sciendo
Control and Cybernetics
vol. **51** (2022) No. 3
pages: 343-387
DOI: 10.2478/candc-2022-0021

Clustering of data represented by pairwise comparisons*

by

Sergey Dvoenko

Tula State University, Tula, Russia

Abstract: In this paper, experimental data, given in the form of pairwise comparisons, such as distances or similarities, are considered. Clustering algorithms for processing such data are developed based on the well-known k-means procedure. Relations to factor analysis are shown. The problems of improving clustering quality and of finding the proper number of clusters in the case of pairwise comparisons are considered. Illustrative examples are provided.

Keywords: clustering, k-means, distance, similarity

1. The cluster analysis paradigm

1.1. The idea of clustering

The well-known research paradigm proposes that a phenomenon under study can be in one of a finite number of hidden states, and can be observed while being in such different states. These states exert the influence on the values of the measured characteristics (features, variables). Therefore, it is important to understand how this influence is realized and how the set of states is organized.

According to the above, it can be assumed that all the observations, represented as points in a feature space are arranged in such a way that they form local concentrations (clusters, classes, taxons, etc.) to be identified as corresponding to the hidden states.

This assumption is also known as an informal so-called “compactness hypothesis” (Aizerman, Braverman and Rozonoer, 1970), in the framework of which such hidden states are named “patterns”. The compactness hypothesis can be used as some sort of fundamental principle, establishing the condition for what we can extract from experimental data in data analysis and, particularly, in clustering. Such an understanding of the situation arose after the impressive

*Submitted: August 2022; Accepted: September 2022.

investigation of F. Rosenblatt (Rosenblatt, 1962) and the post-analysis of his "perceptron" failures.

On the other hand, the phenomenon under study can be not structured internally sufficiently to reveal its properties at a meaningful level. This means, for example, that our measurement capabilities may not allow us to distinguish between results for different hidden states. Unfortunately, in this case, the investigations carried out would be unsuccessful, because the hidden states cannot be recognized.

In cluster analysis, objects $\omega \in \Omega$ are explored simultaneously in order to understand how close are they to each other and how the hidden states get expressed. Usually, this can be done on the basis of comparison of some characteristics (features) of elements ω. The measurement results of these characteristics are usually represented by the data matrix $X(m, n)$ with m as the number of the set elements ω (experiments, objects, etc.) and n as the number of measured characteristics (features, variables, peculiarities, attributes, etc.).

According to the well-known model, used in cluster analysis, the data matrix X is represented by rows, denoted $\mathbf{x}_i = (x_{i1}, \dots x_{in})$, $i = 1, \dots m$, representing vectors in the n-dimensional feature (attribute) space, very often assumed to be Euclidean. Here, $\mathbf{x}_i^T$ is a column according to the notation above. Therefore, each element $\omega \in \Omega$ is represented by means of measurements $\mathbf{x}_i = \mathbf{x}(\omega_i)$ in the corresponding feature space.

It is supposed that elements $\omega_i \in \Omega$ belong to non-intersecting subsets Ω_k, $k = 1, \dots K$, $\Omega_i \cap \Omega_j = \emptyset$, $i \neq j$. These subsets form the corresponding sufficiently well-defined local concentrations in the feature space (clusters). Our goal is to uncover the unknown cluster structure for known K, or, in the more general and also more complex case, to determine K and uncover the unknown cluster structure.

The present paper is organized in the following way. In Section 1 the basic terms of cluster analysis are investigated, namely some criteria, data representation in the form of pairwise comparisons, and the peculiarity of clustering based on pairwise comparisons.

In Section 2 operations on pairwise comparisons are introduced, based on the law of cosines as a foundation for developing clustering algorithms. The crucial idea consists in determining the averages as new objects, not present in the set before.

Section 3 is devoted to the development of new versions of the k-means algorithm for distances and similarities.

In Section 4 relations of cluster analysis to some other problems of data analysis are investigated. Specifically, conditions are defined for solving the problem of factor analysis as the one of clustering.

Section 5 deals with the problem of improving the clustering results. In Owsiński (2020) in the framework of a general approach to clustering, a new bi-partial objective function was proposed. Based on it, in this section, we investigate a specific case of improving the clustering results for the special version of the permutable k-means algorithm.

Section 6 is devoted to the well-known problem of establishing the number of clusters, here on the basis of the specially devised so-called quasi-hierarchical procedure.

1.2. The clustering principle

The clustering principle that we shall refer to throughout the paper is as follows. Let the cluster representatives $\tilde{\mathbf{x}}_k,\ k = 1, \dots K$, be defined in some way. Let all objects be allocated between clusters based on the closest representative for each object $\mathbf{x}_i \in \Omega$. After that, let the cluster centers be calculated as arithmetic averages $\bar{\mathbf{x}}_k,\ k = 1, \dots K$.

If all representatives coincide with centers $\tilde{\mathbf{x}}_k = \bar{\mathbf{x}}_k,\ k = 1, \dots K$, the result is the so-called unbiased clustering (Diday et al., 1979). Otherwise, one deals with a biased clustering. Therefore, if representatives and centers for some clusters do not coincide, the centers are appointed next as new representatives. After that, clusters need to be redefined as above. In general, for unbiased clustering, the center $\bar{\mathbf{x}}_k$ of the cluster Ω_k may not match any element $\mathbf{x}_i \in \Omega_k$ in this cluster. The well-known and widely used k-means algorithm is directly developed based on this principle (see, e.g., Hartigan and Wong, 1979).

As it is known, the k-means algorithm is a locally optimal procedure. The quality of the initial solution (the initial set of representatives) in such procedures is very important. It is known that finding the acceptable initial solution is a standalone and sometimes no less complex problem, than the clustering algorithm itself.

Here and below, we do not consider this problem, but simply assume one of the suitable initial solutions when formulating new algorithms.

Let us remind that the arithmetic average as the cluster center ensures that the cluster dispersion (variance) is minimized. Therefore, the dispersion of the clustering as a whole is minimized, too. The dispersion of the cluster Ω_k is defined by squared distances between the cluster members and its center

$$\sigma_k^2 = (1/m_k) \sum\nolimits_{i=1}^{m_k} |\mathbf{x}_i - \bar{\mathbf{x}}_k|^2 = (1/m_k) \sum\nolimits_{i=1}^{m_k} d^2(\mathbf{x}_i, \bar{\mathbf{x}}_k).$$

The well-known criterion of clustering can also be defined as the weighted average dispersion of clusters to be minimized, in the following form:

$$J(K) = \frac{1}{m} \sum_{k=1}^{K} \sum_{i=1}^{m_k} |\mathbf{x}_i - \bar{\mathbf{x}}_k|^2 = \frac{1}{m} \sum_{k=1}^{K} m_k \sigma_k^2 = \sum_{k=1}^{K} \frac{m_k}{m} \sigma_k^2. \tag{1}$$

It is also known that the cluster dispersion can be calculated without the direct use of the cluster mean, based on pair distances between vectors (see Friedman and Rubin, 1967; Späth, 1983) and can be determined as half of the average of distances squared between the cluster members

$$\eta_k = \frac{1}{2m_k^2} \sum_{i=1}^{m_k} \sum_{j=1}^{m_k} d^2(\mathbf{x}_i, \mathbf{x}_j). \tag{2}$$

Therefore, another criterion of clustering can be defined as the weighted average distances in clusters to be minimized in the following form:

$$\tilde{J}(K) = \frac{1}{m} \sum_{k=1}^{K} m_k \eta_k = \sum_{k=1}^{K} \frac{m_k}{m} \eta_k. \tag{3}$$

It should be noted also that the well-known EM (expectation maximization) procedure is considered as the probability-theoretical counterpart or justification of the k-means algorithm. This procedure was also considered in one of the early publications (Schlesinger, 1965). This iterative procedure consists of two steps: the E-step (expectation) of finding the optimal a posteriori probabilities of separating a mixture of distributions of observations with given parameter distribution and the M-step (maximization) of determining the optimal parameter distribution to maximize the log-likelihood function. In the k-means algorithm, the E-step corresponds to the assignment of observations among clusters and the M-step corresponds to the calculation of new averages of clusters. Assuming a mixture of normal distributions with the same variances, we obtain linear boundaries between regions of disjoint clusters in the feature space. The k-means algorithm implicitly defines exactly such boundaries as a special case of the so-called hyper quadrics (see Duda and Hart, 1973).

1.3. Pairwise comparisons of the set elements

Dually to the representation through objects (rows), the data matrix can be represented by columns $X_j = (x_{1j}, \dots x_{mj})^T$, $j = 1, \dots n$. Here, X_j^T is a row, according to the notation above. This representation is usually admitted in correlation, factor, etc., analysis in exploring the similarity of features (measured characteristics). Feature similarity means the similarity of their behavior relative to the set of objects (acts of measurements).

Based on two representations, by rows and by columns, the data matrix $X(m,n)$ can be transformed into a distance matrix $D(m,m)$ for objects and $D(n,n)$ for features, or the scalar products matrix $S(m,m)$ for objects and $S(n,n)$ for features. Traditionally, $D(m,m)$ and $S(n,n)$ are commonly used, since distances between objects are arising in a very natural manner as the result of comparison in multidimensional space (like in our 3-dimensional world), and normalized scalar products, leading to correlations $R(n,n)$ between features are very natural in the case of comparison of behavior of variation series.

As we can see, distances, as the results of pairwise comparisons, are being usually applied to the set elements as objects. The pairs of objects can also be characterized by the non-negative similarity functions, as opposed to distances. For example, many empirical similarity functions were developed as inverses of the Euclidean distance. At the same time, based on the law of cosines, we can characterize pairs of objects by their scalar products, instead of distances between them. As a result, another “ad hoc” popular idea of similarity consists in using modules or squared scalar products. Such “ad hoc” similarities allow us to assume that the objects are located in the single quadrant of the coordinate space.

If all objects are located only in the single quadrant of the coordinate space, then all scalar products between them are represented by non-negative values. Therefore, such scalar products can be used as similarities.

Here, with the purpose of solving the clustering problems, we talk about the set elements $\omega \in \Omega$, without distinction whether they are objects or features. Nevertheless, in the case of using the squared scalar products for clustering of features, we face a class of algorithms other than based on the k-means algorithm (Dvoenko, 2009a). We discuss below the issue why we get the biased clustering in this case of similarity.

It should be noted also that our assumption of non-negative scalar products between the set elements (objects or features) does not narrow the domain of clustering at all. Indeed, as we show below, to calculate scalar products, it is necessary to determine the position of the origin. It can always be arranged in such a way that all scalar products become non-negative relative to it (for example, the origin of coordinates can be moved outside the convex hull of the set).

On the other hand, there are many cases, when data can be presented by pairwise comparisons only, since, for instance, sometimes it is difficult to decide what characteristics need to be measured for the complex structured entity under investigation. In this case, it is suitable to measure pairwise similarities or dissimilarities directly and use them, for example, in the form of positive scalar products or Euclidean distances, respectively. Therefore, we suppose hypothetical (unknown to us) features were measured, and pair distances or similarities were calculated. Therefore, based on such a hypothetical measurement process, the set of elements is immersed in some coordinate space.[*]

This coordinate space can be characterized as follows. Usually, it is natural to use it as the Euclidean metric space with the dimensionality not more than the set Ω cardinality (e.g., m for objects or n for features themselves).

In the case the space dimensionality is exactly equal to the cardinality of the set Ω, the similarity matrices $S(m, m)$ or $S(n, n)$ are positive definite with the corresponding ranks (m or n, accordingly) (Mercer, 1909; Young and Householder, 1938).

[*] It is obvious that there exist situations, in which pairwise data are the only data available for the task at hand, like with railway distances or road distance in the city, with one-way streets taken into consideration (eds.).

In practice, the coordinate space can be formed by features measured in scales of different types. Therefore, a problem of calculating distances or similarities in such spaces can arise for multiple scale types. In the case of pairwise comparisons, we suppose that the set elements are immersed in the metric space with coordinate axes of the ratio type. Hence, distances or similarities are calculated in the ratio scales.

It should be noted that if the similarity matrix is not positive definite, then the set of elements cannot be correctly immersed in the respective coordinate space. In this case, the similarity matrix has negative eigenvalues. The corresponding immersing problem is discussed in more detail in Dvoenko and Pshenichny (2018).

We should also remark that a similarity function can be represented by some kind of a potential function (so-called "kernels" in modern analysis). It is supposed in this case, according to Mercer's statement (Mercer, 1909), that the set of elements can be immersed in the countably-dimensional metric space in general with a scalar product defined in it (so-called "straightened space"), see Aizerman, Braverman and Rozonoer (1970).

Since pairwise comparisons are represented by distances and similarities, we talk in the further course of this paper about clustering algorithms (here, those based on classical k-means in a feature space) in two forms: as *distance* k-means and in a dual form as *similarity* k-means. This is the basis for developing some other new modifications of the procedure.

1.4. The peculiarity of clustering based on pairwise comparisons

The classical k-means algorithm is formulated for the set of objects represented by feature vectors $\mathbf{x}_i \in \Omega$. Let distances $D(m, m)$ and similarities $S(m, m)$ be calculated based on the data matrix $X(m, n)$ or simply given. We need to develop the so-called *distance* (or *similarity*) k-means algorithm using only distances $D(m, m)$ or similarities $S(m, m)$ without the reference to the matrix $X(m, n)$ at all.

First, to clarify the problem, let us build the natural, but somewhat naïve clustering procedure (a):

(a) Step $s = 0$. Define representatives $\tilde{\omega}_k^s$, $k = 1, \dots K$, perhaps as most distant objects, $s = s + 1$.

Step s. Allocate objects between clusters:

1. $\omega_i \in \Omega_k^s$, $k = \arg\min_{j=1,\dots K} d(\omega_i, \tilde{\omega}_j^s)$,$i = 1, \dots m$.

2. Calculate new centers: $\bar{\omega}_k^s = \arg\min_{\omega_i \in \Omega_k^s} \sum\limits_{\omega_j \in \Omega_k^s} d(\omega_i, \omega_j)$, $k = 1, \dots K$.

3. Stop, if $\tilde{\omega}_k^s = \bar{\omega}_k^s$, $k = 1, \dots K$. Else $\tilde{\omega}_k^{s+1} = \bar{\omega}_k^s$, $k = 1, \dots K$, $s = s + 1$.

As we can see, the problem consists of the following. Namely, in the algorithm (a) we cannot represent objects $\omega \in \Omega$ as vectors $\mathbf{x} = \mathbf{x}(\omega)$, we have only objects as $\omega = \omega(\mathbf{x})$. The centers $\bar{\omega} = \omega(\bar{\mathbf{x}})$ are not present in the distance matrix $D(m, m)$ and we cannot calculate them yet. Hence, it is natural to use as the unknown center $\bar{\omega}_k$ some object closest to all others in the cluster.

As a result, the algorithm (a) stops, when the unbiased clustering is reached for $\tilde{\omega}_k = \bar{\omega}_k$, $k = 1, \dots K$. Nevertheless, the actually obtained clustering would most probably be biased since in the feature space some cluster center may not coincide with the corresponding mean vector $\mathbf{x}(\bar{\omega}_k) \neq \bar{\mathbf{x}}_k$. This algorithm calculates the criterion $J^D(K) = \min_{\bar{\omega}_1, \dots \bar{\omega}_K} J(K)$ instead of the criterion $J^X(K) = \min_{\bar{\mathbf{x}}_1, \dots \bar{\mathbf{x}}_K} J(K)$. Therefore, $J^D(K) \geq J^X(K)$ in general.

Each object is represented in the distance matrix $D(m, m)$ by its distances to others. In order to achieve $J^D(K) = J^X(K)$ in the case the feature space is not available, it is necessary to define cluster centers $\bar{\omega}_k$ as new objects, represented by their distances to the other ones.

In the next section, we introduce the basic issues of immersion of a set of observations in a metric space as a basis for developing the *distance* and *similarity* versions of the k-means algorithm.

2. Immersion of a set in a metric space

2.1. The law of cosines and the origin

Let the elements $\omega_i \in \Omega$ be immersed in the metric space and let them be represented by the distance matrix $D(m, m)$. Let some triangle be formed by the objects $\omega_a \in \Omega$ and $\omega_b \in \Omega$ as two points, and the third object ω_0 being the origin of the metric space, with distances

$d_{0a} = d(\omega_0, \omega_a), \quad d_{0b} = d(\omega_0, \omega_b)$, and

$$d_{ab}^2 = d^2(\omega_a, \omega_b) = d_{0a}^2 + d_{0b}^2 - 2s_{ab}$$

according to the law of cosines, where

$$s_{ab} = \omega_a \circ \omega_b = (d_{0a}^2 + d_{0b}^2 - d_{ab}^2)/2$$

is the scalar product.

Since any element $\omega_k \in \Omega$ can be used as the origin, any pair of elements $\omega_i \in \Omega$, $\omega_j \in \Omega$ is represented relative to it by a scalar product

$(\omega_i \circ \omega_j)_k = s_{ij}^k = (d_{ki}^2 + d_{kj}^2 - d_{ij}^2)/2,$

where $(\omega_i \circ \omega_i)_k = s_{ii}^k = (d_{ki}^2 + d_{ki}^2 - d_{ii}^2)/2 = d_{ki}^2$.

Therefore, the non-normalized scalar products represent the set configuration in the metric space with distances between the set elements and the origin. As a result, we have different scalar product matrices $S_k(m, m)$, $k = 1, \dots m$, with main diagonals, which represent distances squared from corresponding origins ω_k, $k = 1, \dots m$, to other elements ω_i, $i = 1, \dots m$.

Unfortunately, in all cases, each origin ω_k becomes the degenerated ("singular") object since $(\omega_k \circ \omega_k)_k = s_{kk}^k = d_{kk}^2 = 0$. This is not suitable, since the rank of the similarity matrix $S_k(m,m)$ becomes less than its dimensionality $rank\ S_k < m$. Therefore, the matrix becomes positive semidefinite with at least one zero line and column for scalar products of ω_k with the other elements.

On the other hand, let the elements $\omega_i \in \Omega$ be represented by the positive definite scalar product matrix $S(m,m)$. Therefore, such scalar products are calculated relative to some (unknown to us) origin ω_0. At any place in the metric space as the location of the origin, we can get all distances

$$d_{ij}^2 = d_{0i}^2 + d_{0j}^2 - 2s_{ij}^0 = s_{ii}^0 + s_{jj}^0 - 2s_{ij}^0 \quad = s_{ii} + s_{jj} - 2s_{ij}.$$

It is evident that

$$d_{ii}^2 = s_{ii} + s_{ii} - 2s_{ii} = 0.$$

Therefore, it is suitable to define the origin as a new object ω_0, not coinciding with other objects $\omega_i \in \Omega$.

2.2. Representation of the origin as a new object

Let the data matrix $X(m,n)$ be given. It means that some initial origin ω_0 has been defined in the feature space. Let us define a new object ω_α immersed in the feature space as a linear combination

$$\mathbf{x}_\alpha = \sum\nolimits_{i=1}^m \alpha_i \mathbf{x}_i, \quad \sum\nolimits_{i=1}^m \alpha_i = 1,\ \alpha_i \geq 0$$

as it would be the result of some measurement process, $\mathbf{x}_\alpha = \mathbf{x}(\omega_\alpha)$.

Let us center the objects in the matrix $X(m,n)$ relative to $\mathbf{x}_\alpha$, used now as the origin and define vectors $\mathbf{x}_i - \mathbf{x}_\alpha$, $i = 1, ...\ m$. Let us define the new scalar products of centered vectors as

$$\begin{aligned} s_{ij}^\alpha &= \sum_{l=1}^n (x_{il} - x_{\alpha l})(x_{jl} - x_{\alpha l}) = \sum_{l=1}^n (x_{il}x_{jl} - x_{il}x_{\alpha l} - x_{jl}x_{\alpha l} + x_{\alpha l}^2) \\ &= \sum_{l=1}^n x_{il}x_{jl} - \sum_{l=1}^n x_{il} \sum_{p=1}^m \alpha_p x_{pl} - \sum_{l=1}^n x_{jl} \sum_{p=1}^m \alpha_p x_{pl} + \sum_{l=1}^n \left(\sum_{p=1}^m \alpha_p x_{pl} \right) \left(\sum_{q=1}^m \alpha_q x_{ql} \right) \\ &= \sum_{l=1}^n x_{il}x_{jl} - \sum_{p=1}^m \alpha_p \left(\sum_{l=1}^n x_{il}x_{pl} + \sum_{l=1}^n x_{jl}x_{pl} \right) + \sum_{p=1}^m \sum_{q=1}^m \alpha_p \alpha_q \sum_{l=1}^n x_{pl}x_{ql}. \end{aligned}$$

Finally,

$$s_{ij}^\alpha = s_{ij} - \sum_{p=1}^m \alpha_p (s_{ip} + s_{jp}) + \sum_{p=1}^m \sum_{q=1}^m \alpha_p \alpha_q s_{pq}. \tag{4}$$

Basing on the law of cosines, scalar products relative to the initial origin ω_0 are defined as $s_{ij} = (d_{0i}^2 + d_{0j}^2 - d_{ij}^2)/2$. Substituting this in (4) gives scalar products relative to the new origin ω_α as

$$2s_{ij}^\alpha = d_{0i}^2 + d_{0j}^2 - d_{ij}^2 - \sum_{p=1}^{m} \alpha_p (d_{0i}^2 + d_{0p}^2 - d_{ip}^2 + d_{0j}^2 + d_{0p}^2 - d_{jp}^2) \\ + \sum_{p=1}^{m}\sum_{q=1}^{m} \alpha_p \alpha_q (d_{0p}^2 + d_{0q}^2 - d_{pq}^2).$$

After opening of brackets and bringing similar terms together, we get the following:

$$2s_{ij}^\alpha = -d_{ij}^2 + \sum_{p=1}^{m} \alpha_p d_{ip}^2 + \sum_{p=1}^{m} \alpha_p d_{jp}^2 - \sum_{p=1}^{m}\sum_{q=1}^{m} \alpha_p \alpha_q d_{pq}^2.$$

According to the law of cosines, $s_{ii} = d_{0i}^2$ for $i = j$. Therefore, distances squared $d^2(\omega_\alpha, \omega_i) = d_{\alpha i}^2 = s_{ii}^\alpha$ from the new origin ω_α to the other elements in the set Ω are finally defined as

$$d^2(\omega_\alpha, \omega_i) = \sum_{p=1}^{m} \alpha_p d_{ip}^2 - \frac{1}{2}\sum_{p=1}^{m}\sum_{q=1}^{m} \alpha_p \alpha_q d_{pq}^2, \quad i = 1, \dots m. \tag{5}$$

2.3. The Torgerson's origin

It is known that W.S. Torgerson did successfully develop the foundations of the multidimensional scaling theory. Today, his method, constituting the basis for his metric scaling is known as "principal projections" (Torgerson, 1958). His idea consists in calculating the so-called "gravity center" $\bar{\omega}$ of the set Ω and putting the origin in it. His scaling theory was criticized for too strong metric limitations and started a new direction of development of methods of non-metric scaling, as well as further-reaching research.

With our purpose in mind, we should like to generate the new object ω_0 in the metric space as the arithmetic average, in order to put the origin in it and to maintain the scalar products matrix $S(m, m)$ positive definite. We should note that our problem is different from the multidimensional scaling problem, since it does not aim at restoring the so-called "stimuli space." It is sufficient, as it is shown below, to use only pairwise comparisons to build the known clustering algorithms from (but, in general, not limited to) the k-means family.

As it is evident, the new origin ω_α, defined above, can be any point within the convex hull of the set Ω as the appropriate linear combination. Specifically, let $\alpha_i = 1$, $\alpha_{i \neq j} = 0$, $j = 1, \dots m$. We get $d^2(\omega_\alpha, \omega_i) = d_{ij}^2$, $j = 1, \dots m$, since the origin ω_α is the point ω_i.

Therefore, this new object can be not only an element from the set Ω, but also a new one, not belonging to Ω before. On the other hand, we can get the linear combination of any subset from the Ω, and we discuss it below.

Let the new origin be the arithmetic mean $\bar{\omega}$, where $\alpha_i = 1/m$, $i = 1, \dots m$. Immediately, we get this element represented by the distances to other elements as

$$d^2(\omega_\alpha, \omega_i) = d^2(\bar{\omega}, \omega_i) = \frac{1}{m}\sum_{p=1}^{m} d_{ip}^2 - \frac{1}{2m^2}\sum_{p=1}^{m}\sum_{q=1}^{m} d_{pq}^2, \quad i = 1, \dots m. \tag{6}$$

This partial case of the linear combination with equal weights leads us to the known Torgerson's "gravity center" $\bar{\omega}$ as the new origin $\omega_\alpha = \bar{\omega}$ to represent it as a new object (not included in the set Ω before) by its distances to other set elements $\omega_i,\ i = 1, \dots m$.

Let for the data matrix $X(m, n)$ the matrix $S(m, m)$ of scalar products between objects be calculated relative to some initial origin ω_0. Let us define the scalar products of the mean vector

$$\bar{\mathbf{x}} = (\bar{x}_1, \dots \bar{x}_n), \quad \bar{x}_l = (1/m)\sum\nolimits_{p=1}^{m} x_{pl}, \quad l = 1, \dots n$$

as

$$\bar{\omega} \circ \omega_i = \sum_{l=1}^{n} x_{il}\frac{1}{m}\sum_{p=1}^{m} x_{pl} = \frac{1}{m}\sum_{p=1}^{m}\sum_{l=1}^{n} x_{il}x_{pl} = \frac{1}{m}\sum_{p=1}^{m} s_{ip}, \quad i = 1, \dots m.$$

Therefore, the arithmetic average $\bar{\omega}$ is represented as a new object (not included in the set Ω before) also by its scalar products with other set elements $\omega_i,\ i = 1, \dots m$.

Nevertheless, the correct use of scalar products $\bar{\omega} \circ \omega_i$ as similarities requires of them to be positive in order to represent the mean object $\bar{\omega}$. Therefore, it is necessary to move the origin outside the convex hull of the set, so that all scalar products relative to it would become positive.

2.4. Moving the origin outside the convex hull of the set

Let the set Ω be represented by the distance matrix $D(m, m)$. Let the center of the set Ω be considered as the Torgerson's origin ω_0, represented by its distances to other objects as

$$d_{0i}^2 = \frac{1}{m}\sum_{p=1}^{m} d_{ip}^2 - \Delta, \quad \Delta = \frac{1}{2m^2}\sum_{p=1}^{m}\sum_{q=1}^{m} d_{pq}^2, \quad i = 1, \dots m.$$

In the Torgerson's formula, the component $\Delta = \sigma_\Omega^2$ is the set dispersion as a scatter relative to the origin ω_0, since

$$\sigma_\Omega^2 = \frac{1}{m}\sum_{i=1}^{m} d_{0i}^2 = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{m}\sum_{p=1}^{m} d_{ip}^2 - \frac{1}{2m^2}\sum_{p=1}^{m}\sum_{q=1}^{m} d_{pq}^2\right) =$$

$$\frac{1}{m^2}\sum_{i=1}^{m}\sum_{p=1}^{m} d_{ip}^2 - \frac{1}{2m^2}\sum_{p=1}^{m}\sum_{q=1}^{m} d_{pq}^2 = \frac{1}{2m^2}\sum_{p=1}^{m}\sum_{q=1}^{m} d_{pq}^2. \tag{7}$$

Let us define now the new origin ω_θ, represented by its distances to other objects, $d_{\theta i}^2$, $i = 1,...,m$, according to the Torgerson's formula as above, but with $\Delta = 0$. In this case, the distances $d_{\theta i}^2$ become longer than the distances d_{0i}^2. It is evident that the new origin, ω_θ, is not the Torgerson's origin ω_0. It is convenient to assume that the origin ω_0 has been moved to the outside of the convex hull of the set Ω to a new position, ω_θ. According to this reasoning, we can understand $\Delta = 0$ as a very small scatter of elements relative to the new origin ω_θ.

It is necessary to provide at least non-negative scalar products $s_{ij} \geq 0$ between the set elements immersed in the metric space as vectors in order to put them in a single quadrant. Since

$$s_{ij} = \frac{1}{2d_{\theta i}d_{\theta j}}(d_{\theta i}^2 + d_{\theta j}^2 - d_{ij}^2)$$

relative to the new origin ω_θ according to the law of cosines, there should be

$d_{ij}^2 \leq d_{\theta i}^2 + d_{\theta j}^2$ for all $i, j = 1,... m$.

If this condition is violated, it is necessary to put $\Delta < 0$, for example, $\Delta = \min_{ij}(d_{\theta i}^2 + d_{\theta j}^2 - d_{ij}^2)$, so as to get longer distances to the new origin.

In this case, the origin of coordinates is moved outside the convex hull of the set of observations considered. Therefore, the elements of this set are placed in the single quadrant of the metric space. Such transfer of the origin allows for using the similarity function according to the law of cosines.

2.5. Representation of a subset center

According to Torgerson's formula, (6), (7), the center of the set Ω is defined by its distances to other elements as

$$d_{0i}^2 = (1/m)\sum\nolimits_{p=1}^{m} d_{ip}^2 - \sigma_\Omega^2, \quad i = 1,... m.$$

It is a new object, generally not coinciding with other elements in the set Ω.

Let us take some subset $\Omega_{\bar{0}} \subseteq \Omega$. It must be noted that it can be any subset of elements mixed in geometric sense with other elements of the set Ω within the metric space (elements from $\Omega_{\bar{0}}$ are distributed among the elements from Ω/Ω_0 in the space). According to Torgerson's formula, the center $\omega_{\bar{0}}$ of the subset $\Omega_{\bar{0}} \subseteq \Omega$ is represented by its distances to all elements of the whole set Ω in the following manner

$$d_{\bar{0}i}^2 = \frac{1}{m_{\bar{0}}}\sum_{p\in\Omega_{\bar{0}}} d_{ip}^2 - \sigma_{\Omega_{\bar{0}}}^2, \quad \sigma_{\Omega_{\bar{0}}}^2 = \frac{1}{2m_{\bar{0}}^2}\sum_{p\in\Omega_{\bar{0}}}\sum_{q\in\Omega_{\bar{0}}} d_{pq}^2, \quad m_{\bar{0}} = |\Omega_{\bar{0}}|,$$
$$i = 1,... m. \tag{8}$$

Note that, according to (6), Torgerson's formula defines the distances from the center to all elements in the set Ω. Here, distances in (8) to the already defined center $\omega_{\bar{0}}$ are defined also including the elements from $\Omega/\Omega_{\bar{0}}$. Therefore, the second component in (8) remains the same as the dispersion of the subset $\Omega_{\bar{0}} \subseteq \Omega$.

Specifically, let the subset $\Omega_{\bar{0}} = \omega_k$ consist of a single element $\omega_k \in \Omega$. Therefore, as the center itself, it is directly represented by its known distances to other elements $d_{ki},\ i = 1, \dots m$.

In the other important case, let the set Ω be divided into non-intersecting subsets $\Omega_j,\ j = 1, \dots K,\ \Omega_i \cap \Omega_j = \emptyset,\ i \neq j$, as local concentrations. Such subsets usually arise in problems of cluster analysis. Immediately, we get for each subset $\Omega_k,\ k = 1, \dots K$, its center $\bar{\omega}_k,\ k = 1, \dots K$, represented by the distances to other objects in the whole set Ω:

$$d_{ki}^2 = \frac{1}{m_k} \sum_{p \in \Omega_k} d_{ip}^2 - \sigma_{\Omega_k}^2, \quad \sigma_{\Omega_k}^2 = \frac{1}{2m_k^2} \sum_{p \in \Omega_k} \sum_{q \in \Omega_k} d_{pq}^2, \quad m_k = |\Omega_k|,$$
$$i = 1, \dots m. \tag{9}$$

Since the origin is placed in the center of the cluster Ω_k, it is represented also by the scalar products with other elements,

$$s_{ki} = (1/m_k) \sum\nolimits_{p \in \Omega_k} s_{ip}, \quad i = 1, \dots m, \quad k = 1, \dots K.$$

And as mentioned above, it is necessary to move the origin outside the convex hull of the set Ω, providing thereby that all $s_{ij} \geq 0$.

It should be noted that in constructing a new object, not existing in the original set, based on both a linear combination (5) and, in a particular case, on Torgerson's formula (6), it is necessary to put the origin at this new point. In the clustering problem, when solved according to the k-means-like procedure, we must put consecutively the origin in the center of each cluster, representing it by the distances to other elements of the set.

It is the conceptual foundation for clustering and in an extended sense for the machine learning algorithms based on distances or similarities only, when, for example, the k-means, Forel (Zagoruiko, 1999), and B. N. Kozinets's separating hyperplane (Dvoenko, 2009a) procedures are taken as the models. Here we explicitly justify the key principle that allowed us to have developed clustering and machine learning algorithms before, based on distances and similarities, see Dvoenko (2001, 2009a, b, 2011, 2014, 2018), as well as Dvoenko and Owsiński (2019).

3. Clustering based on distances and similarities

3.1. Unbiased clustering by distances

In order to get $J^D(K) = J^X(K)$, it is necessary to define cluster centers $\bar{\omega}_k$ as new objects by means of formula (9):

$$d^2(\bar{\omega}_k, \omega_i) = \frac{1}{m_k} \sum_{p \in \Omega_k} d^2(\omega_i, \omega_p) - \frac{1}{2m_k^2} \sum_{p \in \Omega_k} \sum_{q \in \Omega_k} d^2(\omega_p, \omega_q),$$
$$i = 1, \dots m \tag{10}$$

in the clustering procedure, referred to further on as the *distance* k-means, the procedure (b):

(b) Step $s = 0$. Define the representatives $\tilde{\omega}_k^s$, $k = 1, \dots K$, perhaps as the most distant objects, $s = s + 1$.

Step s. Allocate objects among clusters:

1. $\omega_i \in \Omega_k^s$, $k = \arg\min_{j=1,\dots K} d(\omega_i, \tilde{\omega}_j^s)$, $i = 1, \dots m$.
2. Determine new centers $\bar{\omega}_k^s$, $k = 1, \dots K$, through distances (10): $d^2(\bar{\omega}_k^s, \omega_i)$, $i = 1, \dots m$.
3. Stop, if $\tilde{\omega}_k^s = \bar{\omega}_k^s$, $k = 1, \dots K$. Else $\tilde{\omega}_k^{s+1} = \bar{\omega}_k^s$, $k = 1, \dots K$, $s = s + 1$.

It is evident that when the algorithm (b) stops, then $J^D(K) = J^X(K)$, since $\bar{\mathbf{x}}_k = \mathbf{x}(\bar{\omega}_k)$. A remark ought to be made that the cluster dispersion η_k, (2), is defined here as the direct consequence of (6) in (7), independently of Friedman and Rubin (1967) and Späth (1983) for the distances only. At last, since $\sigma_k^2 = \eta_k$, $k = 1, \dots K$, then $\tilde{J}(K) = J(K)$, and $\tilde{J}^D(K) = \tilde{J}^X(K)$ also.

3.2. Unbiased clustering by similarities

Let us define scalar products of the mean objects $\bar{\omega}$ with other object as similarities $\bar{\omega} \circ \omega_i = s(\bar{\omega}, \omega_i) \geq 0$, $i = 1, \dots m$. Let us define the average similarity of the whole set Ω as compactness, given by

$$\delta_\Omega = (1/m) \sum\nolimits_{i=1}^m s(\bar{\omega}, \omega_i) = (1/m^2) \sum\nolimits_{i=1}^m \sum\nolimits_{p=1}^m s_{ip}.$$

It should be remembered that $s_{ip} \geq 0$, since all the set Ω is located in a single quadrant of the metric space. Therefore, $\delta_\Omega \geq 0$ all the time. The dispersion of the set Ω relative to the origin ω_0 can be represented as

$$\begin{aligned} \sigma_\Omega^2 &= \frac{1}{2m^2} \sum_{p=1}^m \sum_{q=1}^m d_{pq}^2 = \frac{1}{2m^2} \sum_{p=1}^m \sum_{q=1}^m (s_{pp} + s_{qq} - 2s_{pq}) \\ &= \frac{1}{m} \sum_{p=1}^m s_{pp} - \frac{1}{m^2} \sum_{p=1}^m \sum_{q=1}^m s_{pq} \end{aligned}$$

$$= \frac{1}{m}\sum_{p=1}^{m} s_{pp} - \delta_\Omega = C - \delta_\Omega.$$

In the case of the cluster structure, elements $\omega_i \in \Omega$ from the set Ω belong to non-intersecting subsets: $\Omega_k,\ k = 1, \dots K,\ \Omega_i \cap \Omega_j = \emptyset,\ i \neq j$. Therefore, for each cluster $k = 1, \dots K$ we have its dispersion

$$\sigma_k^2 = (1/m_k)\sum\nolimits_{p=1}^{m_k} s_{pp} - \delta_k.$$

For all clusters we get the criterion

$$\begin{aligned} J(K) &= \sum_{k=1}^{K} \frac{m_k}{m}\sigma_k^2 = \sum_{k=1}^{K} \frac{m_k}{m}\left(\frac{1}{m_k}\sum_{p=1}^{m_k} s_{pp} - \delta_k\right) \\ &= \frac{1}{m}\sum_{p=1}^{m} s_{pp} - \sum_{k=1}^{K}\frac{m_k}{m}\delta_k = C - I(K) \end{aligned}$$

where the weighted average compactness of the cluster structure,

$$I(K) = \sum_{k=1}^{K} \frac{m_k}{m}\delta_k \tag{11}$$

is to be maximized; as $I(K) = C - J(K)$ for constant C, then $J(K)$ is to be minimized, see Dvoenko (2009b, 2011). Let us now develop the clustering procedure, which we shall refer to as the *similarity* k-means, (c):

(c) Step $s = 0$. Define the representatives $\tilde{\omega}_k^s,\ k = 1, \dots K$, perhaps as the least similar objects, $s = s + 1$.

Step s. Allocate objects among clusters:

1. $\omega_i \in \Omega_k^s,\ k = \arg\max_{j=1,\dots K} s(\omega_i, \tilde{\omega}_j^s),\ i = 1, \dots m$.
2. Determine new centers $\bar{\omega}_k^s,\ k = 1, \dots K$, with similarities: $s(\bar{\omega}_k^s, \omega_i),\ i = 1, \dots m$.
3. Stop, if $\tilde{\omega}_k^s = \bar{\omega}_k^s,\ k = 1, \dots K$. Else $\tilde{\omega}_k^{s+1} = \bar{\omega}_k^s,\ k = 1, \dots K,\ s = s + 1$.

Based on the reasoning for the *distance* k-means algorithm above, the *similarity* k-means stops exactly for $I^D(K) = I^X(K)$, where $I^D(K) = C - J^D(K)$, $I^X(K) = C - J^X(K)$.

3.3. The permutable k-means algorithm

The well-known k-means algorithm is popular and intuitive, see, e.g. Friedman and Rubin (1967). Its peculiarity is that the optimization criterion is not explicitly present in it and is not recalculated directly, as usually is done in the standard optimization procedures. It is proven only (for example in Dvoenko, 2009b) that the optimization criterion (1) is actually being minimized.

It should be noted that in procedures (b) and (c), just as in the classical k-means algorithm, the optimization criteria $J(K)$ and $I(K)$ are also not explicitly present. As shown above, the criteria $J(K)$ and $\tilde{J}(K)$ are equivalent. Therefore, in order to represent the k-means algorithm as a procedure with an explicit recalculation of the optimization criterion, it is rational to apply the criterion $\tilde{J}(K)$ in order not to explicitly generate new objects as the centers of the corresponding clusters.

The main difference between the here proposed concept of an optimization procedure and the classical k-means algorithm is that while the current object is transferred between clusters, the centers of the corresponding clusters are changed. This happens explicitly in criterion (1) or implicitly in criterion (3). We denote the criterion $\tilde{J}(K)$ at the step s for the cluster structure $\Omega_k^s,\ k = 1,\dots K$ as $\tilde{J}^s$. When the current object ω_i is moved from the cluster Ω_p^s to the cluster Ω_j^s, we get the sets $\Omega_p^s \diagdown \omega_i$ and $\Omega_j^s \bigcap \omega_i$ in the clustering structure $\Omega_1^s, \dots; \Omega_j^s \bigcap \omega_i, \dots \Omega_p^s \diagdown \omega_i, \dots \Omega_K^s$. Let us denote the corresponding value of the criterion $\tilde{J}(K)$ as $\tilde{J}_{ij}^s$.

It is evident that this new so-called permutable algorithm is more complicated than the original k-means algorithm, remaining, however, still a locally optimal procedure.

On the other hand, it would be rational to apply optimal recalculation schemes of the optimization criterion to improve the performance. Let us assume that this can always be done.

We consider here the permutable *distance* k-means algorithm based on recalculation of the criterion $\tilde{J}(K)$ and its dual form as the permutable *similarity* k-means algorithm based on recalculation of the criterion $I(K) = C - \tilde{J}(K)$.

Let us make one more remark here. The permutable k-means algorithm should be presented in the form without the direct use of the cluster centers themselves. Since the goal of the initial decision, determining the starting point of the procedure, is the same as before, we assume that the choice is made of the least scattered clusters. Here it does not matter how we do actually determine these initial clusters. The permutable *distance* k-means algorithm has the form (d):

(d) Step $s = 0$. Define clusters $\Omega_k^s,\ k = 1,\dots K$, possibly as the least scattered ones, $\mathbf{J}^s = \tilde{J}^s$, $s = s + 1$.

Step s. Allocate objects among clusters:

1. $\omega_i \in \Omega_k^s$, $\tilde{J}_{ik}^s = \min_{j=1,\dots K} \tilde{J}_{ij}^s$, $\tilde{J}^s = \tilde{J}_{ik}^s$.

2. $i = i + 1$, reallocate the next object ω_i, until the set Ω is exhausted.

3. Stop, if $\tilde{J}^s = \mathbf{J}^s$. Else $\mathbf{J}^s = \tilde{J}^s$, $s = s + 1$.

Using similar notation and the notion of "most compact" for the initial decision, with in the criterion $I(K)$ (see (11)), we develop here the permutable *similarity* k-means algorithm, (e):

(e) Step $s = 0$. Define clusters $\Omega_k^s,\ k = 1, \ldots K$, possibly as the most compact, $\mathbf{I}^s = I^s$, $s = s + 1$.

Step s. Allocate objects among clusters:

1. $\omega_i \in \Omega_k^s$, $I_{ik}^s = \max_{j=1,\ldots K} I_{ij}^s$, $I^s = I_{ik}^s$.

2. $i = i + 1$, reallocate the next object ω_i, until the set Ω is exhausted.

3. Stop, if $I^s = \mathbf{I}^s$. Else $\mathbf{I}^s = I^s$, $s = s + 1$.

Obviously, the peculiarity of procedures (d) and (e) consists in that the optimization criterion is recalculated every time both during trial permutations of the current object between clusters and during its final transfer to the optimal cluster. Each procedure stops when there is no moving of objects at all.

It is also evident that with such a sequential recalculation of the criterion in procedures (d) and (e), their behavior differs from the behavior of the so-called real-time k-means algorithm, that is – the one used in classification mode. In the case of this real-time use, clusters are redefined after the appearance of a new object. The difference consists in the fact that in the classical real-time k-means algorithm, trial transfers of a new object are performed relative to unchanged cluster centers.

Procedures (d), and (e) imply that the result of the permutable algorithm may differ from the classical result in the general case. This gives a reason for considering them as a separate entity within the class of k-means algorithms. We discuss their novelty in more details below.

Additionally, let us yet consider another version of the permutable algorithm for distances when all clusters are redefined simultaneously after all trial transfers have been tested. Clusters are redefined in the same way as in the classical algorithm relative to unchanged centers. The permutable *distance* k-means algorithm has the following form, (f):

(f) Step $s = 0$. Define clusters $\Omega_k^s,\ k = 1, \ldots K$, perhaps as the least scattered ones, $\mathbf{J}^s = \tilde{J}^s$, $s = s + 1$.

Step s. Allocate objects among clusters:

1. Remember, but do not move $\omega_i \in \Omega_k^s$, $\tilde{J}_{ik}^s = \min_{j=1,\ldots K} \tilde{J}_{ij}^s$, $i = 1, \ldots m$.

2. Reallocate all objects ω_i, $i = 1, \ldots m$ among clusters, calculate $\tilde{J}^s$.

3. Stop, if $\tilde{J}^s = \mathbf{J}^s$.

Stop, if $\tilde{J}^s > \mathbf{J}^s$, cancel all last reallocations, $\tilde{J}^s = \mathbf{J}^s$.

Else $\mathbf{J}^s = \tilde{J}^s$, $s = s + 1$.

It is obvious, in the procedure (f), that the result of the sequential trial permutations relative to one initial cluster structure may differ from the result when all permutations are done simultaneously. Therefore, in order to improve the final result, in general, it would to be rational, after cancellation of all the last permutations for $\tilde{J}^s > \mathbf{J}^s$, to use the step s of the procedure (e) until the end.

4. Relation of cluster analysis to some other problems

4.1. Aggregation problem

Let us consider the heuristic problem of diagonalization of the matrix of connections $A(m, m)$ with non-negative elements $a_{ij} \geq 0$. The solution to this problem is equivalent to identifying the so-called block-diagonal structure of this matrix (Braverman et al., 1971; Braverman and Muchnik, 1983).

Let the set of elements be represented by pairwise relationships. It is assumed that all elements are naturally concentrated in K compact subsets, which can be potentially considered as aggregates. Consequently, by simultaneously rearranging the rows and columns of the relationship matrix, it is possible to distinguish such a block structure that each block along the main diagonal of the matrix consists of elements in the same aggregate.

According to the compactness hypothesis, elements of the same aggregate are more strongly interconnected than elements from different aggregates. Therefore, values of linking quantities for the pairs of elements from the same aggregate are higher than values for pairs from different aggregates.

The procedure of finding the aggregate structure maximizes the weighted relationships within aggregates. The heuristic quality function is calculated as

$$F(K) = \frac{1}{m} \sum_{k=1}^{K} \frac{1}{m_k - 1} \sum_{\substack{\omega_i, \omega_j \in \Omega_k, \\ i \neq j}} a_{ij}.$$

Let us consider the clustering criterion based on similarity:

$$\begin{aligned} I(K) &= \sum_{k=1}^{K} \frac{m_k}{m} \delta_k = \sum_{k=1}^{K} \frac{m_k}{m} \left(\frac{1}{m_k^2} \sum_{p=1}^{m_k} \sum_{q=1}^{m_k} s_{pq} \right) \\ &= \sum_{k=1}^{K} \frac{m_k}{m} \left(\frac{1}{m_k^2} \sum_{p=1}^{m_k} s_{pp} + \frac{1}{m_k^2} \sum_{p=1}^{m_k} \sum_{q=1,\, p\neq q}^{m_k} s_{pq} \right) \end{aligned}$$

$$= \frac{1}{m} \sum_{k=1}^{K} \frac{1}{m_k} \sum_{p=1}^{m_k} s_{pp} + \frac{1}{m} \sum_{k=1}^{K} \frac{1}{m_k} \sum_{p=1}^{m_k} \sum_{q=1,\, p\neq q}^{m_k} s_{pq}.$$

For the normalized relationship matrix $S(m, m)$ with diagonal elements $s_{ii} = 1,\ i = 1, \dots m$, the clustering criterion is calculated as

$$I(K) = \frac{K}{m} + \frac{1}{m} \sum_{k=1}^{K} \frac{1}{m_k} \sum_{p=1}^{m_k} \sum_{q=1,\, p\neq q}^{m_k} s_{pq}.$$

As a result, the functional $F(K)$ is a heuristic version of the clustering criterion $I(K)$, where the contribution of m diagonal elements simply provides a constant added to the criterion value, which does not change for different partitions. Therefore, the diagonalization procedure for a positively semi-definite relationship matrix $A(m, m)$ is a heuristic version of the permutable *similarity* k-means algorithm (e).

Let the elements of the set Ω be the features themselves, represented by a positive definite correlation matrix $R(n, n)$, calculated for the data matrix $X(m, n)$. If all correlations are non-negative, $r_{ij} \geq 0$, then the problem of identifying groups of strongly correlated features can be solved by the permutable *similarity* k-means algorithm (e). This problem can be represented, on the other hand, as a diagonalization problem. In this case, the average attribute of a group becomes the expression of a hidden factor representing this group.

In order for the correlations to be non-negative, $r_{ij} \geq 0$, it is necessary, as shown above, to put the origin of coordinates beyond the convex hull of the set Ω in the metric space. If not, modules or squared correlations can be considered. The problem of grouping features under these conditions is considered in Dvoenko (2009b) as a factor analysis problem.

4.2. Factor analysis

One of the problems of data analysis is to split the set of n features as columns of the data matrix $X(m, n) = (X_1, \dots X_n)$ into groups of similar ones, where $X_j = (x_{1j}, \dots x_{mj})^T$. Features characterize the behavior of the phenomenon under study, where observations of features represent variational series. We show that the problem of calculating centroid factors can be formulated as the cluster analysis problem.

Let the relationships of features be represented by the correlation matrix $R(n, n)$. Let elements of the set Ω be the features themselves, represented by correlations, $-1 \leq r_{ij} \leq 1$, for all pairs of features, where their modules or squares are considered.

Solving the problem of clustering of features by the *similarity* k-means algorithm, when unbiased clustering has been obtained, assumes maximization of two functionals:

$$I'(K) = \frac{1}{n} \sum_{k=1}^{K} n_k \delta'_k = \frac{1}{n} \sum_{k=1}^{K} \sum_{p=1}^{n_k} r^2(\bar{\omega}_k, \omega_p),$$

and

$$I''(K) = \frac{1}{n} \sum_{k=1}^{K} n_k \delta''_k = \frac{1}{n} \sum_{k=1}^{K} \sum_{p=1}^{n_k} |r(\bar{\omega}_k, \omega_p)|.$$

This, in turn, assumes the maximization of functionals

$$I_S(K) = nI'(K) = \sum_{k=1}^{K}\sum_{p=1}^{n_k} r^2(\bar{\omega}_k, \omega_p), \quad \omega_p \in \Omega_k.$$

and

$$I_M(K) = nI''(K) = \sum_{k=1}^{K}\sum_{p=1}^{n_k} |r(\bar{\omega}_k, \omega_p)|, \quad \omega_p \in \Omega_k.$$

Let us consider two algorithms of factor analysis, referred to as algorithms of extreme grouping of parameters (Braverman, 1970; Braverman and Muchnik, 1983; Lumel'sky, 1970), which were developed for the correlation matrix $R(n,n)$ of features. Such algorithms are sometimes also referred to as the so-called Square (S) for solving the Local Principal Component Analysis (LPCA) problem and the Module (M) for solving the Local Centroid Component Analysis (LCCA) problem. The centers of groups are declared as their factors and are built as new features that are most correlated with characteristics of their groups. Such factors of groups are represented only by their correlations with all of the features. The S- and M-algorithms maximize the following functionals

$$J_S = \sum_{k=1}^{K}\sum_{p=1}^{n_k} r^2(\pi_k, \omega_p) \;\text{ and }\; J_M = \textstyle\sum_{k=1}^{K}\sum_{p=1}^{n_k} |r(\mu_k, \omega_p)|, \quad \omega_p \in \Omega_k,$$

where π_k is the principal factor, and μ_k is the centroid factor of the group Ω_k. Such functionals characterize the quality of the separation of features into a given number of K groups, where features are most strongly correlated with their factor in the group. These algorithms find factors, actually solving simultaneously the general tasks of factor analysis: building of general factors and their oblique rotation (Harman, 1976).

Let us consider a normalized similarity matrix $S(n,n)$ as a matrix $R(n,n)$ with elements $s_{ij} = |r_{ij}|$. Let us represent the centers $\bar{\omega}_k$ and the centroid factors μ_k of groups by their similarities with features $\omega_i \in \Omega$:

$$s(\bar{\omega}_k, \omega_i) = \frac{1}{n_k}\sum_{p=1}^{n_k} s_{ip}, \quad s(\mu_k, \omega_i) = \sum_{p=1}^{n_k} s_{ip}, \quad \omega_p \in \Omega_k. \tag{12}$$

Similarly, let us consider a normalized similarity matrix $S(n,n)$ as a matrix $R(n,n)$ with elements $s_{ij} = r_{ij}^2$. Let us represent the principal factors π_k of groups Ω_k by their similarities with features $\omega_i \in \Omega$:

$$s(\pi_k, \omega_i) = \sum_{p=1}^{n_k} \alpha_p^k s_{ip}, \quad \omega_p \in \Omega_k, \tag{13}$$

where $\mathbf{a}_k = (\alpha_1^k, \ldots\ \alpha_{n_k}^k)^T$ is the eigenvector corresponding to the maximal eigenvalue λ_1^k of the similarity submatrix $S(n_k, n_k)$ with eigenvalues in the decreasing order $\lambda_1^k > \ldots\ > \lambda_{n_k}^k > 0$.

Let us also consider a normalized data matrix $X(m, n) = (X_1, \ldots\ X_n)$, consisting of features being columns $X_j = (x_{1j}, \ldots\ x_{mj})^T$, where $\bar{x}_j = 0$ and $\sigma_j^2 = 1$. Let us calculate the feature $Y = (y_1, \ldots\ y_m)^T$ as the average $Y = (1/n) \sum_{j=1}^n X_j$ with components $y_i = (1/n) \sum_{j=1}^n x_{ij}$. Let us calculate the average of the feature Y itself

$$\bar{y} = \frac{1}{m} \sum_{i=1}^m y_i = \frac{1}{n} \sum_{j=1}^n \frac{1}{m} \sum_{i=1}^m x_{ij} = \frac{1}{n} \sum_{j=1}^n \bar{x}_j = 0.$$

The dispersion of the feature Y is

$$\sigma_Y^2 = \frac{1}{m} \sum_{i=1}^m y_i^2 = \frac{1}{n^2} \sum_{p=1}^n \sum_{q=1}^n \frac{1}{m} \sum_{i=1}^m x_{ip} x_{iq} = \frac{1}{n^2} \sum_{p=1}^n \sum_{q=1}^n \frac{1}{m} (X_p \circ X_q)$$
$$= \frac{1}{n^2} \sum_{p=1}^n \sum_{q=1}^n r_{pq}.$$

Let us calculate the similarities of the normalized feature Y with respect to other normalized features X_i as scalar products

$$X_i \circ \frac{Y}{\sigma_Y} = X_i \circ \frac{1}{n\sigma_Y} \sum_{j=1}^n X_j = \sum_{j=1}^n \frac{1}{n\sigma_Y} (X_i \circ X_j). \tag{14}$$

Let the maximal eigenvalue $\lambda = \lambda_{\ 1}$, $\lambda_{\ 1} > \ldots > \lambda_n > 0$ be found for the correlation matrix $R(n, n)$ subject to the condition $R\mathbf{a} = \lambda \mathbf{a}$, where $\mathbf{a} = (\alpha_1, \ldots\ \alpha_n)^T$, $\sum_{j=1}^n \alpha_j^2 = 1$, is the corresponding eigenvector. Let us calculate a feature $Z = (z_1, \ldots\ z_m)^T$ with components $z_i = \mathbf{x}_i \circ \mathbf{a}$, $\mathbf{x}_i = (x_{i1}, \ldots\ x_{in})$. Then, let us calculate the average value of the feature Z:

$$\bar{z} = \frac{1}{m} \sum_{i=1}^m z_i = \frac{1}{m} \sum_{i=1}^m \sum_{j=1}^n x_{ij} \alpha_j = \sum_{j=1}^n \alpha_j \frac{1}{m} \sum_{i=1}^m x_{ij} = \sum_{j=1}^n \alpha_j \bar{x}_j = 0.$$

The dispersion of the feature Z is

$$\sigma_Z^2 = \frac{1}{m} \sum_{i=1}^m z_i^2 = \frac{1}{m} \sum_{i=1}^m \left(\sum_{j=1}^n x_{ij} \alpha_j \right)^2 = \frac{1}{m} \sum_{i=1}^m \sum_{p=1}^n \sum_{q=1}^n x_{ip} x_{iq} \alpha_p \alpha_q =$$

$$\sum_{p=1}^n \sum_{q=1}^n \alpha_p \alpha_q \left(\frac{1}{m} \sum_{i=1}^m x_{ip} x_{iq} \right) = \sum_{p=1}^n \alpha_p \sum_{q=1}^n \alpha_q r_{pq} = \sum_{p=1}^n \alpha_p^2 \lambda = \lambda.$$

Now, let us calculate the similarities of the normalized feature Z with respect to other normalized features X_i as scalar products

$$X_i \circ \frac{Z}{\sigma_Z} = \frac{1}{\sigma_Z}\sum_{p=1}^{m} x_{pi} z_p = \frac{1}{\sigma_Z}\sum_{p=1}^{m} x_{pi} \sum_{j=1}^{n} x_{pj}\alpha_j = \sum_{j=1}^{n} \frac{\alpha_j}{\sigma_Z}(X_i \circ X_j). \quad (15)$$

Following Braverman (1970), Braverman and Muchnik (1983) and Lumel'sky (1970), let us calculate a feature $V = (v_1, \dots v_m)^T$ with components $v_i = \mathbf{x}_i \circ \varepsilon$, $x_i = (x_{i1}, \dots x_{in})$, $\varepsilon_i = (\varepsilon_1, \dots \varepsilon_n)^T$, $\varepsilon = \pm 1$, and then the average value of the feature V:

$$\bar{v} = \frac{1}{m}\sum_{i=1}^{m} v_i = \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij}\varepsilon_j = \sum_{j=1}^{n} \varepsilon_j \frac{1}{m}\sum_{i=1}^{m} x_{ij} = \sum_{j=1}^{n} \varepsilon_j \bar{x}_j = 0.$$

The dispersion of the feature V is

$$\sigma_V^2 = \frac{1}{m}\sum_{i=1}^{m} v_i^2 = \sum_{p=1}^{n}\sum_{q=1}^{n} \varepsilon_p \varepsilon_q \left(\frac{1}{m}\sum_{i=1}^{m} x_{ip} x_{iq} \right) = \sum_{p=1}^{n}\sum_{q=1}^{n} \varepsilon_p \varepsilon_q r_{pq},$$

where r_{pq} is the correlation coefficient of features X_p and X_q. Let us calculate the similarities of the normalized feature V with respect to other normalized features X_i as scalar products

$$X_i \circ \frac{V}{\sigma_V} = \frac{1}{\sigma_V}\sum_{p=1}^{m} x_{pi} v_p = \frac{1}{\sigma_V}\sum_{p=1}^{m} x_{pi} \sum_{j=1}^{n} x_{pj}\varepsilon_j = \sum_{j=1}^{n} \frac{\varepsilon_j}{\sigma_V}(X_i \circ X_j). \quad (16)$$

Let features $X_i = X(\omega_i)$ be the objects ω_i, $i = 1, \dots n$, represented by the positive definite non-normalized similarity matrix $S(n, n)$ with non-negative elements $s_{ij} = X_i \circ X_j = m r_{ij} \geq 0$. Then, dispersions of normalized features Y, Z and V are calculated as

$$\sigma_Y^2 = (1/mn^2) \sum\nolimits_{i=1}^{n} \sum\nolimits_{j=1}^{n} s_{ij},$$

$$\sigma_Z^2 = \lambda'/m = \lambda, \text{ where } S\mathbf{a} = \lambda'\mathbf{a} = m\lambda\mathbf{a},$$

$$\sigma_V^2 = (1/m) \sum\nolimits_{i=1}^{n} \sum\nolimits_{j=1}^{n} s_{ij}, \text{ for all } \varepsilon_i = +1, \text{ where } \sigma_V^2 = n^2\sigma_Y^2.$$

Finally, let the features $Y = Y(\bar{\omega})$, $Z = Z(\pi)$, $V = V(\mu)$ be objects, referred to as $\bar{\omega}$, π and μ. Since features $X_1, \dots X_n$ themselves are not directly available, the scalar products $X_i \circ X_j$ need to be defined by similarities $s(\omega_i, \omega_j)$ in the metric space. Therefore, according to (14) – (16), objects $\bar{\omega}$, π and μ are represented by similarities to all other objects ω_i, $i = 1, \dots n$, as

$$s(\omega_i, \mu) = \frac{1}{\sigma_V}\sum_{j=1}^{n} s(\omega_i, \omega_j), \text{ since all } \varepsilon_j = +1,$$

$$s(\omega_i, \bar{\omega}) = \frac{1}{n\sigma_Y} \sum_{j=1}^{n} s(\omega_i, \omega_j) = \frac{1}{\sigma_V} \sum_{j=1}^{n} s(\omega_i, \omega_j),$$

$$s(\omega_i, \pi) = \frac{1}{\sigma_Z} \sum_{j=1}^{n} \alpha_j s(\omega_i, \omega_j).$$

Let objects $\omega_i \in \Omega$ be distributed among clusters $\Omega_k,\ k = 1, \dots K$. Then, the non-normalized objects $\bar{\omega}_k$, π_k and μ_k of a cluster Ω_k are represented by their similarities to other objects, according to (12) and (13).

Therefore, similarities $s(\omega_i, \bar{\omega}_k)$ and $s(\omega_i, \mu_k)$ coincide with each other up to a constant multiplier. As a result, the grouping, obtained from the M-algorithm is an unbiased clustering. The grouping, obtained from the S-algorithm represents a biased clustering. And finally, the M-algorithm is similar to the *similarity* k-means used for feature grouping.

It should be noted that this result is obtained here as a consequence of the properties of the unbiased partition of a set of elements into non-intersecting subsets. This demonstration is easier than the special proof in Braverman and Muchnik (1983) of the properties of the extreme grouping M-algorithm for the optimization criterion J_M.

The essential issues of factor analysis are discussed in Harman (1976). In particular, one of them is the question of determining the minimum rank of the correlation matrix, which determines the number of common factors.

Let us consider an example of solving a factor analysis problem as a clustering one. The correlation matrix of eight physical variables of a body status is considered, where a preliminary conclusion is made on the basis of the structure of correlations that the rank of the reduced matrix should not be higher than two, see Harman (1976). This means that there are two common factors. The first four variables measure the so-called “lankiness”, while the other four variables measure the so-called “stockiness” (see Table 1).

In this case, all correlations between physical variables as features are positive, which allows for considering them as similarity functions without additional preprocessing. This matrix is positively definite, all its eigenvalues are positive, taking the values of: 4.672880, 1.770983, 0.481035, 0.421441, 0.233221, 0.186674, 0.137304 and 0.096463. The sum of eigenvalues is equal to the size of the matrix and determines its rank as eight. Therefore, this set of eight elements is immersed in eight-dimensional metric space.

As expected, the *similarity* k-means algorithm (c) correctly identifies two clusters, where the averages for them represent the centroid factors of the groups as a solution to the LCCA problem.

In addition, the following remark should be made. In accordance with the linear factor model, a reduced correlation matrix (Table 2) is considered, where the main diagonal shows the commonalities found in Harman (1976), see Table 5.4, page 81.

Table 1. Correlations among eight physical variables for 305 girls

Variables	1	2	3	4	5	6	7	8
1. Height	1							
2. Arm span	0.846	1						
3. Length of forearm	0.805	0.881	1					
4. Length of lower leg	0.859	0.826	0.801	1				
5. Weight	0.473	0.376	0.380	0.436	1			
6. Bitrochanteric diameter	0.398	0.326	0.319	0.329	0.762	1		
7. Chest girth	0.301	0.277	0.237	0.327	0.730	0.583	1	
8. Chest width	0.382	0.415	0.345	0.365	0.629	0.577	0.539	1

Source: Harman (1976), see Table 5.3, p. 80

It is natural that the rank of this reduced matrix is lowered to 5.96 as the sum of commonalities on the main diagonal. This matrix becomes a non-positively definite one with three negative eigenvalues: 4.448400, 1.508262, 0.102580, 0.058149, 0.013292, -0.039344, -0.058103 and -0.073234. The sum of them also determines the rank of this matrix as 5.96. The sum of the positive eigenvalues only is 6.1307. It should be noted that a simple normalization, meant to obtain the unit main diagonal does not eliminate negative eigenvalues, as we obtain: 5.929929, 2.061029, 0.139097, 0.087818, 0.017322, -0.048442, -0.089078 and -0.097676, although it increases the rank of the matrix to its dimensionality, that is – to eight.

It is obvious that this set of elements, represented by their pairwise comparisons in the form of the reduced correlation matrix, is not immersed correctly into the eight-dimensional metric space. It can only be expected that the correct dimensionality may not be higher than five or six, if contributions of all eigenvectors corresponding to negative eigenvalues are eliminated. As noted above, the immersion problems are considered in Dvoenko and Pshenichny (2018).

It should be considered that the above mentioned reduction of this matrix appears to be too strong, since the reduced matrix becomes a non-positively definite one. Therefore, it becomes necessary to increase the commonalities of the features.

As shown above, the diagonal elements of the non-normalized similarity matrix determine the squares of distances of the elements to the origin. Therefore, they determine the natural configuration of the set in a metric space. It should be noted that the normalization destroys this configuration because all elements of the set are located on a hypersphere of a unit radius.

Table 2. Reduced correlation matrix among eight physical variables

Variables	1	2	3	4	5	6	7	8
1. Height	0.842							
2. Arm span	0.846	0.881						
3. Length of forearm	0.805	0.881	0.817					
4. Length of lower leg	0.859	0.826	0.801	0.815				
5. Weight	0.473	0.376	0.380	0.436	0.872			
6. Bitrochanteric diameter	0.398	0.326	0.319	0.329	0.762	0.647		
7. Chest girth	0.301	0.277	0.237	0.327	0.730	0.583	0.584	
8. Chest width	0.382	0.415	0.345	0.365	0.629	0.577	0.539	0.502

Source: Harman (1976), Table 5.4, p.81

After reduction, this correlation matrix becomes the non-normalized similarity matrix. It is known that the correct non-normalized similarity matrix should contain diagonal elements which exceed the non-diagonal ones.

However, Table 2 shows that the reduced correlation matrix for the eight physical variables becomes incorrect. This also does not allow the given set to be correctly immersed into the eight-dimensional metric space.

As it is known, the problem of commonalities in factor analysis does not have an unambiguous solution (see Harman, 1976). It should be noted here that this problem leads, in the general case, again to the problem of the correct immersion of a set in a metric space. Therefore, it is obvious that this requirement may impose additional restrictions on the degree of reduction of the correlation matrix in factor analysis. Such difficulties are not discussed further, because this would go beyond the problems of cluster analysis itself.

5. Improving the quality of clustering

5.1. A bi-partial objective function

The problem of improving the clustering quality with regard to the results, produced by the known algorithms, is still valid nowadays. Therefore, numerous variants of the basic k-means procedure, fuzzy clustering, various concepts of an average, various initial solutions, etc., were proposed. As it was noted above, the possibility of rational data analysis is based on the informal compactness hypothesis. According to this hypothesis, the compact concentrations are formed by objects that are located close in some sense to each other. Hence, the stronger concentration of elements in a subset to be, the stronger tendency for centers from different subsets as representatives of concentrations to be distant from each other.

Here we consider one of the concepts that directly implement such a consequence of the compactness hypothesis. This is the concept of a bi-partial objective function for clustering, see Owsiński (2020). It should be noted that the concept of a generalized two-component optimization criterion was developed first to solve various problems of splitting experimental data into subsets.

One of such important cases is the clustering problem. According to Owsiński (2020), the generalized objective function consists of two parts $Q_S^D(P) = C_S(P)+C^D(P)$, where $C_S(P)$ evaluates the quality of the partition P relative to similarities of elements within subsets in P. The second part $C^D(P)$ evaluates the quality of the partition P relative to distances between elements from different subsets in P. If the partition P represents clusters, then the criterion $Q_S^D(P)$ should be maximized, where both similarities of elements within each cluster $C_S(P)$ and distances between elements from different clusters $C^D(P)$ are maximized. The dual objective function is represented as $Q_D^S(P) = C^S(P)+C_D(P)$, where its minimization for the partition P as composed of clusters means minimization of similarities between clusters $C^S(P)$ and minimization of distances $C_D(P)$ within clusters.

5.2. A permutable k-means for the bi-partial objective function

Within the framework of the clustering problem, let us consider the particular formulation that allows us to apply new properties of the permutable clustering algorithms developed above, see Dvoenko and Owsiński (2019).

Usually, the bi-partial criterion in the form of $Q_S^D(P)$ or $Q_D^S(P)$ require scaling of its parts. Let us propose a criterion of a particular type, in which the proportions of its two parts constitute a linear combination

$$J_\delta(K) = (1-\alpha)\tilde{J}(K) + \alpha\delta(K), \quad 0 \le \alpha \le 1, \tag{17}$$

where according to (3), the criterion $\tilde{J}(K)$ minimizes the cluster dispersion and the criterion $\delta(K)$ minimizes the inter-cluster similarity. Let us develop the function $\delta(K)$.

Let the center ω_0 of the set Ω, as the origin of coordinates, be moved, as shown before in this paper, outside the convex hull of the set so that all pairwise similarities between elements are non-negative, $s(\omega_i, \omega_j) \ge 0$. Then the center $\bar{\omega}_k$ of each cluster also is represented by its non-negative similarities, $s(\omega_i, \bar{\omega}_k) \ge 0$, with the rest of the elements,

$$s(\omega_i, \bar{\omega}_k) = (1/m_k)\sum\nolimits_{p=1}^{m_k} s_{ip}, \quad \omega_p \in \Omega_k, \quad \omega_i \in \Omega, \quad i = 1, \dots m.$$

The cluster compactness is calculated as the average similarity of its center with respect to the objects in the cluster

$$\delta_k = \frac{1}{m_k}\sum_{i=1}^{m_k} s(\omega_i, \bar{\omega}_k) = \frac{1}{m_k^2}\sum_{i=1}^{m_k}\sum_{p=1}^{m_k} s_{ip}, \quad \omega_i \in \Omega_k, \quad \omega_p \in \Omega_k.$$

Let us consider the set of cluster centers $\bar{\omega}_k,\ k = 1,\dots K$. The center $\bar{\omega}_0$ of this set is represented by its non-negative similarities $s(\bar{\omega}_0, \bar{\omega}_k) \geq 0$ with respect to cluster centers relative to the origin ω_0, moved out of the convex hull of the set

$$s(\bar{\omega}_0, \bar{\omega}_k) = \frac{1}{K}\sum_{p=1}^{K} s(\bar{\omega}_k, \bar{\omega}_p), \quad k = 1,\dots K.$$

The average similarity of the center $\bar{\omega}_0$ with respect to other centers, $\bar{\omega}_k,\ k = 1,\dots K$ is calculated as

$$\delta(K) = \frac{1}{K}\sum_{k=1}^{K} s(\bar{\omega}_0, \bar{\omega}_k) = \frac{1}{K^2}\sum_{k=1}^{K}\sum_{l=1}^{K} s(\bar{\omega}_k, \bar{\omega}_l). \tag{18}$$

The disadvantage of (18) is that cluster centers are explicitly represented in it. Hence, it becomes obvious that the criterion (17) cannot be optimized when using the classical version of k-means. Since the cluster centers are represented explicitly, the second part, (18), of the criterion (17) cannot be changed when objects are transferred between clusters. Therefore, when $0 \leq \alpha < 1$, the optimization result corresponds to the classical one for $\alpha = 0$, and when $\alpha = 1$, the criterion simply does not work. Obviously, to optimize such a criterion, the permutable version of the k-means algorithm should be applied. To do this, it is necessary to calculate $\delta(K)$ in some other way.

Let us calculate the similarity of the cluster center $\bar{\omega}_k$ to an object from another cluster $\omega_p \in \Omega_l$ as the average

$$s(\bar{\omega}_k, \omega_p) = (1/m_k)\sum\nolimits_{q=1}^{m_k} s(\omega_p, \omega_q), \quad \omega_q \in \Omega_k.$$

The average similarity of the cluster center $\bar{\omega}_k$ to all objects from another cluster Ω_l is calculated as

$$s(\bar{\omega}_k, \Omega_l) = \frac{1}{m_l}\sum_{p=1}^{m_l} s(\bar{\omega}_k, \omega_p) = \frac{1}{m_l m_k}\sum_{p=1}^{m_l}\sum_{q=1}^{m_k} s(\omega_p, \omega_q),$$

$$\omega_p \in \Omega_l, \quad \omega_q \in \Omega_k.$$

It is easy to see that $s(\bar{\omega}_k, \Omega_l) = s(\bar{\omega}_l, \Omega_k)$, since $s_{pq} = s_{qp}$. Therefore, the equivalent notations can be applied: $s(\bar{\omega}_k, \Omega_l) = s(\bar{\omega}_k, \bar{\omega}_l) = s(\Omega_k, \Omega_l)$. Then, (18) is represented as

$$\delta(K) = \frac{1}{K^2}\sum_{k=1}^{K}\sum_{l=1}^{K} s(\bar{\omega}_k, \bar{\omega}_l) = \frac{1}{K^2}\sum_{k=1}^{K}\sum_{l=1}^{K}\frac{1}{m_l m_k}\sum_{p=1}^{m_l}\sum_{q=1}^{m_k} s_{pq},$$

$$\omega_p \in \Omega_l, \quad \omega_q \in \Omega_k. \tag{19}$$

However, this expression is not altogether correct, since for $k = l$ it includes the compactness of the cluster δ_k. The purpose of the criterion $J_\delta(K)$ is to obtain clusters with minimal cluster variance and minimal similarity between clusters. Obviously, the compactness δ_k of each cluster only increases. Therefore, the final correct expression for the function $\delta(K)$ can be obtained after removing the contribution from the compactness of clusters and taking into account symmetry

$$\delta(K) = \frac{1}{2K(K-1)} \sum_{k=1}^{K} \sum_{l=1,\, l\neq k}^{K} \frac{1}{m_l m_k} \sum_{p=1}^{m_l} \sum_{q=1}^{m_k} s_{pq},$$

$$\omega_p \in \Omega_l, \quad \omega_q \in \Omega_k. \tag{20}$$

Now, it is obvious that to minimize the objective function (17), the permutable algorithms (d) and (f) should be applied. For finding the optimal linear combination, i.e.

$$\alpha* = \arg\min\nolimits_{0\leq\alpha\leq1} J_\delta(K) = \arg\min\nolimits_{0\leq\alpha\leq1} \Big((1-\alpha)\tilde{J}(K) + \alpha\delta(K)\Big),$$

the respective algorithm can be simply iterated along the values of α with a certain step.

It should be noted that the type of the criterion (17) is determined by the classical idea of the quality of clustering based on minimizing the variance of clusters. Now, let us consider its dual form, based on maximizing compactness of clusters and variance between them

$$I_\sigma(K) = (1-\alpha)I(K) + \alpha\sigma^2(K), \quad 0 \leq \alpha \leq 1. \tag{21}$$

The first part of (21) takes the form

$$I(K) = \sum_{k=1}^{K} \frac{m_k}{m}\delta_k = \sum_{k=1}^{K} \frac{m_k}{m}\frac{1}{m_k^2} \sum_{p=1}^{m_k}\sum_{q=1}^{m_k} s_{pq} = \frac{1}{m}\sum_{k=1}^{K}\frac{1}{m_k}\sum_{p=1}^{m_k}\sum_{q=1}^{m_k} s_{pq}.$$

Let us calculate the function $\sigma^2(K)$. The center $\bar{\omega}_0$ of the set of cluster centers is represented by its distances to other centers $\bar{\omega}_k$, according to Torgerson's formula, as

$$d^2(\bar{\omega}_k, \bar{\omega}_0) = \frac{1}{K}\sum_{p=1}^{K} d^2(\bar{\omega}_k, \bar{\omega}_p) - \frac{1}{2K^2}\sum_{p=1}^{K}\sum_{q=1}^{K} d^2(\bar{\omega}_p, \bar{\omega}_q),$$

$$k = 1, \dots K.$$

After substituting and bringing similar expressions together, we calculate the dispersion of the set of cluster centers relative to their center $\bar{\omega}_0$ as

$$\sigma^2(K) = \frac{1}{K}\sum_{k=1}^{K} d^2(\bar{\omega}_k, \bar{\omega}_0) = \frac{1}{2K^2}\sum_{k=1}^{K}\sum_{l=1}^{K} d^2(\bar{\omega}_k, \bar{\omega}_l). \tag{22}$$

Let us determine the distances between the cluster centers. The cluster center $\bar{\omega}_k$ is represented by its distances to all other objects $\omega_i \in \Omega$ and, in particular, to objects from another cluster, $\omega_i \in \Omega_l$. The average square of distances of objects from another cluster $\omega_i \in \Omega_l$ to the center $\bar{\omega}_k$ of this cluster is calculated as

$$d^2(\bar{\omega}_k, \Omega_l) = \frac{1}{m_l}\sum_{i=1}^{m_l} d^2(\omega_i, \bar{\omega}_k) = \frac{1}{m_l}\sum_{i=1}^{m_l}\left(\frac{1}{m_k}\sum_{p=1}^{m_k} d_{ip}^2 - \frac{1}{2m_k^2}\sum_{p=1}^{m_k}\sum_{q=1}^{m_k} d_{pq}^2\right) =$$

$$\frac{1}{m_k m_l}\sum_{i=1}^{m_l}\sum_{p=1}^{m_k} d_{ip}^2 - \frac{1}{2m_k^2}\sum_{p=1}^{m_k}\sum_{q=1}^{m_k} d_{pq}^2 = \frac{1}{m_k m_l}\sum_{i=1}^{m_l}\sum_{p=1}^{m_k} d_{ip}^2 - \sigma_k^2,$$

$$\omega_p \in \Omega_k,\ \omega_q \in \Omega_k.$$

The average square of distances of objects from another cluster $\omega_i \in \Omega_k$ to the center of this cluster $\bar{\omega}_l$ is calculated as

$$d^2(\bar{\omega}_l, \Omega_k) = \frac{1}{m_k}\sum_{i=1}^{m_k} d^2(\omega_i, \bar{\omega}_l) = \frac{1}{m_k m_l}\sum_{i=1}^{m_k}\sum_{p=1}^{m_l} d_{ip}^2 - \frac{1}{2m_l^2}\sum_{p=1}^{m_k}\sum_{q=1}^{m_k} d_{pq}^2 =$$

$$\frac{1}{m_k m_l}\sum_{i=1}^{m_k}\sum_{p=1}^{m_l} d_{ip}^2 - \sigma_l^2, \quad \omega_p \in \Omega_l,\ \omega_q \in \Omega_l.$$

It is easy to see that $d^2(\bar{\omega}_k, \Omega_l) \neq d^2(\bar{\omega}_l, \Omega_k)$, since dispersions of clusters Ω_l and Ω_k are different, $\sigma_k^2 \neq \sigma_l^2$. Therefore, distance between the centers of two clusters is calculated as the average, where $\omega_i \in \Omega_k$, $\omega_p \in \Omega_l$:

$$d^2(\bar{\omega}_k, \bar{\omega}_l) = d^2(\Omega_k, \Omega_l) = \frac{1}{2}\left(d^2(\bar{\omega}_k, \Omega_l) + d^2(\bar{\omega}_l, \Omega_k)\right) =$$

$$\frac{1}{m_k m_l}\sum_{i=1}^{m_k}\sum_{p=1}^{m_l} d_{ip}^2 - \frac{1}{2}(\sigma_k^2 + \sigma_l^2).$$

After substitution of this distance in (22), we obtain the function $\sigma^2(K)$ in the following form:

$$\sigma^2(K) = \frac{1}{2K^2}\sum_{k=1}^{K}\sum_{l=1}^{K} d^2(\bar{\omega}_k, \bar{\omega}_l) = \frac{1}{2K^2}\sum_{k=1}^{K}\sum_{l=1}^{K}\left(\frac{1}{m_k m_l}\sum_{i=1}^{m_k}\sum_{p=1}^{m_l} d_{ip}^2 - \frac{1}{2}(\sigma_k^2 + \sigma_l^2)\right).$$

However, such a function is, again, not fully correct. Obviously, when the criterion $I_\sigma(K)$ is maximized, distances between cluster centers increase, and the clusters themselves become more compact. However, as the function $\sigma^2(K)$ increases, the cluster dispersions decrease, since clusters become more compact. Therefore, it is necessary to remove cluster dispersions from the expression for $\sigma^2(K)$. In addition, when we get the cluster dispersion again for $k = l$, we need also to delete σ_k^2.

The final correct expression for $\sigma^2(K)$, after removing the contribution from the cluster dispersions and taking into account the symmetry, takes the form

$$\sigma^2(K) = \frac{1}{2K(K-1)} \sum_{k=1}^{K} \sum_{l=1,\, l\neq k}^{K} \frac{1}{m_k m_l} \sum_{p=1}^{m_k} \sum_{q=1}^{m_l} d_{pq}^2, \quad \omega_p \in \Omega_k, \quad \omega_q \in \Omega_l.$$

Therefore, in order to find the optimal linear combination of two components in the criterion (21), the permutable algorithm (e) should be applied.

5.3. The experiments

Let us consider for illustration the well-known Fisher's *Iris Data* (Fisher, 1936). These data consist of 150 measurements of 4 quantitative characteristics of flowers (petal length, petal width, sepal length, and sepal width) belonging to three Iris families (Setosa, Versicolor, and Virginica), with 50 measurements for each family.

It is known that the first family is well separated from the other two in the space of four features. The other two families partially overlap each other. Other sets of measurements are also known: they are associated with various adjustments that are not always clear. In the published classical data set, objects nos. 102 and 143 coincide with each other.

The purpose of the experiments, reported in Dvoenko and Owsiński (2019), is to demonstrate various cases of improving the quality of partitioning of these data by the permutable algorithm when solving the optimization problem

$$\alpha* = \arg\min_{0\leq\alpha\leq 1} J_\delta(K) = \arg\min_{0\leq\alpha\leq 1} \Big((1-\alpha)\tilde{J}(K) + \alpha\delta(K)\Big).$$

Different initial solutions, normalization of initial data, etc. were considered as various conditions. All the results are discussed in details in Dvoenko and Owsiński (2019). Here we consider only some of them for illustration, needed for the purposes of the present paper.

In all experiments, the classical result for $\alpha = 0$ first is obtained under certain initial conditions. Further, under the same initial conditions, the parameter α is varied in the range of $0 < \alpha \leq 1$ with the step of 0.01 to find the optimal value among 100 values. In all cases, the first family (Setosa) is always distinguished entirely. Errors occur, as expected, only when separating the second and third families (see Table 3). It is easy to see that *Iris Data* are well-structured because the number of errors cannot be reduced by the classical initial solutions.

Only the bi-partial quality criterion fundamentally allows for reducing separation errors. In Table 3 the first column shows the initial clusters as a partition into Setosa/Versicolor/Virginica: 50/50/50 is the real partition, 50/70/30 refers to 20 samples moved from Virginica to Versicolor, 50/30/70 refers to 20 samples moved from Versicolor to Virginica. Then, the initial clusters in terms of partition into Versicolor/Virginica are as follows: 50/50 is the real partition, 70/30 means 20 samples from Virginica moved to Versicolor, and 30/70 means 20 samples from Versicolor moved to Virginica. Table 3 and Figs. 1 through 3 demonstrate the decreased number of errors relative to the classical case in corresponding intervals of the parameter α. In all cases the first family, Setosa, is well separated without errors. All errors are encountered only for the partially intersecting families Versicolor and Virginica.

Table 3. Separation of the *Iris Data* flower families

Initial clustering	Errors ($\alpha = 0$)	Intervals (for $\alpha*$)	Errors (for $\alpha*$)	Diagrams
50/50/50	16	0.6 – 0.75	15	Fig. 1
50/70/30	16	0.6 – 0.75	15	Fig. 1
50/30/70	16	0.6 – 0.75	15	Fig. 1
50/50	16	0.81 – 0.92	15	Fig. 2
70/30	16	0.81 – 0.92	15	Fig. 2
30/70	16	0.81 – 0.92	15	Fig. 3

In the second experiment, the well-known problem of separating clusters of different sizes is considered. It is known that the k-means algorithm tries to establish clusters of approximately the same size. In the case of clusters of different sizes, the large one is usually diminished (Versicolor and Virginica together), and the smaller one is expanded (Setosa alone).

In the classical case, i.e. for $\alpha = 0$, three errors were obtained for objects 58, 94, and 99, incorrectly assigned to the first family Setosa. When searching for the optimal value $\alpha*$, all errors are eliminated to zero (Fig. 4) in the range of $0.97 \leq \alpha* \leq 1$ for all initial partitions as Setosa vs. Versicolor/Virginica: 50/100 is the real partition, 100/50 means that all 50 samples from Versicolor were moved to Setosa, 30/120 means that 20 samples from Setosa were moved to Versicolor/Virginica.

5.4. Redistribution of dispersion by the bi-partial criterion

Let us try to uncover the mechanism of improving the quality of clustering when solving the optimization problem (17). Thus, suppose a set of m elements is divided into K disjoint subsets (clusters). It is known that the total variance of data in clustering is divided into the intra- and inter-cluster parts (see, e.g., Duda and Hart, 1973, or Duda, Hart and Stork, 2000).

Let S_T be the overall scatter matrix, S_W be the intra-cluster scatter matrix, S_B be the inter-cluster scatter matrix, where, of course, $S_T = S_W + S_B$. Therefore, it is true that $trS_T = trS_W + trS_B$ for diagonal elements, where $m\sigma_T^2 = m\sigma_W^2 + m\sigma_B^2$ and $\sigma_T^2 = \sigma_W^2 + \sigma_B^2$.

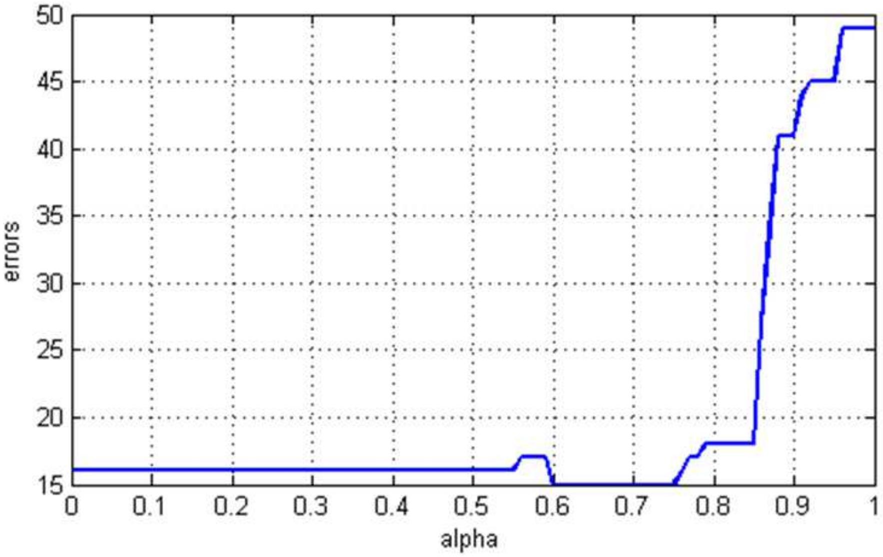

Figure 1. Clustering errors of original *Iris Data*: Setosa/Versicolor/Virginica: 50/50/50, 50/70/30, 50/30/70. All results are the same

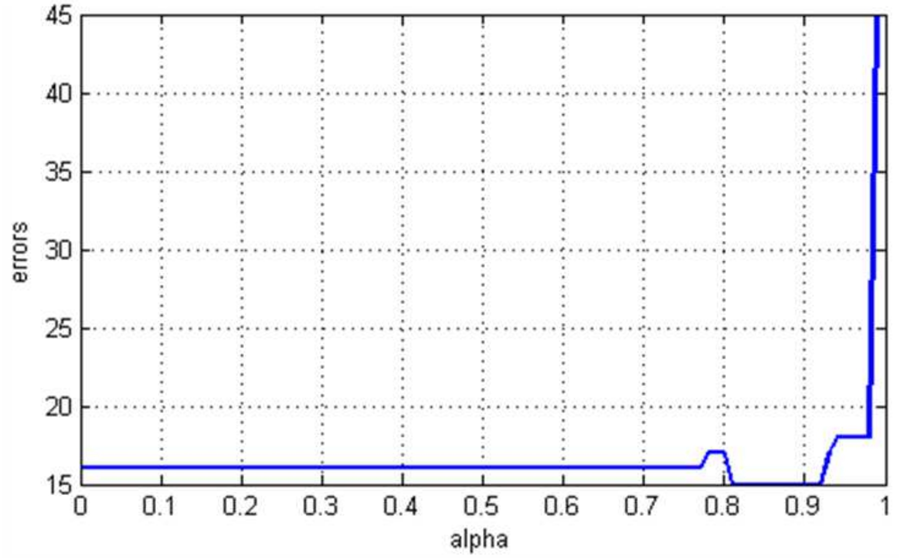

Figure 2. Clustering errors of original *Iris Data*: Versicolor/Virginica: 50/50, 70/30. All results are the same

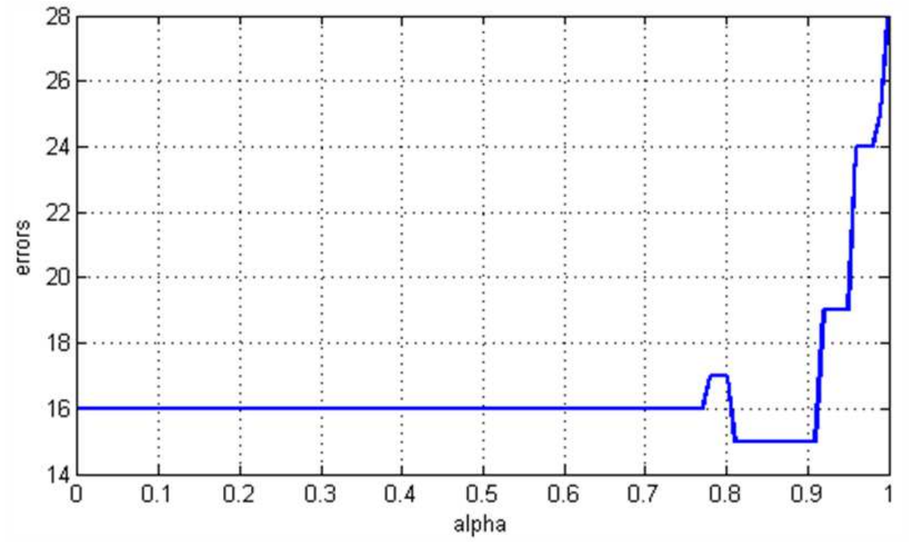

Figure 3. Clustering errors of original *Iris Data*: Versicolor/Virginica: 30/70

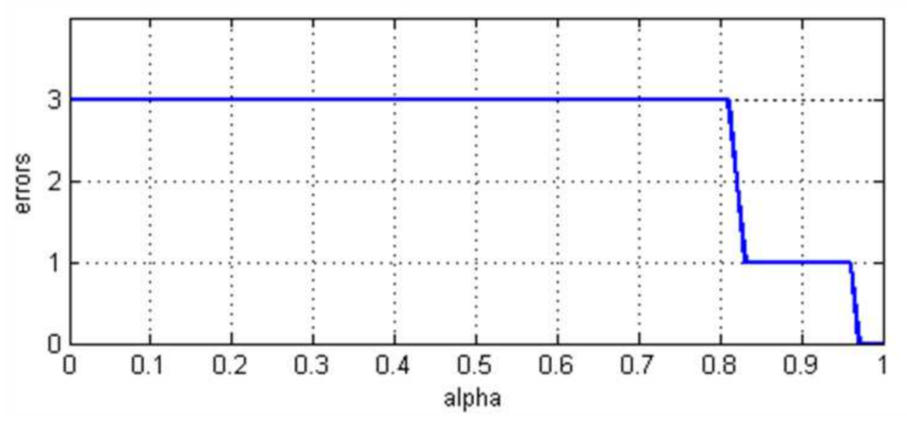

Figure 4. Clustering errors of original *Iris Data*: Setosa vs. Versicolor/Virginica. 50/100, 100/50, 30/120. All results are the same

Let the set $\Omega = \{\omega_1, \ldots \omega_m\}$ be immersed in the metric space and represented by the distance matrix $D(m,m)$, $d_{ij} = d(\omega_i, \omega_j) \geq 0$. Let this set be divided into non-intersecting clusters Ω_k, $k = 1, \ldots K$. Based on Torgerson's formula, the cluster dispersion is calculated as

$$\sigma_k^2 = \frac{1}{2m_k^2} \sum_{p=1}^{m_k} \sum_{q=1}^{m_k} d^2(\omega_p, \omega_q),\ \omega_p \in \Omega_k,\ \omega_q \in \Omega_k, \quad k = 1, \ldots K,$$

the intra-clusters dispersion is calculated as

$$\sigma_W^2 = \sum_{k=1}^{K} \frac{m_k}{m} \sigma_k^2 = \sum_{k=1}^{K} \frac{m_k}{m} \frac{1}{2m_k^2} \sum_{p=1}^{m_k} \sum_{q=1}^{m_k} d^2(\omega_p, \omega_q) =$$

$$\frac{1}{2m} \sum_{k=1}^{K} \frac{1}{m_k} \sum_{p=1}^{m_k} \sum_{q=1}^{m_k} d^2(\omega_p, \omega_q),$$

the general dispersion is calculated as

$$\sigma_T^2 = \frac{1}{2m^2} \sum_{p=1}^{m} \sum_{q=1}^{m} d^2(\omega_p, \omega_q) = \frac{1}{2m^2} \sum_{k=1}^{K} \sum_{l=1}^{K} \sum_{p=1}^{m_k} \sum_{q=1}^{m_l} d^2(\omega_p, \omega_q).$$

The dispersion of cluster centers (inter-cluster) relative to their center $\bar{\omega}_0$ is calculated as

$$\sigma_{IC}^2 = \frac{1}{K} \sum_{k=1}^{K} d^2(\bar{\omega}_k, \bar{\omega}_0) = \frac{1}{2K^2} \sum_{p=1}^{K} \sum_{q=1}^{K} d^2(\bar{\omega}_p, \bar{\omega}_q),$$

where the center $\bar{\omega}_0$ is represented by the distances to centers $\bar{\omega}_k$, based on Torgerson's formula, as

$$d^2(\bar{\omega}_k, \bar{\omega}_0) = \frac{1}{K} \sum_{p=1}^{K} d^2(\bar{\omega}_k, \bar{\omega}_p) - \sigma_{IC}^2.$$

The classical dispersion of the cluster centers is calculated as

$$\sigma_B^2 = \sum_{k=1}^{K} \frac{m_k}{m} d^2(\bar{\omega}_k, \bar{\omega}_0).$$

Therefore, the classical dispersion of cluster centers is calculated as

$$\sigma_B^2 = \sum_{k=1}^{K} \frac{m_k}{m} \left(\frac{1}{K} \sum_{p=1}^{K} d^2(\bar{\omega}_k, \bar{\omega}_p) - \sigma_{IC}^2 \right) =$$

$$\frac{1}{K} \sum_{k=1}^{K} \frac{m_k}{m} \sum_{p=1}^{K} d^2(\bar{\omega}_k, \bar{\omega}_p) - \sigma_{IC}^2 \sum_{k=1}^{K} \frac{m_k}{m} = \frac{1}{K} \sum_{k=1}^{K} \frac{m_k}{m} \sum_{p=1}^{K} d^2(\bar{\omega}_k, \bar{\omega}_p) - \sigma_{IC}^2.$$

As shown above, minimizing the criterion $\tilde{J}(K)$, based on the distance matrix $D(m,m)$ means maximizing the dual criterion $I(K) = C - \tilde{J}(K)$ for the similarity matrix $S(m,m)$, where $s_{ij} \geq 0$. In the criterion $I_\sigma(K) = (1-\alpha)I(K) + \alpha\sigma^2_{IC}$, the first component $I(K)$ is maximized as the weighted average compactness of clusters, and the second component σ^2_{IC} is maximized as the inter-cluster dispersion, which is calculated according to (22). The decomposition of the total dispersion is represented as

$$\sigma^2_T = \sigma^2_W + \frac{1}{K}\sum_{k=1}^{K}\frac{m_k}{m}\sum_{p=1}^{K} d^2(\bar{\omega}_k, \bar{\omega}_p) - \sigma^2_{IC}.$$

Let us denote the union of the classical weighted average variance of cluster centers σ^2_B and the inter-cluster dispersion σ^2_{IC} as

$$\sigma^2_{B\cup IC} = \frac{1}{K}\sum_{k=1}^{K}\frac{m_k}{m}\sum_{p=1}^{K} d^2(\bar{\omega}_k, \bar{\omega}_p).$$

Hence, we get $\sigma^2_B = \sigma^2_{B\cup IC} - \sigma^2_{IC}$. The decomposition of the total variance takes the form

$$\sigma^2_T + \sigma^2_{IC} = \sigma^2_W + \sigma^2_{B\cup IC}. \tag{23}$$

As we can see, the permutable algorithm (d) minimizes according to (1), and (3) the criteria $\tilde{J}(K) = J(K) = \sigma^2_W$. Since $\sigma^2_T = const$, the dispersion $\sigma^2_B = \sigma^2_{B\cup IC} - \sigma^2_{IC}$ is maximized with the balance $\sigma^2_T = \sigma^2_W + \sigma^2_B$ maintained. Then, in the decomposition (23), both parts are increasing while maintaining balance.

In this case, the maximization of the bi-partial objective function $I_\sigma(K)$ affects only the maximization of the dispersion σ^2_{IC}. At the same time, the maximization of the other part of the dispersion $\sigma^2_{B\cup IC}$ by the permutable algorithm is not controlled.

6. Assessing the number of clusters

6.1. Some well-known ideas

As it is well known, the problem of determining the number of subsets, into which a set is split so that they form well justified clusters is theoretically difficult in general. At the same time, in cluster analysis, various practical approaches have been intensively developed based on devising the appropriate criteria and algorithms.

It is known that clustering algorithms, for example, the ones discussed above, from the k-means family, are locally optimal. Therefore, the quality of their results depends not only on the initial solution but also on the number of clusters assumed. Moreover, if the cluster structure is not sufficiently evident, then this also affects the result of processing (insufficiently distant or partially overlapping clusters, etc.). Under such conditions, the type of measure of similarity (or difference) of clusters also plays an essential role. Therefore, in a practical approach, this problem belongs to the class of multi-criteria optimization problems, see Duda and Hart (1973) and Duda, Hart and Stork (2000).

It is easy to see that the criterion (1), discussed above, cannot be optimized relative to the number of clusters. It takes the maximum value of data dispersion when all objects are in the single cluster, and the minimum, zero value, when each object forms its own cluster.

One of the directions of development of methods for determining the unknown number of clusters is associated with the design of special criteria that can contain extremes concerning the number of clusters (see Aivazyan et al., 1989). Note that in the above considered approach, following Owsiński (2020), it is also natural to determine the unknown number of subsets of multidimensional data as clusters based on the generalized objective function $Q_S^D(P)$. When searching for the best approximation, concerning both cluster content and the number of clusters, a combination of intra-cluster similarity with inter-cluster dissimilarity in a single partitioning quality criterion is used.

The practical algorithms usually come in two categories. In the first case, it is necessary to specify first a suitable number of clusters (k-means, M- and S-algorithms, etc.), on the basis of a priori information for building clustering. In the second case, algorithms are often developed for finding a suitable number of clusters in one way or another, with the development and use of the corresponding criteria of clustering quality (Isodata, Forel, hierarchical algorithms, see Duda and Hart, 1973; Duda, Hart and Stork, 2000, or Zagoruiko, 1999). A sufficiently ample review thereof is given in Aivazyan et al. (1989).

6.2. A quasi-hierarchical procedure

It should be noted that the notion of hierarchical search or hierarchical algorithms has a prominent place within the domain of clustering, see, e.g., Ward (1963). The hierarchical procedures of clustering constitute a sequential search for subsets of a given set by both merging, which starts from individual elements (agglomerative procedures), and splitting, which starts from the complete set (divisive procedures), which are trivial clusters (and trivial partitions). The corresponding criterion, which is used together with the procedure, allows for determining a certain level of merging (partitioning), which establishes the obtained subsets as (potentially) non-trivial clusters and their number.

If the search within a hierarchy itself is not in focus, then let a particular procedure be developed based on iterating through subsets to determine an unknown number of clusters (see Dvoenko, 2001, 2009a). It is known that the hierarchical partitioning at a certain level of the dendrogram, i.e. the tree, formed through consecutive aggregations or splits, may not always be optimal for the compactness of the resulting clusters. This means, in particular, that the corresponding partition is biased. Therefore, breaking the hierarchy at this level may restore the unbiased clustering with the more compact clusters.

Let us try to build a divisive dendrogram based on k-means algorithm according to the criterion (1), sequentially increasing the number of clusters, $K = 1, \dots m$. So, based, for example, on the *distance* k-means (b), a divisive quasi-hierarchical procedure is developed as (g):

(g) Step $K = 0$. Define cluster $\Omega_{K+1} = \Omega$, $m = |\Omega|$, $K = K + 1$.

Step K. Increase the number of clusters by one:

1. While $k = 1, \dots K$ find the least compact cluster Ω_k.
2. Define two representatives, $\tilde{\omega}_k$ and $\tilde{\omega}_{K+1}$, as the most distant objects in Ω_k.
3. Split Ω_k into two clusters Ω_k and Ω_{K+1} by *distance* k-means (b).
4. Define $K + 1$ representatives as $\tilde{\omega}_1, \dots \tilde{\omega}_{k-1}, \tilde{\omega}_{k+1}, \dots \tilde{\omega}_K$ with $\tilde{\omega}_k$ and $\tilde{\omega}_{K+1}$ after the preceding point 3.
5. Split Ω into $K + 1$ clusters $\Omega_1, \dots \Omega_{K+1}$ by *distance* k-means (b).
6. $K = K + 1$. Stop, if $K = m$.

Let us consider the sufficiently obvious properties of this procedure. The sequence of partitions begins with a single set $\Omega_1 = \Omega$ and ends with the singleton sets $\Omega_1, \dots \Omega_m$. In general, in the sequence of partitions, not all of them may be included in a hierarchy.

Obviously, two partitions, into K and $K + 1$ subsets, become a hierarchy when splitting the least compact cluster Ω_k into two Ω_k and Ω_{K+1} immediately gives an unbiased partition $\Omega_1, \dots \Omega_{K+1}$. A hierarchy violation occurs when splitting the least compact cluster Ω_k into two subsets, Ω_k and Ω_{K+1}, requires re-splitting all the set Ω into $K + 1$ subsets in such a way that the partition $\Omega_1, \dots \Omega_{K+1}$ becomes unbiased.

Using this algorithm, a set of partial dendrograms is developed, each starting with a partition that violates the current hierarchy. The violation of the hierarchy shows that a better partition is obtained for a given numbers of clusters at a given level of the dendrogram.

We assume that all such violating partitions determine the preferred numbers of clusters. Therefore, the set of partial dendrograms (their initial levels) determines the set of preferred of clusters. Indeed, this may be the case, since the subsequent levels in partial dendrograms simply show the hierarchical splitting of the clusters defined in the violating partition up to the next violation of the hierarchy.

On the other hand, in hierarchical algorithms, the suitable number of clusters is usually determined based on the following empirical rule. The optimal number of clusters is established at the boundary, up to which the dispersion of clusters decreases quickly, and after which its decrease slows down sharply. Therefore, this empirical rule should be implemented also in the sequence of unbiased partitions formed by the sequence of partial dendrograms. In any case, this situation is more correct relative to the criterion (1) regarding the determination of the number of clusters, than for hierarchy, which can contain some biased partitions, in general.

If a single dendrogram is obtained, then it can be shown, Dvoenko (2001), that the quasi-hierarchical procedure for distances is equivalent to the algorithm for finding the minimal spanning tree.

It is obvious that the procedure (g) is a superstructure over clustering algorithms that generate representatives. Since the permutable algorithms do not generate representatives, it is also possible to develop a quasi-hierarchical procedure for the corresponding subsets. It should be noted that in the quasi-hierarchical procedure, the problem of the initial solution for the k-means algorithm is reduced to only one case.

6.3. Experiments

First, let us look at the data on correlations of eight physical variables, characterizing the body measurements, discussed before. Recall that the purpose of factor analysis for this particular data set is to identify two groups of physical characteristics (see Harman, 1976). Here, in this paper, this goal is achieved as a solution to the clustering problem. Such groups have been successfully identified.

In the case of using the *similarity* k-means algorithm, the criterion (11) is used as the weighted average compactness of the cluster structure, $I(K)$, to be maximized. In the case of correlations (Section 4.2), one should maximize the functional

$$I_M(K) = \sum_{k=1}^{K} \sum_{p=1}^{n_k} |r(\bar{\omega}_k, \omega_p)|.$$

The quasi-hierarchical grouping based on the *similarity* k-means algorithm gives a dendrogram, i.e., hierarchy itself, and a sequence of increasing values of the clustering criterion $I_M(K),\ K = 1, \dots 8$, is as follows: 4.63, 6.42, 6.89, 7.27, 7.5, 7.74, 7.88 and 8. The respective diagram (Fig. 5) shows a sharp increase of the clustering criterion to the number of clusters equal two and a subsequent slowdown after it. The empirical rule indicates, therefore, that $K = 2$.

Other data are more complex: Holzinger's data are given in Harman (1976) and represent correlations between 24 psychological tests in a study of mental development of 145 Chicago suburb schoolchildren in 1934. The set of tests is initially divided into five groups, each of which consisted of tests (features) characterizing one of aspects of mental development.

The objective of the first study was to demonstrate the properties of the bi-factor method, in which group factors are developed for predefined groups of psychological tests. As it turned out, the complexity of Holzinger's data did

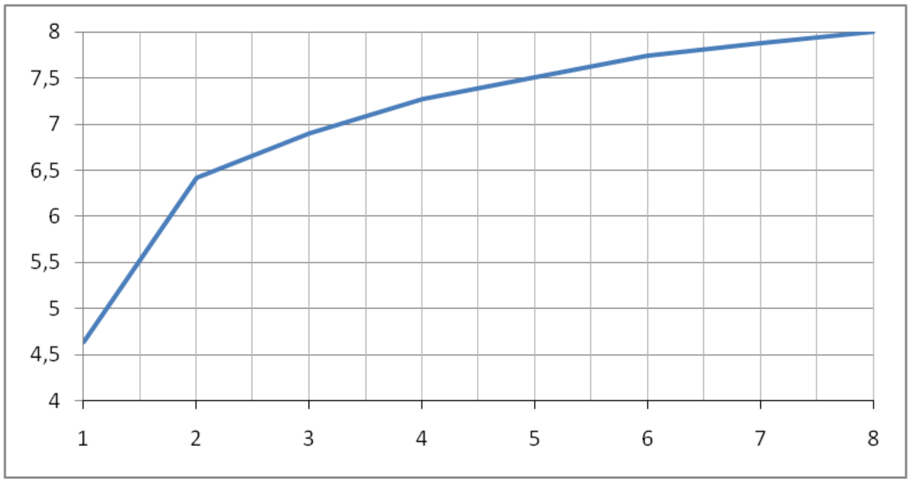

Figure 5. Eight physical variables: $K = 2$. Horizontal axis is the number of clusters, vertical axis is the value of the criterion $I_M(K)$

not allow for separating the group factor for the tests of the fifth group. These tests had sufficient factor loads only as a part of an additional common factor built for all tests.

The next attempt to get an insight into the complexity of these data was made in 1970 to demonstrate the use of methods of an extreme grouping of parameters (LPCA and LCCA), see Braverman (1970), Braverman and Muchnik (1983), and Lumel'sky (1970). However, Holzinger's groups also failed to be restored as ideal ones by these methods. As before, differences in results depend on tests from the 5th group. As a rule, some tests from it fell into other groups. As a consequence, at the same time, some other tests were usually forced out of their groups.

As the result, in general, the M-algorithm was better than the S-algorithm. Under the so-called standard initial conditions (the first K tests form separate groups, the rest join the closest group), an unsatisfactory result was obtained (Table 4). This is easy enough to understand because using the pre-ordered correlated tests creates an inconvenient initial solution.

When the original groups were taken as the initial ones, it was also not possible to restore them, because test 24 fell into group 3, and test 19 moved to group 5. It should be noted that according to this result, the original (ideal) groups of tests should be treated as biased ones.

Here, it seems that peculiarities of the realization of algorithms were also affected by the limited resources of the computing devices at the time the calculations were performed.

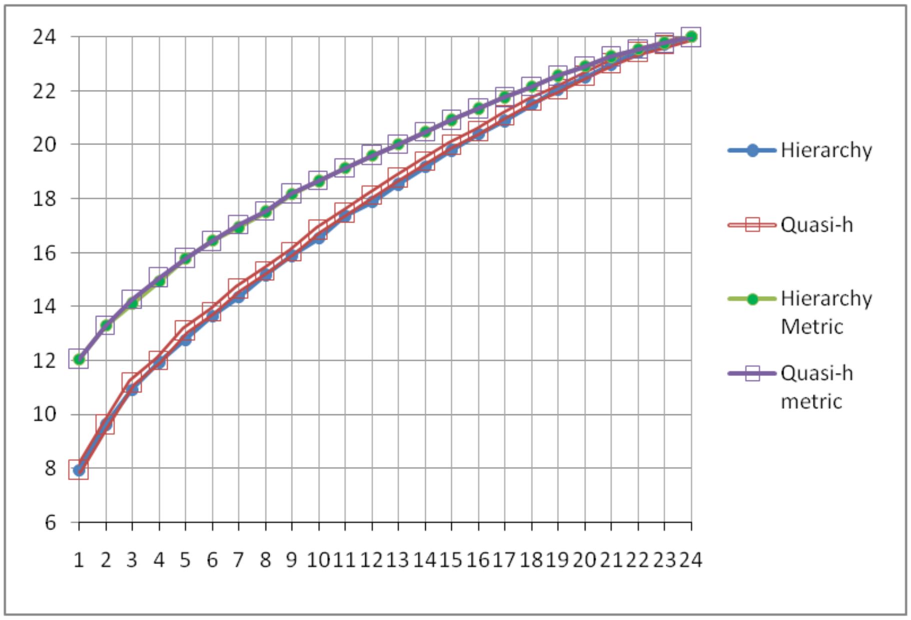

Figure 6. Holzinger's data: K is undefined. Horizontal axis is the number of clusters, vertical axis is the value of the criterion $I_M(K)$

Table 4. Results of the M-algorithm for Holzinger's data

Group			Standard initial conditions	Ideal groups as initial ones
1	Spatial relations	1 – 4	1 – 4, 20, 22, 23	1 – 4
2	Verbal	5 – 9	5 – 9	5 – 9
3	Perceptual speed	10 – 13	10 – 13	10 – 13, 24
4	Memory	14 – 19	14 – 17	14 – 18
5	Deduction	20 – 24	18, 19, 21, 24	19 – 23

The third attempt was made in 2009 to demonstrate the procedure of a quasi-hierarchical clustering, Dvoenko (2009a). Table 5 shows some results from using the k-means algorithm. Increasing values of the criterion $I_M(K)$, $K = 1, \dots 24$, for hierarchy and quasi-hierarchy were obtained. Values $I_M(K)$, $K = 5, \dots 24$, were also obtained starting from the initial (ideal) groups of tests (Fig. 6).

It needs to be noted that all three sequences (left columns in Table 5) of values of the quality indices increase smoothly, without kinks (Fig. 6). This does not allow us to apply a heuristic method of determining the appropriate number of groups.

Also, the hierarchical grouping shows generally the worst quality of results compared to those produced by the quasi-hierarchy (see Table 5). In a quasi-hierarchical grouping, splitting into three groups breaks the hierarchy and gives better quality. The hierarchy is not violated further, and the quality of quasi-hierarchical partitions is systematically better up to the splitting into 21 groups. At the end, all results coincide for obvious reasons, i.e. due to the lack of other variants of partitions.

It is important to note that the so-called ideal partition is unbiased, forming the beginning of the hierarchy (Table 5). Moreover, the quality of such a partition becomes the best. This confirms the hypothesis about the complexity of data and the multi-extreme criterion function. It is obvious that the methods used in the second attempt and earlier to obtain a partition into five groups do not lead to an ideal partition (Table 6). This means that the task of finding an initial solution by itself becomes the non-trivial and comparable in complexity to the basic clustering problem for these data.

The fact that the original Holzinger's partition becomes unbiased means that these groups are separable in a metric space. According to this, we note that the compactness hypothesis, as a philosophical principle, applies not only to single-point manifolds (cluster centers) but also to broader manifolds, such as: linear (regressions, separating hyper-planes), nonlinear (separating hyper-surfaces), etc.

The discussion of linear decision functions takes us beyond the subject of this paper. However, it should be noted that Holzinger's groups are linearly separable in a metric space. This means that the bi-factor analysis task is a problem of learning with a teacher (machine learning). As a result, it becomes clear that each Holzinger's group is linearly separable from the remaining groups, with the high cross-validation quality. The development of the learning algorithm and experiments are discussed in deeper detail in Dvoenko (2009a).

If we remain within the framework of the cluster analysis problem, then the best approximation in terms of the composition of groups gives a quasi-hierarchical separation into 10 groups, see Dvoenko (2009a). In Table 6, for each of Holzinger's groups, the tests in the combined subgroups are shown in parentheses.

Finally, we note that Holzinger's data are represented through a positively defined correlation matrix, consisting of positive values, except for one $r_{3,10} = r_{10,3} = -0.075$. In the studies, reported in Harman (1976), this value is assumed to be zero, while in the studies, reported in Braverman (1970), Braverman and Muchnik (1983), Dvoenko (2009a) and Lumel'sky (1970), the modules of correlations are considered.

As shown previously in this paper, in order to obtain a similarity function based on the law of cosines, it is necessary to move the origin of the coordinates beyond the convex hull of the set, so that all correlations really become positive. To do this, the correlation matrix is transformed into a distance matrix with

Table 5. Comparison of hierarchies and quasi-hierarchies, the values of $I_M(K)$

Number of groups	Hierarchy	Quasi-hierarchy	Ideal initial groups	Origin out of the convex hull	
				Hierarchy	Quasi-hierarchy
1	7.94	7.94	-	12.05	12.05
2	9.64	9.64	-	13.29	13.29
3	10.91	***11.17***	-	14.12	***14.27***
4	11.94	12.00	-	14.94	15.08
5	12.76	13.11	13.34	15.79	15.79
6	13.65	13.81	14.16	16.44	16.44
7	14.35	14.66	14.92	16.94	17.04
8	15.18	15.32	15.62	17.52	17.54
9	15.89	16.03	16.32	18.18	18.20
10	16.53	16.84	16.96	18.66	18.68
11	17.35	17.49	17.65	19.14	19.14
12	17.88	18.14	18.17	19.61	19.61
13	18.53	18.78	18.83	20.02	20.02
14	19.17	19.40	19.47	20.48	20.48
15	19.79	20.00	19.97	20.92	20.92
16	20.39	20.50	20.56	21.35	21.35
17	20.89	21.09	21.11	21.76	21.76
18	21.48	21.64	21.64	22.16	22.16
19	22.05	22.07	22.07	22.56	22.56
20	22.49	22.57	22.57	22.91	22.91
21	22.98	23.03	23.03	23.27	23.27
22	23.45	23.45	23.45	23.55	23.55
23	23.72	23.72	23.72	23.78	23.78
24	24.00	24.00	24.00	24.00	24.00

elements $d_{ij} = \sqrt{2(1 - r_{ij})}$ and the origin is found for $\Delta = 0$, as shown before in Section 2.4. This means that this origin of coordinates is placed outside of the convex hull of the set. Further, the non-normalized similarity matrix is restored with elements

$$s_{ij} = \frac{1}{2d_{0i}d_{0j}}(d_{0i}^2 + d_{0j}^2 - d_{ij}^2),$$

and the normalized similarity matrix is finally obtained by the transformation $s'_{ij} = s_{ij}/\sqrt{s_{ii}s_{jj}}$. The respective results are shown in Table 7.

It is easy to see that these partitions are also not similar to the ideal one for the same reason as that discussed above for clustering. In addition, as

Table 6. Results for the hierarchy and quasi-hierarchy

Group			Hierarchy	Quasi-hierarchy	10 groups
1	Spatial relations	1-4	1-3	1-3	(1,3) (2) (4,22)
2	Verbal	5-9	4-9, 13, 20-23	4-9, 20, 22, 23	(5-9)
3	Perceptual speed	10-13	10-12, 24	10-13, 21, 24	(10-13)
4	Memory	14-19	14-17	14-17	(14,16) (15,17) (18,19)
5	Deduction	20-24	18, 19	18, 19	(20,23) (21,24)

Table 7. Results of hierarchy and quasi-hierarchy for similarities

Group			Hierarchy	Quasi-hierarchy
1	Spatial relations	1 – 4	1 – 3	1 – 3
2	Verbal	5 – 9	4 – 9, 20, 22, 23	4 – 9, 20, 22, 23
3	Perceptual speed	10 – 13	10 – 13, 21, 24	10 – 13, 21, 24
4	Memory	14 – 19	14, 16, 19	14 – 16, 19
5	Deduction	20 – 24	15, 17, 18	17, 18

before, the result for the quasi-hierarchy also becomes better than the result for hierarchy (Table 5, right hand columns). As before, splitting into three groups violates the hierarchy. Naturally, the values of the partitioning quality criterion start with higher values, because in the single quadrant of the metric space all the tests are located more closely (Fig. 6).

This experiment now metrically confirms the previously drawn conclusion about the unbiased original grouping and, consequently, about the linear separability of the Holzinger's groups of tests from each other.

7. Conclusion

The processing of pairwise comparisons continues to be quite an interesting problem, since experimental data are often inconvenient or even impossible to be presented in the traditional form as the results of measurements of some pointwise characteristics. This situation requires the development of appropriate methods and algorithms.

In general, the solution to this problem seems to be achieved in at least three important directions. The first one is the development of new and a modification of known machine learning algorithms. This problem is discussed here on the basis of the clustering problem with the use of the k-means algorithm.

The second one is the immersion of paired comparisons in a metric space. This problem is related to correction of pairwise comparisons, see Bognar (1974), Dvoenko and Pshenichny (2018), Pekalska and Duin (2005). The third one is solving of some specific problems, e.g., immersion of binary relations in a metric space, and is related to increasing the power of non-quantitative measuring scales, see Dvoenko and Pshenichny (2021), Kemeny (1959), Litvak (1982), or Luce (1959).

The author thanks the anonymous referees for their attention to the work, useful comments and suggestions.

Funding

This investigation was supported by the Russian Foundation for Basic Research (RFBR), Project number 20-07-00055.

References

Aivazyan, S. A. et al. (1989) *Applied Statistics. Classification and Reduction of Dimensionality* [in Russian]. FiS, Moscow.

Aizerman, M. A., Braverman, E. M. and Rozonoer, L. I. (1970) *The Method of Potential Functions in Machine Learning Theory* [in Russian]. Nauka, Moscow.

Bognar, J. (1974) *Indefinite Inner Product Spaces.* Springer-Verlag, New York.

Braverman, E. M. (1970) Methods of extremal grouping of parameters and problem of apportionment of essential factors [in Russian]. *Avtomat. i Telemekh.* **1**, 123–132.

Braverman, E. M. et al. (1971) Diagonalization of the relation matrix and detecting hidden factors [in Russian]. *Trans. Inst. of Control Sciences. 1st Issue "Problems of increasing of automata possibilities."* ICS, Moscow, 42–79.

Braverman, E. M. and Muchnik, I. B. (1983) *Structured Methods of Empirical Data Processing* [in Russian]. Nauka, Moscow.

Diday, E., Bochi, S., Brossier, G. and Celeux, G. (1979) *Optimisation en Classification Automatique. 2.* Institut national de recherche en informatique et en automatique (INRIA), Le Chesnay (in French).

Duda, R. O. and Hart, P. E. (1973) *Pattern Classification and Scene Analysis*. Wiley, New York.

Duda, R. O., Hart, P. E. and Stork, D. G. (2000) *Pattern Classification.* Wiley, New York.

Dvoenko, S. D. (2001) Restoration of spaces in data by the method of nonhierarchical decompositions. *Automation and Remote Control.* **62**, 467–473. //doi.org/10.1023/A:1002814429456

Dvoenko, S. D. (2009a) Clustering and separating of a set of members in terms of mutual distances and similarities. *Trans. on MLDM.* IBaI Publishing, **2**(2), 80–99.

Dvoenko, S. D. (2009b) Clustering of a set described by paired distances and closeness between its elements [in Russian]. *Sib. J. of Industr. Math.* **12**(1), 61–73.

Dvoenko, S. D. (2011) On clustering of a set of members by distances and similarities. *Proc. of 11th Int. Conf. on Pattern Recognition and Information Processing (PRIP'2011).* BSUIR, 104–107.

Dvoenko, S. (2014) Meanless k-means as k-meanless clustering with the bipartial approach. *Proc. of 12th Int. Conf. on Pattern Recognition and Information Processing (PRIP'2014).* UIIP NASB, 50–54.

Dvoenko, S. and Owsinski, J. (2019) The permutable k-means for the bipartial criterion. *Informatica.* **43**(2), 253–262. //doi.org/10.31449/inf.v43i2.2090

Dvoenko, S. D. and Pshenichny, D. O. (2018) On metric correction and conditionality of raw featureless data in machine learning. *Pattern Recognit. Image Anal.* **28,** 595–604. //doi.org/10.1134/S1054661818040089

Dvoenko, S. D. and Pshenichny, D. O. (2021) Rank aggregation based on new types of the Kemeny's median. *Pattern Recognit. Image Anal.* **31,** 185–196. //doi.org/10.1134/S1054661821020061

Fisher, R. A. (1936) The use of multiple measurements in taxonomic problems. *Ann. Eugenics.* **7**(2), 179-188.

Friedman, H. P. and Rubin, J. (1967) On some invariant criteria for grouping data. *Journal of the American Statistical Association* **62** (320), 1159–1178. //doi.org/10.1080/01621459.1967.10500923

Harman, H. H. (1976) *Modern Factor Analysis.* University of Chicago Press, Chicago.

Hartigan, J. A. and Wong, M. A. (1979) Algorithm AS 136: A k-means clustering algorithm. *J. Roy. Soc.* **28**(1), 100–108. //doi.org/10.2307/2346830

Kemeny, J. (1959) Mathematics without numbers. *Daedalus*, **88**(4), 577–591.

Litvak, B. G. (1982) *Expert Information: Methods of Acquisition and Analysis* [in Russian]. Radio i Svyaz, Moscow.

Lumel'sky, V. Ya. (1970) Grouping of parameters on the basis of communication matrices [in Russian]. *Avtomat. i Telemekh.* **1**, 133–143.

Luce, R. D. (1959) *Individual Choice Behavior: A Theoretical Analysis.* Wiley, New York.

Mercer, J. (1909) Functions of positive and negative type and their connection with the theory of integral equations. *Philos. Trans. Roy. Soc.*, London.

Owsiński, J. W. (2020) *Data Analysis in Bi-partial Perspective: Clustering and Beyond.* SCI **818**, Springer. //doi.org/10.1007/978-3-030-13389-4

Pekalska, E. and Duin R. P. W. (2005) *The Dissimilarity Representation for Pattern Recognition. Foundations and Applications.* World Scientific, Sngapore.

Rosenblatt, F. (1962) *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms.* Spartan Books, Washington.

Schlesinger, M. (1965) About spontaneous recognition of patterns [in Russian]. *Reading Automations.* Kiev, 38–45.

Späth, H. (1983) *Cluster-Formation und -Analyse: Theorie, FORTRAN-Programme und Beispiele* [in German]. R. Oldenbourg Verlag, München–Wien.

Torgerson, W. S. (1958) *Theory and Methods of Scaling.* Wiley, New York.

Ward, J. (1963) Hierarchical grouping to optimize an objective function. *J. American Statist. Ass.* **58**(301), 236–244. //doi.org/10.1080/01621459.1963.10500845

Young, G. and Householder, A. S. (1938) Discussion of set of points in terms of their mutual distances. *Psychometrica.* **3**, 19–22. //doi.org/10.1007/BF02287916

Zagoruiko, N. G. (1999) *Applied Methods of Data and Knowledge Analysis* [in Russian]. IM SBRAS, Novosibirsk.