K-SAMPLE INDEPENDENCE TEST (OF ARBITRARY VECTORS)

Yves I Ngounou Bakam, Denys Pommmeret

Submitted to the Annals of Statistics

K-SAMPLE INDEPENDENCE TEST (OF ARBITRARY VECTORS)

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

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

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

abstrat

1. Introduction. The problem we consider in this paper is testing the independence between continous random vectors with different dimensions More precisely, for $k = 1, \cdots, K$, let $\mathbf{X}^{(k)} = (X_1^{(k)}, \cdots, X_{p_k}^{(k)})$ be a p_k-dimensional continuous random variable with joint probability distribution function $\mathbf{F}^{(k)}$ and with univariate marginals distribution function denoted by $F_1^{(k)}, \cdots, F_{p_k}^{(k)}$. We consider the following hypotheses

$$H_0 : \mathbf{X}^{(1)} \perp\!\!\!\perp \mathbf{X}^{(2)} \perp\!\!\!\perp \cdots \perp\!\!\!\perp \mathbf{X}^{(K)} \text{ versus } H_1 : \exists l \neq k \qquad \mathbf{X}^{(l)} \not\perp\!\!\!\perp \mathbf{X}^{(k)}. \tag{1}$$

When $K = 2$, there is a large literature for testing independence between two random vectors. We can mention two seminal papers Gieser et al (1997) and Horrell et al (1974). More recently, Feng et al (2020) who used high dimensional rank statistics as Spearman and Kendall'ones. Also Berrett et al (2017) where the statistic is based on entropies, Bodnard et al (2018) resticting their study to high dimension Gaussian vectors, as in Silva et al (2021). And among others Albert et al (2020) and Yin et al (2019).

When $p_1 = \cdots = p_K = 1$, it remains to test the independence of the K components of a vector. In such a case a numerous works attempt to construct efficient statstics, sometimes based on copulas as in Genest et al (2007), Gonz´alez-Barrios et al (2019), Roy et al (2019), or Genest et al (2019). Other approaches can be mentioned as in Mao (2018), using high dimension Kendall's ta, or mor eercently in Drton et al (2020).

In the general case, when $K > 2$ and $p_j > 1$, there is very few general works. Certain restrict ther study to Gaussian vectors, as in Bao et al (2017) or Chen et al (2017), working on the covariance structure in high dimension. The work who seems to tackle the general problem is that of Fan et al (2017) where the authors represented the independence hypothesis through the product of characteristic functions. They deduce a Cramer Von Mises staistic with null asymptpotic distribution related to a process involving eigenvalues of . In addition their appraoch necessitates a weight function to calibrate the numerical integration of the Cramer-Von-Mises statistic. Also the lack of study of alternatives may suggest that such a test could not always detect departures from the null. However, these authors worked in a very general setting since they consider continuous as well as discrete or mixed random vectors, which is outside the scope of our paper, even if we suggest potential extension in the discussion at the end of the paper. In addition Fan et al (2017) proposed a package "" to use their approach. Our work is quite similar in spirit to that of Fan et al (2017) but instead of using the characteristic functions we use copulas. We will see that the resulting results are very easy to use with a chi-square asymptotic distribution under the null and a convergence of the test under alternatives. We will then compare numerically our method with that of Fan et al

MSC2020 subject classifications: Primary ???, ???; secondary ???.

Keywords and phrases: independence test, Asymptotic normality, High-Dimensional and Large-Sample, copula coefficients, data driven, Legendre polynomials.

(2017). Finally, we deduce a clustering method which yield to a data driven classification of populations by depence.

In this paper we restrict our attention to continuous random vectors. According to Sklar's theorem, there exists distribution function $\mathbf{C}^{(1)}, \cdots, \mathbf{C}^{(K)}$, and $\mathbf{C}$, with standard uniform margins, namely copulas, such that for $j = 1, \cdots, K$

$$\mathbf{F}^{(j)}(\mathbf{x}) = \mathbf{C}^{(j)}(F_1^{(j)}(x_1), \cdots, F_{p_j}^{(j)}(x_{p_j})), \tag{2}$$

$$\mathbf{F}^{(i,j)}(\mathbf{z}) = \mathbf{C}^{(i,j)}(F_1^{(i)}(z_1), \cdots, F_{p_i}^{(i)}(z_{p_i}), F_1^{(j)}(z_{p_i+1}), \cdots, F_{p_j}^{(j)}(z_{p_i+p_j})), \tag{3}$$

such that for all $\approxeq \in \mathbb{N}^{p_i}$ and $\gtrapprox \in \mathbb{N}^{p_j}$,

$$\mathbf{C}^{(i,j)}(1, \cdots, 1, \gtrapprox) = \mathbf{C}^{(j)}(\gtrapprox) \text{ and } \mathbf{C}^{(i,j)}(\approxeq, 1, \cdots, 1) = \mathbf{C}^{(i)}(\approxeq)$$

Testing problem (1) can be rewritten as follows

$$H_0 : \mathbf{C}^{(i,j)} = \mathbf{C}^{(i)}.\mathbf{C}^{(j)} \text{ versus } H_1 : \mathbf{C}^{(i,j)} \neq \mathbf{C}^{(i)}.\mathbf{C}^{(j)} \tag{4}$$

2. Two sample case . Let $\mathbf{X}$ a d

Let set $p = p_1 + p_2$. We assume that $K = 2$ and we observe two iid samples from $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}$, denoted by

$$(X_{i,1}^{(1)}, \cdots, X_{i,p}^{(1)})_{i=1,\cdots,n}, (X_{i,1}^{(2)}, \cdots, X_{i,p}^{(2)})_{i=1,\cdots,n}.$$

We want to test

$$H_0 : \mathbf{X}^{(1)} \perp\!\!\!\perp \mathbf{X}^{(2)} \text{ versus } H_1 : \mathbf{X}^{(1)} \not\perp\!\!\!\perp \mathbf{X}^{(2)}. \tag{5}$$

Let us write $\mathbf{X} := (\mathbf{X}^{(1)}, \mathbf{X}^{(2)}) = (X_1^{(1)}, \cdots, X_{p_1}^{(1)}, X_1^{(2)}, \cdots, X_{p_2}^{(2)})$ the p-dimensional random variable formed by $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$.

In the rest of paper we denote any integer vector $\mathbf{j} = (j_1, \cdots, j_{p_l+p_k}) := (\mathbf{j}(l), \mathbf{j}(k))$ where $\mathbf{j}(l) = (j_1, \cdots, j_{p_l})$ and $\mathbf{j}(k) = (j_{p_l+1}, \cdots, j_{p_k})$

2.1. *Formalism of the test.* Write $\mathcal{L} = \{L_n; n \in \mathbb{N}\}$ the set of shifted Legendre polynomials which are orthonormal with respect to uniform measure μ. From L^2 decomposition (see paper of Denys and Yves), the expression of these copulas above are given by: for all $\approxeq^{(1)} = (u_1^{(1)}, \cdots, u_{p_1}^{(1)}) \in I^{p_1}$, $\approxeq^{(2)} = (u_1^{(2)}, \cdots, u_{p_2}^{(2)}) \in I^{p_2}$ and $\approxeq = (\approxeq^{(1)}, \approxeq^{(2)}) \in I^p$

$$\mathbf{C}^{(1)}(\approxeq^{(1)}) = \sum_{\mathbf{m} \in \mathbb{N}^p} \rho_{\mathbf{m}} I_{\mathbf{m}}(\approxeq^{(1)}) \quad \text{with} \quad I_{\mathbf{m}}(\mathbf{u}^{(1)}) = I_{m_1}(u_1^{(1)}) \cdots I_{m_{p_1}}(u_{p_1}^{(1)}),$$

$$\mathbf{C}^{(2)}(\approxeq^{(2)}) = \sum_{\mathbf{k} \in \mathbb{N}^q} \rho_{\mathbf{k}} I_{\mathbf{k}}(\mathbf{u}^{(2)}) \quad \text{with} \quad I_{\mathbf{k}}(\approxeq^{(2)}) = I_{k_1}(u_1^{(2)}) \cdots I_{k_q}(u_{p_2}^{(2)}),$$

$$\mathbf{C}^{(1,2)}(\approxeq) = \sum_{\mathbf{j} \in \mathbb{N}^{p_1+p_2}} \rho_{\mathbf{j}} I_{\mathbf{j}}(\mathbf{w}) \quad \text{with} \quad I_{\mathbf{j}}(\mathbf{w}) = I_{j_1}(u_1^{(1)}) \cdots I_{j_{p_1}}(u_{p_1}^{(1)}) I_{j_{p_1+1}}(u_1^{(2)}) \cdots I_{j_{p+q}}(u_{p_2}^{(2)}),$$

where

$$\rho_{\mathbf{m}}^{(1)} = \mathbb{E}(L_{m_1}(U_1^{(1)}) \cdots L_{m_p}(U_p^{(1)})) \quad \text{with} \quad U_i = F_{X_i}(X_i),$$

$$\rho_{\mathbf{k}}^{(2)} = \mathbb{E}(L_{k_1}(U_1^{(2)}) \cdots L_{k_q}(U_q^{(2)})) \quad \text{with} \quad V_i = F_{Y_i}(Y_i),$$

$$\rho_{\mathbf{j}} = \mathbb{E}(L_{j_1}(U_1^{(1)}) \cdots L_{j_{p_1}}(U_{p_1}^{(1)}) L_{j_{p_1+1}}(U_1^{(2)}) \cdots L_{j_{p_1+p_2}}(U_{p_2}^{(2)})),$$

$$I_i(t) = \int_0^t L_i(x)dx,$$

Write $\mathbf{j} = (\mathbf{j}(1), \mathbf{j}(2))$. Then

$$\begin{aligned} H_0 &\Longleftrightarrow \mathbf{C}^{\mathbf{X}} = \mathbf{C}^{(1)}.\mathbf{C}^{(2)} \\ &\Longleftrightarrow \rho_{\mathbf{j}} - \rho^{(1)}_{\mathbf{j(1)}} \times \rho^{(2)}_{\mathbf{j(2)}} = 0 \qquad \forall \mathbf{j} \in \mathbb{N}^{p_1+p_2} \end{aligned}$$

REMARK 1. 1. Combinaison of remark **??** and properties of multivariate marginal probability distribution function implies

- if the first p_1 coefficients of $\mathbf{j}$ are zero, that is $\mathbf{j} = (0, \cdots, 0, j_{p_1+1}, \cdots, j_{p_1+p_2})$ then

$$\rho_{\mathbf{j}} = \rho^{(2)}_{\mathbf{k}} \text{ and } \rho^{(1)}_{\mathbf{m}} = 1 \Longrightarrow \rho_{\mathbf{j}} - \rho^{(1)}_{\tilde{\mathbf{j}}} \times \rho^{(2)}_{\tilde{\mathbf{j}}} = 0$$

- if the last p_2 coefficients of $\mathbf{j}$ are zero, that is $\mathbf{j} = (j_1, \cdots, j_{p_1}, 0, \cdots, 0)$ then

$$\rho_{\mathbf{j}} = \rho^{(1)}_{\mathbf{m}} \text{ and } \rho^{(2)}_{\mathbf{k}} = 1 \Longrightarrow \rho_{\mathbf{j}} - \rho^{(1)}_{\tilde{\mathbf{j}}} \times \rho^{(2)}_{\tilde{\mathbf{j}}} = 0$$

2. If $\mathbf{j}$ is a zero vector, then

$$\rho_{\mathbf{j}} - \rho^{(1)}_{\tilde{\mathbf{j}}} \times \rho^{(2)}_{\tilde{\mathbf{j}}} = 0$$

3. If only one of the coefficients of $\mathbf{j}$ is nonzero, then

$$\rho_{\mathbf{j}} - \rho^{(1)}_{\tilde{\mathbf{j}}} \times \rho^{(2)}_{\tilde{\mathbf{j}}} = 0$$

Our test procedure is based on the sequences of differences

$$r^{(1,2)}_{\mathbf{j}} := \widehat{\rho}_{\mathbf{j}} - \widehat{\rho}^{(1)}_{\mathbf{j}(1)} \times \widehat{\rho}^{(2)}_{\mathbf{j}(2)} \text{ for } \mathbf{j} \in \mathbb{N}^{p_1+p_2},$$

where

$$\widehat{\rho}_{\mathbf{j}} = \frac{1}{n} \sum_{i=1}^{n} L_{j_1}(\widehat{U}^{(1)}_{i,1}) \times \cdots \times L_{j_{p_1}}(\widehat{U}^{(1)}_{i,p_1}) \times L_{j_{p_1+1}}(\widehat{U}^{(2)}_{i,1}) \times \cdots \times L_{j_{p_1+p_2}}(\widehat{U}^{(2)}_{i,p_2})),$$

with the convention that $r^{(1,2)}_{\mathbf{j}} = 0$ when $\mathbf{j}$ satistify the properties of above remark.

In order to select automatically the number of copula coefficients we consider the L^1 norm of any vector $\mathbf{j} = (j_1, \cdots, j_{p_1+p_2})$ defined by

$$\|\mathbf{j}\| = |j_1| + \cdots + |j_{p_1+p_2}|,$$

and for any integer $d > 1$ we write

$$\mathcal{S}(d) = \{\mathbf{j} \in \mathbb{N}^{p_1+p_2}; \|\mathbf{j}\| = d \text{ and there exists } k \leq p_1 \text{ and } k' > p_1 \text{ such that } j_k > 0 \text{ and } j_{k'} > 0\}.$$

The set $\mathcal{S}(d)$ contains all non null integers $\mathbf{j} = (j_1, \cdots, j_{p+q})$ wich the sum is equal to d and such that the first p_1 element and the last p_2 element are different of zero. We will denote by $c(d) = \binom{d}{d+p_1+p_2-1} - \binom{d}{d+p_1-1} - \binom{d}{d+p_2-1}$ the cardinal of $\mathcal{S}(d)$ ... and we introduce a lexicographic order on $\mathbf{j} \in \mathcal{S}(d)$ as follows:

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

For instance,

- in the univariate case, that is $p_1 = p_2 = 1$, with $d = 2$ we have one possibility: $\mathbf{j} = (j_1, j_2) = (1,1) \Rightarrow ord(\mathbf{j}, d) = 1$ the cases $(0,1)$ and $(1,0)$ are excluded.

- in the bivariate case, that is $p_1 = p_2 = 2$ with $d = 2$, there is only four possibilities:

$$\mathbf{j} = (j_1, j_2, j_3, j_4) = (1,0,1,0) \Rightarrow ord(\mathbf{j}, d) = 1$$
$$\mathbf{j} = (j_1, j_2, j_3, j_4) = (1,0,0,1) \Rightarrow ord(\mathbf{j}, d) = 2$$
$$\mathbf{j} = (j_1, j_2, j_3, j_4) = (0,1,1,0) \Rightarrow ord(\mathbf{j}, d) = 3$$
$$\mathbf{j} = (j_1, j_2, j_3, j_4) = (0,1,0,1) \Rightarrow ord(\mathbf{j}, d) = 4.$$

  The cases $(2,0,0,0)$,$(0,2,0,0)$, $(0,0,2,0)$,$(0,0,0,2)$,$(1,1,0,0)$ or $(0,0,1,1)$ are excluded.

To construct our test statistics, we introduce a series of statistics based on the differences between their copula coefficients and copula coefficients of the vector $\mathbf{Z} = (\mathbf{X}^{(1)}, \mathbf{X}^{(2)})$ as follows: for $1 \leq k \leq c(2)$ we define

$$T_{2,k} = n \sum_{\mathbf{j} \in \mathcal{S}(2); ord(\mathbf{j},2) \leq k} (r_{\mathbf{j}}^{(1,2)})^2, \tag{8}$$

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

$$T_{d,k} = T_{d-1,c(d-1)} + n \sum_{\mathbf{j} \in \mathcal{S}(d); ord(\mathbf{j},d) \leq k} (r_{\mathbf{j}}^{(1,2)})^2. \tag{9}$$

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

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

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

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

When d is large it will make it possible to compare high coefficient orders through $r_{\mathbf{j}}$, while k will permit to visit all the values of $\mathbf{j}$ for this given order. To simplify notation we write such a sequence of statistics as

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

By construction, for all integer $k > 0$ there exists a set $\mathcal{H}(k) \subset \bigcup_{d=2}^{\infty} \mathcal{S}(d)$ with $card(\mathcal{H}(k)) = k$ and such that

$$V_k^{(1,2)} = n \sum_{\mathbf{j} \in \mathcal{H}(k)} (r_{\mathbf{j}}^{(1,2)})^2, \tag{10}$$

and we have the following relation: for all $k \geq 1$ and $j = 1, \cdots, c(k+1)$

$$V_{c(1)+c(2)+\cdots+c(k)+j}^{(1,2)} = T_{k+1,j},$$

with the convention $c(1) = 0$.

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

$$D(n) := \min\Big\{ \underset{1\leq k\leq d(n)}{\operatorname{argmax}} (V_k^{(1,2)} - kq_n)\Big\}, \tag{11}$$

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

(A) $d(n)^{(7\max(p_1,p_2)-4)} = o(q_n)$

A classical choice for q_n is $\log(n)$ initially used in **?** (see for instance the seminal work of **?**). This choice is convenient to detect smooth alternatives (see Section 4) and will be adopted in our simulation. Our first result shows that under the null the least penalized statistic will be selected.

THEOREM 2.1. *Let assumption* **(A)** *holds. Then, under H_0, $D(n)$ converges in Probability towards 1 as $n \to +\infty$.*

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

$$\begin{aligned} r_{\mathbf{j}}^{(1,2)} &= \widehat{\rho}_{\mathbf{j}} - \widehat{\rho}_{\mathbf{j}(\mathbf{1})} \times \widehat{\rho}_{\mathbf{j}(\mathbf{2})} \\ &= \frac{1}{n}\sum_{i=1}^{n} \big(L_1(\widehat{U}_{i,1}^{(1)}) L_1(\widehat{U}_{i,1}^{(2)})\big) - 0 \times 0, \text{ since } \rho_{\mathbf{j}(\mathbf{1})} = \rho_{\mathbf{j}(\mathbf{2})} = 0 \text{ following remark 1.} \end{aligned}$$

It follows that $r_{\mathbf{j}}^{(1,2)}$ is the estimation of kendall's tau between $U_1^{(1)}$ and $U_1^{(2)}$. Under the null, this coefficient is equal to zero and this is the least penalized statistic which is retained by the procedure. Asymptotically, the null distribution reduces to that of $V_1^{(1,2)}$ and is given below.

THEOREM 2.2. *Let assumption* **(A)** *holds and assume that* $\mathbf{j} = (1,0,\cdots,0,1,0,\cdots,0)$. *Then, under H_0, $V_1^{(1,2)} = \sqrt{n} r_{\mathbf{j}}^{(1,2)}$ converges in law towards a central normal distribution with variance*

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

where $F^{(i)}, i = 1, 2$ is the cumulative distribution function of $X_1^{(i)}$.

In order to normalize the test, write

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

with

$$\overline{M} = \frac{1}{n}\sum_{i=1}^{n} M_i$$

where

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

PROPOSITION 1. *Under H_0 we have the following convergence in probability*

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

We then deduce the limit distribution under the null.

COROLLARY 2.3. *Let assumption* **(A)** *holds. 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 $n \to +\infty$.*

3. The *K*-sample case. We assume that we observe K iid samples from $\mathbf{X}^{(1)}, \cdots, \mathbf{X}^{(K)}$, denoted by

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

Our aim is to generalize the two-sample case by considering a series of embedded statistics, each new of them including a new pair of populations to be compared. In this way we introduce the following set of indexes:

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

Clearly $\mathcal{V}(K)$ contains $v(K) = K(K-1)/2$ elements which represent all the pairs of populations that we want to compare and that can be ordered as follows: we write $(\ell, m) <_{\mathcal{V}} (\ell', m')$ if $\ell < \ell'$, or $\ell = \ell'$ and $m < m'$, and we denote by $rank_{\mathcal{V}}(\ell, m)$ the associated rank of (ℓ, m) in $\mathcal{V}(K)$. This can be seen as a natural order (left to right and bottom to top) of the elements of the upper triangle of a $(K-1) \times (K-1)$ matrix as represented below:

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

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

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

Using the previous two-sample statistics we construct an embedded series of statistics as

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

or equivalently,

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

where $D(n)$ is given by (11) and $V^{(\ell,m)}$ is defined as in (10). We have $V_1 < \cdots < V_{v(K)}$. The first statistic V_1 compares the first two populations 1 and 2. The second statistic V_2 compares the populations 1 and 2, and, in addition, the populations 1 and 3. And so on. For each $1 < k < v(K)$, there exists a unique pair (ℓ, m) such that $rank_{\mathcal{V}}(\ell, m) = k$. To choose automatically the appropriate number k we introduce the following penalization procedure, mimicking the Schwarz criteria procedure **?**:

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

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

$$s(\mathbf{n}) = \min\left\{ \underset{1\leq k\leq v(K)}{\operatorname{argmax}} \left(V_k - kp_n\right)\right\}. \tag{12}$$

We discuss this choice in Remark 2. We make the following assumption:

(A') $d(n)^{(7\max(p_1,\cdots,p_K)-4)} = o(p_n)$????????????? Attention: les vecteurs n'ont pas la même taille

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

THEOREM 3.1. *Assume that* **(A)** *and* **(A')** *hold. Then under* H_0, $s(\mathbf{n})$ *converges in probability towards 1 as* $n \to +\infty$.

COROLLARY 3.2. *Assume that* **(A)** *and* **(A')** *hold. Then under* H_0, $V_{s(\mathbf{n})}/\widehat{\sigma}^2(1,2))$ *converges in law towards a* χ_1^2 *distribution.*

Then our final data driven test statistic is given by

$$V = V_{s(\mathbf{n})}/\widehat{\sigma}^2(1,2)). \tag{13}$$

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

$$H_1(1) : \mathbf{X}^{(1)} \not\perp\!\!\!\perp \mathbf{X}^{(2)},$$

and for $k > 1$:

$$H_1(k) : \mathbf{X}^{(i)} \perp\!\!\!\perp \mathbf{X}^{(j)} \text{ for } rank_{\mathcal{V}}(i,j) < k \text{ and } \mathbf{X}^{(i)} \not\perp\!\!\!\perp \mathbf{X}^{(j)} \text{ for } rank_{\mathcal{V}}(i,j) = k,$$

with $1 < k \leq v(K)$. The hypothesis $H_1(k)$ means that the ith and jth populations such that $rank_{\mathcal{V}}(i,j) = k$ are the first dependent on each other (in the sense of the order in $\mathcal{V}(K)$).

We make the following assumption:

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

THEOREM 4.1. *Assume that* **(A)-(A')-(B)** *hold. Then under* $H_1(k)$, $s(\mathbf{n})$ *converges in probability towards* k, *as* $\mathbf{n} \to +\infty$, *and* V *converges to* $+\infty$, *that is,* $\mathbb{P}(V < \epsilon) \to 0$, *for all* $\epsilon > 0$.

REMARK 2. In the classical smooth test approach **?** a standard penalty is $q_n = p_n = \alpha \log(n)$, which is related to the Schwarz criteria **?** as discussed in **?**. In practice, the factor α permits to stabilize the empirical level to be as close as possible to the asymptotic one. Note also that **?** compared this type of Schwarz penalty to the Akaike one where they proposed p_n or q_n to be constant. In our simulation we consider the classical choice $q_n = p_n = \alpha \log(n)$, with an automatic choice of α described in Section **??** which makes it possible to calibrate the test very simply.

5. Simulation study.

5.1. *Two sample.*

5.1.1. $p_1 = 1$ *and* $p_2 = 1$.

5.1.2. $p_1 = 1$ *and* $p_2 = 10$.

5.1.3. $p_1 = 5$ *and* $p_2 = 5$.

5.1.4. $p_1 = 4$ *and* $p_2 = 7$.

5.2. *Ten sample.*

5.2.1. $p_i = 1, i = 1, \cdots, 10$.

5.2.2. $p_1 = \cdots = p_{10} = 3$.

6. Conclusion.

7. Proof of Theorem 4.1. We give the proof for the case $k > 1$, the particular case $k = 1$ being similar. We first show that $\mathbb{P}(s(\mathbf{n}) \geq k)$ tends to 1. Under $H_1(k)$, we have for all $k' < k$:

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

When $rank_{\mathcal{V}}(\ell,m)=k$, under $H_1(k)$, since $\mathbf{X}^{(\ell)} \not\perp\!\!\!\perp \mathbf{X}^{(m)}$, there exists $\mathbf{j}_0$ such that $\rho_{\mathbf{j}_0} \neq \rho^{(\ell)}_{\mathbf{j}_0(1)}\rho^{(m)}_{\mathbf{j}_0(2)}$, that is, $r^{(\ell,m)}_{\mathbf{j}_0} \neq 0$. We can write

$$\mathbb{P}\left(\mathbb{I}_{\{rank_{\mathcal{V}}(\ell,m)=k\}} n \sum_{\mathbf{j}\in\mathcal{H}(D(n))} (r^{(\ell,m)}_{\mathbf{j}})^2 \geq (k-k')p_{\mathbf{n}}\right)$$

$$\geq \mathbb{P}\left(\mathbb{I}_{\{rank_{\mathcal{V}}(\ell,m)=k\}} n \mathbb{I}_{\mathbf{j}_0\in\mathcal{H}(D(n))} (r^{(\ell,m)}_{\mathbf{j}_0})^2 \geq (k-k')p_{\mathbf{n}}\right) \tag{14}$$

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

$$r^{(\ell,m)}_{\mathbf{j}_0} = (\widehat{\rho}_{\mathbf{j}_0} - \rho_{\mathbf{j}_0}) - \left(\rho^{(m)}_{\mathbf{j}_0(2)}\left(\widehat{\rho}^{(\ell)}_{\mathbf{j}_0(1)} - \rho^{(\ell)}_{\mathbf{j}_0(1)}\right) + \widehat{\rho}^{(\ell)}_{\mathbf{j}_0(1)}\left(\widehat{\rho}^{(m)}_{\mathbf{j}_0(2)} - \rho^{(m)}_{\mathbf{j}_0(2)}\right)\right) + \left(\rho_{\mathbf{j}_0} - \rho^{(\ell)}_{\mathbf{j}_0(1)}\rho^{(m)}_{\mathbf{j}_0(2)}\right)$$

According to lemma 8, we get

8. Proof.

Proof of Theorem 2.1. We want to show that $\mathbb{P}(D(n)>1) \to 0$ as n tends to infinity. We have

$$\begin{aligned}
\mathbb{P}_0\Big(D(n)>1\Big) &= \mathbb{P}_0\Big(\exists k\in\{2,\cdots,d(n)\} : V^{(1,2)}_k - k\,q_n \geq V^{(1,2)}_1 - q_n\Big)\\
&= \mathbb{P}_0\Big(\exists k\in\{2,\cdots,d(n)\} : V^{(1,2)}_k - V^{(1,2)}_1 \geq (k-1)q_n\Big)\\
&= \mathbb{P}_0\Big(\exists k\in\{2,\cdots,d(n)\} : n\sum_{\mathbf{j}\in\mathcal{H}^*(k)} (r^{(1,2)}_{\mathbf{j}})^2 \geq (k-1)q_n\Big)\\
&\leq \mathbb{P}_0\Big(n\sum_{\mathbf{j}\in\mathcal{H}^*(d(n))} (r^{(1,2)}_{\mathbf{j}})^2 \geq q_n\Big), \qquad (15)
\end{aligned}$$

with $\mathcal{H}(k)$ satisfying (10) 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 $(r^{(1,2)}_{\mathbf{j}})^2$ in the difference $V^{(1,2)}_k - V^{(1,2)}_1$. For simplification of notation, we write $\mathcal{H}^*$ instead of $\mathcal{H}^*(d(n))$.

Under the null we have $\rho_{\mathbf{j}} = \rho^{(1)}_{\mathbf{j}(1)}\rho^{(2)}_{\mathbf{j}(2)}$ and we can decompose $r^{(1,2)}_{\mathbf{j}}$ as follows

$$\begin{aligned}
r^{(1,2)}_{\mathbf{j}} &= \widehat{\rho}_{\mathbf{j}} - \widehat{\rho}^{(1)}_{\mathbf{j}(1)}\widehat{\rho}^{(2)}_{\mathbf{j}(2)}\\
&= (\widehat{\rho}_{\mathbf{j}} - \rho_{\mathbf{j}}) - \widehat{\rho}^{(1)}_{\mathbf{j}(1)}(\widehat{\rho}^{(2)}_{\mathbf{j}(2)} - \rho^{(2)}_{\mathbf{j}(2)}) - \rho^{(2)}_{\mathbf{j}(2)}(\widehat{\rho}^{(1)}_{\mathbf{j}(1)} - \rho^{(1)}_{\mathbf{j}(1)}). \qquad (16)
\end{aligned}$$

We write

$$\widetilde{\rho}_{\mathbf{j}} = \frac{1}{n}\sum_{s=1}^n L_{j_1(1)}(U^{(1)}_{s,1})\cdots L_{j_{p_1}(1)}(U^{(1)}_{s,p_1}) L_{j_{p_1+1}(2)}(U^{(2)}_{s,1})\cdots L_{j_{p_1+p_2}(2)}(U^{(2)}_{s,p_2})$$

There exists a constant $c>0$ such that, for all $\mathbf{j}\in\mathcal{H}^*$,

$$\max(\rho_{\mathbf{j}}, \widehat{\rho}_{\mathbf{j}}, \widetilde{\rho}_{\mathbf{j}}) \leq c\, d(n)^{p_1+p_2}$$

$$\max(\rho^{(k)}_{\mathbf{j}(k)}, \widehat{\rho}^{(k)}_{\mathbf{j}(k)}, \widetilde{\rho}^{(k)}_{\mathbf{j}(k)})) \leq c\, d(n)^{p_k} \quad k=1,2$$

Proof. If $\mathbf{j}$ belongs to $\mathcal{H}^* = \mathcal{H}^*(d(n))$ we have $\|\mathbf{j}\| \leq d(n)$ and the proof is immediate with property (?) of Legendre polynomials. ■

Since Lemma ?? holds even if $p_1 = 0$ or $p_2 = 0$ we see at once that

$$(r_{\mathbf{j}}^{(1,2)})^2 \leq 3(\widehat{\rho}_{\mathbf{j}} - \rho_{\mathbf{j}})^2 + 3cd(n)^{2p_1}(\widehat{\rho}_{\mathbf{j}(2)}^{(2)} - \rho_{\mathbf{j}(2)}^{(2)})^2 + 3cd(n)^{2p_2}(\widehat{\rho}_{\mathbf{j}(1)}^{(1)} - \rho_{\mathbf{j}(1)}^{(1)})^2,$$

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

$$\begin{aligned}
&\mathbb{P}_0\big(D(n)>1\big) \\
&\qquad \leq \mathbb{P}_0\big(n\sum_{\mathbf{j}\in\mathcal{H}^*}(\widehat{\rho}_{\mathbf{j}} - \rho_{\mathbf{j}})^2 \geq q_n/4\big) \\
&\qquad\quad + \mathbb{P}_0\big(nd(n)^{2p_2}\sum_{\mathbf{j}\in\mathcal{H}^*}(\widehat{\rho}_{\mathbf{j}(1)}^1 - \rho_{\mathbf{j}(1)}^1)^2 \geq q_n/8\big) \\
&\qquad\quad + \mathbb{P}_0\big(nd(n)^{2p_1}\sum_{\mathbf{j}\in\mathcal{H}^*}(\widehat{\rho}_{\mathbf{j}(2)}^2 - \rho_{\mathbf{j}(2)}^2)^2 \geq q_n/8\big).
\end{aligned}$$

For all $\alpha > 0$,

$$\mathbb{P}(n\sum_{\mathbf{j}\in\mathcal{H}^*}(\widehat{\rho}_{\mathbf{j}} - \rho_{\mathbf{j}})^2 \geq \alpha q_n) \to 0$$

$$\mathbb{P}(nd(n)^{2p_k}\sum_{\mathbf{j}\in\mathcal{H}^*}(\widehat{\rho}_{\mathbf{j}(k)}^{(k)} - \rho_{\mathbf{j}(k)}^{(k)})^2 \geq \alpha q_n) \to 0, \quad k=1,2$$

Proof. Write

$$\begin{aligned}
E_{\mathbf{j}} &= \widehat{\rho}_{\mathbf{j}} - \widetilde{\rho}_{\mathbf{j}} \\
G_{\mathbf{j}} &= \widetilde{\rho}_{\mathbf{j}} - \rho_{\mathbf{j}} \\
H_{\mathbf{j}} &= \widehat{\rho}_{\mathbf{j}} - \rho_{\mathbf{j}}
\end{aligned}$$

Clearly

$$\mathbb{P}(n\sum_{\mathbf{j}\in\mathcal{H}^*} H_{\mathbf{j}}^2 \geq \alpha q_n) \leq \mathbb{P}(n\sum_{\mathbf{j}\in\mathcal{H}^*} 2E_{\mathbf{j}}^2 \geq \alpha q_n) + \mathbb{P}(n\sum_{\mathbf{j}\in\mathcal{H}^*} 2G_{\mathbf{j}}^2 \geq \alpha q_n)$$

We first study the quantity involving $E_{\mathbf{j}}$.

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

$$\begin{aligned}
|E_{\mathbf{j}}| &\leq \frac{1}{n}\sum_{s=1}^{n}\left(\sum_{i=1}^{p_1} S_i^{(1)} \sup_x |L'_{j_i(1)}(x)\prod_{u\neq i} L_{j_u(1)}(x)| + \sum_{i=1}^{p_2} S_i^{(2)} \sup_x |L'_{j_i(2)}(x)\prod_{u\neq i} L_{j_u(2)}(x)|\right) \\
&= \qquad\qquad \widetilde{A} \qquad\qquad + \qquad\qquad \widetilde{B}.
\end{aligned}$$

Obviously we have $|E_{\mathbf{j}}|^2 \leq 2\widetilde{A}^2 + 2\widetilde{B}^2$ and then

$$\mathbb{P}_0\big(n\sum_{\mathbf{j}\in\mathcal{H}^*}(E_{\mathbf{j}})^2 \geq \alpha q_n/2\big) \leq \mathbb{P}_0\big(n\sum_{\mathbf{j}\in\mathcal{H}^*}(\widetilde{A})^2 \geq \alpha q_n/4\big) + \mathbb{P}_0\big(n\sum_{\mathbf{j}\in\mathcal{H}^*}(\widetilde{B})^2 \geq \alpha q_n/4\big). \tag{17}$$

From (**??**) and (**??**) (see Appendix **??**) there exists a constant $\widetilde{c} > 0$ such that

$$|\widetilde{A}| \leq \widetilde{c} \sum_{i=1}^{p_1} S_i^{(1)} (j_i(1)^{1/2} \prod_{u \neq i} j_u(1)^{5/2})$$

$$|\widetilde{B}| \leq \widetilde{c} \sum_{i=1}^{p_2} S_i^{(2)} (j_i(2)^{1/2} \prod_{u \neq i} j_u(2)^{5/2}).$$

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

$$\begin{aligned}
&\mathbb{P}_0\big(n \sum_{\mathbf{j} \in \mathcal{H}^*} (\widetilde{A})^2 \geq \alpha q_n/4\big) \\
&\leq \mathbb{P}_0\big(n \sum_{\mathbf{j} \in \mathcal{H}^*} \widetilde{c} \sum_{i=1}^{p_1} \sum_{i'=1}^{p_1} S_i^{(1)} S_{i'}^{(1)} j_i(1)^{1/2} j_{i'}(2)^{1/2} \prod_{s \neq i} j_s(1)^{5/2} \prod_{s' \neq i'} j_{'s}(2)^{5/2} \geq \alpha q_n/4\big) \\
&\leq \mathbb{P}_0\big(\widetilde{c} \sum_{i=1}^{p_1} \sum_{i'=1}^{p_1} n S_i^{(1)} S_{i'}^{(1)} d(n)^{5p_1 - 4} \geq \alpha q_n/4\big),
\end{aligned}$$

(18) $\to 0$ as $n \to \infty$,

since for all $i = 1, \cdots, p_1$, $\sqrt{n} S_i^{(1)}$ converges in law to a Kolmogorov distribution and $d(n)^{5p_1 - 3} = o(q_n)$ by **(A)**. In the same way we otain that

$$\mathbb{P}_0\big(n \sum_{\mathbf{j} \in \mathcal{H}^*} (\widetilde{B})^2 \geq \alpha q_n/4\big) \to 0 \text{ as } n \to \infty, \tag{19}$$

and then

$$\mathbb{P}_0\big(n \sum_{\mathbf{j} \in \mathcal{H}^*} (E_{\mathbf{j}})^2 \geq \alpha q_n/2\big) \to 0 \text{ as } n \to \infty, \tag{20}$$

Coming back to (**??**) we now study the quantity involving $G_{\mathbf{j}}$. First note that $\mathbb{E}(G_{\mathbf{j}}) = 0$. Moreover, $\mathbb{V}(G_{\mathbf{j}}) = \mathbb{V}(\prod_{i=1}^{p_1} L_{j_i(1)}(U_i^{(1)}) \prod_{i=1}^{p_2} L_{j_i(2)}(U_i^{(2)}))/n$. Then, by Markov inequality we have

$$\mathbb{P}_0\big(n \sum_{\mathbf{j} \in \mathcal{H}^*} (G_{\mathbf{j}})^2 \geq \alpha q_n/2\big) \leq \frac{\sum_{\mathbf{j} \in \mathcal{H}^*} \mathbb{V}(\prod_{i=1}^{p_1} L_{j_i(1)}(U_i^{(1)}) \prod_{i=1}^{p_2} L_{j_i(2)}(U_i^{(2)}))}{\alpha q_n/2}.$$

From (**??**) (see Appendix **??**) there exists a constant $c > 0$ such that

$$\mathbb{V}\big(\prod_{i=1}^{p_1} L_{j_i(1)}(U_i^{(1)}) \prod_{i=1}^{p_2} L_{j_i(2)}(U_i^{(2)})\big) \leq c \prod_{i=1}^{p_1 + p_2} j_i.$$

It follows that

$$\begin{aligned}
\mathbb{P}_0\big(n \sum_{\mathbf{j} \in \mathcal{H}^*} (G_{\mathbf{j}})^2 \geq \alpha q_n/2\big) &\leq \frac{c d(n)^{p_1 + p_2}}{\alpha q_n/2} \\
&\to 0 \text{ as } n \to \infty,
\end{aligned} \tag{21}$$

and finally

$$\mathbb{P}(n\sum_{\mathbf{j}\in\mathcal{H}^*} H_{\mathbf{j}}^2 \leq \alpha q_n) \to 0 \text{ as } n\to\infty,$$

■

The proof is complete with Lemma ??. ■

Proof of Theorem 2.2. Let $\mathbf{j} = (1,0,\cdots,0,1,0,\cdots,0)$. We have $V_1^{(1,2)} = T_{2,1}^{(1,2)} = \left(\sqrt{n}r_{\mathbf{j}}^{(1,2)}\right)^2$ and we can decompose $\sqrt{n}r_{\mathbf{j}}^{(1,2)}$ under the null as follows:

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

where under the null

$m = \rho_{\mathbf{j}} - \rho_{\mathbf{j}}^{(1)}\rho_{\mathbf{j}}^{(2)} = \mathbb{E}(L_1(U_{i,1}^{(1)})L_1(U_{i,1}^{(2)})) - \mathbb{E}(L_1(U_{i,1}^{(1)}))\mathbb{E}(L_1(U_{i,1}^{(2)})) = 0$ since $\rho_{\mathbf{j}} = \rho_{\mathbf{j}}^{(1)}\rho_{\mathbf{j}}^{(2)}$

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

$$\begin{aligned}\sqrt{n}r_{\mathbf{j}}^{(1,2)} &= \sqrt{n}(\int\int L_1(\widehat{F}_1^{(1)}(x))L_1(\widehat{F}_1^{(2)}(y))d\widehat{F}_n(x,y) - m)\\ &= \sqrt{n}(\int\int L_1(F_1^{(1)}(x))L_1(F_1^{(2)}(y))d\widehat{F}_n(x,y) - m)\\ &\quad +\sqrt{n}\int\int(\widehat{F}_1^{(1)}(x) - F_1^{(1)}(x))2\sqrt{3}L_1(F_1^{(2)}(y))dF(x,y)\\ &\quad +\sqrt{n}\int\int(\widehat{F}_1^{(2)}(y) - F_1^{(2)}(y))2\sqrt{3}L_1(F_1^{(1)}(x))dF(x,y)\\ &\quad +\sqrt{n}\int\int(\widehat{F}_1^{(1)}(x) - F_1^{(1)}(x))2\sqrt{3}L_1(F_1^{(2)}(y))d(\widehat{F}_n(x,y) - F(x,y))\\ &\quad +\sqrt{n}\int\int(\widehat{F}_1^{(2)}(y) - F_1^{(2)}(y))2\sqrt{3}L_1(F_1^{(1)}(x))d(\widehat{F}_n(x,y) - F(x,y))\\ &:= \sqrt{n}\left(A_{1,n} + A_{2,n} + A_{3,n} + B_{1,n} + B_{2,n}\right).\end{aligned}$$

Since $(\widehat{F} - F)(x) = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{I}(X_i \leq x - F(x)\right)$, we can rewrite

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

$$:= \frac{1}{n}\sum_{i=1}^{n}(Z_{1,i} + Z_{2,i} + Z_{3,i})$$

$$:= \frac{1}{n}\sum_{i=1}^{n} Z_i$$

where Z_i are iid random variables.

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

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

where

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

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

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

where

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

and

$$\begin{aligned}B_{2,k,n} = \iint \Big(\mathbb{1}_{X_{k,1}^{(1)} \leq x} - F_1^{(1)}(x)\Big)L_1\big(F_1^{(2)}(y)\big)dF(x,y) \\ - \frac{1}{n}\sum_{i=1}^{n}\Big(\mathbb{1}_{X_{k,1}^{(1)} \leq X_{i,1}^{(1)}} - U_{i,1}^{(1)}\Big)L_1(U_{i,1}^{(2)}).\end{aligned} \tag{22}$$

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

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

By Glivenko-Cantelli's Theorem we obtain

$$|B_{1,k,n}| \le 2\sqrt{3}S_1^{(2)}+2\sqrt{3}S_1^{(2)}+\sqrt{3}S_1^{(1)} \text{ where } S_1^{(j)} = \sup_x|\widehat{F}_1^{(j)}(x)-F_1^{(j)}(x)|$$

$$= o_{\mathbb{P}}(1). \tag{23}$$

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

$$B_{2,k,n} = \left(\frac{1}{n}\sum_{i=1}^{n}U_{i,1}^{(1)}L_1(U_{i,1}^{(2)})-\iint F_1^{(1)}(x)L_1\big(F_1^{(2)}(y)\big)dF(x,y)\right)$$
$$+\left(\iint \mathbb{1}_{X_{k,1}^{(1)}\le x_1^{(1)}}L_1\big(F_1^{(2)}(y)\big)dF(x,y)-\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{X_{k,1}^{(1)}\le X_{i,1}^{(1)}}L_1(U_{i,1}^{(2)})\right)$$
$$\equiv B_{2,k,n}^1+B_{2,k,n}^2.$$

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

$$B_{2,k,n}^1 = \frac{1}{n}\sum_{s=1}^{n}U_{s,1}^{(1)}L_1(U_{s,1}^{(2)})-\iint F_1^{(1)}(x)L_1\big(F_1^{(2)}(y)\big)dF(x,y)$$
$$=\frac{1}{n}\sum_{s=1}^{n}U_{s,1}^{(1)}L_1(U_{s,1}^{(2)})-\mathbb{E}\Big(U_1^{(1)}L_1(U_1^{(2)})\Big).$$

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

$$B_{2,k,n}^1 = o_{\mathbb{P}}(1). \tag{24}$$

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

$$B_{2,k,n}^2 = \iint \mathbb{1}_{X_{k,1}^{(1)}\le x}L_1\big(F_1^{(2)}(y)\big)dF(x,y)-\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{X_{k,1}^{(1)}\le X_{i,1}^{(1)}}L_1(U_{i,1}^{(2)})$$
$$=\iint \mathbb{1}_{F_1^{(1)}(X_{k,1}^{(1)})\le F_1^{(1)}(x)}L_1\big(F_1^{(2)}(y)\big)dF(x,y)$$
$$-\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{F_1^{(1)}(X_{k,1}^{(1)})\le F_1^{(1)}(X_{i,1}^{(1)})}L_1(U_{i,1}^{(2)})$$
$$=\int_0^1\int_0^1 \mathbb{1}_{U_{k,1}^{(1)}\le u}L_1(v)dC^{(1,2)}(u,v)-\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{U_{k,1}^{(1)}\le U_{i,1}^{(1)}}L_1(U_{i,1}^{(2)})$$

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

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

where

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

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

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

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

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

It implies that

$$B_{2,k,n}^2 = o_{\mathbb{P}}(1), \tag{25}$$

and we conclude that $B_{1,n} = o_{\mathbb{P}}(n^{-1/2})$. The same result occurs for $B_{2,n}$.

Finally, by symmetry we obtain $B_n^{(2)} = o_{\mathbb{P}}(n^{-1/2})$, which proves the theorem. ■

Proof of Proposition 1. Let us define

$$\overline{W} = \frac{1}{n}\sum_{i=1}^{n}W_i$$

where

$$\begin{aligned}W_i = L_1(U_{i,1}^{(1)})L_1(U_{i,1}^{(2)}) &+ 2\sqrt{3}\int\int\left(\mathbb{I}(X_{i,1}^{(1)}\leq x) - F_1^{(1)}(x)\right)L_1(F_1^{(2)}(y))dF(x,y) \\ &+2\sqrt{3}\int\int\left(\mathbb{I}(X_{i,1}^{(2)}\leq y) - F_1^{(2)}(y)\right)L_1(F_1^{(1)}(x))dF(x,y).\end{aligned}$$

By construction $W_1,W_2,\cdots,,W_n$ are iid and we have

$$\frac{1}{n}\sum_{i=1}^{n}\left(W_i-\overline{W}\right)^2 \xrightarrow{\mathbb{P}} \sigma^2(1,2)). \tag{26}$$

According to Slusky's Lemma and (26), the proof is completed by showing that

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

We have

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

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

$$\max(|W_i|,|M_i|)\leq\kappa,$$

which implies that

$$\left|\frac{1}{n}\sum_{i=1}^{n}\Big(W_i-\overline{W}\Big)^2-\widehat{\sigma}^2(1,2))\right|\leq\frac{4\kappa}{n}\sum_{i=1}^{n}|W_i-M_i|.$$

It remains to prove that $W_i-M_i\xrightarrow{\mathbb{P}}0$. We have

$$W_{i,1}-M_{i,1}=I_{i,1}+2\sqrt{3}I_{i,2}+2\sqrt{3}I_{i,3},\tag{27}$$

where

$$\begin{aligned}
I_{i,1}&=L_1(U_{i,1}^{(1)})L_1(U_{i,1}^{(2)})-L_1(\widehat{U}_{i,1}^{(1)})L_1(\widehat{U}_{i,1}^{(2)}),\\
I_{i,2}&=\iint\Big(\mathbb{I}(X_{i,1}^{(1)}\leq x)-F_1^{(1)}(x)\Big)L_1\big(F_1^{(2)}(y)\big)dF(x,y),\\
&\quad-\frac{1}{n}\sum_{k=1}^{n}\Big(\mathbb{I}(X_{i,1}^{(1)}\leq X_{k,1}^{(1)})-\widehat{U}_{k,1}^{(1)}\Big)L_1(\widehat{U}_{k,1}^{(2)})\\
I_{i,3}&=\iint\Big(\mathbb{I}(X_{i,1}^{(2)}\leq x)-F_1^{(2)}(x)\Big)L_1\big(F_1^{(1)}(y)\big)dF(x,y)\\
&\quad-\frac{1}{n}\sum_{k=1}^{n}\Big(\mathbb{I}(X_{i,1}^{(2)}\leq X_{k,1}^{(2)})-\widehat{U}_{k,1}^{(2)}\Big)L_1(\widehat{U}_{k,1}^{(1)}).
\end{aligned}$$

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

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

Next, we remark that $I_{i,2}=B_{2,k,n}$, where $B_{2,k,n}$ is defined in (22). Then $I_{i,2}=o_{\mathbb{P}}(1)$ and similarly $I_{i,3}=o_{\mathbb{P}}(1)$. It follows that $W_i-M_i\xrightarrow{\mathbb{P}}0$ which completes the proof.

■

Proof of Corollary 2.3. The proof is immediate. ■

Proof of Theorem 3.1. ■

Proof of Theorem 4.1. We give the proof for the case $k > 1$, the particular case $k = 1$ being similar. We first show that $\mathbb{P}(s(\mathbf{n}) \geq k)$ tends to 1. Under $H_1(k)$, we have for all $k' < k$:

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

When $rank_{\mathcal{V}}(\ell, m) = k$, under $H_1(k)$, since $\mathbf{X}^{(\ell)} \not\perp\!\!\!\perp \mathbf{X}^{(m)}$, there exists $\mathbf{j}_0$ such that $\rho_{\mathbf{j}_0} \neq \rho_{\mathbf{j}_0(1)}^{(\ell)} \rho_{\mathbf{j}_0(2)}^{(m)}$, that is, $r_{\mathbf{j}_0}^{(\ell,m)} \neq 0$. We can write

$$\begin{aligned}
&\mathbb{P}\left(\mathbb{I}_{\{rank_{\mathcal{V}}(\ell,m)=k\}} n \sum_{\mathbf{j} \in \mathcal{H}(D(n))} (r_{\mathbf{j}}^{(\ell,m)})^2 \geq (k - k')p_{\mathbf{n}}\right) \\
&\geq \mathbb{P}\left(\mathbb{I}_{\{rank_{\mathcal{V}}(\ell,m)=k\}} n \mathbb{I}_{\mathbf{j}_0 \in \mathcal{H}(D(n))} (r_{\mathbf{j}_0}^{(\ell,m)})^2 \geq (k - k')p_{\mathbf{n}}\right) \qquad (28)
\end{aligned}$$

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

$$r_{\mathbf{j}_0}^{(\ell,m)} = (\widehat{\rho}_{\mathbf{j}_0} - \rho_{\mathbf{j}_0}) - \left(\rho_{\mathbf{j}_0(2)}^{(m)} \left(\widehat{\rho}_{\mathbf{j}_0(1)}^{(\ell)} - \rho_{\mathbf{j}_0(1)}^{(\ell)}\right) + \widehat{\rho}_{\mathbf{j}_0(1)}^{(\ell)} \left(\widehat{\rho}_{\mathbf{j}_0(2)}^{(m)} - \rho_{\mathbf{j}_0(2)}^{(m)}\right)\right) + \left(\rho_{\mathbf{j}_0} - \rho_{\mathbf{j}_0(1)}^{(\ell)} \rho_{\mathbf{j}_0(2)}^{(m)}\right)$$

According to lemma 8, we get