The k-means clustering procedure is a way to identify distinct groups within a population. 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.

In the following, let $d : \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty)$ be a distance function or, more generally, a dissimilarity function in $\mathbb{R}^d$. For a probability measure $\mathbb{P}$ on $\mathcal{B}(\mathbb{R}^d)$ and a set $A = \{a_1, \ldots, a_k\}$, $a_i \in \mathbb{R}^d$ for $i = 1, \ldots, k$ and $k \in \mathbb{N}$, one can introduce the averaged distance from any observation to the closest element of A as

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

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 as the integrand and C_n the empirical copula process. 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 for Lebesgue-Stieltjes integral is of prime interest (see 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 $V_{HK}(g) < \infty$ (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 $V_{HK}(J(\cdot, \boldsymbol{v})) < \infty$. *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, 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 of [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 $V_{HK}(J(\cdot, \mathbf{v})) < \infty$, we obtain that $V_{HK}(\min_{\mathbf{a} \in A} J(\cdot, \mathbf{a})) < \infty$ also. So, using the same notation as in [Janßen and Wan, 2020], we have using Equation (4)

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

Hence the statement. □

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, $\forall j \in \{1, \ldots, d\}$, $u_j - v_j$ is BVHK on $[0,1]^d$ since it depends only in one variable and is monotone in this variable. Finally, the absolute value is also BVHK on $[0,1]^d$. Thus $J(\cdot, \mathbf{v})$ is BVHK as a sum of BVHK functions (see Proposition 11 [Owen, 2005]). □

References

[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.

[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.

[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.