Journal of Intelligent Information Systems (2021) 57:509–530
https://doi.org/10.1007/s10844-021-00657-6

Issues in clustering algorithm consistency in fixed dimensional spaces. Some solutions for *k*-means

Mieczysław A. Kłopotek[1] · **Robert A. Kłopotek**[2]

Received: 16 December 2020 / Revised: 16 June 2021 / Accepted: 12 July 2021 /
Published online: 3 November 2021

Abstract
Kleinberg introduced an axiomatic system for clustering functions. Out of three axioms, he proposed, two (scale invariance and consistency) are concerned with data transformations that should produce the same clustering under the same clustering function. The so-called consistency axiom provides the broadest range of transformations of the data set. Kleinberg claims that one of the most popular clustering algorithms, *k*-means does not have the property of consistency. We challenge this claim by pointing at invalid assumptions of his proof (infinite dimensionality) and show that in one dimension in Euclidean space the *k*-means algorithm has the consistency property. We also prove that in higher dimensional space, *k*-means is, in fact, inconsistent. This result is of practical importance when choosing testbeds for implementation of clustering algorithms while it tells under which circumstances clustering after consistency transformation shall return the same clusters. Two types of remedy are proposed: gravitational consistency property and dataset consistency property which both hold for *k*-means and hence are suitable when developing the mentioned testbeds.

Keywords Cluster analysis · Consistency property · Gravitational consistency property · Fixed dimensional Euclidean space consistency · *k*-Means algorithm

1 Introduction

In his heavily cited paper (Kleinberg, 2002), Kleinberg introduced an axiomatic system for clustering functions. A clustering function applied to a dataset S produces a partition Γ.

This is an extended version of a conference paper (Klopotek & Klopotek, 2020).

✉ Mieczysław A. Kłopotek
klopotek@ipipan.waw.pl

Robert A. Kłopotek
r.klopotek@uksw.edu.pl

[1] Institute of Computer Science, Polish Academy of Sciences, 01-248 Warsaw, ul. Jana Kazimierza 5, Poland

[2] Faculty of Mathematics and Natural Sciences, School of Exact Sciences, Cardinal Stefan Wyszyński University in Warsaw, Warsaw, Poland

A *partition* Γ of a set S into k subsets (clusters) is to be understood as the set of subsets $\Gamma = \{C_1, C_2, ..., C_k\}$ such that $\cup_{i=1}^{k} C_i = S$, $C_i \cap C_j = \emptyset$ for any $i \neq j$, $C_i \subseteq S$ and $C_i \neq \emptyset$ for any i. Kleinberg (2002, Section 2) defines clustering function as:

Definition 1 A *clustering function* is a function f that takes a distance function d on [set] S [of size $n \geq 2$] and returns a partition Γ of S. The sets in Γ will be called its *clusters*.

where the distance is understood by him as

Definition 2 With the set $S = \{1, 2, \ldots, n\}$ [...] we define a distance function to be any function $d : S \times S \to \mathbb{R}$ such that for distinct $i, j \in S$ we have $d(i, j) \geq 0$, $d(i, j) = 0$ if and only if $i = j$, and $d(i, j) = d(j, i)$.

Out of three axioms, he proposed, two are concerned with data transformations that should produce the same clustering (partition) under the same clustering function. We can speak here about "clustering preserving transformations" induced by these axioms. The so-called *consistency axiom*, mentioned below, shall be of interest to us here as it provides the broadest range of transformations. Note that, following literature, we use here terms "property" and "axiom" interchangeably.

Property 1 *Let Γ be a partition of S, and d and d' two distance functions on S. We say that d' is a Γ-transformation of d if (a) for all $i, j \in S$ belonging to the same cluster of Γ, we have $d'(i, j) \leq d(i, j)$ and (b) for all $i, j \in S$ belonging to different clusters of Γ, we have $d'(i, j) \geq d(i, j)$. The clustering function f has the consistency property if for each distance function d and its Γ-transformation d' the following holds: if $f(d) = \Gamma$, then $f(d') = \Gamma$*

Subsequently, we will talk about Γ-transformation exchangeably with Γ-based consistency transformation or just consistency transformation. Let us mention also the other clustering preservation axiom of Kleinberg, that is the *scale-invariance axiom.*

Property 2 *A function f has the scale-invariance property if for any distance function d and any $\alpha > 0$, we have $f(d) = f(\alpha \cdot d)$.*

The validity or non-validity of any clustering preserving axiom for a given clustering function is of vital practical importance, as it may serve as a foundation for a testbed of the correctness of the function. Any modern software developing firm creates tests for its software in order to ensure its proper quality. Generators providing versatile test data are therefore of significance because they may detect errors unforeseen by the developers. Thus the consistency axiom may be used to generate new test data from existent one knowing a priori what the true result of clustering should be. The scale-invariance axiom may be used too, but obviously, the diversity of derived sets is much smaller.

Kleinberg defined a class of clustering functions, called the *centroid functions* as follows: for any natural number $k \geq 2$, and any continuous, non-decreasing, and unbounded function $g : \mathbb{R}^+ \to \mathbb{R}^+$, the $(k; g)$-centroid clustering consists of: (1) choosing the set of k centroid points $T \subseteq S$ for which the objective function $\Delta_d^g(T) = \sum_{i \in S} g(d(i, T))$ is minimized, where $d(i, T) = \min_{j \in T} d(i, j)$. (2) a partition of S into k clusters is obtained by assigning each point to the element of T closest to it. He claims that the objective function underlying k-means clustering is obtained by setting $g(d) = d^2$. This is not quite correct because

cluster centers in k-means do not necessarily belong to S, though with a dense set S, the approximation may be relatively good. It would be more appropriate if Kleinberg would speak about k-medoid algorithm

Note that his distance definition (Def. 2) is not a Euclidean one and not even metric, as he stresses. This is of vital importance because based on this he formulates and proves a theorem (his Theorem 4.1)

Theorem 1 *Theorem 4.1 from* Kleinberg (2002). *For every $k \geq 2$ and every function g [...] and for [data set size] n sufficiently large relative to k, the $(k; g)$-centroid clustering function [this term encompassing k-means] does not satisfy the Consistency property.*

which we claim is wrong with respect to k-means for a number of reasons as we will show below. The reasons are:

- The objective function underlying k-means clustering is *not* obtained by setting $g(d) = d^2$ contrary to Kleinberg's assumption (k-medoid is obtained).
- k-means always works in fixed-dimensional space while his proof relies on unlimited dimensional space.
- Unlimited dimensionality implies a serious software testing problem because the algorithm's correctness cannot be established by testing as the number of tests is too vast.
- The consistency property holds for k-means in one-dimensional space.

The last result opens the problem of whether or not the consistency also holds for higher dimensions.

We begin our presentation with recalling basics of the k-means algorithms in Section 2. We recall the Kleinberg's proof of k-means inconsistency and point at its weak points in Section 3. Then we investigate the impact of dimensionality of k-means consistency in Section 4. In Section 5 we discuss the reasons for inconsistency in multi-dimensional spaces and propose a remedy in terms of gravitational consistency and generalized gravitational consistency. In Section 6, we suggest still a different way around the problem by proposing dataset consistency property. Section 7 reports on some experiments illustrating selected insights from the paper. Conclusions are presented in Section 8.

2 *k*-Means algorithm

The popular clustering algorithm, k-means (MacQueen, 1967) strives to minimize the partition quality function (called also *partition cost function*)

$$Q(U, M) = \sum_{i=1}^{m} \sum_{j=1}^{k} u_{ij} \|\mathbf{x}_i - \boldsymbol{\mu}_j\|^2 \tag{1}$$

where $\mathbf{x}_i$, $i = 1, \ldots, m$ are the data points, M is the matrix of cluster centers $\boldsymbol{\mu}_j$, $j = 1, \ldots, k$, and U is the cluster membership indicator matrix, consisting of entries u_{ij}, where u_{ij} is equal to 1 if among all of cluster (gravity) centers $\boldsymbol{\mu}_j$ is the closest to $\mathbf{x}_i$, and is 0 otherwise.

It can be rewritten in various ways while the following are of interest to us here. Let the partition $\Gamma = \{C_1, \ldots, C_k\}$ b a partition of the data set onto k clusters $C_1, \ldots, C_k$. Then

$$Q(\Gamma) = \sum_{j=1}^{k} \sum_{\mathbf{x}_i \in C_j} \|\mathbf{x}_i - \boldsymbol{\mu}(C_j)\|^2 \tag{2}$$

where $\boldsymbol{\mu}(C) = \frac{1}{|C|} \sum_{\mathbf{x}_i \in C} \mathbf{x}_i$ is the gravity center of the cluster C. The above can be presented also as

$$Q(\Gamma) = \frac{1}{2} \sum_{j=1}^{k} \frac{1}{|C_j|} \sum_{\mathbf{x}_i \in C_j} \sum_{\mathbf{x}_l \in C_j} \|\mathbf{x}_i - \mathbf{x}_l\|^2 \tag{3}$$

The problem of seeking the pair (U, M) minimizing J from equation (1) is called *k-means-problem*. This problem is known as NP-hard. We will call *k-means-ideal* such an algorithm that finds a pair (U, M) minimizing Q from equation (1). Practical implementations of k-means usually find some local minima of $Q()$. There exist various variants of this algorithm. For an overview of many of them, see, e.g., Wierzchoń and Kłopotek (2018). An algorithm is said to be from the k-means family if it has the structure described by Algorithm 1. We will use a version with random initialization (randomly chosen initial seeds) as well as an artificial one initialized close to the true cluster center, which mimics k-means-ideal.

Algorithm 1 Structure of a practical algorithm from k-means-family.

Data: the data points $\mathbf{x}_i$, $i = 1, \ldots, m$, the required number of clusters k
Result: $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k$
[1] Initialize k cluster gravity centers $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k$;
while *a stop criterion (no change of cluster membership, or no sufficient improvement of the objective function, or exceeding some maximum number of iterations, or some other criterion) not reached* **do**
  [2] Assign each data element $\mathbf{x}_i$ to the cluster C_j identified by the closest $\boldsymbol{\mu}_j$;
  [3] Update $\boldsymbol{\mu}_j$ of each cluster C_j as the gravity center of the data elements in C_j;
end

3 Kleinberg's proof of Theorem 1 and its unlimited dimensionality deficiency

Kleinberg's proof, delimited to the case of $k = 2$ only, runs as follows: Consider a set of points $S = X \cup Y$ where X, Y are disjoint and $|X| = m$, $|Y| = \gamma m$, where $\gamma > 0$ is "small". $\forall_{i,j \in X} d(i, j) = r$, $\forall_{i,j \in Y} d(i, j) = \epsilon < r$, $\forall_{i \in X, j \in Y} d(i, j) = r + \delta$ where $\delta > 0$ and δ is "small". By choosing $\gamma, \epsilon, r, \delta$ appropriately, the optimal choice of $k = 2$ centroids will consist of one point from X and one from Y. The resulting partition is $\Gamma = \{X, Y\}$. Let divide X into $X = X_0 \cup X_1$ with X_0, X_1 of equal cardinality. Reduce the distances so that $\forall_{c=1,2} \forall_{i,j \in X_c} d'(i, j) = r' < r$ and $d' = d$ otherwise. If r' is "sufficiently small", then the optimal choice of two centroids for S will now consist of one point from each X_c, yielding a different partition of S. But d' is a Γ-transform of d so that a violation of consistency occurs. So far the proof of Kleinberg of the Theorem 1.

The proof cited above is a bit excentric because the clusters are heavily unbalanced (k-means tends to produce rather balanced clusters). Furthermore, the distance function

is awkward because Kleinberg's counter-example would require an embedding in a very high dimensional space, non-typical for k-means applications. It needs to be mentioned that Kleinberg's proof, sketchy in nature, omitted many details. Kleinberg uses a distance definition that is broader than Euclidean and therefore he does not consider space dimensionality. k-means, on the other hand, in its basic version, explicitly assumes an Euclidean space. This is the reason, why we consider Kleinberg's proof in the light of Euclidean space embedding.

We claim in brief:

Theorem 2 *Kleinberg's proof of* Kleinberg (2002) *Theorem 4.1 that k-means ($k = 2$) is not consistent, is not valid in $\mathbb{R}^p$ for data sets of cardinality $n > 2(p+1)$.*

Proof In terms of the concepts used in the Kleinberg's proof, either the set X or the set Y is of cardinality $p+2$ or higher. Kleinberg requires that distances between $p+2$ points are all identical which is impossible in $\mathbb{R}^p$ (only up to $p+1$ points may be equidistant). □

Furthermore Kleinberg's minimized target function

$$\Delta_d^g(T) = \sum_{i \in S} g(d(i, T)) \tag{4}$$

where $d(i, T) = \min_{j \in T} d(i, j)$, differs significantly from the formula (3). For the original set X, the formula (3) would return $\frac{1}{2}(m-1)r^2$, while Kleinberg's would produce $(m-1)r^2$. For a combination of a elements from X and b elements from Y in one cluster we get $\frac{a(a-1)r^2/2+b(b-1)\epsilon^2/2+ab(r+\delta)^2}{a+b}$ from (2) or the minimum of $(a-1)r^2 + b(r+\delta)^2$ and $(b-1)\epsilon^2 + a(r+\delta)^2$ for Kleinberg's $\Delta_d^g(T)$. The discrepancy between these formulas is shown in Fig. 1. We assumed there $r = 10$, $\epsilon = 8$, $\delta = 1$ and $m = 1000$.

We see immediately that

Theorem 3 *Kleinberg's target function does not match the real k-means target function.*

4 The impact of dimensionality of consistency property

As visible from Theorem 2, the dimensionlity of the space impacts the validity of Kleinberg's proof of inconsistency of k-means. However, this does not answer the question whether or not k-means is actually consistent in a fixed dimensional space. In this section we will show that in fact k-means is consistent in one-dimensional space (Theorem 4), but it is inconsistent in 3 or more dimensions (Theorem 5) and also it is inconsistent in 2 dimensions (Theorem 6).

Theorem 4 *k-means is consistent in one dimensional Euclidean space.*

The proof is postponed to the Appendix A.1.

But what about higher dimensions?

Theorem 5 *k-means in 3D is not consistent.*

The proof, by example, is postponed to the Appendix A.2. The example used in that proof is more realistic (balanced, in Euclidean space) than that of Kleinberg and shows that inconsistency of k-means in $\mathbb{R}^m$ is a real problem. With the example used in the proof of Theorem

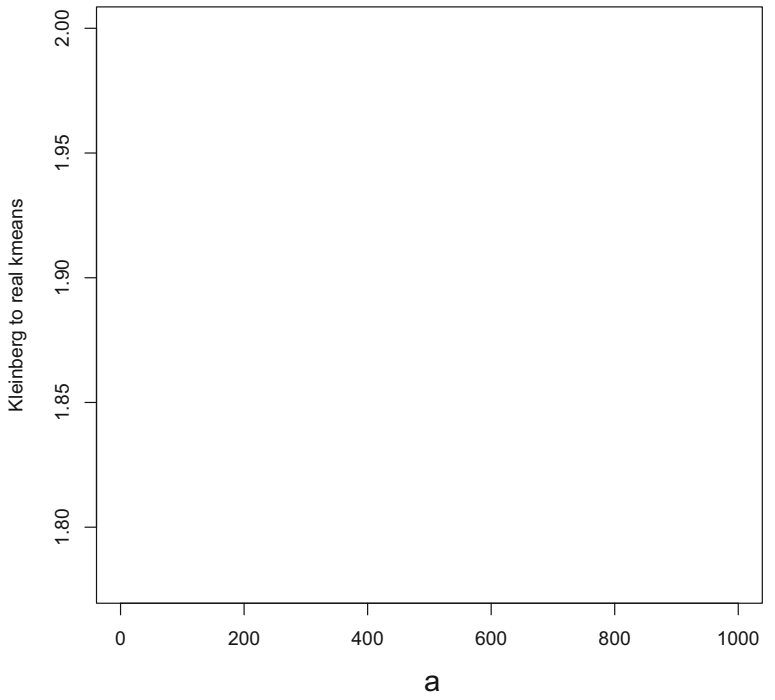

Fig. 1 Quotient of Kleinberg's k-means target (formula (4)) and the real k-means target (formula formula (3))

5 not only *consistency violation* is shown, but also *refinement-consistency violation*. Not only in 3D, but also in higher dimensions (as 3D example may always be embedded in n dimensions, $n > 3$). So what about the case of two dimensions - 2D?

Theorem 6 *k-means in 2D is not consistent.*

Proof The proof of Theorem 6 uses a less realistic example than in Theorem 5, hence Theorem 5 was worthy considering in spite of the fact that it is implied by Theorem 6. Imagine a unit circle with data points arranged as follows (Fig. 2 left): one data point in the center, and the remaining points arranged on the circle with the following angular positions with respect to the circle center.
Set A={$13^o, 14^o, \ldots, 22^o, -13^o, -14^o, \ldots, -22^o$}.
Set B={$133^o, 134^o, \ldots, 142^o, -133^o, -134^o, \ldots, -142^o$}.
k-means with $k = 2$ will merge points of the set B and the circle middle point as one cluster, and the set A as the other cluster. After a Γ transformation (Fig. 2 right) let A turn to A' identical with A and and let B change to B' B'={$162^o, 163^o, \ldots, 171^o, -162^o, -163^o, \ldots, -171^o$}, while the point in the center of the circle remains in its position. Now k-means with $k = 2$ yields one cluster consisting of points of the set B' and the second cluster consisting of the circle middle point and the set A'. The center point of the circle switches the clusters upon Γ transformation (Fig. 2 right). □

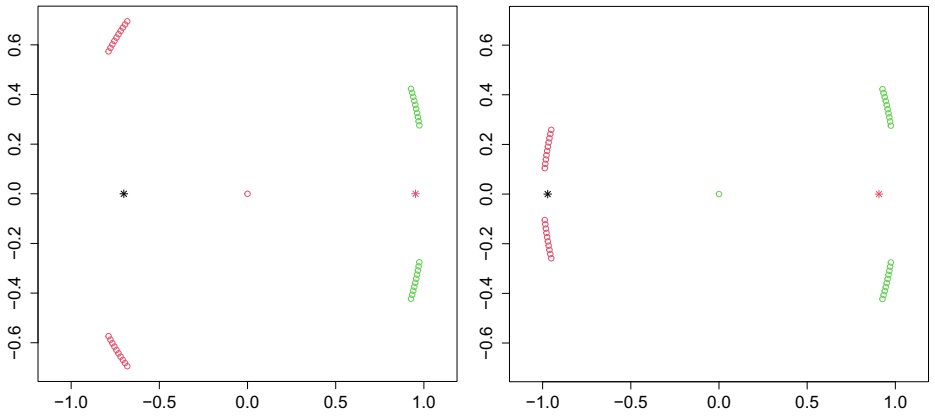

Fig. 2 Inconsistency of k-means in 2D Euclidean space. Left picture - data partition before consistency transform. Right picture - data partition after consistency transform. Cluster elements are marked with blue and green. Red points indicate cluster centers

5 Reasons for multidimensional inconsistency

In order to investigate the reasons for k-means inconsistency in higher dimensions, in analogy to the proof of Theorem 4 from Section 4, let us consider two alternative partitions in a multi-dimensional space:

- the partition $\Gamma_1 = \{C_{1.}, \ldots, C_{k.}\}$ which will be base for the Γ-transform
- and the competing partition $\Gamma_2 = \{C_{.1}, \ldots, C_{.k'}\}$.

Assume further that $C_{ij} = C_{i.} \cap C_{.j}$ are non-empty intersections of clusters $C_{i.} \in \Gamma_1$, $C_{.j} \in \Gamma_2$, of both partitions. Define $minind(C_{i.})$, resp. $maxind(C_{i.})$ as the minimal/maximal index j such that C_{ij} is not empty. The $Q(\Gamma_1)$ will be the sum of centered sums of squares over all C_{ij} plus the squared distances of centers of all C_{ij} to the center of $C_{i.}$ times cardinality of C_{ij}.

We can derive the formula for $Q(\Gamma_1)$ in the same way as in the proof of Theorem 4 in Appendix A.1 (8)

$$
\begin{aligned}
Q(\Gamma_1) &= \left(\sum_{i,j;C_{ij}\neq\emptyset} \sum_{\mathbf{x}\in C_{ij}} \|\mathbf{x} - \boldsymbol{\mu}(C_{ij})\|^2 \right) \\
&\quad + \left(\sum_{C_{i.}\in\Gamma_1} \sum_{j;C_{ij}\neq\emptyset} |C_{ij}| \|\boldsymbol{\mu}(C_{ij}) - \boldsymbol{\mu}(C_{i.})\|^2 \right) \\
&= \left(\sum_{i,j;C_{ij}\neq\emptyset} \sum_{\mathbf{x}\in C_{ij}} \|\mathbf{x} - \boldsymbol{\mu}(C_{ij})\|^2 \right) \\
&\quad + \left(\sum_{C_{i.}\in\Gamma_1} 0.5 \sum_{j;C_{ij}\neq\emptyset} \sum_{j';C_{ij'}\neq\emptyset} \frac{|C_{ij}| \cdot |C_{ij'}|}{|C_{i.}|} \|\boldsymbol{\mu}(C_{ij}) - \boldsymbol{\mu}(C_{ij'})\|^2 \right) \qquad (5)
\end{aligned}
$$

The $Q(\Gamma_2)$ can be derived also in analogy to equation (8) in the proof of Theorem 4 in Appendix A.1 as:

$$Q(\Gamma_2) = \left(\sum_{i,j;C_{ij}\neq\emptyset} \frac{|C_{ij}|}{|C_{.j}|} \sum_{\mathbf{x}\in C_{ij}} \|\mathbf{x}-\boldsymbol{\mu}(C_{ij})\|^2 \right) + \sum_{C_{.j}\in\Gamma_2} \frac{0.5}{|C_{.j}|} \left(\sum_{i';C_{i'j}\neq\emptyset} \sum_{i'';i'\neq i'',C_{i''j}\neq\emptyset} \sum_{\mathbf{x}\in C_{i'j},\mathbf{y}\in C_{i''j}} \|\mathbf{x}-\mathbf{y}\|^2 \right)$$

The first summand of $Q(\Gamma_1)$, that is $\left(\sum_{i,j;C_{ij}\neq\emptyset}\sum_{\mathbf{x}\in C_{ij}}\|\mathbf{x}-\boldsymbol{\mu}(C_{ij})\|^2\right)$ will decrease upon Γ_1 based consistency transformation. The reason is that $\sum_{\mathbf{x}\in C_{ij}}\|\mathbf{x}-\boldsymbol{\mu}(C_{ij})\|^2$ is equivalent to $\frac{0.5}{|C_{ij}|}\left(\sum_{\mathbf{x}\in C_{ij},\mathbf{y}\in C_{ij}}\|\mathbf{x}-\mathbf{y}\|^2\right)$ which decreases because the distances between elements of C_{ij} decreases as they are all in the same cluster $C_{i.}$. As summands of $Q(\Gamma_2)$ are concerned, the first, equal $\left(\sum_{i,j;C_{ij}\neq\emptyset}\frac{|C_{ij}|}{|C_{.j}|}\sum_{\mathbf{x}\in C_{ij}}\|\mathbf{x}-\boldsymbol{\mu}(C_{ij})\|^2\right)$, will therefore also decrease upon Γ_1 transformation. But not by the same absolute value as the first one of $Q(\Gamma_1)$, that is
$\left(\sum_{i,j;C_{ij}\neq\emptyset}\sum_{\mathbf{x}\in C_{ij}}\|\mathbf{x}-\mu(C_{ij})\|^2\right)$, because always $|C_{ij}|\leq|C_{.j}|$. The second summand of $Q(\Gamma_2)$, that is

$$\sum_{C_{.j}\in\Gamma_2}\frac{0.5}{|C_{.j}|}\left(\sum_{i';C_{i'j}\neq\emptyset}\sum_{i'';i'\neq i'',C_{i''j}\neq\emptyset}\sum_{\mathbf{x}\in C_{i'j},\mathbf{y}\in C_{i''j}}\|\mathbf{x}-\mathbf{y}\|^2\right)$$

will increase because $\mathbf{x}$, $\mathbf{y}$ stem from different clusters of Γ_1. If Γ_1 was the optimal clustering for k-means cost function prior to Γ_1 transformation, it would remain so afterward if the second summand of $Q(\Gamma_1)$, that is $\sum_{C_{i.}\in\Gamma_1}0.5\sum_{j;C_{ij}\neq\emptyset}\sum_{j';C_{ij'}\neq\emptyset}\frac{|C_{ij}|\cdot|C_{ij'}|}{|C_{i.}|}\|\boldsymbol{\mu}(C_{ij})-\boldsymbol{\mu}(C_{ij'})\|^2$, would decrease. However, in a multidimensional space, this is not granted anymore, because $\|\boldsymbol{\mu}(C_{ij})-\boldsymbol{\mu}(C_{ij'})\|^2$ may increase when the points of the cluster $C_{i.'}$ are getting closer to one another. An immediate remedy would be then to require that for any two convex subsets C_{ij}, $C_{ij'}$ of $C_{i.}$, $\|\boldsymbol{\mu}(C_{ij})-\boldsymbol{\mu}(C_{ij'})\|^2$ is non-increasing upon Γ_1 transformation. This condition is not easy to check. However, if one decreases all distances within one cluster $C_{i.}$ by the very same factor, then this condition holds. It also holds if, within an orthogonal coordinate system, one decreases all distances within one cluster $C_{i.}$ along each dimension by a factor specific for the dimension and the cluster. Under such circumstances, the distances within a cluster will not be necessarily changed by the same factor.

So, define the gravitational consistency as follows:

Property 3 *Let Γ be a partition of S, and d and d' two distance functions on S. We say that d' is a Γ-gravitational-transformation of d if (a) for all $i, j \in S$ belonging to the same cluster of Γ, we have $d'(i,j)=\alpha d(i,j)$ where $0<\alpha\leq 1$ and α is specified for a given cluster (may be different for different clusters) and (b) for all $i, j \in S$ belonging to different clusters of Γ, we have $d'(i,j)\geq d(i,j)$. The clustering function f has the gravitational consistency property if for each distance function d and its Γ-gravitational-transformation d' the following holds: if $f(d)=\Gamma$, then $f(d')=\Gamma$*

Theorem 7 *k-means ideal has the gravitational consistency property.*

Proof Straightforward from the above. □

Define also the generalized gravitational consistency as follows:

Property 4 *Let Γ be a partition of S, and d and d' two distance functions on S. We say that d' is a Γ-generalized-gravitational-transformation of d if (a) for all $i \in S$ belonging to the same cluster C of Γ, with $\boldsymbol{\mu}(C)$ being its gravity center, and for an orthogonal coordinate CS specific for this cluster, for each coordinate axis $a \in CS$ we have $d'_a(i, \boldsymbol{\mu}(C)) = \alpha(C, a)d_a(i, \boldsymbol{\mu}(C))$ where $0 < \alpha(C, a) \leq 1$, d_a being the length of projection of the vector $(i, \boldsymbol{\mu}(C)$ on the coordinate axis a (same for d') and $\alpha(C, a)$ is specified for a given cluster and coordinate (may be different for different clusters and different coordinates) and (b) for all $i, j \in S$ belonging to different clusters of Γ, we have $d'(i, j) \geq d(i, j)$. The clustering function f has the generalized gravitational consistency property if for each distance function d and its Γ-generalized-gravitational-transformation d' the following holds: if $f(d) = \Gamma$, then $f(d') = \Gamma$*

Theorem 8 *k-means ideal has the generalized gravitational consistency property.*

Proof Straightforward from the above. □

6 Dataset consistency

The gravitational consistency can be viewed as too rigid as there exists a very strict limitation on how the distances between data elements can change. Though generalized gravitational consistency is less restrictive, the variations of distances within a cluster are nonetheless quite restricted, determined by as many factors only as there are dimensions.

Note that we had considered so far the case when any data was clustered by the clustering algorithm. Let us now investigate whether or not we can define data set properties for which Kleinberg's consistency property would hold for k-means. We would speak then about dataset consistency.

The idea we present here is quite simplistic, but nonetheless, it demonstrates that clustering algorithm properties may be implied by data set properties.

Assume we know what properties a dataset needs to possess so that we would know in advance partition Γ_0 for which the absolute minimum of k-means quality function $Q(\Gamma)$ (3) is obtained. Assume that this property depends on the distances between cluster centers, among others. When performing Γ-transformation, the cluster centers can move by at most the distance between the cluster center and the most distant point of the cluster. So it is sufficient to add to the distances between the clusters the maximum relocation for each cluster. Hence after Γ transformation, the distances are still sufficient to ensure the absolute minimum of the k-means target function.

The only task to do now is to identify this property of a dataset, allowing to know in advance the aforementioned absolute minimum of k-means Q-function.

So we will investigate below under what circumstances it is possible to tell, without exhaustive check, that the well-separated clusters are the global minimum of k-means. We will see that the ratio between the largest and the smallest cluster cardinality plays here an important role.

Definition 3 There is a gap g between two clusters A, B, if the distance between (hyper)balls centered at gravity centers of these clusters and enclosing each cluster amounts to g.

Let us consider a set of clusters $\Gamma = \{C_1, \ldots, C_k\}$, where k is the number of clusters, n_i is the number of elements in cluster C_i, r_i is the radius of the (hyper)ball centered at gravity center of cluster C_i and containing all the datapoints of the cluster C_i, $M = \max_i n_i$, $m = \min_i n_i$. Let g be the gap between every two clusters C_i, C_j fulfilling the conditions (6) and (7)

$$\forall_{p,q;p\neq q;p,q=1,\ldots,k} \quad g \geq k\sqrt{n_p + n_q + n}\sqrt{\frac{\sum_{i=1}^{k} n_i r_i^2}{n_p n_q}} \tag{6}$$

$$\forall_{i=1,\ldots,k} \quad g \geq r_i\sqrt{k\frac{M+n}{m}} \tag{7}$$

Theorem 9 *A clustering Γ_0 for which conditions (6) and (7) imposing constraints on the gap between clusters g hold, is optimal clustering that is with the lowest value of $Q(\Gamma)$ among all the partitions of the same cardinality as Γ_0.*

Proof has been postponed to Appendix A.3.
Therefore we may call the above-mentioned *well-separatedness* as *absolute clustering*.

Definition 4 A clustering is called *absolute* if conditions (6) and (7) imposing constraints on the gap between clusters g hold.

One sees immediately that inner cluster consistency is kept, this time in terms of global optimum, under the restraint to k clusters.

Theorem 10 *k-means ideal, applied to a dataset with gaps between intrinsic clusters amounting to the g plus the radii of the clusters between which the gap is measured, has the Kleinberg's consistency property.*

The proof is straightforward.

7 Experiments

7.1 Theorem 4 related experiments

Experiments have been performed to check whether or not the Theorem 4 that denies Kleinberg's findings for one-dimensional space really holds. Samples were generated from uniform distribution (sample size 100, 200, 400, 1000, 2000, 4000, 10000) for each $k = 2, \ldots, floor(\sqrt{samplesize})$. Then the respective sample was clustered into k clusters ($k = 2, \ldots, floor(\sqrt{samplesize})$) and k-means clustering (R package) was performed with $100k$ restarts. Subsequently, Γ transformation was performed where the distances within a cluster were decreased by a randomly chosen factor (a separate factor for each pair of neighboring data points), and at the same time, the clusters were moved away so that the distance between cluster elements of distinct clusters is not decreased. Then k-means clustering was performed with $100k$ restarts in two variants. The first variant was with random

initialization. The second variant was with the initialization of the midpoint of the original (rescaled) cluster interval. Additionally, for control purposes, the original samples were reclustered. The number of partitions was counted for which errors in restoring the original clustering was observed. Experiments were repeated ten times. Table 1 presents the average results obtained.

In this table, looking at the errors for variant 1, we see that more errors are committed with the increasing sample size (and hence increasing the maximum of k). This contrasts with the variant 2 where the number of errors is negligible. The second variant differs from the first in that seeds are distributed so that there is one in each intrinsic cluster.

Clearly the Theorem 4 holds (as visible from the variant 2). At the same time, however, the table shows that k-means with random initialization is unable to initialize properly for a larger number k of clusters in spite of a large number of restarts (variant 1). This is confirmed by the experiments with reclustering original data.

This study also shows how a test data generator may work when comparing variants of k-means algorithm (for one-dimensional data)

7.2 Theorem 5 related experiments

A simulation was performed concerning the relocation of points of the line segments AB, AC from the proof of Theorem 5.

The results are presented in Table 2 The top row, named $\measuredangle C'AB'$ represents the angle between line segments $C'a$ and AB' after rotation of AB and AC line segments upon Γ transformation. The effects of this rotating transformation are measured by the following quantities

– *wrong* Γ - number of k-means clustering errors compared to the original clustering before Γ transformation (consisting in the rotation of AB, AC) (out of 4000 data points in both clusters).

Initially, the angle $\measuredangle CAB$ between the line segments AB, AC was a right angle ($\pi/2$). As shown in Table 2, the angle between these line segments was decreased in steps of $\pi/20$ down to $\pi/20$ and the clustering using k-means (with 50 restarts) was performed.

k-means algorithm, applied to the data set $AB \cup AC \cup DE \cup DF$ returned, as expected two clusters: $AB \cup AC$ and $DE \cup DF$ (the column $\frac{\pi}{2}$). As visible in the row *wrong* Γ, the number of clustering errors compared to the original clustering was increasing up to over 4% of data points being misclassified upon rotation. It is apparent that in fact k-means is not consistent in three dimensions, as claimed in Theorem 5.

In order to illustrate better the importance of the concept of gravitational consistency, an experiment was performed related to equation (5) (first line). As previously, a data set related to $AB' \cup AC'$ subsets of the data for appropriate rotations of the line segments

Table 1 Validation of the Theorem 4

sample size	100	200	400	1000	2000	4000	10000
max. k	10	14	20	31	44	63	100
errors variant 1	0	0	0	2.4	10.2	21.5	62.4
errors variant 2	0	0	0	0	0	0.5	2.0
errors reclustering	0	0	2.3	14.1	30.2	49.5	86.2

Table 2 Validation of the Theorem 5 Explanations of row labels provided in the text

$\measuredangle C'AB'$	$\frac{\pi}{2}$	$0.9\frac{\pi}{2}$	$0.8\frac{\pi}{2}$	$0.7\frac{\pi}{2}$	$0.6\frac{\pi}{2}$	$0.5\frac{\pi}{2}$	$0.4\frac{\pi}{2}$	$0.3\frac{\pi}{2}$	$0.2\frac{\pi}{2}$	$0.1\frac{\pi}{2}$
wrong Γ	0	0.0	27.0	65.0	100.0	118.0	138.0	153.0	172.0	174.0
$\boldsymbol{\mu}$ sc 1	10.97	11.18	11.39	11.58	11.75	11.90	12.03	12.13	12.20	12.25
$\boldsymbol{\mu}$ sc2	24.04	22.47	20.82	19.13	17.42	15.74	14.15	12.73	11.57	10.81
SS sc1	69098	71207	73264	75218	77022	78631	80004	81110	81919	82413
SS sc2	350319	310149	270968	233741	199385	168745	142576	121523	106103	96696

AB, AC was considered. This data set was split into two parts: 1) subcluster Z_1 consisting of points with distance to A not higher than 20, 2) subcluster Z_2 consisting of the remaining points. $Z_1 \cup Z_2 = AB \cup AC$. While the rotation was performed, the following statistics of Z_1', Z_2', that is images of Z_1, Z_2 after rotation were observed:

- $\boldsymbol{\mu}$ *sc 1* - distance between means of the cluster $AB' \cup AC'$ and the mean of subcluster Z_1',
- $\boldsymbol{\mu}$*sc 2* - distance between means of the cluster $AB' \cup AC'$ and the mean of subcluster Z_2',
- *SS sc 1* - contribution of subcluster Z_1' to the sum of squares of the $AB' \cup AC'$. *SS sc 2* - contribution of subcluster Z_2' to the sum of squares of the $AB' \cup AC'$.

When the angle $\measuredangle C'AB'$ was decreased (Γ transformation), the distances between points within both subsets Z_1', Z_2' as well as between both subsets Z_1', Z_2' were decreased. So was the distance between the gravity center of the entire data set $A'B' \cup A'C'$ and the gravity center of the second subset Z_2' was decreasing, as visible in the row $\boldsymbol{\mu}$ *sc 2* of the Table 2. However, the distance between the gravity center of the entire data set $A'B' \cup A'C'$ and the gravity center of the first subset Z_1 was *increasing*, as visible in the row $\boldsymbol{\mu}$ *sc 1* of the Table 2. Also the contribution of this subset to the overall sum of squares of the entire set was increasing, as visible from the row *SS sc 1* of the Table 2. This demonstrates that the Γ transformation, though decreasing the distances between cluster data points, does not necessarily decrease the distance between sub-cluster centers and the cluster center which results in the inconsistency of k-means under Kleinberg's Γ transformation.

7.3 Theorem 7 related experiments

Experiments were also performed referring to the Theorem 7 and the results are summarized in Table 3. The following metrics were used.

- α - the contraction coefficient from Theorem 7
- *wrong* α - number of k-means clustering errors compared to the original clustering before Γ-gravitational transformation of the AB, AC cluster.

The experiments were performed for the same data as in previous subsection. the Γ-gravitational transformation was performed for the (original) cluster $AB \cup AC$ with α as indicated in the row α. The choice of α was based on the requirement that the Γ transformation and the Γ-gravitational transformation should yield a resulting cluster with the same variance of the data points in the cluster after transformation. As visible in the row *wrong* α, no error in data clustering was induced by Γ-gravitational transformation, as expected from the Theorem 7.

Table 3 Validation of the Theorem 7. Explanations of row labels provided in the text

α	1.00	0.95	0.90	0.85	0.80	0.76	0.71	0.68	0.65	0.64
wrong α	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

8 Conclusions

In this paper, we have provided a definite answer to the problem of whether or not k-means algorithm possesses the consistency property. The answer is negative except for one-dimensional space. Settling this problem was necessary because the proof of Kleinberg of this property was inappropriate for real application areas of k-means that it is a fixed-dimensional Euclidean space. The result precludes usage of consistency axiom as a generator of test examples for k-means clustering function (except for one-dimensional data) and implies the need to seek alternatives.

We proposed gravitational consistency, generalized gravitational consistency and dataset consistency as an alternative to Kleinberg's consistency property. Γ-gravitational transformation, as an alternative to Γ transformation, preserves the k-means clustering, but it is a bit rigid, because it keeps the proportions between distances in a single cluster. Generalized Γ-gravitational transformation does not have this disadvantage though there is still some rigidness as the changes in distances are concerned. The dataset consistency transformation is more flexible but requires quite large distances between the clusters. We believe, however, that these three alternatives can still generate a sufficient set of datasets for software tests. Note that an orientation on k-means is not a too serious limitation of usefulness as quite a large number of modern clustering algorithms encompass k-means clustering, just to mention the whole branch of spectral clustering.

Kleinberg's consistency was subject of strong criticism and new variants were proposed like Monotonic consistency (Strazzeri & Sánchez-García, 2018) or MST-consistency (Zadeh, 2010). See also criticism in Carlsson and Mémoli (2010) and Correa-Morrisa (2013). The mentioned new definitions of consistency are apparently restrictions of Γ-consistency, and therefore the Theorem 4 would be valid. The Monotonic consistency seems not to impose restrictions on Kleinberg's proof on k-means violating consistency. Therefore in those cases, the consistency of k-means under higher dimensionality needs to be investigated. Note that we have also challenged the result (Wei, 2017), who claims that Kleinberg's consistency may be achieved by k-means with random initialization (see our Theorem 5). The shift of axioms from clustering function to quality measure (Ben-David & Ackerman, 2008) was suggested to the problems with consistency, but this approach fails to tell what the outcome of clustering should be, which is not useful for the mentioned test generator application.

It should be noted that, beside the Kleinberg axiomatic system, other axiomatic frameworks have been proposed, which may serve as foundations of development of new test data sets from existent ones. For example for unsharp partitioning there was a proposal of an axiomatic system by Wright (1973), for graph clustering by van Laarhoven and Marchiori (2014), for cost function driven algorithms by Ben-David and Ackerman (2009), for linkage algorithms by Ackerman et al. (2010), for hierarchical algorithms by Carlsson and Mémoli (2010), Gower (1990), and Thomann et al. (2015), for multiscale clustering by Carlsson and Mémoli (2008). for settings with increasing sample sizes by (Hopcroft & Kannan, 2012), for community detection by Zeng et al. (2016), for pattern clustering by Shekar (1988). They were not investigated here and are a bit hard to compare because they were proposed for

different classes of clustering algorithms that do not cover the settings relevnt for k-means that is the embedding in the Euclidean space and partition of not only the sample but of the sample space.

Appendix A: Proofs of selected theorems

A.1 Proof of Theorem 4

Proof Consider two alternative partitions in one dimensional space:

- the partition $\Gamma_1 = \{C_{1.}, \ldots, C_{k.}\}$ which will be base for the Γ-transform
- and the competing partition $\Gamma_2 = \{C_{.1}, \ldots, C_{.k'}\}$.

Due to the nature of k-means let each cluster of each partition after Γ-transform be represented as an interval not intersecting with any other cluster of the same partition. For Γ_1, it holds before the transform; therefore, it holds afterward. Γ_2 shall be the competing optimal transform; therefore, it holds for sure afterward. We intend to demonstrate that under the Γ_1 transformation, that is, assuming that the intrinsic partition is Γ_1, the target function of k-means for Γ_1 will decrease not less than that for Γ_2. For simplicity, assume that the indices of clusters grow with the growing value of the cluster center.

For this purpose assume that $C_{ij} = C_{i.} \cap C_{.j}$ are non-empty intersections of clusters $C_{i.} \in \Gamma_1$, $C_{.j} \in \Gamma_2$, of both partitions. Define $minind(C_{i.})$, resp. $maxind(C_{i.})$ as the minimal/maximal index j such that C_{ij} is not empty. The $Q(\Gamma_1)$ will be the sum of centered sums of squares over all C_{ij} plus the squared distances of centers of all C_{ij} to the center of $C_{i.}$ times cardinality of C_{ij} (easily derived from formula (2)).

$$\begin{aligned} Q(\Gamma_1) &= \sum_{C_{i.} \in \Gamma_1} \sum_{j; C_{ij} \neq \emptyset} \left(|C_{ij}| (\mu(C_{ij}) - \mu(C_{i.}))^2 + \sum_{x \in C_{ij}} (x - \mu(C_{ij}))^2 \right) \\ &= \left(\sum_{i,j; C_{ij} \neq \emptyset} \sum_{x \in C_{ij}} (x - \mu(C_{ij}))^2 \right) + \left(\sum_{C_{i.} \in \Gamma_1} \sum_{j; C_{ij} \neq \emptyset} |C_{ij}| (\mu(C_{ij}) - \mu(C_{i.}))^2 \right) \end{aligned} \tag{8}$$

Please note that

$$\begin{aligned} \sum_{j; C_{ij} \neq \emptyset} |C_{ij}| \left(\mu(C_{ij}) - \mu(C_{i.})\right)^2 &= \sum_{j; C_{ij} \neq \emptyset} |C_{ij}| \left(\mu(C_{ij}) - \sum_{j'; C_{ij'} \neq \emptyset} \frac{|C_{ij'}|}{|C_{i.}|} \mu(C_{ij'}) \right)^2 \\ &= \sum_{j; C_{ij} \neq \emptyset} |C_{ij}| \left(\sum_{j'; C_{ij'} \neq \emptyset} \left(\frac{|C_{ij'}|}{|C_{i.}|} \mu(C_{ij}) - \frac{|C_{ij'}|}{|C_{i.}|} \mu(C_{ij'}) \right) \right)^2 \\ &= \sum_{j; C_{ij} \neq \emptyset} |C_{ij}| \left(\sum_{j'; C_{ij'} \neq \emptyset} \frac{|C_{ij'}|}{|C_{i.}|} \left(\mu(C_{ij}) - \mu(C_{ij'})\right) \right)^2 \\ &= 0.5 \sum_{j; C_{ij} \neq \emptyset} \sum_{j'; C_{ij'} \neq \emptyset} \frac{|C_{ij}| \cdot |C_{ij'}|}{|C_{i.}|} \left(\mu(C_{ij}) - \mu(C_{ij'})\right)^2 \end{aligned}$$

The $Q(\Gamma_2)$ can be computed analogously, but let us follow a bit distinct path (starting from formula (3)).

$$
\begin{aligned}
Q(\Gamma_2) &= \sum_{C_{.j}\in\Gamma_2} \frac{0.5}{|C_{.j}|} \sum_{x\in C_{.j}} \sum_{y\in C_{.j}} (x-y)^2 \\
&= \sum_{C_{.j}\in\Gamma_2} \frac{0.5}{|C_{.j}|} \left(\left(\sum_{i;C_{ij}\neq\emptyset} \sum_{x\in C_{ij}} \sum_{y\in C_{ij}} (x-y)^2 \right) \right. \\
&\quad \left. + \left(\sum_{i';C_{i'j}\neq\emptyset} \sum_{i'';i'\neq i'',C_{i''j}\neq\emptyset} \sum_{x\in C_{i'j},y\in C_{i''j}} (x-y)^2 \right) \right) \\
&= \sum_{C_{.j}\in\Gamma_2} \frac{0.5}{|C_{.j}|} \left(\left(\sum_{i;C_{ij}\neq\emptyset} 2|C_{ij}| \sum_{x\in C_{ij}} (x-\mu(C_{ij}))^2 \right) \right. \\
&\quad \left. + \left(\sum_{i';C_{i'j}\neq\emptyset} \sum_{i'';i'\neq i'',C_{i''j}\neq\emptyset} \sum_{x\in C_{i'j},y\in C_{i''j}} (x-y)^2 \right) \right) \\
&= \left(\sum_{i,j;C_{ij}\neq\emptyset} \frac{|C_{ij}|}{|C_{.j}|} \sum_{x\in C_{ij}} (x-\mu(C_{ij}))^2 \right) \\
&\quad + \sum_{C_{.j}\in\Gamma_2} \frac{0.5}{|C_{.j}|} \left(\sum_{i';C_{i'j}\neq\emptyset} \sum_{i'';i'\neq i'',C_{i''j}\neq\emptyset} \sum_{x\in C_{i'j},y\in C_{i''j}} (x-y)^2 \right) \qquad (9)
\end{aligned}
$$

Both summands of $Q(\Gamma_1)$, that is $\left(\sum_{i,j;C_{ij}\neq\emptyset} \sum_{x\in C_{ij}} (x-\mu(C_{ij}))^2)\right)$ and $\left(\sum_{C_{i.}\in\Gamma_1} \sum_{j;C_{ij}\neq\emptyset} (|C_{ij}|(\mu(C_{ij})-\mu(C_{i.}))^2\right)$ will decrease upon Γ_1 based consistency transformation. $(x-\mu(C_{ij}))^2$ decreases because the distance to each of elements of C_{ij} decreases as they are all in the same cluster $C_{i.}$. Each $(\mu(C_{ij})-\mu(C_{ij'}))^2$ decreases because all the elements constituting C_{ij} and $C_{ij'}$ belong to the same cluster $C_{i.}$. Hereby there is always an extreme data point $P_{ij} \in C_{ij}$ separating it from $C_{ij'}$. As the points of both C_{ij} and $C_{ij'}$ get closer to P_{ij} under Γ_1 transformation, so the centers of both C_{ij} and $C_{ij'}$ will get closer to P_{ij}, so that they will move closer to each other. As summands of $Q(\Gamma_2)$ are concerned, the first, equal $\left(\sum_{i,j;C_{ij}\neq\emptyset} \frac{|C_{ij}|}{|C_{.j}|} \sum_{x\in C_{ij}} (x-\mu(C_{ij}))^2\right)$, will also decrease upon Γ_1 transformation. But not by the same absolute value as the first one of $Q(\Gamma_1)$, that is $\left(\sum_{i,j;C_{ij}\neq\emptyset} \sum_{x\in C_{ij}} (x-\mu(C_{ij}))^2\right)$, because always $|C_{ij}| \leq |C_{.j}|$. But the second summand of $Q(\Gamma_2)$, that is

$$
\sum_{C_{.j}\in\Gamma_2} \frac{0.5}{|C_{.j}|} \left(\sum_{i';C_{i'j}\neq\emptyset} \sum_{i'';i'\neq i'',C_{i''j}\neq\emptyset} \sum_{x\in C_{i'j},y\in C_{i''j}} (x-y)^2 \right)
$$

will increase because x, y stem from different clusters of Γ_1. Therefore, if Γ_1 was the optimal clustering for k-means cost function prior to Γ_1 transformation, it will remain so afterward. □

A.2 Proof of Theorem 5

Proof Let A, B, C, D, E, F be points in three-dimensional space with coordinates: $A(1, 0, 0)$, $B(33, 32, 0)$, $C(33, -32, 0)$, $D(-1, 0, 0)$, $E(-33, 0, -32)$, $F(-33, 0, 32)$. Let S_{AB}, S_{AC}, S_{DE}, S_{DF} be sets of say 1000 points each randomly uniformly distributed over line segments (except for endpoints) AB, AC, DE, EF resp. Let $X = S_{AB} \cup S_{AC} \cup S_{DE} \cup S_{EF}$. k-means with $k = 2$ applied to X yields a partition $\Gamma = \{S_{AB} \cup S_{AC}, S_{DE} \cup S_{DF}\}$, as expected (see Fig. 3 left). Let us perform a Γ transformation consisting of rotating line segments AB, AC around the point A in the plane spread by the first two coordinates (X and Y) towards the first coordinate axis (X xis) so that the angle between this axis and AB' and AC' is say one degree. To verify that this is a Γ transformation, consider some points P, Q, P on the line segment AB and Q on the line segment AC. Their distance amounts to $|PQ| = \sqrt{|PA|^2 + |AQ|^2 - 2\cos(\measuredangle BAC)|PA||AQ|}$. The images of P, Q be P', Q' resp., whereby obviously $|P'A| = |PA|$ and $|AQ'| = |AQ|$, and $|\measuredangle B'AC'| < |\measuredangle BAC|$. Therefore

$$\begin{aligned} |P'Q'| &= \sqrt{|P'A|^2 + |AQ'|^2 - 2\cos(\measuredangle B'AC')|P'A||AQ'|} \\ &= \sqrt{|PA|^2 + |AQ|^2 - 2\cos(\measuredangle B'AC')|PA||AQ|} \\ &< \sqrt{|PA|^2 + |AQ|^2 - 2\cos(\measuredangle BAC)|PA||AQ|} = |PQ| \end{aligned}$$

as expected for Γ transformation for points of the same cluster. Let us consider a point R on the line segment DE and the distance $|RP|$ between points from two different clusters. Let R_x, P_x be orthogonal projections of R, P onto the X axis, resp. P, Q lie in two orthogonal planes, spread by X, Z and X, Y axes, resp. Therefore $|RP| =$

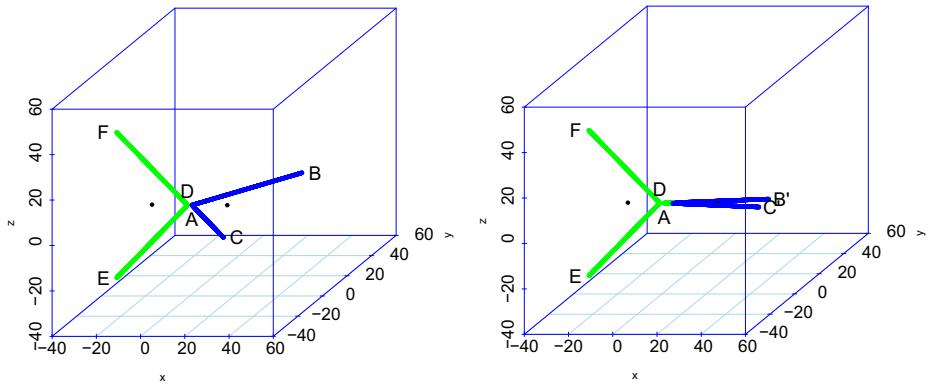

Fig. 3 Inconsistency of k-means in 3D Euclidean space. Left picture - data partition before consistency transform. Right picture - data partition after consistency transform

$\sqrt{|RR_x|^2 + |R_x P_x|^2 + |P_x P|^2}$, whereby $|R_x P_X| = |R_x D| + |DA| + |AP_X|$. Hence

$$\begin{aligned}|RP|^2 = &\left(|RD|\sin\left(\frac{1}{2}|\measuredangle EDF|\right)\right)^2 \\ &+\left(|RD|\cos\left(\frac{1}{2}|\measuredangle EDF|\right) + |DA| + |PA|\cos\left(\frac{1}{2}|\measuredangle BAC|\right)\right)^2 \\ &+\left(|PA|\sin\left(\frac{1}{2}|\measuredangle BAC|\right)\right)^2\end{aligned}$$

Let P'_x be the orthogonal projection of P' on the X xis. Then, after the Γ transformation, the distance of interest $|RP'|$ turns out to be
$|RP'| = \sqrt{|RR_x|^2 + |R_x P'_x|^2 + |P'_x P'|^2}$ that is

$$\begin{aligned}|RP'|^2 = &\left(|RD|\sin\left(\frac{1}{2}|\measuredangle EDF|\right)\right)^2 \\ &+\left(|RD|\cos\left(\frac{1}{2}|\measuredangle EDF|\right) + |DA| + |P'A|\cos\left(\frac{1}{2}|\measuredangle B'AC'|\right)\right)^2 \\ &+\left(|P'A|\sin\left(\frac{1}{2}|\measuredangle B'AC'|\right)\right)^2\end{aligned}$$

$$\begin{aligned}|RP'|^2 = &\left(|RD|\sin\left(\frac{1}{2}|\measuredangle EDF|\right)\right)^2 \\ &+\left(|RD|\cos\left(\frac{1}{2}|\measuredangle EDF|\right) + |DA|\right)^2 \\ &+2\left(|RD|\cos\left(\frac{1}{2}|\measuredangle EDF|\right) + |DA|\right)|PA|\cos\left(\frac{1}{2}|\measuredangle B'AC'|\right) \\ &+\left(|PA|\cos\left(\frac{1}{2}|\measuredangle B'AC'|\right)\right)^2 + \left(|PA|\sin\left(\frac{1}{2}|\measuredangle B'AC'|\right)\right)^2\end{aligned}$$

$$\begin{aligned}|RP'|^2 = &\left(|RD|\sin\left(\frac{1}{2}|\measuredangle EDF|\right)\right)^2 \\ &+\left(|RD|\cos\left(\frac{1}{2}|\measuredangle EDF|\right) + |DA|\right)^2 \\ &+2\left(|RD|\cos\left(\frac{1}{2}|\measuredangle EDF|\right) + |DA|\right)|PA|\cos\left(\frac{1}{2}|\measuredangle B'AC'|\right) + |PA|^2 \\ = &\,|RP|^2 - 2\left(|RD|\cos\left(\frac{1}{2}|\measuredangle EDF|\right) + |DA|\right)|PA|\cos(\frac{1}{2}|\measuredangle BAC|) \\ &+2\left(|RD|\cos\left(\frac{1}{2}|\measuredangle EDF|\right) + |DA|\right)|PA|\cos\left(\frac{1}{2}|\measuredangle B'AC'|\right) > |RP|^2\end{aligned}$$

as expected for Γ transformation for points of two different clusters.

Now the k-means with $k = 2$ yields a different partition, splitting line segments AB' and AC' (see Fig. 3 right). [1] □

A.3 Proof of Theorem 9

Proof In particular, let us consider the set of k clusters $\Gamma = \{C_1, \ldots, C_k\}$ of cardinalities $n_1, \ldots, n_k$ and with radii of balls enclosing the clusters (with centers located at cluster centers) $r_1, \ldots, r_k$.

We are interested in a gap g between clusters such that it does not make sense to split each cluster C_i into subclusters $C_{i1}, \ldots, C_{ik}$ and to combine them into a set of new clusters $\mathcal{S} = \{S_1, \ldots, S_k\}$ such that $S_j = \cup_{i=1}^{k} C_{ij}$.

We seek a g such that the highest possible central sum of squares combined over the clusters C_i would be lower than the lowest conceivable combined sums of squares around respective centers of clusters S_j. Let $Var(C)$ be the variance of the cluster C (average squared distance to cluster gravity center). Let r_{ij} be the distance of the center of subcluster C_{ij} to the center of cluster C_i. Let v_{ilj} be the distance of the center of subcluster C_{ij} to the center of subcluster C_{lj}. So the total k-means function for the set of clusters $(C_1, \ldots, C_k)$ will amount to:

$$Q(\Gamma) = \sum_{i=1}^{k} \sum_{j=1}^{k} (n_{ij} Var(C_{ij}) + n_{ij} r_{ij}^2) \tag{10}$$

And the total k-means function for the set of clusters $(S_1, \ldots, S_k)$ will amount to:

$$Q(\mathcal{S}) = \sum_{j=1}^{k} \left(\left(\sum_{i=1}^{k} n_{ij} Var(C_{ij}) \right) + \left(\sum_{i=1}^{k} n_{ij} \right) \left(\sum_{i=1}^{k-1} \sum_{l=i+1}^{k} \frac{n_{ij}}{\sum_{i=1}^{k} n_{ij}} \frac{n_{lj}}{\sum_{i=1}^{k} n_{ij}} v_{ilj}^2 \right) \right) \tag{11}$$

Should $(C_1, \ldots, C_k)$ constitute the absolute minimum of the k-means target function, then $Q(\mathcal{S}) \geq Q(\mathcal{C})$ should hold, that is:

$$\sum_{j=1}^{k} \left(\left(\sum_{i=1}^{k} n_{ij} Var(C_{ij}) \right) + \left(\sum_{i=1}^{k} n_{ij} \right) \left(\sum_{i=1}^{k-1} \sum_{l=i+1}^{k} \frac{n_{ij}}{\sum_{i=1}^{k} n_{ij}} \frac{n_{lj}}{\sum_{i=1}^{k} n_{ij}} v_{ilj}^2 \right) \right)$$

$$\geq \sum_{i=1}^{k} \sum_{j=1}^{k} (n_{ij} Var(C_{ij}) + n_{ij} r_{ij}^2)$$

This implies:

$$\sum_{j=1}^{k} \left(\sum_{i=1}^{k-1} \sum_{l=i+1}^{k} \frac{n_{ij} n_{lj}}{\sum_{i=1}^{k} n_{ij}} v_{ilj}^2 \right) \geq \sum_{i=1}^{k} \sum_{j=1}^{k} n_{ij} r_{ij}^2 \tag{12}$$

To maximize $\sum_{j=1}^{k} n_{ij} r_{ij}^2$ for a single cluster C_i of enclosing ball radius r_i, note that you should set r_{ij} to r_i. Let $m_j = \arg\max_{j \in \{1,\ldots,k\}} n_{ij}$. If we set $r_{ij} = r_i$ for all j except m_j,

[1] In a test run with 100 restarts, in the first case we got clusters of equal sizes, with cluster centers at (17,0,0) and (-17,0,0), (between_SS / total_SS = 40%) whereas after rotation we got clusters of sizes 1800, 2200 with centers at (26,0,0), (-15,0,0) (between_SS / total_SS = 59%)

then the maximal r_{im_j} is delimited by the relation $\sum_{j=1;j\neq m_j}^{k} n_{ij} r_{ij} \geq n_{im_j} r_{im_j}$. So

$$\sum_{j=1}^{k} n_{ij} r_{ij}^2 \leq \left(\sum_{j=1;j\neq m_j}^{k} n_{ij} \right) r_i^2 \min\left(2, (1 + \frac{\sum_{j=1;j\neq m_j}^{k} n_{ij}}{n_{im_j}})\right) \tag{13}$$
$$\leq 2 \left(\sum_{j=1;j\neq m_j}^{k} n_{ij} \right) r_i^2$$

So if we can guarantee that the gap between cluster balls (of clusters from Γ) amounts to g, then surely

$$\sum_{j=1}^{k} \left(\sum_{i=1}^{k-1} \sum_{l=i+1}^{k} \frac{n_{ij} n_{lj}}{\sum_{i=1}^{k} n_{ij}} v_{ilj}^2 \right) \geq g^2 \sum_{j=1}^{k} \left(\sum_{i=1}^{k-1} \sum_{l=i+1}^{k} \frac{n_{ij} n_{lj}}{\sum_{i=1}^{k} n_{ij}} \right) \tag{14}$$

because in such case $g \leq v_{ilj}$ for all i, l, j.

By combining inequalities (12), (13) and (14) we see that the global minimum is granted if the following holds:

$$g^2 \sum_{j=1}^{k} \left(\sum_{i=1}^{k-1} \sum_{l=i+1}^{k} \frac{n_{ij} n_{lj}}{\sum_{i=1}^{k} n_{ij}} \right) \geq 2 \sum_{i=1}^{k} \left(\sum_{j=1;j\neq m_j}^{k} n_{ij} \right) r_i^2 \tag{15}$$

One can distinguish two cases: either (1) there exists a cluster S_t containing two subclusters C_{pt}, C_{qt} such that $t = \arg\max_j |C_{pj}|$ and $t = \arg\max_j |C_{qj}|$ (maximum cardinality subclasses of their respective original clusters C_p, C_q) or (2) not.

Consider the first case. Let C_p, C_q be the two clusters where C_{pt} and C_{qt} be two subclusters of highest cardinality within C_p, C_q resp. This implies that $n_{pt} \geq \frac{1}{k} n_p, n_{qt} \geq \frac{1}{k} n_q$. Also this implies that for $i \neq p, i \neq q$ $n_{it} \leq n_i/2$.

$$\sum_{j=1}^{k} \sum_{i=1}^{k-1} \sum_{l=i+1}^{k} \frac{n_{ij} n_{lj}}{\sum_{i=1}^{k} n_{ij}} \geq \sum_{i=1}^{k-1} \sum_{l=i+1}^{k} \frac{n_{it} n_{lt}}{\sum_{i=1}^{k} n_{it}} \geq \frac{n_{pt} n_{qt}}{\sum_{i=1}^{k} n_{it}}$$
$$\geq \frac{n_{pt} n_{qt}}{n_p/2 + n_q/2 + \sum_{i=1}^{k} n_i/2} = \frac{n_{pt} n_{qt}}{n_p/2 + n_q/2 + n/2}$$
$$\geq \frac{1}{k^2} \frac{n_p n_q}{n_p/2 + n_q/2 + n/2}$$

Note that

$$2 \sum_{i=1}^{k} \left(\sum_{j=1;j\neq m_j}^{k} n_{ij} \right) r_i^2 \leq 2 \sum_{i=1}^{k} n_i r_i^2$$

So, in order to fulfill inequality (15), it is sufficient to require that

$$g \geq \sqrt{\frac{2 \sum_{i=1}^{k} n_i r_i^2}{\frac{1}{k^2} \frac{n_p n_q}{n_p/2 + n_q/2 + n/2}}} = k \sqrt{n_p/2 + n_q/2 + n/2} \sqrt{\frac{2 \sum_{i=1}^{k} n_i r_i^2}{n_p n_q}} \tag{16}$$
$$= k \sqrt{n_p + n_q + n} \sqrt{\frac{\sum_{i=1}^{k} n_i r_i^2}{n_p n_q}}$$

This of course maximized over all combinations of p, q.

Let us proceed to the second case. Here each cluster S_j contains a subcluster of maximum cardinality of a different cluster C_i. As the relation between S_j and C_i is unique, we can reindex S_j in such a way that actually C_j contains its maximum cardinality subcluster C_{jj}. Let us rewrite the inequality (15).

$$g^2 \sum_{j=1}^{k} \left(\sum_{i=1}^{k-1} \sum_{l=i+1}^{k} \frac{n_{ij} n_{lj}}{\sum_{i=1}^{k} n_{ij}} \right) - 2 \sum_{i=1}^{k} \left(\sum_{j=1; j \neq m_j}^{k} n_{ij} \right) r_i^2 \geq 0$$

This is met if

$$g^2 \sum_{j=1}^{k} \left(\sum_{i=1}^{j-1} \frac{n_{ij} n_{jj}}{\sum_{i=1}^{k} n_{ij}} + \sum_{l=j+1}^{k} \frac{n_{jj} n_{lj}}{\sum_{i=1}^{k} n_{ij}} \right) - 2 \sum_{i=1}^{k} (n_i - n_{ii}) r_i^2 \geq 0$$

This is the same as:

$$g^2 \sum_{j=1}^{k} \left(\sum_{i=1,\ldots,j-1,j+1,\ldots,k} \frac{n_{ij} n_{jj}}{\sum_{i=1}^{k} n_{ij}} \right) - 2 \sum_{i=1}^{k} (n_i - n_{ii}) r_i^2 \geq 0$$

This is fulfilled if:

$$g^2 \sum_{j=1}^{k} \left(\sum_{i=1,\ldots,j-1,j+1,\ldots,k} \frac{n_{ij} n_j / k}{n_j/2 + \sum_{i=1}^{k} n_i/2} \right) - 2 \sum_{i=1}^{k} (n_i - n_{ii}) r_i^2 \geq 0$$

Let M be the maximum over $n_1, \ldots, n_k$. The above holds if

$$g^2 \sum_{j=1}^{k} \left(\sum_{i=1,\ldots,j-1,j+1,\ldots,k} \frac{n_{ij} n_j / k}{M/2 + n/2} \right) - 2 \sum_{i=1}^{k} (n_i - n_{ii}) r_i^2 \geq 0$$

Let m be the minimum over $n_1, \ldots, n_k$. The above holds if

$$g^2 \sum_{j=1}^{k} \left(\sum_{i=1,\ldots,j-1,j+1,\ldots,k} \frac{n_{ij} m / k}{M/2 + n/2} \right) - 2 \sum_{i=1}^{k} (n_i - n_{ii}) r_i^2 \geq 0$$

This is the same as

$$g^2 \frac{m/k}{M/2 + n/2} \left(\sum_{j=1}^{k} \sum_{i=1,\ldots,j-1,j+1,\ldots,k} n_{ij} \right) - 2 \sum_{i=1}^{k} (n_i - n_{ii}) r_i^2 \geq 0$$

$$g^2 \frac{m/k}{M/2 + n/2} \left(\sum_{j=1}^{k} \left(\left(\sum_{i=1}^{k} n_{ij} \right) - n_{jj} \right) - 2 \sum_{i=1}^{k} (n_i - n_{ii}) r_i^2 \right) \geq 0$$

$$g^2 \frac{m/k}{M/2 + n/2} \left(\left(\sum_{j=1}^{k} \sum_{i=1}^{k} n_{ij} \right) - \left(\sum_{j=1}^{k} n_{jj} \right) \right) - 2 \left(\sum_{i=1}^{k} (n_i - n_{ii}) r_i^2 \right) \geq 0$$

$$g^2 \frac{m/k}{M/2 + n/2} \left(\left(\sum_{i=1}^{k} n_i \right) - \left(\sum_{j=1}^{k} n_{jj} \right) \right) - 2 \sum_{i=1}^{k} (n_i - n_{ii}) r_i^2 \geq 0$$

$$g^2 \frac{m/k}{M/2 + n/2} \left(\sum_{i=1}^{k} (n_i - n_{ii}) \right) - 2 \sum_{i=1}^{k} (n_i - n_{ii}) r_i^2 \geq 0$$

$$\sum_{i=1}^{k}(n_i - n_{ii})\left(g^2\frac{m/k}{M/2+n/2} - 2r_i^2\right) \geq 0$$

The above will hold, if for every $i = 1, \ldots, k$

$$g \geq r_i\sqrt{\frac{2}{\frac{m/k}{M/2+n/2}}}$$

$$g \geq r_i\sqrt{k\frac{M+n}{m}} \tag{17}$$

So the inequality (15) is fulfilled, if both inequality (16) and inequality (17) are held by an appropriately chosen g. But relation (17) is identical with (7), and (16) is identical with (6), □

Acknowledgements We would like to acknowledge support for this project from the Polish government fundamental research funds.

Funding This research was funded by the Polish government fundamental research funds.

Availability of data and material Only data generated as described in the paper were used.

Declarations

Conflict of Interests The authors declare that they have no conflict of interest.

References

Ackerman, M., Ben-David, S., & Loker, D. (2010). Characterization of linkage-based clustering. In *COLT 2010* (pp. 270–281).

Ben-David, S., & Ackerman, M. (2008). Measures of clustering quality: A working set of axioms for clustering. In *Proc. Advances in Neural Information Processing Systems*, (Vol. 21 pp. 121–128).

Ben-David, S., & Ackerman, M. (2009). Measures of clustering quality: a working set of axioms for clustering. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.) Advances in neural information processing systems, (Vol. 21 pp. 121–128). Curran Associates Inc.

Carlsson, G., & Mémoli, F. (2008). Persistent clustering and a theorem of J. Kleinberg. arXiv:08082241.

Carlsson, G., & Mémoli, F. (2010). Characterization, stability and convergence of hierarchical clustering methods. *Journal of Machine Learning Research*, *11*, 1425–1470.

Correa-Morrisa, J. (2013). An indication of unification for different clustering approaches. *Pattern Recognition*, *46*, 2548–2561.

Gower, J. C. (1990). Clustering axioms. Classification Society of North America Newsletter, pp 2–3.

Hopcroft, J., & Kannan, R. (2012). Computer science theory for the information age. Chapter 8.13.2. A Satisfiable Set of Axioms, p 272ff.

Kleinberg, J. (2002). An impossibility theorem for clustering. In *Proc. NIPS*, (Vol. 2002 pp. 446–453). http://books.nips.cc/papers/files/nips15/LT17.pdf.

Klopotek, M. A., & Klopotek, R. A. (2020). Clustering algorithm consistency in fixed dimensional spaces. In D. Helic, G. Leitner, M. Stettinger, A. Felfernig, & Z.W. Ras (Eds.) *Foundations of intelligent systems - 25th international symposium, ISMIS 2020, Graz, Austria, September 23–25, 2020, Proceedings, Springer, Lecture notes in computer science*, (Vol. 12117 pp. 352-361), https://doi.org/10.1007/978-3-030-59491-6_33.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proc. fifth Berkeley symp. on math. Statist. and Prob.*, (Vol. 1 pp. 281–297). University of California Press.

Shekar, B. (1988). A knowledge-based approach to pattern clustering. PhD thesis, Indian Institute of Science.

Strazzeri, F., & Sánchez-García, R. J. (2018). Morse theory and an impossibility theorem for graph clustering. arXiv:1806.06142.

Thomann, P., Steinwart, I., & Schmid, N. (2015). Towards an axiomatic approach to hierarchical clustering of measures. *Journal of Machine Learning Research*, *16*, 1949–2002.

van Laarhoven, T., & Marchiori, E. (2014). Axioms for graph clustering quality functions. *Journal of Machine Learning Research*, *15*, 193–215.

Wei, Jh. (2017). Two examples to show how k-means reaches richness and consistency. DEStech Transactions on Computer Science and Engineering https://doi.org/10.12783/dtcse/aita2017/16001.

Wierzchoń, S., & Kłopotek, M. (2018). Modern clustering algorithms. Studies in Big Data 34, Springer.

Wright, W. (1973). A formalization of cluster analysis. *Pattern Rec*, *5*(3), 273–282.

Zadeh, R. (2010). Towards a principled theory of clustering. http://stanford.edu/rezab/papers/principled.pdf.

Zeng, G., Wang, Y., Pu, J., Liu, X., Sun, X., & Zhang, J. (2016). Communities in preference networks: Refined axioms and beyond. In *ICDM*, (Vol. 2016 pp. 599–608).