# On the Consistency of Metric and Non-Metric $K$-medoids

**Ery Arias-Castro**
University of California, San Diego

**He Jiang**
University of California, San Diego

## Abstract

We establish the consistency of $K$-medoids in the context of metric spaces. We start by proving that $K$-medoids is asymptotically equivalent to $K$-means restricted to the support of the underlying distribution under general conditions, including a wide selection of loss functions. This asymptotic equivalence, in turn, enables us to apply the work of Pärna (1986) on the consistency of $K$-means. This general approach applies also to non-metric settings where only an ordering of the dissimilarities is available. We consider two types of ordinal information: one where all quadruple comparisons are available; and one where only triple comparisons are available. We provide some numerical experiments to illustrate our theory.

## 1 INTRODUCTION

Cluster analysis is widely regarded as one of the most important tasks in unsupervised data analysis (Jain et al., 1999; Kaufman and Rousseeuw, 2009). In this paper, we consider several center based clustering methods. Specifically, we show the asymptotic equivalence of $K$-means and $K$-medoids, and use this equivalence to prove the consistency of $K$-medoids in metric and non-metric (i.e., ordinal) settings.

### 1.1 $K$-means and $K$-medoids

The problem of $K$-means can be traced back to the 1960's to early work of MacQueen (1967). As the problem is computationally difficult in higher dimensions or when the number of clusters is large, it is instead most often approached via iterative methods such as Lloyd's algorithm (Lloyd, 1982). Leaving these computational challenges aside, assuming the problem is solved exactly, the consistency of $K$-means as a method has been thoroughly addressed in the literature. Early in this line of work, Pollard (1981) established the consistency of $K$-means in Euclidean spaces. Pärna (1986) extended the result to separable metric spaces, while Pärna (1988, 1990, 1992) examined the particular situation of Hilbert and Banach spaces, where the existence of an optimal solution had been considered by Herrndorf (1983) and Cuesta and Matrán (1988). For more recent results on the consistency of variants of $K$-means, see for example (Gallegos and Ritter, 2013; Terada, 2014; Georgogiannis, 2016; Chakraborty et al., 2020).

The problem of $K$-medoids dates back to the 1980's to work of Kaufman and Rousseeuw (1987), who in the process proposed the Partition Around Medoids (PAM) iterative algorithm. Van Der Laan et al. (2003) discovered that the original PAM has problem with recognizing rather small clusters, and defined a new version of PAM based on maximizing average silhouette, as defined by Kaufman and Rousseeuw (1990). Later Park and Jun (2009) proposed a computationally simpler version of PAM akin to Lloyd's algorithm for $K$-means. See (Kaufman and Rousseeuw, 2009, Ch 2). In a setting where the goal is the clustering of data sequences, Wang et al. (2019) established an exponential consistency result for $K$-medoids itself (when solved exactly). To the best of our knowledge, however, the consistency of $K$-medoids in the more standard setting of clustering points in a metric space has not been previously established.

*We establish the consistency of $K$-medoids by first showing that $K$-medoids is asymptotically equivalent to $K$-means restricted to the support of the underlying distribution, and then leveraging the work of Pärna (1986) on the consistency of $K$-means in metric spaces.*

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. 

### 1.2 Ordinal $K$-medoids

Beyond the more standard setting where the distances are available to us, we also consider ordinal settings where only an ordering of the distances is available. Even when the dissimilarities are available, turning them into ranks, and thus only working with the underlying ordinal information, can be attractive in situations where the numerical value of the dissimilarities has little meaning besides providing an ordering. This is the case, for example, in psychological experiments where human subjects are tasked with rating some items in order of preference. Working with ranks also has the advantage of added robustness to outliers.

Statisticians and other data scientists have dealt with ordinal information for decades. Without going too far afield into rank-based inference (Hájek and Sidák, 1967) or ranking models (Bradley and Terry, 1952), there is non-metric scaling, aka ordinal embedding, which is the problem of embedding a set of items based on an ordering of their pairwise dissimilarities, with pioneering work in the 1960's by Shepard (1962a,b) and Kruskal (1964). The consistency of ordinal embedding — by which we mean any solution to the problem assuming one exists — was already considered by Shepard (1966), and more thoroughly addressed only recently by Kleindessner and Luxburg (2014) and Arias-Castro (2017).

Even closer to our situation, in the area of clustering, we know that hierarchical clustering with either single or complete linkage (or the less popular median linkage) only use the ordinal information, as can be seen from the fact that the output grouping remains the same if the dissimilarities are transformed by the application of a monotonically increasing function. The well-known clustering method DBSCAN of Ester et al. (1996) can be seen as a robust variant of single linkage, in its nearest-neighbor formulation, only relies on ordinal information as well. On the other hand, hierarchical clustering with either average linkage or Ward's criterion does not have that property. The use of $K$-medoids in ordinal settings does not seem nearly as widespread. In fact, we could only find a few references where the idea was proposed, scattered across various fields such as computer vision (Zhu et al., 2011) and data mining (Zadegan et al., 2013). In the context of an application to the clustering of pictures of human faces, Zhu et al. (2011) proposed a rank order distance (ROD) based on a sum of individual ranks, acquired from triple comparisons, and then applied single linkage hierarchical clustering with this distance. They argued that this distance was more appropriate for their particular application than the more standard $L_1$ distance. In a followup work, Huang et al. (2020) proposed a kernel variant of ROD. With the intention of making the clustering result less sensitive to initialization and potential outliers, Zadegan et al. (2013) proposed the concept of hostility index based on a sum of ranks obtained from triple comparisons. Aside from these, Achtert et al. (2006) proposed a dissimilarity based on the distance to the $\ell$-th nearest neighbor, which can therefore be implemented based solely on ordinal information.

> *Besides putting ordinal $K$-medoids in the context of ordinal data, as we just did, we establish its consistency for two types of ordinal information: quadruple comparisons giving an overall ranking of all pairwise dissimilarities; and triple comparisons giving a ranking relative to each sample point.*

### 1.3 Setting and Content

We consider the problem of clustering some data points in a metric space into $k$ clusters, where $k$ is given. The metric space is denoted $(\mathcal{X}, \mathsf{d})$ and assumed to be a locally compact Polish space. The sample is denoted $x_1, \dots, x_n$ and assumed to have been drawn from a Borel probability measure $Q$ assumed to have bounded support[1] containing at least $k$ points. We will let $Q_n$ denote the empirical distribution, namely, $Q_n(B) := \frac{1}{n}\sum_{i=1}^n 1_{\{x_i \in B\}}$ for any set $B \subset \mathcal{X}$. For two sets $A, B \subset \mathcal{X}$, define

$$\mathsf{H}(A|B) := \sup_{a \in A} \inf_{b \in B} \mathsf{d}(a, b), \tag{1}$$

so that the Hausdorff distance between $A$ and $B$ is $\max\{\mathsf{H}(A|B), \mathsf{H}(B|A)\}$.

The organization of the paper will be as follows. In Section 2, we prove the asymptotic equivalence of $K$-means and $K$-medoids, and deduce from that the consistency of $K$-medoids in the metric setting. In Section 3, we consider two ordinal settings, based on quadruple and triple comparisons respectively, and establish the consistency of $K$-medoids in each case using the equivalence result from Section 2. We provide numerical experiments along the way to illustrate our theoretical results. Our work is greatly inspired by that of Pärna (1986), and we will refer to his work often.

**Remark 1.** *We want to mention that all our results apply when $\mathcal{X}$ is a finite dimensional Banach space and $Q$ has a density with respect to the Lebesgue measure which is bounded and has compact support.*

[1] This is for convenience. See (Pärna, 1986).

## 2 EQUIVALENCE OF $K$-MEANS AND $K$-MEDOIDS, AND THE CONSISTENCY OF $K$-MEDOIDS

For a $k$-tuple $A \subset \mathcal{X}$, consider the risk

$$L(A, Q) = \int_{\mathcal{X}} \min_{a \in A} \phi(\mathsf{d}(x, a)) dQ(x), \tag{2}$$

where $\phi : [0, \infty) \to [0, \infty)$ is a loss function assumed to be non-decreasing, continuous, and such that $\phi(d) = 0$ if and only if $d = 0$ — all these assumptions being rather standard. By $K$-means we mean the result of the following optimization problem:

$$\text{minimize } L(A, Q_n) \quad \text{over } A \subset \mathcal{X},\ |A| = k. \tag{3}$$

And by $K$-medoids we mean the same optimization problem but restricted to $k$-tuples made of sample points:

$$\text{minimize } L(A, Q_n) \quad \text{over } A \subset \mathcal{X}_n,\ |A| = k, \tag{4}$$

where $\mathcal{X}_n := \{x_1, \dots, x_n\}$. Note that

$$L(A, Q_n) = \frac{1}{n} \sum_{i=1}^{n} \min_{a \in A} \phi(\mathsf{d}(x_i, a)). \tag{5}$$

It is well-known that, as formulated, (3) and (4) can behave quite differently. Take for example the case of the real line with $Q$ the uniform distribution on $[-2, -1] \cup [1, 2]$. When $k = 1$, in the large-sample limit, the origin is the unique solution to $K$-means problem, while $-1$ and $1$ are the solutions to the $K$-medoids problem. Instead, we consider the following restricted form of $K$-means:

$$\text{minimize } L(A, Q_n) \quad \text{over } A \subset \operatorname{supp}(Q),\ |A| = k, \tag{6}$$

where $\operatorname{supp}(Q)$ denotes the support of $Q$. The analyst cannot consider this problem when the support of $Q$ is unknown, which is typically the case. But this optimization problem is only used as a device to analyze the asymptotic behavior of $K$-medoids.

**Theorem 1.** *In the present context, $K$-medoids (4) is asymptotically equivalent to $K$-means (6), which in turn is asymptotically equivalent to population version of the same problem, namely*

$$\textit{minimize } L(A, Q) \quad \textit{over } A \subset \operatorname{supp}(Q),\ |A| = k. \tag{7}$$

*We conclude that, if $A_n^*$ is a solution to (4), then in probability,*

$$L(A_n^*, Q) \xrightarrow{n \to \infty} \min_{|A| = k} L(A, Q). \tag{8}$$

**Remark 2.** *As discussed in (Cuesta and Matrán, 1988; Pärna, 1990, 1992), a $K$-means problem may not have a solution. In our situation, however, we are assuming that the space is a locally compact Polish space, and a solution can be shown to exist by a simple compactness argument together with our assumptions on $\phi$ (and the fact that the distance function is always continuous in any metric space it equips). This applies to (3), (6) and (7).*

*Proof.* Since everything happens within the support of $Q$, we may assume without loss of generality that $Q$ is supported on the entire space, meaning that $\operatorname{supp}(Q) = \mathcal{X}$. And since we assume $\operatorname{supp}(Q)$ to be bounded, we are effectively assuming that $\mathcal{X}$ is bounded, and therefore compact since it is assumed to be locally compact.

The asymptotic equivalence of (6) and (7) is the consistency result of Pärna (1986). It can be deduced easily from the arguments we present below, which themselves are by-and-large adapted from (Pärna, 1986). So all we are left to do is prove that (4) is asymptotically equivalent to (6). To be sure, by this we mean that, if $A_n^*$ is a solution to the former and $A_n$ a solution to the latter, then

$$\left| L(A_n^*, Q_n) - L(A_n, Q_n) \right| \xrightarrow{n \to \infty} 0, \tag{9}$$

in probability. Because by definition $L(A_n^*, Q_n) \geq L(A_n, Q_n)$, all we need to show is that

$$\limsup_{n \to \infty} L(A_n^*, Q_n) - L(A_n, Q_n) \leq 0. \tag{10}$$

The remaining of the proof consists of three steps. We first show in Lemma 2 below that $L(A, Q_n) \to L(A, Q)$ as $n \to \infty$, uniformly over $A$. We then show in Lemma 4 further down that $A \mapsto L(A, Q)$ is uniformly continuous. The last step consists in using these results in conjunction with the 'squeeze theorem'.

By the uniform convergence established in Lemma 2, we have

$$\lim_{n \to \infty} |L(A_n^*, Q_n) - L(A_n^*, Q)| = 0, \tag{11}$$

as well as

$$\lim_{n \to \infty} |L(A_n, Q_n) - L(A_n, Q)| = 0. \tag{12}$$

Therefore, all we need to show is that

$$\limsup_{n \to \infty} L(A_n^*, Q) - L(A_n, Q) \leq 0. \tag{13}$$

For every point in $A_n$ find the closest sample point, and gather all these in $B_n^*$. Note that by Lemma 1,

$$h_n := \mathsf{H}(A_n | B_n) = \max_{a \in A_n} \min_{b \in B_n^*} \mathsf{d}(a, b) \xrightarrow{n \to \infty} 0, \tag{14}$$

in probability. Hence, by Lemma 4, we have

$$\limsup_{n\to\infty} L(B_n^*, Q) - L(A_n, Q) \le \limsup_{n\to\infty} \omega(h_n) = 0. \quad (15)$$

With the fact that $L(A_n^*, Q_n) \le L(B_n^*, Q_n)$ by definition of $A_n^*$, together with the uniform convergence also giving

$$\lim_{n\to\infty} |L(B_n^*, Q_n) - L(B_n^*, Q)| = 0, \quad (16)$$

we thus conclude that (13) holds. □

**Lemma 1.** *Assuming that $\mathcal{X}$ is compact and that* $\mathrm{supp}(Q) = \mathcal{X}$*, in probability,*

$$\mathsf{H}(\mathcal{X}|\mathcal{X}_n) = \sup_{x\in\mathcal{X}} \min_{i\in[n]} \mathsf{d}(x, x_i) \to 0, \quad n \to \infty. \quad (17)$$

*Proof.* The arguments are standard and follow from the definition of $\mathrm{supp}(Q)$. Indeed, $\mathrm{supp}(Q)$ is the complement of the largest open set $D$ in $\mathcal{X}$ such that $Q(D) = 0$. Since $\mathrm{supp}(Q) = \mathcal{X}$ by assumption, it must be that $Q(B(x,r)) > 0$ for all $x \in \mathcal{X}$ and all $r > 0$, where $B(x,r)$ is defined as the closed ball centered at $x$ with radius $r$. Fix $r > 0$. Because $\mathcal{X}$ is compact there is $y_1, \dots, y_m \in \mathcal{X}$ such that $\mathcal{X} = \bigcup_j B(y_j, r)$. By the triangle inequality,

$$\begin{aligned}
&\mathsf{H}(\mathcal{X}|\mathcal{X}_n) \ge 2r \\
&\Leftrightarrow \ \exists x : \min_i \mathsf{d}(x, x_i) \ge 2r \\
&\Rightarrow \ \exists j : \min_i \mathsf{d}(y_j, x_i) \ge r,
\end{aligned}$$

and by the union bound, this implies that

$$\begin{aligned}
&\mathbb{P}(\mathsf{H}(\mathcal{X}|\mathcal{X}_n) \ge 2r) \\
&\le \sum_j \mathbb{P}(\min_i \mathsf{d}(y_j, x_i) \ge r) \\
&= \sum_j (1 - Q(B(y_j, r))^n \\
&\le m(1 - \min_j Q(B(y_j, r))^n \\
&\to 0, \quad n \to \infty.
\end{aligned}$$

Since $r > 0$ is arbitrary, the claim is established. □

### 2.1 Uniform Convergence Lemma

**Lemma 2.** *Assuming that $\mathcal{X}$ is compact, we have, in probability,*

$$\limsup_{n\to\infty} \sup_{|A|\le k} |L(A, Q_n) - L(A, Q)| = 0. \quad (18)$$

The rest of this subsection is devoted to proving this lemma. It is enough to prove the variant where $|A| \le k$ is replaced by $|A| = k$. The proof is very similar to the proof of (Pärna, 1986, Lemma 1), with some differences. We provide a full proof for the sake of completeness.

Note that, like $L(A, Q)$, $L(A, Q_n)$ can be expressed as an integral:

$$L(A, Q_n) = \int_{\mathcal{X}} \min_{a\in A} \phi(\mathsf{d}(x, a)) dQ_n(x). \quad (19)$$

To each finite set $A$, we associate the following function

$$f_A(x) = \min_{a\in A} \phi(\mathsf{d}(x, a)). \quad (20)$$

Define the following class of functions

$$\mathcal{F} = \{f_A : A \subset \mathcal{X}, |A| = k\}. \quad (21)$$

**Lemma 3** (Theorem 3.2 of (Rao, 1962))**.** *Let $\mathcal{F}$ be a family of continuous functions on a separable metric space $\mathcal{X}$ which is equicontinuous and admits a continuous envelope (there is $g$ continuous such that $|f(x)| \le g(x)$ for all $f \in \mathcal{F}$). In this context, suppose that $(\mu_n)$ is a sequence of measures on $\mathcal{X}$ converging weakly to $\mu$, another measure on $\mathcal{X}$ with $\int g d\mu_n \to \int g d\mu < \infty$. Then we have:*

$$\limsup_{n\to\infty} \sup_{f\in\mathcal{F}} \left| \int f d\mu_n - \int f d\mu \right| = 0. \quad (22)$$

We apply this result with $\mu_n = Q_n$ and $\mu = Q$, for which the weak convergence is satisfied with probability 1 (Varadarajan, 1958). The existence of an envelope function $g$ satisfying the requirements for the function class of interest, $\mathcal{F}$ above, is here immediate since for any $A$,

$$0 \le f_A(x) \le \phi(\mathrm{diam}(\mathcal{X})) < \infty, \quad (23)$$

where $\mathrm{diam}(\mathcal{X})$ denotes the diameter of $\mathcal{X}$, so that we may take $g \equiv \phi(\mathrm{diam}(\mathcal{X}))$. It only remains to show $\mathcal{F}$ is equicontinuous. This amounts to showing that, for any $y_0 \in \mathcal{X}$ and any $\epsilon > 0$, there exists a $\delta > 0$, such that $|f_A(y_0) - f_A(y)| < \epsilon$ for any $k$-tuple $A$ and any $y \in B(y_0, \varepsilon)$.

For the given $y_0$ and $y$, we denote $a(y_0)$ and $a(y)$ closest points in $A$ to them so that

$$\min_{a\in A} \mathsf{d}(y_0, a) - \min_{a\in A} \mathsf{d}(y, a) = \mathsf{d}(y_0, a(y_0)) - \mathsf{d}(y, a(y)).$$

By definition and the triangle inequality,

$$\begin{aligned}
&\mathsf{d}(y_0, a(y_0)) - \mathsf{d}(y, a(y)) \\
&\le \mathsf{d}(y_0, a(y)) - \mathsf{d}(y, a(y)) \\
&\le \mathsf{d}(y_0, y),
\end{aligned}$$

and similarly,

$$\begin{aligned}
&\mathsf{d}(y_0, a(y_0)) - \mathsf{d}(y, a(y)) \\
&\ge \mathsf{d}(y_0, a(y_0)) - \mathsf{d}(y, a(y_0)) \\
&\ge -\mathsf{d}(y_0, y).
\end{aligned}$$

We thus deduce that

$$\left|\min_{a\in A} \mathsf{d}(y_0,a) - \min_{a\in A} \mathsf{d}(y,a)\right| \le \mathsf{d}(y_0,y). \tag{24}$$

Since $\phi$ is assumed to be continuous, it is uniformly continuous on $[0,\mathrm{diam}(\mathcal{X})]$. Let $\omega$ denote its modulus of continuity on that interval so that

$$\left|\phi(d)-\phi(d')\right| \le \omega(|d-d'|), \quad \forall d,d' \in [0,\mathrm{diam}(\mathcal{X})].$$

We then have

$$\begin{aligned}
&|f_A(y_0) - f_A(y)| && (25)\\
&= \left|\min_{a\in A} \phi(\mathsf{d}(y_0,a)) - \min_{a\in A} \phi(\mathsf{d}(y,a))\right| && (26)\\
&= \left|\phi\Big(\min_{a\in A} \mathsf{d}(y_0,a)\Big) - \phi\Big(\min_{a\in A} \mathsf{d}(y,a)\Big)\right| && (27)\\
&\le \omega(\mathsf{d}(y_0,y)), && (28)
\end{aligned}$$

using the monotonicity of $\phi$ along the way. We have proved that $\mathcal{F}$ is indeed equicontinuous. Therefore the proof of Lemma 2 is complete.

### 2.2 Uniform Continuity Lemma

**Lemma 4.** *For any two sets $A, B \subset \mathcal{X}$, we have*

$$L(B,Q) \le L(A,Q) + \omega(\mathsf{H}(A|B)), \tag{29}$$

*where $\omega$ is the modulus of continuity of $\phi$ on $[0,\mathrm{diam}(\mathcal{X})]$.*

The rest of this subsection is devoted to proving this lemma. Fix two sets $A, B \subset \mathcal{X}$, and let $h := \mathsf{H}(A|B)$. For any $a \in A$, define $b_a$ as the closest point in $B$ to $a$. Notice that by definition:

$$\mathsf{d}(a,b_a) \le h, \tag{30}$$

and thus with the triangle inequality, for any point $x$ we have:

$$\mathsf{d}(x,a) \ge \mathsf{d}(x,b_a) - \mathsf{d}(a,b_a) \ge \mathsf{d}(x,b_a) - h. \tag{31}$$

Taking minimums we get:

$$\min_{a\in A} \mathsf{d}(x,a) \ge \min_{a\in A} \mathsf{d}(x,b_a) - h \ge \min_{b\in B} \mathsf{d}(x,b) - h. \tag{32}$$

Using the fact that $\phi$ is non-decreasing, we then have:

$$\begin{aligned}
&\min_{a\in A} \phi(\mathsf{d}(x,a)) - \min_{b\in B} \phi(\mathsf{d}(x,b)) && (33)\\
&= \phi\Big(\min_{a\in A} \mathsf{d}(x,a)\Big) - \phi\Big(\min_{b\in B} \mathsf{d}(x,b)\Big) && (34)\\
&\ge -\omega(h). && (35)
\end{aligned}$$

Therefore, by integrating with respect to $Q$, we obtain:

$$\begin{aligned}
&L(A,Q) - L(B,Q)\\
&= \int \min_{a\in A} \phi(\mathsf{d}(x,a))dQ(x) - \int \min_{b\in B} \phi(\mathsf{d}(x,b))dQ(x)\\
&= \int \Big[\min_{a\in A} \phi(\mathsf{d}(x,a)) - \min_{b\in B} \phi(\mathsf{d}(x,b))\Big]dQ(x)\\
&\ge -\omega(h).
\end{aligned}$$

### 2.3 Simulations

We report on a simple experiment illustrating the asymptotic equivalence established in Theorem 1. To keep a balance between the necessity to probe an asymptotic result ($n$ large enough) and computational feasibility ($n$ not too large), we choose to work with a sample of size $n = 2000$. We generate data from two equally weighted Gaussian distributions in $R^2$, centered at $(-0.5, 0)$ and $(0.5, 0)$, each with covariance $0.05 \times I_2$, and apply Lloyd's $K$-means algorithm (Lloyd, 1982) and the PAM algorithm (Kaufman and Rousseeuw, 1987) . Each setting is repeated 50 times. The result of this experiment is summarized in Table 1. As can be seen from this experiment, although varying according to different metrics and loss functions, the performance of $K$-means and $K$-medoids are indeed very similar.

Table 1: Mean values and standard deviations of the Average Center Error (error) and the Adjusted Rand Index (ARI) of $K$-means and $K$-medoids for various metrics and loss functions.

| | | $K$-means | $K$-medoids |
|---|---|---|---|
| $L_1$ | error [$\times 10^{-2}$] | 1.2 (0.4) | 1.8 (0.7) |
| | ARI | 0.780 (0.017) | 0.778 (0.016) |
| $\sqrt{L_2}$ | error [$\times 10^{-2}$] | 8.9 (1.7) | 11.7 (2.5) |
| | ARI | 0.784 (0.020) | 0.782 (0.022) |
| $L_2$ | error [$\times 10^{-3}$] | 9.4 (3.2) | 12.3 (4.3) |
| | ARI | 0.789 (0.017) | 0.789 (0.017) |
| ${L_2}^2$ | error [$\times 10^{-4}$] | 1.1 (0.9) | 2.3 (1.7) |
| | ARI | 0.785 (0.016) | 0.785 (0.016) |
| $L_\infty$ | error [$\times 10^{-3}$] | 8.9 (3.4) | 12.0 (4.2) |
| | ARI | 0.785 (0.021) | 0.783 (0.020) |

## 3 ORDINAL $K$-MEDOIDS

In this section we consider the problem of clustering with only an ordering of the dissimilarities. We consider two such orderings, one based on quadruple comparisons and another based on triple comparisons. We apply the results from Section 2 to show that, in both cases, $K$-medoids is consistent.

### 3.1 Quadruple Comparisons

First we consider a situation in which all quadruple comparisons of the form 'Is $\mathsf{d}(x_i, x_j)$ larger or smaller than $\mathsf{d}(x_l, x_m)$?' are available. Equivalently, this is a situation in which a complete ordering of the pairwise dissimilarities is available.

For $i \in [n]$, let $R_i(a)$ denote the rank of $\mathsf{d}(x_i, a)$ among $\{\mathsf{d}(x_l, x_m) : l < m\}$, and for a $k$-tuple $A$, define

$$S_{\text{rank}}(A, Q_n) = \frac{1}{n} \sum_{i=1}^{n} \min_{a \in A} \frac{R_i(a)}{\frac{n(n-1)}{2}}. \tag{36}$$

By ordinal $K$-medoids we mean the following optimization problem:

$$\text{minimize } S_{\text{rank}}(A, Q_n) \quad \text{over } A \subset \mathcal{X}_n,\ |A| = k. \tag{37}$$

This problem can be posed with the available information, and thus in principle can be solved. The equivalent restricted variant of ordinal $K$-means corresponds the following optimization problem:

$$\text{minimize } S_{\text{rank}}(A, Q_n) \quad \text{over } A \subset \text{supp}(Q),\ |A| = k. \tag{38}$$

As before, the latter is used as a bridge to show that the former is asymptotically equivalent to the population version of this $K$-means problem, which is given by

$$\text{minimize } S(A, Q) \quad \text{over } A \subset \text{supp}(Q),\ |A| = k, \tag{39}$$

where

$$S(A, Q) := \int \min_{a \in A} G(\mathsf{d}(x, a)) dQ(x), \tag{40}$$

with

$$G(t) := \mathbb{P}(\mathsf{d}(X, X') \le t), \tag{41}$$

$X, X'$ being independent with distribution $Q$.

Here is the missing link between $S_{\text{rank}}$ and $S$.

**Lemma 5.** *The following holds in probability:*

$$\limsup_{n \to \infty} \sup_{|A| \le k} |S_{\text{rank}}(A, Q_n) - S(A, Q_n)| = 0. \tag{42}$$

*Proof.* Let $\hat{G}_n$ denote the empirical distribution function of all the pairwise distances between sample points, meaning,

$$\hat{G}_n(t) := \frac{2}{n(n-1)} \sum_{l<m} 1_{\{\mathsf{d}(x_l, x_m) \le t\}}.$$

By the law of large numbers for $U$-statistics, in probability, $\hat{G}_n(t) \to G(t)$ as $n \to \infty$ for every fixed $t$. The Glivenko–Cantelli lemma does not quite apply as the pairwise distances are not an iid sample, but the two ingredients are there (Van Der Vaart, 1998): pointwise convergence as just stated, and the fact that $\hat{G}_N$ and $G$ are both distribution functions in that they both are non-decreasing from 0 to 1 on $[0, \infty)$. Hence,

$$\varepsilon_n := \sup_t |\hat{G}_n(t) - G(t)| \xrightarrow{n \to \infty} 0,$$

in probability. We then have:

$$\begin{aligned} R_i(a) &= \sum_{l<m} 1_{\{\mathsf{d}(x_l, x_m) \le \mathsf{d}(x_i, a)\}} \\ &= \frac{n(n-1)}{2} \hat{G}_n(\mathsf{d}(x_i, a)) \\ &= \frac{n(n-1)}{2} G(\mathsf{d}(x_i, a)) \pm \frac{n(n-1)}{2} \varepsilon_n, \end{aligned}$$

giving

$$\begin{aligned} S_{\text{rank}}(A, Q_n) &= \frac{1}{n} \sum_{i=1}^{n} \min_{a \in A} G(\mathsf{d}(x_i, a)) \pm \varepsilon_n \\ &= S(A, Q_n) \pm \varepsilon_n, \end{aligned}$$

for any finite $A$, which establishes the result. □

Establishing the consistency of ordinal $K$-medoids is now a straightforward consequence of Theorem 1. We need to assume that $G$ defined above is continuous, which is the case in the canonical situation of Remark 1.

**Theorem 2.** *In the present context, if $A_n^*$ is a solution to ordinal $K$-medoids in the form of (37), then in probability,*

$$S(A_n^*, Q) \xrightarrow{n \to \infty} \min_{|A| = k} S(A, Q). \tag{43}$$

*Proof.* By Lemma 5, we have that ordinal $K$-medoids (37) is asymptotically equivalent to the following problem:

$$\text{minimize } S(A, Q_n) \quad \text{over } A \subset \mathcal{X}_n,\ |A| = k. \tag{44}$$

But $S$ is exactly as $L$ in Section 2, with $G$ replacing $\phi$ there, and since $G$ satisfies the same properties assumed of $\phi$, Theorem 1 applies to yield the claim. □

### 3.2 Triple Comparisons

We turn to a situation in which only triple comparisons of the form 'Is $\mathsf{d}(x_i, x_j)$ larger or smaller than $\mathsf{d}(x_i, x_l)$?' are available. We do assume that all of these comparisons are on hand. Equivalently, this is a situation in which an ordering of the pairwise dissimilarities involving a particular point are available.

Hence, we work here with the ranks (re)defined as follows. For $i \in [n]$, let $R_i(a)$ denote the rank of $\mathsf{d}(x_i, a)$

among $\{\mathsf{d}(x_i, x_j) : j \neq i\}$, and for a $k$-tuple $A$, define

$$S_{\text{rank}}(A, Q_n) = \frac{1}{n}\sum_{i=1}^{n} \min_{a\in A} \frac{R_i(a)}{n-1}. \tag{45}$$

Ordinal $K$-medoids and (restricted) ordinal $K$-means are otherwise defined as before. The population equivalent to ordinal $K$-means is now given by

$$\text{minimize } S(A, Q) \quad \text{over } A \subset \text{supp}(Q),\ |A| = k, \tag{46}$$

where

$$S(A, Q) := \int \min_{a\in A} G^x(\mathsf{d}(x, a))dQ(x), \tag{47}$$

with

$$G^x(t) := \mathbb{P}(\mathsf{d}(x, X') \le t), \tag{48}$$

$X'$ having distribution $Q$.

**Lemma 6.** *The following holds in probability:*

$$\limsup_{n\to\infty} \sup_{|A|\le k} |S_{\text{rank}}(A, Q_n) - S(A, Q_n)| = 0. \tag{49}$$

*Proof.* Define

$$\hat{G}_{n,i}(t) := \frac{1}{n-1}\sum_{j\neq i} 1_{\{\mathsf{d}(x_i,x_j)\le t\}}.$$

By the Dvoretzky–Kiefer–Wolfowitz (DKW) inequality, for each $i$ and any $\varepsilon > 0$, we have:

$$\mathbb{P}\Big(\sup_t |\hat{G}_{n,i}(t) - G^{x_i}(t)| > \epsilon\Big) \le 2\exp(-2(n-1)\epsilon^2).$$

With this, and the union bound, we obtain:

$$\varepsilon_n := \max_i \sup_t |\hat{G}_{n,i}(t) - G^{x_i}(t)| \xrightarrow{n\to\infty} 0,$$

in probability. We then have:

$$\begin{aligned} R_i(a) &= \sum_{j\neq i} 1_{\{\mathsf{d}(x_i,x_j)\le \mathsf{d}(x_i,a)\}} \\ &= (n-1)\hat{G}_{n,i}(\mathsf{d}(x_i, a)) \\ &= (n-1)G^{x_i}(\mathsf{d}(x_i, a)) \pm (n-1)\varepsilon_n, \end{aligned}$$

giving

$$\begin{aligned} S_{\text{rank}}(A, Q_n) &= \frac{1}{n}\sum_{i=1}^{n} \min_{a\in A} G^{x_i}(\mathsf{d}(x_i, a)) \pm \varepsilon_n \\ &= S(A, Q_n) \pm \varepsilon_n, \end{aligned}$$

for any finite $A$, which establishes the result. □

The following is our consistency result for $K$-medoids based on triple comparisons. It is not an immediate consequence of Theorem 1, but the proof arguments are parallel. We need to make additional assumption that $(x, t) \mapsto Q(B(x, t))$ is continuous on $\mathcal{X} \times (0, \infty)$. This is the case in the canonical situation of Remark 1.

**Theorem 3.** *In the present context, if $A_n^*$ is a solution to ordinal $K$-medoids based on triple comparisons, then in probability,*

$$S(A_n^*, Q) \xrightarrow{n\to\infty} \min_{|A|=k} S(A, Q), \tag{50}$$

*now with $S$ defined as in (47).*

*Proof.* By Lemma 6, we have that ordinal $K$-medoids is asymptotically equivalent to the following problem:

$$\text{minimize } S(A, Q_n) \quad \text{over } A \subset \mathcal{X}_n,\ |A| = k. \tag{51}$$

But unlike the situation in Theorem 2, now $S$ is *not* exactly as $L$ in Section 2, complicating matters a little bit. Nevertheless, the proof arguments are parallel to those underlying Theorem 1. As we did there, we need to establish uniform convergence and uniform continuity. As before, we may assume without loss of generality that $\mathcal{X}$ is compact and that $\text{supp}(Q) = \mathcal{X}$. In that case, $(x, t) \mapsto Q(B(x, t))$ is uniformly continuous, and we let $\Omega$ denote its modulus of continuity so that

$$\big|Q(B(x, s)) - Q(B(y, t))\big| \le \Omega(\mathsf{d}(x, y), |s - t|),$$

for all $x, y \in \mathcal{X}$ and all $s, t > 0$.

For the uniform convergence, the proof of Lemma 2 proceeds as before until the very end where instead

$$\begin{aligned} &|f_A(y_0) - f_A(y)| && (52)\\ &= \Big|\min_{a\in A} G^{y_0}(\mathsf{d}(y_0, a)) - \min_{a\in A} G^{y}(\mathsf{d}(y, a))\Big| && (53)\\ &= |G^{y_0}(\mathsf{d}(y_0, A)) - G^{y}(\mathsf{d}(y, A))| && (54)\\ &\le \Omega(\mathsf{d}(y_0, y), |\mathsf{d}(y_0, A) - \mathsf{d}(y, A)|) && (55)\\ &\le \Omega(\mathsf{d}(y_0, y), \mathsf{d}(y_0, y)) && (56)\\ &\to 0, \quad \text{when } \mathsf{d}(y_0, y) \to 0. && (57) \end{aligned}$$

For the uniform continuity, the proof of Lemma 4 proceeds as before except that

$$\begin{aligned} &\min_{a\in A} G^x(\mathsf{d}(x, a)) - \min_{b\in B} G^x(\mathsf{d}(x, b)) && (58)\\ &= G^x(\mathsf{d}(x, A)) - G^x(\mathsf{d}(x, B)) && (59)\\ &\ge -\Omega(0, h), && (60) \end{aligned}$$

with $h := \mathsf{H}(A|B)$ as in that proof, so that the statement of that lemma continues to hold but with $\omega(t) := \Omega(0, t)$. □

### 3.3 Simulations

We again report on a numerical experiment showcasing the results derived in this section in the context of ordinal clustering. We chose to work with a sample of size $n = 750$. We generate data from three equally weighted

Table 2: Mean values and standard deviations of the Average Center Error (error) and the Adjusted Rand Index (ARI) for various metrics and loss functions for $K$-medoids based on triple-comparisons (TC), quadruple-comparisons (QC), and the actual distances (KM).

| | | TC | QC | KM |
|---|---|---|---|---|
| $L_1$ | error [$\times 10^{-2}$] | 4.4 (1.3) | 3.6 (1.3) | 3.7 (1.2) |
| | ARI | 0.924 (0.014) | 0.924 (0.014) | 0.925 (0.014) |
| $\sqrt{L_2}$ | error [$\times 10^{-1}$] | 1.8 (0.3) | 1.6 (0.2) | 1.7 (0.3) |
| | ARI | 0.928 (0.016) | 0.929 (0.015) | 0.929 (0.016) |
| $L_2$ | error [$\times 10^{-2}$] | 3.2 (0.8) | 2.7 (0.9) | 2.7 (0.9) |
| | ARI | 0.934 (0.016) | 0.933 (0.016) | 0.933 (0.016) |
| ${L_2}^2$ | error [$\times 10^{-3}$] | 1.4 (0.7) | 0.9 (0.5) | 0.8 (0.5) |
| | ARI | 0.931 (0.020) | 0.931 (0.019) | 0.930 (0.020) |
| $L_\infty$ | error [$\times 10^{-2}$] | 3.0 (1.1) | 2.6 (1.0) | 2.7 (1.0) |
| | ARI | 0.918 (0.017) | 0.917 (0.017) | 0.918 (0.017) |

Gaussian distributions in two dimensions, centered at $(-0.5, 0)$, $(0.5, 0)$ and $(0, \sqrt{3}/2)$, each with covariance $0.05 \times I_2$ , and apply the PAM algorithm with either the actual distances, the triple comparison ranks, or the quadruple comparison ranks. Each setting is repeated 50 times. The result of our experiment is summarized in Table 2. As can be seen from this experiment, $K$-medoids based on ordinal information performs nearly as well as $K$-medoids based on the full dissimilarity information.

# 4 DISCUSSION

In this paper, we have shown the asymptotic equivalence of $K$-means and $K$-medoids, and used this equivalence to prove the consistency of $K$-medoids in metric and non-metric situations.

## 4.1 Consistency of the Solution

Our consistency results are on the value of the optimization problem defining $K$-medoids in the various settings we considered. Specifically, we showed in each case that $T(A_n^*, Q) \to_{n\to\infty} \min_A T(A, Q)$, in probability, where $T$ is an appropriate criterion (either $L$ or one of the two variants of $S$) and $A_n^*$ is the solution to $K$-medoids. What about the behavior of the solution $A_n^*$ itself?

Here the situation is completely generic: if the solution to the population problem, namely $A_{\text{opt}} := \arg\min_A T(A, Q)$, is unique, then $A_n^* \to_{n\to\infty} A_{\text{opt}}$, again in probability. This is simply due to the fact that in our setting we can reduce the situation to when $\mathcal{X}$ is compact, and in all cases we considered $A \mapsto T(A, Q)$ is continuous.

## 4.2 Clustering After Embedding?

It might be possible to establish the consistency of ordinal $K$-medoids building on the consistency of ordinal embedding. This route appears unnecessarily sophisticated, however, in particular in light of a more straightforward approach that we built on the work of Pärna (1986). And from a computational standpoint, performing $K$-medoids in the ordinal setting has essentially the same complexity as in the regular (i.e., metric) setting, while methods for ordinal embedding tend to be much more demanding in computational resources.

## 4.3 A 'Bad' Variant of $K$-Medoids

In the setting where triple comparisons are available, instead of defining the ranks as we did, we could have worked with the following definition. For $i \in [n]$, let $R_i(a)$ denote the rank of $\mathsf{d}(x_i, a)$ among $\{\mathsf{d}(x_j, a) : j \in [n]\}$. Although the resulting method can be analyzed in very much the same way, it turns out to not be useful for the purpose of clustering. This is due to the fact that the corresponding optimization problem accepts a large range of solutions. To see this, consider the case $k = 1$. With the corresponding definition of $S_{\text{rank}}$, we have that

$$S_{\text{rank}}(a, Q_n) = \frac{1 + 2 + \cdots + n}{n(n-1)}, \tag{61}$$

for all $a \in \{x_1, ..., x_n\}$. And the problem persists for other values of $k$. For another example, consider clustering points distributed uniformly between $[-1, 1]$ into $k = 2$ clusters. It is clear that the correct population centers for $K$-means here are $\{-1/2, 1/2\}$. However, it can be seen that for any $1/2 \le c \le 1$, $A = \{-c, c\}$ also achieves the optimal population risk.

**Acknowledgements**

We are very grateful to Professor Kalev Pärna for sharing with us his papers, which we could not otherwise access. This work was partially supported by the National Science Foundation (DMS 1916071).

## References


Achtert, E., C. Böhm, and P. Kröger (2006). Deli-clu: boosting robustness, completeness, usability, and efficiency of hierarchical clustering by a closest pair ranking. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 119–128. Springer.

Arias-Castro, E. (2017). Some theory for ordinal embedding. *Bernoulli 23*(3), 1663–1693.

Bradley, R. A. and M. E. Terry (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika 39*(3/4), 324–345.

Chakraborty, S., D. Paul, S. Das, and J. Xu (2020). Entropy weighted power $K$-means clustering. In *International Conference on Artificial Intelligence and Statistics*, pp. 691–701. PMLR.

Cuesta, J. and C. Matrán (1988). The strong law of large numbers for $K$-means and best possible nets of Banach valued random variables. *Probability Theory and Related Fields 78*(4), 523–534.

Ester, M., H.-P. Kriegel, J. Sander, and X. Xu (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, Volume 96, pp. 226–231.

Gallegos, M. T. and G. Ritter (2013). Strong consistency of $K$-parameters clustering. *Journal of Multivariate Analysis 117*, 14–31.

Georgogiannis, A. (2016). Robust $K$-means: a theoretical revisit. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 2891–2899.

Hájek, J. and Z. Sidák (1967). Theory of rank tests.

Herrndorf, N. (1983). Approximation of vector-valued random variables by constants. *Journal of Approximation Theory 37*(2), 175–181.

Huang, T., S. Wang, and W. Zhu (2020). An adaptive kernelized rank-order distance for clustering non-spherical data with high noise. *International Journal of Machine Learning and Cybernetics*, 1–13.

Jain, A. K., M. N. Murty, and P. J. Flynn (1999). Data clustering: a review. *ACM Computing Surveys (CSUR) 31*(3), 264–323.

Kaufman, L. and P. Rousseeuw (1987). Clustering by means of medoids. In *Statistical Data Analysis Based on the $L_1$ Norm Conference, Neuchatel, 1987*, pp. 405–416.

Kaufman, L. and P. Rousseeuw (1990). Finding groups in data: An introduction to cluster analysis. *Hoboken NJ John Wiley & Sons Inc 725*.

Kaufman, L. and P. Rousseeuw (2009). *Finding Groups in Data: An Introduction to Cluster Analysis*, Volume 344. John Wiley & Sons.

Kleindessner, M. and U. Luxburg (2014). Uniqueness of ordinal embedding. In *Conference on Learning Theory*, pp. 40–67.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika 29*(1), 1–27.

Lloyd, S. (1982). Least squares quantization in pcm. *IEEE transactions on information theory 28*(2), 129–137.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1, pp. 281–297. Oakland, CA, USA.

Park, H.-S. and C.-H. Jun (2009). A simple and fast algorithm for $K$-medoids clustering. *Expert systems with applications 36*(2), 3336–3341.

Pärna, K. (1986). Strong consistency of $K$-means clustering criterion in separable metric spaces. *Tartu Riikl. Ul. Toimetised 733*, 86–96.

Pärna, K. (1988). On the stability of $K$-means clustering in metric spaces. *Tartu Riikl. Ul. Toimetised 798*, 19–36.

Pärna, K. (1990). On the existence and weak convergence of $K$-centres in Banach spaces. *Tartu Ulikooli Toimetised 893*, 17–287.

Pärna, K. (1992). Clustering in metric spaces: some existence and continuity results for $K$-centers. In *Analyzing and Modeling Data and Knowledge*, pp. 85–91. Springer.

Pollard, D. (1981). Strong consistency of $K$-means clustering. *The Annals of Statistics*, 135–140.

Rao, R. R. (1962). Relations between weak and uniform convergence of measures with applications. *The Annals of Mathematical Statistics*, 659–680.

Shepard, R. N. (1962a). The analysis of proximities: multidimensional scaling with an unknown distance function. i. *Psychometrika 27*(2), 125–140.

Shepard, R. N. (1962b). The analysis of proximities: Multidimensional scaling with an unknown distance function. ii. *Psychometrika 27*(3), 219–246.

Shepard, R. N. (1966). Metric structures in ordinal data. *Journal of Mathematical Psychology 3*(2), 287–315.

Terada, Y. (2014). Strong consistency of reduced $K$-means clustering. *Scandinavian Journal of Statistics 41*(4), 913–931.

Van Der Laan, M., K. Pollard, and J. Bryan (2003). A new partitioning around medoids algorithm. *Journal of Statistical Computation and Simulation 73*(8), 575–584.

Van Der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge University Press.

Varadarajan, V. S. (1958). On the convergence of sample probability distributions. *Sankhyā: The Indian Journal of Statistics (1933-1960) 19*(1/2), 23–26.

Wang, T., Q. Li, D. J. Bucci, Y. Liang, B. Chen, and P. K. Varshney (2019). $K$-medoids clustering of data sequences with composite distributions. *IEEE Transactions on Signal Processing 67*(8), 2093–2106.

Zadegan, S. M. R., M. Mirzaie, and F. Sadoughi (2013). Ranked $K$-medoids: A fast and accurate rank-based partitioning algorithm for clustering large datasets. *Knowledge-Based Systems 39*, 133–143.

Zhu, C., F. Wen, and J. Sun (2011). A rank-order distance based clustering algorithm for face tagging. In *CVPR 2011*, pp. 481–488. IEEE.