Proposition 4. *Under H_0,*

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

We then obtain the following result.

Corollary 3. *Let assumption **(A")** holds. Under H_0, $V^{(1,2)}/\widehat{\sigma}^2(1,2)$ converges in law towards a chi-squared distribution χ_1^2 as $n \to +\infty$.*

B.2. K-sample paired case

The rule (11) becomes

$$s(n) \quad = \quad \min\Big\{ \underset{1\le k\le v(K)}{\text{argmax}} \big(V_k - kp_n\big)\Big\},$$

where p_n satisfies

(A''') $d(n)^{p+5} = o(p_n)$.

In practice we choose $p_n = \alpha \log(n)$. We have

Theorem 6. *Let assumptions **(A'')** and **(A''')** hold. Then under H_0, $s(n)$ converges in probability towards 1 as $n \to +\infty$.*

Corollary 4. *Let assumptions **(A'')** and **(A''')** hold. Then under H_0, $V_{s(n)}/\widehat{\sigma}^2(1,2))$ converges in law towards a χ_1^2 distribution.*

Then, the final data-driven test statistic is given by

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

B.3. Alternative hypotheses in the paired case

We need the following assumption:

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

Proposition 5. *Let assumptions **(A'')**, **(A''')** and **(B')** hold. Then under $H_1(k)$, $s(n)$ converges in probability towards k as $n \to +\infty$, and V converges to $+\infty$, that is, $\mathbb{P}(V < \epsilon) \to 0$ for all $\epsilon > 0$.*

Appendix C: Tuning the test statistic

As evoked in Remark 1, we can choose the penalty $q_{\mathbf{n}} = p_{\mathbf{n}} = \alpha \log(K^{(K-1)} n_1 \cdots n_K/(n_1 + \cdots + n_K)^{K-1})$ by using the following data-driven procedure.

Data-driven tuning procedure:

- *Assume we observe K populations, namely $P_1, \ldots, P_K$*
- *Split randomly each population into $K' > 2$ sub-populations, say $P_{i,j}$, for $i = 1, \ldots, K$, $j = 1, \ldots, K'$.*
- *Clearly, for $i = 1, \ldots, K$, the K' sub-populations $P_{i,1}, \ldots, P_{i,K'}$ have the same copula, that is, the null hypothesis H_0 is satisfied.*
- *We can repeat N times such a procedure to get $K * N$ K' samples under the null.*
- *We then approximate numerically the value of the factor $\alpha > 0$ such that the selection rule retains the first component, that is $s(\mathbf{n}) = 1$, for all the K'-sample tests. From Theorem 3, this is the asymptotic expected value under the null.*
  More precisely, we fix

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

In our simulation, we fixed arbitrarily $K' = 3$, which seems to give a very correct empirical level. Note that the use of this factor α only slightly modified the empirical results.

Appendix D: Legendre polynomials

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

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

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

They satisfy

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

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

Appendix E: Representations of sepals and petals distributions

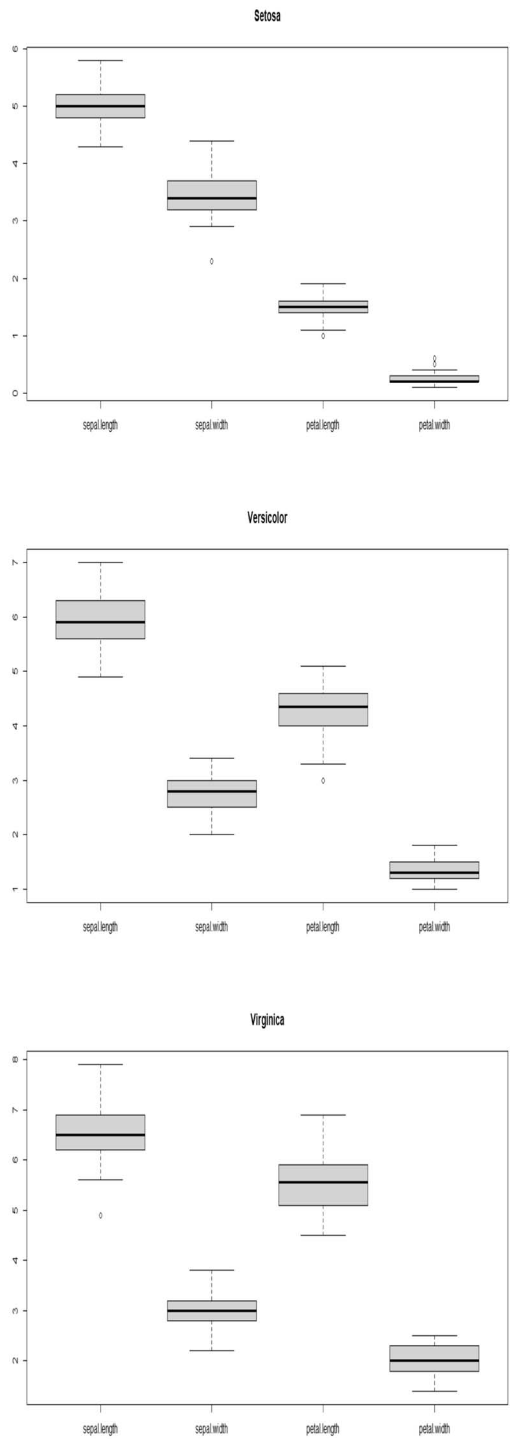

Fig 3. *Lengths and widths for Setosa, Versicolor and Virginica.*

Appendix F: Simulation results in the two-sample case (complements)

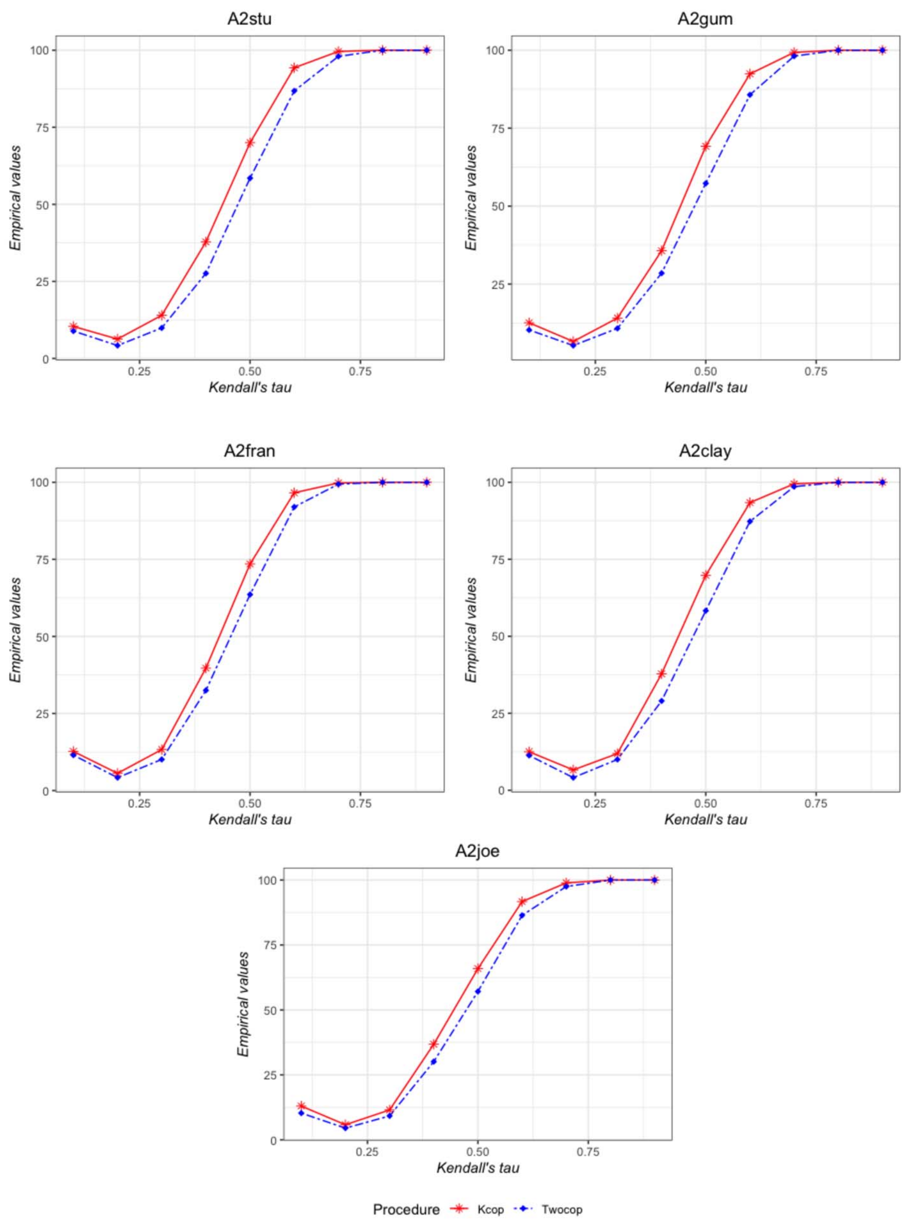

FIG 4. *Two-sample case: empirical powers under alternatives $\mathcal{A}2$:* ***50-50****.*

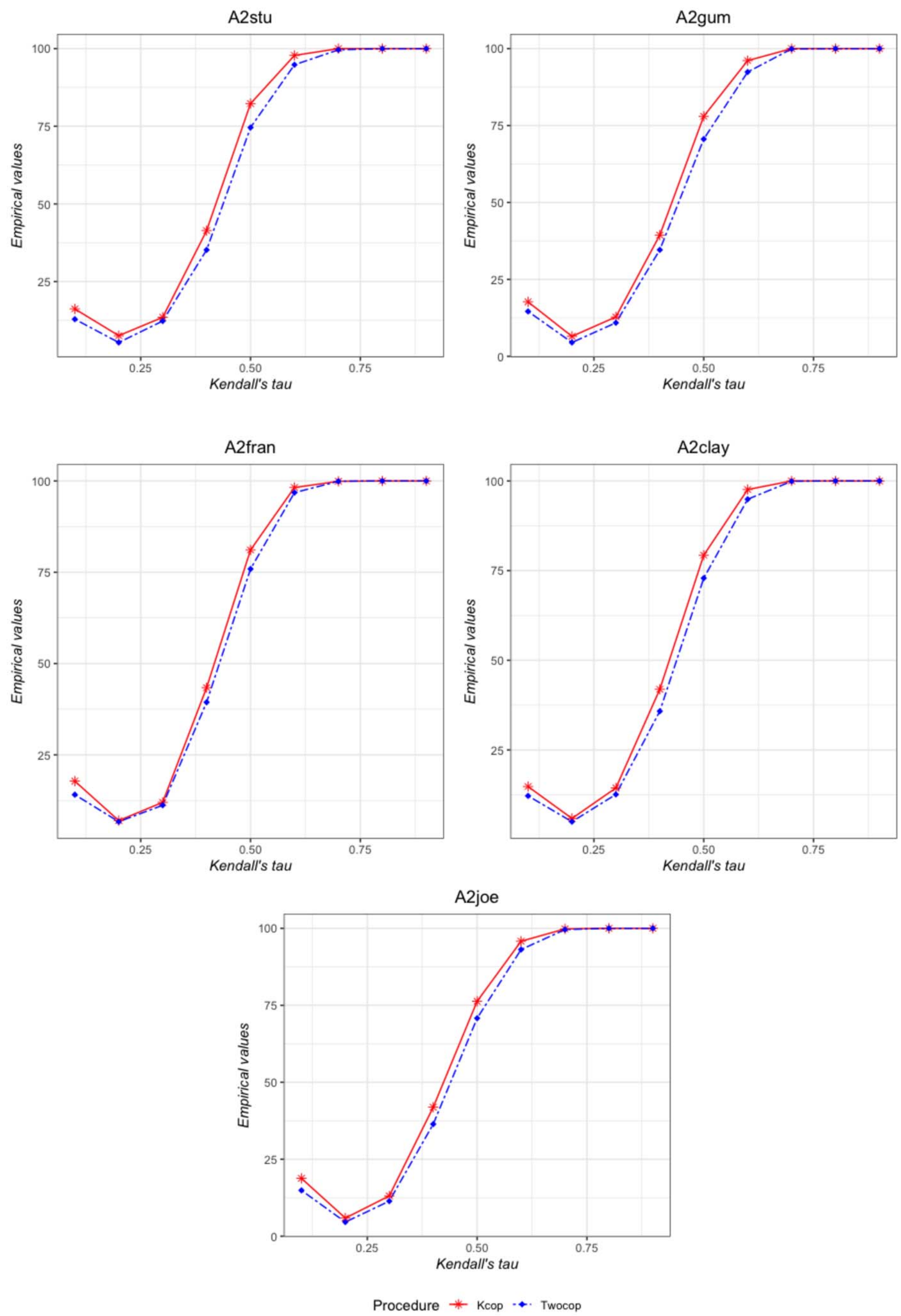

Fig 5. *Two-sample case: empirical powers under alternatives* $\mathcal{A}2$: ***50-100***.

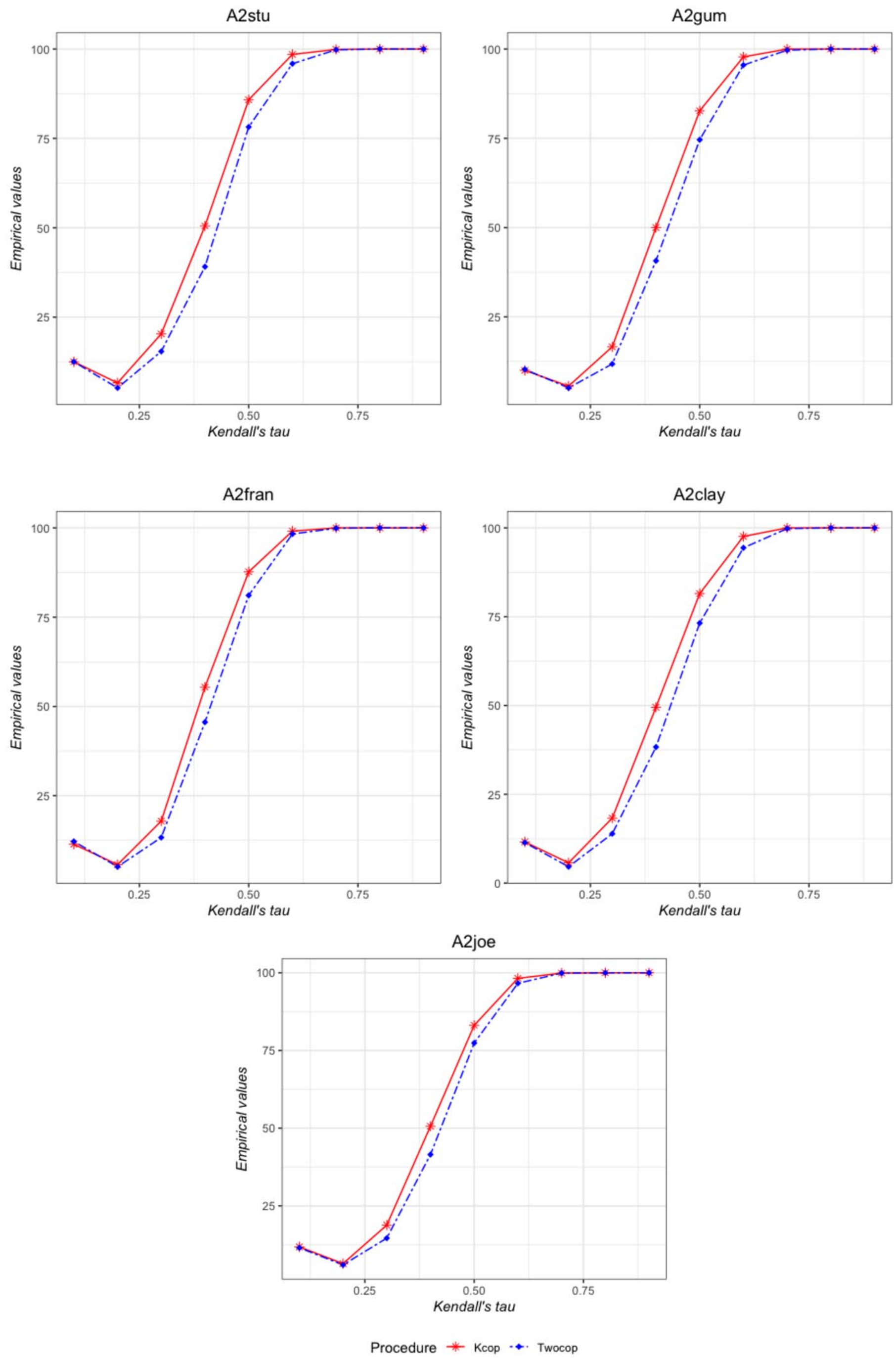

Fig 6. *Two-sample case: empirical powers under alternatives $\mathcal{A}2$: **100-50**.*

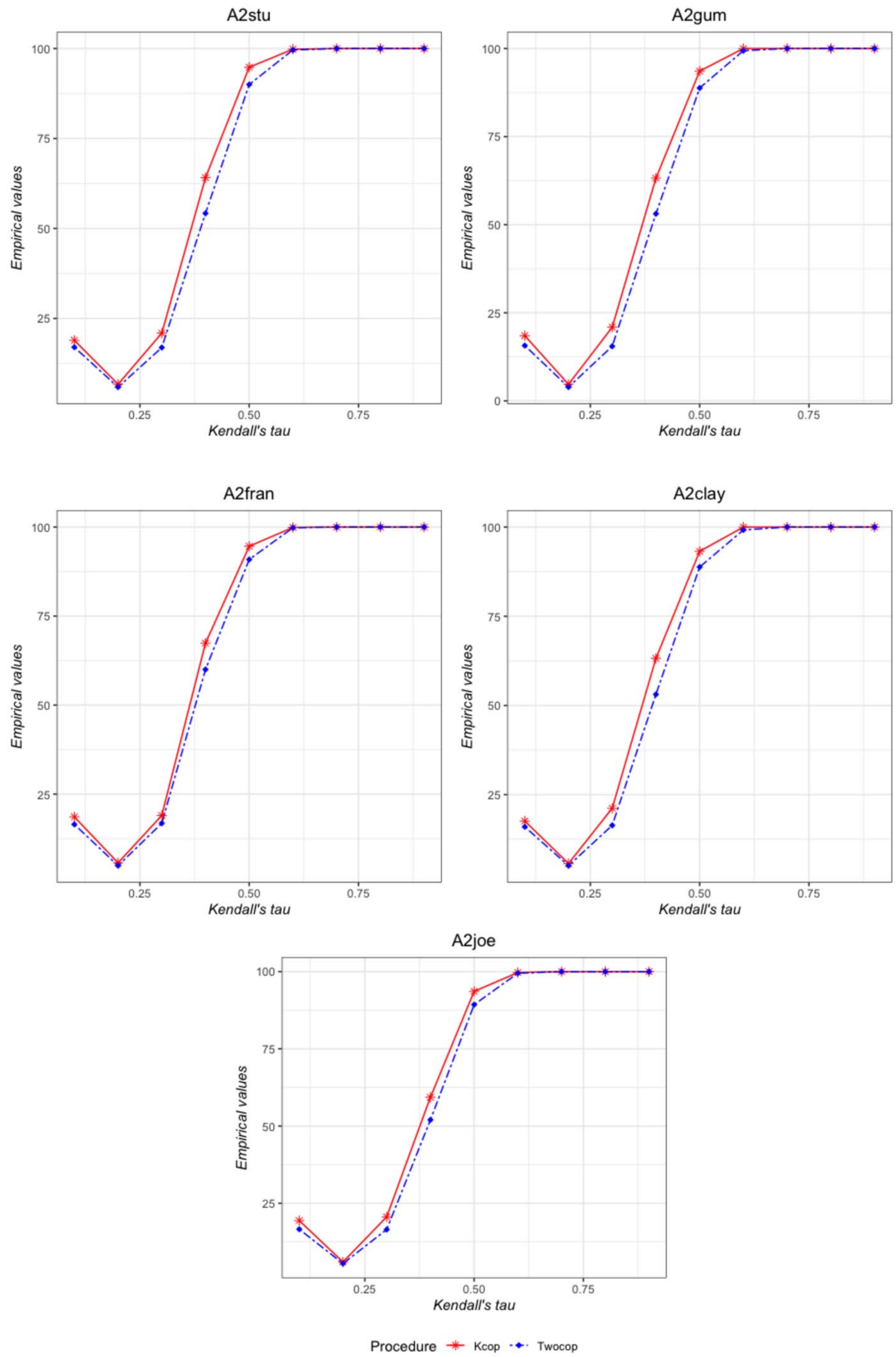

Fig 7. *Two-sample case: empirical powers under alternatives $\mathcal{A}2$:* ***100-100****.*

Appendix G: Insurance data: the two-by-two comparison

TABLE 10
ANOVA test p-values (in bold the cases where the equality is not rejected). Values given in brackets indicate the size of groups (G).

Groups	G_1	G_2	G_3	G_4	G_5	G_6	G_7	G_8	G_9
$G_1(966)$									
$G_2(971)$	**0.794**								
$G_3(996)$	**0.265**	**0.193**							
$G_4(954)$	**0.827**	**0.952**	**0.175**						
$G_5(955)$	**0.397**	**0.588**	**0.051**	**0.519**					
$G_6(915)$	**0.066**	**0.138**	0.003	**0.10**	**0.325**				
$G_7(828)$	0.002	0.009	0.000	0.005	0.028	**0.209**			
$G_8(624)$	0.001	0.005	0.000	0.002	0.017	**0.152**	**0.883**		
$G_9(524)$	0.030	**0.069**	0.001	0.046	**0.179**	**0.700**	**0.389**	**0.304**	
$G_{10}(396)$	0.008	0.020	0.000	0.013	**0.056**	**0.289**	**0.925**	**0.816**	**0.483**

Appendix H: 3-sample with Student copulas

We consider three Student copulas C_1, C_2, C_3, with df=5 and Kendall's tau τ_1, τ_2, τ_3, respectively. The first alternative is a very smooth deviation $(\tau_1, \tau_2, \tau_3) = (0.4, 0.5, 0.6)$ coinciding with three closely related populations. The second alternative is formed of only two populations but with a slightly larger difference $(\tau_1, \tau_2, \tau_3) = (0.4, 0.4, 0.6)$. Table 11 contains empirical powers for $n \in \{50, 100, 200\}$. It appears to be easier to detect the second alternative, which involves two more distinct groups, rather than three groups with a smooth variation. This finding may suggest the possibility of a forward test-based clustering procedure, wherein each population is successively tested before being joined to a cluster. This perspective could be explored further.

TABLE 11
Empirical powers.

	$n = 50$	$n = 100$	$n = 200$
$(\tau_1, \tau_2, \tau_3) = (0.4, 0.4, 0.6)$	56.4	73.2	96.6
$(\tau_1, \tau_2, \tau_3) = (0.4, 0.5, 0.6)$	17.6	29.8	39.5

Appendix I: Empirical levels for the ten-sample case

TABLE 12
Empirical levels for the ten-sample test.

	Models					
n	Gaussian	Student	Gumbel	Frank	Clayton	Joe
	Kendall tau $\tau = 0.1$					
50	11.5	12.1	10.6	10.9	11.0	10.8
100	9.9	9.3	9.3	9.6	8.3	8.3
200	7.8	6.2	7.9	6.2	7.5	7.8
300	6.9	7.5	7.0	7.0	5.7	6.8
400	6.4	5.1	6.7	5.7	5.3	6.0
500	5.2	6.0	5.9	6.2	7.1	5.7
600	5.6	7.4	5.2	6.4	5.7	5.6
700	5.1	6.3	5.4	6.0	5.2	7.2
800	5.1	5.6	6.2	5.8	6.3	5.8
900	5.8	3.4	5.6	6.2	5.3	6.6
1000	6.0	5.9	5.1	4.2	6.4	5.1
	Kendall tau $\tau = 0.5$					
50	5.4	4.0	3.6	3.2	4.4	3.7
100	6.0	4.2	4.0	5.4	5.6	3.6
200	4.9	4.5	5.2	5.1	5.3	4.3
300	5.7	5.5	5.5	4.7	5.4	4.0
400	4.7	5.0	5.4	4.6	5.3	3.2
500	4.4	4.9	4.1	5.5	5.5	4.8
600	4.8	6.5	5.1	6.1	4.8	6.2
700	5.4	5.2	6.1	4.6	4.8	3.9
800	4.9	6.3	4.5	6.1	4.9	4.8
900	4.6	4.0	4.8	5.2	4.9	4.2
1000	4.2	5.5	4.5	4.1	4.8	3.6
	Kendall tau $\tau = 0.8$					
50	1.0	0.6	0.6	0.8	3.1	0.0
100	2.6	2.5	1.8	2.0	4.8	0.7
200	4.1	4.0	4.3	4.0	5.1	2.3
300	4.0	4.5	3.6	4.0	5.7	4.3
400	3.5	4.1	4.9	3.7	5.0	3.3
500	4.9	3.9	3.6	4.8	3.9	4.4
600	4.6	5.2	5.7	4.9	4.9	4.8
700	4.0	5.0	5.5	4.9	4.6	4.0
800	4.5	6.5	3.2	3.7	4.3	3.5
900	4.4	4.6	4.0	5.9	5.8	4.5
1000	3.7	5.5	4.7	4.3	5.2	4.7

Acknowledgments

The authors would like to express their gratitude for the thorough reading, thoughtful comments, and numerous helpful suggestions provided by two anonymous referees and an Associate Editor. Their contributions greatly contributed to the improvement of this paper. The authors would like to extend special thanks to the Associate Editor for his helpful remarks, which led to Proposition 1. The second author would also like to acknowledge the support received from the Research Chair ACTIONS under the aegis of the Risk Foundation, an initiative by BNP Paribas Cardif and the Institute of Actuaries of France.

References

[1] Abramowitz, M. and Stegun, I. A. (1964). *Handbook of mathematical functions with formulas, graphs, and mathematical tables* **55**. US Government Printing Office. MR0167642

[2] Bahraoui, T. and Quessy, J.-F. (2017). Tests of radial symmetry for multivariate copulas based on the copula characteristic function. *Electronic Journal of Statistics* **11** 2066–2096. https://doi.org/10.1214/17-EJS1280 MR3652880

[3] Bhuchongkul, S. (1964). A class of nonparametric tests for independence in bivariate populations. *The Annals of Mathematical Statistics* 138–149. MR0158486

[4] 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. MR2811029

[5] 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. https://doi.org/10.3150/20-BEJ1219 MR4140541

[6] 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. https://doi.org/10.1214/14-AOS1304 MR3325713

[7] Cherubini, U., Luciano, E. and Vecchiato, W. (2004). *Copula methods in finance*. John Wiley & Sons. MR2250804

[8] Dhar, S. S., Chakraborty, B. and Chaudhuri, P. (2014). Comparison of multivariate distributions using quantile–quantile plots and related tests. *Bernoulli* **20** 1484–1506. https://doi.org/10.3150/13-BEJ530 MR3217451

[9] 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. MR2517885

[10] 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.

[11] Grazier, K. L. and G'Sell, W. (2004). *Group Medical Insurance Claims Database Collection and Analysis. Report for public release*. Society of Actuaries.

[12] 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. MR3760445

[13] 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. MR2238600

[14] Joe, H. (2014). *Dependence modeling with copulas*. CRC Press. MR3328438

[15] 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. https://doi.org/10.1214/aos/1176324315 MR1370299

[16] 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.

[17] Kleiber, C. and Stoyanov, J. (2013). Multivariate distributions and the moment problem. *Journal of Multivariate Analysis* **113** 7–18. MR2984352

[18] Ledwina, T. (1994). Data-driven version of Neyman's smooth test of fit. *Journal of the American Statistical Association* **89** 1000–1005. MR1294744

[19] Massart, P. (1990). The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The annals of Probability* 1269–1283. MR1062069

[20] Nelsen, R. B. (2007). *An introduction to copulas*. Springer Science & Business Media. MR2197664

[21] Neyman, J. (1937). "Smooth test" for goodness of fit. *Scandinavian Actuarial Journal* **1937** 149–199.

[22] Ngounou Bakam, Y. I. and Pommeret, D. (2023). Nonparametric estimation of copulas and copula densities by orthogonal projections. *Econometrics and Statistics (in press)*.

[23] 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. https://doi.org/10.1214/08-AOS666 MR2541454

[24] Quessy, J.-F. (2016). A general framework for testing homogeneity hypotheses about copulas. *Electronic Journal of Statistics* **10** 1064–1097. MR3486425

[25] Quessy, J.-F. (2021). A Szekely–Rizzo inequality for testing general copula homogeneity hypotheses. *Journal of Multivariate Analysis* **186**. MR4312922

[26] Remillard, B. and Plante, J.-F. (2012). TwoCop: Nonparametric test of equality between two copulas R package version 1.0.

[27] Rémillard, B. and Scaillet, O. (2009). Testing for equality between two copulas. *Journal of Multivariate Analysis* **100** 377-386. MR2483426

[28] Scaillet, O. (2005). A Kolmogorov-Smirnov type test for positive quad-

rant dependence. *Canadian Journal of Statistics* **33** 415–427. MR2193983

[29] Schmüdgen, K. (2020). Ten Lectures on the Moment Problem. *arXiv: 2008.12698*. https://doi.org/10.48550/ARXIV.2008.12698

[30] Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6** 461–464. MR0468014

[31] Shi, P., Feng, X. and Boucher, J.-P. (2016). Multilevel modeling of insurance claims using copulas. *The Annals of Applied Statistics* **10** 834–863. https://doi.org/10.1214/16-AOAS914 MR3528362

[32] Sinha, B. K. and Wieand, H. (1977). Multivariate nonparametric tests for independence. *Journal of Multivariate Analysis* **7** 572–583. MR0471173