The k-means clustering procedure is a way to identify distinct groups within a population. This procedure prescribes a criterion for partitioning a set of datas into k groups : to divide points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ in $\mathbb{R}^d$ according to this criterion, first choose cluster centres $a_1, \ldots, a_k$ to minimise

$$W_n = \frac{1}{n} \sum_{i=1}^{n} \min_{j \in \{1,\ldots,k\}} d(x_i, a_j),$$

where $d : \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty)$ be a distance function or, more generally, a dissimilarity function in $\mathbb{R}^d$ The motivation is to identify cluster centers such that distances of the observations to their nearest cluster centers are minimized. Accordingly, all observations which are closest to the same cluster center are viewed as belonging to the same group.

For a probability measure $\mathbb{P}$ on $\mathcal{B}(\mathbb{R}^d)$ and a set $A = \{\mathbf{a}_1, \ldots, \mathbf{a}_k\}$, $\mathbf{a}_j \in \mathbb{R}^d$ for $j = 1, \ldots, k$ and $k \in \mathbb{N}$, one can introduce the averaged distance from any observation to the closest element of A as

$$W(A, \mathbb{P}) := \int_{\mathbb{R}^d} \min_{\mathbf{a} \in A} d(\mathbf{x}, \mathbf{a}) \mathbb{P}(d\mathbf{x}).$$

For given $\mathbb{P}$ and k, a set A_k which minimizes $W(A, \mathbb{P})$ among all A with $|A| \leq k$, where $|A|$ stands for the cardinality of the (finite) set A, can be seen as a set of theoretical cluster centers. Note that the set may not necessarily be unique.

If we replace $\mathbb{P}$ by its sample version $\mathbb{P}_n$ (*i.e.* the measure that places mass n^{-1} on each observation $\mathbf{X}_1, \ldots, \mathbf{X}_n$ of a sample) and derive an accordingly optimal set A_k^n, its components minimize the sum of the distances from every observation to its nearest cluster center.

Proofs of consistency theorems for k-means clustering needs a uniform strong law of large numbers (SLLN) stated as (see Section 4 of [Pollard, 1981])

$$\sup_{g \in \mathcal{G}} \left| \int g d(\mathbb{P}_n - \mathbb{P}) \right| \xrightarrow[n \to \infty]{a.s.} 0. \tag{1}$$

Equation (1) is also stated as the class of functions $\mathcal{G}$ is Glivencko-Cantelli (see Chapter 2 of [van der Vaart et al., 1996]). In Theorem 1 of [Janßen and Wan, 2020] where $\mathbb{P}$ and $\mathbb{P}_n$ are replaced respectively by the angular measure and an estimator which has been shown that Equation (1) holds. In our framework, we consider the copula C and C_n the empirical copula process as the measure. The consistency of k-means clustering directly comes down from arguments given in [Janßen and Wan, 2020, Pollard, 1981] given that we are able to state Equation (1).

For this purpose, the notion of bounded variation of functions and in particular the integration by parts formula for Lebesgue-Stieltjes integral is of prime interest (see, for example, Theorem 6 of [Fermanian et al., 2004] or Appendix A.2 in [Fermanian, 1998]). Indeed the integral

$$\int g d(C_n - C) \tag{2}$$

can thus be expressed as the integral of $C_n - C$ with respect to a finite bounded variation function g. Furthermore, this expression is shown to be continuous. We say that g is BVHK if and only if $V_{HK}(g) < \infty$ where V_{HK} denotes that the function is of bounded variation in the Hardy-Krause sense (see references below for a definition). Using this notion, we state the following result.

Theorem 1. *Assume that C is a copula and that $C_n, n \in \mathbb{N}$ is the empirical copula defined on a common probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Furthermore, assume that $J : [0,1]^d \times [0,1]^d \to [0,1]$ is a continuous function with $V_{HK}(J) < \infty$. For each C_n and a given value of $k \in \mathbb{N}$, denote by A_k^n a random set which minimizes*

$$W(A, S_n) := \int_{[0,1]^d} \min_{\boldsymbol{a} \in A} J(\boldsymbol{u}, \boldsymbol{a}) C_n(d\boldsymbol{u}) \tag{3}$$

among all sets $A \subseteq [0,1]^d$ with at most k elements. Accordingly, if we replace C_n by C, denote the optimal set by A_k, and assume that for a given value of k the set A_k is uniquely determined, thus A_n^k converges almost surely to A_k as $n \to \infty$.

Proof As J is a continuous function and $[0,1]^d$ is compact by Tychonov theorem, Theorem 3.1 of [Janßen and Wan, 2020] applies entirely except for Equation (3.3) which has to be proved. This equation is implied by the more general statement :

$$\int_{[0,1]^d} g(\mathbf{u}) C_n(d\mathbf{u}) \xrightarrow[n\to\infty]{a.s.} \int_{[0,1]^d} g(\mathbf{u}) C(d\mathbf{u}),$$

for all continuous function g such that $V_{HK}(g) < \infty$. It is implied by the expression :

$$\int_{[0,1]^d} |g(\mathbf{u})| \, (C_n - C)(d\mathbf{u}) \xrightarrow[n\to\infty]{a.s.} 0. \tag{4}$$

Using integration by parts see proof of Theorem 1 or Theorem 15 in [Radulović et al., 2017], Equation (4) can be rewritten as

$$\Gamma(C_n - C, |g|),$$

where $\Gamma(\cdot, |g|)$ is linear and Lipschitz. As $||C_n - C||_\infty$ converges almost surely to 0, we thus have Equation (4) by the continuous mapping theorem.

Let $\mathbf{v} \in [0,1]^d$. As $V_{HK}(J) < \infty$, we obtain that $V_{HK}(\min_{\mathbf{a} \in A} J(\cdot, \mathbf{a})) < \infty$ since $|A| \leq k$.

Set $\mathcal{E}_k := \{B \subset [0,1]^d, |B| \leq k\}$. Continuity of J and compactness of $[0,1]^d$ imply that $W_S : B \mapsto W(B, S)$ is continuous with respect to the Haussdorf-metric d_H on $\mathcal{E}_k$. As $[0,1]^d$ is compact, we can, for a given $\epsilon > 0$, find m and $B_1, \ldots, B_m \in \mathcal{E}_k$ such that

$$\min_{i=1,\ldots,m} d_H(B, B_i) < \epsilon$$

for all $B \in \mathcal{E}_k$. So, using the same notation as in [Janßen and Wan, 2020], we obtain, using Equation

(4)

$$\max_{i=1,\dots,m} |W(B_i, C_n) - W(B_i, C)| \xrightarrow[n\to\infty]{a.s.} 0. \tag{5}$$

Which is Equation (3.3) in *loc. sit.*, hence the statement. □

Remark 1. *In the above theorem, the convergence of sets if formally meant in the Haussforf distance d_H, but since all involved sets have finitely many elements, it implies pointwise convervence of elements after a suitable reordering.*

In applications, clustering extremes using a probabilistic measure dissimilarity which can be interpreted in the framework of extremes is often used (see, *e.g.* [Bernard et al., 2013, Bador et al., 2015, Saunders et al., 2021]). These algorithms use the madogram $J(\mathbf{u}, \mathbf{v}) = 2^{-1}|\mathbf{u} - \mathbf{v}|_1$, where $|\cdot|_1$ denotes the absolute value norm which can be linked to the pairwise extremal coefficient. This dissimilarity is used to evaluate the proximity between two stations in terms of their extremal behavior. Formally speaking, the statistician has access to a $\mathbf{X} \in \mathbb{R}^{n\times d}$ matrix where n denotes yearly observation period say and d the number of stations. The k-means procedure is thus used on the columns of $\mathbf{X}$ using the madogram as a dissimilarity metric. Theorem 1 thus get out of the scope as we want to cluster n points (years in our example) and not stations. Yet, one can perform k-means with $\mathbf{X}^\top \in \mathbb{R}^{d\times n}$, but again the *i.i.d.* hypothesis would be in struggle as the d stations are not identically nor independently distributed. The problem of clustering $\mathbf{X} = (X_1, \dots, X_d)$ is known as variable clustering (see [Bunea et al., 2020]) and its application to extreme, as far as we know, has not been investigated yet.

Corollary 1. *Kmeans clustering using the madogram as a dissimilarity measure is consistent.*

Proof Take $\mathbf{v} \in [0,1]^d$ fixed. We now have to show that the function $J(\cdot, \mathbf{v}) = \frac{1}{2}\sum_{j=1}^d |u_j - v_j|$ is of bounded variation in the sense of Hardy-Krause and use Theorem 1 to conclude. Indeed, $\forall j \in \{1, \dots, d\}$, u_j and v_j are BVHK on $[0,1]^d \times [0,1]^d$ since it depends only in one variable and is monotone in this variable. As the difference between two BVHK functions are BVHK, it follows that $\forall j \in \{1, \dots, d\}$, $u_j - v_j$ is BVHK on $[0,1]^d \times [0,1]^d$. Finally, the absolute value is also BVHK on $[0,1]^d \times [0,1]^d$. Thus J is BVHK as a sum of BVHK functions (see Proposition 11 [Owen, 2005]). □

When analyzing the extreme value behaviour of bivariate data, information on the pairwise extremal coefficient is crucial wheter extremes are independent or not. However, outside this framework, the sole knowledge of this coefficient is only but a partial information about the dependence between extremes. To overcome this issue, the λ-madogram introduced in [Naveau et al., 2009] is of prime interest as this quantity capture the whole bivariate dependence between two extreme value random variables X and Y with margins F and G respectively. We recall the definition below

$$\nu(\lambda) = \frac{1}{2}\mathbb{E}\left[\left|\{F(X)\}^{\frac{1}{\lambda}} - \{G(Y)\}^{\frac{1}{1-\lambda}}\right|\right], \quad \lambda \in (0,1). \tag{6}$$

We refer to [Marcon et al., 2017] for basic properties of this quantity and its nonparametrical estimation. To obtain a more reliable picture how the λ-madogram evolves over λ, we consider estimators of the integrated madogram

$$I\nu = \int_{[0,1]} \nu(\lambda) d\lambda.$$

Lemma below state the integrated madogram verifies symmetry and triangular inequality.

Lemma 1. *Let (X, Y) be a bivariate random vector with extreme value copula C. The following dissimilarity verifies symmetry and triangular inequality.*

$$IM(X, Y) = I\nu + 1 - \ln(2). \tag{7}$$

Proof We first show where does the link between IM and extreme value theory stems. Using Proposition 1 of [Marcon et al., 2017], we know that :

$$I\nu = \int_{[0,1]} \frac{A(\lambda)}{1 + A(\lambda)} d\lambda - \frac{1}{2} \left[\int_{[0,1]} \frac{\lambda}{1+\lambda} + \int_{[0,1]} \frac{1-\lambda}{1+1-\lambda} \right],$$

where the bracket term is equal to $2(1 - \ln(2))$. Thus

$$IM(X, Y) = \int_{[0,1]} \frac{A(\lambda)}{1 + A(\lambda)} d\lambda. \tag{8}$$

Using classical bounds of the Pickands dependence function, we obtain that

$$\frac{1}{2} \geq IM(X, Y) \geq 1 + 2\ln(3/4),$$

where the upper bound (*resp.* lower bound) is achieved if and only if X and Y are asymptotically independent (*resp.* asymptotically comonotone).

We now have

$$IM(X, Y) - (1 - \ln(2)) = \frac{1}{2} \int_{[0,1]} \mathbb{E}\left[\left| \{F(X)\}^{\frac{1}{\lambda}} - \{G(Y)\}^{\frac{1}{1-\lambda}} \right| \right] d\lambda.$$

Splitting the segment $[0, 1]$ in two parts of same lengths gives

$$\frac{1}{2} \int_{[0,1/2]} \mathbb{E}\left[\left| \{F(X)\}^{\frac{1}{\lambda}} - \{G(Y)\}^{\frac{1}{1-\lambda}} \right| \right] d\lambda + \frac{1}{2} \int_{[1/2,1]} \mathbb{E}\left[\left| \{F(X)\}^{\frac{1}{\lambda}} - \{G(Y)\}^{\frac{1}{1-\lambda}} \right| \right] d\lambda.$$

A simple change of variable for each integral leads to

$$\frac{1}{2} \int_{[1/2,1]} \mathbb{E}\left[\left| \{F(X)\}^{\frac{1}{1-\mu}} - \{G(Y)\}^{\frac{1}{\mu}} \right| \right] d\mu + \frac{1}{2} \int_{[0,1/2]} \mathbb{E}\left[\left| \{F(X)\}^{\frac{1}{1-\mu}} - \{G(Y)\}^{\frac{1}{\mu}} \right| \right] d\mu.$$

We thus obtain

$$IM(X,Y) - (1 - \ln(2)) = IM(Y,X) - (1 - \ln(2)).$$

And the symmetry follows. For the triangular inequality, let us consider Z an extreme value random variable with law H. We have

$$IM(X,Z) = \frac{1}{2}\int_{[0,1]} \mathbb{E}\left[\left|\{F(X)\}^{\frac{1}{\lambda}} \pm \{G(Y)\}^{\frac{1}{1-\lambda}} \pm \{G(Y)\}^{\frac{1}{\lambda}} - \{H(Z)\}^{\frac{1}{1-\lambda}}\right|\right] d\lambda + (1 - \ln(2)).$$

Notice that

$$\frac{1}{2}\int_{[0,1]} \mathbb{E}\left[\left|\{G(Y)\}^{\frac{1}{1-\lambda}} - \{G(Y)\}^{\frac{1}{\lambda}}\right|\right] d\lambda = 2\ln(3/2) - \ln(2) < 1 - \ln(2).$$

Using triangle inequality, we have

$$\begin{aligned} IM(X,Z) &\leq \frac{1}{2}\int_{[0,1]} \mathbb{E}\left[\left|\{F(X)\}^{\frac{1}{\lambda}} - \{G(Y)\}^{\frac{1}{1-\lambda}}\right|\right] d\lambda + 1 - \ln(2) \\ &+ \frac{1}{2}\int_{[0,1]} \mathbb{E}\left[\left|\{G(Y)\}^{\frac{1}{\lambda}} - \{H(Z)\}^{\frac{1}{1-\lambda}}\right|\right] d\lambda + 1 - \ln(2) \\ &= IM(X,Y) + IM(Y,Z). \end{aligned}$$

□

Remark 2. *Bounds given in Equation (8) gives an intuition of k-means clustering will operate with this dissimilarity. More X and Y are asymptotically dependent and more the dissimilarity IM is close to its lower bound. Thus, k-means procedure will tend to cluster points which extremes are similar in behavior.*

Using this dissimilarity measures, the k-means clustering procedure thus states as :

$$W(A,C) = \int_{[0,1]^d} \min_{\mathbf{a}\in A} J(\mathbf{u},\mathbf{a}) dC(\mathbf{u}),$$

with $J : [0,1]^d \times [0,1]^d \to [0,1]$ given by

$$J(\mathbf{u},\mathbf{v}) = \frac{1}{2}\int_{[0,1]} \left|\mathbf{u}^{\frac{1}{\lambda}} - \mathbf{v}^{\frac{1}{1-\lambda}}\right|_1 d\lambda.$$

Proceeding as behind, set $\mathbf{v} \in [0,1]^d$, $\forall j \in \{1,\ldots,d\}$, $u_j^\lambda - v_j^{1-\lambda}$ is BVHK in $[0,1]^d$ so its absolute value is also BVHK for every $\lambda \in (0,1)$. Using that $\int_{[0,1]} d\lambda = 1$, we have that $\int_{[0,1]} |u_j - v_j| d\lambda$ is BVHK. So is J as a sum of BVHK functions. Thus Theorem 1 applies and the k-means procedure with J as a dissimilarity measure is consistent.

References

[Bador et al., 2015] Bador, M., Naveau, P., Gilleland, E., Castellà, M., and Arivelo, T. (2015). Spatial clustering of summer temperature maxima from the cnrm-cm5 climate model ensembles & e-obs over europe. *Weather and Climate Extremes*, 26.

[Bernard et al., 2013] Bernard, E., Naveau, P., Vrac, M., and Mestre, O. (2013). Clustering of maxima: Spatial dependencies among heavy rainfall in france. *Journal of Climate*, 26(20):7929 – 7937.

[Bunea et al., 2020] Bunea, F., Giraud, C., Luo, X., Royer, M., and Verzelen, N. (2020). Model assisted variable clustering: Minimax-optimal recovery and algorithms. *The Annals of Statistics*, 48(1):111 – 137.

[Fermanian, 1998] Fermanian, J.-D. (1998). *Contributions à l'Analyse Nonparamétrique des Fonctions de Hasard sur Données Multivariées et Censurées*. PhD thesis, Univ. Paris VI.

[Fermanian et al., 2004] Fermanian, J.-D., Radulovic, D., and Wegkamp, M. (2004). Weak convergence of empirical copula processes. *Bernoulli*, 10(5):847 – 860.

[Janßen and Wan, 2020] Janßen, A. and Wan, P. (2020). k-means clustering of extremes. *Electronic Journal of Statistics*, 14(1):1211 – 1233.

[Marcon et al., 2017] Marcon, G., Padoan, S., Naveau, P., Muliere, P., and Segers, J. (2017). Multivariate nonparametric estimation of the pickands dependence function using bernstein polynomials. *Journal of statistical planning and inference*, 183:1–17.

[Naveau et al., 2009] Naveau, P., Guillou, A., Cooley, D., and Diebolt, J. (2009). Modelling pairwise dependence of maxima in space. *Biometrika*, 96(1):1–17.

[Owen, 2005] Owen, A. B. (2005). Multidimensional variation for quasi-monte carlo. In *Contemporary Multivariate Analysis And Design Of Experiments: In Celebration of Professor Kai-Tai Fang's 65th Birthday*, pages 49–74. World Scientific.

[Pollard, 1981] Pollard, D. (1981). Strong Consistency of K-Means Clustering. *The Annals of Statistics*, 9(1):135 – 140.

[Radulović et al., 2017] Radulović, D., Wegkamp, M., and Zhao, Y. (2017). Weak convergence of empirical copula processes indexed by functions. *Bernoulli*, 23(4B):3346 – 3384.

[Saunders et al., 2021] Saunders, K., Stephenson, A., and Karoly, D. (2021). A regionalisation approach for rainfall based on extremal dependence. *Extremes*, 24(2):215–240.

[van der Vaart et al., 1996] van der Vaart, A., van der Vaart, A., van der Vaart, A., and Wellner, J. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer.