draft Clustering Independence in high dimensions

Yves I Ngounou Bakam

draft
Clustering Independence in high dimensions

Yves I. Ngounou Bakam

Abstract

1. Introduction and motivation

For $k = 1, \cdots, K$, let $\boldsymbol{X}^{(k)} = (X_1^{(k)}, \cdots, X_{p_k}^{(k)})$ be a p_k-dimensional random variable with univariate marginals probability distribution function (pdf) denoted by $F_1^{(k)}, \cdots, F_{p_k}^{(k)}$.

The goal is to partition the set $S = \{\boldsymbol{X}^{(1)}, \cdots, \boldsymbol{X}^{(K)}\}$ of random vectors, each potentially of different dimensions, into clusters. Within each cluster, the variables should exhibit independence, while across clusters, there should be some level of dependence. As far as we know, addressing this problem in its complete generality remains unexplored.

This clustering of independence among random vectors offers several advantages and points of interest, particularly in various scientific, technological, and practical applications:

- Promoting Ethical Practices: As machine learning technologies become more pervasive and influential in society, addressing ethical considerations such as fairness, bias, and accountability becomes paramount. Clustering independence can help mitigate risks associated with data dependencies, model biases, and unintended consequences, promoting ethical practices and responsible AI development.
- Resource Optimization: In various applications such as machine learning and optimization problems, clustering independence helps in efficiently allocating resources by identifying independent components that can be processed or analyzed separately.
- Efficient Algorithms: Algorithms designed to leverage clustering independence can be more computationally efficient, reducing time and computational resources required for tasks like data processing, analysis, and modeling.
- Resilient Systems: Understanding the independence structure ensures that systems or models are built on robust foundations, minimizing the risk of overfitting and enhancing stability across different scenarios or environments.
- Relevant Features: By understanding which variables or features are independent of each other, one can prioritize and select the most relevant features for modeling and prediction tasks, improving model efficiency and accuracy.

Partitioning a set of random vectors into clusters characterized by intra-cluster independence and inter-cluster dependence offers a structured, flexible, and insightful approach to analyzing complex systems and datasets, providing numerous advantages in terms of modeling, efficiency, interpretability, and applicability across diverse domains and applications.

The paper is organised as follows:

2. Copula coefficients test of independence

We consider the following null hypothesis of independence

$$H_0 : \boldsymbol{X}^{(1)} \perp\!\!\!\perp \boldsymbol{X}^{(2)} \perp\!\!\!\perp \cdots \perp\!\!\!\perp \boldsymbol{X}^{(K)} \text{ versus } H_1 : \neg H_0 \tag{1}$$

When $K = 2$, there is a large literature for testing independence between two random vectors. We can mention the two seminal papers of Gieser and Randles [17] and Horrell and Lessig [19]. More recently, Feng et al. [12] considered the problem of multivariate tests of independence using high dimensional rank statistics as Spearman and Kendall'ones. We can also mention Berrett and Samworth [5] where the test statistic is based on entropies, Bodnar et al. [6] resticting their study to high dimension Gaussian vectors, as in Silva et al. [27]. And among others Albert et al. [1] and Yin and Yuan [29].

When $p_1 = \cdots = p_K = 1$, (1) remains to test the independence of the K components of a vector. In such a case a numerous works attempt to construct efficient statistics, with a series of papers based on copulas as in Genest et al. [15], González-Barrios et al. [18], Roy et al. [25], or Genest et al. [13]. Other approaches can be mentioned as in Mao [23], using high dimension Kendall's tau, as more recently in Drton et al. [10].

In the general case, when $K > 2$ and $p_j \geq 1$, there is very few general works. One reason for this is that the null distributions for the proposed test statistics become difficult to evaluate. We can mention Chakraborty and Zhang [7] where their approach is based on the notion of distance covariance and Roy and Ghosh [24] who uses the ideas of maximun mean discrepancy and ranks of nearest neighbors. Certain restrict their study to Gaussian vectors, as in Bao et al. [4] or Chen and Liu [8], working on the covariance structure in high dimension. The work who seems to tackle the general problem is that of Fan et al. [11] where the authors represented the independence hypothesis through the product of characteristic functions. They deduce a Cramer-Von-Mises statistic with null asymptotic distribution related to a process involving eigenvalues of covariance function. In addition their approach necessitates a weight function to calibrate the numerical integration of the Cramer-Von-Mises statistic. However, these authors worked in a very general setting since they consider continuous as well as discrete or mixed random vectors.

Our approach can be considered as a competitor of Fan et al. [11] in the continuous case and is quite similar in spirit but instead of using the characteristic functions we use copulas.

By definition, a p_k-dimensional copula $\boldsymbol{C}^{(k)}$ is the joint cumulative distribution of a random vector $\boldsymbol{U}^{(k)} = (U_1^{(k)}, \cdots, U_{p_k}^{(k)})$ with uniform marginals over $[0, 1]$, that is,

$$\begin{aligned} \boldsymbol{C}^{(k)} &: [0,1]^{p_k} \to [0,1] \\ \boldsymbol{u} = (u_1 \cdots, u_{p_k}) &\mapsto \boldsymbol{C}^{(k)}(\boldsymbol{u}) = \mathbb{P}\Big(U_1^{(k)} \leq u_1, \cdots, U_{p_k}^{(k)} \leq u_{p_k}\Big). \end{aligned}$$

The *copula coefficients* associated to $\boldsymbol{C}^{(k)}$, denoted $\rho_{\mathbf{j}}^{(k)}$, are defined to be (see **?**])

$$\rho_{\mathbf{j}}^{(k)} = \int_{[0,1]^{p_k}} L_{j_1}(u_1) \cdots L_{j_{p_k}}(u_{p_k}) d\boldsymbol{C}^{(k)}(u_1 \cdots, u_{p_k}), \tag{2}$$

where L_j is the Legendre polynomial on $[0, 1]$ of degree j satisfying

$$\int_{[0,1]} L_j(x) L_k(x) dx = \delta_{jk}$$

where $\delta_{jk} = 1$ if $j = k$ and 0 otherwise. These polynomials are defined by

$$\begin{aligned} &L_0 = 1, L_1(x) = \sqrt{3}(2x - 1), \text{ and for } n > 1 : \\ &(n+1)L_{n+1}(x) = \sqrt{(2n+1)(2n+3)}(2x-1)L_n(x) - \frac{n\sqrt{2n+3}}{\sqrt{2n-1}} L_{n-1}(x). \end{aligned}$$

According to Proposition 1 of Bakam and Pommeret [2], the series of copula coefficients entirely characterize the copulas, that is

$$\boldsymbol{C}^{(k)} = \boldsymbol{C}^{(\ell)} \text{ if and only if } \rho_{\mathbf{j}}^{(k)} = \rho_{\mathbf{j}}^{(\ell)} \text{ for all } \mathbf{j} \tag{3}$$

where for all $j \in \{1, \cdots, p_k\}$, $\quad I_j = \displaystyle\int_0^1 L_j(s) ds.$

2.1. Notation

Write $p = p_1 + \cdots + p_K$. For $\mathbf{j} \in \mathbb{N}^p$ and for $\mathbf{x} \in \mathbb{R}^p$ we decompose $\mathbf{j} = (\mathbf{j}^{(1)}, \cdots, \mathbf{j}^{(K)})$, $\mathbf{x} = (\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(K)})$ and $\mathbf{u} = (u_1^{(1)}, \cdots, u_{p_K}^{(K)})$, where $\mathbf{j}^{(k)} = (j_1^{(k)}, \cdots, j_{p_k}^{(k)})$, $\mathbf{x}^{(k)} = (x_1^{(k)}, \cdots, x_{p_k}^{(k)})$ and $\mathbf{u}^{(k)} = (u_1^{(k)}, \cdots, u_{p_k}^{(k)})$, for all $k = 1, \cdots, K$. $\mathbf{u}^{(k)} = (u_1^{(k)}, \cdots, u_{p_k}^{(k)})$ and $\mathbf{u} = (u_1^{(1)}, \cdots, u_{p_K}^{(K)})$. We also write $\mathbf{X} = (\mathbf{X}^{(1)}, \cdots, \mathbf{X}^{(K)})$ the p-vector obtained from the K populations. We use the classical transformations: $\mathbf{U} = (\mathbf{U}^{(1)}, \cdots, \mathbf{U}^{(K)})$ where $\mathbf{U}^{(k)} = (U_1^{(k)}, \cdots, U_{p_k}^{(k)}) = (F_1^{(k)}(X_1^{(k)}), \cdots, F_{p_k}^{(k)}(X_{p_k}^{(k)})$, for all $k = 1, \cdots, K$. The empirical versions based of n iid observations of such variables are obtained by replacing $U_j^{(k)}$ by $\widehat{U}_j^{(k)}$ where

$$\widehat{U}_j^{(k)} = \frac{1}{n}\sum_{i=1}^{n} \widehat{F}_j^{(k)}(X_{j,i}^{(k)}),$$

where $X_{j,i}^{(k)}$ is the ith observation of the jth component from population K.

By Sklar's theorem (Sklar [28]), the continuity assumption on the marginals implies that the the testing problem (1) can be rewritten as follows

$$H_0 : \boldsymbol{C} = \boldsymbol{C}^{(1)} \cdots \boldsymbol{C}^{(K)} \text{ versus } H_1 : \boldsymbol{C} \neq \boldsymbol{C}^{(1)} \cdots \boldsymbol{C}^{(K)} \tag{4}$$

where for $k = 1, \cdots, K$, $\boldsymbol{C}^{(k)}$ and $\boldsymbol{C}$ are the unique copulas associated to$\mathbf{X}^{(k)}$ and $\mathbf{X}$ respectively such that

$$\boldsymbol{C}(1, \ldots, 1, \boldsymbol{u}^{(k)}, 1, \ldots, 1) = \boldsymbol{C}^{(k)}(\boldsymbol{u}^{(k)})$$

Rather than comparing copulas, our procedure is based on the copula coefficients defined in (2). According to (3), the null hypothesis changes in the following way

$$\widetilde{H}_0 : \quad \rho_{\mathbf{j}} = \rho_{\mathbf{j}^{(1)}}^{(1)} \cdots \rho_{\mathbf{j}^{(K)}}^{(K)}, \qquad \forall \mathbf{j} \in \mathbb{N}^p, \tag{5}$$

where $\rho_{\mathbf{j}}$ is the $\mathbf{j}$th copula coefficient associated to $\mathbf{C}$.

2.2. Data-driven smooth test

Our procedure consists in detecting the largest differences between the copula coefficients involved in (5). We consider K iid samples from $\mathbf{X}^{(1)}, \cdots, \mathbf{X}^{(K)}$, denoted by

$$(X_{1,i}^{(1)}, \cdots, X_{p_1,i}^{(1)})_{i=1,\cdots,n}, \cdots, (X_{1,i}^{(K)}, \cdots, X_{p_K,i}^{(K)})_{i=1,\cdots,n}.$$

We want to detect a difference between $\rho_{\mathbf{j}}$ and $\rho_{\mathbf{j}^{(1)}}^{(1)} \cdots \rho_{\mathbf{j}^{(K)}}^{(K)}$, for some $\mathbf{j} \in \mathbb{N}^p$. For any vector $\mathbf{j} = (j_1, \cdots, j_p)$ we denote the $L1$ norm by

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

Remark 1. From the orthogonality of the Legendre polynomials we obtain:

- If $\mathbf{j}$ is a zero vector, then $\rho_{\mathbf{j}} - \rho_{\mathbf{j}^{(1)}}^{(1)} \cdots \rho_{\mathbf{j}^{(K)}}^{(K)} = 0$
- If only one of the coefficients of $\mathbf{j}$ is non null, then $\rho_{\mathbf{j}} - \rho_{\mathbf{j}^{(1)}}^{(1)} \cdots \rho_{\mathbf{j}^{(K)}}^{(K)} = 0$

Considering the null hypothesis $\widetilde{H}_0$ as expressed in (**??**), our test procedure is based on the sequences of differences

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

where

$$\widehat{\rho}_{\mathbf{j}} = \begin{cases} 1 & \text{if } \mathbf{j} = \mathbf{0}, \\ 0 & \text{if exactly one component of } \mathbf{j} \text{ is non nul}, \\ \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} L_{j_1^{(1)}}(\widehat{U}_{1,i}^{(1)}) \times \cdots \times L_{j_{p_1}^{(1)}}(\widehat{U}_{p_1,i}^{(1)}) \times L_{j_1^{(2)}}(\widehat{U}_{1,i}^{(2)}) \times \cdots \times L_{j_{p_K}^{(K)}}(\widehat{U}_{p_K,i}^{(K)}), & \text{else.} \end{cases}$$

$$\widehat{\rho}_{j^{(k)}}^{(k)} = \begin{cases} 1 & \text{if } \mathbf{j}^{(1)} = \mathbf{0}, \\ 0 & \text{if exactly one component of } \mathbf{j}^{(1)} \text{ is non null}, \\ \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} L_{j_1^{(k)}}(\widehat{U}_{1,i}^{(k)}) \times \cdots \times L_{j_{p_k}^{(k)}}(\widehat{U}_{p_k,i}^{(k)}), & \text{else.} \end{cases}$$

Note that by construction $r_{\mathbf{j}} = 0$ when $\mathbf{j}$ satisfies one of the conditions of Remark 1. Such estimators of copula coefficients have been studied in Bakam and Pommeret [3] where it is shown their excellent behavior.

In order to select automatically the number of copula coefficients we introduce the following set for any integer $p > 1$:

$$\mathcal{S}(p) \quad = \quad \{\mathbf{j} \in \mathbb{N}^p;\ \text{and at least two components are non null}\}.$$

We introduce the following order on $\mathcal{S}(p)$.

If $\mathbf{j}$ and $\widetilde{\mathbf{j}}$ are two vectors in $\mathcal{S}(p)$ we define

$$k(\mathbf{j}, \widetilde{\mathbf{j}}) \quad = \quad \min\{k \in \{1, \cdots, K\}; \mathbf{j}^{(k)} \neq \widetilde{\mathbf{j}}^{(k)}\}$$

the first index such that the subvectors of $\mathbf{j}$ and $\widetilde{\mathbf{j}}$ are different. In the same way

$$k(\mathbf{j}^{k)}, \widetilde{\mathbf{j}}^{(k)}) \quad = \quad \min\{\ell \in \{1, \cdots, p_k\}; \mathbf{j}_\ell^{(k)} \neq \widetilde{\mathbf{j}}_\ell^{(k)}\}$$

Then the order on $\mathcal{S}(p)$ is defined by

$$\mathbf{j} < \widetilde{\mathbf{j}} \quad \text{if} \quad \begin{cases} ||\mathbf{j}||_1 < ||\widetilde{\mathbf{j}}||_1 \\ ||\mathbf{j}||_1 = ||\widetilde{\mathbf{j}}||_1 \text{ and } \mathbf{j}_r^{(R)} < \widetilde{\mathbf{j}}_r^{(R)} \end{cases}$$

where $R = k(\mathbf{j}, \widetilde{\mathbf{j}})$ and $r = k(\mathbf{j}^{(R)}, \widetilde{\mathbf{j}}^{(R)})$.

For instance, the first element of $\mathcal{S}(p)$ is such that $\mathbf{j}^{(1)} = (1, 0, \cdots, 0)$, $\mathbf{j}^{(2)} = (1, 0, \cdots, 0)$, and $\mathbf{j}^{(k)} = (0, \cdots, 0)$, for $k = 3, \cdots, p$. In that case we say that $Ord(\mathbf{j}) = 1$.

To construct our test statistics, we consider a sequence $d(n)$ which tends to infinity as $n \to \infty$, and we introduce a series of statistics based on the sequence $(r_{\mathbf{j}})_{\mathbf{j} \in \mathbb{N}^p}$ as follows: for $1 \leq k \leq d(n)$, we define

$$T_k = n \sum_{\mathbf{j} \in \mathcal{S}(p); ord(\mathbf{j}) \leq k} (r_{\mathbf{j}})^2. \tag{6}$$

Clearly all these statistics are embedded.

When $d(n)$ 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. Notice that we need to compare all copula coefficients and then to let $d(n)$ tend to infinity to detect all possible alternatives. However, choosing too large parameters tends to power dilution of the test. Following [21], we suggest a data driven procedure to select automatically the number of coefficients to test the hypothesis $\widetilde{H}_0$. Namely, we set

$$D(n) := \min\{\underset{1 \leq k \leq d(n)}{\operatorname{argmax}}(T_k - kq_n)\}, \tag{7}$$

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 $\boldsymbol{C}$ and $\boldsymbol{C}^{(1)} \cdots \boldsymbol{C}^{(p)}$ is $T_{D(n)}$ and we consider the following rate for the number of components in the statistic:

(A) $d(n)^{(7p-4)} = o(q_n)$

A classical choice for q_n is $q_n = \log(n)$ initially used in Schwarz [26] (see for instance the seminal work of Ledwina [22]). This choice is convenient to detect smooth alternatives (see Section 2.3) and will be adopted in our simulation. Our first result shows that under the null the least penalized statistic will be selected.

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

Theorem 2. *Let assumption* **(A)** *holds. Then, under H_0, $T_{D(n)}$ converges in law towards a central normal distribution with variance*

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

In order to normalize the test, write

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

where

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

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

$$\widehat{\sigma}^2 \xrightarrow{\mathbb{P}} \sigma^2.$$

We then deduce the limit distribution under the null.

Corollary 1. *Let assumption* **(A)** *holds. Then under H_0, $T_{D(n)}/\widehat{\sigma}^2$ converges in law towards a chi-squared of one degree of freedom distribution χ_1^2 as $n \to +\infty$.*

2.3. Convergence under alternative hypotheses

We consider the general alternative hypothesis: $H_1 : \boldsymbol{C} \neq \boldsymbol{C}^{(1)} \cdots \boldsymbol{C}^{(p)}$ and we make the following assumption:

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

Theorem 3. *Assume that* **(A)-(B)** *hold. Then under H_1, $T_{D(n)}$ converges to $+\infty$, that is, $\mathbb{P}(T_{D(n)} < \epsilon) \to 0$, for all $\epsilon > 0$.*

Remark 2. In the classical smooth test approach [22] a standard penalty is $q_n = b_n = \log(n)$, which is related to the Schwarz criteria [26] as discussed in [21]. Note also that [20] compared this type of Schwarz penalty to the Akaike one where they proposed b_n or q_n to be constant. In our simulation we consider the classical choice $q_n = b_n = \log(n)$.

3. Clustering approach

The purpose of this section is to partition any set $S = \{\boldsymbol{X}^{(1)}, \cdots, \boldsymbol{X}^{(K)}\}$ of random vectors of arbitrary dimension into clusters in such a way that inside each cluster the variables are independent and the clusters between themselves are dependent. To our knowledge, the problem in its full generality has not yet been addressed.

This clustering procedure can solve several complex problems in a very short time and is useful in practice, for instance: i) in finance to build a well-diversified portfolio whose stock options would be less subject to a systemic crisis; ii) in the world of actuarial science by making it possible to practice a price segmentation strategy; iii) in biology to classify interdependent and intra-independent network of genes; iv) in a general regression framework where it is of importance to retain a set of independent covariables.

3.1. Clustering principle

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

More precisely, we consider a multivariate variable $\mathcal{S} = \{\boldsymbol{X}^{(1)}, \cdots, \boldsymbol{X}^{(K)}\}$ of p dimension in which $\boldsymbol{X}^{(\ell)}$ is a $p_\ell \geq 1$-dimensional random vectors for $\ell \in \{1, 2, \cdots, K\}$, such that $p_1 + p_2 + \cdots + p_K = p$. For $l = 1, \cdots K$, we consider a sample $\{\boldsymbol{X}_{i,1}^{(l)}, \cdots, X_{i,p_l}^{(l)})_{i=1,\cdots,n}\}, l = 1, \cdots K$ from $X^{(\ell)}$. The clustering algorithm starts by choosing the two variables of $\mathcal{S}$ that are less distant in terms of independence. In this way, it chooses the smallest two-sample statistic. If the independence of both associated vectors is not rejected, these two variables form the first cluster. Then the algorithm proposes a new variable closest to this cluster in terms of independence. While the test accepts the simultaneous independence of the vectors, the cluster growths. If the test is rejected then the cluster is closed. We iterate this several times until every vector is associated with a cluster.

3.2. Clustering algorithm

We can summarize the clustering algorithm as follows:

Algorithm: Independence clustering

Initialization: $c = 1$, $\mathcal{S} = \{\boldsymbol{X}^{(1)}, \cdots, \boldsymbol{X}^{(K)}\}$, and $S_0 = \emptyset$;
Select $\{\ell^\star, m^\star\} = \text{argmin}\{V^{(\ell,m))}_{D(n)}; \ell \neq m \in S \setminus \bigcup_{k=1}^{c} S_{k-1}\}$;
Test H_0 between $\boldsymbol{X}^{(\ell^\star)}$ and $\boldsymbol{X}^{(m^\star)}$;
if *H_0 is not rejected* **then**
    $S_1 = \{\boldsymbol{X}^{\ell^\star}, \boldsymbol{X}^{m^\star}\}$;
else
    STOP. There is no cluster.
end
while $S \setminus \bigcup_{k=1}^{c} S_k \neq \emptyset$ **do**
    Select $\{j^\star\} = \text{argmin}\{V^{(i,j)}_{D(n)}; i \in S_c, j \in S \setminus \bigcup_{k=1}^{c} S_k\}$;
    Test H_0 the simultaneous independence of all $\boldsymbol{X}^{(i)}$, $i \in S_c$ and $\boldsymbol{X}^{(j^\star)}$;
    if *H_0 is not rejected* **then**
        $S_c = S_c \bigcup \{\boldsymbol{X}^{(j^\star)}\}$;
    else
        $S_{c+1} = \{\boldsymbol{X}^{(j^\star)}\}$;
        $c = c + 1$;
    end
end

4. Simulation study

5. Real datasets Application

5.1. Geology data

We analyse the well-known uranium exploration dataset of Cook and Johnson [9]. The same data set has been analyzed by Genest & Rivest [16] and Genest, Quessy & Rémillard [14] to demonstrate a semiparametric inference for Archimedean copulas and a goodness-of-fit test. The data consists of concentrations of seven chemical elements in 655 water samples collected from the Montrose quad-rangle of Western Colorado. Concentrations of Uranium (U), Lithium (Li), Cobalt (Co), Potassium (K), Cesium (Cs), Scandium (Sc) and Titanium (Ti) were measured. The comparison of these chemical elements in the context of mutual dependence or groupwise dependence is a major concern for geologists.

5.2. Finance data

In an ever-evolving global landscape shaped by globalization and the outsourcing of major corporations, often through a growing nexus of cooperation and economic integration among countries and financial markets, making astute investments and achieving success becomes progressively complex. In this context, investors are keen on examining the interplay among significant global financial markets to construct a diversified portfolio with substantial potential.

5.3. Machine learning

6. Conclusion

7. Proofs

References

[1] M. Albert, B. Laurent, A. Marrel, A. Meynaoui, Adaptive test of independence based on hsic measures, arXiv preprint arXiv:1902.06441 (2019).
[2] Y. I. N. Bakam, D. Pommeret, K-sample test for equality of copulas, arXiv preprint arXiv:2112.05623 (2021).
[3] Y. I. N. Bakam, D. Pommeret, Nonparametric estimation of copulas and copula densities by orthogonal projections, Econometrics and Statistics (2023).
[4] Z. Bao, J. Hu, G. Pan, W. Zhou, Test of independence for high-dimensional random vectors based on freeness in block correlation matrices, Electronic Journal of Statistics 11 (2017) 1527–1548.
[5] T. B. Berrett, R. J. Samworth, Nonparametric independence testing via mutual information, Biometrika 106 (2019) 547–566.
[6] T. Bodnar, H. Dette, N. Parolya, Testing for independence of large dimensional vectors, The Annals of Statistics 47 (2019) 2977–3008.
[7] S. Chakraborty, X. Zhang, Distance metrics for measuring joint dependence with application to causal inference, Journal of the American Statistical Association (2019).
[8] X. Chen, W. Liu, Testing independence with high-dimensional correlated samples, The Annals of Statistics 46 (2018) 866–894.
[9] R. D. Cook, M. E. Johnson, Generalized burr-pareto-logistic distributions with applications to a uranium exploration data set, Technometrics 28 (1986) 123–131.
[10] M. Drton, F. Han, H. Shi, High-dimensional consistent independence testing with maxima of rank correlations, The Annals of Statistics 48 (2020) 3206–3227.
[11] Y. Fan, P. L. de Micheaux, S. Penev, D. Salopek, Multivariate nonparametric test of independence, Journal of Multivariate Analysis 153 (2017) 189–210.
[12] L. Feng, X. Zhang, B. Liu, Multivariate tests of independence and their application in correlation analysis between financial markets, Journal of Multivariate Analysis 179 (2020) 104652.
[13] C. Genest, J. Nešlehová, B. Rémillard, O. Murphy, Testing for independence in arbitrary distributions, Biometrika 106 (2019) 47–68.
[14] C. Genest, J.-F. Quessy, B. Rémillard, Goodness-of-fit procedures for copula models based on the probability integral transformation, Scandinavian Journal of Statistics 33 (2006) 337–366.
[15] C. Genest, J.-F. Quessy, B. Rémillard, Asymptotic local efficiency of cramér–von mises tests for multivariate independence, The Annals of Statistics 35 (2007) 166–191.
[16] C. Genest, L.-P. Rivest, Statistical inference procedures for bivariate archimedean copulas, Journal of the American statistical Association 88 (1993) 1034–1043.
[17] P. W. Gieser, R. H. Randles, A nonparametric test of independence between two vectors, Journal of the American Statistical Association 92 (1997) 561–567.
[18] J. M. González-Barrios, E. Gutiérrez-Peña, J. D. Nieves, R. Rueda, A novel characterization and new simple tests of multivariate independence using copulas, arXiv preprint arXiv:1906.02196 (2019).
[19] J. F. Horrell, V. P. Lessig, A note on a nonp$\alpha\rho\alpha\alpha$metric test of independence between two vectors, Journal of Marketing Research 11 (1974) 106–108.
[20] T. Inglot, T. Ledwina, Towards data driven selection of a penalty function for data driven neyman tests, Linear Algebra and its Applications 417 (2006) 124–133.
[21] W. C. Kallenberg, T. Ledwina, Consistency and monte carlo simulation of a data driven version of smooth goodness-of-fit tests, The Annals of Statistics (1995) 1594–1608.
[22] T. Ledwina, Data-driven version of neyman's smooth test of fit, Journal of the American Statistical Association 89 (1994) 1000–1005.
[23] G. Mao, Testing independence in high dimensions using kendall's tau, Computational Statistics & Data Analysis 117 (2018) 128–137.
[24] A. Roy, A. K. Ghosh, Some tests of independence based on maximum mean discrepancy and ranks of nearest neighbors, Statistics & Probability Letters 164 (2020) 108793.
[25] A. Roy, A. K. Ghosh, A. Goswami, C. Murthy, Some new copula based distribution-free tests of independence among several random variables, Sankhya A (2020) 1–41.
[26] G. Schwarz, Estimating the dimension of a model, The annals of statistics (1978) 461–464.
[27] I. R. Silva, Y. Zhuang, J. C. Junior, Kronecker delta method for testing independence between two vectors in high-dimension, Statistical Papers (2021) 1–23.
[28] M. Sklar, Fonctions de repartition an dimensions et leurs marges, Publ. inst. statist. univ. Paris 8 (1959) 229–231.
[29] X. Yin, Q. Yuan, A new class of measures for testing independence, Statistica Sinica 30 (2020) 2131–2154.