An Approximation Algorithm for a Problem of Partitioning a Sequence into Clusters with Restrictions on Their Cardinalities

Alexander Kel'manov[1,2], Ludmila Mikhailova[1], Sergey Khamidullin[1], and Vladimir Khandeev[1,2(✉)]

[1] Sobolev Institute of Mathematics, 4 Koptyug Ave., 630090 Novosibirsk, Russia
{kelm,mikh,kham,khandeev}@math.nsc.ru
[2] Novosibirsk State University, 2 Pirogova St., 630090 Novosibirsk, Russia

Abstract. We consider the problem of partitioning a finite sequence of points in Euclidean space into a given number of clusters (subsequences) minimizing the sum of squared distances between cluster elements and the corresponding cluster centers. It is assumed that the center of one of the desired clusters is the origin, while the centers of the other clusters are unknown and determined as the mean values over clusters elements. Additionally, there are a few structural restrictions on the elements of clusters with unknown centers: (1) clusters form non-overlapping subsequences of the input sequence, (2) the difference between two consecutive indices is bounded from below and above by prescribed constants, and (3) the total number of elements in these clusters is given as an input. It is shown that the problem is strongly NP-hard. A 2-approximation algorithm which runs in polynomial time for a fixed number of clusters is proposed for this problem.

Keywords: Clustering · Structural constraints · Euclidean space · Minimum sum-of-squared distances · NP-hardness · Guaranteed approximation factor

1 Introduction

The subject of this study is a problem of partitioning a finite sequence of points in Euclidean space into subsequences. Our goal is to find out the computational complexity of the problem and to provide a polynomial-time factor-2 approximation algorithm.

The research is motivated by insufficient study of the problem and its relevance, in particularly, to problems of approximation, clustering, sequence (time series) analysis as well as to many natural science and engineering applications that require classification of results of chronologically sorted numerical experiments and observations on the state of some objects (see, for example, [1–4] and references therein). Some applications (sources) of the problem are presented in the next section.

Y. Kochetov et al. (Eds.): DOOR 2016, LNCS 9869, pp. 171–181, 2016.
DOI: 10.1007/978-3-319-44914-2_14

This is the incremental work to the results previously obtained in [5–7]. Each of the cited works is an essential building-block in the algorithm presented in this work — the first algorithm with a guaranteed approximation factor.

2 Problem Formulation, Complexity, and Related Problems

Everywhere below $\mathbb{R}$ denotes the set of real numbers, $\|\cdot\|$ denotes the Euclidean norm, and $\langle\cdot,\cdot\rangle$ denotes the scalar product.

Formally, we consider the following problem.

Problem 1. Given a sequence $\mathcal{Y} = (y_1, \ldots, y_N)$ of points from $\mathbb{R}^q$ and some positive integers $T_{\min}$, $T_{\max}$, L, and M. *Find* nonempty disjoint subsets $\mathcal{M}_1, \ldots, \mathcal{M}_L$ of $\mathcal{N} = \{1, \ldots, N\}$, i.e. subsets of indices of the elements from the sequence $\mathcal{Y}$, such that

$$F(\mathcal{M}_1, \ldots, \mathcal{M}_L) = \sum_{l=1}^{L} \sum_{j \in \mathcal{M}_l} \|y_j - \overline{y}(\mathcal{M}_l)\|^2 + \sum_{i \in \mathcal{N} \setminus \mathcal{M}} \|y_i\|^2 \longrightarrow \min\,, \qquad (1)$$

where $\mathcal{M} = \bigcup_{l=1}^{L} \mathcal{M}_l$, and $\overline{y}(\mathcal{M}_l) = \frac{1}{|\mathcal{M}_l|} \sum_{j \in \mathcal{M}_l} y_j$ is the centroid of subset $\{y_j | \, j \in \mathcal{M}_l\}$, under the following constraints: (i) the cardinality of $\mathcal{M}$ is equal to M, (ii) concatenation of elements of subsets $\mathcal{M}_1, \ldots, \mathcal{M}_L$ is an increasing sequence, provided that the elements of each subset are in ascending order, (iii) the following inequalities for the elements of $\mathcal{M} = \{n_1, \ldots, n_M\}$ are satisfied:

$$T_{\min} \le n_m - n_{m-1} \le T_{\max} \le N, \quad m = 2, \ldots, M\,. \qquad (2)$$

From the above formulation, it is clear that Problem 1 belongs to the class of clustering problems with a quadratic criterion. Clusters are the unknown index subsets $\mathcal{M}_1, \ldots, \mathcal{M}_L, \mathcal{N} \setminus \mathcal{M}$ and the corresponding subsequences of the input sequence.

One of the sources of Problem 1 is the next problem which is typical for many natural science and technical applications, in particular, for noise-proof remote monitoring, electronic intelligence, analysis and recognition of biomedical and speech signals, data mining, machine learning, and others.

There is a series of N chronologically ordered measurements $y_1, \ldots, y_N$ of a q-tuple y of numerical characteristics of some object. The object has $L+1$ states. Among them L states are active and one state is passive. In the passive state all the numerical characteristics in the tuple equal zero, while, in each active state the value of at least one characteristic is nonzero. The data contains some measurement errors. It is known that for some time the object is located in one of the active states, and then switches to a different active state. At that all the active states of the object are accompanied by a switching into the passive state for some unknown time interval which is bounded from above and below. In addition we are given the natural numbers $T_{\min}$ and $T_{\max}$, which correspond

to the minimum and maximum time interval between any two successive active states of the object. The correspondence of the sequence element to some state of the object is not known in advance. It is required to find the sequence of active states of the object and to estimate the characteristics of the object in each of the active states (which correspond to the respective cluster centers).

Formalization of this problem with respect to the criterion of the minimum sum of squared deviations induces the following approximation problem. Given a sequence $\mathcal{Y} = (y_1, \ldots, y_N)$ of points from $\mathbb{R}^q$ and some positive integers $T_{\min}$, $T_{\max}$, L, and M. Find an approximating sequence $z_1, \ldots, z_N$ having the following structure

$$z_n = \begin{cases} x_1, & n \in \mathcal{M}_1 \ , \\ \ldots & \\ x_L, & n \in \mathcal{M}_L \ , \\ 0, & n \in \mathcal{N} \backslash \mathcal{M} \ , \end{cases} \tag{3}$$

where $x_1, \ldots, x_L$ are unknown points from $\mathbb{R}^q$, such that

$$\sum_{i \in \mathcal{N}} \|y_i - z_i\|^2 \longrightarrow \min \ , \tag{4}$$

under the same constraints on the numbers from subsets $\mathcal{M}_l, \ldots, \mathcal{M}_L$, and $\mathcal{M}$ as in Problem 1.

Schematically, the segment of sequence z_n, $n \in \mathcal{N}$, has the following structure

$$\ldots 0 x_{l-1} 0 \ldots 0 x_{l-1} 0 \ldots\ldots 0 x_l 0 \ldots 0 x_l 0 \ldots \ . \tag{5}$$

Here $x_{l-1}, x_l \in \mathbb{R}^q$ are unknown nonzero points corresponding to the $(l-1)$-th and l-th active states of the object, 0 corresponds to the passive state of the object. The number of zero points between the nonzero points is unknown and lies within the admissible range from $T_{\min} - 1$ to $T_{\max} - 1$ in accordance with the constraints (2).

Relying on (3), expanding the sum (4) and grouping the terms, it is easy to verify by differentiation that the values $x_l = \overline{y}(\mathcal{M}_l)$, $l = 1, \ldots, L$, are optimal in the sense of (4), and thus the formulated approximation problem induces Problem 1. Herein in the optimal approximating sequence, the segment (5) has the following form

$$\ldots 0 \overline{y}(\mathcal{M}_{l-1}) 0 \ldots 0 \overline{y}(\mathcal{M}_{l-1}) 0 \ldots\ldots 0 \overline{y}(\mathcal{M}_l) 0 \ldots 0 \overline{y}(\mathcal{M}_l) 0 \ldots$$

For all $l = 1, \ldots, L$ in this sequence, the indices from the set $\mathcal{M}_l$, the cluster $\{y_j \,|\, j \in \mathcal{M}_l\}$, and its centroid $\overline{y}(\mathcal{M}_l)$ are determined as the result of solving Problem 1. Centroid $\overline{y}(\mathcal{M}_l)$ is an estimate for the point x_l.

From the above mentioned schematic record of sequences in the string form, it is evident that each of them can be interpreted as a sequence containing the segments with some quasiperiodic (because of the constraints (2)) repetitions. If we define the boundaries of the series on the first or the last repetition, then one can interpret all of the above problems as problems of partitioning a sequence

into segments with quasiperiodic repetitions of a priori unknown points, estimating these points, and finding their positions in the sequence.

The next statement establishes the complexity status of Problem 1.

Proposition 1. *The Problem 1 is strongly NP-hard.*

Proposition 1 follows from the fact that the special case of Problem 1 with $L = 1$ is strongly NP-hard [5]. Thus, Problem 1 belongs to the class of computationally intractable problems.

3 Known and Obtained Results

Problem 1 is among the poorly studied discrete optimization problems. It is closely related to the problem (see [7]) in which the input sequence $\mathcal{Y}$ is one-dimensional, i.e. $q = 1$. The points from tuple $(x_1, \ldots, x_L)$ belong to $\mathbb{R}^d$, where $d \geq 1$, and they are given at the problem input, at that $T_{\min} \geq d$ in the restrictions (2). In the objective function of the problem instead of the centroids $\overline{y}(\mathcal{M}_1), \ldots, \overline{y}(\mathcal{M}_L)$ of the desired subsets appear the elements from the given tuple $(x_1, \ldots, x_L)$. The unknown variables are the sets $\mathcal{M}_1, \ldots, \mathcal{M}_L$. This problem can be interpreted as a problem searching a sequence for non-overlapping segments with quasiperiodic repetitions of points from the tuple together with the positions of these points in the sequence. It was shown in [7] that this problem is solvable in polynomial time using dynamic programming. Below we apply a simplification of this dynamic program in our algorithm.

Except for the special case with $L = 1$ in Eq. (1), no algorithms with guaranteed approximation factor are known at the moment for Problem 1. For this special case, the following results were obtained.

In [5], the variant of Problem 1 in which $T_{\min}$ and $T_{\max}$ are the parameters was analyzed. In the cited work it was shown that in the case when $L = 1$, this parameterized variant is strongly NP-hard for any $T_{\min} < T_{\max}$. In the trivial case when $T_{\min} = T_{\max}$, the problem is solvable in polynomial time.

In [6], for the same case of Problem 1, when $L = 1$, a 2-approximation polynomial-time algorithm running in $\mathcal{O}(N^2(MN + q))$ time was presented.

In addition, in [8,9], two special cases of the case $L = 1$ were studied. In both subcases the dimension q of the space is fixed. For the subcase with integer inputs in [8] an exact pseudopolynomial algorithm was constructed. The time complexity of this algorithm is $\mathcal{O}(MN^2(MD)^q)$, where D is the maximum absolute in any coordinate of the input points. For the subcase with real inputs in [9] a fully polynomial-time approximation scheme was proposed which, given a relative error ε, finds a $(1 + \varepsilon)$-approximate solution of Problem 1 in $\mathcal{O}(MN^3(1/\varepsilon)^{q/2})$ time.

The main result of this paper is an algorithm that allows to find a 2-approximate solution of Problem 1 in $\mathcal{O}(LN^{L+1}(MN + q))$ time, which is polynomial if the number L of clusters is fixed.

4 Fundamentals of Algorithm

To construct the algorithm we need a few basic assertions, an auxiliary problem and an exact polynomial algorithm for its solution.

The geometrical foundations of the algorithm are given by the following lemmas.

Lemma 1. *For any point $u \in \mathbb{R}^q$ and any finite nonempty set $\mathcal{Z} \subset \mathbb{R}^q$ the following equality holds*

$$\sum_{z \in \mathcal{Z}} \|z - u\|^2 = \sum_{z \in \mathcal{Z}} \|z - \overline{z}\|^2 + |\mathcal{Z}| \cdot \|u - \overline{z}\|^2 , \tag{6}$$

where $\overline{z} = \frac{1}{|\mathcal{Z}|} \sum_{z \in \mathcal{Z}} z$ is the centroid of $\mathcal{Z}$.

Lemma 1 has quite simple proof and is well-known. Its proof has been given in several publications (for example, in [10]).

Lemma 2. *Assume that the conditions of Lemma 1 hold. Then, if some point $u \in \mathbb{R}^q$ is closer (with respect to the Euclidean distance) to the centroid $\overline{z}$ of $\mathcal{Z}$ than all points in $\mathcal{Z}$, then*

$$\sum_{z \in \mathcal{Z}} \|z - u\|^2 \leq 2 \sum_{z \in \mathcal{Z}} \|z - \overline{z}\|^2 .$$

Lemma 2 follows from (6), because by the assumption for every point $z \in \mathcal{Z}$ we have the inequality $\|u - \overline{z}\| \leq \|z - \overline{z}\|$.

From now on we use $f^x(y)$ to denote a function $f(x, y)$ for which x is fixed.

Lemma 3. *Let*

$$S(\mathcal{M}_1, \ldots, \mathcal{M}_L, x_1, \ldots, x_L) = \sum_{l=1}^{L} \sum_{j \in \mathcal{M}_l} \|y_j - x_l\|^2 + \sum_{i \in \mathcal{N} \setminus \mathcal{M}} \|y_i\|^2 , \tag{7}$$

$$G(\mathcal{M}_1, \ldots, \mathcal{M}_L, x_1, \ldots, x_L) = \sum_{l=1}^{L} \sum_{j \in \mathcal{M}_l} (2\langle y_j, x_l \rangle - \|x_l\|^2) ,$$

where $x_1, \ldots, x_L$ are points from $\mathbb{R}^q$, and elements of the sets $\mathcal{M}_l, \ldots, \mathcal{M}_L$, and $\mathcal{M}$ satisfy restrictions of Problem 1. Then the following statements are true:

(1) for any nonempty fixed subsets $\mathcal{M}_1, \ldots, \mathcal{M}_L$ the minimum of function (7) over $x_1, \ldots, x_L$ is reached at the points $x_l = \overline{y}(\mathcal{M}_l)$, $l = 1, \ldots, L$, and is equal to $F(\mathcal{M}_1, \ldots, \mathcal{M}_L)$;

(2) for any tuple $x = (x_1, \ldots, x_L)$ of fixed points from $\mathbb{R}^q$ the minimum of function $S^x(\mathcal{M}_1, \ldots, \mathcal{M}_L)$ over $\mathcal{M}_1, \ldots, \mathcal{M}_L$ is reached at the subsets $\mathcal{M}_1^x, \ldots, \mathcal{M}_L^x$ that maximize function $G^x(\mathcal{M}_1, \ldots, \mathcal{M}_L)$.

Proof. The first statement of this lemma is easily verified by differentiation and also follows from Lemma 1. To prove the second statement it is sufficient to note that the following equality holds

$$S^x(\mathcal{M}_1, \ldots, \mathcal{M}_L) = \sum_{j \in \mathcal{N}} \|y_j\|^2 - G^x(\mathcal{M}_1, \ldots, \mathcal{M}_L) \,, \tag{8}$$

where the sum on the right-hand side is independent of $\mathcal{M}_1, \ldots, \mathcal{M}_L$. □

The main ingredient to our algorithm is an exact polynomial-time algorithm for solving the following auxiliary problem.

Problem 2. Given a sequence $\mathcal{Y} = (y_1, \ldots, y_N)$ and a tuple $x = (x_1, \ldots, x_L)$ of points from $\mathbb{R}^q$, and some positive integers $T_{\min}$, $T_{\max}$, and M. *Find* nonempty disjoint subsets $\mathcal{M}_1, \ldots, \mathcal{M}_L$ of $\mathcal{N} = \{1, \ldots, N\}$ that maximize the objective function $G^x(\mathcal{M}_1, \ldots, \mathcal{M}_L)$, under the same constraints on the optimized variables as in Problem 1.

To explain the algorithm for solving this auxiliary problem, we define the function

$$g_l^x(n) = 2\langle y_n, x_l \rangle - \|x_l\|^2, \quad n \in \mathcal{N}, \quad l = 1, \ldots, L \,, \tag{9}$$

where x_l is a point from tuple x, and y_n is an element of sequence $\mathcal{Y}$.

In accordance with the definition (9), for the objective function $G^x(\mathcal{M}_1, \ldots, \mathcal{M}_L)$ we have

$$G^x(\mathcal{M}_1, \ldots, \mathcal{M}_L) = \sum_{l=1}^{L} \sum_{n \in \mathcal{M}_l} g_l^x(n) \,.$$

In addition, we note that Lemma 3 yields the following equalities

$$\begin{aligned} (\mathcal{M}_1^x, \ldots, \mathcal{M}_L^x) &= \arg \min_{\mathcal{M}_1, \ldots, \mathcal{M}_L} S^x(\mathcal{M}_1, \ldots, \mathcal{M}_L) \\ &= \arg \max_{\mathcal{M}_1, \ldots, \mathcal{M}_L} G^x(\mathcal{M}_1, \ldots, \mathcal{M}_L) \,. \end{aligned} \tag{10}$$

In the next lemma and its corollary we give a dynamic programming scheme. This scheme guarantees finding the optimal solution $\mathcal{M}_1^x, \ldots, \mathcal{M}_L^x$ of Problem 2 and (according to the Eq. (10)) the optimal solution of the problem of minimizing the function $S^x(\mathcal{M}_1, \ldots, \mathcal{M}_L)$. The presented scheme follows from the results obtained in [7] and is given here for completeness.

Lemma 4. *Let the conditions of Problem 2 hold. Then for any positive integers L and M such that $(M-1)T_{\min} < N$ and $L \le M$, the optimal value $G_{\max}^x$ of the objective function of Problem 2 is given by the formula*

$$G_{\max}^x = \max_{n \in \{1+(M-1)T_{\min}, \ldots, N\}} G_{L,M}^x(n) \,; \tag{11}$$

here, the values of $G^x_{L,M}(n)$ *are calculated using the recurrence formula*

$$G^x_{l,m}(n) = g^x_l(n) + \begin{cases} 0, & \text{if } l = 1, m = 1, \\ \max\limits_{j\in\gamma_{m-1}(n)} G^x_{1,m-1}(j), & \text{if } l = 1, m = 2, \ldots, M-(L-1), \\ \max\limits_{j\in\gamma_{m-1}(n)} G^x_{l-1,m-1}(j), & \text{if } l = 2, \ldots, L,\ m = l, \\ \max\{\max\limits_{j\in\gamma_{m-1}(n)} G^x_{l,m-1}(j), \max\limits_{j\in\gamma_{m-1}(n)} G^x_{l-1,m-1}(j)\}, & \text{if } l = 2, \ldots, L,\ m = l+1, \ldots, M-(L-l), \end{cases} \tag{12}$$

where

$$\gamma_{m-1}(n) = \{j \mid \max\{1+(m-2)T_{\min}, n-T_{\max}\} \le j \le n - T_{\min}\}, \quad m = 2, \ldots, M, \tag{13}$$

for every $n = 1+(m-1)T_{\min}, \ldots, N-(M-m)T_{\min}$.

Corollary 1. *Let the conditions of Lemma 4 hold. In addition, let*

$$r^x_{l,m}(n) = \begin{cases} 1, & \text{if } l = 1,\ m = 2, \ldots, M-(L-1), \\ l-1, & \text{if } l = 2, \ldots, L,\ m = l, \\ l-1, & \text{if } \max\limits_{j\in\gamma_{m-1}(n)} G^x_{l,m-1}(j) < \max\limits_{j\in\gamma_{m-1}(n)} G^x_{l-1,m-1}(j), \\ & \quad l = 2, \ldots, L,\ m = l+1, \ldots, M-(L-l), \\ l, & \text{if } \max\limits_{j\in\gamma_{m-1}(n)} G^x_{l,m-1}(j) \ge \max\limits_{j\in\gamma_{m-1}(n)} G^x_{l-1,m-1}(j), \\ & \quad l = 2, \ldots, L,\ m = l+1, \ldots, M-(L-l), \end{cases}$$

$$I^x_{l,m}(n) = \arg\max_{j\in\gamma_{m-1}(n)} G^x_{l,m-1}(j),\ \ l = 1, \ldots, L,\ \ m = l+1, \ldots, M-(L-l),$$

for every $n = 1+(m-1)T_{\min}, \ldots, N-(M-m)T_{\min}$*;*

$$n^x(m) = \begin{cases} \arg\max\limits_{n\in\{1+(M-1)T_{\min}, \ldots, N\}} G^x_{L,M}(n), & \text{if } m = M, \\ I^x_{k^x(m),m+1}(n^x(m+1)), & \text{if } m = M-1, \ldots, 1, \end{cases}$$

$$k^x(m) = \begin{cases} L, & \text{if } m = M, \\ r^x_{k^x(m+1),m+1}(n^x(m+1)), & \text{if } m = M-1, \ldots, 1; \end{cases}$$

$$J^x(l) = \begin{cases} 0, & \text{if } l = 0, \\ \left|\left\{m \in \{1, \ldots, M\} \mid k^x(m) \le l\right\}\right|, & \text{if } l = 1, \ldots, L. \end{cases}$$

Then the sets $\mathcal{M}^x_1, \ldots, \mathcal{M}^x_L$ *are given by the formula*

$$\mathcal{M}^x_l = \left\{n \mid n = n^x(m),\ m = J^x(l-1)+1, \ldots, J^x(l)\right\} \tag{14}$$

for every $l = 1, \ldots, L$.

A step-by-step description of the algorithm implementing the above scheme is given in the following.

Algorithm $\mathcal{A}_1$.

Input: sequence $\mathcal{Y}$, tuple $(x_1, \ldots, x_L)$ of points, numbers $T_{\min}$, $T_{\max}$, and M.

`Step 1.` Compute the values $g_l^x(n)$ for $l = 1, \ldots, L$, and $n = 1 + (l-1)T_{\min}, \ldots, N - (L-l)T_{\min}$ using Formula (9).

`Step 2.` Using Formulae (12) and (13), compute the values $G_{l,m}^x(n)$ for each $l = 1, \ldots, L$, $m = l, \ldots, M - (L-l)$, $n = 1 + (m-1)T_{\min}, \ldots, N - (M-m)T_{\min}$.

`Step 3.` Find the maximum $G_{\max}^x$ of the objective function G^x by Formula (11), and the optimal subsets $\mathcal{M}_l^x$ by Formula (14).

Output: the family $\{\mathcal{M}_1^x, \ldots, \mathcal{M}_L^x\}$ of subsets.

Remark 1. Before the start of the algorithm it is required to verify the two conditions of Lemma 4. These necessary conditions provide the consistency of the constraints in Problems 1 and 2, as well as the correctness of the input data of the algorithm.

Remark 2. In [7], it was found that Algorithm $\mathcal{A}_1$ finds the optimal solution of Problem 2 in $\mathcal{O}(LN(M(T_{\max} - T_{\min} + 1) + q))$ time. In this expression, the value of $T_{\max} - T_{\min} + 1$ is at most N. Therefore, the algorithm running time can be estimated as $\mathcal{O}(LN(MN + q))$.

5 Approximation Algorithm

Our approach to Problem 1 is as follows. For each ordered set (tuple) containing L elements of the sequence $\mathcal{Y}$, we find an exact solution of the auxiliary Problem 2, i.e. a family containing disjoint subsets of indices of the input sequence, which is a feasible solution of the original Problem 1.

The found family of subsets we declare a solution candidate for Problem 1 and include this family in the set of solution candidates.

From the obtained set as the final solution we choose a family of subsets which yields the largest value for the objective function of Problem 2.

Let us formulate an algorithm that implements the described approach. Below, in the step-by-step description, it is assumed that the input positive integers satisfy the conditions of Lemma 4 (see Remark 1).

Algorithm $\mathcal{A}$.

Input: sequence $\mathcal{Y}$, numbers $T_{\min}$, $T_{\max}$, M, and L.

`Step 1.` For every tuple $x = (x_1, \ldots, x_L) \in \mathcal{Y}^L$ of elements of the sequence $\mathcal{Y}$, using Algorithm $\mathcal{A}_1$, find the optimal solution $\{\mathcal{M}_1^x, \ldots, \mathcal{M}_L^x\}$ of Problem 2.

`Step 2.` Find a tuple $x(A) = \arg\max_{x \in \mathcal{Y}^L} G^x(\mathcal{M}_1^x, \ldots, \mathcal{M}_L^x)$ and a family $\{\mathcal{M}_1^A, \ldots, \mathcal{M}_L^A\} = \{\mathcal{M}_1^{x(A)}, \ldots, \mathcal{M}_L^{x(A)}\}$. If the optimum is taken by several tuples, we choose any of them.

Output: the family $\{\mathcal{M}_1^A, \ldots, \mathcal{M}_L^A\}$ of subsets.

Lemma 5. *Let* $\{\mathcal{M}_1^*, \ldots, \mathcal{M}_L^*\}$ *be the optimal solution of Problem 1, and* $\{\mathcal{M}_1^A, \ldots, \mathcal{M}_L^A\}$ *be the solution found by Algorithm* $\mathcal{A}$. *Then*

$$F(\mathcal{M}_1^A, \ldots, \mathcal{M}_L^A) \le 2F(\mathcal{M}_1^*, \ldots, \mathcal{M}_L^*) .$$

Proof. The optimal solution $\{\mathcal{M}_1^*, \ldots, \mathcal{M}_L^*\}$ of Problem 1 corresponds to the tuple $(\overline{y}(\mathcal{M}_1^*), \ldots, \overline{y}(\mathcal{M}_L^*))$ of centroids, where $\overline{y}(\mathcal{M}_l^*) = \frac{1}{|\overline{y}(\mathcal{M}_l^*)|} \sum_{y \in \mathcal{M}_l^*} y$, $l = 1, \ldots, L$. Let us consider the point $t_l = \arg \min_{y \in \mathcal{M}_l^*} \|y - \overline{y}(\mathcal{M}_l^*)\|$, $l = 1, \ldots, L$, from the subset $\mathcal{M}_l^*$, closest to the centroid of this subset. This point in the set $\mathcal{M}_l^*$ and the set $\mathcal{M}_l^*$ itself satisfy the conditions of Lemma 2. Therefore, by applying the inequality of Lemma 2 to every subset $\mathcal{M}_l^*$, $l = 1, \ldots, L$, we can estimate the sum

$$\begin{aligned} S(\mathcal{M}_1^*, \ldots, \mathcal{M}_L^*, t_1, \ldots, t_L) &= \sum_{l=1}^{L} \sum_{y \in \mathcal{M}_l^*} \|y - t_l\|^2 + \sum_{i \in \mathcal{N} \setminus \mathcal{M}^*} \|y_i\|^2 \\ &\leq 2 \sum_{l=1}^{L} \sum_{y \in \mathcal{M}_l^*} \|y - \overline{y}(\mathcal{M}_l^*)\|^2 + \sum_{i \in \mathcal{N} \setminus \mathcal{M}^*} \|y_i\|^2 \\ &\leq 2 \sum_{l=1}^{L} \sum_{y \in \mathcal{M}_l^*} \|y - \overline{y}(\mathcal{M}_l^*)\|^2 + 2 \sum_{i \in \mathcal{N} \setminus \mathcal{M}^*} \|y_i\|^2 = 2F(\mathcal{M}_1^*, \ldots, \mathcal{M}_L^*) \,, \end{aligned} \tag{15}$$

where $\mathcal{M}^* = \cup_{l=1}^{L} \mathcal{M}_l^*$.

On the other hand, we notice that the tuple $t = (t_1, \ldots, t_L)$ is among the tuples from $\mathcal{Y}^L$ that have been examined at Step 1 of Algorithm $\mathcal{A}$. Let $\{\mathcal{M}_1^t, \ldots, \mathcal{M}_L^t\}$ be the optimal solution found at Step 2 of Algorithm $\mathcal{A}$ for Problem 2 at $x = t$. Then according to statement 2 of Lemma 3, i.e. according to (10), the family $\{\mathcal{M}_1^t, \ldots, \mathcal{M}_L^t\}$ supplies the minima to the function $S^x(\mathcal{M}_1, \ldots, \mathcal{M}_L)$ at $x = t$. Consequently the bound

$$S(\mathcal{M}_1^t, \ldots, \mathcal{M}_L^t, t_1, \ldots, t_L) \leq S(\mathcal{M}_1^*, \ldots, \mathcal{M}_L^*, t_1, \ldots, t_L) \tag{16}$$

is valid for the left-hand side of (15).

Furthermore, by the definition of Step 2 and according to (8) we have the bound

$$S(\mathcal{M}_1^A, \ldots, \mathcal{M}_L^A, x_1^A, \ldots, x_L^A) \leq S(\mathcal{M}_1^t, \ldots, \mathcal{M}_L^t, t_1, \ldots, t_L) \,, \tag{17}$$

where $(x_1^A, \ldots, x_L^A) = x(A)$. Additionally, from the first statement of Lemma 3 we have the inequality

$$F(\mathcal{M}_1^A, \ldots, \mathcal{M}_L^A) \leq S(\mathcal{M}_1^A, \ldots, \mathcal{M}_L^A, x_1^A, \ldots, x_L^A) \,. \tag{18}$$

Finally, by combining (15)–(18) we get the chain of estimation inequalities

$$\begin{aligned} F(\mathcal{M}_1^A, \ldots, \mathcal{M}_L^A) &\leq S(\mathcal{M}_1^A, \ldots, \mathcal{M}_L^A, x_1^A, \ldots, x_L^A) \\ &\leq S(\mathcal{M}_1^t, \ldots, \mathcal{M}_L^t, t_1, \ldots, t_L) \leq S(\mathcal{M}_1^*, \ldots, \mathcal{M}_L^*, t_1, \ldots, t_L) \\ &\leq 2F(\mathcal{M}_1^*, \ldots, \mathcal{M}_L^*) \,, \end{aligned}$$

which proves Lemma 5. □

We finally prove the running time of the algorithm and that the bound of 2 on its approximation factor is tight.

Theorem 1. *Algorithm $\mathcal{A}$ finds a 2-approximate solution of Problem 1 in $\mathcal{O}(LN^{L+1}(M(T_{\max} - T_{\min} + 1) + q))$ time. The performance guarantee 2 of the algorithm is tight.*

Proof. The 2-accuracy bound of the algorithm follows from Lemma 5. We bound the time complexity of the algorithm using its step-by-step description.

The computation time is determined by the time complexity of Step 1, at which Problem 2 is solved $\mathcal{O}(N^L)$ times by applying Algorithm $\mathcal{A}_1$, whose time complexity is $\mathcal{O}(LN(M(T_{\max} - T_{\min} + 1) + q))$ (see Remark 2). In addition, it needs $\mathcal{O}(N^L)$ comparisons for searching a largest value of the objective function of Problem 2 at Step 2. By summing all these times we obtain the final bound for the algorithm time complexity.

The tightness of the performance guarantee of Algorithm $\mathcal{A}$ follows from the tightness of the performance guarantee of the 2-approximation algorithm for the case of Problem 1 when $L = 1$ (see [6]). □

Remark 3. According to Remark 2, the running time of Algorithm $\mathcal{A}$ is $\mathcal{O}(LN^{L+1}(MN + q))$, which is polynomial if the number L of clusters is fixed.

6 Conclusion

In this paper we have shown the strong NP-hardness of one problem of partitioning a finite sequence of points of Euclidean space into clusters with restrictions on their cardinalities. We also have shown an approximation algorithm for this problem. The proposed algorithm allows to find a 2-approximate solution of the problem in a polynomial time if the number of clusters is fixed.

In our opinion, the presented algorithm would be useful as one of the tools for solving problems in applications related to data mining, and analysis and recognition of time series (signals).

Of considerable interest is the development of faster polynomial-time approximation algorithms for the case when the number of clusters is not fixed. An important direction of study is searching subclasses of this problem for which faster polynomial-time approximation algorithms can be constructed.

Acknowledgments. This work was supported by Russian Science Foundation, project no. 16-11-10041.

References

1. Fu, T.: A review on time series data mining. Eng. Appl. Artif. Intell. **24**(1), 164–181 (2011)
2. Kuenzer, C., Dech, S., Wagner, W.: Remote Sensing Time Series. Remote Sensing and Digital Image Processing, vol. 22. Springer, Switzerland (2015)
3. Warren Liao, T.: Clustering of time series data – a survey. Pattern Recogn. **38**(11), 1857–1874 (2005)
4. Aggarwal, C.C.: Data Mining: The Textbook. Springer, Switzerland (2015)

5. Kel'manov, A.V., Pyatkin, A.V.: On complexity of some problems of cluster analysis of vector sequences. J. Appl. Ind. Math. **7**(3), 363–369 (2013)
6. Kel'manov, A.V., Khamidullin, S.A.: An approximating polynomial algorithm for a sequence partitioning problem. J. Appl. Ind. Math. **8**(2), 236–244 (2014)
7. Kel'manov, A.V., Mikhailova, L.V.: Joint detection of a given number of reference fragments in a quasi-periodic sequence and its partition into segments containing series of identical fragments. Comput. Math. Math. Phys. **46**(1), 165–181 (2006)
8. Kel'manov, A.V., Khamidullin, S.A., Khandeev, V.I.: An exact pseudopolynomial algorithm for a sequence bi-clustering problem (in Russian). In: Book of Abstract of the XVth Russian Conference "Mathematical Programming and Applications", pp. 139–140. Inst. Mat. Mekh. UrO RAN, Ekaterinburg (2015)
9. Kel'manov, A.V., Khamidullin, S.A., Khandeev, V.I.: A fully polynomial-time approximation scheme for a sequence 2-cluster partitioning problem. J. Appl. Indust. Math. **10**(2), 209–219 (2016)
10. Kel'manov, A.V., Romanchenko, S.M.: An FPTAS for a vector subset search problem. J. Appl. Indust. Math. **8**(3), 329–336 (2014)